Single-cell Analysis Reveals the Molecular Roadmap of Mouse Embryogenesis

Chengxiang Qiu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:
Jay Shendure, Chair
Cole Trapnell
William Stafford Noble

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Single-cell Analysis Reveals the Molecular Roadmap
of Mouse Embryogenesis

Chengxiang Qiu

Chair of the Supervisory Committee:
Jay Shendure
Department of Genome Sciences

Mammalian embryogenesis is a rapid and complex process that involves the proliferation and diversification of cells. Within a few weeks, a single-cell zygote gives rise to hundreds of millions of cells that express a wide range of molecular programs. These cells eventually give rise to all of the tissues and organs in the adult body. There are two major questions in mammalian embryogenesis: how cells transition from one stage of development to the next, and what the molecular factors that control this process are. Our current understanding of these questions, particularly in the *in vivo* context, is still incomplete. Although intensively studied, a systematically ascertained delineation of the major cell lineages and trajectories that comprise *in vivo* mammalian development, as well as the regulatory transcription factors (TFs) that drive it, remains elusive.

Recently, we and other researchers have used a variety of single-cell methods to profile millions of cells from a whole mouse embryo. This allowed us to characterize the transcriptional profiles (single-cell RNA-seq, or scRNA-seq), chromatin accessibility (single-cell ATAC-seq, or scATAC-seq), and other modalities of individual cells, from a series of "snapshots" of the embryo at different timepoints, spanning the pre-gastrulation, gastrulation, and organogenesis periods. For example, Pijuan-Sala *et al.* captured the transcriptional profiles of over 100,000 individual cells from the mouse embryo during gastrulation, from embryonic

day (E) 6.5 to E8.5. Cao *et al.*, on the other hand, developed a new single-cell combinatorial indexing technology (sci-RNA-seq3) to profile the transcriptomes of around 2 million cells derived from mouse embryos staged between stages E9.5 and E13.5. I hypothesized that these and other single-cell datasets, especially if they were combined, would offer significant possibilities for obtaining a comprehensive understanding of mouse embryogenesis. Therefore, this thesis will introduce three different projects that leverage scRNA-seq datasets with comprehensive computational analysis to address this question.

In the first project, we integrated multiple published scRNA-seq datasets to define cell states at 19 stages of mouse gastrulation and early organogenesis, from E3.5 to E13.5. We then heuristically connected these cell states to their pseudo-ancestors and pseudo-descendants based on their transcriptional similarity, creating a graph of cellular trajectories. We then leveraged this graph to identify TFs and TF motifs that regulate the emergence of new cell types.

In the second project, we applied sci-RNA-seq3 to fill the data gap between late gastrulation and birth, by profiling the transcriptional states of 12.4 million nuclei from 83 precisely staged embryos spanning late gastrulation (E8) to birth (postnatal day 0 or P0), with a temporal resolution of at least 6 hours. We identified hundreds of cell types and performed deeper analyses of the development of the posterior embryo during somitogenesis, as well as the kidney, mesenchyme, retina, and early neurons. We leveraged these data, together with other published datasets, to construct a rooted tree of cell type relationships that spans mouse development from zygote to pup, followed by identifying sets of TFs and other genes as candidate drivers of cell type differentiation.

In the third project, we aimed to establish scRNA-seq of the whole embryo as a scalable platform for the systematic phenotyping of mouse genetic models. We used sci-RNA-seq3 to profile and analyze the gene expression of 1.6 million nuclei from 101 embryos of 22 mutants and 4 wild-type genotypes at E13.5. We then developed and applied several analytical

frameworks to detect differences in the composition and/or gene expression of 52 cell types or trajectories across genotypes.

This study takes a comprehensive approach to understanding mammalian embryogenesis. By examining the cellular trajectories that cells take during development and the molecules that drive these trajectories, we aim to answer fundamental questions about this process. We believe that this work will provide a foundation for a deeper understanding of mammalian development.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

The Chinese proverb *"One tree doesn't make a forest"* is true for many things in life, including a PhD journey. I could not have completed my PhD without the help and support of a large community of people. My mentor, Jay Shendure, provided me with guidance and mentorship. My friends in the lab and PhD program offered support and encouragement. The Genome Sciences department at the University of Washington provided me with the resources and facilities I needed to succeed. And my family was always there for me, cheering me on. The COVID-19 pandemic began during my second year of PhD studies, and it lasted for almost two years. I spent most of my time at home, joining meetings and lectures online. This was not an ideal situation, but I never lost my connection with my colleagues and continued to work on my research. The support I received from my community went beyond research. It was also present in the everyday details of my life, which made my PhD journey more colorful and enjoyable. I am grateful for the people who made it possible.

First of all, I am deeply grateful to my PhD mentor, Jay Shendure, for his guidance, mentorship, and support. He has had a profound impact on my scientific career and personal development. I still remember when we were working on our first project, I identified several hundred possible ancestor-descendant cell states relationships across timepoints. This was purely data-driven, and I was excited to share my findings with Jay. I sent him the results, and the next morning, I found that he had already put his personal comments and references on each of the entities. I was amazed that he was so busy, but he still took the time to carefully review my work and provide valuable feedback. Jay has a gift for building positive and

supportive lab cultures. He creates an environment where everyone feels welcome, respected, and valued. He is also a great mentor, always willing to share his knowledge and experience. I am confident that the skills and knowledge I have learned from Jay will serve me well in my future career.

Second, I would like to thank my collaborators from the lab who directly or indirectly contributed to the projects that I would introduce in this thesis. Beth Martin and Riza Daza performed all the experiments to generate the scRNA-seq and scATAC-seq datasets, which were critical to my data analysis and research project. Their hard work and dedication made my research possible. Junyue Cao mentored me during my lab rotation and directed me to my current research field. I am grateful for his guidance and support. Xingfang (Fanny) Huang, Diego Calderon, and Silvia Domcke helped me a lot with my computational analysis. I am grateful for their expertise and patience. I would also like to thank my other colleagues in the lab, who helped me complete my research projects and inspired me with many conversations. Sanjay Srivatsan, Tony Li, Nobuhiko Hamazaki, Eva Nichols, Truc-Mai Le, Diana R. O'Day, Megan Taylor, and Olivia Fulton were great collaborators who helped me with my experiments and data analysis. I am grateful for their hard work and dedication. Sam Regalado, Junhong Choi, Wei Yang, Xiaoyi Li, JB Lalanne, Aidan Keith, Flo Chardon, Chase Suiter, Connor Kubo, David Lee, Hanna Liao, Jenny Nathans, Shruti Jain, Riddhiman Garge, Haedong Kim, Troy McDiarmid, Sudarshan Pinglay, Elizabeth Vincent, Val Browning, and Kina Atkin-Yamaguchi were also great collaborators who inspired me with their insights and ideas. I am so glad to have had the opportunity to work with you all. I also want to thank former lab members Wei (Will) Chen, Alexander Boulgakov, Anna Minkina, Jacob Tome, Anh Leith, Sereno Lopez-Darwin, Ruolan Qiu, Jase Gehring, Yi Yin, Ronnie Blecher-Gonen, Vikram Agarwal, Molly Gasperini, Seungsoo Kim, Andrew Hill, Aaron McKenna, Bridget Kulasekara, and Hannah Pliner. I am so glad to have had

the opportunity to learn from you all. Finally, I want to thank Olga Oseth, Melissa Gillies, and Choli Lee for their support of the whole lab team. I am grateful for their administrative and technical assistance.

Third, I would like to thank all my collaborators outside of the lab. My committee members, Cole Trapnell, William Stafford Noble, Robert Waterston, and Cecilia Moens, gave me a lot of valuable advice on my research work and future career development. I am grateful for their insights and guidance. Ian Welsh and Stephen Murray from The Jackson Laboratory helped us collect and precisely stage mouse embryos. I am grateful for their technical expertise and support. Malte Spielmann, Jana Henck, and Varun Sreenivasan collaborated with me and Fanny on the mutant mice project. I am grateful for their insights and contributions. Many people from the Trapnell lab, including Xiaojie Qiu, Hyeon-Jin Kim, Maddy Duran, Eliza Barkan, Andrew Mullen, Amy Tresenrider, Lauren Saunders, Michael Dorrity, David Kimelman, Brent Ewing, and Dana Jackson, helped me with computational analysis and interpreting results. I am grateful for their expertise and patience. People from the Noble lab, including Ran Zhang, Gang Li, Gesine Cauer, Giancarlo Bonora, Yang Lu, Ritambhara Singh, and Gurkan Yardimci, helped me during my rotation or later collaboration. I am grateful for their mentorship and support. I am grateful to Christine Disteche and Xixian Deng for their valuable collaboration on the allele-specific chromatin accessibility projects. I am grateful for their insights and contributions. I want to thank the Genome Sciences department, the Chair Stanley Fields, the PhD program manager Brian Giebel, and the full support human resource team. I am grateful for their support and guidance throughout my PhD journey. I would also like to express my gratitude to the GSIT team for their patience and helpfulness. I cannot remember how many emails I have sent them, but they have always replied promptly and gone out of their way to help me resolve my problems.

I would also like to thank my previous PIs, Qinghua Cui and Katalin Susztak, for their mentorship and guidance. Qinghua Cui, my master's PI at Peking University, helped me to develop my interest in bioinformatics and computational biology. He provided me with the foundation and knowledge that I needed to pursue a career in this field. I am grateful for his guidance and support during my master's studies. Katalin Susztak, my PI at Upenn, was a great mentor and role model. She taught me a lot about good science and how to conduct high-quality research. She also helped me to develop my presentation and communication skills. I am grateful for the opportunity to have worked with her and to have learned from her.

Finally, I would like to express my heartfelt gratitude to my family for their endless love and support. My parents, Tiancai Qiu and Chunmei Jiang, and my brother, Chengbin Qiu, have always been there for me, through good times and bad. My wife, Shizheng Huang, is always there to pick me up when I am down, and I am so lucky to have her in my life. My son, Moyan Qiu, and his best friend, QTPi, are the light of my life. They bring me so much joy and happiness. I am so lucky to have such a loving and supportive family. They are my everything.

Chapter 1

INTRODUCTION

The questions that people often ask about themselves — *who am I, where did I come from, and where am I going?* — can also be asked about individual cells during development. What is it? Where does it come from? And what will it give rise to? These three fundamental questions about cell identity, origin, and fate can be summarized as the lineage relationship between cells or cell states involved in mammalian development. In other words, *how do cells transition from earlier to later stages of development?* During development, a single zygote cell divides rapidly and forms three germ layers: the ectoderm, mesoderm, and endoderm. These germ layers give rise to all of the body's tissues and organs. The temporary cell states undergo a series of sequential differentiation steps to form the final mature cell types with precisely assigned functions. For example, the eye is a relatively small tissue, but it contains six distinct types of retinal neurons that form highly organized layers. Each type of retinal neuron plays a unique role in transmitting light signals. More broadly, previous research has identified over one thousand different types of neurons in the human or mouse brain. These neurons have a wide range of functions, including carrying out sensory perception, controlling movement, and processing information. During development, the process of cell differentiation is highly regulated and controlled by a variety of molecular factors. Some cells, such as neural-mesodermal progenitors (NMPs), have the potential to differentiate into multiple cell types. NMPs are located at the posterior embryo during late gastrulation and early somitogenesis. They have bipotent differentiation capabilities, meaning that they can differentiate into either neuroectoderm (*Sox2*+) or mesoderm (*Tbx6*+) [1]. The precisely organized and modified process of development is like a well-designed machine. This raises

the second important question - *what are the key molecular factors that underlie each potential course of differentiation?* More specifically, how does a cell determine its fate during a particular developmental stage? In principle, developmental programs can be comprehensively described. For example, Sulston and colleagues reconstructed the complete embryonic lineage of the roundworm *C. elegans* through visual observation in 1983 [2]. However, *C. elegans* is a small, translucent, and developmentally invariant organism, which makes it the only model organism for which such a complete description has been achieved. Our current understanding of the two fundamental questions of mammalian development is fragmentary. Most studies have focused on relatively short stages of development or a particular lineage, and a systematic delineation of the major cell lineages and trajectories that comprise *in vivo* mammalian development, as well as the involved regulatory transcription factors (TFs), remains elusive.

1.1 Genome editing and light-sheet microscopy shed light on embryogenesis

Recent advances in technologies, such as cell lineage tracing and light-sheet microscopy, have enabled researchers to answer long-standing questions about cellular dynamics in embryogenesis. In 2016, McKenna *et al.* developed a technique called GESTALT (genome editing of synthetic target arrays for lineage tracing), which enables tracking cell lineage in whole, complex organisms [3]. It uses CRISPR/Cas9 to introduce and accumulate mutations in a DNA barcode over multiple rounds of cell division. The barcode can be used to mark cells and track cell lineages by analyzing the patterns of mutations shared between cells. McKenna *et al.* applied the technology to track the lineages of hundreds of thousands of cells in zebrafish and demonstrated that most cells in adult organs are derived from a small number of embryonic progenitors. GESTALT has the potential to be used to map the complete cell lineage in diverse organisms, and it could also be adapted to link cell lineage information to molecular profiles of the same cells.

McDole *et al.* improved adaptive light-sheet microscopy by making it adaptive to the dramatic changes in size, shape, and optical properties of the embryo over time, enabling high-resolution imaging of cells from gastrulation to early organogenesis in developing mouse embryos *in vivo* [4]. To analyze the data collected by their microscope, the researchers developed a computational framework that can reconstruct long-term cell tracks, cell divisions, dynamic fate maps, and maps of tissue morphogenesis across the entire embryo. This study pioneered the direct visualization of the dynamic changes of the mouse embryo during early development, such as cell divisions and cell movements.

GESTALT and adaptive light-sheet microscopy are both powerful tools for advancing our understanding of mouse embryogenesis and could help identify new genes and pathways involved in gastrulation and early organogenesis. However, they each have limitations. GESTALT is not able to track cells for long periods of time, and it is not as effective in larger and more complex organisms like the mouse. Adaptive light-sheet microscopy is also limited by the size of the embryo that can be imaged at once. These limitations have prevented researchers from reconstructing the cellular trajectories spanning embryogenesis in the mouse.

1.2 Single-cell sequencing revolutionizes embryogenesis research

The above technologies, which either leverages "recording" or microscopy, can be summarized as "direct" strategies for studying mammalian embryogenesis. The direct strategy involves following the development of a single embryo over time, *e.g.* this can be done using time-lapse microscopy, which allows researchers to track the movements and changes of cells in an embryo over a period of hours or days. This is a powerful strategy that provides a precise, dynamic, and comprehensive picture of the embryo during development. However, due to technical challenges, it is difficult to apply to large embryos (*e.g.* late-stage mouse

embryos). Additionally, it can be expensive and time-consuming to track a single embryo over an extended period.

Alternatively, in recent years, a more "indirect" strategy for studying mammalian embryogenesis has become increasingly popular. This strategy involves sampling cells from different embryos at multiple timepoints, rather than following the development of a single embryo over time. This strategy takes advantage of the rapid development of single-cell technologies, which have become more affordable and easier to use. Scientists have used single-cell RNA sequencing (scRNA-seq) or single-cell ATAC sequencing (scATAC-seq) to study the gene expression or chromatin accessibility of individual cells in a whole embryo or a specific tissue at multiple timepoints. This has allowed them to identify the cellular features of individual cells at a single stage, as well as the similarities between cells across timepoints. The results of these studies can then be used to computationally reconstruct the cellular trajectories that led to the development of the embryo. This process is similar to how the first movies were made, where individual frames were captured and then stitched together to create a moving image. Similarly, researchers using the "sampling" strategy take a series of snapshots of cells at different stages of development, and then use computational methods to reconstruct the continuous roadmap of the embryo's developmental trajectory.

The success of the sampling strategy for studying mammalian embryogenesis depends on four factors:

- **The number of cells profiled.** The number of cells that need to be profiled depends on the size of the embryo and the stage of development at which it is profiled. To ensure that all cell types, including rare cell types (*e.g.* germ cells), are likely to be captured from each stage, a sufficient number of cells must be profiled.

- **The depth of profiling.** This refers to the number of genes or transcripts (for scRNA-seq) that are measured for each cell. The deeper the profiling, the more likely it is that

sufficient variation will be captured from the different cell types in the embryo.

- **The temporal resolution.** This refers to the time interval between successive measurements. The shorter the temporal resolution, the more detailed intermediate cell states will be captured, and the more accurate the reconstruction of the cellular trajectories will be. This is especially important for stages of development like gastrulation and early somitogenesis, where cells are undergoing rapid and dynamic changes.

- **The span of development covered.** This refers to the range of developmental stages that are sampled. The broader the range of developmental stages that are studied, the better the understanding of the overall process of cell differentiation and lineage commitment.

The single-cell sampling strategy is an efficient and versatile tool for studying mammalian embryogenesis. It is less expensive and less time-consuming than traditional methods, such as live imaging, and it can be used to study embryos that are difficult or impossible to follow over time. This strategy has been used to collect and study genomic information from large numbers of cells during embryogenesis in multiple species, including worms, flies, zebrafish, frogs, mice, and humans (**Fig. 1.1**). The size of embryos and the duration of their development increase from worms to humans, making it technically challenging to profile whole embryos with single-cell technology. Most of the data generated using this approach have come from scRNA-seq, as transcriptomes are the most important factor that specifies cell state identities. Here I will provide a brief overview of some of the milestone studies and their most notable features.

Caenorhabditis elegans (*C. elegans*), despite its limited cell number, contains a wide diversity of cell types and a complex anatomy. Its cell lineage has been fully mapped, making *C. elegans* an ideal animal to study the transcriptome dynamics during cell differentiation

Species	Developmental Stages	Background	# of Embryos	# of Cells	scRNA-seq technology	Reference
C. elegans	300, 400, 500 mins post bleach	N2 strain	---	86,024	10X Genomics	Packer et al.
Drosophila	0 - 20 hours	Canton-S	---	547,805	sci-RNA-seq3 (nuclei)	Calderon et al.
Zebrafish	hpf3.3, 3.8, 4.3, 4.8, 5.3, 6, 7, 8, 9, 10, 11, 12	wild-type (Tupfel longfin/AB) crosses	694	38,731	Drop-seq	Farrell et al.
Zebrafish	hpf6, 8, 10, 14, 18, 24	AB and TU wild-type strains	"50-100 [per] sample"	36,727	inDrops	Wagner et al.
Zebrafish	hpf10 - hdf10	EKW strain	40	120,444	10X Genomics	Lange et al.
Zebrafish	hpf3 - hpf120	wild-type (Tupfel longfin/AB) crosses	"10-12 [per] sample"	489,686	MULTI-seq and Drop-seq	Sur et al.
Xenopus	S8, 10, 11, 12, 13, 14, 16, 18, 20, 22	Xenopus tropicalis embryos	"5-10 embryos at a time"	123,633	inDrops	Briggs et al.
Mouse	E3.5, E4.5, E5.5, E6.5	C57BL/6Babr mice	32	509	manual isolation and G&T-seq	Mohammed et al.
Mouse	E5.25, E5.5, E6.25, E6.5	C57BL/6J and CAST/EiJ female mice	28	1,724	manual isolation and smart-seq2	Cheng et al.
Mouse	E6.5 - E8.5	C57BL/6 female mice	347	108,857	10X Genomics	Pijuan-Sala et al.
Mouse	E6.5 - E8.25	C57BL/6JRccHsd or Hsd:ICR(CD-1)	153	33,700	MARS-Seq	Mittnenzweig et al.
Mouse	E9.5 - E13.5	C57BL/6 mice	61	2,058,652	sci-RNA-seq3 (nuclei)	Cao et al.
Human	PCW4-6	head, trunk, viscera, and limb	7	185,140	10X Genomics	Xu et al.
Human	PCW3-12	whole embryos, head, and brain	14	430,808	10X Genomics	Zeng et al.
Human	72 to 129 days	15 organs	121	4,062,980	sci-RNA-seq3 (nuclei)	Cao et al.

Figure 1.1: **A summary of scRNA-seq datasets of embryos in multiple species, including worms, flies, zebrafish, frogs, mice, and humans.**

and development. In 2019, Packer *et al.* profiled the transcriptomes of 86,024 single embryonic cells from *C. elegans*, identifying 502 terminal and preterminal cell types as well as their molecular basis for specification [5]. More importantly, they mapped most single-cell transcriptomes to their exact position in the invariant lineage of *C. elegans*, to explore the relationships between transcriptomes and lineages for cells. Some of their findings, which serve as a proof of concept, help us understand the major question in the single-cell sampling strategy: does transcriptome similarity between cells really reflect their lineages during development? For example, they found that the correlation between a cell's lineage and its transcriptome increased from middle to late gastrulation, but then fell substantially as cells in the nervous system and pharynx adopted their terminal fates. They also found that genes that distinguish sister cells are often co-expressed in the parent cell and then selectively retained in one daughter cell. Additionally, they found that most lineages that produce the same anatomical cell type converge to a homogeneous transcriptomic state. Finally, they raised concerns that computational methods that rely solely on single-cell transcriptomic data to reconstruct developmental trajectories are often inaccurate.

As mentioned earlier, the temporal resolution of sampling is essential for the accuracy

of trajectory reconstruction. Ideally, we would profile embryos continuously rather than at discrete timepoints to obtain a complete picture of embryogenesis *in vivo*. However, this is not feasible for most model organisms like mice or humans. In *Drosophila melanogaster*, it is possible to do this because it is small and its development is rapid. In 2022, Calderon *et al.* profiled chromatin accessibility in almost 1 million nuclei and gene expression in half a million nuclei from 11 overlapping windows spanning the entire period of *Drosophila* embryogenesis (0 to 20 hours) [6]. Instead of using the initial collected time windows (which are discrete), they applied deep neural network-based predictive modeling to predict a more accurate developmental age (which is continuous) of each nucleus in the dataset. This resulted in continuous, multimodal views of cellular transitions in absolute time. Although they did not apply scRNA-seq and scATAC-seq in a co-assay manner, they were still able to explore the dynamics of enhancer usage and gene expression together within and across lineages at the scale of minutes. The inclusion of predicted nuclear ages will make it possible to explore the precise time points at which genes become active in distinct tissues, as well as how chromatin is remodeled over time.

Danio rerio (zebrafish) is perhaps the most well-studied model organism in this field because it is relatively easier to study than mice or humans (small size, transparent embryos, and rapid development) while still having highly conserved developmental processes. Additionally, zebrafish has been intensively studied using traditional methods, which makes it easier to annotate cell types and identify trajectories. In 2018, two independent studies, by Farrell *et al.* and Wagner *et al.*, each profiled approximately 40,000 cells from pooled zebrafish embryos during gastrulation and early organogenesis [7, 8]. Although the studies focused on different stages of embryogenesis and their temporal resolution was different (every 0.5-1 hour from 3.3 hours post-fertilization (hpf) to 12 hpf in the Farrell *et al.* study, every 2-6 hours from 4 hpf to 24 hpf in the Wagner *et al.* study), a few timepoints coincided, making it possible to integrate the two datasets. Both studies innovatively used computational methods to reconstruct a tree-like graph of cells or cell states. The tree structure is

a convenient way (sometimes imprecise though) to describe developmental trajectories over time because it clearly shows the differentiation branchpoints of cell states. Of note, concurrently, Briggs *et al.* profiled over 100,000 cells from *Xenopus tropicalis* embryos over the first day of life (from 5 hpf to 24 hpf) [9]. Based on the scRNA-seq data, they identified cell states from each single timepoint and then computationally connected the cell states between adjacent timepoints, resulting in a tree-like graph of cell states. They then aligned this graph to the tree reconstructed from zebrafish by applying the same strategy to the data from Wagner *et al.*, followed by systematically identifying the cell type orthologs between the two species (frogs vs. zebrafish). The details of the computational methods used to reconstruct the developmental trajectories in those three studies will be discussed in the next section.

In the last two years, several new studies generated scRNA-seq datasets from zebrafish embryos, either applying more dense temporal sampling, investigating different cellular features, or capturing different periods of development. Saunders *et al.* applied scRNA-seq to study 1812 individually resolved zebrafish embryos [10]. Their data spans 19 time points, 23 genetic perturbations (F0 knockouts, generated by CRISPR-Cas9 mutagenesis), and a total of 3.2 million cells. Lange *et al.* posted Zebrahub, a dynamic atlas of zebrafish embryonic development that integrates single-cell sequencing and light-sheet microscopy data [11]. It provides high-resolution molecular insights into zebrafish embryos and enables *in silico* fate mapping experiments (*e.g.* NMPs have been explored in their study). Sur *et al.* extended the range of covered stages from 12 hpf in the Farrell *et al.* study to 120 hpf [12]. The resulting atlas, created by integrating two datasets, maps transcriptionally distinct populations across 62 stages of development, ranging from 3.3 hpf to 120 hpf. Those newly generated data, together, provide another opportunity to reconstruct more comprehensive and precise trajectories for zebrafish development.

The house mouse, *Mus musculus*, is an exceptional model system, combining genetic

tractability with close homology to human biology. Within a few weeks, a single-cell zygote undergoes rapid cell division and differentiation, giving rise to hundreds of millions of cells that express a wide range of molecular programs. However, the long timespan, large embryo size, and complex cellular states of mouse embryogenesis make it challenging to study developmental trajectories using single-cell technology. There are three major datasets that have captured the transcriptomes of individual cells at the whole-embryo scale from different stages of mouse embryogenesis. The first two studies focused on the gastrulation stage (embryonic day (E) 6.5 to E8.5), while the third study focused on the early organogenesis stage (E9.5 to E13.5). Pijuan-Sala *et al.* profiled the transcriptomes of 108,857 single cells from over 300 mouse embryos collected at nine sequential time points from E6.5 to E8.5, with a temporal resolution of six hours [13]. They constructed a comprehensive molecular map of cellular differentiation from pluripotency to all major embryonic lineages, and particularly explored several complex events, such as the convergence of visceral and definitive endoderm (primitive streak-derived). The dataset generated by this study is of high quality and well-annotated, and has been widely used by researchers studying mouse gastrulation. Similarly, during gastrulation stages, Mittnenzweig *et al.* profiled the transcriptomes of 33,700 cells from 153 mouse embryos [14]. They predicted the precise developmental timepoints for individual embryos based on morphological and transcriptional features, resulting in embryos being grouped into 13 time windows from E6.5 to E8.1. Most interestingly, they applied an optimal transport model to estimate transition probabilities between cell states from adjacent timepoints. This allowed them to reconstruct the developmental trajectories of mouse gastrulation. The method details will be discussed in the next section.

Cao *et al.* studied early mouse organogenesis by profiling the transcriptomes of approximately 2 million nuclei from embryos collected at 1-hour intervals from E9.5 to E10.5 [15]. They applied Monocle/v3 algorithms to identify 12 major developmental trajectories as well as 64 sub-trajectories. Compared to the first two studies, mouse embryos at the organogenesis stage are much larger and require more cells to be profiled. Cao *et al.* addressed

this challenge by developing their own three-level single-cell combinatorial indexing RNA-seq (sci-RNA-seq3) technology, which is more cost-effective and allowed them to generate such a comprehensive data atlas. Additionally, this is the first study to profile over 1 million cells in a single experiment, opening a new stage of single-cell study that enables the implementation of single-cell technology on larger organisms.

Human embryogenesis is a long and complex process, and it is difficult to collect human embryos for research. As a result, most single-cell studies on human embryogenesis have focused on relatively shorter stages or specific tissues, and the number of cells profiled and the transcripts that recovered from each cell has been relatively lower than studies on other species. For example, Xu *et al.* profiled the transcriptomes of 180,000 single cells from post-conceptional weeks (PCW) 4 to 6 human embryos [16]. Zeng *et al.* profiled over 400,000 cells from 14 human samples (whole embryos, heads, and brains) from PCW 3 to 12 [17]. The largest dataset to date is from Cao *et al.* in 2020, who applied sci-RNA-seq3 to profile the gene expression of 4 million single cells from 121 human fetal samples (15 organs; ranging from 72 to 129 days) [18]. Overall, it is still challenging to reconstruct the comprehensive cellular trajectories of human embryogenesis *in vivo*, compared to zebrafish or mice, due to the paucity of data.

All of the studies mentioned above primarily focused on profiling whole embryos. However, some other studies have focused on specific tissues or organs during development. For example, Tyser *et al.* used a combination of transcriptomic, imaging, and genetic lineage labeling approaches to map the origin of the embryonic mouse heart at single-cell resolution [19]. They applied microdissection to isolate 3,105 cells from the anterior cardiac region of mouse embryos at six different stages of development, from E7.75 to E8.25. This allowed them to obtain high-resolution data on the cellular differentiation of the heart. Manno *et al.* reported a comprehensive single-cell transcriptomic atlas of the embryonic mouse brain from gastrulation to birth (E6 to E18) [20]. They identified almost 800 cellular states based

on 292,495 cells from 215 pooled embryos. These cellular states describe a developmental program for the functional elements of the brain and its enclosing membranes, such as the early neuroepithelium, region-specific secondary organizers, and both neurogenic and gliogenic progenitors. Focusing on a single tissue can provide higher resolution, but it might lose some necessary cell types during tissue dissection and the global view of cell compositions within a whole embryo.

When analyzed together, single-cell atlases can be used to explore the properties of individual cell states and reconstruct cellular lineage relationships. However, this also raises some new challenges in terms of choosing the best computational approach. In the following section, I will discuss various approaches to analyzing one or more datasets to reconstruct the cellular trajectories of embryogenesis and their advantages and limitations.

1.3 Computational methods for studying cellular trajectories in embryogenesis

Many studies have successfully characterized cellular lineages and trajectories using single-cell transcriptomics. The choice of method depends on the specific goals of the study, the type of data available, and the computational resources available. Some factors to consider include the species, the temporal resolution and time span of the data, and the profiling depth of the cells. Some of these methods do not require cells to be sampled from a series of timepoints, such as UMAP, pseudotime, PAGA, and RNA velocity.

UMAP (uniform manifold approximation and projection) is a dimensionality reduction algorithm that is commonly used in single-cell data analysis [21]. It works by first constructing a similarity graph of the data points (*i.e.* cells), and then iteratively optimizing the layout of the points in the lower-dimensional space (2D or 3D) such that the local dis-

tances between points are preserved as much as possible. Comparing to other widely-used non-linear dimensionality reduction methods, such as Multidimensional Scaling (MDS) and t-distributed Stochastic Neighbor Embedding (t-SNE), UMAP uses a different way of calculating distances that is more efficient and can preserve the global and local structures of the data better. As a result, it is a versatile tool that can be used for both visualizing large datasets and inferring developmental trajectories, especially in less complex systems. However, it is important to be careful not to over-interpret UMAP results, as some studies have shown that it can be misleading in some cases [22]. Additionally, the visualization of data using UMAP can vary depending on the parameters used (*e.g.* min_dist and n_neighbors). Therefore, it is important to be careful when selecting these parameters..

Pseudotime inference is a method used to infer the ordering of cells ("pseudotime", a time-like variable) along a lineage based on their gene expression profiles measured by scRNA-seq. The predicted pseudotime of individual cells is often visualized using UMAP. Pseudotime is a powerful strategy to infer developmental trajectories because it estimates the progression of individual cells through development, rather than the time at which they were collected, which can vary between replicates or cell types. However, the one-dimensional nature of pseudotime makes it challenging to capture the complexity of multiple independent trajectories during development. For example, the pseudotime method implemented in Monocle/v3, which is commonly used, can still have difficulty estimating pseudotime across disjointed cell partitions [15].

PAGA (partition-based graph abstraction) is another commonly used method for estimating trajectories [23]. It differs from pseudotime in that it focuses on cell clusters instead of individual cells. PAGA works by first clustering cells into groups based on their gene expression profiles. These clusters are then connected together in a graph, where the edges between clusters represent the similarities between them. The main disadvantage of PAGA is that it is highly dependent on the way cell clusters are defined. If too many clusters are

defined, the trajectories will be too complex and difficult to interpret. If too few clusters are defined, the heterogeneity within each cluster will be too high.

Unlike UMAP and pseudotime, which only use gene expression data, RNA velocity is based on the idea that the abundance of unspliced and spliced mRNA can be used to infer the transcription and splicing rates of genes [24]. The RNA velocity of each gene is the rate of change of its expression level over time. Cells with similar RNA velocities are likely to be in the same state of differentiation. RNA velocity is a powerful tool that can be used to gain insights into the dynamics of cell differentiation. However, it has some limitations. The estimated velocity must be projected into a low-dimensional embedding (*e.g.* UMAP) for visualization, and the results can be difficult to interpret. Additionally, the performance of RNA velocity is sensitive to the choice of parameters and the amount of noise in the data.

For methods designed to analyze a series of timepoints with sampled cells, I have summarized them into two main categories: those that mainly focus on local structures and those that mainly focus on global structures. Local structure methods identify the local neighbors of individual cells and then build cell-cell transition maps across timepoints. Global structure methods, on the other hand, try to find the minimal global loss for cell-cell transitions across timepoints (*e.g.* optimal transport).

In the first category, the URD approach was used to reconstruct developmental trajectories for early zebrafish embryos in the study by Farrell *et al.* This approach uses a combination of the pseudotime and random walks algorithms [7]. It first computes the transition probabilities between cells based on the similarity of their gene expression profiles followed by assigning each cell a pseudotime as the average number of transitions required to reach that cell from user-defined "root" cells (*e.g.* the cells from the earliest timepoint). Next, it estimates the trajectories from user-defined "tip" cells (*e.g.* the cells from the final timepoint) by simulating random walks biased towards younger pseudotime. The trajectories

are joined to recover a tree-like structure, which is finally visualized using a force-directed layout. Technically, the URD algorithm does not consider the initial timepoint at which the cells were profiled. It only uses this information to assign cells as either roots or tips. However, collecting cells from a wide range of timepoints provided insights into cell state heterogeneity over time during development.

Another approach in the first category, a graph-based approach, was used to map a cell-state landscape in the studies by Wagner *et al.* (on zebrafish) and Briggs *et al.* (on frogs) [8, 9]. Unlike the above methods, this approach heavily relies on the initial timepoints at which the cells were profiled. It first splits the entire dataset of cells into groups based on their initial timepoints, and then performs embedding and cell state annotation independently for each group. Next, the algorithm co-embeds the cells from each two adjacent timepoints into a common space, followed by using k-nearest neighbors (k-NN) to link the cell states, where the cell state with the most abundant neighboring cells is assigned as the ancestor cell state. The method finally constructs a tree-like structure to clearly interpret the developmental process in either zebrafish or frogs, allowing for the easy identification of key regulators along the emergence of new cell types. The two graphs, from zebrafish and frogs, were reconstructed using the same strategy, so it is straightforward to align the developmental trajectories between the two species. However, the model has some potential limitations. First, developmental trajectories do not always follow a simple binary differentiation structure. There are many complex cases, such as the interactions between different germ layers or the convergence of different trajectories. For example, the gut was identified as being derived from both the visceral endoderm (extraembryonic) and the definitive endoderm (embryonic). Second, the model may be challenging to apply to later stages of embryogenesis, when some clusters are identified by sub-clustering but the cell state connections are still identified by global embedding.

The second category of methods is mainly based on optimal transport (OT), a mathe-

matical theory that studies the problem of moving one distribution of mass to another as efficiently as possible. It is a powerful tool for finding a mapping between two sets of points such that the total distance between the corresponding points is minimized, even if the two sets of points have different feature spaces (*i.e.* the set of features that are used to represent data). Schiebinger *et al.* developed a computational tool called Waddington-OT (WOT) to reconstruct the cellular lineages from cells collected at half-day intervals across 18 days during the reprogramming of induced pluripotent stem cells [25]. WOT estimates the transition probability between cells from each pair of adjacent timepoints using optimal transport and gene expression features. The cellular lineages are then tracked over time by following the transition probabilities. Moreover, the transition probability between any two timepoints, not just the two adjacent ones, can be predicted by multiplying the transition probability matrices of the timepoints between them in sequence. This can be used to identify the ancestors of any state at the terminal stages, starting from the earliest stages. However, there are some potential concerns that need to be further considered when applying WOT to other datasets. First, WOT requires a high temporal resolution. If there are large "gaps" in the data, the estimation of transition probabilities may be inaccurate. Second, the application presented in the study is a relatively simple differentiation example. Embryogenesis is a much more complex process, and it spans a much longer time. It is not clear whether WOT will be able to accurately estimate transition probabilities in such a complex and extended setting.

Mittnenzweig *et al.* constructed an improved OT model (termed "flow" model) for estimating the transition of cells on the transcriptional manifold [14]. In brief, the researchers used a combination of transcriptome and morphological data to create a continuum of transcriptional transformation for individual embryos. They then grouped cells from the same embryo into 13 temporal groups (E6.5 to E8.1) and aggregated transcriptionally similar cells into metacells (MCs). Finally, they used an OT-based method to identify the transition probability between metacells from each pair of adjacent temporal groups. The flow model

has three main features which makes it different from other OT-based methods. First, it uses metacells as the basic unit of analysis. This means that annotations or transition estimations are performed on each metacell, rather than on single cells or cell types. Second, the model adds a constraint that the fraction of cells within a metacell at a given timepoint must be equal to the total number of cells that enter the metacell from the earlier timepoint and the total number of cells that exit the metacell to the later timepoint. Third, the model also takes into account cell growth rate, which further improves the accuracy of the estimations. The flow model is a robust method, but it has some drawbacks, such as the graph it produces being challenging to interpret and the size of the metacells being somewhat subjective. Additionally, since the flow model has only been applied to embryos during gastrulation, its performance in later stages of development, such as organogenesis, remains to be tested.

In a recent study, Klein *et al.* introduced Moscot, a new OT-based framework that overcomes the limitations of existing methods [26]. Moscot supports multimodal data and joint cellular representations, making it more powerful than unimodal methods. It is also scalable and can be used for large-scale datasets. Additionally, Moscot unifies previous single-cell applications of OT in the temporal and spatial domain, and introduces a novel spatiotemporal application.

While many computational methods for cellular trajectory reconstruction based on scRNA-seq data have been introduced, there are several key factors to consider when choosing the appropriate method. These include:

- There is a trade-off between the accuracy of cellular trajectories and the resolution of cell states (*i.e.* either single-cell, metacell, or cell type). Here, a cell state refers to a group of cells with the least transcriptional heterogeneity, and a cellular trajectory refers to the developmental relationship between two cell states, such as progenitor cells giving rise to differentiated cells. For example, iterative clustering is commonly used

to identify cell subtypes or even sub-subtypes. However, these subtypes, especially for rare types, may be difficult to accurately identify with trajectories when we perform a systematic analysis of the global embedding.

- Similarly, there is a trade-off between the accuracy of cellular trajectories and their ease of interpretation. For example, the tree-like model from Briggs *et al.* is easier to understand because it provides a clear and concise overview of the differentiation process. In contrast, the flow model introduced by Mittnenzweig *et al.* is more accurate because it divides cells into more detailed groups (*i.e.* metacells) and considers more realistic factors (*e.g.* growth rate).

- The cellular complexity of an embryo development process can vary depending on the species, tissue, or developmental stage. Therefore, it can be challenging to apply the identical approach to data generated by different studies or cells profiled from different stages (*e.g.* gastrulation vs. late organogenesis). It is better to select the appropriate method based on the specific data and the research question being asked.

- The increasing size of scRNA-seq datasets is driving the development of optimized methods for their analysis. These methods, such as GPU acceleration and downsampling, can significantly reduce the computational cost of analyzing scRNA-seq data [26].

- In addition, careful attention should also be paid to technical factors during data preprocessing. These factors can have an unexpected impact on the results. For example, Scrublet is the current main method for detecting doublet cells in scRNA-seq [27]. It works by randomly combining gene expression from two real cells to create simulated doublets. The doublet cells are then identified as those that are closest to the simulated doublets in the co-embedding. However, this method may also remove some cells in intermediate state of differentiation. Although no systematic studies have

yet been conducted to confirm this, it is important to be aware of this potential pitfall when using Scrublet or other doublet removal methods.

After discussing the conventional data and methods used to study embryogenesis, the following sections will briefly discuss some other modalities of data or alternative strategies for studying embryogenesis.

1.4 Spatial mapping of cell states

The rapid increase in cell number during embryo development poses a complex challenge: how does the embryo organize its physical structure of cells? The spatial distribution of cell states can be used to identify cell-cell interactions and tissue neighborhoods, which can provide insights into the developmental process. However, spatial origins are often lost during regular scRNA-seq profiling. Spatial transcriptomics is a technique that combines spatial information (the location of each cell) with transcriptomic information (the expression of genes in each cell), to resolve this problem. There are two main types of spatial transcriptomics methods: imaging-based and sequencing-based. Imaging-based methods, such as SeqFISH and MERFISH [28, 29], can be used to label individual genes in cells with fluorescent probes. This provides high-resolution spatial information, but multiplexing multiple genes can be challenging. Sequencing-based methods, such as Slide-seq and CITE-seq [30, 31], extract RNA from cells and then sequence it. This allows for multiplexing of many genes and is more widely used recently. Spatial transcriptomes have been used in many studies to help us understand embryogenesis. For example, Chen *et al.* used Stereo-seq to profile the spatial transcriptomes of 53 sections of mouse embryos at 8 timepoints from E9.5 to E16.5 [32]. Peng *et al.* collected GEO-seq samples from distinct spatial positions in the mouse embryo with mixed cell populations from E5.5, E6, E6.5, E7, and E7.5 [33].

Spatial transcriptomics data provide a coarse-grained view of gene expression in specific regions of embryos (*e.g.* at the voxel level). This information can be used to predict the spatial locations of cells profiled in a regular scRNA-seq study, which usually has a larger scale. There are two main types of computational tools used for spatial mapping: Deconvolution-based tools, such as Cell2location and CIBERSORTx [34, 35], first compute a gene expression signature for individual cell types, followed by using this signature to estimate the cell-type composition of each spatial voxel in the spatial transcriptome. Optimal transport-based tools, such as NovaSpaRc and Tangram [36, 37], globally map the individual cells and individual spatial voxels, by minimizing the transportation cost between the cells and the voxels. The first type of method is dependent on the cell-type definition itself, which can affect its robustness. The second type of method is more computationally expensive, but both have been shown to be effective in many studies.

1.5 Other omics data in embryo development

Other omics data at the single-cell level, such as chromatin accessibility, have been less characterized for embryos than transcriptomic data due to the technical challenges involved, especially for later stages. Argelaguet *et al.* developed a single-cell co-assay technique called scNMT-seq (combining transcriptome, methylome, and nucleosome) to profile 1,105 single cells from mouse embryos at four developmental stages: E4.5, E5.5, E6.5, and E7.5 [38]. By comparing the differential RNA expression, methylation, and accessibility of cells at different stages of development, the researchers identified the key regulators of the transition from pluripotency to the specification of the three primary germ layers. Pijuan-Sala *et al.* mapped chromatin accessibility in 19,453 single nuclei from mouse embryos at E8.25 [39]. This allowed them to identify cell-type-specific regions of open chromatin, which they then used to pinpoint two TAL1-bound endothelial enhancers. A larger scale study from Sarropoulos *et al.* created a single-cell atlas of chromatin accessibility in the mouse cerebellum,

spanning 11 developmental stages from E10 to P63 [40]. They identified cell-type-specific and timepoint-specific CRE (*cis*-regulatory element) activity, as well as exploring the conservation of CRE sequences during development. Overall, chromatin accessibility and other epigenetic information are essential for understanding the molecular mechanisms underlying cell type specification during embryogenesis. However, the available data and computational tools are still limited, making this an important research direction for the future.

1.6 Perturbation datasets for understanding molecular function

The scRNA-seq assay can be used to study the molecular functions of regulators that specify cellular trajectories by analyzing perturbation datasets. However, it is important to preserve replicate variation for individual cells in the perturbation datasets. This requires specific technology. In a recent pair of studies from the same group focused on zebrafish embryos, Saunders *et al.* applied scRNA-seq to profile 1812 individually resolved zebrafish embryos with 23 genetic perturbations (F0 knockouts, generated by CRISPR-Cas9 mutagenesis) [10]. The study is notable for its use of a specific technology (sci-Plex) that allows the researchers to save the sample identity for individual cells, enabling the detection of cellular changes between a large number of replicates with different conditions. Dorrity *et al.* applied the same technology to characterize hundreds of individual zebrafish embryos exposed to temperature stress in order to identify the cell types and molecular programs within them that drive phenotypic variation [41].

There are two main types of computational methods used to identify cellular changes in single-cell perturbation datasets: those that are based on annotated cell clusters and those that are not. In Saunders *et al.*'s study, they leveraged well-annotated cell clusters, performing beta-binomial regression to explore which cell clusters changed in composition between mutant and wildtype embryos. This strategy is powerful, but it has some limitations. It's

heavily dependent on the quality of cell-type annotations, necessitates an adequate number of replicates within each group, and may be less effective in detecting alterations in rare cell types or genetic perturbations with minimal effect sizes.

In contrast, some other methods do not require pre-annotated cell clusters. For example, Dann *et al.* developed Milo, a novel method that obviates the need for clustering, utilizes k-nearest neighbor graphs to model cellular states as overlapping neighborhoods. This allows for more accurate identification of perturbed states and enables complex experimental designs [42]. Burkhardt *et al.* developed a method called MELD that models single-cell experiments as points on a probability distribution over the underlying transcriptomic cell state space [43]. MELD quantifies the effect of an experimental perturbation as the change in the probability distribution between the experiment condition and the control condition. Lotfollahi *et al.* developed scGen, a model that combines variational autoencoders and latent space vector arithmetic to predict single-cell perturbation responses from high-dimensional gene expression changes [44]. Of note, methods that do not rely on cell annotations face two key limitations: firstly, their interpretation is considerably more challenging, and secondly, specifying contrasts can be particularly cumbersome, especially when aiming to test for differences between groups while controlling for extraneous factors such as time, temperature, or batch effects. In summary, investigating the single-cell perturbation dataset represents an active area of research, with a continuous influx of innovative computational methods being developed. [45, 46, 47]. It would be particularly interesting to connect cellular trajectory reconstruction with perturbation analysis.

1.7 Identifying conserved trajectories through cross-species comparisons

As described above, the availability of transcriptome profiles of embryos collected from a series of timepoints for different species provides a valuable opportunity to align developmental

trajectories across species. For example, Cardoso-Moreira *et al.* profiled the transcriptomes of seven organs (cerebrum, cerebellum, heart, kidney, liver, ovary, and testis) in seven species (human, rhesus macaque, mouse, rat, rabbit, opossum, and chicken) at different developmental stages, from early organogenesis to adulthood [48]. By comparing gene expression patterns at the organ level (bulk RNA-seq), they revealed correspondences of developmental stages across species, as well as differences in the timing of key events during the development of the gonads. In Briggs *et al.*'s study, they reconstructed the developmental trajectories of zebrafish and frogs during gastrulation and early organogenesis stages, followed by systematically identifying the aligned cell states and trajectories between the two species [9]. In Cao *et al.*'s study, they integrated scRNA-seq data from mouse early organogenesis (E9.5-E13.5) and human fetal stages (72 to 129 days) [18]. The developmental trajectories of the cells were aligned well in the co-embedding, suggesting that the two species share many similarities in early organogenesis. However, systematic analyses of multiple species, especially those with a wide evolutionary distance, are still limited. This may be due to the lack of effective computational strategies.

1.8 In vitro models for studying sell trajectories and regulators

In vitro models are becoming increasingly common in the study of mouse and human embryogenesis due to the practical challenges of collecting sufficient numbers of embryos from these species. For example, gastruloids are self-organizing 3D aggregates of pluripotent stem cells that resemble the early stages of embryonic development for either mice or humans. Van den Brink *et al.* applied scRNA-seq to analyze the gene expression profiles of 25,202 individual cells from 100 mouse gastruloids at 120 hours after aggregation, which were generated using the mouse embryonic stem (ES) cell lines E14-IB10 or LfngT2AVenus [49]. They identified 13 distinct cell types, arranged along an anterior-posterior axis from the spinal cord to the cardiac mesoderm, corresponding to the process of somitogenesis in the embryo. Amadei

et al. assembled stem cell-based embryo models *in vitro* from mouse embryonic stem cells (mESCs), trophoblast stem (TS) cells, and induced extraembryonic endoderm stem (iXEN) cells [50]. These embryo models developed in a similar way to natural mouse embryos in utero up to E8.5. The embryo model had a well-defined head, with forebrain and midbrain regions, a beating heart-like structure, a trunk with a neural tube and somites, and other features. *In vitro* models have a promising future for studying mouse or human embryogenesis, especially for the study of molecular functions by introducing perturbations to these synthetic models.

1.9 Limitations of reconstructing cellular trajectories from single-cell data

There are several general limitations to the current strategy of computationally reconstructing the cellular trajectories of mammalian development based on single-cell data.

- The trajectories are estimated indirectly, mainly based on identifying similarities of cell states across timepoints. This means that the results can be biased by the different replicates that are sampled, the different technologies that are applied, or the different cell-recovery depths.

- It is challenging to perform experiments with cells or nuclei collected from multiple timepoints in a single study, especially for large organisms like mice and humans. This means that data from different timepoints often need to be integrated from different studies or different experiment batches. Thus, batch effects need to be removed while preserving biological meaning. Some integration methods, such as Seurat [51] and BBKNN [52], have been shown to be effective in doing this. However, there is still room for improvement in developing more accurate and robust methods for single-cell data integration.

- Most methods for reconstructing lineage relationships use a systematic approach, in which the same method is applied to cells regardless of their stage or lineage. However, this may not always be effective for a particular example, especially for later developmental stages, where cell lineages can be complex. In these cases, it may be necessary to manually revise the results of the automated methods. For organisms other than *C. elegans*, it is still challenging to verify or benchmark the computational results of single-cell transcriptomics studies. The absence of a gold standard method for validation makes it difficult to compare and evaluate different approaches.

- Most current computational methods are only effective for relatively simple cell differentiation, such as the differentiation of stem cells or the major germ layers of an embryo (*i.e.* gastrulation). However, those methods may be difficult to use for reconstructing cellular trajectories of mouse or human embryogenesis, especially after early organogenesis.

- The similarity of transcriptomes between cells does not necessarily mean that the cells have a similar lineage. For example, in Cao *et al.*'s study [15], they observed that different epithelial cells, such as olfactory epithelial cells and gut epithelial cells, which have different developmental lineages, exhibit high transcriptional similarity and cluster together. This limitation can also be explained by technological limitations. Cells can usually only be captured by a single feature, such as transcriptional profiles. There are some studies that have used advanced technologies to profile cells with spatial-transcriptome or lineage-transcriptome, but these methods are limited in throughput or resolution.

1.10 Thesis outline

In summary, current single-cell datasets and computational tools provide valuable knowledge and insights into cellular trajectories and key regulators during embryogenesis. However, most studies have focused on model organisms such as flies and zebrafish. More complex organisms, such as mice and humans, have been less well-studied in a systematic way, resulting in a lack of comprehensive knowledge of their developmental trajectories. To address this, we have recently generated new data and developed new computational tools. In my thesis, I will introduce three different projects that focus on different mouse embryogenesis stages or different genotype backgrounds (wildtype or mutant type). These projects aim to study the two important questions mentioned earlier: how cells transition from earlier to later stages during mouse embryogenesis, and what the molecular mechanisms are that drive the cell differentiation process.

In **Chapter 2**, my colleagues and I developed and applied a computational strategy to reconstruct the cellular trajectories spanning mouse gastrulation and early organogenesis. Recent advances in single-cell technology have enabled the collection of multiple kinds of genomic information from large numbers of cells during mouse embryogenesis. I hypothesized that distances between cells in a low-dimensional embedding space, such as the UMAP learned from their transcriptional profiles, can be used to accurately relate cell states to one another across developmental time. To this end, I integrated several published scRNA-seq datasets that collectively span mouse gastrulation and early organogenesis, and supplemented them with a newly generated dataset to bridge technological gaps, to generate a graph of developmental cell states, by heuristically connecting these cell states to their pseudo-ancestors and pseudo-descendants. This tree-like graph describes the comprehensive cellular trajectories of early mouse embryogenesis, which can be used to better understand the distinct characteristics of every cell state and the cell state transitions across a series of successive

developmental stages. Then, I leveraged this graph to nominate TFs and TF motifs as key regulators of each branchpoint at which a new cell type emerges. Finally, I also applied the same procedures to scRNA-seq datasets of zebrafish and frog embryogenesis. By comparing the cell state trajectories of these three species, I nominated "cell type homologs" based on shared regulators and transcriptional states.

In **Chapter 3**, my colleagues and I used sci-RNA-seq3 to fill the large data gap between gastrulation and birth. We profiled the transcriptional states of 12.4 million nuclei from 83 precisely staged embryos spanning late gastrulation (E8) to birth (P0), with a temporal resolution of 2 hours during somitogenesis, 6 hours through to birth, and 20 minutes during the immediate postpartum period. From these data, we annotated hundreds of cell types and performed deeper analyses of the unfolding of the posterior embryo during somitogenesis, as well as the ontogenesis of the kidney, mesenchyme, retina, and early neurons. Finally, we leveraged the depth and temporal resolution of these whole-embryo snapshots, together with other published data, to construct and curate a rooted tree of cell-type relationships that spans mouse development from zygote to pup. Throughout this tree, we systematically nominated sets of TFs and other genes as candidate drivers of the *in vivo* differentiation of hundreds of mammalian cell types. Remarkably, the most dramatic shifts in transcriptional state were observed in a restricted set of cell types in the hours immediately following birth. These shifts presumably underlie the massive changes in physiology that must accompany the successful transition of a placental mammal to extrauterine life.

In **Chapter 4**, my colleagues and I set out to establish scRNA-seq of the whole embryo as a scalable platform for the systematic phenotyping of mouse genetic models. We applied sci-RNA-seq3 to profile 101 embryos of 22 mutant and 4 wildtype genotypes at embryonic stage E13.5, altogether profiling over 1.6M nuclei. The 22 mutants represent a range of anticipated severities, from established multisystem disorders to deletions of individual enhancers. We developed and applied several analytical frameworks for detecting differences in composition

and/or gene expression across 52 cell types or trajectories. We also identify differences between widely used wildtype strains, compare phenotyping of gain vs. loss of function mutants, and characterize deletions of topological associating domain (TAD) boundaries.

In the last chapter, I discuss some of the limitations of the three projects that I introduced. I also raised some future directions for research in this area.

Chapter 2

SYSTEMATIC RECONSTRUCTION OF THE CELLULAR TRAJECTORIES OF MOUSE EMBRYOGENESIS

This Chapter is adopted from published work with minimum changes:

Chengxiang Qiu #, Junyue Cao, Beth K. Martin, Tony Li, Ian C. Welsh, Sanjay Srivatsan, Xingfan Huang, Diego Calderon, William Stafford Noble, Christine M. Disteche, Stephen A. Murray, Malte Spielmann, Cecilia B. Moens, Cole Trapnell, Jay Shendure # "Systematic reconstruction of the cellular trajectories of mouse embryogenesis", *Nature Genetics*, 2022.

#: corresponding authors

I began this project in 2019 during my rotation in the Shendure lab. Jay proposed several candidate projects for me to choose from, and I ultimately selected the one with Jun. Jun had just published a major paper in *Nature*, profiling transcriptomes from 2 million nuclei in mouse embryogenesis. Some of the trajectories identified in the paper (*e.g.*, the epithelial trajectory) inspired us to systematically reconstruct the cellular trajectories of mouse embryogenesis and identify the key transcription factors specifying the trajectories. During my rotation, I diligently learned these biological concepts and tried different methods for data analysis. I had no prior knowledge of mouse development or single-cell data analysis, so I was initially stressed about the project. However, Jay was always encouraging and supportive, and he helped me to overcome my challenges.

After joining the lab, I continued working on this project, combining our data with other published datasets profiling cells of mouse embryos from the early stages. The project quickly expanded, and we were able to create a comprehensive roadmap of cell states for mouse embryogenesis from E3.5 to E13.5. We also extended our analysis to zebrafish and frogs, finding homologs of cell types across species. Finally, we wrote a paper and published it in *Nature Genetics*. Through this project, I gained a deeper understanding of the basic knowledge involved in mouse embryogenesis, such as how the three germ layers form and give rise to different types of lineages. I thoroughly enjoyed working with Jay on the manuscript, both during the initial submission and the revision process. He took every question seriously, no matter how seemingly unremarkable. For example, he carefully went through each pair of cell state ancestors and descendants, as well as the key TFs, to search for evidence and provide comments. I learned so much from this process, much more than just getting a publication. The scientific method is a process of rigorous inquiry and experimentation. It's important to respect this process and to avoid taking shortcuts. Working on this project made me love this research topic even more. Development is such an interesting and mysterious process. It has been studied for hundreds of years, yet there are still new findings being made every day.

More formally, the author contributions are listed in the manuscript as follows: C.Q., J.C. M.S. and J.S. designed the research. I.C.W. and S.A.M. collected and staged E8.5 mouse embryos. B.K.M. developed the optimized/simplified sci-RNA-seq3 protocol and applied it to these embryos. C.Q. and T.L. performed computational analyses. J.C., X.H., D.C., S.S., W.S.N., and C.T. assisted with data analysis. M.S., C.M.D., and C.B.M. assisted with results interpretation. C.Q. and J.S. wrote the paper, with input from all authors.

Abstract

Mammalian embryogenesis is characterized by rapid cellular proliferation and diversifi-

cation. Within a few weeks, a single cell zygote gives rise to millions of cells expressing a panoply of molecular programs, including much of the diversity that will subsequently be present in adult tissues. Although intensively studied, a comprehensive delineation of the major cellular trajectories that comprise mammalian development in vivo remains elusive. Here we set out to integrate several single cell RNA-seq datasets (scRNA-seq) that collectively span mouse gastrulation and organogenesis. To bridge technologies, these datasets were supplemented with new, intensive profiling of 150,000 nuclei from a series of E8.5 embryos in 1-somite increments with an improved combinatorial indexing protocol. Overall, we define cell states at each of 19 successive stages spanning E3.5 to E13.5, heuristically connect them to their pseudo-ancestors and pseudo-descendants, and for a subset of stages, deconvolve their approximate spatial distributions. Despite being constructed through automated procedures, the resulting trajectories of mammalian embryogenesis (TOME) are largely consistent with our contemporary understanding of mammalian development. We leverage TOME to nominate transcription factors (TF) and TF motifs as key regulators of each branch point at which a new cell type emerges. Finally, we apply the same procedures to single cell datasets of zebrafish and frog embryogenesis, and nominate "cell type homologs" based on shared regulators and transcriptional states.

2.1 Introduction

A fundamental goal of developmental biology is to understand the relationships of cell types to one another during embryogenesis, as well as the molecular programs that underlie each cell type's emergence. In principle, developmental programs can be comprehensively described, *e.g.* Sulston and colleagues' reconstruction of the complete embryonic lineage of the roundworm *C. elegans* through visual observation [2]. However, *C. elegans* – small, translucent, and developmentally invariant – remains the only model organism for which such a complete description has been realized.

Within the past five years, we and others have developed and applied new technologies for single cell molecular profiling to developing model organisms at the "whole animal" scale, including the worm, fly, zebrafish, frog, and mouse [13, 15, 8, 9, 7, 53]. Such studies lay the foundations for global views of metazoan development, including, for example, populating the Sulston lineage of *C. elegans* with the gene expression programs of each cell type [53, 5].

For mouse embryogenesis in particular, we and others have performed single cell or single nucleus RNA-seq data (scRNA-seq) during implantation [54, 55], gastrulation [13] and organogenesis [15]. Collectively, these four studies span the development of the mouse embryo from dozens of cells of a few types (E3.5) to millions of cells of hundreds of types (E13.5). However, the data associated with these studies has yet to be systematically integrated in a manner that permits their robust exploration. Such integration is challenging, both for technical reasons (*e.g.* different studies, different technologies, batch effects, *etc.*) as well as because of the sheer complexity of mouse development.

Here we set out to systematically reconstruct the major cellular trajectories of mammalian embryogenesis from E3.5 to E13.5. Our primary strategy is inspired by Briggs and colleagues [9] and makes several assumptions: 1) Although mouse development is variable, key patterns will be invariant across wild-type animals; 2) *"Omnis cellula e cellula"* also applies to cell states, *i.e.* cell states observed at a given timepoint must have arisen from cell states present at the preceding timepoint; 3) We are sampling frequently and deeply enough that newly detected cell states will not arise from antecedent cell states that were undetected at the preceding timepoint; 4) Provided that the interval between successive timepoints is small enough, transcriptional similarity is an effective means of linking related cell states observed at adjacent timepoints.

A caution is that in contrast to the Sulston's seminal map of *C. elegans*, we focus here on reconstructing cellular trajectories [56], a concept related, but by no means equivalent, to cell

lineage. Although it is a reasonable expectation that closely related cells (*e.g.* siblings) will be transcriptionally similar [5], the converse is not necessarily true. For example, molecular states can be insufficiently divergent, or even convergent, both of which obscure lineage relationships [57]. Furthermore, even the expectation that lineally closely related cells will be transcriptionally similar is not always met, as rapidly changing molecular states can lead to "gaps" in trajectories [5]. In sum, our goal here is a continuous, navigable roadmap of the molecular states of cell types during mouse development. Such a roadmap may constrain the potential lineage relationships amongst constituent cell types, but it does not explicitly specify them.

2.2 Intensive scRNA-seq of individual embryos during early somitogenesis

Relative to [9], a technical challenge that we faced is that the datasets that we sought to integrate were generated by different groups at different times using different scRNA-seq technologies (**Supplementary Table 1**). To address this, we performed anchor-based batch correction [51] prior to integration, which proved quite effective including across technologies (**Supplementary Fig. 1**). However, probably because there was no overlapping timepoint, the integration of scRNA-seq data generated at E8.5 (cells, 10X Genomics) and E9.5 (nuclei, sci-RNA-seq3) was particularly challenging. Numerous cell types appeared or disappeared between these timepoints [15, 13], and it was unclear which of these changes were due to technical differences vs. *bona fide* developmental progression (**Supplementary Fig. 2a**). It was similarly unclear whether changes in gene expression levels were technical or biological in nature. To address this, we set out to generate new data at E8.5 that might serve to "bridge" these two datasets (**Fig. 1a-b**).

Because of how quickly changes are occurring during this window of development, we focused on individual, somite-resolved E8.5 embryos using a simplified, optimized version of

sci-RNA-seq3 (**Methods**) [58]. We selected 12 embryos from 2 separate litters harvested at E8.5, including a single primitive streak stage embryo (prior to somitogenesis) and 11 embryos staged in 1-somite increments from 2 to 12 somites (**Fig. 1c**). Nuclei from each embryo were deposited to individual wells (8 wells per embryo) for sci-RNA-seq3, such that the first index identified the originating embryo of any given cell. This contrasts with [13], where multiple E8.5 embryos were pooled prior to profiling. The optimized sci-RNA-seq3 method markedly improved data quality, with 9-fold higher UMIs and 6-fold higher gene detection per nucleus, relative to (Cao *et al.* 2019) (**Supplementary Fig. 3a**). Even after deeper sequencing of the original libraries from (Cao *et al.* 2019) to a similar duplication rate, the improvement remained substantial (4-fold higher for UMI counts per nucleus; **Supplementary Fig. 3a**). After quality filtering, we obtained profiles for 154,313 somite-staged E8.5 nuclei (median UMI count 7,672; median genes detected 3,463) (**Supplementary Fig. 3b-c**).

Anchor-based batch correction and integration of profiles of E8.5 cells from [13] (termed "E8.5a") and newly generated profiles of E8.5 nuclei (termed "E8.5b") worked very well with the exception of primitive erythroid cells, which we suspect may be due to more extensive differences between cells vs. nuclei in this cell type (**Supplementary Fig. 2b**). As expected because they were generated on nuclei with the same technology, integration of E8.5b and E9.5 profiles also worked well (**Supplementary Fig. 2c**).

These new data, generated via optimized sci-RNA-seq3 at E8.5, enabled the identification of the same 30 cell types as we identified with E8.5 data from [13] (**Fig 1b; Supplementary Fig. 2; Supplementary Table 2**). However, the depth of the new data, together with the fact that we separately processed individual somite-resolved embryos, facilitated the resolution of substantial substructure. Examples include:

- **Floor Plate**: We observe two, clearly distinct subpopulations that express floor plate markers *Foxa2* and *Shh* (**Fig. 1b; Supplementary Fig. 4**) [59]. Although these

appear to be converging towards a common transcriptional state, an anterior subpopulation (*Bmp7+*) arises from the forebrain/midbrain, while a posterior subpopulation arises from the spinal cord [60].

- **Heart fields**: We observe subpopulations arising from the splanchnic mesoderm that correspond to the first (*Tbx5+*, *Hcn4+*) and second (*Isl1+*, *Tbx1+*) heart fields (**Fig. 1b; Supplementary Fig. 5**) [61, 62, 63, 64]. Similar to the floor plate, although these appear to converge towards a common transcriptional state, the heart fields remain distinguished by these and other markers throughout early somitogenesis.

- **Rhombomeres**: We observe four subpopulations of the hindbrain, as well as two additional subpopulations within cell types annotated as the midbrain and spinal cord, that appear to correspond to rhombomeres *r1* to *r6* (**Fig. 1b; Supplementary Fig. 6**). These annotations are based on distinct combinations of Hox markers and other genes. For example, rhombomeres 3 and 5 specifically express *Egr2*, while rhombomere 5 further expresses *Hoxa3*, *Hoxb3*, and *Mafb* [65, 66]. Each rhombomere includes cells from embryos spanning somitogenesis, consistent with roughly concurrent, rather than sequential, differentiation of these rhombomeres. However, a subset of cells from rhombomere 4 are from the earliest embryos of the series and express *Hoxa1* and *Hoxb1*, consistent with the possibility that rhombomere 4 begins to develop first (**Fig. 1d-e**) [67, 68]. Although we must be cautious about interpreting UMAP topologies, the rhombomeres are ordered along a rostral-caudal axis in relation to other major aspects of neuroectoderm regionalization, with *Wnt1* and *Nkx6-1* expression further marking dorsal and ventral regions, respectively (**Supplementary Fig. 6**) [69, 70].

- **Neural crest**: In the global embedding, we observe three distinct subpopulations of neural crest cells (NCC) that appear to derive from different subsets of neuroectoderm (**Fig. 1b**). Reanalysis with RNA velocity and examination of *Hox* gene expression suggests that these three populations may correspond to mesencephalic & pharyngeal

arch 1 (PA1) NCC; PA2 NCC; and PA3 NCC (**Fig. 1d; Supplementary Fig. 7**). Differential patterns of early neural crest marker expression (*e.g. Foxd3*) as well as their distribution in relation to somitogenesis are consistent with these subpopulations emerging asynchronously (**Fig. 1e; Supplementary Fig. 7**) [71].

We next sought to systematically explore the extent to which the transcriptional dynamics of individual cell types are coordinated with the timing of somite formation. For each cell type, we calculated the correlation between cells' somite counts and those of their five nearest neighbors in a global 3D UMAP embedding. In this framing, high correlations are consistent with rapid, "within cell type" changes in transcriptional state that are synchronized with somite counts. Consistent with our earlier analyses (**Fig. 1e; Supplementary Fig. 4c**), the highest such correlations were for neuroectodermal cell types (*e.g.* hindbrain, neuromesodermal progenitors (NMPs), floor plate, neural crest, *etc.*), rather than the somites themselves (**Fig. 1f**). Focusing on NMPs, whose heterogeneous states bridge paraxial mesoderm and spinal cord neuroectoderm, we observed that the top principal components of transcriptional variation are strongly correlated with mesodermal (*T* (*Brachyury*)+, *Tbx6*+) vs. neuroectodermal (*Sox2*+) state (PC1; 23.7% of variation), cell cycle index (PC2; 15.1% of variation) and somite count (PC3; 8.4% of variation) (**Supplementary Fig. 8; Supplementary Table 3**) [72, 73]. The genes most highly correlated with these PCs are shown in **Fig. 1g**. For example, key regulators of mesoderm (*T*) [74], the somite segmentation clock (*Hes7*) [75] and Wnt signaling (*Wnt3a*, *Rspo3*, *Ptk7*) [76, 77] are positively correlated with PC1, while regulators or effectors of neural adhesion or neurite outgrown (*Ptprz1*, *Nrcam*, *Ptn*) [78, 79, 80] as well as retinoic acid signaling (*Rarb*) are negatively correlated.

2.3 Systematic reconstruction of the cellular trajectories of mouse embryogenesis from E3.5 to E13.5

We collated published data from three studies spanning E3.5 to E8.5 [55, 54, 13], the new E8.5 data described above (**Fig. 1a-b**), and published data from one study spanning E9.5 to E13.5 but with deeper sequencing of those libraries, as also described above (**Supplementary Fig. 3**) [15]. Altogether, these data were derived from 480 samples (where each sample is an individual embryo or small pool of mouse embryos) from 19 timepoints or stages spanning E3.5 to E13.5, with successive stages separated by as few as 6 hours but no more than 1 day (**Supplementary Table 1**). The number of single cell or nucleus profiles used totalled 1,658,968 and ranged from 67 to 455,124 per stage (**Supplementary Fig. 9a-c**). For each stage, we performed pre-processing followed by Louvain clustering and manual annotation of individual clusters based on marker gene expression (**Supplementary Fig. 10; Supplementary Table 2**). Here we use "cell state" to mean an annotated cluster at a given stage. Altogether, we identified 473 cell states across the 19 timepoints, each of which received one of 94 cell type annotations.

For each pair of adjacent stages, we performed anchor-based batch correction followed by projecting cells into a shared embedding space [51]. After co-embedding, we applied a k-nearest neighbor (k-NN) based heuristic to connect cell states between adjacent stages. Briefly, for each cell state at the later timepoint, we identified the 5 closest cells from the antecedent timepoint in the co-embedding. Bootstrapping to obtain a robust estimate (500 iterations with 80% subsampling), we then calculated the median proportion of such neighbors derived from each potential antecedent cell state, and treated this as the weight of the corresponding edge. Because these are inferred relationships based on transcriptional similarity, analogous to pseudotime, we use the terms pseudo-ancestor and pseudo-descendant to refer to cell states at the immediately preceding or immediately following cell state to which

an edge has been drawn.

