Characterizing cell state and cell fate by high-throughput single cell genomics

Junyue Cao

A dissertation

Submitted in partial fulfillment of the

Requirement for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Jay Shendure, Chair

Cole Trapnell

Robert Waterston

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Characterizing cell state and cell fate by high-throughput single cell genomics

Junyue Cao

Chair of the Supervisory Committee:

Jay Shendure

Professor, Department of Genome Sciences, University of Washington

Howard Hughes Medical Institute

Animal development is one of the greatest sources of wonders in science. Development of multicellular organism is characterized by the differentiation of a fertilized egg into diverse cell types of the body in a programed temporal spatial order. The process of development includes fertilization, cleavage, gastrulation, organogenesis, metamorphosis, regeneration, and senescence. Characterizing cell differentiation in each step, by resolving the cell state diversity and cell fate dynamics, is the key to fully understanding developmental process. The expression levels of mRNA species are readily linked to cellular function, and therefore profiling the transcriptome of individual cells has emerged as a powerful strategy for resolving cell state heterogeneity. However, current methods for single cell RNA sequencing all rely on the isolation of individual cells within physical compartments and thus have problems such as low throughput, high cost, and information lost from other molecular layers. During the three and

half years of my graduate study, I developed four novel high-throughput single cell genomic techniques to get over these limitations and applied them to profiling cell state heterogeneity and dynamics in development at single cell resolution.

To resolve cellular state heterogeneity, I developed a combinatorial indexing strategy to profile the transcriptome across tens of thousands of single cells (sci-RNA-seq: Single cell Combinatorial Indexing RNA sequencing), and applied sci-RNA-seq to generate **the first catalog of single cell transcriptomes at the scale of whole organism *Caenorhabditis elegans*** (Cao. J., Jonathan. P., et al, Comprehensive single-cell transcriptional profiling of a multicellular organism, **Science**, 2017). I profiled over 50,000 cells from the nematode *C. elegans* at the L2 stage, which is over 50-fold "shotgun cellular coverage" of its somatic cell composition. This is the first study to show that single cell transcriptome alone is sufficient to separate all major cell types from whole animal. Cell type specific genes for 27 distinct cell types are identified, including for some fine-grained cell types that are present in only one or two cells per individual. Given that *C.elegans* is the only organism with a fully mapped cellular lineage, these data represent a rich resource for future research aimed at defining cell types and states. The dataset will advance our understanding of developmental biology, and constitute a major step towards a comprehensive, single-cell molecular atlas of a whole animal.

To further characterize cellular state across multiple molecular layers, I developed sci-CAR, the first high throughput single cell genomic approach that can jointly profile epigenome (chromatin accessibility) and transcriptome in each of 1000s of single cells (Cao. J. et al, Joint profiling of chromatin accessibility and gene expression in thousands of single cells, **Science**, 2018). I applied sci-CAR to 11,233 cells from whole mouse kidney and linked cis-regulatory sites to their

putative target genes based on the covariance of chromatin accessibility and transcription at the single-cell level. To the best of our knowledge, **this represents the first joint profiling of the epigenome and transcriptome in individual cells at the scale and complexity of a whole mammalian organ.**

One critical challenge in development is to characterize the cell differentiation path for all major cell types forming our body. During mammalian organogenesis, the cells of the three germ layers transform into an embryo that includes most major internal and external organs. The key regulators of developmental defects can be studied during this critical window, but conventional approaches lack the throughput and resolution to obtain a global view of the molecular states and trajectories of a rapidly diversifying and expanding number of cell types. To investigate cell state dynamics in this critical window, I developed another single cell transcriptome profiling technique (sci-RNA-seq3), the first single cell RNA-seq technique capable of profiling millions of single cells in a single experiment, with over one hundred times higher throughput and lower cost compared with conventional approaches. I applied sci-RNA-seq3 to profiling ~ 2 million cells derived from 61 mouse embryos staged between 9.5 and 13.5 days of gestation (Cao. J., Spielmann. M., et al, The single-cell transcriptional landscape of mammalian organogenesis, **Nature**, 2018). **This is by far the most comprehensive cell atlas of mammalian development as well as the largest single cell RNA-seq data set in the world.** By unsupervised clustering analysis, I characterized hundreds of expanding, contracting and transient cell types, many of which are only detected because of the depth of cellular coverage obtained here, and defined the corresponding sets of cell type-specific marker genes, several of which are validated by whole mount *in situ* hybridization. With a new single cell RNA-seq analysis package Monocle 3, I further delineated and annotated 56 single cell developmental trajectories of mouse

organogenesis, spanning all major systems such as central nervous system and reproductive system. The dynamics of cell proliferation and key gene regulators within each cell lineage are further identified. These data comprise a foundational resource for single cell genomic field and mammalian developmental biology.

To further characterizing the mechanism regulating cell state dynamics, I developed sci-fate, the first strategy to recover whole transcriptome temporal dynamics across thousands of single cells (Cao. J., et al, Characterizing single cell temporal dynamics with sci-fate, manuscript in preparation, 2019). I applied sci-fate to a model system of cortisol response and developed a computation strategy to identify key driving transcription factors regulating cell state changes. Based on the data, I built a cell state transition network for future cell state prediction, and illustrate key factors regulating cell state transition dynamics. **This is the first study to quantitatively characterize cell state dynamics at whole transcriptome level** and constitutes a major step to fully understanding mechanisms in cell fate determination.

Acknowledgements

My graduate school career would not have been possible without the contributions and support from so many people. First, I would like to thank my Ph. D. mentor, Jay Shendure, for being an amazing and outstanding mentor for the past four years. His enthusiasm for science, depth of knowledge, superb working efficiency, and most of all, incredible creativity, have been my major source of inspiration through my whole graduate school study and will continue to be so hereafter. Most importantly, his encouragement has given me the faith and confidence to pursue science as my future career, and develop novel techniques to tackle biological mysteries for the benefits of humanity.

I would like to thank the scientific mentors with whom I have been fortunate to work through my young scientific career. Luhua Lai, my undergraduate mentor in molecular prediction and design, taught me the wonders of asking and answering my own scientific questions, and led me to the amazing field of scientific research. Chengkai Dai, my scientific mentor in Jackson Laboratory, trained me with diverse biological experiment skills, as well as the passion for science and persistence in doing research and pushing project forward. Michael Elowitz, Long Cai, and Raymond Deshaies, my mentors of graduate rotations in Caltech, taught me to view biological questions from different perspectives, and nurtured my interest in single cell science field and system biology.

I would also like to thank Cole Trapnell, my mentor in computation analysis during graduate study, for his incredibly support and encouragement with me as I struggled to develop computation skills in the pursuit of novel scientific discoveries. I would also like to thank Robert

Waterston and David Beier, for their guidance in my graduate career and awakened my passion for developmental biology. I would like to thank my committee members, Celeste Berg and Daniel Promislow, for their generous support through my graduate research.

I am fortunate to join the Shendure lab and thanks to all members in our lab, for the guidance and support in my graduate school study, and most importantly their friendships. In particular, Darren Cusanovich and Vijay Ramani were my mentors during my rotations in UW and provided invaluable instructions from experiment design to technique troubleshooting. I am also grateful to have enormous technique and computation support from all members of the Shendure lab, especially from Malte Spielmann, Choli Lee, Riza Daza, Hannah Pliner, Andrew Hill and Xingfan Huang and Ruolan Qiu. I am also grateful to the support from all my collaborators, especially from Jonathan Packer, Delasa Aghamirzaie from Trapnell lab, Chau Huynh from Waterston lab, Frank Steemers, Fan Zhang, Lena Christiansen from illumina, and Andrew Adey from Adey lab. I would also like to thank the MCB program for offering me the chance for transferring to UW from Caltech to continue my graduate study, and incredible support during the last four years.

At last, I want to thank my family. My parents, Suzhen Liu and Genping Cao, are the constant source of support in my life and career. They sacrificed so much for my career and their love has been invaluable to me. My dear wife, my love, Wei Zhou, supports and shares every breath, and every heat-beating of my life. As a scientist herself, Wei contributes to the synthesis and maturation of every novel idea and every exciting moment in my science career and everyday life. I also want to thank to my brilliant son, Jayden Cao and my lovely new-born daughter, Sonia Cao, for motivating me to make full use of every minutes of my life and devote myself

into the science career. You all are my direct driving force to tackle existing scientific mysteries, my deepest motivation to solve all diseases and aging, and make our future world a better place.

TABLE OF CONTENTS

Introduction 12

Chapter 1. comprehensive characterization of cell states by combinatorial indexing based single cell RNA sequencing 29

Abstract 30

Introduction 30

Results and discussion 31

Supplementary methods and materials 55

Reference 105

Chapter 2. cell state characterization by single cell chromatin accessibility and transcriptome co-assay 122

Abstract 123

Introduction 123

Results and discussion 123

Supplementary methods and materials 139

Reference 180

Chapter 3. cell fate characterization of mammalian organogenesis 186

Abstract 187

Introduction 187

Results and discussion 188

Supplementary methods and materials 209

Reference 259
Chapter 4. characterize cell state dynamics by sci-fate 269
Abstract 269
Introduction 269
Results and discussion 271
Supplementary methods and materials 289
Reference 308
Chapter 5. Conclusions and perspectives 315

Introduction

"As an embryo, you had to build yourself from a single cell. You had to respire before you had lungs, digest before you had a gut, build bones when you were pulpy, and form orderly arrays of neurons before you knew how to think. One of the critical differences between you and a machine is that a machine is never required to function until after it is built.(*1*)"

Animal development is one of the greatest sources of wonders in science. Development of multicellular development is characterized by the differentiation of a fertilized egg into diverse cell types of the body in a programed temporal spatial order. The process of development include fertilization, cleavage, gastrulation, organogenesis, metamorphosis, regeneration, and senescence (*1*). Characterizing cell differentiation in each step, by resolving the cell state diversity and cell fates via which hundreds of different cell types are generated, is the key to fully solve the mysteries of development.

Characterizing cell state diversity by single cell genomics

Since the first discovery of cell more than three hundred years ago(*2*), biologist have sought to characterize and classify cell states in development, driven by the advances of technology. The invention of light microscope techniques enables cell state separation by morphology and the classification of distinct neuron types from the brain in 1887(*3*). The advent of fluorescent microscopy(*4*) and fluorescence-activated cell sorting techniques(*5*), together with the advances in fluorescent labeling techniques for protein and nucleic acids, remarkably improves our ability to separate distinct cell types based on selected molecular features. These efforts altogether

reveal an impressive cell state diversity in development, especially in the immune and neuron system.

Despite of these achievement, however, our understanding of cell diversity is still limited. Each cell can be characterized by the state of millions of chromatin accessible sites, tens of thousands of genes, and even more dynamically regulated proteome and metabolome. Characterizing cell states by morphology only or several ad hoc protein markers can be highly biased, depending on the pre-selected panel of gene features. In addition, cell state separation by different criteria (morphology or functions) can give non-related or even controversial results.

The key to resolve the above limitations lies in the advent of genomic profiling techniques. Due to the invention of the first next generation sequencing platform in 2005(*6–9*), together with molecular techniques to convert diverse cellular features into DNA sequence(*10–14*), genomic sequencing has been widely applied to comprehensively profiling the bulk biological information flow along central dogma, including genomic DNA, epigenome, transcriptome and proteome. Historically, these bulk -omics techniques require input with hundreds to millions of cells, thus only capture ensemble averages of cell states. This gap is quickly filled by the development of single cell genomic techniques, allowing high-throughput and less biased molecular surveys at individual cell level.

The first single cell genomic approach for deciphering cellular heterogeneity is single cell RNA-seq demonstrated in 2009 (*15*), shortly after the first application of RNA-seq to bulk samples in 2006(*16*, *17*). Initial single cell RNA-seq techniques mostly isolate single cells by manual picking, or by flow-sorting to microwell plate(*18*, *19*), followed by bulk RNA-seq procedures

including first strand cDNA generation by reverse transcription, double strand DNA synthesis, fragmentation and PCR amplification. Even with robotic systems to automate the process, these single cell RNA-seq approaches by isolation of individual cells within physical compartments are normally limited in throughput (100s of cells per experiment) and high cost. During the following several years, multiple droplet-based scRNA-seq techniques are developed to further improve the throughput and reduce the cost(*20–22*), by capturing individuals cells in nanolitre-sized droplets loaded with reagents for indexed reverse transcription, followed by pooling thousands of droplets containing cells for further amplification and sequencing.

The developments of high-throughput single cell RNA-seq techniques lead to the generation of vast amounts of data. Meanwhile, many computation methods and tools have been developed to apply the data in answering biological questions(*23*). Generally, there are four major steps for processing single cell RNA-seq data sets: **1) Gene quantification**: reads are aligned to the genome or transcriptome and single cell gene expressions are quantified, similarly with bulk RNA-seq. Cells with poor quality metrics (e.g. high proportion of reads mapping to mitochondrial genes or low number of detected genes) are filtered out. 2) **Cell level analysis**: dimensionality reduction techniques such as Principal Component Analysis (PCA)(*23*, *24*), Multidimensional Scaling (MDS)(*25*) and t-Distributed Stochastic Neighbor Embedding (t-SNE)(*26*) are used to visualize the data. Different cell groups are identified based on clustering techniques such as K-medoids method(*27*) or local density (*27*, *28*). Alternatively, trajectory analysis is also used as dimension reduction techniques to analyze cell differentiation process. For example, Monocle, the first package for trajectory analysis, infer the developmental pseudotime and branches in cell differentiation, and order cells by the distance to the start cell based on tree-embedding strategy(*29*, *30*). **3)Gene level analysis**: differentially expressed genes

across different clusters or pseudotime can be identified by significant testing, such as fitting a generalized linear model (GLM) for each gene followed by Wald (*31*) or likelihood ratio test(*32*). Furthermore, gene regulatory networks can be inferred based on gene-to-gene expression correlations by graphical LASSO (GLASSO) (*33*). 4) **Multi-omics analysis**: combined with other -omics data set, single cell RNA-seq can be applied to tackle novel questions: combining cell type specific markers recovered from single cell RNA-seq and immunohistochemistry (IHC) database (e.g. The Human Protein Atlas (*34*)) enable us to pinpoint the location of different cell types in tissues. Combining transcriptome and cell lineage information improve our understanding of gene expression regulation in development(*35*). Integrating cell type specific gene expression and disease specific genes (GWAS genes) help infer the cell origin of diseases (*36*). Other routes of analysis may depend on more specific biological questions.

The proliferation of scRNA-seq techniques accompanied a variety of single cell genomic approaches to profile the other molecular layers in individual cells, including single cell measurement of genome sequence(*37*, *38*), histone modification(*39*), chromatin accessibility(*40*, *41*), chromosome conformation(*42*, *43*), DNA methylation(*44*, *45*), and lineage history information(*46*, *47*). More recently, single-cell multimodal measurements are developed to associate multiple molecular features in a single cell. For example, the transcriptome can be simultaneously measured together with other cellular features, including genomic DNA(*48*), DNA methylation(*49–51*), chromatin accessibility(*52*), surface protein(*53*, *54*) or lineage history(*35*), so that the association of variation between different molecular layers can be assessed. The multimodal measurement can be further expanded to link single cell transcriptome with multiplexed perturbation conditions by CRISPR-Cas9 (*55*)(*56*).

Single cell genomics has been widely used in characterizing cell state diversity and discovery of novel cell types across model systems. The first single cell transcriptome atlas of whole multicellular organism (*c.elegans*) was published in 2017(*57*), followed by profiling of more complex model systems including drosophila embryo(*58*), zebrafish embryo(*59*, *60*), Xenopus embryos(*61*) and Planarians(*62*). The application of single cell genomics in mammalian system is limited to major organs initially, such as lung epithelium (*63*), intestinal epithelium(*64*), kidney(*36*), liver(*65*), and nervous system(*66*, *67*). This limitation is quickly resolved by the advent of more advanced single cell profiling techniques, featured by broad atlas study across major mouse organs with single cell transcriptome (*68*, *69*) as well as by single cell epigenome profiling (*70*).

Characterizing cell fate by single cell genomics

Another amazing feature of multicellular organism development is the reproducibility of the cell state diversification process, with associated cell state dynamics in both space and time. For example, in the development of C.elegans, the outcome of every cell division is largely predictable based on their relative location(*71*). This reproducibility suggests the existence of one (or multiple) underlying trajectory, or cell fate, for each final cell state, programmed by gene regulatory networks (*72*). Characterizing the cell state transition path, or cell fate, as well as the internal molecular driving force, is the core in understanding development and applications such as cell engineer.

Classically, cell fate is regarded as a smooth and continuous process, represented by Waddington's epigenetic landscape(*73*). In this assumption, a cell, like a pebble, rolls down from the top of a hill and follows existing paths to its final state. Once cells determine its direction, its future fate is modeled as a smooth process down the trail. Based on this assumption, computation methods (*74*, *75*) are developed to order cells into a smooth transcriptome trajectory, or pseudotime trajectory, generally regarded as cell state transition path, by single cell transcriptome similarity. From the ordered trajectory, starting and end states of cells, as well as intermediate cell states and branching points can be identified. Moreover, key gene factors underlying cell state progression and differentiation can be inferred from the expression dynamics along the trajectory, revealing the mechanism in cell state regulation.

Single cell genomics, together with pseudotime ordering techniques, has been widely applied to reconstructing developmental cell state dynamics. For example, the initial pseudotime analysis(*74*) recovered the development path for myoblast development, validated by the expression dynamics of known gene markers, and identified novel gene regulators. The similar strategy is then applied to other in vitro cell differentiation process such as the development of human definitive endoderm cells(*76*), and mesodermal lineage cell differentiation(*77*). With more advanced cell ordering techniques(*78–81*) as well as the advent of single cell RNA-seq techniques with higher throughput and reduced cost, single cell genomics is used to infer the cell state transition trajectories of all major cellular lineages during in-vivo development, including zebrafish embryo(*59*, *60*), Xenopus embryos(*61*) and Planarians(*62*), and reveals a global view of the dynamic molecular progress underway in the diverse, rapidly changing cell populations during embryo development. Additionally, cell fate characterization is not limited in single cell RNA-seq. For example single cell chromatin accessibility profiling (scATAC-seq) has been

applied to characterize cell state dynamics in myoblast differentiation(*82*) and drosophila embryo development(*83*).

Limitation of current single cell genomic approaches

Despite of the enormous power of single cell genomics in characterizing cell state diversity and deciphering cell fates in development, it has several key limitations: first, most single cell RNA-seq methods rely on the isolation of individual cells within physical compartments. Consequently, preparing single cell RNA-seq libraries with these methods can be low throughput (100s - 1000s of cells), and expensive, the cost scaling linearly with the numbers of cells processed, thus limited in profiling complex developmental system, especially in mammalian development with millions of cells in each individual. Second, although methods are developed to characterize cell state by genome-wide measurement of transcriptome and epigenome (e.g. chromatin accessibility, methylation) at single cell resolution, however, nearly all such methods assay just one aspect of cellular biology, limiting our ability to understand how these aspects fit together to govern gene regulation. Third, almost all current single cell genomic approaches only profile a "snapshot" of cell state. The pseudo-trajectory result can be very different depending on the input gene modules and algorithms for ordering. With temporal information lost during sample preparation, single cell genomics is inefficient to determine the quantitative features in cell fate transition, such as state transition speed and transition probability, which are critical for future cell state prediction and cell engineer.

Organization of Thesis

In the thesis, I describe the development as well as application of novel high throughput single cell genomic approaches to resolve the above limitations, and to better understand cell state diversity and cell fate dynamics in the development of model systems.

The first chapter describes a new strategy to profile the transcriptomes of single cells (**sci-RNA-seq**: Single cell Combinatorial Indexing RNA sequencing) (J.C., J. P., et al, Science, 2017). sci-RNA-seq enables the profiling of tens-of-thousands of single cell transcriptome per experiment with low cost ($0.03-$0.20 per cell) and is readily compatible with cell fixation or nuclei. I applied sci-RNA-seq to profile nearly 50,000 cells from the nematode *Caenorhabditis elegans* at the L2 stage, which is over 50-fold "shotgun cellular coverage" of its somatic cell composition. From these data, we defined consensus expression profiles for 27 main cell types, and recovered rare neuronal cell types corresponding to as few as one or two cells in the L2 worm.

The second chapter describes a new approach to jointly profile chromatin accessibility and mRNA in each of thousands of single cells (sci-CAR, J.C. et al, Science, 2018). As a proof-of-concept, I applied sci-CAR to jointly profile chromatin accessibility and transcription in 11,233 cells from the mouse kidney. I also demonstrated how these data can be used to link cis-regulatory sites to their target genes based on the covariance of chromatin accessibility and transcription across large numbers of single cells.

The third chapter describes the development of a novel high throughput single cell RNA-seq technique (**sci-RNA-seq3**), the first techniques for profiling millions of cells in a single experiment. I applied sci-RNA-seq3 to investigate the transcriptional landscape of mammalian organogenesis, profiled ~2 million cells derived from 61 mouse embryos staged between 9.5 and

13.5 days of gestation. Hundreds of expanding, contracting and transient cell types are identified, together with transcriptome developmental trajectories to all major cell types. We have also used these data to explore the dynamics of proliferation and gene expression within cell types over time, including focused analyses of the apical ectodermal ridge, limb mesenchyme and skeletal muscle.

The fourth chapter describes the first strategy to characterize cell state transition dynamics on whole transcriptome level. The strategy depends on sci-fate, a novel combinatorial indexing based high throughput single cell RNA-seq technique, capable of profiling both whole and newly synthesised transcriptome in thousands of cells. I further developed a computation pipeline to estimate newly synthesised RNA capture rate and gene degradation rate from sci-fate data, and inferred differential trajectories at single cell level. As a proof of concept, I applied sci-fate to a model system of cortisol response, and characterized over 6,000 single cell state transition events, consistent with known cell cycle dynamics upon glucocorticoid receptor activation. From the analysis, I showed the cell state transition direction and probabilities are regulated by inter-state distances and state instability landscape.

The final chapter of this thesis discuss key conclusions from this work and future directions.

Reference:

1. S. F. Gilbert, M. J. F. Barresi, DEVELOPMENTAL BIOLOGY, 11TH EDITION 2016. *Am. J. Med. Genet. A*. **173**, 1430–1430 (2017).

2. R. Hooke, Jo Martyn And, *Micrographia, or, Some physiological descriptions of minute bodies made by magnifying glasses :with observations and inquiries thereupon /by R.*

Hooke . (1665).

3. Santiago Ramon y Cajal, 1852-1934. *Am. J. Psychiatry*. **152**, 914–914 (1995).

4. F. W. D. Rost, *Fluorescence Microscopy* (Cambridge University Press, 1992).

5. M. J. Fulwyler, in *Automation in Hematology* (1981), pp. 69–80.

6. J. Shendure *et al.*, Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*. **309**, 1728–1732 (2005).

7. M. Margulies *et al.*, Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. **437**, 376–380 (2005).

8. G. Sablok, S. Kumar, S. Ueno, J. Kuo, C. Varotto, *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches* (Springer, 2015).

9. J. Shendure *et al.*, DNA sequencing at 40: past, present and future. *Nature*. **550**, 345–353 (2017).

10. D. Baltimore, Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of RNA Tumour Viruses. *Nature*. **226**, 1209–1211 (1970).

11. H. M. Temin, S. Mizutani, Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of Rous Sarcoma Virus. *Nature*. **226**, 1211–1213 (1970).

12. G. T. Hermanson, in *Bioconjugate Techniques* (2008), pp. 783–823.

13. B. E. Bernstein *et al.*, Genomic Maps and Comparative Analysis of Histone Modifications in Human and Mouse. *Cell*. **120**, 169–181 (2005).

14. A. P. Boyle *et al.*, High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*. **132**, 311–322 (2008).

15. F. Tang *et al.*, mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*. **6**, 377–382 (2009).

16. M. N. Bainbridge *et al.*, Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*. **7**, 246 (2006).

17. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*. **5**, 621–628 (2008).

18. D. A. Jaitin *et al.*, Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*. **343**, 776–779 (2014).

19. D. Ramsköld *et al.*, Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).

20. A. M. Klein *et al.*, Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*. **161**, 1187–1201 (2015).

21. E. Z. Macosko *et al.*, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. **161**, 1202–1214 (2015).

22. G. X. Y. Zheng *et al.*, Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

23. R. Rostom, V. Svensson, S. A. Teichmann, G. Kar, Computational approaches for interpreting scRNA-seq data. *FEBS Lett.* **591**, 2213–2225 (2017).

24. K. Pearson, LIII.On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*. **2**, 559–572 (1901).

25. J. E. Jackson, J. Edward Jackson, The User's Guide to Multidimensional Scaling. *Technometrics*. **27**, 87–88 (1985).

26. J. Leps, P. Smilauer, in *Multivariate Analysis of Ecological Data using CANOCO*, pp. 149–167.

27. W. R. Fox, L. Kaufman, P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. *Appl. Stat.* **40**, 486 (1991).

28. A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks. *Science*. **344**, 1492–1496 (2014).

29. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

30. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell developmental trajectories (2017), , doi:10.1101/110668.

31. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 (2014), , doi:10.1101/002832.

32. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods*. **14**, 979–982 (2017).

33. G. I. Allen, Z. Liu, in *2012 IEEE International Conference on Bioinformatics and Biomedicine* (2012; http://dx.doi.org/10.1109/bibm.2012.6392619).

34. M. Uhlén *et al.*, Proteomics. Tissue-based map of the human proteome. *Science*. **347**, 1260419 (2015).

35. B. Raj *et al.*, Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain by scGESTALT (2017), , doi:10.1101/205534.

36. J. Park *et al.*, Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science*. **360**, 758–763 (2018).

37. N. Navin *et al.*, Tumour evolution inferred by single-cell sequencing. *Nature*. **472**, 90–94 (2011).

38. S. A. Vitak *et al.*, Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods*. **14**, 302–308 (2017).

39. A. Rotem *et al.*, Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* **33**, 1165–1172 (2015).

40. D. A. Cusanovich *et al.*, Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*. **348**, 910–914 (2015).

41. J. D. Buenrostro *et al.*, Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*. **523**, 486–490 (2015).

42. V. Ramani *et al.*, Massively multiplex single-cell Hi-C (2016), , doi:10.1101/065052.

43. T. Nagano *et al.*, Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*. **502**, 59–64 (2013).

44. C. Luo *et al.*, Single-cell methylomes identify neuronal subtypes and regulatory elements in

mammalian cortex. *Science*. **357**, 600–604 (2017).

45. R. M. Mulqueen *et al.*, Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018).

46. A. McKenna *et al.*, Whole organism lineage tracing by combinatorial and cumulative genome editing (2016), , doi:10.1101/052712.

47. K. L. Frieda *et al.*, Synthetic recording and in situ readout of lineage information in single cells. *Nature*. **541**, 107–111 (2017).

48. Y. Hou *et al.*, Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* **26**, 304–319 (2016).

49. S. J. Clark *et al.*, scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).

50. Y. Hu *et al.*, Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol.* **17**, 88 (2016).

51. S. Pott, Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *Elife*. **6** (2017), doi:10.7554/eLife.23203.

52. C. Angermueller *et al.*, Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods*. **13**, 229–232 (2016).

53. M. Stoeckius *et al.*, Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods*. **14**, 865–868 (2017).

54. V. M. Peterson *et al.*, Multiplexed quantification of proteins and transcripts in single cells.

Nat. Biotechnol. **35**, 936–939 (2017).

55. A. Dixit *et al.*, Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*. **167**, 1853–1866.e17 (2016).

56. M. Gasperini *et al.*, A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*. **176**, 377–390.e19 (2019).

57. J. Cao *et al.*, Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*. **357**, 661–667 (2017).

58. N. Karaiskos *et al.*, The embryo at single-cell transcriptome resolution. *Science*. **358**, 194–199 (2017).

59. D. E. Wagner *et al.*, Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*. **360**, 981–987 (2018).

60. J. A. Farrell *et al.*, Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*. **360**, eaar3131 (2018).

61. M. Klymkowsky, F1000Prime recommendation of The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *F1000 - Post-publication peer review of the biomedical literature* (2019), , doi:10.3410/f.733108089.793555548.

62. M. Plass *et al.*, Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*. **360** (2018), doi:10.1126/science.aaq1723.

63. B. Treutlein *et al.*, Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*. **509**, 371–375 (2014).

64. A. L. Haber *et al.*, A single-cell survey of the small intestinal epithelium. *Nature*. **551**, 333–339 (2017).

65. K. B. Halpern *et al.*, Erratum: Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. *Nature*. **543**, 742–742 (2017).

66. A. Zeisel *et al.*, Molecular Architecture of the Mouse Nervous System. *Cell*. **174**, 999–1014.e22 (2018).

67. B. B. Lake *et al.*, Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).

68. X. Han *et al.*, Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*. **173**, 1307 (2018).

69. The Tabula Muris Consortium *et al.*, Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*. **562**, 367–372 (2018).

70. D. A. Cusanovich *et al.*, A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell*. **174**, 1309–1324.e18 (2018).

71. J. E. Sulston, E. Schierenberg, J. G. White, J. N. Thomson, The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev. Biol.* **100**, 64–119 (1983).

72. N. Moris, C. Pina, A. M. Arias, Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* **17**, 693–703 (2016).

73. M. Allen, Compelled by the Diagram: Thinking through C. H. Waddington's Epigenetic Landscape. *Contemporaneity: Historical Presence in Visual Culture*. **4**, 119–142 (2015).

74. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by

pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

75. S. C. Bendall *et al.*, Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*. **157**, 714–725 (2014).

76. L.-F. Chu *et al.*, Single-cell RNA-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm. *Genome Biol.* **17** (2016), doi:10.1186/s13059-016-1033-x.

77. K. M. Loh *et al.*, Mapping the Pairwise Choices Leading from Pluripotency to Human Bone, Heart, and Other Mesoderm Cell Types. *Cell*. **166**, 451–467 (2016).

78. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell developmental trajectories (2017), , doi:10.1101/110668.

79. M. Setty *et al.*, Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat. Biotechnol.* **34**, 637–645 (2016).

80. L. Haghverdi, M. Buettner, F. Alexander Wolf, F. Buettner, F. J. Theis, Diffusion pseudotime robustly reconstructs lineage branching (2016), , doi:10.1101/041384.

81. K. R. Campbell, C. Yau, Probabilistic modeling of bifurcations in single-cell gene expression data using a Bayesian mixture of factor analyzers. *Wellcome Open Res*. **2**, 19 (2017).

82. H. Pliner *et al.*, Chromatin accessibility dynamics of myogenesis at single cell resolution (2017), , doi:10.1101/155473.

83. D. A. Cusanovich *et al.*, The cis-regulatory dynamics of embryonic development at single cell resolution (2017), , doi:10.1101/166066.

Chapter 1: comprehensive characterization of cell states by single cell combinatorial indexing RNA sequencing

*Modified from article Comprehensive single cell transcriptional profiling of a multicellular organism, Junyue Cao, Jonathan Packer, et al, Science, 2017

Authors: Junyue Cao[1,2,†], Jonathan S. Packer[1,†], Vijay Ramani[1,††], Darren A. Cusanovich[1,††], Chau Huynh[1], Riza Daza[1], Xiaojie Qiu[1,2], Choli Lee[1], Scott N. Furlan[3,4,5], Frank J. Steemers[6], Andrew Adey[7,8], Robert H. Waterston[1,*], Cole Trapnell[1,*], Jay Shendure[1,9,*]

Affiliations:

[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA.

[2]Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA.

[3]Ben Towne Center for Childhood Cancer Research, Seattle Children's Research Institute, Seattle, WA, USA.

[4]Department of Pediatrics, University of Washington, Seattle, WA, USA.

[5]Fred Hutchinson Cancer Research Center, Seattle, WA, USA.

[6]Illumina Inc., Advanced Research Group, San Diego, CA, USA.

[7]Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR, USA.

[8]Knight Cardiovascular Institute, Portland, OR, USA.

[9]Howard Hughes Medical Institute, Seattle, WA, USA.

†These authors contributed equally to this work

††These authors contributed equally to this work

*Correspondence to: coletrap@uw.edu (CT), watersto@uw.edu (RHW) & shendure@uw.edu (JS).

Abstract: To resolve cellular heterogeneity, we developed a combinatorial indexing strategy to profile the transcriptomes of single cells or nuclei (sci-RNA-seq: Single cell Combinatorial Indexing RNA sequencing). We applied sci-RNA-seq to profile nearly 50,000 cells from the nematode *Caenorhabditis elegans* at the L2 stage, which is over 50-fold "shotgun cellular coverage" of its somatic cell composition. From these data, we define consensus expression profiles for 27 cell types, and recover rare neuronal cell types corresponding to as few as one or two cells in the L2 worm. We integrate these profiles with whole animal ChIP sequencing data to deconvolve the cell type specific effects of transcription factors. These data generated by sci-RNA-seq constitute a powerful resource for nematode biology, and foreshadow similar atlases for other organisms.

One Sentence Summary: We applied single cell combinatorial indexing RNA-seq to achieve >50-fold "shotgun cellular coverage" of the somatic cell composition of L2 *C. elegans*.

Introduction:

Individual cells are the natural unit of form and function in biological systems. However, conventional methods for profiling the molecular content of biological samples mask cellular heterogeneity, likely present even in ostensibly homogenous tissues (*1*). Recently, profiling the transcriptome of individual cells has emerged as a powerful strategy for resolving such heterogeneity. The expression levels of mRNA species are linked to cellular function, and therefore can be used to classify cell types (*2–10*) and to order cell states (*11*). Although methods

for single cell RNA-seq have proliferated, they rely on the isolation of individual cells within physical compartments (*12–20*). Consequently, preparing single cell RNA-seq libraries with these methods can be expensive, the cost scaling linearly with the numbers of cells processed (*21*, *22*).

We recently developed combinatorial indexing, a method using split-pool barcoding of nucleic acids to uniquely label a large number of single molecules or single cells. Single *molecule* combinatorial indexing can be used for haplotype-resolved genome sequencing and *de novo* genome assembly (*23*, *24*), while single *cell* combinatorial indexing ("sci") can be used to profile chromatin accessibility (sci-ATAC-seq) (*25*), genome sequence (sci-DNA-seq) (*26*), genome-wide chromosome conformation (sci-Hi-C) (*27*), and DNA methylation (sci-MET) (*28*) in large numbers of single cells.

Here we developed a combinatorial indexing method to uniquely label the transcriptomes of large numbers of single cells or nuclei (sci-RNA-seq). We then applied sci-RNA-seq to deeply profile single cell transcriptomes in the nematode *C. elegans* at the L2 stage. *C. elegans* is the only multicellular organism for which all cells and cell types are defined, as is its entire developmental lineage (*29*, *30*). However, despite its modest cell count (*e.g.* 762 somatic cells per L2 larva), our knowledge of the molecular state of each cell and cell type remains fragmentary. We therefore saw an opportunity to generate a powerful resource for nematode biologists as well as for the single cell genomics community.

Results:

Overview of sci-RNA-seq

In its current form, sci-RNA-seq relies on the following steps (Fig. 1A): 1) Cells are fixed and permeabilized with methanol (alternatively, cells are lysed and nuclei recovered), and then split across 96- or 384-well plates. 2) A first molecular index is introduced to the mRNA of cells within each well with *in situ* reverse transcription (RT) incorporating a barcode-bearing, well-specific polyT primer containing unique molecular identifiers (UMI). 3) All cells are pooled and redistributed by fluorescence activated cell sorting (FACS) to 96- or 384-well plates in limiting numbers (*e.g.* 10-100 per well). Cells are gated on the basis of DAPI (4',6-diamidino-2-phenylindole) staining to discriminate single cells from doublets during sorting. 4) Second strand synthesis, transposition with Tn5 transposase, lysis, and PCR amplification are performed. The PCR primers target the barcoded polyT primer on one end, and the Tn5 adaptor insertion on the other end, such that resulting PCR amplicons preferentially capture the 3' ends of transcripts. These primers introduce a second barcode, specific to each well of the PCR plate. 5) Amplicons are pooled and subjected to massively parallel sequencing, resulting in 3'-tag digital gene expression profiles, with each read associated with two barcodes corresponding to the first and second rounds of cellular indexing (Fig. 1B). In a variant of the method described below, we introduce a third round of cellular indexing during Tn5 transposition of double-stranded cDNA.

The majority of cells pass through a unique combination of wells, resulting in a unique combination of barcodes for each cell that tags its transcripts. The rate of two or more cells receiving the same combination of barcodes can be tuned by adjusting how many cells are distributed to the second set of wells (*25*). Increasing the number of barcodes used during each round of indexing leads boosts the number of cells that can be profiled while reducing the effective cost per cell (fig. S1). Additional levels of indexing can potentially offer even greater complexity and lower costs. Multiple samples (*e.g.* different cell populations, tissues,

individuals, time-points, perturbations, replicates, etc.) can be concurrently processed within one experiment, using different subsets of wells for each sample during the first round of indexing.

Scalability of sci-RNA-seq

We tested 262 sci-RNA-seq conditions with mammalian cells, optimizing the protocol and reaction conditions. We demonstrate scalability with 384 x 384 well sci-RNA-seq. During the first round of indexing, half of 384 wells contained pure populations of either human (HEK293T or HeLa S3) or mouse (NIH/3T3) cells, and the other half mixed human and mouse cells. After barcoded RT, cells were pooled and then sorted to a new 384 well plate for the second round of barcoding and deep sequencing of pooled PCR amplicons. We recovered 15,997 single cell transcriptomes and readily assigned cells as human or mouse (Fig. 1C).

Optimization of sci-RNA-seq and application to nuclei

We performed optimized 96 x 96 well sci-RNA-seq on five cell populations, each present in distinct subsets of wells during the first round of barcoding: HEK293T cells (8 wells); HeLa S3 cells (8 wells); an intraspecies mixture of HEK293T and HeLa S3 cells (32 wells); and interspecies mixtures of HEK293T and NIH/3T3 cells (24 wells) or nuclei (24 wells). We deeply sequenced the resulting library (~250,000 reads per cell; ~210,000 reads per nucleus; ~88% duplication rate), profiling 744 single cell and 175 single nucleus transcriptomes.

Transcriptomes in the 24 wells containing an interspecies mixture of human and mouse cells overwhelmingly mapped to the genome of one species or the other (289 of 294 cells), with only 5 ‘collisions’ (where collisions likely represent coincidental passage through the same wells by two or more cells) (Fig. 1D). Excluding collisions, we observed an average of 24,454 UMIs (5,604 genes) per human cell and 17,665 UMIs (4,065 genes) per mouse cell, with 1.9% and 3.3% of reads per cell mapping to the incorrect species.

Transcriptomes originating in the 24 wells containing an interspecies mixture of human and mouse nuclei also overwhelmingly mapped to the genome of one species or the other (172 of 175 nuclei), with only 3 collisions (fig. S2A). Excluding collisions, we observed an average of 32,951 UMIs (5,737 genes) per human nucleus and 20,123 UMIs (4,107 genes) per mouse nucleus (fig. S2B-C), with 2.2% and 1.9% of reads per cell mapping to the incorrect species. The greater UMI counts in nuclei are potentially due to the higher amounts of mRNA in cells resulting in a reduced RT efficiency per molecule. Consistent with this, optimizing the number of cells per RT reaction increased UMI counts per cell (*31*).

Estimates of gene expression from the aggregated transcriptomes of nuclei versus cells were well correlated (Pearson: 0.96 for HEK293T, 0.97 for NIH/3T3; Fig. 1E, fig. S2D). From cells, 81% of reads mapped to the expected strand of genic regions (47% exonic, 34% intronic), and 19% to intergenic regions or the unexpected strand of genic regions. From nuclei, 84% of reads mapped to the expected strand of genic regions (35% exonic, 49% intronic) and 16% to intergenic regions or the unexpected strand of genic regions, similar to previous studies (*32*). Whereas exonic reads show an expected enrichment at the 3' ends of gene bodies, intronic reads do not, and may be the result of poly(dT) priming from poly(dA) tracts in heterogeneous nuclear RNA (fig. S3).

Transcriptomes originating in the 48 wells containing pure or an intraspecies mixture of HEK293T and HeLa S3 cells were readily separated into two clusters by t-stochastic neighbor embedding (t-SNE) (Figs. 1F and S4). Estimates of gene expression from the aggregated transcriptomes of all identified HEK293T cells versus a related bulk RNA-seq workflow (Tn5-RNA-seq (*33*)) without methanol fixation were well correlated (Pearson: 0.94, Fig. 1G).

Robustness of sci-RNA-seq

After optimizing the number of cells per RT reaction, we fixed a mixture of HEK293T and NIH/3T3 cells, and performed 16 x 84 well sci-RNA-seq (*31*). We recovered 185 human cells and 109 mouse cells with 22 collisions (Fig. 2A). At ~240,000 reads per cell (73% duplication rate), we observed an average of 49,043 UMIs (7,563 genes) per human cell and 36,737 UMIs (6,263 genes) per mouse cell (Fig. 2B, fig. S5A), with 0.9% and 1.2% of reads per cell mapping to the incorrect species. Although this and the previous experiment were performed two months apart on independently grown and fixed cells, the aggregated transcriptomes were well correlated (Pearson: 0.98 for HEK293T, 0.98 for NIH/3T3; Figs. 2C and S5B).

We stored a portion of the methanol-fixed mixture of HEK293T and NIH/3T3 cells at -80C for 4 days and repeated sci-RNA-seq. At ~200,000 reads per cell (73% duplication rate), we observed an average of 30,024 UMIs (5,965 genes) per human cell and 21,393 UMIs (4,503 genes) per mouse cell, with comparable purity (fig. S5C). The aggregated transcriptomes of the fixed-fresh vs. fixed-frozen cells were well correlated (Pearson: 0.99 for HEK293T cells, 0.98 for NIH/3T3 cells; Figs. 2D and S5D).

sci-RNA-seq with three levels of indexing

Two-level combinatorial indexing enables routine profiling of ~10^4 single cells per experiment. We tested an additional level of indexing during Tn5 transposition of double-stranded cDNA (*25*). We performed 16 x 6 x 16 well sci-RNA-seq on mixed HEK293T and NIH/3T3 cells after methanol fixation. After RT with 16 barcodes and second strand synthesis, cells were pooled and distributed to 6 wells for tagmentation with indexed Tn5 (6 barcodes), then pooled again and sorted to 16 wells for PCR with indexed primers. At ~20,000 reads per cell (51% duplication rate), we recovered 119 human and 62 mouse cells with 5 collisions (fig. S6A). The aggregated transcriptomes of three-level vs. two-level sci-RNA-seq were well correlated (Pearson: 0.96 for HEK293T, 0.94 for NIH/3T3; fig. S6B-C). Downsampling to 15,000 reads per

cell, three-level indexing recovered fewer UMIs per cell than two-level indexing (3-level: on average, 6,033 for HEK293T, 3,640 for NIH/3T3; 2-level: 9,942 for HEK293T, 8,611 for NIH/3T3; fig. S6D-G), possibly due to lower efficiency of indexed vs. unindexed Tn5. This limitation notwithstanding, three-level combinatorial indexing has the potential to enable routine profiling of $>10^6$ single cells per experiment (fig. S6H; (*31*)).

Single cell RNA profiling of C. elegans

We next applied sci-RNA-seq to *C. elegans*. Of note, the cells in *C. elegans* larvae are much smaller, more variably sized, and have lower mRNA content than the mammalian cell lines on which we optimized the protocol. We pooled ~150,000 larvae synchronized at the L2 stage and dissociated them into single-cell suspensions. We then performed *in situ* RT across six 96-well plates (576 first-round barcodes), each well containing ~1,000 *C. elegans* cells and also ~1,000 human cells (HEK293T) as internal controls. After pooling all cells, we sorted the mixture of *C. elegans* and HEK293T cells to 10 new 96-well plates for PCR barcoding (960 second-round barcodes), gating on DNA content to distinguish between *C. elegans* and HEK293T cells. This sorting resulted in 96% of wells harboring only *C. elegans* cells (140 each), and 4% of wells harboring a mix of *C. elegans* and HEK293T cells (140 *C. elegans* and 10 HEK293T each).

This experiment yielded 42,035 *C. elegans* single-cell transcriptomes (UMI counts per cell for protein-coding genes ≥ 100). 94% of reads mapped to the expected strand of genic regions (92% exonic, 2% intronic). At a sequencing depth of ~20,000 reads per cell and a duplication rate of 80%, we identified a median of 575 UMIs mapping to protein-coding genes per cell (mean 1,121 UMIs and 431 genes per cell) (fig. S7A). Importantly, control wells containing both *C. elegans* and HEK293T cells demonstrated clear separation between species (fig. S7B), with 3.1% and 0.2% of reads per cell mapping to the incorrect species, respectively.

Identifying cell types

Semi-supervised clustering analysis segregated the cells into 29 distinct groups, the largest containing 13,205 (31.4%) and the smallest only 131 (0.3%) cells (Fig. 3A). Somatic cell types comprised 37,734 cells. We identified genes that were expressed specifically in a single cluster, and by comparing those genes to expression patterns reported in the literature, assigned the clusters to cell types (figs. S15-S23). Twenty-six cell types were represented in the 29 clusters: 19 represented exactly one literature-defined cell type, 7 contained multiple distinct cell types, 2 contained cells of a specific cell type but had abnormally low UMI counts, and 1 could not be readily assigned. Neurons, which were present in 7 clusters in the global analysis, were independently reclustered, initially revealing 10 major neuronal subtypes.

Intestine cells were not represented in any cluster. Intestine cells comprise 2.5% of the somatic cells but are polyploid in *C. elegans* larvae (*34*) and also autofluorescent in the DAPI channel used to measure DNA content (*35*). We speculated they may have been excluded by how we gated on DNA content. We therefore performed a second 384 x 144 well *C. elegans* experiment, collecting all cells including polyploid cells on the basis of DAPI fluorescence (96 wells), or gating to enrich for polyploid cells (48 wells). Intestine cells were present (as compared with their absence in the previous experiment) and 2-fold enriched in wells gated for polyploidy. This experiment yielded 7,325 cells (UMI counts per cell for protein-coding genes ≥ 200), of which 6,335 were somatic and 511 intestine cells (fig. S8A).

Gene expression patterns in hypodermal cells suggested that the worm cells from the second *C. elegans* experiment were more tightly synchronized, overlapping but not identical in developmental timing to the first experiment (fig. S8B-F). *C. elegans* larvae feature pervasive oscillations in gene expression within each larval stage (*36*), making it difficult to distinguish biological variation from batch effects. However, the aggregated transcriptomes of human

HEK293T cells from these same experiments were well correlated (Pearson: 0.97) and not readily separated by tSNE (fig. S9). This suggests that the variation observed is primarily due to differences in the developmental timing or preparation of the *C. elegans* larvae and cells, rather than technical variation in the sci-RNA-seq protocol. Regardless of its source, to minimize confounding by this variation, we only included the intestine cells from the second *C. elegans* experiment in subsequent analyses, with all other cell types being represented by the first experiment only.

The global and neuron-specific clustering analyses from the first *C. elegans* experiment, supplemented with intestine cells from the second experiment, allowed us to construct aggregate expression profiles for 27 cell types (Tables S2-S4; a 28th cell type, dopaminergic neurons, is excluded due to small cell numbers). These profiles are available online via GExplore (http://genome.sfu.ca/gexplore/gexplore_search_tissues.html; fig. S14). Comparing the observed proportions of each cell type to their known frequencies in L2 larvae showed that sci-RNA-seq captured many cell types at or near expected frequencies (Fig. 3B; 15/28 types had abundance ≥ 50%, and 27/28 had abundance ≥ 20%, of expectation).

Transcriptional programs can be readily distinguished within single cell transcriptome datasets at shallow sequencing depths (*37*). Thus, despite being able to distinguish many distinct cell types in the worm, our molecular definition for each would be incomplete. However, we observed that half of all *C. elegans* protein-coding genes were expressed in at least 100 cells in the full dataset, and 66% of protein-coding genes in at least 20 cells. This compares favorably with the estimates of expressed genes at the L2 stage from whole animal RNA-seq (69%) (*38*). The “whole worm” expression profile derived by aggregating all sci-RNA-seq reads correlated well with whole animal bulk RNA-seq (*38*) for L2 *C. elegans* (Fig. 3C; Spearman: 0.796 with cells from the first experiment only, 0.824 including intestine cells from the second experiment).

Furthermore, 3,925 genes were enriched in a single tissue (differential expression at least five-fold greater than the 2nd-highest expressing tissue; Fig. 3D), and 1,939 genes were enriched for expression in a single cell type (Fig. 3E). Thus, despite the fact that sci-RNA-seq captures a minority of transcripts in each cell, our 'oversampling' of the cellular composition of the organism enables us to construct representative expression profiles for individual cell types (Fig. 3F).

Neuronal cell types

Because the transcripts of tissue or cell type clusters suggested subclasses within groups (Fig. 4A), we examined expression within several tissues in more detail. We confirmed and extended findings that anterior and posterior body wall muscle have distinct expression patterns (fig. S10A-B, (*39*)), and also observed distinct expression patterns for posterior vs. other intestine cells (fig. S10C-D) and amphid vs. phasmid sheath cells (fig. S10E-F). But gene expression patterns were particularly diverse in neuronal cell types.

By morphological criteria, the 302 neurons of worm are classified into 118 distinct types (*40*) and from the database of reporter transgene expression patterns, most of these are postulated to have unique molecular signatures (*41*). Our initial re-clustering of neuronal cells divided them into 10 broad classes (Fig. 4A). Most classes of neurons were represented by several small but highly distinct clusters in the t-SNE plot. Further analysis of cluster-specific gene expression showed that many clusters corresponded to highly specific subsets of neurons in the L2 worm (Fig. 4B). Three clusters corresponded to sets of four neurons in an individual worm, 8 clusters corresponded to a single pair of neurons (AFD, ASG, ASK, AWA, BAG, CAN, RIA, and RIC), and 3 clusters corresponded to exactly one neuron (ASEL, ASER, and DVA). Hierarchical clustering analysis showed that of the most of 917 genes highly enriched in neurons, compared to other tissues, were expressed in only a minority of neuronal clusters (Fig. 4C). 73% of neuron-

enriched genes had no more than 10 neuron clusters (out of 40 total) in which they were expressed at ≥ 10% of the level of the highest-expressed cluster. 155 genes were highly enriched in a single neuron cluster relative to all others (Fig. 4D).

Expression of marker genes, such as *gcy-3* and *gcy-6,* were key in identifying two neuronal clusters as left ASE (ASEL) and right ASE (ASER) gustatory neurons, respectively (Fig. 4E). These neurons have asymmetry in gene expression (*42*), and we observe 44 genes to be differentially expressed (Fig. 4F, fold difference > 3, FDR < 5%). mRNA from these neurons has previously been profiled with co-immunoprecipitation of RNA and a transgenic poly(A)-binding protein expressed specifically in ASEL or ASER, followed by microarray analysis (*43*). The differentially expressed genes we observe are consistent with this study (fig. S11), highlighting the ability of sci-RNA-seq to facilitate the analysis of cell types as rare as a single cell per individual.

Two neuronal clusters correspond to sister cells, the AWA and ASG neurons, (Fig. 4G), which arise from the same parental cell in the last round of *C. elegans* embryonic cell divisions. Their differentiation has previously been used as a model for the study of the regulation of cell fate decisions (*44*). In our data, 136 genes were differentially expressed between these two cell types (Fig. 4H, fold difference > 3, FDR < 5%). The divergent transcriptomes of the AWA and ASG neurons, along with the left and right ASE neurons, highlight the potential of cells that are extremely closely related in morphology and developmental lineage to feature distinct programs of gene regulation.

Integration with transcription factor binding sites

We hypothesized that correlating transcription factor (TF) binding patterns—profiled in ChIP-seq experiments from the modENCODE (*45*) and modERN (*46*) consortia—with cell type

gene expression profiles could give insights into the regulatory programs underlying the gene expression profiles. For each of 27 cell types, we constructed regularized regression models to predict each gene's expression as a function of the TF ChIP peaks present in its promoter (Fig. 5). We restricted a cell type's model to those TFs that were detectably expressed within it (>10 transcripts per million (TPM)), increasing the proportion of TF-to-cell-type associations that are likely to reflect causal gene regulation. Our regression analysis predicted gene expression by selecting numerous regulators critical for development or proper function-specific cell types, including *hlh-1* and *unc-120* in body wall muscle (*47*), *pha-4* in pharyngeal cell types (*48*), *hlh-8* (CeTwist) in sex myoblasts (*49*), *blmp-1* and *nhr-25* in hypodermis (*50*, *51*), *elt-2* in the intestine (*52*), and *xnd-1* in the germline (*53*, *54*).

The regression identified several putative novel regulators of cell-type specific expression. For example, *fkh-8,* which is expressed specifically in ciliated sensory neurons (our data and reporter construct from (*55*)) was predictive of their gene expression program (fig. S12). The uncharacterized TF *F49E8.2* is expressed specifically in the germline and associated with germline gene expression (fig. S12). F49E8.2 is an ortholog of the human gene "E2F-associated phosphoprotein" (EAPP) (*56*), and F49E8.2 ChIP-seq peaks co-localize with germline-specific EFL-1 peaks (ortholog of E2F, data from (*57*)) more often than could be expected due to chance (fig. S13ab, χ^2-test, $p = 2.8 \times 10^{-21}$), suggesting that these proteins may physically interact. The hypodermis-associated TFs *blmp-1* and *nhr-25* were also associated with gene expression in socket cells, excretory cells, and rectal cells. *nhr-25* is expressed 4.5-fold higher in socket cells than in seam cells (560 vs. 124 TPM) and 8.7-fold more than in the non-seam hypodermis (560 vs. 64 TPM), suggesting a role in glial development.

Discussion

Our method for single cell RNA-seq combinatorial indexing of cells or nuclei can be applied to profile the transcriptomes of tens-of-thousands of single cells per experiment through a library construction completed by a single individual in two days at a cost of $0.03-$0.20 per cell. sci-RNA-seq is compatible with cell fixation, which can minimize perturbations to cell state or RNA integrity before or during processing and facilitates the concurrent processing of multiple samples within a single experiment, potentially reducing batch effects relative to platforms requiring serial processing, an area of concern for the single cell RNA-seq field (*58*). Given that the second barcode is introduced after flow sorting, it is also possible to associate wells on the PCR plate with FACS-defined subpopulations. sci-RNA-seq is also compatible nuclei, which may be important for tissues for which unbiased cell disaggregation protocols are not well established (possibly most tissues). Lastly, sci-RNA-seq is scalable. We demonstrate up to 576 x 960 indexing, which enabled the generation of ~4 x 10^4 single cell transcriptomes in one experiment. However, processing of more cells with sub-linear cost scaling is possible by using more barcoded RT and PCR primers (*e.g.* 1,536 x 1,536 combinatorial indexing) and/or introducing additional rounds of indexing. With 384 x 384 x 384 combinatorial indexing, one can hypothetically profile the transcriptomes of over 10 million cells per experiment.

With sci-RNA-seq we generated a catalog of single cell transcriptomes with over 50-fold "shotgun cellular coverage" of the L2 *C. elegans* soma. We detect 18 non-neuronal cell types and a multitude of neuronal cell types, which we grouped into either 10 broad classes or 40 fine-grained clusters from an unsupervised analysis, highlighting the potential of an organism's gene regulatory programs to be enacted at a fine-grained level. We anticipate these data will be a rich resource for nematode biology – a starting point for an atlas that leverages Sulston's lineage map to define the molecular state of every cell throughout the life cycle of *C. elegans.* Furthermore, as illustrated by our experience with intestinal cells, the greater knowledge of "ground truth" for

C. elegans may further the refinement of experimental and computational methods for recovering and distinguishing cell types and states.

sci-RNA-seq expands the repertoire of single cell molecular phenotypes that can be resolved by combinatorial indexing (*25–28*). Provided that multiple aspects of cellular biology can be concurrently barcoded, combinatorial indexing may also facilitate the scalable generation of 'joint' single cell molecular profiles (*e.g.* RNA-seq and ATAC-seq from each of many single cells). We also envision that large-scale, integrated profiling of the molecular states and lineage histories (*59*) of single cells in other organisms will begin to give shape to "global views" of their developmental biology.

ACKNOWLEDGMENTS

The raw data have been deposited with the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) under accession code GSE98561 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=wduvwooqvrynngt&acc=GSE98561). These data are also made available as gene-by-cell matrices, along with a "vignette" in the form of a Jupyter notebook that shows examples of how to work with the data (http://waterston.gs.washington.edu/sci_RNA_seq_gene_count_data). We thank members of the Shendure, Trapnell and Waterston labs for helpful discussions and feedback, particularly A. Hill, V. Agarwal, M. Gasperinin, L. Starita, Y. Yin and B. Martin; S. Zimmerman and C. Berg for helpful technique suggestions; the modERN consortium for allowing us to use their ChIP-seq data; D. Prunkard, and L. Gitari in the Pathology Flow Cytometry Core Facility for their exceptional assistance in flow sorting; J. Rose, D. Maly, L. VandenBosch, and T. Reh lab for sharing the NIH/3T3 cell line; H. Hutter for adding our tissue-specific expression profiles on gExplore. HeLa S3 cells were used as part of this study. H. Lacks, and the HeLa cell line that was established from her tumor cells in 1951, have made significant contributions to scientific

progress and advances in human health. We are grateful to H. Lacks, now deceased, and to her surviving family members for their contributions to biomedical research. This work was funded by grants from the NIH (DP1HG007811 and R01HG006283 to JS; U41HG007355 and R01GM072675 to RHW, DP2 HD088158 to CT) and the Paul G. Allen Family Foundation (to JS), W. M. Keck Foundation (to CT and JS), Dale. F. Frey Award for Breakthrough Scientists (to CT) and Alfred P. Sloan Foundation Research Fellowship (to CT) and by the William Gates III Endowed Chair in Biomedical Sciences (RHW). DAC was supported in part by T32HL007828 from the National Heart, Lung, and Blood Institute. JS is an Investigator of the Howard Hughes Medical Institute.

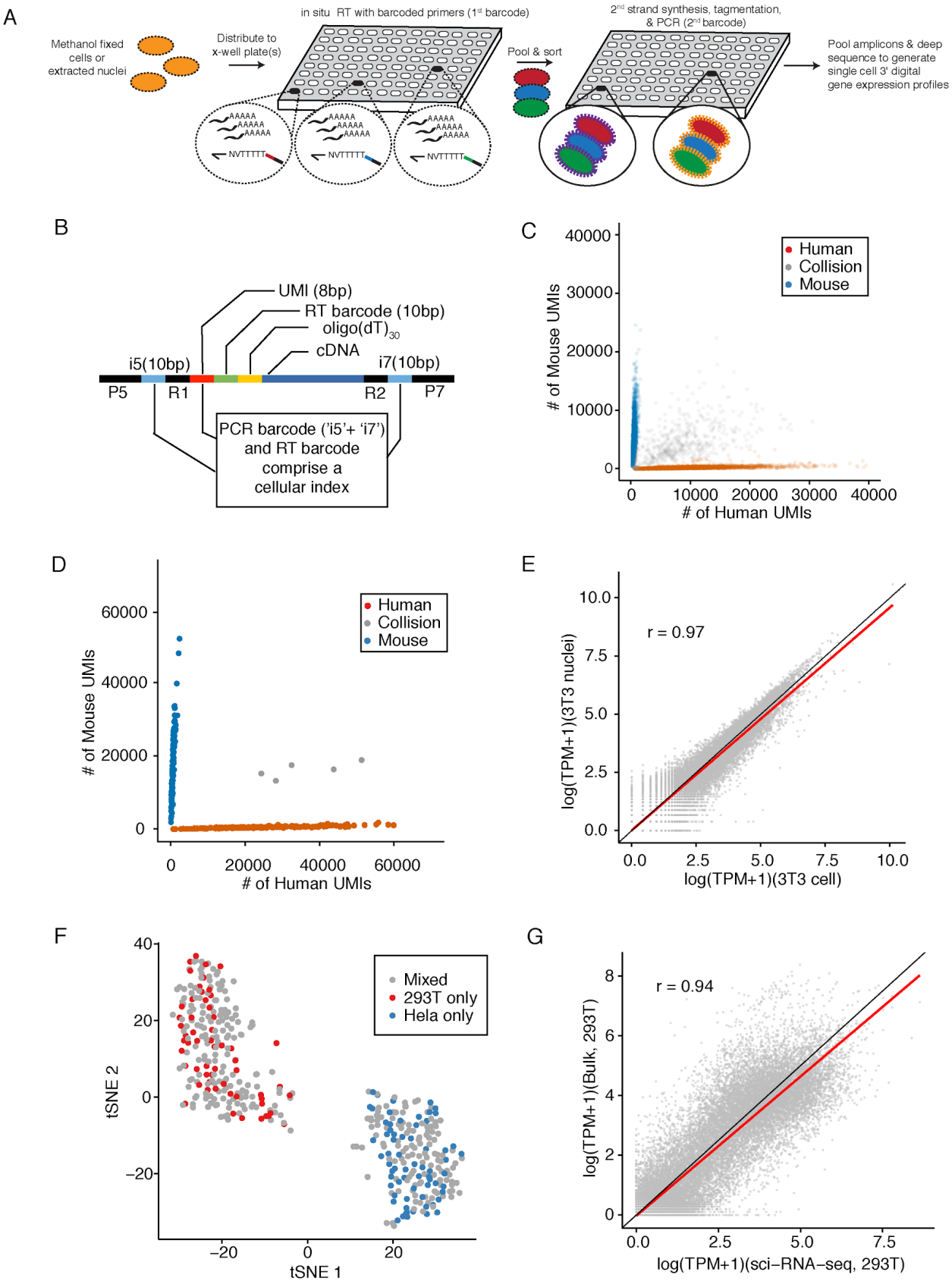

A
Methanol fixed cells or extracted nuclei
Distribute to x-well plate(s)
in situ RT with barcoded primers (1st barcode)
Pool & sort
2nd strand synthesis, tagmentation, & PCR (2nd barcode)
Pool amplicons & deep sequence to generate single cell 3' digital gene expression profiles
B
UMI (8bp)
RT barcode (10bp)
oligo(dT)30
cDNA
i5(10bp)
i7(10bp)
P5
R1
R2
P7
PCR barcode ('i5'+ 'i7') and RT barcode comprise a cellular index
C
Human
Collision
Mouse
# of Mouse UMIs
# of Human UMIs
D
Human
Collision
Mouse
# of Mouse UMIs
# of Human UMIs
E
r = 0.97
log(TPM+1)(3T3 nuclei)
log(TPM+1)(3T3 cell)
F
Mixed
293T only
Hela only
tSNE 2
tSNE 1
G
r = 0.94
log(TPM+1)(Bulk, 293T)
log(TPM+1)(sci−RNA−seq, 293T)

Fig. 1. sci-RNA-seq enables multiplex single cell transcriptome profiling. (**A**) Schematic of sci-RNA-seq workflow. (**B**) Schematic of sci-RNA-seq library amplicons. Index2 and read1 covers the i5 index, UMI and RT barcode. Index1 and read2 covers the i7 index and cDNA fragment. (**C**) Scatter plot of unique human and mouse UMI counts from 384 x 384 sci-RNA-seq. Blue: inferred mouse cells (n = 5953). Red: inferred human cells (n = 3967). Grey: collisions (n = 884). (**D**) Scatter plot of unique human and mouse cell UMI counts from 96 x 96 sci-RNA-seq with optimized protocol. Blue: inferred mouse cells (n = 129). Red: inferred human cells (n = 160). Grey: collisions (n = 5). In (C) and (D), only cells originating from wells containing mixed human and mouse cells are shown. (**E**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of NIH/3T3 cells (n = 238) vs. nuclei (n = 124). (**F**) tSNE plot of cells originating in wells containing HEK293T (red) (n = 60), HeLa S3 (blue) (n = 69) or a mixture (grey) (n = 321). (**G**) Correlation between gene expression measurements from aggregated sci-RNA-seq data vs. bulk RNA-seq data from a related protocol (*33*). (E) and (G) include linear regression (red) and y=x (black) lines.

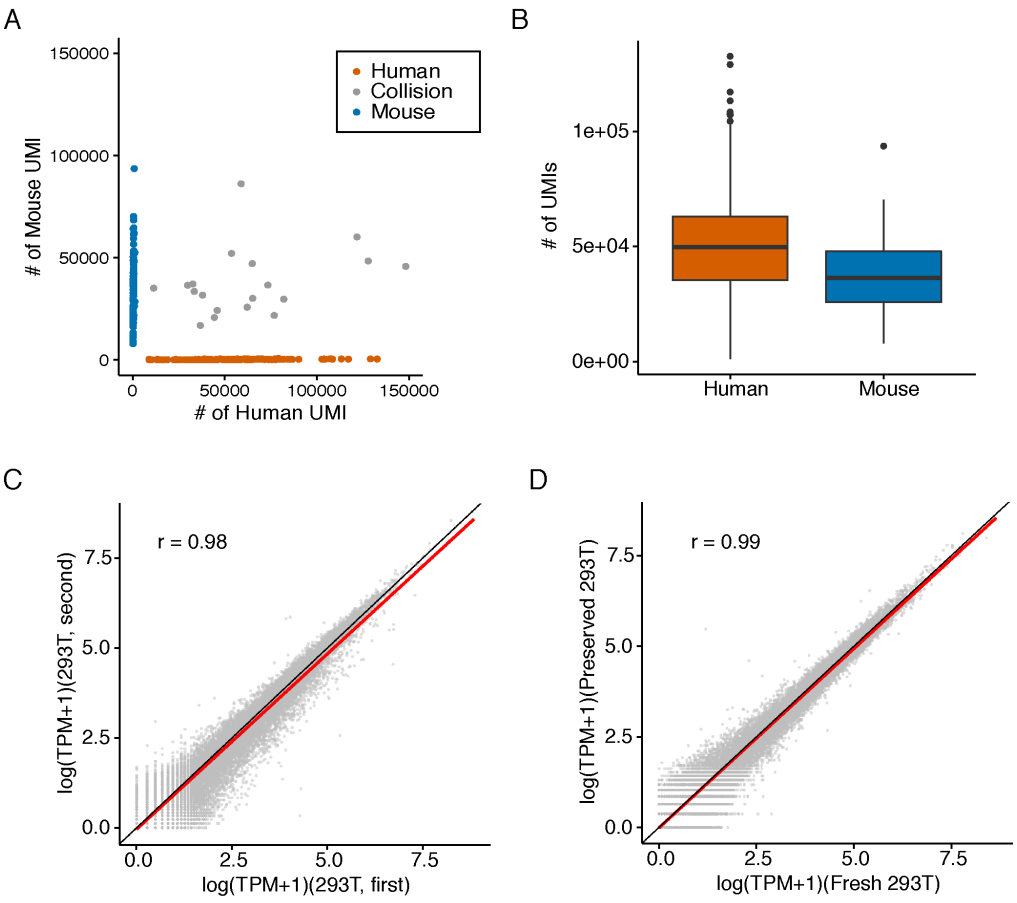


Fig. 2. sci-RNA-seq shows robust gene expression measurements. (**A**) Scatter plot of unique human and mouse UMI counts from a 16 x 84 sci-RNA-seq experiment on mixed HEK293T and NIH/3T3 cells. Blue: inferred mouse cells (n = 109). Red: inferred human cells (n = 168). Grey: collisions (n = 19). (**B**) Boxplots showing number of UMIs detected per cell. (**C**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles from experiments performed two months apart on independently grown and fixed cells. (**D**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of fixed-fresh vs. fixed-frozen cells. (C) and (D) include linear regression (red) and y=x (black) lines.

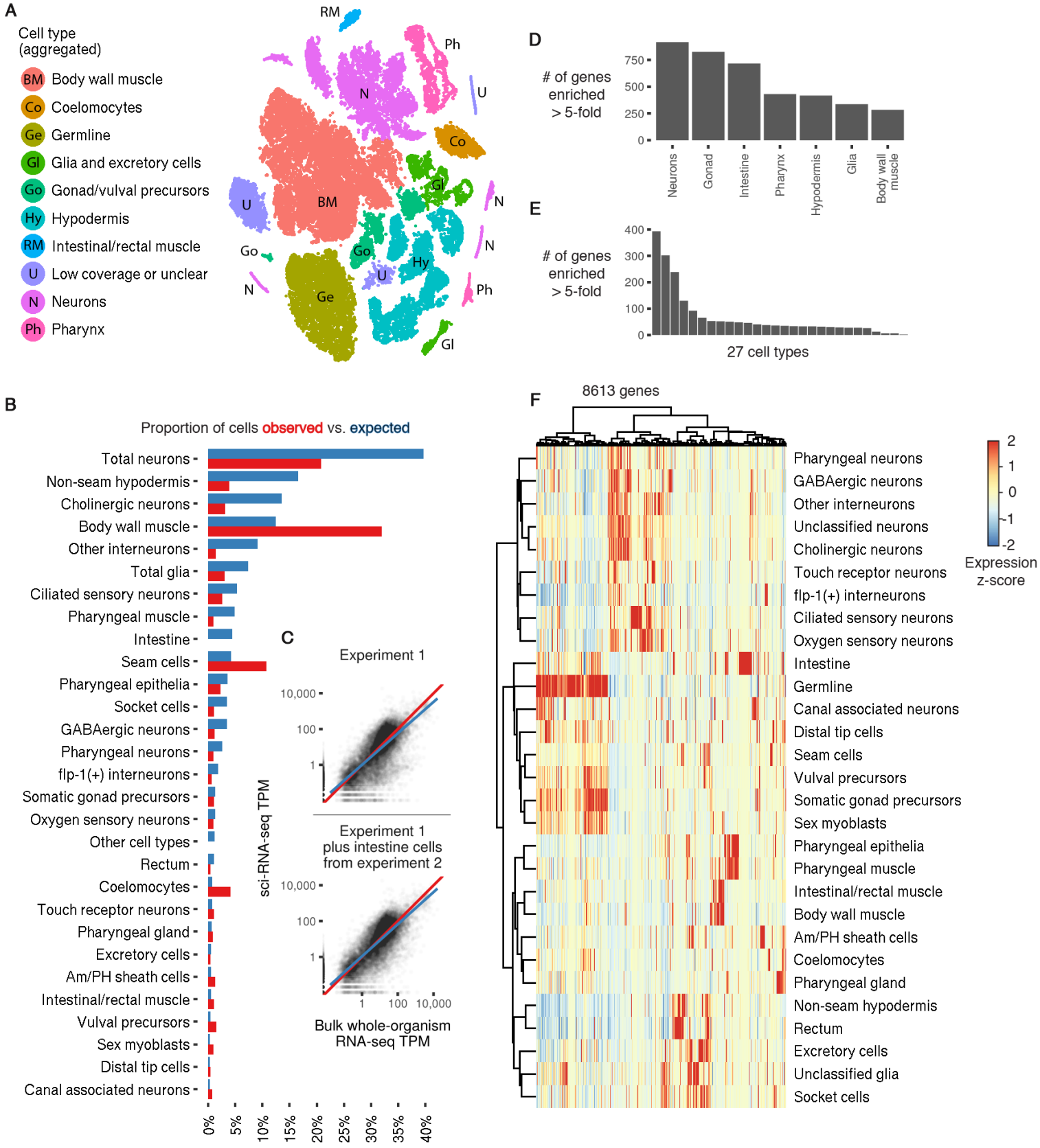

A
Cell type (aggregated)
BM Body wall muscle
Co Coelomocytes
Ge Germline
Gl Glia and excretory cells
Go Gonad/vulval precursors
Hy Hypodermis
RM Intestinal/rectal muscle
U Low coverage or unclear
N Neurons
Ph Pharynx
D
# of genes enriched > 5-fold
Neurons
Gonad
Intestine
Pharynx
Hypodermis
Glia
Body wall muscle
E
# of genes enriched > 5-fold
27 cell types
B
Proportion of cells observed vs. expected
Total neurons
Non-seam hypodermis
Cholinergic neurons
Body wall muscle
Other interneurons
Total glia
Ciliated sensory neurons
Pharyngeal muscle
Intestine
Seam cells
Pharyngeal epithelia
Socket cells
GABAergic neurons
Pharyngeal neurons
flp-1(+) interneurons
Somatic gonad precursors
Oxygen sensory neurons
Other cell types
Rectum
Coelomocytes
Touch receptor neurons
Pharyngeal gland
Excretory cells
Am/PH sheath cells
Intestinal/rectal muscle
Vulval precursors
Sex myoblasts
Distal tip cells
Canal associated neurons
C
Experiment 1
Experiment 1 plus intestine cells from experiment 2
sci-RNA-seq TPM
Bulk whole-organism RNA-seq TPM
F
8613 genes
Expression z-score

Fig. 3. A single sci-RNA-seq experiment highlights the single cell transcriptomes comprising the *C. elegans* larva. (**A**) t-SNE visualization of the high-level cell types identified. (**B**) Bar plot showing the proportion of somatic cells profiled in the first sci-RNA-seq *C. elegans* experiment that could be identified as belonging to each cell type (red) compared to the proportion of cells from that type present in an L2 *C. elegans* individual (blue). (**C**) Scatter plots showing the log-scaled transcripts per million (TPM) of genes in the aggregation of all sci-RNA-seq reads (x axis) or in bulk RNA-seq (y axis; geometric mean of 3 experiments). Top plot includes only the first sci-RNA-seq experiment. Bottom plot also includes intestine cells from the second sci-RNA-seq experiment. (**D**) Number of genes that are enriched at least 5-fold in a specific tissue relative to the 2nd-highest-expressing tissue, excluding genes for which the differential expression between the 1st and 2nd-highest expressing tissues is not significant (q-value > 0.05). (**E**) Same as (D) except comparing cell types instead of tissues. (**F**) Heatmap showing the relative expression of genes in consensus transcriptomes for each cell type estimated by sci-RNA-seq. Genes are included if they have a size-factor-normalized mean expression of >0.05 in at least one cell type (8,613 genes in total). The raw expression data (UMI count matrix) is log-transformed, column centered and scaled (using the R function scale), and the resulting values are clamped to the interval [-2, 2].

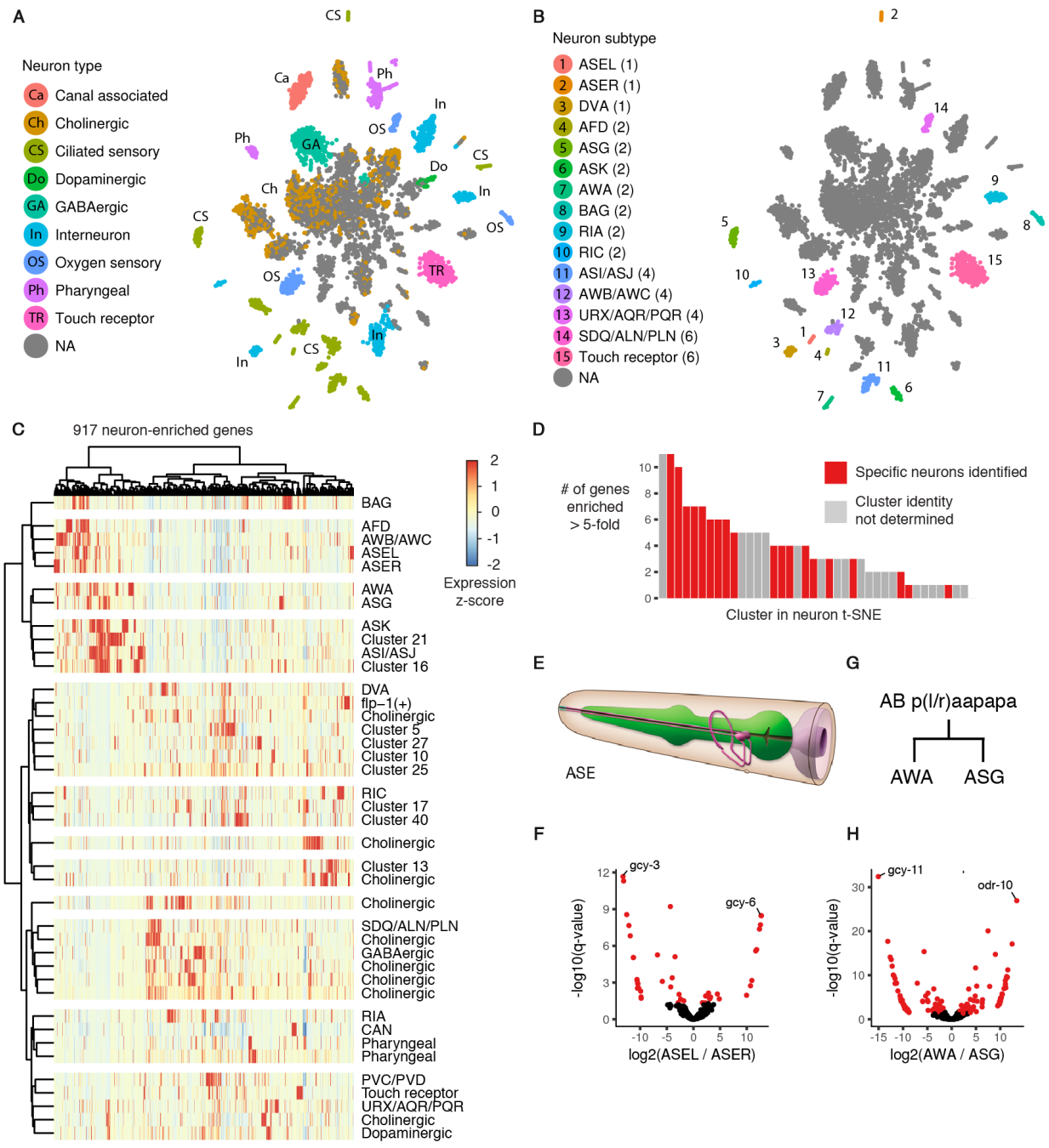


Fig. 4. sci-RNA-seq reveals the transcriptomes of fine-grained anatomical classes of *C. elegans* neurons. (**A**) t-SNE visualization of high-level neuronal subtypes. Cells identified as neurons from the t-SNE clustering shown in Fig. 3A were re-clustered with t-SNE. (**B**) Clusters in the neuron t-SNE that can be identified as corresponding to one, two, or four specific neurons

in an individual *C. elegans* larva. The number of neurons of each type are shown in parentheses. (**C**) Heatmap showing the relative expression of neuron-enriched genes across 40 neuron clusters identified by t-SNE and density peak clustering. Genes are included if their expression in the aggregate transcriptome of all neurons in our data is >5-fold higher than their expression in any other tissue, excluding cases where the differential expression is not significant (q-value > 0.05). (**D**) Distribution for each neuron cluster of the number of genes that are expressed >5-fold higher in that cluster than in the 2nd-highest expressing neuron cluster (q-value for differential expression < 0.05). (**E**) Cartoon illustrating the position of the left and right ASE neurons (pink) relative to the pharynx (green); reproduced with permission from www.wormatlas.org (*60*). (**F**) Volcano plot showing differentially expressed genes between the left and right ASE neurons. Points in red correspond to genes that are differentially expressed (q-value < 0.05) with a > 3-fold difference between the higher- and lower-expressing neuron(s). (**G**) The left AWA and ASG neurons arise from the embryonic cell AB plaapapa; the right AWA and ASG neurons arise from

AB praapapa. (**H**) Volcano plot showing differentially expressed genes between the AWA and ASG neurons.