As a simple example, clustering and annotation of scRNA-seq data from two adjacent timepoints, E6.25 and E6.5, identified 5 and 6 cell states, respectively (**Fig. 2a**). If we co-embed these data and follow the aforedescribed procedure, we strongly link 5 cell states at E6.5 to 5 cell states bearing the same annotations at E6.25. The new cell state at E6.5, which corresponds to the primitive streak, is strongly linked to E6.25 epiblast, which we assign as its pseudo-ancestor (**Fig. 2a**). Upon applying this procedure to E6.5 - E6.75 and E6.75 - E7.0, the primitive streak is further assigned as the pseudo-ancestor of the nascent mesoderm, anterior primitive streak and primordial germ cells (**Supplementary Fig. 11**).

We applied this approach to each pair of adjacent timepoints (**Supplementary Fig. 12**; E8.5a and E8.5b were treated as distinct, adjacent timepoints). Although the resultant edge weights were bimodally distributed, a cutoff of 0.2 was selected towards being more inclusive of weaker relationships as well as to ensure connectivity of the overall graph (**Fig. 2b; Supplementary Fig. 9d-e**). Of note, we introduced 4 "dummy nodes", corresponding to morula at E3.0 (as a root for trophectoderm and inner cell mass), trophectoderm at E3.5 and E4.5 (which had been removed at these timepoints by immunosurgery [55]) and parietal endoderm at E6.75 (undetected, likely due to undersampling). For technical reasons (see above), we also introduced an edge between primitive erythroid cells at E8.5a and E8.5b. The resulting representation is a directed acyclic graph with 477 nodes and 577 edges that captures trajectories of mammalian embryogenesis (TOME) (**Fig. 2c**).

2.4 Do molecular trajectories recapitulate cellular phylogenies?

To reiterate, the graph shown in **Fig. 2c** (TOME) does not reflect cell lineage but rather relationships between cell states that were inferred on the basis of transcriptional similarity.

Nonetheless, under the supposition that lineally related cell states diverge from one another through a succession of continuous molecular states, we can ask whether or not established lineage relationships are respected by TOME. Of the 578 edges with weights greater than 0.2, 381 (66%) are between cell states bearing the same annotation, while 196 (34%) are between cell states bearing different annotations. In **Supplementary Table 4**, we show all edge weights and comment on inferred transitions. Several observations merit emphasis.

First, the graph largely respects germ layers, which are indicated by node colors in **Fig. 2c**. There are no edges between extraembryonic and embryonic cell states, and relatively few edges between embryonic cell states of different germ layers. Among the strongest edges that cross between germ layers "boundaries" are edges that connect E8.5 neural crest to two subtypes of E9.5 osteoblast progenitors, presumably corresponding to the well-established neural crest contribution to bones [81]; an edge between caudal lateral epiblast and a subset of paraxial mesoderm at E7.5 - E8.0, also previously described [82]; and multiple edges between epithelia derived from different germ layers (*e.g.* renal epithelium (mesoderm), gut and lung epithelium (endoderm), and branchial arch epithelium (surface ectoderm)), probably consequent to transcriptional similarity rather shared lineage [83, 15];

Second, 80% of cell types are strongly linked (edge weight > 0.7) to a single pseudo-ancestor when they first appear. These strong edges generally respect established lineage relationships, *e.g.* parietal and visceral endoderm arising from hypoblast [84], notochord and definitive endoderm arising from the anterior primitive streak [85, 86], the first and second heart fields successively arising from splanchnic mesoderm [87], and many others.

Third, apparent convergences — instances wherein we assign more than one pseudo-ancestor to a cell state — sometimes correspond to a given cell type persisting and "contributing" to another cell type over several consecutive timepoints (*e.g.* hemoendothelial progenitors are recurrently assigned as pseudo-ancestors of endothelial cells at E7.75 - E8.25). In

other cases, apparent convergences may reflect incomplete separation between highly related cell types, rather than ongoing differentiation (*e.g.* the several edges between notochord and definitive endoderm; recurring edges between different subtypes of mesoderm). However, yet other cases reflect bonafide convergence of transcriptional states, *i.e.* where a cell type has multiple origins. For example and as also noted above, the two subtypes of E9.5 osteoblast progenitors have edges back to both E8.5 neural crest and E8.5 paraxial mesoderm, consistent with the literature [81], while a subtype of paraxial mesoderm has edges back to nascent mesoderm and caudal lateral epiblast [82]. Of note, not all established convergences are captured, *e.g.* the known contribution of embryonic visceral endoderm to the gut at E7.5-E7.75 [88] is detected at E7.5-E7.75, but falls short of the 0.2 edge weight threshold (**Supplementary Table 4**).

Fourth, an important limitation of this heuristic approach, made apparent by a few clear inaccuracies in the graph, is that true lineage relationships for a given cell state can be obscured by the presence of a highly similar cell state at the preceding timepoint. For example, E9.5 neuron progenitor cells are assigned as the pseudo-ancestor of multiple neuronal subtypes that appear at E10.5, but we do not observe these same relationships to recur at subsequent timepoints, although neuronal differentiation is surely ongoing. This is probably because at timepoints subsequent to E10.5, each derivative neuronal subtype is most similar to itself at the preceding timepoint, such that it fails to be linked back to the persisting neuron progenitors. This same phenomenon probably explains another error, wherein when definitive erythroid cells first appear at E10.5, they are linked to E9.5 blood progenitors (expected) but also to E9.5 primitive erythroid cells (unexpected). Another example involves motor neurons, which are most closely related to the hindbrain and spinal cord when E9.5 is looked at in isolation (expected), but to the forebrain/midbrain when integrating with E8.5 (unexpected). In this case, the error would likely require sampling at higher temporal resolution in order to correct. For a more exhaustive consideration of the ways in which trajectory-based inference can be misleading about cell lineage histories, see [57]. Of note,

at least some of the inaccuracies noted above are resolvable by focused analyses that leverage the distinction between nascent and spliced transcripts, *i.e.* RNA velocity [24]. For example, if we reanalyze these problematic subsets of TOME with scVelo [89], the heterogeneity and ongoing contributions of neuron progenitors are more evident [90] (**Fig. 3a**), and primitive and definitive hematopoiesis are much more clearly separated (**Fig. 3b**). To approach this more systematically, we calculated edge weights between cell states at adjacent timepoints with an alternative heuristic that was based on RNA velocity (**Methods**). We observed that out of 515 edges with weights > 0.2 that were nominated by the k-NN strategy, 392 had velocity-based transition probabilities > 0.2 (76%) (**Supplementary Fig. 13-14; Supplementary Table 5**). However, there were also 123 edges nominated by the k-NN strategy only, and 75 edges nominated by the RNA velocity strategy only (**Supplementary Fig. 14c**). Although we may assign greater confidence to edges nominated by both methods, edges supported by one method or the other may include both true and false positives. As an example of a likely true positive supported by RNA velocity only, the connection between embryonic visceral endoderm (E8.0) and gut (E8.25), fell short of the edge threshold by the k-NN strategy (weight 0.14) but was strongly supported by the RNA velocity strategy (weight 0.96).

Fifth, a further limitation is that our reliance on discrete entities, *i.e.* cell states, obscures aspects of developmental biology that are inherently continuous. For example, spatial transcriptional heterogeneity, which often manifests as continuous gradients, is obscured by cell type or cell state discretization. Here, we have sometimes represented aspects of spatial heterogeneity in a limited way through distinct nodes, though this is far from ideal. Although it is challenging to reduce to a graph-based representation, continuous aspects of spatial heterogeneity might be retained in co-embeddings across timepoints. For example, for neural tube-derived cells from E8.5b and E9.5, the co-embedding is potentially informative in both directions (*e.g.* to identify the subset of E8.5 diencephalon cells which are most related to E9.5 retinal primordium; or the subsets of E9.5 hindbrain cells which are most related to

specific E8.5-annotated rhombomeres) (**Fig. 3c**).

In summary, molecular trajectories often recapitulate well-documented cellular phylogenies, but there are clear limitations. Nonetheless, the graph is largely consistent with our contemporary understanding of mammalian development, despite being constructed through automated procedures. To facilitate its exploration, we created an interactive website in which the nodes and edges shown in **Fig. 2c** can be navigated (http://tome.gs.washington.edu).

2.5 Inference of the approximate spatial locations of cell states during mouse gastrulation

The spatial relationships of cells are a crucial aspect of development, but this information is lost while profiling disaggregated cells or nuclei. Towards addressing this, several groups have developed *in silico* methods for integrating scRNA-seq data with spatially resolved gene expression profiles obtained by fluorescence in situ hybridization (FISH) or other means [91, 92]. Here we sought to leverage data recently generated by Peng and colleagues, who applied cryosectioning and bulk RNA-seq (GEO-seq) to obtain spatially resolved transcriptomes for precise territories of the mouse embryo from E5.5 to E7.5 [33]. Inspired by an analysis by Peng *et al.* estimating the regionalization of endodermal subclusters across E7.0 GEO-seq territories, we leveraged TOME to estimate the abundance of individual cell types within each GEO-seq territory (**Fig. 4a; Supplementary Fig. 15 & 16; Supplementary Table 6**) [35]. For many cell types and territories, this approach appeared to work quite well. For example, the GEO-seq territories inferred to be composed of rostral and caudal neuroectoderm, caudal lateral epiblast, and surface ectoderm are clearly distinguishable at E7.5, in a pattern consistent with expectation (**Fig. 4b**) [93]. Also at E7.5, what we had annotated prior to this analysis as different subsets of paraxial mesoderm (A & B) are also regionalized to the anterior and posterior embryo, respectively (**Fig. 4c**). Finally, we observe the antici-

pated convergence of embryonic visceral endoderm and definitive endoderm cells during gut development, although the overlap is not complete [88] (**Fig. 4d; Supplementary Fig. 16**).

2.6 Inferring the molecular histories of individual cell types

We next sought to infer continuous expression levels for individual genes over the course of each cellular trajectory, focusing on derivatives of the epiblast from E6.25 onwards. First, we leveraged the fact that individual embryos do not correspond precisely to their intended timepoints. Using pseudotime, we ordered the pseudobulk expression profiles of individual embryos (or pools of embryos comprising each sample, in the case of [13]). The resulting ordering, which is robust to downsampling, corresponds well with developmental age but may additionally distinguish earlier vs. later individuals/pools at each intended timepoint (**Supplementary Fig. 17**).

Next, for each epiblast-derived cell type that was detectable at E13.5, we calculated a smoothed expression profile along its inferred history, as illustrated in **Supplementary Fig. 18** for selected genes in one cell type from each germ layer. Despite including the data source as a covariate, these inferred trajectories remained modestly confounded by batch effects across E8.5a - E8.5b, *i.e.* the switch from cell-based 10X Genomics data to nucleus-based sci-RNA-seq3 data (**Supplementary Fig. 19**). Nonetheless, at least anecdotally, TFs with established roles in a given cell type were often upregulated in association with its first appearance (**Supplementary Fig. 18**).

2.7 Systematic nomination of key transcription factors for cell type specification

Inspired by these anecdotal examples, we next sought to systematically identify TFs that are strong candidates for specifying each newly emerging cell type throughout early mammalian development [94, 95]. First, we collated 1,636 mouse proteins that are putative TFs from AnimalTFDB/v3 [96]. Then, for each branchpoint in TOME at which a given cell type first emerged, we heuristically defined key TF candidates as those: 1) significantly upregulated in the newly emerged cell type, relative to the pseudo-ancestor; 2) detected in at least 10% of cells in the newly emerged cell type; and 3) not significantly upregulated at any "sister" edges, relative to the newly emerged cell type (**Fig. 5a**). For each such key TF candidate, we calculated a normalized score based on the fold-difference of its expression between the new cell type and its ancestor/sister(s).

Altogether, we identified 632 candidate key TFs associated with the emergence of one or more of 92 cell types (27 +/- 18 per cell type; **Fig. 5b; Supplementary Table 7**). 49% of candidate key TFs were specific to one or two cell types. For example, *Gsc* (*goosecoid*) was identified as a candidate key TF for the emergence of the anterior primitive streak, but no other cell type, and *Srf* for the first heart field, but no other cell type [97, 98, 99]. On the other hand, a few TFs, such as *Meis2* and *Dach1*, were associated with the emergence of dozens of cell types (**Fig. 5c**). In **Fig. 5d**, we show the top scoring candidate key TFs for selected trajectories. Despite our automated approach that relied on a handful of datasets, many of these TFs are established as playing critical roles in the emergence of the corresponding cell types. For example, for the first heart field, the top 3 TFs identified are *Nkx2-5*, *Mef2c*, and *Gata5* [100, 101, 102]; for notochord, *Foxj1*, *T* (*Brachyury*), and *Noto* [103, 104, 105]; for neural crest, *Sox9*, *Msx1*, and *Id2* [106, 107, 108, 109]; and for hematoendothelial progenitors, *Etv2*, *Tal1*, and *Gata2* [110, 111, 112]. In fact, when we

performed a brief search for literature corresponding to the top 5 TFs nominated for each cell type, we found relevant references for 494 of 533 (93%) of them (**Supplementary Table 8**).

Multilineage priming (MLP) has extensively been documented in hematopoietic lineages and more recently in *C. elegans* [5, 113]. As one form of MLP, we also sought to identify TFs whose reduced expression was associated with cell type emergence, which we defined as those: 1) detected in at least 10% of cells in the pseudo-ancestor; 2) significantly downregulated in the newly emerged cell type, relative to the pseudo-ancestor; and 3) both detected in at least 10% of cells and not significantly downregulated at any "sister" edges, relative to the newly emerged cell type. Altogether, we identified 482 candidate key TF whose reduced expression is associated with the emergence of one or more of 90 cell types (23 +/- 26 per cell type; **Supplementary Table 9**). For example, at the split from inner cell mass to epiblast and hypoblast at E4.5, *Gata6* and *Nanog* are identified as decreasing in the respective emergence of the epiblast and hypoblast, consistent with the literature [114, 115]. Also, *Pou5f1* (*Oct4*) is identified as a key TF with reduced expression in association with 20 cell types, but increased expression with only 1, consistent with its established role in stemness (**Supplementary Fig. 20a**) [116, 117]. In sharp contrast, *Nfia* and *Nfib* (nuclear factors I/a and I/b) are nominated as key TFs at the emergence of 15 and 11 cell types, respectively, but in all cases upregulated, consistent with broad roles in lineage progression [118, 119].

2.8 Identification of core promoter cis-regulatory motifs involved in in vivo cell type specification

Although single cell chromatin accessibility profiling (*e.g.* sc-ATAC-seq) is increasingly enabling the ascertainment of *cis*-regulatory programs in embryonic and fetal tissues [120, 121,

39], such data is not yet available for a dense timecourse of early development for any of the three species considered here. As a step forward with sc-RNA-seq data alone, we sought to identify DNA sequence motifs that are enriched in the core promoters of developmentally regulated genes in TOME. First, we extended the approach described above to nominate key TFs whose upregulation or downregulation is associated with the emergence of each cell type, to all genes. Altogether, this yielded 8,307 key genes associated with the emergence of one or more of 92 cell types (470 +/- 433 per cell type; **Supplementary Fig. 20b; Supplementary Table 10**). Second, for each cell type, we applied HOMER [122] to discover DNA sequence motifs that are specifically enriched in the core promoters of key genes (-300 to +50 bp of annotated TSSs). Finally, we estimated q-values for discovered motifs by data label permutation. At an FDR of 10%, we implicated 119 de novo promoter motifs in the emergence of 57 mouse cell types (**Supplementary Table 11**), as well as an additional 235 previously documented promoter motifs (some overlapping with the de novo set) in the emergence of 34 mouse cell types (**Supplementary Table 12**).

We then asked whether the sequence motifs identified in the core promoters of developmentally regulated genes correspond to the binding sites of candidate key TFs for the same cell types, which would provide a plausible confirmation of their role. We identify 38 such instances, 33 of which are positive correlations (*i.e.* consistent directionality between TF expression and target gene expression) and 5 of which are negative correlations (**Supplementary Table 13**). For example, the transcriptional activator *Rfx3* is sharply upregulated at the emergence of the notochord at E7.25, and its cognate motif is strongly enriched at the promoters of key genes upregulated in these same cells (**Supplementary Fig. 21a-c**) [103, 123]. In contrast, the transcriptional repressor *Snai1* (*Snail*) is upregulated at the emergence of nascent mesoderm at E6.75, but its cognate motif is strongly enriched in the promoters of downregulated key genes (**Supplementary Fig. 21d-f**) [124, 125]. Interestingly, RFX3 motifs are very strongly enriched near the TSSs of notochord-upregulated genes, while SNAIL1 motifs are more diffusely enriched across the promoters of

nascent mesoderm-downregulated genes (**Supplementary Fig. 21b,e**).

A limitation of these analyses is that we restricted our search for enriched sequence motifs to the core promoters of up- or down-regulated key genes. As single cell, genome-wide chromatin accessibility datasets spanning mouse development are generated, such analyses can be extended to enhancer-mediated regulation.

2.9 Comparison of the cellular trajectories of mouse, zebrafish and frog embryogenesis

The origins and evolution of vertebrate cell types are fascinating topics on which the single cell profiling of embryogenesis may shed much needed light [126]. However, even if we adopt an evolutionary definition of cell types, it remains unclear how best to identify "cell type homologs" across vast evolutionary distances. To facilitate the alignment of cell types across vertebrates, we applied the same strategy used for TOME to zebrafish (*D. rerio*) and frog (*X. tropicalis*) embryogenesis, again relying on publicly available single cell RNA-seq datasets. For zebrafish, we integrated data from two studies that used different technologies but together included 15 developmental stages, beginning at the high stage (hpf 3.3) and ending at the early pharyngula stage (hpf 24), essentially spanning epiboly and segmentation (**Fig. 6a; Supplementary Table 1**) [8, 7]. The resulting graph contains 221 nodes, each assigned one of 63 cell type annotations, and 257 edges with weights greater than 0.2 (**Fig. 6b**). Marker genes used to annotate cell types are provided in **Supplementary Table 14**, and all edge weights in **Supplementary Table 15**. We also nominated key upregulated and downregulated TFs using the same approach described for mouse development above (**Supplementary Tables 16-17**).

For frog, we re-analyzed one dataset spanning 10 developmental stages, from S8 and

S22 [9], spanning gastrulation and neurulation (**Fig. 6a; Supplementary Table 1**). The resulting graph contains 192 nodes, each assigned one of 60 cell type annotations, and 221 edges with weights greater than 0.2 (**Fig. 6c**). Marker genes used to annotate cell types are provided in **Supplementary Table 18**, all edge weights in **Supplementary Table 19**, and candidate key TFs in **Supplementary Tables 20-21**.

We next sought to align cell types from each species to their "cell type homologs" in the other two species. Because *M. musculus* is separated from *D. rerio* and *X. tropicalis* by 450 million and 360 million years of evolution, respectively, these alignments proved much more challenging than integrating data from more closely related species such as mouse and human [18, 127]. We attempted three strategies.

As a first strategy, treating cells of each state from each timepoint as a "pseudo-cell", we integrated data from all three species with anchor-based batch correction [51]. Within the resulting UMAP co-embedding of 825 pseudo-cells, we could identify 15 major groups — epiblast & germline, early gastrulation, neuroectoderm, surface ectoderm & epithelium, mesoderm, notochord & notoplate, endoderm & gut, retinal primordium, neural crest, brain & spinal cord, neurons, endothelium, myocytes & cardiomyocytes, white blood cells and erythroid cells — each containing cell states from all three species (**Supplementary Fig. 22**). However, within each such major group, the homology between specific cell types generally remained ambiguous.

As a second strategy, we performed all possible pairwise comparisons between the transcriptomes of cell types of each pair of species, excluding extraembryonic lineages [128]. First, we performed cell type correlation analysis [15], which uses a regression framework to ask, between each pair of species, which cell types are the best reciprocal best matches to one another (**Supplementary Fig. 23; Supplementary Table 22; Methods**). We then manually reviewed the highest ranking cell type pairings for biological plausibility. For mouse

vs. zebrafish, out of 5,133 pairings tested, 138 were highly ranked, of which we selected 44 as the most biologically plausible (**Supplementary Table 23**). Exclusion criteria included the cell types arising from different germ layers or major groups (as defined in **Supplementary Fig. 22**), arising at very different temporal stages, or if a cell type was exclusive to one species. In cases where multiple related matches were observed, we generally selected the match with the highest beta score. Applying this same approach to mouse vs. frog and zebrafish vs. frog, we identified 28 and 48 plausible cell type homologs pairings, respectively (**Supplementary Fig. 24a; Supplementary Table 23**).

As a third strategy, we focused on overlaps between the candidate key TFs associated with the emergence of each cell type in each species. For each possible interspecies pairing of cell types, we identified orthologous TFs that were nominated in both, and then adopted a permutation approach to identify instances in which an excess of orthologous candidate key TFs were shared between the cell types. For mouse vs. zebrafish, out of 5,046 pairings tested, 75 exhibited more sharing than over 99% of permutations, of which we retained 25 as the most biologically plausible (**Supplementary Table 24**). Applying this same approach to mouse vs. frog and zebrafish vs. frog, we identified 18 and 10 plausible cell type homolog pairings, respectively (S**upplementary Fig. 24b; Supplementary Table 24**).

Some candidate cell type homologs overlapped between these second and third strategies (**Fig. 7a; Supplementary Table 25**). Overall, we were able to assign at least one cell type homolog to 52 of 87 embryonic mouse cell states, 49 of 59 zebrafish embryonic cell states, and 45 of 60 frog embryonic cell states. Some loosely annotated cell types were resolved through homology. For example, zebrafish *eomesa*+ and *dlx1a*+ differentiating neurons were homologous to mouse intermediate progenitor cells and inhibitory interneurons, respectively. In certain cases, we observed "three way" pairwise homology and nominated regulators (**Fig. 7b**). For example, *Gsc*, a canonical TF of the Spemann organizer [129], was nominated as a key regulator of the anterior primitive streak (mouse), dorsal margin involuted (zebrafish),

and dorsal marginal zone (frog), cell types that were also identified as homologs of one another. Other such "three way" nominated TF regulators and associated cell types include *Sox7* for haemogenic endothelium [130], *Tbx2* for the otic placode [131, 132] and *Six1* for myocytes [133] (**Fig. 7b**).

2.10 Discussion

Nearly forty years ago, Sulston and colleagues painstakingly mapped out the entirety of the invariant embryonic cell lineage of *C. elegans*, comprising 671 cells [2]. The Sulston map provided a foundational scaffold for the integration of future experimental results, as well as a precise nomenclature for the discussion of specific subsets of cells within the developing worm. Recently, Packer and colleagues intersected the Sulston lineage with the mRNA profiles of the same cells, shedding fresh light on the relationship between cell states and fates [5].

Can equivalently global views of development be achieved for the developing mouse? For reasons including scale, complexity, variance and accessibility, this is an extraordinary challenge and one that may take decades to fully come to fruition, if indeed it ever does. However, given the pace at which relevant technologies are emerging and evolving, it feels increasingly possible.

Here, towards a scaffold for such an undertaking, we sought to leverage recently published and newly generated single cell RNA-seq data to construct a "roadmap" of molecular trajectories that cells traverse during the peri-implantation, gastrulation and organogenesis stages of mouse development (**Fig. 2**). Our approach — constructing a directed acyclic graph wherein each node corresponds to a group of related cells at a given timepoint, and each edge to similarity between groups observed at adjacent timepoints — is highly reduc-

tionist. However, we believe that this framing provides a useful entry point for analyses that benefit from a global view of developmental processes. For example, in addition to nominating specific TFs as key regulators of the initial emergence of each cell type, we are able to systematically assess which TFs and genes appear to have relatively specific vs. general roles in development (**Fig. 5; Supplementary Fig. 20a-b**), as well as other characteristics (*e.g.* upregulated key TFs are associated with broad H3K27me3 domains; **Supplementary Fig. 20c**) [134]. Furthermore, by constructing developmental graphs for additional vertebrate species through the same method, we can identify "cell type homologs" through approaches that consider all cell types in each pair of species, analogous to the comparison of genomes (**Figs. 6-7**). Of note, the set of apparent cell type homologs was noisy prior to manual filtering; fully automating these assignments remains an outstanding goal.

Although "cell type" is a useful concept, a limitation of this terminology is that it obscures continuous aspects of heterogeneity — *e.g.* as might be expected for spatial gradients or during the maturation of a cell type. This framing also forces us to make decisions about the level of resolution at which to define cell types. Although we have made some progress in relating TOME to spatial information (*e.g.* **Fig. 3c; Fig. 4**), new nomenclature that facilitates the discussion of precise subsets of cells within spatially or otherwise heterogeneous cell types is sorely needed.

Finally, as discussed above, molecular trajectories are not equivalent to cellular phylogenies, but are likely to constrain them. Indeed, the cell type relationships inferred here on the basis of gene expression are largely consistent with our understanding of the bona fide ancestors and descendants of cell types in mouse development. Exceptions and ambiguities may represent errors that can be clarified through deeper analysis (**Fig. 3**), or potentially novel relationships that require validation. To that end, in vivo, organism-scale lineage recording, originally developed in zebrafish, has recently been adapted to the mouse [3, 135, 136, 137]. Although such methods remain very far from delivering anything approaching the resolution

of the Sulston lineage, they can likely be applied in their current form to support or reject potential lineage relationships suggested here. Additionally, even within cell states, such lineage data might shed light on the patterns of cell division that underlie the differentiation and proliferation of each cell type.

TOME provides a scaffold onto which additional single-cell gene expression datasets from mouse development can be further layered. In this vein, the remarkable consistency and robustness of in vivo mouse development is a terrific feature, and contrasts with in vitro differentiation time courses, which may vary by lab, operator, cell line, *etc.* Of note, by profiling individual embryos staged around E8.5, we observe dramatic changes in gene expression for some cell types (*e.g.* hindbrain, NMPs) occurring within short periods of time (1 somite or 2 hour increments). Particularly at later timepoints, it remains possible that our daily temporal sampling compromises our assumption that antecedent states will be relatable to descendent states. As such, as mouse development is further profiled, improving temporal resolution should be a high priority. Although it remains unclear exactly how frequent sampling needs to be in order to fully mitigate confounding by rapid transitions in transcriptional states, near-term goals that we are pursuing include consistent sc-RNA-seq sampling of mouse development at least every 6 hours, from fertilization to birth. We also anticipate additional single cell data types (*e.g.* chromatin accessibility, methylation, histone modifications, transcription factor binding, *etc.*) can be generated by independent groups and/or technologies and layered onto TOME as well. We are particularly excited about the possibility of linking the temporal unfolding of combinatorial TF expression to enhancer accessibility, and then enhancer accessibility to the expression of *cis*-regulated genes.

2.11 Supplementary Materials

Data reporting

For newly generated E8.5b data, no statistical methods were used to predetermine sample size. Embryos used in experiments 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 different researchers in different locations.

Generating new E8.5 data using an optimized version of sci-RNA-seq3

For newly generated E8.5b data, C57BL/6 mice were obtained at The Jackson Laboratory. In brief, timed matings of mice were performed via standard husbandry procedures. On the morning of E8.5, individual decidua were removed and placed in ice cold PBS during the harvest. Individual embryos were dissected free of extraembryonic membranes, imaged, and the number of somites present were noted prior to snap freezing in liquid nitrogen (**Fig. 1c**). Samples were stored at -80C until further processing.

We performed a simplified version of sci-RNA-seq3, further optimized for "tiny" samples [15]. Briefly, to each tube, 100ul of a hypotonic, PBS-based lysis buffer was added with DEPC as an RNase inhibitor. The resulting nuclei were then fixed with 4 volumes of a mix of methanol and dithiobis (succinimidyl propionate) (DSP). After rehydrating and washing the nuclei carefully in a sucrose/PBS/triton buffer (SPBST), the nuclei were distributed to a 96-well plate for reverse transcription, allocating 8 wells per embryo. After reverse transcription, nuclei were pooled, washed in SPBST and redistributed to a fresh plate for ligation of the second index primer with T4 DNA ligase. Nuclei were then again pooled, washed, and redistributed to 5 final plates for second strand synthesis, extraction, tagmentation, and PCR to add the third index plus a plate index. Products were pooled by PCR plate, size-selected and sequenced on an Illumina NovaSeq. A more detailed version of the streamlined, "tiny" sci-RNA-seq3 protocol is available at [58].

Processing of sequencing reads of new E8.5 data

For newly generated E8.5b data, read alignment and gene count matrix generation was performed using the pipeline that we developed for sci-RNA-seq3 [15] with minor modifications: base calls were converted to fastq format using Illumina's bcl2fastq/v2.20 and demultiplexed based on PCR i5 and i7 barcodes using maximum likelihood demultiplexing package deML [138] with default settings. Downstream sequence processing and single cell digital expression matrix generation were similar to sci-RNA-seq [53] 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 (Levenshtein edit distance (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/v0.6.5 with default settings. Trimmed reads were mapped to the mouse reference genome (mm10) for mouse embryo nuclei, using STAR/v2.6.1d [139] with default settings and gene annotations (GENCODE VM12 for mouse). 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, hairpin ligation adaptor index and read 2 end-coordinate (*i.e.* reads with UMI sequence less than 2 edit distance, 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). 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 [140]. 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.

After the single cell gene count matrix was generated, cells with low quality (UMI $<$

200 or detected gene $<$ 100 or unmatched_rate $\geq$ 0.4) were filtered out and 239,533 cells were left. Each cell was assigned to its original mouse embryo on the basis of the reverse transcription barcode. For the detection of potential doublet cells, we first split the dataset into subsets for each individual, and then applied the scrublet/v0.1 pipeline [27] 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 scores over 0.2 were annotated as detected doublets. We detected 2% potential doublet cells in the whole data set.

For detection of doublet-derived subclusters for cells, we used an iterative clustering strategy based on Scanpy/v.1.6.0 [141]. Briefly, gene count mapping to sex chromosomes were removed before clustering and dimensionality reduction, and then genes with no count were filtered out and each cell was normalized by the total UMI count per cell. The top 1,000 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 UMAP, followed by Louvain clustering performed on the 30 principal components with default parameters. For Louvain clustering, we first fitted the top 30 PCs to compute a neighborhood graph of observations with local neighborhood number of 50 by scanpy.pp.neighbors. We then cluster the cells into sub-groups using the Louvain algorithm implemented as scanpy.tl.louvain function. For UMAP visualization, we directly fit the PCA matrix into scanpy.tl.umap function with min_distance of 0.1. For subcluster identification, we selected cells in each major cell type and applied PCA, UMAP, Louvain clustering similarly to the major cluster analysis. Subclusters with a detected doublet ratio (by Scrublet) over 15% were annotated as doublet-derived subclusters.

For data visualization, cells labeled as doublets (by Scrublet) or from doublet-derived subclusters were filtered out. For each cell, we only retain protein-coding genes, lincRNA

genes and pseudogenes. Genes expressed in less than 10 cells and cells in which fewer than 100 genes were detected were further filtered out. The downstream dimension reduction and clustering analysis were done with Monocle/3-alpha. The dimensionality of the data was reduced by PCA (50 components) first on the top 5,000 most highly dispersed genes and then with UMAP (max_components = 2, n_neighbors = 50, min_dist = 0.01, metric = 'cosine'). Cell clusters were identified using the Louvain algorithm implemented in Monocle/3 (res = 1e-06). We found that the above Scrublet and iterative clustering based approach is limited in marking cell doublets between abundant cell clusters and rare cell clusters (*e.g.* less than 1% of total cell population). To further remove such doublet cells, we took the cell clusters identified by Monocle/3, downsampled each cell cluster to 2,500 cells, and computed differentially expressed genes across cell clusters with the top_markers function of Monocle/3 (reference_cells=1000). We then selected a gene set combining the top ten gene markers for each cell cluster (filtering out genes with fraction_expressing < 0.1 and then ordering by pseudo_R2). Cells from each main cell cluster were selected for dimension reduction by PCA (10 components) first on the selected gene set of top cluster specific gene markers, and then by UMAP (max_components = 2, n_neighbors = 50, min_dist = 0.1, metric = 'cosine'), followed by clustering identification using the Louvain algorithm implemented in Monocle/3 (res = 1e-04 for most clustering analysis). Subclusters showing low expression of target cell cluster specific markers and enriched expression of non-target cell cluster specific markers were annotated as doublets derived subclusters and filtered out in visualization and downstream analysis. We further filtered out the potential low-quality cells by investigating the numbers of UMIs and the proportion of reads mapping to the exonic regions per cell (**Supplementary Fig. 3b-c**), resulting in a set of 154,313 cells (median UMI count 7,672; median genes detected 3,463) that were used for reconstructing cellular trajectories.

Deeper sequencing of previously reported libraries (E9.5 - E13.5)

To obtain higher quality data across E9.5-E13.5, we performed a deeper sequencing

(specifically, three additional Novaseq runs) of previously reported libraries [15]. We merged the new reads with the previous reads and performed the same strategy of data-processing that we applied to the newly created E8.5 data. After the single cell gene count matrix was generated, cells with low quality (UMI < 200 or detected gene < 100 or unmatched_rate ≥ 0.4) were filtered out and 2,432,186 cells remained. Compared to the previous data [15], the median UMI count per cell improved from 671 to 1,434, while the median genes detected per cell improved from 518 to 735 (**Supplementary Fig. 3a**).

Each cell was assigned to its original mouse embryo on the basis of the reverse transcription barcode. After removing doublets, we further filtered out potential low-quality cells based on UMI counts and the proportion of reads mapping to the exonic regions per cell (**Supplementary Fig. 3b-c**), resulting in 1,393,565 cells (median UMI count 1,744; median genes detected 851) that were used for reconstructing cellular trajectories.

Decoding the transcriptional heterogeneity of NMP cells

To systematically identify cell types whose transcriptional dynamics are most highly correlated with somite counts, we first manually excluded cell types with fewer than 100 cells, and then for each cell type, we calculated the Pearson correlation between cells' somite counts and those of their top 5 nearest neighbors in a global 3D UMAP embedding.

We applied two different strategies to identify the genes (among the top 5,000 highly variable genes) which were significantly correlated with the top 3 PCs of NMP cells. As the first strategy, we performed a generalized linear regression using the fit_models function in Monocle/3 across the NMP cells. As the second strategy, we performed a Pearson regression between each individual PC and the gene expression values, which were calculated from original UMI counts normalized to total UMI per cell, followed by natural-log transformation. The PCs were calculated on NMP cells only. The significant results (FDR < 0.05 and absolute

coefficients > 0.2 by Pearson correlation) are shown in **Supplementary Table 3**.

Systematic reconstruction of the cellular trajectories of mouse embryogenesis

Single cell or single nucleus RNA-seq data were collected from three studies from other labs [55, 39, 54] and supplemented with the new E8.5 data ("E8.5b") as well as data from [15] but supplemented and reanalyzed after deeper sequencing of the same libraries, as described above. These data span 19 timepoints between E3.5 and E13.5 of mouse embryogenesis, collectively 1,658,968 cells/nuclei from 480 samples, where each sample consists of either a single mouse embryo or a small pool of embryos from the same timepoint. Further details are provided in **Supplementary Table 1**. For each dataset, we took the unique molecular identifiers (UMI) count matrix (feature X cell) from the data source and separated cells by timepoint. For each timepoint, we performed conventional single-cell RNA-seq data processing using Seurat/v3: 1) normalizing the UMI counts by the total count per cell followed by log-transformation; 2) selecting the 2,500 most highly variable genes and scaling the expression of each to zero mean and unit variance; 3) applying PCA and then using the top 30 PCs to create a k-NN graph, followed by Louvain clustering (resolution = 1); 4) performing UMAP visualization in 2D space (dims = 1:30, min. dist = 0.75) [51]. For some timepoints, we observed obvious batch effects with respect to either study or sample identity. We therefore performed an additional batch correction before the PCA, following the standard pipeline for dataset integration in Seurat/v3 (https://satijalab.org/seurat/v3.2/integration.html), using either the study or sample identity to split datasets, followed by identifying "anchors" between pairs of post-splitting subsets of the datasets (features = 2500, k.filter = 200, dims = 1:30) (**Supplementary Fig. 1a-b**).

For cell clustering, we manually adjusted the resolution parameter towards modest over-clustering, and then manually merged adjacent clusters if they had a limited number of

differentially expressed genes (DEGs) relative to one another (for this purpose, DEGs were defined as genes expressed at mean > 0.5 UMIs per cell across the pair of clusters with a > 4-fold difference between the clusters) or if they both highly expressed the same literature-nominated marker genes. Subsequently, we annotated individual cell clusters using 2 to 5 literature-nominated marker genes per cell type label (**Supplementary Table 2**). Many of the cell type labels and associated marker genes were obtained from the four studies that generated the data. However, we double-checked each cell type assignment, often with additional marker genes. Importantly, we revisited and revised some of the cell type or trajectory annotations of [15], *e.g.* "Ependymal cell" - "Roof plate"; "Isthmic organizer cells" - "Mesencephalon/MHB". A full list of these annotation revisions is provided in **Supplementary Table 26**. To benchmark the robustness of cell type annotations, we applied the sklearn.svm.LinearSVC function in scikit-learn/1.0 with 5-fold cross-validation, using the expression values of all genes as predictors (**Supplementary Fig. 10**).

To connect each cell state observed at a given timepoint with its "pseudo-ancestors", we first merged all cells from that timepoint and the preceding timepoint using Seurat/v3. Integration and batch correction were performed as described above, except that we also split based on timepoint identity (features = 2500, k.filter = 200, dims = 1:30). Because of the very large number of cells, we used a reciprocal PCA-based space [51] to find anchors for pairs of timepoints that included data from [15]. After integration, we performed PCA and then used the top 30 PCs to co-embed cells as a 3D UMAP (min. dist = 0.75), from which we calculated Euclidean distances between individual cells from the earlier and later timepoints.

We then determined edge weights between cell states of the successive timepoints using a bootstrapping strategy. For cells of each cell state at the later timepoint, we identified their five closest neighbor cells from the earlier timepoint and then calculated the proportion of these neighbors derived from each potential antecedent cell state. We repeated these

steps 500 times with 80% subsampling from the same embedding. We then took the median proportions as the set of weights for edges between a cell state and its potential antecedents. To evaluate the robustness of this approach to the choice of co-embedding space, we repeated it using Euclidean distances between cells in PCA space (dims = 30) instead of UMAP space (dims = 3). The resulting edge weights were highly correlated (Pearson correlation coefficient = 0.993). We evaluated the above approach with k parameters (for the k-NN) other than five, and found the resulting edge weights to be highly correlated with those obtained with k = 5 (Pearson correlation coefficients from 0.9994 to 0.9999 for k = 8, 10, 15, 20). Edge weights > 0.2 from the UMAP embedding were retained for the resulting acyclic directed graph shown in **Fig. 2c**.

We repeated this strategy to generate similar graphs for zebrafish (*D. rerio*) and frog (*X. tropicalis*) embryogenesis, again relying on publicly available scRNA-seq datasets. For zebrafish, we integrated data from two studies that overlapped at three timepoints (hpf6, hpf8, hpf10); we excluded cells from hpf4 because of excessive batch effects [8, 7]. For frog, we used cells from a single study [9]. Further details regarding data sources are available in **Supplementary Table 1**.

RNA velocity analysis

Three datasets were used in performing RNA velocity analysis – the Pijuan-Sala *et al.* dataset, the newly generated E8.5 dataset and the dataset resulting from deeper sequencing of Cao *et al.* 2019 libraries [13, 15]. For the Pijuan-Sala *et al.* dataset, which was generated on the 10X Genomics platform, we downloaded the raw data (E-MTAB-6967) and reprocessed it using kb-python (Melsted *et al.* 2019). For the new E8.5 data as well as the deeper sequencing of Cao *et al.* 2019 libraries, both generated with sci-RNA-seq3, we processed the raw data using the basic sci-RNA-seq pipeline followed by extracting the spliced reads and unspliced reads for each cell using velocyto [24, 15]. The RNA velocity

analysis and UMAP visualization were performed with Scanpy/v.1.6.0 and scVelo [89, 141]. Briefly, genes with low expression were filtered out, and each cell's counts were normalized towards the median UMI counts per cell by a scaling factor. The 3000 genes with the highest variance were selected, and the data were log-transformed after adding a pseudocount. The spliced and unspliced count matrices were similarly filtered and normalized. We then applied scvelo.pp.memoments and scvelo.tl.velocity for velocity estimation, followed by scvelo.tl.velocity_graph and scvelo.tl.umap for data visualization.

To infer the cell-state transitions between adjacent timepoints based on RNA velocity, cells from each pair of adjacent timepoints were integrated, and this was followed by applying the RNA velocity analysis using scVelo [89]. Of note, we did not perform RNA velocity analysis for cell states before E6.5 and during the transition between E8.5a vs. E8.5b because of limited numbers of cells or the major technological transition, respectively. For cell states from E8.5b onward, we performed a random downsampling on each cell state to 1,500 cells prior to RNA velocity analysis, in order to reduce computational costs. The resulting transition probabilities between individual cells, were calculated using cosine correlation between the potential cell-to-cell transitions and the inferred velocity vector (ranging from 0 to 1). To calculate the transition probability from cell state A at the earlier timepoint to cell state B at the later timepoint, we summed the transition probabilities of all cells within A to all cells within B, followed by normalizing the total cell number of B. Finally, the edge weight from A to B was further calculated by normalizing their transition probability to the total transition probabilities which originated from A.

Inferring the molecular histories of individual cell types

For this particular analysis, because one dataset did not include the extraembryonic tissues [15], we excluded cells annotated as derived from the extraembryonic lineages (embryonic visceral endoderm, extraembryonic visceral endoderm, parietal endoderm, and extraembry-

onic ectoderm). For E6.5, the sequencing depths were very different between datasets, so we only used cells from the Pijuan-Sala *et al.* dataset. In addition, the Pijuan-Sala *et al.* dataset pooled multiple embryos per sample, so we used sample identity instead of embryo identity. In the end, four samples from the Cheng *et al.* dataset, 34 samples from the Pijuan-Sala *et al.* dataset, 12 samples from the new E8.5 data ("E8.5b"), and 61 samples from the deeper sequencing of Cao *et al.* libraries, were used for the pseudobulk analysis. UMI counts mapping to each sample were aggregated to generate a pseudobulk RNA-seq profile for each sample. We then applied the fit_models function of Monocle/3 to identify genes that were highly correlated with the embryos'/samples' staged age. To mitigate major batch effects between cell vs. nucleus-derived subsets of the data, we separately performed DEG analysis on the samples from before and including E8.5a (n = 34, from Pijuan-Sala *et al.* dataset) vs. including and after E8.5b (n = 73), and then took the union of the top 3000 genes with the lowest q values identified in each subset. We then filtered out genes that were significantly different between the pre- and post-E8.5a/b subsets (p_value < 0.05). This left 534 genes, which were used to construct a pseudotime trajectory using DDRTree as implemented in Monocle/v2 [142]. Each embryo/sample was assigned a pseudotime value on the basis of its position along the trajectory. Of note, this ordering was highly robust to 80% subsampling (all Pearson correlation coefficients were > 0.99 between pseudotimes derived from 100 iterations of 80% subsampling vs. the full dataset).

Deconvolution of cell composition of GEO-seq sample using CIBERSORTx

This analysis was performed by running deconvolution on each GEO-seq sample using CIBERSORTx with default parameters [35, 33]. GEO-seq samples were collected from distinct spatial positions in the mouse embryo with mixed cell populations from E5.5, E6, E6.5, E7, and E7.5 [33]. For each stage, we first learned a gene expression signature for each cell state at the corresponding timepoint. Because single cell profiles from E6 were missing from the scRNA-seq data integrated here, we used data from E6.25 instead.

Systematic nomination of key transcription factors for cell type specification

The list of 1,636 mouse proteins that are putatively TFs was collated from AnimalTFDB/v3 (http://bioinfo.life.hust.edu.cn/AnimalTFDB/) [96]. For each edge in TOME at which a given cell type first emerged, we used three criteria to identify key TF candidates: 1) its expression significantly increased in the newly emerged cell type, relative to the pseudo-ancestral cell state (Seurat/v3; p_val_adj < 0.05, non-parametric Wilcoxon rank-sum test); 2) it was significantly more highly expressed in the newly emerged cell type, relative to its "sister" edges deriving from the same pseudo-ancestor (by the same test and threshold); 3) it was detected in at least 10% of cells of the newly emerged cell type. For each such candidate key TF, we scaled its log fold-change calculated by either criteria #1 or #2 to unit variance and zero mean (across the set of candidate key TF identified for a given newly emerged cell type) and then averaged these scaled fold-change values to determine a score intended to convey its importance relative to other candidate key TFs for the same cell type.

To identify TFs whose reduced expression was associated with the emergence of each cell type, we looked for those that: 1) are detected in at least 10% of cells of the pseudo-ancestral cell type; 2) are significantly downregulated in the newly emerged cell type, relative to the pseudo-ancestor (Seurat/v3; p_val_adj < 0.05, non-parametric Wilcoxon rank-sum test); and 3) are both detected in at least 10% of cells and significantly more highly expressed at "sister" edges, relative to the newly emerged cell type (by the same test and threshold).

The list of 2,547 zebrafish TFs and 1,236 frog TFs was collated from AnimalTFDB/v3 (http://bioinfo.life.hust.edu.cn/AnimalTFDB/) [96]. Candidate key TFs for each cell type emergence in these species were identified and scored as described above for mouse.

Co-embedding of cell states from three species

We first created a list of orthologous genes across the three species by liftover of all gene identities from the three species to the corresponding human gene identities, either based on BioMart (Ensembl Genes 102) [143] or the original study in the case of frog [9]. A list of 22,815 genes was compiled, wherein each of the genes was orthologous in at least two species. Of note, we retained all of the possible orthologous gene pairs learned from the BioMart, including "1-to-1", "1-to-many", and "many-to-many" categories. To create the transcriptional features of each cell state, we first averaged cell-state-specific UMI counts, normalized by the total count, multiplied by 100,000 and natural-log-transformed after adding a pseudocount. We then divided all the cell states from three species into four groups: the mouse single-cell group (n = 151), the mouse single-nucleus group (n = 277), the zebrafish group (n = 205), and the frog group (n = 192). We treated each cell state as a pseudo-cell, performing the anchor-based batch correction approach implemented by Seurat/v3 (nfeatures = 5,000, k.filter = 100, dims = 1:30, min.dist = 0.6) [51]. For cell states spanning multiple timepoints, cells from each timepoint were treated as a separate pseudo-cell for the purposes of this analysis.

Identification of interspecies correlated cell types using non-negative least-squared (NNLS) regression

We first created a list of orthologous genes between each pair of species (n = 17,333 for mm vs. zf, n = 14,249 for mm vs. xp, and n = 13,326 for zf vs. xp), either based on BioMart (Ensembl Genes 102) [143] or the original study in the case of frog [9]. Of note, we retained all of the possible orthologous gene pairs learned from the BioMart, including "1-to-1", "1-to-many", and "many-to-many" categories. To identify correlated cell types between each pair of species, we first calculated an expression value for each gene in each cell type by averaging the log-transformed normalized UMI counts of all cells of that type across all timepoints at which the cell type appeared. Extraembryonic cell types (inner cell mass, hypoblast, parietal endoderm, extraembryonic ectoderm, visceral endoderm, embryonic visceral endoderm, and

extraembryonic visceral endoderm for the mouse; blastomere, EVL, periderm, forerunner cells for the zebrafish) were excluded from this analysis. For mouse E6.5, we only used cells from a single study [13]. For each pair of species, we took homologous genes and applied non-negative least squares (NNLS) regression to predict gene expression in target cell type (T_a) in dataset A based on the gene expression of all cell types (M_b) in dataset B: $T_a = \beta_{0a} + \beta_{1a}M_b$, based on the union of the 1,200 most highly expressed genes and 1,200 most highly specific genes in the target cell type. We then switched the roles of datasets A and B, *i.e.* predicting the gene expression of target cell type (T_b) in dataset B from the gene expression of all cell types (M_a) in dataset A: $T_b = \beta_{0b} + \beta_{1b}M_a$. Finally, for each cell type a in dataset A and each cell type b in dataset B, we combined the two correlation coefficients: $\beta = 2(\beta_{ab} + 0.001)(\beta_{ba} + 0.001)$ to obtain a statistic for which high values reflect reciprocal, specific predictivity.

To identify candidate cell type homologs, we manually reviewed pairings with a beta score $>$ 1e-4 and that ranked highly from the perspective of both species, *i.e.* where cell type B was one of the top five matches for cell type A and vice versa. We next performed a manual selection based on the following criteria: 1) excluding pairs of cell types which derive from different germ layers or major groups (**Supplementary Fig. 22**) (*e.g.* blood progenitors (mm) vs. optic cup (zf)); 2) excluding pairs of cell types which emerged at very different temporal stages (*e.g.* rostral neuroectoderm (mm) vs. DEL (zf)); 3) excluding cell types only expected in one species or the other (*e.g.* hatching gland (zf) is not expected in mouse); 4) for cell types which were correlated with multiple cell types with ancestor-descendant relationships in the other species, we selected the one which was more ancestral (*e.g.* hindbrain (mm) was correlated with both hindbrain ventral (zf) and hindbrain (zf), and we assigned it to hindbrain (zf)); 5) for cell types which were correlated with multiple cell types in the other species that lacked a clear ancestor-descendant relationship, we selected the pair with the highest beta score. The details of manual selection are provided in **Supplementary Table 23**.

Identification of correlated cell types between species based on overlapping key TF candidates

For each possible interspecies pairing of cell types, we identified orthologous TFs that were nominated in both species and then calculated, as an estimate of relative likelihood, the product of the frequencies in which each of these TFs were nominated as key in their respective species (to account for the fact that some TFs are nominated in many cell types and therefore more likely to overlap; **Fig. 5c**). To identify which such instances were potentially significant, we repeated these procedures after taking random samples of key TFs without replacement (10,000 times) and retained pairings with estimated relative likelihoods more extreme than 99% of permutations. We then performed a similar manual selection, details of which are provided in **Supplementary Table 24**.

Of note, we also attempted interspecies cell type pairing using key genes instead of key TFs for each cell type (**Supplementary Table 27**). However, the correlated cell types identified by overlapping key genes were noisier than other approaches. For example, anterior floor plate (mm) was correlated to diencephalon (*aplnr2*+) (zf) as expected, but it was also correlated to seven other cell types from zebrafish, including erythroid, midbrain ventral, myotome, diencephalon, roof plate, mesoderm lateral plate (*tbx1*+), dorsal margin involuted. As the other strategies appeared less noisy and therefore easier to manually curate, we did not carry this third approach forward.

We compared our cell-type alignments between zebrafish vs. frog to a recent study [144] that also sought to align the same datasets. We could find consistent alignments for 35 of 46 pairs of cell types which they identified (Supplementary Table 28). Note that neither we nor they simply used the original data and annotations, but rather we re-processed them in different ways. For example, we combined sc-RNA-seq data from two zebrafish studies [8, 7] followed by reannotation of the merged set of cells from each individual timepoint, while the

other study sometimes merged multiple cell types into one (optic cup and retina pigmented epithelium - optic). These differences make a full comparison challenging. Nonetheless, at least on a high-level check, these entirely independent efforts are mostly in agreement, which is encouraging.

Identification of cis-regulatory motifs involved in in vivo cell type specification

As a first step towards identifying *cis*-regulatory motifs involved in cell type identification, we extended to all genes the approach described above to nominate key TFs whose upregulation or downregulation is associated with the emergence of each cell type. For each edge in TOME at which a given cell type first emerged, we used three criteria to identify key gene candidates: 1) its expression significantly increased in the newly emerged cell type, relative to the pseudo-ancestral cell state (Seurat/v3; p_val_adj < 0.05, non-parametric Wilcoxon rank-sum test); 2) it was significantly more highly expressed in the newly emerged cell type, relative to its "sister" edges deriving from the same pseudo-ancestor (by the same test and threshold); 3) it was detected in at least 10% of cells of the newly emerged cell type. To identify genes whose reduced expression was associated with the emergence of each cell type, we looked for those that: 1) are detected in at least 10% of cells of the pseudo-ancestral cell type; 2) are significantly downregulated in the newly emerged cell type, relative to the pseudo-ancestor (Seurat/v3; p_val_adj < 0.05, non-parametric Wilcoxon rank-sum test); and 3) are both detected in at least 10% of cells and significantly more highly expressed at "sister" edges, relative to the newly emerged cell type (by the same test and threshold).

We used HOMER/v4.11 [122] to identify DNA sequence motifs that are specifically enriched in the core promoters of key genes (-300 to +50 bp of annotated TSSs). Running the findMotifs.pl function with default parameters, each test set was defined as the core promoters of either upregulated or downregulated key genes at specific cell edges (excluding sets

with fewer than 5 key genes), and compared to a background set of core promoters of key genes from all edges not in the test set. Motif quality was evaluated based on a q-value, which was calculated for each motif by 100 iterations of randomizing data labels and re-running HOMER. In addition, motifs were aligned to known motif binding sequences based on the JASPAR and internal HOMER databases with default parameters [145]. Mapping of specific motif positions around the TSS was assessed with the HOMER function annotatePeaks.pl using the following parameters: tss mm10 -hist 10 -ghist.

Data availability

All data and code used here have been made freely available via http://tome.gs.washington.edu. The data generated in this study can be downloaded in raw and processed forms from the NCBI Gene Expression Omnibus under accession number GSE186069 (new E8.5 data) & GSE186068 (deeper sequencing of Cao *et al.* libraries). The supplementary tables can be downloaded from here:
https://shendure-web.gs.washington.edu/content/members/cxqiu/public/nobackup/tmp/

Acknowledgments

We thank the members of the Shendure and Trapnell labs for helpful discussions. This work was funded by Paul G. Allen Frontiers Foundation (Allen Discovery Center grant to J.S. and C.T.), the National Institutes of Health (UM1HG011531 to W.S.N., J.S., and C.M.D., R01 NS109425 to C.B.M.), and the Bonita and David Brewer Fellowship (to C.Q.). D.C. was further supported by T32HL007828 from the National Heart, Lung, and Blood Institute. J.S. is an investigator of the Howard Hughes Medical Institute. M.S. is a DZHK principal investigator and supported by grants from the Deutsche Forschungsgemeinschaft (DFG) (SP1532/3-1, SP1532/4-1, and SP1532/5-1) and the Deutsches Zentrum für Luft- und Raumfahrt (DLR 01GM1925).

Competing Financial Interests Statement

The authors declare that they have no competing financial interests.

2.12 Figures

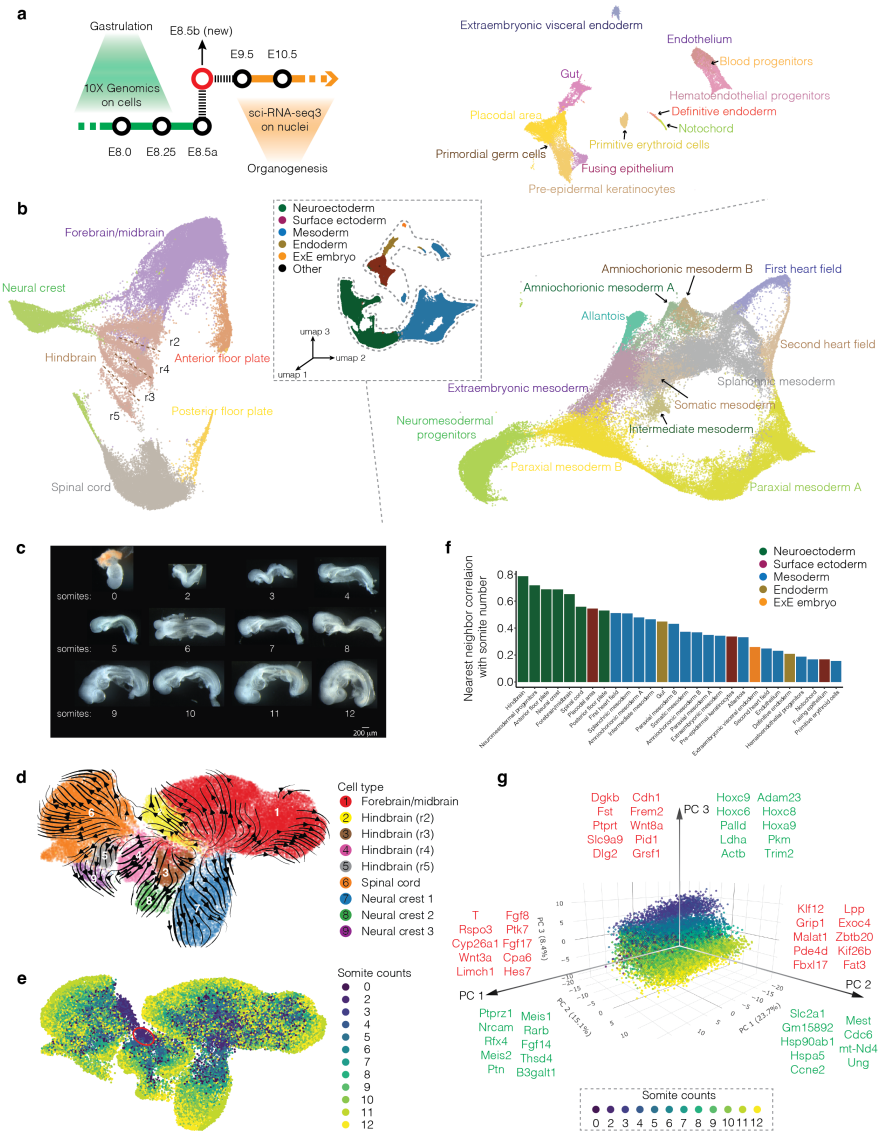

Figure 2.1: **Figure 1. Intensive scRNA-seq of somite-resolved E8.5 mouse embryos. a,** A new scRNA-seq dataset was generated from nuclei derived from individual E8.5 mouse embryos via an optimized sci-RNA-seq3 protocol, in order to "bridge" existing data. **b,** 3D UMAP visualizations of the new E8.5 dataset ("E8.5b"). **c,** Twelve mouse embryos, including a single primitive streak stage embryo and 11 embryos staged in 1-somite increments from 2 to 12 somites were collected. **d,** Re-embedded 2D UMAP of cells annotated as forebrain, midbrain, hindbrain, spinal cord, and neural crest. Arrows correspond to RNA velocity trends. **e,** The same UMAP as in panel d, colored by somite counts. **f,** We calculated the Pearson correlation coefficient between the somite number of each cell of that type and the average somite number of its five nearest neighbors in the global 3D UMAP embedding. **g,** 3D visualization of the top three principal components (PCs) of gene expression variation in NMPs, calculated on the basis of the 2,500 most highly variable genes.

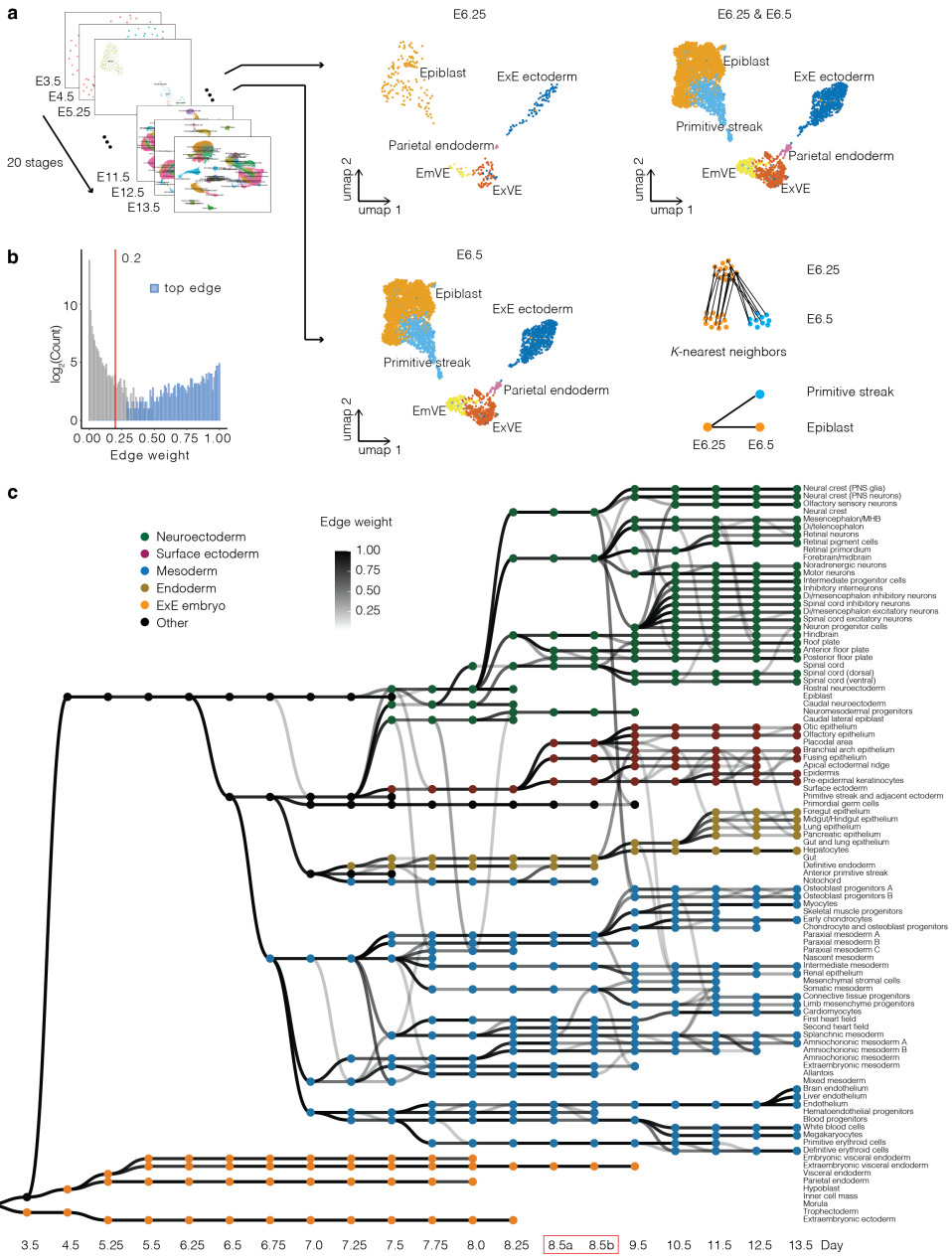

Figure 2.2: **Figure 2. Systematic reconstruction of the cellular trajectories of mouse embryogenesis. a,** Overview of approach. Cells from each pair of adjacent stages were projected into the same embedding space [51]. **b,** Histogram of all calculated edge weights. The y-axis is on a log2 scale. Edges with weights above 0.2 (red line) were retained. **c,** Directed acyclic graph showing inferred relationships between cell states across early mouse development. Each row corresponds to one of 94 cell type annotations, columns to developmental stages spanning E3.5 to E13.5, nodes to cell states, and node colors to germ layers. All edges with weights above 0.2 are shown in grey scale.

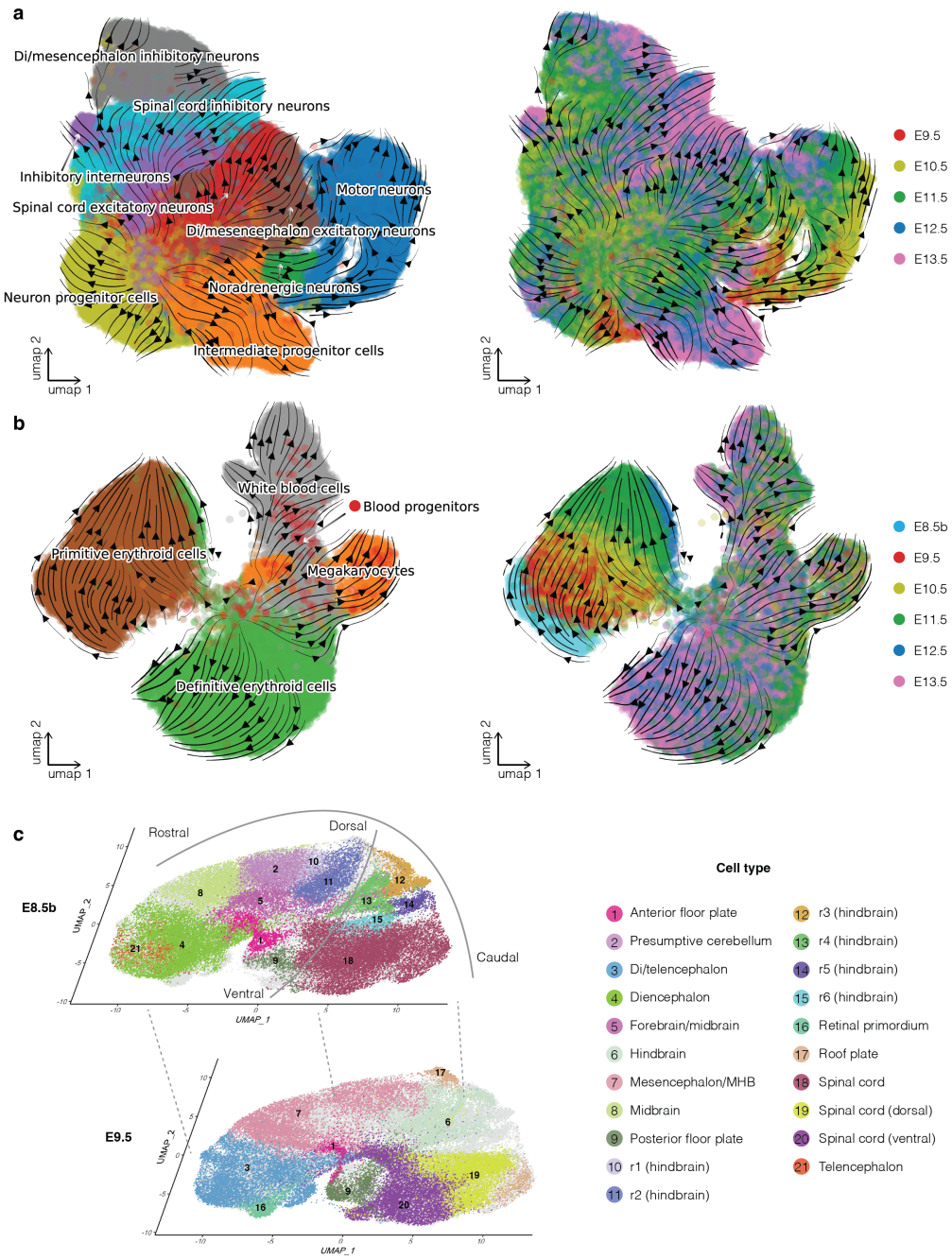

Figure 2.3: **Figure 3. RNA velocity and spatially correlated co-embeddings clarify relationships between cell types during neuronal differentiation, hematopoiesis and neural tube development. a,** RNA velocity was estimated on the basis of the proportion of reads mapping to exonic vs. intronic portions of genes using scVelo [89], corresponding to neuronal differentiation. **b,** Same as panel a, but for cells corresponding to hematopoiesis. **c,** UMAP visualization of co-embedded cells from neural tube derivatives from E8.5b and E9.5 data after batch correction. The same UMAP is shown twice for both, with colors highlighting cells and corresponding annotations from either E8.5b (top) or E9.5 (bottom).

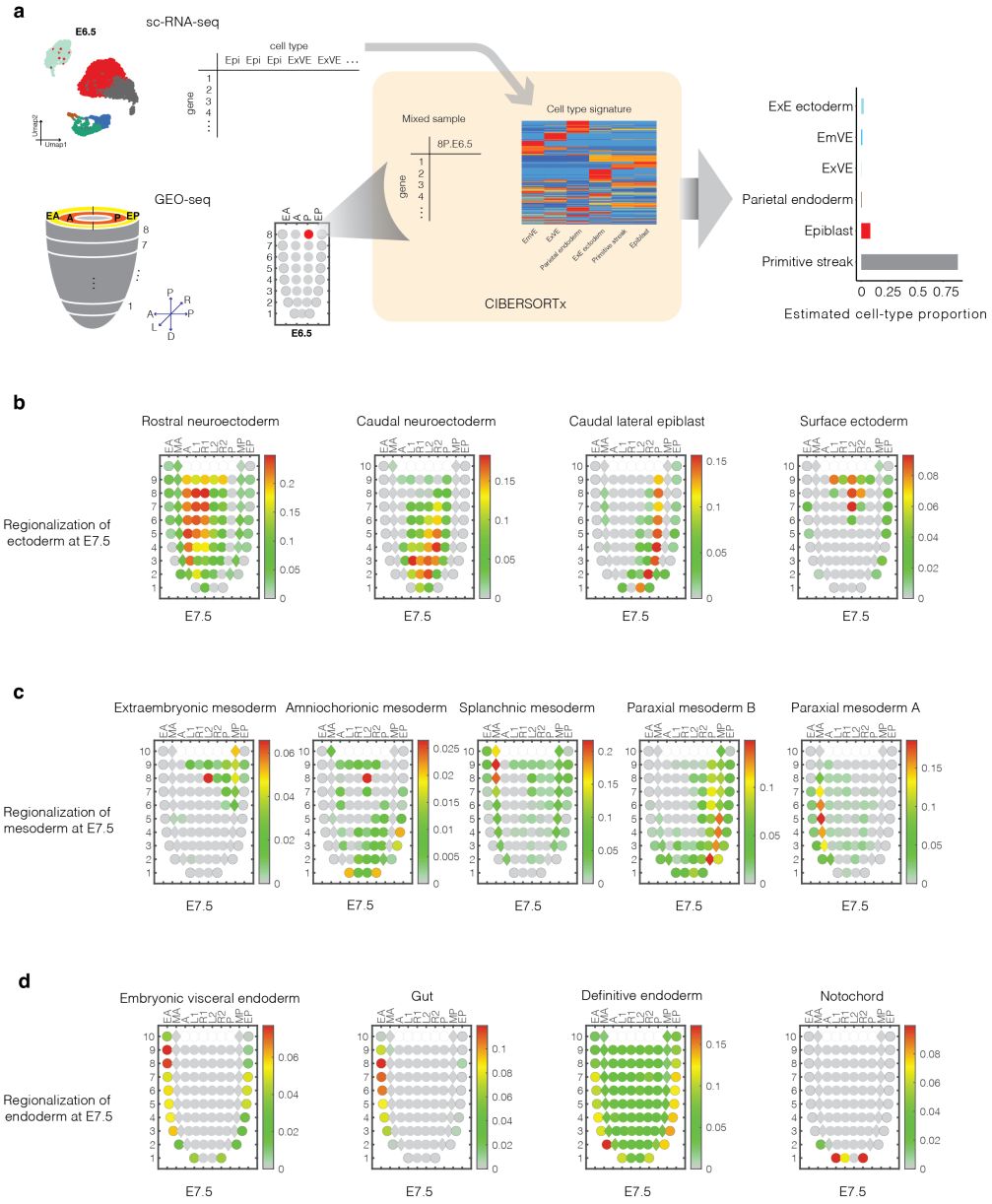

Figure 2.4: **Figure 4. Inference of the approximate spatial locations of cell states during mouse gastrulation.** **a,** Inference of cell type contributors to each spatial territory of the gastrulating mouse embryo based on the application of CIBERSORTx to GEO-seq data [35, 33]. **b,** Corn plots [33] showing the spatial pattern of inferred contributions of various ectodermal cell types at E7.5. **c,** Corn plots showing the spatial pattern of inferred contributions of various mesodermal cell types at E7.5. **d,** Corn plots showing the spatial pattern of inferred contributions of various endodermal cell types at E7.5, as well as notochord. In each corn plot, each circle or diamond refers to a GEO-seq sample, and its weighted color to the estimated cell type composition. Corn plot nomenclature from [33].