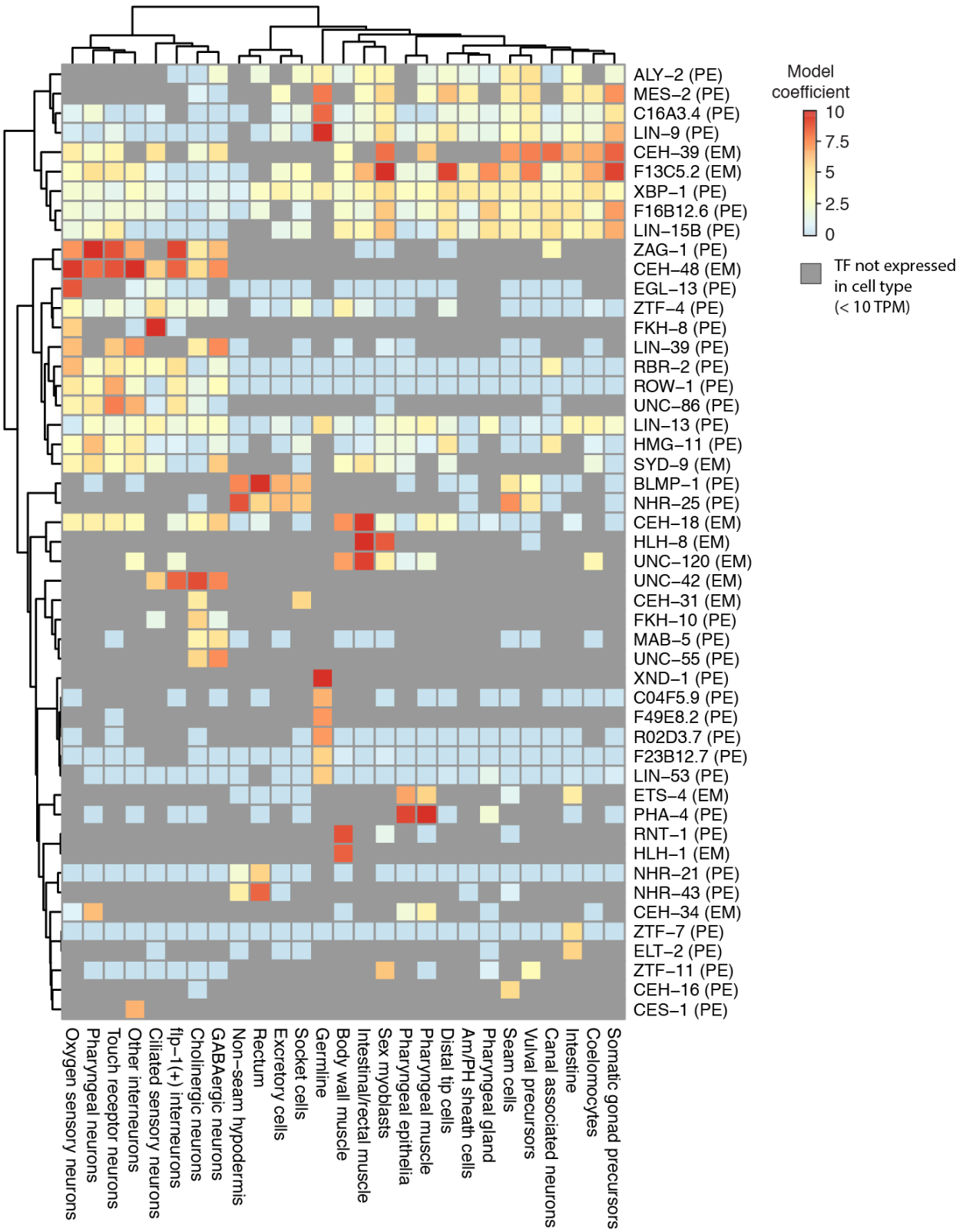

ALY–2 (PE)
MES–2 (PE)
C16A3.4 (PE)
LIN–9 (PE)
CEH–39 (EM)
F13C5.2 (EM)
XBP–1 (PE)
F16B12.6 (PE)
LIN–15B (PE)
ZAG–1 (PE)
CEH–48 (EM)
EGL–13 (PE)
ZTF–4 (PE)
FKH–8 (PE)
LIN–39 (PE)
RBR–2 (PE)
ROW–1 (PE)
UNC–86 (PE)
LIN–13 (PE)
HMG–11 (PE)
SYD–9 (EM)
BLMP–1 (PE)
NHR–25 (PE)
CEH–18 (EM)
HLH–8 (EM)
UNC–120 (EM)
UNC–42 (EM)
CEH–31 (EM)
FKH–10 (PE)
MAB–5 (PE)
UNC–55 (PE)
XND–1 (PE)
C04F5.9 (PE)
F49E8.2 (PE)
R02D3.7 (PE)
F23B12.7 (PE)
LIN–53 (PE)
ETS–4 (EM)
PHA–4 (PE)
RNT–1 (PE)
HLH–1 (EM)
NHR–21 (PE)
NHR–43 (PE)
CEH–34 (EM)
ZTF–7 (PE)
ELT–2 (PE)
ZTF–11 (PE)
CEH–16 (PE)
CES–1 (PE)
Oxygen sensory neurons
Pharyngeal neurons
Touch receptor neurons
Other interneurons
Ciliated sensory neurons
flp–1(+) interneurons
Cholinergic neurons
GABAergic neurons
Non–seam hypodermis
Rectum
Excretory cells
Socket cells
Germline
Body wall muscle
Intestinal/rectal muscle
Sex myoblasts
Pharyngeal epithelia
Pharyngeal muscle
Distal tip cells
Am/PH sheath cells
Pharyngeal gland
Seam cells
Vulval precursors
Canal associated neurons
Intestine
Coelomocytes
Somatic gonad precursors
Model coefficient
10
7.5
5
2.5
0
TF not expressed in cell type (< 10 TPM)

Fig. 5. Cell type specific expression profiles from sci-RNA-seq enable the deconvolution of whole-animal transcription factor ChIP-seq data. For each of 27 cell types, a regularized regression model was fit to predict log-transformed gene expression levels in that cell type on the basis of ChIP-seq peaks in gene promoters (*31*). The ChIP-seq data was generated by the modENCODE (*61*) and modERN consortia (*46*), profiling transcription factor binding in whole *C. elegans* animals. "EM" next to a TF label indicates the ChIP-seq data for the TF is from an embryonic stage, while "PE" indicates the data is from a post-embryonic stage. Colors in the heatmap show the extent to which having a ChIP-seq peak for a given TF in a gene promoter correlates with increased expression in a given cell type. Peaks in "HOT regions" (*31*) are excluded. Grey cells in the heatmap correspond to cases where a TF is not expressed in a cell type (< 10 TPM), in which case ChIP-seq data for that TF is not considered by the regression model.

Supplementary Materials:

Materials and Methods

Figures S1-S24

Supplementary Materials:

Materials and Methods:

Mammalian cell culture

All mammalian cells were cultured at 37°C with 5% CO_2, and were maintained in high glucose DMEM (Gibco cat. no. 11965) supplemented with 10% FBS and 1X Pen/Strep (Gibco cat. no. 15140122; 100U/ml penicillin, 100µg/ml streptomycin). Cells were trypsinized with 0.25% typsin-EDTA (Gibco cat. no. 25200-056) and split 1:10 three times a week.

Generation of whole *C. elegans* cell suspensions

A *C. elegans* strain (RW12139 *stIs11435(unc-120::H1-Wcherry;unc-119(+));unc-119(tm4063)*) carrying an integrated Punc-120::mCherry gene in a wild type background was used in all experiments. A synchronized L2 population was obtained by two cycles of bleaching gravid adults to isolate fertilized eggs allowing the eggs to hatch in the absence of food to generate a population of starved L1 animals. Around 150,000 L1 larvae were plated on each 100 mm petri plate seeded with NA22 bacteria and incubated at 24°C for 15 hr to produce early L2 larvae. Dissociated cells were recovered following a published protocol (*62*) with modification. Specifically, L2 stage worms were collected by adding 10 ml sterile ddH2O to each plate. The collected L2s were pelleted by centrifugation at 1300 g for 1 min. The larval pellet was washed five times with sterile ddH2O to remove bacteria. The resulting pellet was transferred to a 1.6 ml

microcentrifuge tube. Around 40 µl of the final compact pellet was used for each cell dissociation experiment. The worm pellet was treated with 250 µl of SDS-DTT solution (20 nM HEPES pH8, 0.25% SDS, 200 mM DTT, 3% sucrose) for 4 min. Immediately after SDS-DTT treatment, egg buffer (118 mM NaCl. 48 mM KCl. 3 mM CaCl2. 3 mM MgCl2. 5 mM HEPES (pH 7.2) was added to the SDS-DTT treated worms. Worms were pelleted at 500 g for 1 min, then washed 5 times with egg buffer). Pelleted SDS-DTT treated worms were digested with 200 µl of 15 mg/ml pronase (Sigma-Aldrich, St. Louis, MO) for 20 min. The treated worms were broken up to release cells by aspirating up and down through 21G1 ¼ needle. When sufficient single cells were observed the reaction was stopped by adding 900 µl L-15 medium containing 10% fetal bovine serum. Cells were separated from worm debris by centrifuging the pronase-treated worms at 150 g for 5 min at 4°C. The supernatant was transferred to 1.6 ml microcentrifuge tube and centrifuged at 500 g for 5 min at 4°C. The cell pellet was washed twice with egg-buffer containing 1% BSA.

Sample processing

All cell lines were trypsinized, spun down at 300x**g** for 5 min (4°C) and washed once in 1X PBS. *C. elegans* cells were dissociated as described above.

For sci-RNA-seq on whole cells, 5M cells were fixed in 5 mL ice-cold 100% methanol at -20°C for 10 min, washed twice with 1 ml ice-cold 1X PBS containing 1% diethyl pyrocarbonate (0.1% for *C. elegans* cells) (DEPC; Sigma-Aldrich), washed three times with 1 mL ice-cold PBS containing 1% SUPERase In RNase Inhibitor (20 U/µL, Ambion) and 1% BSA (20 mg/ml, NEB). Cells were resuspended in wash buffer at a final concentration of 5000 cells/ul. For all washes, cells were pelleted through centrifugation at 300x**g** for 3 min, at 4°C.

For sci-RNA-seq on nuclei, 5M cells were combined and lysed using 1 mL ice-cold lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2 and 0.1% IGEPAL CA-630 from (*63*)), modified to also include 1% SUPERase In and 1% BSA). The isolated nuclei were then pelleted, washed twice with 1 mL ice-cold 1X PBS containing 1% DEPC, twice with 500 µL cold lysis buffer, once with 500 µL cold lysis buffer without IGEPAL CA-630, and then resuspended in lysis buffer without IGEPAL CA-630 at a final concentration of 5000 nuclei/µL. For all washes, nuclei were pelleted through centrifugation at 300x**g** for 3 min. at 4°C).

For cell-mixing experiments, trypsinized cells were counted and the appropriate number of cells from each cell line were combined prior to fixation or lysis. Fixed cells or nuclei were then distributed into 96- or 384-well plates. For each well, 1,000-10,000 cells or nuclei (2 µL) were mixed with 1 µl of 25 µM anchored oligo-dT primer (5′-ACGACGCTCTTCCGATCTNNNNNNNN[10bp index]TTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3′, where "N" is any base and "V" is either "A", "C" or "G"; IDT) and 0.25 µL 10 mM dNTP mix (Thermo), denatured at 55°C for 5 min and immediately placed on ice. 1.75 µL of first-strand reaction mix, containing 1 µL 5X Superscript IV First-Strand Buffer (Invitrogen), 0.25 µl 100 mM DTT (Invitrogen), 0.25 µl SuperScript IV reverse transcriptase (200 U/µl, Invitrogen), 0.25 µL RNaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen), was then added to each well. Of note, the RT efficiency was affected by the number of cells (or nuclei) per reaction and too many cells (>4,000) per reaction resulted in lower reaction efficiency and higher impurity. For optimized efficiency, we use 2,000 mammalian cells or 5,000 mammalian nuclei per well for RT reaction. Reverse transcription was carried out by incubating plates at 55°C for 10 min, and was stopped by adding 5 µl 2X stop solution (40 mM EDTA, 1 mM spermidine) to each well. All cells (or nuclei) were then pooled, stained with 4',6-diamidino-2-phenylindole (DAPI, Invitrogen) at a final concentration of 3 µM,

and sorted at varying numbers of cells/nuclei per well (depending on experiment) into 5 uL buffer EB using a FACSAria III cell sorter (BD). Cells are gated based on DAPI stain such that singlets are discriminated from doublets and sorted into the each well. 0.5 µl mRNA Second Strand Synthesis buffer (NEB) and 0.25 µl mRNA Second Strand Synthesis enzyme (NEB) were then added to each well, and second strand synthesis was carried out at 16°C for 150 min. The reaction was then terminated by incubation at 75°C for 20 min.

Tagmentation was carried out on double-stranded cDNA using the Nextera DNA Sample Preparation kit (Illumina). Each well was mixed with 5 ng Human Genomic DNA (Promega), as carrier to avoid over-tagmentation and reduce losses during purification, 5 µL Nextera TD buffer (Illumina) and 0.5 µL TDE1 enyzme (Illumina), and then incubated at 55°C for 5 min to carry out tagmentation. Note that because the PCR primers used to amplify libraries are specific to the RT products, tagmented carrier genomic DNA are not appreciably amplified or sequenced. The reaction was then stopped by adding 12 µL DNA binding buffer (Zymo) and incubating at room temperature for 5 min. Each well was then purified using 36 uL AMPure XP beads (Beckman Coulter), eluted in 16 µL of buffer EB (Qiagen), then transferred to a fresh multi-well plate.

For PCR reactions, each well was mixed with 2µL of 10 µM P5 primer (5′-AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′; IDT), 2 µL of 10 µM P7 primer (5′-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3′; IDT), and 20 µL NEBNext High-Fidelity 2X PCR Master Mix (NEB). Amplification was carried out using the following program: 72°C for 5 min, 98°C for 30 sec, 18-22 cycles of (98°C for 10 sec, 66°C for 30 sec, 72°C for 1 min) and a final 72°C for 5 min. After PCR, samples were pooled and purified using 0.8 volumes of AMPure XP beads. Library concentrations were determined by Qubit (Invitrogen) and the libraries were visualized by electrophoresis on a 6% TBE-PAGE

gel. Libraries were sequenced on the NextSeq 500 platform (Illumina) using a V2 75 cycle kit (Read 1: 18 cycles, Read 2: 52 cycles, Index 1: 10 cycles, Index 2: 10 cycles).

sci-RNA-seq with three-level indexing

Cells were harvested and processed for reverse transcription following the same procedure as sci-RNA-seq with two-level indexing. After reverse transcription, each well was mixed with 0.66 μL second strand synthesis buffer (NEB), 0.33 μL second strand synthesis enzyme (NEB), and incubated at 16°C for 2 hours. Cells from all wells were pooled and distributed to a new 96 well plate (4.5 μL per well). 5 μL Nextera TD buffer (Illumina) and 0.5 μL indexed TDE1 enzyme (Illumina) were added to each well. Tagmentation was performed at 55°C for 10 min and stopped by adding 5 μl 2X stop solution (40 mM EDTA, 1 mM spermidine) to each well. All cells (or nuclei) were then pooled, stained with 4',6-diamidino-2-phenylindole (DAPI, Invitrogen) at a final concentration of 3 μM, and sorted at varying numbers of cells/nuclei per well (depending on experiment) into 5 μL buffer (4.6 μL EB buffer, 0.2 μL 1% SDS, 0.2 μL BSA (NEB)) using a FACSAria III cell sorter (BD). Cells are gated based on DAPI stain such that singlets are discriminated from doublets and sorted into the each well. After sorting, each well was mixed with 1 μL of 10 μM P7 primer (5′-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3′, IDT) and incubated at 55°C for 15 min. Then each well was added with 1 μL 10% Tween-20, 1 μL nuclease-free water, 1μL of 10 μM indexed P5 primer (5′-AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′; IDT), and 10 μL NEBNext High-Fidelity 2X PCR Master Mix (NEB). Amplification program and following steps were the same with sci-RNA-seq with two-level indexing.

Read alignments and construction of gene expression matrix

Base calls were converted to fastq format and demultiplexed using Illumina's bcl2fastq/ 2.16.0.10 tolerating one mismatched base in barcodes (edit distance (ED) < 2). Data were processed with GNU Parallel (*64*). Demultiplexed reads were then adaptor clipped using trim_galore/0.4.1 with default settings. Trimmed reads were mapped to the human reference genome (hg19), mouse reference genome (mm10), *C.elegans* reference genome (PRJNA13758) or a chimeric reference genome of hg19, mm10 and PRJNA13758, using STAR/v 2.5.2b (*65*) with default settings and gene annotations (GENCODE V19 for human; GENCODE VM11 for mouse, WormBase PRJNA13758.WS253.canonical_gene set for *C.elegans*). Uniquely mapping reads were extracted, and duplicates were removed using the unique molecular identifier (UMI) sequence (ED < 2, including insertions and deletions), reverse transcription (RT) index, and read 2 end-coordinate (*i.e.* reads with identical UMI, RT index, and tagmentation site were considered duplicates). Finally, mapped reads were split into constituent cellular indices by further demultiplexing reads using the RT index (ED < 2, including insertions and deletions). For mixed-species experiment, the percentage of uniquely mapping reads for genomes of each species was calculated. Cells with over 85% of UMIs assigned to one species were regarded as species-specific cells, with the remaining cells classified as mixed cells or "collisions". The collision rate was calculated as twice the ratio of mixed cells (as we are blind to collisions involving cells of the same species). For gene body coverage analysis of exonic reads, the split human and mouse single cell SAM files were concatenated and exonic reads were selected and analyzed using RSEQC/2.6.1, using BED annotation files downloaded from the UCSC Golden Path. For read position analysis for intronic reads, the split human and mouse single cell SAM files were concatenated and intronic reads were selected; the fractional position of each intronic read along the genomic distance between the TSS and transcript terminus was calculated, and these values used to generate a density plot.

To generate digital expression matrices, we calculated the number of strand-specific UMIs for each cell mapping to the exonic and intronic regions of each gene with python HTseq package (*66*). Generally, fewer than 3% of total UMIs strand-specifically mapped to multiple genes. For multi-mapped reads, reads were assigned to the closest gene, except in cases where another intersected gene fell within 100 bp to the end of the closest gene, in which case the read was discarded. For most analyses we included both expected-strand intronic and exonic UMIs in per-gene single-cell expression matrices.

For sci-RNA-seq with three-level indexing, reads were analyzed with the same procedure, except that RT index was combined with Tn5 index, and thus the mapped reads were split into constituent cellular indices by demultiplexing reads using both the RT index and Tn5 index (ED < 2, including insertions and deletions).

t-SNE visualization of HEK293T cells and HeLa S3 cells

We visualized the clustering of sci-RNA-seq data from populations of pure HEK293T, pure HeLa S3 and mixed HEK293T + HeLa S3 cells using t-Distributed Stochastic Neighbor Embedding (t-SNE). Cells with more than 100,000 UMIs were discarded. The top 3,000 genes with the highest variance in the digital gene expression matrix for these cells were first given as input to Principal Components Analysis (PCA). The top 10 principal components were then used as the input to t-SNE, resulting in the two-dimensional embedding of the data shown in Fig. 1F. The process was repeated using only intronic reads (fig. S4C). For this analysis, the top 2,000 (instead of 3,000) highly variable genes were used as input to PCA; all other parameters remained unchanged.

Genotyping of single HeLa cells by 3’ tag sequences

HeLa S3 cell identity was verified on the basis of homozygous alleles not present in the hg19 assembly, using a callset derived from (*67*). Single-cell BAM files (with cellular indices encoded in the "read_id" field) were concatenated, and then processed as follows using a python wrapper of the samtools API (*i.e.* pysam). For each homozygous alternate SNV overlapping with a GENCODE V19 defined gene (n = 865,417) in the HeLa S3 variant callset, we computed the fraction of matching (*i.e.* HeLa S3 specific) alleles, and computed this value for all cells where at least 1 read containing a polymorphic site. We then re-plotted in R the tSNE visualization shown in fig. S4B, now colored by the relative fraction of homozygous alternate alleles called for each cell.

Comparing sci-RNA-seq and bulk RNA-seq data for HEK293T cells

To compare aggregated sci-RNA-seq single cell transcriptomes with bulk RNA-seq, we performed bulk RNA-seq using a modified protocol (*33*). In brief, 500 ng total RNA extracted from three biological replicate HEK293T samples (extraction using RNeasy kit (Qiagen)) with the RNeasy kit (Qiagen) were used for reverse transcription following the standard SuperScript II protocol. 500 ng total RNA (in 9 µL water) was mixed with 2 µL 25 uM oligo-dT(VN) (5′-ACGACGCTCTTCCGATCTNNNNNNNN[10bp index]TTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3′, where "N" is any base and "V" is either "A", "C" or "G"; IDT) and 1 µL 10 mM dNTPs, then incubated at 65°C for 5 min Following incubation, 8 µL reaction mix (4 µL 5X Superscript II First-Strand Buffer, 2 µl 100 mM DTT, 1 µl SuperScript II reverse transcriptase, 1 µL RnaseOUT) was added. Reactions were incubated at 42°C for 50 min and terminated at 70°C for 15 min. For second strand synthesis, 2 µL RT product was mixed with 6.5 µL water, 1 µL mRNA Second Strand Synthesis buffer (NEB) and 0.25 µl mRNA Second Strand Synthesis enzyme (NEB). Second strand synthesis was carried out at 16°C for 150 min, followed by 75°C for 20 min. Tagmentation was carried out by

adding 10 μL Nextera TD buffer, 1 μL Nextera Tn5 enyzme and incubating at 55°C for 5 min. Tagmented cDNA was purified using a Clean & Concentrator^{TM}-100 kit (Zymo) and eluted in 16 uL buffer EB. PCR, purification, and quantification were then performed as detailed above.

For comparing single cell RNA-seq and bulk RNA-seq, single cell gene counts of exonic reads and intronic reads were added for the same gene from sci-RNA-seq of pure HEK293T cells as well as HEK293T cells identified from HEK293T and NIH/3T3 mixed cells. Counts for bulk RNA-seq of HEK293T cells were extracted based on the RT barcode and aggregated separately, again adding exonic and intronic read counts per gene. Transcript counts were converted to transcripts per million (TPM) and then transformed to log(TPM + 1). Pearson correlation coefficients were calculated between the aggregated sci-RNA-seq and bulk RNA-seq data using R.

Analysis of *C. elegans* whole-organism sci-RNA-seq experiments

Both *C. elegans* sci-RNA-seq experiments were processed identically except as noted. A digital gene expression matrix was constructed from the raw sequencing data as described above. Cells with UMI count for protein-coding genes < 100 (experiment 1) or < 200 (experiment 2; higher threshold to compensate for slightly more leakage between cells) were excluded from the analysis. The dimensionality of this matrix was reduced first with PCA (40 components) and then with t-SNE, giving a two-dimensional representation of the data. This t-SNE was performed using the implementation in Monocle version 2.3.5 (*68*). Similar to the approach in (*69*), cells in this two-dimensional representation were clustered using the density peak algorithm (*70*) as implemented in Monocle 2.3.5. Genes specific to each cluster were identified and compared to microscopy-based expression profiles reported in the literature (fig. S15-23), allowing the distinct cell types represented in each cluster to be identified. Based on these results, in experiment 1, we manually merged two clusters that both corresponded to body wall muscle, and

manually split two clusters that included hypodermis, somatic gonad cells, and glia. Seven clusters exclusively contained neurons. We identified neuronal subtypes applying PCA, t-SNE, and density peak clustering to this subset of cells using the same approach as for the global cluster analysis.

In addition to neurons, body wall and intestinal/rectal muscle cells, pharyngeal cells, hypodermal cells, glial cells, intestinal cells (from experiment 2), gonad cells, and coelomocytes were each independently sub-clustered. Clusters from these iterative t-SNE analyses that featured expression of marker genes from multiple tissues were identified as likely doublets. These cells, which comprised ~2.5% of the total, were excluded from all downstream analyses.

Consensus expression profiles for each cell type except intestine were constructed by first dividing each column in the gene-by-cell digital gene expression matrix for experiment 1 by the cell's size factor and then for each cell type, taking the mean of the normalized UMI counts for the subset of cells assigned to that cell type. These mean normalized UMI counts were then re-scaled to transcripts per million. Cells that had a UMI count of less than one quarter of the median for their assigned cell type were excluded from the consensus expression profiles. The intestine consensus expression profile was generated in the same manner, but used cells from experiment 2 instead of experiment 1.

95% confidence intervals for the mean expression of each gene in each cell type were estimated using a normal approximation to the negative binomial distribution. For each cell type, the expression of a given gene was assumed to follow a negative binomial distribution, with a mean μ and dispersion parameter α estimated using Monocle's estimateDispersions function (using only cells of that particular cell type). The variance of this random variable is equal to $\mu + \mu^2\alpha$. By the central limit theorem, the values of the estimate for the mean will asymptotically approach a distribution $N(\mu, (\mu + \mu^2\alpha) / n)$, where n is the number of cells of the cell type in

question. Confidence intervals for the true value of μ are computed based on this normal approximation.

Genes with expression patterns highly enriched in a single tissue were identified as follows. For each gene (excluding those expressed in fewer than 10 cells), the tissue in which it is expressed highest and the tissue in which it is expressed second-highest (relative to other tissues) are enumerated. The gene is considered enriched in the highest expressing tissue if it is both expressed at a >5-fold greater level than in the second-highest expressing tissue and the differential expression of this gene between the highest and second-highest expressing tissues is non-zero at a false detection rate of < 5%. The differential expression tests are performed with the differentialGeneTest function of Monocle 2 (*68*). The false detection rates are computed based on the tests for all genes, not just the genes with a given highest/second-highest expressing tissue. Genes with expression patterns enriched in a single cell type or a single neuron cluster were identified using the same method (*i.e.* comparing the highest and second-highest expressing cell type instead of tissue).

Differential expression tests for analyses presented in Fig. 4F,H and fig. S10B,D,F were also conducted using the differentialGeneTest function of Monocle 2, excluding genes expressed in fewer than 10 cells total among the cell types being compared (*e.g.* when comparing the ASEL vs. ASER neurons, genes are considered if they are expressed in at least 10 ASEL/R cells).

Integration of sci-RNA-seq expression profiles and modENCODE (*61*)/modERN (*46*) ChIP-seq data

Transcription factor (TF) ChIP-seq datasets were downloaded from the ENCODE data portal. The ChIP-seq data included experiments conducted on whole embryos or whole larvae at different developmental stages. ChIP peaks for the same TF were merged if they overlapped and

were either both from an embryonic stage experiment or both from a post-embryonic stage experiment. If a TF had both embryonic and post-embryonic data available, only the post-embryonic data was used.

A ChIP-seq peak was considered to be associated with a gene if: 1) the peak summit was within 2 kb of the canonical transcription start site (TSS) for the gene, 2) the distance from the peak summit to the second closest TSS (regardless of strand) was at least 50% greater than the distance to the closest TSS, and 3) the peak overlapped peaks for < 20% of assayed TFs from the same broad developmental stage (embryonic or post-embryonic). This excludes so-called "HOT regions" which are likely to reflect either non-sequence-specific TF binding or an artifact of the ChIP-seq assay (*71*).

Each gene-associated ChIP-seq peak is assigned a score equal to 0.2 minus the proportion of assayed TFs from the same broad developmental stage (embryonic or post-embryonic) that have peaks which overlap the peak in question. This serves to further down-weight peaks in marginally HOT regions. Each gene is assigned a score for each TF that is equal to the maximum peak score of all peaks for the TF that are assigned to the gene (or zero, if no such peaks exist). These scores are referred to as "TF association scores" below.

For each of the 27 cell types with sci-RNA-seq consensus expression profiles, a regression model was constructed to predict the expression levels of genes in the given cell type based on the TF association scores for each individual gene. The response in these models was log2(transcripts per million + 1) for each gene. The features are the TF association scores for each gene; however, only scores for TFs that are expressed with at least 10 transcripts per million in the cell type in question are included as features. The models are fit using elastic net regularization as implemented in the R package *glmnet*. Model coefficients shown in Fig. 5 are from models fit with the largest regularization parameter that gives a mean squared error (MSE)

less than 1 standard error from the MSE of a model with the optimal regularization parameter, as inferred by cross validation ("lambda.1se").

To identify pairs of TFs that have co-localized binding patterns more often than could be expected by chance (fig. S13), peaks were first clustered by recursively merging those with summits within 150 bp of each other. This analysis was limited to TFs with ChIP-seq data from post-embryonic worms, and also included germline-specific ChIP-seq for EFL-1 and DPL-1 produced by (*57*). Peak clusters that contained peaks for >20% of the TFs ("HOT regions") were excluded from further analysis. Peak clusters were associated with genes using the same criteria as used for individual peaks (described above, treating the midpoint of the cluster's genomic interval as the "summit" of the cluster). Peak clusters that could not be associated with a gene were excluded from further analysis. From the remaining peak clusters, a matrix was constructed where the rows are identifiers for each peak cluster and the columns are binary variables with value 1 if the cluster includes at least one peak for a given TF, 0 otherwise.

This matrix was used as input to the Graphical LASSO (*72*), an algorithm which provides robust estimates of partial correlations between a set of random variables given a limited number of observations and under the assumption that most variables are conditionally independent from another. In this context, the partial correlation between two columns of the input matrix is equal to the correlation of the events ">0 peaks for TF 1 are present in this peak cluster" and ">0 peaks for TF 2 are present in this peak cluster", conditioned on the presence or absence of peaks for all other TFs. The Graphical LASSO was applied to either the full matrix (fig. S13D) or the subset of rows in the matrix that corresponded to peak clusters in the promoters of gonad-enriched genes (fig. S13A) or neuron-enriched genes (fig. S13C). From the partial correlations outputted by each Graphical LASSO, we constructed a network where the nodes are TFs (columns in the

matrix) and undirected edges connect each pair of TFs for which the partial correlation in either direction (TF 1 → TF 2 or TF 2 → TF 1) is > 0.01.

The Graphical LASSO model requires a regularization parameter to be set by the user, with increasing values. We set this parameter to the smallest value that satisfied the requirement that the probability that a non-zero partial correlation in the output is in fact zero in the "true" model—the false detection rate—is less than 5%. To find a mapping between regularization parameter values and the false detection rate, we constructed a null model by shuffling the values of the input matrix in a manner that preserves both row and column sums, using the CurveBall algorithm (*73*). In a shuffled matrix, all non-zero partial correlations reported by the Graphical LASSO are false detections. We therefore estimate the false detection rate of a given regularization parameter value to be equal to the mean number of non-zero partial correlations reported by the Graphical LASSO for shuffled matrices (averaging over 50 shuffles) divided by the number of non-zero partial correlations reported by the Graphical LASSO on the unshuffled input data.

Cost estimation

Using the 576 x 960 sci-RNA-seq experiment as an example, reagent costs are largely enzyme-driven and include SuperScript IV reverse transcriptase ($934), second strand synthesis mix ($750), Nextera Tn5 enzyme ($5,000), NEBnext master mix ($1,150), FACS sorting ($250) and other reagents and plates ($250). If we sort 60 cells per well (assuming recovery rate is 100%) for 960 wells (5% collision rate), then the reagent cost of library preparation is around $0.14 per cell (expected yield of around 55,000 cells). However, it is worth noting that simply increasing the number of cells sorted per well decreases costs (*e.g.* sorting 150 cells to each well would yield around 140,000 cells at a cost of $0.05 per cell), but also results in an increased collision rate (12%). Alternatively, by increasing to 1,536 barcodes during the first (RT-based)

round of indexing, we can sort up to 320 cells per well at a 10% collision rate, thereby reducing the cost per cell to less than $0.025 per cell. Straightforward reductions in reaction volumes and/or in-house enzyme production at all steps may also lead to further reductions in costs, as would additional rounds of molecular indexing. For example, with 384 x 384 x 384 combinatorial indexing, we can potentially uniquely barcode the transcriptomes of around 12 million cells at a 10% collision rate, corresponding to >200-fold increase in detection capacity relative to the 576 x 960 experiment, without much increase in reagent costs.

Supplementary Figures

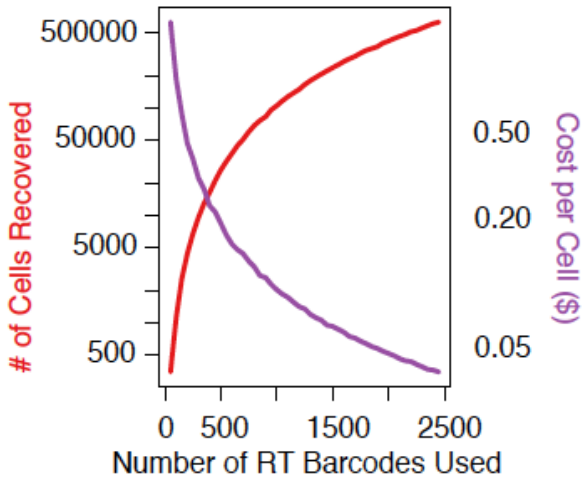


Fig. S1

Combinatorial indexing with increasing numbers of reverse transcription (RT) barcodes enables sublinear scaling of cost per cell. Plot assumes two-level indexing and estimates how detection capacity (i.e. the number of cells detected in a sci-RNA-seq experiment, red) and cost per cell (blue) vary as a function of the number of RT barcodes used, assuming a collision rate of 5%.

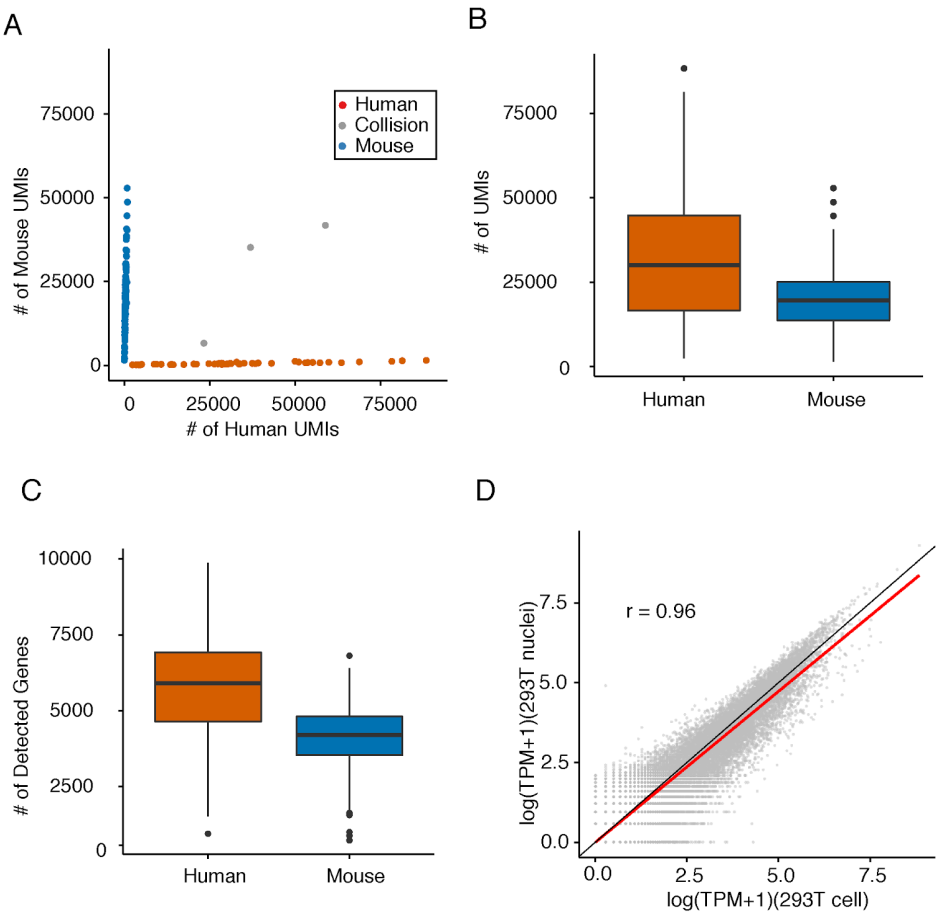


Fig. S2

sci-RNA-seq is compatible with isolated nuclei as starting material. (**A**) Scatter plot of unique human and mouse nuclei UMI counts from a 96 x 96 sci-RNA-seq experiment. This experiment included different cell populations, but only cells originating from a mixture of human (HEK293T) and mouse (NIH/3T3) nuclei are plotted here. Inferred mouse cells (n = 124) are colored in blue; inferred human cells (n = 48) are colored in red, and "collisions" (n = 3) are colored in grey. (**B** to **C**) Boxplots showing the number of UMIs (B) and genes (C) detected per cell in nuclear sci-RNA-seq experiments. (**D**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of HEK293T cells (n = 328) vs. HEK293T nuclei (n = 48), together with a linear regression line (red) and y=x line (black).

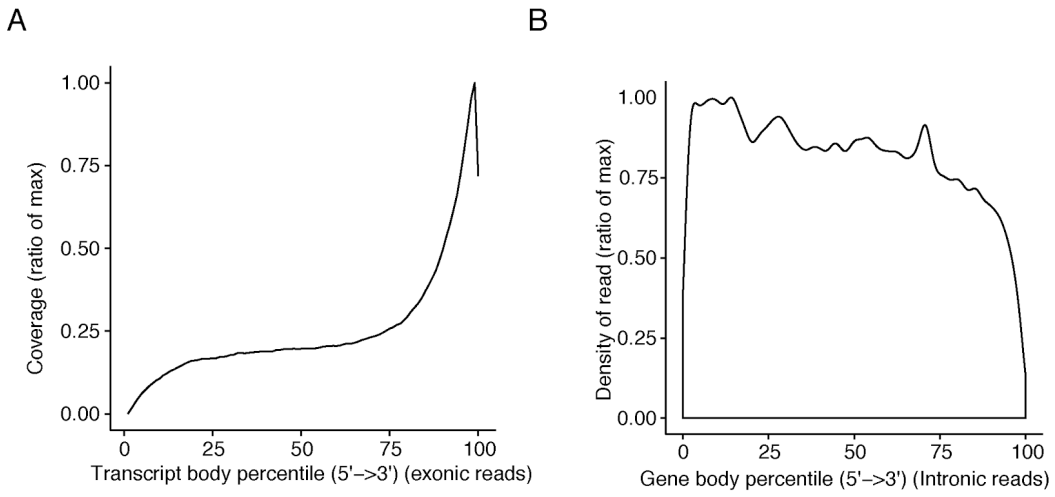


Fig. S3

Positional bias of exonic and intronic sci-RNA-seq reads. (**A**) Density plot showing that as expected, sci-RNA-seq reads mapping to exons are strongly biased to originate near the 3' ends of transcripts (intronic regions excluded from percentile scaling). (**B**) Density plot showing that in contrast, sci-RNA-seq reads mapping to introns do not exhibit 3' bias (intronic regions included in percentile scaling). Y-axis is scaled to the ratio of max.

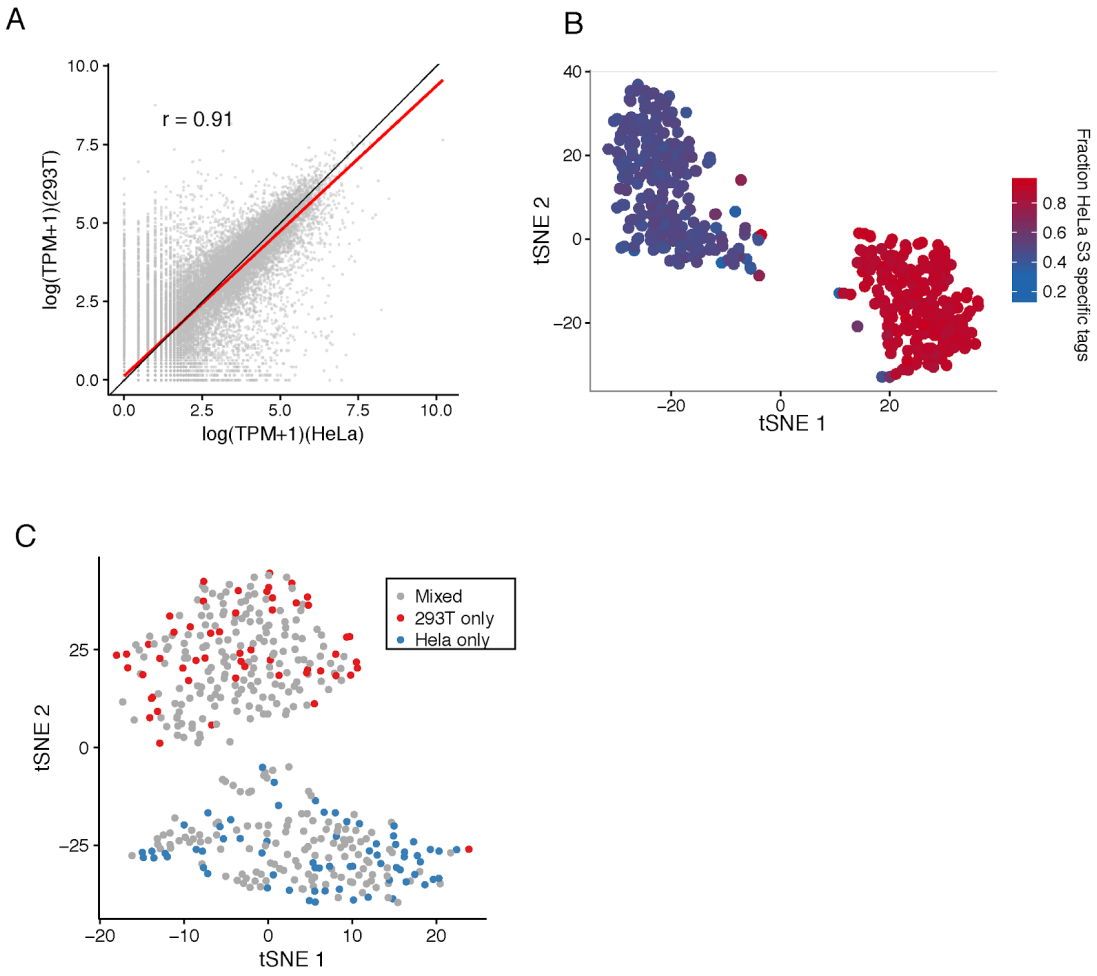


Fig. S4

Quality control for sci-RNA-seq on mixed populations of HeLa S3 and HEK293T cells. (**A**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of HeLa S3 vs. HEK293T cells, together with a linear regression line (red) and y=x line (black). (**B**) tSNE plot (as in Fig. 1F), with cells colored by fraction of reads harboring HeLa S3 specific SNVs (single nucleotide variants) relative to hg19 assembly. (**C**) tSNE using digital gene expression matrices constructed from only intronic reads. Cells are colored by the population from which they derived, with pure HEK293T in red, pure HeLa S3 in blue, and mixed cells in grey.

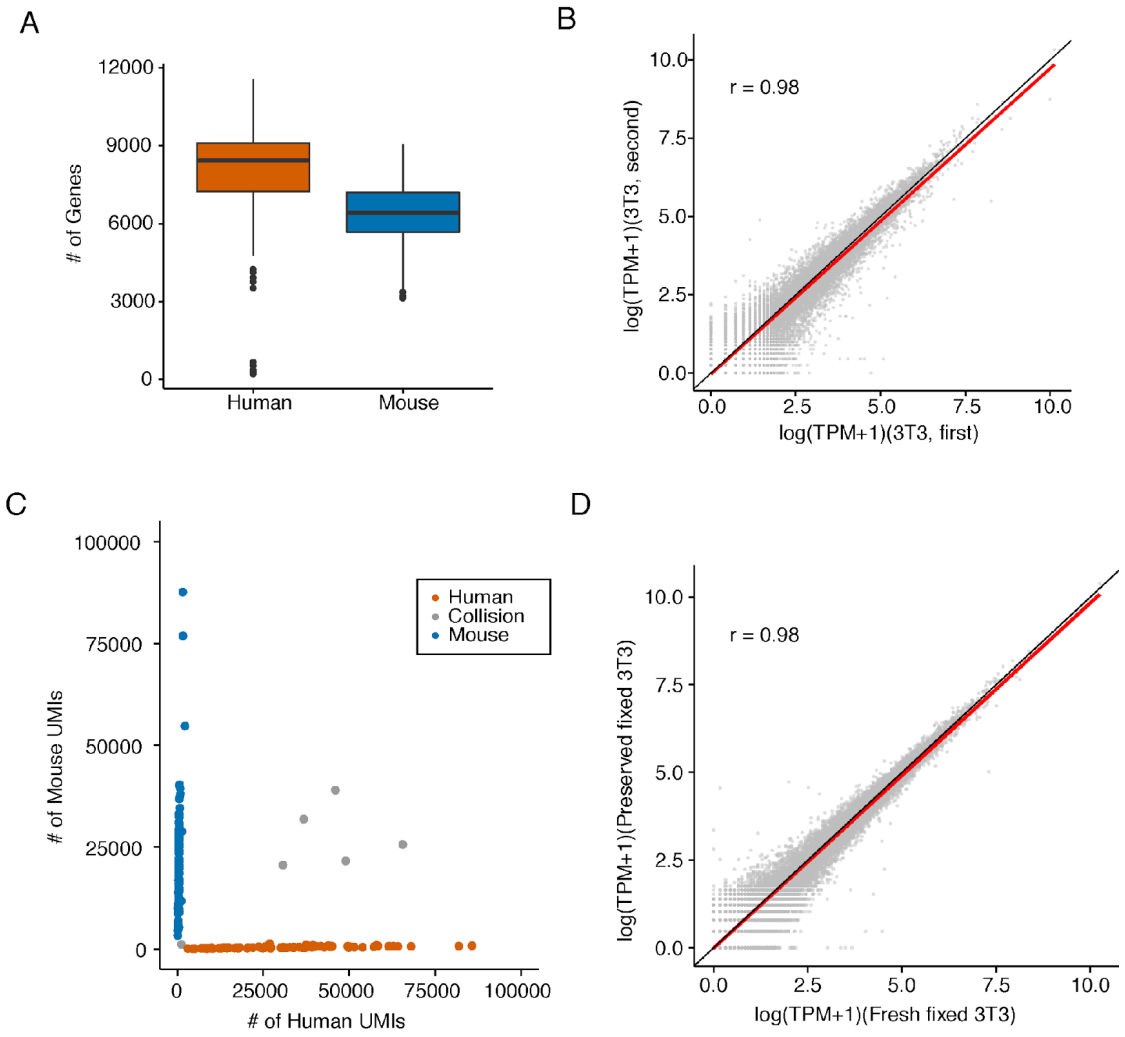


Fig. S5

sci-RNA-seq shows robust gene expression measurements. (**A**) Boxplots showing the number of genes detected per cell in a 16 x 84 well sci-RNA-seq experiment. (**B**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of NIH/3T3 cells from two sci-RNA-seq experiments, performed two months apart and on independently grown and fixed cells, together with a linear regression line (red) and y=x line (black). (**C**) Scatter plot of unique human and mouse UMI counts from a 16 x 84 sci-RNA-seq experiment on mixed HEK293T and

NIH/3T3 cells after methanol fixation and freezing at -80°C for 4 days. Inferred mouse cells (n = 90) are colored in blue; inferred human cells (n = 89) are colored in red, and "collisions" (n = 6) are colored in grey. (**D**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of fixed-fresh vs. fixed-frozen NIH/3T3 cells, together with a linear regression line (red) and y=x line (black).

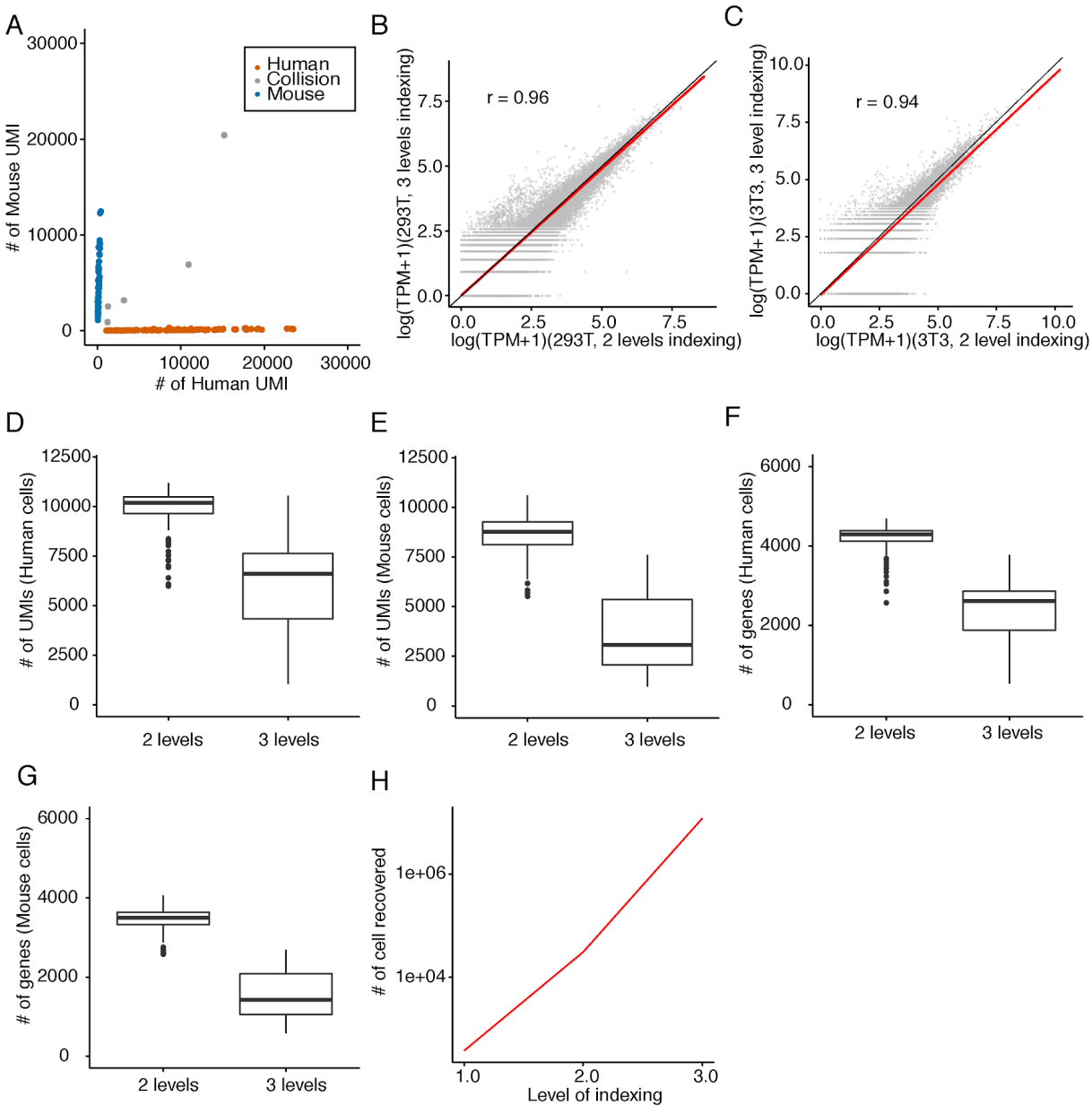


Fig. S6

Representative result from sci-RNA-seq with 3-level indexing. (**A**) Scatter plot of unique human and mouse UMI counts from a 16 x 6 x 16 sci-RNA-seq experiment on mixed HEK293T and NIH/3T3 cells. Inferred mouse cells (n = 62) are colored in blue; inferred human cells (n = 119) are colored in red, and "collisions" (n = 5) are colored in grey. (**B**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of HEK293T cells with 2-level vs. 3-level indexing, together with a linear regression line (red) and y=x line (black). (**C**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of NIH/3T3 cells in sci-RNA-seq with 2-level vs. 3-level indexing, together with a linear regression line (red) and y=x line (black). (**D** to **E**) Boxplots showing the number of UMIs detected per HEK293T cell (D) and NIH/3T3 cell (E) in sci-RNA-seq with 2-level or 3-level indexing, sampling 15,000 total reads per cell. (**F** to **G**) Boxplots showing the number of genes detected per HEK293T cell (F) and NIH/3T3 cell (G) in sci-RNA-seq with 2-level or 3-level indexing, sampling 15,000 total reads per cell. (**H**) Plot illustrating how estimated detection capacity (*i.e.* the number of cells detected in a sci-RNA-seq experiment, red) varies as a function of number of rounds of indexing used, assuming a collision rate of 10% and 384 indexes at each level.

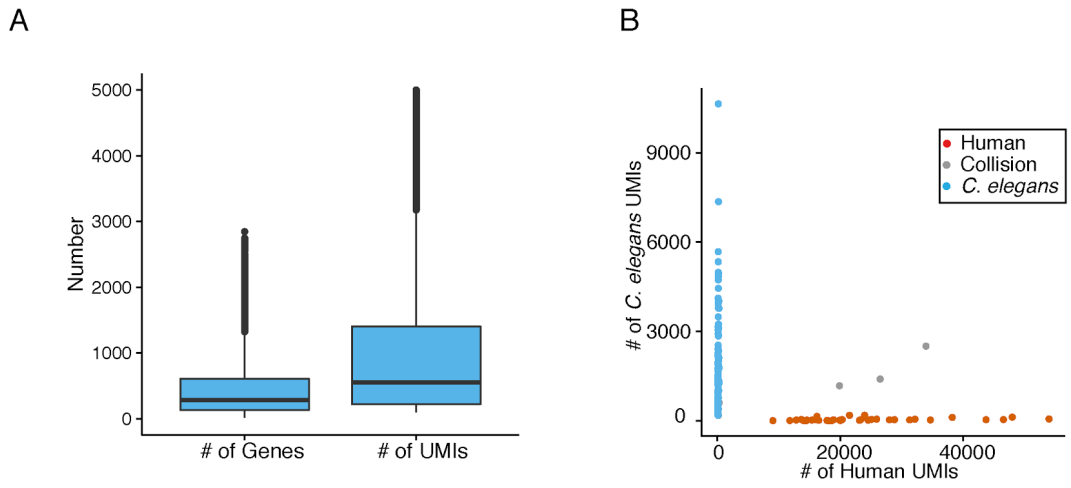


Fig. S7

Quality control metrics for *C. elegans* sci-RNA-seq experiments. (**A**) Distribution of number of protein-coding genes and UMI counts (mapping to protein-coding genes) detected per *C. elegans* cell. (**B**) Scatter plot of unique UMI counts per cell from a sci-RNA-seq experiment performed on mixture of HEK293T (human) and *C. elegans* cells.

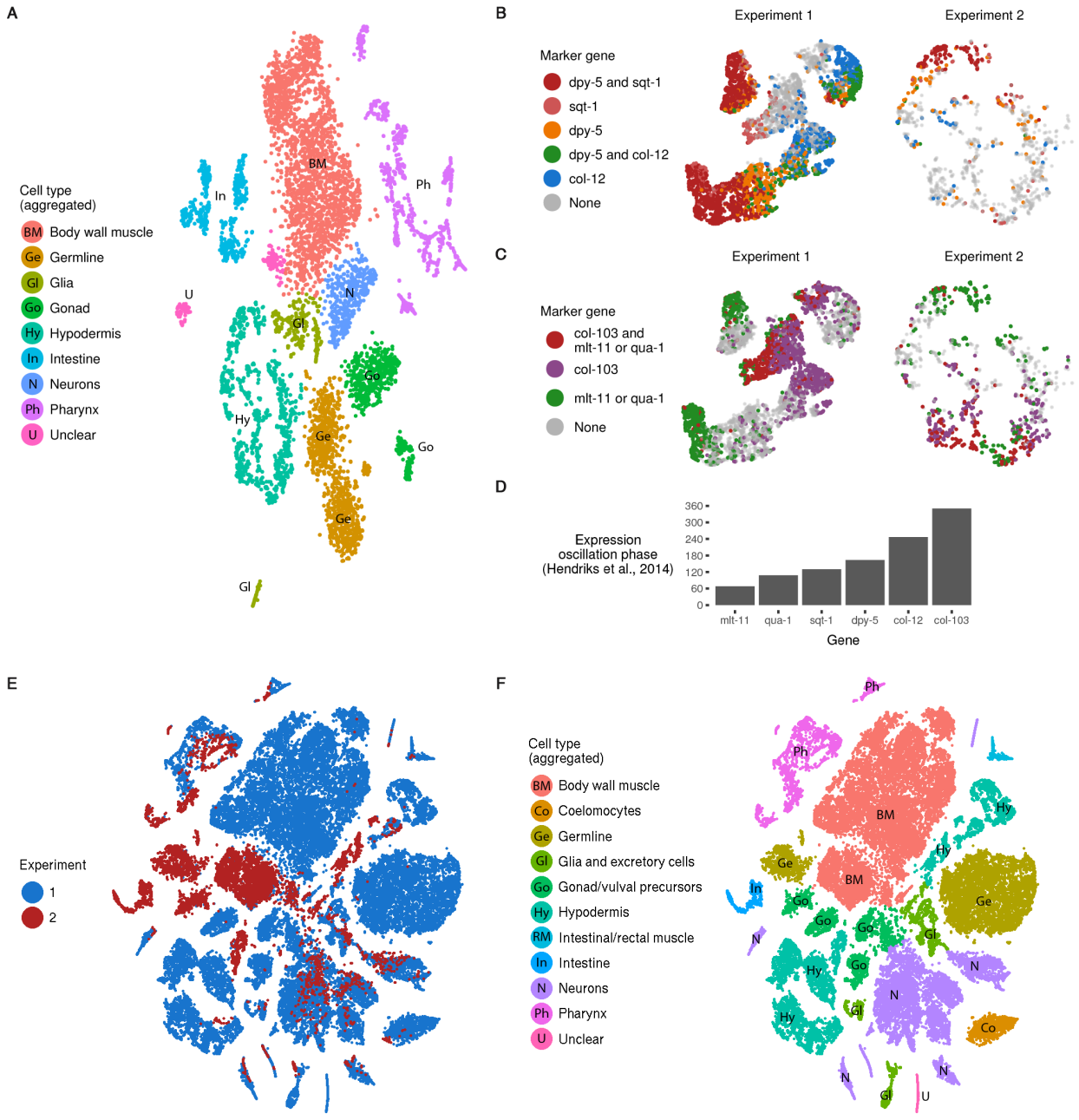


Fig. S8

A second *C. elegans* sci-RNA-seq experiment recovers intestine cells. (**A**) t-SNE visualization of cells from the second *C. elegans* experiment, which included all cells (96 wells) or only cells with high DAPI stain (48 wells). 511 intestine cells were successfully recovered. (**B**) Expression of the cuticle collagens *dpy-5*, *sqt-1*, and *col-12* in cells from experiments 1 and 2. t-SNE

coordinates for cells are the same as in Fig. 3A (for experiment 1) and (A) (for experiment 2), but only hypodermal cells are shown. *dpy-5* and *sqt-1* are expressed during the synthesis of new cuticle preceding each larval molt, while *col-12* is expressed during molting and ecdysis (*74*). (**C**) Expression of the signaling gene *qua-1*, the protease inhibitor *mlt-11*, and the collagen *col-103*, in experiments 1 and 2. *qua-1* and *mlt-11* are expressed at the initiation of new cuticle synthesis (*75*). *col-103* is expressed in the intermolt, after ecdysis but before new cuticle synthesis begins (*36*). Taken together with (B), the expression patterns suggest that the worms in experiment 1 spanned a range of developmental sub-stages from late L2 to around the L3 molt, while worms from experiment 2 had greater synchrony and were mostly from the early L2 stage. (**D**) Phase of the molting-cycle associated gene expression oscillations of selected genes, as reported by (*36*)**.** The values are modulo 360, *i.e.* 360 is the same as 0 and equidistant from 90 and 270**.** (**E** to **F**) t-SNE visualizations of cells from both *C. elegans* experiments processed together.

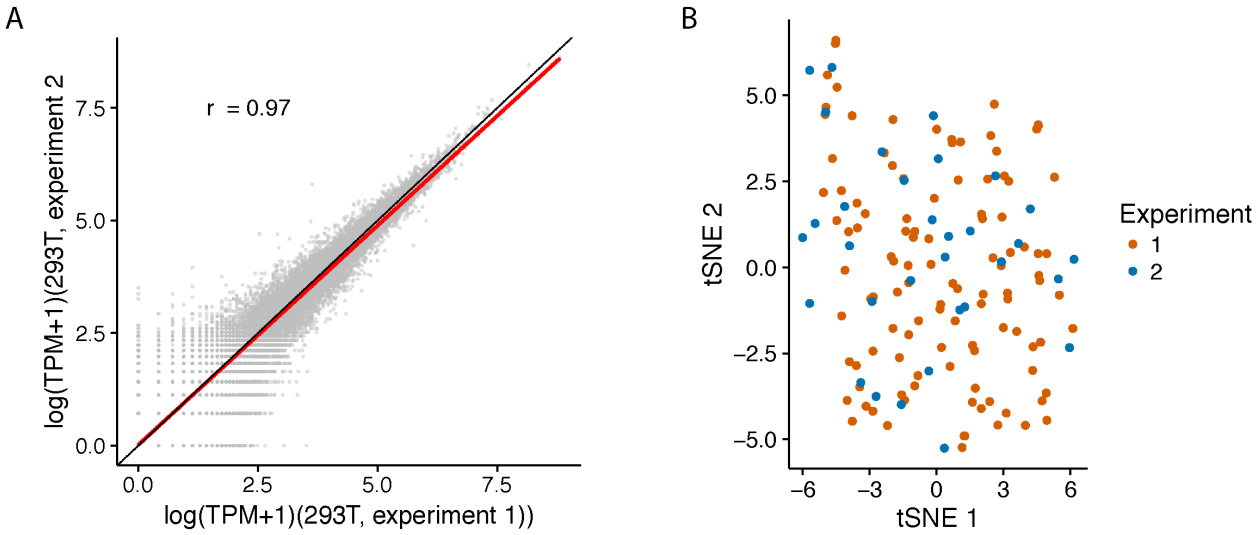


Fig. S9

Evaluation of technical variance between the two *C. elegans* experiments. (**A**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of HEK293T cells spiked in with in the first *C. elegans* experiment (n = 32) vs. the second experiment (n = 111), together with a linear regression line (red) and y=x line (black). (**B**) t-SNE clustering of HEK293T cells recovered from the two experiments. Cells are colored by the experiment from which they derived.

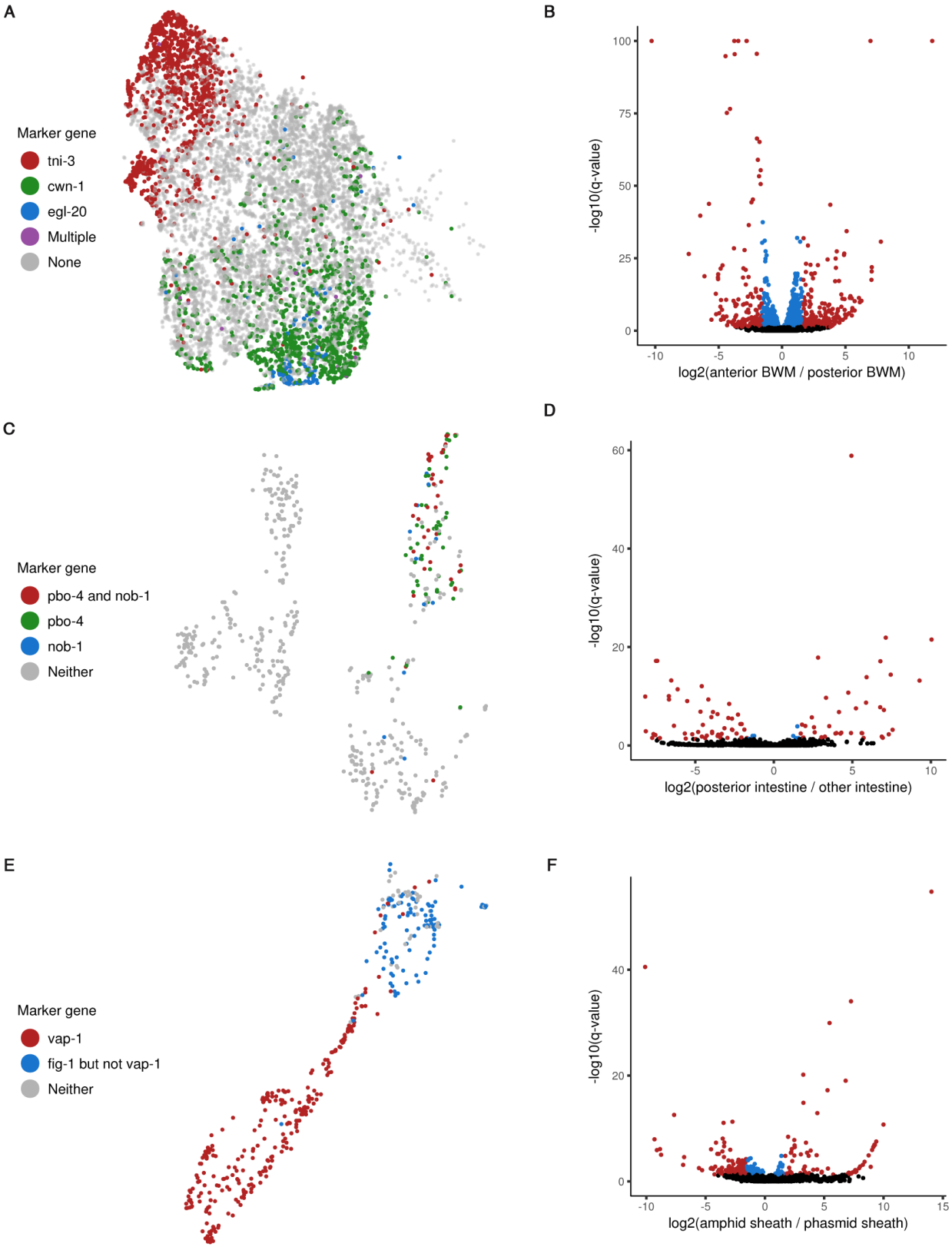

A
Marker gene
tni-3
cwn-1
egl-20
Multiple
None
B
-log10(q-value)
log2(anterior BWM / posterior BWM)
C
Marker gene
pbo-4 and nob-1
pbo-4
nob-1
Neither
D
-log10(q-value)
log2(posterior intestine / other intestine)
E
Marker gene
vap-1
fig-1 but not vap-1
Neither
F
-log10(q-value)
log2(amphid sheath / phasmid sheath)

Fig. S10

sci-RNA-seq reveals genes differentially expressed between anterior and posterior cells for three cell types. (**A**) Expression of anterior/posterior marker genes in body wall muscle cells. Cell t-SNE coordinates are the same as in Fig. 3A, except only BWM cells are shown. *tni-3* (red) is specific to the head (*39*), while *cwn-1* (green) and *egl-20* (blue) are specific to the posterior and tail respectively (*76*). (**B**) Volcano plot showing genes differentially expressed between anterior [*tni-3*(+)] and posterior [*cwn-1*(+) or *egl-20*(+)] body wall muscle. -log10 q-values (y-axis) are capped at 100. Genes with differential expression q-value < 0.05 are colored red if the fold difference in expression is >3, blue otherwise. (**C**) Expression of posterior marker genes in intestine cells. Cell t-SNE coordinates are the same as in fig. S10A, except only intestine cells are shown. *pbo-4* and *nob-1* are specific to the posterior (*77*, *78*). (**D**) Volcano plot showing genes differentially expressed between posterior [*pbo-4*(+) or *nob-1*(+)] intestine and other intestine. Colors are the same as in (B). (**E**) Expression of amphid/phasmid (anterior/posterior) marker genes in amphid/phasmid sheath cells. Cell t-SNE coordinates are the same as in Fig. 3A, except only amphid/phasmid sheath cells are shown. *fig-1* is expressed in both amphid and phasmid sheath cells, while *vap-1* is specific to the amphid sheath cells. (*79*, *80*). (**F**) Volcano plot showing genes differentially expressed between amphid [*vap-1*(+)] and phasmid [*fig-1*(+) *vap-1*(-)] sheath cells. Colors are the same as in (**B**).

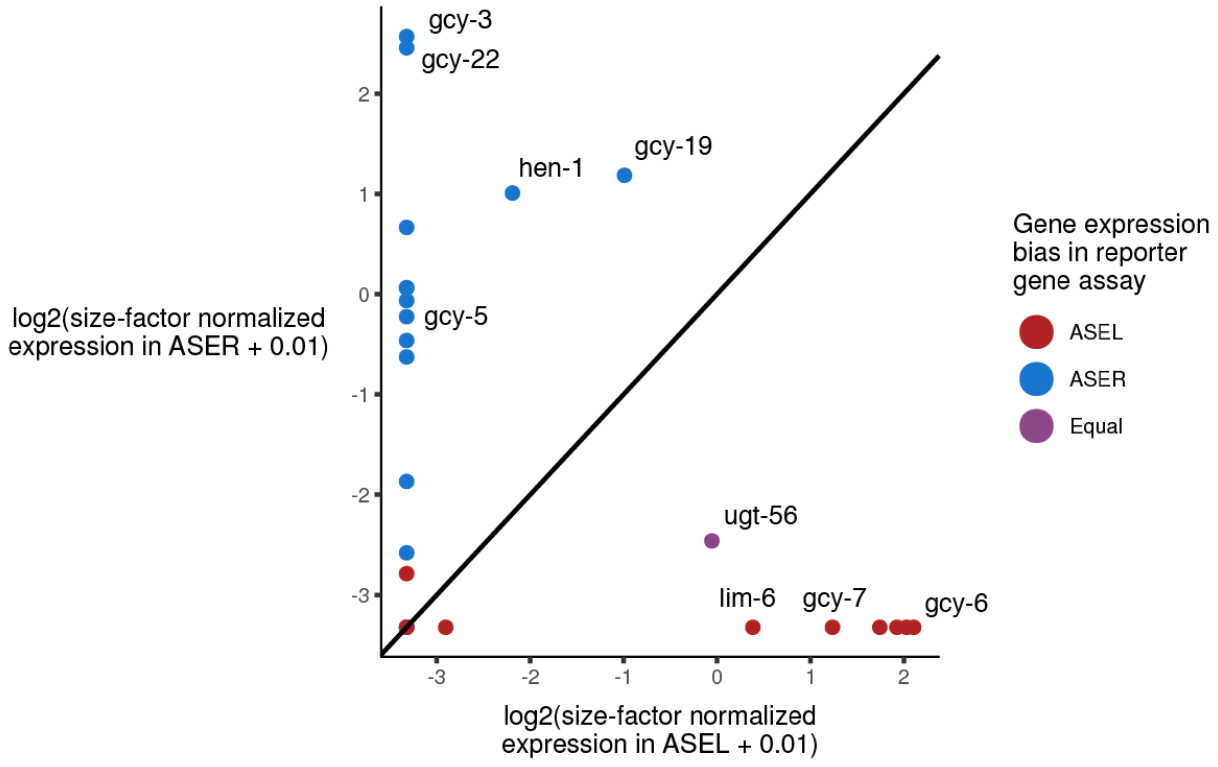


Fig. S11

sci-RNA-seq expression profiles for the ASEL and ASER neurons are consistent with reporter gene assays for asymmetric gene expression. Points represent genes which were tested for asymmetric expression between the ASEL and ASER neurons in promoter-fusion reporter gene assays, as reported by (*43*). Point colors show the expression bias observed in the reporter gene assay for a given gene. The x-axis and y-axis show the log-transformed, size-factor normalized mean number of unique molecular identifiers observed for a given gene per ASEL and ASER cell respectively in the sci-RNA-seq data.

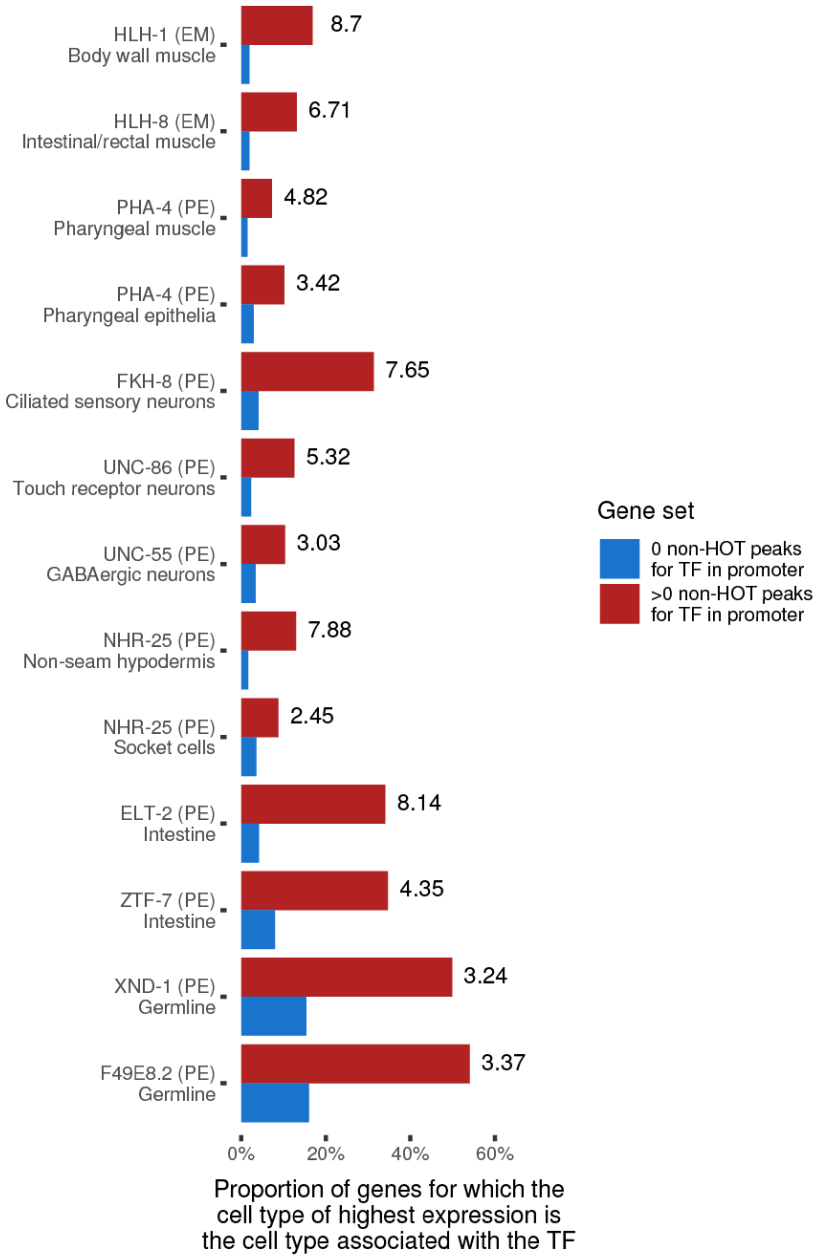


Fig. S12

Transcription factor ChIP-seq peaks predict cell type enriched gene expression. For many TF-to-cell-type associations, the presence of a ChIP-seq peak for the TF in the promoter of a given gene substantially increases the likelihood of the associated cell type being the cell type in which the gene is most highly expressed. Red bars show this probability for genes with at least one peak for the listed TF in their promoter; blue bars show the probability for genes with no

peak for the TF in their promoter. Numbers next to the red bars show the ratio of the probabilities for genes with >0 vs. 0 peaks for the TF in their promoter. The associations here are selected examples, each having a positive coefficient in Fig 5. A “PE” following a TF name indicates that the ChIP-seq dataset(s) for that TF are from post-embryonic worms; “EM” indicates that they are from embryos. “HOT region” peaks, defined as those which overlap peaks >20% of all TFs assayed in the same broad developmental stage (embryonic or post-embryonic), are excluded from the analysis.

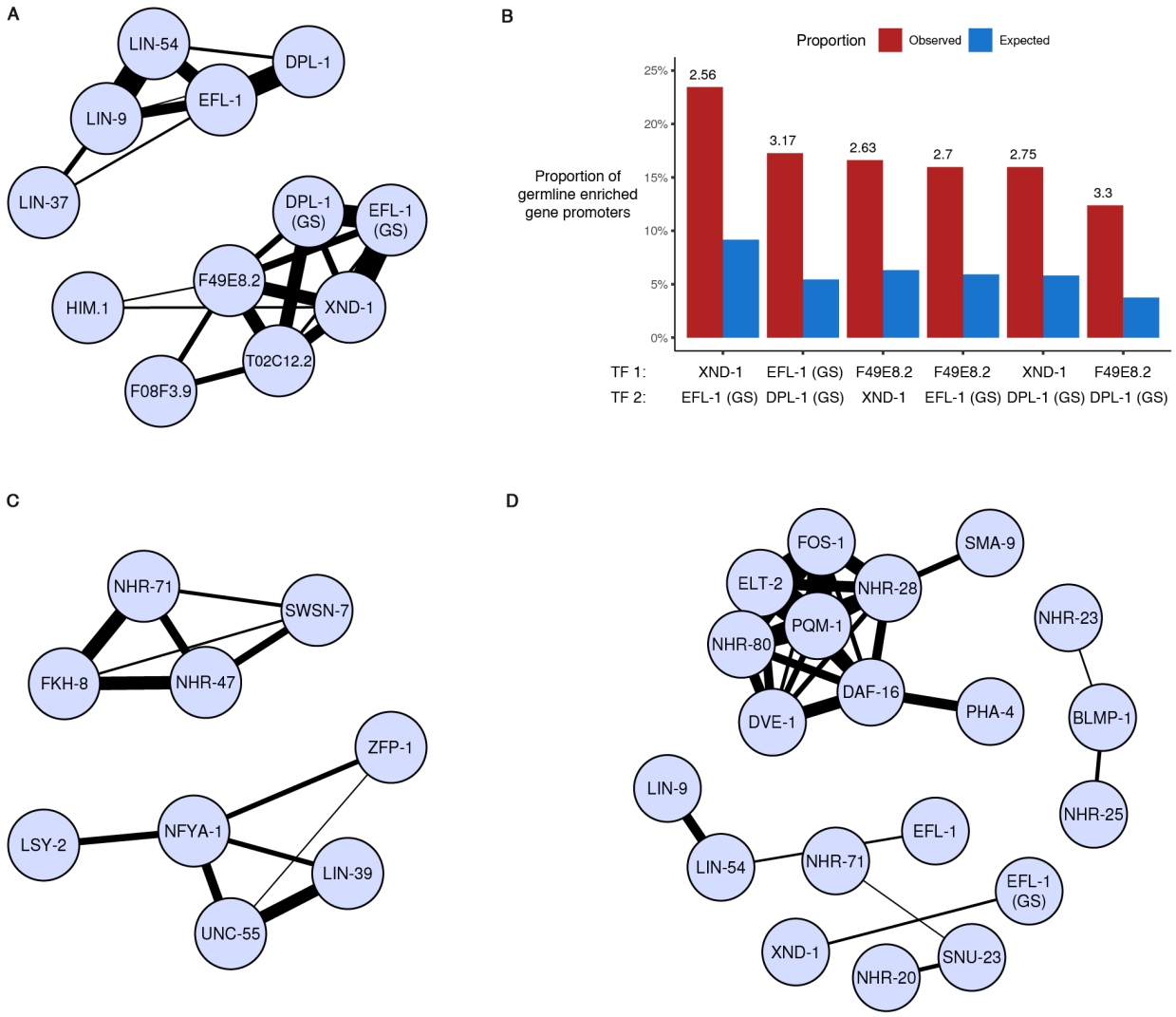


Fig. S13

Transcription factor ChIP-seq peaks have distinct co-localization patterns in the promoters of genes with tissue-enriched expression patterns. (**A**, **C** and **D**) A Graphical LASSO model (**Methods**) is used to find pairs of transcription factors which have overlapping ChIP-seq peaks more often than could be expected by chance, in the context of (A) the promoters of genes with gonad-enriched expression (>5-fold greater in gonad than in any other tissue), (C) the promoters of genes with neuron-enriched expression, or (D) the promoters of all genes. All TF ChIP-seq in this analysis is from post-embryonic stages. EFL-1 (GS) and DPL-1 (GS) refer to peaks from

germline-specific ChIP-seq datasets from (*57*). EFL-1, DPL-1, LIN-9, LIN-37, and LIN-54 are members of the DRM complex (*C. elegans* ortholog of the mammalian DREAM complex), which activates a subset of genes in the germline while repressing them in soma (*57*, *81–83*). (**B**) The observed proportion of germline-enriched genes (those with germline expression >5-fold higher than in any other cell type) that have peaks for both listed TFs in their promoter (in red), compared to the proportion that would be expected if the TF binding patterns were independent conditional on being in a germline-enriched gene promoter (in blue). The numbers above each red bar is the ratio of observed / expected. The conditioning of these statistics on the context of being in a germline-enriched gene promoter rules out the possibility that the co-localizations observed in (A) are simply due to each TF independently being associated with germline-specific genes.

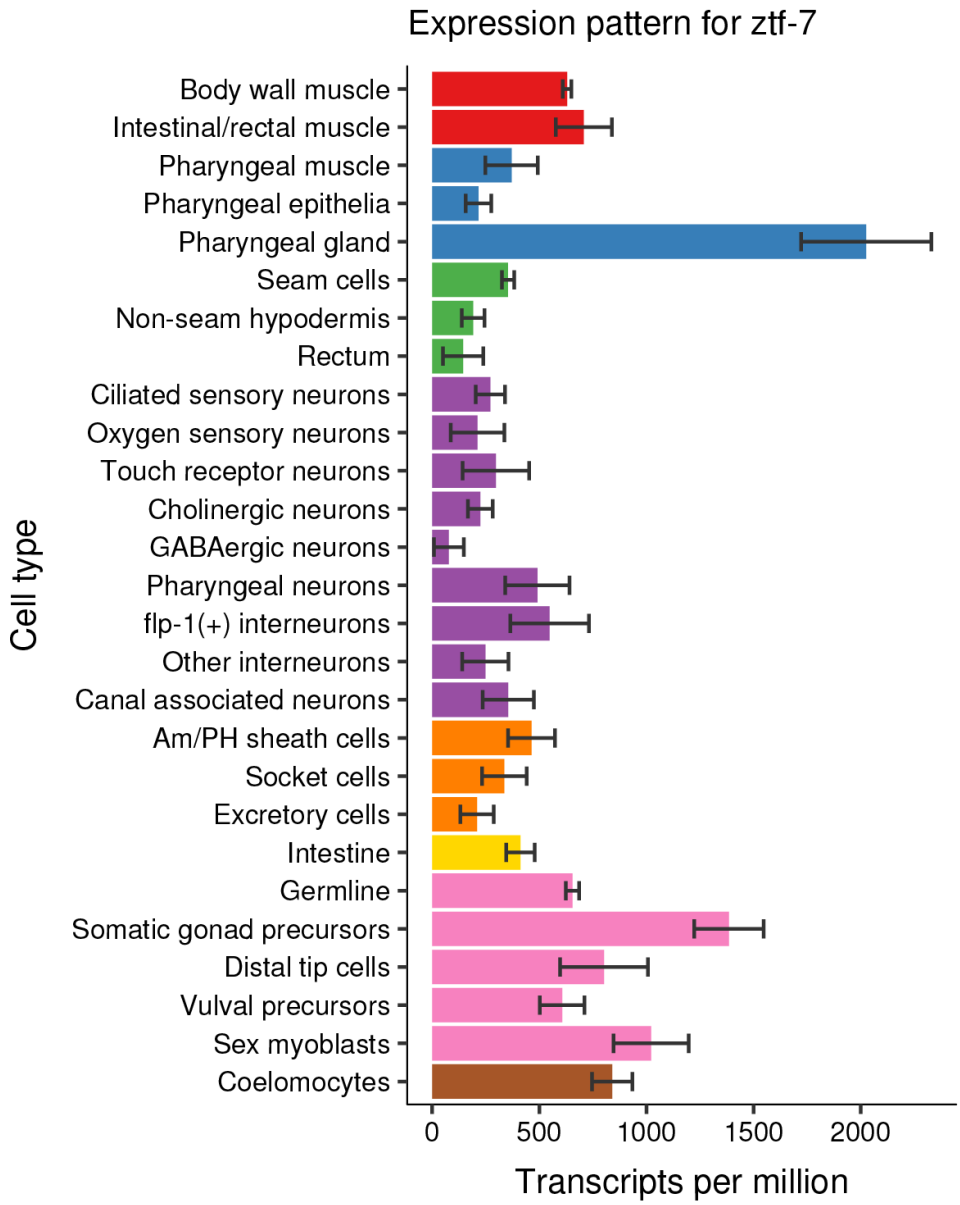


Fig. S14

Example of "gene expression report" image, with full set hosted on GExplore. For a given gene, mean expression values are shown for each of 27 cell types. Black bars indicate the 95% confidence interval. All gene profiles are viewable at: http://genome.sfu.ca/gexplore/gexplore_search_tissues.html.

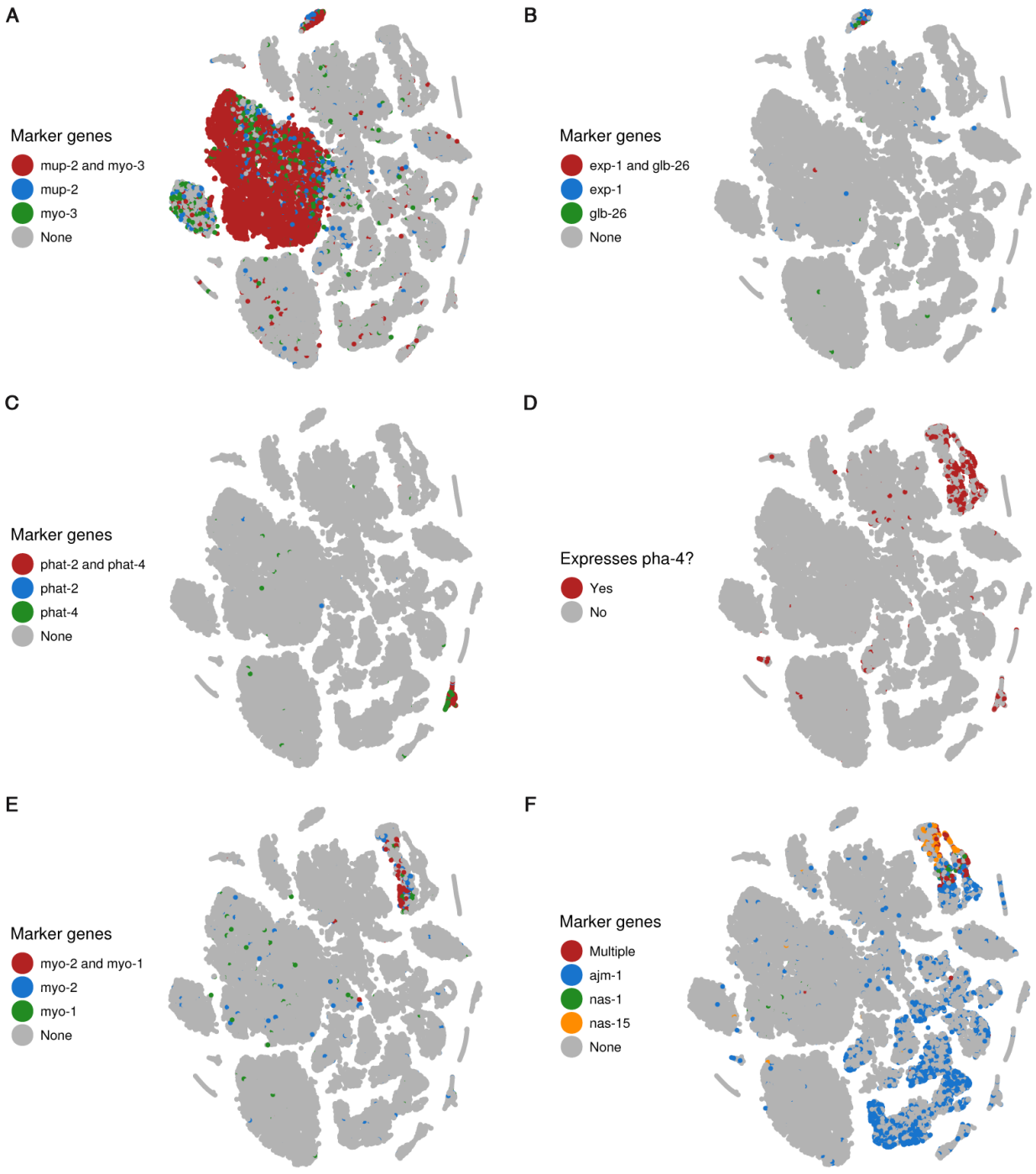


Fig. S15

Expression patterns of marker genes for body wall muscle, intestinal/rectal muscle, and pharynx. (**A**) *mup-2* (troponin T) and *myo-3* (myosin heavy chain A) expression identifies body wall muscle and intestinal/rectal muscle cells (*84*). The cluster to the left of the large muscle

cluster are low UMI-count cells that we believe to be damaged body wall muscle cells. They were excluded from downstream analysis. (**B**) *exp-1* and *glb-26* expression distinguishes intestinal/rectal muscle cells from body wall muscle (*85*, *86*). (**C**) *phat-2* and *phat-4* expression identifies pharyngeal gland cells (*87*). (**D**) *pha-4* expression identifies a cluster (top right) of non-gland pharyngeal cells (*48*). The small *pha-4*(+) cluster on the left are distal tip cells (see fig. S18B). (**E**) *myo-1* and *myo-2* expression identifies pharyngeal muscle cells (*88*). For the purpose of constructing consensus expression profiles, cells in this t-SNE cluster were considered pharyngeal muscle if they expressed at least two of *myo-1*, *myo-2*, *myo-5*, *tnt-4*, *mlc-1* or *mlc-2*. (**F**) *ajm-1*, *nas-1*, and *nas-15* expression identifies non-muscle epithelial cells in the pharyngeal t-SNE cluster. *ajm-1* is expressed in all epithelial cells, while *nas-1* and *nas-15* are specific to the pharynx (*89*, *90*). For the purpose of constructing consensus expression profiles, cells in the pharyngeal muscle/epithelial t-SNE cluster were considered to be epithelial if they do not express any of the markers listed in (E) and expressed at least one of *ajm-1*, *sma-1*, *nas-1*, *nas-15*, or *ifa-1*.

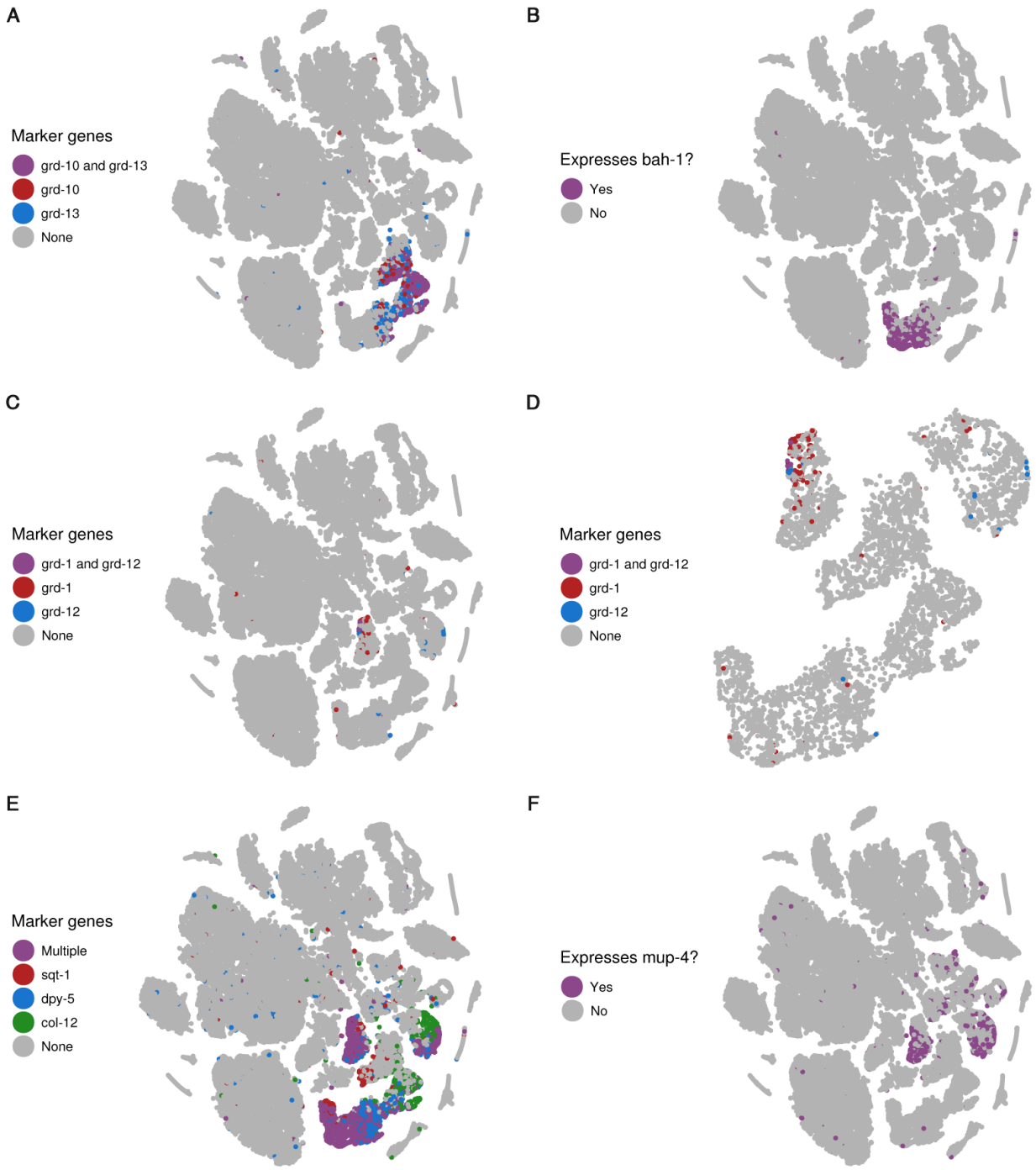


Fig. S16

Expression patterns for marker genes for hypodermis and the rectum. (**A**) *grd-10* and *grd-13* expression identifies seam cells (*91*). (**B**) *bah-1* expression identifies additional seam cells (*92*) and shows that the t-SNE cluster with *grd-10/13* expression is likely to be entirely seam

cells. This cluster also expresses seam cell specific transcription factors including *ceh-18* and *nhr-73*. (**C** to **D**) *grd-1* and *grd-12* expression identifies rectal cells. *grd-1* is expressed in the rectal gland cells (*93*), while *grd-12* is expressed in the B and Y rectal epithelial cells (*91*) (D) is a zoomed-in view of the hypodermal cell clusters in (C). **E**) Expression of the cuticle collagen genes *sqt-1*, *dpy-5*, *col-12* identify hypodermal cells (*94*), including two clusters of non-seam hypodermal cells. We were unable to clearly identify the anatomical differences between the cells in the two non-seam hypodermal clusters. **F**) Expression of *mup-4* is exclusive to non-seam hypodermis and glia, consistent with previous reports of its expression in the circumferential rings of the cuticle (*95*).

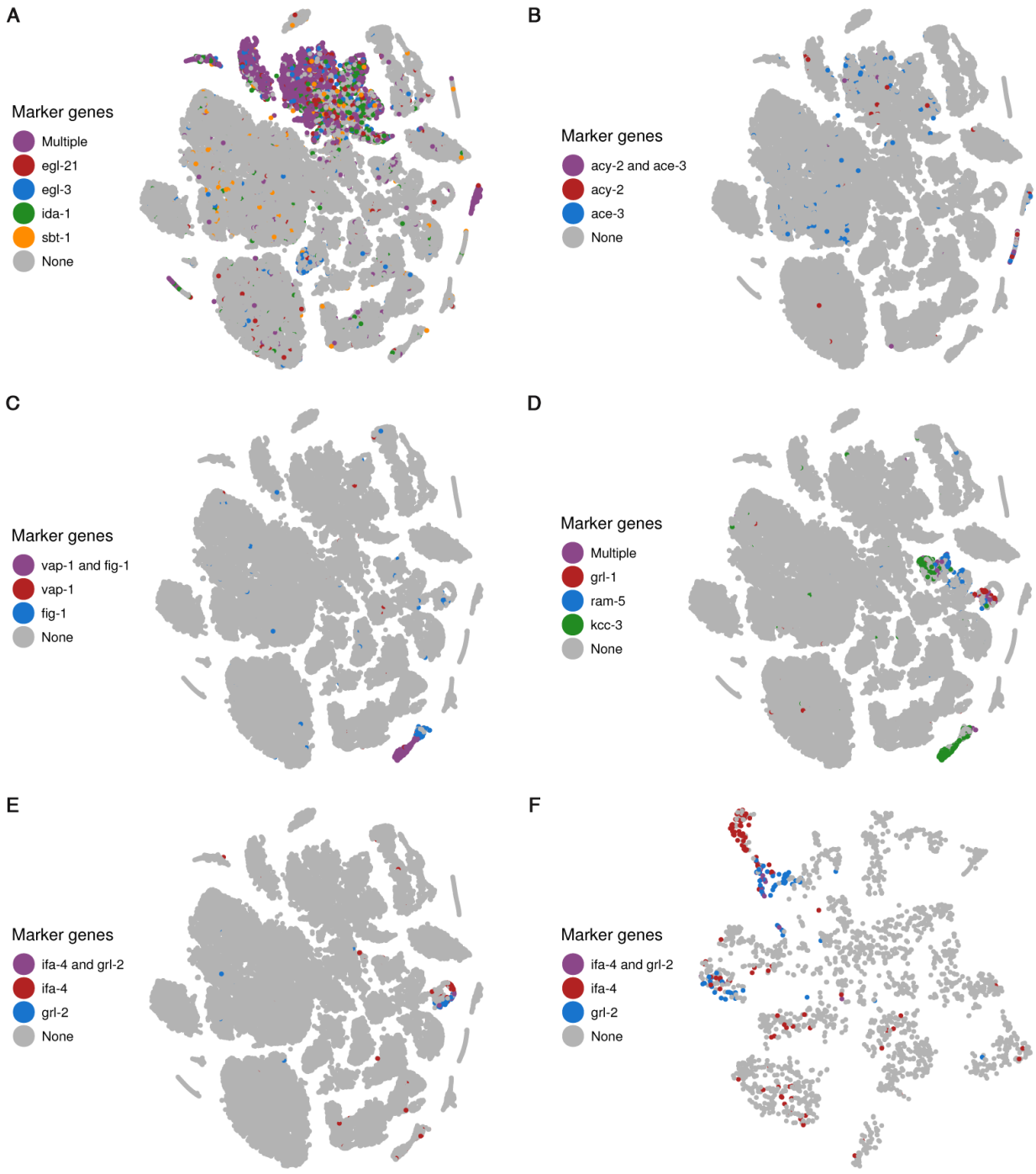


Fig. S17

Expression patterns of marker genes for neurons, glia, and excretory cells. (**A**) Expression of *egl-21*, *egl-3*, *ida-1*, and *sbt-1* identifies neuronal cells (*96–99*). (**B**) The canal associated neurons do not express the marker genes listed in (A), but are identified by their expression of

acy-2 and *ace-3* (*100*, *101*). (**C**) Expression of *vap-1* and *fig-1* identifies the amphid and phasmid sheath cells (*79*). (**D**) Expression of *grl-1* and *ram-5* identifies socket cells (*91*, *102*). Expression of *kcc-3* outside the amphid/phasmid sheath cell cluster identifies additional sheath cells (*103*). For the purpose of constructing consensus expression profiles, cells in the non-amphid/phasmid-sheath glial t-SNE clusters were considered to be socket cells if they were not identified to be excretory cells, expressed at least one of *grl-1*, *grd-15*, *daf-6*, or *ram-5*, and did not express *kcc-3*. (**E**) Expression of *ifa-4* and *grl-2* identifies excretory cells (*91*, *104*). (**F**) *ifa-4*(+) and *grl-2*(+) cells cluster together in a t-SNE of only cells from the glial/excretory cell clusters. We suspect that the *ifa-4*(+) cluster at the top corresponds to the excretory canal cell, while the *grl-2*(+) cluster corresponds to the excretory duct, pore, and/or gland cells.

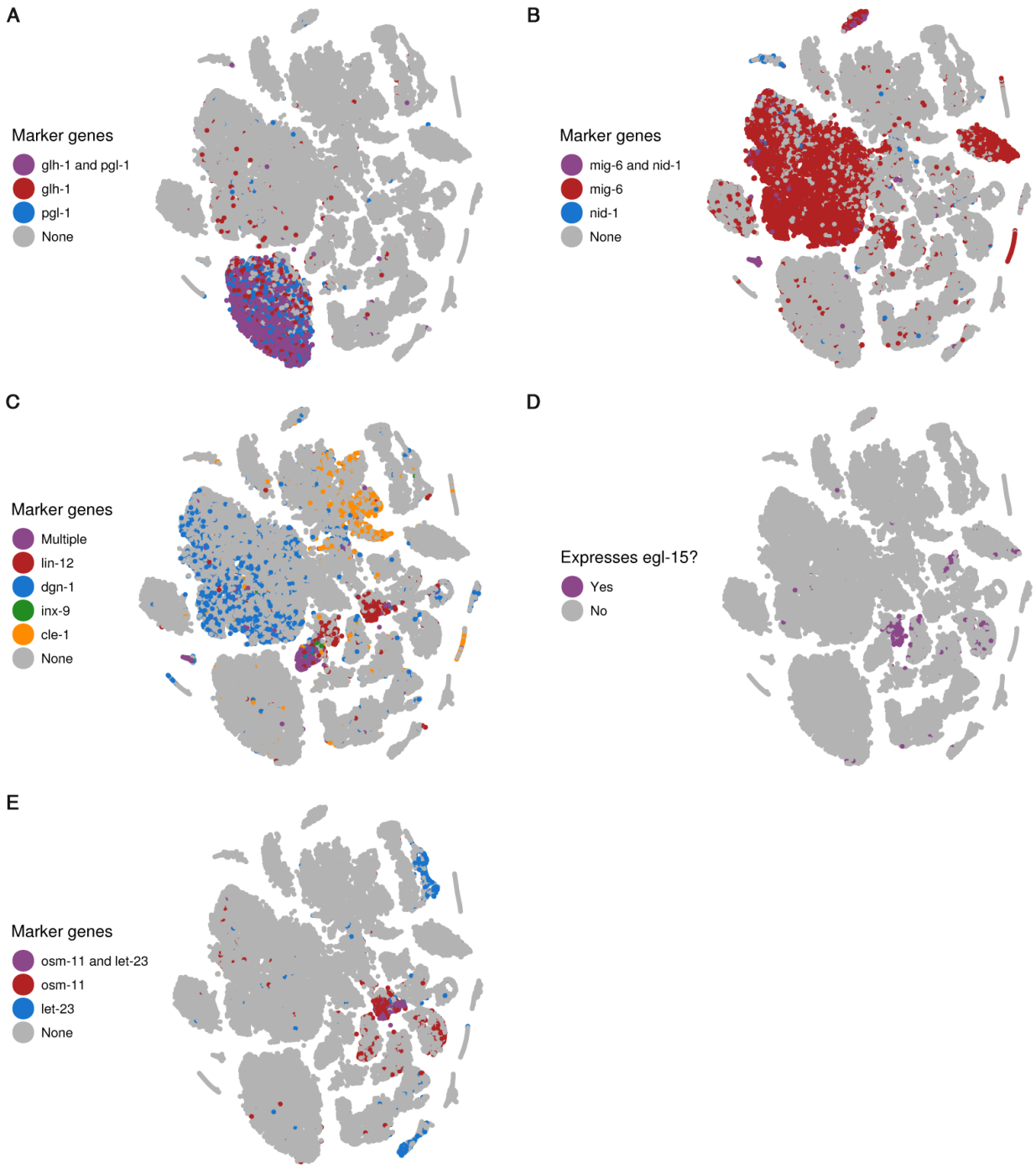


Fig. S18

Expression of marker genes for the germline, somatic gonad, and other sex-related tissues.

(**A**) Expression of *glh-1* and *pgl-1* identifies germline cells (*105*, *106*). (**B**) Co-expression of *mig-6* and *nid-1* identifies the distal tip cells of the somatic gonad (small purple cluster on the lower

left; (*107*, *108*)). (**C**) Co-expression of at least two of *lin-12*, *dgn-1*, *inx-9*, and *cle-1* identifies the somatic gonad precursor cells (*109–112*). (**D**) Expression of *egl-15* identifies sex myoblasts (*113*). (**E**) Expression of *osm-11* and *let-23* identifies vulval precursor cells (*114*, *115*).

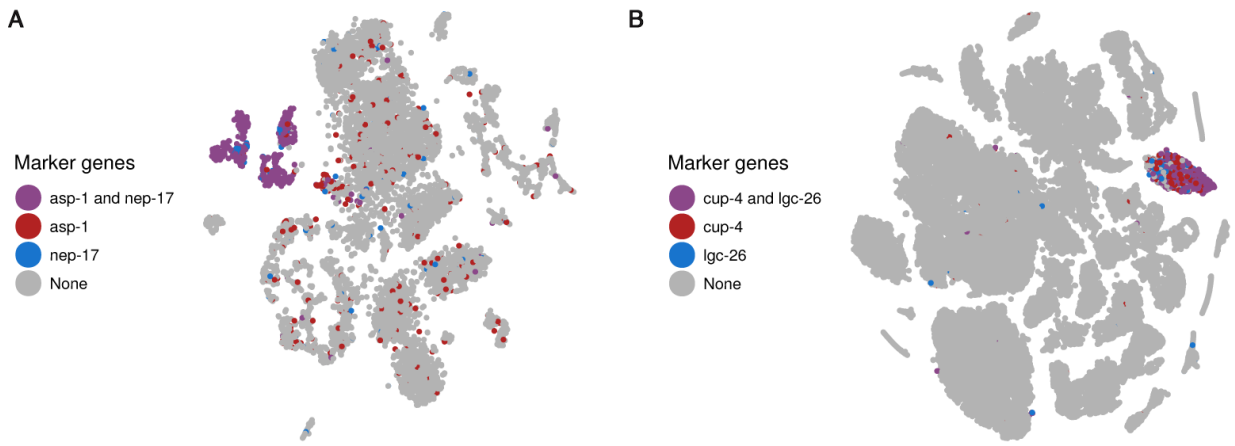


Fig. S19

Expression of marker genes for the intestine and coelomocytes. (**A**) Expression of *asp-1* and *nep-17* identifies intestine cells from the second *C. elegans* experiment. (*116*, *117*). (**B**) Expression of *cup-4* and *lgc-26* identifies coelomocytes (*118*).

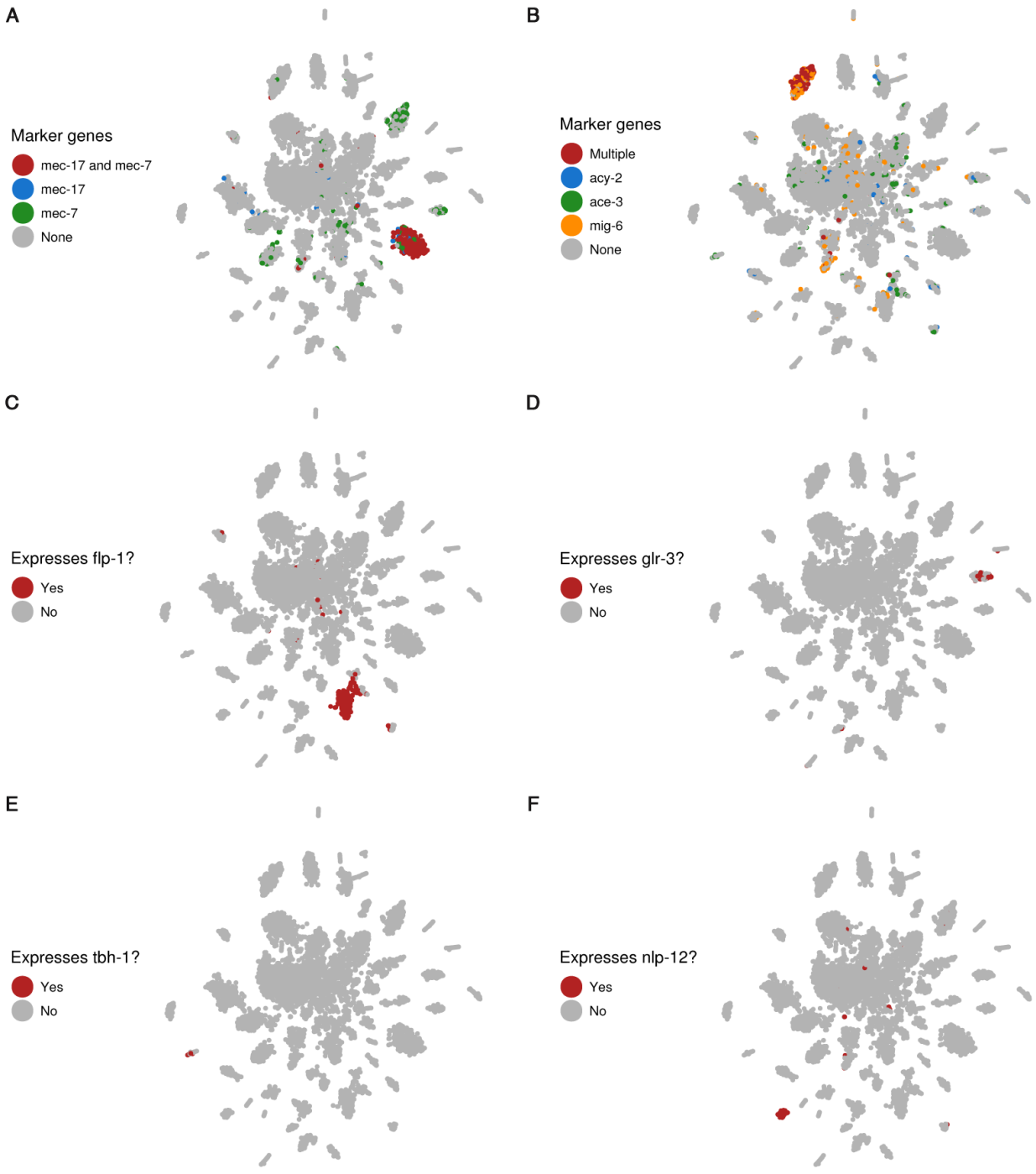


Fig. S20

Expression patterns of marker genes for touch receptor neurons and interneuron subtypes.

t-SNE plots shown are from a clustering of just neuronal cells (identified in fig. S17A,B). (**A**)

Expression of *mec-17* and *mec-7* identifies touch receptor neurons (*119*). (**B**) Expression of *acy-2* and *ace-3* identifies canal associated neurons (*100*, *101*). The canal associated neurons are also the only neuron class that expresses *mig-6* (*120*). (**C**) *flp-1* expression identifies interneurons of the anatomical classes AVK, AVA, AVE, RIG, RMG, AIY, AIA (*121*). *flp-1* has also been reported to be expressed in the M5 pharyngeal motor neuron. (**D**) *glr-3* is expressed exclusively in the RIA interneurons (*122*). (**E**) Among neurons, *tbh-1* is expressed exclusively in the RIC interneurons (*123*). (**F**) *nlp-12* expression identifies the DVA tail interneuron (*124*).

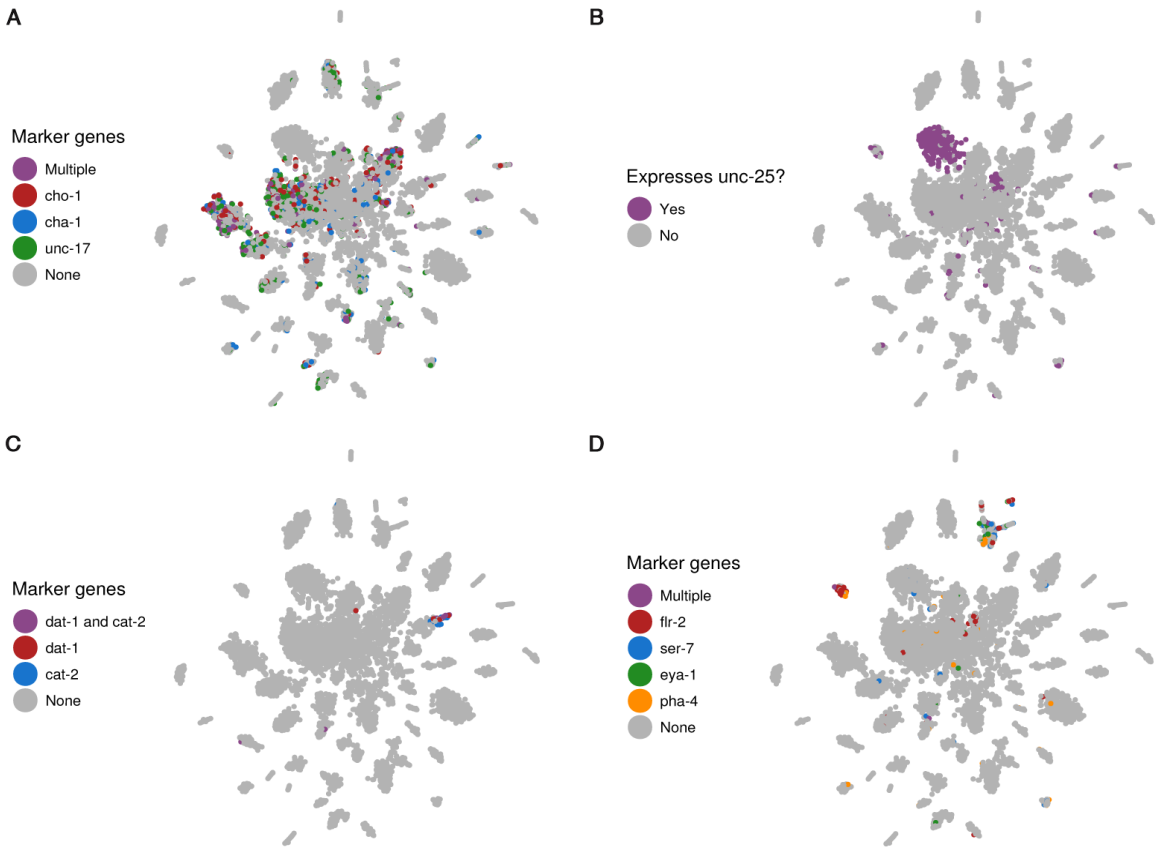


Fig. S21

Expression of marker genes for cholinergic, GABAergic, dopaminergic, and pharyngeal neurons. t-SNE plots shown are from a clustering of just neuronal cells (identified in fig. S17A,B). (**A**) Expression of *cho-1*, *cha-1*, and *unc-17* identifies cholinergic neurons (*125*). For the purpose of constructing consensus expression profiles, neuronal cells were identified as cholinergic if they were not part of a t-SNE cluster identified as any other neuronal subtype and they expressed at least one of *cho-1*, *cha-1*, *unc-17*, *acr-15*, or *acr-18*. (**B**) *unc-25* expression identifies GABAergic neurons (*126*). (**C**) Expression of *dat-1* and *cat-2* identifies dopaminergic neurons (*127*, *128*). (**D**) While no single marker is both highly expressed and specific to pharyngeal neurons, the expression patterns of *flr-2*, *ser-7*, *eya-1*, and *pha-4* together identify two clusters as highly likely to correspond to pharyngeal neurons (*48*, *129–131*).

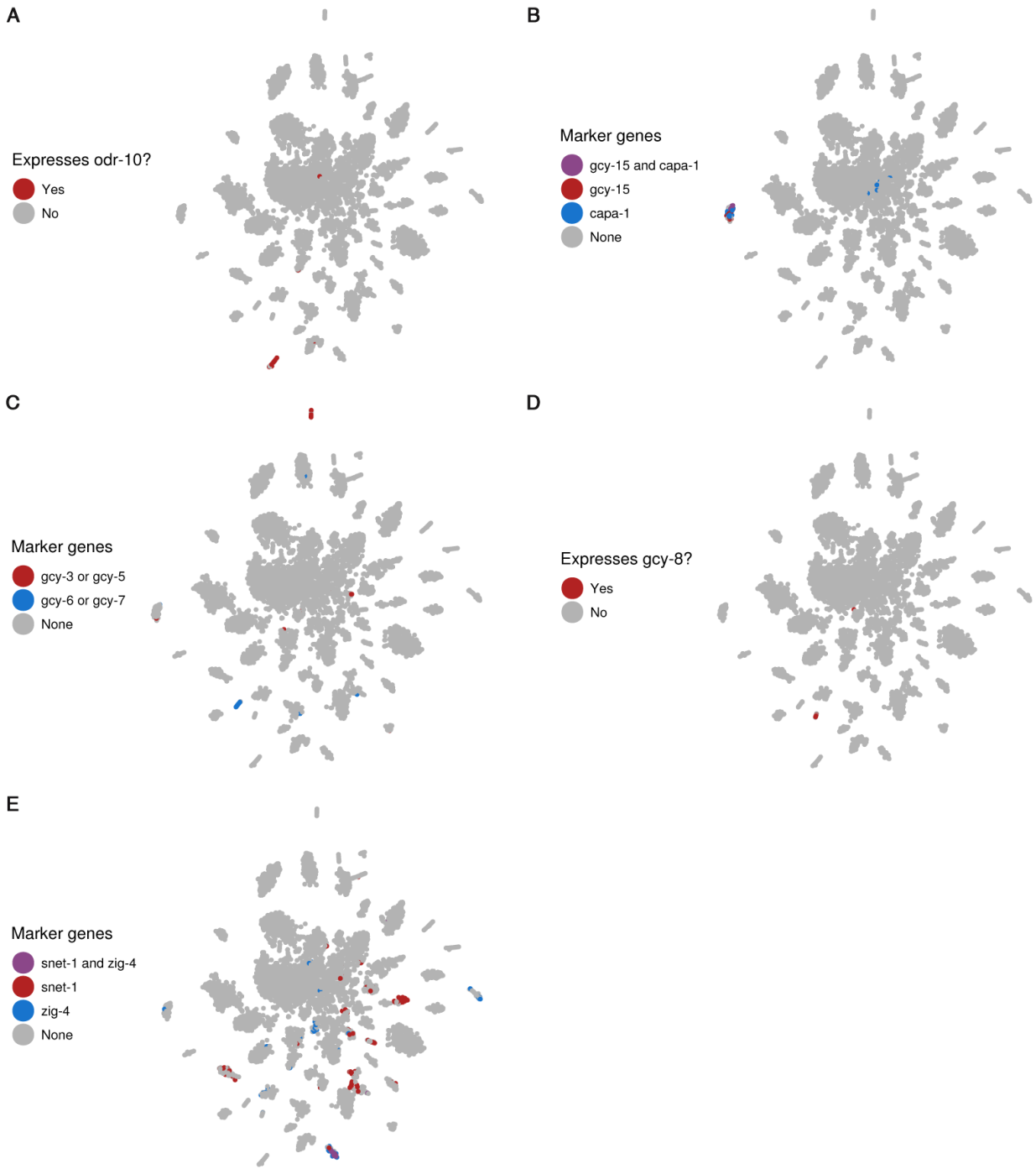


Fig. S22

Expression patterns of marker genes for the AWA, ASG, ASE, AFD, and ASK neurons. t-SNE plots shown are from a clustering of just neuronal cells (identified in fig. S17A,B). (**A**) *odr-10* expression identifies the AWA neurons (*132*). (**B**) *gcy-15* expression identifies the ASG

neurons (*133*). *capa-1* has also been reported to be expressed in two specific but unidentified pairs of neurons in the head (*134*); in our data it is expressed predominantly in the same cluster as *gcy-15*. (**C**) Expression of *gcy-3* and *gcy-5* identifies the ASER neuron, while expression of *gcy-6* and *gcy-7* identifies the ASEL neuron (*42*, *43*). (**D**) *gcy-8* expression identifies the AFD neurons (*135*). (**E**) Co-expression of *snet-1* and *zig-4* identifies the ASK neurons (*136*, *137*).

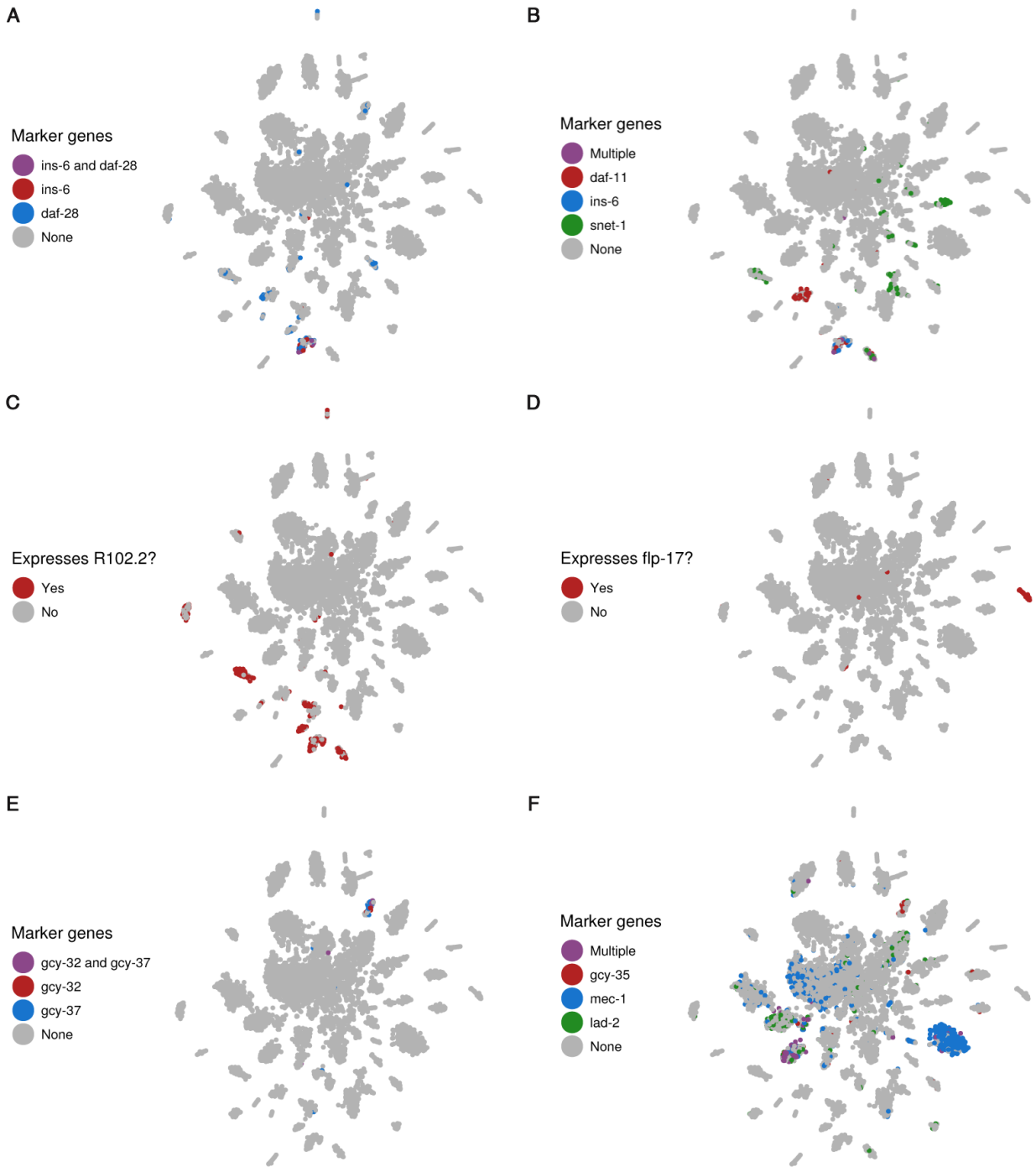


Fig. S23

Expression patterns of marker genes for ASI/ASJ, AWB/AWC, BAG, URX, SDQ, and other ciliated sensory neurons. t-SNE plots shown are from a clustering of just neuronal cells (identified in fig. S17A,B). (**A**) Expression of *ins-6* and *daf-28* identifies a neuron cluster that

consists of the ASI and ASJ neurons (*138*, *139*). (**B**) Based on reported expression patterns, a neuron cluster that expresses *daf-11* but not *ins-6* or *snet-1* can only correspond to the AWB and/or AWC neurons (*136*, *138*, *140*). (**C**) Beyond those identified in fig. S22**,** and (A) of this figure, three additional neuron clusters express *R102.2*. Based on the expression patterns reported by (*141*), these clusters correspond to the ciliated sensory neurons classes ADF, ASH, PHA, and/or PHB. We could not precisely identify them however. For the purpose of constructing consensus expression profiles, neuronal cells were considered ciliated sensory neurons if they either were part of a cluster that was identified as a ciliated sensory neuron class or were part of a cluster that could not be conclusively identified but expressed high levels of *R102.2*, *dyf-2*, *che-3*, or *nphp-4*. (**D**) *flp-17* expression identifies the BAG neurons (*121*). (**E**) Expression of *gcy-32* and *gcy-37* identifies a neuron cluster that consists of the URX, AQR, and PQR neurons (*142*, *143*). (**F**) Among neurons, *gcy-35* is expressed in the URX, AQR, PQR, SDQ, ALN, PLN O_2-sensory neurons, as well as the AVM and BDU neurons (*143*). *mec-1* was reported to be expressed in the touch receptor neurons, SDQ/ALN/PLN O_2-sensory neurons, and PVT neurons (*144*). *lad-2* was reported to be expressed in the SDQ/ALN/PLN O_2-sensory neurons and some sublateral motor neurons (*145*). Based on these expression patterns, a neuron cluster enriched for expression of all three of these genes is likely to correspond to the SDQ/ALN/PLN O_2-sensory neurons.

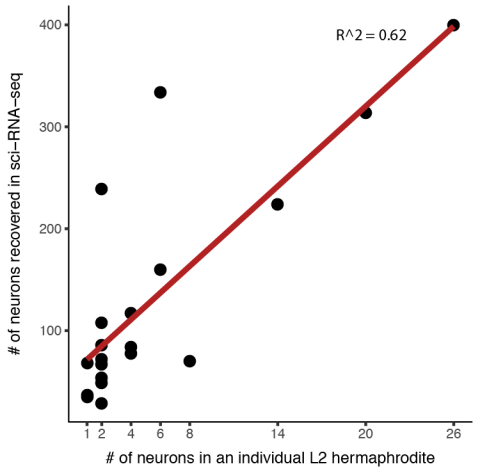


Fig. S24

Recovery rates of neuron types in sci-RNA-seq. The observed number of cells identified in sci-RNA-seq for a given neuron type (y axis) is compared to the number of neurons of that type in an individual L2 hermaphrodite *C. elegans* (x-axis). The plot includes all specific neuron types that we were able to identify, excluding cholinergic neurons, which were not limited to distinct t-SNE clusters and therefore may be under-counted as we only considered a cell cholinergic if we observed expression of at least one cholinergic marker gene (see **Fig. S21**). The neuron types included in the plot are: ASEL, ASER, DVA, AFD, ASG, ASK, AWA, BAG, CAN, RIA, RIC, ASI/ASJ, AWB/AWC, URX/AQR/PQR, SDQ/ALN/PLN, touch receptor neurons (ALM/PLM/AVM/PVM), dopaminergic neurons (CEP/ADE/PDE), flp-1(+) neurons (excluding the pharyngeal neuron M5), pharyngeal neurons, and GABAergic neurons.

Reference of chapter 1

1. C. Trapnell, Defining cell types and states with single-cell genomics. *Genome Res.* **25**, 1491–1498 (2015).

2. D. Ramsköld *et al.*, Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).

3. A. K. Shalek *et al.*, Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*. **498**, 236–240 (2013).

4. Q. F. Wills *et al.*, Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat. Biotechnol.* **31**, 748–752 (2013).

5. G. X. Y. Zheng *et al.*, Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

6. A. A. Pollen *et al.*, Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32**, 1053–1058 (2014).

7. A. Zeisel *et al.*, Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*. **347**, 1138–1142 (2015).

8. B. B. Lake *et al.*, Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*. **352**, 1586–1590 (2016).

9. I. Tirosh *et al.*, Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*. **352**, 189–196 (2016).

10. W. Zeng *et al.*, Single-nucleus RNA-seq of differentiating human myoblasts reveals the extent of fate heterogeneity. *Nucleic Acids Res.* (2016), doi:10.1093/nar/gkw739.

11. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

12. D. Ramsköld *et al.*, Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).

13. G. X. Y. Zheng *et al.*, Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

14. B. B. Lake *et al.*, Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*. **352**, 1586–1590 (2016).

15. F. Tang *et al.*, mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*. **6**, 377–382 (2009).

16. S. Picelli *et al.*, Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods*. **10**, 1096–1098 (2013).

17. R. V. Grindberg *et al.*, RNA-sequencing from single nuclei. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19802–19807 (2013).

18. H. Christina Fan, G. K. Fu, S. P. A. Fodor, Combinatorial labeling of single cells for gene expression cytometry. *Science*. **347**, 1258367 (2015).

19. E. Z. Macosko *et al.*, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. **161**, 1202–1214 (2015).

20. A. M. Klein *et al.*, Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*. **161**, 1187–1201 (2015).

21. A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, S. A. Teichmann, The technology and biology of single-cell RNA sequencing. *Mol. Cell*. **58**, 610–620 (2015).

22. S. Liu, C. Trapnell, Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Res.* **5** (2016), doi:10.12688/f1000research.7223.1.

23. A. Adey *et al.*, In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res.* **24**, 2041–2049 (2014).

24. S. Amini *et al.*, Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343–1349 (2014).

25. D. A. Cusanovich *et al.*, Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*. **348**, 910–914 (2015).

26. S. A. Vitak *et al.*, Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods*. **14**, 302–308 (2017).

27. V. Ramani *et al.*, Massively multiplex single-cell Hi-C. *Nat. Methods*. **14**, 263–266 (2017).

28. R. M. Mulqueen *et al.*, Scalable and efficient single-cell DNA methylation sequencing by combinatorial indexing (2017), , doi:10.1101/157230.

29. J. E. Sulston, E. Schierenberg, J. G. White, J. N. Thomson, The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev. Biol.* **100**, 64–119 (1983).

30. J. E. Sulston, H. R. Horvitz, Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Dev. Biol.* **56**, 110–156 (1977).

31. Supplemental online materials.

32. R. V. Grindberg *et al.*, RNA-sequencing from single nuclei. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19802–19807 (2013).

33. J. Gertz *et al.*, Transposase mediated construction of RNA-seq libraries. *Genome Res.* **22**, 134–141 (2012).

34. E. M. Hedgecock, J. G. White, Polyploid tissues in the nematode Caenorhabditis elegans. *Dev. Biol.* **107**, 128–133 (1985).

35. G. V. Clokey, L. A. Jacobson, The autofluorescent "lipofuscin granules" in the intestinal cells of Caenorhabditis elegans are secondary lysosomes. *Mech. Ageing Dev.* **35**, 79–94 (1986).

36. G.-J. Hendriks, D. Gaidatzis, F. Aeschimann, H. Großhans, Extensive oscillatory gene expression during C. elegans larval development. *Mol. Cell*. **53**, 380–392 (2014).

37. G. Heimberg, R. Bhatnagar, H. El-Samad, M. Thomson, Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Syst*. **2**, 239–250 (2016).

38. M. E. Boeck *et al.*, The time-resolved transcriptome of C. elegans. *Genome Res.* **26**, 1441–1450 (2016).

39. R. Ruksana *et al.*, Tissue expression of four troponin I genes and their molecular interactions with two troponin C isoforms in Caenorhabditis elegans. *Genes Cells*. **10**, 261–276 (2005).

40. J. G. White, E. Southgate, J. N. Thomson, S. Brenner, The structure of the nervous system of the nematode Caenorhabditis elegans. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **314**, 1–340 (1986).

41. O. Hobert, L. Glenwinkel, J. White, Revisiting Neuronal Cell Type Classification in Caenorhabditis elegans. *Curr. Biol.* **26**, R1197–R1203 (2016).

42. O. Hobert, R. J. Johnston, S. Chang, Left–right asymmetry in the nervous system: the Caenorhabditis elegans model. *Nat. Rev. Neurosci.* **3**, 629–640 (2002).

43. J. Takayama, S. Faumont, H. Kunitomo, S. R. Lockery, Y. Iino, Single-cell transcriptional analysis of taste sensory neuron pair in Caenorhabditis elegans. *Nucleic Acids Res.* **38**, 131–142 (2010).

44. T. R. Sarafi-Reinach, T. Melkman, O. Hobert, P. Sengupta, The lin-11 LIM homeobox gene specifies olfactory and chemosensory neuron fates in C. elegans. *Development*. **128**, 3269–3281 (2001).

45. C. L. Araya *et al.*, Corrigendum: Regulatory analysis of the C. elegans genome with spatiotemporal resolution. *Nature*. **528**, 152 (2015).

46. modERN consortia. *ENCODE*, (available at http://encodeproject.org/).

47. T. Fukushige, T. M. Brodigan, L. A. Schriefer, R. H. Waterston, M. Krause, Defining the transcriptional redundancy of early bodywall muscle development in C. elegans: evidence for a unified theory of animal muscle development. *Genes Dev.* **20**, 3395–3406 (2006).

48. J. Gaudet, S. E. Mango, Regulation of organogenesis by the Caenorhabditis elegans FoxA protein PHA-4. *Science*. **295**, 821–825 (2002).

49. B. D. Harfe *et al.*, Analysis of a Caenorhabditis elegans Twist homolog identifies conserved and divergent aspects of mesodermal patterning. *Genes Dev.* **12**, 2623–2635 (1998).

50. M. Horn *et al.*, DRE-1/FBXO11-dependent degradation of BLMP-1/BLIMP-1 governs C. elegans developmental timing and maturation. *Dev. Cell*. **28**, 697–710 (2014).

51. C. R. Gissendanner, A. E. Sluder, nhr-25, the Caenorhabditis elegans ortholog of ftz-f1, is required for epidermal and somatic gonad development. *Dev. Biol.* **221**, 259–272 (2000).

52. T. Fukushige, M. G. Hawkins, J. D. McGhee, The GATA-factor elt-2 is essential for formation of the Caenorhabditis elegans intestine. *Dev. Biol.* **198**, 286–302 (1998).

53. C. R. Wagner, L. Kuervers, D. L. Baillie, J. L. Yanowitz, xnd-1 regulates the global recombination landscape in Caenorhabditis elegans. *Nature*. **467**, 839–843 (2010).

54. R. Mainpal, J. Nance, J. L. Yanowitz, A germ cell determinant reveals parallel pathways for germ line development in Caenorhabditis elegans. *Development*. **142**, 3571–3582 (2015).

55. I. A. Hope, A. Mounsey, P. Bauer, S. Aslam, The forkhead gene family of Caenorhabditis elegans. *Gene*. **304**, 43–55 (2003).

56. D. D. Shaye, I. Greenwald, OrthoList: a compendium of C. elegans genes with human orthologs. *PLoS One*. **6**, e20085 (2011).

57. M. Kudron *et al.*, Tissue-specific direct targets of Caenorhabditis elegans Rb/E2F dictate distinct somatic and germline programs. *Genome Biol.* **14**, R5 (2013).

58. P.-Y. Tung *et al.*, Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7**, 39921 (2017).

59. A. McKenna *et al.*, Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*. **353**, aaf7907 (2016).

60. A. Z. and Hall, *WormAtlas* (2017), (available at http://www.wormatlas.org/neurons/Individual%20Neurons/ASEframeset.html).

61. C. L. Araya *et al.*, Corrigendum: Regulatory analysis of the C. elegans genome with spatiotemporal resolution. *Nature* (2015), doi:10.1038/nature16075.

62. S. Zhang, D. Banerjee, J. R. Kuhn, Isolation and culture of larval cells from C. elegans. *PLoS One*. **6**, e19505 (2011).

63. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*. **10**, 1213–1218 (2013).

64. O. Tange, Others, Gnu parallel-the command-line power tool. *The USENIX Magazine*. **36**, 42–47 (2011).

65. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).

66. S. Anders, P. T. Pyl, W. Huber, HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, btu638 (2014).

67. A. Adey *et al.*, The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*. **500**, 207–211 (2013).

68. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell developmental trajectories. *bioRxiv* (2017), p. 110668.

69. N. Habib *et al.*, Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science*. **353**, 925–928 (2016).

70. A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks. *Science*. **344**, 1492–1496 (2014).

71. M. B. Gerstein *et al.*, Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*. **330**, 1775–1787 (2010).

72. J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*. **9**, 432–441 (2008).

73. G. Strona, D. Nappo, F. Boccacci, S. Fattorini, J. San-Miguel-Ayanz, A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nat. Commun.* **5**, 4114 (2014).

74. I. L. Johnstone, J. D. Barry, Temporal reiteration of a precise gene expression pattern during nematode development. *EMBO J.* **15**, 3633–3639 (1996).

75. A. R. Frand, S. Russel, G. Ruvkun, Functional genomic analysis of C. elegans molting. *PLoS Biol.* **3**, e312 (2005).

76. M. Harterink *et al.*, Neuroblast migration along the anteroposterior axis of C. elegans is controlled by opposing gradients of Wnts and a secreted Frizzled-related protein. *Development*. **138**, 2915–2924 (2011).

77. K. Nehrke, J. E. Melvin, The NHX family of Na+-H+ exchangers in Caenorhabditis elegans. *J. Biol. Chem.* **277**, 29036–29044 (2002).

78. J. I. Murray *et al.*, Multidimensional regulation of gene expression in the C. elegans embryo. *Genome Res.* **22**, 1282–1294 (2012).

79. T. Bacaj, M. Tevlin, Y. Lu, S. Shaham, Glia Are Essential for Sensory Organ Function in C. elegans. *Science*. **322**, 744–747 (2008).

80. E. A. Perens, S. Shaham, C. elegans daf-6 encodes a patched-related protein required for lumen formation. *Dev. Cell*. **8**, 893–906 (2005).

81. M. M. Harrison, C. J. Ceol, X. Lu, H. R. Horvitz, Some C. elegans class B synthetic multivulva proteins encode a conserved LIN-35 Rb-containing complex distinct from a NuRD-like complex. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 16782–16787 (2006).

82. T. M. Tabuchi *et al.*, Chromosome-biased binding and gene regulation by the Caenorhabditis elegans DRM complex. *PLoS Genet.* **7**, e1002074 (2011).

83. I. Latorre *et al.*, The DREAM complex promotes gene body H2A.Z for target repression. *Genes Dev.* **29**, 495–500 (2015).

84. D. G. Moerman, B. D. Williams, Sarcomere assembly in C. elegans muscle. *WormBook*, 1–16 (2006).

85. A. A. Beg, E. M. Jorgensen, EXP-1 is an excitatory GABA-gated cation channel. *Nat. Neurosci.* **6**, 1145–1152 (2003).

86. L. Tilleman *et al.*, An N-myristoylated globin with a redox-sensing function that regulates the defecation cycle in Caenorhabditis elegans. *PLoS One*. **7**, e48768 (2012).

87. V. Ghai, R. B. Smit, J. Gaudet, Transcriptional regulation of HLH-6-independent and subtype-specific genes expressed in the Caenorhabditis elegans pharyngeal glands. *Mech. Dev.* **129**, 284–297 (2012).

88. J. P. Ardizzi, H. F. Epstein, Immunochemical localization of myosin heavy chain isoforms and paramyosin in developmentally and structurally diverse muscle cell types of the nematode Caenorhabditis elegans. *J. Cell Biol.* **105**, 2763–2770 (1987).

89. M. Labouesse, Epithelial junctions and attachments. *WormBook*, 1–21 (2006).

90. F. Möhrlen, H. Hutter, R. Zwilling, The astacin protein family in Caenorhabditis elegans. *Eur. J. Biochem.* **270**, 4909–4920 (2003).

91. L. Hao, R. Johnsen, G. Lauter, D. Baillie, T. R. Bürglin, Comprehensive analysis of gene expression patterns of hedgehog-related genes. *BMC Genomics*. **7**, 280 (2006).

92. K. Drace, S. McLaughlin, C. Darby, Caenorhabditis elegans BAH-1 is a DUF23 protein expressed in seam cells and required for microbial biofilm binding to the cuticle. *PLoS One*. **4**, e6741 (2009).

93. G. Aspöck, H. Kagoshima, G. Niklaus, T. R. Bürglin, Caenorhabditis elegans has scores of hedgehog-related genes: sequence and expression analysis. *Genome Res.* **9**, 909–923 (1999).

94. A. P. Page, I. L. Johnstone, The cuticle. *WormBook*, 1–15 (2007).

95. L. Hong *et al.*, MUP-4 is a novel transmembrane protein with functions in epithelial cell adhesion in Caenorhabditis elegans. *J. Cell Biol.* **154**, 403–414 (2001).

96. T. C. Jacob, J. M. Kaplan, The EGL-21 carboxypeptidase E facilitates acetylcholine release at Caenorhabditis elegans neuromuscular junctions. *J. Neurosci.* **23**, 2122–2130 (2003).

97. J. Kass, T. C. Jacob, P. Kim, J. M. Kaplan, The EGL-3 proprotein convertase regulates mechanosensory responses of Caenorhabditis elegans. *J. Neurosci.* **21**, 9265–9272 (2001).

98. T. R. Zahn, M. A. Macmorris, W. Dong, R. Day, J. C. Hutton, IDA-1, a Caenorhabditis elegans homolog of the diabetic autoantigens IA-2 and phogrin, is expressed in peptidergic neurons in the worm. *J. Comp. Neurol.* **429**, 127–143 (2001).

99. D. Sieburth *et al.*, Systematic analysis of genes required for synapse structure and function. *Nature*. **436**, 510–517 (2005).

100. H. C. Korswagen, A. M. van der Linden, R. H. Plasterk, G protein hyperactivation of the Caenorhabditis elegans adenylyl cyclase SGS-1 induces neuronal degeneration. *EMBO J.* **17**, 5059–5065 (1998).

101. D. Combes, Y. Fedon, J.-P. Toutant, M. Arpagaus, Multiple ace genes encoding acetylcholinesterases of Caenorhabditis elegans have distinct tissue expression. *Eur. J. Neurosci.* **18**, 497–512 (2003).

102. R. Y. Yu, C. Q. Nguyen, D. H. Hall, K. L. Chow, Expression of ram-5 in the structural cell is required for sensory ray morphogenesis in Caenorhabditis elegans male tail. *EMBO J.* **19**, 3542–3555 (2000).

103. A. Yoshida *et al.*, A glial K+/Cl- cotransporter modifies temperature-evoked dynamics in Caenorhabditis elegans sensory neurons. *Genes Brain Behav.* (2015) (available at http://onlinelibrary.wiley.com/doi/10.1111/gbb.12260/pdf).

104. A. Karabinos, E. Schulze, J. Schünemann, D. A. D. Parry, K. Weber, In vivo and in vitro evidence that the four essential intermediate filament (IF) proteins A1, A2, A3 and B1 of the nematode Caenorhabditis elegans form an obligate heteropolymeric IF system. *J. Mol. Biol.* **333**, 307–319 (2003).

105. M. E. Gruidl *et al.*, Multiple potential germ-line helicases are components of the germ-line-specific P granules of Caenorhabditis elegans. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 13837–13842 (1996).

106. I. Kawasaki *et al.*, The PGL family proteins associate with germ granules and function redundantly in Caenorhabditis elegans germline development. *Genetics*. **167**, 645–661 (2004).

107. E. J. Cram, H. Shang, J. E. Schwarzbauer, A systematic RNA interference screen reveals a cell migration gene network in C. elegans. *J. Cell Sci.* **119**, 4811–4818 (2006).

108. S. H. Kang, J. M. Kramer, Nidogen is nonessential and not required for normal type IV collagen localization in Caenorhabditis elegans. *Mol. Biol. Cell*. **11**, 3911–3923 (2000).

109. H. A. Wilkinson, I. Greenwald, Spatial and temporal patterns of lin-12 expression during C. elegans hermaphrodite development. *Genetics*. **141**, 513–526 (1995).

110. R. P. Johnson, S. H. Kang, J. M. Kramer, C. elegans dystroglycan DGN-1 functions in epithelia and neurons, but not muscle, and independently of dystrophin. *Development*. **133**, 1911–1921 (2006).

111. T. A. Starich, D. H. Hall, D. Greenstein, Two classes of gap junction channels mediate soma-germline interactions essential for germline proliferation and gametogenesis in Caenorhabditis elegans. *Genetics*. **198**, 1127–1153 (2014).

112. B. D. Ackley *et al.*, The NC1/endostatin domain of Caenorhabditis elegans type XVIII collagen affects cell migration and axon guidance. *J. Cell Biol.* **152**, 1219–1232 (2001).

113. S. A. Kostas, A. Fire, The T-box factor MLS-1 acts as a molecular switch during specification of nonstriated muscle in C. elegans. *Genes Dev.* **16**, 257–269 (2002).

114. H. Komatsu *et al.*, OSM-11 facilitates LIN-12 Notch signaling during Caenorhabditis elegans vulval development. *PLoS Biol.* **6**, e196 (2008).

115. C. W. Whitfield, C. Bénard, T. Barnes, S. Hekimi, S. K. Kim, Basolateral localization of the Caenorhabditis elegans epidermal growth factor receptor in epithelial cells by the PDZ protein LIN-10. *Mol. Biol. Cell*. **10**, 2087–2100 (1999).

116. I. Tcherepanova, L. Bhattacharyya, C. S. Rubin, J. H. Freedman, Aspartic proteases from the nematode Caenorhabditis elegans. Structural organization and developmental and cell-specific expression of asp-1. *J. Biol. Chem.* **275**, 26359–26369 (2000).

117. J. D. McGhee *et al.*, The ELT-2 GATA-factor and the global regulation of transcription in the C. elegans intestine. *Dev. Biol.* **302**, 627–645 (2007).

118. A. Patton *et al.*, Endocytosis function of a ligand-gated ion channel homolog in Caenorhabditis elegans. *Curr. Biol.* **15**, 1045–1050 (2005).

119. Y. Zhang *et al.*, Identification of genes expressed in C. elegans touch receptor neurons. *Nature*. **418**, 331–335 (2002).

120. T. Kawano *et al.*, C. elegans mig-6 encodes papilin isoforms that affect distinct aspects of DTC migration, and interacts genetically with mig-17 and collagen IV. *Development*. **136**, 1433–1442 (2009).

121. K. Kim, C. Li, Expression and regulation of an FMRFamide-related neuropeptide gene family in Caenorhabditis elegans. *J. Comp. Neurol.* **475**, 540–550 (2004).

122. P. J. Brockie, D. M. Madsen, Y. Zheng, J. Mellem, A. V. Maricq, Differential expression of glutamate receptor subunits in the nervous system of Caenorhabditis elegans and their regulation by the homeodomain protein UNC-42. *J. Neurosci.* **21**, 1510–1522 (2001).

123. S. Suo, Y. Kimura, H. H. M. Van Tol, Starvation induces cAMP response element-binding protein-dependent gene expression through octopamine-Gq signaling in Caenorhabditis elegans. *J. Neurosci.* **26**, 10082–10090 (2006).

124. T. Janssen *et al.*, Discovery of a cholecystokinin-gastrin-like signaling system in nematodes. *Endocrinology*. **149**, 2826–2839 (2008).

125. J. B. Rand, Acetylcholine. *WormBook*, 1–21 (2007).

126. E. M. Jorgensen, GABA. *WormBook*, 1–13 (2005).

127. R. Nass *et al.*, A genetic screen in Caenorhabditis elegans for dopamine neuron insensitivity to 6-hydroxydopamine identifies dopamine transporter mutants impacting transporter biosynthesis and trafficking. *J. Neurochem.* **94**, 774–785 (2005).

128. S. Suo, N. Sasagawa, S. Ishiura, Cloning and characterization of a Caenorhabditis elegans D2-like dopamine receptor. *J. Neurochem.* **86**, 869–878 (2003).

129. A. Oishi *et al.*, FLR-2, the glycoprotein hormone alpha subunit, is involved in the neural control of intestinal functions in Caenorhabditis elegans. *Genes Cells*. **14**, 1141–1154 (2009).

130. R. J. Hobson *et al.*, SER-7, a Caenorhabditis elegans 5-HT7-like receptor, is essential for the 5-HT stimulation of pharyngeal pumping and egg laying. *Genetics*. **172**, 159–169 (2006).

131. M. Furuya, H. Qadota, A. D. Chisholm, A. Sugimoto, The C. elegans eyes absent ortholog EYA-1 is required for tissue differentiation and plays partially redundant roles with PAX-6. *Dev. Biol.* **286**, 452–463 (2005).

132. P. Sengupta, J. H. Chou, C. I. Bargmann, odr-10 encodes a seven transmembrane domain olfactory receptor required for responses to the odorant diacetyl. *Cell*. **84**, 899–909 (1996).

133. C. O. Ortiz *et al.*, Searching for neuronal left/right asymmetry: genomewide analysis of nematode receptor-type guanylyl cyclases. *Genetics*. **173**, 131–149 (2006).

134. M. Lindemans *et al.*, A neuromedin-pyrokinin-like neuropeptide signaling system in Caenorhabditis elegans. *Biochem. Biophys. Res. Commun.* **379**, 760–764 (2009).

135. H. Inada *et al.*, Identification of guanylyl cyclases that function in thermosensory neurons of Caenorhabditis elegans. *Genetics*. **172**, 2239–2252 (2006).

136. K. Yamada *et al.*, Olfactory plasticity is regulated by pheromonal signaling in Caenorhabditis elegans. *Science*. **329**, 1647–1650 (2010).

137. O. Aurelio, D. H. Hall, O. Hobert, Immunoglobulin-domain proteins required for maintenance of ventral nerve cord organization. *Science*. **295**, 686–690 (2002).

138. A. Cornils, M. Gloeck, Z. Chen, Y. Zhang, J. Alcedo, Specific insulin-like peptides encode sensory information to regulate distinct developmental processes. *Development*. **138**, 1183–1193 (2011).

139. W. Li, S. G. Kennedy, G. Ruvkun, daf-28 encodes a C. elegans insulin superfamily member that is regulated by environmental cues and acts in the DAF-2 signaling pathway. *Genes Dev.* **17**, 844–858 (2003).

140. D. A. Birnby *et al.*, A transmembrane guanylyl cyclase (DAF-11) and Hsp90 (DAF-21) regulate a common set of chemosensory behaviors in caenorhabditis elegans. *Genetics*. **155**, 85–104 (2000).

141. J. F. Etchberger *et al.*, The molecular signature and cis-regulatory architecture of a C. elegans gustatory neuron. *Genes Dev.* **21**, 1653–1674 (2007).

142. S. Yu, L. Avery, E. Baude, D. L. Garbers, Guanylyl cyclase expression in specific sensory neurons: a new family of chemosensory receptors. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 3384–3387 (1997).

143. J. M. Gray *et al.*, Oxygen sensation and social feeding mediated by a C. elegans guanylate cyclase homologue. *Nature*. **430**, 317–322 (2004).

144. L. Emtage, G. Gu, E. Hartwieg, M. Chalfie, Extracellular proteins organize the mechanosensory channel complex in C. elegans touch receptor neurons. *Neuron*. **44**, 795–807 (2004).

145. X. Wang *et al.*, The C. elegans L1CAM homologue LAD-2 functions as a coreceptor in MAB-20/Sema2–mediated axon guidance. *J. Cell Biol.* **180**, 233–246 (2008).

Chapter 2: cell state characterization by single cell chromatin accessibility and transcriptome co-assay

*Modified from article Joint profiling of chromatin accessibility and gene expression in thousands of single cells, Junyue Cao, et al, Science, 2018

Authors: Junyue Cao[1,2], Darren A. Cusanovich[1†‡], Vijay Ramani[1†], Delasa Aghamirzaie[1], Hannah A. Pliner[1], Andrew J. Hill[1], Riza M. Daza[1], Jose L. McFaline-Figueroa[1], Jonathan S. Packer[1], Lena Christiansen[3], Frank J. Steemers[3], Andrew C. Adey[4, 5], Cole Trapnell[1,6,7,*], Jay Shendure[1,6,7,8,*]

Affiliations:

[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA

[2]Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA

[3]Illumina Inc., CA, USA

[4]Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR, USA

[5]Knight Cardiovascular Institute, Portland, OR, USA

[6]Allen Discovery Center for Cell Lineage Tracing, Seattle, WA, USA

[7]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA

[8]Howard Hughes Medical Institute, Seattle, WA, USA

†These authors contributed equally to this work

‡Present address: Asthma & Airway Disease Research Center and Department of Cellular and Molecular Medicine, The University of Arizona, Tucson, AZ, USA

*Correspondence to: coletrap@uw.edu (CT) & shendure@uw.edu (JS)

Abstract:

Although we can increasingly measure transcription, chromatin, methylation, etc. at single cell resolution, most assays survey only one aspect of cellular biology. Here we describe sci-CAR, a combinatorial indexing-based co-assay that jointly profiles chromatin accessibility and mRNA in each of thousands of single cells. As a proof-of-concept, we apply sci-CAR to 4,825 cells comprising a time-series of dexamethasone treatment, as well as to 11,296 cells from the adult mouse kidney. With the resulting data, we compare the pseudotemporal dynamics of chromatin accessibility and gene expression, reconstruct the chromatin accessibility profiles of cell types defined by RNA profiles, and link cis-regulatory sites to their target genes on the basis of the covariance of chromatin accessibility and transcription across large numbers of single cells.

One Sentence Summary:

We developed and applied sci-CAR to jointly profile the epigenome and transcriptome of thousands of single cells in systems including cortisol response and whole mouse kidney.

Introduction:

The concurrent profiling of multiple classes of molecules, *e.g.* RNA and DNA, within single cells has the potential to reveal causal regulatory relationships and to enrich the utility of organism-scale single cell atlases. However, to date, nucleic acid 'co-assays' rely on physically isolating each cell, limiting their throughput to a few cells per study (**Fig. S1A**) (*1–6*).

Result:

Single-cell combinatorial indexing ("sci") methods use split-pool barcoding to uniquely label the nucleic acid contents of single cells or nuclei (*7–13*). Here we describe sci-CAR, which jointly profiles single cell chromatin accessibility and mRNA in a scalable fashion. Sci-CAR

effectively combines sci-ATAC-seq and sci-RNA-seq into a single protocol (**Fig. 1**): (i) Nuclei are extracted, with or without fixation, and distributed to wells. (ii) A first RNA-seq 'index' is introduced by *in situ* reverse transcription (RT) with a poly(T) primer bearing a well-specific barcode and a unique molecular identifier (UMI). (iii) A first ATAC-seq index is introduced by *in situ* tagmentation with Tn5 transposase bearing a well-specific barcode. (iv) All nuclei are pooled and redistributed by FACS to multiple plates. (v) After second-strand synthesis of cDNA, nuclei in each well are lysed, and the lysate split to RNA and ATAC-dedicated portions. (vi) To provide a second priming site for amplification of 3' cDNA tags, the RNA-dedicated lysate is subjected to transposition with unindexed Tn5 transposase. 3' cDNA tags are amplified with primers corresponding to the Tn5 adaptor and RT primer. These primers also bear a well-specific barcode that is the second RNA-seq index. (vii) The ATAC-seq-dedicated lysate is amplified with primers specific to the barcoded Tn5 adaptors from step iii. These primers also bear a well-specific barcode that is the second ATAC-seq index. (viii) Amplicons from RNA-seq and ATAC-seq-dedicated lysates are respectively pooled and sequenced. Each sequence read is associated with two barcodes corresponding to each round of indexing. As with other sci- protocols, most nuclei pass through a unique combination of wells, receiving a unique combination of barcodes that can be used to group reads derived from the same cell. Because the barcodes introduced to RNA-seq and ATAC-seq libraries correspond to specific wells, we can link the mRNA and chromatin accessibility profiles of individual cells.

We applied sci-CAR to a cell culture model of cortisol response, wherein dexamethasone (DEX), a synthetic mimic of cortisol, activates glucocorticoid receptor (GR), which binds to thousands of locations across the genome, altering the expression of hundreds of genes (*14–17*). We collected lung adenocarcinoma-derived A549 cells after 0, 1 or 3 hrs of 100 nM DEX treatment, and performed a 96 x 576 well sci-CAR experiment. The three timepoints were each represented

in 24 wells during the first round of indexing, while the remaining 24 wells contained a mixture of HEK293T (human) and NIH3T3 (mouse) cells (**Fig. S1B**).

We obtained sci-RNA-seq profiles for 6,093 cells (median 3,809 UMIs) and sci-ATAC-seq profiles for 6,085 cells (median 1,456 unique reads) (**Fig. S1C-E**). For both data types, reads assigned to the same cell overwhelmingly mapped to one species (**Fig. S1F-G**). We obtained roughly equivalent UMIs per cell from 'RNA-only' plates processed in parallel, albeit at a lower sequencing depth per cell. Aggregated transcriptomes of co-assayed vs. RNA-only plates were well-correlated (r = 0.97-0.98; **Fig. S2**). In contrast, although co-assayed vs. 'ATAC-only plates' were comparable in quality and well-correlated in aggregate (**Fig. S3**), ATAC-only plates had ~10-fold higher complexity. The lower efficiency of the co-assay for ATAC is likely explained by factors including buffer modifications and our use of only half the lysate.

There were 4,825 cells (70% of either set) for which we recovered both transcriptome and chromatin accessibility data. To confirm that paired profiles truly derived from the same cells, we asked whether cells from mixed human-mouse wells were consistently assigned as human or mouse. Indeed, 1,423/1,425 (99%) of co-assayed cells from those wells were assigned the same species label from both sci-RNA-seq and sci-ATAC-seq profiles (**Fig. 2A**).

We next examined the time course of GR activation. DEX treatment of A549 cells increased both transcription and promoter accessibility of markers of GR activation, including *NFKBIA*, *SCNN1A*, *CKB*, *PER1* and *CDH16 (14, 16)* (**Fig. S4A-B**). Unsupervised clustering or t-SNE visualization of either sci-RNA-seq or sci-ATAC-seq profiles readily separated clusters corresponding to untreated and DEX-treated cells (**Fig. 2B-C**). Reassuringly, cells from co-assay plates and single-assay plates of either type were intermixed (**Figs. S4C**).

88% and 93% of co-assayed cells in clusters 1 and 2 of sci-ATAC-seq data were found in corresponding sci-RNA-seq clusters (**Fig. S4D-E**). Cells with concordant vs. discordant assignments did not significantly differ in read depth (*P*-value > 0.1, Welch two-sample t-test), but notably fell on the border between clusters 1 and 2 in either t-SNE (**Figs. 2D**, **S4F**). While most discordant cells (70%) were from 0 hrs, the remainder tended to derive from 1 hrs rather than 3 hrs (5% of 1 hr vs. 1% of 3 hr cells, *P*-value = 2.2e-16, Fisher's Exact Test). Although we cannot rule out that this is due to imperfect clustering, these discordantly assigned cells potentially reflect transitional states in GR activation.

Differential expression (DE) analysis of sci-RNA-seq data revealed significant changes in 2,613 genes (5% FDR). For comparison, a similar analysis with bulk RNA-seq data of DEX treatment in A549 cells at 0 vs. 3 hrs (*18*) identified 870 DE genes, 536 of which were also DE here. Log_2 fold changes were well-correlated between the datasets for DE genes (r = 0.86, **Fig. S4G**).

Differential accessibility (DA) analysis of sci-ATAC-seq profiles identified significant changes at 4,763 sites (5% FDR). For comparison, a similar analysis of bulk DNase-seq data from DEX-treated A549 cells at 0 vs. 3 hrs (*18*) identified 672 DA sites, 544 of were also DA here. Log_2 fold changes were well-correlated between the datasets for DA sites (rho = 0.68, **Fig. S4H**).

Of our DA sites, 701 (15%) were promoters, of which 175 overlapped with DE transcripts. Transcripts for genes with DA promoters that were not DE were detected in significantly fewer cells than genes with DA promoters that were DE (median 10% vs. 25%, *P*-value < 5e-5, unpaired two sample permutation test based on 20,000 simulations), suggesting we may be insufficiently powered to detect DE at many genes with DA promoters. For the 175 genes that are both DA and DE, the log_2 fold changes were modestly correlated (rho = 0.63, **Fig. S4I**), with 130/175 (74%) exhibiting directional concordance (exact two-sided binomial test, *P*-value = 9e-11).

We ordered cells along a pseudotime trajectory with Monocle (*19*) based on the top 1,000 DE genes (**Fig. S5A**). Cells were ordered consistently with the time course (**Fig. 2E**). Of note, the aforementioned cells from 1 hrs whose cluster assignments were discordant (**Figs. 2D**, **S4F**) occurred significantly earlier in pseudotime than cells with concordant assignments (*P*-value = 3e-5, Wilcoxon rank sum test, **Fig. S5B**). Of the 2,613 DE genes, 979 (37%) increased and 1,111 (43%) decreased in expression along pseudotime, while 523 (20%) exhibited transient changes (**Fig. S5C-D**, **Tables S2, S4**). We exploited the co-assay to examine the dynamics of chromatin accessibility across RNA-defined pseudotime, identifying opening (47%), closing (32%) and transient (21%) DA sites (**Fig. S5E**, **Tables S3, S5**). There were eleven genes that showed significant changes in *both* gene expression and promoter accessibility along pseudotime (5% FDR for both), with well-correlated dynamics (**Figs. 2F, S5F-H**).

We converted the (cell x site) matrix to a (cell x transcription factor (TF) motif) matrix, simply by counting occurrences of each motif in all accessible sites for each cell (*20*). The motifs of 91/399 (23%) of expressed TFs were DA across the treatment conditions (5% FDR) (**Tables S6-S7**). Where ChIP-seq data was available for the same time course (*18*), we observed consistent dynamics of increasing motif-associated accessibility (**Fig. S6A**) and TF binding to accessible sites (**Fig. S6B**). Motif accessibility dynamics across expression-defined pseudotime are summarized in **Fig. S6C**. The motif of the canonical glucocorticoid receptor *NR3C1* was the most activated, even though its expression decreased (**Figs. 2G**), consistent with its activation by recruitment from the cytosol rather than by increased expression. In contrast, *KLF9* is a direct target of GR activation via a feed forward loop (*21*). Consistent with this, we observe that both its expression and its motif accessibility increase along pseudotime (**Fig. 2G, Fig. S6D-E**).