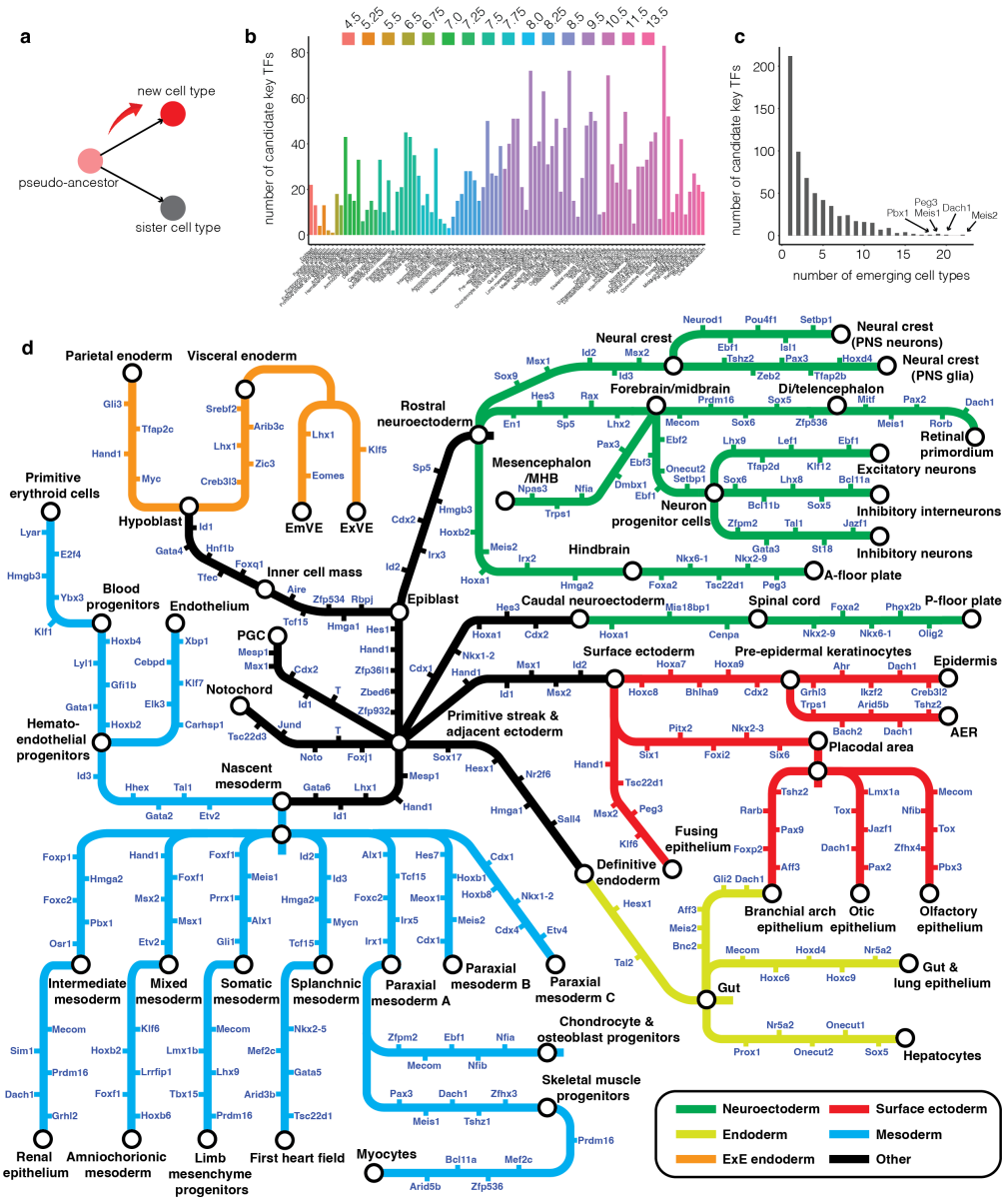

Figure 2.5: **Figure 5. Systematic nomination of candidate key transcription factors for cell type specification. a,** We heuristically defined candidate key TFs as those that are expressed in the pseudo-ancestral cell state, are significantly upregulated in the newly emerged cell type, and are not significantly upregulated at any sister edges. **b,** Histogram of the number of candidate key TFs for each cell type at the timepoint of its first emergence. **c,** The histogram of the number of cell types in which each TF was nominated as a candidate key TF. **d,** Diagram illustrating selected cellular trajectories from TOME, decorated with the top 5 scoring candidate key TFs for each edge. EmVE: embryonic visceral endoderm. ExVE: extraembryonic visceral endoderm. PNS: peripheral nervous system. MHB: midbrain-hindbrain boundary. Style inspired by [146].

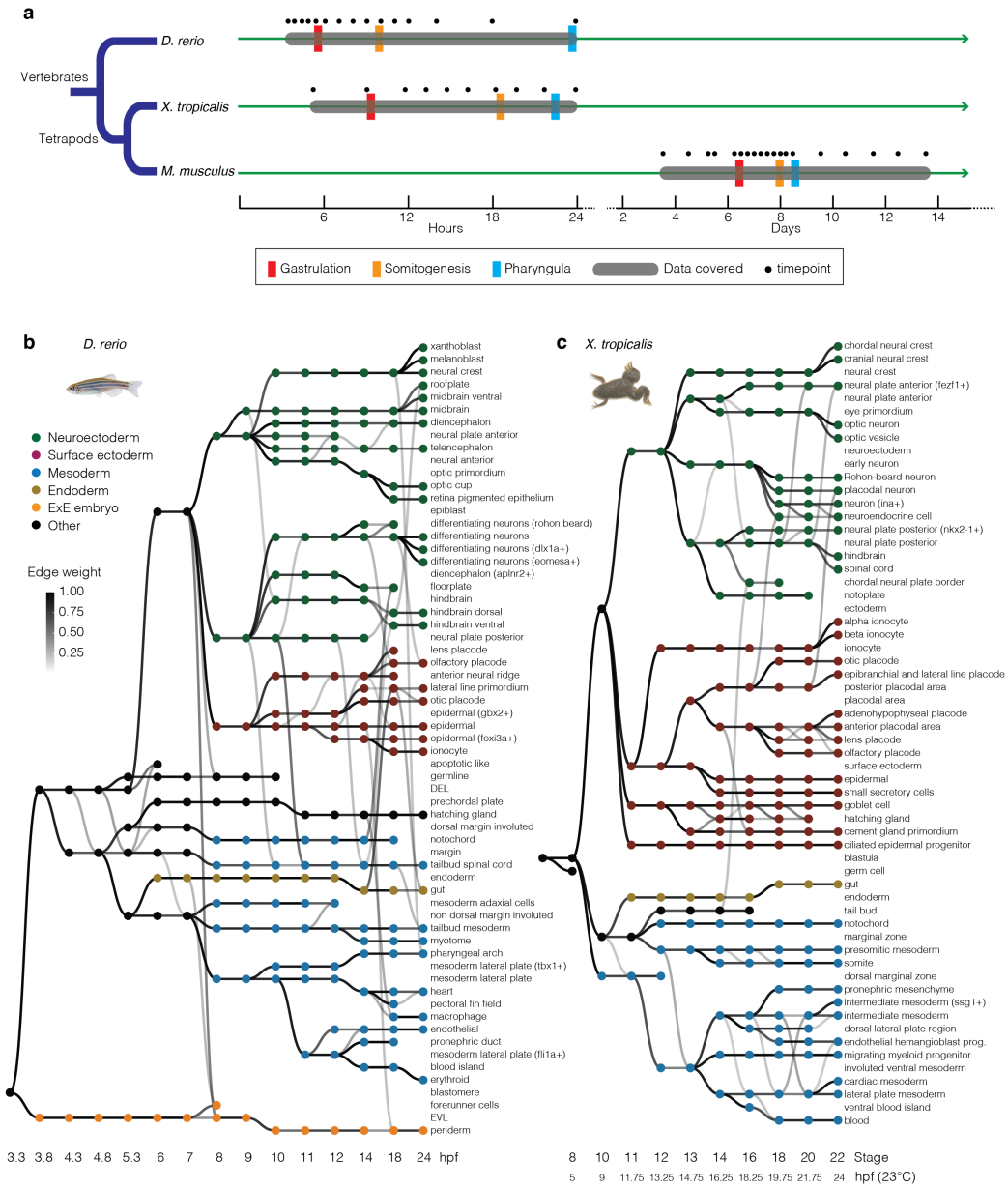

Figure 2.6: **Figure 6. Reconstruction of the cellular trajectories of zebrafish and frog embryogenesis. a,** Comparative developmental timelines for mouse, zebrafish, and frog, spread over two time scales, and approximate (as temperature-dependent, particularly for frog). **b,** Directed acyclic graph showing inferred relationships between cell states across early zebrafish development. Each row corresponds to one of 63 cell type annotations, and columns to developmental stages spanning hpf3.3 to hpf24. **c,** Directed acyclic graph showing inferred relationships between cell states across early frog development. Each row corresponds to one of 60 cell type annotations, columns to developmental stages spanning S8 (hpf5, 23C) to S22 (hpf24, 23C).

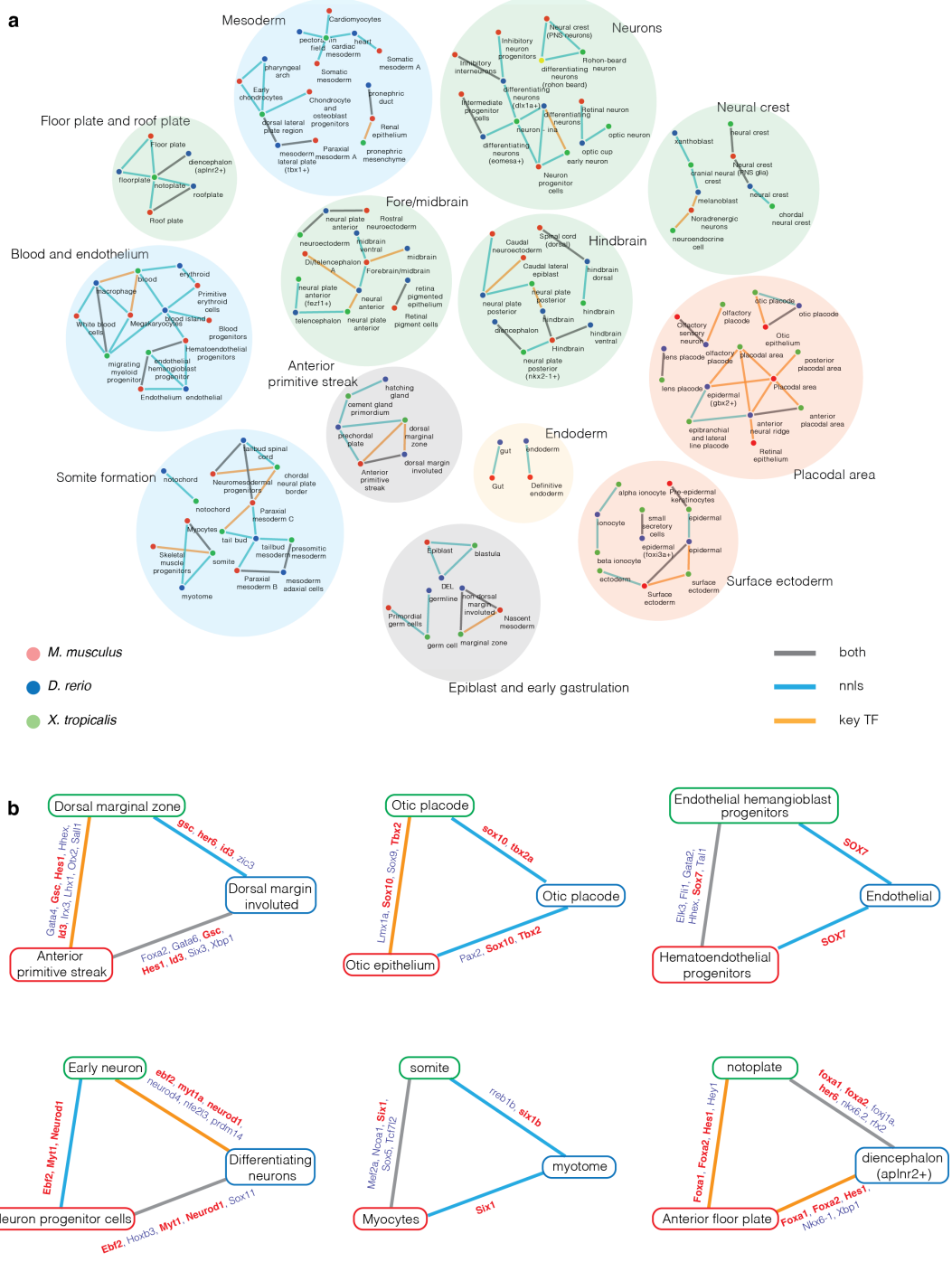

Figure 2.7: **Figure 7. The union of candidate cell type homologs, identified between three species (mouse, zebrafish, frog) by two strategies. a,** Candidate cell type homologs were identified either by comparison of transcriptomes via non-negative least squares ("nnls") regression or by examining overlap between upregulated candidate key TFs ("key TF"). Nominated pairings were manually reviewed, and a subset retained based on biological plausibility. **b,** Selected examples of "three way" pairwise cell type homology from different germ layers in the above network.

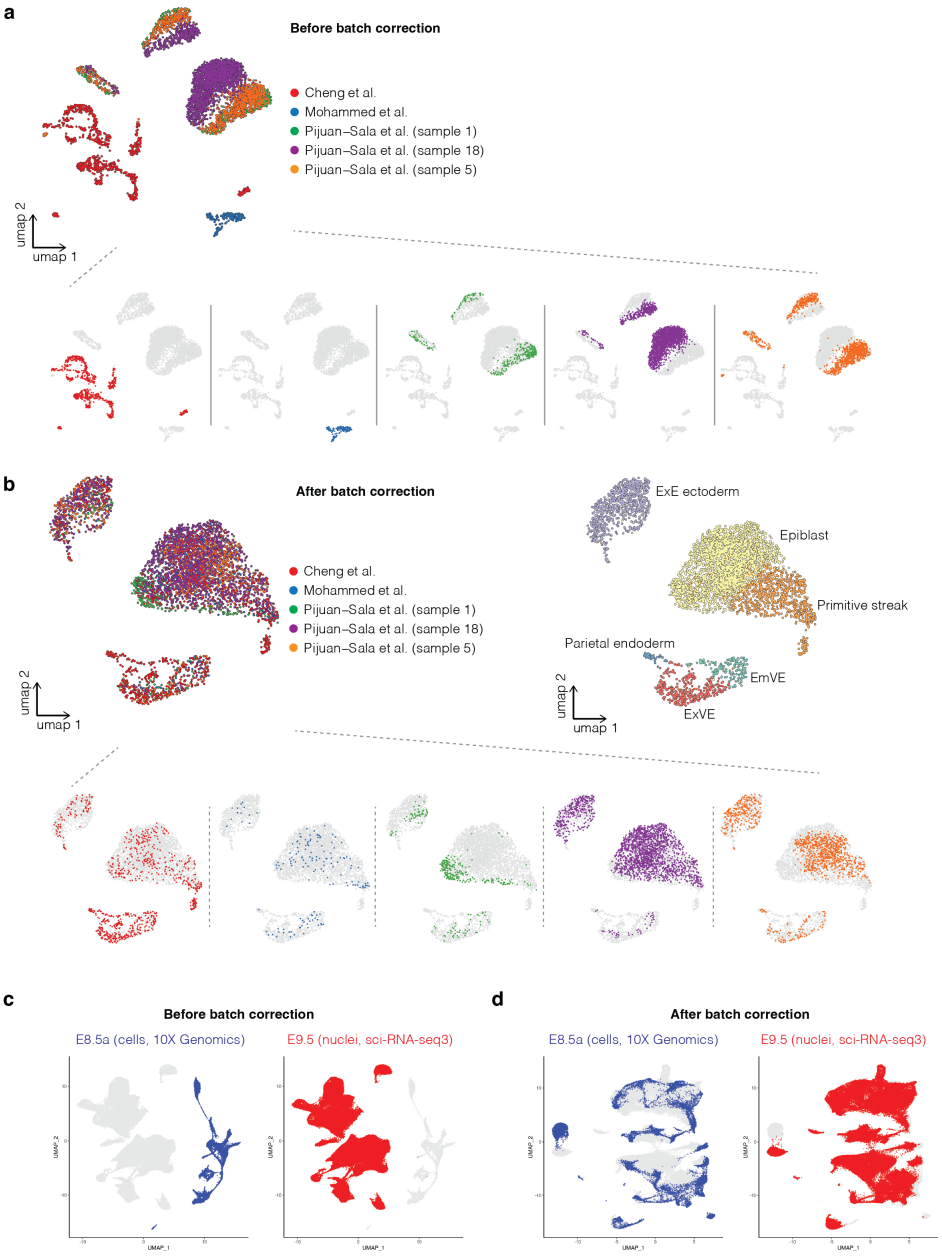

Figure 2.8: **Supplementary Figure 1. Integration of datasets generated by different groups using different scRNA-seq technologies.** **a,** As illustrated by a UMAP of co-embedded E6.5 cells, batch effects are observed between three studies, as well as different embryos from the same study. **b,** UMAP of the same cells as in panel a with batch correction prior to integration [51]. **c,** UMAP visualization of co-embedding of data from E8.5a (cells) generated on the 10x Genomics platform [13] and E9.5 (nuclei) generated using sci-RNA-seq3 [15], before batch correction [51]. **d,** UMAP of the same cells as in panel c but with batch correction prior to integration [51].

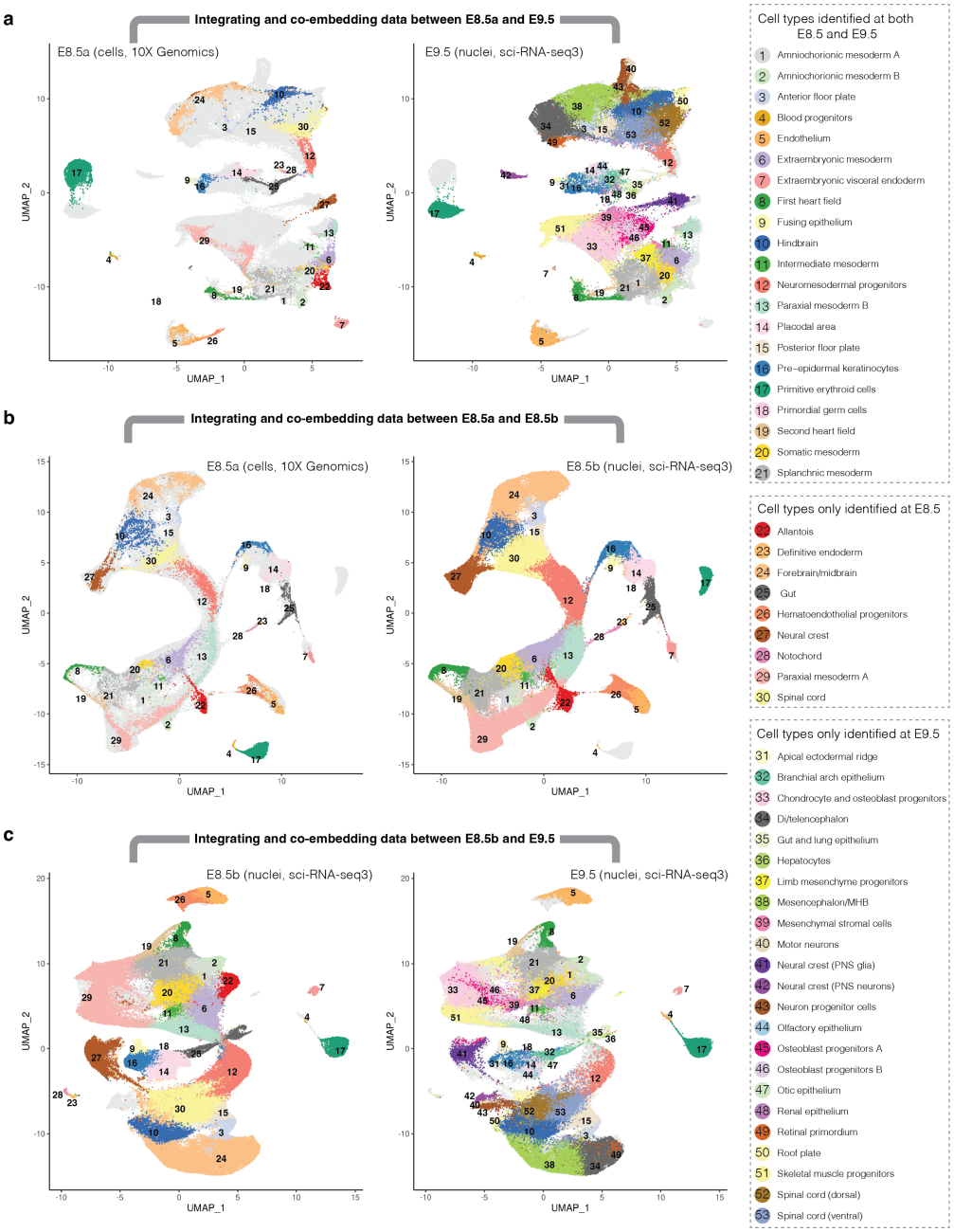

Figure 2.9: **Supplementary Figure 2. Integrating and co-embedding cells from E8.5a, E8.5b, and E9.5. a,** UMAP visualization of co-embedded cells at E8.5a generated on the 10x Genomics platform and nuclei at E9.5 generated using sci-RNA-seq3 after batch correction. **b,** UMAP visualization of co-embedded cells at E8.5a generated on the 10x Genomics platform and nuclei at E8.5b generated using sci-RNA-seq3 after batch correction. **c,** UMAP visualization co-embedded nuclei at E8.5b and nuclei at E9.5, both generated with sci-RNA-seq3, after batch correction.

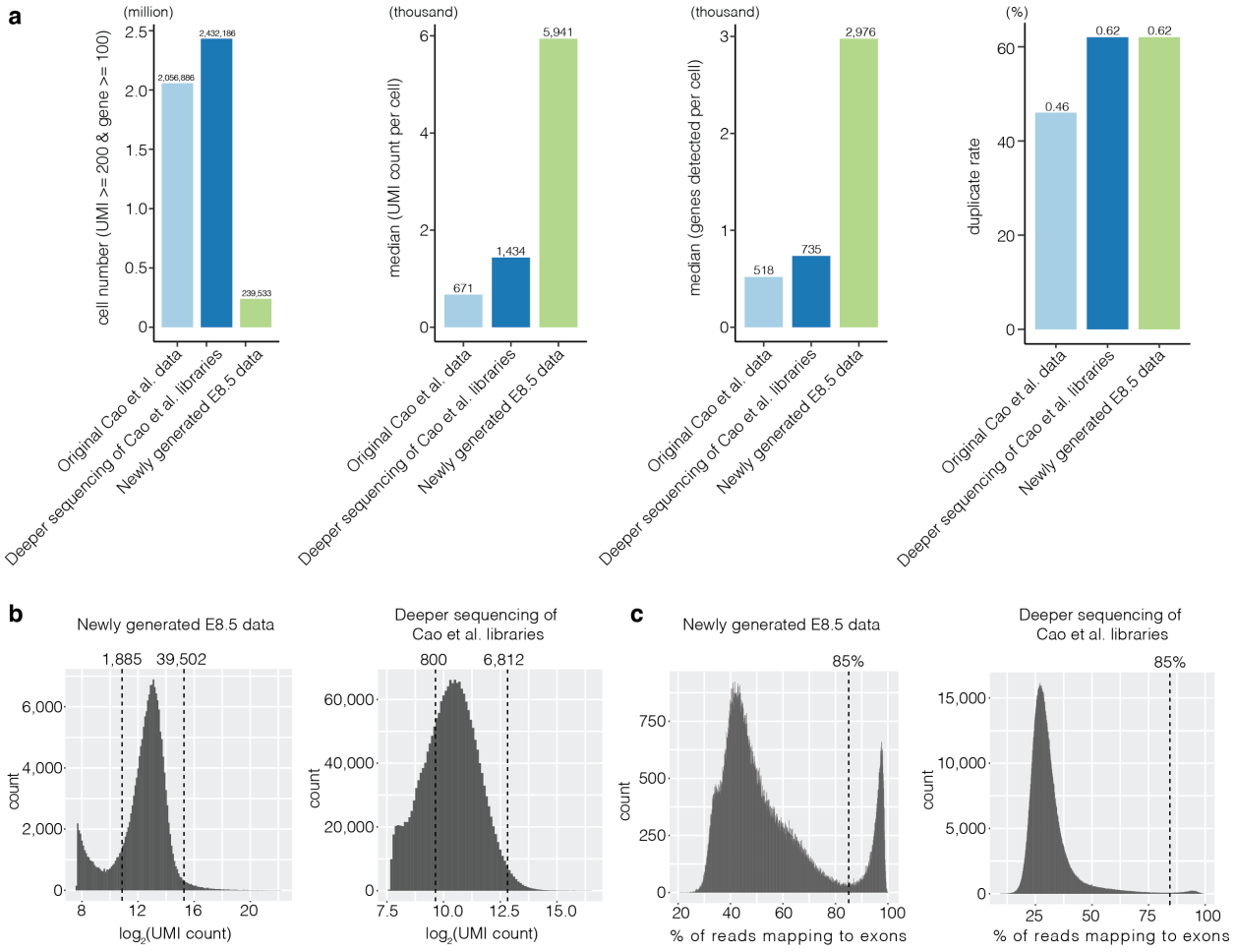

Figure 2.10: **Supplementary Figure 3. Higher quality sci-RNA-seq3 data generated by either application of an optimized protocol (E8.5b) or deeper sequencing of previously reported libraries (E9.5 - E13.5). a,** The cell number, median UMI count per cell, median genes detected per cell, and duplicate rate, are shown for a previously published dataset on E9.5 - E13.5 embryos (light blue bars) [15], deeper sequencing and reanalysis of those same sequencing libraries (dark blue bars) or data newly generated on E8.5 embryos using an optimized sci-RNA-seq3 protocol (green bars). **b,** Histograms of log2(UMI count) per cell for the newly created E8.5 dataset (left) and more deeply sequenced [15] libraries (right). **c,** Histograms of the proportion of reads mapping to the exonic regions per cell for the newly created E8.5 dataset (left) and more deeply sequenced [15] libraries (right).

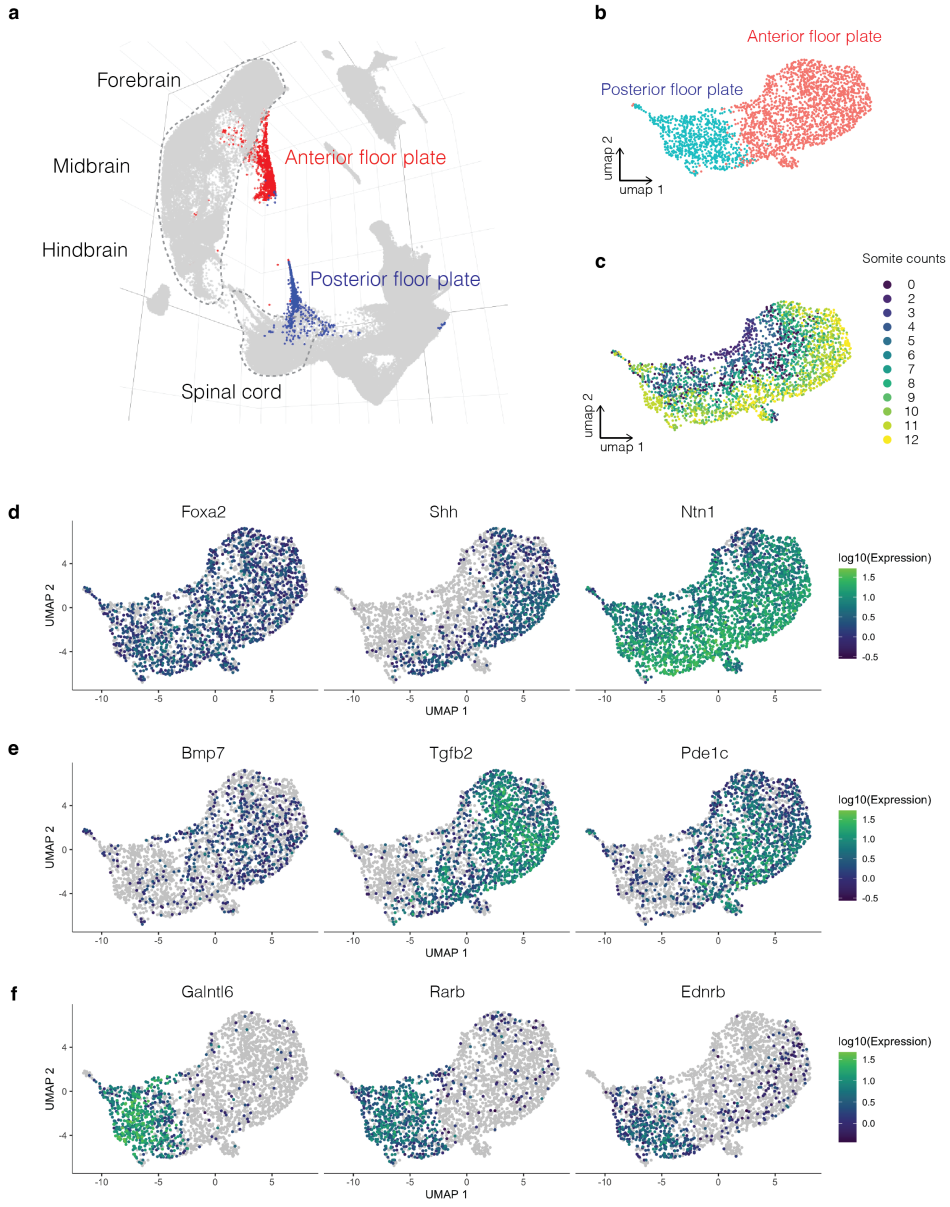

Figure 2.11: **Supplementary Figure 4. Anterior and posterior floor plate subpopulations emerging from forebrain/hindbrain and spinal cord, respectively. a,** Subview of global 3D UMAP visualization highlighting anterior (red) and posterior (blue) floor plate subpopulations in E8.5 data generated with optimized sci-RNA-seq3 protocol. **b,** Re-embedded 2D UMAP of cells from anterior floor plate and posterior floor plate. **c,** The same UMAP as in panel b, colored by somite counts. **d,** The same UMAP as in panel b, colored by gene expression of marker genes which are shared by anterior and posterior floor plate subpopulations. **e,** The same UMAP as in panel b, colored by gene expression of marker genes which appear specific to anterior floor plate subpopulation. **f,** The same UMAP as in panel b, colored by gene expression of marker genes which are specific to posterior floor plate subpopulation.

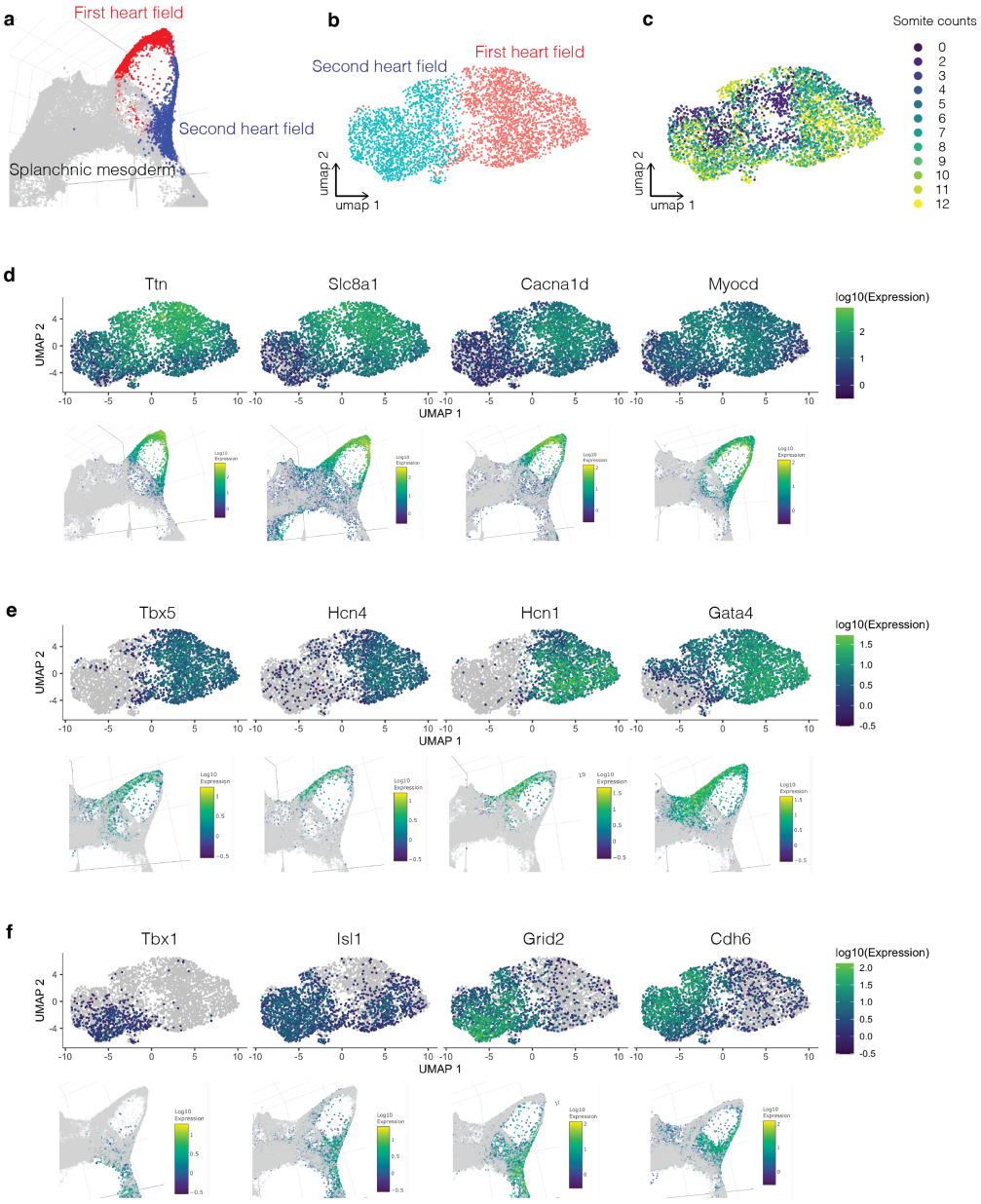

Figure 2.12: **Supplementary Figure 5. Resolution of the first and second heart fields during early somitogenesis.** **a,** Subview of global 3D UMAP visualization highlighting the first (red) and second (blue) heart fields in E8.5 data generated with optimized sci-RNA-seq3 protocol. **b,** Re-embedded 2D UMAP of cells from the first and second heart fields. **c,** The same UMAP as in panel b, colored by somite counts. **d,** The same 2D UMAP as in panel b, colored by gene expression of marker genes which are shared by the first and second heart fields. **e,** The same UMAP as in panel b, colored by gene expression of marker genes which are specific to the first heart field. **f,** The same UMAP as in panel b, colored by gene expression of marker genes which are specific to the second heart field.

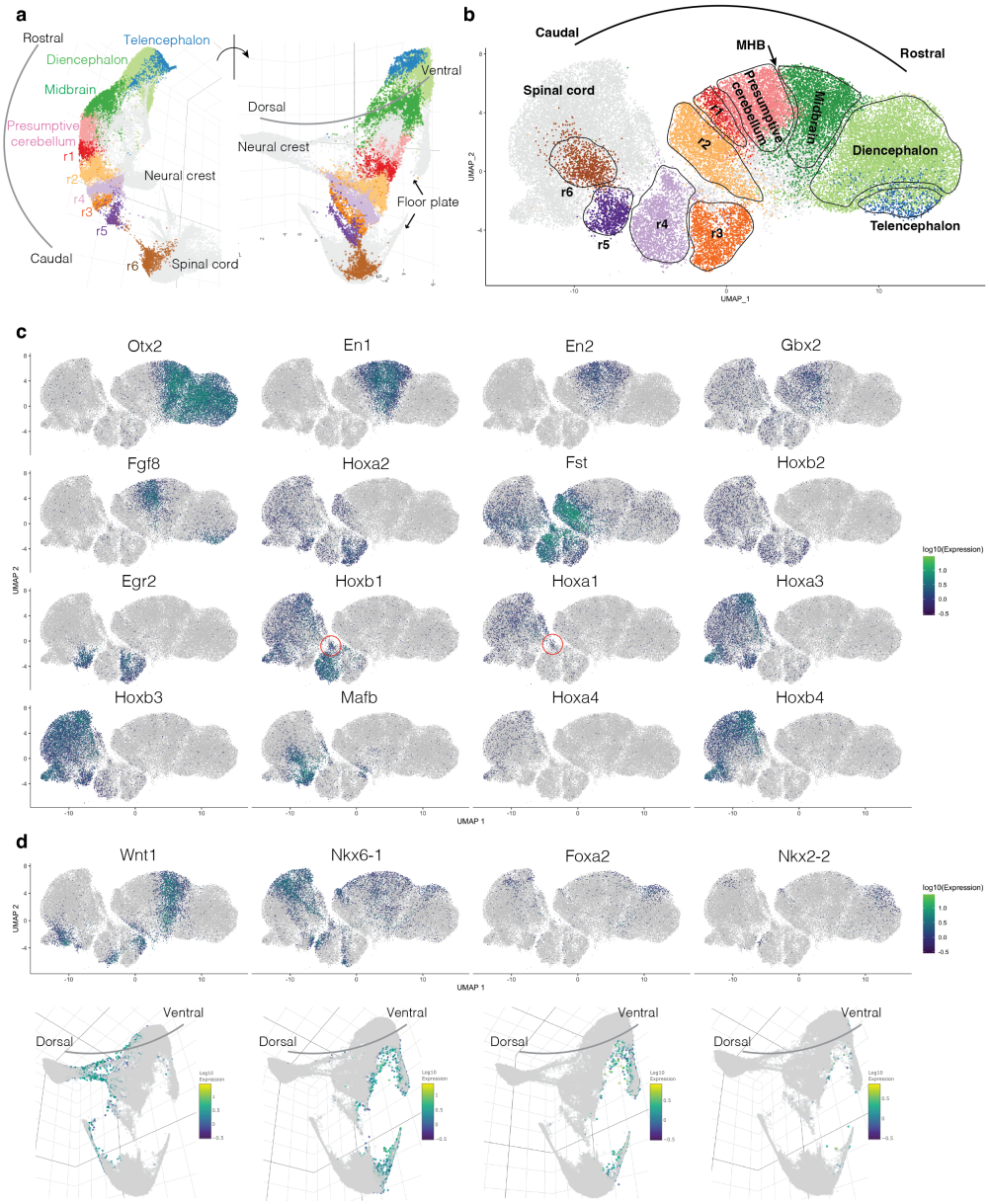

Figure 2.13: **Supplementary Figure 6. Resolution of hindbrain segmentation in newly created E8.5 dataset. a,** Subview of global 3D UMAP visualization highlighting subsets of cells annotated as rhombomeres 1 - 6 (r1 - 6) in E8.5 data generated with optimized sci-RNA-seq3 protocol. **b,** Re-embedded 2D UMAP of cells annotated as forebrain, midbrain, presumptive cerebellum, r1 - r6, spinal cord and neural crest. **c,** The same UMAP as in panel b, colored by gene expression of marker genes used for annotation of anatomical regions [65, 147, 148, 149, 150, 66]. **d,** The same UMAP as in panel b, colored by gene expression of marker genes for the dorsal-ventral axis [69, 70].

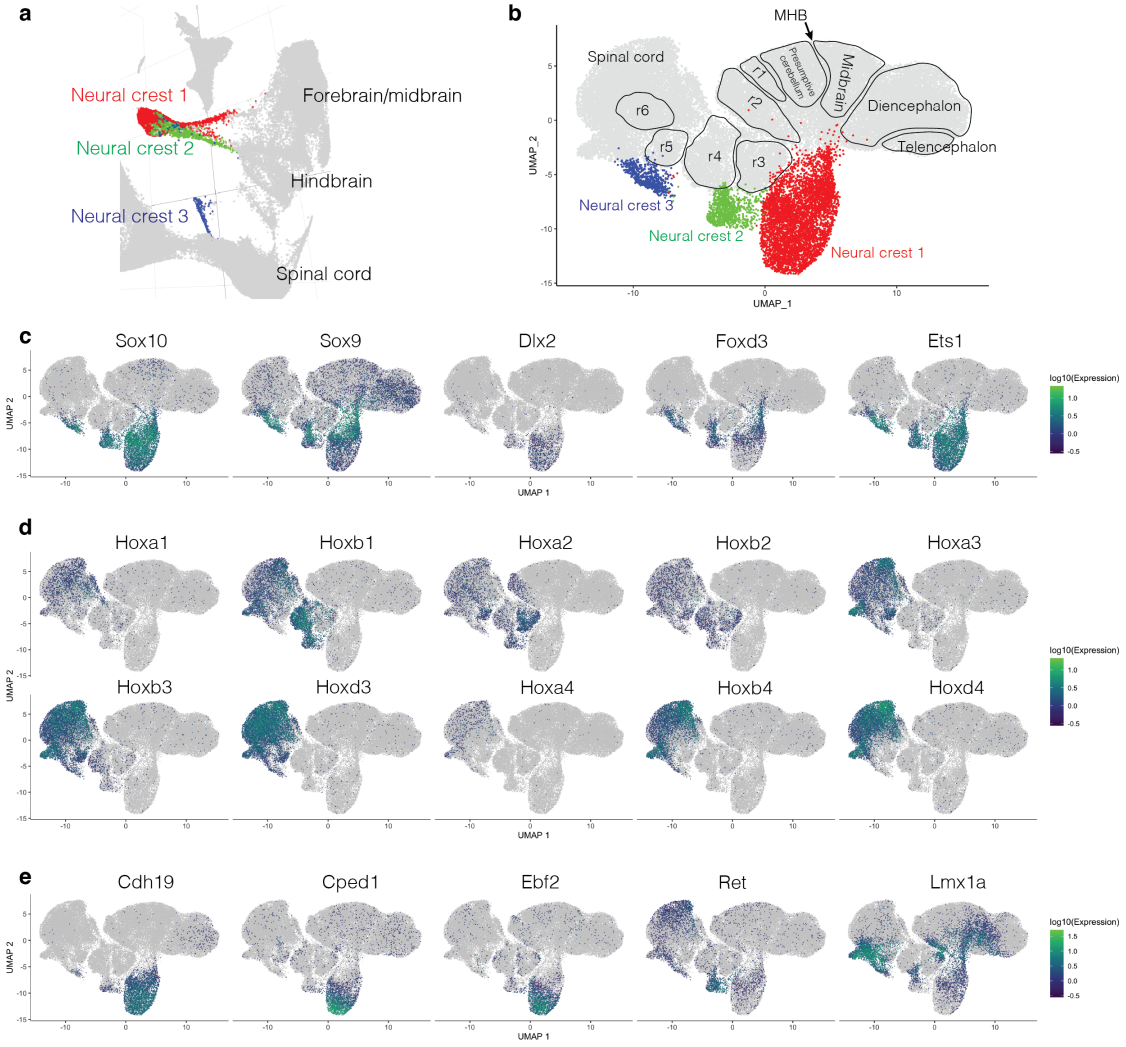

Figure 2.14: **Supplementary Figure 7. Three distinct subpopulations of neural crest cells (NCCs) may correspond to mesencephalic NCCs and pharyngeal arch (PA) contributions. a,** Subview of global 3D UMAP visualization highlighting the three subpopulations of NCCs. **b,** Re-embedded 2D UMAP of cells annotated as forebrain, midbrain, presumptive cerebellum, r1 - r6 rhombomeres, spinal cord and neural crest, with highlighting of the three subpopulations of NCCs. **c,** The same 2D UMAP as in panel b, colored by gene expression of marker genes which are shared by the three subpopulations of NCCs. **d,** The same UMAP as in panel b, colored by gene expression of *Hox* genes used for rough annotation of three subpopulations of NCCs [151]. **e,** The same UMAP as in panel b, colored by gene expression of marker genes which are specific to NC1 (Cdh19, Cped1, Ebf2), NC2 (Ret), or NC3 (Lmx1a).

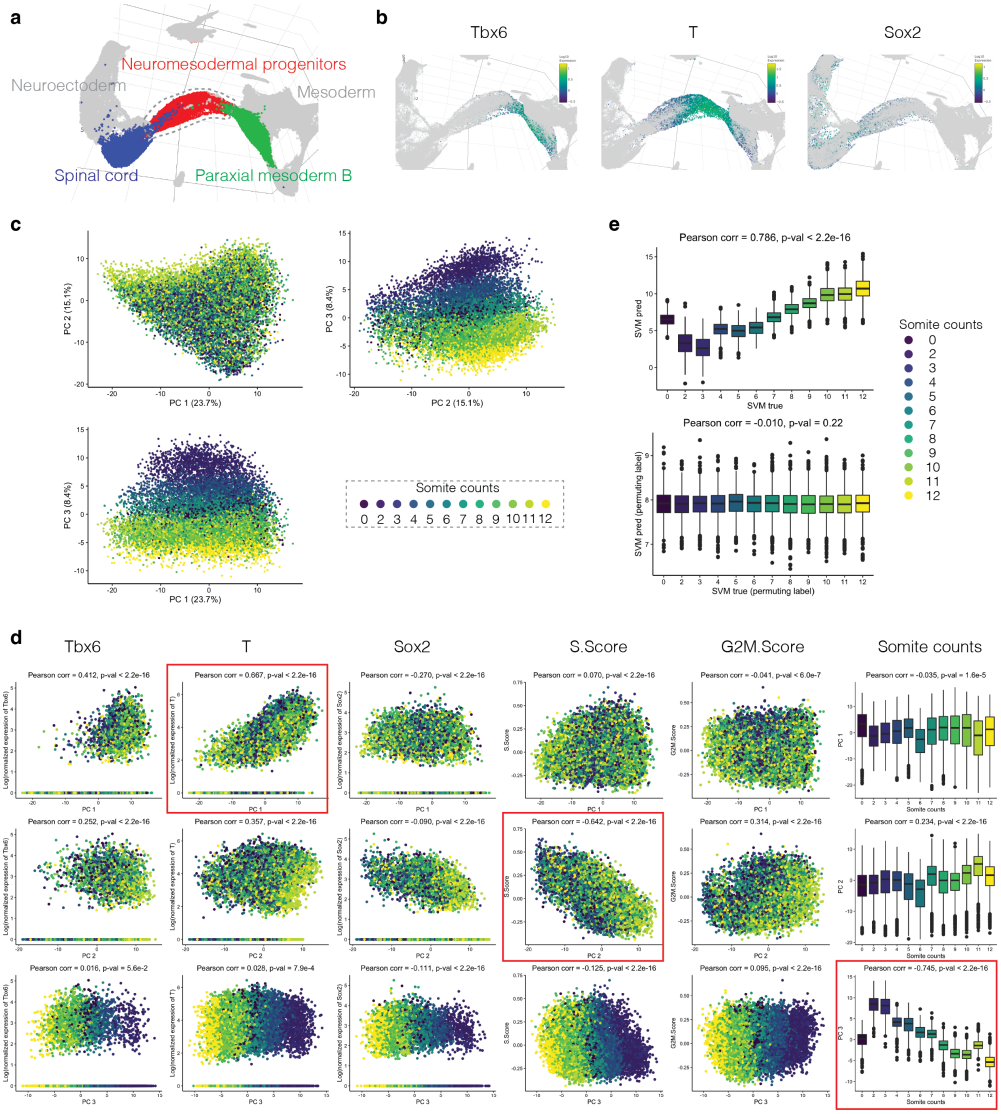

Figure 2.15: **Supplementary Figure 8. Decoding of transcriptional heterogeneity within neuromesodermal progenitors (NMP). a,** Subview of global 3D UMAP visualization highlighting spinal cord (blue), neuromesodermal progenitors (red), and paraxial mesoderm B (green). **b,** The same 3D UMAP as panel a but zooming in to highlight NMP cells, colored according to expression levels of markers of mesodermal (*Tbx6*, *T*) or neuroectodermal (*Sox2*) state [72, 73]. **c,** Embeddings of NMP cells in PCA space with visualization of top three PCs. Cells are colored by the somite count of the originating embryo. **d,** Correlations between top three PCs (rows 1-3) and the normalized expression of selected genes (*Tbx6*, *T*, *Sox2*; columns 1-3), cell cycle indices (columns 4-5) or somite counts (column 6). Red boxes highlight the strongest absolute correlation in each row. **e,** The 114 genes most strongly correlated with PC3 (which appears to correlate to somite counts) were identified using Pearson correlation. We applied the sklearn.svm.LinearSVR function in scikit-learn/1.0 with 5-fold cross-validation, using the expression values of those genes as predictors.

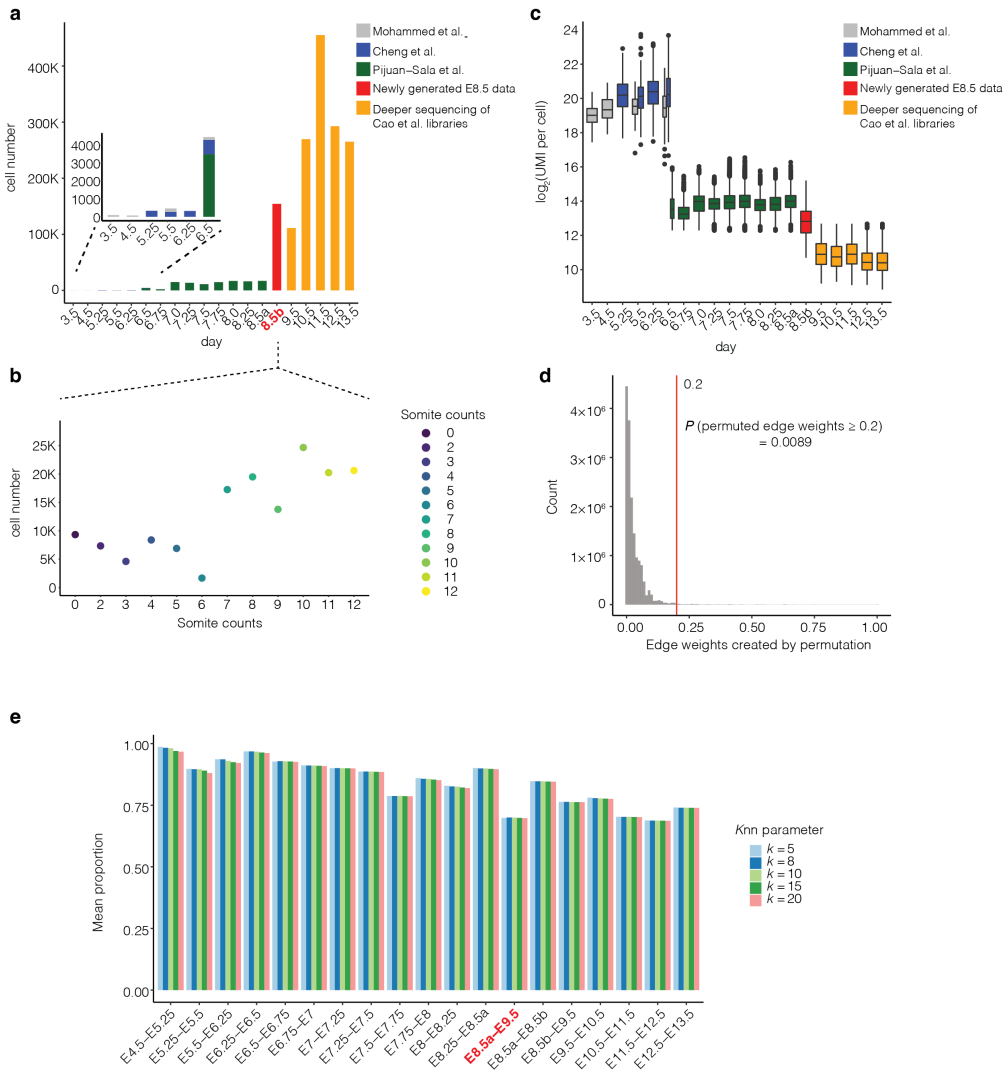

Figure 2.16: **Supplementary Figure 9. Integration of datasets spanning E3.5 to E13.5 of mouse development. a,** The number of cells per stage obtained from three previous studies, new E8.5 data obtained via optimized sci-RNA-seq3, and deeper sequencing of (Cao *et al.* 2019). **b,** The number of cells per embryo corresponding to specific somite counts from new E8.5 data. **c,** Box plot of log2(UMI counts) per cell across the stages and studies. **d,** The same strategy of creating the edges between adjacent timepoints was performed after randomly shuffling the cell-state annotations for cells within each timepoint, followed by repeating this process 1,000 times, resulting in a null distribution of edge weights. After permutation, less than 1% of potential edges are assigned weights greater than 0.2 (red line). **e,** To quantify the quality of the integration between adjacent timepoints, we focused on cells at the later timepoint assigned to annotations that were also present at the earlier timepoint. We then calculated the fraction of these cells' ancestral k-nearest neighbors (in the global 3D UMAP co-embedding) that were assigned the identical annotation. The mean proportion for different values of k are reported in the histogram.

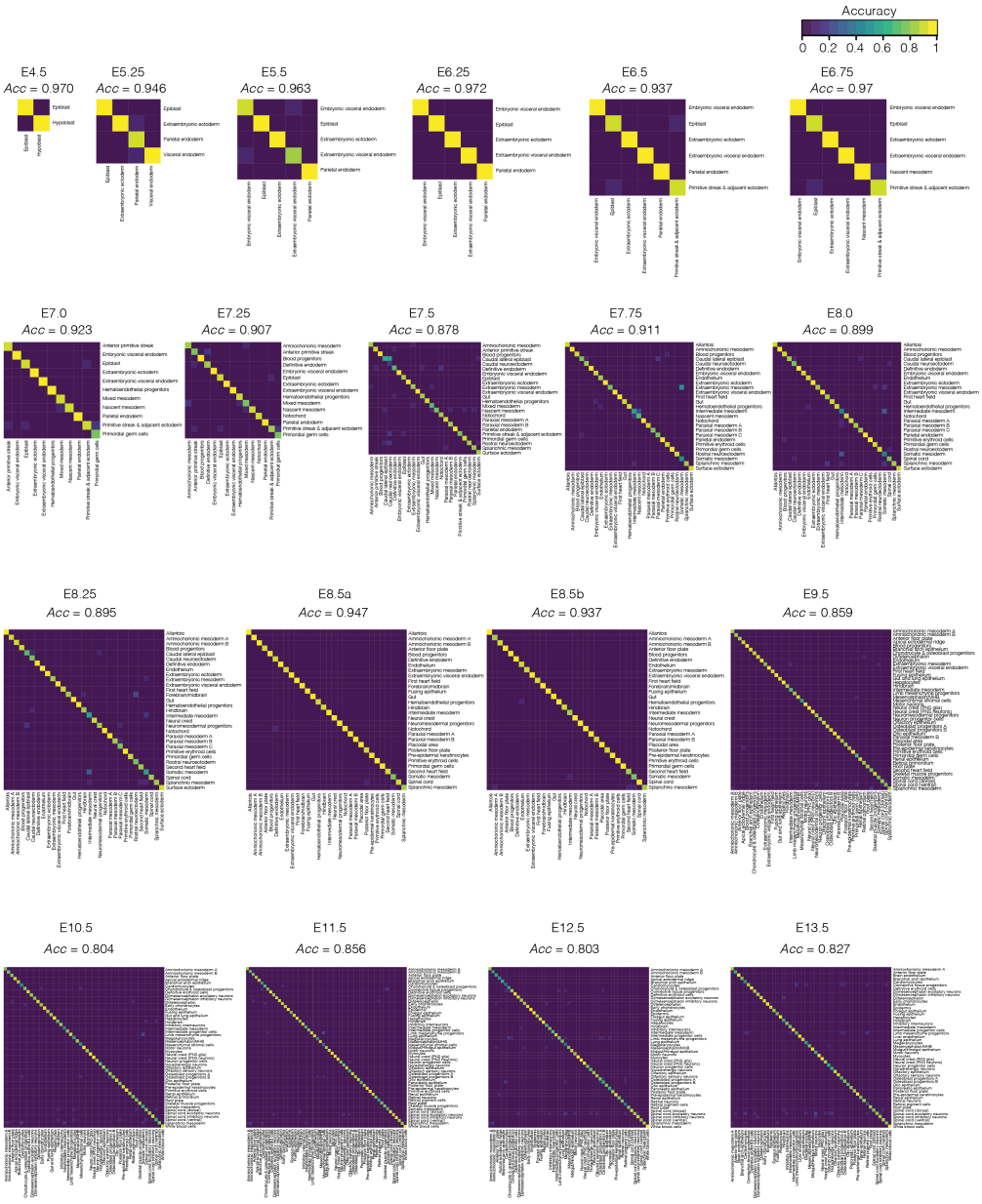

Figure 2.17: **Supplementary Figure 10. Benchmarking of the robustness of cell type annotations.** We applied the sklearn.svm.LinearSVC function in scikit-learn/1.0 with 5-fold cross-validation, using the expression values of all genes as predictors. Each heatmap shows the confusion matrix between true cell-type labels (rows) and predicted cell-type labels (columns) for cells within each individual timepoint, normalized to total counts per column (*i.e.* each column sums to one). The accuracy (Acc) across the whole matrix is shown above each heatmap. PNS: peripheral nervous system. MHB: midbrain-hindbrain boundary. Di: Diencephalon.

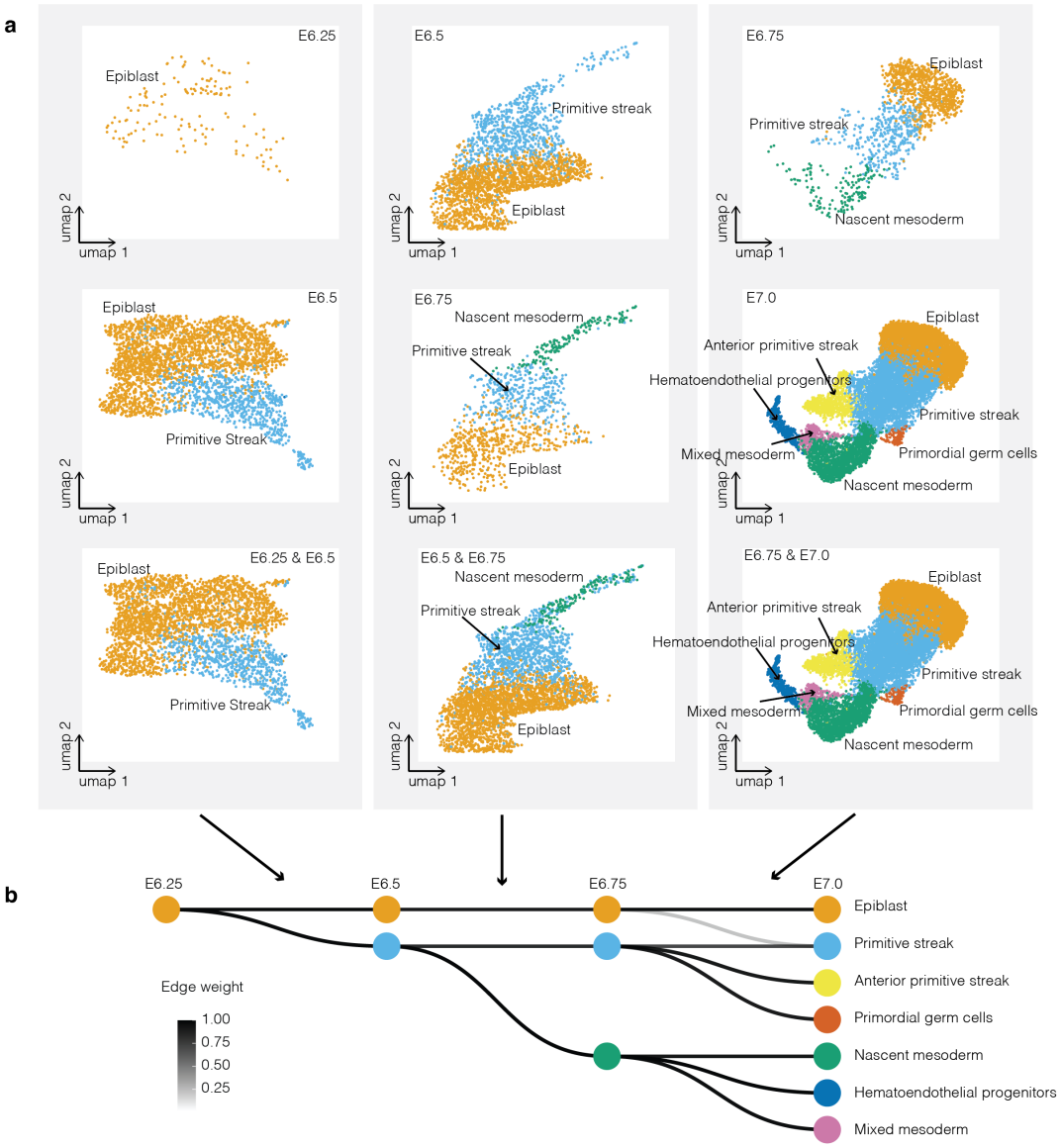

Figure 2.18: **Supplementary Figure 11. Inference of epiblast derivatives between E6.25 and E7.0. a,** A portion of the UMAP corresponding to the epiblast and its inferred derivatives is shown for co-embeddings of E6.25 - E6.5 (left column), E6.5 - E6.75 (middle column) and E6.75 - E7.0 (right column). Within each column is the same UMAP visualization, but showing only cells from the earlier timepoint (top row), the later timepoint (middle row) or both timepoints (bottom row). **b,** Directed acyclic graph showing inferred relationships between cell states amongst early epiblast derivatives. All edges with weights above 0.2 are shown in grey scale.

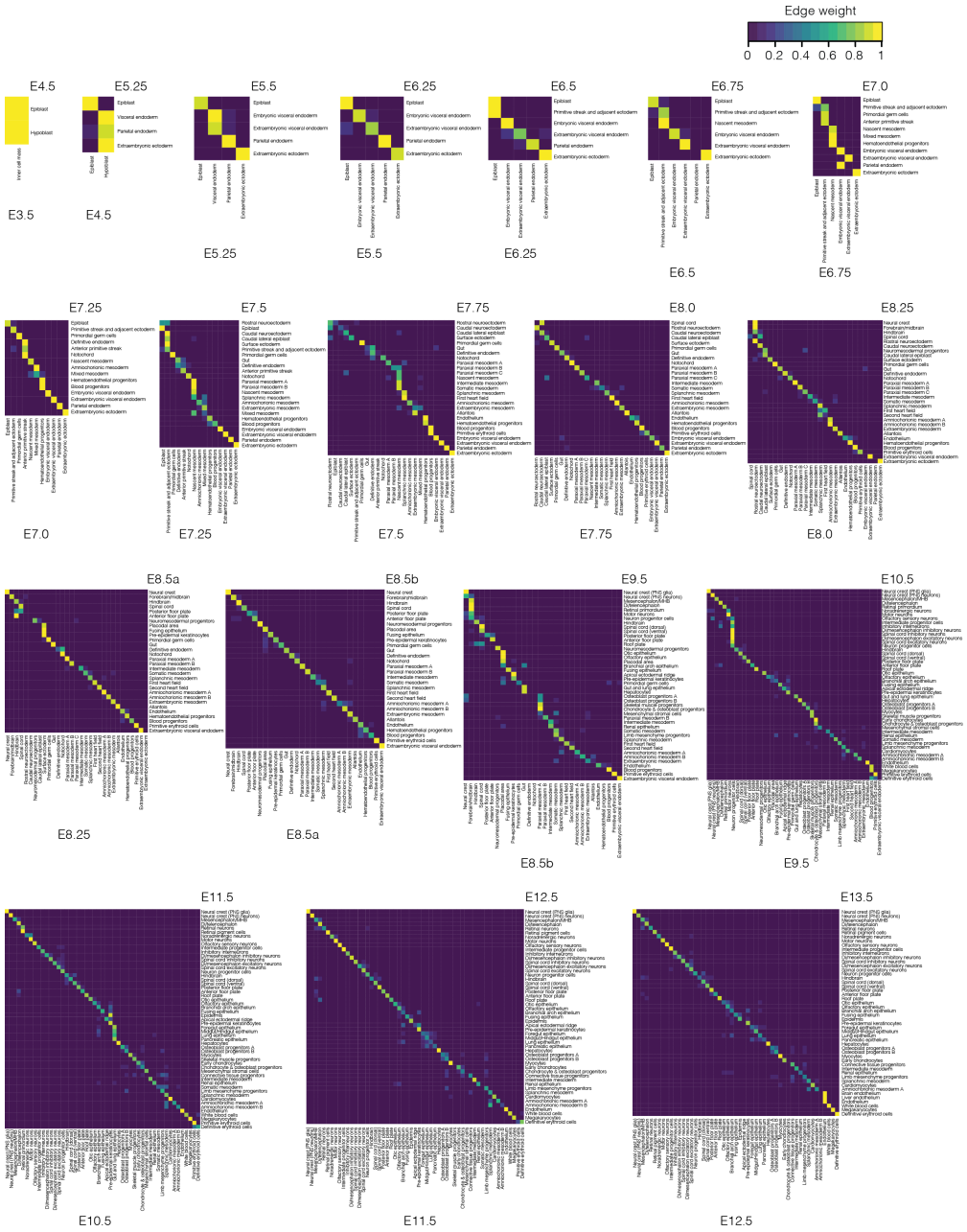

Figure 2.19: **Supplementary Figure 12. Heatmap of edge weights between cell states at each pair of adjacent timepoints.** Each heatmap shows edge weights between all cell states at a given timepoint (rows) and potential pseudo-ancestral cell states from the immediately preceding timepoint (columns). Edge weights were calculated based on a k-nearest neighbor (k-NN) based heuristic that was applied to a co-embedding of separately annotated cells from the adjacent timepoints. The edge weights range from 0 to 1, and edges with weights greater than 0.2 were carried forward. PNS: peripheral nervous system. MHB: midbrain-hindbrain boundary. Di: Diencephalon.

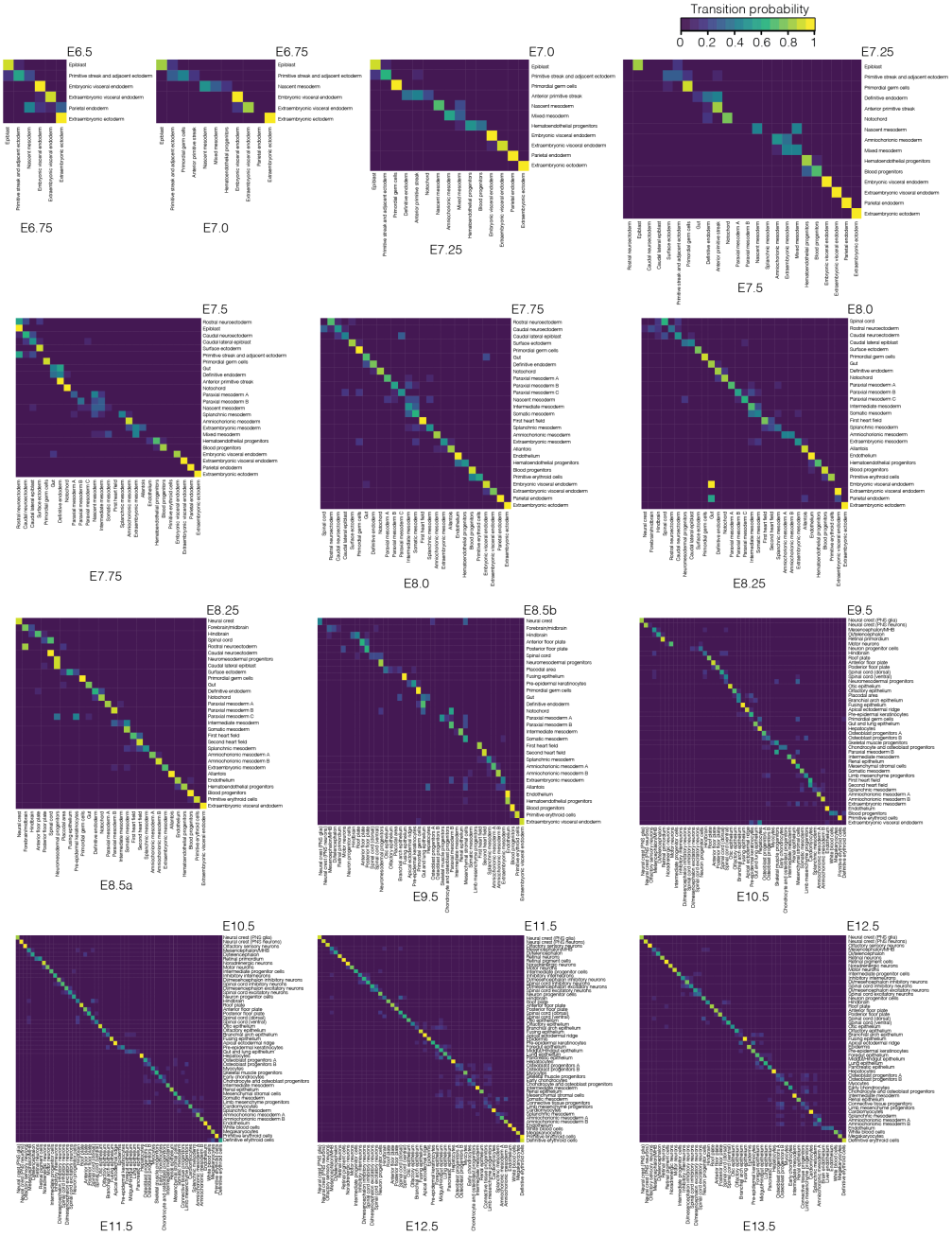

Figure 2.20: **Supplementary Figure 13. RNA velocity-based inference of potential cell state relationships across pairs of adjacent timepoints.** Briefly, we integrated cells from each pair of adjacent timepoints, and then applied RNA velocity analysis using scVelo [89]. The resulting transition probabilities between individual cells, were calculated using cosine correlation between the potential cell-to-cell transitions and the inferred velocity vector. To calculate the transition probability from cell state A at the earlier timepoint to cell state B at the later timepoint, we summed the transition probabilities of all cells within A to all cells within B, followed by normalizing the total cell number of B.

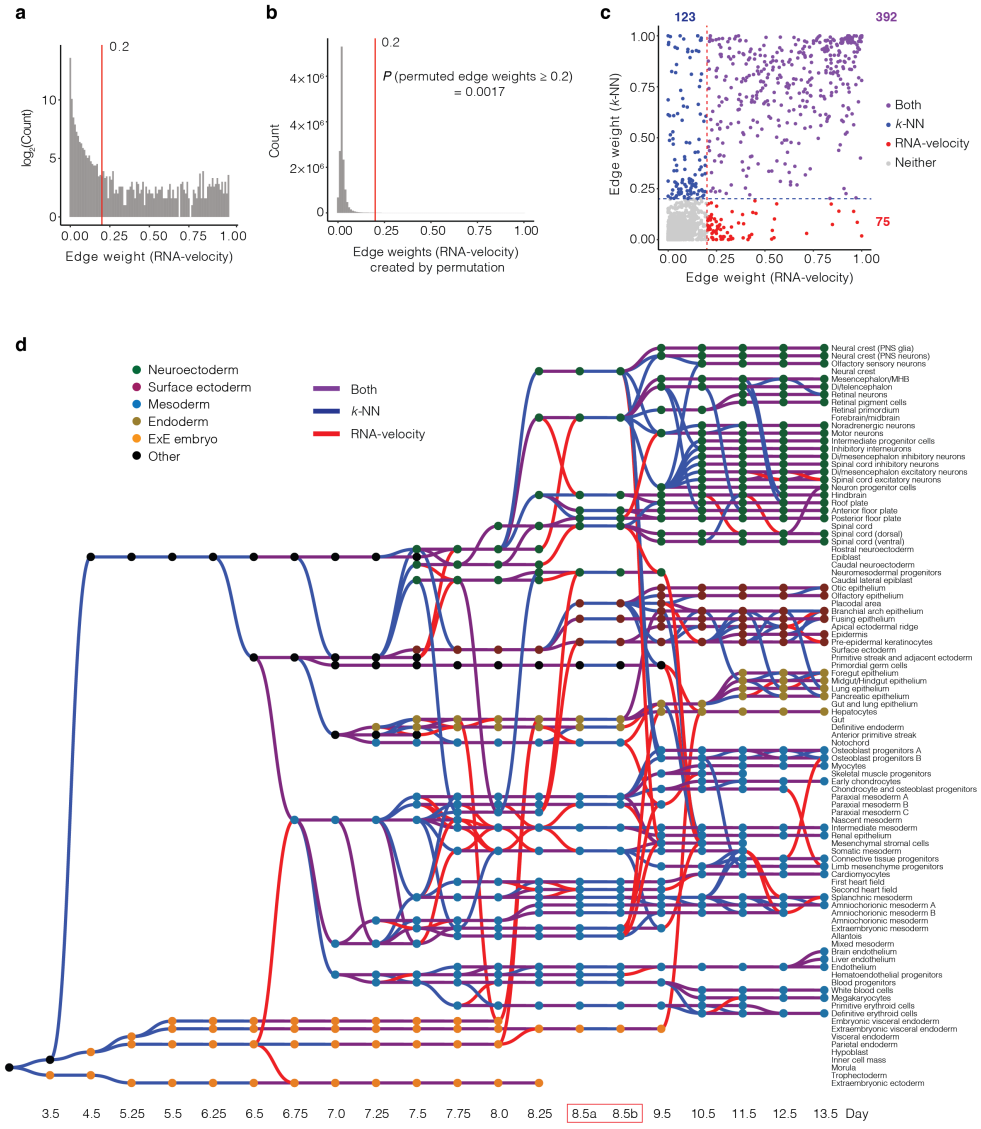

Figure 2.21: **Supplementary Figure 14. TOME edges nominated by *k*-NN vs. RNA velocity-based heuristics are largely concordant. a,** Histogram of all potential edge weights calculated by RNA-velocity. **b,** After calculating the transition probability for individual cells between adjacent timepoints using scVelo [89], the same strategy of creating the edges was performed after randomly shuffling the cell-state annotations for cells within each timepoint, followed by repeating this process 1,000 times, resulting in a null distribution of edge weights. **c,** Out of 15,261 potential edges, there were 123 edges nominated by the k-NN strategy only (weight > 0.2), and 75 edges nominated by the RNA velocity strategy only (weight > 0.2), and 392 nominated by both strategies. **d,** Directed acyclic graph showing inferred relationships between cell states across early mouse development.

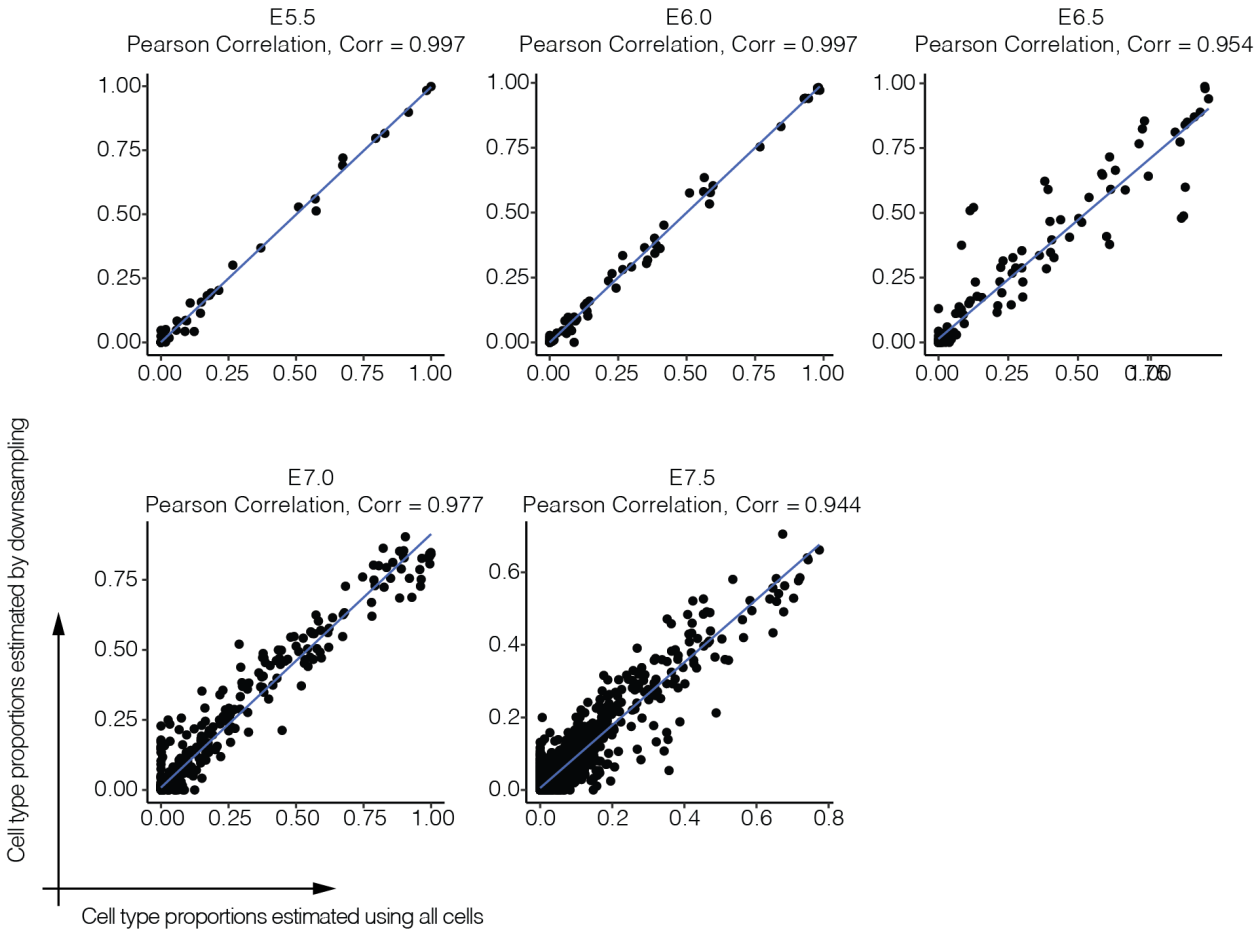

Figure 2.22: **Supplementary Figure 15. The inferred cell state proportions of each GEO-seq territory are robust to downsampling.** For timepoint which GEO-seq data was available (E5.5, E6.0, E6.5, E7.0, and E7.5), we estimated a gene expression signature for each cell state from scRNA-seq data, either by using all the cells or by downsampling to a maximum of 50 cells per state, and then repeated the inference of cell type contributors to each spatial territory of the gastrulating mouse embryo based on the application of CIBERSORTx to GEO-seq data. The Pearson correlation of resulting estimated cell state proportions for each GEO-seq territory with downsampling (y-axes) or without downsampling (x-axes) are shown. Of note, we did not use downsampling in the results shown in **Fig 4**.

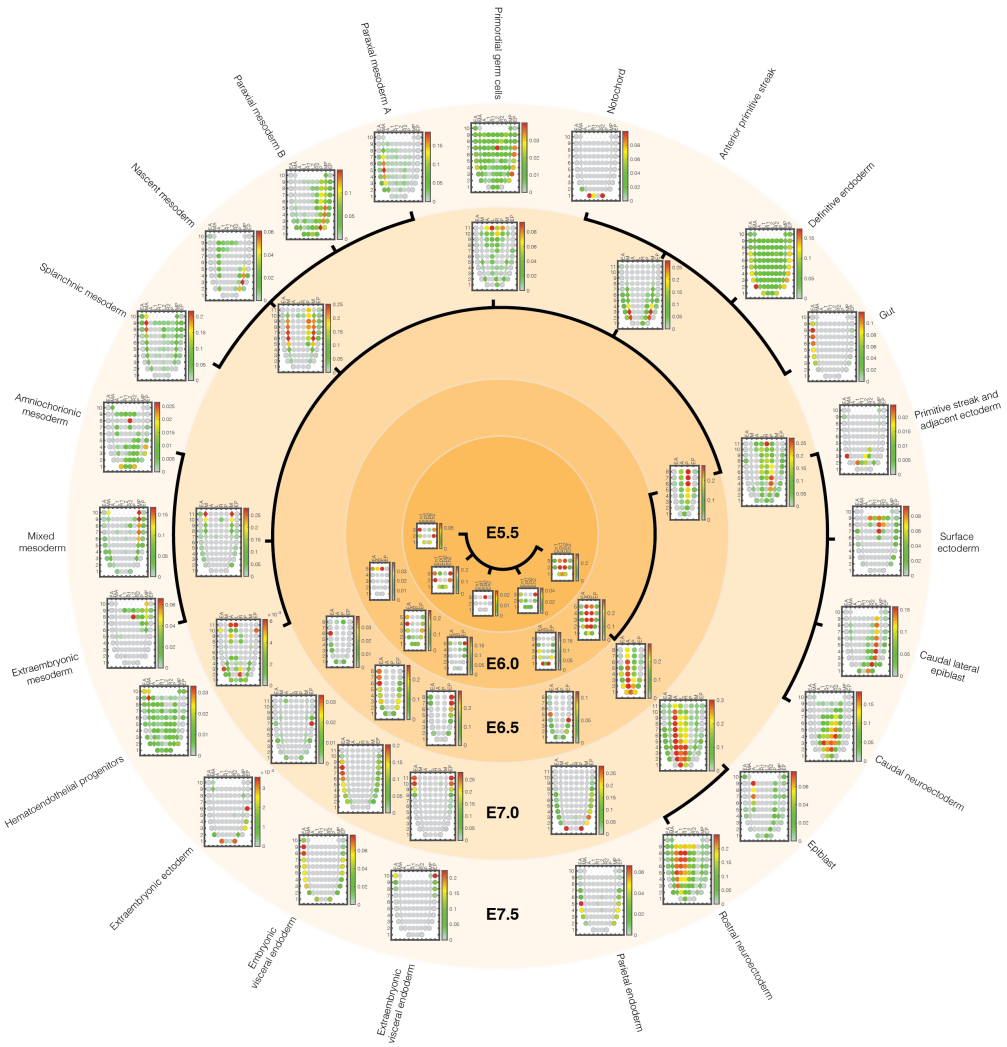

Figure 2.23: **Supplementary Figure 16. Estimated cell type proportions for different regions of the gastrulating mouse embryo, arranged by inferred cell type relationships over time.** As described in **Fig. 4a**, inference of cell type contributor(s) to each spatial territory of the gastrulating mouse embryo based on the application of CIBERSORTx to GEO-seq data [35, 33]. As scRNA-seq data from E6.0 was unavailable, we used data from E6.25 instead. Black edges correspond to edges between cell states over time estimated by TOME (only edges with the largest weights are shown). In each corn plot, each circle or diamond refers to a GEO-seq sample, and its weighted color to the estimated cell type composition. Corn plot nomenclature from [33].

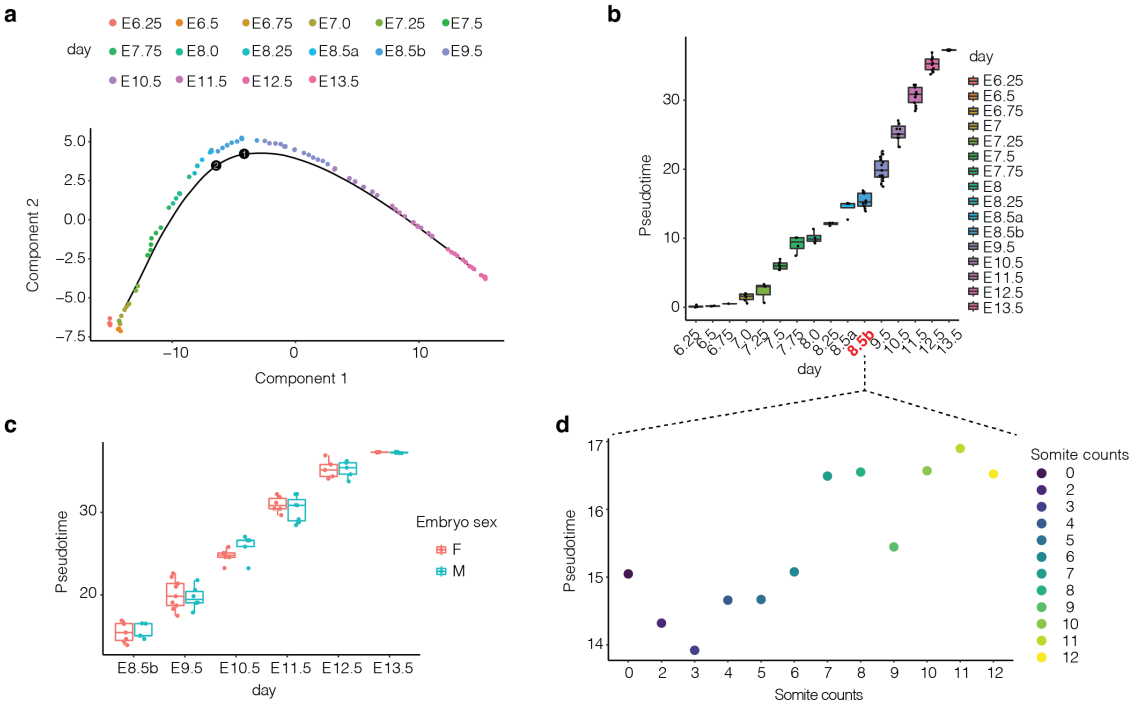

Figure 2.24: **Supplementary Fig 17. Inferring continuous molecular histories of individual cell types. a,** Pseudotime trajectory analysis of pseudobulk RNA-seq profiles of mouse embryos. Briefly, epiblast-derived cells from individual embryos (or pools of embryos comprising each sample, in the case of [13]) were aggregated to create 111 pseudobulk samples, on which we performed pseudotime trajectory analysis. Each point in the resulting 2D embedding corresponds to an embryo, and the curve to pseudotime trajectory. **b,** Pseudotime of embryos from staged timepoints between E6.25 and E13.5. **c,** Pseudotime of male and female embryos from staged timepoints between E8.5b to E13.5. 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*). **d,** Pseudotime of individual embryos with different somite counts from E8.5b.

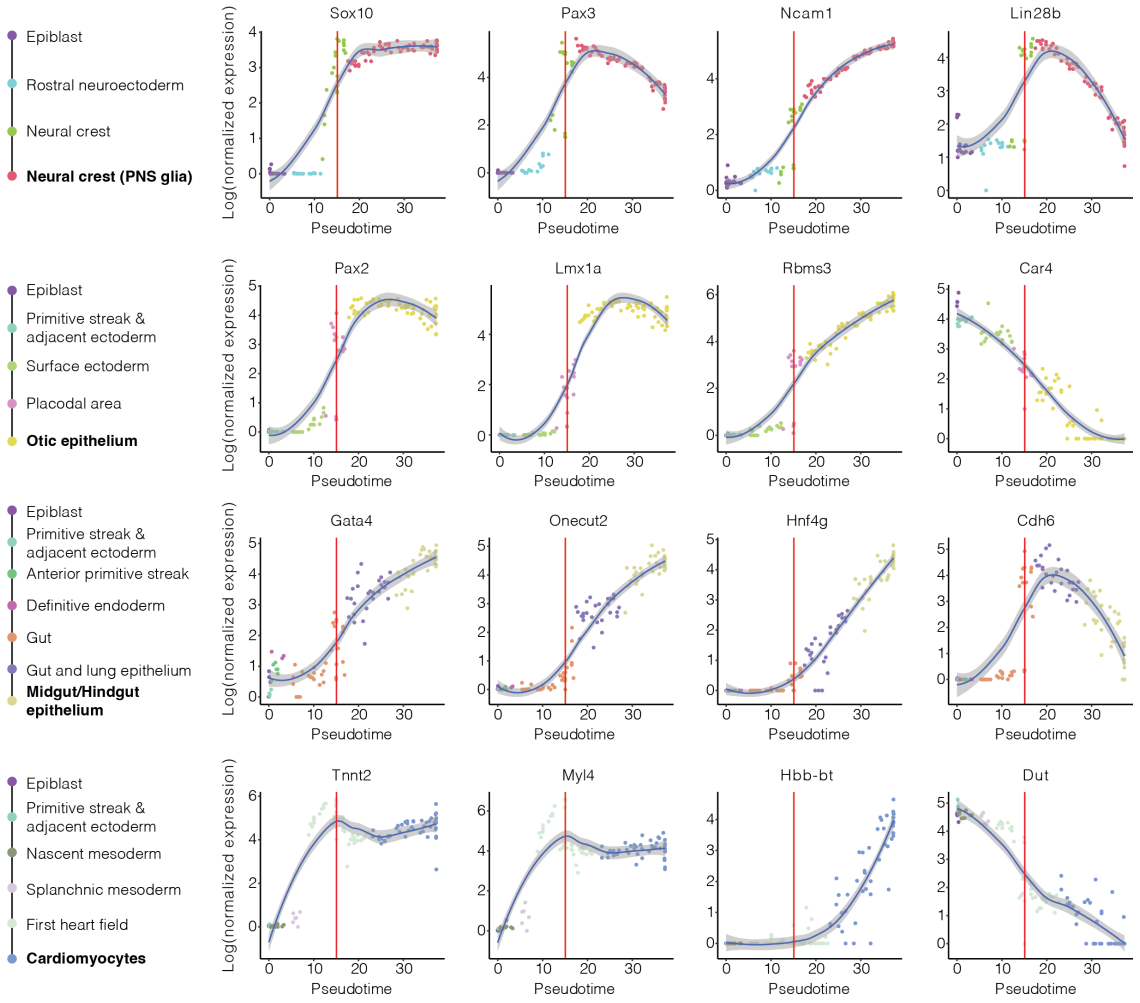

Figure 2.25: **Supplementary Figure 18. Smoothed expression profiles for four selected genes for each of four selected cell types (rows; one from each germ layer), along their inferred trajectories (key at left).** We selected linear paths corresponding to strongest pseudo-ancestor edges, working back from each E13.5 cell state to the E6.25 epiblast cell state. The first and second columns of plots correspond to key regulators or marker genes, and the third and fourth columns to the genes most positively and negatively correlated with pseudotime, respectively. Each plotted point corresponds to gene expression within a cell state for an individual embryo.

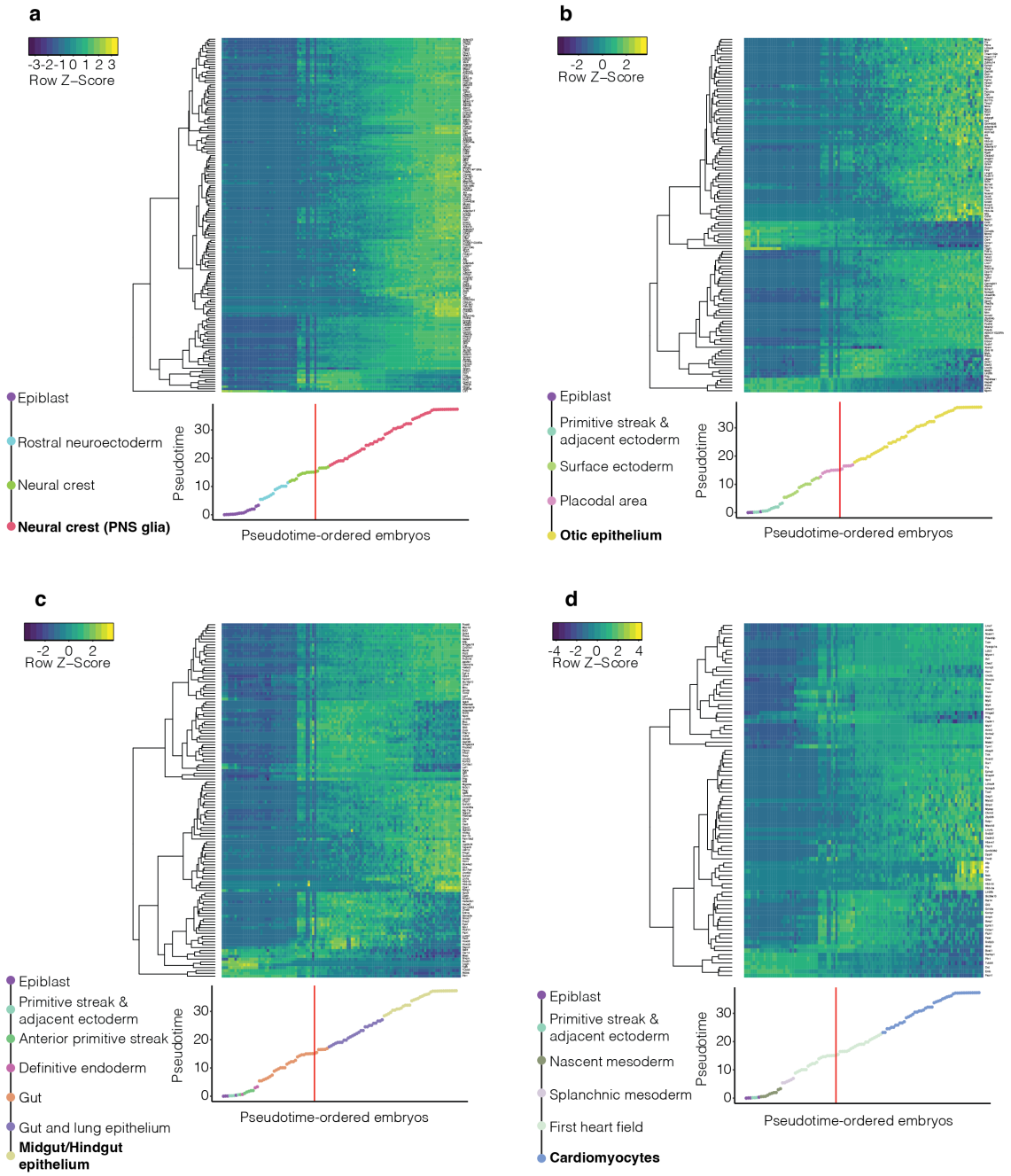

Figure 2.26: **Supplementary Figure 19. Gene dynamics across the inferred molecular trajectories of four selected cell types. a,** 155 genes were identified as significantly associated with pseudotime of the neural crest (PNS glia) trajectory, based on linear regression with the origin of the data as a covariate. **b,** 122 genes were identified as significantly associated with pseudotime of the otic epithelium trajectory. **c,** 124 genes were identified as significantly associated with pseudotime of the midgut/hindgut epithelium trajectory. **d,** 85 genes were identified as significantly associated with pseudotime of the cardiomyocyte trajectory.

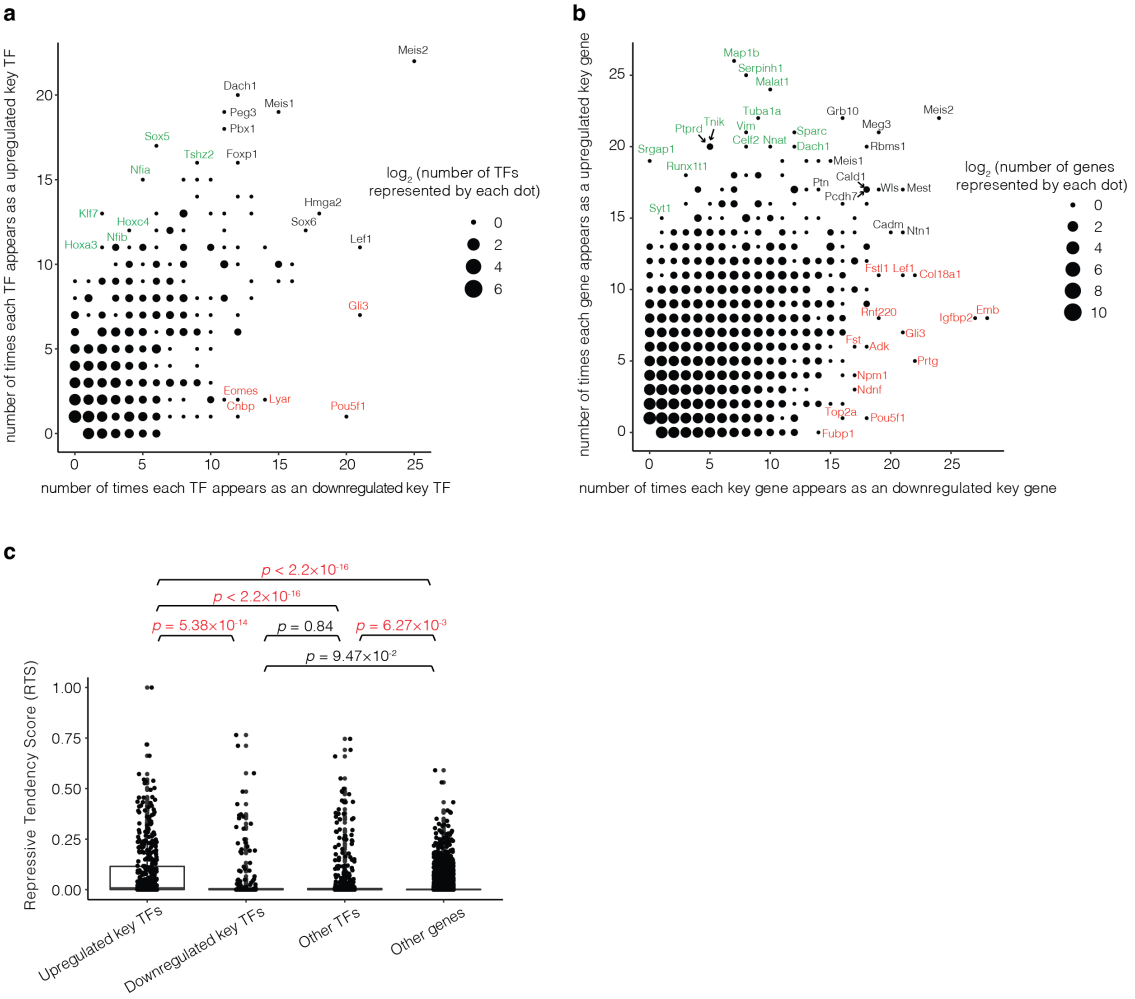

Figure 2.27: **Supplementary Figure 20. Recurrence of individual TFs or genes as candidate upregulated or downregulated key TFs or genes for mouse cell type specification. a,** TFs that are most often nominated as downregulated key TFs, *e.g. Pou5f1* (*Oct4*) are identified with red labels, while those most often nominated as upregulated key TFs, *e.g. Klf7*, are identified with green labels. **b,** Genes that are most often nominated as downregulated key genes, *e.g. Fubp1*, are identified with red labels, while those most often nominated as upregulated key genes, *e.g. Srgap1*, are identified with green labels. **c,** A file containing repressive tendency scores (RTS) was downloaded from [134]. It includes 16,298 mouse genes with RTS which define the association between each gene and broad H3K27me3 domains. Here the distribution of RTS is compared between 462 upregulated key TFs, 194 downregulated key TFs, 590 non-key TFs, and 15,052 non-TF genes.

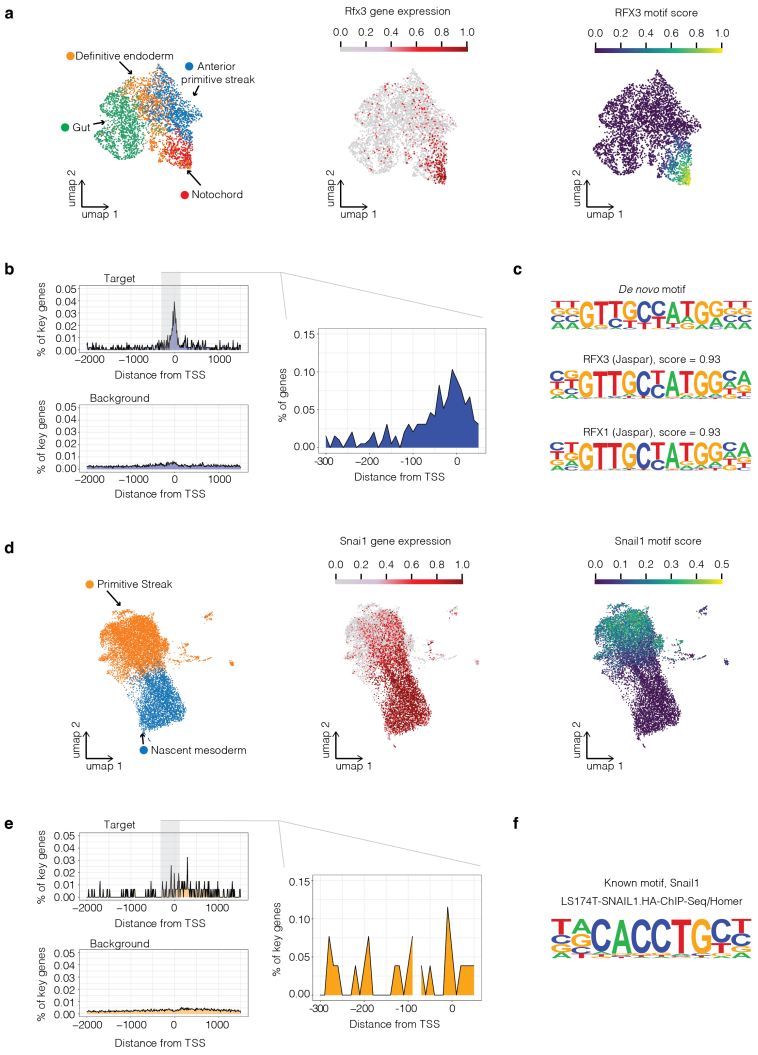

Figure 2.28: **Supplementary Figure 21. Correlation between key TF expression and up- or down-regulation of putative targets of regulation. a,** UMAP visualization of co-embedded cells from cell states including anterior primitive streak, definitive endoderm, gut, and notochord colored by cell type (left), *Rfx3* gene expression (middle) or RFX3 motif score (right), respectively. **b,** Positional bias of RFX3 binding motif along the core promoters of key genes for notochord emergence (right panel), an expanded region for key genes for notochord emergence (left top panel), or an expanded region for background (left bottom panel). **c,** The motif logo of the top de novo motif for notochord emergence and its two best alignments in the known motif database. **d,** UMAP visualization of co-embedded cells from cell states including primitive streak and nascent mesoderm colored by cell types (left), *Snai1* gene expression (middle) or SNAIL1 motif score (right), respectively. **e,** Positional bias of SNAI1 binding motif along the core promoters of key genes for nascent mesoderm emergence (right panel), an expanded region for key genes for nascent mesoderm emergence (left top panel), or an expanded region for background (left bottom panel). **f,** The known motif logo of SNAIL1.

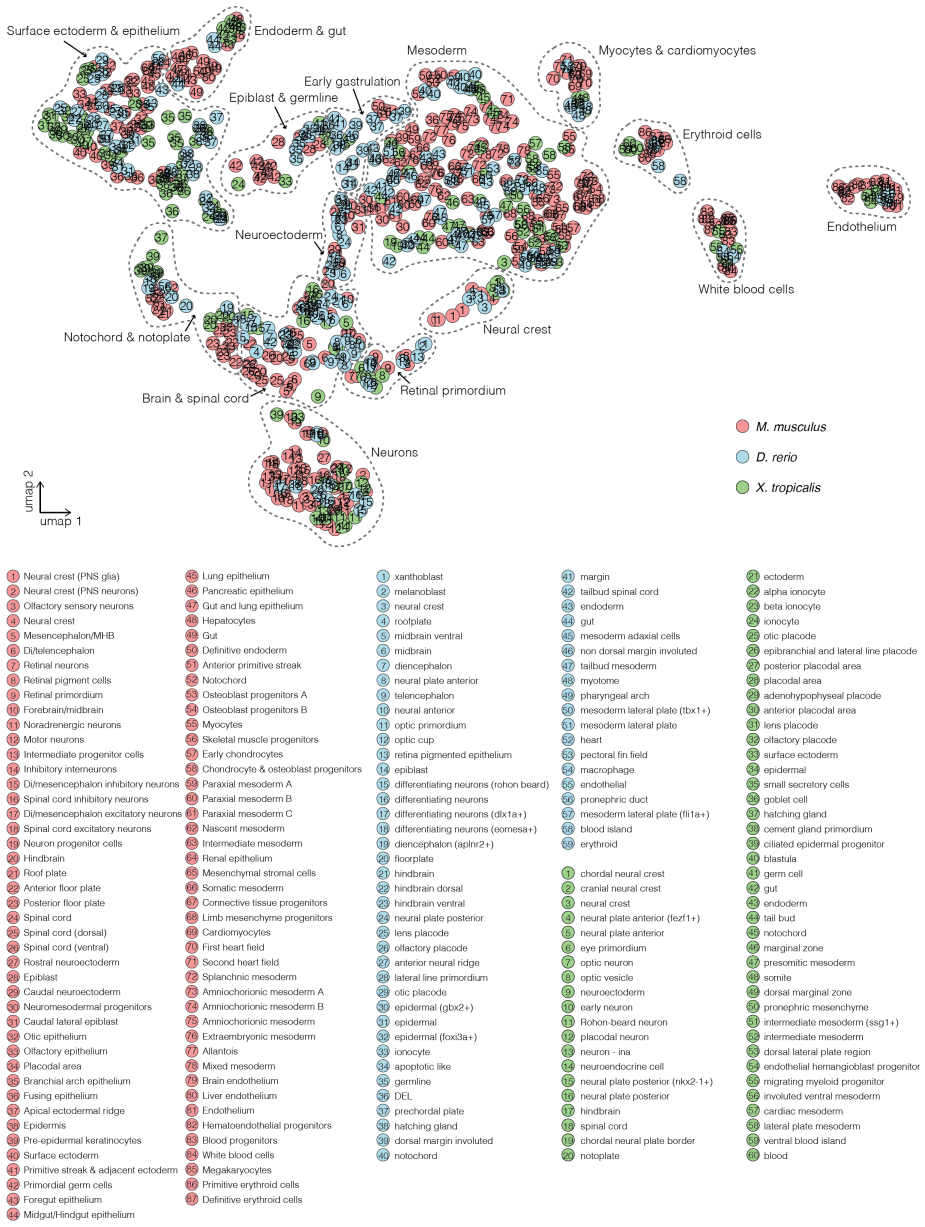

Figure 2.29: **Supplementary Fig 22. Co-embedding of 825 cell states from three species by integrating their transcriptional features.** For cell states spanning multiple timepoints, cells from each timepoint were treated separately for the purposes of this analysis. To create a transcriptional feature corresponding to each cell state (*i.e.* a pseudo-cell), we first averaged cell-state-specific UMI counts, normalized by the total count, multiplied by 100,000 and natural-log-transformed after adding a pseudocount. We then divided all resulting 825 pseudo-cells from the three species into four groups: the mouse single-cell group (n = 151), the mouse single-nucleus group (n = 277), the zebrafish group (n = 205), and the frog group (n = 192), and performed the anchor-based batch correction [51]. UMAP visualization shows co-embedded pseudo-cells from the mouse (red), the zebrafish (blue), and the frog (green).

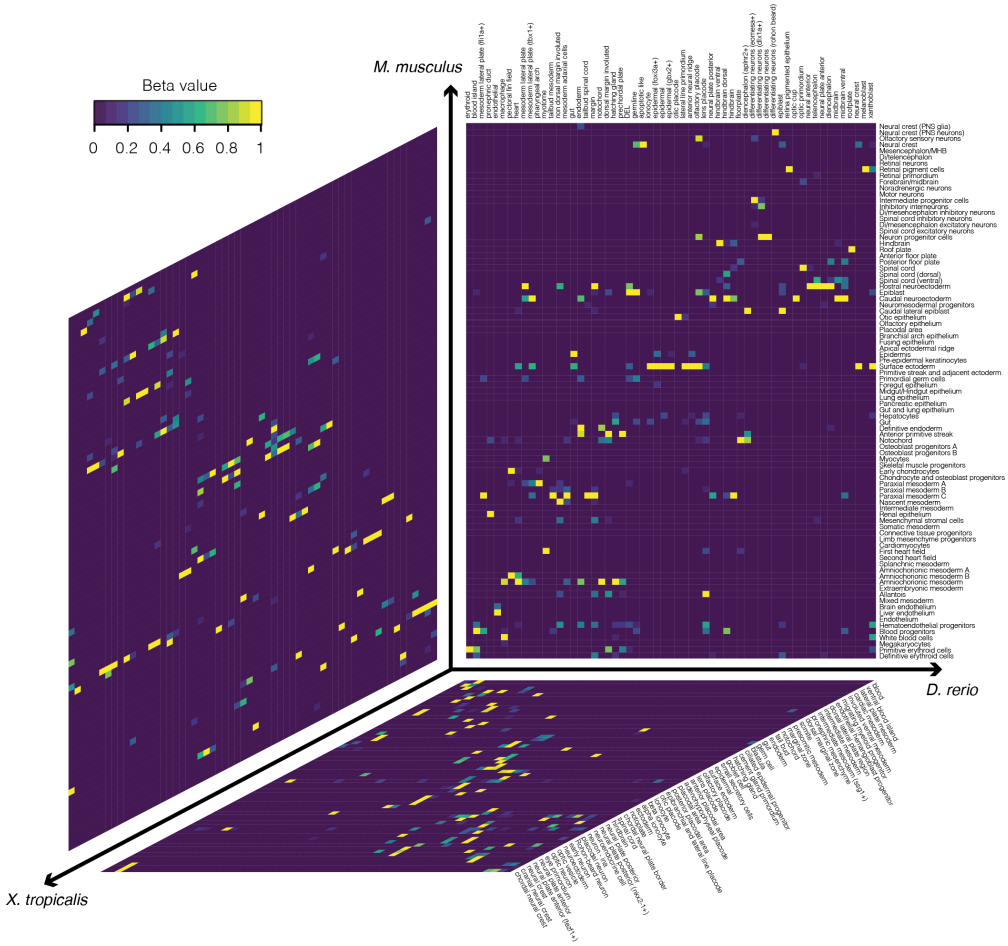

Figure 2.30: **Supplementary Figure 23. Correlated cell types between species based on non-negative least-squares (NNLS) regression.** To identify correlated cell types between each pair of species, we averaged cell-type-specific UMI counts, normalized by the total count, multiplied by 100,000 and natural-log-transformed after adding a pseudocount. We then applied NNLS regression to predict the gene expression of target cell type (T_a) in dataset A with the gene expression of all cell types (M_b) in dataset B: $T_a = \beta_{0a} + \beta_{1a} M_b$, based on the union of the 1,200 most highly expressed genes and 1,200 most highly specific genes in the target cell type. Shown here is a heat map of the normalized b values between 87 cell types from the mouse, 59 cell types from the zebrafish, and 60 cell types from the frog.

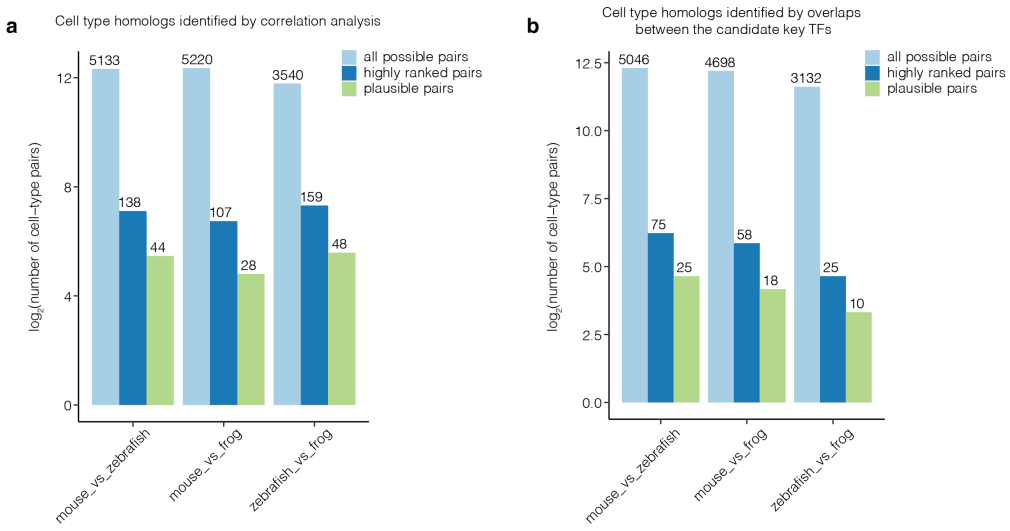

Figure 2.31: **Supplementary Figure 24. The log-scaled number of all possible pairs, highly ranked pairs, and biologically plausible pairs of cell types identified in pairwise comparisons of three species (mouse, zebrafish, frog) by two strategies. a,** The log2-scaled number of all possible pairs, highly ranked pairs, and biologically plausible pairs of cell types evaluated by non-negative least-squared (NNLS) regression. **b,** The log2-scaled number of all possible pairs, highly ranked pairs, and biologically plausible pairs of cell types evaluated on the basis of overlapping, orthologous candidate key TFs.

Chapter 3

A SINGLE-CELL TRANSCRIPTIONAL TIMELAPSE OF MOUSE EMBRYONIC DEVELOPMENT, FROM GASTRULA TO PUP

This Chapter is adopted from published work with minimum changes:

Chengxiang Qiu ∗ #, Beth K. Martin ∗, Ian C. Welsh ∗, Riza M. Daza, Truc-Mai Le, Xingfan Huang, Eva K. Nichols, Megan L. Taylor, Olivia Fulton, Diana R. O'Day, Anne Roshella Gomes, Saskia Ilcisin, Sanjay Srivatsan, Xinxian Deng, Christine M. Disteche, William Stafford Noble, Nobuhiko Hamazaki, Cecilia B. Moens, David Kimelman, Junyue Cao, Alexander F. Schier, Malte Spielmann, Stephen A. Murray, Cole Trapnell, Jay Shendure # "A single-cell transcriptional timelapse of mouse embryonic development, from gastrula to pup", *bioRxiv*, 2023.

∗: co-first authors; #: corresponding authors

We had been planning this project for many years, but it wasn't until early 2022 that we were able to make it happen. Several factors came together to make it possible. First, our work on the TOME project highlighted the importance of data quality, especially in terms of cell number, UMI count, and temporal resolution. We found that the trajectories obtained from the gastrulation stage were more reliable than those from the organogenesis stage. This is because the increasing embryo size and complex cell states required much higher quality scRNA-seq data. Second, Beth optimized the sci-RNA-seq3 protocol during the revision of the TOME paper. This improvement led to a significant increase in data quality. The

UMI per cell is now around 8,000, which is very impressive. Third, Ian collected and staged mouse embryos at the early somitogenesis stage, with somites from 0 to 12. The dynamic transcriptome changes in selected cell types, such as NMPs, motivated us to extend this time series.

I am deeply grateful to Jay, Beth, and Ian for making this project possible. As a developmental biologist, I am incredibly excited to have this opportunity to analyze the data and contribute to our understanding of mouse embryogenesis. I believe that this project could be a milestone in our understanding of this complex process. After we posted the paper on *bioRxiv*, I received a lot of emails from people who are interested in particular developmental stages or cell types. I imagine that what we have done on the data analysis side is relatively shallow. We encourage experts from all fields to dig deeper into the data and discover new insights. Actually, we had many thoughts or initial analysis results that we couldn't put in the paper due to space limitation.

The large size of the data (the first scRNA-seq dataset with >10 million cells from a single study) posed new challenges for computational analysis. I had to think about the most efficient way to analyze the data. For example, analyzing 10,000 cells is very different from analyzing 10 million cells. I needed to consider which package or function I should use, whether downsampling is appropriate, and how to store the data more efficiently, since every test would take much longer. During this process, I switched my preference from R to Python, which is much faster and more memory efficient. The new BPCells package in R saves a lot of memory and overcomes the previous language limitation (vector size), but the analysis time is still much longer (*e.g.* performing dimension reduction or creating a nearest neighbor graph). All of these experiences will be valuable for my future work.

During this process, I came to understand that data is not knowledge. High-quality data is essential, but better data analysis is more important for gaining new and useful insights

from the data. This is similar to wet lab work, where every unreasonable observation needs to be investigated further. If every step of the analysis makes sense, then the unreasonable observation may be true and could lead to some really cool findings.

More formally, the author contributions are listed in the manuscript as follows: J.S., M.S., C.Q. designed the research. I.C.W. collected and staged the mouse embryos, and wrote the corresponding parts of the manuscript. B.K.M. developed the optimized sci-RNA-seq3 protocol and generated the data (with assistance from R.M.D., T-M.L., E.N., M.T., O.F., D.R.O., A.R.G., S.I.), and wrote the corresponding parts of the manuscript. C.Q. performed all computational analyses. X.H., S.S., W.S.N., and C.T. assisted with data analysis. I.C.W., E.N., X.D., C.M.D., N.H., J.C., C.B.M., D.K., A.F.S., M.S., S.A.M., C.T. assisted with results interpretation. C.Q. and J.S. collaboratively explored and annotated the data, and wrote the manuscript, except for sections corresponding to mouse collection/staging and data generation. J.S. supervised the project.

Abstract

The house mouse, *Mus musculus*, is an exceptional model system, combining genetic tractability with close homology to human biology. Gestation in mouse development lasts just under three weeks, a period during which its genome orchestrates the astonishing transformation of a single cell zygote into a free-living pup composed of >500 million cells. Towards a global framework for exploring mammalian development, we applied single cell combinatorial indexing (sci-$*$) to profile the transcriptional states of 12.4 million nuclei from 83 precisely staged embryos spanning late gastrulation (embryonic day 8 or E8) to birth (postnatal day 0 or P0), with 2-hr temporal resolution during somitogenesis, 6-hr resolution through to birth, and 20-min resolution during the immediate postpartum period. From these data (E8 to P0), we annotate hundreds of cell types and perform deeper analyses of the unfolding of the posterior embryo during somitogenesis as well as the ontogenesis of the

kidney, mesenchyme, retina, and early neurons. Finally, we leverage the depth and temporal resolution of these whole embryo snapshots, together with other published data, to construct and curate a rooted tree of cell type relationships that spans mouse development from zygote to pup. Throughout this tree, we systematically nominate sets of transcription factors (TFs) and other genes as candidate drivers of the *in vivo* differentiation of hundreds of mammalian cell types. Remarkably, the most dramatic shifts in transcriptional state are observed in a restricted set of cell types and initiate in the first hour following birth, and presumably underlie the massive changes in physiology that must accompany the successful transition of a placental mammalian fetus to extrauterine life.

3.1 Introduction

The last ten years have witnessed the development and application of single-cell molecular profiling technologies to characterize biological development at the scale of the whole organism[5, 6, 7, 8, 9, 13, 15, 53, 121, 14]. Most of these studies comprise time series, in which each embryo profiled with single-cell RNA-seq (scRNA-seq) or ATAC-seq (scATAC-seq) captures one moment in developmental time, yielding "snapshots" that must be pieced together analogous to a timelapse movie. Inevitably, there are tradeoffs between the span of development studied and both the frame rate and resolution of the snapshots taken. For example, in the mouse, two studies intensely profiled gastrulating embryos, together quantifying gene expression in 150,000 cells from over 500 embryos spanning E6.5 to E8.5[13, 14], while a study from our group profiled 2 million nuclei from 61 embryos staged at roughly 24-hr intervals spanning E9.5 to E13.5[15]. We systematically integrated these and other scRNA-seq snapshots of whole mouse embryos, overcoming various batch effects to obtain a coarse tree relating mammalian transcriptional states from E3.5 to E13.5[152]. However, these studies did not examine later stages of prenatal development, which are more difficult to subject to whole embryo single cell profiling due to the sheer number of cells as well as

the emergence of bone, which is difficult to pulverize.

Here we set out to deeply profile the molecular states of single nuclei derived from precisely staged mouse embryos spanning late gastrulation (E8) to birth (P0). The resulting dataset greatly improves upon our previously reported single cell atlas of early organogenesis in the mouse[15] with respect to: (a) the number of single nuclei profiled (2 million to 12.4 million single nuclei), (b) the depth of profiling (median 671 to 2,545 unique molecular identifiers [UMIs] per nucleus), (c) temporal resolution (24-hr to 20-min, 2-hr or 6-hr intervals), and (d) the span of development covered (E9.5-E13.5 to E8-P0).

We focus here on describing the dataset and our early analyses, which include the preliminary annotation of hundreds of cell types and deeper dives into the ontogenesis of selected systems. Improving upon our previously described approach[152], we also leverage these and other published scRNA-seq snapshots of whole mouse embryos to curate a data-driven tree of mouse development that relates cell types throughout prenatal development, as well as to nominate sets of TFs and other genes as candidate drivers of the *in vivo* differentiation of each mammalian cell type. Surprisingly, we find that birth (E18.75 to P0) is associated with a dramatic shift in transcriptional state in a restricted subset of cell types. A more detailed time-course of additional embryos at 20-min intervals in the immediate postpartum period confirmed narrowed the onset of these transitions to the first hour after birth, and highlighted ways in which rapidly activated gene expression programs in specific cell types may support the adaptation from fetal to extrauterine life. Altogether, our studies provide a framework for exploring the entirety of mammalian embryogenesis, from single cell zygote to newborn pup.

3.2 Ontogenetic staging and embryo selection

Towards a more continuous view of single cell transcriptional dynamics throughout embryonic development, we sought to profile whole mouse embryos at a higher "frame rate" than the 24-hr intervals of our previous study of early organogenesis[15]. To enable rigorous comparisons between embryos within our sample set, we distinguish between gestational stage and absolute developmental stage of individual embryos harvested. Mouse gestational age, based upon the observation of a vaginal plug for which noon on that day is counted as E0.5, provides a loose approximation of the time elapsed from conception to the time a litter is harvested. Importantly, stochastic differences in timing of mating or fertilization, and genetic factors such as strain/species-dependent gestational length and litter size, can result in significant intra- and inter-litter variation of embryos of an identical gestational stage[153]. Conversely, embryonic morphogenesis is highly ordered, reproducible, and inherently reflective of an embryo's developmental age with respect to absolute position within a morphogenetic trajectory and the dynamic progression of underlying cellular transcriptional states[154, 155, 5]. Therefore, we sought to stage individual embryos based on well-defined morphological criteria. Each embryo was assigned to one of 45 temporal bins at 6-hr increments from E8 to P0 (**Fig. 1a; Supplementary Figs. 1-2**). For the earlier stages (before E10), embryos were staged based on morphological features (**Supplementary Fig. 1**) and somites counted at the time of harvest, facilitating temporal binning at roughly 2-hr increments. For E10.25 to E14.75 embryos, developmental age was determined by measurements of hindlimb bud development using the Embryonic Mouse Ontogenetic Staging System (eMOSS), which leverages dynamic changes in hindlimb bud morphology and a pattern recognition algorithm to estimate the absolute stage of a sample[156, 157]. Due to the increased complexity of limb morphology at later stages, automated staging beyond E15 is not possible. Therefore, harvests for all remaining embryonic samples (E15 to E18.75) were performed precisely at 00:00, 06:00, 12:00, and 18:00 on the targeted gestational day. From close inspection of limbs in

this set, we defined additional dynamics related to digit morphogenesis that allowed further binning of samples collected on days 15 and 16 (**Supplementary Fig. 2**). The remaining timepoints (E17 and beyond) were staged based on hour of harvest and gestational age.

From a total of 523 staged embryos, we selected 75 for whole embryo scRNA-seq, aiming for one embryo for every somite count from 0 to 34, and then one embryo for every 6-hr bin from E10 to P0 (**Supplementary Table 1**). Embryos with somite counts 1, 13, and 19 are missing from the series, while a handful of post-E10 timepoints are represented by 2-3 individual embryos. Wherever possible, we sought to alternate between male and female embryos in neighboring bins; for the final P0 timepoint, both sexes were deeply profiled.

3.3 Single nucleus transcriptional profiling of whole embryos

Frozen embryos were processed via a recently optimized protocol for three-level single cell transcriptional profiling by combinatorial indexing (sci-RNA-seq3) [58]. In each of 15 sci-RNA-seq3 experiments, we generally sought to include embryos from adjacent stages of development (**Supplementary Table 1**). Sequencing data was generated across 21 runs of the Illumina NovaSeq instrument, with all libraries sequenced to either a PCR duplication rate of >50% or a median UMI count of >2,500 per nucleus (**Supplementary Table 2**). Reads from each sci-RNA-seq3 experiment were demultiplexed, trimmed and mapped to the mouse reference genome (mm10). This was followed by the removal of PCR duplicates. Finally, poor quality nuclei and potential doublets were aggressively filtered (**Methods**)[15, 152]. As nuclei from only one embryo are deposited to each well during the first indexing round, we retain the identity of the embryo from which each single nucleus profile is derived[15].

The cell-by-gene count matrix that we carried forward combines profiles from all 15 experiments, and includes transcriptional profiles for 11,441,407 nuclei from 74 embryos

spanning E8 to P0 (**Fig. 1b**). Of note, 1% of these single nucleus profiles (data from embryos with somite counts 0-12) were reported in a previous publication from our group[152]. On average, 154,614 nuclei were profiled per embryo, with recovery generally correlated with stage (**Fig. 1c; Supplementary Table 1;** median 142,854; range = 1,679 to 1,613,834). The median UMI count obtained per nucleus was 2,700 and the median number of genes detected was 1,574 (**Supplementary Fig. 3a**).

We did not perform batch correction across experiments, as various analyses suggested that batch effects were relatively minimal. In particular, cells from the same time window but profiled by different experiments, or cells from adjacent timepoints but profiled by different experiments, were well-integrated (**Supplementary Fig. 3b; Supplementary Fig. 4**). Consistent with this, a principal component analysis (PCA) of pseudo-bulked RNA-seq profiles corresponding to each timepoint resulted in a major first component (PC1; 77.3%) strongly correlated with developmental time (**Fig. 1d**). We also checked for ambient noise (*e.g.* as might be due to transcript leakage) by examining highly abundant, highly cell-type-specific genes such as hemoglobins and collagens, and found it present at low levels, *e.g.* the mean number of UMIs for *Hbb-bs* was 10.8 in definitive erythroid cells and 0.26 in all other cells, and for *Col1a1* was 186 in pre-osteoblasts vs. 1.23 in all other cells (**Supplementary Fig. 5**).

What kind of "coverage" of all cells in the mouse embryo are we achieving here? As the total number of cells in developing mouse embryos is not well documented, we sought to experimentally estimate this by quantifying total DNA content across a series of staged embryos. In brief, we estimate that the embryo grows 3,000-fold between E8.5 and P0 (210K to 670M cells), with the cellular 'doubling time' of the embryo slowing from around 6 hours to 1.5 days across the same interval (**Fig. 1e; Supplementary Table 3; Methods**). Thus, even with the large number of nuclei profiled here, our "cellular coverage" of the embryo remains modest, ranging from 0.5-fold for early stages (summing across 6 embryos with

somite counts 7 to 12) to 0.002-fold immediately before birth (summing across 6 embryos staged E17.5 to E18.75).

3.4 Preliminary annotation of major cell clusters and cell types

To get our bearings on this very large cell-by-gene count matrix, we used Scanpy[141] to generate a global embedding of the 11,441,407 nuclei (hereafter referred to as cells) from all experiments and timepoints, and then annotated 26 major cell clusters based on marker genes (**Fig. 1e-f; Supplementary Table 4**). The major cell clusters assigned the most cells were the mesoderm (n = 3,267,338), central nervous system (CNS) neurons (n = 2,106,206), neuroectoderm & glia (n = 1,733,663) and definitive erythroid (n = 1,033,409) lineages.

We emphasize that the resolution of these major cell clusters is somewhat arbitrary and impacted by abundance, *e.g.* we have separate major cell clusters for B, T and mast cells, but the remainder of blood cells, including hematopoietic stem cells (HSCs), are lumped into 'white blood cells'; intermediate neuronal progenitors and their derivatives are a major cell cluster, while nearly all other CNS neurons are lumped into a separate major cell cluster. Although most major cell clusters were very straightforward to annotate based on the literature and our previous experience with mouse scRNA-seq datasets up to E13.5[152, 15], a few first appear only late in organogenesis, including adipocytes, testes & adrenal cells (**Fig. 1e**). As expected, cell clusters whose proportions decline over time either stream towards derivatives (*e.g.* neuroectoderm & glia to CNS neurons, intermediate neuronal progenitors) or are displaced by a functionally analogous but developmentally distinct lineage (*e.g.* primitive erythroid to definitive erythroid).

We next performed sub-clustering of cells within each of the 26 major cell clusters, followed by annotation of each cluster based on marker genes, resulting in 190 labeled cell types

(**Supplementary Fig. 6; Supplementary Table 5**). These annotations are preliminary, and subject to refinement or correction as we and others further explore the data[152, 15].

We next sought to leverage this single cell timelapse of mouse development from gastrula to pup to perform deeper dives into the unfolding of the posterior embryo during somitogenesis, as well as the ontogenesis of the kidney, mesenchyme, retina, and early neurons.

3.5 The posterior embryo during somitogenesis

Neuromesodermal progenitors (NMPs) are a fascinating population of bipotent cells that give rise to the spinal cord as well as trunk and tail somites[158]. We previously profiled late-gastrulation embryos staged in one-somite increments (0 to 12 somites) and investigated transcriptional heterogeneity among early NMPs[152]. With the goal of extending this analysis to later time points and also relating NMPs to other physically coincident cell types in the posterior embryo, we extracted and re-embedded cells from all somite-staged embryos (0 to 34 somites) annotated as NMPs & spinal cord progenitors, mesodermal progenitors (*Tbx6+*), notochord or gut (**Fig. 2a-b**).

Focusing first on the subset of cells annotated as either 'NMPs & spinal cord progenitors' or 'mesodermal progenitors (*Tbx6+*)' (cluster 1 in **Fig. 2a**), we performed PCA on highly variable genes. The top three PCs, which explain nearly half of transcriptional variation in these cells, appear to correspond to the differences between neuroectodermal vs. mesodermal fates (PC1), developmental stage (PC2), and the differentiation of bipotential NMPs towards either fate (PC3) (**Fig. 2c-f; Supplementary Table 6**). Assuming that PC3 tracks the progression of differentiation consistently between neuroectodermal and mesodermal fates, the data suggest that the state of being *T* (*Brachyury*)+, *Meis1*- (**Fig. 2f-g**) may be a better indicator of bi-potency than being *T*+, *Sox2*+[159], consistent with two recent *in*

vivo studies of the genetic dependencies of NMPs[160, 161]. In addition to *T*, other markers of NMPs that were strongly correlated with the undifferentiated state include *Cyp26a1* and *Wnt3a*[162].

We observe striking contrasts in gene expression between NMPs derived from earlier (0-12 somites) vs. later (14-34 somites) embryos (**Fig. 2c-g**). Although we initially worried that this was a batch effect (as these embryos were processed in separate experiments), the observation is consistent with a study in which NMPs obtained from microdissected E8.5 and E9.5 embryos exhibited strong differences[163], with many of the same genes exhibiting sharp differential expression here (*Cdx1* (early); *Hoxa10* (late); **Fig. 2d; Supplementary Table 7**). Overall, our results are consistent with these early vs. late NMPs underlying the "trunk-to-tail" transition[164]. Given that early and late somite count embryos were processed in different experiments, we profiled an additional 12 mouse embryos ranging in somite counts from 8 to 21. This new sci-RNA-seq3 experiment validated and refined the estimated timing of this transition (**Supplementary Fig. 7**).

In addition to NMPs, another cell type marked by *T* is the notochord (cluster 2 in **Fig. 2a**). In the earliest embryos of this time course (0-12 somites), we observe two transcriptionally distinct subsets of cells within the notochord cluster, one more strongly *Noto*+ and the other more strongly *Shh*+ (**Fig. 2h-i**). As embryogenesis progresses, these subsets transition to a continuum, but the distinctions are preserved and reinforced, with the inferred derivatives of one early subpopulation marked by the expression of *Noto*, posterior *Hox* genes, and genes involved in Notch signaling, Wnt signaling and mesodermal differentiation (**Supplementary Fig. 8a**). Within this first early subpopulation, we also detect a highly distinct cluster of 60 cells that strongly express *Foxj1* and genes involved in motile ciliogenesis; these ciliated nodal cells are transient, peaking in terms of their contribution to the overall embryo at the 2-somite stage (**Fig. 2h-i; Supplementary Fig. 8b**).

In contrast, the inferred derivatives of second early subpopulation, marked by stronger Shh expression, is enriched for genes involved in neurogenesis and synaptogenesis, notably including *Sox10*, *Bmp3*, *Nrg1* and *Erbb4* (**Supplementary Fig. 8c**). One possibility is that the Noto+ subpopulation corresponds to the posterior notochord, which arises from the node, and the *Shh*++ subpopulation corresponds to the anterior mesendoderm (*i.e.* anterior head process and possibly the prechordal plate), which arises by condensation of dispersed mesenchyme and may contribute to forebrain patterning[165, 166, 167, 168]. Performing PCA (after excluding ciliated nodal cells), we find that these presumably A-P differences are the predominant source of transcriptional heterogeneity in the notochord cluster (PC1; 28.7% of variation; **Supplementary Table 8**).

Turning to the gut (cluster 3 in **Fig. 2a**), we once again observe subsets of transcriptionally distinct progenitors at early somite stages that give rise to a continuum at later somite stages (**Fig. 2j**). A major axis of this continuum reflects A-P patterning, with subsets corresponding to lung (*Nkx2-1*+), hepatocytes (*Afp*+, *Hhex*+), pancreas (*Nkx2-2*+), foregut (*Sox2*+, *Gata4*+), midgut (*Gata4*+, *Onecut2*+) and hindgut (*Hoxc8*+, *Cdx2*+) progenitors (**Fig. 2k**; PC1; 19.6% of variation; **Supplementary Table 9**). As *T* expression is classically associated with the notochord and posterior mesoderm in the mouse literature, we were initially surprised to see strong *T* expression in the inferred posterior hindgut, coincident with the expression of posterior *Hox* genes (**Supplementary Fig. 8d**). To our knowledge, this expression pattern was only recently documented[169], and is consistent with the ancestral role of *T* in the closing of the blastopore[170] as well as hindgut defects in *Drosophila brachyenteron* and *Caenorhabditis elegans* mab-9 mutants[171, 172].

Finally, we sought to explore whether there are overlaps between genes associated with spatial patterning (A-P axis) or developmental progression (somite counts) across germ layers. In comparing genes whose expression patterns are highly correlated with the inferred A-P axis between notochord (PC1; n=591) and gut (PC1; n=502), we observe overlap and

directional concordance (198 overlapping genes, 86% of which are consistently associated with the inferred anterior or posterior aspect of the notochord and gut; p<1e-28, χ2-test; **Fig. 2l; Supplementary Table 10**). Concordant, posterior-associated genes are highly enriched for genes involved in Wnt signaling as well as posterior *Hox* genes. One model to explain these overlaps between germ layers is that they are residual to the common origin of anterior mesendodermal derivatives (anterior head process, prechordal plate, anterior endoderm) from the early & mid-gastrula organizers vs. posterior mesendodermal derivatives (notochord and posterior endoderm) from the node[165]. Alternatively, these overlaps could be explained by physically coincident progenitors of these germ layers being exposed to similar patterns of Wnt signaling (*e.g.* strength, timing).

A second overlap between germ layers involves genes highly correlated with early vs. late somite counts in NMPs (n=257) vs. the gut (PC2; n=502). Once again, we observe overlap and directional concordance (82 overlapping genes, 70 (85%) of which are consistently associated with early or late somite counts; p<1e-15, χ2-test) (**Fig. 2m; Supplementary Table 11**). Given that early and late somite count embryos were processed in different experiments, we examined the additional 12 mouse embryos ranging in somite counts from 8 to 21. In this replication dataset, 77% of the initially overlapping, concordant genes replicated in terms of directionality-of-change between early (<13) vs. late (>13) somite count embryos in both NMPs and the gut (54/70; expected 25%; **Supplementary Fig. 7**). Genes reproducibly associated with early somite counts in both NMPs and the gut were strongly enriched for *Myc* targets, and also included *Lin28a*, a deeply conserved regulator of developmental timing[173, 174] (**Fig. 2m; Supplementary Table 11**). Other genes such as *Npm1* and *Hsp90* isoforms are plausibly associated with batch effects. However, analysis of a module of genes correlated with *Npm1* found it to be declining with developmental time across the entire time series, rather than correlated with batch variables (**Supplementary Fig. 9**).

3.6 Diversification of the intermediate and lateral plate mesoderm

The mesoderm, which includes axial, paraxial, intermediate and lateral plate components, is a complex germ layer, with the ontogenesis of some derivatives (*e.g.* heart, blood) much better understood than others (*e.g.* the mesenchyme that plays a major role in patterning each organ). In the previous section, we considered aspects of the axial (notochord) and paraxial (somite-forming) mesoderm. In this section, we focus on the transition from intermediate mesoderm to the kidney, as well as on the transition from the lateral plate mesoderm (LPM) to organ-specific mesenchyme.