Single-cell RNA sequencing studies have recently characterized the transcriptomes of diverse cell types represented in the mammalian kidney (*22–24*). However, little is known about

the epigenetic landscapes that underlie these cell type-specific gene expression programs. To investigate this, we isolated and fixed nuclei from whole kidneys of two 8-week male mice (**Fig. S7A**). From one sci-CAR experiment, we obtained sci-RNA-seq profiles for 13,893 nuclei (median 1,011 UMIs; **Fig. S7B**) and sci-ATAC-seq profiles for 13,395 nuclei (median 7,987 unique reads; **Fig. S7C**). There were 11,296 cells for which we recovered both transcriptome and chromatin accessibility profiles.

We compared sci-CAR transcriptomes with a recently published single cell RNA-seq dataset of the same tissue generated by Drop-seq (*24*). After correcting for gene length biases (Drop-seq is biased towards shorter transcripts, and sci-RNA-seq towards longer transcripts) aggregated transcriptomes were reasonably well correlated (r = 0.73, **Fig. S7D**). Semi-supervised clustering of 10,727 sci-CAR transcriptomes (>500 UMIs) identified 14 groups, ranging in size from 74 (0.7%) to 2,358 (22.0%) cells (**Figs. 3A, S7E-F**). Established markers identified nearly all cell types (**Fig. S8A-B**). The expression profiles of proximal tubule cells separate them into three subtypes including S1/S2 cells (*Slc5a12*+, *Gatm*+, *Alpl*+, *Slc34a1*+), S3 type 1 cells (*Slc34a1*+, *Atp11a*+), and S3 type 2 cells (*Atp11a*+, *Rnf24*+) (**Fig. S8C**) (*25*, *26*). The smallest cluster is positive for cell cycle progression markers (*Mki67* and *Cenpp*), and may represent an actively proliferating subpopulation (**Fig. S8D**) (*25*, *26*). Cell type proportions were well-correlated between replicate kidneys, with the exception of paranephric body adipocytes (1.2% vs. 0.4%), likely due to technical variation in kidney dissection as these reside superficial to the renal fascia (**Fig. S7E**).

We identified 8,774 genes that were DE across the 14 cell types (5% FDR), including 1,771 with >2-fold greater expression in the highest vs. second highest cell type (**Fig. S9A-B, Tables S8, S9**). New marker genes were identified, such as *Daam2* for renal pericytes and *Calcr* for collecting duct intercalated cell B (**Fig. S9C-D**) (*25*, *26*). We examined expression of solute carrier

transporters (SLCs), as these correspond to a principal function of the kidney. 208/345 (60%) of these were DE in subsets of renal tubule cell types, many corresponding to known and potentially novel reabsorption specificities (**Figs. 3B**, **S9E**).

We compared aggregated sci-CAR chromatin accessibility profiles with published bulk ATAC-seq data on adult mouse kidney (*18*), and found them to be reasonably well correlated (r = 0.75; **Fig. S10A-B**). Across all genes, aggregate promoter accessibility correlated with aggregate gene expression (rho = 0.26; **Fig. S10C**). Nonetheless, a significant challenge for single cell ATAC-seq data, relative to single cell RNA-seq data, is the sparsity of the resulting matrices (*8*). Thus, our initial efforts to cluster co-assayed cells based solely on their ATAC-seq profiles failed to discover the expected diversity of cell types. We therefore sought to leverage the co-assay aspect of these data to recover the chromatin landscapes of individual cell types.

As a first approach, we simply annotated cell types from transcriptional profiles for ~96% of the 11,296 cells that were successfully co-assayed. We then aggregated ATAC-seq signal for each cell type separately, followed by peak calling (*27*). As a second approach, we also developed an algorithm to combine the ATAC-seq profiles of cells with highly similar RNA-seq profiles prior to clustering (**Fig. S7A**). For cells from each RNA-seq-defined cell type, we identified subsets of cells with highly similar expression profiles (a mean of 50 cells assigned to each of 222 'pseudo-cells'). We then aggregated the ATAC-seq profiles of each pseudo-cell, and performed t-SNE on these. In contrast with single-cell ATAC-seq data, pseudo-cell chromatin accessibility profiles corresponding to the same cell types clustered together (**Fig. 3C**). Overall, these analyses illustrate how co-assay data can be leveraged to overcome the relative sparsity of single cell ATAC-seq data and define chromatin accessibility profiles even for closely related cell types.

We identified 22,026 DA sites across the 14 mouse kidney cell types, including 2,096 promoters and 19,930 distal sites (5% FDR; **Figs. 3D, S10D-E**; **Tables S11, S12**). In some cases,

DA at a gene's promoter was concordant with DE (**Fig. S11A-B**), but this was the exception rather than the rule. Out of 2,096 genes with a DA promoter in at least one cell type, 132 genes were also DE (1% FDR) with a >2-fold difference between the first and second ranked cell type. Although promoter accessibility and expression of these genes across cell types are positively correlated (median rho = 0.17), the majority (112/132 or 85%) exhibited maximal promoter accessibility and gene expression in different cell types (**Fig. S11C**). The relatively weaker correlation compared with what we observed in the A549 dexamethasone time series (rho = 0.63; **Fig. S4I**) is potentially a consequence of the fact that in the A549 cells, we were comparing changes in promoter accessibility vs. expression, whereas here we are comparing absolute enrichment of accessibility at promoters vs. expression.

We sought to link distal *cis*-regulatory elements to their target genes based on the covariance of chromatin accessibility and gene expression across large numbers of co-assayed cells. As the sparsity of our single cell profiles makes this challenging, we worked with the aforedescribed 222 pseudo-cells (**Fig. S12A**). For each gene, we computed correlations between its expression and the adjusted accessibility of all sites within 100 kilobases (kb) of its transcriptional start site (TSS) using LASSO (least absolute shrinkage and selection operator).

Within the top 2,000 DE genes (ranked by q-value), we linked 1,260 distal sites to 321 genes (median 3 sites per gene, out of median 19 sites within 100 kb of TSS tested; **Fig. S12BC**). 44% of sites were linked to the nearest TSS, and 21% to the second nearest TSS (**Fig. S12D**). Distal site-gene linkages were significantly closer than all possible pairs tested (mean 41 kb for links vs. 48 kb for all pairs tested; *P*-value < 5e-5, unpaired permutation test based on 20,000 simulations; **Fig. S12E**).

To evaluate the possibility that the links were artifacts of regularized regression, we permuted the sample IDs of the chromatin accessibility matrix and performed the same analysis.

After this permutation, only 4 links were identified (**Fig. S12B**). To control for correlations between closely located accessible sites in the genome, we separately permuted the peak IDs. This yielded 216 links, or just 17% as many links as without permutation (**Fig. S12B**).

The 321 genes with linked distal sites were specifically expressed in a variety of cell types (**Fig. S12F**). For example, the link with the highest correlation is between distal convoluted tubule cell marker gene Slc12a3 and a site 36 kb downstream of its TSS and overlapping its last exon (**Fig. S13**). The accessibility of this linked site was modestly more specific to distal convoluted tubule cells than the Slc12a3 promoter. In contrast, the accessible site closest to the Slc12a3 promoter (only 216 bp away) was not linked to the Slc12a3 promoter by our approach, nor is its accessibility specific to distal convoluted tubule cells. Similarly, a marker gene for Loop of Henle cells, Slc12a1, is linked to two distal sites (**Fig. S14**), both of which exhibit accessibility specific to Loop of Henle cells. In contrast, the nearest accessible site (9 kb from the TSS), which was not linked, does not exhibit this specificity.

Links between distal *cis*-regulatory elements and their target genes can be useful for explaining differential expression across cell types. For example, the cell type-specific expression of Slc6a18, a marker gene for type 2 proximal tubule S3 cells, is not mirrored by cell type-specific promoter accessibility (**Fig. S11C**). However, from our covariance approach, its TSS is linked to a site 16 kb away whose accessibility is correlated with Slc6a18 expression (**Fig. 4A**). To quantify the utility of the links between distal *cis*-regulatory elements and their target genes identified from sci-CAR data, we constructed a linear regression model to predict gene expression differences based on chromatin accessibility at promoters only vs. promoters together with linked distal sites. Including linked distal sites improved predictions by four-fold (P-value < 5e-5, paired permutation test based on 20,000 simulations; **Fig. 4B**).

Our analyses illustrate the advantages of a single cell co-assay over assays that solely profile transcription or chromatin accessibility. Sci-CAR is compatible with fresh or fixed nuclei, and like other sci-seq techniques, can encode multiple samples per experiment. Its throughput can potentially be increased by additional rounds of split-pool indexing (*13*). With 384 x 384 x 384 sci-CAR, one could potentially co-assay millions of single cells per experiment. A limitation of sci-CAR is the sparsity of the resulting data, particularly with respect to chromatin accessibility. This can potentially be overcome in the future through protocol optimizations, particularly of crosslinking conditions. A second limitation is that although we were able to link distal elements and target genes on the basis of covariance of accessibility and expression, these data remain correlative and involve a minority of DE genes and DA elements.

Notwithstanding these limitations, sci-CAR expands the potential of combinatorial indexing for scalably profiling single cell molecular phenotypes, and may be particularly useful in the context of organism-scale single cell atlases. With further development, we anticipate that additional DNA/RNA co-assays may be realized by simply integrating other sci-seq protocols together with sci-RNA-seq (*e.g.* methylation + transcripts; chromosome conformation + transcripts; DNA sequence + transcripts) (*8–13*). A longer-term goal is to adapt single cell combinatorial indexing to span the Central Dogma, such that aspects of DNA, RNA and protein species can be concurrently assayed from each of many single cells.

Acknowledgements: We thank members of the Shendure and Trapnell labs for helpful discussions and feedback, particularly B. Martin, X. Qiu, A. Leith, A. Minkina, Y. Yin, Z. Duan and R. Qiu; as well as R. Hunter, and R. Rualo in the Transgenic Resources Program of University of Washington for their exceptional assistance.

Competing interests: L.C. and F.J.S. declare competing financial interests in the form of stock ownership and paid employment by Illumina, Inc. One or more embodiments of one or more patents and patent applications filed by Illumina may encompass the methods, reagents, and data disclosed in this manuscript.

Data and materials availability: Processed and raw data can be downloaded from NCBI GEO (GSE117089). All methods for making the transposase complexes are described in (*7*); however, Illumina will provide transposase complexes in response to reasonable requests from the scientific community subject to a material transfer agreement.

Funding: This work was funded by the Paul G. Allen Frontiers Foundation (Allen Discovery Center grant to JS and CT), grants from the NIH (DP1HG007811 and R01HG006283 to JS; DP2 HD088158 to CT; R35GM124704 to AA), the W. M. Keck Foundation (to CT and JS), the Dale. F. Frey Award for Breakthrough Scientists (to CT), the Alfred P. Sloan Foundation Research Fellowship (to CT), and the Brotman Baty Institute for Precision Medicine. DAC was supported in part by T32HL007828 from the National Heart, Lung, and Blood Institute. JS is an Investigator of the Howard Hughes Medical Institute.

Author contributions: J.S. and C.T. designed and supervised the research; J.C. developed technique and performed experiments with assistance from D.C., V.R., R.D., J.M., L.C., F.S. and A.A.; J.C. performed computation analysis with assistance from D.C., V.R., D.A., H.P., A.H. and J.P.; J.S., C.T. and J.C. wrote the paper.

FIGURES

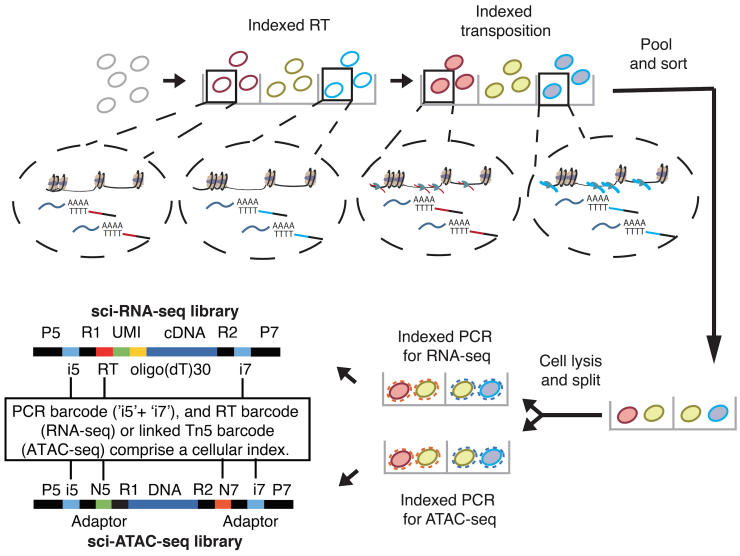


Fig. 1. sci-CAR workflow. Key steps outlined in text. RNA-seq: index2 and read1 cover the i5 index, UMI and RT barcode; index1 and read2 cover the i7 index and cDNA fragment. ATAC-seq: read1 and read2 cover genomic DNA sequence. Index 1 and index 2 cover the Tn5 and PCR barcodes.

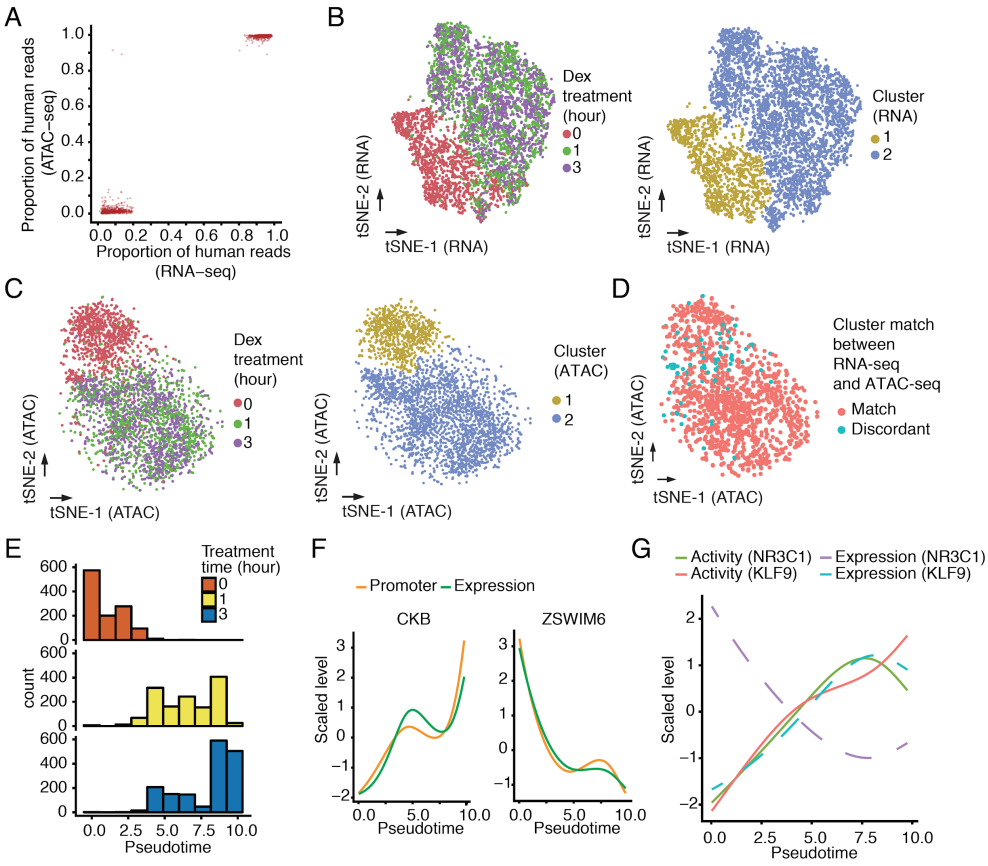


Fig. 2. Joint profiling of chromatin accessibility and transcription in dexamethasone treated A549 cells. (**A**) Scatter plot showing the proportion of human reads, out of all reads mapping uniquely to the human or mouse reference genomes, for cells in which both RNA-seq profiles and ATAC-seq profiles were obtained. Only HEK293T (human) and NIH/3T3 (mouse) cells are plotted. (**B**) t-SNE visualization of A549 cells (RNA-seq) including cells from both sci-CAR and sci-RNA-seq-only plates, colored by DEX treatment time (left) or unsupervised clustering id (right). (**C**) t-SNE visualization of A549 cells (ATAC-seq) including cells from both sci-CAR and sci-ATAC-seq-only plates, colored by DEX treatment time (left) or unsupervised clustering id (right). (**D**) t-SNE visualization of A549 cells (ATAC-seq) with linked RNA-seq profiles. If the cell is in cluster 1 (or cluster 2) in both RNA-seq and ATAC-seq, then it is labeled as "Match", otherwise it is labeled "Discordant". (**E**) Distribution of cells from different DEX treatment

timepoints in gene expression pseudotime inferred by trajectory analysis. (**F**) Smoothed line plot showing scaled (with the R function scale) gene expression and promoter accessibility of *CKB* and *ZSWIM6* across pseudotime. Unscaled, unsmoothed data shown in Fig. S5F-G. (**G**) Smoothed line plot showing the scaled mRNA level and activity change of transcription factors *NR3C1* and *KLF9* across pseudotime. Unscaled, unsmoothed data shown in Fig. S6D-E.

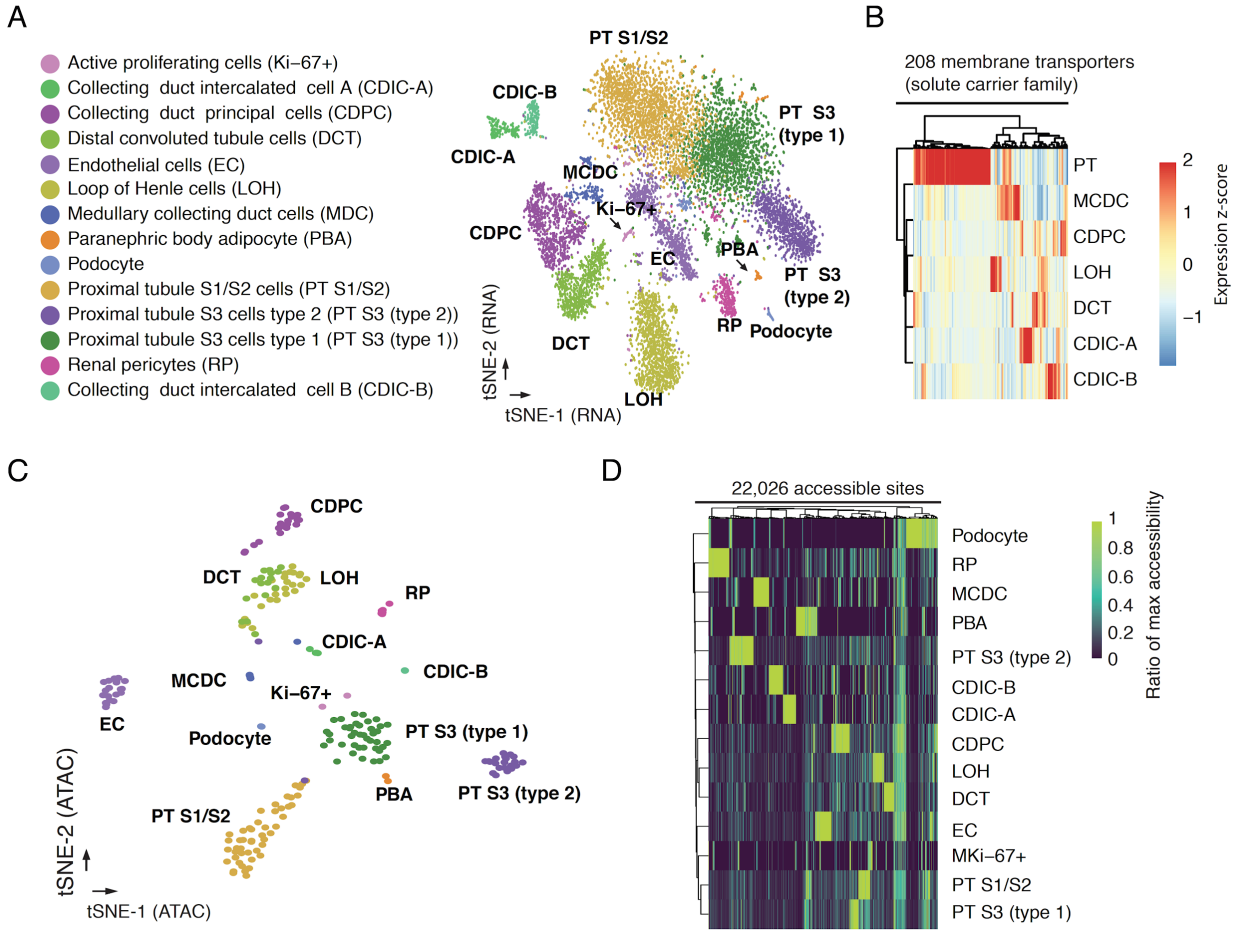


Fig. 3. sci-CAR enables joint profiling of chromatin accessibility and transcription in mouse kidney. (**A**) t-SNE visualization of mouse kidney nuclei (RNA-seq). Cell types are assigned based on established marker genes. (**B**) Heatmap showing the relative expression of genes from the solute carrier group of membrane transport proteins in consensus transcriptomes of each cell type

estimated by RNA-seq data from the co-assay. The raw expression data (UMI count matrix) was log-transformed, column centered and scaled (using the R function scale), and the resulting values clamped to [-2, 2]. (**C**) t-SNE visualization of mouse kidney nuclei (ATAC-seq) after aggregating cells with highly similar transcriptomes ('pseudocells'), colored by cell types identified from RNA-seq. (**D**) Heatmap showing the relative chromatin accessibility of cell type-specific sites for each cell type estimated by ATAC-seq data from the co-assay. The raw aggregated ATAC-seq data (read count matrix) was normalized first by the total number of reads for each cell type then by the maximum accessibility score across all cell types.

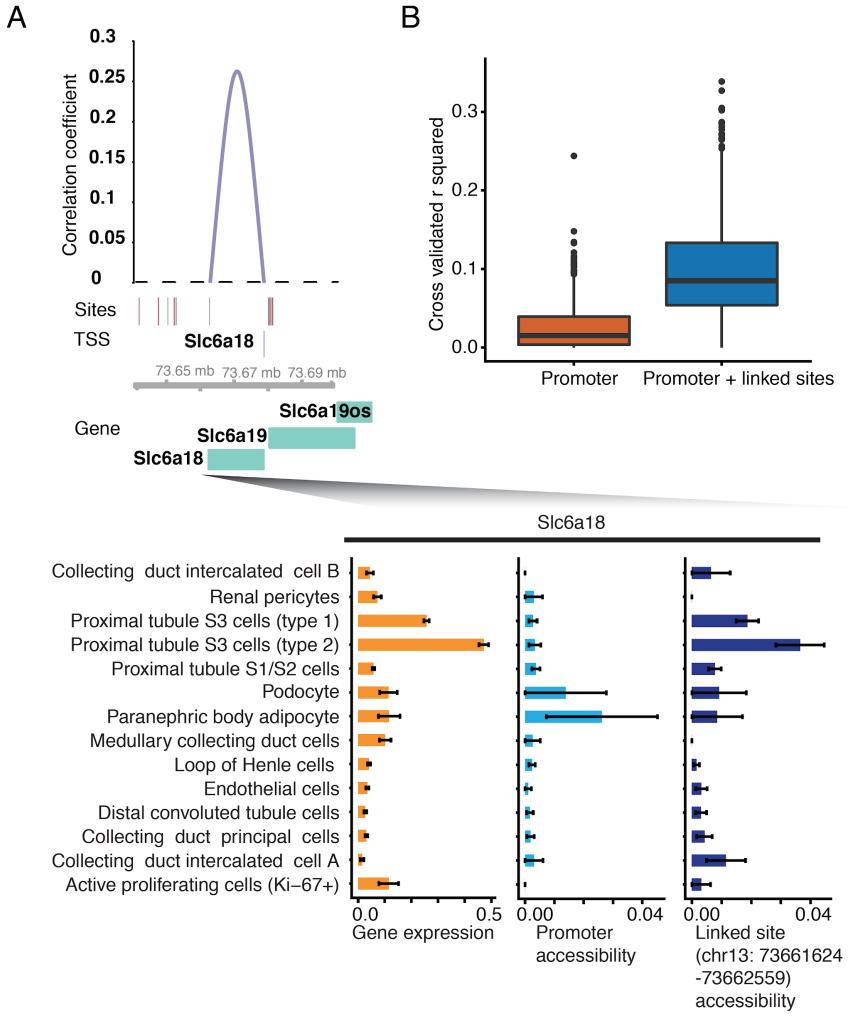


Fig. 4. Linking cis-regulatory elements to regulated genes based on covariance in single cell co-assay data. (**A**) Top: genome browser plot showing links between accessible distal regulatory sites and the gene Slc6a18. The height corresponds to the correlation coefficient. Bottom: barplots showing the average expression, promoter accessibility and linked site accessibility for cell type-specific marker gene Slc6a18 across different cell types. Gene expression values for each cell were calculated by dividing the raw UMI count by cell-specific size factors. Site accessibilities for each cell were calculated by dividing the raw read count by cell-specific size factors. Error bars represent standard errors of the means. (**B**) Two linear regression models were built to predict gene expression differences between cell types. The first model predicts changes on the basis of promoter accessibility alone. The second model predicts changes based on the chromatin accessibility of the promoter and distal sites that are linked to it. The boxplot shows the cross-validated r-squared calculated for each gene from the two models.

Supplementary Materials:

Materials and Methods

Figures S1-S14

References

Supplementary Materials:

Materials and Methods:

Mammalian cell culture

All mammalian cells were cultured at 37°C with 5% CO_2, and were maintained in high glucose DMEM (Gibco cat. no. 11965) for HEK293T and NIH/3T3 cells or DMEM/F12 medium for A549 cells, both supplemented with 10% FBS and 1X Pen/Strep (Gibco cat. no. 15140122; 100U/ml penicillin, 100 µg/ml streptomycin). Cells were trypsinized with 0.25% typsin-EDTA (Gibco cat. no. 25200-056) and split 1:10 three times per week.

Mouse tissues

All animal experiments were approved by the University of Washington, Institutional Animal Care and Use Committee. Two sacrificed mice (male, wild type, B6 background, 8 weeks) were purchased from the preclinical research and translational services core at the University of Washington. Kidney tissues were dissected and flash frozen in liquid nitrogen separately.

Sample processing for cultured cells

A549 cells were treated with 100 nM DEX for 1 hrs or 3 hrs before harvest. Control cells were treated with same volume of EtOH. All cell lines (A549, HEK293T and NIH/3T3 cells) were

trypsinized, spun down at 300x**g** for 5 min (4°C) and washed once in 1X ice-cold PBS. 5M cells were combined and lysed using 1 mL ice-cold cell lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2 and 0.1% IGEPAL CA-630 from (*28*), modified to also include 1% SUPERase In RNase inhibitor and 1% BSA). The isolated nuclei were then pelleted, washed twice with 500 µL cold lysis buffer without IGEPAL CA-630, and resuspended in lysis buffer without IGEPAL CA-630 at a final concentration of 5,000 nuclei/µL. For all washes, nuclei were pelleted by centrifugation at 500x**g** for 5 min (4°C). Nuclei were then distributed into one 96-well plate. For each well, 5,000 nuclei (2 µL) were mixed with 1 µl of 25 µM anchored oligo-dT primer (5′-ACGACGCTCTTCCGATCTNNNNNNNN[10bp index]TTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3′, where "N" is any base and "V" is either "A", "C" or "G"; IDT) and 0.25 µL 10 mM dNTP mix (Thermo), denatured at 55°C for 5 min and immediately placed on ice. 1.75 µL of first-strand reaction mix, containing 1 µL 5X Superscript IV First-Strand Buffer (Invitrogen), 0.25 µl 100 mM DTT (Invitrogen), 0.25 µl SuperScript IV reverse transcriptase (200 U/µl, Invitrogen), 0.25 µL RNaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen), was then added to each well. Reverse transcription was carried out by incubating plates at the following temperature gradient: 4°C 2 minutes, 10°C 2 minutes, 20°C 2 minutes, 30°C 2 minutes, 40°C 2 minutes, 50°C 2 minutes and 55°C 10 minutes. 3 µL cell lysis buffer without IGEPAL CA-630, 10 µL Nextera TD buffer (Illumina) and 1 to 2 µL indexed TDE1 enzyme (Illumina) were added to each well. Tagmentation was performed at 55°C for 30 min and stopped by adding 20 µl 2X stop solution (40 mM EDTA, 1 mM spermidine) to each well. All cells (or nuclei) were then pooled, stained with 4',6-diamidino-2-phenylindole (DAPI, Invitrogen) at a final concentration of 3 µM, and sorted at 25 nuclei per well into 5 µL EB buffer. Cells were gated based on DAPI stain such that singlets were discriminated from doublets and sorted into each well. 0.66 µl mRNA Second Strand Synthesis buffer (NEB) and 0.34 µl

mRNA Second Strand Synthesis enzyme (NEB) were then added to each well, and second strand synthesis was carried out at 16°C for 180 min. The reaction was stopped by adding 6 μL DNA binding buffer (Zymo) and incubating at room temperature for 5 min. Each well was then purified using 24 uL AMPure XP beads (Beckman Coulter), eluted in 12.5 μL of buffer EB (Qiagen). Eluted product was split (6 μL each) and transferred to two new 96-well plates.

For one plate (RNA-seq dedicated portion), each well was mixed with 5 μL Nextera TD buffer (Illumina) and 1 μL i7 only TDE1 enzyme (25 nM, Illumina), and then incubated at 55°C for 5 min to carry out tagmentation. After tagmentation, each well was mixed with 0.4 μL 1% SDS, 0.4 μL BSA (NEB), and 2 μL of 10 μM P7 primer (5′-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3′, IDT), and incubated at 55°C for 15 min. Then, to each well, we added 2 μL 10% Tween-20, 1.2 μL nuclease-free water, 2μL of 10 μM indexed P5 primer (5′-AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′; IDT), and 20 μL NEBNext High-Fidelity 2X PCR Master Mix (NEB). Amplification was carried out using the following program: 72°C for 5 min, 98°C for 30 sec, 18-22 cycles of (98°C for 10 sec, 66°C for 30 sec, 72°C for 1 min) and a final 72°C for 5 min. For the 'sci-RNA-seq only' branch of the workflow, after second strand synthesis, instead of cell lysis and AMPure bead purification, all cells were used for tagmentation and the subsequent PCR reaction, using the same procedure as the co-assay. After PCR, samples were pooled and purified using 0.8 volumes of AMPure XP beads. Library concentrations were determined by Qubit (Invitrogen) and the libraries were visualized by electrophoresis on a 6% TBE-PAGE gel. All RNA-seq libraries were sequenced on the NextSeq 500 platform (Illumina) using a V2 75 cycle kit (Read 1: 18 cycles, Read 2: 52 cycles, Index 1: 10 cycles, Index 2: 10 cycles). The co-assay RNA-seq library was

sequenced to saturation (~100,000 reads per cell) and the sci-RNA-seq library was sequenced to ~16,000 reads per cell.

For the other plate (ATAC-seq dedicated portion), each well was mixed with 1 μL of 10 μM indexed P5 primer (5′-AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTC-3′; IDT), 1 μL of 10 μM P7 primer (5′-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3′, IDT), 2 μL nuclease-free water, and 10 μL NEBNext High-Fidelity 2X PCR Master Mix (NEB). Amplification was carried out using the following program: 72°C for 3 min, 98°C for 30 sec, 18-22 cycles of (98°C for 10 sec, 63°C for 30 sec, 72°C for 1 min) and a final 72°C for 5 min. For the 'sci-ATAC-seq only' branch of the workflow, sorted cells were processed similarly to the published method (*29*) with minor modifications: each well was mixed with 7 μL buffer EB (Qiagen), 0.4 μL 1% SDS, 0.4 μL BSA (NEB), and 2 μL of 10 μM P7 primer (5′-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3′, IDT), and incubated at 55°C for 15 min. Then, to each well, we added 2 μL 10% Tween-20, 1.2 μL nuclease-free water, 2 μL of 10 μM indexed P5 primer (5′-AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTC-3′; ID), and 20 μL NEBNext High-Fidelity 2X PCR Master Mix (NEB). The amplification procedure was the same for 'ATAC-seq only' and for co-assay plates. After PCR, all samples were pooled and purified using 1 volume of AMPure XP beads. Amplified libraries were then gel extracted to remove amplicons shorter than 200 bp, presumably primer dimers. Library concentrations were determined by Qubit (Invitrogen) and the libraries were visualized by electrophoresis on a 6% TBE-PAGE gel. Libraries were sequenced on a NextSeq 500 platform (Illumina) using a mid-output 300 cycle kit and a custom recipe (paired end 51 bp reads with index reads that covered both the Tn5 barcode

and the library amplification barcode) and custom primers as previously described in (*29*). ATAC-seq libraries was sequenced to ~55,000 reads per cell.

Mouse kidney nuclei extraction and fixation

Mouse kidney was minced to small pieces by blade in cell lysis buffer and transferred to the top of a 40 um cell strainer (Falcon). Kidney tissues were homogenized with the rubber tip of a syringe plunger (5 ml, BD) in 4ml cell lysis buffer. The filtered nuclei were then transferred to a new 15 ml tube (Falcon) and pelleted by centrifuge at 500x**g** for 5 min at 4°C and washed once with 1 ml ice-cold cell lysis buffer. The nuclei were fixed in 4 ml ice cold 4% paraformaldehyde (EMS) for 15 min on ice. After fixation, the nuclei were washed twice in 1 ml nuclei wash buffer (cell lysis buffer without IGEPAL), and re-suspended in 500 ul nuclei wash buffer. The samples were split to 5 tubes with 100 ul in each tube and flash frozen in liquid nitrogen.

Sample processing with fixed nuclei

Fixed nuclei (by paraformaldehyde) were processed via similar steps to the cultured cells with minor modifications: Thawed nuclei were permeabilized with 0.2% tritonX-100 for 3 minutes on ice, then washed twice with nuclei wash buffer. After sorting, each well was mixed with 6 μL EB buffer, 0.5 μL 1% SDS, 0.5 μL protease K (Qiagen), and incubated at 65°C for 16 hours to reverse crosslinking. 2 μL 10% tween-20 was added to each well, and then 6 μL of mix was transferred to a fresh 96-well plate for ATAC-seq library preparation: each well was mixed with 2 μL of 10 μM indexed P5 primer (5′-AATGATACGGCGACCACCGAGATCTACAC[i5]TCGTCGGCAGCGTC-3′; IDT), 2 μL of 10 μM P7 primer (5′-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3′, IDT), 8 μL nuclease-free water, 1 μL 10% tween-20, 1 μL BSA (NEB) and 20 μL NEBNext High-

Fidelity 2X PCR Master Mix (NEB). The amplification program and subsequent steps for ATAC-seq were the same as with fresh nuclei.

The cell lysate for RNA-seq library preparation was purified using 2 volumes of AMPure XP beads and eluted in 5 μL elution buffer. 0.66 μl mRNA Second Strand Synthesis buffer (NEB) and 0.34 μl mRNA Second Strand Synthesis enzyme (NEB) were then added to each well, and second strand synthesis was carried out at 16°C for 180 min. Tagmentation and following steps were the same as with fresh nuclei.

Read alignments and downstream processing

Read alignment and gene count matrix generation for the single cell RNA-seq aspect of the co-assay was performed using the pipeline that we developed for sci-RNA-seq (*13*) with minor modifications. Reads were first mapped to a reference genome with STAR/v2.5.2b (*30*), with gene annotations from GENCODE V19 for human, and GENCODE VM11 for mouse. For experiments with A594, HEK293T and NIH/3T3 cells, we used an index combining chromosomes from both human (hg19) and mouse (mm10). For the mouse kidney experiment, we used mouse genome build mm10.

For the mixed-species experiments (HEK293T, NIH/3T3, A549 cells), the number of UMIs and the percentage of mapping reads for genomes of each species was calculated based on uniquely mapped reads after removing duplicates. Cells with over 80% of UMIs assigned to one species were regarded as species-specific cells, with the remaining cells classified as mixed cells or "collisions". Barcodes with more than 1,000 UMIs were deemed to correspond to cells, while those with fewer than 1,000 were excluded from further analysis. We recovered transcriptomes from 4,277 A549 cells, 812 HEK293T cells, 868 NIH3T3 cells, and 136 human-mouse 'collisions', i.e. transcriptomes that derive from multiple cells that traversed through the same combination of wells.

For comparison, we obtained a median of 3,331 UMIs (1,524 genes) per A549 cell from a 'sci-RNA-seq only' version of the workflow performed in parallel, albeit at a much lower sequencing depth per cell.

For the mouse kidney experiment, the number of UMIs was calculated based on number of reads strand-specifically mapped to gene exons or introns after removing duplicates. In total, we obtained sci-RNA-seq profiles for 13,893 cells (min. 200 UMIs per cell).

To process reads from single cell ATAC-seq, base calls were first converted to fastq format using Illumina's bcl2fastq/2.16.0.10. Reads were demultiplexed according to PCR barcodes using a custom python script that tolerated one mismatched base in the barcode. The indexed Tn5 barcode for each reads was corrected to its nearest barcode (edit distance (ED) < 2) and reads with uncorrected barcodes (ED > 2) were removed. Demultiplexed reads were then adaptor-clipped using trim_galore/0.4.1 with default settings. Trimmed reads were mapped to a chimeric reference genome of human (hg19) and mouse for ATAC-seq from A594, HEK293T and NIH/3T3 cells, and mouse (mm10) only for ATAC-seq from mouse kidney experiment with STAR/v2.5.2b (*30*). Mapped reads were split into constituent cellular indices by further demultiplexing reads using the Tn5 index (ED < 2, including insertions and deletions). Duplicates for each cell were removed by samtools/v1.3 (*31*).

For the mixed-species experiment, ATAC barcodes with more than 200 unique reads were identified as real cells, and those with fewer than that discarded. Cells with over 80% of unique reads assigned to one species were regarded as species-specific cells, with the remaining cells classified as mixed cells or "collisions". The 6,085 cells for which we recovered chromatin accessibility profiles included 4,258 A549 cells, 868 HEK293T cells, 877 NIH3T3 cells, and 82 human-mouse collisions. Identified HEK293T and NIH3T3 cells from RNA-seq and ATAC-seq

were extracted, and cells with linked profiles were used to calculate the ratio of co-assay cells assigned with the same species label. Excluding clear interspecies collisions, we obtained a median of 1,307 unique reads per HEK293T cell, 1,065 unique reads per NIH3T3 cell, and 1,566 unique reads per A549 cell, on par with results from the original sci-ATAC-seq protocol (*8*). The sci-ATAC-seq data exhibited higher 'species purity' than the sci-RNA-seq data, with a median of 0.5% (HEK293T), 1.0% (NIH3T3) and 0.4% (A549) of reads mapping to the incorrect species. The aggregate fragment length distribution exhibited the nucleosome periodicity characteristic of ATAC-seq data (**Fig. S3BC**). When sci-RNA/ATAC-seq data are compared to data from a 'sci-ATAC-seq only' version of the workflow performed in parallel, a similar proportion of reads mapped to accessible sites (**Fig. S3A**), but with a substantially lower library complexity for the joint protocol (sci-ATAC-seq only: median 13,144 unique reads per cell, sci-ATAC-seq in sci-CAR: median of 1,456 unique reads per cell). In the mouse kidney experiment, we obtained sci-ATAC-seq profiles for 13,395 nuclei (min. 300 unique reads intersecting with accessible sites per cell).

Analysis of sci-RNA-seq data from the A549 experiment

A digital gene expression matrix was constructed from the raw sequencing data as described above. Cells from both 'sci-RNA-seq only' and co-assay plates were combined. Cells with fewer than 500 UMIs or more than 9,100 UMIs were discarded. The dimensionality of the the data was reduced with t-SNE (initialized by projecting each cell onto the top 10 principal components), followed by unsupervised clustering via the densityPeak algorithm implemented in Monocle 2.6.3 (*19*).

To calculate the change in each gene's expression between cells from different treatment conditions, its UMI count in each cell was first divided by that cell's library size factor (as

computed by Monocle). Mean (size-factor-normalized) expression and standard error was calculated across all cells in each treatment condition.

Cells were organized into pseudotemporal trajectories according to their RNA-seq profiles using Monocle 2.6.3 (*19*). Briefly, differentially expressed genes across three treatment conditions were identified with the differentialGeneTest() function of Monocle 2 (*19*). As cells with low RNA-seq signal showed higher noise in trajectory analysis, cells with low numbers of genes detected (<= 1000 genes) were filtered out. After ranking genes by significance (as reported by differentialGeneTest), the top 1,000 most significant genes were used to construct the pseudotime trajectory (*19*), with log-transformed total UMI counts per cell as a covariate in the tree construction. Each cell was assigned a pseudotime value based on its position along the trajectory tree. Smoothed gene expression kinetics were visualized using the plot_genes_in_pseudotime function in Monocle 2 (*19*). Cells in the trajectory were grouped in the same method as (*32*). Briefly, cells were grouped first at similar positions in pseudotime by k-means clustering along the pseudotime axis (k = 10). These clusters were subdivided into groups containing at least 50 and no more than 100 cells. We then aggregated the transcriptome, chromatin accessibility and transcription factor activity profiles of cells within each group. The gene expression (chromatin accessibility or TF activity) along pseudotime was calculated in the same approach as (*32*). Briefly, genes passing significant test (5% FDR) across different treatment conditions were selected and a natural spline was used to fit the gene expression along pseudotime, with mean_number_genes included as a covariate. The gene expression for each gene was subtracted by the lowest expression and then divided by the highest expression. Genes with max expression within the earliest 25% of pseudotime were labeled as activated genes. Genes with max expression in the last 25% of pseudotime were labeled as repressed genes. Other genes were labeled as transient genes.

Analysis of sci-ATAC-seq data from the A549 experiment

Single cell ATAC-seq profiles were generated as described above. To define peaks of accessibility across all sites, we used MACS/v2.1.1(*27*). ATAC-seq reads of cells from different treatment conditions were aggregated and peaks were called on each group separately. These peaks were then merged with bedtools (*33*), together with gene promoter regions (annotated transcription start site (TSS) in GENCODE V19 for human and GENCODE VM11 for mouse minus 500 base pairs in strand specific manner). Each read alignment was extended by 50 bp upstream and downstream from the insertion site of tagmentation. Cells were determined to be accessible at a given peak if a read from a cell overlapped with the peak. The peak count matrix was generated by custom python script with the HTseq package (*34*).

Dimensionality reduction and t-SNE analysis for ATAC-seq data from A549 cells was performed similarly to (*32*) with minor modifications. Briefly, A549 cells from both ATAC-seq only and co-assay were combined and the peak count matrix was binarized. Low signal cells (peak count < 300) were filtered out. Peaks within 1 kb were merged and reads in merged peaks were aggregated to generate a matrix with merged peaks. The dimension of the data was reduced with t-SNE (initialized using the top 15 PCs). Cells were then clustered using the unsupervised densityPeak algorithm implemented in Monocle 2.6.3 (*19*). Differentially accessible peak across different treatment conditions in A549 cells were calculated by likelihood ratio test similar to (*32*).

To calculate the change in chromatin accessibility between cells from different treatment conditions, a site's accessibility in each cell was calculated by dividing the cell's raw read count by cell specific size factor estimated by estimateSizeFactors function in Monocle 2 (*19*). Mean chromatin accessibility and standard error were calculated based on all cells in each treatment condition.

Chromatin accessibility changes along pseudotime were calculated similarly with gene expression. Briefly, cells with linked transcriptome and chromatin accessibility profiles were selected, and assigned with the same pseudotime and sub-group identified in trajectory analysis of RNA-seq. ATAC-seq profiles (after merging nearby peaks) from cells were aggregated for each sub-group. Sites detected in more than 5 cells and significant accessible sites across different treatment conditions were selected. A natural spline was used to fit the site accessibility along pseudotime, with mean number of peaks detected per cell as a covariate. The chromatin accessibility for each site was subtracted by the lowest accessibility and then divided by the highest signal. Sites with max accessibility within 20% to 80% of pseudotime were labeled as transient sites. For the remaining sites, the sites with higher average accessibility in the earliest 20% vs. the latest 20% of pseudotime were labeled as opening sites. Otherwise, they were labeled as closing sites.

To select the differentially accessible promoters along pseudotime, we first selected the genes that were differentially expressed (5% FDR, likelihood ratio test) across pseudotime, and then fitted a negative-binomial regression model to promoter accessibility of selected genes, with the number of peaks detected per cell as a covariate. Differential accessible scores (p-value) for selected genes were calculated with the likelihood ratio test by comparing two models with or without pseudotime included as covariate. P-values were converted to q-values by the Benjamini-Hochberg procedure after filtering out lowly accessible sites (based on max accessibility level across cells) with R package genefilter (*35*). Chromatin accessibility along pseudotime was smoothed by genSmoothCurves function in monocle2 (*19*) with pseudotime and the number of peaks included as covariates.

Transcription factor analysis in the A549 experiment

For estimating transcription factor activity in single cells, we started with the binarized cell by site matrix and the list of significantly differentially accessible sites ($qvalue \leq 0.05$, 4763 sites) in different DEX treatments. Then for each site i in each condition k, we calculated the following *S*score:

$$S_{site\ i}^{condition\ k} = max(|log_2(\frac{n_i^k}{n_i^{j \neq k}})|)$$

in which n_i^k is the number of cells in condition k in which site i was observed accessible. Basically, the S score shows the maximum $|log2|$fold change of a site in each condition compared with the rest of the conditions. We kept the sites with $S \geq 1$to restrict the motif analysis only to significant sites where the number of cells with that site accessible changed with at least with 2 fold. This resulted in 1,696 sites that were used for estimating transcription factor activity in single cells.

We scanned 250 bps around the midpoint of the sites for occurance of motifs in CIS-BP database using FIMO (*36*, *37*). We kept the motifs with $pvalue < 1e-4$ and formed a binary matrix of sites by motifs based on presence or absence of a motif in each site. The cell by motif matrix (M) was generated using a matrix multiplication between matrices of cell by sites (C) and sites by motifs (S):

$$M = C \times S$$

Cell-motif matrix was further normalized by each cell's size factor to correct for their difference in read depth. Cells with less than 1,000 counts were removed. Differential tests to identify TFs with significant changes across different treatment conditions were performed by a likelihood ratio test similar to the gene differential test in Monocle 2 (*19*), with cell type specific size factor

estimated by estimateSizeFactors function in Monocle 2 (*19*) as a covariate. TF activity along pseudotime was calculated similarly to the chromatin accessibility analysis described above.

We downloaded ChIP-Seq data for A549 cell line for the same treatment conditions from ENCODE database (data released between year 2016 and 2018) (*18*). The datasets with genome version GRCh38 were lifted over to hg19 genome. These included the 15 tested TF targets in over 32 experiments. We next intersected the 1,696 significantly accessible sites with $S \geq 1$ with sites that were present in at least one of the ChIP-Seq experiment. We made a binary matrix of accessible sites occupied by TF targets and then performed a matrix multiplication between the accessibility matrix (cells by sites) and occupancy matrix (sites by TF targets). The cells by TF target matrix was normalized by each cell's size factor.

Analysis of sci-RNA-seq data from the mouse kidney experiment

A digital gene expression matrix was constructed from the raw sequencing data as described above. Cells with less than 500 UMIs were discarded. Genes expressed in less than 10 cells were filtered out. Downstream analyses were performed with Monocle 2 (*19*) and the python package scanpy (*38*). Briefly, the dimensionality of the data was reduced by PCA (30 components) first and then with t-SNE, followed by louvain clustering performed on the 30 principal components. After unsupervised clustering analysis, the intercalated cells cluster were separated to two cell clusters corresponding to intercalated cell type A and type B, based on marker gene expression in t-SNE. Two cell clusters, both identified as Loop of Henle cells, as well as two other cell clusters, both identified as endothelial cells, were merged. Three small clusters were merged to proximal tubule S3 type 1 cells, as they were highly correlated with one other (spearman correlations of aggregated transcriptome > 0.98) and enriched in proximal tubule S3 gene markers. To calculate the gene expression change between different cell types, the gene expression of each cell was calculated by

dividing the raw UMI count by cell specific size factor estimated by estimateSizeFactors function in Monocle2 (*19*). Mean expression and standard error was calculated across all cells in each cell type. Differentially expressed gene across different cell types were identified similarly as in A549 cell experiment. Consensus expression profiles for each cell type were constructed as in (*13*). To identify cell type specific gene marker, we first selected gene that were differentially expressed across different cell types (5% FDR, likelihood ratio test), then selected genes that has maximum expression in each cell type with at least 2-fold increases compared to other cell type with the second maximum expression.

Analysis of sci-ATAC-seq data from the mouse kidney experiment

Kidney cell ATAC profiles were analyzed by the same procedures used to analyze the A549 data, with minor modifications. Each cell identified in ATAC-seq was assigned to a cell type based on their linked RNA-seq profile, where available. Reads of cells from each cell type were aggregated and peaks were called separately. The peaks for each cell type were merged to form a single catalog of sites used to analyze all cells in downstream steps. To ensure all genes were represented in the catalog, this catalog was augmented with peaks corresponding to the promoter regions of each gene (gene start site annotated in GENCODE VM11 minus 200 base pairs in strand-specific manner). Each read was resized by extending 100 bp upstream and downstream from its start site. Cells were determined to be accessible at a given peak if the read from the cell overlapped with the peak. The peak count matrix was generated as for the A549 experiment. As the ATAC-seq data for mouse kidney is very sparse, we developed an approach to enrich ATAC-seq signal by aggregating cells with similar transcriptomes: for each cell type identified by RNA-seq, we performed dimensionality reduction on transcriptome by t-SNE (on the top 10 principal components) and sub-clustering by k-means clustering on t-SNE coordinates. The k is selected based on the number of cells in each cell type so that the average cell number per sub-clusteris 50.

k is set to 2 for cell types with less than 100 cells. After aggregating, we obtained a total of 222 pseudo-cells with at least 2 pseudo-cells from each cell type.

t-SNE visualization for ATAC-seq data was done in Monocle 2 (*19*). Briefly, sites accessible in fewer than 5 pseudo-cells were filtered out. The top 12 principal components were used as input to t-SNE, resulting in the two-dimensional embedding of the data.

To identify regulatory elements that were accessible in individual kidney cell types, as measured by the ATAC-seq portion of the coassay, we used a logistic regression framework to test whether cells of a given type were more likely to have Tn5 insertions at a given site relative all other cells in the dataset. We used the differential test implemented in the Monocle 2 (*19*) using the "binomialff" family with the following model:

$$\text{logit}(p_{i,j}) = \mu_i + \alpha_j + \beta_j + \varepsilon_i.$$

where p is the probability that the ith site is accessible in the jth cell, μ is the total proportion of cells that are accessible at the ith site, α indicates the membership of the jth cell in the cluster being tested, β is the log10(total number of sites observed as accessible for the jth cell), and ε is an error term for the ith site. We used a likelihood ratio test framework (as implemented in Monocle 2) to determine if the full model (including cell cluster membership) provided a significantly better fit of the data than a model that only accounts for the intercept and the log10(number of sites observed in each cell). This was carried out using the function differentialGeneTest in Monocle 2 with a fullModelFormula of "~α+β" and a reducedModelFormula of "~β", where α and β correspond to columns for the terms defined above as included in the pData table of the CellDataSet object. Note that we also modified this function to return the coefficient of the group membership term as an estimate of effect size. For each cell type, significant score (p-value) was calculated for each sites comparing the cell type to other cells. To increase the power to identify cell type specific accessible

sites, we filtered out low accessible sites in the target cells type before Benjamini-Hochberg correction with R package genefilter (*35*). Cell type specific sites for each cell type were identified as sites with false discovery rate of 5%.

To calculate the chromatin accessibility change across different cell types, the site accessibility of each cell was calculated by dividing the raw read count by cell specific size factor estimated by estimateSizeFactors function in Monocle 2 (*19*). Mean chromatin accessibility and standard error were calculated based on all cells in each cell type.

Analysis for linking cis-regulatory elements to regulated genes in the mouse kidney experiment

We aimed to identify links between chromatin accessible sites and regulated genes based on their covariance. To reduce measurement noise, which was driven largely by read depth per cell, we applied the same approach used for the ATAC-seq analysis mentioned above, constructing pseudo-cells by aggregating the RNA-seq and ATAC-seq profile of highly similar cells. After aggregating, we obtained a total of 222 pseudo-cells with at least 2 pseudo-cells from each cell types. Accessible sites detected in less than 5% of pseudo-cells were removed. The aggregated RNA-seq and ATAC-seq read counts per cell was normalized by cell-specific library size factors computed separately for each layer by estimateSizeFactorsForMatrix() function in DESeq2 (*39*), log transformed, centered, then scaled by scale() function in R. For each gene of the top 2,000 most significantly differentially expressed protein coding genes across different cell types in the mouse kidney, a LASSO regression model was constructed with package glmnet (*40*) to predict theexpression levels based on the normalized chromatin accessibility of its promoter or promoter plus nearby accessible sites (less than 100 kb from the gene start site or end site) by fitting the following model:

$$G_i = \beta_0 + \beta_p P_i + \beta_d D_i$$

where G_i is the adjusted gene expression value for gene i. It is calculated by aggregating RNA-seq count for all cells ($g_{1,2,3\ldots j}$) within the same "pseudo-cell", normalized by cell specific size factor (SG_i) estimate by estimateSizeFactorsForMatrix function in DESeq2 (*39*) and log transformed:

$$G_i = ln(\frac{\sum_{1}^{j} g_{ij}}{SG_i} + 1)$$

As G_i is created by summing counts across many cells (median of 47 cells, median of 4 counts for the top 2,000 DE genes across all pseudo-cells), it can reasonably be approximated by Gaussian distribution based on central limit theorem. To simplify downstream comparison between genes, we standardize the response G_i prior to fitting the model for each gene *i* with the scale() function in R.

Similar with G_i, P_i and D_i are the summed ATAC-seq read counts for promoter accessibility (P_i) and nearby (within 100 kb) accessible sites (D_i) of gene i: they are calculated by aggregating ATAC-seq counts for all cells ($p_{1,2,3\ldots j}$) within the same "pseudo-cell". For genes with multiple promoters, ATAC-seq read counts for all possible promoters are added together. Then they are normalized by cell specific size factor (SP_i) estimated with the estimateSizeFactorsForMatrix function in DESeq2 (*39*), and log transformed:

$$P_i = ln(\frac{\sum_{1}^{j} p_{ij}}{SP_i} + 1)$$

$$D_i = ln(\frac{\sum_{1}^{j} d_{ij}}{SP_i} + 1)$$

Prior to fitting, P_i and D_i are are standardized with the scale() function in R.

Our approach aims to identify sites that may regulate each gene, by finding the subset that can be used to predict its expression in a regression model. However, a site with accessibility correlated with a gene's expression does not guarantee it is regulating that gene. In particular, sites with accessibility that is inversely correlated with expression are unlikely to be genuine enhancers. We aimed to enrich links with bona fide enhancers by excluding those sites from the model that were not sufficiently correlated with its expression. We determined this threshold through a permutation-based procedure detailed below.

Specifically, we permuted the cell id or peak id of ATAC-seq matrix and redid the same analysis. Under the permutation, glmnet reported some links between sites and genes, but the magnitude of the coefficients attached to these links was small and often negative, likely reflecting minute amounts of covariance not ablated by the permutation procedure, *e.g.* if gene A is specifically expressed in cell type 1 and site B is specifically accessible in cell type 2. Although negative correlations between a site's accessibility and a gene's expression could reflect the activity of a transcriptional repressor, we felt that the more likely explanation for negative links reported by glmnet was mutually exclusive patterns of cell-type specific expression and accessibility. We thus used the permutation procedure to define a minimum threshold of correlation ($r = 0.04$) between a site and a gene when deciding whether to exclude the site prior to running the LASSO. If one site was linked to multiple genes, only the link with the highest correlation coefficient was preserved.

To quantify the predictive power of identified cis-regulatory and gene links, we constructed a linear regression model fitted with promoter accessibility or promoter accessibility plus linked distal sites accessibility to predict the expression of regulated genes across "pseudo-cells" (z score). Predicted gene expression was calculated based on ten-fold cross validation and r squared was measured to compare the prediction accuracy of two models.

Supplementary Figures

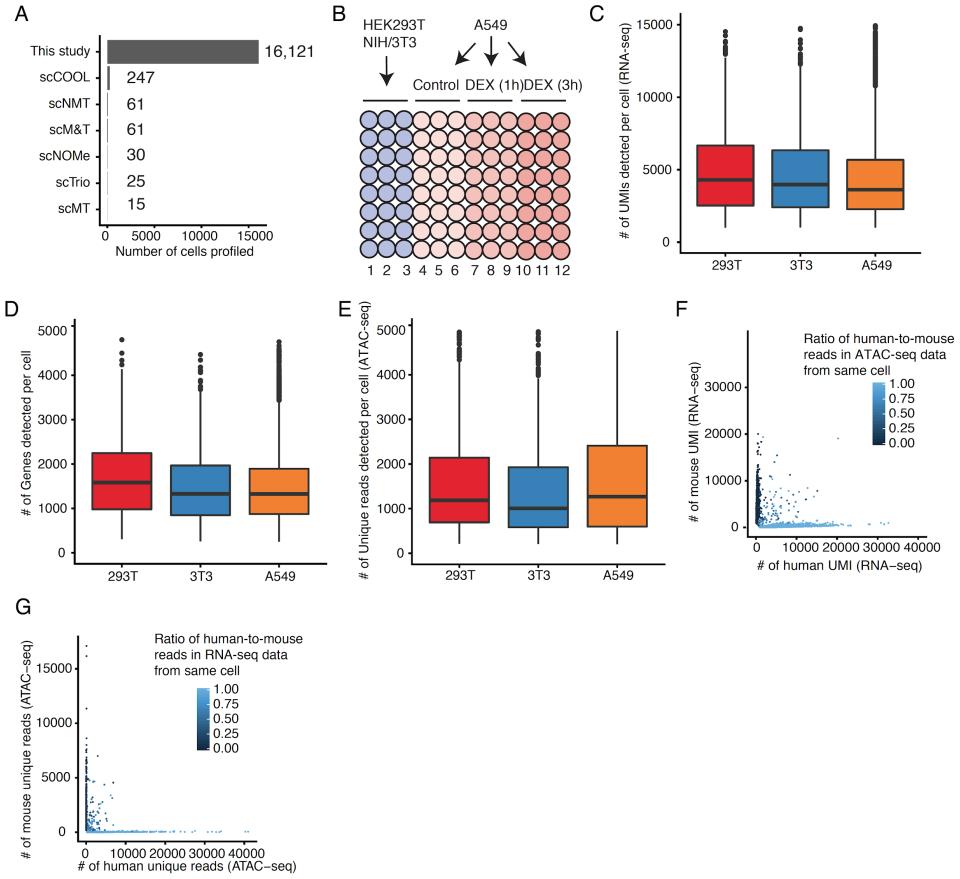


Fig. S1. sci-CAR links gene expression and chromatin accessibility information at the single cell level. (**A**) Comparison of throughput demonstrated in this study vs. other recent nucleic acid co-assay studies. (**B**) Layout of 96-well plate from first round of indexing of sci-CAR of human-mouse cell mixtures and dexamethasone-treated A549 cells. (**C, D**) Boxplots showing the number of UMIs (C) and genes (D) detected per cell in single cell RNA-seq profiles from sci-CAR experiment. (**E**) Boxplot showing the number of unique reads per cell in single cell ATAC-seq profiles sci-CAR experiment. (**F**) Scatter plot of unique human and mouse UMI counts (RNA-seq) from 96 x 576 sci-CAR. Only cells with both RNA-seq and ATAC-seq profiles are plotted, colored

by the ratio of human-to-mouse reads in linked ATAC-seq data. (**G**) Scatter plot of unique human and mouse reads (ATAC-seq) from 96 x 576 sci-CAR. Only cells with both RNA-seq and ATAC-seq profiles are plotted, colored by the ratio of human-to-mouse reads in linked RNA-seq data.

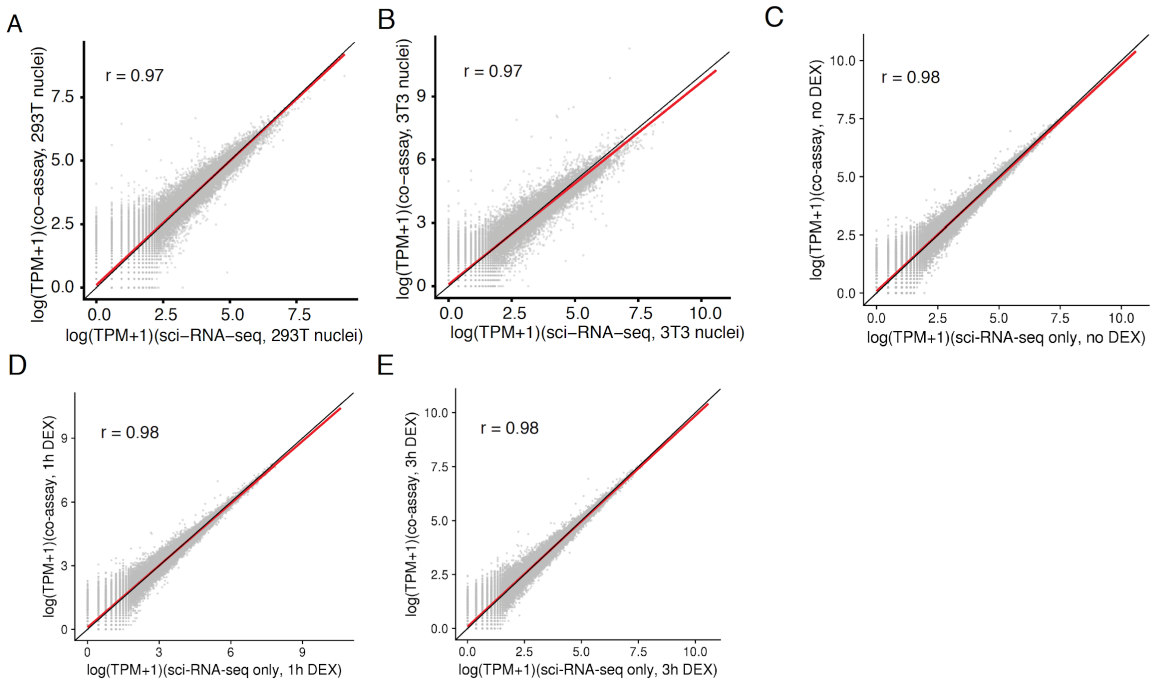


Fig. S2. Quality control of sci-CAR single cell transcriptomes. (**A, B**) Correlation between gene expression measurements in aggregated profiles of HEK293T nuclei (A) and NIH/3T3 nuclei (B) from sci-CAR vs. sci-RNA-seq from (*13*), together with a linear regression line (red) and y=x line (black). (**C-E**) Correlation between gene expression measurements of A549 cells from no DEX treatment (C), 1 hour DEX treatment (D) and 3 hour DEX treatment (E) in sci-CAR vs. sci-RNA-seq-only plates of the same experiment, together with a linear regression line (red) and y=x line (black).

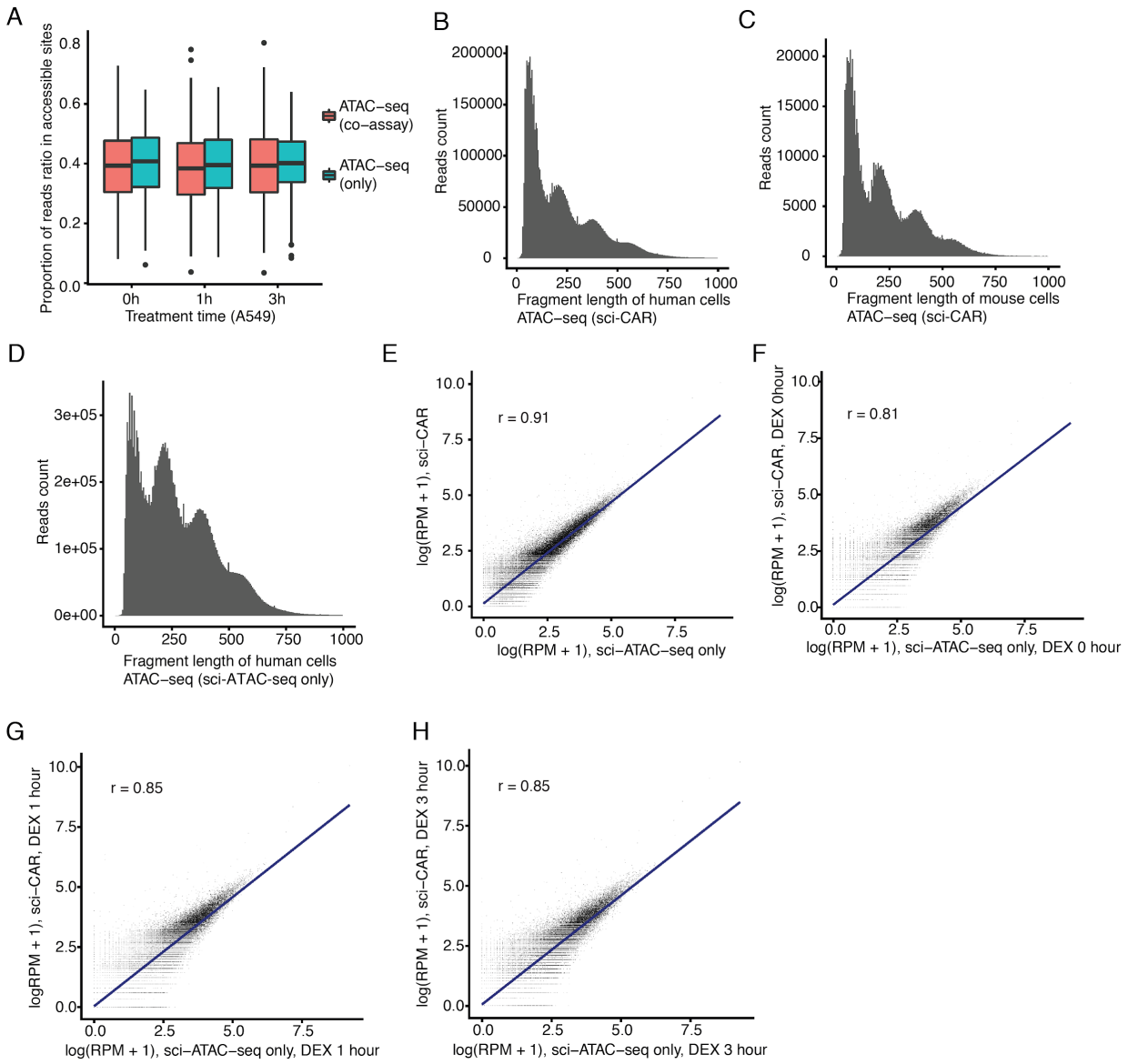


Fig. S3. Quality control of sci-CAR single cell chromatin accessibility profiles. (**A**) Box plot showing the proportion of unique ATAC reads intersecting with accessible sites in A549 nuclei from sci-CAR vs. sci-ATAC-seq-only plates. (**B-D**) Fragment length distribution in ATAC-seq from sci-CAR of human cells (B) or mouse cells (C), or human cells from sci-ATAC-seq-only plates (D). (**E-H**) Correlation between aggregated ATAC-seq signal (RPM, reads per million) of A549 cells (all treatment conditions (E), DEX 0 hour treatment (F), 1 hour treatment (G) and 3 hour treatment (H) from sci-CAR vs. sci-ATAC-seq-only plates, together with a linear regression

line (dark blue). The peak reference is based on all available ATAC-seq data from all timepoints for this experiment, and all peaks are plotted in each scatter plot.

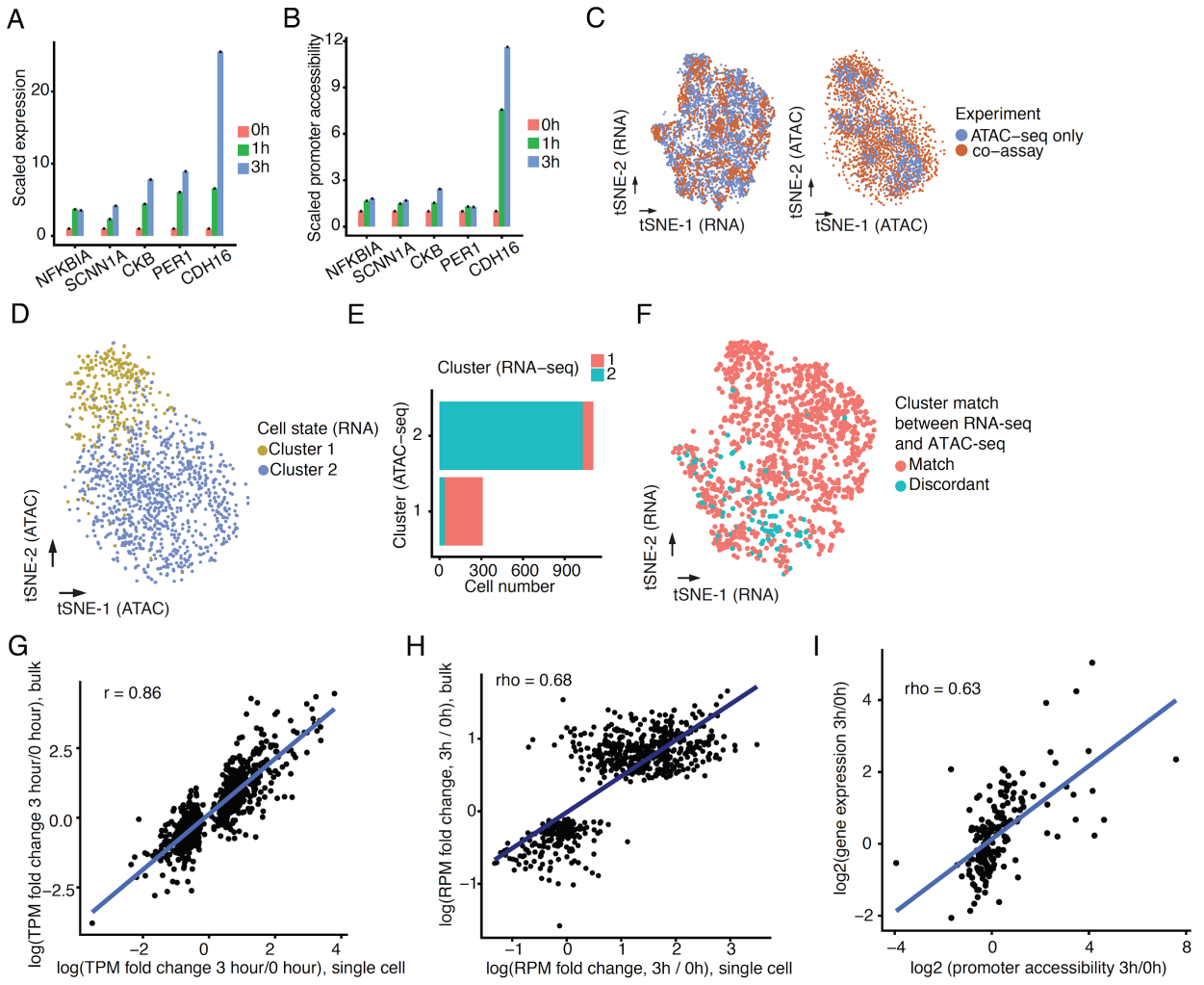


Fig. S4. sci-CAR profiles single cell expression and chromatin accessibility in dexamethasone treated A549 cells. (**A**) Bar plot showing the average expression of five known targets of GR upregulation across different timepoints of DEX treatment of A549 cells (normalized to 0 hr group). The gene expression profiles of each cell was calculated by dividing the raw UMI counts by a cell-specific size factor. Error bars represent standard errors of the means calculated based on all cells in each group. (**B**) Bar plot showing the average promoter accessibility of five known targets of GR upregulation across different timepoints of DEX treatment of A549 cells (normalized

to 0 hrs group). Error bars represent standard errors of the means calculated based on all cells in each group. (**C**) t-SNE visualization of A549 cells (left: RNA-seq; right: ATAC-seq) including cells from both sci-CAR and sci-RNA-seq only (left) or sci-ATAC-seq-only (right) plates, colored by experiment. (**D**) t-SNE visualization of A549 cells (ATAC-seq) with linked RNA-seq profiles, colored by cluster id identified from RNA-seq data. (**E**) Bar plot showing the number of cells in ATAC-seq cluster 1 or cluster 2 whose linked transcriptome is assigned to RNA-seq cluster 1 or cluster 2. (**F**) t-SNE visualization of A549 cells (RNA-seq) with linked ATAC-seq profiles. If the cell is in cluster 1 (or cluster 2) in both RNA-seq and ATAC-seq, then it is labeled as "Match", otherwise it is labeled as "Discordant". (**G**) Scatter plot showing the correlation between log transformed fold change (between 3 hour DEX treatment and 0 hour DEX treatment) of bulk RNA-seq (from ENCODE) and aggregated single cell RNA-seq across 870 DE genes identified by bulk RNA-seq, together with the linear regression line (blue). (**H**) Scatter plot showing the correlation between log transformed fold change (between 3 hour DEX treatment and 0 hour DEX treatment) of bulk DNase-seq (from ENCODE) and aggregated single cell ATAC-seq (RPM, reads per million) across 672 DA sites identified from bulk DNase-seq data, together with the linear regression line (blue). The peak reference from sci-CAR was used in reanalyzing bulk DNase-seq data. (**I**) Scatter plot showing the correlation between log2 transformed fold change (between 3 hour DEX treatment and 0 hour DEX treatment) of promoter accessibility and gene expression for 175 genes that are both DA and DE, together with the linear regression line (blue).

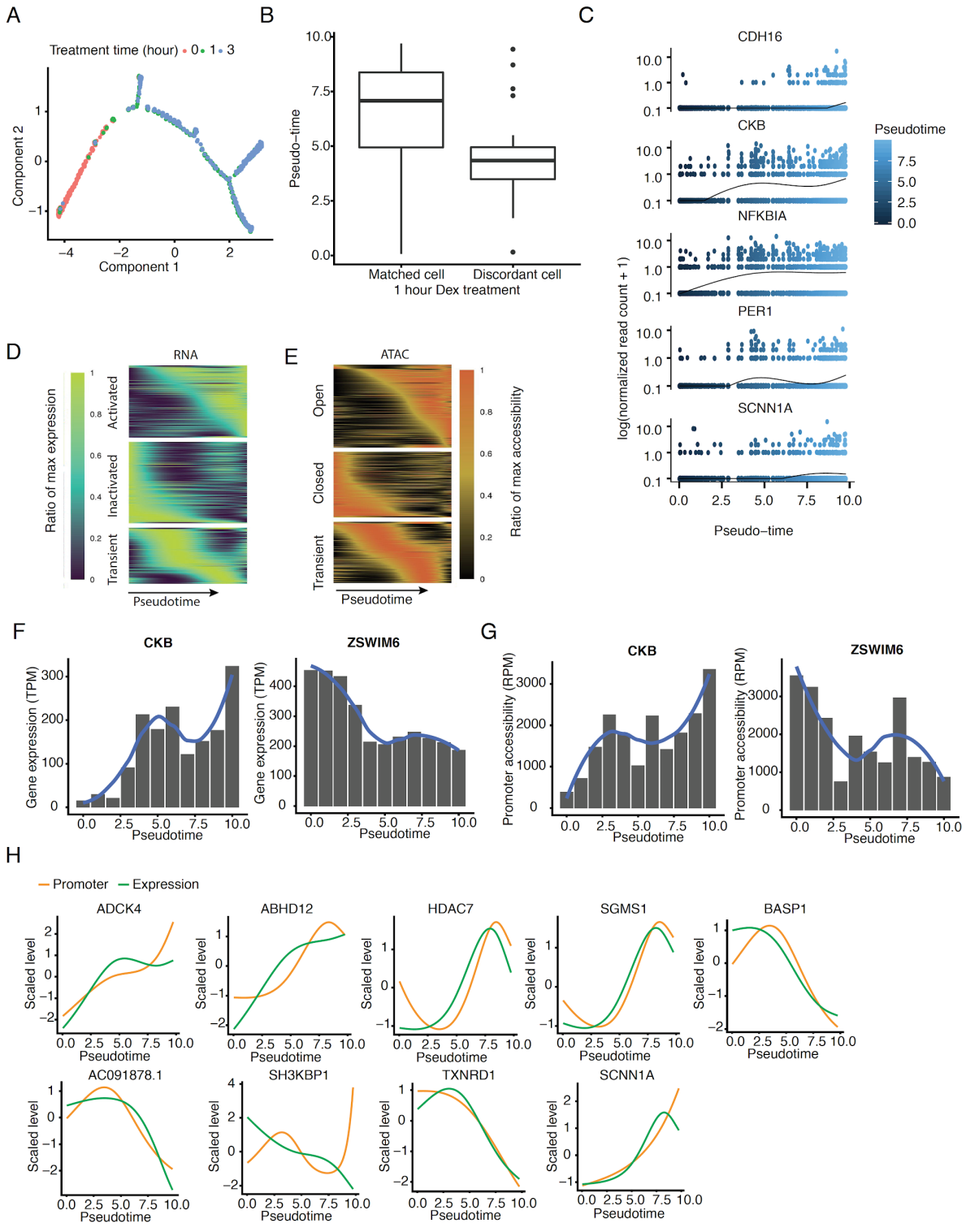


Fig. S5. Pseudotime analyses of dexamethasone treated A549 cells. (**A**) The Monocle-based single cell pseudotime trajectory of sci-RNA-seq profiles from A549 cells, including cells from

both sci-CAR and sci-RNA-seq-only plates. (**B**) Boxplot showing the pseudo-time of cells from matched vs. discordant cells from the 1 hr DEX treatment group. If the cell is in cluster 1 (or cluster 2) in both RNA-seq and ATAC-seq, then it is labeled as "Matched", and otherwise as "Discordant". (**C**) Relative expression of GR target gene markers in cells ordered by pseudotime. Black line indicates the pseudotime-dependent average from a smoothed binomial regression. The relative expression is calculated by first normalizing the read count by cell size factor estimated by estimateSizeFactors function in Monocle 2, and then log transformed after adding a pseudocount 1. (**D, E**) Smoothed pseudotime-dependent gene expression (D) and chromatin accessibility (E) curves, generated by a negative binomial regression and scaled as a percent of maximum gene expression (D) or chromatin accessibility (E). Each row indicates a different gene (D) or DNA element (E). (**F, G**) Similar to Fig. 2F; bar plots showing unscaled aggregate gene expression in transcripts per million (F) and aggregate promoter accessibility in reads per million (G) of genes CKB and ZSWIM6, together with loess smoothed line. (**H**) Similar to Fig. 2F; smoothed pseudotime-dependent gene expression and promoter accessibility of genes with significant change (FDR 5%) in both gene expression and promoter accessibility along pseudotime.

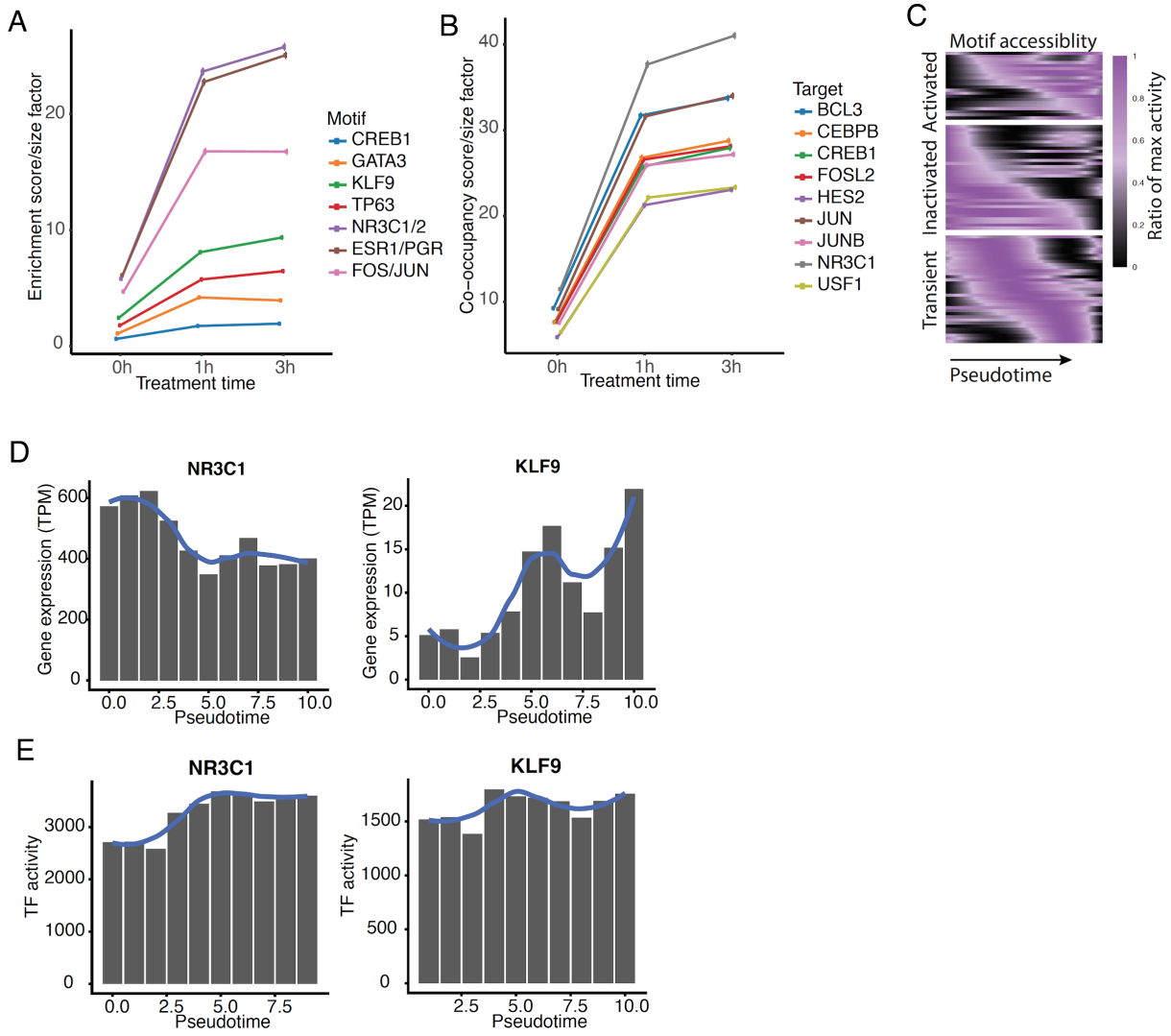


Fig. S6. Analyses of single cell transcription factor activity change in dexamethasone treated A549 cells. (**A, B**) Dynamics of transcription factor activity during DEX treatment represented by (A) aggregate instances of motifs present in accessible sites and (B) accessible sites occupied by targeted transcription factors captured by publicly available ChIP-Seq data for this time course. (**C**) Similar to Fig. S5D and S5E; smoothed pseudotime-dependent transcription factor activity curves, scaled as percentage of maximum motif accessibility. (**D, E**) Similar to Fig. 2G; bar plots

showing unscaled aggregate gene expression in transcripts per million (D) and aggregate TF motif intersected reads per million (E) of gene NR3C1 and KLF9, together with loess smoothed line.

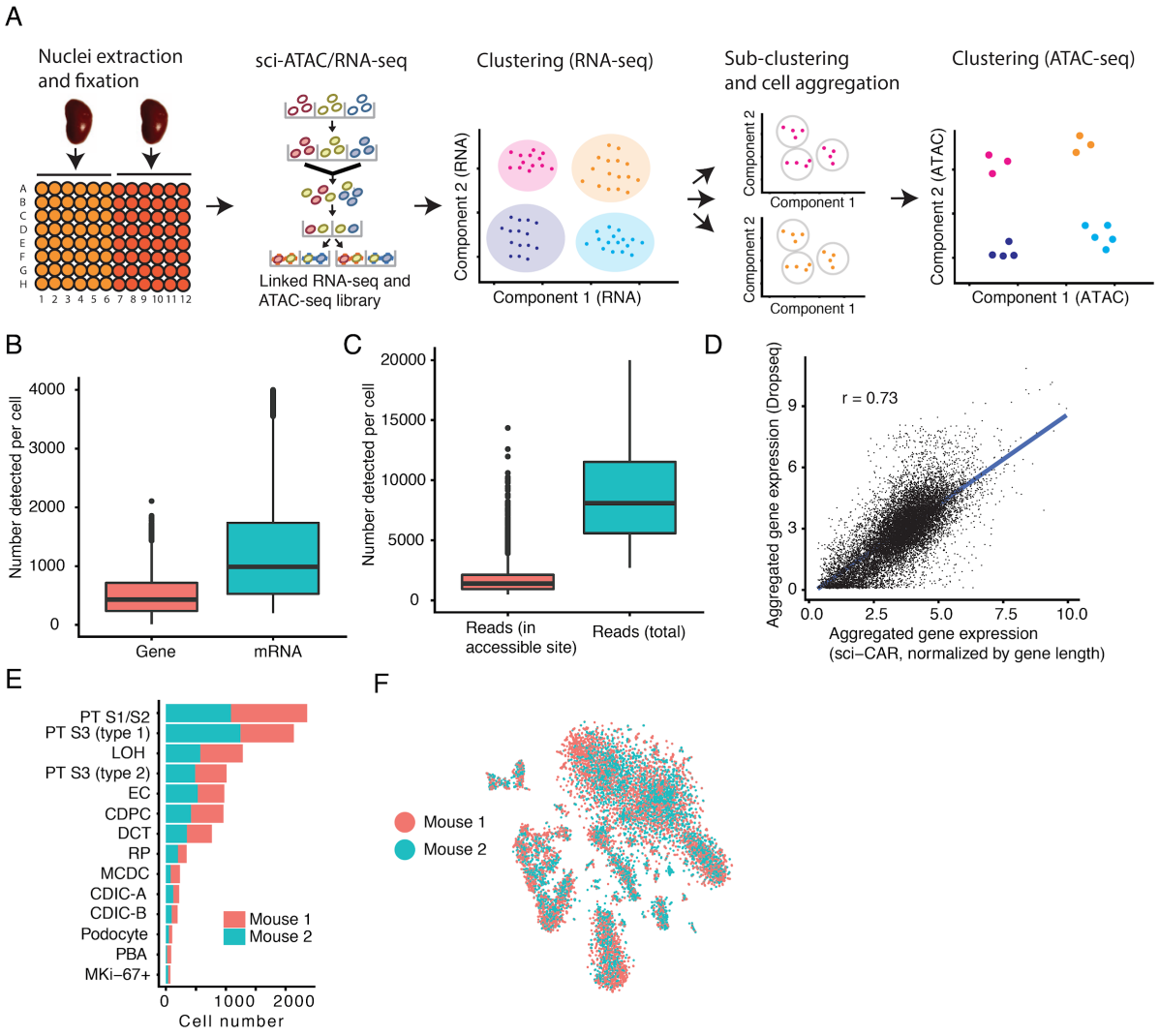


Fig. S7. Application of sci-CAR to the mouse kidney. (**A**) Layout of 96-well plate from first round of indexing of sci-CAR, followed by schematic of key analysis steps for RNA-seq informed clustering of ATAC-seq profiles via 'pseudo-cells'. (**B**) Box plot showing the number of UMIs and genes detected per cell in mouse kidney nuclei by sci-CAR. (**C**) Box plot showing the number of ATAC-seq unique reads and reads intersected with accessible sites per cell in mouse kidney nuclei by sci-CAR. (**D**) Correlation between aggregated gene expression (log(TPM + 1))

measurements of mouse kidney cells from Drop-seq vs. sci-CAR normalized by gene length, together with a linear regression line (blue). (**E**) Bar plot showing the number of cells of each cell type profiled by sci-CAR, colored by replicate id. (**F**) t-SNE visualization of mouse kidney nuclei (based on single RNA-seq profiles from sci-CAR), colored by replicate id.

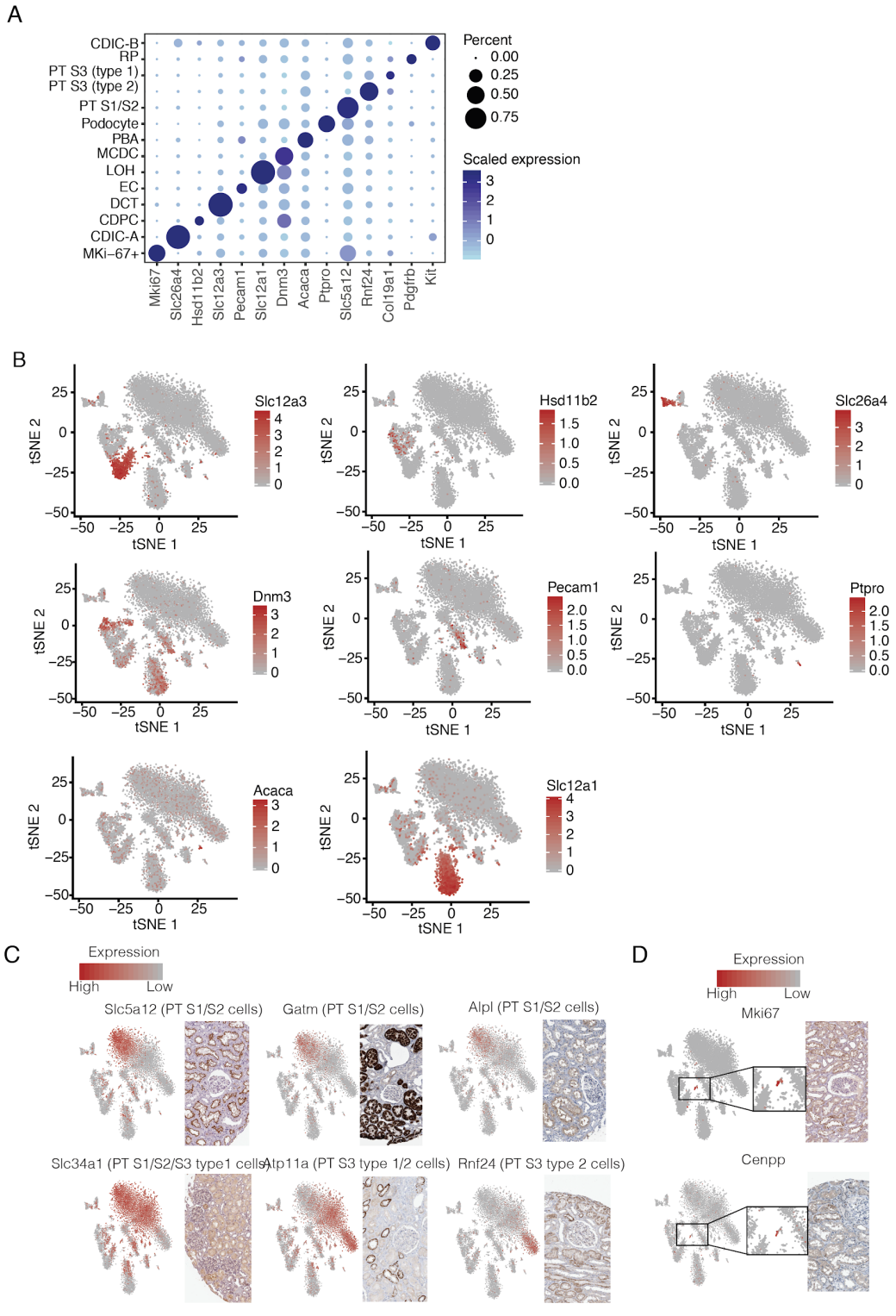


Fig. S8. sci-CAR expression data identifies major cell types of mouse kidney. (**A**) Dot plot visualization of each cell type in RNA-seq. The size of the dot represents the percentage of cells

within a cell type expressing the gene (UMI count > 0) and the color indicates the average expression level. (**B**) t-SNE visualization of mouse kidney nuclei (RNA-seq aspect of sci-CAR data), colored by the single cell expression of gene markers. Expression value were calculated by dividing raw UMI count by cell specific size factors. (**C, D**) Left panels are the t-SNE visualization of mouse kidney nuclei (RNA-seq), colored by cell type specific gene markers for proximal tubule sub-types (C) and active proliferating sub-population (D). Right panels are the immunohistochemistry (IHC) data for kidney tissue from The Human Protein Atlas (*25*, *26*, *41–52*).

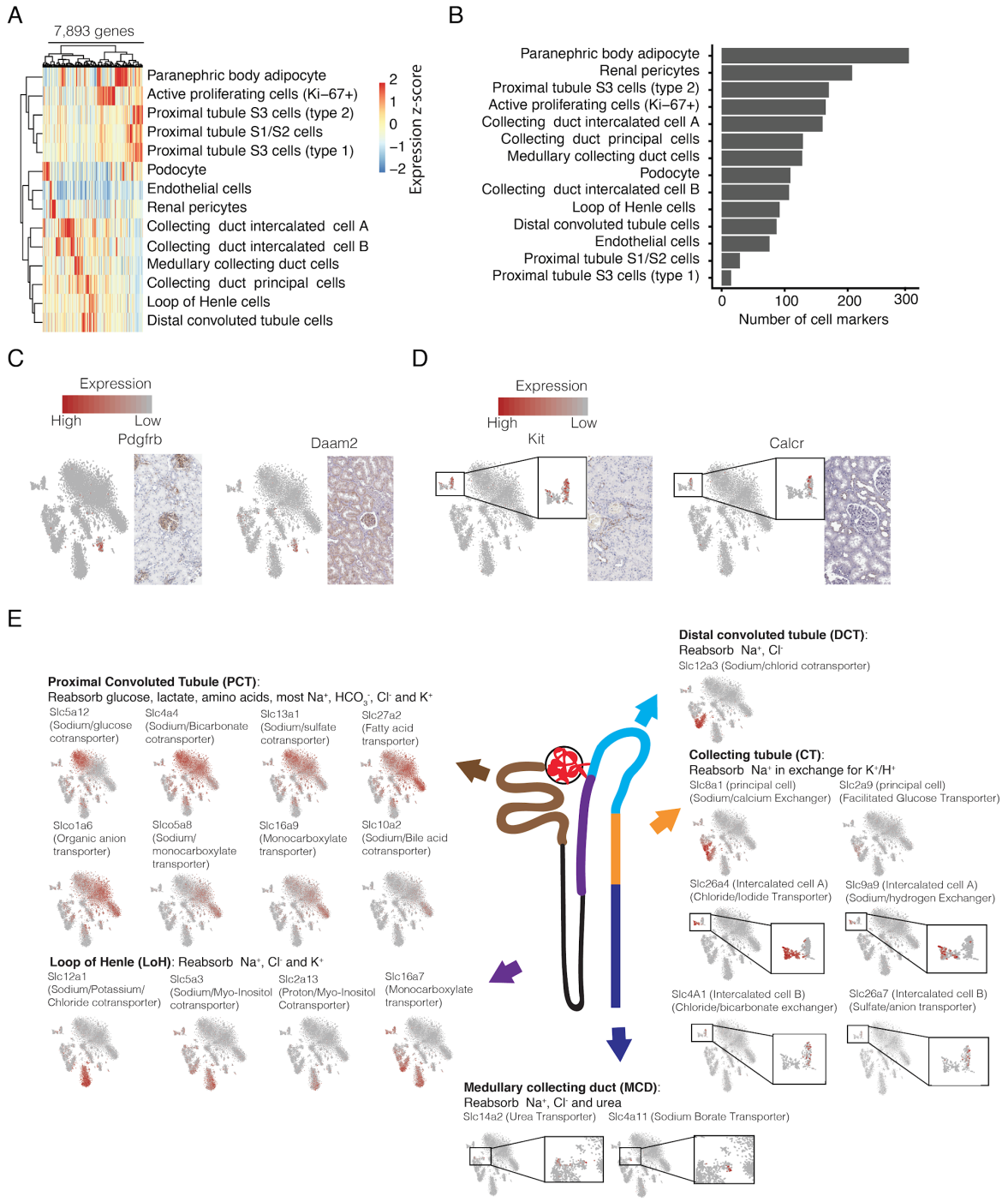


Fig. S9. sci-CAR identified cell type specific genes markers of mouse kidney. (**A**) Heatmap showing the relative expression of genes in consensus transcriptomes of each cell type estimated by single cell RNA-seq data from sci-CAR. Genes are included if they have a size-factor-

normalized mean expression of >0.05 in at least one cell type (7,893 genes in total). The raw expression data (UMI count matrix) is log-transformed, column centered and scaled (using the R function scale), and the resulting values are clamped to the interval [-2, 2]. (**B**) Number of genes that are enriched at least 2-fold in a specific cell type relative to the 2nd-highest-expressing cell type, excluding genes for which the differential expression all cell types is not significant (q-value > 0.05). (**C, D**) Left panels are the t-SNE visualization of mouse kidney nuclei (RNA-seq), colored by cell type specific gene markers for renal pericytes (C) and collecting duct intercalated cell B (D). Right panels are the immunohistochemistry (IHC) data for kidney tissue from The Human Protein Atlas (*25*, *26*, *41–52*). (**E**) t-SNE visualization of mouse kidney nuclei (RNA-seq), colored by gene expression of cell type-specific transporters. The expression of cell type-specific transporters correlates with the reabsorption functions in different regions of kidney tubule.

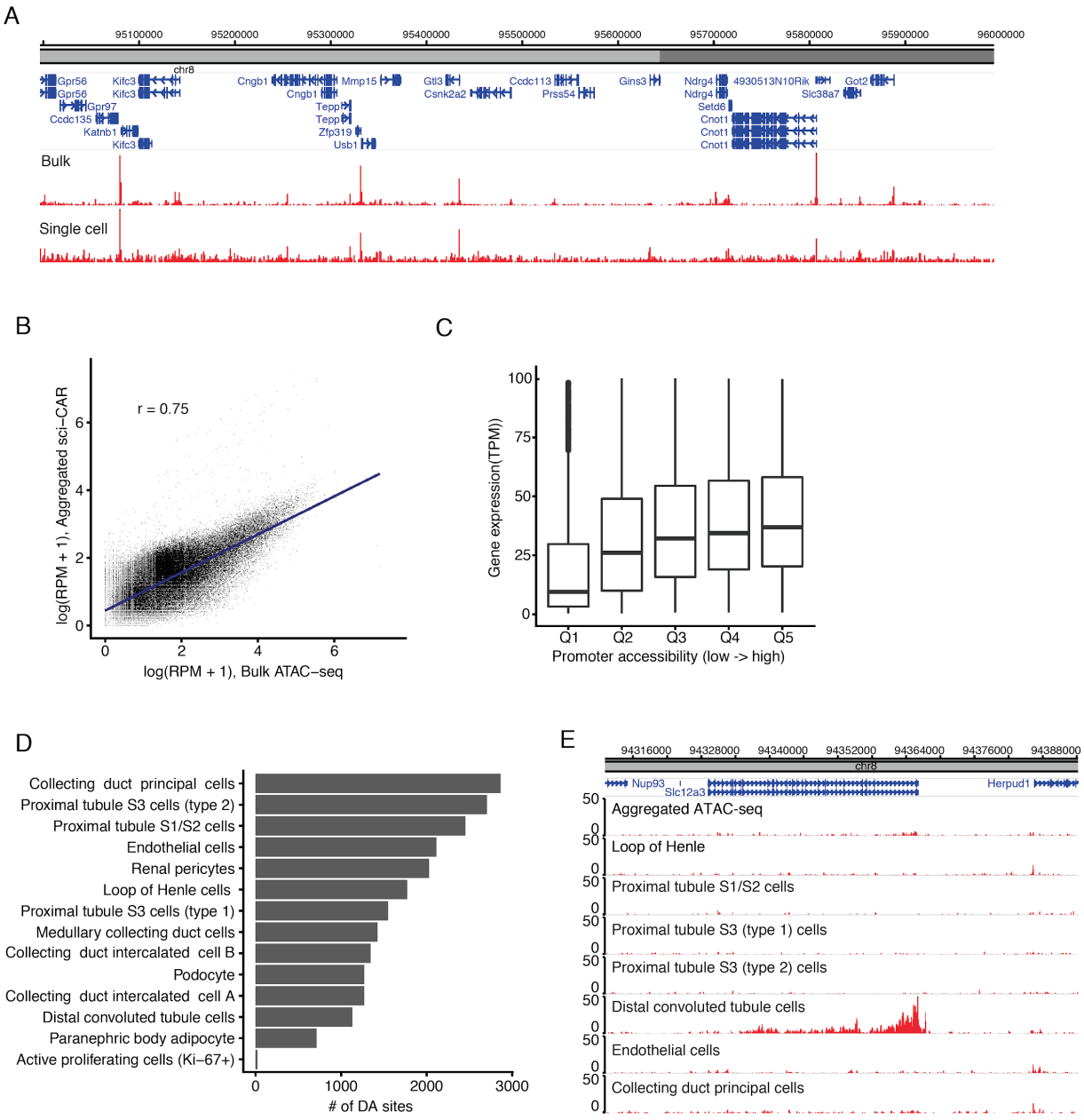


Fig. S10. sci-CAR chromatin accessibility data identifies cell-type specific differentially accessible sites. (**A**) Aggregate single cell chromatin accessibility profiles from sci-CAR recapitulate bulk ATAC-seq profiles of the adult mouse kidney (*18*). 5M reads were sampled from each data set for density plot visualization with EpiGenome Browser (*53*). (**B**) Correlation between aggregated gene expression (log(RPM + 1)) measurements of mouse kidney cells from sci-CAR vs. published bulk ATAC-seq data on adult mouse kidney (*18*), together with a linear regression

line (blue). The peak reference (MACS2-called peaks and promoters of all genes) from sci-CAR was used in reanalyzing bulk ATAC-seq data. (**C**) Genes were divided into five equal groups based on aggregate promoter accessibility. Boxplot showing the aggregate expression (TPM, transcripts per million) of genes within each group. (**D**) Bar plot showing the number of enriched differentially accessible sites identified for different cell types in mouse kidney. (**E**) Coverage density plot of aggregated ATAC-seq profile for different cell types at the *Slc12a3* locus. 5M reads were sampled from each data set for density plot visualization with EpiGenome Browser (*53*). Cell types with less than 5M reads are not shown.

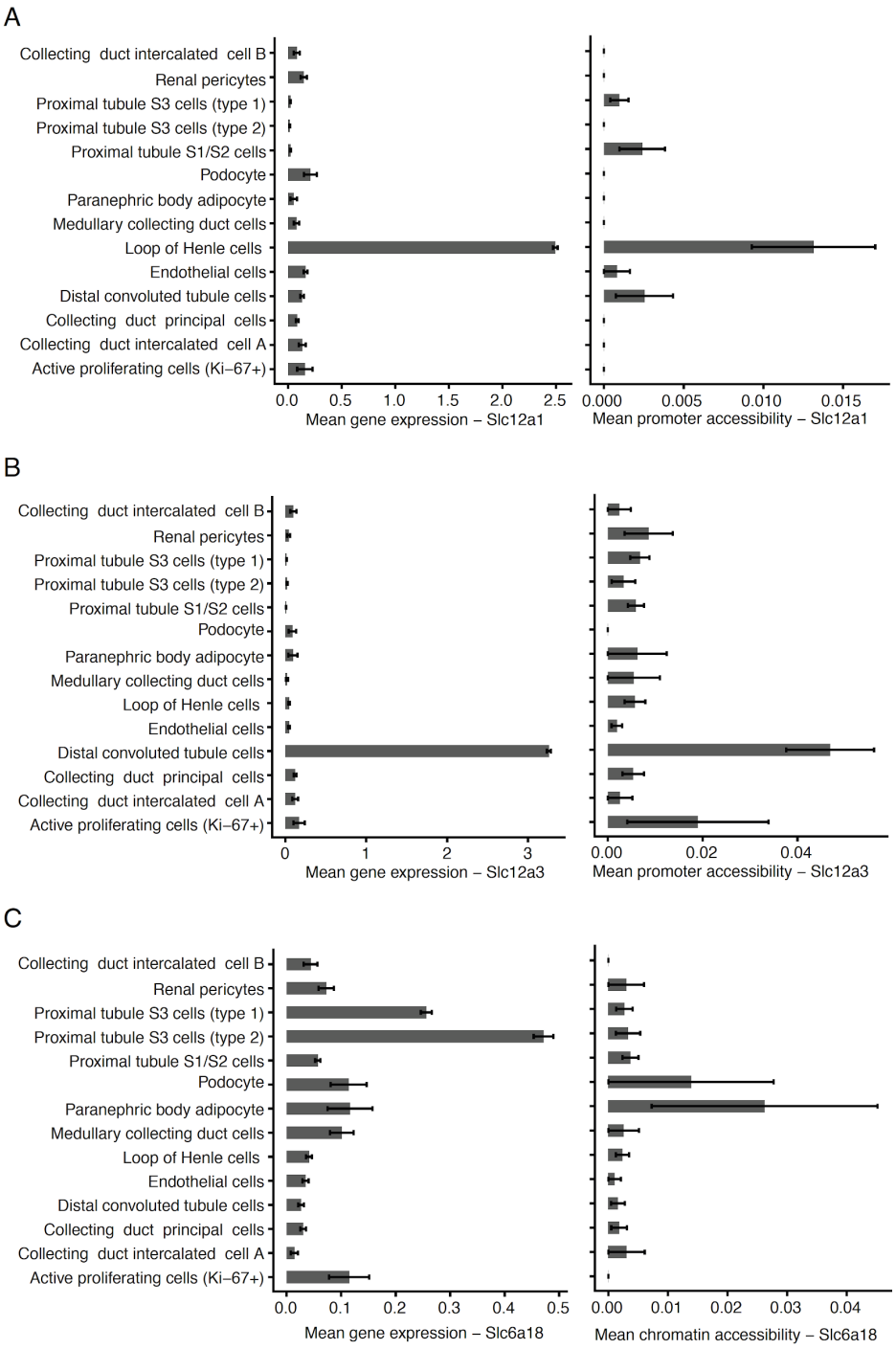


Fig. S11. sci-CAR profiles promoter accessibility and gene expression across different cell types. (**A-C**) Left panels: bar plot showing the average expression of cell type specific marker genes Slc12a1 (A), Slc12a3 (B) and Slc6a18 (C) across different cell types in mouse kidney. The

gene expression of each cell was calculated by dividing the raw UMI count by cell specific size factor. Error bars represents standard errors of the means. Right panels: bar plots showing the average promoter accessibility of cell type specific marker genes Slc12a1 (A), Slc12a3 (B) and Slc6a18 (C) across different cell types in mouse kidney. The promoter accessibility of each cell was calculated by dividing the raw read count by cell specific size factor. Error bars represents standard errors of the means.

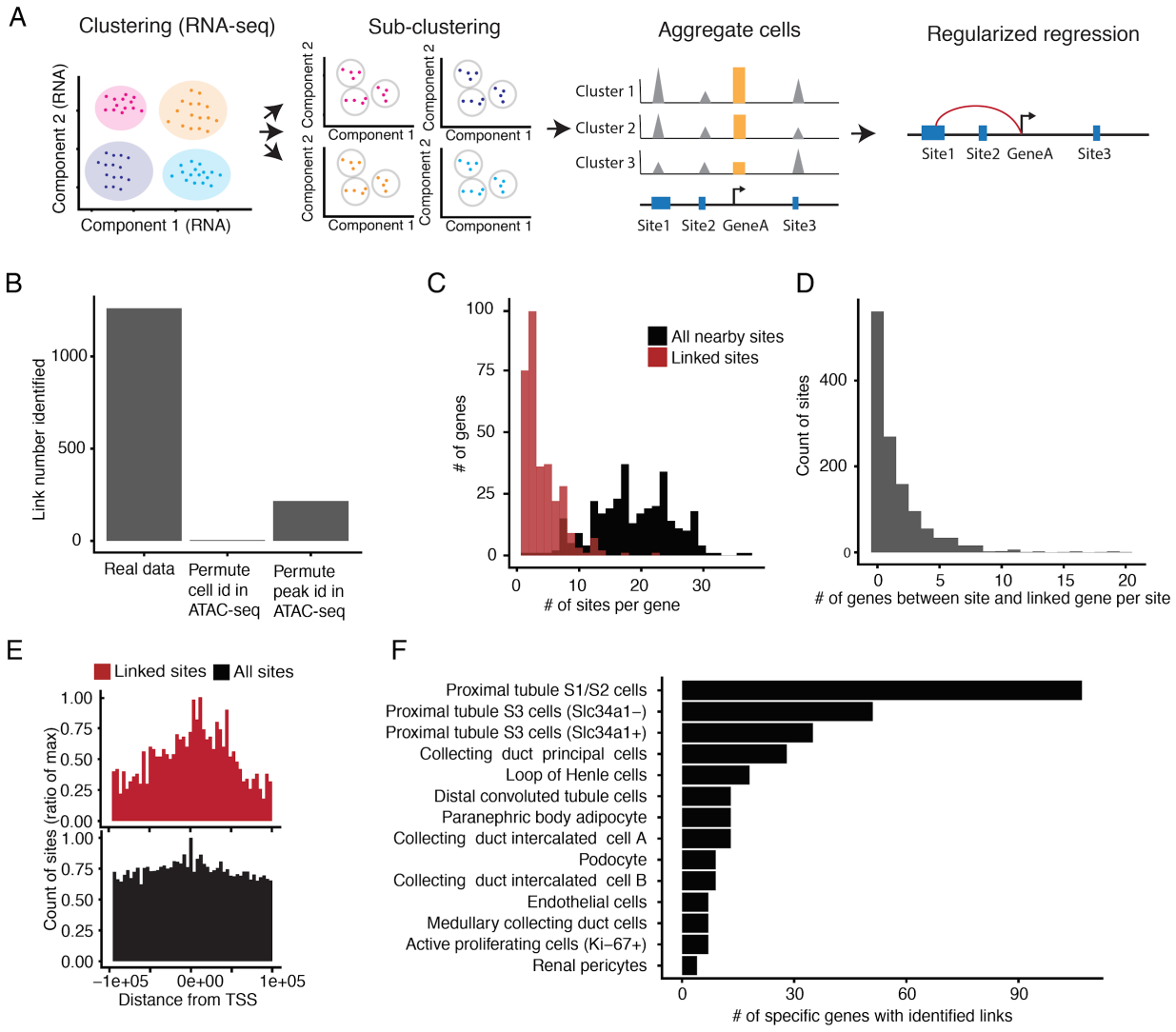


Fig. S12. sci-CAR links cis-regulatory elements and regulated genes. (**A**) Overview of analytical steps. (**B**) Bar plot showing the number of site-gene links identified from real data vs. after permuting cell id or peak id of ATAC-seq data. (**C**) Histogram of the number of accessible sites per gene that were linked (red) vs. tested (black), for genes with 1+ links only. (**D**) Histogram of the number of genes occurring in the genomic region between the cis-regulatory element and regulated gene comprising each link. (**E**) Histogram showing the distance distribution between sites and transcription start sites (TSS) for linked sites (red) vs. all sites within 100 kb (black). (**F**) Bar plot showing the number of genes (with linked sites identified) specific to each cell type. A

gene is specific to one cell type when it is differentially expressed across cell types (FDR < 1%) with highest expression in the cell type.

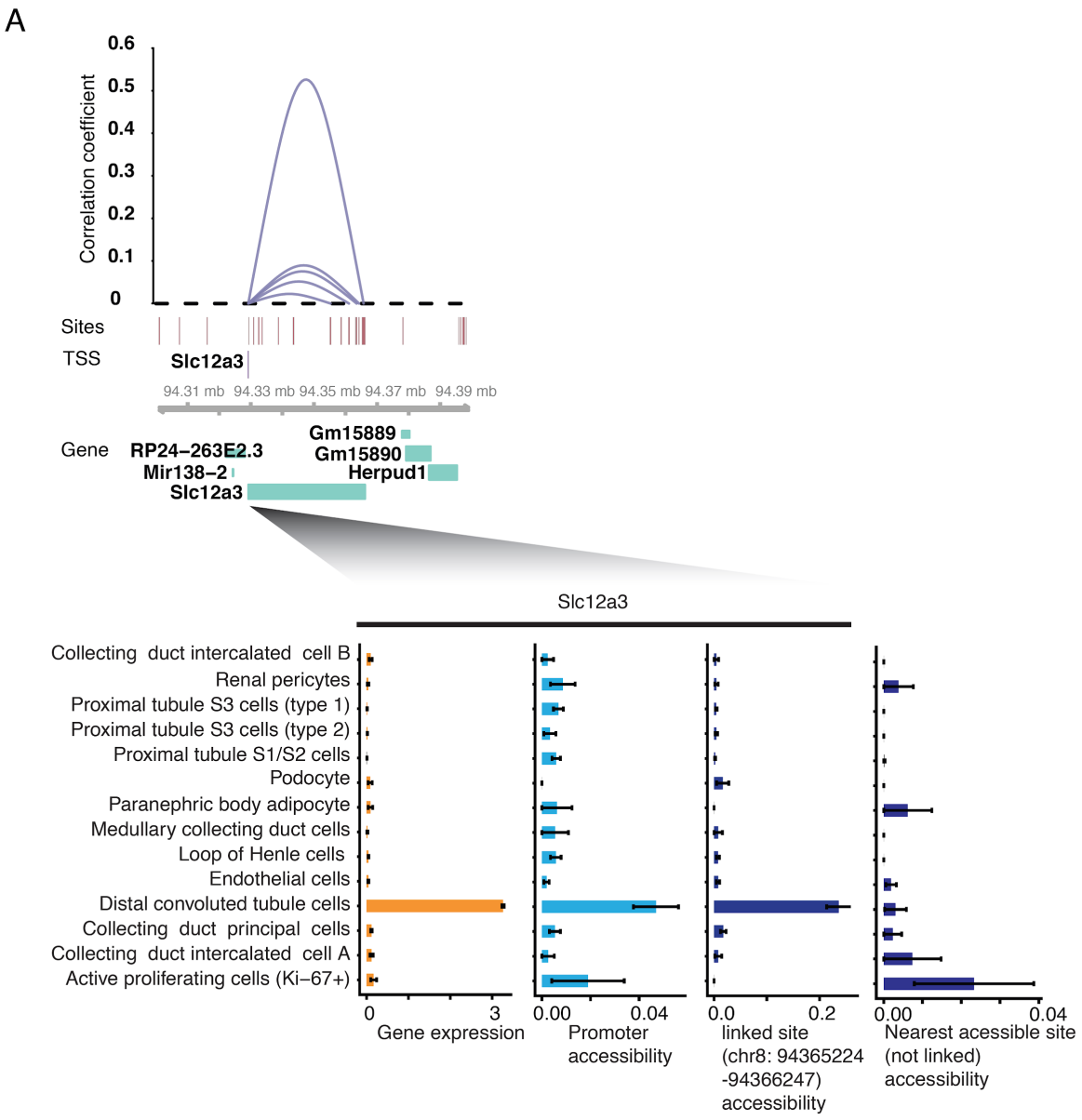


Fig. S13. Linked cis-regulatory elements and regulated genes show similar cell type specificities. (**A**) top: genome browser plot showing links between accessible distal regulatory sites and the gene Slc12a3. The height corresponds to the correlation coefficient. Bottom: Similar to Fig. 4A (bottom), barplots showing the average expression, promoter accessibility, linked site accessibility, and the accessible site nearest to the gene start site for cell type-specific marker gene Slc12a3 across different cell types. Gene expression values for each cell were calculated by

dividing the raw UMI count by cell-specific size factors. Site accessibilities for each cell were calculated by dividing the raw read count by cell-specific size factors. Error bars represent standard errors of the means.

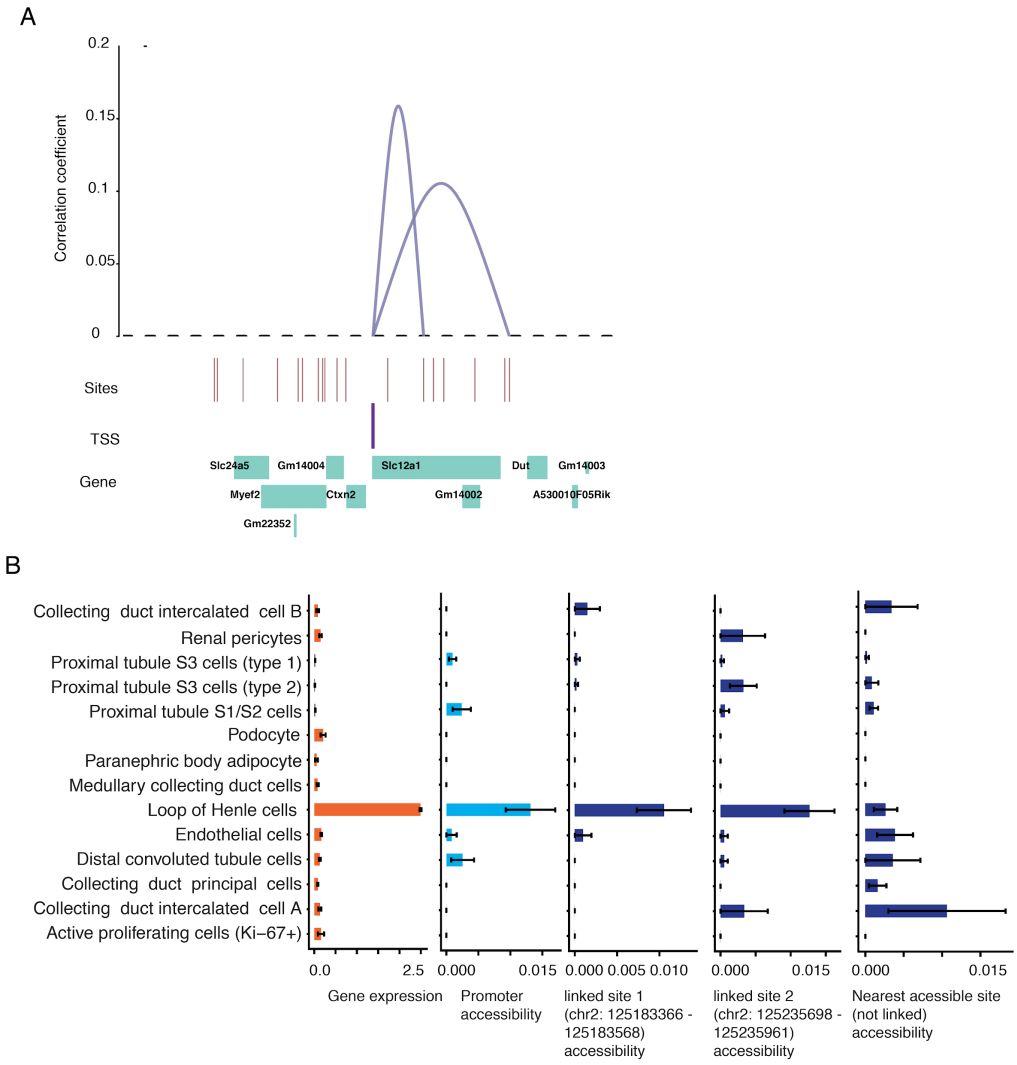


Fig. S14. Linked cis-regulatory elements and regulated genes show similar cell type specificities. (**A**) Similar to Fig. 4A (top), genome browser plot showing links between accessible distal regulatory sites and Slc12a1. The height corresponds to the correlation coefficient from the LASSO regression. (**B**) Similar to Fig. 4A (bottom), barplots showing the average expression, promoter accessibility and linked or unlinked site accessibilities for Slc12a1 across different cell

types. Gene expression values for each cell were calculated by dividing the raw UMI count by cell-specific size factors. Site accessibilities for each cell were calculated by dividing the raw read count by cell-specific size factors. Error bars represent standard errors of the means.

Reference of chapter 2:

1. S. J. Clark *et al.*, scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).

2. C. Angermueller *et al.*, Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat. Methods*. **13**, 229–232 (2016).

3. Y. Hou *et al.*, Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res.* **26**, 304–319 (2016).

4. Y. Hu *et al.*, Simultaneous profiling of transcriptome and DNA methylome from a single cell. *Genome Biol.* **17**, 88 (2016).

5. S. Pott, Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *Elife*. **6** (2017), doi:10.7554/eLife.23203.

6. F. Guo *et al.*, Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Res.* **27**, 967–988 (2017).

7. S. Amini *et al.*, Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343 (2014).

8. D. A. Cusanovich *et al.*, Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*. **348**, 910–914 (2015).

9. V. Ramani *et al.*, Massively multiplex single-cell Hi-C. *Nat. Methods*. **14**, 263–266 (2017).

10. Y. Yin *et al.*, High-throughput mapping of meiotic crossover and chromosome mis-segregation events in interspecific hybrid mice. *bioRxiv* (2018), p. 338053.

11. S. A. Vitak *et al.*, Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods*. **14**, 302–308 (2017).

12. R. M. Mulqueen *et al.*, Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018).

13. J. Cao *et al.*, Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*. **357**, 661–667 (2017).

14. T. E. Reddy *et al.*, Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Res.* **19**, 2163–2171 (2009).

15. S. John *et al.*, Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268 (2011).

16. T. E. Reddy, J. Gertz, G. E. Crawford, M. J. Garabedian, R. M. Myers, The Hypersensitive Glucocorticoid Response Specifically Regulates Period 1 and Expression of Circadian Genes. *Mol. Cell. Biol.* **32**, 3756–3767 (2012).

17. C. M. Vockley *et al.*, Direct GR Binding Sites Potentiate Clusters of TF Binding across the Human Genome. *Cell*. **166**, 1269–1281.e19 (2016).

18. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature*. **489**, 57–74 (2012).

19. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods*. **14**, 979–982 (2017).

20. J. D. Buenrostro *et al.*, Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*. **523**, 486–490 (2015).

21. Y. Chinenov, M. Coppo, R. Gupte, M. A. Sacta, I. Rogatsky, Glucocorticoid receptor coordinates transcription factor-dominated regulatory network in macrophages. *BMC Genomics*. **15**, 656 (2014).

22. L. Chen *et al.*, Transcriptomes of major renal collecting duct cell types in mouse identified by single-cell RNA-seq. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E9989–E9998 (2017).

23. X. Han *et al.*, Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*. **172**, 1091–1107.e17 (2018).

24. J. Park *et al.*, Comprehensive single cell RNAseq analysis of the kidney reveals novel cell types and unexpected cell plasticity (2017), , doi:10.1101/203125.

25. Human Protein Atlas, (available at www.proteinatlas.org).

26. M. Uhlen *et al.*, A pathology atlas of the human cancer transcriptome. *Science*. **357** (2017), doi:10.1126/science.aan2507.

27. Y. Zhang *et al.*, Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

28. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*. **10**, 1213–1218 (2013).

29. D. A. Cusanovich *et al.*, The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature*. **555**, 538–542 (2018).

30. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).

31. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

32. H. A. Pliner *et al.*, Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data. *Mol. Cell* (2018), doi:10.1016/j.molcel.2018.06.044.

33. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841–842 (2010).

34. S. Anders, P. T. Pyl, W. Huber, HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*. **31**, 166–169 (2015).

35. Gentleman R, Carey V, Huber W and Hahne F, *genefilter: genefilter: methods for filtering genes from high-throughput experiments* (2017).

36. C. E. Grant, T. L. Bailey, W. S. Noble, FIMO: scanning for occurrences of a given motif. *Bioinformatics*. **27**, 1017–1018 (2011).

37. M. T. Weirauch *et al.*, Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*. **158**, 1431–1443 (2014).

38. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

39. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

40. J. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33** (2010), doi:10.18637/jss.v033.i01.

41. Cenpp. *The Human Protein Atlas*, (available at https://www.proteinatlas.org/ENSG00000188312-CENPP/tissue/kidney#img).

42. Pdgfrb. *The Human Protein Atlas*, (available at https://www.proteinatlas.org/ENSG00000113721-PDGFRB/tissue/kidney#img).

43. Slc5a12. *The Human Protein Atlas*, (available at https://www.proteinatlas.org/ENSG00000148942-SLC5A12/tissue/kidney#img).

44. GATM. *The Human Protein Atlas*, (available at https://www.proteinatlas.org/ENSG00000171766-GATM/tissue/kidney#img).

45. Alpl. *The Human Protein Atlas*, (available at https://www.proteinatlas.org/ENSG00000162551-ALPL/tissue/kidney#img).

46. Slc34a1. *The Human Protein Atlas*, (available at https://www.proteinatlas.org/ENSG00000131183-SLC34A1/tissue/kidney#img).

47. Atp11a. *The Human Protein Atlas*, (available at https://www.proteinatlas.org/ENSG00000068650-ATP11A/tissue/kidney#img).

48. Rnf24. *The Human Protein Atlas*, (available at https://www.proteinatlas.org/ENSG00000101236-RNF24/tissue/kidney#img).

49. Mki67. *The Human Protein Atlas*, (available at https://www.proteinatlas.org/ENSG00000148773-MKI67/tissue/kidney#img).

50. Kit. *The Human Protein Atlas*, (available at https://www.proteinatlas.org/ENSG00000157404-KIT/tissue/kidney#img).

51. Daam2. *The Human Protein Atlas*, (available at https://www.proteinatlas.org/ENSG00000146122-DAAM2/tissue/kidney#img).

52. Calcr. *The Human Protein Atlas*, (available at https://www.proteinatlas.org/ENSG00000004948-CALCR/tissue/kidney#img).

53. X. Zhou, T. Wang, in *Current Protocols in Bioinformatics* (2012).

Chapter 3: cell fate characterization of mammalian organogenesis

*Modified from an article the single cell transcriptional landscape of mammalian organogenesis, Junyue Cao, Malte Spielmann, et al, Nature, 2019

Authors: Junyue Cao[1, 2†], Malte Spielmann[1†], Xiaojie Qiu[1,2], Xingfan Huang[1,3], Daniel M. Ibrahim[4,5], Andrew J. Hill[1], Fan Zhang[6], Stefan Mundlos[4, 5], Lena Christiansen[6], Frank J. Steemers[6], Cole Trapnell[1,7]*, Jay Shendure[1,7,8]*

Affiliations:

1. Department of Genome Sciences, University of Washington, Seattle, Washington, USA
2. Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA.
3. Department of Computer Science, University of Washington, Seattle, Washington, USA
4. Max Planck Institute for Molecular Genetics, RG Development & Disease, Berlin, Germany
5. Institute for Medical and Human Genetics, Charité Universitätsmedizin Berlin, Berlin, Germany
6. Illumina, San Diego, California, USA
7. Brotman Baty Institute for Precision Medicine, Seattle, WA 98195
8. Howard Hughes Medical Institute, Seattle, Washington, USA

† Equally contributing

* Corresponding authors

Correspondence to: coletrap@uw.edu, shendure@uw.edu

Abstract:

Mammalian organogenesis is an astonishing process. Within a short window of time, the cells of the three germ layers transform into an embryo that includes most major internal and external organs. Here we set out to investigate the transcriptional dynamics of mouse organogenesis at single cell resolution. With sci-RNA-seq3, we profiled ~2 million cells, derived from 61 embryos staged between 9.5 and 13.5 days of gestation, in a single experiment. The resulting 'mouse organogenesis cell atlas' (MOCA) provides a global view of developmental processes during this critical window. We identify hundreds of cell types and 56 trajectories, many of which are detected only because of the depth of cellular coverage, and collectively define thousands of corresponding marker genes. With Monocle 3, we explore the dynamics of gene expression within cell types and trajectories over time, including focused analyses of the apical ectodermal ridge, limb mesenchyme and skeletal muscle.

Introduction:

Most studies of mammalian organogenesis rely on model organisms, and in particular, the mouse. Mice develop quickly, with just 21 days between fertilization and birth. The implantation of the blastocyst (E4.0) is followed by gastrulation and the formation of germ layers (E6.5-E7.5)(*1*, *2*). At the early-somite stages, the embryo transits from gastrulation to early organogenesis, forming the neural plate and heart tube (E8.0–E8.5). In the ensuing days (E9.5-E13.5), the embryo expands from hundreds-of-thousands to over ten million cells, and concurrently develops nearly all major organ systems. Unsurprisingly, these four days have been intensively studied. Indeed, most genes underlying major developmental defects can be studied in this window(*3*, *4*).

The transcriptional profiling of single cells (scRNA-seq) represents a promising avenue for obtaining a global view of developmental processes(*5–7*). For example, scRNA-seq recently revealed remarkable heterogeneity in neurons and myocardiocytes during mouse development(*8*, *9*). However, although two scRNA-seq atlases of mouse were recently released(*10*, *11*), they are mostly restricted to adult organs, and do not attempt to characterize the emergence and dynamics of cell types during development.

Results:

Single cell RNA-seq of 2 million cells

Single cell combinatorial indexing ('sci-') is a methodological framework involving split-pool barcoding of cells or nuclei(*12–19*). We previously developed sci-RNA-seq and applied it to generate 50-fold shotgun coverage of the cellular content of L2 stage *Caenorhabditis elegans(17)*. A conceptually identical method was recently termed Split-Seq(*20*). To increase the throughput, we explored >1,000 experimental conditions (**Extended Data Fig. 1ab**; **Methods**). The major improvements of the resulting method, sci-RNA-seq3, include: (i) Nuclei are extracted directly from fresh tissues without enzymatic treatment, then fixed and stored. (ii) For the third level of indexing(*17*), we switched from Tn5 tagmentation to hairpin ligation. (iii) Individual enzymatic reactions were optimized. (iv) FACS sorting was replaced by dilution, and sonication and filtration steps added to minimize aggregation. Even without automation, sci-RNA-seq3 library preparation can be completed through the intensive effort of a single individual in one week at a cost of less than $0.01 per cell.

We collected 61 C57BL/6 mouse embryos at E9.5, E10.5, E11.5, E12.5 or E13.5, and snap froze them in liquid nitrogen. Nuclei from each embryo were isolated and deposited to different wells, such that the first index identified the originating embryo of any given cell. As a control, we

spiked a mixture of human HEK293T and mouse NIH/3T3 nuclei into two wells. The resulting sci-RNA-seq3 library was sequenced in a single Illumina NovaSeq run, yielding 11 billion reads (**Fig. 1a**; **Extended Data Fig. 1cd**).

From one experiment, we recovered 2,058,652 cells from mouse embryos and 13,359 cells from HEK293T or NIH/3T3 cells (UMI (unique molecular identifier) count ≥ 200). Transcriptomes from human/mouse control wells were overwhelmingly species-coherent (3% collisions), with performance similar to previous experiments(*17*) (**Extended Data Fig. 1e-i**). A limitation is that only ~7% of cells entering the experiment were ultimately profiled, with losses largely consequent to filtration steps intended to remove aggregates of nuclei.

We profiled a median of 35,272 cells per embryo (**Fig. 1b**; **Extended Data Fig. 1j**). Despite shallow sequencing (~5,000 raw reads per cell; 46% duplicate rate), we recovered a median of 671 UMIs (519 genes) per cell (**Extended Data Fig. 1k**). 3.7-fold deeper sequencing of a subset of wells nearly doubled complexity (to a median of 1,142 UMIs per cell; 87% duplicate rate). As we are profiling RNA in nuclei, 59% of UMIs per cell strand-specifically mapped to introns, and 25% to exons. Our profiles may thus primarily reflect nascent transcription, temporally offset but also predictive(*21*) of the cellular transcriptome. Later stage embryos exhibited somewhat reduced UMI counts, possibly reflecting decreasing nuclear mRNA content (**Extended Data Fig. 1l**). We used Scrublet(*22*) to detect 4.3% likely doublet cells, corresponding to a doublet estimate of 10.3% including both within- and between-cluster doublets (**Extended Data Fig. 1mn**).

Based on our rough estimates of the number of cells per embryo at each timepoint (**Methods**), our 'shotgun cellular coverage' of the mouse embryo is 0.8x at E9.5 (200K cells/embryo; 152K profiled across all replicates), 0.3x at E10.5 (1.1M cells/embryo; 378K profiled), 0.2x at E11.5 (2.6M cells/embryo; 616K profiled), 0.08x at E12.5 (6M cells/embryo; 475K profiled), and 0.03x

at E13.5 (13M cells/embryo; 437K profiled). Thus, although we are not yet oversampling(*17*), our depth of profiling is equivalent to 3-80% of the cellular content of an individual mouse embryo.

Embryos were readily identified as male (n = 31) or female (n = 30) (**Extended Data Fig. 1op**). Applying t-stochastic neighbor embedding (t-SNE) to "pseudo-bulk" profiles (aggregating the transcriptomes of each embryo's cells) resulted in five tightly clustered groups corresponding to developmental stages (**Extended Data Fig. 1q**). We also ordered the mouse embryos along a pseudotime trajectory(*23*) (**Fig. 1c**). Two prominent gaps (E9.5-E10.5 and E11.5-E12.5) suggest particularly dramatic changes during these windows (**Extended Data Fig. 1rs**). In these pseudo-bulk profiles, 12,236 genes were differentially expressed across developmental stages.

Identification of cell types and subtypes

We subjected the 2,058,652 single cell transcriptomes to Louvain clustering and t-SNE visualization (**Fig. 2a**). Reassuringly, cells from replicate embryos of the same developmental stage were similarly distributed, whereas cells from different stages were not (**Extended Data Figs. 2a-f**). Based on genes specific to each of 40 clusters, we manually annotated cell types. Merging two clusters both corresponding to the definitive erythroid lineage and discarding a putative doublet cluster (detected doublet rate of 52%) yielded 38 major cell types (**Fig. 2b**; **Extended Data Fig. 2g**).

In general, highly specific marker genes made the annotation of these major cell types straightforward (**Fig. 2b**). For example, cluster 6 (epithelial cells) specifically expressed *Epcam* and *Trp63*(*24*)(*25*), while cluster 29 (hepatocytes) specifically expressed *Afp* and *Alb*(*10*). Smaller clusters were readily annotated as well. For example, cluster 36 (melanocytes) specifically expressed *Tyr* and *Trpm1*(*26, 27*), while cluster 37 (lens) specifically expressed *Cryba2*. Some

markers, although observed in a substantial proportion of cells in many clusters, were much more highly expressed in one cluster (*e.g. Hbb-bh1* in primitive erythroid cells). For clusters corresponding to the embryonic mesenchyme and connective tissue, annotation was more challenging because fewer markers are known (*e.g. Fndc3c1* in early mesenchyme; **Extended Data Fig. 2h**)

17,789 of 26,183 genes (68%) were differentially expressed across the major cell types (5% FDR). Amongst these, we identified 2,863 cell type-specific marker genes (mean 75; those with >2-fold expression difference between first and second ranked cell type; a cutoff of >5-fold yielded 932 marker genes; **Extended Data Fig. 2i**). The vast majority of these markers are novel. For example, we detect the highest expression of sonic hedgehog (*Shh*)(*28*) in the notochord (cluster 30), together with *Ntn1, Slit1, and Spon1*, all known to be expressed in the cells of the notochord and floor plate during development(*29–31*). However, *Tox2*, *Stxbp6*, *Schip1*, *Frmd4b,* not previously been described as markers of the notochord, were also markers of cluster 30. Whole-mount *in situ* hybridization (WISH) of *Shh* (known) and *Tox2* (novel) confirmed both genes are expressed in notochord at E10.5 (**Extended Data Fig. 2j**).

We observed marked changes in the proportions of cell types during organogenesis. While most major cell types proliferated exponentially, a few were transient and disappeared by E13.5 (**Extended Data Fig. 2kl**). For example, at E9.5, we detect cells corresponding to the primitive erythroid lineage, originating from the yolk sack (cluster 26; marked by *Hbb−bh1*). However, the definitive erythroid lineage, originating from the fetal liver (cluster 22; marked by *Hbb−bs*), progressively displaces it to become the exclusive red cell lineage by E13.5 (**Fig. 2a**; **Extended Data Fig. 2m**).

The 38 major cell types are represented by a median of 47,073 cells, the largest containing 144,648 cells (connective tissue progenitors), and the smallest only 1,000 cells (neutrophils). As additional heterogeneity was readily apparent, we adopted an iterative strategy, repeating Louvain clustering on each major cell type. After subclusters dominated by a few embryos were removed and highly similar subclusters merged (**Methods**), 655 subclusters were identified (**Extended Data Fig. 3**). As an operational definition specific to this manuscript, we refer to the 38 major clusters as cell types, and the 655 subclusters as subtypes. Notably, our sensitivity to detect cell types and subtypes in this study was dependent on the large number of cells profiled (**Extended Data Fig. 4a-d**).The 655 subtypes consist of a median of 1,869 cells, and range from 51 (a subtype of notochord) to 65,894 (a subtype of connective tissue progenitors) cells (**Extended Data Fig. 4e-g**).

We annotated 13% of subtypes as likely artifacts (>10% of cells in these subtypes are predicted doublets; **Extended Data Fig. 4h**). For the remaining 572 subtypes, we identified a median of 20 subtype-specific markers (>2-fold expression difference between first and second ranked cell subtypes of the corresponding major cell type; **Extended Data Fig. 4ij**). Furthermore, most subtypes can be distinguished from all 571 other non-doublet subtypes based on marker gene sets and >4-fold expression differences (63% with 2 markers, 95% with 4 markers; **Methods**; **Extended Data Fig. 4k**).

As there are presently no comparable single cell atlases of E9.5-E13.5, we compared MOCA subtypes to 130 fetal cell types (E14.5) of a recent mouse cell atlas (MCA)(*10*). With a new inter-study cross-matching method, we matched 96 MCA cell types to 58 MOCA subtypes (**Methods**; **Extended Data Fig. 5a-c**). As expected, MOCA subtypes that failed to match MCA cell types tended to derive from earlier stages (*e.g.* neural tube) or were rare (*e.g.* lens), while MCA cell types that failed MOCA subtypes were mostly tissue-specific immune or epithelial cells, potentially

because they emerge after E13.5. Nonetheless, the atlases unquestionably inform one another, as the MCA's anatomical resolution is useful for localizing MOCA subtypes, while MOCA's developmental focus informs the embryonic origin of MCA cell types (**Extended Data Fig. 5b**). As an example of the former, a subcluster of endocrine epithelial cells in MOCA mapped to both the acinar and endocrine cells of the fetal stomach in the MCA. As an example of the latter, "cells in cell cycle" in the MCA's fetal kidney mapped to a subtype of intermediate mesoderm in MOCA, plausibly corresponding to progenitors of the kidney. A similar analysis matched 48 cell types annotated in a recent mouse brain atlas (BCA)(*32*) to 68 MOCA subtypes with high specificity (**Extended Data Fig. 5d**).

Characterization of the apical ectodermal ridge

We annotated all subtypes of epithelium and endothelium (clusters 6 and 20, respectively; **Fig. 3a**; **Extended Data Fig. 6a-c**). For example, epithelial subtype 6.8 was marked by *Oc90*, exclusively expressed in the epithelium of the otic vesicle(*33*); epithelial subtype 6.23 by *Fgf8, Msx2,* and *Rspo2*, known markers of the apical ectodermal ridge (AER)(*34*); and endothelial subtype 20.12 by *Tbx20* and *Tmem108*, specific to endocardial cells and cardiac valve endothelium(*35*, *36*).

To investigate a subtype in more detail, we focused on the AER, a highly specialized epithelium involved in digit development(*37*). In addition to known markers for AER, subtype 6.23 (1,237 cells; 0.06% of MOCA) was distinguished by expression of *Fndc3a, Adamts3, Slc16a10, Snap91,* and *Pou6f2*. WISH of *Fgf8* (known), *Fndc3a, Adamts3,* and *Snap91* (all novel) confirmed expression specific to the most distal tip of the limb bud representing the AER at E10.5 or E11.5 (**Fig. 3b-e**).

We next examined the dynamics of AER proliferation and gene expression. Although detected at all timepoints and nearly all embryos, the estimated number of AER cells per embryo peaks between E10.5 and E11.5 (**Fig. 3f**), consistent with a previous report(*38*) and our validations (**Fig. 3c**). We performed pseudotemporal ordering of AER cells, yielding a simple early-to-late trajectory and 710 differentially expressed genes (5% FDR; **Fig. 3gh**; **Extended Data Fig. 6d**). For example, *Fgf8, Fgf9(39)* and *Rspo2(34)* are preceded in their activation dynamics by *Fndc3a*. Genes whose expression significantly decreased include *Mki67* and *Igf2*, which have roles in promoting cellular proliferation(*40*, *41*). Pathway-level analyses also showed the downregulation of proliferative programs in this window (**Extended Data Fig. 6ef**).

Reconstructing developmental trajectories

We next sought to investigate the developmental trajectories that cell types traverse during mammalian organogenesis. Most contemporary algorithms for trajectory reconstruction assume a continuous manifold (whereas our data begin at E9.5, and therefore are missing at least some ancestral states) and do not allow for convergence of cell fates (whereas some cell types are known to derive from multiple transcriptionally distinct lineages). To overcome these limitations while also enabling scaling to millions of cells, we developed a new version of Monocle(*42*). Monocle 3 first projects cells onto a low-dimensional space encoding transcriptional state using UMAP(*43*). It then groups mutually similar cells using the Louvain community detection algorithm, and merges adjacent groups into 'supergroups'(*44*). Finally, it resolves the paths or trajectories that individual cells can take during development, identifying the locations of branches and convergences within each supergroup.

Subsequent to a focused application of Monocle 3 to cells corresponding to the limb bud mesenchyme (**Supplementary Note 1**; **Extended Data Fig. 7**; **Supplementary Tables 7-9**), we applied it to identify major developmental trajectories across the entire dataset. Monocle 3 organized 1,524,792 high-quality cells (UMI > 400) into twelve groups. We merged two groups corresponding to sensory neurons, and another two corresponding to blood cells. Nearly all of the 38 major cell types fall almost exclusively in one of the ten resulting trajectories (**Fig. 4a-b**; **Extended Data Fig. 8ab**). The two most complex structures are the *neural tube/notochord trajectory*, which includes the notochord, neural tube, progenitor and developing neuronal and glial cell types, and the *mesenchymal trajectory*, which includes all mesenchymal and muscle cell types. There are three *neural crest trajectories*, corresponding to sensory neurons, Schwann cell precursors and melanocytes. The *hematopoietic trajectory* includes megakaryocytes, erythrocytes and white blood cells, while the remaining four trajectories (*endothelial, epithelial, hepatic, lens*) each correspond to a single major cell type (**Fig. 4b**). The discontinuity between these ten major trajectories likely reflects the lack of representation of some intermediate or ancestral states, consequent to our study beginning at E9.5. Although the estimated number of cells per embryo in each trajectory increases exponentially, their proportions remain relatively stable, with the exception of hepatocytes, which markedly increase their contribution from 0.3% at E9.5 to 2.8% at E13.5 (**Extended Data Fig. 8c**).

Unlike t-SNE, UMAP places related cell types near one another. For example, cell types found at later developmental timepoints such as inhibitory neurons are connected to early CNS precursors (radial glia) by a 'bridge' of neural progenitor cells; however, the same radial glial cells project in a different direction towards increasingly mature oligodendrocytes (**Fig. 4a**, left). Similarly, early mesenchymal cells radiate from a defined region into myocytes, limb mesenchyme, chondrocytes/osteoblasts and connective tissues (**Fig. 4a**, right).

After removing 12% of cells corresponding to doublet-annotated cells and/or subclusters, we iteratively reanalyzed the ten major trajectories (**Fig. 5**; **Extended Data Fig. 9**). For example, the epithelial trajectory breaks into several discontinuous subtrajectories, each emanating from a focal concentration of E9.5-derived cells and projecting in one or more directions, through cells corresponding to progressively later timepoints (**Fig. 4c**; **Extended Data Fig. 8d**). Notably, the AER subtrajectory projects out of surface ectoderm and then back into epidermis, consistent with its transitory nature.

We mapped the 572 subtypes defined by t-SNE and Louvain clustering to the developmental subtrajectories defined by Monocle 3 (**Extended Data Fig. 9**). The vast majority of subtypes mapped to a single subtrajectory, often as temporally restricted subsetsWe annotated the subtrajectories on the basis of marker genes of subtypes mapping to them. The resulting 56 developmental subtrajectories span all major systems including the CNS, PNS, respiratory, digestive, cardiovascular, immune, lymphatic, urinary, endocrine, integumentary, skeletal, muscular and reproductive systems (**Fig. 5**; **Extended Data Fig. 10**).

In some cases, we observe a single, simple linear trajectory. However, we also observe many examples of branching trajectories, as well as of cell types that appear to be generated via multiple parallel paths. As an example of the latter, both CNS excitatory and inhibitory neurons appear to develop through multiple, convergent trajectories, possibly due to the maturation in multiple anatomical locations. Other subtrajectories exhibited even more complex features, including multiple starting and ending points within a continuous structure (*e.g.* intermediate mesoderm trajectory).

Although Monocle 3 did not have access to these labels, the subtrajectories are highly consistent with developmental time (*i.e.* cells ordered from E9.5 to E13.5, **Extended Data Fig. 9-10**). To orient subtrajectories, we identified one or several starting points as focal concentrations

of E9.5 cells, and then computed developmental pseudotime for cells present along various paths (**Extended Data Fig. 11**; **Methods**). We also annotated each subtype according to the subtrajectory to which it maps, as well as its relative temporal position within that subtrajectory (*e.g.* subtype 6.14 = "auditory epithelial trajectory.1-of-3"). These representations provide a starting point for more detailed explorations of the 572 subtypes and 56 subtrajectories.

Reconstructing skeletal myogenesis

To investigate a developmental process in greater detail, we focused on developing muscle, which is comprised of distinct mesodermal lineages that form prior to E9.5(*45*). We hypothesized that the myogenic trajectory would feature multiple entry points that feed cells into a common path corresponding to activation of the core gene expression program shared by myotubes.

To test this, we *in silico* isolated myocytes and their putative 'ancestral' cells from the mesenchyme trajectory (**Fig. 6a**; **Methods**). Next, we used Monocle 3 to construct a myogenesis-specific trajectory, which featured multiple focal concentrations of E9.5 cells, with cells from later stages distributed over several paths radiating outward (**Fig. 6a**). *Pax3* and *Pax7*, which mark skeletal muscle progenitors, were expressed over a broad swath of the principal graph (**Fig. 6b**). Cells expressing *Myf5* co-localized with a subset of *Pax7*+ cells, consistent with the role of *Myf5* in embryonic myogenesis(*46*). From this region of the trajectory, two parallel linear segments emanated, on which cells expressed either *Myf5* or *Myod*. Both paths terminate with cells expressing *Myog* or *Myh3*, markers of myocytes and myotubes, respectively. The cells on the *Myf5*+ path, largely from early time points, also expressed higher levels of genes in the Robo/Slit signaling pathway, which has been implicated in driving "pioneer myoblasts" to form embryonic myofibers(*47*) (**Extended Data Fig. 12**). An additional path traversed by cells from E9.5, which expressed *Lhx2*, *Tbx1*, and *Pitx2* but very low levels of *Pax3*, feeds into the trajectory just upstream

of the *Myf5* and *Myod1* segments, and possibly corresponds to pharyngeal mesoderm(*45*). Overall, the trajectory is consistent with the view that different mesodermal lineages use distinct factors to converge on a core program of muscle genes (**Fig. 6c**). Globally, we detected 2,908 genes expressed in a trajectory-dependent manner (FDR < 0.05 and Moran's $I > 0.01$) that grouped into 14 distinct patterns (**Extended Data Fig. 12**).

Discussion

Here, to obtain a global view of mammalian organogenesis, we profiled the transcriptomes of ~2 million cells from mouse embryos spanning E9.5 to E13.5. In the resulting atlas (MOCA), we identify over 500 subtypes of cells and 56 developmental subtrajectories, each distinguished by multiple marker genes and collectively spanning essentially every organ system. With sci-RNA-seq3, we introduce a technical framework for individual labs to generate datasets corresponding to millions of single cells. With Monocle 3, we introduce a computational framework for trajectory inference that operates at this same scale. These data constitute a potentially foundational resource for the mammalian developmental biology field. MOCA and the underlying data are made freely available, together with a website to facilitate their further exploration (http://atlas.gs.washington.edu/mouse-rna/).

MOCA has limitations. First, although not sequenced to saturation, the cell-by-gene matrix is sparse. Nonetheless, our results support the view that cell types are readily distinguishable despite hundreds rather than thousands of UMIs per cell(*48*). Of course, the tradeoff between breadth and depth depends on one's goals. As an example supporting the 'many cells, few UMIs per cell' approach, consider primordial germ cells, which were readily identifiable despite their rarity (subtypes 16.13 and 6.27, which sum to 269/2,058,652 cells or 0.01% of MOCA). Nonetheless,

despite its unprecedented depth, our study does not exceed 1-fold coverage of the mouse embryo at any timepoint, and it is possible that we are missing extremely rare cell types.

Second, although we are reasonably confident in our annotations, they should be regarded as preliminary. Mid-gestational mouse development has not previously been extensively studied at single cell resolution, and many published markers have limited specificity. Furthermore, because we studied disaggregated whole embryos, the assignment of anatomical specificity is challenging. We anticipate that the comprehensive annotation of MOCA will benefit from community input and domain expertise, and to that end created an interactive wiki (http://atlas.gs.washington.edu/mouse-rna/). Inevitably, however, additional experiments (*e.g. in situ* analyses of marker genes) will be necessary to resolve ambiguities. Importantly for future atlasing efforts, we found the annotation of temporally-resolved developmental trajectories to be much more straightforward than that of cell types.

A long-standing dream, perhaps at last within sight from a technical perspective, is a comprehensive, spatiotemporally-resolved molecular atlas of mammalian development at single cell resolution. To this end, the mouse has several advantages, including its small size, the accessibility of early developmental timepoints, an inbred genetic background, and genetic manipulability. It also seems likely that 'whole organism' profiling of small mammals will be essential for identifying the inevitable gaps in any efforts to generate a comprehensive atlas of human cell types.

Single cell atlases of the development of wild-type mice may also represent an important step towards understanding pleiotropic developmental disorders at the organismal scale, and for detailed investigations of subtle roles for genes and regulatory sequences in development. For example, many knockouts of both coding and conserved regulatory sequences do not show any

abnormalities with conventional phenotyping(*49*). We anticipate that 'whole organism' sc-RNA-seq will empower reverse genetics, *e.g.* potentially enabling the discovery of subtle defects in the molecular programs or the relative proportions of specific cell types(*50*).

Endnotes

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank members of the Shendure and Trapnell labs, especially D. Cusanovich, R. Daza, G. Findlay, A. McKenna, H. Pliner and V. Ramani, as well as L. McInnes, D. Beier, N. Ahituv and S. Tapscott, for helpful discussions and feedback. We also thank R. Hunter, and R. Rualo in the Transgenic Resources Program of University of Washington and N. Brieske and A. Stiege at the Max Planck Institute for Molecular Genetics for their exceptional assistance. We thank S. Geuer for the *Fndc3a* probe. M.S. was supported by a grant from the Deutsche Forschungsgemeinschaft (SP1532/2-1). This work was funded by the Paul G. Allen Frontiers Group (Allen Discovery Center grant to J.S. and C.T.), grants from the NIH (DP1HG007811 and R01HG006283 to J.S.; DP2 HD088158 to C.T.), the W. M. Keck Foundation (to C.T. and J.S.). J.S. is an Investigator of the Howard Hughes Medical Institute.

Author Contributions J.C. developed techniques and performed sci-RNA-seq3 experiments with assistance from M.S., F.Z., L.C., F.S.; M.S. performed embryo collection and in-situ hybridization validations with assistance from D.I. and S.M.; J.C. and C.T. performed computation analysis with assistance from M.S., X.Q. and A.H.; X.Q. and C.T. developed Monocle 3. X.H. developed website with assistance from J.C.; J.S. and C.T. supervised the project; J.S., C.T., J.C. and M.S. conceived the project and wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. L.C., F.Z. and F.S. declare competing financial interests in the form of stock ownership and paid employment by Illumina, Inc. One or more embodiments of one or more patents and patent applications filed by Illumina may encompass the methods, reagents, and data disclosed in this manuscript. Some work in this study may be related to technology described in the following exemplary published patent applications: WO2010/0120098 and WO2011/0287435. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.S. (shendure@uw.edu) or C.T. (coletrap@uw.edu).

Main figures

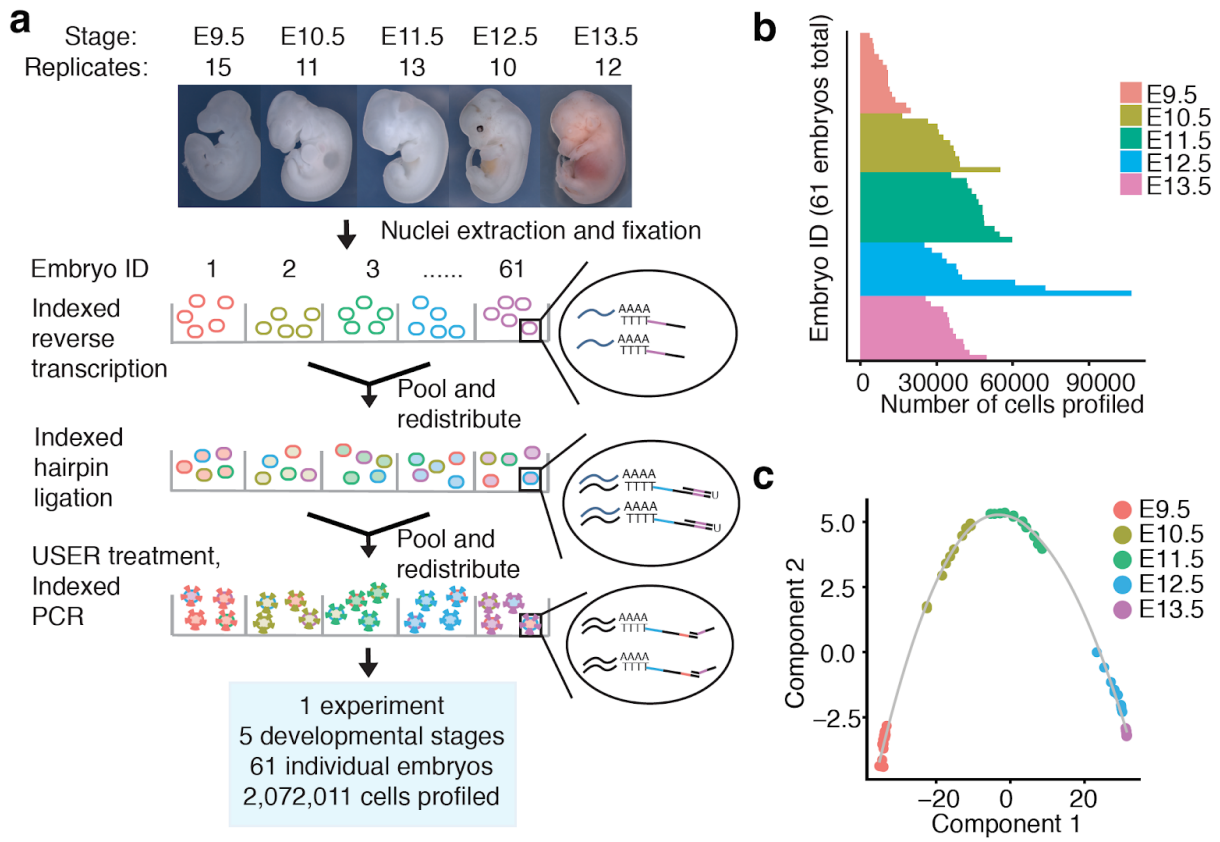


Fig. 1. sci-RNA-seq3 enables profiling of 2,072,011 cells from 61 mouse embryos across 5 developmental stages in a single experiment. (**a**) sci-RNA-seq3 workflow and experimental scheme. (**b**) Bar plot showing number of cells profiled from each of 61 mouse embryos. (**c**) Pseudotime trajectory of pseudobulk RNA-seq profiles of mouse embryos.

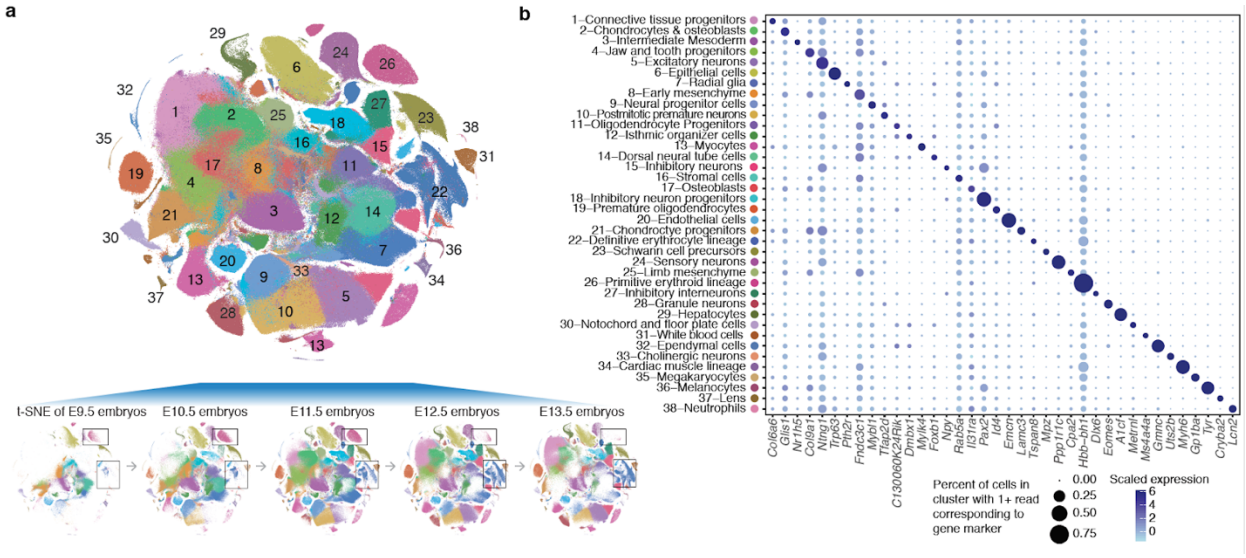


Fig. 2. Identifying the major cell types of mouse organogenesis. (**a**) t-SNE visualization of 2,026,641 mouse embryo cells, colored by cluster id from Louvain clustering (in Fig. 2b), and annotated based on marker genes. The same t-SNE is plotted below, showing only cells from each stage (cell numbers from left to right: n = 151,000 for E9.5; 370,279 for E10.5; 602,784 for E11.5; 468,088 for E12.5; 434,490 for E13.5). Primitive erythroid (transient) and definitive erythroid (expanding) clusters are boxed. (**b**) Dot plot showing expression of one selected marker gene per cell type. The size of the dot encodes the % of cells within a cell type in which that marker was detected, and its color encodes the average expression level.

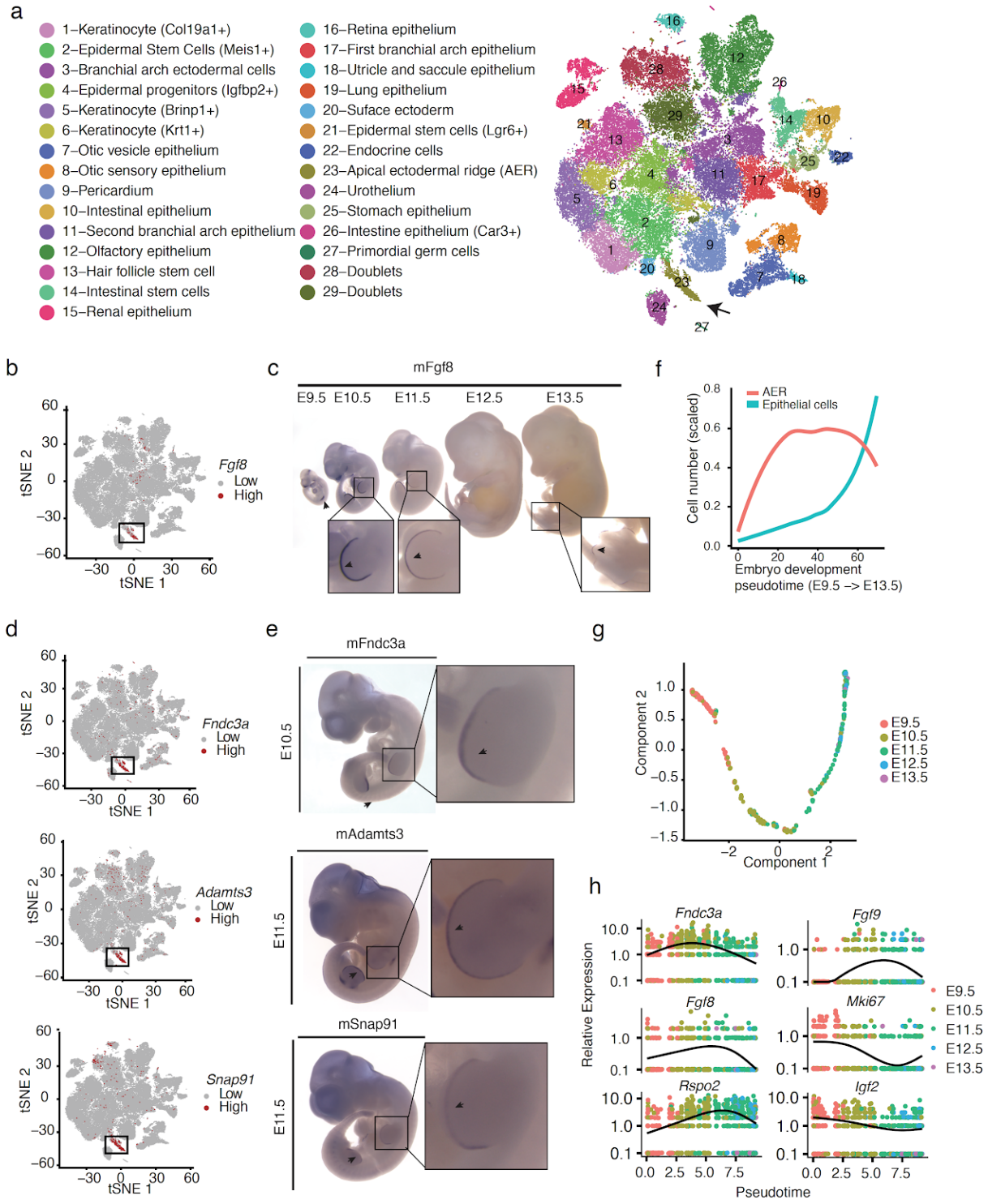


Fig. 3. Identification and characterization of epithelial cell subtypes and the limb apical ectodermal ridge (AER). (**a**) t-SNE visualization and marker-based annotation of epithelial cell

subtypes (74,651 cells). (**b**) t-SNE visualization of all epithelial cells colored by expression level of *Fgf8*. "High" indicates cells with UMI count for *Fgf8* > 1. (**c**) *In situ* hybridization images of *Fgf8* in embryos from E9.5 to E13.5. Arrow: site of gene expression. n = 5 (**d, e**) t-SNE visualization of all epithelial cells colored by expression level (d) and whole *in situ* hybridization images (e) of *Fndc3a* (top), *Adamts3* (middle) and *Snap91* (bottom). n = 5 "High" indicates cells with UMI count for *Fndc3a* > 3, *Adamts3* > 1, *Snap91* > 1. Arrow: site of gene expression. (**f**) Line plot showing the estimated relative cell numbers for epithelial cells and AER cells, calculated as in Extended Data Fig. 2m. Data points for individual embryos were ordered by development pseudotime and smoothed by loess method. (**g**) Pseudotime trajectory of AER single cell transcriptomes (cell number n = 1,237), colored by development stage. (**h**) Kinetics plot showing relative expression of AER marker genes across developmental pseudotime.