What is the continuum of transcriptional states that span the transition from intermediate mesoderm to a functional nephron? When we generate a new embedding from relevant cell types, we observe two major trajectories. The first corresponds to the maturation of the posterior intermediate mesoderm to metanephric mesenchyme to nephron progenitors to renal tubule; while the second corresponds to the maturation of the anterior intermediate mesoderm to ureteric bud to collecting duct (**Fig. 3a-c**). The posterior and anterior trajectories in late gastrulation are marked by *Gdnf* and *Ret* expression, respectively, critical genes for normal kidney development[175, 176]. These trajectories then progress to the metanephric mesenchyme and ureteric bud, respectively, around E10.25, and then to specific functional components of the nephron (**Fig. 3a-c; Supplementary Fig. 10a-c**). Even as their transcriptomes mature with time, the metanephric mesenchyme and ureteric bud persist through P0 (clusters 6 & 3 in **Fig. 3a-b**; **Supplementary Fig. 10b**), presumably providing an ongoing source of progenitors for nephrogenesis, which continues for a few days after birth[177]. The apparent bifurcation of the proximal tubule corresponds to major differences in the transcriptional state of cells from embryos obtained before birth (E18.75 or earlier) vs. after birth (P0) (cluster 9 in **Fig. 3a-b; Supplementary Fig. 10d**). We return to this observation in the final section of the manuscript.

In the anterior intermediate mesoderm, both tip and stalk cells are identified within the ureteric bud, the former of which gives rise to the collecting duct, and the latter to the ureter[178, 179] (**Supplementary Fig. 10e**). Of note, we observe "convergence" of the posterior and anterior trajectories in collecting duct intercalated cells (cluster 4 in **Fig. 3a-b**). More detailed investigation suggests that the posterior intermediate mesoderm may also contribute to the collecting duct, consistent with lineage tracing experiments demonstrating the dual origin of intercalated cell types from the distal nephron and ureteric lineages[180] (**Fig. 3d-e; Supplementary Fig. 10f**).

The LPM is arguably much more complex than the axial, paraxial and intermediate mesoderms, giving rise to a remarkable diversity of cell types[181]. Although some LPM derivatives are intensely studied (*e.g.* heart), others remain poorly understood, in particular the mesoderm that lines the body wall and internal organs. Not only will this aspect of the LPM go on to give rise to important cell types and structures (*e.g.* fibroblasts, smooth muscle, mesothelium, pericardium, adrenal cortex, genital ridge, *etc.*), its interactions with other germ layers play a key role in organogenesis, *e.g.* reciprocal signaling between the mesoderm and endoderm serves as the basis of foregut patterning[182, 183].

To annotate derivatives of the splanchnic mesoderm in particular (*i.e.* the visceral layer of the LPM), we leveraged published spatial transcriptome data from E9.5-E16.5 mice to impute spatial coordinates for cells in our data[32, 37]. Through a combination of spatial inference and marker gene analysis, we were able to assign annotations to 22 subtypes of the LPM & intermediate mesoderm major cell type (**Fig. 3f-g; Supplementary Fig. 11; Supplementary Table 12**). Many of these assignments were supported by publicly available in situ hybridization images (**Supplementary Fig. 12**). For example, we can define subsets of the splanchnic mesoderm that lines specific organs, including the heart (proepicardium), brain (meninges), lung, liver, foregut and gut, as well as organ-specific smooth muscle (airway vs. gastrointestinal vs. vascular) (**Fig. 3g**). We can also distinguish

two subpopulations of LPM-derivatives mapping to the kidney, one to the cortex and the other more heterogeneously distributed within the renal mesenchyme, which we believe correspond to renal pericytes & mesangial cells and renal stromal cells, respectively. Although both subpopulations express Foxd1, supporting their assignment to the kidney, focused analyses are consistent with their having distinct origins (**Supplementary Fig. 13**). However, lineage tracing experiments would be necessary to test this hypothesis. Of note, renal stromal cells exhibited gene expression heterogeneity along what may be the cortical-medullary spatial axis, of genes including *Foxd1* (cortical), *Netrin-1* (cortical) and *Zeb2* (medullary) (**Supplementary Fig. 14**).

The high resolution of our time course during early somitogenesis enables us to narrow the temporal windows during which many subtypes of organ-specific mesenchyme are specified (**Supplementary Fig. 15a**). We also applied a mutual nearest neighbors (MNN)-based approach to identify putative precursor cells of each subtype. We performed this analysis separately for subtypes first detected earlier (5-20 somites; **Supplementary Fig. 15b-d**) vs. later (25-34 somites; **Supplementary Fig. 15e-g**) in development. For example, we can identify distinct subsets of splanchnic mesoderm that are most highly related to foregut mesenchyme, hepatic mesenchyme and proepicardium, and may correspond to the 'territories' in which these organ-specific mesenchyme are induced (**Supplementary Fig. 15b-d**). These annotations may provide a starting point for deeper dives into each mesenchymal subset and how it patterns (and is reciprocally patterned by) the organ with which it interfaces, as has recently been done for the foregut[182]. For example, the hepatic and foregut mesenchyme are sharply distinguished from one another, as well as from the splanchnic territories from which they arise, by many genes including expected TFs (*e.g. Gata4* and *Barx1*, respectively[184, 185]). However, the inferred splanchnic territories of origin (labels 15 vs. 14 in Supplementary **Fig. 15c**) are also clearly distinct from one another, with hepatic mesenchymal progenitors expressing a program of epithelial-mesenchymal transition (EMT) and foregut mesenchymal progenitors expressing multiple guidance cue programs

(*e.g.* semaphorins, ephrins, SLITs, netrins) (**Supplementary Table 13**).

3.7 The timing and trajectories of retinal diversification

In the next two sections, we turn from the mesoderm to the neuroectoderm, first considering the retina and then neuronal diversification more broadly. In mouse development, neural epithelium arises as early as E9, giving rise to ciliary and pigment epithelium as well as multipotent retinal progenitor cells (RPCs) by E10. Through the remainder of fetal development and continuing postnatally, RPCs give rise to seven major types of retinal neurons in a conserved order[186]. Here we sought to leverage the depth and temporal resolution of these data to more precisely define the developmental intervals and rates at which retinal cell types emerge, proliferate and diversify.

We re-embedded and re-annotated 160,834 cells with relevant preliminary annotations across all timepoints (**Fig. 4a; Supplementary Fig. 16a-b**). We observe that eye field is already detectable in our earliest embryo (early head fold stage; 0 somite embryo in E8.5 bin; *Pax2*+, n = 782 cells), diversifying towards retinal progenitors (as early as E9.75) and retinal pigment epithelium (RPE) (as early as E10), as well as a third branch that appears as early as E9.5, sharply downregulates *Rax*, and ceases proliferating by E14.5, likely corresponding to the optic stalk (**Supplementary Fig. 16c-d**). This branch is undetectable in later time points, but pathway analysis suggests this is due to terminal differentiation in the context of a rapidly growing embryo, rather than apoptosis. Among retinal neurons, differentiation towards retinal ganglion cells (RGCs) begins as early as E11.75, and towards cone photoreceptors as early as E13. As development progresses, we observe a succession of retinal neuron types appearing in the expected order (**Fig. 4b-c; Supplementary Fig. 16e**), except for Müller glia, which emerge postnatally[187]. Among RPCs, the succession of sampled timepoints fills out a continuum of transcriptional states associated with diversi-

fication towards most major retinal neuron types (**Fig. 4a-b**)[188]. In contrast, the ciliary marginal zone (CMZ), identified as early as E11.25, remains most similar to a rapidly expanding pool of naive retinal progenitors. Strikingly, the CMZ appears to give rise to a second wave of pigment epithelium in the perinatal period, entirely separated in terms of both its transcriptional trajectory and timeframe from the branch leading to RPE, likely corresponding to the iris pigment epithelium (IPE; **Fig. 4a; Supplementary Fig. 16f-g**).

Reanalyzing RGCs, we identify 15 clearly distinguishable subtypes, mainly diversifying in late gestation and well-defined by specific combinations of TFs (**Fig. 4d-e**). This extent of detected RGC diversity is on par with expectation for P0[189], suggesting that the improved performance of sci-RNA-seq3 has substantially improved its ability to discriminate neuronal subtypes.

3.8 The emergence of neuronal subtypes from the patterned neuroectoderm

The neuroectoderm and its derivatives comprise about 40% of cells profiled here (4.9M nuclei; **Fig. 1e-f**), of which the eye field and its derivatives are only about 3%. In this section, we describe the broader outlines of early neurogenesis in the post-gastrulation embryo, with deeper dives into the emergence of specific neuronal trajectories from the patterned neuroectoderm, as well as the transcriptional heterogeneity of spinal interneurons.

In the earliest embryos of this series (0-12 somite stage), we previously defined a continuum of cell states that correlated with anatomical patterning of the "pre-neurogenesis" neuroectoderm[152]. Performing a similar analysis here that extends through early organogenesis (E8-E13), we observe clusters corresponding to territories that will give rise to the major regions of the mammalian brain (**Fig. 5a; Supplementary Fig. 17**). As post-gastrulation development further unfolds, we observe numerous, distinct neurogenic trajectories arising

from these territories (**Fig. 5b-c**; note that patterned neuroectoderm, direct neurogenesis and indirect neurogenesis largely correspond to the 'neuroectoderm & glia', 'CNS neurons' and 'intermediate neuronal progenitors' major cell cluster annotations, respectively)[190].

Beginning as early as E8.75 (or more precisely, the 16 somite stage), most neuronal diversity in the prenatal mouse embryo is derived from direct neurogenesis (**Fig. 5d**), including motor neurons, cerebellar Purkinje cells, Cajal-Retzius cells, thalamic neuronal precursors and many other neuronal subtypes (see 'CNS neurons' sub-panel of **Supplementary Fig. 6** for full list). Indirect neurogenesis[190] has a slower start, with intermediate neuronal progenitors (*Eomes*+, *Pax6*+) first detected at E10.25, later giving rise to deep layer neurons, upper layer neurons, subplate neurons, and cortical interneurons (**Fig. 5d; Supplementary Fig. 18a-b**). Although many neuronal subtypes deriving from direct neurogenesis are easily distinguished, the majority of these cells (55%) could initially only be coarsely annotated as glutamatergic/GABAergic neurons or dorsal/ventral spinal cord progenitors (collectively 1.1M cells). To leverage the greater heterogeneity evident as these trajectories "launch" from the patterned neuroectoderm, we reanalyzed the subset of these cells derived from early embryos (stages earlier than E13), which greatly facilitated their more refined annotation (**Fig. 5e; Supplementary Fig. 18c; Supplementary Table 12**), while also highlighting the bases for components of this heterogeneity (*e.g.* anterior vs. posterior; excitatory vs. inhibitory; **Supplementary Fig. 18d**).

Among these more refined annotations are 11 subtypes of spinal interneurons. These assignments were based on subtype-specific TFs in these data (**Fig. 5f; Supplementary Table 15**) vs. the literature[191]. To systematically investigate transcriptional heterogeneity among spinal interneuron subtypes, we first applied the same PCA-based approach that we previously applied to NMPs (**Fig. 2e-f**). For differentiating spinal interneurons, PC1 and PC2 (together nearly 40% of variation) appear to correspond to neuronal differentiation (**Fig. 5g-h; Supplementary Table 16**), PC3 to glutamatergic vs. GABAergic identity,

and PC4 to dorsal vs. ventral identity (**Fig. 5h**).

We next sought to infer the progenitor cells in the patterned neuroectoderm from which various neuronal and non-neuronal cell types derive. To this end, we took pre-E13 cells annotated as intermediate neuronal progenitors, astrocytes, choroid plexus, or any of the derivatives of direct neurogenesis, and co-embedded them with cells of the patterned neuroectoderm (**Fig. 5a**). Next, for each derivative cell type in this co-embedding, we selected 500 cells from the earliest stage embryos and identified their mutual nearest neighbor (MNN) cells in the patterned neuroectoderm. Finally, we located these inferred progenitors in our original embedding of the pre-E13 patterned neuroectoderm (**Fig. 5i-j**).

The resulting distribution of inferred progenitors in the patterned neuroectoderm is considerably more granular than our annotations of anatomical territories (**Fig. 5j** vs. **Fig. 5a**). The inferred progenitors of the choroid plexus are overwhelmingly in the anterior roof plate (91%), with a minor subset in the dorsal diencephalon (5%), although this balance is likely impacted by the temporal window in which this analysis is focused (E8-E13)[192]. Inferred astrocyte progenitors exhibit a more complex distribution, with inferred VA2 progenitors primarily assigned to the spinal cord/r7/r8 (83%) and hindbrain (16%), and inferred VA3 progenitors to the spinal cord/r7/r8 (57%) and floorplate & p3 domain (32%)[193] (**Supplementary Fig. 19**). VA1 astrocytes appear to arise later than VA2 and VA3 astrocytes, and were not present in sufficient numbers for their progenitors to be inferred here.

The inferred progenitors of neuronal subtypes also largely fall within the expected territories, with additional sub-structure within those. For example, the inferred progenitors of dorsal and ventral spinal interneurons cluster distinctly (**Fig. 5j**). Of note, the progenitors of three neuronal subtypes (cerebellar Purkinje neurons, precerebellar neurons, spinal dI6 interneurons) were not clearly mappable. To address the possibility that this is because

they share progenitors with other neuronal subtypes, we repeated the MNN analysis in an iterative fashion. Although the origins of precerebellar neurons and spinal dI6 interneurons remain ambiguous, this analysis suggests that cerebellar Purkinje neurons and dl2 spinal interneurons may have transcriptionally similar progenitors (**Supplementary Fig. 20a**).

How do neuronal subtypes' identities arise, and how are they maintained[194]? As described above, we identified TFs that were specific to each of the 11 spinal interneuron subtypes (median 53 TFs per subtype; top 3: **Fig. 5f**; full list: **Supplementary Table 15**). These subtype-specific TFs exhibit diverse temporal dynamics (**Supplementary Fig. 20b**). Can we pinpoint which of these TFs might be involved in the initial specification of each subtype? Focusing on the dorsal spinal interneurons for which we successfully inferred progenitors (*dl1-dl5*), we identified TFs that are specific to the inferred progenitors of that subtype in the patterned neuroectoderm, relative to the inferred progenitors of other dorsal spinal interneurons (**Supplementary Fig. 20c**, top), most of which were bHLH or homeodomain TFs[195]. However, their expression was not maintained (**Supplementary Fig. 20c**, bottom), in line with the complex temporal dynamics of other subtype-specific TFs expressed after neuronal specification (**Supplementary Fig. 20b**)

Finally, we sought to systematically delineate the timing at which each of these neuronal subtypes initiates differentiation (**Supplementary Fig. 20d**). This analysis suggests that the differentiation of each neuronal subtype from the patterned neuroectoderm is modestly asynchronous, but also clearly subtype-specific. For example, about 95% of inferred progenitors of *dl2* spinal interneurons are from 20 somite to E11 stage embryos, while 95% of *dl4* spinal interneurons inferred progenitors are from 27 somite to E11.75 stage embryos.

Taken together, these analyses are consistent with a model articulated by Sagner & Briscoe in which both spatial and temporal factors contribute to the specification of neuronal subtypes as they emerge from the patterned neuroectoderm[194]. Furthermore, they

highlight the complexity of this process not only at the initiation of each neuronal subtype, but also over the course of their early maturation, *e.g.* at 6-hr resolution we can observe individual spinal interneuron subtypes expressing a dynamic succession of developmentally potent TFs (**Supplementary Fig. 20b**).

3.9 A rooted tree of cell type relationships spanning E0 to P0

A primary objective of developmental biology is to delineate the lineage relationships among cell types. Transcriptional profiles of single cells do not explicitly contain lineage information. However, under the assumption that the transcriptional states of closely related cell types are more similar than those of distantly related cell types, one can envision a developmental tree based solely on scRNA-seq data[57]. Indeed, we and others have reconstructed scRNA-seq-based trees for portions of worm, fly, fish, frog and mouse development[5, 6, 7, 8, 9, 15, 14, 13]. Returning to the analogy of a timelapse movie, the correspondence between the inferred and actual lineage relationships among cell types is anticipated to be a function of the quality and quantity of the snapshots taken (*e.g.* frame-rate of embryos sampled, number of cells sampled per time point, depth of sampling of each cell). As such, particularly for organogenesis & fetal development, the temporal resolution, breadth and depth of data reported here creates an opportunity to refine and extend a tree that relates cell types throughout mouse development.

In this section, we summarize our efforts to construct a rooted tree of cell types that spans mouse development from zygote to pup, *i.e.* E0 to P0, based on four published datasets[55, 196, 54, 13] (110,000 cells spanning E0 to E8.5) and the main dataset described here (11.4M cells spanning E8 to P0) (**Supplementary Table 17**). As with our initial attempt at reconstructing such a tree for mouse development[152], a major challenge is that these datasets are based on different scRNA-seq profiling technologies. Further challenges include the fact that

cells' transcriptional states are only loosely synchronized with developmental time[5, 6, 7, 8, 9, 15, 14, 13], the multiple scenarios by which cell state manifolds may be misleading[57], and finally, the sheer complexity of this mammalian organism.

To address these challenges, we adopted a heuristic approach, in which the tree reconstruction process was primarily data-driven, but with some degree of manual curation applied. First, based on data source, developmental window and cell type annotations, we split cells into fourteen subsystems which could be separately analyzed and subsequently integrated. The first two subsystems correspond to the pre-gastrulation and gastrulation phases of development and are based on the external datasets[55, 196, 54, 13]. The remaining twelve subsystems derive from the data reported here, and collectively encompass organogenesis & fetal development (**Supplementary Tables 17-18**).

Second, dimensionality reduction was performed separately on cells from each of the fourteen subsystems. Manual reexamination of each subsystem led to some corrections or refinements of cell type annotations, ultimately resulting in 283 annotated cell type nodes, some with only a handful of cells (*e.g.* 60 ciliated nodal cells) and others with vastly more (*e.g.* 650,000 fibroblasts) (**Supplementary Table 19-20**). Of note, each of these annotated cell type nodes derives from one data source, such that there are some redundant annotations that facilitate "bridging" between datasets (**Supplementary Fig. 21**). In contrast with our previous strategy in which nodes were stage-specific[152], each cell type node here is temporally asynchronous, and of course may also contain other kinds of heterogeneity (*e.g.* spatial, differentiation, cell cycle, *etc.*).

Third, we sought to draw edges between nodes (**Fig. 6a-f**). Within each subsystem, we identified pairs of cells that were mutual nearest neighbors (MNN) in 30-dimensional PCA space (k = 10 neighbors for pre-gastrulation and gastrulation subsystems, k = 15 for organogenesis & fetal development subsystems). Although the overwhelming majority of MNNs

occurred within cell type nodes, some MNNs spanned nodes and are presumably enriched for bona fide cell type transitions. To approach this systematically, we calculated the total number of MNNs that spanned each possible pair of cell type nodes within a given subsystem, normalized by the total number of possible MNNs between those nodes, and ranked all possible intra-subsystem edges based on this metric (**Supplementary Table 21**). Of note, due to its complexity, this was done in two stages for the "Brain & spinal cord" subsystem, first applying the heuristic to the subset of cell types corresponding to the patterned neuroectoderm, and then again to identify edges between the patterned neuroectoderm and its derivatives (*i.e.* neurons, glial cells, *etc.*).

Fourth, we manually reviewed the ranked list of 1,155 candidate edges for biological plausibility (those with a normalized MNN score > 1; **Supplementary Fig. 21a**), resulting in 452 edges which we manually annotated as more likely to correspond to either "developmental progression" or "spatial continuity" (**Supplementary Table 22**). Where nodes were connected to more than one other node, distinct subsets of cells were generally involved in each edge (**Fig. 6a-b; Fig. 6d-e**), and internode MNN pairs exhibited temporal coincidence (**Fig. 6c,f**). As only a handful of cells were profiled in the pre-gastrulation subsystem, those edges were added manually.

Finally, to bridge subsystems, we performed batch correction and co-embedding of selected timepoints from either the pre-gastrulation and gastrulation datasets, or the gastrulation and organogenesis & fetal development datasets, to identify equivalent cell type nodes, resulting in a third category of "dataset equivalence" edges (**Supplementary Fig. 21b-e**). Most of the 12 organogenesis & fetal development subsystems originate in cell type nodes for which equivalent nodes are already present at gastrulation (**Supplementary Fig. 21e**). The exceptions, presumably due to undersampling of this transition, were the "blood" and "PNS neuron" subsystems, for which we manually added edges to connect them with biologically plausible pseudo-ancestors. Altogether, we added 55 inter-subsystem edges.

The resulting developmental cell type tree, spanning E0 to P0 or zygote-to-pup, can be represented as a rooted, directed graph (**Fig. 6g**), in which dataset equivalence edges are oriented forward in time, developmental progression edges are manually oriented, and spatial continuity edges are left bidirectional. In practice, a small number of nodes in the tree have more than one parent, so the "tree" is formally a rooted, directed graph.

3.10 Systematic nomination of transcription factors and other genes for cell type specification

We previously annotated the edges of a more limited cell type tree of mouse development by identifying TFs exhibiting sharp changes in gene expression at the developmental timepoint where a given cell type first appears, through comparisons of all cells of a new type to all cells of its inferred pseudoancestor as well as to any sister nodes[152]. Here, in the course of similarly annotating the zygote-to-pup cell type tree shown in **Fig. 6g**, we sought to take a more nuanced approach.

In particular, we stratified each cell type transition into four phases (**Supplementary Fig. 22a**). Given a directional edge between two nodes, AtoB, we identified the subset of cells within each node that were either "inter-node" MNNs of the other cell type (groups 2 & 3 in **Supplementary Fig. 22b**) or "intra-node" MNNs of those cells (groups 1 & 4 in **Supplementary Fig. 22b**). If AtoB, this approach effectively models the transition as group 1 to 2 to 3 to 4. Next, we identified differentially expressed TFs (DETFs) and genes (DEGs) across each portion of the modeled transition, *i.e.* early (1 to 2), inter-node (2 to 3), and late (3 to 4). In contrast with our previous heuristic[152], this strategy highlights differences between cells that are most proximate to the cell type transition itself. Moreover, DETFs and DEGs identified in the early (1 to 2) and late (3 to 4) phases correspond to changes within node A or B, respectively, rather than between nodes A and B, which may

facilitate the identification of early changes in gene expression, *i.e.* those that precede the cell type transition itself.

We applied this heuristic to 436 edges of the rooted tree shown in **Fig. 6g**, excluding only dataset equivalence edges and the pre-gastrulation subsystem. Of note, the directionality of many of these edges was not immediately obvious (*i.e.* those annotated as "spatial continuity" edges in **Supplementary Table 22**). In these cases the orientation of the "early" and "late" phases is arbitrary. For edges with a relatively small number of MNN pairs, we expanded each group to at least 200 cells by iteratively including their MNNs within the same cell type, to increase statistical power. Across all 436 edges, this heuristic resulted in the nomination of ranked lists of median 28 (IQR 12-51) DETFs and median 171 (IQR 76-389) DEGs per edge (mean), 5% (DETFs) and 7% (DEGs) of which were exclusively nominated in either the early or late phase of the transition, respectively. The most significant DETFs and DEGs for each edge are provided in **Supplementary Tables 23 and 24**, respectively. Most TFs and genes are only nominated in the context of one or a few edges, but there are a number of outliers that presumably have more general roles in cell type specification (**Supplementary Fig. 22c-d**).

On a cursory review, there are many instances in which the top-ranked upregulated DETF for the early phase of the transition corresponds to a well-established, early regulator of that cell type (*e.g. MITF* for melanocytes[197]; *Ebf1* & *Pax5* for B-cell progenitors[198]; *Lef1* for B-cells[199]; *Zfpm1* for megakaryocyte-erythroid progenitors[200]), but also many nominations of potentially novel regulatory relationships that may warrant further investigation (*e.g. Ltf* for monocytic myeloid-derived suppressor cells; *Tcf7l2* for Kupffer cells, *Esrrg* for dorsal telencephalon-derived choroid plexus; *Zfp536* for myelinating Schwann cells; *Rreb1* for adipocyte progenitors, *etc.*) (**Supplementary Table 23**).

Digging further into a well-studied transition, *Sox17* is the sole upregulated DETF during

the early phase of the transition from anterior primitive streak to definitive endoderm (group 1to2), while a broader set of TFs (*Elf3*, *Sall4*, *Hesx1*, *Lin28a*, *Hmga1*, *Ovol2*, but notably not *Sox17*) are upregulated during the transition itself (group 2 to 3) (**Supplementary Table 23**). Non-TF DEGs specific to the early phase of the transition include *Cer1* (encoding Cerberus, a well-established signaling molecule and marker of definitive endoderm)[201], *Slc25a4* (encoding an adenine nucleotide translocator specific to the inner mitochondrial membrane)[202], and *Slc2a3* (encoding Glut3, a glucose transporter specific to cell types with high energy needs)[203] (**Supplementary Table 24**). To examine this further, we extracted all cells participating in groups 1-4 of this transition and subjected them to conventional pseudotime analysis[15]. In brief, this analysis supported the upregulation of *Sox17* as preceding that of other nominated key TFs, and further highlighted *Cer1* as the only non-TF DEG with similar kinetics to *Sox17* (**Supplementary Fig. 22e-f**).

A more complex example involves hematopoietic stem cells (*Cd34+*), which in the developmental graph are the node-of-origin of a dozen cell types (**Supplementary Fig. 22g**). An important point is that although we treat hematopoietic stem cells (HSC) (*Cd34+*) as a single node for the purposes of the zygote-to-pup developmental graph, the cells assigned to this node are quite heterogeneous. For example, the cells participating in the MNNs that support the edges to various lymphoid, myeloid and erythroid derivatives are distinct subpopulations of the HSC node (**Supplementary Fig. 22h-i**), and these subpopulations differentially express TFs nominated as early regulators, *e.g. Ebf1* for B cells[198] and *Id2*[204] and *Nfatc2* for conventional dendritic cells (**Supplementary Fig. 22j**). Establishing a level of resolution for a developmental tree that hits the right balance between the complexity of mammalian development and the desire for interpretability is an ongoing tension.

3.11 Rapid shifts in transcriptional state occur in a restricted subset of cell types upon birth

In the course of our analyses of these data, we anecdotally noted that for certain cell types, cells derived from P0 pups appeared very well separated from their fetal pseudoancestors. This sharply contrasts with the remainder of the dataset, in which cells of a given cell type were generally well mixed between adjacent timepoints. The proximal tubule is one example of this phenomenon, discussed briefly above (cluster 9 in **Fig. 3a-b**; **Supplementary Fig. 10d**). However, a similar pattern was also noted for hepatocytes, adipocytes, and various cell types of the lungs and airways (**Fig. 7a**). As we worried that this could be due to batch effects or the well-documented pitfalls of over-interpreting UMAPs[22], we conducted a time point correlation analysis. In brief, for each cell type, we performed dimensionality reduction (30 dimensions) on cells of that type from multiple, late-gestational timepoints (E16 to P0). We then identified k-nearest neighbors (k-NN, k = 10) to ask whether neighboring cells were derived from the same or different timepoints. In this framing, a low proportion of neighbors from different timepoints corresponds to a relatively abrupt change in transcriptional state. In general, cells from P0 pups were closer neighbors to other P0 cells than was the case for all other timepoints (**Fig. 7b**), which might be explained in part by a longer temporal interval between E18.75 and P0 than 6 hours. However, for many cell types, the difference was extreme, consistent with dramatic changes in a transcriptional state upon birth. These include not only the aforementioned cell types, but also various endothelial and blood lineages (**Fig. 7b**). In contrast, neuronal cell types did not exhibit such marked differences between animals collected shortly before vs. after birth (**Fig. 7b**).

To validate and further resolve periparturition transcriptional dynamics, we collected nine pups that were part of a single litter. After 3 pups were delivered vaginally, the remaining six pups were delivered by Cesarean section (C-section) and sacrificed either immediately (0 min;

2 pups), or 20, 40, 60 or 80 min after C-section (1 pup each) (**Fig. 7c; Supplementary Fig. 23a**). Nuclei from these nine pups, together with residual nuclei from three samples from the original experiments, were subjected to a new sci-RNA-seq3 experiment, which yielded nearly 1M additional single cell profiles (**Supplementary Fig. 23b; Supplementary Table 1-2**).

Focusing first on the six pups delivered by C-section and 24 major cell clusters for which we were reasonably powered, we applied timepoint correlation analysis, as described above except treating the C-section timepoint as a continuous variable. Consistent with our earlier results (**Fig. 7b**), the outliers are the hepatocyte, adipocyte and lung & airway major cell clusters (**Fig. 7d; Supplementary Fig. 23c-d**). This experiment replicates our initial finding and also narrows the window in which these abrupt changes are first evident to the first hour of extrauterine life (**Fig. 7e**).

Although we cannot fully rule out technical variables associated with sacrificing pups, we took care to minimize handling and stress prior to euthanasia and immediate snap-freezing, both for naturally and C-section delivered pups. Moreover, it is plausible that rapid changes in transcriptional programs might be physiologically necessary due to the profound differences between the placental and extrauterine environments. One can imagine why these might be triggered by birth rather than developmentally timed, because of uncertainty with respect to precisely when birth will occur. Finally, we speculate that the proper development of certain organs might require the continuation of a precisely timed program (*e.g.* the brain), whereas for other organs, survival depends on an extremely rapid change in their function in the immediate wake of birth (*e.g.* the lungs & airways) (**Fig. 7b**). In examining DEGs of rapidly changing cell types, either between E18.75 and P0 embryos or across the 20 minute time-series of pups delivered by C-section, we see clues in regards to at least some of the specific physiological functionalities that these rapid changes might be serving (**Supplementary Tables 25-26**).

For example, in hepatocytes, genes involved in gluconeogenesis are sharply upregulated, including *Ppargc1a*, which encodes Pgc-1a, which in the liver serves as a master regulator of hepatic gluconeogenesis[205], as well as *Pck1*, *G6Pc* and *Got1*, which encode key enzymes in this pathway (**Fig. 7f**). Aspects of these changes have previously been linked to changes in key nutritional hormones (rising glucagon, falling insulin) immediately after birth and presumably are necessary for maintaining normoglycemia in the wake of being abruptly cut off from maternal nutrients[206]. In brown adipocytes, we observe sharp upregulation of *Irf4*, a cold-induced master regulator of thermogenesis, and again of *Ppargc1a*, which in adipocytes plays a different role than in the liver, as Pgc-1a partners with *Irf4* to drive the expression of *Ucp1* and uncoupled respiration[207], presumably to maintain body temperature upon transition to the extrauterine environment[208] (**Fig. 7f**).

The exact amount of time that lapsed between the birth of the vaginally birthed pups and their harvesting was not precisely captured in the replication experiments. However, on co-embedding cells derived from vaginally birthed pups with those delivered by C-section for the three most relevant major cell clusters, timepoint correlation analysis suggested that they were harvested within an hour of their birth (**Supplementary Fig. 23e**). However, this assumes similar kinetics for these rapid transcriptional changes in C-section vs. vaginally delivered pups. On more detailed inspection, the patterns are considerably more complex, with certain clusters (most notably, a subset of hepatocytes) appearing specific to vaginally birthed pups (**Supplementary Fig. 23f; Supplementary Table 27**).

What might explain this? The transition from the placenta to extrauterine life, as experienced by the neonate (and perhaps even by specific organs within the neonate), is very different between C-section vs. vaginal delivery, and various studies have shown that human babies delivered by each route have differences in physiology and health outcomes[209]. It is possible that aspects of these postnatal phenotypic differences have their roots in how the massive, abrupt, cell type-specific changes documented here are influenced by mode of

delivery.

3.12 Discussion

We present a rich dataset in which we profiled the transcriptional states of 12.4 million nuclei sampled from 83 precisely staged embryos spanning late gastrulation (E8) to birth (P0), with 2-hr temporal resolution during somitogenesis, 6-hr resolution through to birth, and 20-min resolution during the immediately postpartum period. We note that despite its large scale, the entire project was driven by a small number of individuals. For example, the mouse embryos were staged by a single individual (I.W.), the vast majority of the sci-RNA-seq3 experiments were carried out by a single individual (B.M.) and the computational analyses were conducted by a single individual (C.Q.), with nearly all experiments and analyses completed within one year. We estimate that the direct costs of all reagents and labor invested in processing frozen embryos to sequencing libraries totaled to about $70,000, while the cost of sequencing totaled to about $300,000. Notably, our single dataset is equivalent to about 30% of the aggregated corpus of the Human Cell Atlas Data Portal as of March 2023.

Three broad concepts supported our ability to generate, analyze and integrate such a large dataset with a small team at a modest cost: First, multiplexing, which fundamentally underlies the exponential scalability of single cell combinatorial indexing as well as that of massively parallel DNA sequencing. Second, open science, as we have taken abundant advantage of many freely released software packages for single cell data analysis released by the community[141, 15, 51, 37]. Third, our focus on mouse development, an eminently reproducible process through which we could access all mammalian cell types (or their predecessors) within a series of physically compact samples.

Our main goal in initiating this study was not to learn a specific piece of biology, but

rather to lay a foundation for a comprehensive understanding of the development of a mammalian gastrula into a free-living pup (E8 to P0). By annotating hundreds of cell types and conducting deeper dives into selected developmental systems, we have sought to illustrate the value of this foundation. Although the dataset is a rich source of hypotheses (*e.g.* the nomination of specific TFs as *in vivo* regulators of myriad cell types), the largest surprise to us was the discovery of rapid changes in transcriptional state in a restricted subset of cell types in the minutes to hours following birth. There is immense evolutionary pressure on the transition from placental to extrauterine life, which is arguably as fraught a moment as gastrulation in terms of physiological peril[210]. In specific cases, the genes sharply upregulated in specific cell types can be matched to specific adaptations (*e.g.* gluconeogenesis in hepatocytes, uncoupled respiration and thermogenesis in brown adipocytes). On the other hand, dozens of additional genes are sharply upregulated or downregulated in hepatic, adipose, and lung & airway tissues immediately after birth (**Supplementary Tables 25-26**). For these and many other tissues and cell types in which rapid periparturitional transcriptional changes are also documented (**Fig. 7b,d**), the adaptive functions served, not to mention the mechanisms underlying their rapid induction as well as differences shaped by the circumstances of delivery, warrant further exploration.

A limitation of our sampling strategy is that we only profiled a single embryo for most timepoints, such that we are unable to conduct a systematic analysis of interindividual variation at any given timepoint. We do observe hints of such variation, *e.g.* multiple different types of renal cells were not detected at E12.25, which may reflect aberrant renal development in that individual embryo (**Supplementary Fig. 10b**). However, such analyses may be better pursued through other datasets, *e.g.* our profiling of 101 embryos (of 26 genotypes) staged at E13.5, also by sci-RNA-seq3[211]. On a related point, although both sexes are represented in the dataset (as we alternated between adjacent timepoints), we have not yet delved into sex differences, and this remains one of many avenues of investigation for which we hope researchers in the field will find these data useful.

We recently proposed the concept of a "consensus ontogeny" of cell types, inclusive of both lineage histories and molecular states, as a potential structure for a "reference cell tree"[212]. The cell type tree constructed here, which spans mouse development, from a single cell zygote to the half-billion cells that make up a free-living pup (E0 to P0), represents a step in that direction. As with previous such steps by us and others[14, 152, 15, 13], both cell type annotations and the tree structure itself almost certainly contain errors and are subject to corrections and refinement. Furthermore, as additional methods for single cell molecular profiling (*e.g.* chromatin accessibility, spatial mapping, *etc.*) or organism-scale lineage tracing[3, 213, 136, 214] are applied to the mouse, we envision that a progressively improving consensus ontogeny will provide a framework for an increasingly rich, navigable roadmap of mouse development.

Because it touches all aspects of prenatal development of the embryo proper from gastrulation to birth, the dataset may also be useful in ways that we did not anticipate at the project's outset, *e.g.* as input for pre-training large language models of mammalian biology[215]. Finally, just as Sulston reconstructed both the embryonic and post-embryonic lineages of *C. elegans*[2, 216], we note that mouse development does not end at P0. We envision that the methods described here can also be applied to postnatal timepoints at the scale of the entire organism, with the long term goal of generating a timelapse of the entire life cycle of a mammal at single cell resolution, from a single cell zygote to a natural death (E0 to D0).

3.13 Supplementary Materials

Data reporting

For newly generated mouse embryo data, no statistical methods were used to predeter-

mine sample size. Embryo collection and sci-RNA-seq3 data generation were performed by different researchers in different locations.

Mouse embryos collection and staging

The details of collecting 12 mouse embryos around E8.5 with somites ranging from 0 to 12 were described in a previous publication[152]. Briefly, C57BL/6NJ (strain 005304) mice were obtained at The Jackson Laboratory and mice were maintained via standard husbandry procedures. Timed matings were set in the afternoon and plugs were checked the following morning. Noon of the day a plug was found was defined as embryonic day (E) E0.5. On the morning of E8.5, individual decidua were removed and placed in ice cold PBS during the harvest. Individual embryos were dissected free of extraembryonic membranes, imaged, and the number of somites present were noted prior to snap freezing in liquid nitrogen (**Supplementary Fig. 1**). A portion of yolk sac from each embryo was collected for sex based genotyping and samples were stored at -80C until further processing.

For the newly generated mouse embryo data (ranging from E8.75 to P0), we employed a combination of staging methodologies depending on gestational age of harvest (**Supplementary Fig. 2**). To maximize temporal coherence, resolution, and accuracy, we sought to stage individual embryos based on well-defined morphological criteria, rather than by gestational day alone. Embryos harvested between E8.0 - E10.0 were staged based upon the number of somites counted at the time of harvest and further characterized by morphological features (**Supplementary Fig. 1**). For E10.25-E14.75 embryos, developmental age was determined using the embryonic Mouse Ontogenetic Staging System (eMOSS, https://limbstaging.embl.es/), which leverages dynamic changes in hindlimb bud morphology and landmark-free based morphometry to estimate the absolute developmental stage of a sample[156, 157]. A modified staging tool, implemented in python, with increased performance on E14.0-E15.0 samples was used to confirm staging of samples within this window (documentation and python

scripts available at: https://github.com/marcomusy/welsh_embryo_stager). To distinguish samples staged via eMOSS, these samples are designated with "mE" for morphometric embryonic day (*e.g.* mE13.5, **Supplementary Fig. 2**). Due to the increased complexity of limb morphology at later stages automated staging beyond E15.0 is not possible. As a consequence, harvests for all remaining embryonic samples (E15.0-E18.75) were performed precisely at 00:00, 06:00, 12:00, and 18:00 on the targeted day. From close inspection of limbs in this sample set we defined additional dynamics related to digit morphogenesis that allowed further binning of samples collected on Days 15 and 16 (**Supplementary Fig. 2**). Therefore, amongst samples profiled in this study only the E17.0-E18.75 samples were staged solely by gestational age. Lastly, P0 samples were harvested from litters at noon of the day of birth (parturition for C57BL/6NJ occurs between E18.75 and E19.0).

Collecting mouse pups immediately after birth

Samples for analysis of periparturition transcriptional dynamics were collected from a plugged female that was monitored for signs of labor beginning at E18.75. Following the natural delivery of 3 pups the dam was euthanized and following removal from the uterus and extraembryonic membranes the remaining pups were either harvested immediately or placed in a warming chamber to monitor respiratory response and collected at 20 minute intervals. We collected nine new pups, to perform another sci-RNA-seq3 experiment. The first three pups are estimated to be between 1-2 hours old, although this was not precisely timed (samples 1-3 in **Fig. 7c; Supplementary Fig. 23a**). None of these pups had nursed at the time of harvest. The next two pups were taken by C-section, decapitated and snap frozen immediately; no breaths were taken (samples 4-5 in **Fig. 7c; Supplementary Fig. 23a**). The next four pups were taken by C-section and used for a "pink up" time course, harvesting one pup every 20 min (*i.e.* 20 mins, 40 mins, 60 mins, and 80 mins; samples 6-9 in **Fig. 7c; Supplementary Fig. 23a**). During this time, all pups remained very active and working to establish a breathing rhythm. Pup 6 had not fully pinked up at time of harvest,

but pups 7-9 had. Pups 8 and 9 had visible lungs in their chest cavities at 60 min. The last pup harvested at 80 min was fully pink with a reasonably stable breathing rhythm. No vocalization was heard from any pups during this collection. Of note, for additional quality control, we put nuclei from previously profiled E18.75 and P0 embryos into a small number of wells of the sci-RNA-seq3 experiment in which nuclei from this series were processed.

Generating data using an optimized version of sci-RNA-seq3

Together with E8.5 data which has been reported before[152], a total of 15 sci-RNA-seq3 experiments were performed on 75 individual mouse embryos. At least one sample was included for every 6 hour time interval from E8.0 to P0, and we also included a fine-scale sample set of embryos with distinct somite counts from 0-34 somites. Multiple samples were selected for a few timepoints (*e.g.* two samples for E13.0) to boost cell numbers. Meanwhile, we tried to ensure that both male and female mice roughly appear at adjacent timepoints (**Fig. 1d**). A detailed summary & images of individual embryos can be found in **Supplementary Fig. 1-2** and **Supplementary Table. 1**.

To generate the dataset, we used the optimized sci-RNA-seq3 protocol [58] as written, adjusting the volume and type of lysis buffer to the size of the embryos. Briefly, frozen embryos were pulverized on dry ice and cells were lysed with a phosphate-based, hypotonic lysis buffer containing magnesium chloride, Igepal, diethyl pyrocarbonate (DEPC) as an RNase inhibitor, and either sucrose or bovine serum albumin (BSA). Lysate was passed over a 20um filter, and the nuclei-containing flow-through was fixed with a mixture of methanol and dithiobis (succinimidyl propionate) (DSP). Nuclei were rehydrated and washed in a sucrose/PBS/triton/magnesium chloride buffer (SPBSTM), then counted and distributed into 96-well plates for reverse transcription with indexed oligo-dT primers.

Age-specific adaptations were as follows:

- E10-E13 embryos use 5mL BSA lysis buffer, E14 embryos use 10mL BSA lysis buffer, E15-E18 embryos use 20mL sucrose-based lysis buffer. Each of these samples were split over 48-96 wells for reverse transcription and the first round of indexing. A newborn P0 mouse requires 40mL of sucrose-based lysis buffer, and the lysate is divided into 4 fractions for filtration and fixing because of the amount of tissue involved. The two P0 mice were each processed as an individual experiment and were each split over 384 wells for reverse transcription.

- For the mouse samples E8.0-E9.75, we used the "Tiny Sci" adaptation of the optimized sci-RNA-seq3[58]. Frozen embryos were gently resuspended in 100ul lysis buffer to free the nuclei, then 400ul of DSP-methanol fixative was added. In the same tube, fixed nuclei are rehydrated, washed, then put directly into 8-32 wells for reverse transcription.

- After reverse transcription, nuclei were pooled, washed, and redistributed into fresh 96-well plates to attach a second index sequence by ligation. Then the nuclei were pooled again, washed and redistributed into the final plates. There, the nuclei undergo second-strand synthesis, extraction, tagmentation with Tn5 transposase and finally PCR to add the final indexes. The PCR products were pooled, size-selected, and then the library was sequenced on an Illumina NovaSeq. For some experiments, a second NovaSeq run was necessary to capture the extent of the library complexity, so we would add more sequencing reads until the PCR duplication rate met a threshold of 50% or the median UMI count per cell went over 2,500. The validation dataset (**Supplementary Fig. 7**) generated from 8-21 somites embryos was sequenced on an Illumina NextSeq.

Processing of sci-RNA-seq3 sequencing reads

Data from each individual sci-RNA-seq3 experiment was processed independently. For

each experiment, read alignment and gene count matrix generation was performed using the pipeline that we developed for sci-RNA-seq3[15] (https://github.com/JunyueC/sci-RNA-seq3_pipeline) with minor modifications: base calls were converted to fastq format using Illumina's bcl2fastq/v2.20 and demultiplexed based on PCR i5 and i7 barcodes using maximum likelihood demultiplexing package deML[138] with default settings. Downstream sequence processing and single cell expression matrix generation were similar to sci-RNA-seq[53] 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 (Levenshtein edit distance (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/v0.6.5 (https://github.com/FelixKrueger/TrimGalore) with default settings. Trimmed reads were mapped to the mouse reference genome (mm10) for mouse embryo nuclei using STAR/v2.6.1d[139] with default settings and gene annotations (GENCODE VM12 for mouse). 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, hairpin ligation adaptor index and read 2 end-coordinate (*i.e.* reads with UMI sequence less than 2 edit distance, 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). 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[140]. For multi-mapping reads (*i.e.* those mapping to multiple genes), the read were assigned to the gene for which the distance between the mapped location and the 3' end of that gene was smallest, except in cases where the read mapped to within 100 bp of the 3' end of more than one 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. After the single cell

gene count matrix was generated, cells with low quality (UMI $<$ 200 or detected genes $<$ 100 or unmatched_rate (proportion of reads not mapping to any exon or intron) $\geq$ 0.4) were filtered out. Each cell was assigned to its originating mouse embryo on the basis of the reverse transcription barcode.

Doublet removal

Here, we performed three steps with the aim of exhaustively detecting and removing potential doublets. Of note, all these analyses were performed separately on data from each experiment. First, we used Scrublet to detect doublets directly. In this step, we first randomly split the dataset into multiple subsets (six for most of the experiments) in order to reduce the time and memory requirements. We then applied the scrublet/v0.1 pipeline[27] 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 scores over 0.2 were annotated as detected doublets.

Second, we performed two rounds of clustering and used the doublet annotations to identify subclusters that are enriched in doublets. The clustering was performed based on Scanpy/v.1.6.0[141]. Briefly, gene counts mapping to sex chromosomes were removed, and genes with zero counts were filtered out. Each cell was normalized by the total UMI count per cell, and the top 3,000 genes with the highest variance were selected, followed by renormalizing the gene expression matrix. 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), followed by Louvain clustering with default parameters (resolution = 1). For the Louvain clustering, we first computed a neighborhood graph using a local neighborhood number of 50 using scanpy.pp.neighbors. We then clustered the cells into sub-groups using the Louvain algorithm implemented by the scanpy.tl.louvain function.

For each cell cluster, we applied the same strategies to identify subclusters, except that we set resolution = 3 for Louvain clustering. Subclusters with a detected doublet ratio (by Scrublet) over 15% were annotated as doublet-derived subclusters. We then removed cells which are either labeled as doublets by Scrublet or that were included in doublet-derived subclusters. Altogether, 2.7% to 16.8% of cells in each experiment were removed by this procedure.

We found that the above Scrublet and iterative clustering based approach has difficulty identifying doublets in clusters derived from rare cell types (*e.g.* clusters comprising less than 1% of the total cell population), so we applied a third step to further detect and remove doublets. This step uses a different strategy to cluster and subcluster the data, and then looks for subclusters whose differentially expressed genes differ from those of their associated clusters. This step consists of a series of ten substeps. 1) We reduced each cell's expression vector to retain only protein-coding genes, lincRNAs, and pseudogenes. 2) Genes expressed in fewer than 10 cells and cells in which fewer than 100 genes were detected were further filtered out. 3) The dimensionality of the data was reduced by PCA (50 components) first on the top 5,000 most highly dispersed genes and then with UMAP (max_components = 2, n_neighbors = 50, min_dist = 0.1, metric = 'cosine') using Monocle/3-alpha[15]. 4) Cell clusters were identified in UMAP 2D space using the Louvain algorithm implemented in Monocle/3-alpha (resolution = 1e-06). Cell partitions were detected using the partitionCells function implemented in Monocle/3-alpha. This function applies algorithms that automatically partition cells to learn disjoint or parallel trajectories based on concepts from "approximate graph abstraction"[23]. 5) We took the cell partitions identified by Monocle/3-alpha (cell clusters were used instead for three experiments that profiled embryos before E10), downsampled each partition to 2,500 cells, and computed differentially expressed genes across cell partitions with the top_markers function of Monocle/3 (reference_cells=1000). 6) We selected a gene set combining the top ten gene markers for each cell partition (filtering out genes with fraction_expressing < 0.1 and then ordering by pseudo_R2). 7) Cells from each main cell par-

tition were subjected to dimensionality reduction by PCA (10 components) on the selected set of top partition-specific gene markers. 8) Each cell partition was further reduced to 2D using UMAP (max_components = 2, n_neighbors = 50, min_dist = 0.1, metric = 'cosine'). 9) The cells within each partition were further subclustered using the Louvain algorithm implemented in Monocle/3-alpha (res = 1e-04 for most clustering analysis). 10) Subclusters that expressed low levels of the genes that were found to be differentially expressed in step 5, had high levels of markers specific to a different partition, and had relatively high doublet scores, were labeled as doublet-derived subclusters and removed from the analysis. On average, this procedure eliminated 3.4% of cells from each experiment (range 0.5-13.2%) of the cells in each experiment (**Supplementary Figs. 24-25**).

Cell clustering and cell-type annotations

For data from individual experiments, after removing the potential doublets detected by the above three steps, we further filtered out the potential low-quality cells by investigating the numbers of UMIs and the proportion of reads mapping to the exonic regions per cell (**Supplementary Fig. 3**). Then, we merged cells from individual experiments, to generate the final dataset, which included 15 sci-RNA-seq3 experiments with 21 runs of NovaSeq. When we took a first round of cell-embedding, we noticed that one mouse embryo at E14.5 had a grossly reduced proportion of neuronal cells. This particular sample had been divided during pulverization, and we suspect that large portions of the frozen embryo did not make it into the experiment. We removed cells from this E14.5 embryo, and we further filtered out cells from the whole dataset with doublet score (by Scrublet) > 0.15 (0.3% of the whole dataset), as well as cells with either the percentage of reads mapping to ribosomal chromosome (Ribo%) > 5 or the percentage of reads mapping to mitochondrial chromosome (Mito%) > 10 (0.1% of the whole dataset). Finally, 11,441,407 cells from 74 embryos were retained, of which the median UMI count per cell is 2,700 and median gene count detected per cell is 1,574. For the final dataset, the number of cells recovered by each embryo and the

basic quality information for cells from each sci-RNA-seq3 experiment was summarized in the **Supplementary Table. 1 & 2**. For sex separation and confirmation of embryos with or without sex genotyping, 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*).

We then applied Scanpy/v.1.6.0[141] to this final dataset, performing conventional single-cell RNA-seq data processing: 1) retaining protein-coding genes, lincRNA, and pseudogenes for each cell and removing gene counts mapping to sex chromosomes; 2) normalizing the UMI counts by the total count per cell followed by log-transformation; 3) selecting the 2,500 most highly variable genes and scaling the expression of each to zero mean and unit variance; 4) applying PCA and then using the top 30 PCs to calculate a neighborhood graph (n_neighbors = 50), followed by leiden clustering (resolution = 1); 4) performing UMAP visualization in 2D or 3D space (min.dist = 0.1). For cell clustering, we manually adjusted the resolution parameter towards modest overclustering, and then manually merged adjacent clusters if they had a limited number of differentially expressed genes (DEGs) relative to one another or if they both highly expressed the same literature-nominated marker genes. For each of the 26 major cell clusters identified by the global embedding, we further performed a sub-clustering with the similar strategies, except setting n_neighbors = 30 when calculating the neighbor graph and min_dist = 0.3 when performing the UMAP. Subsequently, we annotated individual cell clusters identified by the sub-clustering analysis using at least two literature-nominated marker genes per cell type label (**Supplementary Table 5**).

To be clear, we have hierarchically nominated three levels of cell-type annotations in the manuscript.

- In the global embedding involving all 11.4 M cells we identified 26 major cell clusters

(**Fig 1.e-f; Supplementary Table 4**).

- For individual major cell clusters, we performed sub-clustering, resulting in 190 cell types (**Supplementary Fig. 6; Supplementary Table 5**).

- For a handful of cell types, in specific parts of the manuscript, we performed further sub-clustering, to identify cell subtypes. For example: 1) we re-embedded 745,494 cells from the lateral plate & intermediate mesoderm derivatives, identifying 22 subtypes, most of which are corresponding to different types of mesenchymal cells (**Fig. 3f; Supplementary Table 12**). 2) we re-embedded 296,020 cells (glutamatergic neurons, GABAergic neurons, spinal cord dorsal progenitors, and spinal cord ventral progenitors) from stages <E13, identifying 18 different neuron subtypes (**Fig. 5e; Supplementary Table 12**).

Of note, we processed and analyzed the "birth-series" dataset (n = 962,697 nuclei after removing low quality cells and potential doublets cells) and the "early vs. late somites" data (n = 104,671 nuclei after removing low quality cells and potential doublets cells) using exactly the same strategy, except without performing sub-clustering on each major cell cluster.

Whole mouse embryo analysis

Each cell was assigned to the mouse embryo from which it derived based on its reverse transcription barcode. For each of the 74 samples, UMI counts mapping to the sample were aggregated to generate a pseudo-bulk RNA-seq profile for the sample. Each cell's counts were then normalized by dividing by its estimated size factor. The data were then log2-transformed after adding a pseudocount, and PCA was performed on the transformed data using the 3,000 most highly variable genes. The normalization and dimension reduction were performed using Monocle/v3.

Quantitatively estimating cell number for individual mouse embryo at any stage during organogenesis

To estimate the cell number of individual embryos, we selected a representative embryo from 12 timepoints at 1 day increments, from E8.5 to P0 (roughly considered as E19.5). Each embryo was digested with proteinase K overnight, and total genomic DNA was isolated with a Qiagen puregene tissue kit (Qiagen cat. no. 158063). DNA was quantitated and cell number was estimated by taking the total ng recovered, and assuming 2.5 billion base pairs per mouse genome (times two for a diploid cell), 650g per mole of a base pair. Estimating cell number this way does not include any losses due to the DNA preparation, and does not count non-nucleated cells.

Based on the experimentally estimated cell numbers of those 12 embryos, we applied polynomial regression (degree = 3) to fix a curve across embryos between the embryonic day and log2-scaled cell number (adjusted $R^2 > 0.98$) (**Supplementary Fig. 26a**). P0 was treated as E19.5 in the model. Then, the total cell number of a whole mouse embryo at any day between E8.5 and P0 is predicted using the below formula: Log_2(Cell number) = 0.011369 $\times$ day^3 -0.583861 $\times$ day^2 + 10.397036 $\times$ day - 35.469755

To further estimate the "doubling time" of the total cell number in a whole mouse embryo, at a given timepoint (day), we took the derivatives from the above formula as the log2-scaled proliferation rate p(day), and then calculated $24 \times 2 / 2^{p(day)}$, resulting in an estimate of the number of hours required for the mouse embryo to duplicate its total cell number (**Supplementary Fig. 26b**).

Spatial mapping with Tangram

To infer the spatial origin of each lateral plate & intermediate mesoderm derivative, we

used a public dataset called Mosta[32], which profiles spatial transcriptomes for 53 sections of mouse embryos spanning 8 timepoints from E9.5 to E16.5. We combined this data with our own data to perform spatial mapping analysis using Tangram[37]. First, we first randomly downsampled the total number of voxels within each section from Mosta to 9,000 for computational efficiency. Then, we combined the scRNA-seq data from three adjacent timepoints from our dataset with the Mosta data for each timepoint. For example, for the E16.5 timepoint of the Mosta data, we combined the scRNA-seq data from E16.25, E16.5, and E16.75 from our dataset. We used the Tangram with default parameters to estimate the spatial coordinates of cells from each cell type in the scRNA-seq data, and then visualized the results on the coordinates provided by Mosta. The Tangram model was trained in GPU mode using a NVIDIA A100 GPU. After applying Tangram to each section, a cell-by-voxel matrix with mapping probabilities was returned. This matrix shows the probability that each cell originated from each voxel in the section. To reduce noise, we smoothed the mapping probabilities for each voxel by averaging the values of their k nearest neighboring voxels. The value of k was calculated by using the natural log of the total number of voxels on that section. Finally, we scaled the smoothed mapping probabilities to 0 to 1 across voxels of each section.

Generating tree of cell types for mouse development

We collected and combined scRNA-seq data from four published datasets, which consisted of 110,000 cells spanning E0 to E8.5, and the main dataset described in this paper, which consisted of 11.4 million cells spanning E8 to P0 (**Supplementary Table 17**). We generated the tree of cell types for mouse development by following these steps:

- First, we divided the cells into 14 subsystems based on their data source, developmental window, and cell type annotations. The first two subsystems, Pre-gastrulation and Gastrulation, were based on external datasets and corresponded to the pre-gastrulation

and gastrulation phases of development. The remaining 12 subsystems, Blood, Brain & spinal cord, Endothelium, Epithelial cells, Eye, Gut, Lateral plate mesoderm, Mesoderm, Notochord, PNS glia, PNS neurons, and Kidney, were derived from the data reported in this paper and collectively encompassed organogenesis & fetal development.

- Second, we performed dimensionality reduction separately on cells from each of the 14 subsystems. This allowed us to correct or refine some cell type annotations, ultimately identifying 283 annotated cell type nodes. Each of these annotated cell type nodes is derived from a single data source, and some of them are redundant, which can help to bridge between datasets. For example, there are midbrain nodes deriving from both the Gastrulation and Brain & spinal cord subsystems.

- Third, within each subsystem, we identified pairs of cells that were mutual nearest neighbors (MNN) in 30-dimensional principal component analysis (PCA) space. We used k = 10 neighbors for the pre-gastrulation and gastrulation subsystems, and k = 15 for the organogenesis & fetal development subsystems. While the majority of MNNs occurred within cell type nodes, some MNNs spanned nodes and are likely to be enriched for bona fide cell type transitions. To systematically approach this, we calculated the total number of MNNs that spanned each possible pair of cell type nodes within a given subsystem, normalized by the total number of possible MNNs between those nodes. We then ranked all possible intra-subsystem edges based on this metric. Notably, due to the complexity of the "Brain & spinal cord" subsystem, we applied the heuristic in two stages. First, we applied it to the subset of cell types corresponding to the patterned neuroectoderm. Then, we repeated the process to identify edges between the patterned neuroectoderm and its derivatives (*i.e.* neurons, glial cells, *etc.*).

- Fourth, we manually reviewed the ranked list of 1,155 candidate edges for biological plausibility. We selected edges with a normalized MNN score greater than 1, resulting

in 452 edges. We manually annotated these edges as more likely to correspond to either developmental progression or spatial continuity. We added edges to the pre-gastrulation subsystem manually, as only a handful of cells were profiled in this subsystem.

- Finally, to bridge subsystems, we performed batch correction and co-embedding of selected timepoints from either the pre-gastrulation and gastrulation datasets, or the gastrulation and organogenesis & fetal development datasets. This allowed us to identify equivalent cell type nodes, resulting in a third category of "dataset equivalence" edges. For example, we performed anchor-based batch correction[51] followed by integration between cells from E6.5 to E8.5 generated on the 10x Genomics platform[13] (n = 108,857 cells) and the earliest 1% of this dataset (0-12 somite stage embryos) generated by sci-RNA-seq3 (n = 153,597 nuclei) (**Supplementary Fig. 21b-c**). This allowed us to identify 36 cell types from the integrated dataset, which we used to identify "bridging" edges between the gastrulation subsystem and the later subsystems (**Supplementary Fig. 21d-e**). Most of the 12 organogenesis & fetal development subsystems originate in cell type nodes for which equivalent nodes are already present at gastrulation. However, the "Blood" and "PNS neuron" subsystems did not have equivalent nodes at gastrulation, presumably due to undersampling of this transition. We manually added edges to connect these subsystems with biologically plausible pseudo-ancestors. Altogether, we added 55 inter-subsystem edges.

- We created a rooted, directed graph that represents mouse development from E0 to P0. The graph was created using yFiles Hierarchical layout in Cytoscape/v3.9.1. For presentation purposes, we removed most of the spatial continuity edges, except for those between spinal cord dorsal and ventral progenitors after E13.0 and GABAergic and glutamatergic neurons after E13.0. We also merged nodes with redundant labels derived from different datasets (*i.e.*, dataset equivalence edges). This resulted in a rooted graph with 262 cell type nodes and 338 edges (**Fig. 6g**). We further evaluated the robustness of our approach to technical factors or parameter choices (**Supplementary**

Note 1; Supplementary Fig. 27).

Nominating key TFs and genes

The list of 1,636 mouse proteins that are putatively TFs was collated from AnimalTFDB/v3 (http://bioinfo.life.hust.edu.cn/AnimalTFDB/)[96]. For each edge in the cell type tree, we stratified each cell type transition into four phases. Specifically, we identified the subset of cells within each node that were either "inter-node" MNNs of the other cell type or "intra-node" MNNs of those cells. If A to B, this approach effectively models the transition as group 1to2to3to4 (**Supplementary Fig. 22a-b**). Next, we identified differentially expressed transcription factors (DETFs) and genes (DEGs) across each portion of the modeled transition, *i.e.* early (1 to 2), inter-node (2 to 3), and late (3 to 4), by applying FindMarkers function in Seurat/v3 with parameters (logfc.threshold = 0, min.pct = 0). This strategy highlights differences between cells that are most proximate to the cell type transition itself.

After excluding dataset equivalence edges and the "Pre-gastrulation" subsystem, we nominated key TFs and genes that specify cell types for each of the 436 edges. Of note, the directionality of many of these edges was not immediately obvious (*i.e.* those annotated as "spatial continuity" edges). In these cases, the orientation of the "early" and "late" phases is arbitrary. For edges with a relatively small number of mutual nearest neighbors (MNN) pairs, we expanded each group to at least 200 cells by iteratively including their MNNs within the same cell type, to increase statistical power.

Identifying cell types with abrupt transcriptional changes before vs. after birth

To systematically identify which cell types exhibit abrupt transcriptional changes before

versus after birth, we performed the following steps for each of the 190 cell types:

- The 71 cell types with at least 200 cells from P0 and at least 200 cells from at least five timepoints prior to P0 were retained for the following analysis.

- We combined cells from animals harvested subsequent to E16 and performed PCA based on the top 2,500 highly variable genes.

- Timepoints with at least 200 cells were selected and cells were downsampled from each timepoint to the median number of cells across those selected timepoints.

- The k-nearest neighbors (k was adjusted for different cell types, by taking the log2-scaled median number of cells across the selected timepoints) were searched in PCA space (n = 30 dimensions).

- We calculated the average proportion of nearest neighbor cells that were from a different timepoint for cells within each cell type. In this framing, a low proportion of neighbors from different timepoints corresponds to a relatively abrupt change in transcriptional state.

We performed a similar analysis in the birth-series dataset. For each major cell cluster in the birth-series dataset, we took cells from the six pups delivered by C-section and calculated the Pearson correlation coefficient between the timepoint of each cell and the average timepoints of its 10 nearest neighbors identified from the global PCA embedding (n = 30 dimensions). In this framing, a high correlation indicates that the cell and its nearest neighbors all underwent rapid, synchronized changes in transcriptional state.

Supplementary Note 1

To evaluate whether our approach is robust to technical factors or parameter choices, we took the following three approaches. First, we examined whether the MNNs that we identified between different cell types were enriched for cells from the same embryo. Since the data from pre-gastrulation and gastrulation were generated from pooled samples, we only investigated this phenomenon for later stages, *i.e.* E8-P0 data generated via sci-RNA-seq3. Overall, we found that only 16.4% of MNNs from different cell types were between cells from the same embryo. However, we notably only profiled one embryo for most timepoints, which may inflate this value relative to what it might have been if we had profiled multiple embryos per timepoint. This is supported by the fact that when we look at windows with multiple embryos profiled per timepoint (E8-E10 and E13-E13.75), the proportion of MNNs from different cell types that connect cells from the same embryo was only 10.5% for E8-E10, and only 2.4% for E13-E13.75 (**Supplementary Fig. 27a**). Overall, the fact that MNNs spanning cell types overwhelmingly connect cells from different embryos (and different timepoints) is reassuring. Second, to assess the robustness of MNNs to cell sampling, we

randomly subsampled 80% of cells from each developmental system during organogenesis & fetal development (except for notochord, which is a relatively rare cell type). We then repeated our MNN approach on the subsamples and compared the resulting numbers of MNNs obtained for each edge to those obtained when using the full dataset. This process was repeated 100 times for each developmental system. The resulting correlation coefficients ranged from 0.92 to 0.99, with an average of 0.98 (**Supplementary Fig. 27b**). This suggests that the MNNs we identified are robust to cell sampling.

Third, the k parameter is critical when using kNNs to identify MNNs between cell types. The original k value was selected based on the log2-transformed median number of cells across cell types (k = 10 neighbors for pre-gastrulation and gastrulation subsystems, k = 15 for organogenesis & fetal development subsystems). To determine the effect of k parameter choice on the MNNs identified between cell types, we examined different k values (k = 5, 10,

20, 30, 40, 50) for kNN to identify MNNs for each developmental system during organogenesis & fetal development. We then compared the results to the original result, which was based on k = 15. The resulting Spearman correlation coefficients ranged from 0.92 to 0.99, with an average of 0.98 (**Supplementary Fig. 27c**). This suggests that the MNNs we identified are robust to the choice of k parameter.

Data availability

All data used here have been made freely available via https://atlas.gs.washington.edu/jax/. The data generated in this study can be downloaded in raw and processed forms from the NCBI Gene Expression Omnibus under accession number GSE186069 and GSE228590. The code used here has been provided from https://github.com/ChengxiangQiu/JAX_code. The supplementary tables can be downloaded from here:
https://shendure-web.gs.washington.edu/content/members/cxqiu/public/nobackup/tmp/

Acknowledgments

We thank the members of the Shendure lab for helpful discussions. This work was supported by the Brotman Baty Institute for Precision Medicine, a grant from Paul G. Allen Frontiers Group (Allen Discovery Center for Cell Lineage Tracing to J.S.) and the National Institutes of Health (1UM1HG011586 to W.N.S, J.S and C.M.D..; R01HG010632 to J.S. and C.T.). I.W. and S.A.M were supported by N.I.H. grant UM1OD023222 and the JAX Director's Innovation Fund. J.S. is an Investigator of the Howard Hughes Medical Institute.

Competing Financial Interests Statement

J.S. is a scientific advisory board member, consultant and/or co-founder of Scale Biosciences, Prime Medicine, Cajal Neuroscience, Guardant Health, Maze Therapeutics, Camp4

Therapeutics, Phase Genomics, Adaptive Biotechnologies, Sixth Street Capital and Pacific Biosciences. C.T. is a co-founder of Scale Biosciences. All other authors declare no competing interests.

3.14 Figures

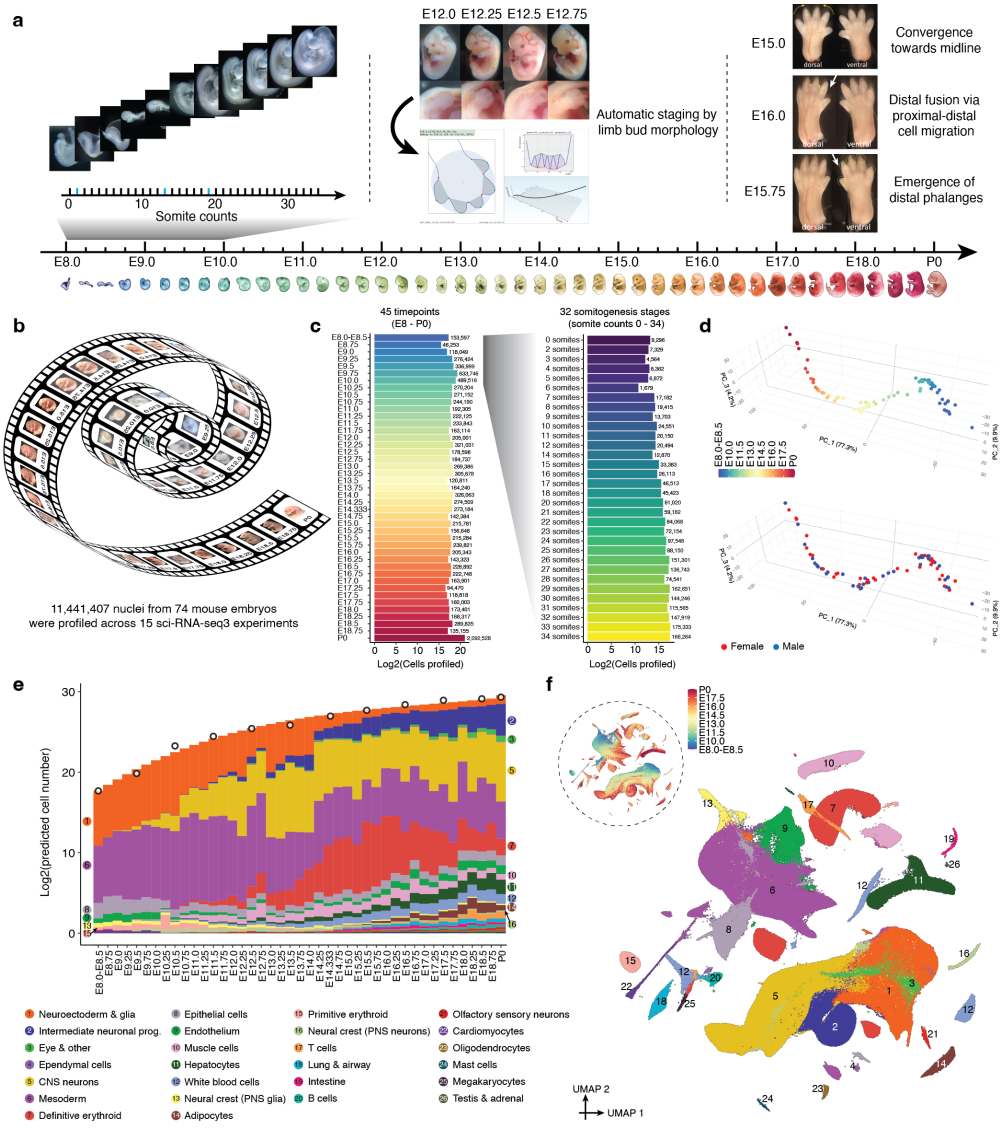

Figure 3.1: **Figure 1. A single cell transcriptional timelapse of mouse development, from gastrula to pup. a,** Embryos were collected and precisely staged based on morphological features, including by counting somite numbers (up to E10) and an automated process that leverages limb bud geometry (E10-E15). **b,** Across 15 sci-RNA-seq3 experiments, we generated single nucleus profiles for 11.4M cells from 74 embryos spanning mouse development from late gastrulation to birth. **c,** The number (log2 scale) of nuclei profiled at each timepoint. **d,** Embeddings of pseudo-bulk RNA-seq profiles of 74 mouse embryos in PCA space with visualization of top three PCs. **e,** Composition of embryos from each 6-hr bin by major cell cluster. The y-axis is scaled to the estimated cell number (log2 scale) at each timepoint (**Methods**). **f,** 2D UMAP visualization of the whole dataset. Colors and numbers correspond to 26 major cell cluster annotations.