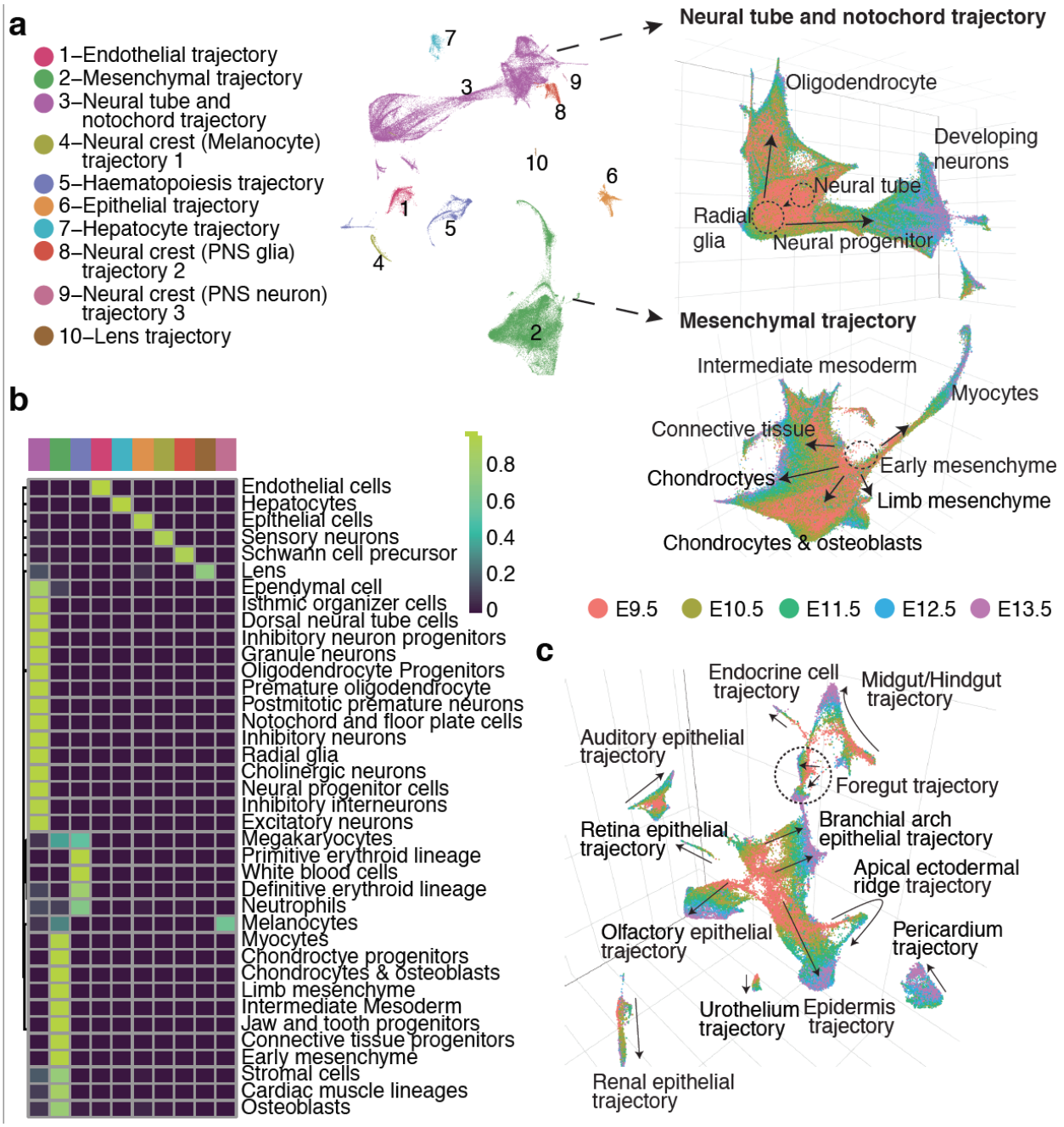


Fig. 4. Characterization of ten major developmental trajectories present during mouse organogenesis. (**a**) UMAP 3D visualization of our overall dataset; left: views from one direction; bottom: zoomed view of neural tube/notochord (top) and mesenchymal (bottom) trajectories, colored by development stage. PNS: peripheral nervous system. (**b**) Heatmap showing the proportion of cells from each of the 38 major cell types (rows) assigned to each of the 10 major

trajectories (columns; color key in left panel of a). (**c**) UMAP 3D visualization of epithelial subtrajectories colored by development stage (color key in right panel of a).

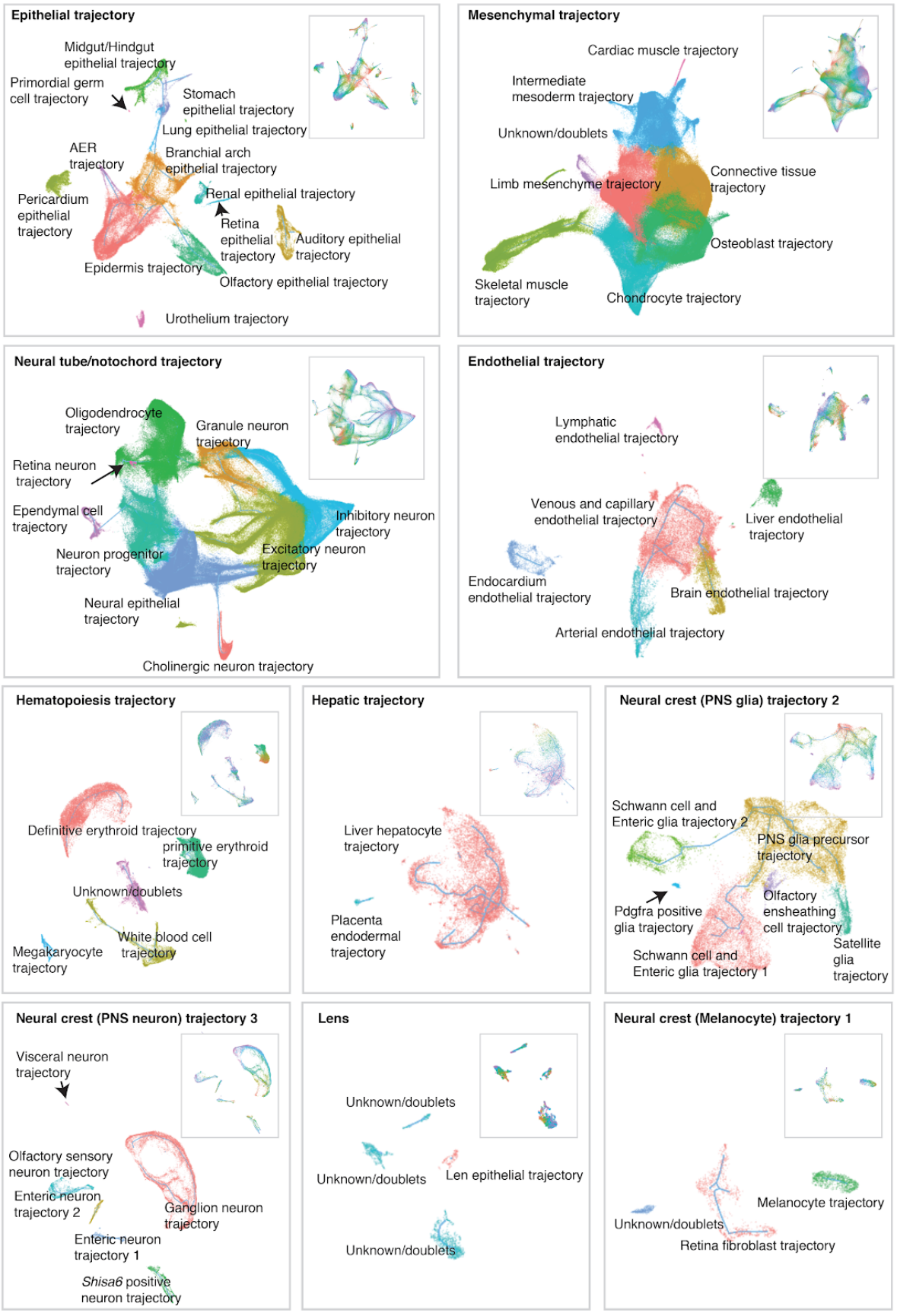

Epithelial trajectory
Midgut/Hindgut epithelial trajectory
Primordial germ cell trajectory
Stomach epithelial trajectory
Lung epithelial trajectory
AER trajectory
Branchial arch epithelial trajectory
Pericardium epithelial trajectory
Renal epithelial trajectory
Retina epithelial trajectory
Auditory epithelial trajectory
Epidermis trajectory
Olfactory epithelial trajectory
Urothelium trajectory
Mesenchymal trajectory
Cardiac muscle trajectory
Intermediate mesoderm trajectory
Unknown/doublets
Limb mesenchyme trajectory
Connective tissue trajectory
Osteoblast trajectory
Skeletal muscle trajectory
Chondrocyte trajectory
Neural tube/notochord trajectory
Oligodendrocyte trajectory
Granule neuron trajectory
Retina neuron trajectory
Ependymal cell trajectory
Inhibitory neuron trajectory
Neuron progenitor trajectory
Excitatory neuron trajectory
Neural epithelial trajectory
Cholinergic neuron trajectory
Endothelial trajectory
Lymphatic endothelial trajectory
Venous and capillary endothelial trajectory
Liver endothelial trajectory
Endocardium endothelial trajectory
Brain endothelial trajectory
Arterial endothelial trajectory
Hematopoiesis trajectory
Definitive erythroid trajectory
primitive erythroid trajectory
Unknown/doublets
White blood cell trajectory
Megakaryocyte trajectory
Hepatic trajectory
Liver hepatocyte trajectory
Placenta endodermal trajectory
Neural crest (PNS glia) trajectory 2
Schwann cell and Enteric glia trajectory 2
PNS glia precursor trajectory
Pdgfra positive glia trajectory
Olfactory ensheathing cell trajectory
Schwann cell and Enteric glia trajectory 1
Satellite glia trajectory
Neural crest (PNS neuron) trajectory 3
Visceral neuron trajectory
Olfactory sensory neuron trajectory
Enteric neuron trajectory 2
Ganglion neuron trajectory
Enteric neuron trajectory 1
Shisa6 positive neuron trajectory
Lens
Unknown/doublets
Unknown/doublets
Len epithelial trajectory
Unknown/doublets
Neural crest (Melanocyte) trajectory 1
Melanocyte trajectory
Unknown/doublets
Retina fibroblast trajectory

Fig. 5. UMAP visualization of individual major trajectories. After removing doublet-annotated cells and subclusters, we iteratively reanalyzed each of the ten major trajectories. Colored by subtrajectory name (main plots) or developmental stage (insets; colors as in Fig. 4c). Edges in the principal graphs that define trajectories reported by Monocle 3 are shown as light blue line segments.

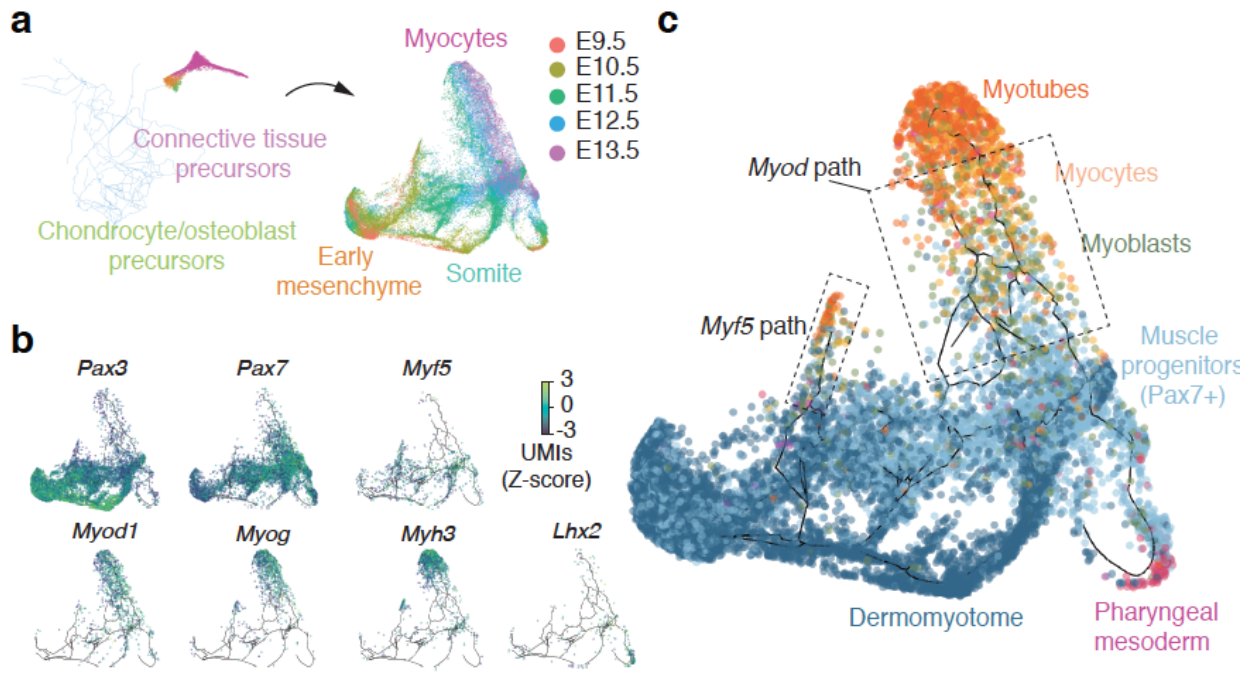


Fig. 6. Resolving cellular trajectories in myogenesis. Edges in the principal graphs that define trajectories reported by Monocle 3 are shown as light blue line segments. (**a**) Cells putatively involved in myogenesis were isolated from the mesenchymal cell trajectory *in silico* and then used to construct a myocyte subtrajectory. Principal graph nodes with more than 50% occupied by cells from cluster 13 were taken as "seed nodes" and then cells on any nodes within 20 edges of these seed nodes were selected for subtrajectory analysis. Cells in the myocyte subtrajectory (left) colored by developmental stage (right). (**b**) Cells in the myocyte trajectory, colored by their expression of selected transcriptional regulators of myogenesis. Cells with no detectable expression for a given gene are omitted from its plot. Values are log10-transformed, standardized

UMI counts. (**c**) Cells classified by developmental stage according to the markers shown in panel c (Dermomyotome: *Pax3*+, *Pax7*-; Muscle progenitors: *Pax7*+; Myoblasts: *Myf5*+ or *Myod*+ and *Myog*-; Myocytes: *Myog*+; Myotubes: *Myh3*+).

METHODS

Data reporting

No statistical methods were used to predetermine sample size. Embryos used in experiment were randomized before sample preparation. Investigators were blinded to group allocation during data collection and analysis: embryo collection and sci-RNA-seq3 analysis were performed by two different researchers.

Embryo dissection

The C57BL/6 mice were obtained from The Jackson Laboratory (Bar Harbor, ME) and plug matings were set up. Noon on the day of the vaginal plug was considered as embryonic day (E) 0.5. Dissections were done as previously described(*52*) and all embryos were immediately snap frozen in liquid nitrogen. Embryos were collected from at least three independent litters per development stage. All animal procedures were in accordance with institutional, state, and government regulations and approved by the Office of Animal Welfare (OAW) under the IACUC protocol 4378-01.

Whole-mount in situ hybridization

The mRNA expression in E9.5-E13.5 mouse embryos was assessed by whole mount *in situ* hybridisation (WISH) using a digoxigenin-labeled antisense riboprobe transcribed from a cloned gene specific probes (PCR DIG Probe Synthesis Kit, Roche). Whole embryos were fixed overnight

in 4% PFA/PBS. The embryos were washed in PBST (0.1% Tween), and dehydrated stepwise in 25%, 50% and 75% methanol/PBST and finally stored at -20°C in 100% methanol. The WISH protocol was as follows: Day 1) Embryos were rehydrated on ice in reverse methanol/PBST steps, washed in PBST, bleached in 6% H_2O_2/PBST for 1 hour and washed in PBST. Embryos were then treated in 10 µg/ml Proteinase K/PBST for 3 minutes, incubated in glycine/PBST, washed in PBST and finally re-fixed for 20 minutes with 4% PFA/PBS, 0.2% glutaraldehyde and 0.1% Tween 20. After further washing steps with PBST, embryos were incubated at 68°C in L1 buffer (50% deionised formamide, 5x SSC, 1% SDS, 0.1% Tween 20 in DEPC; pH 4.5) for 10 minutes. Next, embryos were incubated for 2 hours at 68°C in hybridisation buffer 1 (L1 with 0.1% tRNA and 0.05% heparin). Afterwards, embryos were incubated o.n. at 68°C in hybridisation buffer 2 (hybridisation buffer 1 with 0.1% tRNA and 0.05% heparin and 1:500 DIG probe). Day 2) Removal of unbound probe was done through a series of washing steps 3x30 minutes each at 68°C: L1, L2 (50% deionised formamide, 2x SSC pH 4.5, 0.1% Tween 20 in DEPC; pH 4.5) and L3 (2x SSC pH 4.5, 0.1% Tween 20 in DEPC; pH 4.5). Subsequently, embryos were treated for 1 hour with RNase solution (0.1 M NaCl, 0.01 M Tris pH 7.5, 0.2% Tween 20, 100 µg/ml RNase A in H_2O), followed by washing in TBST 1 (140mM NaCl, 2.7mM KCl, 25mM Tris-HCl, 1% Tween 20; pH 7.5). Next, embryos were blocked for 2 hours at RT in blocking solution (TBST 1 with 2% calf-serum and 0.2% BSA), followed by incubation at 4°C o.n. in blocking solution containing 1:5000 Anti-Digoxigenin-AP (catalog number: Roche-11093274910) . Day 3) Removal of unbound antibody was done through a series of washing steps 8x 30 min at RT with TBST 2 (TBST with 0.1% Tween 20, and 0.05% levamisole/tetramisole) and left o.n. at 4°C. Day 4) Staining of the embryos was initiated by washing at RT with alkaline phosphatate buffer (0.02 M NaCl, 0.05 M $MgCl_2$, 0.1% Tween 20, 0.1 M Tris-HCl, and 0.05% levamisole/tetramisole in H_2O)

3x 20 minutes, followed by staining with BM Purple AP Substrate (Roche). The stained embryos were imaged using a Zeiss Discovery V.12 microscope and Leica DFC420 digital camera.

Mammalian cell culture

All mammalian cells were cultured at 37°C with 5% CO2, and were maintained in high glucose DMEM (Gibco cat. no. 11965) for HEK293T (from ATCC) and NIH/3T3 (a gift from T. Reh's lab at the University of Washington) cells, both supplemented with 10% FBS and 1X Pen/Strep (Gibco cat. no. 15140122; 100U/ml penicillin, 100 µg/ml streptomycin). Cells were trypsinized with 0.25% typsin-EDTA (Gibco cat. no. 25200-056) and split 1:10 three times a week.

Mouse embryo nuclei extraction and fixation

Mouse embryos from different development stages were processed together to reduce batch effects. Each mouse embryo was minced into small pieces by blade in 1 mL ice-cold cell lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2 and 0.1% IGEPAL CA-630 from(*53*), modified to also include 1% SUPERase In and 1% BSA) and transferred to the top of a 40 µm cell strainer (Falcon). Tissues were homogenized with the rubber tip of a syringe plunger (5 ml, BD) in 4 ml cell lysis buffer. The filtered nuclei were then transferred to a new 15 ml tube (Falcon) and pelleted by centrifuge at 500x**g** for 5 min and washed once with 1 ml cell lysis buffer. The nuclei were fixed in 4 ml ice cold 4% paraformaldehyde (EMS) for 15 min on ice. After fixation, the nuclei were washed twice in 1 ml nuclei wash buffer (cell lysis buffer without IGEPAL), and re-suspended in 500 µl nuclei wash buffer. The samples were split to two tubes with 250 µl in each tube and flash frozen in liquid nitrogen. We estimated the nuclei extraction efficiency based on the extracted nuclei number vs. expected total nuclei number in each embryo. The estimated nuclei extraction efficiency ranged from 60% to 85%.

As a quality control, HEK293T and NIH/3T3 cells were trypsinized, spun down at 300x**g** for 5 min (4°C) and washed once in 1X PBS. Equal numbers of HEK293T and NIH/3T3 cells were combined and lysed using 1 mL ice-cold cell lysis buffer followed by the same fixation and storage conditions as used for the mouse embryos.

Mouse embryo cell counts

3-5 embryos per developmental stage were microdissected in PBS at room temperature. Each mouse embryo was minced into small pieces by blade and a single cell suspension was obtained by incubating the tissue in 4 ml Trypsin-EDTA 0.05% (Gibco) at 37°C for 10 min vortexing every other minute. The cells of each embryo were diluted in 4 ml medium and transferred to the top of a 40µ m cell strainer (Falcon). Cell numbers was then determined by counting cells using a hemocytometer.

sci-RNA-seq3 library preparation and sequencing

Thawed nuclei were permeabilized with 0.2% TritonX-100 (in nuclei wash buffer) for 3 min on ice, and briefly sonicated (Diagenode, 12 sec on low power mode) to reduce nuclei clumping. The nuclei were then washed once with nuclei wash buffer and filtered through 1 ml Flowmi cell strainer (Flowmi). Filtered nuclei were spun down at 500x**g** for 5 min and resuspended in nuclei wash buffer.

Nuclei from each mouse embryo were then distributed into several individual wells in four 96-well plates. The links between well id and mouse embryo were recorded for downstream data processing. For each well, 80,000 nuclei (16 µL) were mixed with 8 µl of 25 µM anchored oligo-

dT primer (5′- /5Phos/CAGAGCNNNNNNNN[10bp barcode]TTTTTTTTTTTTTTTTTTTTTTTTTTTTT-3′, where “N” is any base; IDT) and 2 µL 10 mM dNTP mix (Thermo), denatured at 55°C for 5 min and immediately placed on ice. 14 µL of first-strand reaction mix, containing 8 µL 5X Superscript IV First-Strand Buffer (Invitrogen), 2 µl 100 mM DTT (Invitrogen), 2 µl SuperScript IV reverse transcriptase (200 U/µl, Invitrogen), 2 µL RNaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen), was then added to each well. Reverse transcription was carried out by incubating plates by gradient temperature (4°C 2 minutes, 10°C 2 minutes, 20°C 2 minutes, 30°C 2 minutes, 40°C 2 minutes, 50°C 2 minutes and 55°C 10 minutes).

After ligation reaction, 60 µL nuclei dilution buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2 and 1% BSA) was added into each well. Nuclei from all wells were pooled together and spun down at 500x**g** for 10 min. Nuclei were then resuspended in nuclei wash buffer and redistributed into another four 96-well plates with each well including 4 µL T4 ligation buffer (NEB), 2 µL T4 DNA ligase (NEB), 4 µL Betaine solution (5M, Sigma-Aldrich), 6 µL nuclei in nuclei wash buffer, 8µL barcoded ligation adaptor (100 uM, 5’- GCTCTG[9 bp or 10 bp barcode A]/dideoxyU/ACGACGCTCTTCCGATCT[reverse complement of barcode A]-3’) and 16 µL 40% PEG 8000 (Sigma-Aldrich). The ligation reaction was done at 16°C for 3 hours.

After RT reaction, 60 µL nuclei dilution buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2 and 1% BSA) was added into each well. Nuclei from all wells were pooled together and spun down at 600x**g** for 10min. Nuclei were washed once with nuclei wash buffer and filtered with 1 ml Flowmi cell strainer (Flowmi) twice, counted and redistributed into eight 96-well plates with each well including 2,500 nuclei in 5 µL nuclei wash buffer and 5 µL elution buffer (Qiagen). 1.33

µl mRNA Second Strand Synthesis buffer (NEB) and 0.66 µl mRNA Second Strand Synthesis enzyme (NEB) were then added to each well, and second strand synthesis was carried out at 16°C for 180 min.

For tagmentation, each well was mixed with 11 µL Nextera TD buffer (Illumina) and 1 µL i7 only TDE1 enyzme (62.5 nM, Illumina), and then incubated at 55°C for 5 min to carry out tagmentation. The reaction was then stopped by adding 24 µL DNA binding buffer (Zymo) per well and incubating at room temperature for 5 min. Each well was then purified using 1.5x AMPure XP beads (Beckman Coulter). In the elution step, each well was added with 8 µL nuclease free water, 1 µL 10X USER buffer (NEB), 1 µL USER enzyme (NEB) and incubated at 37°C for 15 min. Another 6.5 µL elution buffer was added into each well. The AMPure XP beads were removed by magnetic stand and the elution product was transferred into a new 96-well plate.

For PCR amplification, each well (16 µL product) was mixed with 2 µL of 10 µM indexed P5 primer (5′-AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′; IDT), 2 µL of 10 µM P7 primer (5′-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3′, IDT), and 20 µL NEBNext High-Fidelity 2X PCR Master Mix (NEB). Amplification was carried out using the following program: 72°C for 5 min, 98°C for 30 sec, 12-14 cycles of (98°C for 10 sec, 66°C for 30 sec, 72°C for 1 min) and a final 72°C for 5 min.

Of note, for a single experiment, we have 384 barcodes introduced at the reverse transcription step, 384 barcodes introduced by hairpin ligation and 768 barcodes introduced by PCR. This corresponds to 384 * 384 * 768 = ~113 million possible combinations.

After PCR, samples were pooled and purified using 0.8 volumes of AMPure XP beads. Library concentrations were determined by Qubit (Invitrogen) and the libraries were visualized by electrophoresis on a 6% TBE-PAGE gel. All libraries were sequenced on one NovaSeq platform (Illumina) (Read 1: 34 cycles, Read 2: 52 cycles, Index 1: 10 cycles, Index 2: 10 cycles).

Processing of sequencing reads

Base calls were converted to fastq format using Illumina's bcl2fastq/v2.16 and demultiplexed based on PCR i5 and i7 barcodes using maximum likelihood demultiplexing package deML(*54*) with default settings. Downstream sequence processing and single cell digital expression matrix generation were similar to sci-RNA-seq(*17*) except that RT index was combined with hairpin adaptor index, and thus the mapped reads were split into constituent cellular indices by demultiplexing reads using both the RT index and ligation index (ED < 2, including insertions and deletions). Briefly, demultiplexed reads were filtered based on RT index and ligation index (ED < 2, including insertions and deletions) and adaptor clipped using trim_galore/v0.4.1 with default settings. Trimmed reads were mapped to the mouse reference genome (mm10) for mouse embryo nuclei, or a chimeric reference genome of human hg19 and mouse mm10 for HEK293T and NIH/3T3 mixed nuclei, using STAR/v 2.5.2b(*55*) with default settings and gene annotations (GENCODE V19 for human; GENCODE VM11 for mouse). Uniquely mapping reads were extracted, and duplicates were removed using the unique molecular identifier (UMI) sequence, reverse transcription (RT) index, hairpin ligation adaptor index and read 2 end-coordinate (*i.e.*

reads with identical UMI, RT index, ligation adaptor index and tagmentation site were considered duplicates). Finally, mapped reads were split into constituent cellular indices by further demultiplexing reads using the RT index and ligation hairpin (ED < 2, including insertions and deletions). For mixed-species experiment, the percentage of uniquely mapping reads for genomes of each species was calculated. Cells with over 85% of UMIs assigned to one species were regarded as species-specific cells, with the remaining cells classified as mixed cells or "collisions". To generate digital expression matrices, we calculated the number of strand-specific UMIs for each cell mapping to the exonic and intronic regions of each gene with python/v2.7.13 HTseq package(*56*). For multi-mapped reads, reads were assigned to the closest gene, except in cases where another intersected gene fell within 100 bp to the end of the closest gene, in which case the read was discarded. For most analyses we included both expected-strand intronic and exonic UMIs in per-gene single-cell expression matrices.

Because of the marked increase in processing time that it would entail, we note that we did not perform a UMI error correction step. However, to confirm that our failure to do so would not inflate UMI counts, we compared results with vs. without UMI error correction (edit distance of 1) for a subset of wells. Compared with skipping UMI error correction, 99.4% of reads remain after UMI error correction, which indicated to us that the error correction step only has minor impact on the estimated UMI counts per cell (less than 1%). This is likely due to the high quality of sequencing data that we obtained on the NovaSeq, the low number of PCR amplification steps, and the low duplication rate. We emphasize that groups implementing sci-RNA-seq3 should either perform UMI error correction or a similar data quality check.

Whole mouse embryo analysis

After the single cell gene count matrix was generated, each cell was assigned to its original mouse embryo based on the RT barcode. Reads mapping to each embryo were aggregated to generate "bulk RNA-seq" for each embryo. For sex separation of embryos, we counted reads mapping to a female-specific non-coding RNA (*Xist*) or chrY genes (except *Erdr1* which is in both chrX and chrY). Embryos were readily separated into females (more reads mapping to *Xist* than chrY genes) and males (more reads mapping to chrY genes than *Xist*).

Pseudotemporal ordering of whole mouse embryos was done by Monocle 2(*57*). Briefly, an aggregated gene expression matrix was constructed as described above. Differentially expressed genes across different development conditions were identified with differentialGeneTest function of Monocle 2(*57*). The top 2,000 genes with the lowest q value were used to construct the pseudotime trajectory using Monocle 2(*57*). Each embryo was assigned a pseudotime value based on its position along the trajectory.

Cell clustering, t-SNE visualization and marker gene identification

A digital gene expression matrix was constructed from the raw sequencing data as described above. Cells with fewer than 200 UMIs or over 3,172 UMIs (two standard deviation above the mean UMI count) were discarded. Downstream analysis were performed with Monocle2/v2.6.0(*57*) and python package scanpy/v1.0(*58*). Briefly, gene count mapping to sex chromosomes were removed before clustering and dimensionality reduction. Preprocessing steps were similar to the approach used by ref (*59*). Briefly, genes with no count were filtered out and each cell was normalized by the total UMI count per cell. The top 2,000 genes with the highest variance were selected and the digital gene expression matrix was renormalized after gene filtering. The data was log transformed after adding a pseudocount, and scaled to unit variance and zero mean. The dimensionality of the

data was reduced by PCA (30 components) first and then with t-SNE, followed by Louvain clustering performed on the 30 principal components (resolution=1.5). For Louvain clustering, we first fitted the top 30 PCs to compute a neighborhood graph of observations with local neighborhood number of 15 by scanpy.api.pp.neighbors function in scanpy/v1.0(*60*). We then cluster the cells into sub-groups using the Louvain algorithm implemented as scanpy.api.tl.louvain function(*60*). For tSNE visualization, we directly fit the PCA matrix into scanpy.api.tl.tsne function(*60*) with perplexity of 30. 40 clusters were identified. We then sampled 1,000 cells from each cluster and differentially expressed genes across different clusters were identified with differentialGeneTest function of Monocle 2/v2.6.0(*57*). Genes specific to each cluster were identified similar as previously described(*61*). Clusters were assigned to known cell types based on cluster-specific markers. One cluster had abnormally high UMI counts but no strongly cluster-specific genes, suggesting that it may be a technical artifact of cell doublets and therefore was removed. This was confirmed upon analysis for doublets with Scrublet (see next paragraph). Another two clusters both appeared to correspond to the definitive erythroid lineage and were merged. Consensus expression profiles for each cell type were constructed as previously described(*61*). Differentially expressed genes across cell types were identified with the differentialGeneTest() function of Monocle 2/v2.6.0(*62*). To identify cell type-specific gene markers, we selected genes that were differentially expressed across different cell types (FDR of 5%, likelihood ratio test) and also with a >2-fold expression difference between first and second ranked cell types.

For the detection of potential doublet cells, we first split the dataset of ~2 million cells into four equally sized subsets, and then applied the scrublet/v0.1 pipeline(*63*) to each subset with parameters (min_count = 3, min_cells = 3, vscore_percentile = 85, n_pc = 30,

expected_doublet_rate = 0.06, sim_doublet_ratio = 2, n_neighbors = 30, scaling_method = 'log') for doublet score calculation. Cells with doublet score over 0.25 are annotated as detected doublets. We detected 4.3% potential doublet cells in the whole data set, which corresponds to an overall estimated doublet rate of 10.3% (including both within- and between-cluster doublets). The aforementioned major cluster with abnormally high UMI counts but no strongly cluster-specific genes (see last paragraph) had a high detected doublet proportion (52%), confirming it as a doublet-related artifact. For detection of doublet derived subclusters, we redid the above analysis on the whole dataset after removing the doublet-derived main cluster. Subclusters with detected doublet proportion of >10% were annotated as doublet derived subclusters.

For subcluster identification, we selected high quality cells (UMI > 400) in each major cell type and applied PCA, t-SNE, Louvain clustering similarly to the major cluster analysis. Subclusters were filtered out if most cells (>50%) of the cluster derived from a single embryo. Highly similar subclusters were merged if their aggregated transcriptomes were highly correlated (Pearson correlation coefficient > 0.95) and the two clusters were close with each other in t-SNE space. Genes differentially expressed across subclusters were identified for each major cell type as described above. Subclusters with a detected doublet ratio (by Scrublet) over 10% are annotated as doublet-derived subclusters.

To identify a distinguishing set of gene markers for each of the 572 subclusters (those of the 655 with detected doublet cell ratio <= 10%), we used the following algorithm: 1) We selected genes detected in at least 5% of cells in the target subcluster; 2) From these, we identified genes with a >4-fold greater expression in the target cluster than all 571 other subclusters; 3) If there was no such gene, the algorithm tried to identify a gene (“marker A”) such that subclusters with low

expression of marker A (less than 25% of its expression in the target cluster) are readily distinguished from the target cluster based on this difference, and are therefore removed from the comparison set. Gene marker A is selected to maximize the number of subclusters removed from the comparison set. 4) To identify markers that separate the target cluster from the remaining subclusters, we repeat steps 2-3 until a marker with a >4-fold expression difference between the target cluster and all remaining subclusters is identified. The set of markers identified through this heuristic is sufficient to distinguish the target subcluster from all 571 other non-doublet subclusters on the basis of >4-fold expression differences.

For identifying correlated cell types between two cell atlas datasets, we first aggregate the cell type specific UMI counts, normalized by the total count, multiplied by 100,000, and log transformed after adding a pseudo-count. We then applied non-negative least squares (NNLS) regression to predict the gene expression of target cell type (Ta) in dataset A with the gene expression of all cell types (Mb) in dataset B:

$$Ta=\beta 0a+\beta 1aMb$$

where Taand Mb represent filtered gene expression for target cell type from data set A and all cell types from data set B, respectively. To improve accuracy and specificity, we selected cell type-specific genes for each target cell type by: 1) ranking genes based on the expression fold-change between the target cell type vs. the median expression across all cell types, and then selecting the top 200 genes. 2) ranking genes based on the expression fold-change between the target cell type vs. the cell type with maximum expression among all other cell types, and then selecting the top 200 genes. 3) Merge the gene lists from step (1) and (2). β1ais the correlation coefficient computed by NNLS regression.

Similarly, we then switch the order of datasets A and B, and predict the gene expression of target cell type (Tb) in dataset B with the gene expression of all cell types (Ma) in dataset A:

$$Tb=\beta 0b+\beta 1bMa$$

Thus, each cell type a in dataset A and each cell type b in dataset B are linked by two correlation coefficients from the above analysis: βab for predicting cell type a using b, and βbafor predicting cell type b using a. We combine the two values by:

$$\beta = 2 * (\beta ab + 0.01) * (\beta ba + 0.01)$$

and find βreflects the matching of cell types between two data sets with high specificity (**Extended Data Fig. 5a**). For each cell type in dataset A, all cell types in dataset B are ranked by βand the top cell type (with $\beta > 0.01$) is identified as the matched cell type. For validation, we first applied cell type correlation analysis to independently generated and annotated analyses of the adult mouse kidney (sci-RNA-seq component of sci-CAR(*19*) vs. Microwell-seq(*10*)). We subsequently compared cell subclusters from this study (with detected doublet cell ratio <= 10%) to fetus-related cell types (those with annotations including the term “fetus”) from the Microwell-seq-based Mouse Cell Atlas (MCA)(*10*). A similar comparison was performed against cell types annotated in a recent mouse brain atlas (BCA)(*32*).

For estimation of the number of cells of each cell type (or cell subtype), we first calculated the proportion of each cell type in individual embryos, and then multiplied the proportion by the estimated total cell number for each embryo (E9.5: 200,000, E10.5: 1,100,000; E11.5: 2,600,000; E12.5: 6,100,000; E13.5: 13,000,000).

Assuming that ~100 cells are required to detect a cell type and that the cell type in question is only present at one timepoint, we note that the power of this study would be limited to detecting cell types whose 'population size' per embryo is >125 cells at E9.5, >333 cells at E10.5, >500 cells at E11.5, >1,250 cells at E12.5, or >3,400 cells at E13.5. However, our power may be greater than that for cell types that are present across timepoints. For example, the primordial germ cell subcluster 16.13, which includes just 88 of 2,058,652 cells in the dataset, is contributed to by cells from all five timepoints.

AER and limb mesenchyme pseudo-time analysis

Pseudotemporal ordering of AER cells, forelimb or hindlimb was done with Monocle 2(*57*). Briefly, differentially expressed genes across five development stages were identified with the differentialGeneTest function of Monocle 2(*57*). The top 500 genes with the lowest q value were used to construct the pseudotime trajectory using Monocle 2(*57*), with UMI count per cell as a covariate in the tree construction. Each cell was assigned a pseudotime value based on its position along the trajectory. Smoothed gene marker expression change along pseudotime were generated by plot_genes_in_pseudotim function in Monocle 2(*57*). Cells in the trajectory were grouped in the same method as a previous study(*64*). Briefly, cells were grouped first at similar positions in pseudotime by k-means clustering along the pseudotime axis (k = 10). These clusters were subdivided into groups containing at least 50 and no more than 100 cells. We then aggregated the transcriptome profiles of cells within each group. The gene expression along pseudotime was calculated in the same approach as a previous study(*64*). Briefly, genes passing significant test (FDR of 5%) across different treatment conditions were selected and a natural spline was used to fit the gene expression along pseudotime, with mean_number_genes included as a covariate. The

gene expression for each gene was subtracted by the lowest expression and then divided by the highest expression. Genes with max expression within the early 20% of pseudotime were labeled as repressed genes. Genes with max expression in the last 20% of pseudotime were labeled as activated genes. Other genes were labeled as transient genes. Enriched reactome terms (Reactome_2016) and transcription factors (ChEA_2016) were identified using EnrichR/v1.0 package(*65*).

Trajectory inference with Monocle 3

The Monocle 3 workflow consists of 3 core steps to organize cells into potentially discontinuous trajectories, followed by optional statistical tests to find genes that vary in expression over those trajectories. Monocle 3 also includes visualization tools to help explore trajectories in three dimensions.

Dimensionality reduction with Uniform Manifold Approximation and Projection (UMAP)

Monocle 3 first projects the data into a low-dimensional space, which facilitates learning a principal graph that describes how cells transit between transcriptomic states. Monocle 3 does so with UMAP/v0.3.2, a recently proposed algorithm based on Riemannian geometry and algebraic topology to perform dimension reduction and data visualization(*66*). Its visualization quality is competitive with the popular t-SNE (t-stochastic neighbor embedding) method used widely in single-cell transcriptomics. However, where t-SNE mainly aims to place highly similar cells in the same regions of a low-dimensional space, UMAP also preserves longer-range distance relationships. The UMAP algorithm itself is also more efficient (the algorithm complexity of UMAP is $O(N)$ vs. $O(Nlog(N))$ for t-SNE). Briefly, UMAP first constructs a topological representation of the high dimensional data with local manifold approximations and patches

together their local fuzzy simplicial set representations. UMAP then optimizes the lower dimension embedding, minimizing the cross-entropy between the low dimensional representation and the high dimensional one.

The computational efficiency of UMAP dramatically accelerated the analysis of the mouse embryo data. We found that UMAP finished processing two million cells dataset in around 3 CPU hours while t-SNE took more than 64 CPU hours. A few implementation details lead to the effectiveness of UMAP. Two major steps are involved in both the UMAP and t-SNE algorithms: first, the preprocessing step before UMAP is similar to Monocle 2. Briefly, genes expressed in fewer than 10 cells (or fewer than 5 cells in datasets with fewer than 1,000 cells) were filtered out. The digital gene count matrix was first normalized by cell specific size factor estimated by "estimateSizeFactors" function in Monocle 3, log transformed after adding a pseudocount, and then scaled to unit variance and zero mean. The top 5,000 most highly dispersed genes (2,000 genes for datasets with fewer than 5,000 cells, 300 genes for datasets with fewer than 1,000 cells) were selected. The matrix was then projected into 50 top PCs (30 top PCs for trajectory analysis of the ten supergroups, 10 top PCs for data sets with fewer than 5,000 cells, 5 top PCs for data sets with fewer than 1,000 cells) by partial SVD. Thus an intermediate structure from the high dimension space (here, we used the top 50 principal components constructed from the 5,000 most highly dispersed genes) is built and then a low dimensional embedding is found to represent the intermediate structure. For the second step, both methods used stochastic grid descent approach with differing loss functions to embed the data into low dimension space. While t-SNE needs a loss function for global normalization, UMAP uses a different objective function that avoids that need. This step essentially enables UMAP to scale linearly with the number of data samples.

Dimensionality reduction was implemented with the reduceDimension() function in Monocle 3. This function calls the UMAP/v0.3.2 python implementation (https://github.com/lmcinnes/umap) from Leland McInnes and John Healy through the *reticulate/v1.10* package (https://cran.r-project.org/web/packages/reticulate/index.html). To process all the cells together, we set UMAP parameters as follows: (n_neighbors = 50, min_dist = 0.01, cosine distance metric). To more finely resolve subtrajectories, we adjusted these as such: (n_neighbors = 15, min_dist = 0.1, cosine distance metric).

Partitioning cells into discontinuous trajectories

Recently, Wolf and colleagues proposed the idea to organize single-cell transcriptome data into a "partitioned approximate graph abstraction" (PAGA) that relates clusters of cells that might be developmentally related to one another. Briefly, their algorithm constructs a k-nearest neighbor graph on cells and then identifies "communities" of cell via the Louvain method, similar to previous methods for analyzing CyTOF or single-cell RNA-seq data(*67*). PAGA then constructs a graph in which the vertices are Louvain communities. Two vertices are linked with an edge in the PAGA graph when the cells in the respective communities are neighbors in the kNN graph more frequently than would be expected under a simple binomial model(*68*). Similar methods were also recently developed and applied in analyzing zebrafish and xenopus cell atlas datasets(*5*, *6*).

Monocle 3 draws from these ideas, first constructing a kNN graph (k=20) on cells in the UMAP space, then grouping them into Louvain communities, and testing each pair of communities for a significant number of links between their respective cells. Those communities that have more links than expected under the null hypothesis of spurious linkage (FDR < 1%) remain connected in the PAGA graph, and those links that fail this test are severed. The resulting PAGA graph will have

one or more components, each of which is passed to the next step (learning the principal graph) as a separate group of cells that will be organized in a trajectory. The PAGA algorithm essentially stops at this stage, presenting the PAGA graph as a kind of coarse-grained trajectory in each community reflects a different state cells can adopt as they develop. In contrast, as described in the next section, Monocle 3 uses the PAGA graph to constrain the space of principal graphs that can form the final trajectory. That is, Monocle 3 uses the coarse-grained PAGA graph to learn a fine-grained trajectory.

Monocle 3's implementation of the above procedures (in the partitionCells() function) scales to millions of cells. Briefly, it uses the *clustering_louvain* function from the igraph package to perform community detection. Next, the core PAGA calculations from Wolf *et al.* are computed via a series of sparse matrix operations. Let X be a (sparse) matrix representing the community membership of the cells. Each column of X represents a Louvain community and each row of X corresponds to a particular cell. $X_{ij} = 1$ if cell i belongs to Louvain community j, otherwise 0. We can further obtain the adjacency matrix A of the kNN graph used to perform the louvain clustering where $A_{ij} = 1$ if cell i connects to j in the kNN graph. Then the connection matrix M between each cluster is calculated as,

$$M = X^{\intercal} A X$$

Once M is constructed, we can then follow *Supplemental Note 3.1* from ref. (*68*) to calculate the significance of the connection between each louvain clustering and consider any clusters with p-value larger than 0.05 by default as not disconnected.

Learning the principal graph

Monocle 3 learns a principal graph (via the learnGraph() function) that resides in the same low-dimensional space as the data to represent the possible paths cells can take as they develop. Monocle uses a principal graph embedding procedure that is based on the SimplePPT algorithm(*69*, *70*), with several key enhancements that accelerate graph embedding, support large datasets, allow for loops, and smooth the graph to eliminate noisy branches.

The first enhancement is that Monocle 3 learns the principal graph in the (by default, 3 dimensional) UMAP space using a fast reduced-representation approach to avoid dealing directly with many thousands of cells. It first selects a set of "landmark" cells using by first running the kmeans() clustering algorithm in R with k equal to the value of the "ncenter" argument, which can be passed to learnGraph() by the user. The landmark cells are then selected by first mapping each cell to its nearest kmeans point, and then selecting the cell for each kmeans point with the highest local density. By default, Monocle 3 uses a data-dependent policy for adjusting ncenter automatically(*62*). Here, unless otherwise specified, we override the default policy and use ncenter = 2000 in the analyses of the embryo data. Monocle 3 will then learn a principal graph within these landmarks cells rather than the full dataset to accelerate the optimization. Running time and fine detail in the trajectory will depend on the number principal graph nodes; more nodes generally results in a more accurate tree but at increased running time.

The second enhancement is a procedure to smooth and refine the principal graph to exclude small branches. In order to capture smaller fine details of a trajectory such as complex branching architecture, SimplePPT requires that the principal graph contain hundreds or even thousands of principal graph nodes. Consequently, the principal tree reported by SimplePPT often contains very small branches to which a very small percentage of cells project. Although SimplePPT does

provide tuning parameters that control graph smoothness to a certain extent, we have found that a simple heuristic pruning procedure is effective and easier for users to understand how to control. The procedure operates via a depth-first visitation of the graph nodes in the principal tree. At nodes with degree ≤ 2, no action is taken. For nodes with degree > 2, the diameter path for each subtree rooted at a neighbor not yet visited in the search is computed. If the path is less than a user-specified length (by default, 10 principal tree nodes), the whole subtree is pruned.

The third major enhancement is that Monocle 3 can learn principal graphs with loops instead of requiring that the trajectory be a tree. This is achieved by augmenting the principal tree reported by SimplePPT with additional edges meant to close loops in the trajectory. The algorithm considers adding an edge between two leaf nodes a and b in the principal graph if the pair meet several criteria. The first criteria is that the geodesic distance between a and b along the principal tree should be at least a certain minimum distance (by default, ⅓ of the tree's diameter path). That is, when the nodes are close in (euclidean) UMAP space, but distant in the graph, they ought to be linked. The second criteria is that they shouldn't be linked if doing so would create an especially long edge. By default, a and b cannot not be farther apart in UMAP space than the longest edge in the principal tree. The third criteria is based on the same test of connectivity used when partitioning the cells: consider leaf nodes a and b, which serve as proxies between two clusters of cells (those for which a and b are their nearest k-medioid). If cells near a have an unexpectedly high number of cells near b amongst their k nearest neighbors ($p < 0.05$ by default), then learnGraph() will link a and b in the principal graph, provided the other two criteria discussed above are also met.

For analysis of the ten major trajectories, we used ncenter = 5,000 for neural tube/notochord trajectory, and ncenter = 2,000 for epithelial and mesenchymal trajectories. For the other

supergroups, we used ncenter = (number of cells) / 25 and minimal_branch_len = 20. For analysis of the 56 subtrajectories, we mostly used ncenter = (number of cells in the trajectory) / 30 [2,000 maximum], and minimal_branch_len = 20. Each subtrajectory was manually checked and the parameters (ncenter and minimal_branch_len) for about a quarter of these were adjusted, mostly to further prune branches such that the principal graph follows cell transition path from early to late development stages.

The principal graph offers users a means of selecting subsets of cells that lead to particular lineages for further analysis. For example, to isolate cells leading to the myocyte fate, we first quantified the fraction of cells at each principal graph node that were classified as myocytes (cluster 13). From all 'majority myocyte' nodes, we then used the principal graph's edges to expand this set of nodes into wider 'neighborhood' of cells.

Computing pseudotimes

In order to calculate cell-wise pseudotime, we developed a projection strategy which is applicable to datasets with millions of cells. This strategy works by constructing a graph ψ on all cells using the principal graph as a guide, and then computing each cell's pseudotime as its geodesic distance back to one or more user-selected "root" nodes in the trajectory. In more detail, we first map each cell to its nearest principal point based on euclidean distance in the UMAP space. Then, for each principal graph edge, retrieve all the cells that map to its endpoints a and b. Next, orthogonally project each cells to the nearest point on the principal graph edge as previously described(*71*), so that each cell c_i can be ordered along the edge according to its projection $p(c_i)$. Without loss of generality, suppose this order is $a < p(c_i) < p(c_j) < b$. We then add edges (a, c_i) and

(c_j, b) to ψ. If c_i and c_j are in the same louvain component or connected louvain components (as determined during partitionCells), we also add (c_i, c_j) to ψ. Given ψ and a set of user-specified principal graph nodes, we can then assign pseudotime values to all cells. Monocle provides several ways to specify these nodes, either by name (i.e. programmatically) or interactively. Each cell's pseudotime is taken as the geodesic distance along ψ to the closest of these root nodes.

For root node selection of the mesenchymal and neural tube/notochord trajectories, we first assigned each principal point to a subcluster with the maximum cell proportion. We then selected the subcluster with the earliest average developmental stage, and use the earliest principal point assigned to this sub-cluster as the root state for pseudotime computation. For the other major trajectories, we assigned root nodes to the earliest principal point in each subtrajectory (except in neural crest trajectory 2, where we assigned the root node to the earliest principal points in PNS glia precursor cell trajectory and *Pdgfra*-positive glia trajectory). Some cells from complex trajectories (mesenchymal trajectory, neural tube/notochord trajectory, epithelial trajectory and endothelial trajectory) show outlier pseudotime values (more than 3 standard deviation higher than the mean values). These extreme values are clipped to the max value after excluding the outliers. For root node selection of the 56 cell type specific trajectories, we first computed the average development stage for each principal point. As the root state features the earliest development stage, we compared the average development stage of each node and its k-nearest neighbors (k = 10). We then manually checked each trajectory and selected root nodes from principal points with earlier development stage than all its nearby neighbours.

Identifying genes with complex trajectory-dependent expression

In order to identify genes that vary in expression over a developmental trajectory, we borrow a statistical test commonly used in analyzing spatial data. Moran's *I* statistic is a measure of multi-directional and multi-dimensional spatial autocorrelation. The statistic encodes spatial relationships between data-points via a nearest neighbor graph, making it particularly well suited for analyzing large single-cell RNA-seq datasets.

Moran's *I* test(*72*) is defined as:

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

where N is the number of cells indexed by i and j; x is the expression value of gene of interest; $\bar{x}_i$ ($\bar{x}_j$) is the mean of the gene expression for cell i's (or j's) nearest neighbors; w_{ij} is a matrix of weights defined by a nearest neighbor graph with zero on the diagonal (i.e., $w_{ii} = 0$) and $w_{ij} = 1/k_i$ where k_i is the number of nearest neighbors; and W is the sum of all w_{ij}.

To identify the nearest neighbors used for creating the weight matrix W, we first build a k (default to be 25) nearest neighbor graph (kNN) for all cells in the UMAP space. We also project each cell to its nearest node in the principal graph. Then we remove all edges from the kNN graph that connect cells that project onto principal graph nodes do not share an edge.

In Monocle 3, we implemented the principalGraphTest() function to identify correlated genes on the complex trajectory embedded in the manifold which relies on modified versions of routines from *spdep* package for performing the Moran's *I* test.

Reporting summary

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

Code availability

Scripts for processing sci-RNA-seq3 sequencing were written in python and R with code available at https://github.com/JunyueC/sci-RNA-seq3_pipeline. Trajectory analysis was done with Monocle 3 with setup instructions and tutorial available at http://cole-trapnell-lab.github.io/monocle-release/monocle3/.

Data availability

sci-RNA-seq3 protocol and all data are made freely available, including through a cell type wiki to facilitate their ongoing annotation by the research community (http://atlas.gs.washington.edu/mouse-rna/). The data generated by this study can be downloaded in raw and processed forms from the NCBI Gene Expression Omnibus (GSE119945).

Supplementary Note 1

We first sought to apply Monocle 3 to a single major cell type, cluster 25, whose 26,559 cells we annotate as limb bud mesenchyme on the basis of *Hoxd13, Fgf10* and *Lmx1b* expression. Visualizing the trajectory of cells of this cluster illustrates the dramatic expansion of limb mesenchymal cells over developmental time, with the main outgrowth between E10.5 and E12.5 (**Extended Data Fig. 7a**). Gene expression is highly dynamic during this expansion, with the levels of 4,763 protein-coding genes changing (FDR of 1%). The early stages of limb mesenchyme development are characterized by expression of some expected genes such as *Tbx15(73)*, and

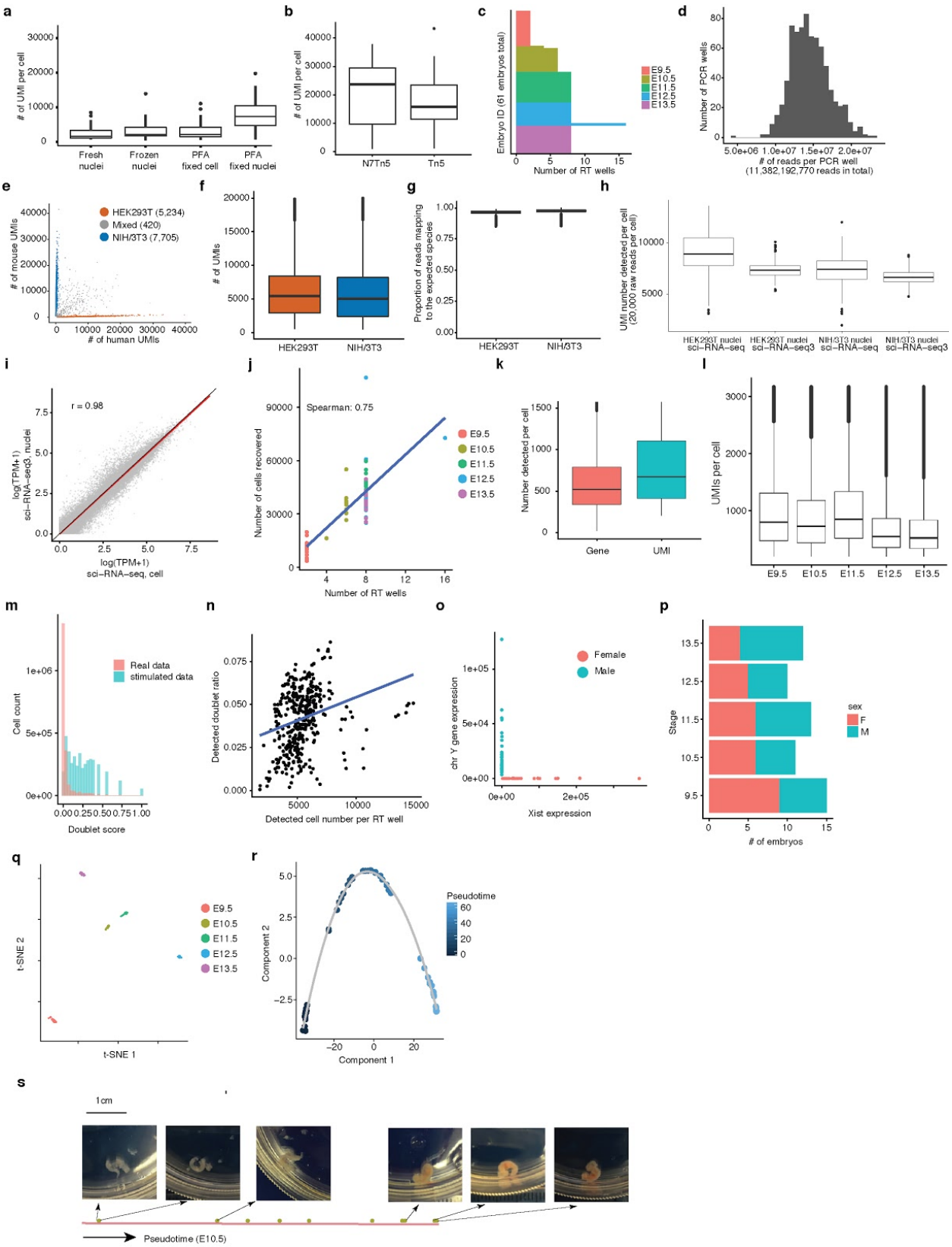
Gpc3(74) and the later stages by *Msx1(75)*, *Epha4(76)* and *Dach1(77)* (**Extended Data Fig. 7b**), but the vast majority of dynamically expressed genes are novel. Transcription factors significantly upregulated during limb mesenchyme development included those with roles in chondrocyte differentiation (*e.g. Sox9(78)* and *Yap1(79)*), muscle differentiation (*e.g. Tead4(80)*), and wound healing and limb regeneration (*e.g. Smarcd1(81)*) (**Extended Data Fig. 7c**).

Interestingly, forelimb and hindlimb cells were not obviously separated by unsupervised clustering (**Extended Data Fig. 7d**) or trajectory analysis (**Extended Data Fig. 7e**), but could be distinguished by the mutually exclusive expression of *Tbx5* in forelimb (2,085 cells, 7.9% of all limb mesenchyme cells) and *Pitx1* in hindlimb (1,885 cells, 7.1% of all limb mesenchyme cells) with only 22 cells expressing both markers (0.08% of all limb mesenchyme cells vs. ~0.6% expected if they were independent; **Extended Data Fig. 7f**)(*82*). 285 genes were differentially expressed between cells assigned to the forelimb and hindlimb in this way (**Extended Data Fig. 7g**). Known marker genes such as *Tbx4* and the genes of the Hoxc cluster (*Hoxc4-10*)(*83*) were upregulated in hindlimb cells as expected, but we also identified genes not previously shown to be differentially expressed. For example, we observed *Epha3* and *Hs3st3b1* to be 5-fold enriched in forelimb, and *Pcdh17* and *Igf1* to be 3-fold enriched in hindlimb.

Although developmental time is a major axis of variation in the limb mesenchyme trajectory (**Extended Data Fig. 7a**), there is clearly additional structure. At least some of this appears to correspond to the two main spatial axes of limb development: the proximal-distal axis (the primary direction of outgrowth) and the anterior-posterior axis (corresponding to the five digits)(*82*). With Monocle 3, we applied Moran's I test(*72*) to detect genes exhibiting autocorrelation across the limb mesenchyme trajectory (*i.e.* genes expressed in similar regions of the principal graph). We found, for example, that cells expressing *Sox6* and *Sox9* (proximal markers)(*84*, *85*), *Hoxd13* and *Tfap2b* (distal markers)(*37*), *Pax9* and *Alx4* (anterior markers), and *Shh* and *Hand2* (posterior

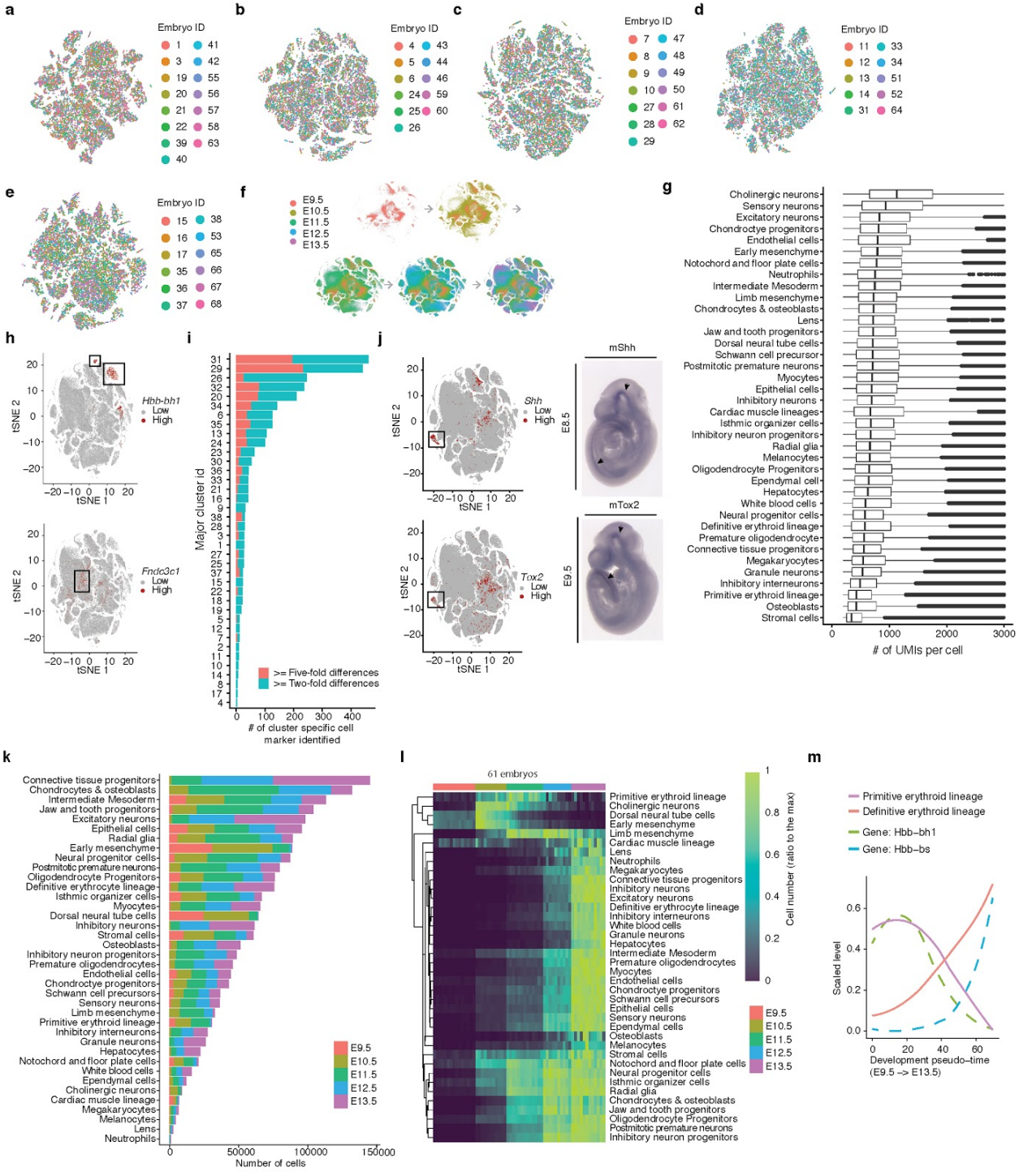
markers), were differentially distributed across the trajectory (**Extended Data Fig. 7h**, **Extended Data Fig. 7i**). Whole-mount *in situ* hybridization of *Hoxd13* (a known distal marker) and *Cpa2* (a novel marker whose distribution in the Monocle 3 trajectory was similar to that of known distal markers), confirmed that both genes are expressed in the distal limb mesenchyme between E10.5 and E13.5 (**Extended Data Fig. 7j-l**). Altogether, we identified 1,783 genes exhibiting variable expression across the limb mesenchymal trajectory (FDR of 1%; Moran's I > 0.01). These genes clustered into eight patterns of expression, several of which matched the distributions of known markers for the proximal-distal and anterior-posterior axes (**Extended Data Fig. 7m**). These analyses illustrate how this single cell atlas of mouse organogenesis can be used to characterize the spatiotemporal dynamics of gene expression in specific systems.

Extended Data Figures


a
# of UMI per cell
Fresh nuclei
Frozen nuclei
PFA fixed cell
PFA fixed nuclei
b
# of UMI per cell
N7Tn5
Tn5
c
Embryo ID (61 embryos total)
Number of RT wells
E9.5
E10.5
E11.5
E12.5
E13.5
d
Number of PCR wells
# of reads per PCR well
(11,382,192,770 reads in total)
e
HEK293T (5,234)
Mixed (420)
NIH/3T3 (7,705)
# of mouse UMIs
# of human UMIs
f
# of UMIs
HEK293T
NIH/3T3
g
Proportion of reads mapping to the expected species
HEK293T
NIH/3T3
h
UMI number detected per cell (20,000 raw reads per cell)
HEK293T nuclei sci-RNA-seq
HEK293T nuclei sci-RNA-seq3
NIH/3T3 nuclei sci-RNA-seq
NIH/3T3 nuclei sci-RNA-seq3
i
r = 0.98
log(TPM+1) sci-RNA-seq3, nuclei
log(TPM+1) sci-RNA-seq, cell
j
Spearman: 0.75
Number of cells recovered
Number of RT wells
E9.5
E10.5
E11.5
E12.5
E13.5
k
Number detected per cell
Gene
UMI
l
UMIs per cell
E9.5
E10.5
E11.5
E12.5
E13.5
m
Real data
stimulated data
Cell count
Doublet score
n
Detected doublet ratio
Detected cell number per RT well
o
Female
Male
chr Y gene expression
Xist expression
p
Stage
sex
F
M
# of embryos
q
E9.5
E10.5
E11.5
E12.5
E13.5
t-SNE 2
t-SNE 1
r
Pseudotime
Component 2
Component 1
s
1cm
Pseudotime (E10.5)

Extended Data Fig. 1. Performance and QC-related analyses for sci-RNA-seq3. (**a**) Comparison of fixation conditions in human HEK293T cells. Paraformaldehyde (PFA) fixed nuclei yielded the highest numbers of UMIs. Cell number n = 21 for fresh nuclei, 17 for frozen nuclei, 32 for PFA fixed cell, 31 for PFA fixed nuclei. (**b**) Tn5 transposomes loaded only with N7 adaptor (cell number n = 13) increased UMI counts by over 50%, relative to the standard Nextera Tn5 (cell number n = 11), in human HEK293T cells. (**c**) Bar plot showing the number of RT wells used for each of 61 mouse embryos. (**d**) Histogram showing the distribution of raw sequencing reads from each PCR well in sci-RNA-seq3. (**e**) Scatter plot of mouse (NIH/3T3) vs. human (HEK293T) UMI counts per cell. (**f-g**) Box plot showing the number of UMIs and purity (proportion of reads mapping to the expected species) per cell from HEK293T (cell number n = 7,943) and NIH/3T3 cells (cell number n = 10,914). At a sequencing depth of 23,207 reads per cell, we observed a median of 5,461 UMIs per HEK293T cell and 5,087 UMIs per NIH/3T3 cell, with 3.9% and 2.9% of reads per cell mapping to incorrect species, respectively. (**h**) Box plot comparing the number of UMIs per cell (downsampled to 20,000 raw reads per cell) for sci-RNA-seq3 (cell number n = 689 for HEK293T and 997 for NIH/3T3) vs. sci-RNA-seq (cell number n = 47 for HEK293T and 120 for NIH/3T3). (**i**) Correlation (Spearman's correlation) between gene expression measurements in aggregated profiles of HEK293T from sci-RNA-seq3 nuclei vs. sci-RNA-seq cells. (**j**) Scatter plot showing correlation between number of RT wells used and number of cells recovered per embryo. (**k**) Box plot showing the number of genes and UMIs detected per cell. (**l**) Box plot showing the number of UMIs detected per cell from embryos across five developmental stages. Cell number n = 152,120 for E9.5; 378,427 for E10.5; 615,908 for E11.5; 475,047 for E12.5; 437,150 for E13.5. (**m**) Histogram showing the distribution of the cell doublet score for the actual mouse embryo data vs. doublets stimulated by Scrublet. (**n**) Scatter plot of the number of cells profiled per RT well and the detected doublet cell ratio. Blue line showing the

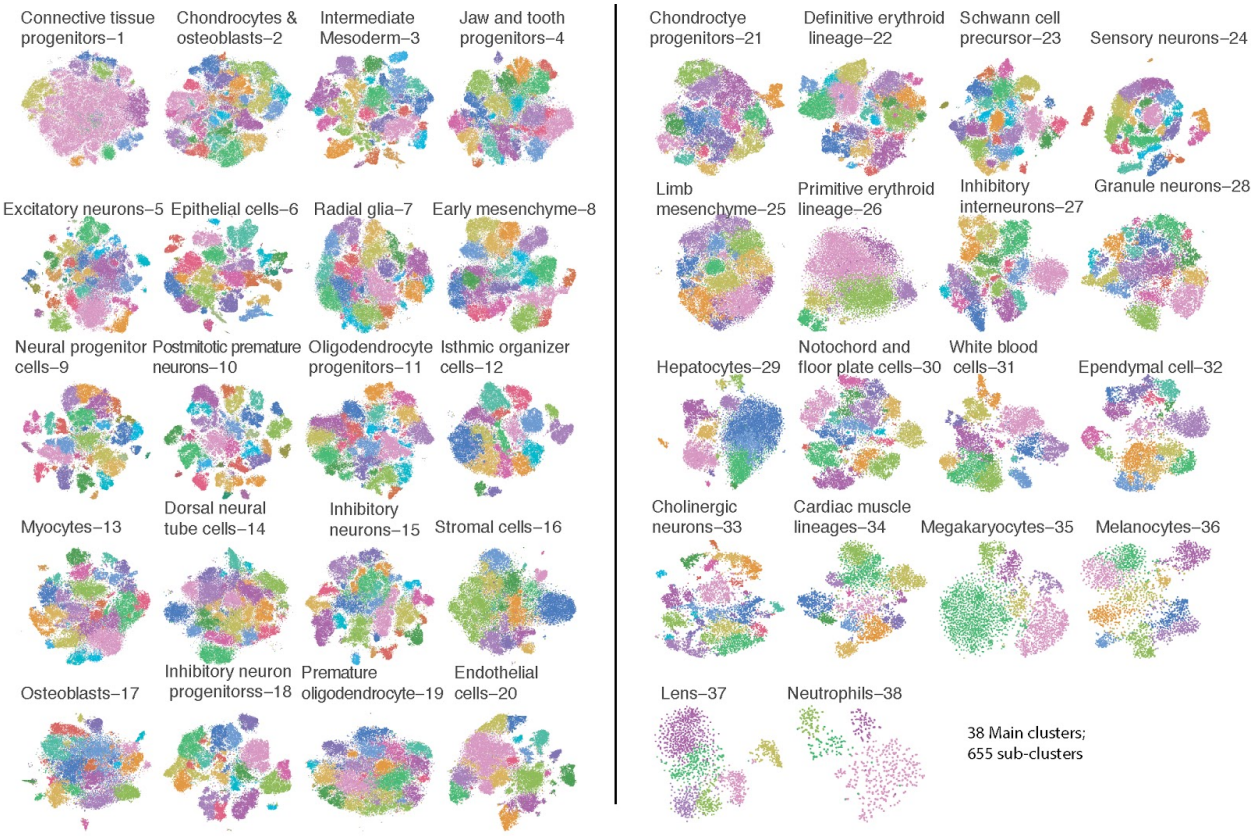
linear regression line. The detected doublet cell rate was modestly correlated with number of cells profiled per well during reverse transcription (Spearman's rho: 0.35). (**o**) Scatter plot of unique reads aligning to *Xist* (female-specific) vs. chrY transcripts (male-specific) per mouse embryo. Sex assignments of individual embryos inferred from these data. (**p**) Bar plot showing the number of male and female embryos profiled at each developmental stage. (**q**) t-SNE of the aggregated transcriptomes of single cells derived from each of 61 mouse embryos results in five tightly clustered groups perfectly matching their developmental stages (embryo number n = 61). (**r**) Pseudotime trajectory of pseudobulk RNA-seq profiles of mouse embryos (embryo number n = 61); identical to Fig. 1f, but colored by pseudotime. (**s**) The 61 profiled embryos were ordered by pseudotime. The three earliest vs. three latest (in pseudotime) E10.5 embryos are shown in photos, and appear to potentially be morphologically distinct. Notably, the distinct coloring of E10.5 embryos positioned earlier vs. later in developmental pseudotime is potentially due to different levels of hemoglobin. For all box plots: thick horizontal lines, medians; upper and lower box edges, first and third quartiles, respectively; whiskers, 1.5 times the interquartile range; circles, outliers.



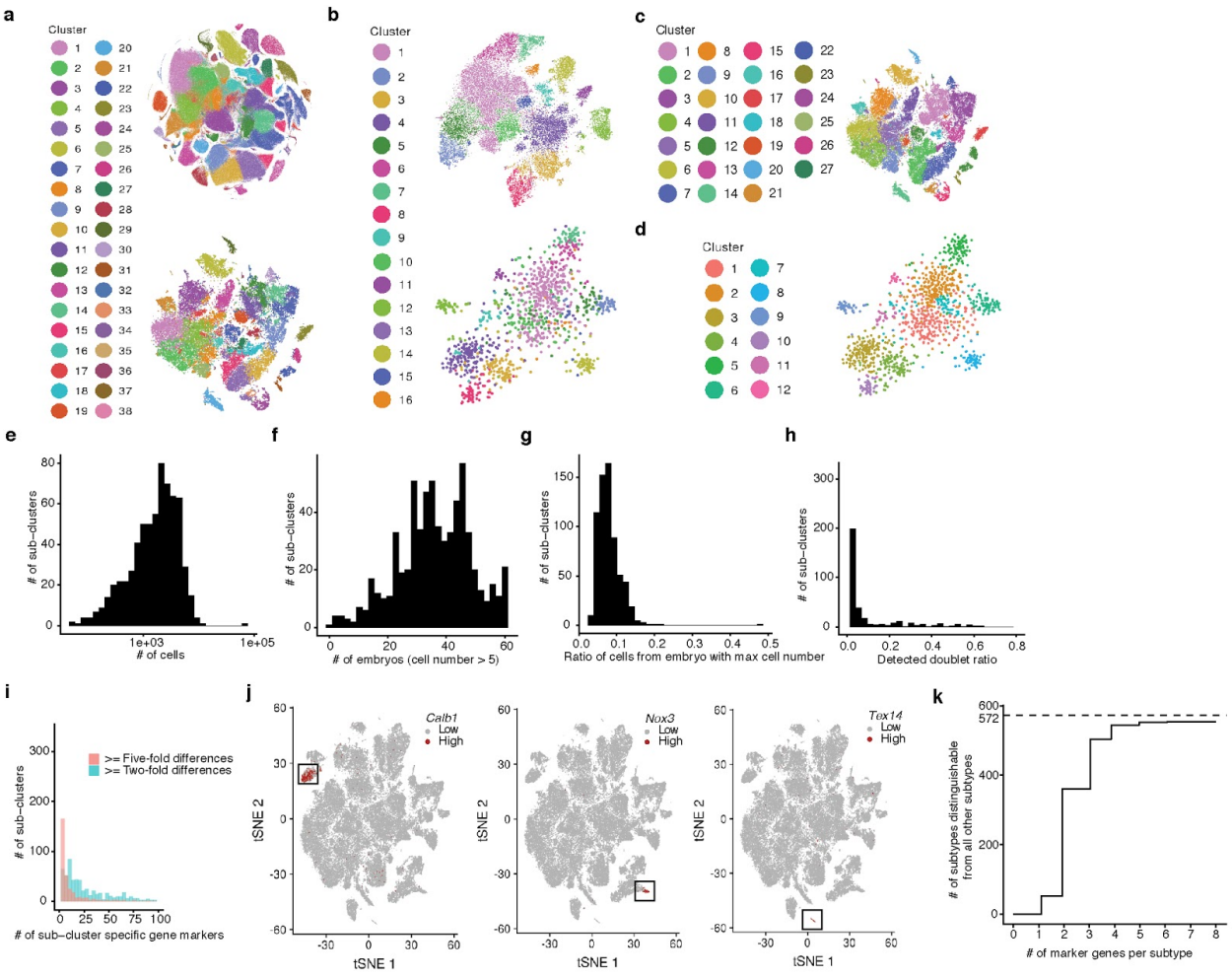
Extended Data Fig. 2. Identifying the major cell types and cell composition dynamics during mouse organogenesis. (**a-e**) t-SNE visualization of mouse embryo cells from different developmental stages, as shown in lower portion of Fig. 2a, but sampling 10,000 cells per stage

and coloring by embryo ID: E9.5 (a), E10.5 (b), E11.5 (c), E12.5 (d), E13.5 (e). We consistently observe that cells derived from independent embryos at the same timepoint are similarly distributed. (**f**) The same t-SNE as Fig. 2a is shown, with subsets of cells highlighted. The first panel only shows cells from E9.5 embryos, and cells from subsequent developmental stages are progressively added. (**g**) Box plot showing the number of UMIs detected per cell for major cell types. Thick horizontal lines, medians; upper and lower box edges, first and third quartiles, respectively; whiskers, 1.5 times the interquartile range; circles, outliers. (**h**) t-SNE visualization of a randomly sampled 100,000 cells colored by expression level of *Hbb-bh1* (top) or *Fndc3c1* (bottom). "High" indicates cells with UMI count for *Hbb-bh1* > 3, *Fndc3c1* > 1. (**i**) Bar plot showing the number of marker genes in each major cell type, defined as differentially expressed genes (5% FDR) with a >2-fold (green) or >5-fold (red) expression difference between first and second ranked cell types. (**j**) Left: t-SNE visualization of a randomly sampled 100,000 cells colored by expression level of *Shh* (top) or *Tox2* (bottom). Right: whole mount *in situ* hybridization images of *Shh* (top) or *Tox2* (bottom) in embryos. n = 5 "High" indicates cells with UMI count for *Shh* > 0, *Tox2* > 1. Arrow: site of gene expression. (**k**) Bar plot showing the number of cells profiled for each cell type, split out by development stage. (**l**) Heatmap showing the estimated relative number of each cell type (rows) in 61 mouse embryos (columns). An estimate of the absolute cell number per cell type per embryo was calculated by multiplying the proportion that cell type contributed to a given embryo by the estimated total number of cells at that development stage. For presentation, these estimates are normalized in each row by the maximum estimated cell count for that cell type across all 61 embryos. Embryos are sorted left-to-right by developmental pseudotime. (**m**) Line plot showing the estimated relative cell numbers for primitive erythroid and definitive erythroid lineages, calculated as in panel b. Dashed lines show relative expression of marker genes for primitive erythroid (*Hbb-bh1*) and definitive erythroid (*Hbb-bs*) major cell types.

Data points for individual embryos were ordered by development pseudotime and smoothed by the loess method.

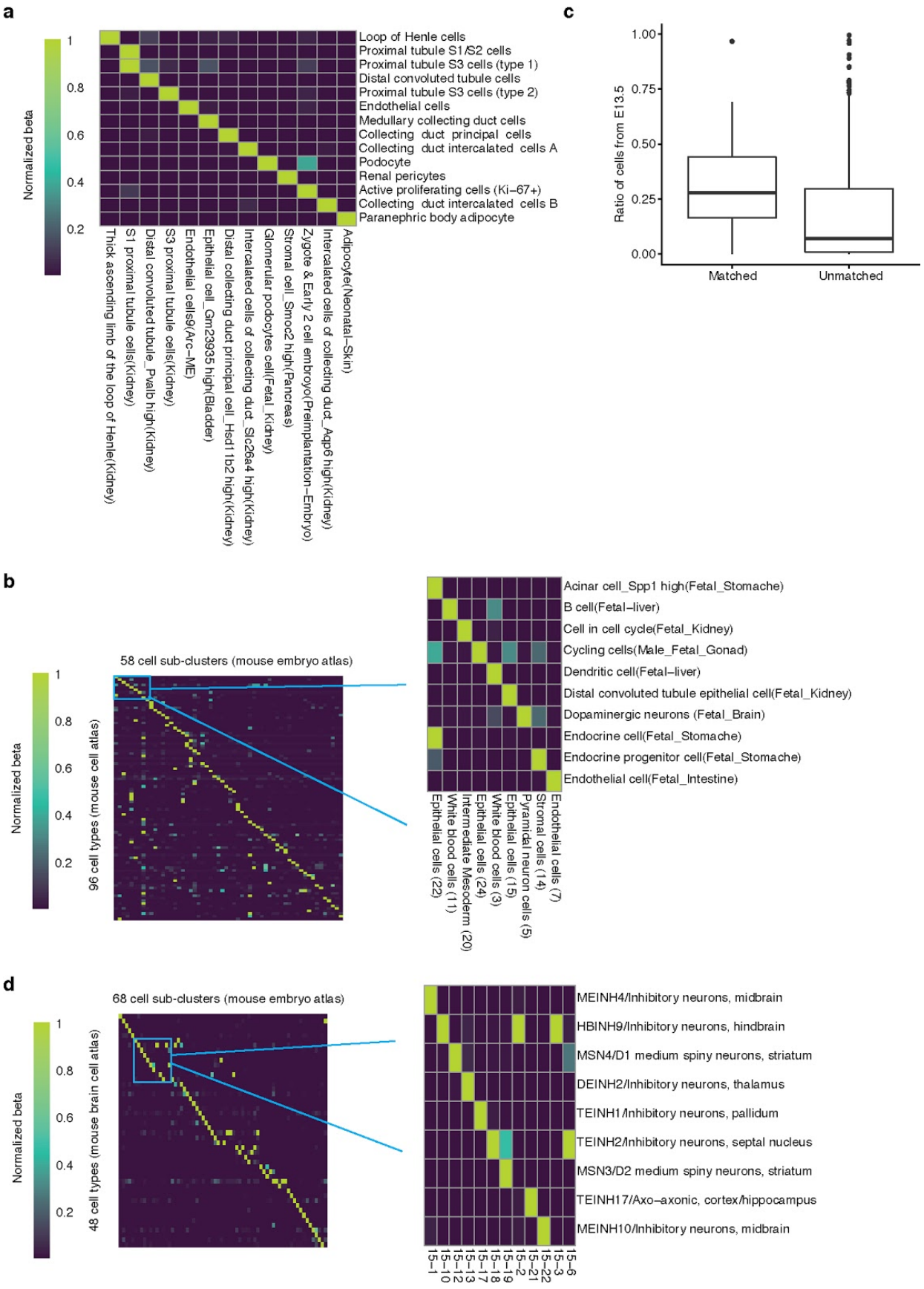


Extended Data Fig. 3. Louvain clustering and t-SNE visualization of subclusters of the each of 38 major cell types. As cell type heterogeneity was readily apparent within many of the 38 clusters shown in Fig. 2a, we adopted an iterative strategy, repeating Louvain clustering on each main cell type to identify subclusters. After subclusters dominated by one or two embryos were removed and highly similar subclusters merged, a total of 655 subclusters (also termed ‘subtypes’ to distinguish them from the 38 major cell types identified by the initial clustering).

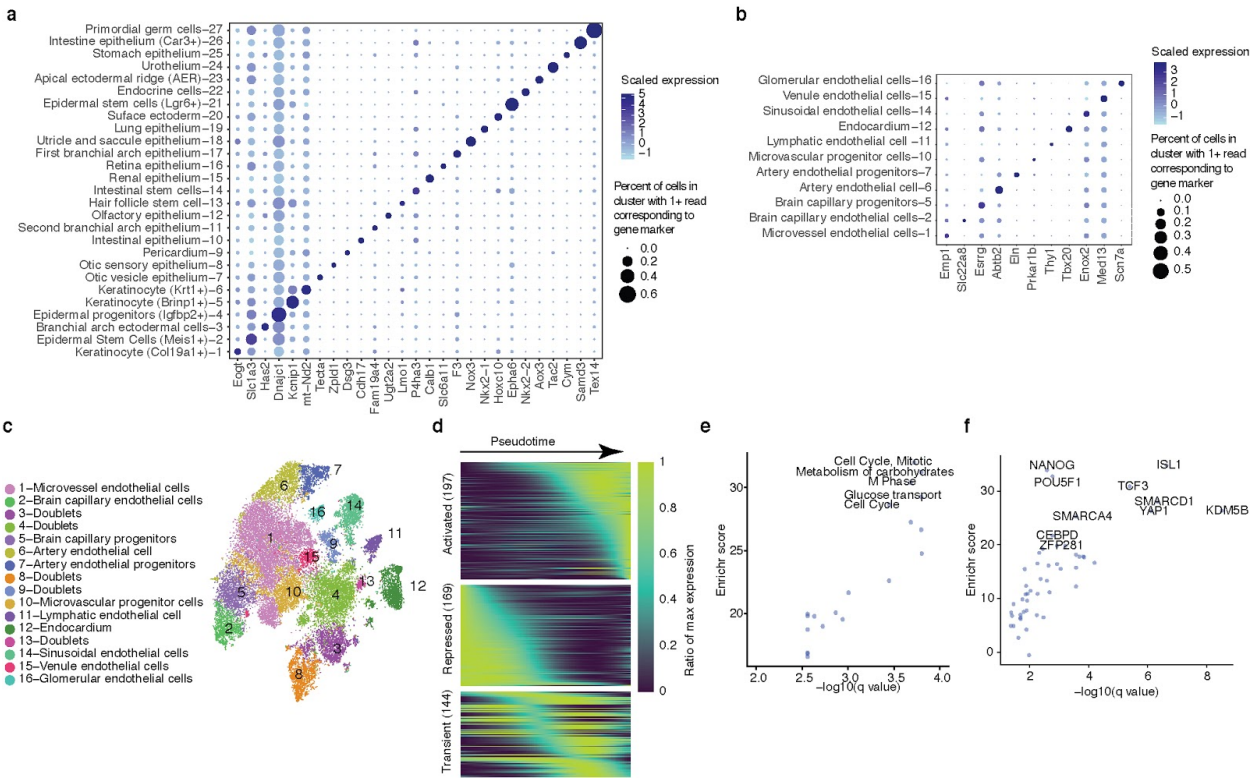


Extended Data Fig. 4. Analysis of cell subtypes during mouse organogenesis. (**a**) t-SNE visualization of all cells (top plot, n = 2,026,641) and downsampled subset of high-quality cells (bottom plot, n = 50,000, UMI > 400), colored by Louvain cluster IDs from Fig. 2a. (**b**) t-SNE visualization of all endothelial cells (top plot, n = 35,878) and those from the downsampled subset (bottom plot, n = 1,173), colored by Louvain cluster ID computed based on the 35,878 endothelial cells. (**c-d**) t-SNE visualization of the downsampled subset of 50,000 cells (c), and 1,173 endothelial cells (d), colored by Louvain cluster ID computed based on sampled cells only. The number of clusters and subclusters identified with the same parameters drops from 38 (a, bottom plot) to 27 (c) and 16 (b, bottom plot) to 12 (c), respectively. (**e**) Histogram showing the distribution of subclusters with respect to cell number (median 1,869; range 51-65,894). (**f**) Histogram showing

the distribution of subclusters with respect to the number of contributing embryos (>5 cells to qualify as a contributor). (**g**) Histogram showing the distribution of subclusters with respect to the ratio of cells derived from the most highly contributing embryo. (**h**) Histogram showing the distribution of subclusters with respect to the ratio of doublet cells detected by Scrublet. (**i**) Histogram showing the distribution of subclusters with respect to the number of marker genes (at least 2-fold (blue) or 5-fold (red) higher expression when compared with the second highest expressing cell subtype within the same main cluster; 5% FDR). 644 of 655 sub-clusters (98%) have at least one such gene marker with a 2-fold difference, and 441 of 655 (67%) have at least one such marker with a 5-fold difference. (**j**) t-SNE visualization of subcluster specific marker expression (as example, cell number n = 74,651): *Calb1* (left), *Nox3* (middle) and *Tex14* (right) are gene markers for three endothelial subclusters. “High” indicates cells with UMI count for *Calb1* > 0, *Nox3* > 0, *Tex14* >1. (**k**) Cumulative histogram showing how many subtypes (out of a total of 572 non-doublet-artifact subtypes) can be distinguished from all other subtypes on the basis of one or several markers and >4-fold expression differences (Methods).

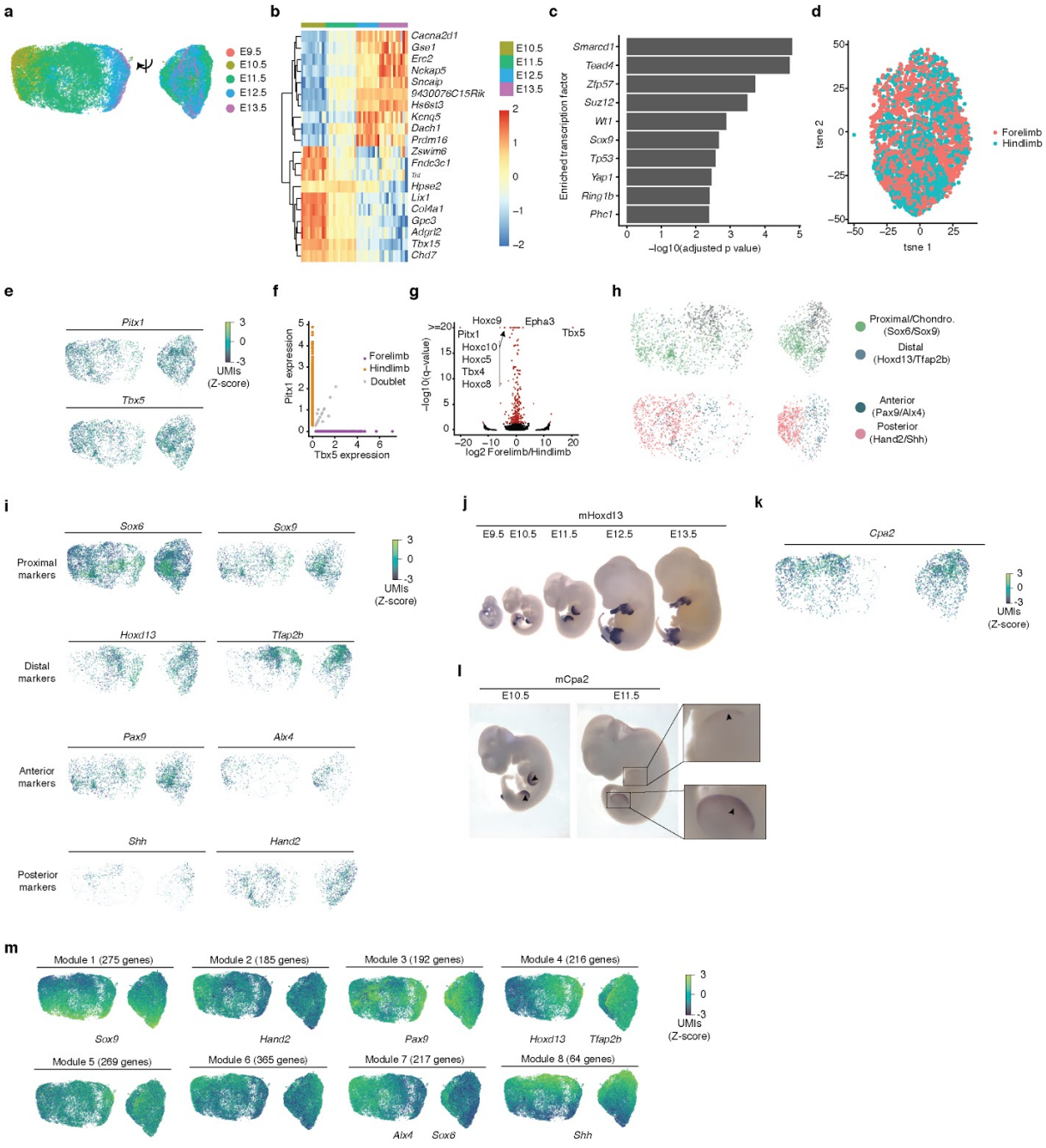

a
Normalized beta
1
0.8
0.6
0.4
0.2
Loop of Henle cells
Proximal tubule S1/S2 cells
Proximal tubule S3 cells (type 1)
Distal convoluted tubule cells
Proximal tubule S3 cells (type 2)
Endothelial cells
Medullary collecting duct cells
Collecting duct principal cells
Collecting duct intercalated cells A
Podocyte
Renal pericytes
Active proliferating cells (Ki−67+)
Collecting duct intercalated cells B
Paranephric body adipocyte
Thick ascending limb of the loop of Henle(Kidney)
S1 proximal tubule cells(Kidney)
Distal convoluted tubule_Pvalb high(Kidney)
S3 proximal tubule cells(Kidney)
Endothelial cells9(Arc−ME)
Epithelial cell_Gm23935 high(Bladder)
Distal collecting duct principal cell_Hsd11b2 high(Kidney)
Intercalated cells of collecting duct_Slc26a4 high(Kidney)
Glomerular podocytes cell(Fetal_Kidney)
Stromal cell_Smoc2 high(Pancreas)
Zygote & Early 2 cell embryo(Preimplantation−Embryo)
Intercalated cells of collecting duct_Aqp6 high(Kidney)
Adipocyte(Neonatal−Skin)
c
Ratio of cells from E13.5
1.00
0.75
0.50
0.25
0.00
Matched
Unmatched
b
58 cell sub-clusters (mouse embryo atlas)
Normalized beta
1
0.8
0.6
0.4
0.2
96 cell types (mouse cell atlas)
Acinar cell_Spp1 high(Fetal_Stomache)
B cell(Fetal−liver)
Cell in cell cycle(Fetal_Kidney)
Cycling cells(Male_Fetal_Gonad)
Dendritic cell(Fetal−liver)
Distal convoluted tubule epithelial cell(Fetal_Kidney)
Dopaminergic neurons (Fetal_Brain)
Endocrine cell(Fetal_Stomache)
Endocrine progenitor cell(Fetal_Stomache)
Endothelial cell(Fetal_Intestine)
Epithelial cells (22)
White blood cells (11)
Intermediate Mesoderm (20)
Epithelial cells (24)
White blood cells (3)
Epithelial cells (15)
Pyramidal neuron cells (5)
Stromal cells (14)
Endothelial cells (7)
d
68 cell sub-clusters (mouse embryo atlas)
Normalized beta
1
0.8
0.6
0.4
0.2
48 cell types (mouse brain cell atlas)
MEINH4/Inhibitory neurons, midbrain
HBINH9/Inhibitory neurons, hindbrain
MSN4/D1 medium spiny neurons, striatum
DEINH2/Inhibitory neurons, thalamus
TEINH1/Inhibitory neurons, pallidum
TEINH2/Inhibitory neurons, septal nucleus
MSN3/D2 medium spiny neurons, striatum
TEINH17/Axo−axonic, cortex/hippocampus
MEINH10/Inhibitory neurons, midbrain
15−1
15−10
15−12
15−13
15−17
15−18
15−19
15−2
15−21
15−22
15−3
15−6

Extended Data Fig. 5. Cell type correlation analysis between single cell mouse atlases. (**a**) Cell type correlation analysis (Methods) matched cell types between independently generated and annotated analyses of the adult mouse kidney (sci-RNA-seq component of sci-CAR(*19*) (rows) vs. Microwell-seq(*10*) (columns)). All cell types identified by sci-RNA-seq are shown, but we only show Microwell-seq cell types that are top matches for 1+ sci-RNA-seq cell types. Colors correspond to beta values, normalized by the maximum beta value per row. (**b**) Left: We compared our subtypes against 130 fetal cell types annotated in the MCA(*10*) with cell type correlation analysis, matching 96 MCA-defined cell types (rows) to 58 subtypes in our mouse embryo atlas (columns). Colors correspond to beta values, normalized by the maximum beta value per row. All MCA cell types with maximum beta of matched cell type > 0.01 are shown (rows; n = 96), as are mouse embryo atlas cell types that are top matches for 1+ displayed MCA cell types (columns; n = 58). Right: zoom-in to a subset of matches shown on the left. Cell types annotations are from MCA (rows) or our study (columns; major cell type annotation and sub-cluster id). (**c**) Box plot showing the ratio of cells from E13.5 for subclusters with (sub-cluster number n = 58) vs. without (sub-cluster number n = 514) a matched cell type in the MCA. Thick horizontal lines, medians; upper and lower box edges, first and third quartiles, respectively; whiskers, 1.5 times the interquartile range; circles, outliers. (**d**) Left: We compared our subtypes against 265 cell types annotated by a recent mouse brain cell atlas (BCA)(*32*) with cell type correlation analysis, matching 48 BCA-defined cell types (rows) to 68 subtypes in our data (columns). Colors correspond to beta values, normalized by the maximum beta value per row. All mouse embryo cell types with maximum beta of matched cell type > 0.01 are shown (column; n = 68), as are BCA cell types that are top matches for 1+ displayed mouse embryo cell types (rows; n = 48). Right: zoom-in to a subset of matches shown on the left. Cell types annotations are from BCA (rows) or our study (columns; major cell cluster and sub-cluster id).



Extended Data Fig. 6. Analysis of mouse epithelium, endothelium and limb apical ectodermal ridge cells. (**a-b**) Dot plot showing expression of one selected marker gene per epithelial (a) or endothelial (b) subtype. Doublet-derived subclusters (2/29 epithelial subtypes and 5/16 endothelial subtypes) are excluded from these plots, but are still shown in Fig. 3a and panel c, respectively. The size of the dot encodes the percentage of cells within a cell type, and its color encodes the average expression level. (**c**) t-SNE visualization and marker-based annotation of endothelial cell subtypes (n = 35,878). (**d**) Heatmap showing smoothed pseudotime-dependent differential gene expression (169 genes at FDR of 1%) in AER cells, generated by a spline fitting with generalized linear model (assuming gene expression following the negative binomial distribution) and scaled as a percent of maximum gene expression. Each row indicates a different gene, and these are split into subsets that are activated (top), repressed (middle) or exhibit transient dynamics (bottom) between E9.5 and E13.5. (**e-f**) Plots showing the -log10 transformed q value and Enrichr based

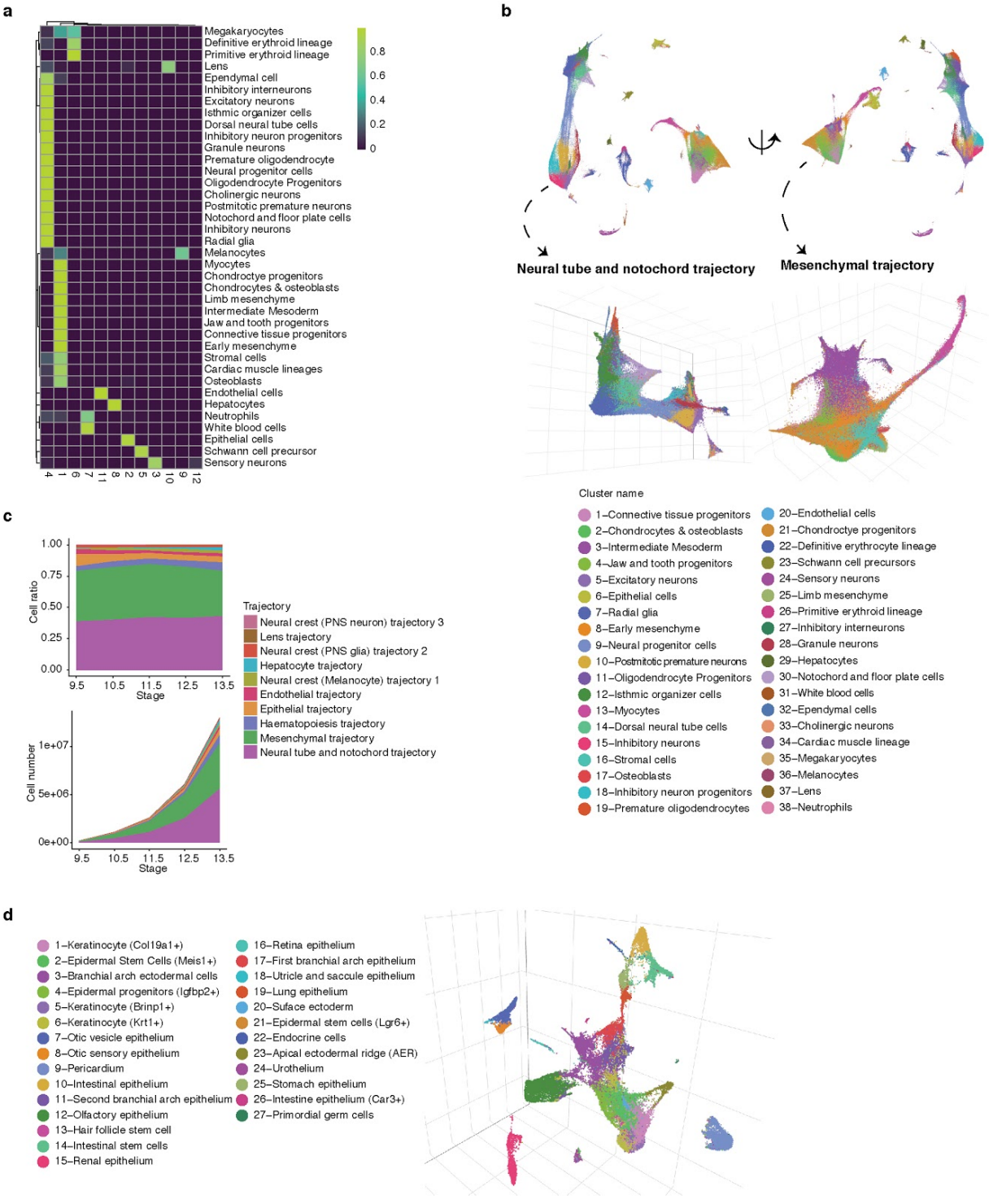
combined score of enriched Reactome terms (e) and transcription factors (f) for genes whose expression significantly decreases in AER development. The top enriched pathway terms (Reactome2016) for significantly decreasing genes include cell cycle progression (Mitotic Cell Cycle, qval = 0.0002, one-sided Fisher exact test with multiple comparisons adjusted) and glucose metabolism (Metabolism of carbohydrates, qval = 0.0002, one-sided Fisher exact test with multiple comparisons adjusted). The top enriched TFs with targets from decreasing genes include pluripotent factors such as *Isl1* (qval < 1e-5), *Pou5f1* (qval = 0.002, one-sided Fisher exact test with multiple comparisons adjusted) and *Nanog* (qval = 0.003, one-sided Fisher exact test with multiple comparisons adjusted).



Extended Data Fig. 7. Characterizing cellular trajectories during limb mesenchyme differentiation. (**a**) UMAP 3D visualization of limb mesenchymal cells colored by development stage (cell number n = 26,559, left and right represent views from two directions). (**b**) Heatmap showing top differentially expressed genes between different developmental stages for limb

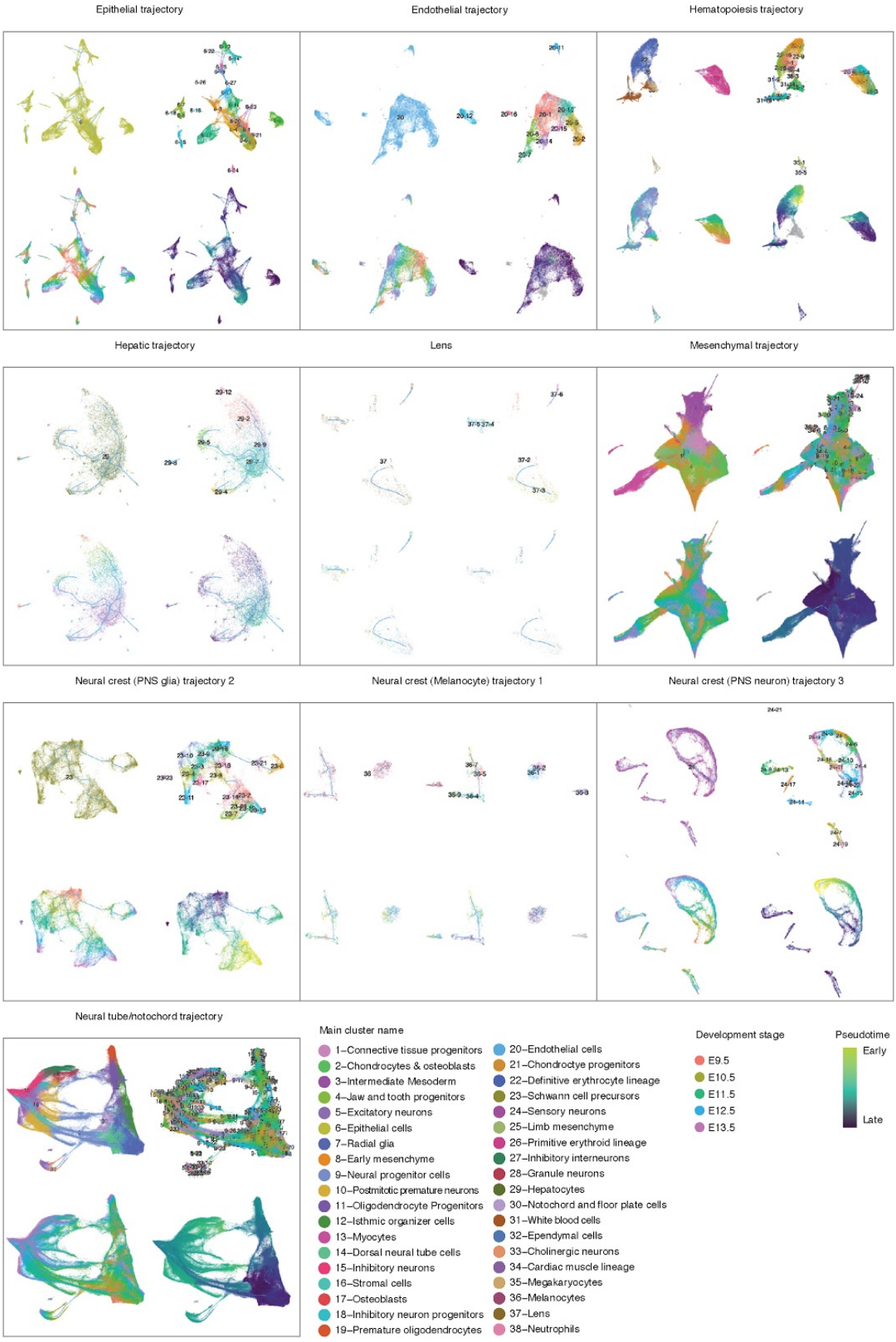
mesenchyme cells. (**c**) Bar plot showing the -log10 transformed adjusted p value (one-sided Fisher exact test with multiple comparisons adjusted) of enriched transcription factors for significantly up-regulated genes during limb mesenchyme development. (**d**) t-SNE visualization of limb mesenchyme cells colored by forelimb (Tbx5 +, cell number n = 2,085) and hindlimb (Pitx1+, cell number n = 1,885). Cells with no expression or both expression in Tbx5 and Pitx1 are not shown. (**e, h, i, k**) Each panel illustrates a different marker gene. Colors indicate UMI counts that have been scaled for library size, log-transformed, and then mapped to Z-scores to enable comparison between genes. Cells with no expression of a given marker are excluded to prevent overplotting. (**e**) Hindlimb marker *Pitx1* and forelimb marker *Tbx5*. (**f**) Scatter plot showing the normalized expression of *Pitx1* and Tbx5 in limb mesenchyme cells. Only cells in which *Pitx1* and/or *Tbx5* detected were shown. (**g**) Volcano plot showing the differentially expressed genes (FDR of 5%, one-sided likelihood ratio test with multiple comparisons adjusted, colored by red) between forelimb (cell number n = 2,085) and hindlimb (cell number n = 1,885). Top differentially expressed genes are labeled. X axis: log2 transformed fold change between forelimb and hindlimb for each gene. Y axis: -log10 transformed qval from differential gene expression test. (**h**) Same visualization as panel e, colored by normalized gene expression of proximal/chondrocyte (*Sox6, Sox9*), distal (*Hoxd13*, *Tfap2b*), anterior (*Pax9*, *Alx4*), or posterior (*Hand2*, *Shh*) markers. Only cells with the gene marker expressed are plotted. (**i**) Same visualization as panel e. First row: proximal limb markers *Sox6* (which also marks chondrocytes) and *Sox9*. Second row: distal limb markers *Hoxd13* and *Tfap2b*. Third row: Anterior limb markers(*51*) *Pax9* and *Alx4*. Fourth row: posterior limb markers *Shh* and *Hand2*. (**j**) *In situ* hybridization images of *Hoxd13* in E9.5 to E13.5 embryos, n = 5. (**k**) Same visualization as panels e, colored by normalized gene expression of *Cpa2*. Only cells with positive UMI counts are shown. Values are log10-transformed, standardized UMI counts. Its expression pattern within this trajectory led us to predict that *Cpa2* is a distal marker of

the developing limb mesenchyme, like *Hoxd13*. (**l**) *In situ* hybridization images of *Cpa2* in E10.5 and E11.5 embryos, n = 5. Arrow: site of gene expression. (**m**) Modules of spatially restricted genes in the limbs. A total of 1,783 genes were clustered via hierarchical clustering. The dendrogram was cut into 8 modules using the cutree function in R, and the aggregate expression of genes in each module was computed. Colors indicate aggregate UMI counts for each module that have been scaled for library size, log-transformed, and then mapped to Z-scores to enable comparison between modules. Cells with no expression of a given module are excluded to prevent overplotting.

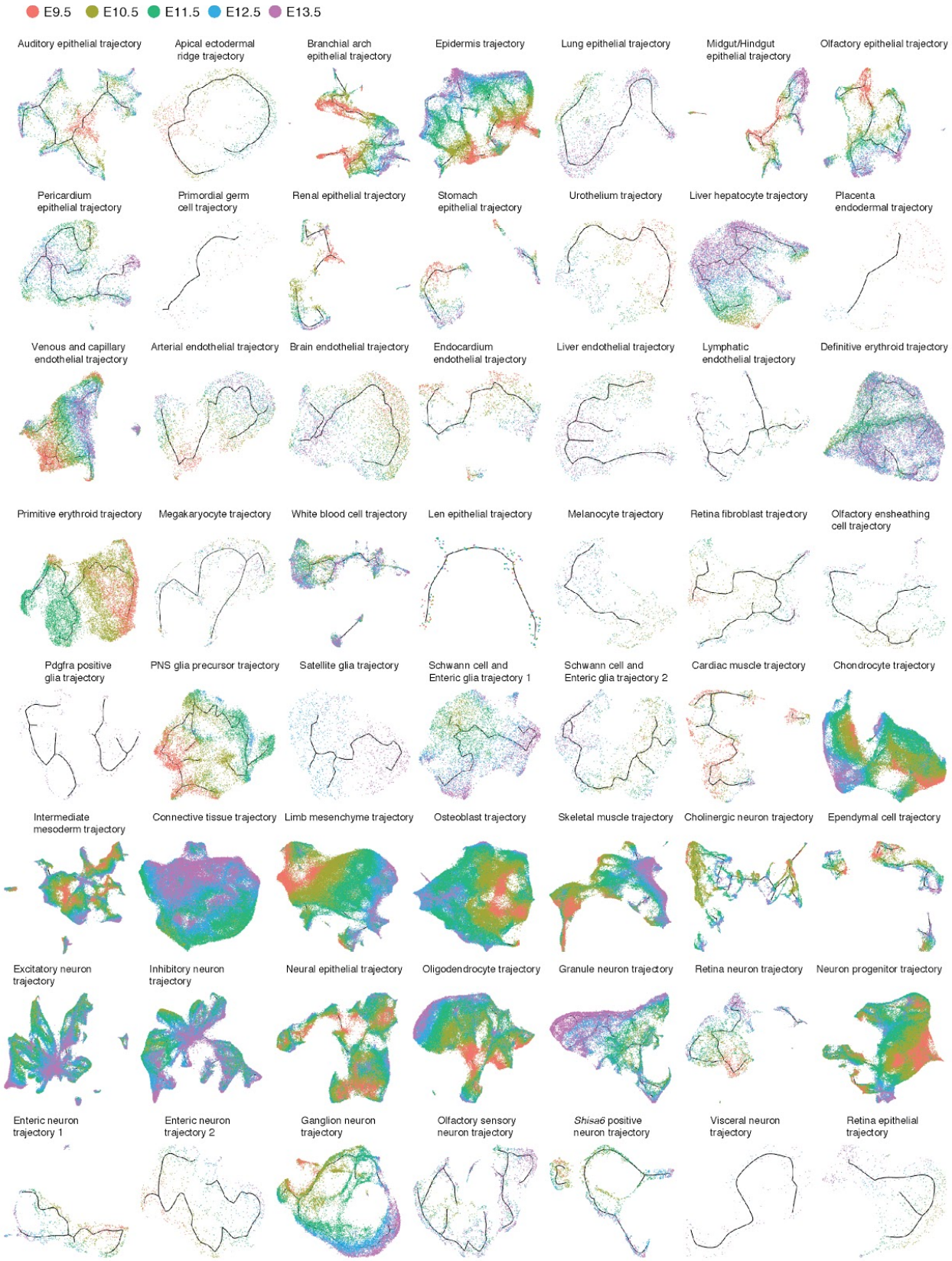


Extended Data Fig. 8. Characterization of ten major developmental trajectories present during mouse organogenesis. (**a**) Heatmap showing the proportion of cells from each of the 38 major cell types assigned to each of the twelve PAGA algorithm-identified groups. We merged

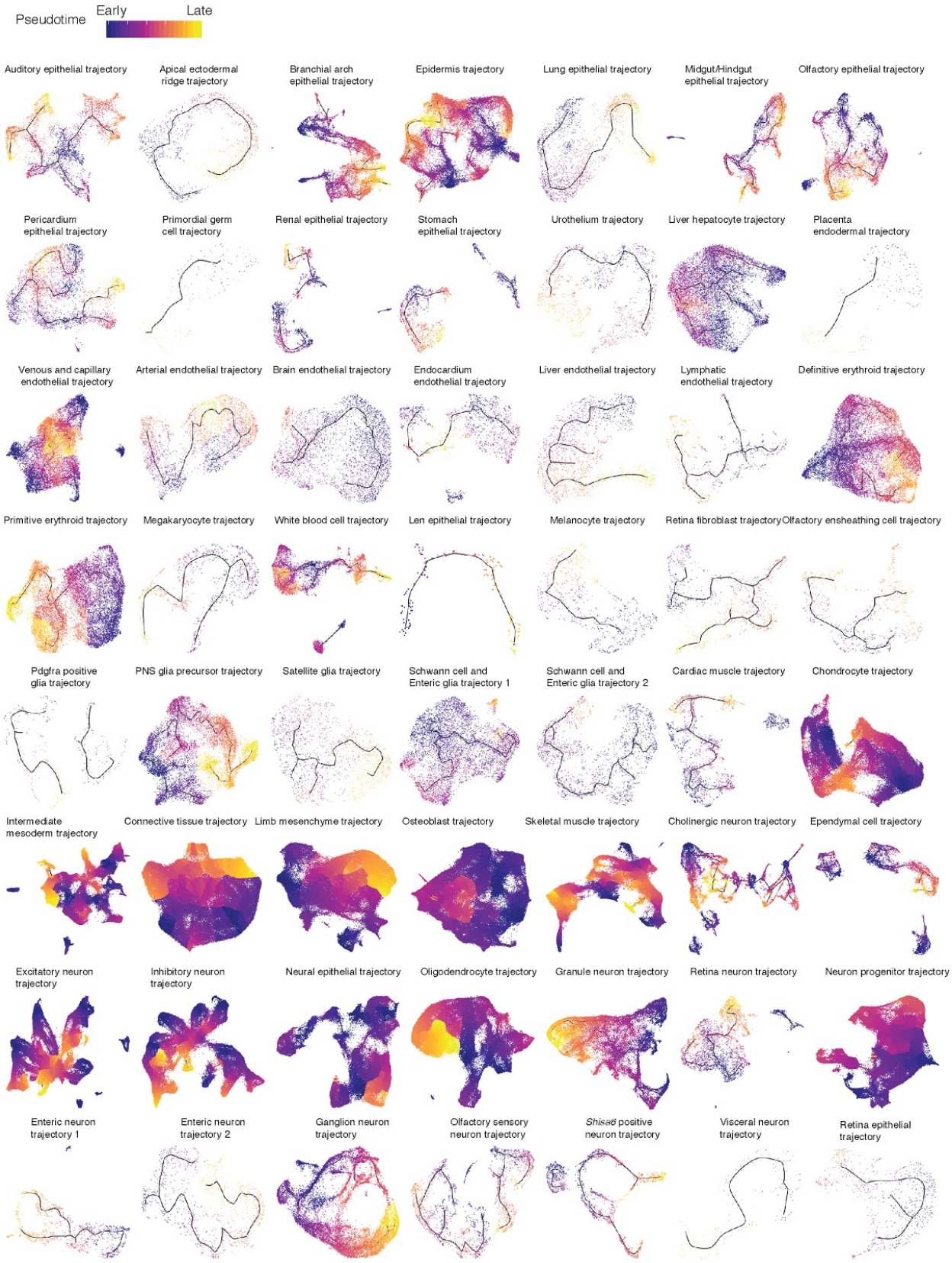
two groups corresponding to sensory neurons (12 & 3), and another two groups corresponding to blood cells (6 & 7), as each pair was closely located in UMAP space upon visual inspection, yielding the ten supergroups shown in a similar heatmap in Fig. 4b. (**b**) Same as Fig. 4a, but with colors corresponding to the 38 major cell clusters. (**c**) Area plot showing the estimated proportion (top) and estimated absolute number (bottom) of cells per embryo derived from each of the ten major cell trajectories from E9.5 to E13.5. Although the estimated number of cells per embryo in each of these supergroups increases exponentially, their proportions remain relatively stable, with the exception of hepatocytes which expand their contribution by nearly ten-fold during this developmental window (from 0.3% at E9.5 to 2.8% at E13.5). (**d**) UMAP 3D visualization of epithelial subtrajectories (as in Fig. 4c), colored as per the epithelial subtypes shown in Fig. 3a.


Epithelial trajectory
Endothelial trajectory
Hematopoiesis trajectory
Hepatic trajectory
Lens
Mesenchymal trajectory
Neural crest (PNS glia) trajectory 2
Neural crest (Melanocyte) trajectory 1
Neural crest (PNS neuron) trajectory 3
Neural tube/notochord trajectory
Main cluster name
1–Connective tissue progenitors
2–Chondrocytes & 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–Chondroctye 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
Development stage
E9.5
E10.5
E11.5
E12.5
E13.5
Pseudotime
Early
Late

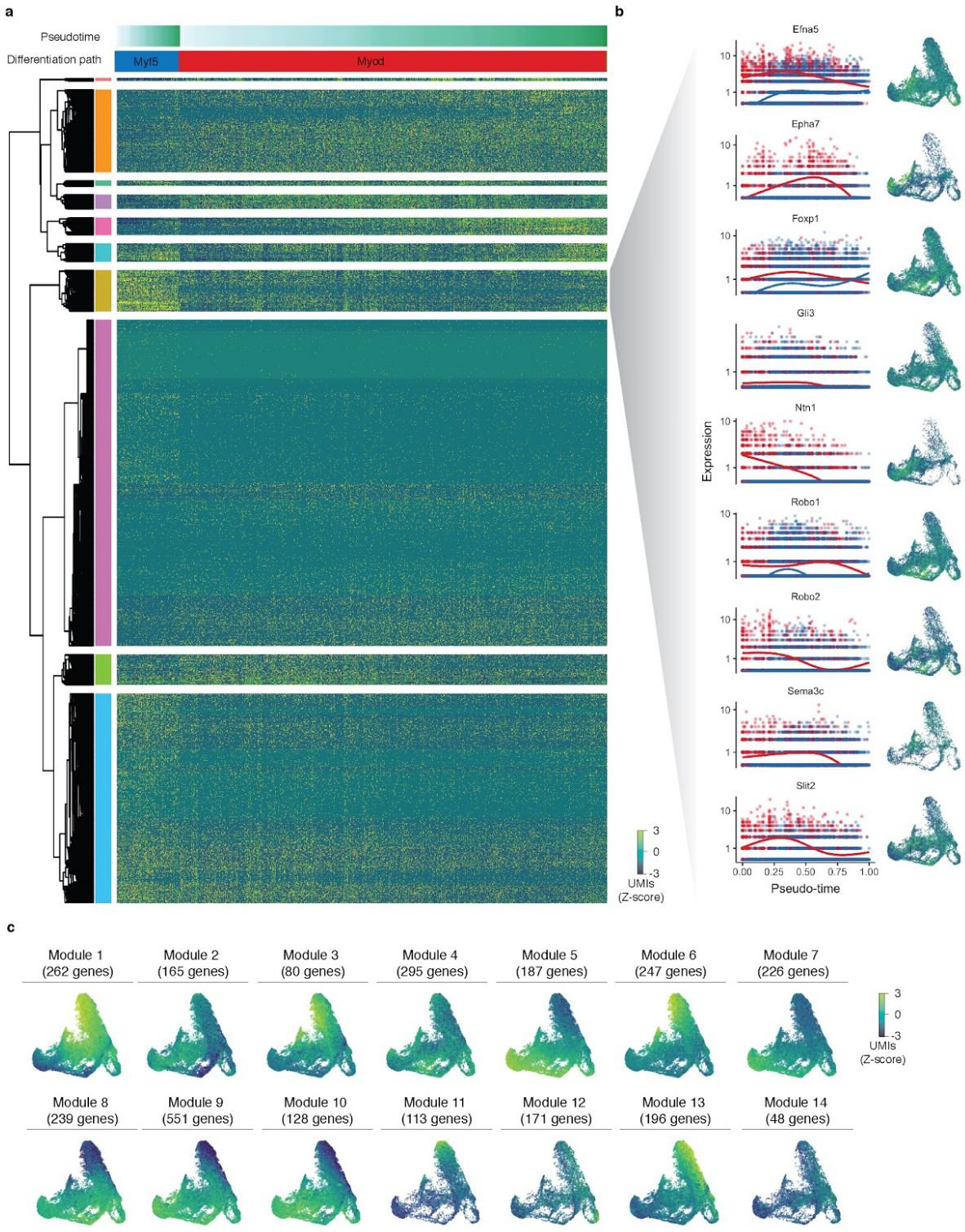
Extended Data Fig. 9. UMAP visualization of the ten major cell trajectories. We iteratively reanalyzed each of the ten major trajectories, nearly all of which further resolved into multiple subtrajectories. The ten major cell trajectories are visualized with UMAP (as in Fig. 5) but colored: as per the 38 major cell clusters (top left), sub-cluster id (top right), developmental stage (bottom left) and pseudotime (bottom right). The lines correspond to the principal graph learned by Monocle 3. These images are also available at http://atlas.gs.washington.edu/mouse-rna/3dplot/ as manipulatable 3D renderings.


E9.5
E10.5
E11.5
E12.5
E13.5
Auditory epithelial trajectory
Apical ectodermal ridge trajectory
Branchial arch epithelial trajectory
Epidermis trajectory
Lung epithelial trajectory
Midgut/Hindgut epithelial trajectory
Olfactory epithelial trajectory
Pericardium epithelial trajectory
Primordial germ cell trajectory
Renal epithelial trajectory
Stomach epithelial trajectory
Urothelium trajectory
Liver hepatocyte trajectory
Placenta endodermal trajectory
Venous and capillary endothelial trajectory
Arterial endothelial trajectory
Brain endothelial trajectory
Endocardium endothelial trajectory
Liver endothelial trajectory
Lymphatic endothelial trajectory
Definitive erythroid trajectory
Primitive erythroid trajectory
Megakaryocyte trajectory
White blood cell trajectory
Len epithelial trajectory
Melanocyte trajectory
Retina fibroblast trajectory
Olfactory ensheathing cell trajectory
Pdgfra positive glia trajectory
PNS glia precursor trajectory
Satellite glia trajectory
Schwann cell and Enteric glia trajectory 1
Schwann cell and Enteric glia trajectory 2
Cardiac muscle trajectory
Chondrocyte trajectory
Intermediate mesoderm trajectory
Connective tissue trajectory
Limb mesenchyme trajectory
Osteoblast trajectory
Skeletal muscle trajectory
Cholinergic neuron trajectory
Ependymal cell trajectory
Excitatory neuron trajectory
Inhibitory neuron trajectory
Neural epithelial trajectory
Oligodendrocyte trajectory
Granule neuron trajectory
Retina neuron trajectory
Neuron progenitor trajectory
Enteric neuron trajectory 1
Enteric neuron trajectory 2
Ganglion neuron trajectory
Olfactory sensory neuron trajectory
Shisa6 positive neuron trajectory
Visceral neuron trajectory
Retina epithelial trajectory

Extended Data Fig. 10. UMAP visualization of the 56 subtrajectories, colored by development stage. We further iteratively reanalyzed and visualized with UMAP each of the 56 subtrajectories. Although Monocle 3 did not have access to these labels, the subtrajectories are highly consistent with developmental time (*i.e.* cells ordered from E9.5 to E13.5). The lines correspond to the principal graph learned by Monocle 3.


Pseudotime
Early
Late
Auditory epithelial trajectory
Apical ectodermal ridge trajectory
Branchial arch epithelial trajectory
Epidermis trajectory
Lung epithelial trajectory
Midgut/Hindgut epithelial trajectory
Olfactory epithelial trajectory
Pericardium epithelial trajectory
Primordial germ cell trajectory
Renal epithelial trajectory
Stomach epithelial trajectory
Urothelium trajectory
Liver hepatocyte trajectory
Placenta endodermal trajectory
Venous and capillary endothelial trajectory
Arterial endothelial trajectory
Brain endothelial trajectory
Endocardium endothelial trajectory
Liver endothelial trajectory
Lymphatic endothelial trajectory
Definitive erythroid trajectory
Primitive erythroid trajectory
Megakaryocyte trajectory
White blood cell trajectory
Len epithelial trajectory
Melanocyte trajectory
Retina fibroblast trajectory
Olfactory ensheathing cell trajectory
Pdgfra positive glia trajectory
PNS glia precursor trajectory
Satellite glia trajectory
Schwann cell and Enteric glia trajectory 1
Schwann cell and Enteric glia trajectory 2
Cardiac muscle trajectory
Chondrocyte trajectory
Intermediate mesoderm trajectory
Connective tissue trajectory
Limb mesenchyme trajectory
Osteoblast trajectory
Skeletal muscle trajectory
Cholinergic neuron trajectory
Ependymal cell trajectory
Excitatory neuron trajectory
Inhibitory neuron trajectory
Neural epithelial trajectory
Oligodendrocyte trajectory
Granule neuron trajectory
Retina neuron trajectory
Neuron progenitor trajectory
Enteric neuron trajectory 1
Enteric neuron trajectory 2
Ganglion neuron trajectory
Olfactory sensory neuron trajectory
Shisa6 positive neuron trajectory
Visceral neuron trajectory
Retina epithelial trajectory

Extended Data Fig. 11. UMAP visualization of the 56 subtrajectories, colored by inferred pseudotime. To orient each subtrajectory (same projections as Extended Data Fig. 10), we identified one or several starting points as focal concentrations of E9.5 cells, and then computed developmental pseudotime for cells present along various paths. The lines correspond to the principal graph learned by Monocle 3.



Extended Data Fig. 12. Gene dynamics in the myogenic trajectory. (**a**) Genes that are differentially expressed between the Myf5 path and the Myod path highlighted in Fig. 6. Cells

along each path were compared via Monocle's differentialGeneTest function. Pseudotimes along each path were scaled from 0 to 100 independently. The "full model" formula was "~path * sm.ns(Pseudotime, df=3)", while the "reduced model" was "~sm.ns(Pseudotime, df=3)". Differentially expressed genes (FDR < 1%, one-sided likelihood ratio test with multiple comparisons adjusted) were clustered via Ward's method and visualized as a heatmap via the pheatmap package. (**b**) Pseudotemporal kinetics for selected genes involved in Robo/Slit signaling. Red indicates cells on the Myod path, while blue corresponds to the Myf5 path. Next to the expression curves for each are shown the standardized expression scores for each gene on the original myogenic trajectory. Only cells with detectable expression are rendered to prevent overplotting. (**c**) Modules of genes differentially expressed over the myogenic trajectory. A total of 2,908 genes were clustered via hierarchical clustering. The dendrogram was cut into 14 modules using the cutree function in R, and the aggregate expression of genes in each module was computed. Colors indicate aggregate UMI counts for each module that have been scaled for library size, log-transformed, and then mapped to Z-scores to enable comparison between modules. Cells with no expression of a given module are excluded to prevent overplotting.

References of chapter 3:

1. Y. Kojima, O. H. Tam, P. P. L. Tam, Timing of developmental events in the early mouse embryo. *Semin. Cell Dev. Biol.* **34**, 65–75 (2014).

2. P. P. L. Tam, D. A. F. Loebel, Gene function in mouse embryogenesis: get set for gastrulation. *Nat. Rev. Genet.* **8**, 368–381 (2007).

3. M. E. Dickinson *et al.*, High-throughput discovery of novel developmental phenotypes. *Nature*. **537**, 508–514 (2016).

4. T. F. Meehan *et al.*, Disease model discovery from 3,328 gene knockouts by The International Mouse Phenotyping Consortium. *Nat. Genet.* **49**, 1231–1238 (2017).

5. D. E. Wagner *et al.*, Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*. **360**, 981–987 (2018).

6. J. A. Briggs *et al.*, The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science*. **360** (2018), doi:10.1126/science.aar5780.

7. J. A. Farrell *et al.*, Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*. **360** (2018), doi:10.1126/science.aar3131.

8. C. Mayer *et al.*, Developmental diversification of cortical inhibitory interneurons. *Nature*. **555**, 457–462 (2018).

9. F. Lescroart *et al.*, Defining the earliest step of cardiovascular lineage segregation by single-cell RNA-seq. *Science* (2018), doi:10.1126/science.aao4174.

10. X. Han *et al.*, Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*. **172**, 1091–1107.e17 (2018).

11. The Tabula Muris Consortium, S. R. Quake, T. Wyss-Coray, S. Darmanis, Transcriptomic characterization of 20 organs and tissues from mouse at single cell resolution creates a Tabula Muris (2017), , doi:10.1101/237446.

12. S. Amini *et al.*, Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343–1349 (2014).

13. A. Adey *et al.*, In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res.* **24**, 2041–2049 (2014).

14. D. A. Cusanovich *et al.*, Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*. **348**, 910–914 (2015).

15. S. A. Vitak *et al.*, Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods*. **14**, 302–308 (2017).

16. V. Ramani *et al.*, Massively multiplex single-cell Hi-C. *Nat. Methods*. **14**, 263–266 (2017).

17. J. Cao *et al.*, Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*. **357**, 661–667 (2017).

18. R. M. Mulqueen *et al.*, Scalable and efficient single-cell DNA methylation sequencing by combinatorial indexing (2017), , doi:10.1101/157230.

19. J. Cao *et al.*, Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*. **361**, 1380–1385 (2018).

20. A. B. Rosenberg *et al.*, Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* (2018), doi:10.1126/science.aam8999.

21. G. La Manno *et al.*, RNA velocity of single cells. *Nature*. **560**, 494–498 (2018).

22. S. L. Wolock, R. Lopez, A. M. Klein, Scrublet: computational identification of cell doublets in single-cell transcriptomic data (2018), , doi:10.1101/357368.

23. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell developmental trajectories (2017), , doi:10.1101/110668.

24. A. Yang *et al.*, p63 is essential for regenerative proliferation in limb, craniofacial and epithelial development. *Nature*. **398**, 714–718 (1999).

25. J. L. McQualter, K. Yuen, B. Williams, I. Bertoncello, Evidence of an epithelial stem/progenitor cell hierarchy in the adult mouse lung. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 1414–1419 (2010).

26. M. Cichorek, M. Wachulska, A. Stasiewicz, A. Tymińska, Skin melanocytes: biology and development. *Advances in Dermatology and Allergology*. **1**, 30–41 (2013).

27. M. Tomihari, S.-H. Hwang, J.-S. Chung, P. D. Cruz Jr., K. Ariizumi, Gpnmb is a melanosome-associated glycoprotein that contributes to melanocyte/keratinocyte adhesion in a RGD-dependent fashion. *Exp. Dermatol.* **18**, 586–595 (2009).

28. M. Varjosalo, J. Taipale, Hedgehog: functions and mechanisms. *Genes Dev.* **22**, 2454–2472 (2008).

29. U. Strähle, C. S. Lam, R. Ertzer, S. Rastegar, Vertebrate floor-plate specification: variations on common themes. *Trends Genet.* **20**, 155–162 (2004).

30. G. P. Holmes *et al.*, Distinct but overlapping expression patterns of two vertebrate slit homologs implies functional roles in CNS development and organogenesis. *Mech. Dev.* **79**, 57–72 (1998).

31. V. Akle *et al.*, F-spondin/spon1b expression patterns in developing and adult zebrafish. *PLoS One*. **7**, e37593 (2012).

32. A. Zeisel *et al.*, Molecular Architecture of the Mouse Nervous System. *Cell*. **174**, 999–1014.e22 (2018).

33. B. H. Hartman, R. Durruthy-Durruthy, R. D. Laske, S. Losorelli, S. Heller, Identification and characterization of mouse otic sensory lineage genes. *Front. Cell. Neurosci.* **9**, 79 (2015).

34. E. Szenker-Ravi *et al.*, RSPO2 inhibition of RNF43 and ZNRF3 governs limb development independently of LGR4/5/6. *Nature*. **557**, 564–569 (2018).

35. X. Cai *et al.*, Tbx20 acts upstream of Wnt signaling to regulate endocardial cushion formation and valve remodeling during mouse cardiogenesis. *Development*. **140**, 3176–3187 (2013).

36. R. A. Miller, N. Christoforou, J. Pevsner, A. S. McCallion, J. D. Gearhart, Efficient array-based identification of novel cardiac genes through differentiation of mouse ESCs. *PLoS One*. **3**, e2176 (2008).

37. F. Petit, K. E. Sears, N. Ahituv, Limb development: a paradigm of gene regulation. *Nat. Rev. Genet.* **18**, 245–258 (2017).

38. Q. Guo, C. Loomis, A. L. Joyner, Fate map of mouse ventral limb ectoderm and the apical ectodermal ridge. *Dev. Biol.* **264**, 166–178 (2003).

39. E. al Lewandoski M, Fgf8 signalling from the AER is essential for normal limb development. - PubMed - NCBI, (available at https://www.ncbi.nlm.nih.gov/pubmed/11101846).

40. J. Gerdes, U. Schwab, H. Lemke, H. Stein, Production of a mouse monoclonal antibody reactive with a human nuclear antigen associated with cell proliferation. *Int. J. Cancer*. **31**, 13–20 (1983).

41. D. Bergman, M. Halje, M. Nordin, W. Engström, Insulin-like growth factor 2 in development and disease: a mini-review. *Gerontology*. **59**, 240–249 (2013).

42. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

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

44. F. Alexander Wolf *et al.*, Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *bioRxiv* (2017), p. 208819.

45. T. Braun, M. Gautel, Transcriptional mechanisms regulating skeletal muscle differentiation, growth and homeostasis. *Nat. Rev. Mol. Cell Biol.* **12**, 349–361 (2011).

46. G. Comai, R. Sambasivan, S. Gopalakrishnan, S. Tajbakhsh, Variations in the efficiency of lineage marking and ablation confound distinctions between myogenic cell populations. *Dev. Cell*. **31**, 654–667 (2014).

47. O. Halperin-Barlev, C. Kalcheim, Sclerotome-derived Slit1 drives directional migration and differentiation of Robo2-expressing pioneer myoblasts. *Development*. **138**, 2935–2945 (2011).

48. G. Heimberg, R. Bhatnagar, H. El-Samad, M. Thomson, Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Syst*. **2**, 239–250 (2016).

49. M. Osterwalder *et al.*, Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature*. **554**, 239–243 (2018).

50. D. E. Dickel *et al.*, Ultraconserved Enhancers Are Required for Normal Development. *Cell*. **172**, 491–499.e15 (2018).

51. D. Li *et al.*, Formation of proximal and anterior limb skeleton requires early function of Irx3 and Irx5 and is negatively regulated by Shh signaling. *Dev. Cell*. **29**, 233–240 (2014).

52. K. Kraft *et al.*, Deletions, Inversions, Duplications: Engineering of Structural Variants using CRISPR/Cas in Mice. *Cell Rep.* (2015), doi:10.1016/j.celrep.2015.01.016.

53. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*. **10**, 1213–1218 (2013).

54. G. Renaud, U. Stenzel, T. Maricic, V. Wiebe, J. Kelso, deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics*. **31**, 770–772 (2015).

55. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).

56. S. Anders, P. T. Pyl, W. Huber, HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*, btu638 (2014).

57. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell developmental trajectories (2017), , doi:10.1101/110668.

58. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

59. G. X. Y. Zheng *et al.*, Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

60. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

61. J. Cao *et al.*, Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*. **357**, 661–667 (2017).

62. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods*. **14**, 979–982 (2017).

63. S. L. Wolock, R. Lopez, A. M. Klein, Scrublet: computational identification of cell doublets in single-cell transcriptomic data (2018), , doi:10.1101/357368.

64. H. Pliner *et al.*, Chromatin accessibility dynamics of myogenesis at single cell resolution (2017), , doi:10.1101/155473.

65. M. V. Kuleshov *et al.*, Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–7 (2016).

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

67. J. H. Levine *et al.*, Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*. **162**, 184–197 (2015).

68. F. A. Wolf *et al.*, Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells (2017), , doi:10.1101/208819.

69. Q. Mao, L. Wang, I. Tsang, Y. Sun, Principal Graph and Structure Learning Based on Reversed Graph Embedding. *IEEE Trans. Pattern Anal. Mach. Intell.* (2016), doi:10.1109/TPAMI.2016.2635657.

70. Q. Mao, L. Yang, L. Wang, S. Goodison, Y. Sun, in *Proceedings of the 2015 SIAM International Conference on Data Mining*, pp. 792–800.

71. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods*. **14**, 979–982 (2017).

72. P. A. P. Moran, Notes on continuous stochastic phenomena. *Biometrika*. **37**, 17–23 (1950).

73. M. K. Singh *et al.*, The T-box transcription factor Tbx15 is required for skeletal development. *Mech. Dev.* **122**, 131–144 (2005).

74. S. Paine-Saunders, B. L. Viviano, J. Zupicich, W. C. Skarnes, S. Saunders, glypican-3 controls cellular responses to Bmp4 in limb patterning and skeletal development. *Dev. Biol.* **225**, 179–187 (2000).

75. K. Hara, H. Ide, Msx1 expressing mesoderm is important for the apical ectodermal ridge (AER)-signal transfer in chick limb development. *Dev. Growth Differ.* **39**, 705–714 (1997).

76. D. G. Lupiáñez *et al.*, Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*. **161**, 1012–1025 (2015).

77. R. J. Davis *et al.*, Dach1 mutant mice bear no gross abnormalities in eye, limb, and brain development and exhibit postnatal lethality. *Mol. Cell. Biol.* **21**, 1484–1490 (2001).

78. H. Akiyama, M.-C. Chaboissier, J. F. Martin, A. Schedl, B. de Crombrugghe, The transcription factor Sox9 has essential roles in successive steps of the chondrocyte differentiation pathway and is required for expression of Sox5 and Sox6. *Genes Dev.* **16**, 2813–2828 (2002).

79. Y. Deng *et al.*, Yap1 Regulates Multiple Steps of Chondrocyte Differentiation during Skeletal Development and Bone Repair. *Cell Rep.* **14**, 2224–2237 (2016).

80. S. Joshi *et al.*, TEAD transcription factors are required for normal primary myoblast differentiation in vitro and muscle regeneration in vivo. *PLoS Genet.* **13**, e1006600 (2017).

81. D. Knapp *et al.*, Comparative transcriptional profiling of the axolotl limb identifies a tripartite regeneration-specific gene program. *PLoS One*. **8**, e61352 (2013).

82. R. Zeller, J. López-Ríos, A. Zuniga, Vertebrate limb bud development: moving towards integrative analysis of organogenesis. *Nat. Rev. Genet.* **10**, 845–858 (2009).

83. S. Nishimoto, C. Minguillon, S. Wood, M. P. O. Logan, A combination of activation and repression by a colinear Hox code controls forelimb-restricted expression of Tbx5 and reveals Hox protein specificity. *PLoS Genet.* **10**, e1004245 (2014).

84. N. Vargesson, V. Luria, I. Messina, L. Erskine, E. Laufer, Expression patterns of Slit and Robo family members during vertebrate limb development. *Mech. Dev.* **106**, 175–180 (2001).

85. J. Chimal-Monroy *et al.*, Analysis of the molecular cascade responsible for mesodermal limb chondrogenesis: Sox genes and BMP signaling. *Dev. Biol.* **257**, 292–301 (2003).

Chapter 4: characterize cell state dynamics by sci-fate

*Modified from a manuscript in preparation.

Abstract:

The beauty of development lies in the generation of diverse cell states in strictly organized temporal order. Despite of the proliferation in single cell genomic techniques, it has remained challenging to quantitatively determine cell state transition dynamics. Here we introduce sci-fate, a combinatorial indexing-based high throughput assay for profiling both whole and newly synthesised transcriptome in each of thousands of single cells. As a proof of concept, we applied sci-fate to a model system of cortisol response, and characterized over 6,000 single cell state transition events, consistent with known cell cycle dynamics upon glucocorticoid receptor activation. From the analysis, we showed the cell state transition direction and probabilities are regulated by inter-state distances and state instability landscape. The technique and computational approaches are readily applicable to other biological systems to quantitatively characterize cell state dynamics, and decipher the internal mechanism for cell fate determination.

One sentence summary:

We developed and applied sci-fate to construct >6,000 single cell whole transcriptome trajectories in a model system of cortisol response, and characterized the features regulating cell state transition dynamics.

Introduction:

Cell transits across functional and molecularly distinct state during multicellular organism development. Characterizing the cell state transition path, or cell fate, is the core in understanding development and applications such as cell engineer. While methods for single cell genomic

techniques have proliferated, they only capture a snapshot of cell state, thus cannot provide information on cell transition dynamics (*1*). Although time-lapse microscopy based single cell tracing can be used to characterize cell state transitions (*2*, *3*), they are limited in throughput and can only track the changes of several genes, and thus has low capacity to decipher complex systems.

Here we describe a novel strategy to infer quantitative cell state transition dynamics at the level of whole transcriptome. This strategy depends on a new combinatorial indexing based single cell RNA-seq technique, sci-fate. By labeling newly synthesised mRNA with 4-thiouridine (*4*, *5*) which will generate C > T point mutations during reverse transcription, sci-fate captures both whole transcriptome and newly synthesised transcriptome at single cell level, together with the degraded transcriptome information from its past state (past state memory). The past state memory of each cell is then corrected by mRNA degradation rate (memory correction technique), such that each cell can be characterized by transcriptome dynamics between two time points.

To characterize cell state transition dynamics regulated by intrinsic and extrinsic factors, we applied sci-fate to a model system of cortisol response, in which cell fate was driven by two major forces: intrinsic cell cycle program and extrinsic drug induced glucocorticoid receptor (GR) activation. GR activation influences the activity of almost every cell in the body, and regulates genes controlling development, metabolism and immune response (*6*). With sci-fate, we profiled whole transcriptome dynamics for over 6,000 single cells. Based on the similarity between past and current transcriptome states, we built thousands of cell state transition trajectories spanning five time points, which can be clustered into three types of cell fates consistent with known cell cycle progress patterns in GR activation. We further characterized cellular hidden states by functional TF modules activity, and inferred a cell transition network for cell state prediction.

Finally we showed the cell state transition direction and probability are regulated by transcriptome similarity and instability landscape of its nearby states. The theoretical, computational and experimental approaches developed here should be readily applicable to other biological systems in which cell transition dynamics are still unknown.

Results:

Overview of sci-fate

sci-fate relies on the following steps (**Fig. 1A**): (i) cells were first incubated with 4-thiouridine (S4U), a widely used thymidine analog to label newly synthesised RNA(*7–13*). (ii) Cells are harvested, fixed by 4% paraformaldehyde, followed by thiol(SH)-linked alkylation reaction which covalently attaches a carboxyami-domethyl group to S4U by nucleophilic substitution(*4*). (iii) Cells were distributed in bulk to each well of 4x96 well plates. The first RNA-seq molecular index is introduced to the mRNA of cells in each well via *in situ* reverse transcription (RT) with a poly(T) primer bearing both a well-specific barcode and a degenerate unique molecular identifier (UMI). During cDNA synthesis, the mRNA labeled with modified S4U mimic thymine-to-cytosine (T > C) conversions and result in mutated first strand cDNA. (iv) Cells from all wells are pooled and then redistributed by fluorescence-activated cell sorting (FACS) to multiple 96-well plates. Cells are gated on DAPI (4',6-diamidino-2-phenylindole) staining to discriminate single cell from doublets during sorting. Double-stranded cDNA is generated by RNA degradation and second-strand synthesis, and is subjected to transposition with Tn5. cDNA is then amplified via the polymerase chain reaction (PCR) with a combination of primers recognizing the Tn5 adaptor on the 5' end and the RT primer on the 3' end. These primers also bear a well-specific barcode that introduces the second RNA-seq molecular index. (v) Amplicons from the PCR are pooled and subjected to massively parallel sequencing. As with other "sci-" protocols(*14–21*), most nuclei pass through a unique combination of wells and therefore each cell's contents are marked by a

unique combination of barcodes that can be used to group reads that derive from the same cell. Newly synthesised mRNA out of the whole transcriptome is identified by background error corrected "T > C" conversions (**Method**).

As quality control, we first tested the technique in a mixture of HEK293T (human) and NIH/3T3 (mouse) cells under four conditions: with or without S4U labeling (200nM, 6 hrs), and with or without IAA treatment (**Fig. S1A-D**). With S4U labeling and IAA treatment (sci-fate condition), transcriptomes from human/mouse cells were overwhelmingly species-coherent (> 99% purity for both human and mouse cells, 2.6% collisions) with high ratio of T > C mutated reads detected (46% for human and 31% for mouse cells in sci-fate condition vs. 0.8% for human and 0.8% for mouse cells in no treatment condition). We obtained roughly equivalent cell purity across four conditions, albeit slightly lower UMIs detected in IAA treatment groups. Aggregated transcriptomes of sci-fate vs. normal sci-RNA-seq were highly-correlated (Spearman's correlation r = 0.99; **Fig. S1EF**), suggesting the short term labeling and conversion process have minimal effect on cell state.

Joint profiling of total and newly synthesised transcriptome in dexamethasone treated A549 cells

We then applied sci-fate to a model of cortisol response, wherein dexamethasone (DEX), a synthetic mimic of cortisol, activates glucocorticoid receptor (GR), which binds to thousands of locations across the genome, and significantly alters cell state within a short term (*22–25*). We treated lung adenocarcinoma-derived A549 cells for 0, 2, 4, 6, 8 or 10 hrs with 100nM DEX. In each condition, cells were incubated with S4U (200nM) for the last two hours before harvest for 384 x 192 well sci-fate (**Fig. 1B**). The six conditions were each represented in 64 wells during the

first round of indexing so that the treatment condition could be recovered based on the first index of each cell.

After filtering out low quality cells, potential doublets and a small subgroup of differentiated cells (**Method**), we obtained single cell profiles for 6,680 cells (median of 26,176 mRNAs detected per cell) with a median of 20% labeled UMIs per cell (**Fig. 1C, Fig. S2AB**). The intronic reads shows significantly higher newly synthesised rate than exonic reads (65% in intronic reads vs. 13% in exonic reads, p-value < 2.2e-16, Wilcoxon signed rank test; **Fig. 1D**), consistent with the expectation that the intronic reads are enriched in newly synthesised transcriptome.

We first asked if the whole transcriptome and newly synthesised transcriptome convey different information in cell state characterization. We aggregated the the whole transcriptome and newly synthesised transcriptome for each treatment conditions and checked their correlations. Different from the whole transcriptome, the newly synthesised transcriptome showed a sharp difference between no DEX treatment (0h) and treated groups (**Fig. S2C**). Consistent with this, dimension reduction with Uniform Manifold Approximation and Projection (UMAP)(*26*) on whole or newly synthesised transcriptome gives different results (**Fig. 1E**): whole transcriptome cannot separate no DEX treatment (0h) and early DEX treatment (2h) cells while newly synthesised transcriptome aggregates all DEX treated cells into a single group. Cell clusters identified by whole or newly synthesised transcriptome do not fully match with each other (**Fig. 1F, Fig. S2DE**). This is expected as the newly synthesised transcriptome directly reflects the gene promoter activity, or epigenetic response to external environment, while the whole transcriptome is mostly determined by the leftover mRNA from its past state.

To characterize cell states with joint information, we combined the top principal components (PCs) from whole and newly synthesised transcriptome for UMAP analysis. Joint information separates cells into no DEX treatment (0h), early treatment (2h) and late treatment (>2h) (**Fig. 1E)**. Interestingly, two clusters (cluster 1 and 4) characterized by whole transcriptome were split into four separate groups by joint information (**Fig. 1F)**. We checked the expression level and newly synthesis rate of cell cycle related gene markers (*27*) (**Fig. 1G, Fig. S2FG**): the newly separated clusters by joint information correspond to G2/M phase (high expression and high synthesis rate of G2/M markers) and early G0/G1 phase cells (high expression and low synthesis rate of G2/M markers). This suggests newly synthesised transcriptome convey different cell state information compared with the whole transcriptome, and joint information potentially enables higher resolution in cellular state characterization.

Characterizing functional TF modules driving cell fate determination

We next sought to characterize TF modules driving cell state transition. The links between transcription factors (TF) and their regulated genes were identified by two steps: for each gene, we computed correlations between mRNA synthesis rate during the last two hours and TF expression level across over 6,000 cells using LASSO (least absolute shrinkage and selection operator). These identified links were further filtered by either published CHIP-seq data(*28*) or motif enrichment analysis(*29*) (**Method**). In total we identified 986 links between 29 TFs and 532 genes (**Fig. 2A**), based on TF-gene covariance and validated by DNA binding data. To evaluate the possibility that the links were artifacts of regularized regression, we permuted the sample IDs of the TF expression matrix and performed the same analysis. No links were identified after this permutation.

TF modules driving GR response are identified, including known GR response effectors such as CEBPB(*30*) (**Fig. S3AB**), FOXO1(*31*), and JUNB(*32*) (**Fig. 2A**). We also found several novel GR response related TF modules including YOD1 and GTF2IRD1, with both upregulated expression and activity in DEX treated cells (**Fig. S3CD**). Main TF modules driving cell cycle progression are identified, and these include E2F1, E2F2, E2F7, BRCA1, and MYBL2 (*33*). Compared with total expression level, the new RNA synthesis rate of regulated genes by cell cycle TF modules displays higher correlation with the target TF expression (**Fig. S3E**). Additionally, we also found TF modules related with cell differentiation such as GATA3, mostly expressed in a group of quiescent population of cells (*34*), and TF modules related with oxidative stress response such as NRF1 (*35*) and NFE2L2 (NRF2) (*36*).

We next characterized TF activity by aggregating the new RNA synthesis rate of genes within each TF module, and computed the absolute correlation coefficient between each TF pairs (**Fig. S3F**). Highly correlated TF activity suggests they may function in a linked process. Hierarchical clustering segregate these 29 TF modules into five major modules (**Fig. S3F**): the first module are all cell cycle related TF modules such as E2F1 and FOXM1 (*33*), and represents the driving force for cell cycle progression. The third module are all GR response related TF modules such as FOXO1, CEBPB, JUNB and RARB(*30*)(*31*)(*32*). The other TF module groups include three TFs (KLF6, TEAD1, and YOD1) co-regulated by both cell cycle and GR response (module 2), an internal differentiation pathway including GATA3 and AR (module 3), and stress response related TFs such as NRF1 and NFE2L2 (module 5).

To identify different cell cycle states, we first ordered cells by cell cycle linked TF module activity. Cells are ordered into a smooth trajectory of cell cycle, validated by the synthesis rate of known

cell cycle markers (*27*) (**Fig. 2B**). We observed a gap between G2/M phase and G0/G1 phase, consistent with the dramatic cell state change during cell division. By unsupervised clustering, we identified nine cell cycle states spanning G0/G1, S and G2/M cell cycle phases based on cell cycle marker expression (**Fig. 2B**). Cells can be ordered into another smooth trajectory by GR reponse linked TF modules. The trajectory correlates well with DEX treatment time and dynamics of known GR activation regulated TF activity (**Fig. 2C**). By unsupervised clustering analysis, we identified three cell clusters along GR response, corresponding to no/low/high GR response state (**Fig. 2C**).

We next sought to quantitatively characterize hidden cell states in the system (**Fig. S4A**). Nine cell cycle states and three GR reponse states were identified in **Fig. 2BC**. All possible combinatorial states were identified, with the smallest group including 1.1% (74) of all cells (**Fig. 2D**). The observed cell state proportion is close to the expected proportion assuming independent assortment. This is consistent with the low correlation coefficient (Pearson's correlation r = 0.004) between the activity of these two functional TF modules across over 6,000 cells. For comparison, by dimension reduction and clustering analysis on whole and newly synthesised transcriptome, we identified 6 main clusters (**Fig. S4B**). These main clusters can be readily defined by combined groups of these 27 cell states (**Fig. 2E**).

Characterizing single cell transition trajectory and state transition network

With both whole transcriptome and newly synthesised transcriptome characterized for each cell, we can infer the single cell transcriptome state before S4U labeling (**Fig. 3A**). The recovery of past cellular transcriptome depends on two parameters: the detection rate of newly synthesised

reads in sci-fate, and the degradation rate (or half time) of each mRNA (**Method**). Both two parameters can both be estimated from the same experiment in sci-fate.

We first estimated the detection rate of sci-fate. We assume the mRNA half life is stable across different DEX treatmentment conditions. This assumption is further validated by self-consistency check later. Under this assumption, the partly degraded bulk transcriptome before the 2 hour S4U labeling should be the same between no DEX and 2 hour DEX treated cells. Thus their differences in whole transcriptome (bulk) should equal with their differences in the newly synthesised transcriptome (bulk) corrected by technique detection rate. As whole and newly synthesised transcriptome are both profiled in our experiment, we can directly compute the detection rate of sci-fate. The differences in newly synthesised mRNA correlates well with the differences in mRNA expression level (Pearson's r = 0.93, **Fig. S5A**), suggesting the new RNA detection rate is rather stable across genes. We thus used the median of new RNA capture rate (82%) for downstream analysis.

We next computed the mRNA degradation rate in 2 hours. As A549 cell population can be regarded stable without external perturbation, for cells after 2 hour DEX treatment, its past state (before 2 hour S4U labeling) should be the same with the 0 hour DEX treated cells. Similarly, the past state (before S4U labeling) for T = 0/2/4/6/8/10 hour DEX treated cells should be similar to the profiled T = 0/0/2/4/6/8 hour cells. With whole transcriptome and newly synthesised transcriptome profiled for all treatment conditions, mRNA degradation rate across thousands of genes in each 2 hour time interval can be estimated. As a self-consistency check mentioned above, the gene degradation rates are highly correlated across different DEX treatment time (**Fig. S5B**). We then used the averaged gene degradation rate for downstream analysis. With both new mRNA

detection rate and gene degradation rate available, we estimated single cell past transcriptome state so that each cell can be characterized by transcriptome dynamics in a two-hour interval.

To recover cell state dynamics for a longer interval (i.e. 10 hours), we developed a cell linkage pipeline to link parent and child cells in the same cell state transition trajectory (**Fig. 3A**): for each cell A (e.g. 2 hour DEX treated cells), we identified a cell B profiled in the earlier time point (e.g. no DEX treated cells) and B had its current state similar with A's past state, based on a recently developed alignment strategy to identify common cell states between two data sets (*27*). B can be regarded as the parent state of A. Similarly, we also identified another cell C profiled in the later time point (e.g. 4 hour DEX treated cells) and C had its past state similar with A's current state. Cell C can be regarded as the A's future state. By extending the same strategy to all past and future state identified for each cell, we constructed 6,680 single cell transition trajectory across 10 hours and five time points (**Fig. 3AB**). Of note, this analysis is based on an assumption that the past and current state of each cell (except cells at the start and end time points) are comprehensively detected, which holds true in our data sets as over 6,000 cells are profiled (over 1,000 cells per condition), or a cell for less than one min during cell cycle. Multiple cells (>50) are profiled at each cell state, thus stochastic cell state transition process can also be captured.

To validate the result, we applied dimension reduction and unsupervised clustering analysis to these 6,680 single cell trajectories, which grouped into three trajectory clusters. We checked the dynamics of cell states characterized in **Fig. 3C**. As expected, all three trajectories showed cell state transition from no GR response to low/high GR response states over time (**Fig. 3D**). We observed distinct cell cycle dynamics across these three trajectories(**Fig. 3D**): trajectory 1 showed decreased G2/M phase and consistently increased G0/G1 phase, and represented cell state

transition from G2/M and G1 intermediate states to G1 phase. Trajectory 2 showed cell state transition from S and G2/M intermediate states to G2/M phase. In the trajectory 3, we observed cell state transition from G1 and S intermediate phase to early S phase during early DEX treatment (0-2 hour), but the transition is inhibited in late DEX treatment conditions (>2 hour DEX treatment), suggesting long term DEX treatment results in G1 phase arrest. This is consistent with cell state proportion changes along treatment time and previous research (*37*, *38*)(**Fig. 3D**). These suggests the single cell transition paths characterized by sci-fate can recover general cell state transition directions.

With multiple cells (>70) profiled at each state, we computed the cell state transition probability across all 27 hidden states. Cell state transitions with low transition probabilities (< 0.1) are potentially due to rare events or noise, and thus filtered out. The cell state transition network can be defined by 27 cell states as nodes, and links showing the potential transition paths (**Fig. 3E**). The direction of cell cycle progression is readily characterized by at least three transition stages with irreversible transition directions along cell cycle (**Fig. 3E**). In late G1 phase and late G2/M phase, we also found several states showing reversible transitions dynamics, which potentially reflect two cell cycle checkpoints in G1/S and G2/M phases(*33*). As expected, cells on similar cell cycle but different GR responses states showed dramatically different transition dynamics, and cells with high GR response state tend to be arrested in G1 or G2/M phase.

As a consistency check to validate whether the cell state transition network captures cell state transition dynamics, we evaluated if the transition probabilities can recover the real cell state distributions across different time points. Indeed, although cell state proportions are dynamically changed across 10 hours (**Fig. 3F**), the state transition network accurate predicts the 27 cell state

ratios across all five later time points from cell state proportion in 0 hour DEX treated cells (**Fig. 3G, Fig. S6A**). We also computed the cell state transition network with only part of the data (0 hour to 6 hour), which gave highly correlated transition probabilities with the full data, and accurately predict cell states at 10 hours (**Fig. 3H, Fig. S6B**).

Characterizing factors regulating cell state transition directions

To characterize the factors regulating cell state transition probability, we first calculated cell state distance, by the pearson's distance of aggregated transcriptome (whole and newly synthesised) between each state pairs. As expected, cell state transition probability negatively correlates with transition distance (Spearman's correlation coefficient = -0.38, **Fig. 4A**). We also computed state instability, defined by the proportion of cells moving out of the state within two hours (**Fig. 4B**). The state instability landscape matches well with cell transition directions (**Fig. 4B**): states in no GR response show higher instability compared with high GR response states. In high GR response states, cells at early G1 phase has the lowest instability, while cells at G1/S intermediate states showed a high unstable peak, consistent with the G1 phase arrest in late DEX treatment.