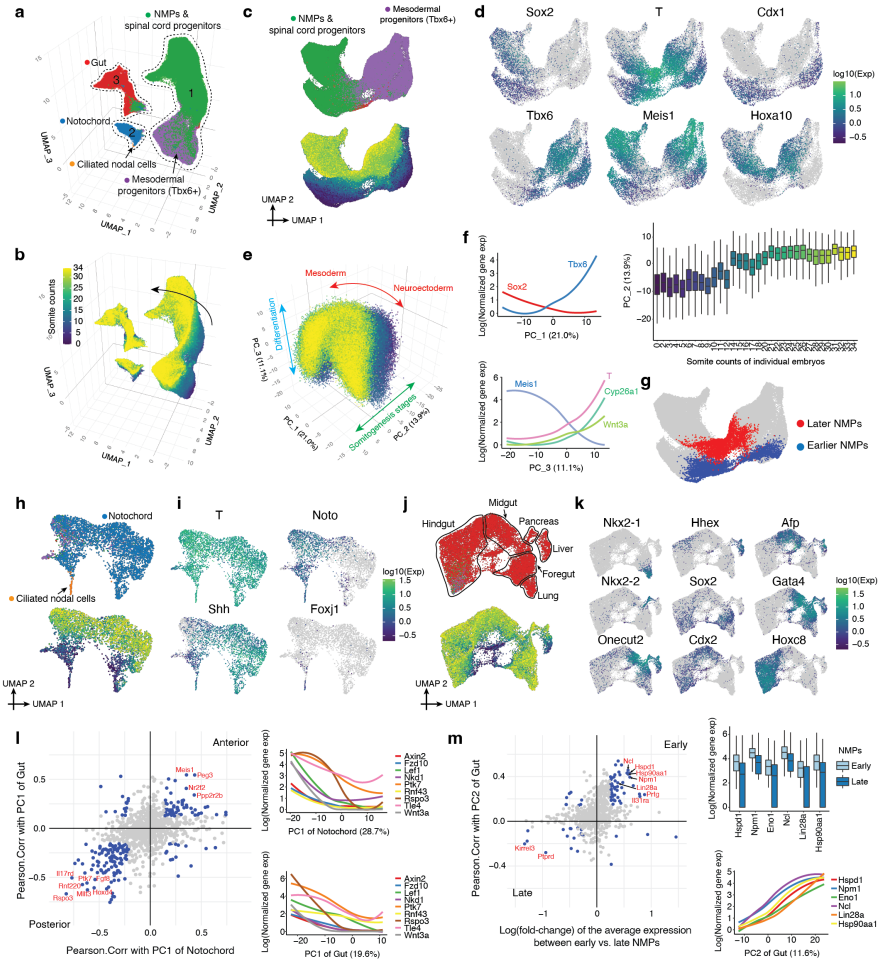

Figure 3.2: **Figure 2. Transcriptional heterogeneity in the posterior embryo during the early somitogenesis. a,** Re-embedded 3D UMAP of 121,118 cells which were corresponding to posterior embryo development during the early somitogenesis (somite counts 0-34; E8-E10). **b,** The same UMAP as in panel a, colored by somite counts. **c,** Re-embedded 2D UMAP of cells from cluster 1 in panel a. **d,** The same UMAP as in panel c, colored by gene expression of marker genes. **e,** 3D visualization of the top three principal components (PCs) of gene expression variation in cells from cluster 1. **f,** Correlations between top three PCs and the normalized expression of selected genes or somite counts. **g,** The same UMAP as in panel c, with earlier and later NMPs highlighted. **h,** Re-embedded 2D UMAP of cells from cluster 2 in panel a. **i,** The same UMAP as in panel h, colored by gene expression of marker genes. **j,** Re-embedded 2D UMAP of cells from cluster 3 in panel a. **k,** The same UMAP as in panel j, colored by gene expression of marker genes. **l,** Left: the correlation between gene expression for the top highly variable genes and either PC1 of notochord or PC1 of gut. Right: gene expression of selected genes involved in Wnt signaling are plotted over PC1 of notochord or PC1 of gut. **m,** Left: the fold-change of the average expression for the top highly variable genes between early vs. late NMPs, and the correlation between gene expression for the top highly variable genes and PC2 of gut. Right: gene expression of selected genes (several *Myc* targets[217], *Lin28a*, *Hsp90aa1*) are plotted between early vs. late NMPs or over PC2 of gut.

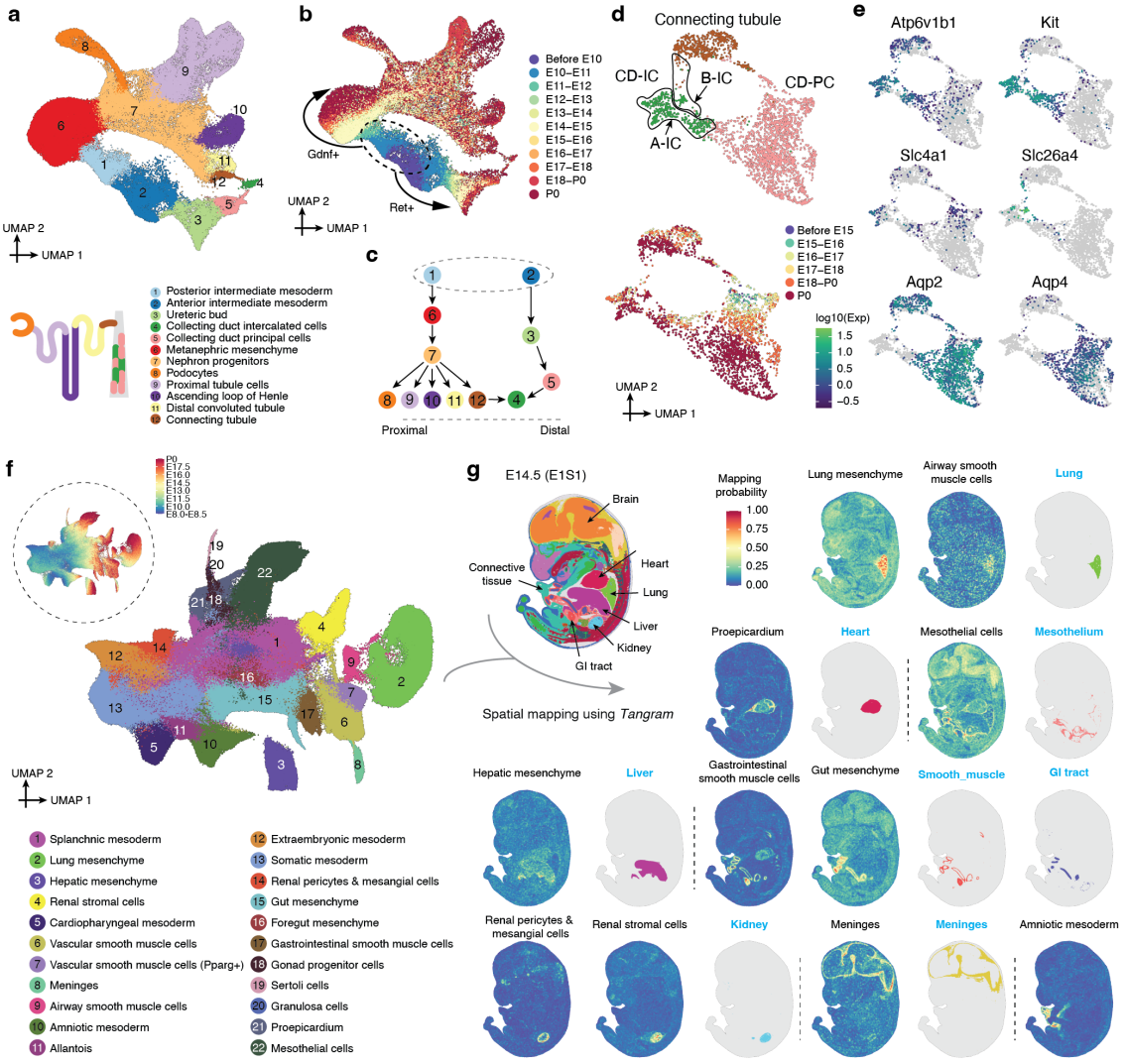

Figure 3.3: **Figure 3. Diversification of the intermediate and lateral plate mesoderm. a,** Re-embedded 2D UMAP of 95,226 cells corresponding to renal development. **b,** The same UMAP as in panel a, colored by developmental stage. **c,** Manually inferred relationships between annotated renal cell types. **d,** Re-embedded 2D UMAP of 2,894 cells from connecting tubule cells and collecting duct cells. **e,** The same UMAP as in panel d, colored by expression of marker genes. **f,** Re-embedded 2D UMAP of 745,494 cells from lateral plate and intermediate mesoderm derivatives. **g,** To infer the spatial origin of each lateral plate and intermediate mesoderm derivative, we leveraged a public spatial transcriptomes dataset, Mosta [32], together with our data and the Tangram algorithm[37]. The spatial mapping probabilities across voxels on the E14.5 section for selected subtypes within the lateral plate and intermediate mesoderm derivatives are shown.

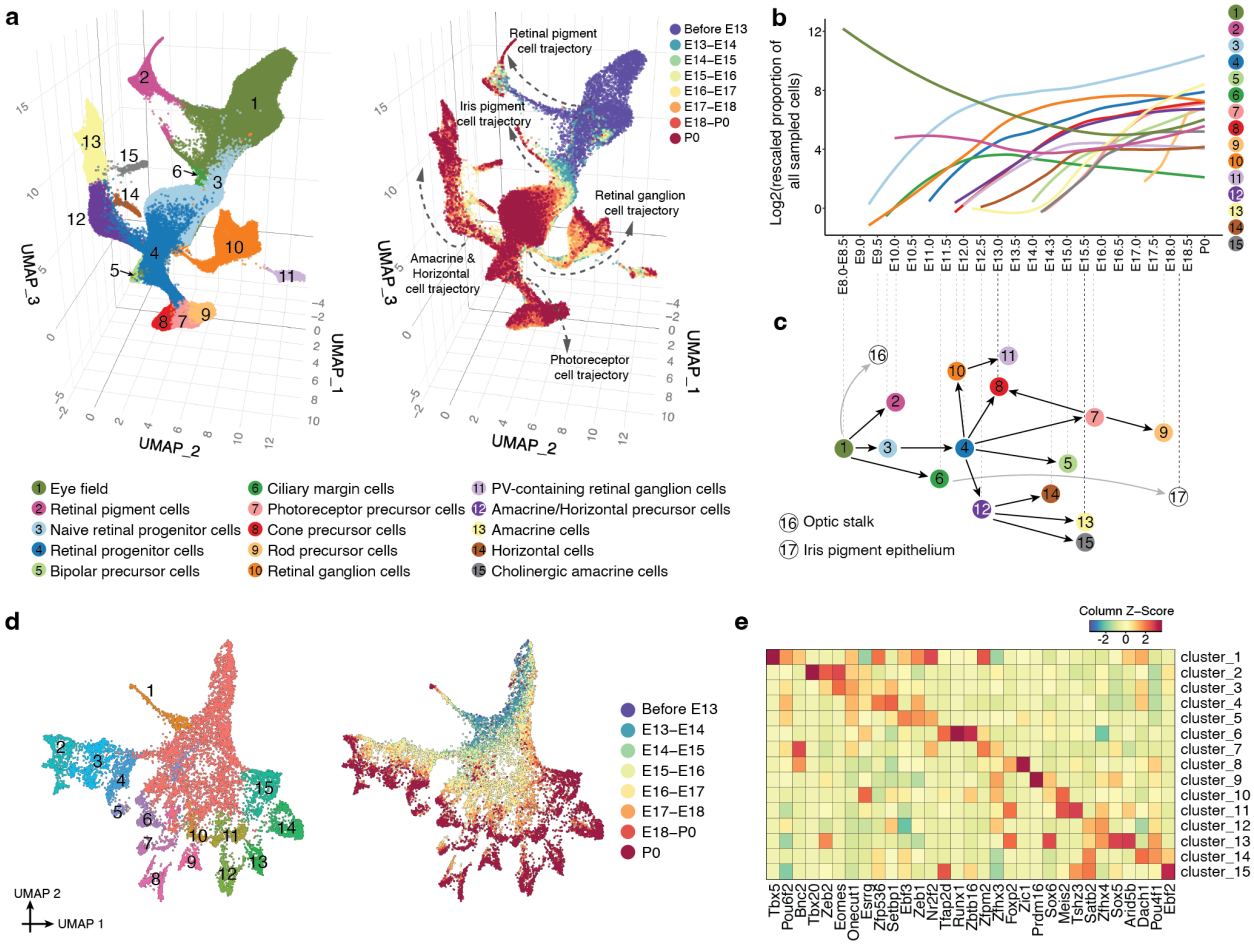

Figure 3.4: **Figure 4. The timing and trajectories of retinal development.** **a,** Re-embedded 3D UMAP of 160,834 cells corresponding to the retinal development from E8 to P0. **b,** Rescaled proportion of profiled cells (log2; y-axis) for each cell type shown in panel a, as a function of developmental time (x-axis). **c,** Schematic of retinal cell types emphasizing the timing at which they first appear and their inferred developmental relationships from E8-P0. **d,** Re-embedded 2D UMAP of retinal ganglion cells. Cells are colored by either clusters or timepoint. **e,** The top 3 TF markers of the 15 clusters shown in panel d. Marker TFs were identified using the FindAllMarkers function of Seurat/v3[51].

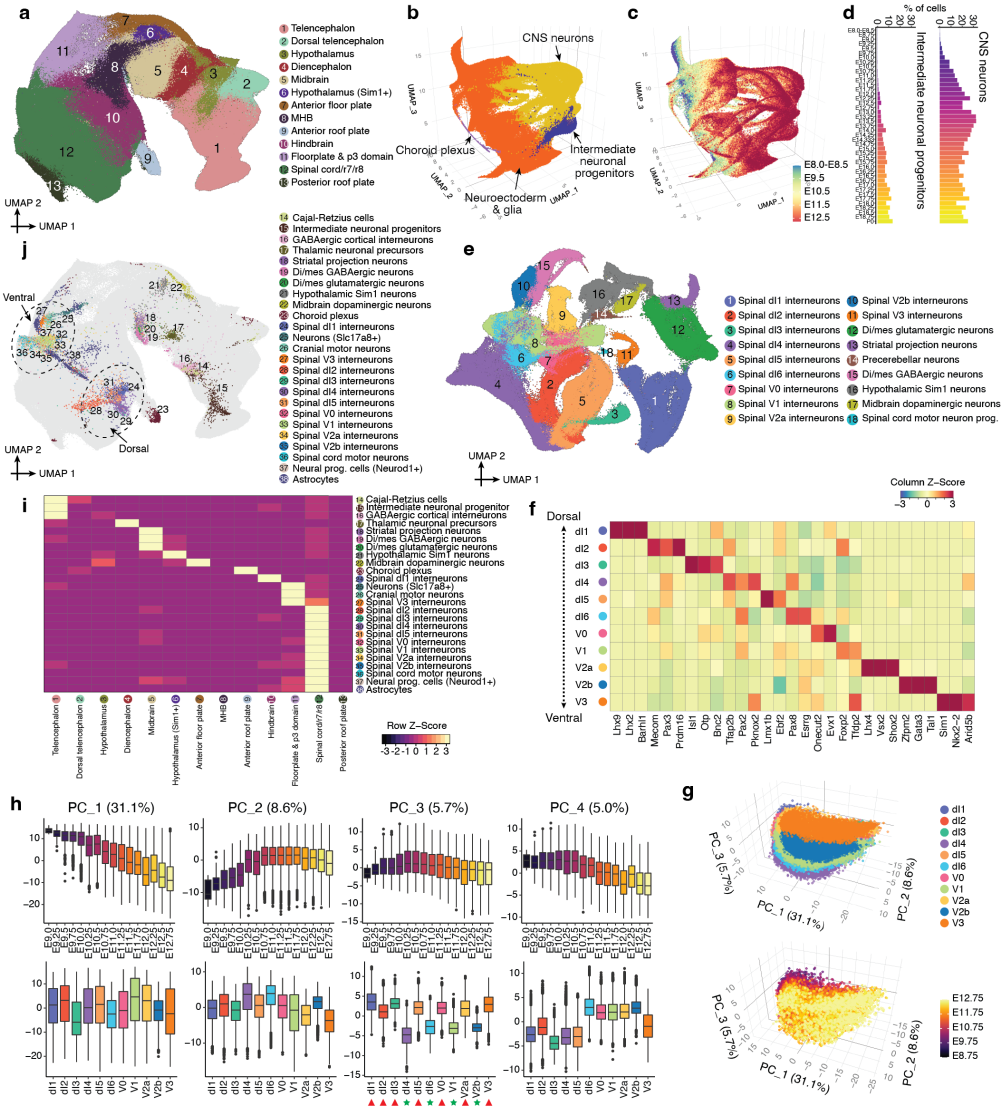

Figure 3.5: **Figure 5. The emergence of neuronal subtypes from the patterned neuroectoderm.** **a,** Re-embedded 2D UMAP of 1,185,052 cells, corresponding to different neuroectodermal territories. **b,** Re-embedded 3D UMAP of 1,772,567 cells from neuroectodermal territories together with derived cell types. **c,** The same UMAP as in panel b, colored by timepoint. **d,** Composition of embryos from each 6-hr bin by intermediate neuronal progenitor (left) and CNS neuron (right) major cell clusters. **e,** Re-embedded 2D UMAP of 296,020 cells (glutamatergic neurons, GABAergic neurons, spinal cord dorsal progenitors, and spinal cord ventral progenitors) from stages < E13. **f,** The top 3 TF markers of the 11 spinal interneurons. **g,** 3D visualization of the top three PCs of gene expression variation in 11 spinal interneurons. **h,** Correlations between top four PCs and timepoints (top row) or cell types (bottom row). **i,** The number of mutual nearest neighbors (MNN) pairs between pairwise neuroectodermal territories (column) and their derivative cell types (row). **j,** The same UMAP as in panel a, but with inferred progenitor cells colored by derivative cell type with the most frequent MNN pairs.

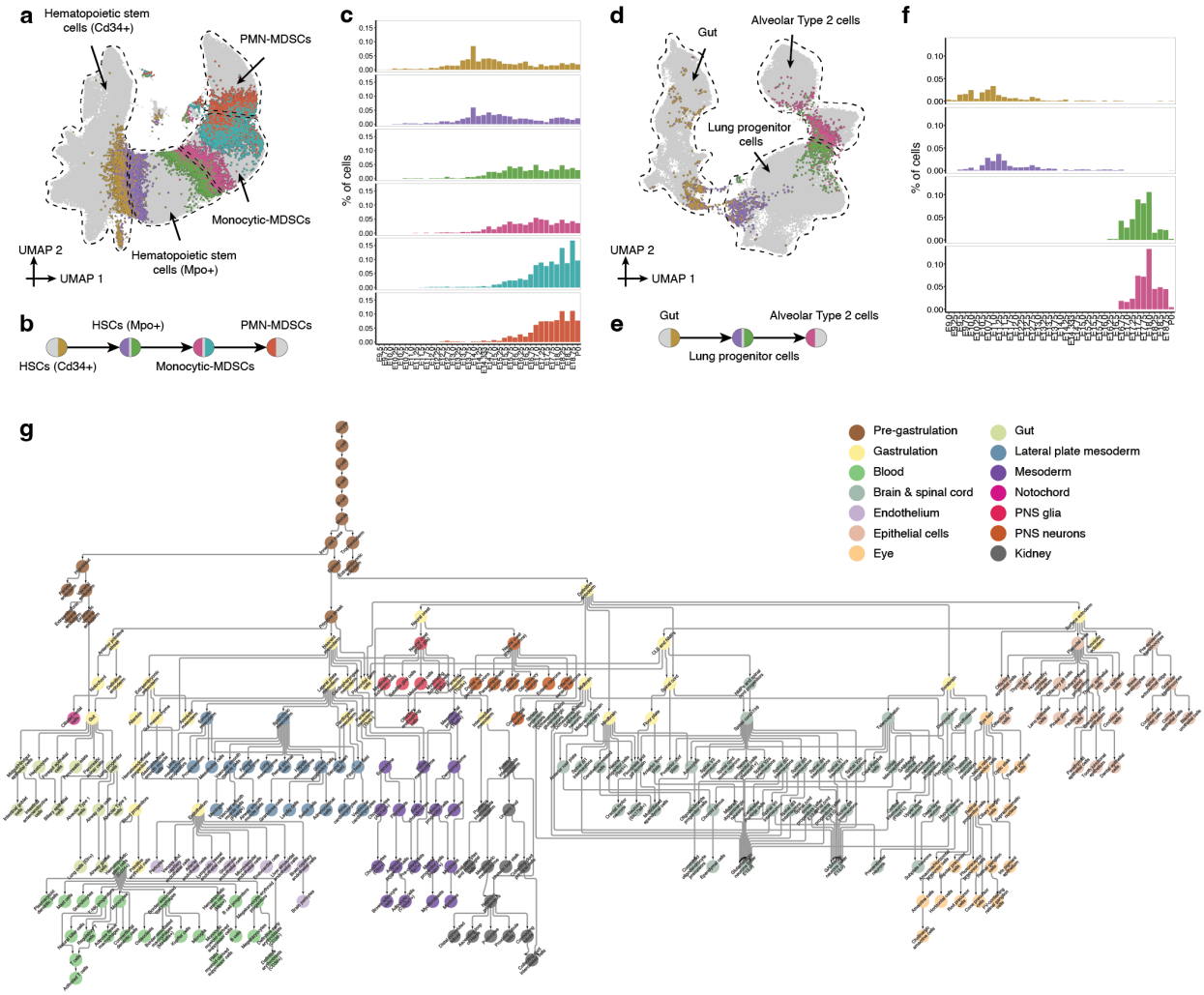

Figure 3.6: **Figure 6. A data-driven tree relating cell types throughout mouse development, from zygote to pup. a,** Illustration of basis for edge inference heuristic. Re-embedded 2D UMAP of 101,001 cells from selected cell types within the "Blood" subsystem. Cells involved in MNN pairs that bridge cell types are colored. **b,** Inferred lineage relationships between annotated cell types in panel a, with corresponding color scheme. **c,** The % of inter-cell-type MNN cells (y-axis) over the total number of cells profiled from embryos from the corresponding time bin. **d,** Additional illustration of basis for edge inference heuristic. Re-embedded 2D UMAP of 71,718 cells from selected cell types within the "Gut" subsystem. **e,** Inferred lineage relationships between annotated cell types in panel d, with corresponding color scheme. **f,** The % of inter-cell-type MNN cells (y-axis) over the total number of cells profiled from embryos from the corresponding time bin. **g,** A rooted, directed graph corresponding to a mouse development, spanning E0 to P0 (yFiles Hiearchic layout in Cytoscape/v3.9.1). It includes 262 cell type nodes and 338 edges. Nodes are colored and labeled by each of the 14 subsystems.

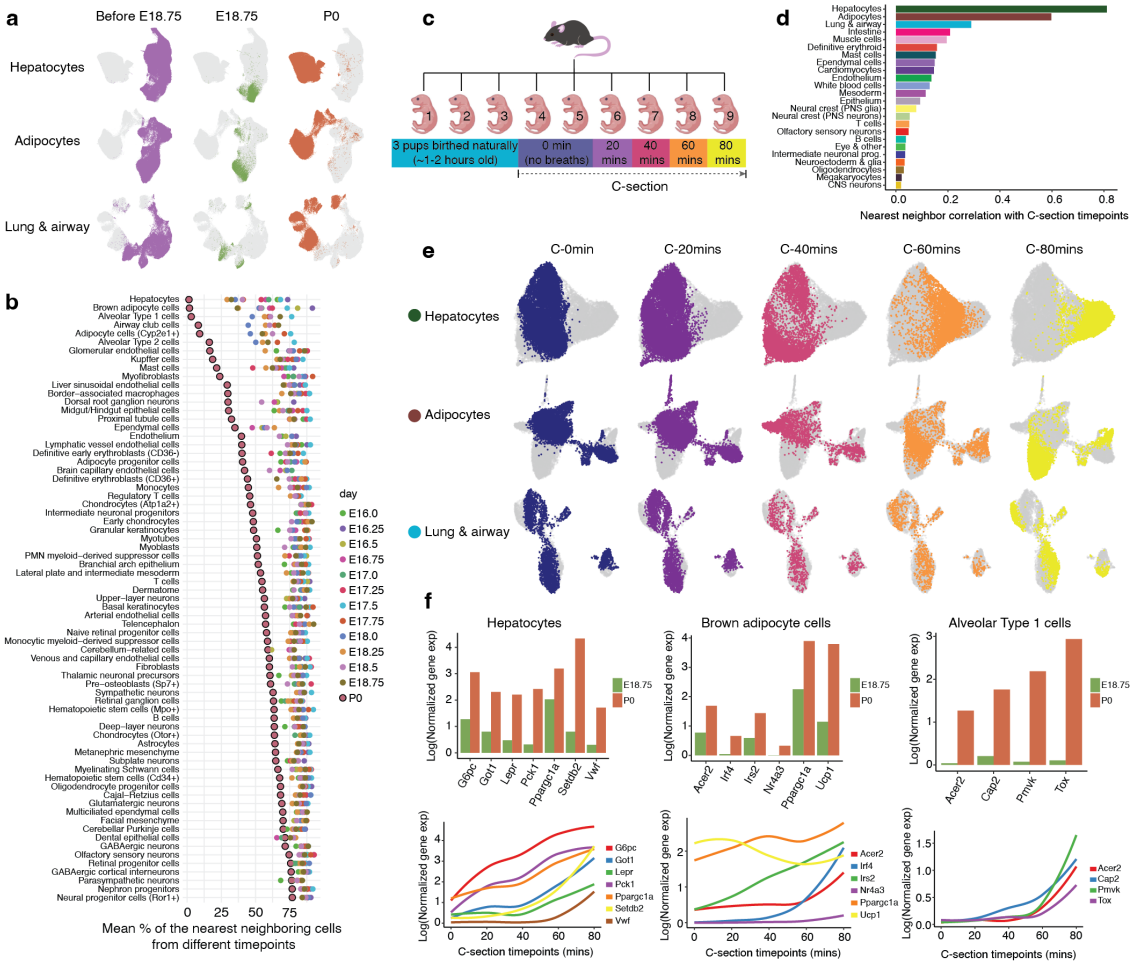

Figure 3.7: **Figure 7. Rapid shifts in transcriptional state occur in a restricted subset of cell types upon birth. a,** Re-embedded 2D UMAP of cells from three major cell clusters: hepatocytes, adipocytes, and lung and airway. **b,** Systematically exploring which cell types exhibit abrupt transcriptional changes before vs. after birth. **c,** A new scRNA-seq dataset ("birth-series") was generated from nuclei derived from nine individual pups from a single litter harvested shortly after delivery. **d,** For each major cell cluster in the birth-series dataset, we took cells from the six pups delivered by C-section and calculated a Pearson correlation coefficient between the timepoint of each cell and the average timepoints of its 10 nearest neighbors. **e,** Re-embedded 2D UMAP of cells from three major cell clusters, based on cells from six pups delivered by C-section in birth series dataset: hepatocytes (n = 41,122 cells), adipocytes (n = 19,696 cells), and lung and airway (n = 7,986 cells). **f,** Average normalized gene expression of selected genes are plotted between E18.75 vs. P0 in the original data (top), and normalized gene expression of the same genes are plotted as a function of C-section timepoints (bottom), for hepatocytes (left), brown adipocyte cells (middle), and alveolar type 1 cells (right), respectively.

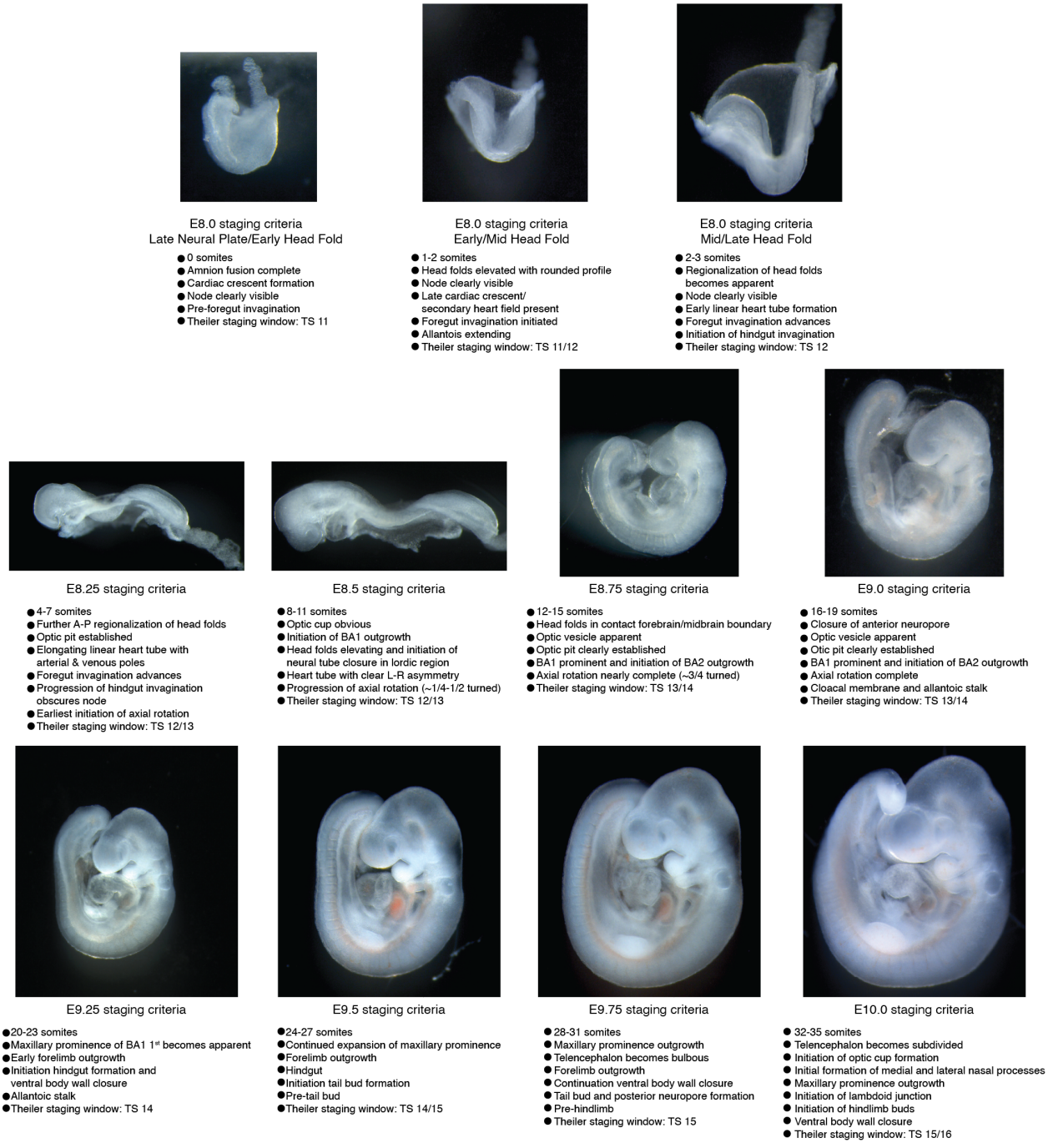

Figure 3.8: **Supplementary Figure 1. Embryos harvested between E8 and E10 were precisely staged based upon somite counting.** Harvested embryos were grouped into bins based on somite counting and further characterized based upon morphological features. Stage-representative images are shown with details of the main staging criteria for each coarse temporal bin listed. The approximately overlapping Theiler Stage (TS) is also noted for reference.

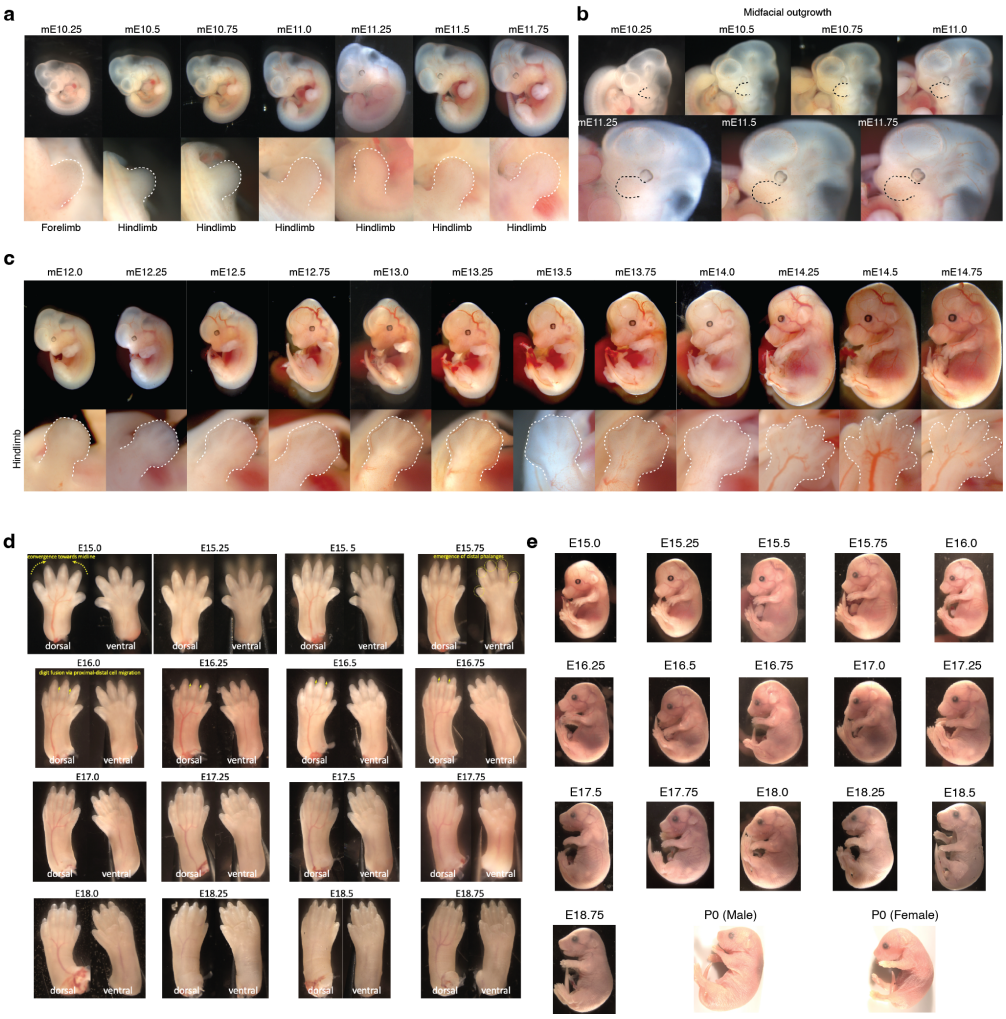

Figure 3.9: **Supplementary Figure 2. After E10, embryos were precisely staged based on morphological features.** This was mainly done using the embryonic mouse ontogenetic staging system (eMOSS), an automated process that leverages limb bud geometry to infer developmental stage[156, 157]. **a,** For each temporal bin at 6-hr increments from E10.25-E11.75, an image of a stage-representative embryo is shown. **b,** View of the craniofacial region of embryos shown in panel a demonstrates that limb bud staging also recreates the ordered ontogenetic progression of craniofacial morphogenesis. **c,** For each temporal bin at 6-hr increments from E12.0-E14.25, an image of a randomly selected embryo is shown. **d,** eMOSS is able to stage E10.25-E4.75, after which limb morphology becomes too complex. The remaining timepoints (E17.0-E18.75) were staged based upon gestational age. For each temporal bin at 6-hr increments from E15.0-E18.75, an image of the hindlimbs of a randomly selected embryo is shown. **e,** For each temporal bin at 6-hr increments from E15.0-P0, an image of a stage-representative embryo is shown.

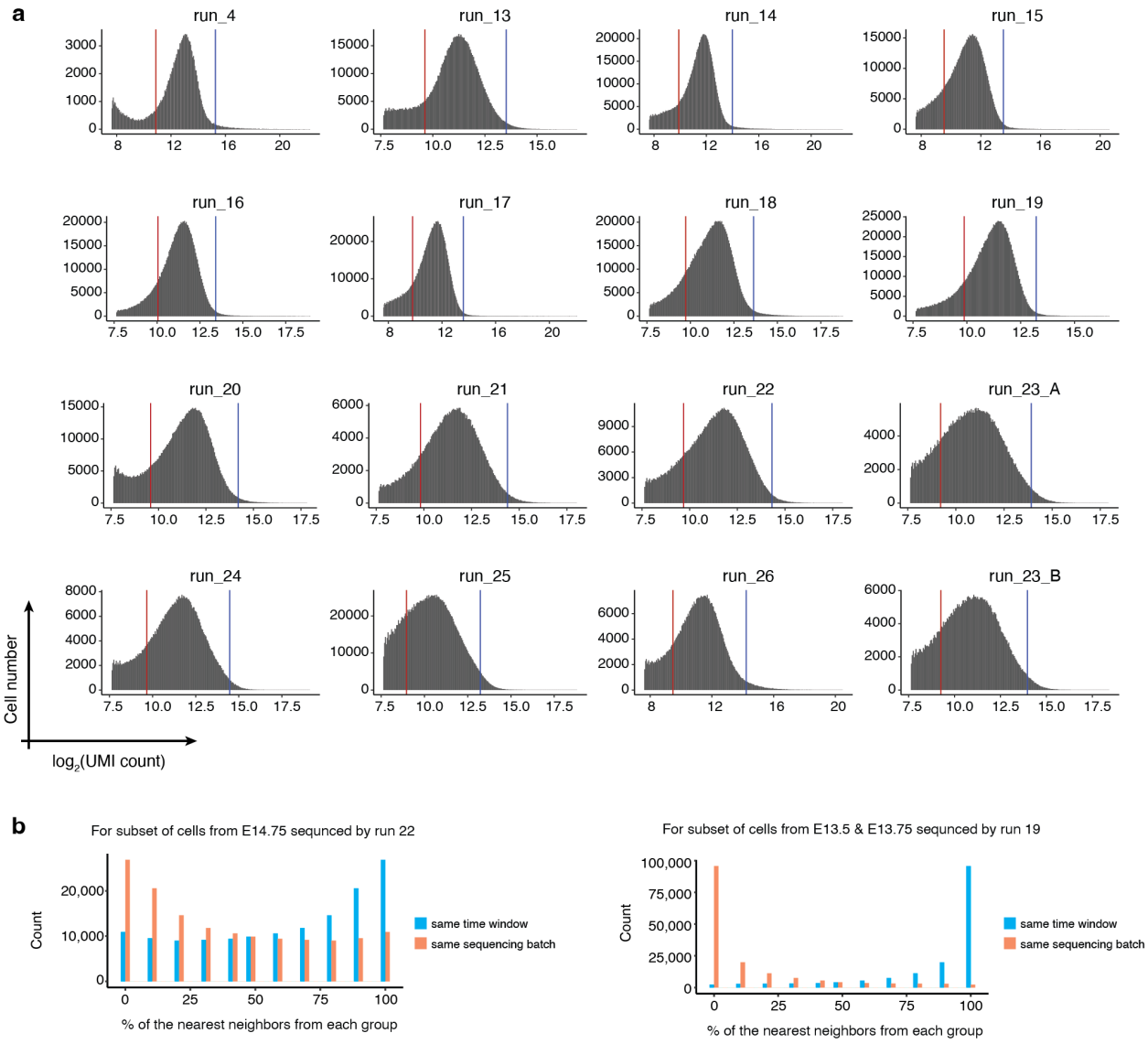

Figure 3.10: **Supplementary Figure 3. Higher quality sci-RNA-seq3 data as generated by an optimized protocol. a,** Histograms of log2(UMI count) per single nucleus for each of 15 sci-RNA-seq3 experiments. **b,** Although most of the embryos from the same approximate stage (*e.g.* E14.0-E14.75) were included in the same sci-RNA-seq3 experiment, we profiled extra nuclei in some experiments for a handful of timepoints to ensure sufficient coverage. Here we sought to leverage those instances to check for potential batch effects across experiments. For this, on the embedding learned from all of the data, we asked whether these cells' profiles are more similar to cells from the same experiment or, alternatively, cells from the same time window. The percentages of the nearest neighboring cells from the two groups for individual cells are presented in the histogram. In both examples, we observe that nearest neighbors are overwhelmingly cells from a different experiment (but the same time window), rather than cells from the same experiment (but a different time window).

Figure 3.11: **Supplementary Figure 4. Cells processed in different experiments are well-integrated without batch correction. a,** To further check for potential batch effects, we generated co-embeddings of samples processed from adjacent timepoints in different experiments, without batch correction. **b,** We also generated a co-embedding of cells from run 23 A (red) and run 23 B (green), which derived from the same sci-RNA-seq3 experiment but were sequenced on different NovaSeq runs.

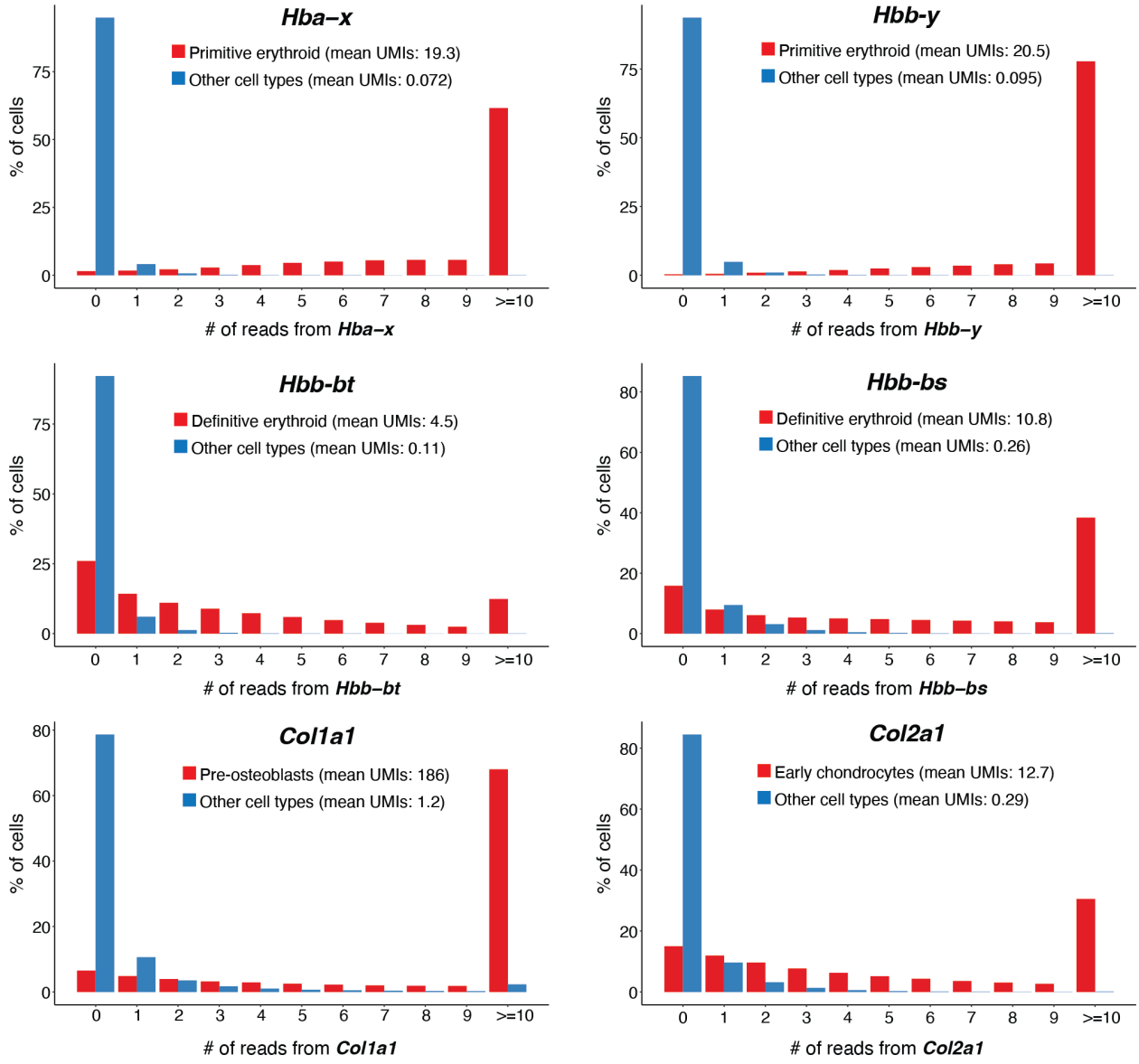

Figure 3.12: **Supplementary Figure 5.Ambient noise (*e.g.* as might be due to transcript leakage) was assessed by examining hemoglobin and collagen transcripts.** The distribution of the number of reads mapping to each selected hemoglobin or collagen gene across cells, for the cell type that is expected to express that gene at high levels (red) vs. all other cell types (blue). The mean UMI counts of cells in each group are also reported.

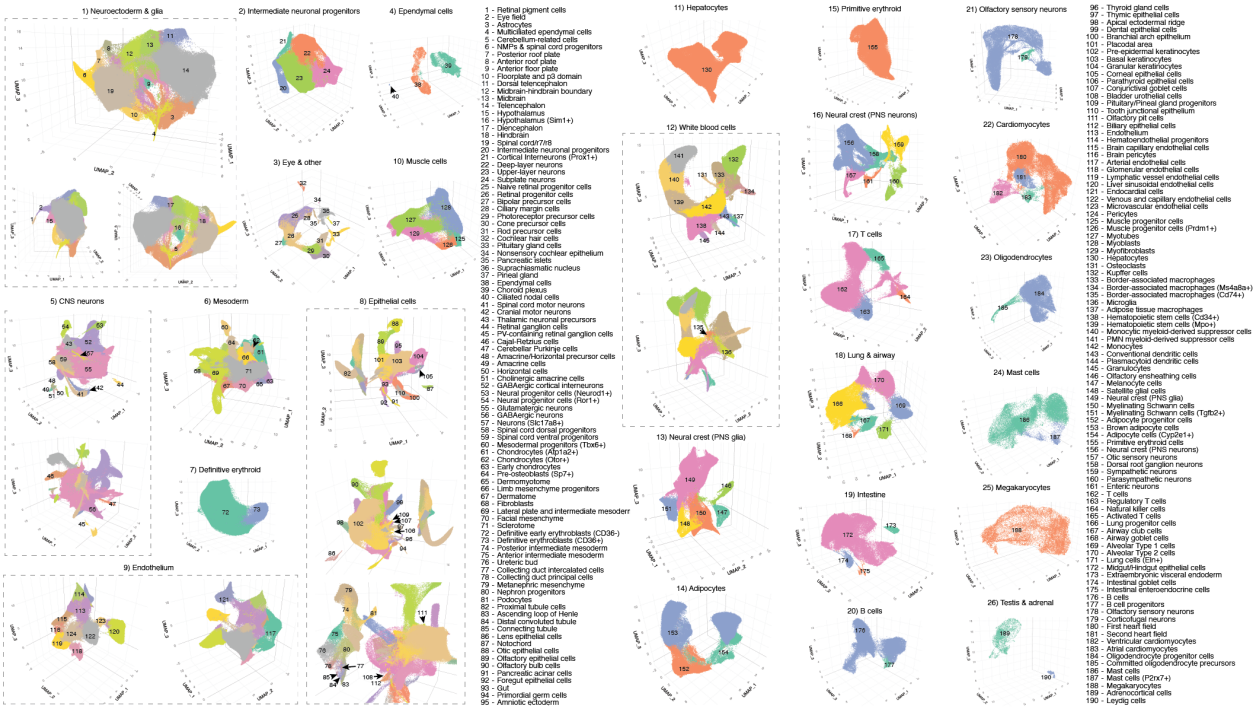

Figure 3.13: **Supplementary Figure 6. Cell type annotations.** For each of the 26 major cell clusters, we performed subclustering and then annotated each of 190 subclusters using at least two literature-nominated marker genes per cell type label (**Supplementary Table 5**).

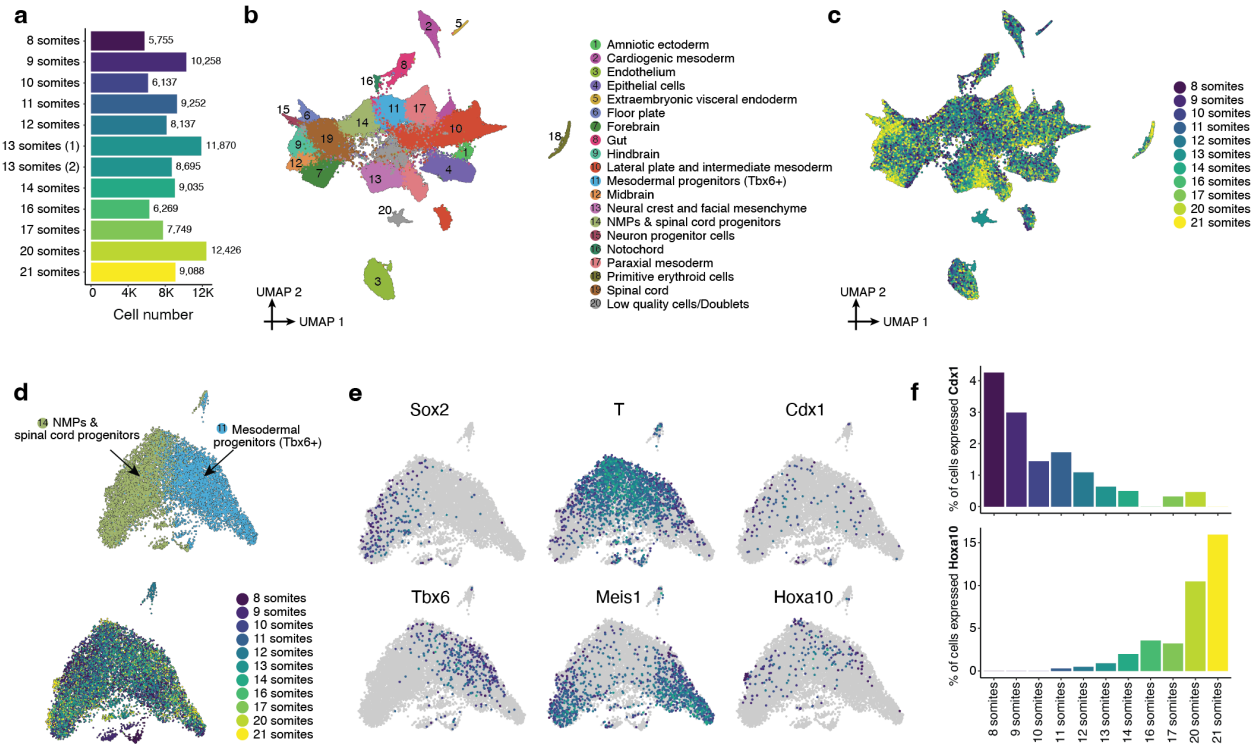

Figure 3.14: **Supplementary Figure 7. A validation sci-RNA-seq3 dataset of mouse embryos from somites 8 to 21.** To validate findings related to differences between embryos staged with early vs. late somite counts, particularly in NMPs, we profiled another 12 precisely staged mouse embryos, ranging from 8 to 21 somites, in an independent sci-RNA-seq3 experiment. **a,** The number of cells profiled from each embryo. **b,** 2D UMAP visualization of the validation dataset (all cell types). **c,** The same UMAP as in panel b, with cells colored by somite count of the originating embryo. **d,** Re-embedded 2D UMAP of 9,686 cells from NMPs and spinal cord progenitors (cluster 11) and mesodermal progenitors (*Tbx6+*) (cluster 14) in panel b. Cells are colored by either the original annotation (top) or somite count (bottom). **e,** The same UMAP as in panel d, colored by gene expression of marker genes: the neuroectodermal (*Sox2+*) vs. mesodermal (*Tbx6+*) fate[72]; the differentiation of bipotential NMPs (*T+*, *Meis1-*) towards either fate[218, 219]; earlier (*Cdx1+*) vs. later (*Hoxa10+*) NMPs[163]. **f,** Within the cells shown in panel d, the proportion of cells (y-axis) which express either *Cdx1* (top) or *Hoxa10* (bottom) are plotted as a function of somite count of the originating embryo.

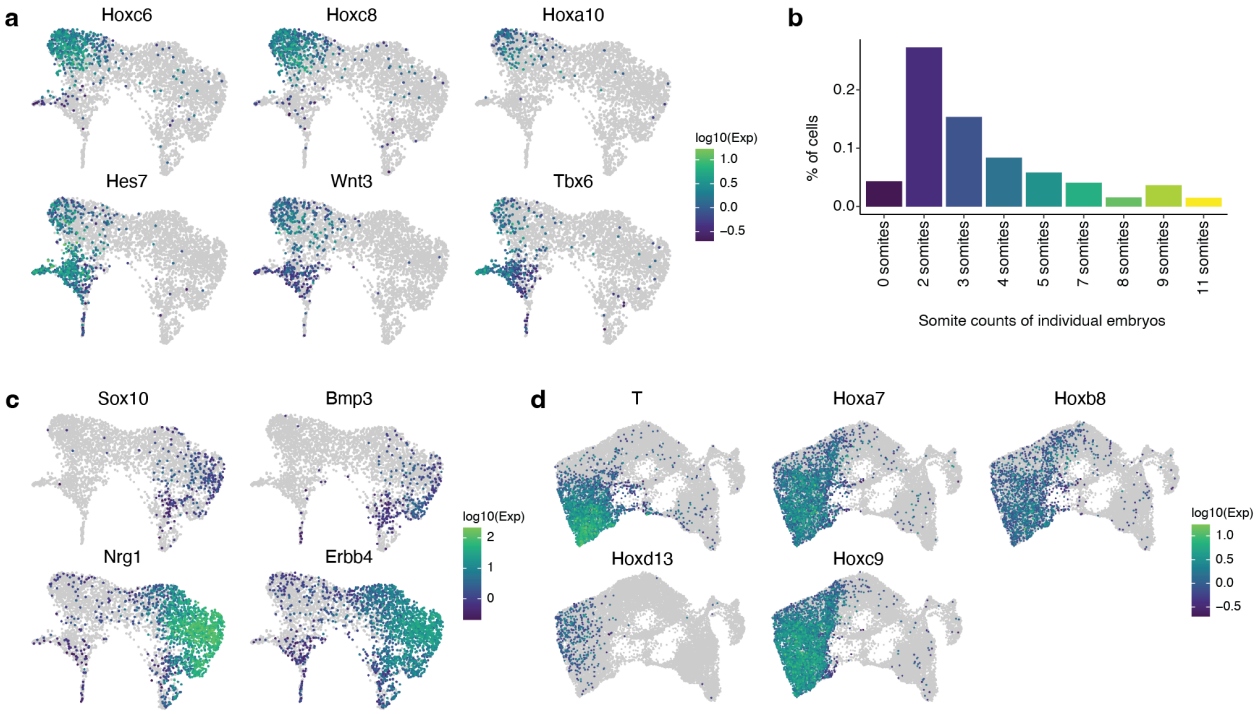

Figure 3.15: **Supplementary Figure 8. Transcriptional heterogeneity in the posterior embryo during the early somitogenesis. a,** The same UMAP as in **Fig. 2h**, colored by gene expression of marker genes which appear specific to the subpopulation of notochord cluster that is *Noto*+, including posterior *Hox* genes (*Hoxc6*, *Hoxc8*, *Hoxa10*), and genes involved in Notch signaling (*Hes7*), Wnt signaling (*Wnt3*) and mesodermal differentiation (*Tbx6*). **b,** Cell proportions falling into the ciliated nodal cell cluster for embryos with different somite counts. **c,** The same UMAP as in **Fig. 2h**, colored by gene expression of marker genes which appear specific to the subpopulation of the notochord *Noto*- and more strongly *Shh*+, including *Sox10*, *Bmp3*, *Nrg1*, and *Erbb4*. **d,** The same UMAP as in **Fig. 2j**, colored by gene expression of marker genes which appear specific to the posterior gut endoderm, including *T*, *Hoxa7*, *Hoxb8*, *Hoxd13*, and *Hoxc9*.

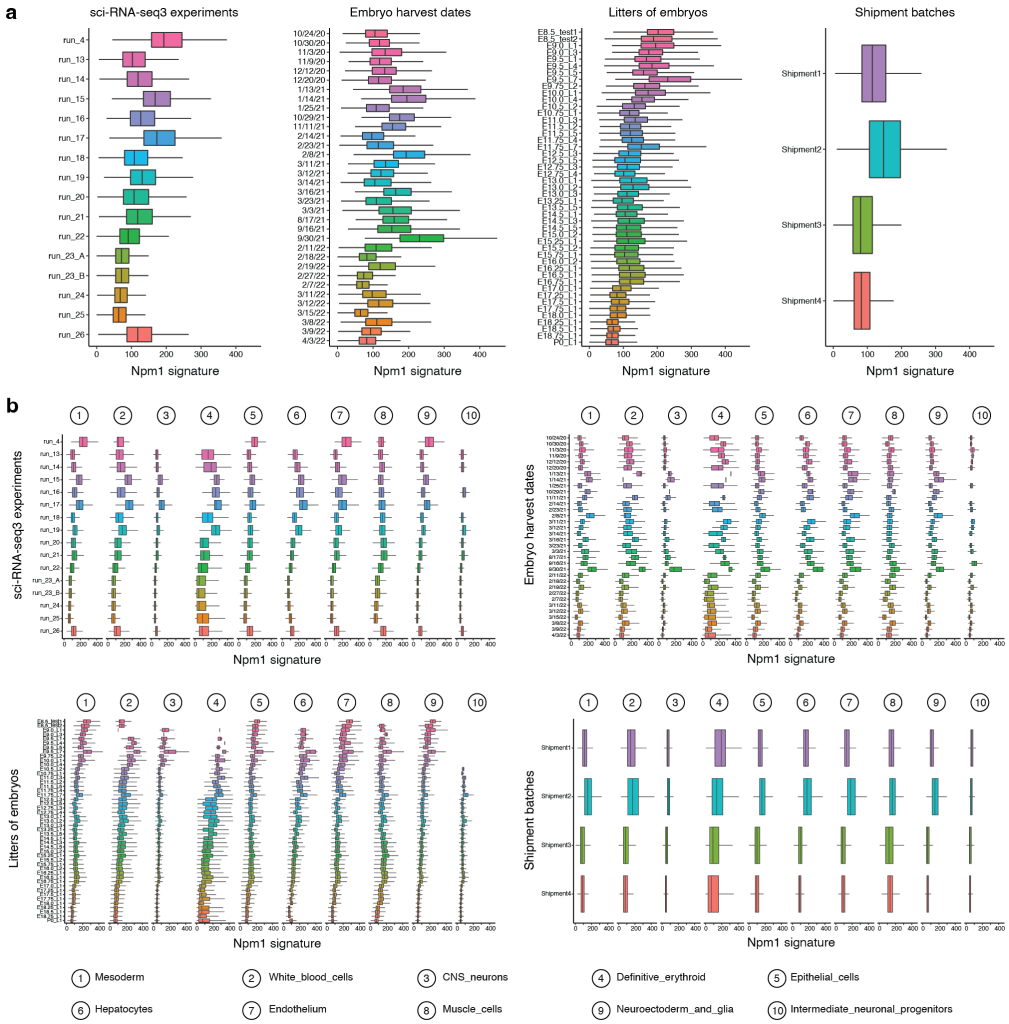

Figure 3.16: **Supplementary Figure 9. Checking the consistency of Npm1 signatures across different batches. a,** First, we downsampled the dataset to 1M cells using geosketch[220] and performed k-means clustering to ensure that each cluster contained roughly 500 cells. Second, we aggregated UMI counts for cells within each cluster to generate 2,289 meta-cells, and normalized the UMI counts for each meta-cell followed by log2-transformation. Third, we performed Pearson correlation between each protein-coding gene and *Npm1*, and selected genes with correlation coefficients > 0.6 (738 genes). Finally, we summed the normalized UMI counts of these genes to calculate a *Npm1* signature for individual cells. The resulting *Npm1* signatures are subsetted in four plots, from left to right: by sci-RNA-seq3 experiment, embryo harvest date, litter of embryos, or shipment batch. **b,** Same as panel a, but further stratified by the top 10 abundant major cell clusters.

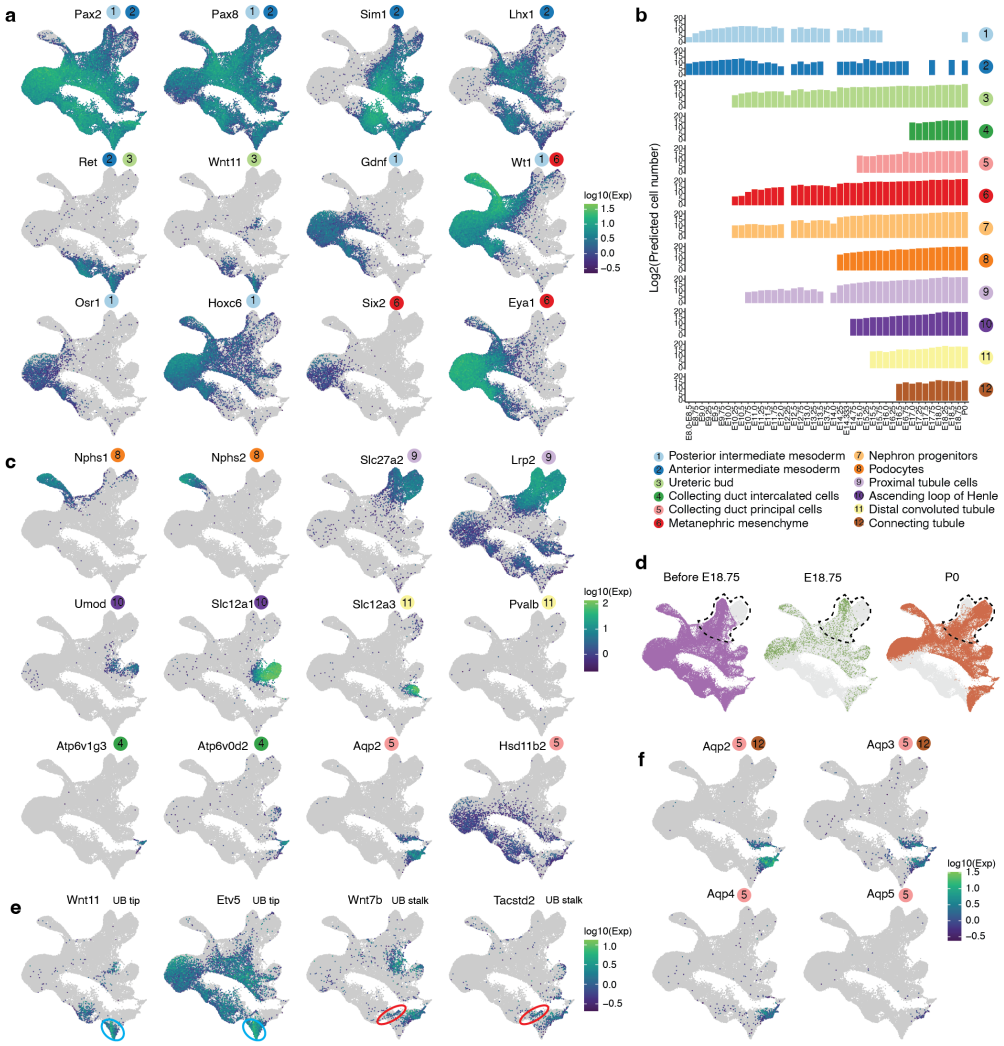

Figure 3.17: **Supplementary Figure 10. Transcriptional heterogeneity in renal development.** **a,** The same UMAP as in **Fig. 3a**, colored by expression of marker genes which appear specific to anterior intermediate mesoderm, posterior intermediate mesoderm, ureteric bud or metanephric mesenchyme. **b,** The predicted absolute number (log2 scale) of cells of each renal cell type at each timepoint. **c,** The same UMAP as in **Fig. 3a**, colored by expression of marker genes which appear specific to podocytes, proximal tubule cells, ascending loop of Henle, distal convoluted tubule, collecting duct intercalated cells or collecting duct principal cells. **d,** The same UMAP as **Fig. 3a** is shown three times, with colors highlighting cells from before E18.75 (left), E18.75 (middle), or P0 (right). **e,** The same UMAP as in **Fig. 3a**, colored by expression of marker genes which appear specific to the ureteric bud tip or stalk [178]. **f,** The same UMAP as in **Fig. 3a**, colored by expression of marker genes which appear specific to connecting tubule cells or collecting duct cells [221].

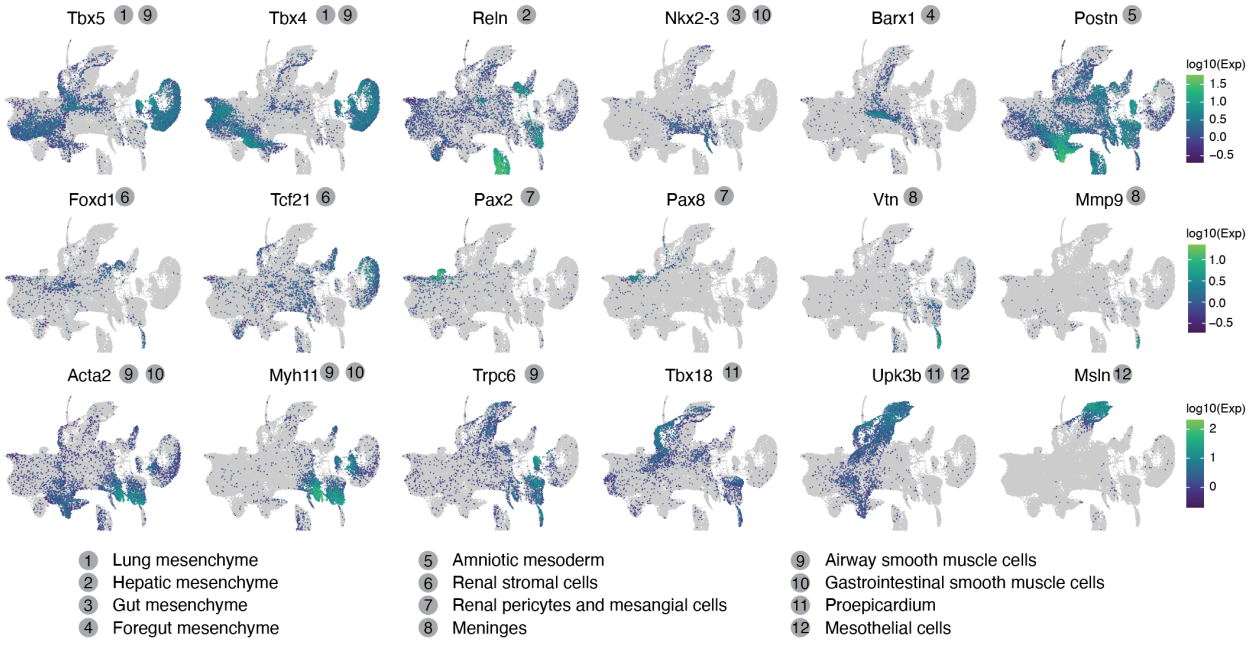

Figure 3.18: **Supplementary Figure 11. Transcriptional heterogeneity in mesenchyme.** The same UMAP as in **Fig. 3f**, colored by expression of marker genes which appear specific to lung mesenchyme (*Tbx5*+, *Tbx4*+), hepatic mesenchyme (*Reln*+), gut mesenchyme (*Nkx2-3*+), foregut mesenchyme (*Barx1*+), amniotic mesoderm (*Postn*+), renal stromal cells (*Foxd1*+, *Tcf21*+), renal pericytes and mesangial cells (*Pax2*+, *Pax8*+), meninges (*Vtn*+), airway smooth muscle cells (*Trpc6*+, *Tbx5*+), gastrointestinal smooth muscle cells (*Nkx2-3*+), proepicardium or mesothelium (*Msln*+). References for marker genes are provided in **Supplementary Table 12**.

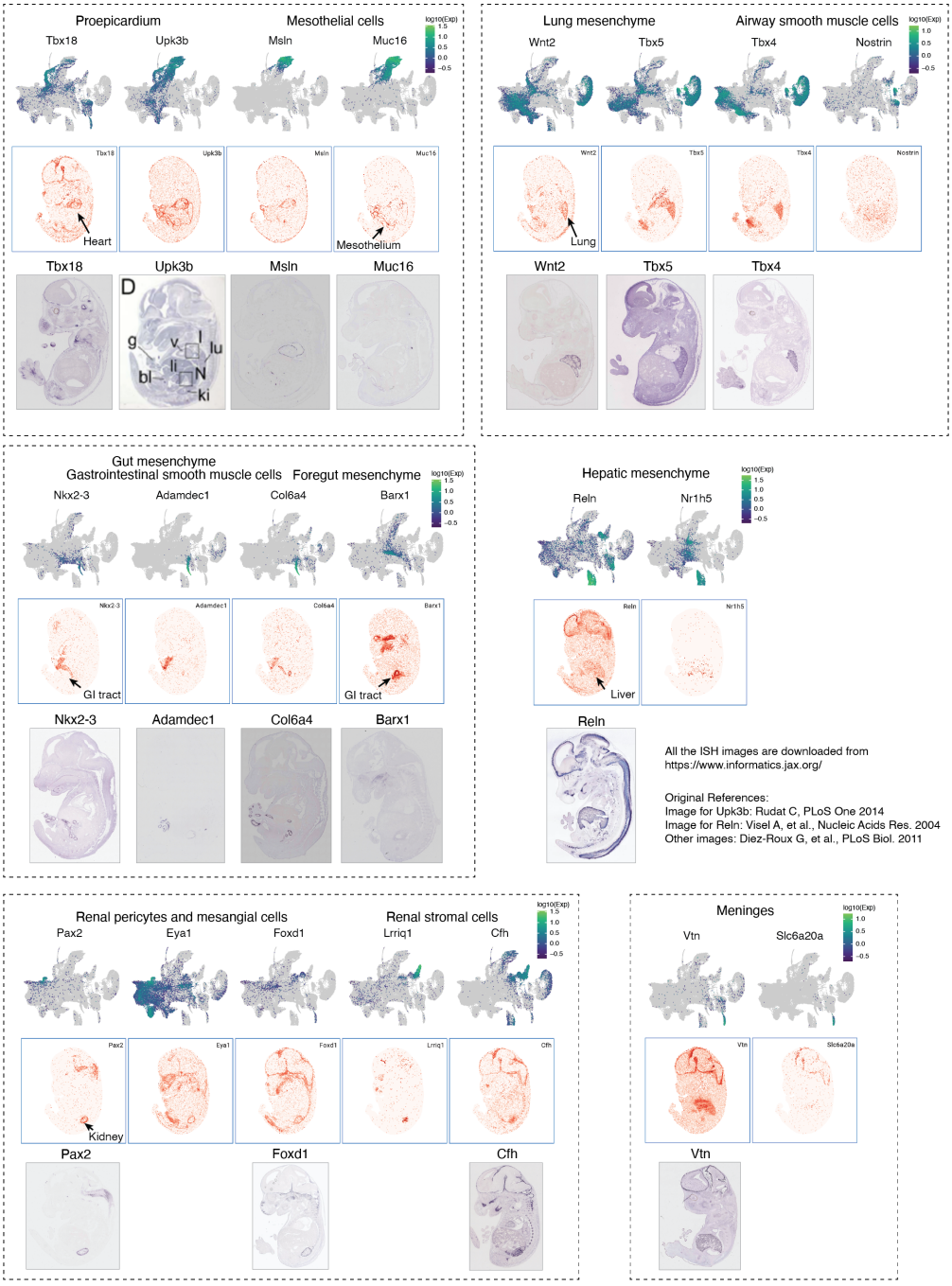

Figure 3.19: **Supplementary Figure 12. Published in situ hybridization (ISH) images support our annotations of lateral plate and intermediate mesoderm derivatives.** In each subpanel, three rows are shown for one or two lateral plate and intermediate mesoderm derivative cell types. Notably, each of these cell types was annotated based on spatial mapping analysis, as shown in **Fig. 3g**. Top: The same UMAP as in **Fig. 3f**, colored by gene expression of marker genes which appear specific to the given cell type. Middle: Virtual *in situ* hybridization (ISH) images of individual genes from one selected section (E1S1) from E14.5 of the Mosta data. Bottom: In situ hybridization (ISH) images of individual genes were obtained from the Jackson Laboratory Mouse Genome Informatics (MGI) website.

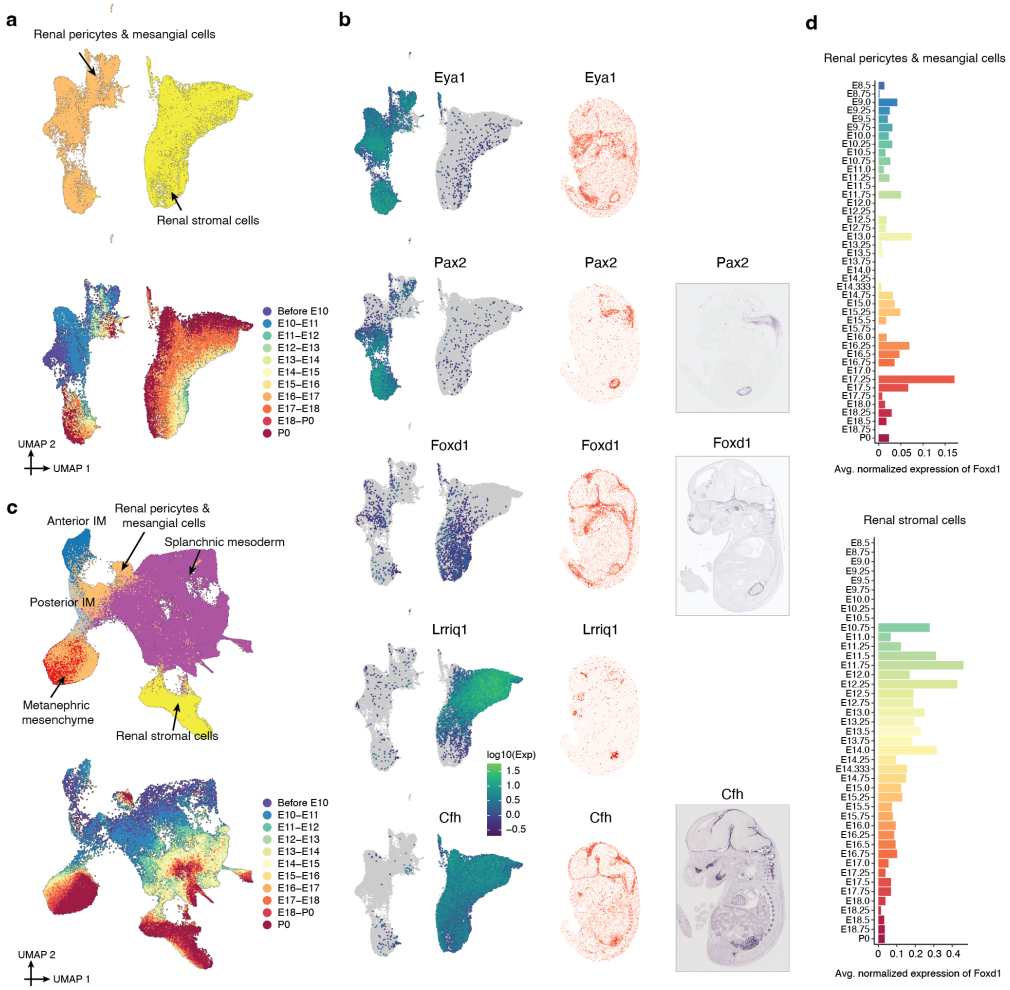

Figure 3.20: **Supplementary Figure 13. Assessing the potential origins of LPM subsets annotated as renal pericytes and mesangial cells and renal stromal cells. a,** Re-embedded 2D UMAP of 39,468 cells from renal pericytes and mesangial cells and renal stromal cells. **b,** Left: The same UMAP as in panel a, colored by gene expression of marker genes which appear specific to renal pericytes and mesangial cells (*Pax2*+, *Eya1*+) or renal stromal cells (*Lrriq1*+, *Cfh*+). *Foxd1* is expressed in a subset of both cell types. Middle: Virtual *in situ* hybridization (ISH) images of individual genes from one selected section (E1S1) from E14.5 of the Mosta data. Right: *In situ* hybridization (ISH) images of individual genes were obtained from the Jackson Laboratory Mouse Genome Informatics (MGI) website. **c,** Re-embedded 2D UMAP of 206,908 cells from renal pericytes and mesangial cells, renal stromal cells, anterior intermediate mesoderm, posterior intermediate mesoderm, metanephric mesenchyme, and splanchnic mesoderm. **d,** The average normalized expression of *Foxd1* over time is shown for renal pericytes and mesangial cells (top) and renal stromal cells (bottom).

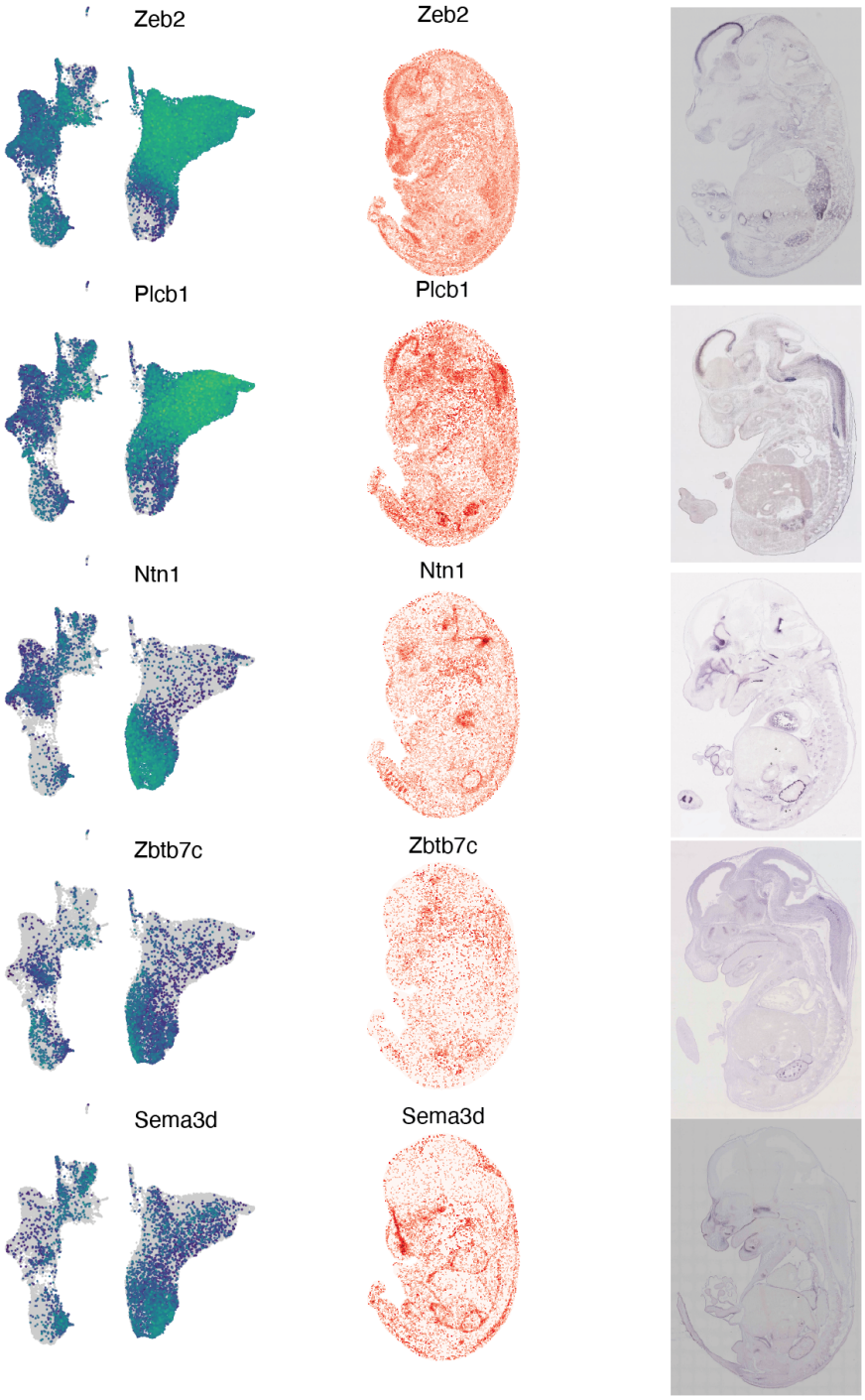

Figure 3.21: **Supplementary Figure 14. Spatial heterogeneity within the renal stromal cells.** Left: The same UMAP as in **Supplementary Fig. 13a**, colored by gene expression of marker genes which appear specific to two subsets of renal stromal cells: medullary renal stromal cells (*Zeb2*+, *Plcb1*+) and cortical renal stromal cells (*Ntn1*+, *Zbtb7c*+, *Sema3d*+), respectively. Middle: Virtual *in situ* hybridization (ISH) images of individual genes from one selected section (E1S1) from E14.5 of the Mosta data. Right: *In situ* hybridization (ISH) images of individual genes were obtained from the Jackson Laboratory Mouse Genome Informatics (MGI) website.

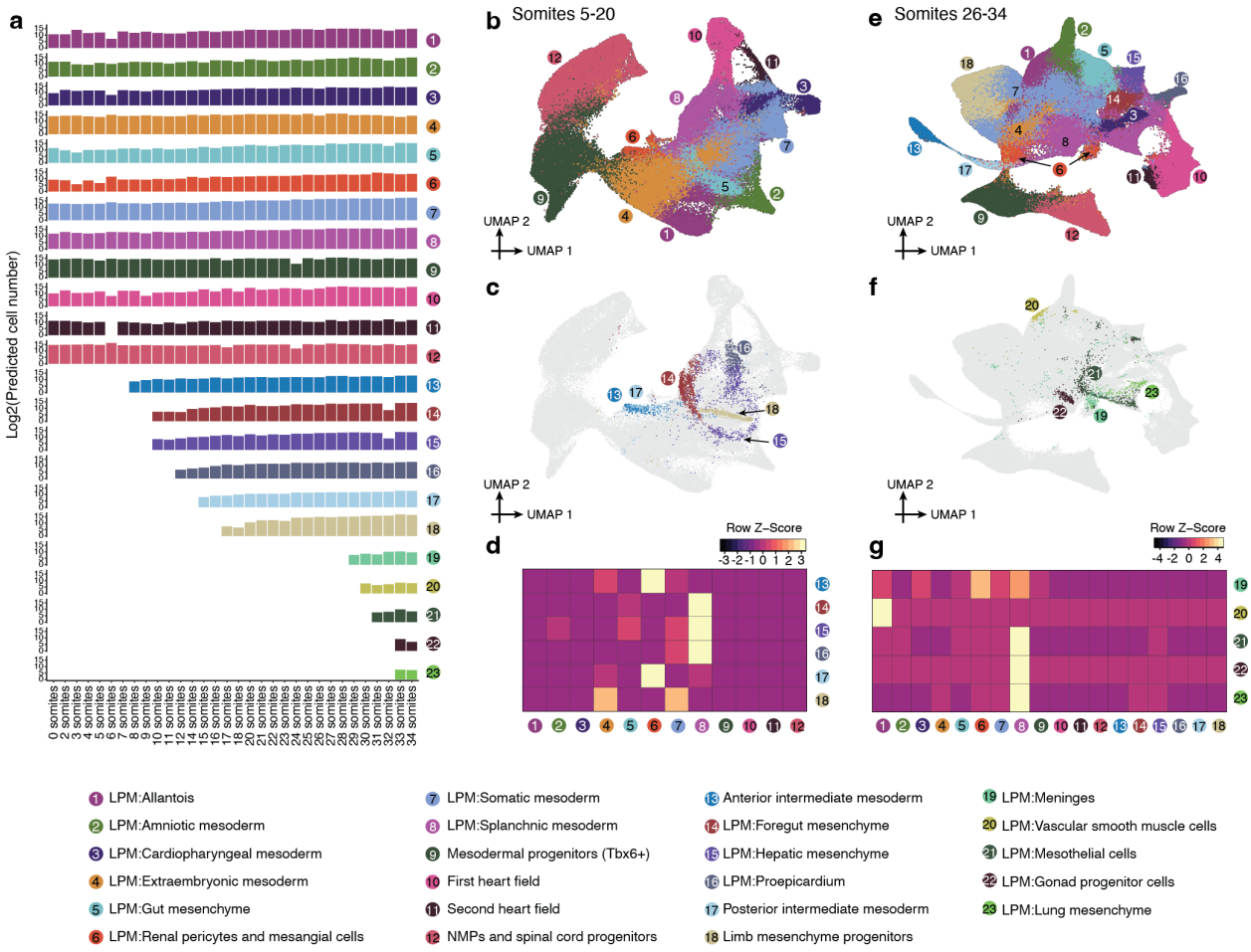

Figure 3.22: **Supplementary Figure 15. The emergence of mesenchymal subtypes from the patterned mesoderm. a,** The predicted absolute number (log2 scale) of cells of each mesoderm cell type at each somite count. **b,** Re-embedded 2D UMAP of 110,753 cells from the selected cell types of mesoderm (clusters 1-12 as listed in panel a) from 5-20 somite stage embryos. **c,** The same UMAP as in panel b, but with inferred progenitor cells colored by derivative cell type with the highest mutual nearest neighbors (MNN) pairing score. **d,** Normalized MNN pairing score between mesodermal territories (column) and their inferred derivative cell types (row) from 5-20 somite stage embryos. **e,** Re-embedded 2D UMAP of 275,000 cells from the selected cell types of mesoderm (clusters 1-12 as listed in panel a) from 26-34 somite stage embryos. **f,** The same UMAP as in panel e, but with inferred progenitor cells colored by derivative cell type with the highest MNN pairing score. **g,** Normalized MNN pairing score between mesodermal territories (column) and their inferred derivative cell types (row) from 26-34 somite stage embryos.

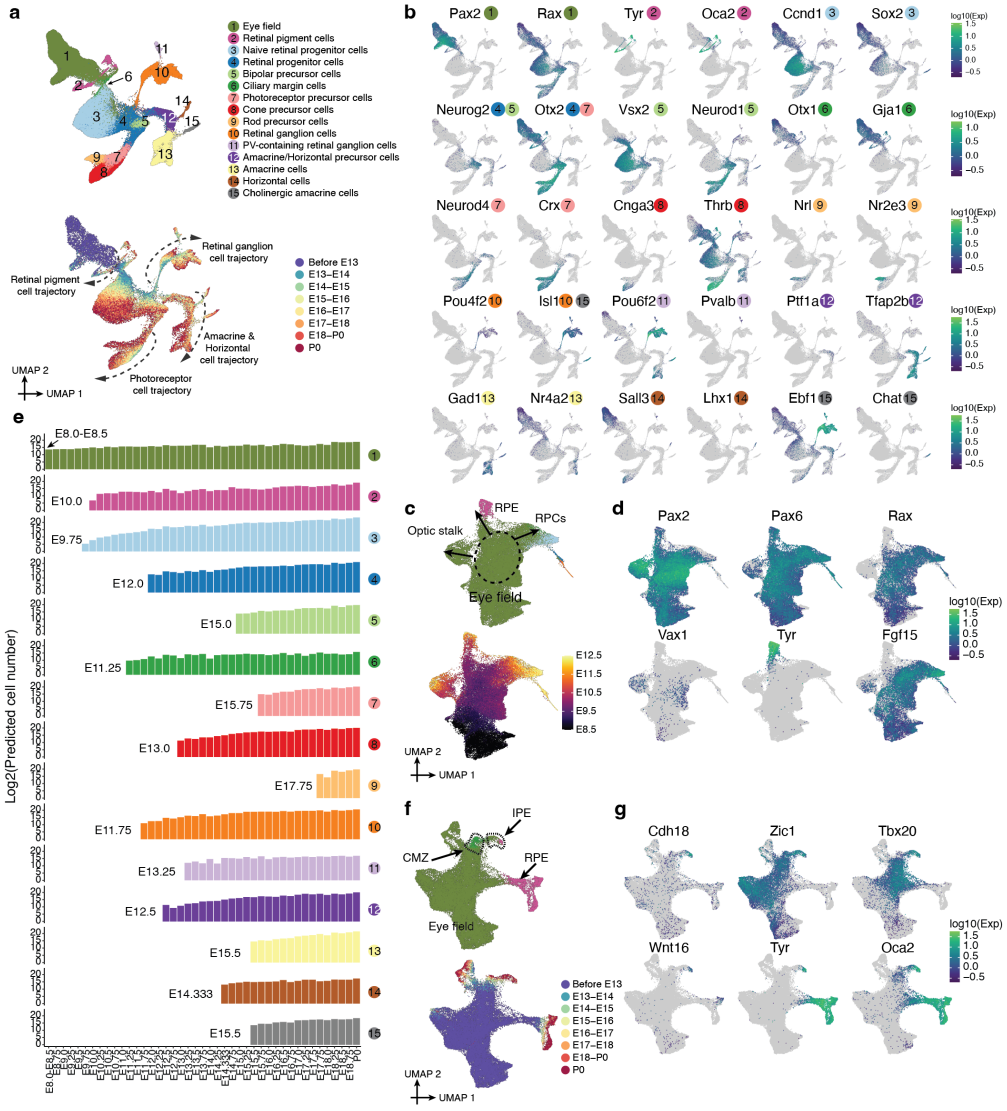

Figure 3.23: **Supplementary Figure 16. The timing and trajectories of retinal development.** **a,** Re-embedded 2D UMAP of 160,834 cells corresponding to the retinal development from E8 to P0. **b,** The same UMAP as in panel a, colored by gene expression of marker genes for each annotated retinal cell type. **c,** Re-embedded 2D UMAP of the subset of cells in panel a from stages before E12.5. **d,** The same UMAP as in panel c, colored by gene expression of markers of retinal progenitor cells RPCs (*Pax2*+, *Pax6*+, *Rax*+, *Fgf15*+)[222], RPE (*Tyr*+)[223], and the optic stalk (*Pax2*+, *Vax1*+, *Rax*-)[224]. **e,** The predicted absolute number (log2 scale) of cells of each retinal cell type at each timepoint. **f,** Re-embedded 2D UMAP of a subset of cells in panel a corresponding to eye field, RPE and CMZ. **g,** The same UMAP as in panel f, colored by gene expression of marker genes for IPE[225] or pigment epithelium more generally (*Tyr*, *Oca2*).

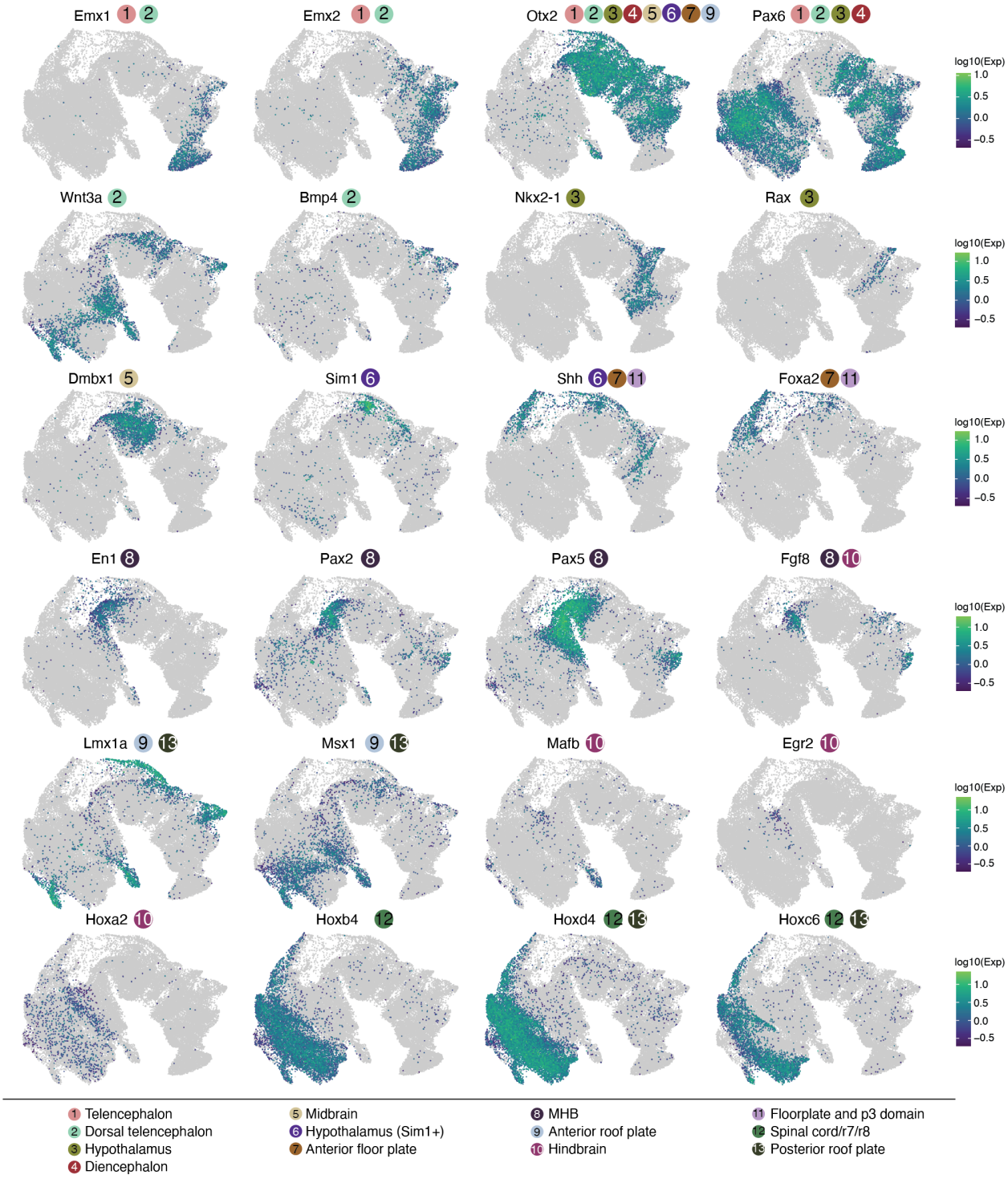

Figure 3.24: **Supplementary Fig 17. Marker gene expression for different neuroectodermal territories.** The same UMAP as in **Fig. 5a**, colored by gene expression of marker genes for each neuroectodermal territory. References for marker genes are provided in **Supplementary Table 5**.

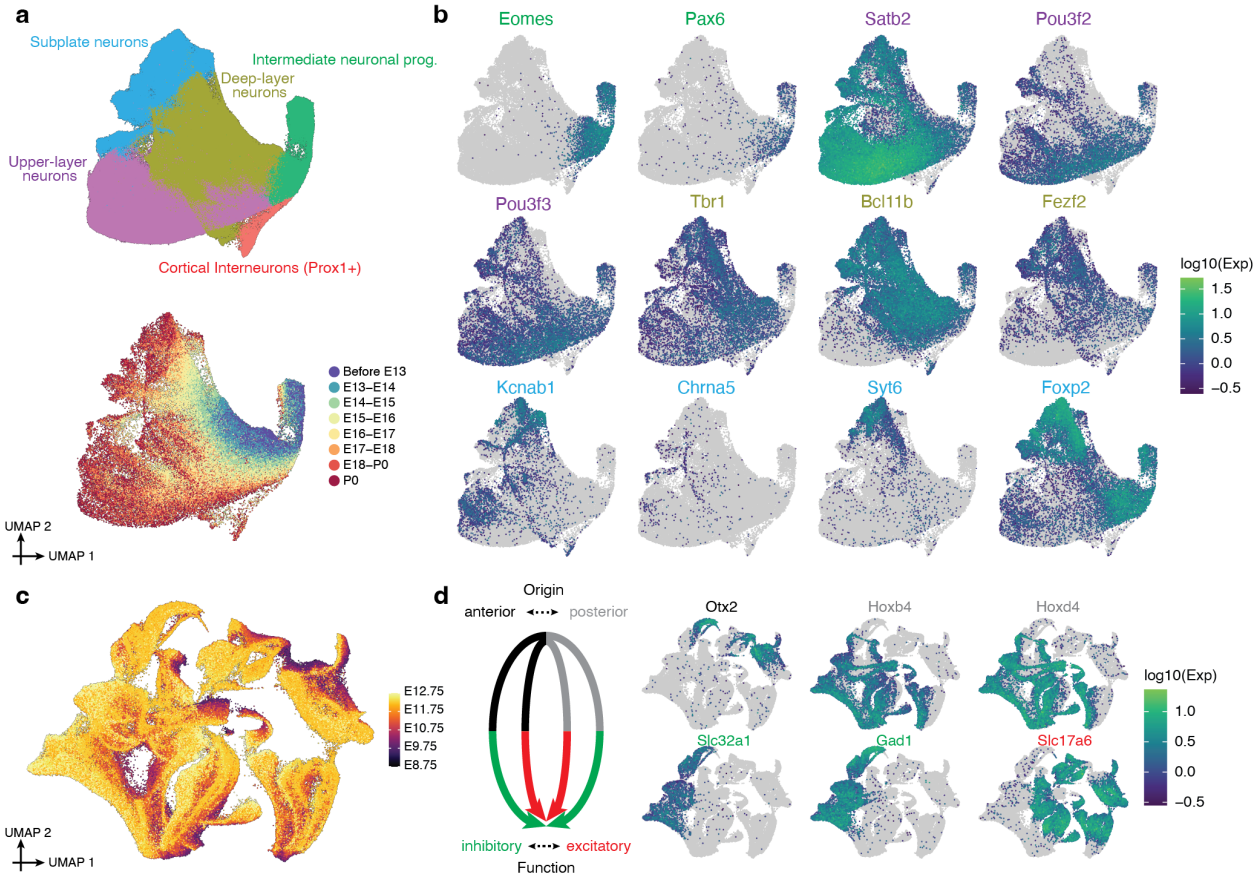

Figure 3.25: **Supplementary Figure 18. Subtypes of intermediate neuronal progenitors, glutamatergic and GABAergic neurons. a,** Re-embedded 2D UMAP of 628,251 cells within the intermediate neuronal progenitors major cell cluster, colored by either cell type (top) or developmental stage (bottom). **b,** The same UMAP as in panel a, colored by gene expression of marker genes which appear specific to intermediate neuronal progenitors (*Eomes*+, *Pax6*+), upper-layer neurons (*Satb2*+, *Pou3f2*+, *Pou3f3*+), deep-layer neurons (*Tbr1*+, *Bcl11b*+, *Fezf2*+), or subplate neurons (*Kcnab1*+, *Chrna5*+, *Syt6*+, *Foxp2*+). **c,** The same UMAP as in **Fig. 5e**, with cells colored by timepoints. **d,** Left: Neuronal subtypes shown in **Fig. 5e** originate from anterior vs. posterior of neuroectoderm, and then subsequently display inhibitory vs. excitatory functions after differentiation. Right: The same UMAP as in **Fig. 5e**, colored by gene expression of marker genes which appear specific to anterior (*Otx2*+)[226] vs. posterior (*Hoxb4*+, *Hoxd4*+)[227] origins, or inhibitory (*Slc32a1*+, *Gad1*+)[228] vs. excitatory (*Slc17a6*+)[229] functions.

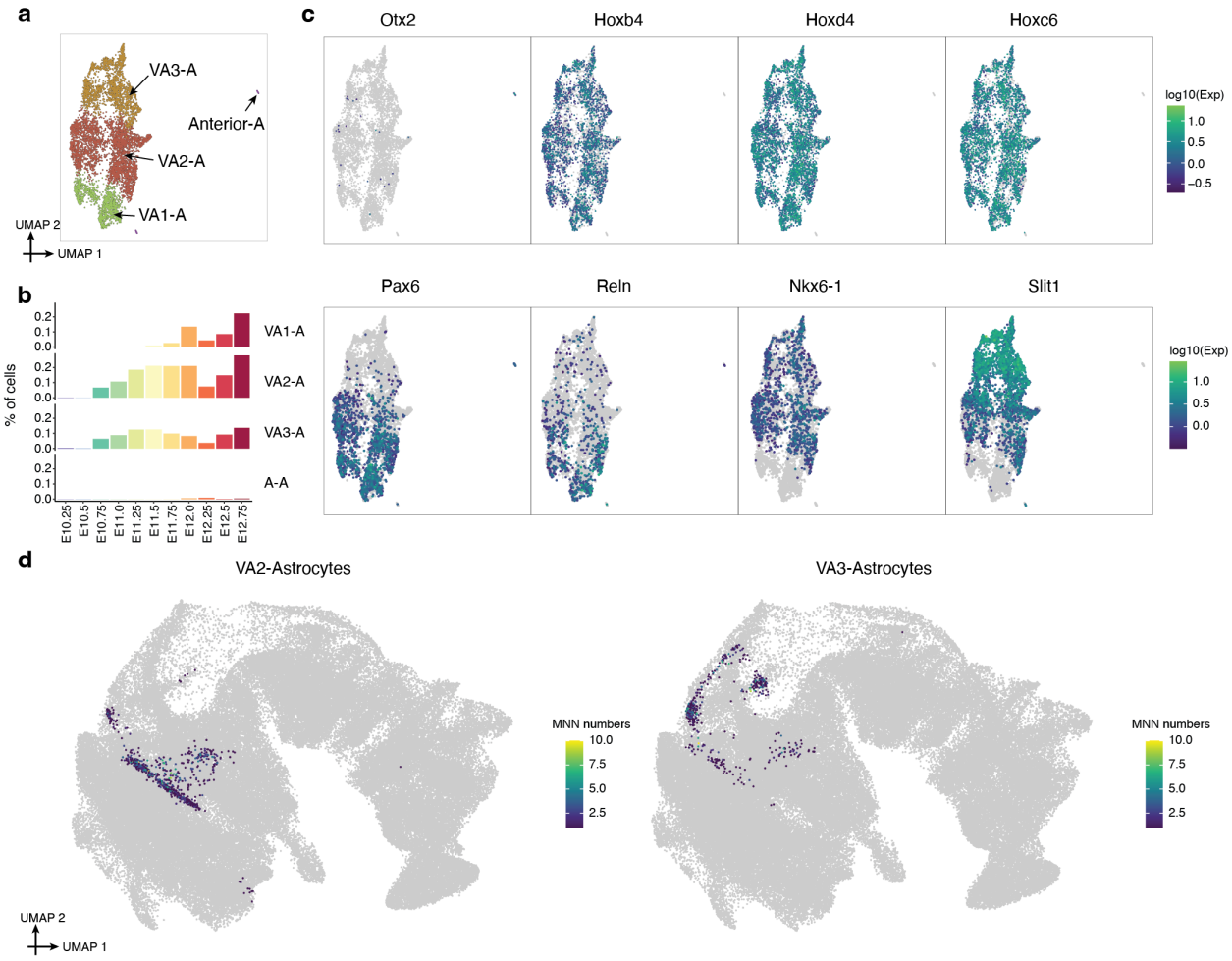

Figure 3.26: **Supplementary Figure 19. Subtypes of early astrocytes and their inferred progenitors. a,** Re-embedded 2D UMAP of 5,928 cells within the astrocytes from stages < E13. **b,** Composition of embryos from each 6-hr bin by different subpopulations of astrocytes. **c,** The same UMAP as in panel a, colored by gene expression of marker genes which appear specific to anterior (*Otx2*+) or posterior (*Hoxb4*+, *Hoxd4*, *Hoxc6*+) astrocytes, VA1-astrocytes (*Pax6*+, *Reln*+), VA2-astrocytes (*Pax6*+, *Reln*+, *Nkx6-1*+, *Slit1*+), and VA3-astrocytes (*Nkx6-1*+, *Slit1*+)[193]. **d,** The same UMAP of the patterned neuroectoderm as in **Fig. 5a**, with inferred progenitor cells of astrocytes colored by the frequency that has been identified as a MNN with either VA2-astrocytes (left) or VA3-astrocytes (right).

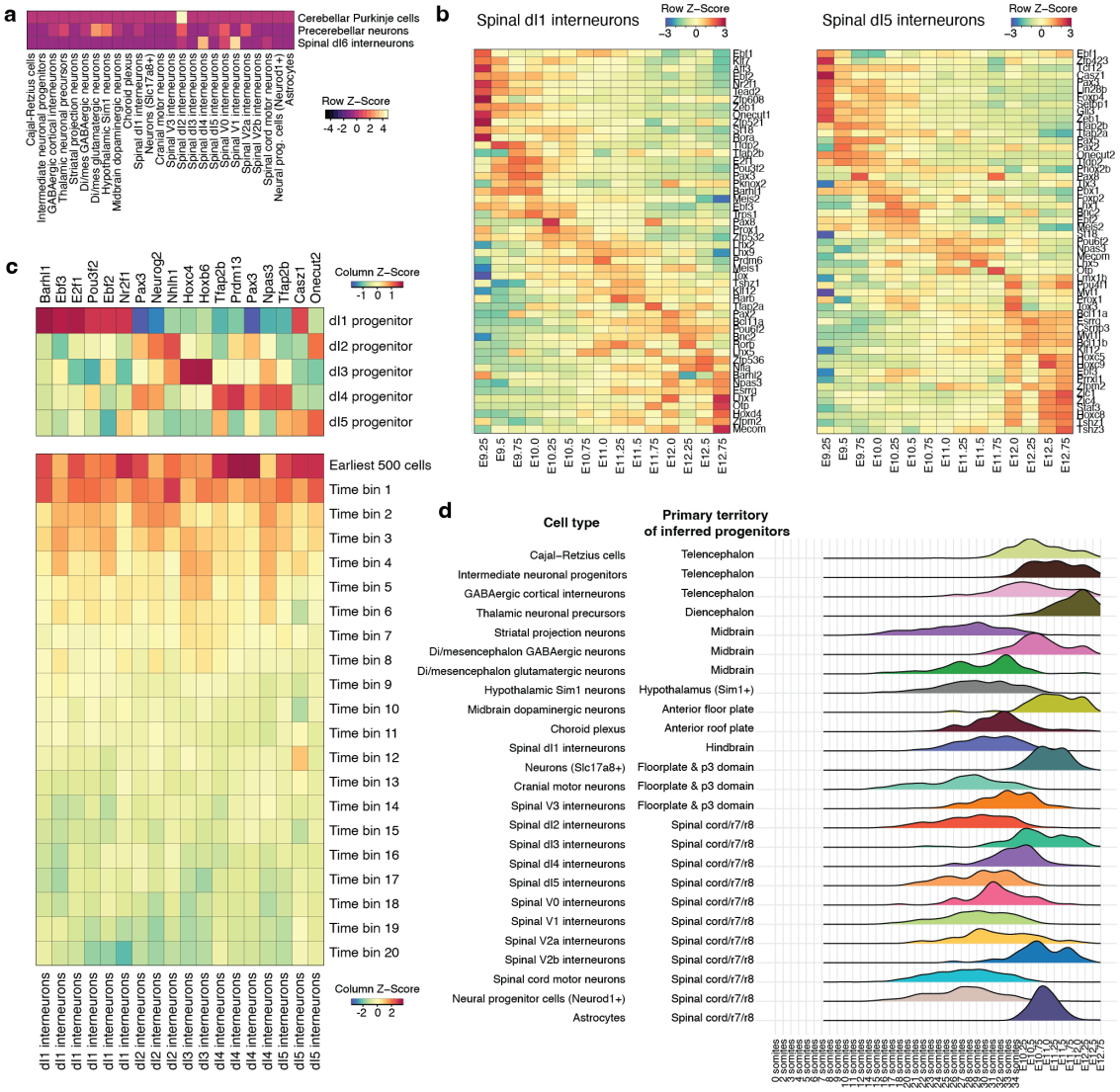

Figure 3.27: **Supplementary Figure 20. The timing of neuronal subtype differentiation from the patterned neuroectoderm.** **a,** For those three cell types (cerebellar Purkinje cells, precerebellar neurons, spinal dI6 interneurons) which were excluded in **Fig. 5d-e** due to having fewer than 50 MNN pairs, we performed a recursive mapping. **b,** Gene expression across timepoints, for the specific TF markers of spinal *dI1* (left) or spinal *dI5* (right) interneurons. **c,** Top: gene expression for 18 selected TFs, across progenitor cells of *dI1-5* from the neuroectodermal territories. Bottom, gene expression for 18 selected TFs across 21 time bins for *dl1-5* spinal interneurons in which the TF has been nominated as marker TF. **d,** For each neuronal subtype in **Fig. 5i-j**, we selected the annotation in the patterned neuroectoderm to which the most inferred progenitors had been assigned, and plotted the distribution of timepoints for that subset of inferred progenitors.

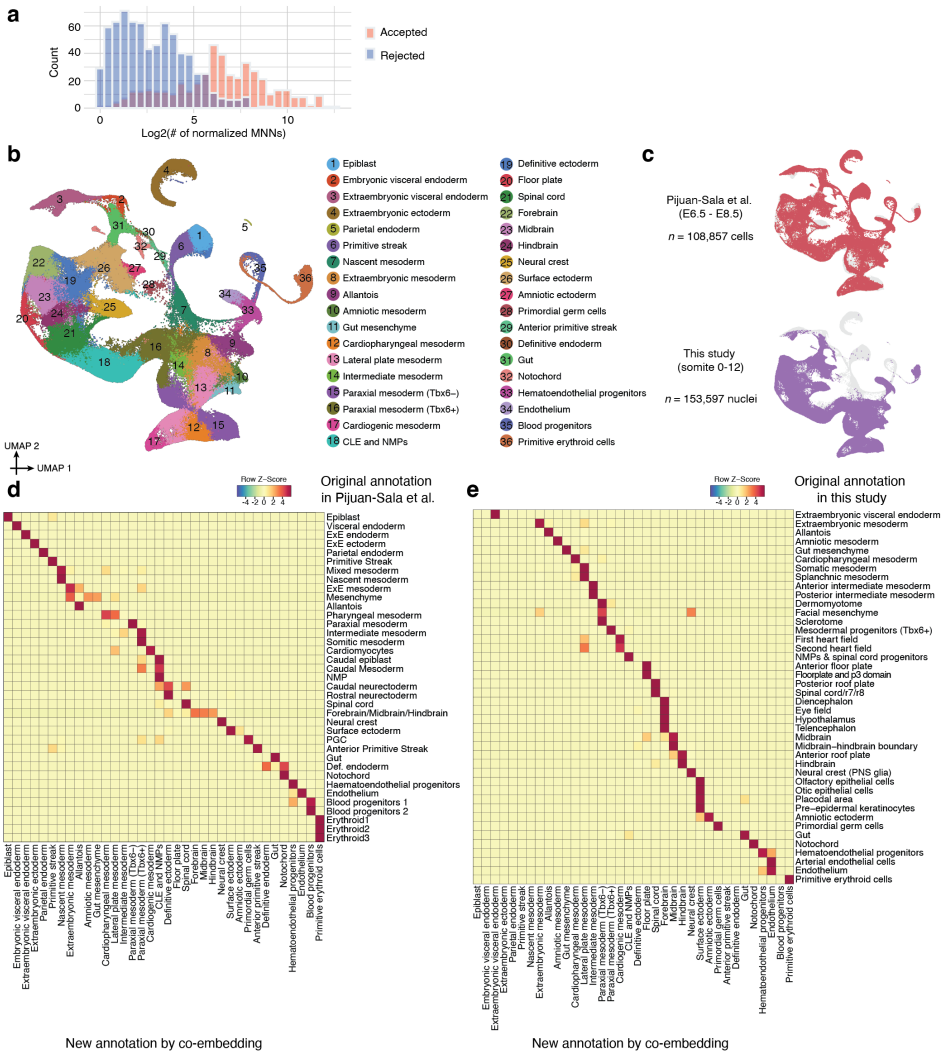

Figure 3.28: **Supplementary Figure 21. Integration of scRNA-seq profiles from gastrulation and early somitogenesis to identify equivalent cell type nodes across datasets generated by distinct technologies. a,** 1,155 edges with the number of normalized MNNs > 1 were manually reviewed for biological plausibility. Histogram of edges that were accepted or rejected as a function of normalized MNN score. **b,** 2D UMAP visualization of co-embedded cells, derived both from a gastrulation dataset based on cells from E6.5 to E8.5 generated on the 10x Genomics platform[13] (n = 108,857 cells) and the earliest 1% of this dataset (0-12 somite stage embryos) generated by sci-RNA-seq3 (n = 153,597 nuclei), after batch correction[51]. **c,** The same UMAP as in panel b is shown twice, with colors highlighting cells/nuclei from Pijuan-Sala's dataset[13] (top) or early somitogenesis[152] (bottom). **d,** For cells from the original Pijuan-Sala's dataset[13], we quantify and display the overlap between the original annotations and the new annotations shown in panel b. **e,** For nuclei from the early somitogenesis embryos[152], we quantify and display the overlap between the original annotations and the new annotations shown in panel b.

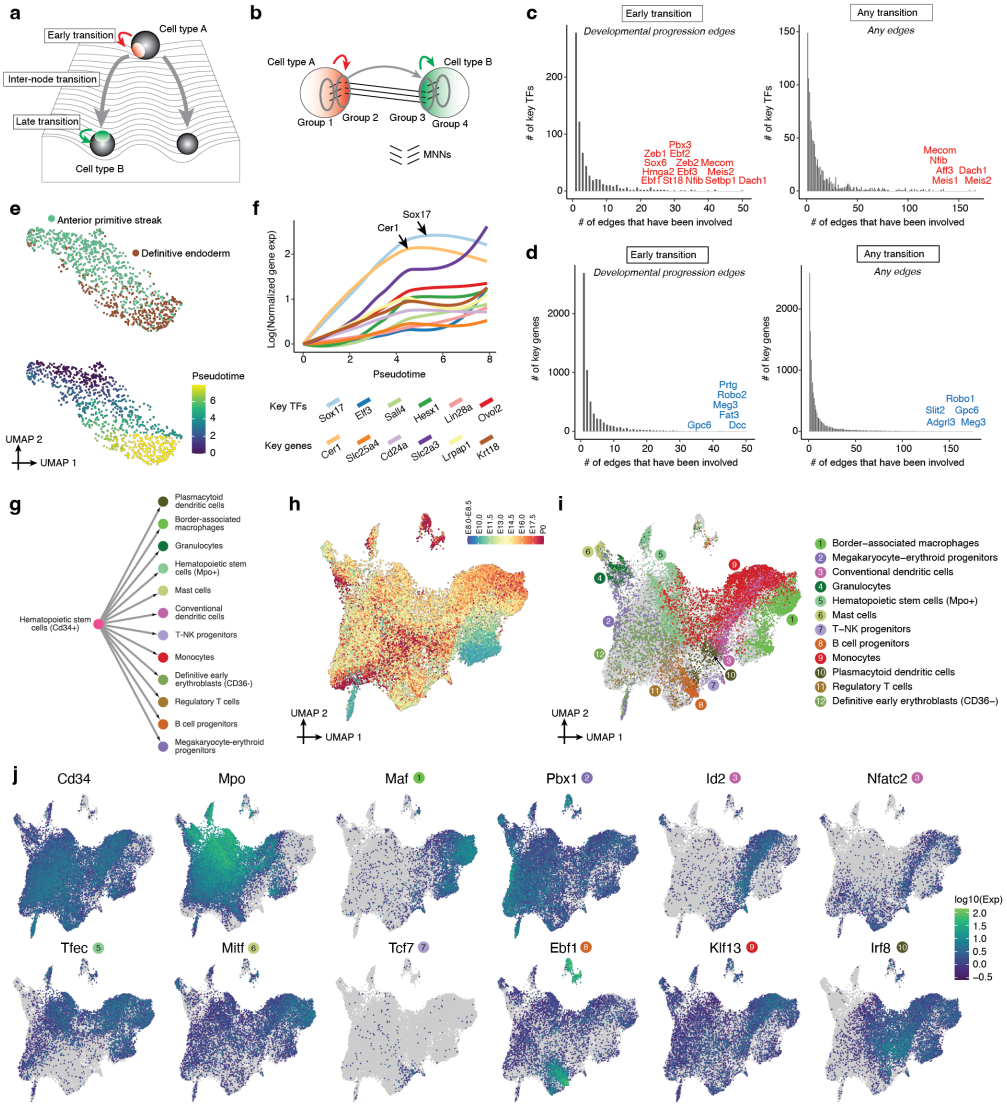

Figure 3.29: **Supplementary Fig 22. Systematic nomination of TFs and other genes for cell type specification. a,** A Waddington landscape cartoon illustrating how a cell type transition might be broken into three phases. **b,** Given a directional edge between two nodes, we identified the inter-node and intra-node MNNs. **c,** Histograms of the number of edges in which TFs are differentially expressed. **d,** Same as panel c, but for all genes rather than only TFs. **e,** Re-embedded 2D UMAP of 988 cells participating in groups 1-4 of the transition from anterior primitive streak to definitive endoderm. **f,** For cells in panel e, normalized gene expression of selected genes are plotted as a function of estimated pseudotime. **g,** A sub-graph of **Fig. 6g**, including hematopoietic stem cells (*Cd34+*) and 12 cell type nodes which appear derived from it. **h,** Re-embedded 2D UMAP of 37,750 cells from hematopoietic stem cells (*Cd34+*), colored by developmental stage. **i,** The same UMAP as in panel h, but with inferred progenitor cells. **j,** The same UMAP as in panel h, colored by gene expression of selected top key TFs which were upregulated during the "early transition" for each derivative.

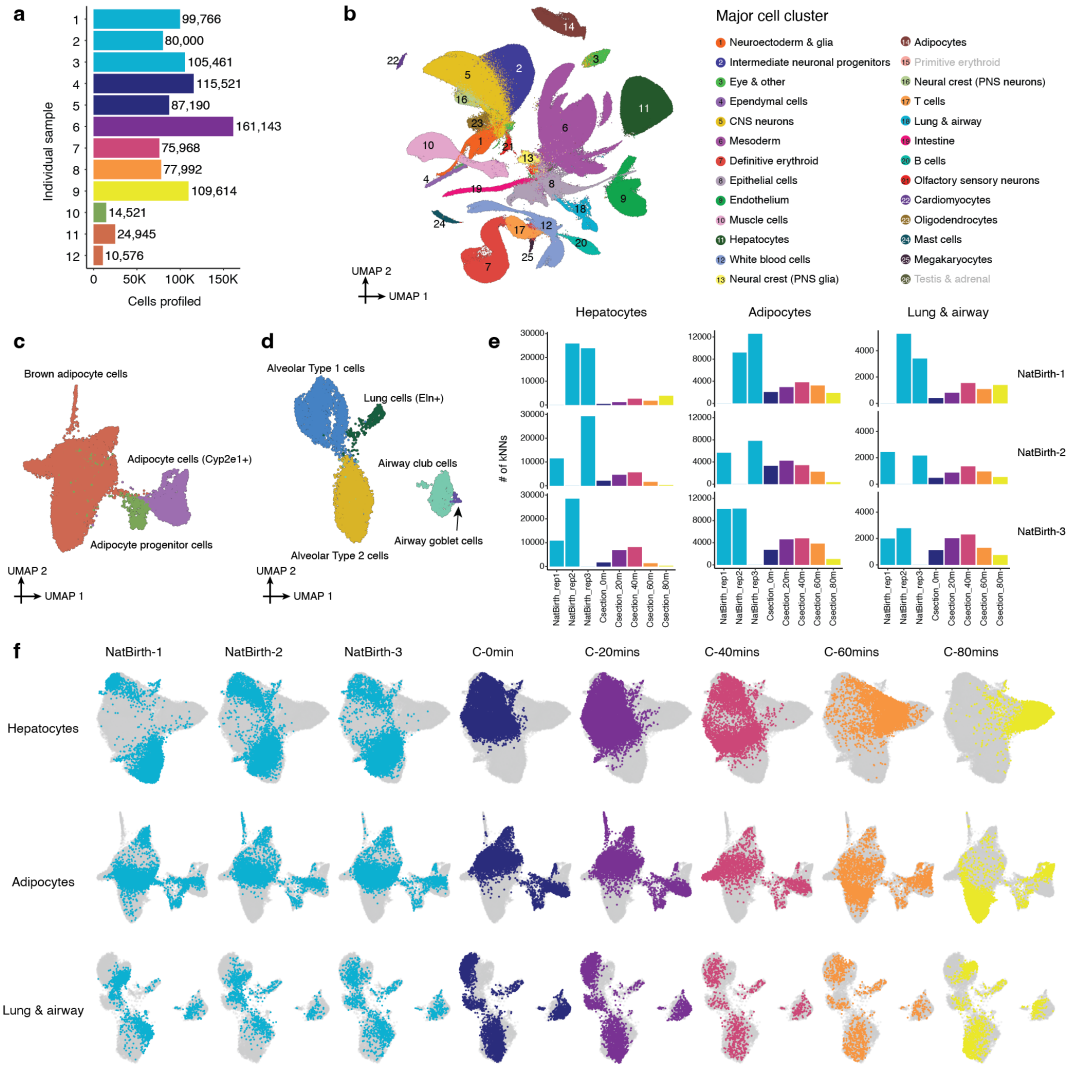

Figure 3.30: **Supplementary Figure 23. Rapid shifts in transcriptional state occur in a restricted subset of cell types upon birth, and differ between vaginally and C-section delivered pups. a,** The number of nuclei profiled for each animal shown in **Fig. 7c**. **b,** 2D UMAP visualization of the birth-series dataset (n = 962,697 cells). **c,** Re-embedded 2D UMAP of 19,696 cells of the adipocyte major cell cluster. **d,** Re-embedded 2D UMAP of 7,986 cells of the lung and airway major cell cluster. **e,** For these three major cell clusters, we co-embedded cells from three vaginally delivered pups and six pups delivered by C-section. For cells from each of the three vaginally delivered pups, we calculated the number of their 10 nearest neighbors from other samples. **f,** Re-embedded 2D UMAP of cells from these three major cell clusters, based on cells from three vaginally delivered pups and six pups delivered by C-section.

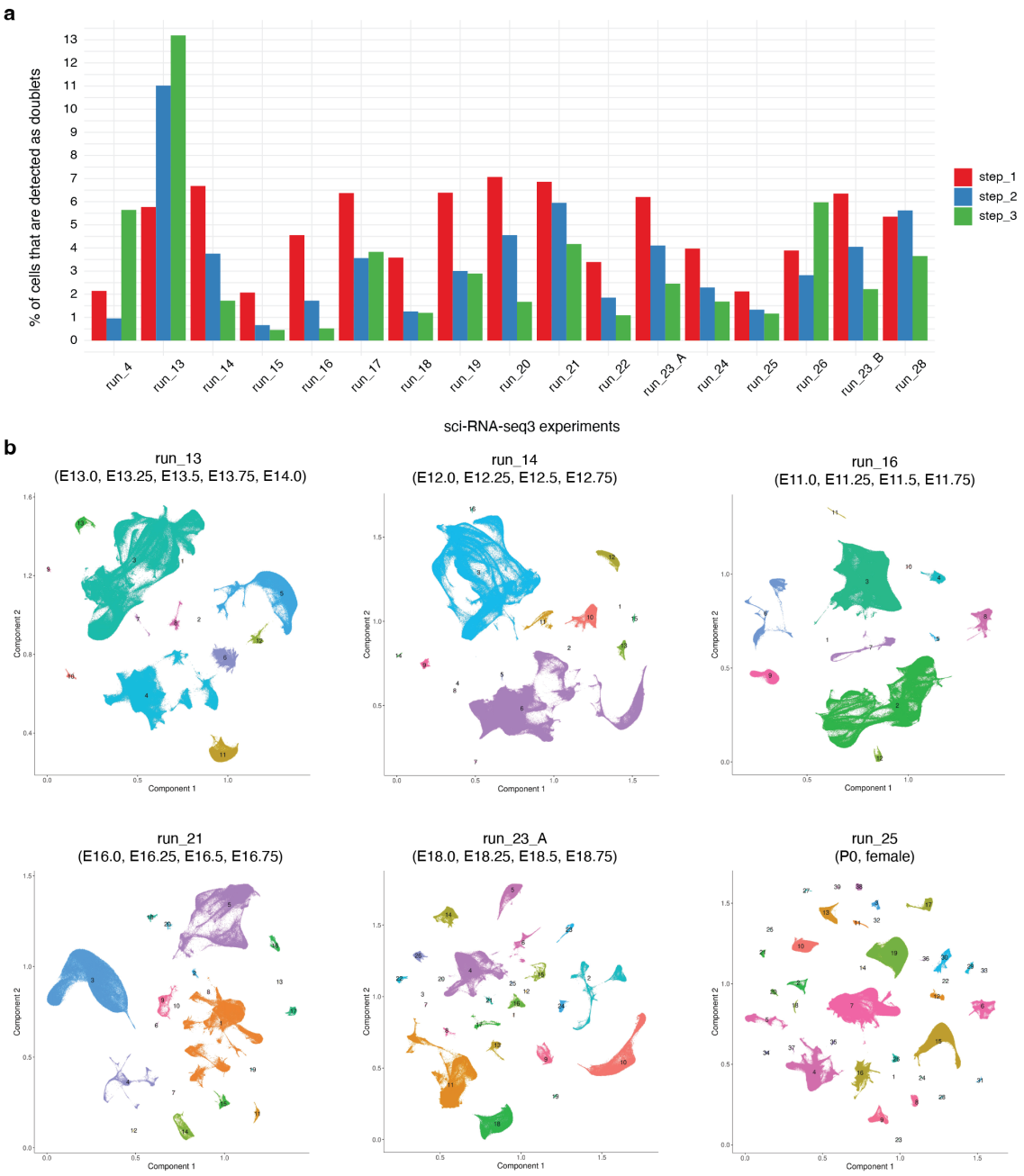

Figure 3.31: **Supplementary Figure 24. Three-step doublet detection workflow for sci-RNA-seq3 experiments. a,** We performed three steps to detect and remove potential doublets from each single sci-RNA-seq3 experiment (**Methods**). The percentage of cells detected and removed as doublets by each of the three steps in individual sci-RNA-seq3 experiments is shown. **b,** The labeled cell partitions for each of six selected experiments are shown, after removing doublets from the first two steps.

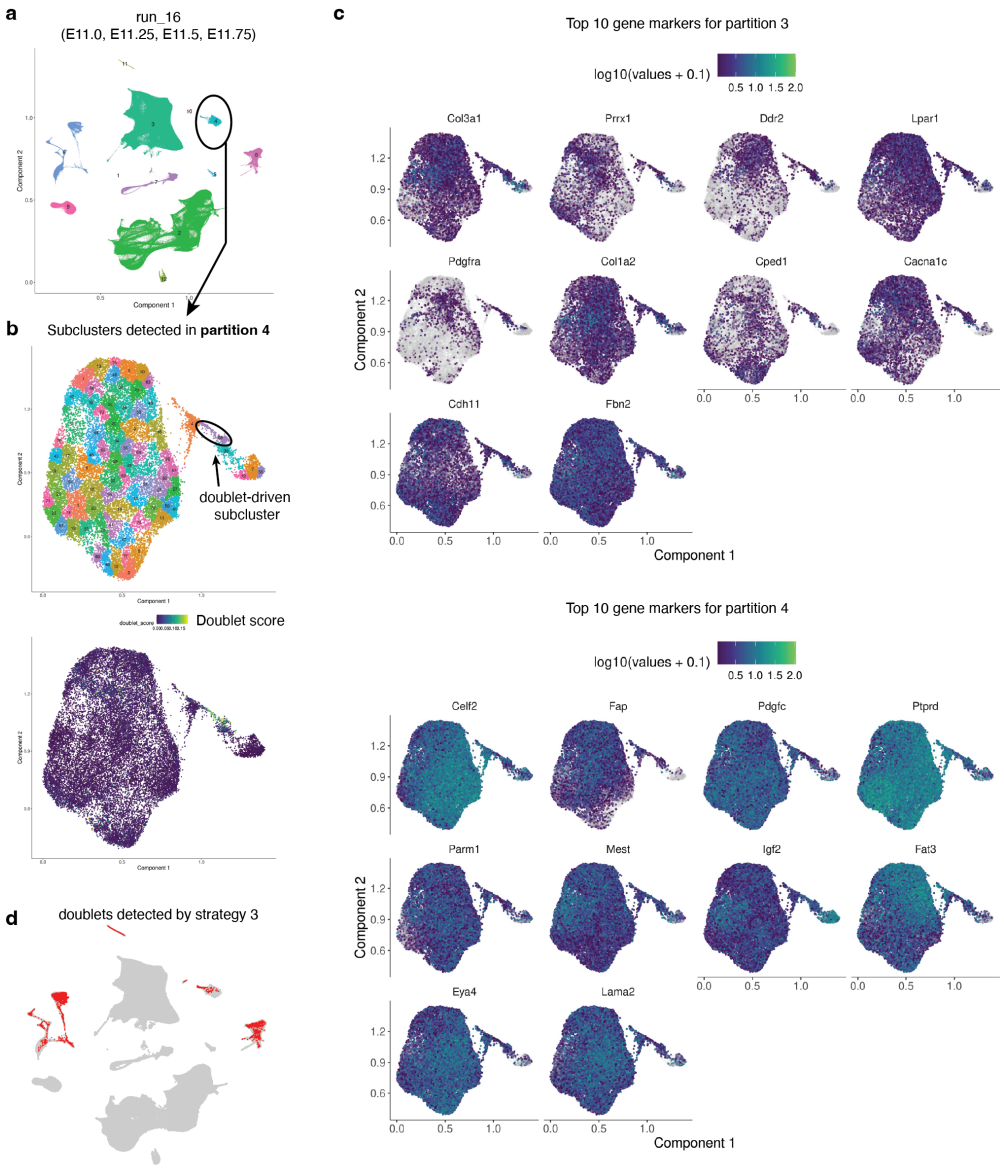

Figure 3.32: **Supplementary Figure 25. Example of detection of doublet-driven subclusters via step 3. a,** Re-embedded 2D UMAP of 986,264 cells from experiment run 16, after removing doublets detected in the first two steps. Cells were colored by each of the 12 partitions detected by the partitionCells function implemented in Monocle/3-alpha. **b,** Re-embedded 2D UMAP of cells from partition 4, with cells colored by subclusters. The same UMAP is shown below, with cells colored by doublet score calculated by Scublet. **c,** The same UMAP as in panel b, colored by the normalized gene expression of the top 10 differentially expressed genes in either partition 3 (top) or partition 4 (bottom). **d,** The same UMAP as in panel a, highlighted by doublets detected in step 3 (red).

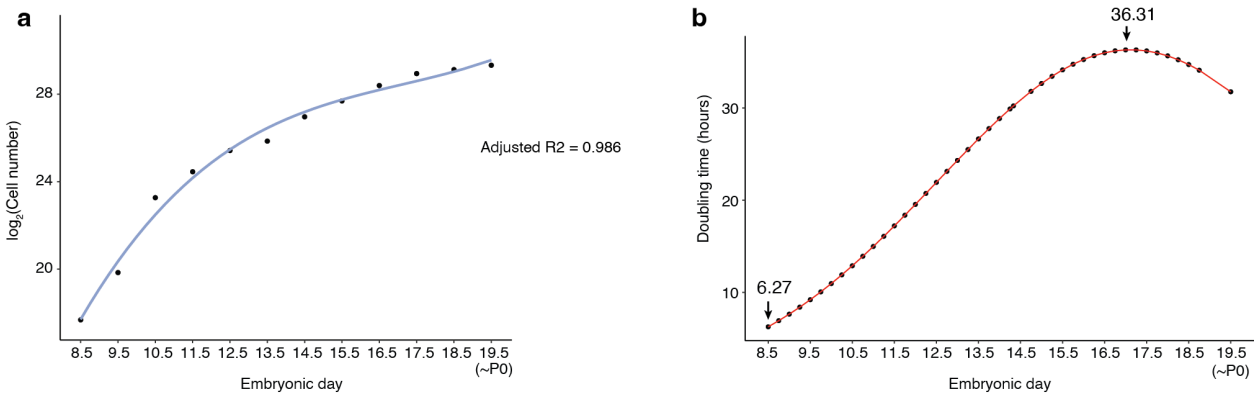

Figure 3.33: **Supplementary Figure 26. Quantitatively estimating cell number for individual mouse embryos as a function of developmental stage. a,** Based on the experimentally estimated cell numbers of the 12 embryos (ranging from E8.5 to P0), we applied polynomial regression (degree = 3) to fix a curve across embryos between the embryonic day and log2-scaled cell number. P0 was treated as E19.5 in the model. **b,** The estimated "doubling time" of the total cell number in a whole mouse embryo are plotted as a function of timepoints. The timepoints with the longest (E17.0) and shortest (E8.5) estimated "doubling times" are highlighted.

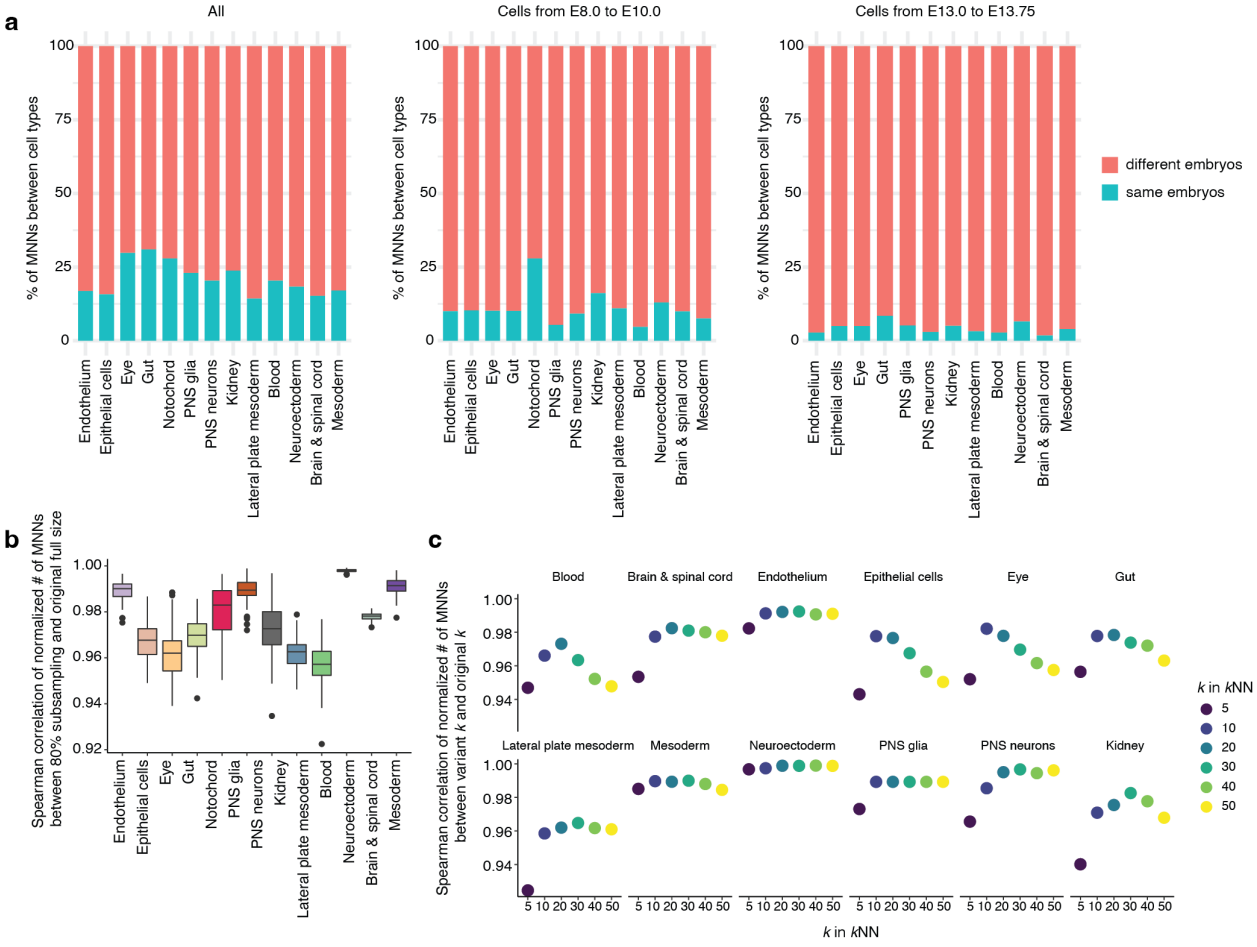

Figure 3.34: **Supplementary Figure 27. The MNN approach used for graph construction is robust to subsampling and choice of the k parameter. a,** The percentage of MNNs between different cell types, from the same embryo (blue) or from different embryos (red), is shown for each developmental system during organogenesis and fetal development, for all cells (left), cells from E8.0 to E10.0 (middle), or cells from E13.0 to E13.75 (right). **b,** The Spearman correlation coefficients of the normalized number of MNNs between cell types, comparing random subsampling of 80% of the cells to the full set of cells. The subsampling was repeated 100 times. The number of MNNs between cell types were normalized by the total number of possible MNNs between them. **c,** The Spearman correlation coefficients of the normalized number of MNNs between cell types, comparing various choices for k parameter (k = 5, 10, 20, 30, 40, 50) and the choice of k parameter (k = 15) when applying kNN to the developmental systems during organogenesis and fetal development. The number of MNNs between cell types were normalized by the total number of possible MNNs between them.

Chapter 4

SINGLE CELL, WHOLE EMBRYO PHENOTYPING OF MAMMALIAN DEVELOPMENTAL DISORDERS

This Chapter is adopted from published work with minimum changes:

Xingfan Huang ∗, Jana Henck ∗, **Chengxiang Qiu** ∗, Varun K. A. Sreenivasan, Saranya Balachandran, Oana V Amarie, Martin Hrabe de Angelis, Rose Behncke, Wing-Lee Chan, Alexandra Despang, Diane E. Dickel, Madeleine Duran, Annette Feuchtinger, Helmut Fuchs, Valerie Gailus-Durner, Natja Haag, Rene Hägerling, Nils Hansmeier, Friederike Hennig, Cooper Marshall, Sudha Rajderkar, Alessa Ringel, Michael Robson, Lauren Saunders, Patricia da Silva-Buttkus, Nadine Spielmann, Sanjay R. Srivatsan, Sascha Ulferts, Lars Wittler, Yiwen Zhu, Vera M. Kalscheuer, Daniel Ibrahim, Ingo Kurth, Uwe Kornak, Axel Visel, Len A. Pennacchio, David R. Beier, Cole Trapnell, Junyue Cao #, Jay Shendure #, Malte Spielmann #, "Single cell, whole embryo phenotyping of mammalian developmental disorders ", *Nature (in press)*, 2023.