The cell state proportion changes after 10 hours correlates well with cell state instability (Spearman's correlation coefficient = -0.88, **Fig. 4C**), suggesting cell state dynamics are regulated by the cell state instability landscape. The state instability also correlates well with state transition probability entropy, which reflects the diversity of state transition targets (Pearson's correlation r = 0.73, **Fig. 4D**). To validate whether the inter-state transition probability can be inferred by nearby state instability, we fit nearby state instability and distance into a neural network model, to predict state transition probability from each state to the other states. Combining both nearby state

instability and distances achieved more than ten folds higher performance in predicting inter-state transition probability, compared with using state distances alone (median cross validated r squared is 0.58 by using both information vs. 0.046 by using state distance only, p-value = 4.5e-10, two sided wilcoxon rank sum test, **Fig. 4E**), suggesting the cell state transition directions and probabilities are regulated by nearby state stability landscape. And cells prefer to moving to a more stable nearby state over just the nearest position.

Discussion

Here we developed the first strategy to characterize cell state transition dynamics on whole transcriptome level. The strategy depends on sci-fate, a novel combinatorial indexing based high throughput single cell RNA-seq technique, capable of profiling both whole and newly synthesised transcriptome in thousands of cells. Similar with other "sci-" techniques, sci-fate is readily scaled up to millions of cells(*39*), and potentially compatible with profiling both transcriptome and epigenome(*40*). This enables sci-fate to characterize cell state dynamics in a much complexed system (i.e. whole embryo development) where the real cell transition path to hundreds of cell types are still unknown. We further developed a computation pipeline to estimate newly synthesised RNA capture rate and gene degradation rate from sci-fate data (memory correction), and infer thousands of differential trajectories for each single cell, linked by shared past and current transcriptome state at each time point.

To validate the techniques and examine how cell state dynamic are regulated by internal and external factors, we applied the strategy to a model system of cortisol response, in which cell fate were dynamically regulated by internal cell cycle and extrinsic drug induced GR activation. We showed the newly synthesised transcriptome directly links to the epigenome response to

environmental stimuli, and joint analysis of both whole and newly synthesised transcriptome enables higher resolution in cell state separation. By co-variance between TF expression and new RNA synthesis rate across thousands of cells, we identified up to one thousand links between TFs and regulated genes, validated by DNA binding data. We further identified 27 "hidden cell states" characterized by the combinatorial state of functional TF modules in cell cycle progression and GR response, compared with only 6 states by conventional clustering analysis.

By memory correction and cell linkage analysis, we built over 6,000 single cell transition trajectories spanning 10 hours, with the main trajectories consistent with known cell state dynamics in cell cycle and GR response. Cell state transition network are characterized by the transition probability across all cell states, validated by the recovery of 27 cell state dynamics across all five time points. Finally, we found the cell state transition probabilities are regulated by two key features of cell state transition network: inter-state distance and state instability landscape, both of which can be potentially estimated by conventional single cell RNA-seq techniques.

While powerful, this strategy has several limitations. First, to faithfully build single cell trajectory, we need comprehensive cell state characterization at each time point. Also multiple observations for each states are needed to robustly estimate the transition probability. These limitations can be readily resolved by the combinatorial strategy of sci-fate, which is capable of profiling millions of cells in a single experiment. Another caveat is that most S4U labeling experiments are applied to in vitro systems. However, recent research has shown that S4U can stably label cell type specific RNA transcription in multiple mouse tissues (i.e. brain, intestine and adipose tissue)(*41*, *42*), suggesting sci-fate, with further optimizations to enhance S4U incorporation and detection rate, can be applied to profile in vivo single cell transcriptome dynamics.

sci-fate opens a new avenue for applying “static” single cell genomic techniques to characterizing dynamic systems. Compared with traditional imaging based techniques, sci-fate profiles cell state dynamics at whole transcriptome level, and enables comprehensive cell state characterization without marker selection and discovery of key driving force in cell differentiation. Finally, we anticipate that sci-fate can be readily combined with alternative lineage tracing techniques(*43–45*), to decode the detailed cell state transition dynamics to every final cell state within hundreds of developmental lineages.

Main figures

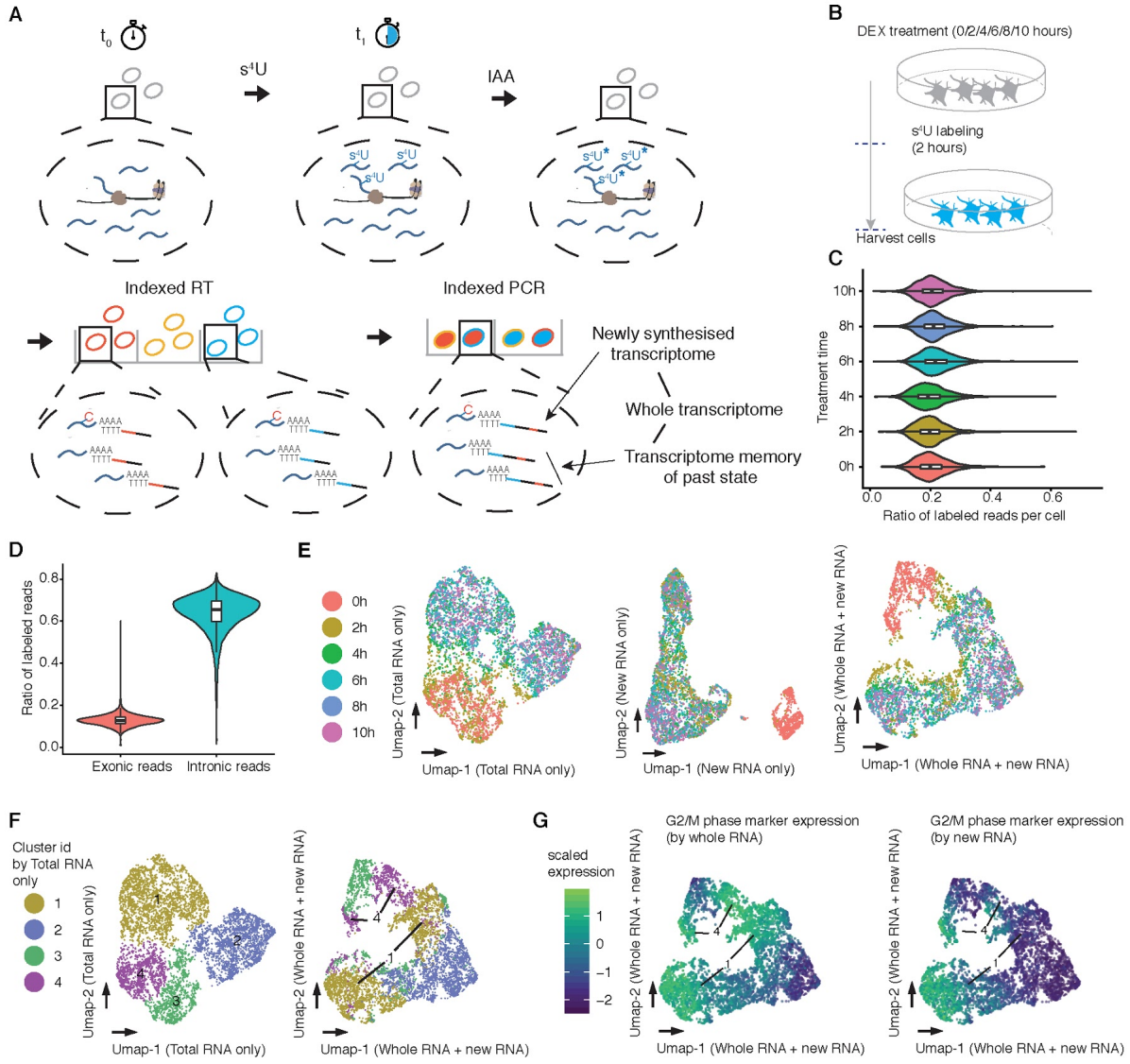


Fig. 1. Joint profiling of total and newly synthesised transcriptome by sci-fate. (**A**) sci-fate workflow with key steps outlined in text. (**B**) Experiment scheme. A549 cells were treated with dexamethasone time dependently. Cells from all treatment conditions were labeled with S4U two hours before harvest for sci-fate. (**C**) Violin plot showing the ratio of S4U labeled reads per cell in six treatment time. (**D**) Violin plot showing the ratio of S4U labeled reads in exonic and intronic reads. For all box plots: thick horizontal lines, medians; upper and lower box edges, first and third quartiles, respectively; whiskers, 1.5 times the interquartile range; circles, outliers. (**E**) UMAP visualization of A549 cells by whole transcriptome (left), newly synthesised transcriptome (middle)

and both (right). (**F**) Similar with (**E**), colored by cluster id identified by whole transcriptome. (**G**) UMAP visualization of A549 cells by joint information, colored by normalized expression of G2/M marker genes by RNA level (left) and newly synthesised RNA level (right). UMI counts for these genes are scaled by library size, log-transformed, aggregated and then mapped to Z-scores.

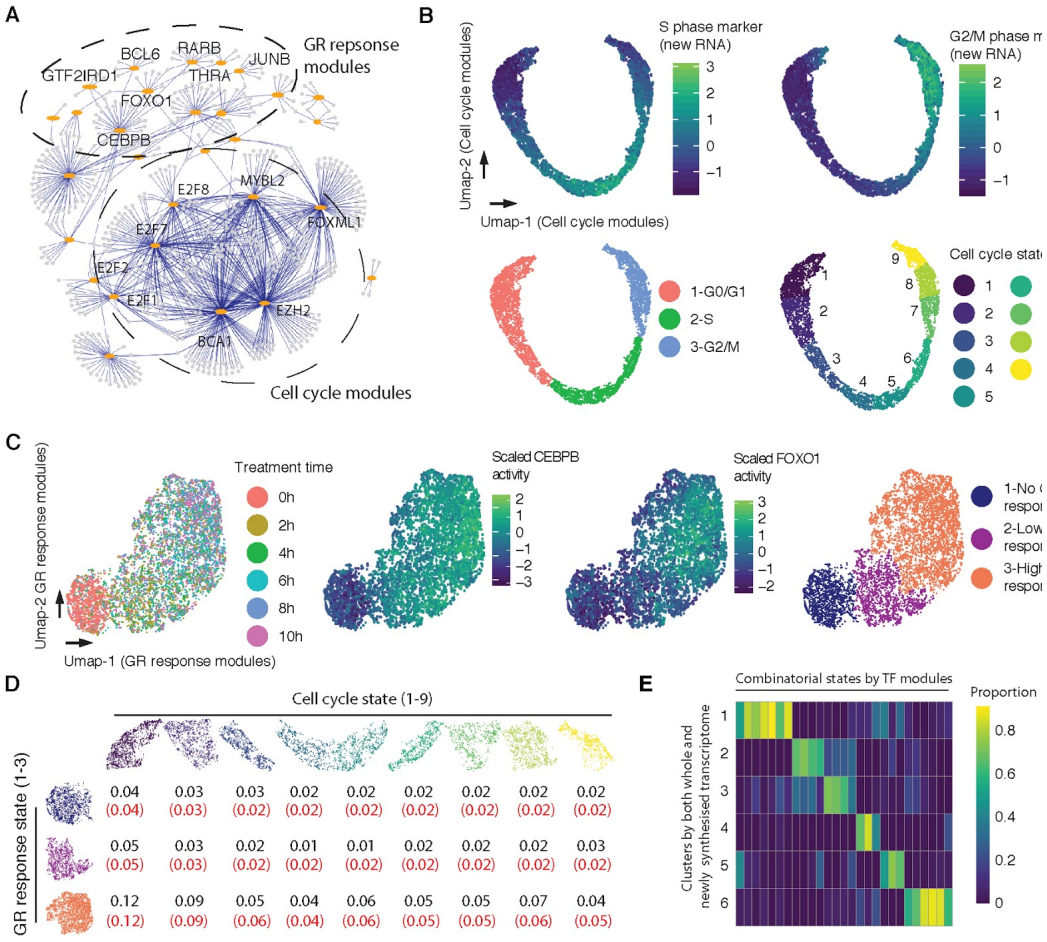


Fig. 2. Characterizing TF modules driving cell state transition. (**A**) Identified links (blue) between transcription factors (orange) and regulated genes (grey). TF modules related with cell cycle progression or GR response are labeled. (**B**) UMAP visualization of A549 cells ordered by cell cycle TF modules, colored by newly synthesised mRNA of S phase and G2/M phase markers (top), three cell cycle phases (bottom left), and nine cell cycle states by unsupervised clustering analysis (bottom right). (**C**) UMAP visualization of A549 cells ordered by GR reponse TF modules,

colored by DEX treatment time (left), CEBPB and FOXO1 activity (middle) and cluster id from unsupervised clustering analysis (right). To calculate TF activity, newly synthesised UMI counts for these genes are scaled by library size, log-transformed, aggregated and then mapped to Z-scores. (**D**) A table showing the observed ratio (black) of cell state by the combinatorial state of cell cycle modules (x axis) and GR response modules (y axis). The red number is the expected ratio assuming independent assortment. (**E**) Heatmap showing the proportion of cell states defined by the combinatorial states of TF modules in each main clusters identified by clustering analysis based on joint whole and newly synthesised transcriptome.

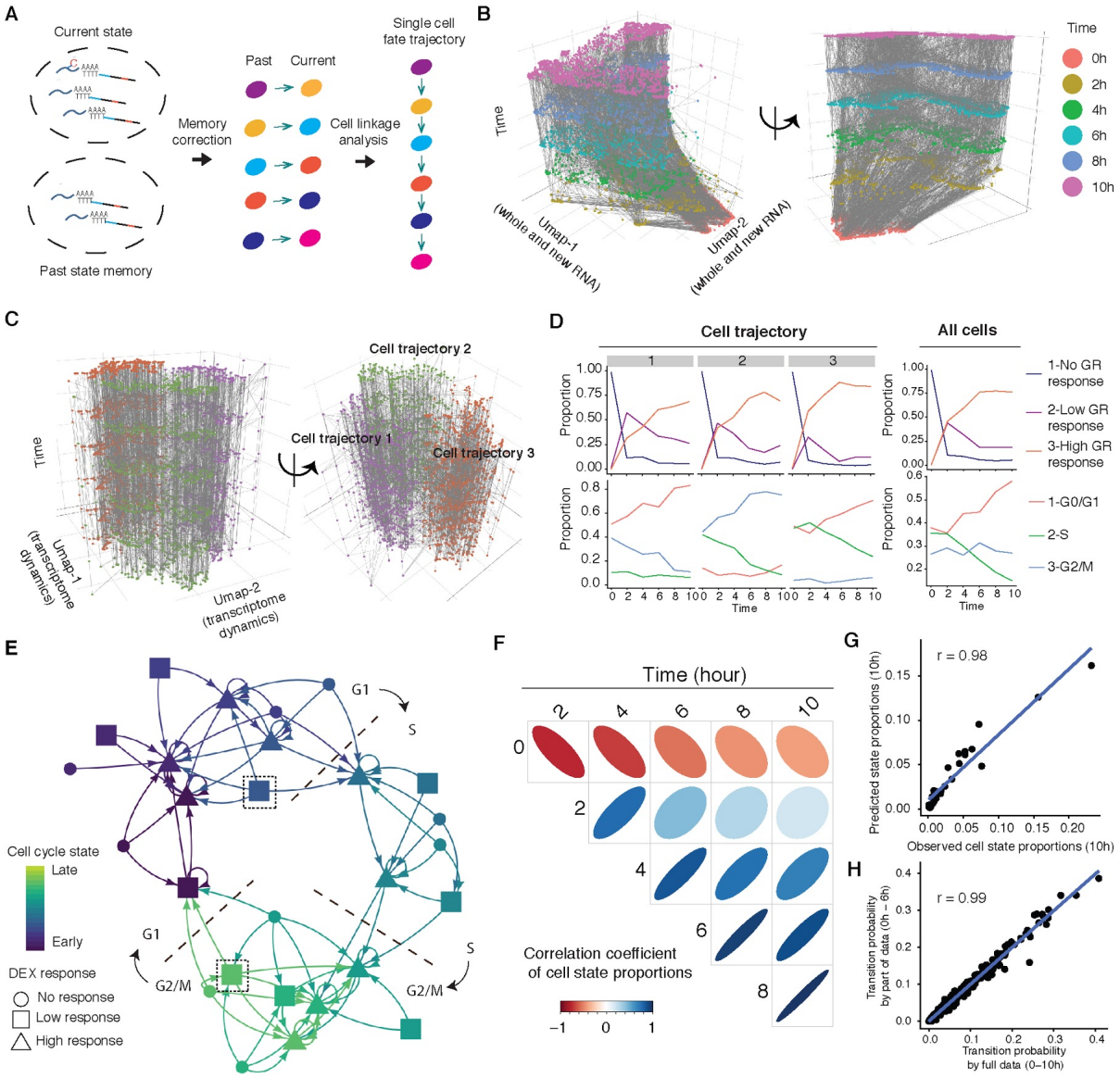


Fig. 3. Characterizing >6,000 single cell state transition trajectories. (**A**) Scheme showing memory correction and cell linkage analysis to construct single cell transition trajectory with details outlined in text and method. (**B**) 3D plot of cells colored by DEX treatment time (also as z coordinates). The x and y coordinates correspond to the UMAP space by whole and newly synthesised transcriptome in **Fig. 1E** (left). Linked parent and child cells are connected with grey lines. (**C**) Similar with (**B**), except the x and y coordinates correspond to the UMAP space by single cell transcriptome dynamics across six time points. (**D**) Line plots showing the cell state dynamics

of different GR reponse states (top) and cell cycle states (bottom) in each cell trajectory clusters (left) or all cells (right) independent of cell linkage analysis. (**E**) Cell state transition network. The nodes are 27 cell states characterized in **Fig. 2D** and the links are identified transition paths between cell states. Links with low transition probabilities (< 0.1) are filtered out. Squares with dashed lines showing the example states with reversible transition dynamics. (**F**) Correlation plot showing the correlation of cell state proportions between treatment conditions. Positive correlations are displayed in blue and negative correlations in red color. The shape of the ellipse are correlated with the correlation coefficients (on the ellipse). (**G**) Scatter plot showing the correlation of cell state proportions between observed 10 hour DEX treatment groups and predicted cell state proportions. The prediction is based on the cell state transition probabilities and cell state proportion in no DEX treatment group. The blue line represent the linear regression line. (**H**) Scatter plot showing the correlation of cell state transition probabilities calculated by full data (0-10 hours) or part data (0-6 hours), along with the linear regression line.

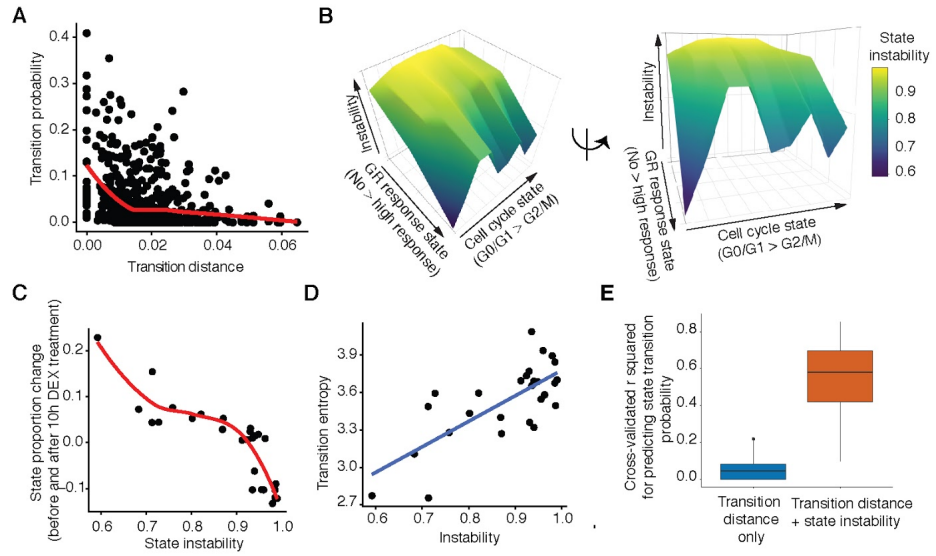


Fig. 4. Cell state transition probabilities are regulated by nearby state stability landscape. (**A**) Scatter plot showing the relationship between transition distance (Pearson's distance) and transition probability between cell states, together with the red LOESS smooth line by ggplot2 (*46*). (**B**) 3D plot showing the instability landscape of cell states. X-axis represents GR response states (from no to low to high response state). Y-axis represent the cell cycle states ordered from G0/G1 to G2/M states. Z-axis represent the cell state instability, defined by the probability of cells within each cell state jumping to other states after 2 hours. (**C**) Scatter plot showing the relationship between cell state instability and cell proportion change before and after 10 hour DEX treatment, together with the red LOESS smooth line by ggplot2 (*46*). (**D**) Scatter plot showing the correlation between state instability and state transition entropy with the linear regression line (blue). (**E**) Box plot showing the cross-validated r squared for predicting inter-state transition probability by transition distance only or combining transition distance and state instability landscape by densely connected neural network.

Materials and Methods:

Mammalian cell culture

All mammalian cells were cultured at 37°C with 5% CO_2, and were maintained in high glucose DMEM (Gibco cat. no. 11965) for HEK293T and NIH/3T3 cells or DMEM/F12 medium for A549 cells, both supplemented with 10% FBS and 1X Pen/Strep (Gibco cat. no. 15140122; 100U/ml penicillin, 100 µg/ml streptomycin). Cells were trypsinized with 0.25% typsin-EDTA (Gibco cat. no. 25200-056) and split 1:10 three times per week.

Sample processing for sci-fate

A549 cells were treated with 100 nM DEX for 0 hrs, 2 hrs, 4 hrs, 6 hrs, 8 hrs and 10 hrs. Cells in all treatment conditions were incubated with 200uM S4U for the last two hours before cell harvest. For HEK293T and NIH/3T3 cells, cells were incubated with 200uM S4U for 6 hours before cell harvest.

All cell lines (A549, HEK293T and NIH/3T3 cells) were trypsinized, spun down at 300x**g** for 5 min (4°C) and washed once in 1X ice-cold PBS. All cells were fixed with 4ml ice cold 4% paraformaldehyde (EMS) for 15 min on ice. After fixation, cells were pelleted at 500x**g** for 3 min (4°C) and washed once with 1ml PBSR (1 x PBS, pH 7.4, 1% BSA, 1% SuperRnaseIn, 1% 10mM DTT). After wash, cells were resuspended in PBSR at 10 million cells per ml, and flash frozen and stored in liquid nitrogen. Paraformaldehyde fixed cells were thawed on 37 degree water bath, spun down at 500x**g** for 5 min, and incubated with 500ul PBSR including 0.2% Triton X-100 for 3min on ice. Cells were pelleted and resuspended in 500ul nuclease free water including 1% SuperRnaseIn. 3ml 0.1N HCl were added into the cells for 5min incubation on ice (*21*). 3.5ml Tris-HCl (pH = 8.0) and 35ul 10% Triton X-100 were added into cells to neutralize HCl. Cells were pelleted and washed with 1ml PBSR. Cells were resuspended in 100ul PBSR. 100ul PBSR with fixed cells were incubated with mixture including 40ul Iodoacetamide (IAA, 100mM), 40ul sodium phosphate buffer (500mM, pH = 8.0), 200ul DMSO and 20ul H2O, at 50°C for 15min. The reaction was quenched by 8ul DTT (1M) and 8.5ml PBS(*47*). Cells were pelleted and resuspended in 100ul PBSI (1 x PBS, pH 7.4, 1% BSA, 1% SuperRnaseIn). For all later washes, nuclei were pelleted by centrifugation at 500x**g** for 5 min (4°C).

The following steps are similar with sci-RNA-seq protocol with paraformaldehyde fixed nuclei (*15*, *16*). Briefly, cells were distributed into four 96-well plates. For each well, 5,000 nuclei (2 μL) were mixed with 1 μl of 25 μM anchored oligo-dT primer (5′-ACGACGCTCTTCCGATCTNNNNNNNN[10bp

index]TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3′, where “N” is any base and “V” is either “A”, “C” or “G”; IDT) and 0.25 µL 10 mM dNTP mix (Thermo), denatured at 55°C for 5 min and immediately placed on ice. 1.75 µL of first-strand reaction mix, containing 1 µL 5X Superscript IV First-Strand Buffer (Invitrogen), 0.25 µl 100 mM DTT (Invitrogen), 0.25 µl SuperScript IV reverse transcriptase (200 U/µl, Invitrogen), 0.25 µL RNaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen), was then added to each well. Reverse transcription was carried out by incubating plates at the following temperature gradient: 4°C 2 minutes, 10°C 2 minutes, 20°C 2 minutes, 30°C 2 minutes, 40°C 2 minutes, 50°C 2 minutes and 55°C 10 minutes. All cells (or nuclei) were then pooled, stained with 4',6-diamidino-2-phenylindole (DAPI, Invitrogen) at a final concentration of 3 µM, and sorted at 25 nuclei per well into 5 µL EB buffer. Cells were gated based on DAPI stain such that singlets were discriminated from doublets and sorted into each well. 0.66 µl mRNA Second Strand Synthesis buffer (NEB) and 0.34 µl mRNA Second Strand Synthesis enzyme (NEB) were then added to each well, and second strand synthesis was carried out at 16°C for 180 min. Each well was then mixed with 5 µL Nextera TD buffer (Illumina) and 1 µL i7 only TDE1 enzyme (25 nM, Illumina, diluted in Nextera TD buffer), and then incubated at 55°C for 5 min to carry out tagmentation. The reaction was stopped by adding 10 µL DNA binding buffer (Zymo) and incubating at room temperature for 5 min. Each well was then purified using 30 uL AMPure XP beads (Beckman Coulter), eluted in 16 µL of buffer EB (Qiagen), then transferred to a fresh multi-well plate.

For PCR reactions, each well was mixed with 2µL of 10 µM P5 primer (5′-AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′; IDT), 2 µL of 10 µM P7 primer (5′-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3′; IDT), and 20 µL NEBNext High-Fidelity 2X PCR Master Mix (NEB). Amplification was carried out using the

following program: 72°C for 5 min, 98°C for 30 sec, 18-22 cycles of (98°C for 10 sec, 66°C for 30 sec, 72°C for 1 min) and a final 72°C for 5 min. After PCR, samples were pooled and purified using 0.8 volumes of AMPure XP beads. Library concentrations were determined by Qubit (Invitrogen) and the libraries were visualized by electrophoresis on a 6% TBE-PAGE gel. Libraries were sequenced on the NextSeq 500 platform (Illumina) using a V2 150 cycle kit (Read 1: 18 cycles, Read 2: 130 cycles, Index 1: 10 cycles, Index 2: 10 cycles).

Read alignments and downstream processing

Read alignment and gene count matrix generation for the single cell RNA-seq was performed using the pipeline that we developed for sci-RNA-seq (*48*) with minor modifications. Reads were first mapped to a reference genome with STAR/v2.5.2b (*49*), with gene annotations from GENCODE V19 for human, and GENCODE VM11 for mouse. For experiments with HEK293T and NIH/3T3 cells, we used an index combining chromosomes from both human (hg19) and mouse (mm10). For the A549 experiment, we used human genome build hg19.

The single cell sam files were first converted into alignment tsv file using sam2tsv function in jvarkit(*50*). Next, for each single cell alignment file, mutations matching the background SNPs were filtered out. For background SNP reference of A549 cells, we downloaded the paired-end bulk RNA-seq data for A549 cells from ENCODE (*28*) (sampled name: ENCFF542FVG, ENCFF538ZTA, ENCFF214JEZ, ENCFF629LOL, ENCFF149CJD, ENCFF006WNO, ENCFF828WTU, ENCFF380VGD). Each paired-end fastq files were first adaptor-clipped using trim_galore/0.4.1(*51*) with default settings, aligned to human hg19 genome build with STAR/v2.5.2b (*49*). Unmapped and multiple mapped reads were removed by samtools/v1.3 (*52*). Duplicated reads were filtered out by MarkDuplicates function in picard/1.105(*53*). De-duplicated reads from all samples were combined and sorted with samtools/v1.3 (*52*). Background SNPs were called by mpileup function in samtools/v1.3(*52*) and mpileup2snp function in VarScan/2.3.9(*54*).

For HEK293T and NIH/3T3 test experiment, background SNP reference was generated in a similar pipeline above, with the aggregated single cell sam data from control condition (no S4U labeling and no IAA treatment condition).

For each single cell alignment file, all mutations with quality score <= 13 were removed. Mutations at the both ends of each reads were mostly due to sequencing errors, and thus also got filtered out. For each read, we checked if there are T > C mutations (for sense strand) or A > G mutations (for antisense strand), and labeled these mutated reads as newly synthesised reads.

Each cell was characterized by two digital gene expression matrixes from the full sequencing data and newly synthesised RNA data as described above. Genes with expression in equal or less than 5 cells were filtered out. Cells with fewer than 2000 UMIs or more than 80,000 UMIs were discarded. Cells with doublet score > 0.2 by doublet analysis pipeline Scrublet/0.2(*55*) were removed.

The dimensionality of the data was first reduced with PCA (after selecting the top 2,000 genes with highest variance) on digital gene expression matrixes on either full gene expression data or the newly synthesised gene expression data by Monocle 3 (*56*, *57*). The top 10 PCs were selected for dimension reduction analysis with uniform manifold approximation and projection (UMAP/0.3.2), a recently proposed algorithm based on Riemannian geometry and algebraic topology to perform dimension reduction and data visualization (*26*). For joint analysis, we combined top 10 PCs calculated on the whole transcriptome and top 10 PCs on the newly synthesised transcriptome for each single cell before dimension reduction with UMAP. Cell clusters were done via densityPeak algorithm implemented in Monocle 3 (*56*, *57*). We first performed UMAP analysis on joint information of all processed cells, and identified an outlier cluster (724 out of 7,404 cells). These cells were marked by high level expression of GATA3, a marker of differentiated cells (*34*), and were filtered out before downstream analysis.

Analysis for linking transcription factor (TF) to regulated genes

We aimed to identify links between TFs and regulated genes based on their covariance. Cells with more than 10,000 UMI detected, and genes with newly synthesis reads detected in more than 10% of all cells were selected. The full gene expression and newly synthesised gene count per cell were normalized by cell-specific library size factors computed on the full gene expression matrix by estimateSizeFactors in Monocle 3 (*56*, *57*), log transformed, centered, then scaled by scale() function in R. For each gene detected, a LASSO regression model was constructed with package glmnet (*58*) to predict the normalized expression levels, based on the normalized expression of 853 TFs annotated in the "motifAnnotations_hgnc" data from package RcisTarget(*29*), by fitting the following model:

$$G_i = \beta_0 + \beta_t T_i$$

where G_i is the adjusted gene expression value for gene i. It is calculated by the newly synthesised mRNA count for each cell, normalized by cell specific size factor (SG_i) estimate by estimateSizeFactors in Monocle 3 (*56*, *57*) on the full expression matrix of each cell, and log transformed:

$$G_i = ln(\frac{g_i}{SG_i} + 0.1)$$

To simplify downstream comparison between genes, we standardize the response G_i prior to fitting the model for each gene *i* with the scale() function in R.

Similar with G_i, T_i is the adjusted TF expression value for each cell. It is calculated by the full TF expression count for each cell, normalized by cell specific size factor (SG_i) estimate by

estimateSizeFactors in Monocle 3 (*56*, *57*) on the full expression matrix of each cell, and log transformed:

$$T_i = ln(\frac{t_i}{SG_i} + 0.1)$$

Prior to fitting, T_i are are standardized with the scale() function in R.

Our approach aims to TFs that may regulate each gene, by finding the subset that can be used to predict its expression in a regression model. However, a TF with expression correlated with a gene's expression does not guarantee it is regulating that gene: if gene A is specifically expressed in cell state 1 and TF B is specifically expressed in cell type 2. Although negative correlations between a TF's expression and a gene's newly synthesis rate could reflect the activity of a transcriptional repressor, we felt that the more likely explanation for negative links reported by glmnet was mutually exclusive patterns of cell-state specific expression and TF activity. Thus during prediction, we excluded TFs with negative correlated expression with the gene's synthesis rate and also low correlation coefficient (<= 0.03) links. We identified a total of 6,103 links between TFs and regulated genes.

To identify putative direct-binding targets,, we intersected the links with TFs profiled in ENCODE Chip-seq experiment(*28*). Out of 1,086 links with TFs characterized in ENCODE, 807 links were validated by TF binding sites near gene promoters (*59*), a 4.3 folds enrichment in odd ratio (number of validated links over non-validated links) compared with background (odd ratio = 2.89 in links identified in LASSO regression vs. 0.67 in background, p-value < 2.2e-16, Fisher's Exact test). Only gene sets with significantly enrichment of the correct TF Chip-seq binding sites are retained (Fish's Exact test, False discovery rate of 5%), and pruned to remove indirect target genes without TF binding data support. 591 links were retained in this approach.

To expand the validated TF-gene links, we further applied package SCENIC(*29*), a pipeline to construct gene regulatory networks based on the enrichment of target TF motifs around genes' promoters (10kb). Each co-expression module identified by LASSO regression was analyzed using cis-regulatory motif analysis using RcisTarget(*29*). Only modules with significant motif enrichment of the correct TF regulator were retained, and pruned to remove indirect target genes without motif support. We filtered the TF-gene links by three correlation coefficient threshold (0.3, 0.4 and 0.5), and combined all links validated by RcisTarget(*29*). In total, there were 509 links validated by motif analysis approach. Combining both approaches, we identified a total 986 TF-gene regulatory links by the covariance between TF expression and gene synthesis rate, validated by DNA binding data or motif analysis. To evaluate the possibility that the links were artifacts of regularized regression, we permuted the sample IDs of the TF expression matrix and performed the same analysis. No links were identified after this permutation.

Ordering cells by functional TF modules

To calculate TF activity in each cell, newly synthesised UMI counts for genes within the target TF module were scaled by library size, log-transformed, aggregated and then mapped to Z-scores. As TFs with highly correlated or anti-correlated activity suggest they may function in linked biological process, we calculated the absolute Pearson's correlation coefficient between each pair of TF activity, and based on this we clustered TFs by ward.d2 clustering method in package pheatmap/1.0.12(*60*). Five functional TF modules were identified and annotated based on their functions.

To characterize cell states on the dimension of each functional TF modules, cells were ordered by the activity of cell cycle related TFs (TF module 1) or GR response related TFs (TF module 3) with UMAP (metric = "cosine", n_neighbors = 30, min_dist = 0.01). The cell cycle progression

trajectory were validated by cell cycle gene markers in Seurat/2.3.4(*27*). Three cell cycle phases were identified by densityPeak algorithm implemented in Monocle 3 (*56*, *57*), on the UMAP coordinates ordered by cell cycle TF modules. As each main cell cycle phase still showed variable TF activity and cell cycle marker expression, we segmented each phase to early/middle/late states by k-means clustering (k = 3), and recovered a total of nine cell cycle states. Three GR reponse states were identified by densityPeak algorithm implemented in Monocle 3 (*56*, *57*).

Past transcriptome state recovery from sci-fate

To identify the past transcriptome state (the cell state before S4U labeling), we assume the mRNA half life is stable across different DEX treatmentment conditions. This assumption is further validated by self-consistency check later. Under this assumption, the partly degraded bulk transcriptome before the 2 hour S4U labeling should be the same between no DEX and 2 hour DEX treated cells. Thus their differences in whole transcriptome (bulk) should equal with their differences in the newly synthesised transcriptome (bulk) corrected by technique detection rate:

$$A_{0h} / S_{0h} - (N_{0h} / S_{0h}) /\alpha = A_{2h} / S_{2h} - (N_{2h} / S_{2h}) /\alpha$$

A_{0h} is the aggregated UMI count for all cells in no DEX treatment group; S_{0h} is the library size (total UMI count of cells) at no DEX treatment; N_{0h} is the aggregated newly synthesised UMI count for all cells in no DEX treatment group; A_{2h} is the aggregated UMI count for all cells in 2 hour DEX treatment group; S_{2h} is the library size (total UMI count of cells) in 2 hour DEX treatment group; N_{2h} is the aggregated newly synthesised UMI count for all cells in 2 hour DEX treatment group; α is the detection rate for sci-fate. In theory, one detection rate can be calculated for each gene. However, for genes with minor differences of newly synthesis rate between two

conditions, the estimated α is dominated by noise. We thus selected genes showing higher differences in normalized newly synthesis rate between two conditions: we first tested a series of threshold for gene filtering and calculated the α for each gene. We then plotted the relationship between threshold and the ratio of genes with out-range α values (< 0 or > 1). We selected the threshold that was at the knee point of the plot with 186 genes selected. The differences in newly synthesised mRNA of these genes highly correlates with the differences in mRNA expression level (Pearson's r = 0.93, **Fig. S4A**), suggesting the new RNA detection rate is rather stable across genes. There is a median of 82% newly synthesised RNA captured by sci-fate.

We next computed the mRNA degradation rate across each 2 hours. As A549 cell population can be regarded stable without external perturbation, for 2 hour DEX treated cells, its past state (before 2 hour S4U labeling) should be the same with the 0 hour DEX treated cells. Similarly, the past state (before S4U labeling) for T = 0/2/4/6/8/10 hour DEX treated cells should be similar to the profiled T = 0/0/2/4/6/8 hour cells:

$$A_{t1} / S_{t1} - (N_{t1} / S_{t1}) / \alpha = A_{t0} / S_{t0} * \beta$$

A_{t1} is the aggregated UMI count for all cells in t1; S_{t1} is is the library size (the total UMI count of cells) at t1; N_{t1} is the aggregated newly synthesised UMI count for all cells at t1; α is the estimated detection rate of sci-fate; A_{t0} is the aggregated UMI count for all cells in t0; S_{t0} is is the library size (the total UMI count of cells) at t0; β is 1 - gene specific degradation rate between t0 and t1, and is related with the mRNA half life γ by:

$$\beta = 1 - (1 / 2)^{(t1 - t0) / \gamma}$$

The gene degradation rate β can be calculated on each 2 hour interval of DEX treatment. As a self-consistency check mentioned above, the gene degradation rates are highly correlated across different DEX treatment time (**Fig. S4B**). We then used the averaged gene degradation rate for downstream analysis.

With the detection rate and gene degradation rate estimated, the past transcriptome state of each cell can be estimated by:

$$a_{t1} - n_{t1} / \alpha = a_{t0} * \beta$$

a_{t1} is the single cell UMI count in t1; n_{t1} is the single cell newly synthesised UMI count at t1; α is the estimated detection rate of sci-fate; $\beta$is 1 - gene specific degradation rate between t0 and t1. a_{t0} is the estimated single cell UMI count in a past time point t0, with all negative values converted to 0.

Linkage analysis to build single cell state trajectory

By linkage analysis, we aim to identify linked parent and child cells in the same cell trajectory. Technically, for cells at t1, we combines their past state transcriptome state (before S4U labeling, 2 hours before t1 in our experiment) as one group 1, and the full transcriptome state of t0 (2 hours before t1) as another group 2. Assuming there is no apparent cell apoptosis, these two groups should have similar cell state distribution. We applied a manifold alignment strategy to identify common cell states between two data sets, based on common sources of variation(*27*). This analysis is based on another assumption that the past and current state of each cell (except cells at the start and end time points) are comprehensively detected, which holds true in our data sets as over 6,000 cells are profiled (over 1,000 cells per condition), or a cell for less than one min during cell cycle. As a result of the pipeline, cell states from t0 and past cell states from t1 are aligned in the same UMAP space. Violation of the assumptions above can be detected by outliers during

alignment of the two data sets. For each cell A in t1, we selected its nearest neighbour in t0 as its parent state in the alignment UMAP space. Similarly, for each cell in t0, we selected its nearest neighbour in t1 as its child cell state. Of note, the link is not necessary to be bi-directional: the parent state of one cell may be linked to a different child cell. As the parent state and child state was identified for each cell (except the cells at 0 hour and 10 hour), we then identified the linked parent cell of each cell's parent, and similarly the linked child cell of each cell's child. Thus each single cell can be characterized by a single cell state transition path across all five time points spanning 10 hours. As multiple cells (>50) are profiled at each cell state, stochastic cell state transition process can also be captured.

Dimension reduction and clustering analysis for single cell transcriptome dynamics

For dimension reduction on single cell transcriptome dynamics, top 5 PCs for full transcriptome and top 5 PCs for newly synthesised transcriptome were selected for each state, and combined in temporal order along single cell state trajectory for UMAP analysis. Main cell trajectory types were identified by density peak clustering algorithm(*61*).

With cell state proportion at the beginning time point (0 hour treatment) and cell state transition probabilities estimated from the data, we first predicted the cell state distribution after 2 hours, assuming the cell state transition process in DEX treatment is a cell-autonomous, time-independent, Markovian dynamics. Similarly, the cells state distribution at later time point can be calculated based on the predicted cell state distribution 2 hours before.

Inter-state transition probability prediction by state instability

Cell state instability is defined as the probability of each state moving to other states after 2 hours. To calculate cell state distance, we first sampled equal number (n = 50) of cells at each state, and aggregated the full transcriptome and newly synthesised transcriptome of all cells within the state.

Each cell state can be defined by the joint information combining the whole and newly synthesised transcriptome. The cell state distance is calculated as the Pearson's correlation coefficient of the joint information between two states.

To predict inter-state transition probability, we constructed a 3 layer neural network (units number: 128, 128, 26 with relu activation at each layer; loss function: cosine_proximity, batch size: 128, epochs: 80) with Keras/2.2.4(*62*). For input, we used state instability of current state, the normalized state instability of the other 26 states (scaled by the instability of current state), and transition distance (squared) from current state to the other 26 states (in the same order of states in state instability vector). To avoid over-fitting, we permuted the state orders in state instability 200 times for each input, while still keeping the state order of state transition distance the same with the state instability. To evaluate the model performance, we apply leave-one-out validation by training the model on 26 states, and validate the model on the left state on predicting the state transition probabilities to all the other 26 states. For predicting the inter-state probability with state transition distance only, the same model is used for training and validation with all input state instabilities replaced with 1.

Supplementary Figures

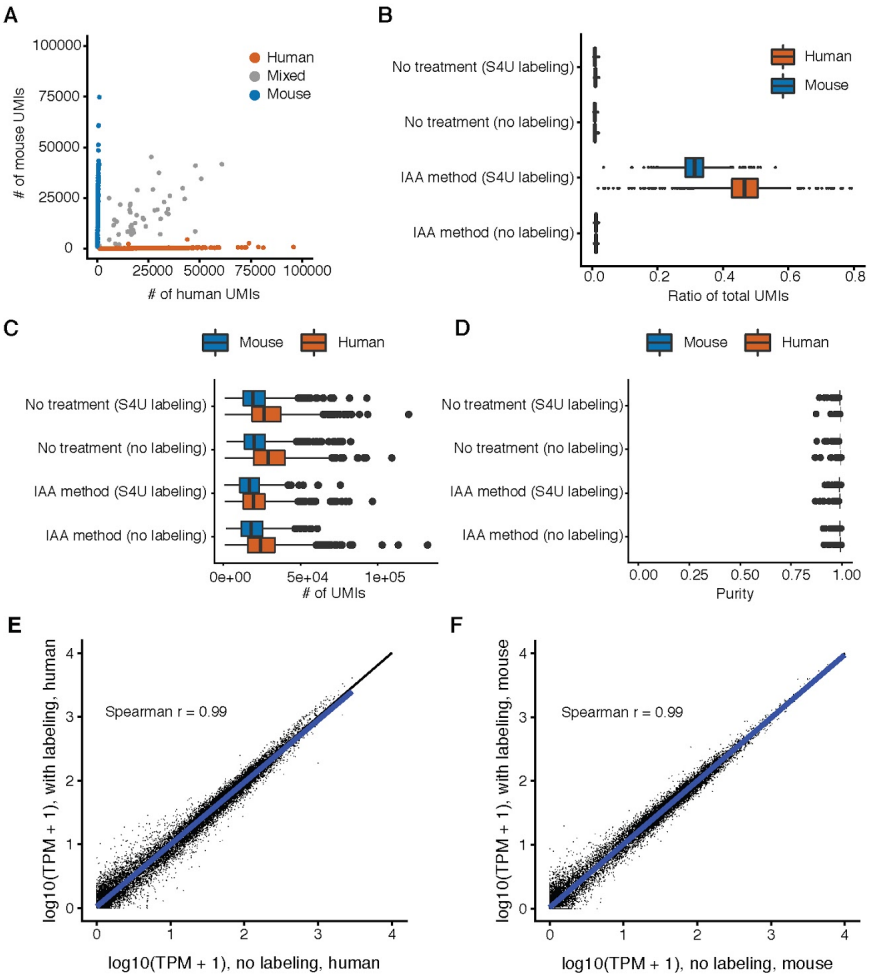


Fig. S1. Performance and QC-related analyses for sci-fate. (**A**) Scatter plot of mouse (NIH/3T3) vs. human (HEK293T) UMI counts per cell in the condition of sci-fate. (**B-D**) Boxplot showing the ratio of S4U labeled reads, number of UMIs, and purity (proportion of reads mapping to the expected species) per cell from HEK293T (cell number n = 932) and NIH/3T3 cells (cell number n = 438). For all box plots: thick horizontal lines, medians; upper and lower box edges, first and third quartiles, respectively; whiskers, 1.5 times the interquartile range; circles, outliers. (**E-F**) Correlation (Spearman's correlation) between gene expression measurements in aggregated profiles of HEK293T (**E**) and NIH/3T3 cells (**F**) from sci-fate (y axis) vs. sci-RNA-seq cells (x axis).

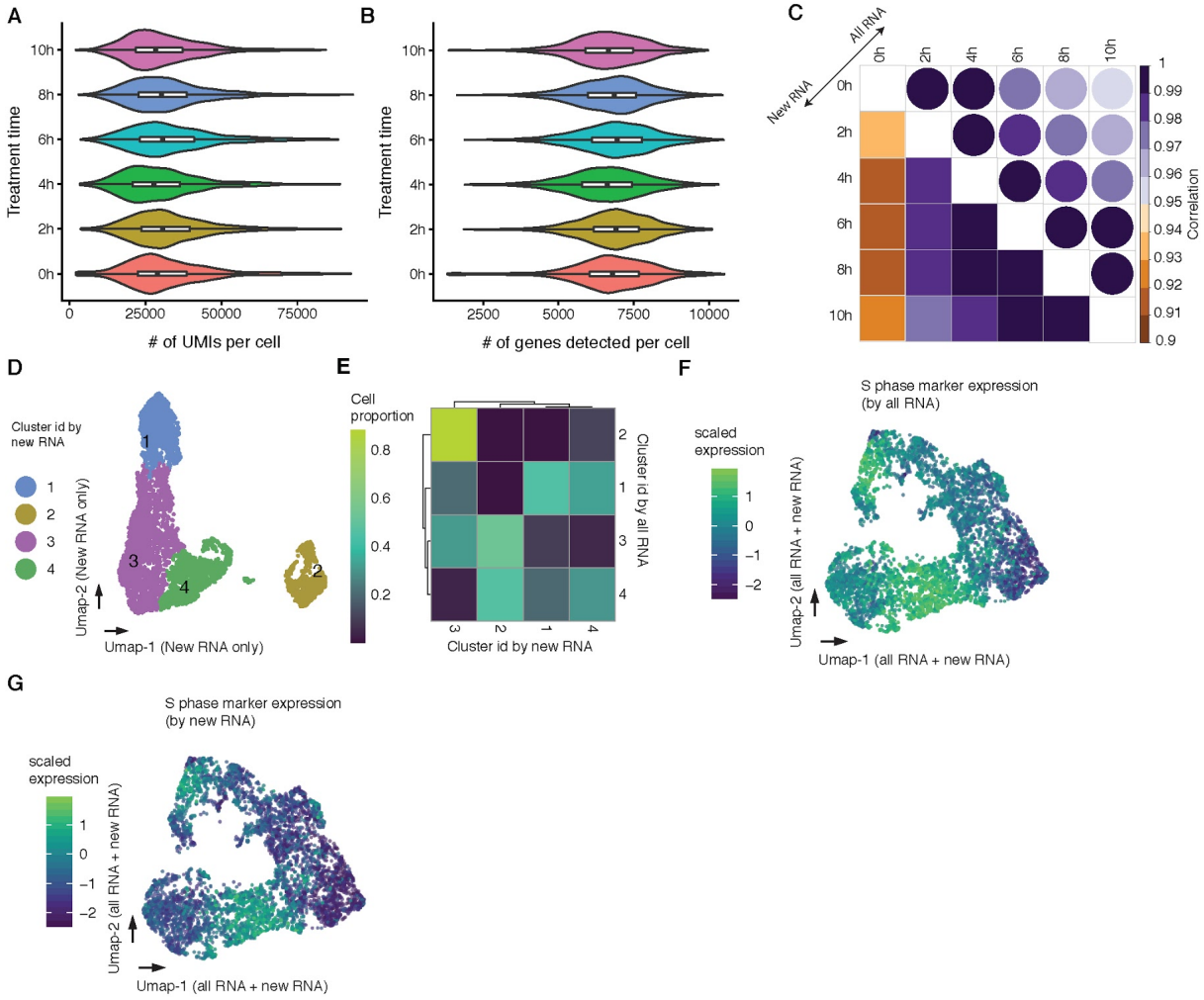


Fig. S2. Performance of sci-fate on dexamethasone treated A549 cells. (**A, B**) Violin plot showing the number of UMIs (**A**) and genes (**B**) per cell in six treatment conditions. For all box plots: thick horizontal lines, medians; upper and lower box edges, first and third quartiles, respectively; whiskers, 1.5 times the interquartile range; circles, outliers. (**C**) Correlation plot showing the Pearson correlation coefficient between different treatment conditions for aggregated whole transcriptome (top right) and newly synthesised transcriptome (down left). (**D**) UMAP visualization of A549 cells by newly synthesised transcriptome, colored by cluster id identified by newly synthesised transcriptome. (**E**) heatmap showing the proportion of cells from each clusters defined by whole transcriptome, falling into each cell cluster by newly synthesised transcriptome.

(**F-G**) UMAP visualization of A549 cells by both total and newly synthesised transcriptome, colored by normalized expression of S phase marker genes by total RNA expression (**F**) and newly synthesised RNA (**G**). UMI counts for these genes are scaled for library size, log-transformed, aggregated and then mapped to Z-scores.

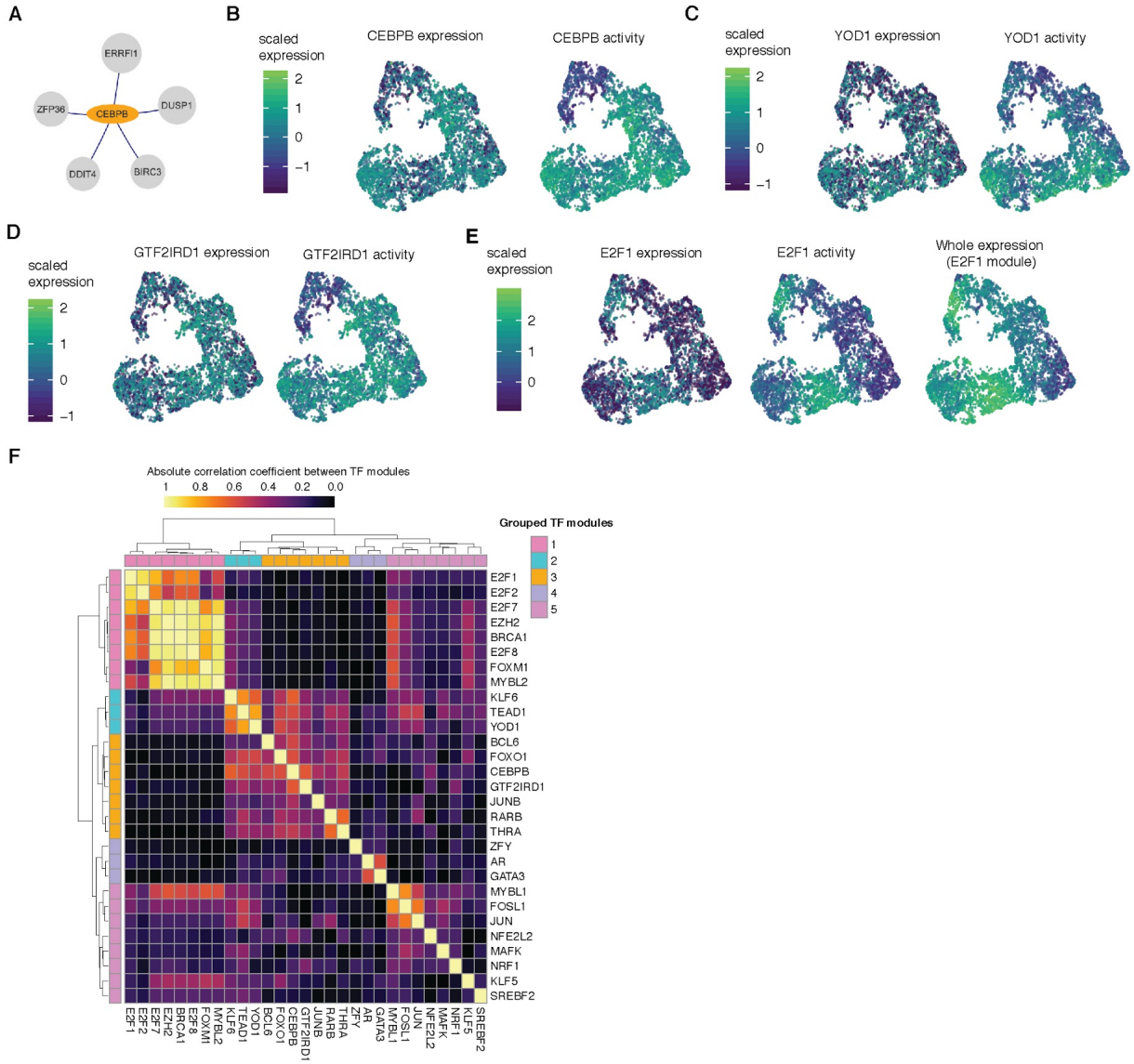


Fig. S3. TF modules driving cell state transition in DEX treated A549 cells. (**A**) Identified gene targets (grey) of CEBPB (orange). Only links with regularized correlation coefficient from

LASSO > 0.6 are shown. (**B**) UMAP visualization of A549 cells by whole and newly synthesised transcriptome colored by CEBPB expression (left) and activity (right). (**C**) similar with (**B**), colored by the YOD1 expression (left), and YOD1 activity (right). (**D**) similar with (**B**), colored by the GTF2IRD1 expression (left), and GTF2IRD1 activity (right). (**E**) similar with (**B**), colored by the E2F1 expression (left), E2F1 activity (middle) and aggregated expression of whole transcriptome for E2F1 linked genes (right). (**F**) Heatmap showing the absolute value of Pearson's correlation coefficient between TF modules. 29 TF modules were grouped into five groups by hierarchical clustering analysis.

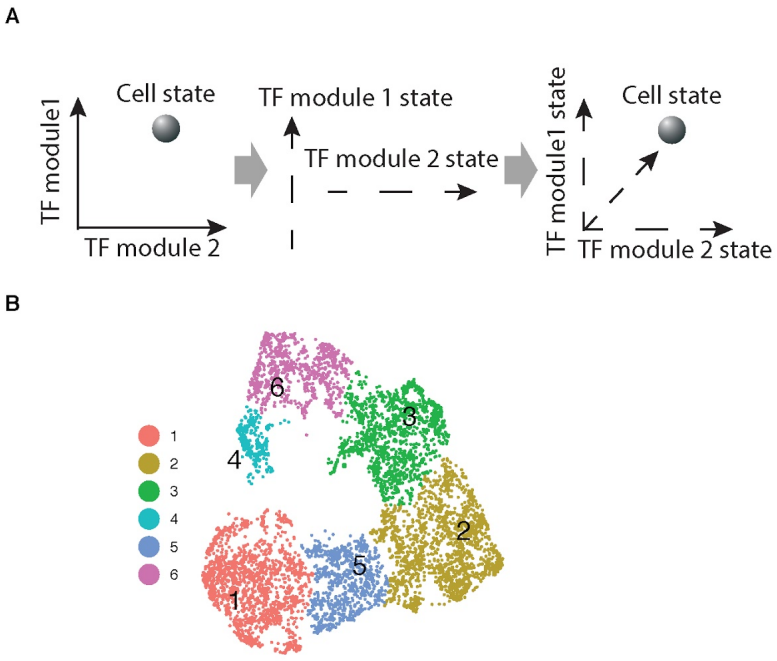


Fig. S4. cell states are characterized by combinatorial states of functional TF modules. (**A**) Scheme showing the strategy for characterizing cell states by combinatorial states of functional TF modules. (**B**) Umap visualization of all cells by both whole and newly synthesised transcriptome, colored with main cluster id identified by density peak clustering algorithm on the UMAP space.

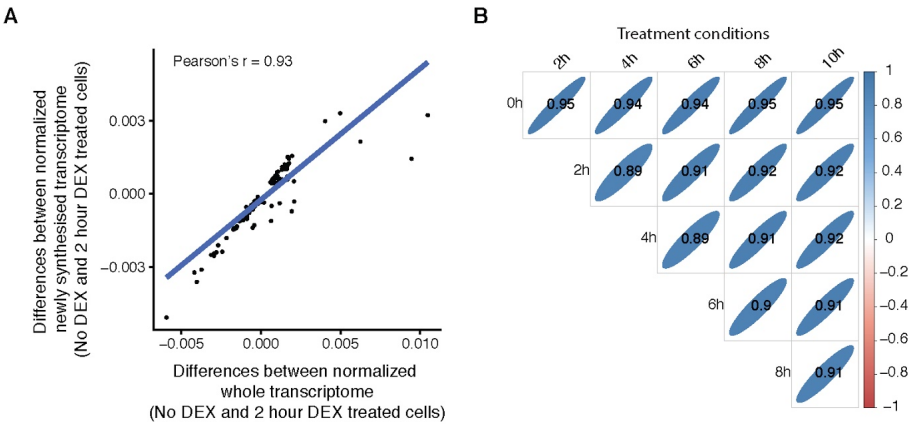


Fig. S5. New RNA detection rate and RNA degradation rate estimation. (**A**) scatter plot showing the correlation between x axis: differences of normalized whole transcriptome between no DEX and 2 hour DEX treated cells, and y axis: differences of normalized newly synthesised transcriptome between no DEX and 2 hour DEX treated cells. Blue line is the linear regression line. Both whole transcriptome and newly synthesised transcriptome of each time point are normalized by the library size of whole transcriptome of the time point. (**B**) Correlation plot showing the correlation of estimated gene degradation rate between treatment conditions. Positive correlations are displayed in blue and negative correlations in red color. The shape of the ellipse are correlated with the correlation coefficients (on the ellipse).

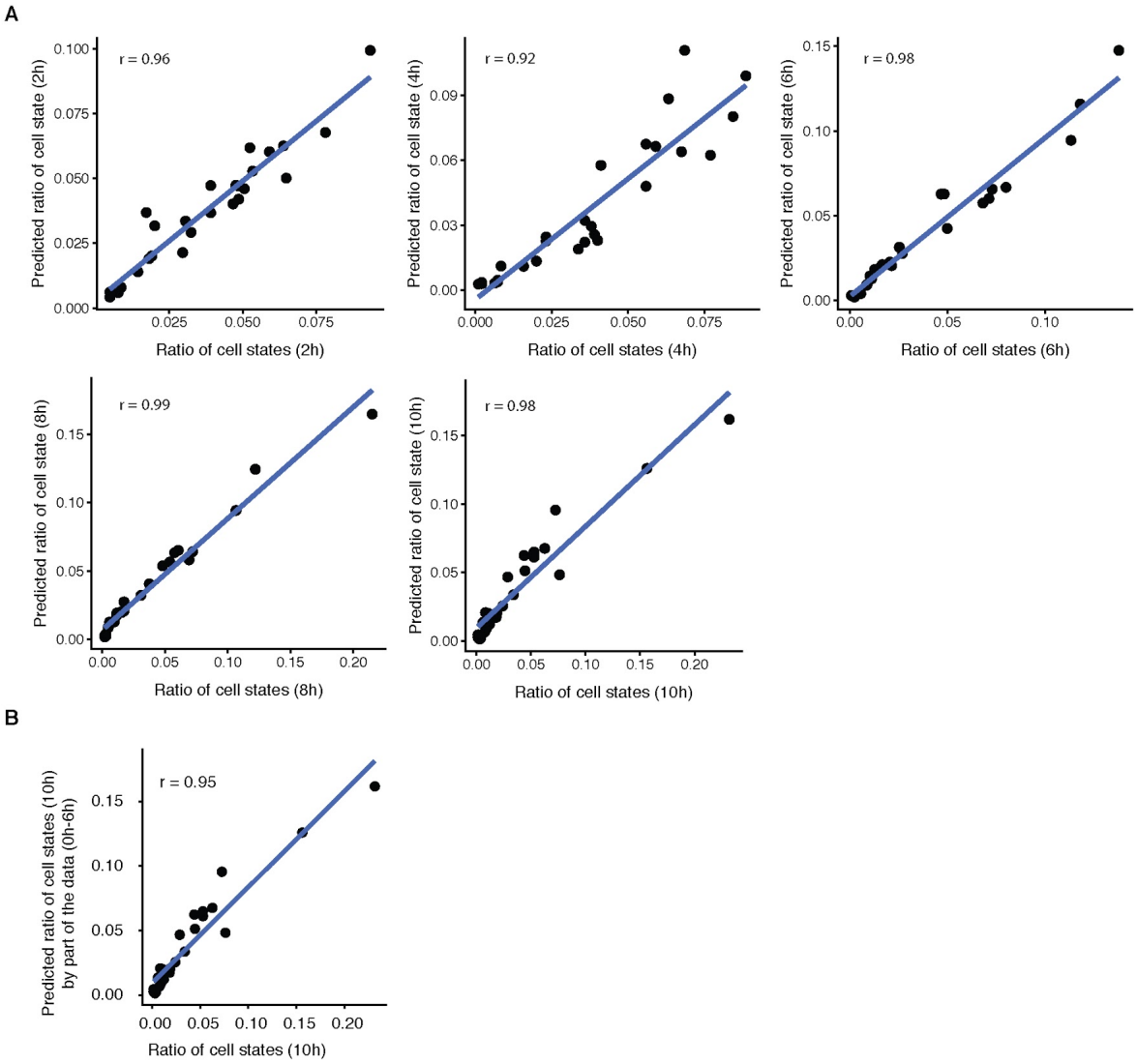


Fig. S6. cell state transition network for cell state prediction. (**A**) Scatter plot showing the correlation between observed cell states at each treatment time and predicted cell state by cell state transition probabilities and cell state proportion in no DEX treatment group. The blue line represent linear regression line. (**B**) Scatter plot showing the correlation of cell state proportions between observed 10 hour DEX treatment groups and predicted values. The predicted values is based on cell state transition probabilities estimated by part data (0-6 hours) and cell state proportion in no DEX treatment group. The blue line represent the linear regression line.

Reference of chapter 4:

1. N. Moris, C. Pina, A. M. Arias, Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* **17**, 693–703 (2016).

2. A. Filipczyk *et al.*, Network plasticity of pluripotency transcription factors in embryonic stem cells. *Nat. Cell Biol.* **17**, 1235–1246 (2015).

3. S. Hormoz *et al.*, Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements. *Cell Syst*. **3**, 419–433.e8 (2016).

4. V. A. Herzog *et al.*, Thiol-linked alkylation of RNA to assess expression dynamics. *Nat. Methods*. **14**, 1198–1204 (2017).

5. J. A. Schofield, E. E. Duffy, L. Kiefer, M. C. Sullivan, M. D. Simon, TimeLapse-seq: adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nat. Methods*. **15**, 221–225 (2018).

6. J. C. Buckingham, Glucocorticoids: exemplars of multi-tasking. *Br. J. Pharmacol.* **147**, S258 (2006).

7. M. D. Cleary, C. D. Meiering, E. Jan, R. Guymon, J. C. Boothroyd, Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nat. Biotechnol.* **23**, 232–237 (2005).

8. L. Dolken *et al.*, High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA*. **14**, 1959–1972 (2008).

9. C. Miller *et al.*, Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol. Syst. Biol.* **7**, 458–458 (2014).

10. E. E. Duffy *et al.*, Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry. *Mol. Cell*. **59**, 858–866 (2015).

11. B. Schwalb *et al.*, TT-seq maps the human transient transcriptome. *Science*. **352**, 1225–1228 (2016).

12. M. Rabani *et al.*, Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* **29**, 436–442 (2011).

13. M. R. Miller, K. J. Robinson, M. D. Cleary, C. Q. Doe, TU-tagging: cell type–specific RNA isolation from intact complex tissues. *Nat. Methods*. **6**, 439–441 (2009).

14. D. A. Cusanovich *et al.*, Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*. **348**, 910–914 (2015).

15. J. Cao *et al.*, Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*. **357**, 661–667 (2017).

16. J. Cao *et al.*, Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*. **361**, 1380–1385 (2018).

17. V. Ramani *et al.*, Massively multiplex single-cell Hi-C (2016), , doi:10.1101/065052.

18. R. M. Mulqueen *et al.*, Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018).

19. S. A. Vitak *et al.*, Sequencing thousands of single-cell genomes with combinatorial indexing.

Nat. Methods. **14**, 302–308 (2017).

20. Y. Yin *et al.*, High-throughput mapping of meiotic crossover and chromosome mis-segregation events in interspecific hybrid mice (2018), , doi:10.1101/338053.

21. A. B. Rosenberg *et al.*, Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*. **360**, 176–182 (2018).

22. T. E. Reddy *et al.*, Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Res.* **19**, 2163–2171 (2009).

23. S. John *et al.*, Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268 (2011).

24. T. E. Reddy, J. Gertz, G. E. Crawford, M. J. Garabedian, R. M. Myers, The Hypersensitive Glucocorticoid Response Specifically Regulates Period 1 and Expression of Circadian Genes. *Mol. Cell. Biol.* **32**, 3756–3767 (2012).

25. C. M. Vockley *et al.*, Direct GR Binding Sites Potentiate Clusters of TF Binding across the Human Genome. *Cell*. **166**, 1269–1281.e19 (2016).

26. L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*. **3**, 861 (2018).

27. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

28. The ENCODE Project Consortium, The ENCODE (ENCyclopedia Of DNA Elements)

Project. *Science*. **306**, 636–640 (2004).

29. S. Aibar *et al.*, SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods*. **14**, 1083–1086 (2017).

30. M. Boruk, J. G. A. Savory, R. J. G. Haché, AF-2-Dependent Potentiation of CCAAT Enhancer Binding Proteinβ -Mediated Transcriptional Activation by Glucocorticoid Receptor. *Mol. Endocrinol.* **12**, 1749–1763 (1998).

31. W. Qin *et al.*, Identification of functional glucocorticoid response elements in the mouse FoxO1 promoter. *Biochem. Biophys. Res. Commun.* **450**, 979–983 (2014).

32. C. S. Sheela Rani, N. Elango, S.-S. Wang, K. Kobayashi, R. Strong, Identification of an Activator Protein-1-Like Sequence as the Glucocorticoid Response Element in the Rat Tyrosine Hydroxylase Gene. *Mol. Pharmacol.* **75**, 589 (2009).

33. M. Fischer, G. A. Müller, Cell cycle transcription control: DREAM/MuvB and RB-E2F complexes. *Crit. Rev. Biochem. Mol. Biol.* **52**, 638–662 (2017).

34. J. Chou, S. Provot, Z. Werb, GATA3 in development and cancer differentiation: cells GATA have it! *J. Cell. Physiol.* **222**, 42–49 (2010).

35. J. Y. C. Madhurima Biswas, Role of Nrf1 in antioxidant response element-mediated gene expression and beyond. *Toxicol. Appl. Pharmacol.* **244**, 16 (2010).

36. I.-G. Ryoo, M.-K. Kwak, Regulatory crosstalk between the oxidative stress-related transcription factor Nfe2l2/Nrf2 and mitochondria. *Toxicol. Appl. Pharmacol.* **359**, 24–33 (2018).

37. J. M. Harmon, M. R. Norman, B. J. Fowlkes, E. B. Thompson, Dexamethasone induces irreversible G1 arrest and death of a human lymphoid cell line. *J. Cell. Physiol.* **98**, 267–278 (1979).

38. A. K. Greenberg *et al.*, Glucocorticoids inhibit lung cancer cell growth through both the extracellular signal-related kinase pathway and cell cycle regulators. *Am. J. Respir. Cell Mol. Biol.* **27**, 320–328 (2002).

39. J. Cao *et al.*, Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*. **357**, 661–667 (2017).

40. J. Cao *et al.*, Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*. **361**, 1380–1385 (2018).

41. W. Matsushima *et al.*, SLAM-ITseq: sequencing cell type-specific transcriptomes without cell sorting. *Development*. **145** (2018), doi:10.1242/dev.164640.

42. U. Sharma *et al.*, Small RNAs are trafficked from the epididymis to developing mammalian sperm (2017), , doi:10.1101/194522.

43. A. McKenna *et al.*, Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*. **353**, aaf7907 (2016).

44. B. Raj *et al.*, Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).

45. K. L. Frieda *et al.*, Synthetic recording and in situ readout of lineage information in single cells. *Nature*. **541**, 107–111 (2017).

46. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).

47. M. Muhar *et al.*, SLAM-seq defines direct gene-regulatory functions of the BRD4-MYC axis. *Science*. **360**, 800–805 (2018).

48. J. Cao *et al.*, Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*. **357**, 661–667 (2017).

49. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).

50. P. Lindenbaum, JVarkit: java-based utilities for Bioinformatics. *figshare* (2015).

51. FelixKrueger, FelixKrueger/TrimGalore. *GitHub*, (available at https://github.com/FelixKrueger/TrimGalore).

52. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

53. Picard Tools - By Broad Institute, (available at http://broadinstitute.github.io/picard/).

54. D. C. Koboldt *et al.*, VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).

55. S. L. Wolock, R. Lopez, A. M. Klein, Scrublet: computational identification of cell doublets in single-cell transcriptomic data (2018), , doi:10.1101/357368.

56. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods*. **14**, 979–982 (2017).

57. cole-trapnell-lab, cole-trapnell-lab/monocle-release. *GitHub*, (available at https://github.com/cole-trapnell-lab/monocle-release).

58. J. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33** (2010), doi:10.18637/jss.v033.i01.

59. Dataset - ENCODE Transcription Factor Binding Site Profiles, (available at http://amp.pharm.mssm.edu/Harmonizome/dataset/ENCODE+Transcription+Factor+Binding+Site+Profiles).

60. raivokolde, raivokolde/pheatmap. *GitHub*, (available at https://github.com/raivokolde/pheatmap).

61. A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks. *Science*. **344**, 1492–1496 (2014).

62. keras-team, keras-team/keras. *GitHub*, (available at https://github.com/keras-team/keras).

Chapter 5: Conclusion and perspectives

In this thesis, I described the development of four novel single cell genomic techniques as well as the analysis framework to characterize cell state diversity and cell fate dynamics in whole organism development.

Technique development remains the driving force for new biological discoveries. Here I showed the development of the first combinatorial indexing based single cell RNA-seq technique (sci-RNA-seq, Chapter 1), to profile tens of thousands of cell states by whole transcriptome, and after thousands of optimization conditions tested, the first single cell RNA-seq technique which enables profiling of over 2 million cells in a single experiment (sci-RNA-seq3, Chapter 3). To associate transcriptome and epigenome for cell state characterization, I developed a new high-throughput single cell assay (sci-CAR, Chapter 2) to profile both chromatin accessibility and gene expression across thousands of cells in a single experiment. To link cell state with temporal information, I developed a novel combinatorial indexing based high-throughput single cell RNA-seq (sci-fate, Chapter 4), to assay both whole and newly synthesized transcriptome to recover single cell transcriptome dynamics across different time points.

Accompany these techniques, the thesis also described new computation pipelines and analysis frameworks for cell state and fate analysis. These include downstream sequencing processing scripts for all techniques described above, as well as computation strategies to associate different molecular layers in multi-modal scRNA-seq analysis. For example, I developed a regression-based strategy to link hundreds of distal cis-regulatory elements and their target genes by covariance between single cell chromatin accessibility and gene expression across thousands of cells (Chapter 2), and a novel strategy to link transcription factor and their regulated genes by

both DNA binding data and covariance between TF expression and gene synthesis rate from sci-fate (Chapter 4). In chapter 3, I tested and compared different algorithms for processing extra-large single cell data set with millions of cells, and described a full analysis pipeline from data filtering, pre-processing, dimension reduction, clustering, to trajectory inference and gene marker identification, for characterizing hundreds of cell states as well as tens of development trajectories in a single experiment (Chapter 3).

More importantly, these molecular and computation techniques are developed to answer key biological questions: What is cell state? Traditionally, cell state, or cell type, are defined by the level of pre-selected gene markers and/or specific spatial locations. In this thesis, I showed major and rare cell types of whole organism can be comprehensively identified based on single cell transcriptome, by profiling nearly 50,000 cells from the nematode *Caenorhabditis elegans* at the L2 stage, which is over 50-fold "shotgun cellular coverage" of its somatic cell composition (Chapter 1). Furthermore, cell state can be determined by multiple molecular layers. For example, I characterized major cell types of whole mouse kidney, by profiling > 10,000 single cell chromatin accessibility and transcriptome, both showing highly correlated results in separating cell states (Chapter 2).

The second part of this thesis focus on characterizing cell fate dynamics in development. Commonly, cell fate is regarded as a smooth and continuous process, represented by Waddington's epigenetic landscape. Based on this assumption, I applied a novel high throughput single cell RNA-seq technique (sci-RNA-seq3) to profile over 2 million cells spanning the main mammalian organogenesis stages. With this data, I characterized > 500 cell states, delineated and annotated 56 single cell developmental trajectories of mouse organogenesis. We have also used

these data to explore the dynamics of proliferation and gene expression within cell types over time, including focused analyses of the apical ectodermal ridge, limb mesenchyme and skeletal muscle. These data are the largest dataset as far as we know, comprise a foundational resource for both the single cell genomics and mammalian developmental biology fields.

Despite of its broad application, most current single cell genomic techniques only take a "snapshot" of cell state with temporal information lost during sample preparation. This would introduce a paradox: a cell growing in vivo, and another cell frozen in liquid nitrogen may have identical epigenome or transcriptome information, but they should be totally different from each other with distinct development potentials. This example can be further expanded to show the paucity of "static" single cell genomics in separating cells with similar molecular states but different environment stimuli and thus different development directions. To solve this problem, I developed a new single cell RNA-seq technique (sci-fate, Chapter 4) to characterize cell states by transcriptome dynamics at single cell level. Joint information including both whole and newly synthesized transcriptome identified new cell states compared with conventional single cell RNA-seq. Different from conventional views that cell state is determined by static -omics information, this research showed cell state can be defined by the transition probability distribution to its future states within a fixed short time (i.e. 2 hours). Two cells with similar potential to all possible future states can be regarded from the same cell state. This principle is rather straightforward in our normal life, e.g. the health status of one individual should not be determined by a single check on born, but the occurrence potential of different diseases in his future life, affected by genetic variants and environments.

To further characterize the quantitative features of cell state dynamics, I developed a strategy to

recover single cell transcriptome dynamics across different temporal points. Instead of a smooth and continuous trajectory, cell fate can be characterized by a cell state transition network with nodes of cell states connected by edges representing inter-state transition probabilities. Future cell states can be accurately predicted based on cell state transition network with initial cell state known. With cell state transition probability across all cell states profiled by sci-fate, I further characterized two key factors regulating cell state transition directions: the first is inter-state transition distance, defined by the Pearson's distance of transcriptome between two states. As expected, inter-state distance negatively regulates cell state transition probabilities. The second factor is state instability, defined by the probability of one state transiting to other states within a short time (i.e. 2 hours). Combining cell state instability and state distance greatly increases the prediction accuracy of inter-state transition probability, compared with state distance alone, suggesting nearby state instability is the key factor in cell fate determination.

Currently we are at the starting line of a fascinating journey to fully solve all mysteries in developmental biology, with every cell state and every cellular development trajectory to be characterized by advanced single cell genomics. Moving forward, single cell genomic techniques will continue to grow and mature within the next few years, including higher efficiency to profile low signal cells, higher throughput to detect rare cell types, and the ability to profile new and more molecular layers within the same cell, to uncover the underlying causal relationship between molecular layers. Another next major focus would be to characterize spatiotemporal dynamics of single cells in vivo. Currently this can be partly achieved by combining single cell genomics with imaging-based approaches such as high throughput single molecular FISH and time-lapse imaging, or molecular tracing techniques, though more advanced techniques are still needed to fill this gap. Accompany the technique development, there is also a growing desire for

more advanced computation and statistic methods to incorporate information from different molecular layers, spatial information, and lineage history to fully characterize cell state diversity and more accurate cell fate trajectory recovery.

The advancement of next generation sequencing techniques stimulates a revolutionary research area, genomics science. Similarly, the proliferation of single cell profiling techniques will lead to a new and even more exciting field: cell-omics. Different from conventional single cell genomics which aim to bridge genomics with cell phenotypes and dynamics, cell-omics would focus on the characterization and manipulation of cell function and interaction in higher order organizations such as tissues and organs, to link cell composition and structure with the whole organism phenotypes. One of the keys in cell-omics would be the generation of the first human cell reference within the next few years, including the description of all major cell states by single cell epigenome, transcriptome and proteome as well as their spatiotemporal dynamics. The rising of this new field will spur a wave of revolution in traditional biomedical areas such as pathology and system biology. Looking forward, with the development of more advanced cell-omic techniques to detect, engineer and manipulate every single cell in human body, we may have the ability to fulfill the ultimate goal in biomedical research, a dream we have pursued for the last thousands of years, a world free of any disease and aging.

VITA

Junyue Cao

Education:

Ph. D. Molecular and Cellular Biology, March 2019, University of Washington, Seattle, Washington, U.S.A.

Bachelor of Science in Biology, 2010, Peking University, Beijing, China

Publications:

* denotes equal contribution

2018

Junyue Cao*, Malte Spielmann*, Xiaojie Qiu, Daniel M. Ibrahim, Xingfan Huang, Andrew J. Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J. Steemers, Cole Trapnell, Jay Shendure "The dynamic transcriptional landscape of mammalian organogenesis at single cell resolution" **Nature** (*in press for publication on Feb.20th 2019*)

Junyue Cao, Darren A Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A Pliner, Andrew J Hill, Riza M Daza, Jose L McFaline-Figueroa, Jonathan S Packer, Lena Christiansen, Frank J Steemers, Andrew C Adey, Cole Trapnell, Jay Shendure "Joint profiling of chromatin accessibility and gene expression in thousands of single cells" **Science** 10.1126/science. aau0730 (2018).

2017

Junyue Cao*, Jonathan S Packer*, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, Andrew Adey, Robert H Waterston, Cole Trapnell, Jay Shendure **"**Comprehensive single-cell transcriptional profiling of a multicellular organism" **Science** 357, 661–667 (2017)

2016

Kuo-Hui Su, **Junyue Cao**, Zijian Tang, Siyuan Dai, Yishu He, Stephen Byers Sampson, Ivor J Benjamin, Chengkai Dai **"**HSF1 critically attunes proteotoxic stress sensing by mTORC1 to combat stress and promote growth" **Nature cell biology** 18, pages 527–539 (2016)

Ruzbeh Mosadeghi, Kurt M Reichermeier, Martin Winkler, Anne Schreiber, Justin M Reitsma, Yaru Zhang, Florian Stengel, **Junyue Cao**, Minsoo Kim, Michael J Sweredoski, Sonja Hess, Alexander Leitner, Ruedi Aebersold, Matthias Peter, Raymond J Deshaies, Radoslav I Enchev "Structural and kinetic analysis of the COP9-Signalosome activation and the cullin-RING ubiquitin ligase deneddylation cycle" **eLife** 2016;5:e12102 (2016)

Qi Shen, Changsheng Zhang, Hongbo Liu, Yuting Liu, **Junyue Cao**, Xiaolin Zhang, Yuan Liang, Meiping Zhao, Luhua Lai "De novo design of helical peptides to inhibit tumor necrosis factor-α by disrupting its trimer formation" **Med. Chem. Commun** 2016, 7, 725-729 (2016)

2014

Siyuan Dai*, Zijian Tang*, **Junyue Cao*,** Wei Zhou*, Huawen Li, Stephen Sampson, Chengkai Dai **"**Suppression of the HSF1-mediated proteotoxic stress response by the metabolic stress sensor AMPK" **The EMBO Journal** (2014) e201489062

2012

Chengkai Dai, Siyuan Dai, Junyue Cao **"**Proteotoxic stress of cancer: Implication of the heat-shock response in oncogenesis" **Journal of cellular physiology** 227: 2982–2987(2012)