∗: co-first authors; #: corresponding authors

This is the most collaborative project I have worked on during my PhD. It was a "multinational" collaboration involving people from Seattle, New York, and Germany. We met every Thursday at 8am for three years. I am deeply grateful to everyone who contributed to this project, including Jay, Malte, and Jun, who designed and supervised the project, and Fanny, Jana, and Varun, who analyzed the data and performed validation experiments. Their insights and hard work made this project possible.

Perturbation experiments and data are an important research direction for the future. They can help us understand how cellular lineages and trajectories change in response to regulators or other factors (*e.g.* the environment). This requires new technologies that can profile a sufficient number of replicates, efficiently test many candidates in an experiment, and avoid potential technical noise (*e.g.* batch effects). On the other hand, it also requires new computational methods. As I introduced in the first chapter, there are many new methods that have been developed to address this question. However, I believe there is still room for improvement. Finally, I learned that validation experiments are essential for the interpretation of single-cell data analysis results.

More formally, the author contributions are listed in the manuscript as follows: J.C., M.S. and J.S. conceptualized, supervised and funded the project. D.R.B., W.C., A.D., D.E.D., N.Haag, D.I., I.K., F.H., V.M.K., U.K., L.A.P., S.R., A.R., M.R., A.V. L.W. and Y.Z. provided mouse embryos. J.C. and J.H. extracted and fixed the nuclei from embryos and performed the sci-RNA-seq experiment. S.U., R.B., R.H., N.Hans and J.H. performed RNAscope experiment and image analysis. M.H.A., V.G-D. and H.F. supervised and coordinated the Ttc21b and Gli2 validation experiment; O.V.A. and P.S.B. performed embryo staining, O.V.A., P.S.B and J.H. performed data analysis and interpretation of the Ttc21b and Gli2 validation experiment. X.H., C.Q., J.H., V.S. and S.B. performed all computational analyses. C.M. created the interactive webpage with guidance from X.H. and J.S. M.D., L.S., S.S. and C.T. provided assistance with data analysis and results interpretation. X.H., C.Q., J.H. and V.S. wrote the first draft of the manuscript, which was finalized together with J.C., M.S. and J.S. and input from all authors.

Abstract

Mouse models are a critical tool for studying human diseases, particularly developmental disorders[230]. However, conventional approaches for phenotyping may fail to detect sub-

tle defects throughout the developing mouse[231]. Here we set out to establish single cell RNA sequencing (sc-RNA-seq) of the whole embryo as a scalable platform for the systematic phenotyping of mouse genetic models. We applied combinatorial indexing-based sc-RNA-seq[15] to profile 101 embryos of 22 mutant and 4 wildtype genotypes at embryonic stage E13.5, altogether profiling over 1.6M nuclei. The 22 mutants represent a range of anticipated severities, from established multisystem disorders to deletions of individual enhancers[232, 233]. We developed and applied several analytical frameworks for detecting differences in composition and/or gene expression across 52 cell types or trajectories. Some mutants exhibit changes in dozens of trajectories while others only in a few cell types. We also identify differences between widely used wildtype strains, compare phenotyping of gain vs. loss of function mutants, and characterize deletions of topological associating domain (TAD) boundaries. Intriguingly, some changes are shared among mutants, suggesting that developmental pleiotropy might be "decomposable" through further scaling of this approach. Overall, our findings show how single cell profiling of whole embryos can enable the systematic molecular and cellular phenotypic characterization of mouse mutants with unprecedented breadth and resolution.

4.1 Introduction

For over 100 years, the laboratory mouse (*Mus musculus*) has served as the quintessential animal model for studying human diseases[230]. For developmental disorders in particular, mice have been transformative, as a mammalian system that is nearly ideal for genetic analysis and in which the embryo is readily accessible[234].

At its inception, mouse genetics relied on spontaneous or induced mutations resulting in visible physical defects that could then be mapped. However, gene-targeting techniques later paved the way for "reverse genetics", *e.g.* analyzing the phenotypic effects of intentionally

engineered mutations[234]. Through systematic efforts such as the International Knockout Mouse Consortium, knockout models are now available for thousands of genes[235]. Furthermore, with genome editing[236, 237], it is increasingly practical to delete individual regulatory elements[238].

Phenotyping has also grown more sophisticated. Conventional investigations of developmental syndromes typically focus on one organ at a specific stage, *e.g.* combining expression analyses, histology, and imaging to investigate a visible malformation[230]. The Mouse Clinic, involving a battery of standardized tests, reflects a more systematic approach[239], but phenotypes detected through such tests (*e.g.*, behavioral, electrophysiological) may require years of additional work to link to molecular and cellular correlates. Furthermore, it is often the case that an intentionally engineered mutation results in no detectable abnormality[240]. In such instances, it remains unknown whether there is truly no phenotype, or whether the methods used are simply insufficiently sensitive. In sum, phenotyping has become "rate limiting" in mouse genetics.

Single cell molecular profiling offers a potential path to overcome such barriers. As a first step, we and others have applied sc-RNA-seq to profile wildtype mouse development at the scale of the whole embryo[15, 54, 55, 13, 14]. Applying sc-RNA-seq to mouse mutants, several groups have successfully unraveled how specific mutations affect transcriptional networks and lead to altered cell fate decisions in individual organs[241]. However, there is still no clear framework for analyzing such data at the whole embryo scale.

4.2 Single-cell RNA-seq of 101 mouse embryos

We set out to establish whole embryo sc-RNA-seq as a scalable framework for the systematic molecular and cellular phenotyping of mouse genetic models. We collected 103 mouse

embryos, including 22 different mutants and four wildtype strains at embryonic stage E13.5, and generally four replicates per strain (**Fig. 1a**). Mutants were chosen to represent a spectrum of phenotypic severity ranging from established pleiotropic disorders to knockouts of individual regulatory elements.

We grouped mutants into four categories (**Supplementary Table 1, Fig. 1a**):

1. pleiotropic mutants: knockouts of developmental genes expressed in multiple organs (*Ttc21b* KO, *Carm1* KO, *Gli2* KO), and two mutations of the *Sox9* regulatory landscape suspected to have pleiotropic effects (*Sox9* TAD boundary KI; *Sox9* regulatory INV)[233, 242, 243, 244].

2. developmental disorder mutants: intended to model specific human diseases (*Scn11a* GOF, *Ror2* KI, *Gorab* KO, *Cdkl5* -/Y)[245, 246, 247].

3. mutations of loci associated with human disease (*Scn10a*/*Scn11a* DKO, *Atp6v0a2* KO, *Atp6v0a2* R755Q, *Fat1* TAD KO)[248, 249].

4. prospective deletions of cis-regulatory elements, including of TAD boundaries near developmental transcription factors (*Smad3*, *Tbx5*, *Neurog2*, *Sim1*, *Smad7*, *Dmrt1*, *Tbx3*, *Twist1*)[232].

As a positive control, this category includes a ZRS distal enhancer (Zone of polarizing activity Regulatory Sequence) KO mutant, which specifically fails to develop distal limb structures[250]. Except for *Scn11a* GOF, all mutants were homozygous.

The 103 flash-frozen embryos (26 genotypes x 4 individuals; one embryo lost in transport), were sent by five groups to a single site and subjected to sci-RNA-seq3[15]. After doublet filtering, we profiled 1,671,245 nuclei (16,226 +/- 9,289 per embryo; 64,279 +/- 18,530 per

strain; median UMI count: 843/cell; median genes detected: 534; 75% duplication rate). Below we refer to this dataset as the mouse mutant cell atlas (MMCA).

Applying principal components analysis (PCA) to "pseudobulk" profiles of the embryos resulted in two groups corresponding to genetic background (**Fig. 1b**), with FVB embryos clustering separately from C57BL/6J, G4, and BALB/C embryos. However, embryos corresponding to individual mutants did not cluster separately, suggesting none were affected with severe, global aberrations. A single outlier (104) was aberrant with respect to cell recovery (n = 1,047) and appearance (**Supplementary Fig. 1a**).

To validate staging, we leveraged our previous mouse organogenesis cell atlas (MOCA), which spans E9.5 to E13.5[15]. PCA of pseudobulk profiles of 61 wildtype embryos from MOCA resulted in a first component (PC1) strongly correlated with developmental age (**Fig. 1c**). Projecting pseudobulk profiles of the 103 MMCA embryos to this embedding resulted in the vast majority of MMCA embryos clustering with E13.5 MOCA embryos along PC1, consistent with accurate staging. However, five MMCA embryos appeared closer to E11.5 or E12.5 MOCA embryos. Four of these were retained as their delay might be explained by their mutant genotype, while one wildtype embryo (C57BL/6; 41) was designated a second outlier. We removed cells from the two outlier embryos (104; 41) as well as cells with high proportions of reads mapping to the mitochondrial genome (>10%) or ribosomal genes (>5%). This left 1,627,857 cells, derived from 101 embryos (**Fig. 1d**).

To facilitate data integration, we projected cells from all genotypes to a wildtype-derived "reference embedding" (**Supplementary Fig. 1b,c; Methods**). Altogether, we identified 13 major trajectories, 8 of which were further stratified into 59 sub-trajectories (**Fig. 1e; Supplementary Fig. 2a; Supplementary Table 2**), generally covering the expected cell trajectories at this stage of development. These were also generally consistent with our annotations of MOCA, albeit with some corrections as described elsewhere[18, 152]. Greater

granularity for some cell types is likely a consequence of the deeper sampling of E13.5 cells in these new data (**Supplementary Fig. 2b**).

4.3 Mutant-specific differences in cell type composition

In analyzing these data, we pursued three approaches: 1) quantification of gross differences in cell type composition (this section); 2) investigation of more subtle differences in the distribution of cell states within annotated trajectories and sub-trajectories; and 3) analysis of the extent to which phenotypic features are shared between mutants.

To systematically assess cell type compositional differences, we first examined the proportions of cells assigned to each of 13 major trajectories. These proportions were mostly consistent across genotypes (**Supplementary Fig. 3a**), but some mutants exhibited substantial differences. For example, compared to C57BL/6 wildtype, the proportion of cells in the neural tube trajectory decreased from 37.3% to 33.7% and 32.6% in the *Gli2* KO and *Ttc21b* KO strains, respectively, while the proportion of cells in the mesenchymal trajectory decreased from 44.1% to 37.1% in the *Gorab* KO strain. These changes are broadly consistent with the gross phenotypes associated with these mutations[251, 242, 247], but are caveated by substantial interindividual heterogeneity (**Supplementary Fig. 3b**). We also observe differences in major trajectory composition between the four wildtype strains. For example, FVB and G4 wildtype mice consistently had fewer mesenchymal and more neural tube cells than BALB/C and C57BL/6 wildtype embryos (**Supplementary Fig. 3c**). We further checked for technical effects (*e.g.* experimental batch) that might confound cell type recovery rates (**Supplementary Fig. 4a-c**).

We next sought to investigate compositional differences at the level of sub-trajectories. For each combination of background and sub-trajectory, we performed regression to iden-

tify mutations that were nominally predictive of the proportion of cells falling in that sub-trajectory (uncorrected p-value < 0.05; beta-binomial regression; **Methods**). Across 22 mutants, this analysis highlighted 300 nominally significant changes (**Fig. 2a; Supplementary Table 3**). Due to the limited number of replicate embryos per strain, our power to definitively call such changes is limited, particularly in the smaller trajectories (**Supplementary Fig. 4d; Methods**). Nevertheless, two patterns are noteworthy:

First, *Atp6v0a2* KO and *Atp6v0a2* R755Q, distinct mutants of the same gene[248], exhibit highly consistent patterns of change, both with respect to which sub-trajectories are nominally significant as well as the direction and magnitude of changes. The consistency supports the validity of this analytical approach.

Second, mutants varied considerably with respect to the number of sub-trajectories nominally significant for compositional differences. At the higher extreme, 30 of 54 sub-trajectories were nominally altered by the *Sox9* regulatory INV mutation, consistent with the wide-ranging roles of *Sox9* in development[252, 253]. At the lower extreme, TAD boundary knockouts exhibited very few changes, consistent with the paucity of gross phenotypes in such mutants[232]. Nonetheless, all TAD boundary knockouts did show some nominal changes, including specific ones, *e.g.* the lung epithelial and liver hepatocyte trajectories were specifically decreased in *Dmrt1* and *Tbx3* TAD boundary KOs, respectively.

There were a few extreme examples, *e.g.* where a sub-trajectory appeared to be fully lost. For example, *Ttc21b*, which encodes a cilial protein and whose knockout is associated with brain, bone and eye phenotypes[254, 242], exhibited a dramatic reduction in retinal neuron (log2(ratio) = -7.16) (**Fig. 2b**), lens (log2(ratio) = -2.40) and retina epithelium (log2(ratio) = -1.65) trajectories (**Supplementary Fig. 5a-c**). Validations via H&E staining support these patterns, as the homozygous *Ttc21b* mutant exhibits a visible collapse in structures that are detectable within the wildtype eye at E13.5. Specifically, the retinal neurons, lens

and optic nerve were missing in the homozygous mutant (**Fig. 2c**). The retinal epithelium was delocalized and reduced as well (**Fig. 2c; Supplementary Fig. 5c**).

We next examined the positive control, the ZRS limb enhancer KO, a well-studied mutant which shows a loss of the distal limb structure at birth[250]. Eight sub-trajectories were nominally altered in this mutant, mostly mesenchymal. Although the reduction in limb mesenchymal cells was modest (24% or log2(ratio) = -0.39), co-embedding of limb mesenchyme cells from ZRS limb enhancer KO and FVB wildtype embryos identified a subpopulation that specifically expressed markers of the distal mesenchyme of the early embryonic limb bud, such as *Hoxa13* and *Hoxd13*, that was markedly affected (**Fig. 2d-e; Supplementary Fig. 5d**). Such heterogeneity was not observed for the seven other nominally altered sub-trajectories (**Supplementary Fig. 5e**), consistent with the specificity of this phenotype.

4.4 LochNESS analysis reveals differences in transcriptional state within cell type trajectories

We next sought to develop a more sensitive approach for detecting deviations in transcriptional programs within cell type trajectories. For this, we developed "lochNESS" (local cellular heuristic Neighborhood Enrichment Specificity Score), a score calculated based on the "neighborhood" of each cell in a sub-trajectory co-embedding of a given mutant (all replicates) vs. a pooled wildtype (all replicates of all backgrounds). Briefly, lochNESS takes aligned PC features of each sub-trajectory and finds k-NNs for each cell from other embryos. For each mutant cell, we compute the fold-change of the observed vs. expected number of mutant cells in its neighborhood (**Fig. 3a; Methods**; similar methods developed independently by Dann and colleagues[255]).

Visualization of lochNESS in the embedded space highlights areas with enrichment or

depletion of mutant cells. For example, returning to the ZRS limb enhancer KO embryos, we observe markedly low lochNESS in the distal mesenchyme of the early embryonic limb bud (**Fig. 3b; Fig. 2d**). This highlights the value of lochNESS, as within a sub-trajectory (limb mesenchyme), an effect is both detected and assigned to a subset of cells in a label-agnostic fashion.

Globally, the distribution of lochNESS is unremarkable for some mutants (*e.g.* most TAD boundary KOs) but aberrant for others (*e.g.* pleiotropic mutants such as *Sox9* regulatory INV) (**Supplementary Fig. 6a**). After performing additional quality control checks (**Supplementary Fig. 6b-d; Methods**), we examined lochNESS for each mutant in each sub-trajectory. Consistent with earlier analyses, we observe low lochNESS for the retinal neuron sub-trajectory in *Ttc21b* KO mice (**Fig. 3c; Supplementary Fig. 6e**). We also observe a strong shift towards low lochNESS for the floor plate sub-trajectory in *Gli2* KO mice, and a subtle change for the roof plate trajectory, which is forming opposite to the floor plate along the D-V axis of the developing neural tube (**Fig. 3c; Supplementary Fig. 6e**)[256].

To explore this further, we extracted and reanalyzed cells corresponding to the floor plate and roof plate (**Supplementary Fig. 7a**). Within the floor plate, *Gli2* KO cells consistently exhibited low lochNESS (**Fig. 3d**). However, there were only a handful of differentially expressed genes between wildtype and mutant, and no significantly enriched pathways. For example, genes like *Robo1* and *Slit1*, involved in neuronal axon guidance, are specifically expressed in the floor plate relative to the roof plate (**Fig. 3e**), but are not differentially expressed between wildtype and *Gli2* KO cells of the floor plate. Alternatively, our failure to detect substantial differential expression may be due to power, as there were fewer floor plate cells in the *Gli2* KO (60% reduction). Overall, these observations are consistent with the established role of *Gli2* in floor plate induction, its role as activator of Shh in dorso-ventral patterning of the neural tube and the previous demonstration that *Gli2*

knockouts fail to properly induce a floor plate[257, 256].

Less expectedly, we identified two subpopulations of roof plate-derivative cell types, one depleted and the other enriched in *Gli2* KO embryos (**Fig. 3d; Supplementary Fig. 7a-c**). To annotate these subpopulations, we examined genes whose expression was predicted by lochNESS (Methods). The mutant-enriched group of roof plate cells was marked by cilial genes and *Ttr*, a marker for choroid plexus, while the mutant-depleted group was marked by Wnt-related genes (*e.g. Rspo1/2/3* and *Wnt3a/8b/9a*) indicating it to be a region close to the choroid plexus of the lateral ventricle, namely the cortical hem (**Fig. 3e; Supplementary Fig. 7d; Supplementary Table 4-5**). We also mapped the three clusters shown in **Supplementary Fig. 7a** to spatial transcriptomic data from E13.5 mouse embryos (E13.5)[32] (**Supplementary Fig. 7e**). Supporting our annotations, cluster 1 mapped to the floor of the neural tube, cluster 2 next to the lateral ventricle choroid plexus (ChP), and cluster 3 to the ChP, both in the lateral (anterior) and 4th (posterior) ventricles. We then examined marker genes that further separate lateral ventricle and 4th ventricle of ChP and found that in addition to roof plate marker *Lmx1a*, cluster 3 expresses 4th ventricle marker *Meis1* while cluster 2 expresses lateral ventricle markers *Otx1* and *Emx2* (**Supplementary Fig. 7f; Supplementary Table 4**).

To experimentally validate these observations, we examined developmental progression of the neural tube and brain in E13.5 *Gli2* KO mutant and wildtype embryos. In coronal sections of the mutant, we observed severe developmental defects including deformed forebrain lobes and delayed neural tube development (**Supplementary Fig. 8a**). Immunofluorescence imaging of *Pax6* revealed a severely disturbed shape of the neural tube, confirming the well-described "dorsalization" phenotype of the neural tube (**Supplementary Fig. 8b**), and consistent with marked reductions in the proportion of floor plate cells in the *Gli2* KO mutant (**Fig. 3d**). Turning to the less expected observation of increased ChP, we found the lateral ventricle as well as the 4th ventricle displayed a disturbed *Ttr* staining pattern.

While the wildtype shows inner and outer *Ttr* signal within the single cell layer, the mutant displayed a 'double dapi' layer, indicating a disordered tissue organization (**Fig. 3f; Supplementary Fig. 8c,d**). Adjusting for the overall smaller size of *Gli2* KO mutants at E13.5, we quantified *Ttr*-positive cells in the lateral and 4th ventricle, and found a proportional increase in the mutant relative to wildtype (**Supplementary Table 6**), again consistent with the marked increase in the proportion of ChP cells in this mutant (**Fig. 3d**). In summary, we could confirm both the expected reduction in floor plate and the unexpected increase in roof plate-derived ChP in the mutant. Of note, the relatively subtle and opposing effects on these roof plate subpopulations were missed by our original analysis of cell type proportions, and only uncovered by the granularity of lochNESS.

LochNESS distributions can be systematically screened to identify sub-trajectories exhibiting mutant-specific shifts. For example, while all TAD boundary KO mutants have similarly unremarkable global lochNESS distributions, when we plot these distributions by sub-trajectory, a handful of shifted distributions are evident (**Supplementary Fig. 9a,b**). For example, multiple epithelial sub-trajectories, including pre-epidermal keratinocyte, epidermis, branchial arch, and lung epithelial trajectories, are most shifted in *Tbx3* TAD boundary KO cells, with further analyses preliminarily supporting a role for *Tbx3* in epidermal and lung development (**Supplementary Fig. 9c,d; Supplementary Table 7; Methods**)[258].

4.5 Identification of mutant-specific and mutant-shared effects

Pleiotropy, wherein a single gene influences multiple, unrelated traits, is a pervasive phenomenon in developmental genetics, and yet remains poorly understood[259]. Although here we have "whole embryo" molecular profiling of just 22 mutants, we sought to investigate whether we could distinguish between mutant-specific and mutant-shared effects within each

major trajectory. In brief, within a co-embedding of cells from all embryos from a given background, we computed k-NNs as in **Fig. 3a**, and then calculated the observed vs. expected ratio of each genotype among a cell's k-NNs. The "similarity score" between one genotype vs. all others is defined as the mean of these ratios across cells of the genotype (**Methods**). To assess whether any observed similarities or dissimilarities are robust, we can also calculate similarity scores between individual embryos. For example, for the mesenchymal trajectory of C57BL/6 mutants, similarity scores are generally higher for pairwise comparisons of individuals with the same genotype (**Fig. 4a; Supplementary Fig. 10a-c**). Pairs of individuals with the *Scn11a* GOF mutation exhibited the most extreme similarity scores, consistent with our earlier observation that they clustered with E12.5 rather than E13.5 embryos (**Fig. 1c**). Upon further analysis, we believe that the most parsimonious explanation is incorrect staging of these litters, rather than mutation-specific, global developmental delay (**Supplementary Note 1; Supplementary Fig. 10d-g**).

We also observed that the similarity scores between three mutants (*Atp6v0a2* KO, *Atp6v0a2* R755Q, *Gorab* KO) were consistent with shared effects, in the mesenchymal, epithelial, endothelial, hepatocyte and neural crest (PNS glia) trajectories in particular; in other main trajectories, such as neural tube and hematopoiesis, *Atp6v0a2* KO and *Atp6v0a2* R755Q exhibited high similarity scores with one another, but not with *Gorab* KO (**Fig. 4a; Supplementary Fig. 10a,c,f**). In human patients, mutations in ATP6V0A2 and GORAB cause overlapping connective tissue disorders, which is reflected in the misregulation of the mesenchymal trajectory of *Atp6v0a2* and *Gorab* mutants[247, 248]. However, only the ATP6V0A2-related disorder displays a prominent CNS phenotype, consistent with the changes in the neural tube trajectory seen only in *Atp6v0a2* mutants (**Supplementary Fig. 10a,c,f**).

To further explore phenotypic sharing between these mutants, we co-embedded cells of the lateral plate & intermediate mesoderm sub-trajectory from C57BL/6 strains. We resolved

the identity of most sub-clusters using marker genes and spatial mapping, identifying multiple subsets where *Atp6v0a2* KO, *Atp6v0a2* R755Q and *Gorab* KO mice are similarly distributed compared to other C57BL/6 genotypes (**Fig. 4b; Supplementary Fig. 11**). Some subsets are enriched for cells from these mutants (*e.g.* proepicardium, hepatic mesenchyme, lung mesenchyme) while others are depleted (*e.g.* gastrointestinal smooth muscle) (**Fig. 4c,d; Supplementary Table 4**). Although individually subtle, the consistent shifts in cell type proportions between the two *Atp6v0a2* and *Gorab* KO mutants across these subsets of lateral plate mesoderm-derived mesenchyme presumably underlie their high mesenchymal similarity scores (**Fig. 4c**).

Altogether, these analyses illustrate how the joint analysis of mutants subjected to whole embryo sc-RNA-seq can reveal sharing of molecular and cellular phenotypes. This includes global similarity (*Atp6v0a2* KO vs. *Atp6v0a2* R755Q) as well as instances in which specific aspects of phenotypes are shared between previously unrelated mutants (*Atp6v0a2* mutants vs. *Gorab* KO).

4.6 Global developmental defects in Sox9 regulatory mutant

About half of the mutants profiled here model disruptions of regulatory, rather than coding, sequences. Among these, the *Sox9* regulatory INV mutant stands out in having a dramatically shifted lochNESS distribution, particularly in mesenchyme (**Fig. 5a; Supplementary Fig. 6a**). *Sox9* encodes a pleiotropic transcription factor crucial for development of the skeleton, the brain, sex determination, and other systems, orchestrated by a complex regulatory landscape[260, 261, 262]. This particular mutant features an inversion of a 1Mb upstream region bearing several distal enhancers and a TAD boundary, essentially relocating these elements into a TAD with *Kcnj2*, which encodes a potassium channel (**Fig. 5b**)[233]. Like the *Sox9* KO, the homozygous *Sox9* regulatory INV is perinatally lethal, with exten-

sive skeletal phenotypes including digit malformation, a cleft palate, bowing of bones and delayed ossification. In addition to the loss of 50% of *Sox9* expression, the inversion causes pronounced missexpression of *Kcnj2* in the digit anlagen in a wildtype *Sox9* pattern[233]. However, the extent to which *Kcnj2* and *Sox9* are mis-expressed elsewhere, as well as the molecular and cellular correlates of the widespread skeletal phenotype, have yet to be deeply investigated.

At the level of mesenchymal sub-trajectories, shifts in the lochNESS distribution for *Sox9* regulatory INV were consistently observed, but limb mesenchyme and connective tissue were particularly enriched for cells with extremely high lochNESS (**Fig. 5a**, right). Notably, 2 of 3 major enhancers (E250, E195) known to drive *Sox9*-mediated chondrogenesis in mesenchymal stem cells are located within the inverted region (**Fig. 5b**)[260]. Cell type composition analysis (**Fig. 2a**) showed that *Sox9* regulatory INV mutants harbor considerably larger numbers of cells classified as limb mesenchyme, at the expense of osteoblasts, lateral plate & intermediate mesoderm, chondrocytes and connective tissue trajectory. This shift can also be seen in a UMAP embedding (**Fig. 5c**), a topic we revisit further below.

These changes in cell type composition were accompanied by reduced expression of *Sox9* and increased expression of *Kcnj2* in bone (aggregate of chondrocyte, osteoblast, limb mesenchyme; **Supplementary Fig. 12a**), although the number of cells expressing *Kcnj2* was generally low. This suggests that the *Sox9* regulatory inversion is resulting in increased *Kcnj2* expression (via *Sox9* enhancer adoption) and *Sox9* reduction (via boundary repositioning) not only in the digit anlagen, but in skeletal mesenchyme more generally. To validate this, we performed RNA in situ hybridization (RNAscope) on sections of developing bones of the rib cage at E13.5, comparing a heterozygous *Sox9* regulatory INV mouse with a wildtype littermate. Consistent with our sc-RNA-seq data derived from homozygous mutants, we observe a *Sox9*-patterned increase in *Kcnj2* levels, together with losses in *Sox9* expression, in the developing bone (**Fig. 5d; Supplementary Fig. 12b**).

Since the inverted *Sox9* regulatory region also hosts multiple enhancers active in other tissues (*e.g.* E161 in lung; E239 in cerebral cortex)[260], we wondered whether these patterns were also seen in other tissues. Indeed, both sc-RNA-seq and RNAscope quantification show increased *Kcnj2* levels in all other tissues examined. While reductions in *Sox9* expression, clear in bone, were not observed in most other tissues by sc-RNA-seq, RNAscope showed *Sox9* reductions in telencephalon and lung as well (**Supplementary Fig. 12a,b**). Taken together, these data suggest marked changes in mesenchyme due to reduced *Sox9*, together with broader increases in *Kcnj2* expression. As expected based on the role of *Sox9* in chondrogenesis, hallmark pathways related to chondrocyte proliferation and differentiation[263, 264, 265, 266] were downregulated; less expectedly, several immune-related pathways were upregulated (**Supplementary Fig. 12c**).

To explore the apparent accumulation of limb mesenchyme in the *Sox9* regulatory INV (**Fig. 5c, Supplementary Fig. 12d**) in more detail, we reanalyzed mutant and wildtype cells from the limb mesenchyme sub-trajectory, which revealed subpopulations of condensing mesenchyme, perichondrium, and undifferentiated mesenchyme (**Supplementary Fig. 12e,f**). RNA velocity analyses suggested the vast majority of limb mesenchyme "accumulation" in mutant embryos is due to cells that are delayed or stalled in an undifferentiated or stem-like state (**Fig. 5c,e; Supplementary Fig. 12e**). This accumulation is even more apparent in integrated views of the limb mesenchyme sub-trajectory, where we observe branches that are highly enriched for *Sox9* regulatory INV mutant cells, within undifferentiated mesenchyme (**Fig. 5e; Supplementary Fig. 12g,h**).

To investigate these branches further, we sub-clustered undifferentiated mesenchyme cells from mutant and wildtype (**Fig. 5f,g**). Interestingly, the most differentially expressed genes in "branch 2" were largely neuronal (*e.g.* several neurexins; neuregulin 3), an observation supported by gene set enrichment analysis (**Supplementary Fig. 12i,j**). A cellular composition analysis revealed that these neuronal-like cells were not restricted to the Sox9

regulatory INV mutant, but also found in wildtype embryos, albeit much less frequently (**Supplementary Fig. 12g,h**). To validate this unexpected 'neural-like' branch of mesenchymal cells as well as to assess their anatomical distribution, we mapped these cells to spatial transcriptomic data from E13.5 mouse embryos[32]. Strikingly, this analysis placed "branch 2" cells along the neural tube and the brain regions (**Supplementary Fig. 13a**). To address concerns that artifacts might arise from mapping single-cell data onto non-single-cell spatial maps, we also integrated our data with sci-space[267] spatial transcriptomic data (E14.5), as these retain single nucleus resolution. The results are consistent, in that branch 2 mesenchymal cells are enriched in brain regions, branch 0 cells are enriched in limb bud regions, and branch 1 & 3 cells are diffusely distributed but largely excluded from brain regions (**Supplementary Fig. 13b**).

Taken together, these analyses support the validity of this neural-like subset of mesenchyme (present in wildtype and increased in *Sox9* regulatory INV mutants). The observation is consistent with the reports that mesenchymal stem cells can be differentiated to neuronal states *in vitro*[268].

4.7 Discussion

Here we set out to establish whole embryo sc-RNA-seq as a new paradigm for the systematic, scalable phenotyping of mouse developmental mutants. On data obtained for 22 mutants in a single experiment, we developed analytical approaches to identify deviations in cell type composition, subtle differences in gene expression within cell types ("lochNESS"), and sharing of sub-phenotypes between mutants ("similarity scores"). Overall, the results are encouraging, and show how systematic, outcome-agnostic computational analyses of data obtained at the whole embryo scale may in some cases reveal molecular and cellular phenotypes that are missed by conventional phenotyping.

We emphasize that the concurrent analysis of many mutants proved essential to the contextualization of particular observations, *e.g.* to understand how specific or non-specific any apparent deviation really was, against a background of dozens of genotypes and over 100 embryos. This also enabled us to discover shared aspects of phenotypes between previously unrelated genotypes, *e.g.* between *Gorab* and *Atp6v0a2* mutants. Looking forward, profiling of additional mouse mutants might enable the further "decomposition" of developmental pleiotropy, a poorly understood phenomenon, into "basis vectors".

The diverse mutants analyzed yielded a variety of results that speak to the utility of whole embryo sc-RNA-seq for phenotyping. For example, an abnormal eye phenotype in *Ttc21b* mutants was previously described, but considered likely to be secondary to a more general craniofacial defect[254, 242]. The sc-RNA-seq analysis of E13.5 *Ttc21b* mutants demonstrated multiple retinal cell trajectories were essentially absent. Detailed histological analysis confirmed this, suggesting that the eye abnormality is likely not a secondary effect, but rather that the overactive SHH signaling has a primary effect on retinal development in this mutant.

The utility of pursuing whole embryo sc-RNA-seq was also demonstrated by an unexpected finding of both a depleted and an enriched cell population of roof plate cells derivatives in the *Gli2* KO mutant. The "dorsalization" of the neural tube in the absence of SHH signaling is well-described[242, 257, 256] and was confirmed in our histological analysis of this line (**Supplementary Fig. 8**). However, there have been no described changes in the roof plate or its derivatives to date in *Gli2* KO mice [257]. In contrast, whole embryo sc-RNA-seq uncovered that derivatives of the roof plate depict changes in composition (a primary finding) and tissue development (based on secondary validation) in the mutant, illustrating how this approach can potentially yield new insight into even well-studied developmental pathways. However, due to our dataset only capturing one timepoint, whether *Gli2* misexpression causes the structural change directly in the derivative tissue or earlier

during roof plate formation, remains elusive.

Our mouse mutant cell atlas (MMCA) has limitations. First, we only profiled 4 replicates per mutant at a single developmental time point. Based on a simulation analysis of the analytical approach that considers cell proportions only, four replicates of each mutant is likely sufficient to detect modest changes in abundant cell types (*e.g.* a 10% change for cell types at 10% abundance) but only large changes in rarer cell types (*e.g.* a 25% change in cell types at 1% abundance; **Supplementary Fig. 4b**). As such, to detect more subtle changes in model organisms like mouse where very large numbers of replicates are not feasible, more sophisticated strategies like lochNESS, which is not based on counts of cell types but rather directly considers the distribution of cells derived from different genotypes in a complex embedding, may be essential. It is important to note that our cell-composition analysis, which includes both wildtype and mutant cells from the same strain to generate a pooled reference, assumes that the cell type proportions of non-wildtype genotypes are roughly consistent, at least on the whole, with those of wildtype cells. This assumption may be more problematic in studies of biologically related mutants.

Second, profiling only a small fraction of cells present in E13.5 embryos potentially limits sensitivity. On the other hand, for any given mutant, we had over 1.5M cells from other genotypes (wildtype or other mutants), which facilitated the detection of mutant-specific phenotypes for rare cell types, *e.g.* in the retina (*Ttc21b* KO) and roof plate (*Gli2* KO).

Third, we were not able to explore all mutants in detail, nor to thoroughly investigate other aspects of the data (*e.g.* the differences between wildtype strains). In the future, we anticipate that community input and domain expertise will be essential to extract full value from these data. To facilitate this, we created an interactive browser that allows exploration of mutant-specific effects on gene expression in trajectories and sub-trajectories, together with the underlying data (https://atlas.gs.washington.edu/mmca_v2/). Additionally, some

of the phenotypes identified here have probably not been described before due to the lack of resolution of conventional phenotyping. New secondary validations strategies need to be developed to confirm subtle defects in molecular programs or subtle changes in the relative proportions of specific cell types. A very promising approach would be to complement whole embryo sc-RNA-seq with rapidly advancing methods for whole mouse body antibody labeling and 3D imaging[269].

Fourth, our results emphasize the importance of a well-matched control; although data from our wildtype embryos could be re-used as control data for future studies of additional mutants, that risks batch effects, and a safer strategy would be to always include a well-matched, "in-line" wild-type control while profiling mutant embryos.

In 2011, the International Mouse Phenotyping Consortium (IMPC) set out to drive towards the "functionalization" of every protein-coding gene in the mouse, by generating thousands of knockout mouse lines[270]. In principle, the whole embryo sc-RNA-seq phenotyping approach presented here could be extended to all Mendelian genes or even to all 20,000 mouse gene KOs.

4.8 Supplementary Materials

Supplementary Note 1: Potentially incorrect staging of Scn11a *GOF mutants*

The *Scn11a* GOF mutant exhibited the most extreme similarity scores, in terms of both similarity between replicates and dissimilarity with other genotypes (**Fig. 4a; Supplementary Fig. 10a**). The *Scn11a* GOF mutant carries a missense mutation in the Scn11a locus which is reported to result in reduced pain sensitivity both in mice and men without obvious signs of neurodegeneration, suggesting altered electrical activity of peripheral pain-sensing

neurons and impaired synaptic transmission to postsynaptic neurons[245]. However, at least grossly, the mutant does not seem to be associated with mesenchymal phenotypes. Noting that the *Scn11a* GOF mutant embryos clustered with E12.5 embryos instead of E13.5 embryos in our pseudobulk analysis (**Fig. 1c**), we speculated that its extreme similarity scores might be attributable either to developmental delay of the *Scn11a* GOF mutant at the scale of the whole embryo or incorrect staging. To investigate this further, we co-embedded *Scn11a* GOF mutant cells with pooled wildtype cells and MOCA cells from the neural tube trajectory. While wildtype cells were distributed near E13.5 cells from MOCA, the *Scn11a* GOF cells were embedded closer to cells from earlier developmental timepoints (**Supplementary Fig. 10d**). As a more systematic approach, we calculated a "time score" for each cell from the MMCA dataset by taking the k-NNs of each MMCA cell in the MOCA dataset and calculating the average of the developmental time of the MOCA cells. The relative time score distributions of *Scn11a* GOF cells and wildtype cells suggest that *Scn11a* GOF cells are significantly delayed in all major trajectories examined (single sided student's t-test, raw p-value < 0.01; **Supplementary Fig. 10e**). To further follow up on the possibility of earlier time points being inadvertently harvested for this mutant, we examined 2 mixed litters of wildtype and heterozygous mutants at the stage E13.5 (**Supplementary Fig. 10g**). As the heterozygous mutants did not exhibit signs of developmental delay such as smaller size, difference in eye formation or limb development, the theory of general developmental delay seems unlikely, as a more parsimonious explanation is incorrect staging of these litters.

Data reporting

No statistical methods were used to predetermine sample size. Embryos used in experiments 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 different researchers in different locations.

Embryo collection

Mutants were generated through conventional gene editing tools and breeding or tetraploid aggregation and collected at the embryonic stage E13.5, calculated from the day of vaginal plug (noon = E0.5). Collection and whole embryo dissection was performed as previously described[271]. The embryos were immediately snap-frozen in liquid nitrogen and shipped to the Shendure Lab (University of Washington) in dry ice. Sets of animals with the same genotype were either all male or half male-half female. All animal procedures were in accordance with institutional, state, and government regulations.

Nuclei isolation and fixation

Snap frozen embryos were processed as previously described[15]. Briefly, the frozen embryos were cut into small pieces with a blade and further dissected by resuspension in 1 ml ice cold cell lysis buffer (CLB, 10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-630, 1% SUPERase In and 1% BSA) in a 6 cm dish. adding another 3ml CLB, the sample was strained (40 µm) into a 15 ml Falcon tube and centrifuged to a pellet (500g, 5 min). Resuspending the sample with another 1 ml CLB, the isolation of nuclei was ensured. Pelleting the isolated nuclei again (500g, 5 min) was followed by a washing step by fixation in 10 ml 4% Paraformaldehyde (PFA) for 15 minutes on ice. The fixed nuclei were pelleted (500g, 3 min) and washed twice in the nuclei suspension buffer (NSB) (500g, 5 min). The nuclei finally were resuspended in 500µl NSB and split into 2 tubes, each containing 250 µl sample. The tubes were flash frozen in liquid nitrogen and stored in a -80°C freezer, until further use for library preparation. The embryo preparation was preceded randomly for nuclei isolation in order to avoid batch effects.

sci-RNA-seq3 library preparation and sequencing

The fixed nuclei were permeabilized, sonicated and washed. Nuclei from each mouse embryo were then distributed into several individual wells into 4 96-well plates. We split samples into four batches (25 samples randomly selected in each batch) for sci-RNA-seq3 processing. The ID of the reverse transcription well was linked to the respective embryo for downstream analysis. In a first step the nuclei were then mixed with oligo-dT primers and dNTP mix, denatured and placed on ice, afterwards they were proceeded for reverse transcription including a gradient incubation step. After reverse transcription, the nuclei from all wells were pooled with the nuclei dilution buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 1% SUPERase In and 1% BSA), spun down and redistributed into 96-well plates containing the reaction mix for ligation. The ligation proceeded for 10 min at 25°C. Afterwards, nuclei again were pooled with nuclei suspension buffer, spun down and washed and filtered. Next, the nuclei were counted and redistributed for second strand synthesis, which was carried out at 16°C for 3h. Afterwards tagmentation mix was added to each well and tagmentation was carried out for 5 minutes at 55°C. To stop the reaction, DNA binding buffer was added and the sample was incubated for another 5 minutes. Following an elution step using AMPure XP beads and elution mix, the samples were subjected to PCR amplification to generate sequencing libraries.

Finally after PCR amplification, the resulting amplicons were pooled and purified using AMPure XP beads. The library was analyzed by electrophoresis and the concentration was calculated using Qubit (Invitrogen). The library was sequenced on the NovaSeq platform (Illumina) (read 1: 34 cycles, read 2: 100 cycles, index 1: 10 cycles, index 2: 10 cycles).

Processing of sequencing reads

Read alignment and cell-x-gene expression count matrix generation was performed based on the pipeline that we developed for sci-RNA-seq3[15] with the following minor modifications: base calls were converted to fastq format using Illumina's bcl2fastq/v2.20 and demul-

tiplexed based on PCR i5 and i7 barcodes using maximum likelihood demultiplexing package deML[138] with default settings. Downstream sequence processing and cell-x-gene expression count matrix generation were similar to sci-RNA-seq[53] except that the 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 the RT index and ligation index (Levenshtein edit distance (ED) < 2, including insertions and deletions). Briefly, demultiplexed reads were filtered based on the RT index and ligation index (ED < 2, including insertions and deletions) and adaptor-clipped using trim_galore/v0.6.5 with default settings. Trimmed reads were mapped to the mouse reference genome (mm10), using STAR/v2.6.1d[139] with default settings and gene annotations (GENCODE VM12 for mouse). 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, hairpin ligation adaptor index and read 2 end-coordinate (*e.g.* reads with UMI sequence less than 2 edit distance, 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). To generate the cell-x-gene expression count matrix, 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[140]. 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 the cell-x-gene expression count matrix.

The single cell gene count matrix included 1,941,605 cells after cells with low quality (UMI ≤ 250 or detected gene ≤ 100) were filtered out. Each cell was assigned to its original mouse embryo on the basis of the reverse transcription barcode. We applied three strategies to detect potential doublet cells. As the first strategy, we split the dataset into subsets for each

individual, and then applied the scrublet/v0.1 pipeline[27] 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 scores over 0.2 were annotated as detected doublets (5.5% in the whole data set).

As the second strategy, we used an iterative clustering strategy based on Seurat/v3[51] to detect the doublet-derived subclusters for cells. Briefly, gene count mapping to sex chromosomes was removed before clustering and dimensionality reduction, and then genes with no count were filtered out and each cell was normalized by the total UMI count per cell. The top 1,000 genes with the highest variance were selected. The data was log transformed after adding a pseudo count, and scaled to unit variance and zero mean. The dimensionality of the data was reduced by PCA (30 components) first and then with UMAP, followed by Louvain clustering performed on the 10 principal components (resolution = 1.2). For Louvain clustering, we first fitted the top 10 PCs to compute a neighborhood graph of observations (k.param = 50) followed by clustering the cells into sub-groups using the Louvain algorithm. For UMAP visualization, we directly fit the PCA matrix with min_distance = 0.1. For subcluster identification, we selected cells in each major cell type and applied PCA, UMAP, Louvain clustering similarly to the major cluster analysis. Subclusters with a detected doublet ratio (by Scrublet) over 15% were annotated as doublet-derived subclusters.

We found the above Scrublet and iterative clustering-based approach is limited in marking cell doublets between abundant cell clusters and rare cell clusters (*e.g.* less than 1% of the total cell population), thus, we applied a third strategy to further detect such doublet cells. Briefly, cells labeled as doublets (by Scrublet) or from doublet-derived subclusters were filtered out. For each cell, we only retain protein-coding genes, lincRNA genes, and pseudogenes. Genes expressed in less than 10 cells and cells expressing less than 100 genes were further filtered out. The downstream dimension reduction and clustering analysis were done

with Monocle/v3[15]. The dimensionality of the data was reduced by PCA (50 components) first on the top 5,000 most highly variable genes and then with UMAP (max_components = 2, n_neighbors = 50, min_dist = 0.1, metric = 'cosine'). Cell clusters were identified using the Leiden algorithm implemented in Monocle/v3 (resolution = 1e-06). Next, we took the cell clusters identified by Monocle/v3 and first computed differentially expressed genes across cell clusters with the top_markers function of Monocle/v3 (reference_cells=1000). We then selected a gene set combining the top ten gene markers for each cell cluster (filtering out genes with fraction_expressing < 0.1 and then ordering by pseudo_R2). Cells from each main cell cluster were selected for dimension reduction by PCA (10 components) first on the selected gene set of top cluster-specific gene markers, and then by UMAP (max_components = 2, n_neighbors = 50, min_dist = 0.1, metric = 'cosine'), followed by clustering identification using the Leiden algorithm implemented in Monocle/v3 (resolution = 1e-04). Subclusters showing low expression of target cell cluster-specific markers and enriched expression of non-target cell cluster-specific markers were annotated as doublets derived subclusters and filtered out in visualization and downstream analysis. Finally, after removing the potential doublet cells detected by either of the above three strategies, 1,671,270 cells were retained for further analyses.

Whole mouse embryo analysis

As described previously[15], each cell could be assigned to the mouse embryo from which it derived on the basis of its reverse transcription barcode. After removing doublet cells and another 25 cells which were poorly assigned to any mouse embryo, 1,671,245 cells from 103 individual mouse embryos were retained (a median of 13,468 cells per embryo). UMI counts mapping to each sample were aggregated to generate a pseudobulk RNA-seq profile for each sample. Each cell's counts were normalized by dividing its estimated size factor, and then the data were log2-transformed after adding a pseudocount followed by performing the PCA. The normalization and dimension reduction were done in Monocle/v3.

We previously used sci-RNA-seq3 to generate the MOCA dataset, which profiled 2 million cells derived from 61 wild-type B6 mouse embryos staged between stages E9.5 and E13.5. The cleaned dataset, including 1,331,984 high quality cells, was generated by removing cells with <400 detected UMIs as well as doublets (http://atlas.gs.washington.edu/mouse-rna). UMI counts mapping to each sample were aggregated to generate a pseudobulk RNA-seq profile for each embryo. Each cell's counts were normalized by dividing its estimated size factor, and then the data were log2-transformed after adding a pseudocount, followed by PCA. The PCA space was retained and then the embryos from the MMCA dataset were projected onto it.

Cell clustering and annotation

After removing doublet cells, genes expressed in less than 10 cells and cells expressing less than 100 genes were further filtered out. We also filtered out low-quality cells based on the proportion of reads mapping to the mitochondrial genome (MT%) or ribosomal genome (Ribo%) (specifically, filtering cells with MT% > 10 or Ribo% > 5). We then removed cells from two embryos that were identified as outliers based on the whole-mouse embryo analysis (embryo 41 and embryo 104). This left 1,627,857 cells (median UMI count 845; median genes detected 539) from 101 individual embryos that were retained for all subsequent analyses.

To eliminate the potential heterogeneity between samples due to different mutant types and genotype backgrounds, we sought to perform the dimensionality reduction on a subset of cells from the wildtype mice (including 15 embryos with 215,575 cells, 13.2% of all cells) followed by projecting all remaining cells, derived from the various mutant embryos, onto this same embedding. These procedures were done using Monocle/v3. In brief, the dimensionality of the subset of data from the wildtype mice was reduced by PCA, retaining 50 components, and all remaining cells were projected onto that PCA embedding space. Next, to mitigate potential technical biases, we combined all cells from wildtype and mutant mice and applied

the align_cds function implemented in Monocle/v3, with MT%, Ribo%, and log-transformed total UMI of each cell as covariates. We took the subset of cells from wildtype mice, using their "aligned" PC features to perform UMAP (max_components = 3, n_neighbors = 50, min_dist = 0.01, metric = 'cosine') by uwot/v0.1.8, followed by saving the UMAP space. Cell clusters were identified using the Louvain algorithm implemented in Monocle/v3 on three dimensions of UMAP features, resulting in 13 isolated major trajectories (**Fig. 1e**). We then projected all of the remaining cells from mutant mouse embryos onto the previously saved UMAP space and predicted their major-trajectory labels using a k-nearest neighbor (k-NN) heuristic. Specifically, for each mutant-derived cell, we identified its 15 nearest neighbor wildtype-derived cells in UMAP space and then assigned the major trajectory with the maximum frequency within that set of 15 neighbors as the annotation of the mutant cell. We calculated the ratio of the maximum frequency to the total as the assigned score. Of note, over 99.9% of the cells from the mutant mice had an assigned score greater than 0.8. The cell-type annotation for each major trajectory was based on expression of the known marker genes (**Supplementary Table 2**).

Within each major trajectory, we repeated a similar strategy, but with slightly adjusted PCA and UMAP parameters. For the major trajectories with more than 50,000 cells, we reduced the dimensionality by PCA to 50 principal components; for the other major trajectories of more than 1,000 cells, we reduced the dimensionality by PCA to 30 principal components; for the remaining major trajectories, we reduced the dimensionality by PCA to 10 principal components. UMAP was performing with max_components = 3, n_neighbors = 15, min_dist = 0.1, metric = 'cosine'. For the mesenchymal trajectory, we observed a significant separation of cells by their cell-cycle phase in the UMAP embedding. We calculated a *g2m* index and a *s* index for individual cells by aggregating the log-transformed normalized expression for marker genes of the G2M phase and the S phase and then included them in align_cds function along with the other factors. Applying these procedures to all of the main trajectories, we identified 64 sub-trajectories in total. Similarly, after assigning

each cell from the mutant mice with a sub-trajectory label, we calculated the ratio of the maximum frequency to the total as the assigned score. Of note, over 96.7% of the cells from the mutant mice had an assigned score greater than 0.8. The cell-type annotation for each sub-trajectory was also based on the expression of known marker genes (**Supplementary Table 2**).

Identification of inter-datasets correlated major and sub trajectories using non-negative least-squares (NNLS) regression

To identify correlated cell trajectories between MOCA and MMCA datasets, we first calculated an aggregate expression value for each gene in each cell trajectory by summing the log-transformed normalized UMI counts of all cells of that trajectory. For consistency during the comparison to MOCA, we manually regrouped the cells from the MMCA dataset into 10 cell trajectories, by merging the olfactory sensory neuron trajectory into the neural crest (PNS neuron) trajectory, merging the myotube trajectory, the myoblast trajectory, and the cardiomyocyte trajectory into the mesenchymal trajectory, splitting the hepatocyte trajectory into the lens epithelial trajectory and the liver hepatocyte trajectory. Next, for the two datasets, we applied non-negative least squares (NNLS) regression to predict gene expression in a target trajectory (T_a) in dataset A based on the gene expression of all trajectories (M_b) in dataset B: $T_a = \beta_{0a} + \beta_{1a}M_b$, based on the union of the 3,000 most highly expressed genes and 3,000 most highly specific genes in the target trajectory. We then switched the roles of datasets A and B, *e.g.* predicting the gene expression of target trajectory (T_b) in dataset B from the gene expression of all trajectories (M_a) in dataset A: $T_b = \beta_{0b} + \beta_{1b}M_a$. Finally, for each trajectory a in dataset A and each trajectory b in dataset B, we combined the two correlation coefficients: $\beta = 2(\beta_{ab} + 0.001)(\beta_{ba} + 0.001)$ to obtain a statistic, where high values reflect reciprocal, specific predictivity. We repeated this analysis on sub-trajectories within each major trajectories.

Identification of significant cell composition changes in mutant mice using beta-binomial regression

A cell number matrix of all 64 developmental sub-trajectories (rows) and 101 embryos (columns) was created and the cell number were then normalized by the size factor of each column which was estimated by estimate_size_factors function in Monocle/v3. 10 sub-trajectories with a mean of cell number across individual embryo < 10 were filtered out. The beta-binomial regression was performed using the VGAM package of R, based on the model "(trajectory specific cell number, total cell number of that embryo - trajectory specific cell number) genotype". Of note, embryos from the four different mouse strain backgrounds were analyzed independently.

We hypothesize that the power of our strategy to detect the cell proportion changes between different genotypes is affected by three factors: a) the abundance of a given cell type; b) the number of replicates in each genotype group; and c) the effect size. To evaluate power, we performed a simulation analysis that varied these factors, implemented as follows:

We selected the 20 most abundant cell types in wildtype embryos. Their abundances ranged from 1% to 20%. The proportions of these cell types served as the basis for our simulations. We simulated ten groups of "wildtype" samples with 4, 8, 16, ..., 40 replicates in each group, wherein each sample consisted of cells drawn from the 20 cell types. For each replicate, the simulated number of cells of each cell type was calculated as the product of: a) the cell-type proportions, simulated by fitting a dirichlet model based on the real proportions from step 1; and b) the total number of cells recovered for that replicate, simulated based on the mean (n = 15,000) and standard deviation of the cell numbers across replicates in the real dataset.

We simulated ten groups of "mutant" samples by repeating the above step except adding

shifts to the numbers of cells within each cell type. The shifting scales were based on different effect sizes. For instance, effect size = 0.1 represents a 10% reduction in the number of cells.

We performed beta-binomial regression (the same test used in **Fig. 2a**) to test if the cell type proportions were significantly changed between simulated "wildtype" and "mutant" samples, further checking the results as stratified by cell type (with different abundances), the number of replicates, and the effect size.

The results are in line with our hypothesis that the detection power of our strategy varies among comparisons with different effect sizes, sample sizes, or cell-type abundances (**Supplementary Fig. 4**). The main "take-home" messages are summarized below:

- 25% changes are robustly detectable, even for rare cell types like <2%, with modest numbers of replicates.
- 10% changes are possible to detect, but only for abundant cell types (*e.g.* >5%). More replicates can help in this zone.
- 1% changes are almost impossible to detect with a cell proportions approach, even with very large numbers of replicates.

In general, at the level of single cell sampling performed in our study, four samples (corresponding to the number of samples used in the manuscript) would be sufficient to detect a 25% effect size for those cell types that is present at a 1% proportion in wildtype embryos.

Defining and calculating lochNESS

To identify local enrichments or depletions of mutant cells, we aim to define a metric for

each single cell to quantify the enrichments or depletions of mutant cells in its surrounding neighborhood. For these analyses, we consider a mutant and a pooled wildtype combining all 4 background strains in a main trajectory as a dataset. For each dataset, we define "lochNESS" as: $lochNESS = \frac{\#ofmutantcellsinkNNs}{k} / \frac{\#ofmutantcellsindataset}{N} - 1$, where N is the total number of cells in the dataset, $k = \frac{\sqrt{N}}{2}$ scales with N and the cells from the same embryo as the cell are excluded from the k-NNs. Note that this value is equivalent to the fold change of mutant cell percentage in the neighborhood of a cell relative to in the whole main trajectory. For implementation, we took the aligned PCs in each sub-trajectory as calculated above and for each cell in an embryo we find the k-NNs in the remaining mutant embryo cells and wildtype cells. We plot the lochNESS in a red-white-blue scale, where white corresponds to 0 or the median lochNESS, blue corresponds to high lochNESS or enrichments, and red corresponds to low lochNESS or depletions.

Currently we calculate lochNESS using a pooled wildtype combining all 4 background strains to include larger numbers of cells in constructing the k-NN graph. If the numbers of cells are sufficient, a wildtype from the matched background strain can be used. Additionally, if the numbers of cells are sufficient, one set of lochNESS can be calculated for each wildtype sample separately and the variability between samples can be considered.

Examining global distributions of lochNESS

Plotting the global distributions of lochNESS for each mutant across all sub-trajectories, we further observed that some mutants (*e.g.* most TAD boundary knockouts; *Scn11a* GOF) exhibit unremarkable distributions (**Supplementary Fig. 6a**). However, others (*e.g. Sox9* regulatory INV; *Scn10a/11a* DKO) are associated with a marked excess of high lochNESS, consistent with mutant-specific effects on transcriptional state across many developmental systems. For reference, we simultaneously create a null distribution of lochNESS using random permutation of the mutant and wildtype cell labels, simulating datasets in which the

cells are randomly mixed. Of note, we confirmed that repeating the calculation of lochNESS after random permutation of mutant and wildtype labels resulted in bell-shaped distributions centered around zero (**Supplementary Fig. 6b**). As such, the deviance of lochNESS can be summarized as the average euclidean distance between lochNESS versus lochNESS under permutation (**Supplementary Fig. 6c**). In addition, we computed lochNESS between wildtypes from different background strains and observed minimal variation in cell distribution between wildtype from G4, FVB and BALB/C strains and potential strain-specific distributions in C57BL/6 wildtype mice (**Supplementary Fig. 6d**).

Comparing lochNESS with batch mixing score LISI

LochNESS shares conceptual similarities with batch correcting measurement scores like LISI[272], which quantifies the amount of mixing in a cell's neighborhood by counting the number of batches represented in the neighborhood. As a direct comparison, we calculated LISI on each mutant with a pooled wildtype reference in PCA space. We calculated LISI with a dynamic perplexity based on the dataset size, similar to our strategy for determining the neighborhood size for lochNESS. Focusing on the G4 mutants as an example, the results show a correlation between LISI and lochNESS, where LISI values close to 1 correspond to the more extreme positive or negative values of lochNESS as expected (**Supplementary Fig. 6f**). LochNESS has several conceptual advantages compared to LISI. First, lochNESS can easily determine whether the mutant sample is enriched or depleted in an area that is not well mixed using the sign of the value (positive = enrichment, negative = depletion), while LISI can only separate mixed (scores approaching 2) vs. separated (scores approaching 1). Second, lochNESS can be easily extended to comparisons between multiple samples, while LISI is relatively restricted to pairwise comparisons. Third, lochNESS considers a dataset specific neighborhood size and baseline proportions.

Systematic screening of lochNESS distributions

LochNESS distributions can be systematically screened to identify sub-trajectories exhibiting substantial mutant-specific shifts. For example, while all TAD boundary KO mutants have similarly unremarkable global lochNESS distributions, when we plot these distributions by sub-trajectory, a handful of shifted distributions are evident (**Supplementary Fig. 9a**). Such deviations, summarized as the average euclidean distances between lochNESS and lochNESS under permutation, are visualized in **Supplementary Fig. 9b**. For example, multiple epithelial sub-trajectories, including pre-epidermal keratinocyte, epidermis, branchial arch, and lung epithelial trajectories, are most shifted in *Tbx3* TAD boundary KO cells. Co-embeddings of mutant and wildtype cells of these sub-trajectories, together with regression analysis, identify multiple keratin genes as positively correlated with lochNESS, consistent with a role for *Tbx3* in epidermal development (**Supplementary Fig. 9c,d; Supplementary Table 7**)[258]. The lung epithelial cells were separated into two clusters, with the cluster more depleted in *Tbx3* TAD boundary KO cells marked by *Etv5*, a transcription factor associated with alveolar type II cell development, as well as Bmp signaling genes that regulate *Tbx3* during lung development (*Bmp1/4*), and distal airway markers *Sox9* and *Id2* (**Supplementary Table 4**). Of note, the shifts we observed in *Tbx3* TAD boundary KO cells remain preliminary and would need to be confirmed by further validation experiments.

Spatial mapping with Tangram

We computationally map our dataset onto a spatially resolved transcriptomics dataset, the mouse organogenesis spatiotemporal transcriptomics atlas (MOSTA) generated with Stereo-seq[32]. The atlas has a total of 53 sagittal sections from C57BL/6 mouse embryos at from E9.5 to E16.5 in 1 day intervals, and we obtained one section from the most relevant E13.5 data (E13.5_E1S1.MOSTA.h5ad) from the data sharing website associated with the manuscript: https://db.cngb.org/stomics/mosta/download/. To map the cells for each single cell cluster on the spatially resolved transcriptomics dataset, we used a machine

learning-based method called Tangram[37]. Briefly, Tangram is a computational tool that uses a Bayesian approach to infer the spatial locations of cells in a single-cell transcriptomics dataset based on their transcriptomic profiles and the spatial patterns of gene expression in the spatially resolved dataset. The relevant subset of the MMCA data was preprocessed in Scanpy, but the metadata was inherited from the results generated in the "Cell clustering and annotation" section above. We used Tangram with default parameters to estimate the spatial coordinates of cells from each cluster in the single cell dataset and visualized results on the coordinates provided by MOSTA. We trained the Tangram model in 'gpu' mode using a NVIDIA A100 GPU. Overall, Tangram provided a powerful method for mapping the cells from the single cell RNA-seq dataset onto MOSTA, enabling us to infer the spatial locations of different cell clusters of interest within the tissue.

Calculating mutant and embryo similarity scores

We can extend the lochNESS analysis, which is computed on each mutant and its corresponding wildtype mice, to compute "similarity scores" between all pairs of individual embryos from the same background strain. We consider all embryos in the same background in a main trajectory as a dataset. For each dataset, we take define a "similarity score" between $cell_n$ and $embryo_j$ as:

$$similarityscore_{cell_n,embryo_j} = \frac{\#ofcellsfromembryo_iinkNNsofcell_n}{k} / \frac{\#ofcellsfromembryo_jindataset}{N}$$

Where N is the total number of cells in the dataset and $k = \frac{\sqrt{N}}{2}$. We take the mean of the similarity scores across all cells in the same embryo, resulting in an embryo similarity score matrix where entries are:

$$similarityscore_{embryo_i,embryo_j} = \frac{1}{n_i} \sum_{n=1}^{n_i} similarityscore_{cell_n,embryo_i}$$

Where n_i is the number of cells in $embryo_i$.

RNAscope in situ Hybridization

For RNAscope, embryos were collected at stage E13.5 and fixed for 4 hours in 4% PFA/PBS at room temperature. The embryos were washed twice in PBS before incubation in a sucrose series (5%, 10% and finally 15% sucrose (Roth) /PBS) each for an hour or until the embryos sank to the bottom of the tube. Finally, the embryos were incubated in 15% sucrose/PBS and O.C.T. (Sakura) in a 1:1 solution before embedding the embryos in O.C.T in a chilled ethanol bath and put into -80°C for sectioning. The embryos were cut into 5 μm thick sections on slides for RNAscope.

Simultaneous RNA in situ hybridization was performed using the RNAscope® technology (Advanced Cell Diagnostics [ACD]) and the following probes specific for Mm-Kcnj2 (Cat. No. 476261, ACD) and Mm-Sox9 (Cat. No. 401051-C2, ACD) on 5 μm sections of the mouse embryos. RNAscope probes were purchased by ACD and designed as described by Wang et al.[273]. The RNAscope® assay was run on a HybEZ™II Hybridization System (Cat. No. 321720, ACD) using the RNAscope® Multiplex Fluorescent Reagent Kit v2 (Cat. No. 323100, ACD) and the manufacturer's protocol for fixed-frozen tissue samples with target retrieval on a hotplate for 5 minutes. Fluorescent labeling of the RNAscope® probes was achieved by using OPAL 520 and OPAL 570 dyes (Cat. No. FP1487001KT + Cat. No. FP1488001KT, Akoya Biosciences, Marlborough, MA, USA) and stained sections were scanned at 25x magnification using a LSM 980 with Airyscan 2 (Carl Zeiss AG, Oberkochen, DE).

Image analysis

For quantitative analysis of the RNAscope images, representative fields of view for each stained section were analyzed using the image processing software Fiji[274]. Each organ of interest mRNA signal was counted in a defined area (1 x 1 mm^2) with an n=6 per condition. Statistics were calculated using student t-Test.

Ttc21b and Gli2 Mutant fixation for H&E and Immunofluorescence

Ttc21b homozygous, heterozygous mutants and wild-type E13.5 mouse embryos were fixed overnight in 4% PfA at 4 degrees. To stop fixation, the samples were transferred into 70% ethanol, washed twice and dehydrated. In the following, the embryos were embedded in paraffin, and cut into 2,5 μm thick sections.

Ttc21b Mutant H&E staining

Histochemical staining was performed on the eyes of the embryos using haematoxylin and eosin. Slides were scanned with a digital slide scanner (NanoZoomer 2.0HT, Hamamtsu, Japan) and analyzed using NDP.view2 software (Hamamatsu Photonics). Numbers of processed embryos: wild type = 2, *Ttc21b* heterozygous=2, *Ttc21b* homozygous = 4.

Stained embryo sections were scanned with an AxioScan 7 digital slide scanner (Zeiss, Jena, Germany).

Fluorescence quantification

Quantification of prealbumin expression cells was performed using image analysis software Definiens Developer XD2 (Definiens AG, Germany). The region of interests (ROI 1-4) within the fourth and lateral ventricle ChP, were annotated manually in serial sections. The calculated parameter was the ratio of the total number of prealbumin-positive cells over the embryo section area (µm).

Statistics and Reproducibility

H&E staining of the developing eye (**Fig. 2c**) was performed on homozygous *Ttc21b*

mutants (n=4), heterozygous *Ttc21b* mutants (n=2) and wildtype E13.5 embryos (n=2). Experiments on the sections were performed in parallel to ensure consistency.

H&E staining of *Gli2* mutant and wildtype embryo sections (**Supplementary Fig. 8a**) were performed on homozygous *Gli2*-/- (n=4) and wildtype (n=2) samples. Experiments on the sections were performed in parallel to ensure consistency.

Immunofluorescence stainings of the choroid plexus marker *Ttr* and neural tube marker *Pax6* (**Fig. 3f, Supplementary Fig.8b-d**) were performed on sections of homozygous *Gli2*-/- (n=4) and wildtype (n=2) samples. Immunofluorescence of the same antibody was performed on all mutants in parallel to ensure consistency.

RNAscope images quantification of *Sox9* and *Kcnj2* expression of heterozygous E13.5 wildtype and *Sox9* regulatory INV mutant embryos (n=6 embryos for each condition) was counted in a defined area (1 x 1 mm^2). Statistics were calculated using two-sided student t-test. RNAscope of the tissue was performed on all samples in parallel to ensure consistency.

Clustering and annotation limb mesenchyme trajectory

Seurat/v4.0.6 was used for the analysis. Wildtype cells in the limb mesenchyme trajectory from all wild-type mice (n = 15 mice, n = 25,211 cells) were used to first annotate the cells. The raw counts were log-normalized after which PCA was performed with default parameters on top 2000 highly variable genes selected using the "vst" method. Nearest neighbors were computed on the PCA space, with default parameters, except that all the principal components computed earlier were used. Clustering was performed using the Louvain community detection algorithm with a resolution of 0.1, resulting in three clusters. Positive marker genes for these clusters were identified using the Wilcoxon Rank Sum test, where only the genes expressed in at least 20% of the cells in either cell groups were considered. The clusters were

annotated based on biologically relevant markers (**Supplementary Fig. 12f**). The newly assigned cell annotations for the limb mesenchyme trajectory cells in the wildtype dataset were transferred to the corresponding cells in the Sox9 regulatory INV mutant using the FindTransferAnchors and TransferData functions using default parameters, except that all the computed principal components were used. 92.3% of the transferred annotations had a score (prediction.score.max) greater than or equal to 0.8.

Density visualisation and RNA velocity analysis

Using Seurat/v4.0.6, the raw counts were log-normalized, and PCA was performed with default parameters on top highly variable genes 2000 genes, selected using the "vst" method. Dimensionality reduction was performed using PCA using default parameters, after which the UMAP embedding was carried out on all computed PC components. Density plots were created using the stat_2d_density_filled function in ggplot2/v3.3.5. For RNA velocity analysis using scVelo/v0.2.4, the total, spliced, and unspliced count matrices, along with the UMAP embeddings were exported as an h5ad file using anndata/v0.7.5.2 for R. The count matrices were filtered and normalized using scv.pp.filter_and_normalize, with min_shared_counts=20 and n_top_genes=2000. Means and variances between 30 nearest neighbors were calculated in the PCA space (n_pcs=50, to be consistent with default value in Seurat). The velocities were calculated using default parameters and projected onto the UMAP embedding exported from Seurat.

Single sample Gene Set Enrichment Analysis

Single-sample Gene Set Enrichment Analysis (ssGSEA) was applied to sc-RNA-seq data using the escape R-package[275]. The msigdbr and getGeneSets functions were used to fetch and filter the entire Hallmark (H, 50 sets) or the Signature Cell Type (C8, 700 sets) *Mus musculus* gene sets from the MSigDB[276, 277]. enrichIt with default parameters,

except for using 10000 groups and variable number of cores, was performed on the seurat-object containing data corresponding to the undifferentiated mesenchyme cells from the *Sox9* regulatory INV mutant, after converting the feature names to gene symbols as necessitated by the escape package. The obtained enrichment scores for each gene set were compared between the two branches (**Fig. 5f**) using the two sample Wilcoxon test (wilcox_test) with default parameters and adjusted for multiple comparisons using Bonferroni correction.

Integration and spatial mapping with sci-space data

We integrated our dataset with a spatial transcriptomics dataset on mid-gestational mice (E14.5), based on the sci-space method[267], in which a subset of transcriptionally profiled nuclei have known physical locations in sagittal sections within which they were mapped prior to sc-RNA-seq. We used anchor-based integration as implemented by Seurat for a co-embedding of a subset of MMCA and sci-space. For cells in the subset of MMCA, we find the nearest neighbor in sci-space data in the integrated co-embedding, and plot the location of the neighboring sci-space cell where it is known.

Data availability

The data generated in this study can be downloaded in raw and processed forms from the NCBI Gene Expression Omnibus under accession number GSE199308. Other intermediate data files and an interactive app to explore our dataset are made freely available via https://atlas.gs.washington.edu/mmca_v2/. All code is made freely available through a public GitHub repository at https://github.com/shendurelab/MMCA. The supplementary tables can be downloaded from here:
https://shendure-web.gs.washington.edu/content/members/cxqiu/public/nobackup/tmp/

Acknowledgments

We thank Stefan Mundlos and Cesar Prada for helpful discussions around data processing and analysis and results interpretation, as well as all members of the Cao, Shendure, Spielmann labs for continuous support and helpful input. N.Haag and I.K. thank Matthias Ebbinghaus for help with breeding of *Scn11a* GOF mice. We thank Vanessa Suckow for genotyping the *Ror2* KI and *Cdkl5* -/Y mice. We thank Scott Houghtaling and Tzu-Hua Ho for breeding and embryo harvest of *Ttc21b*, *Carm1* and *Gli2* mice. X.H. thanks Gwen the Cat for support and cheer ups during meetings. J.S. and work in the Shendure Lab was supported by the Paul G. Allen Frontiers Foundation (Allen Discovery Center grant to J.S. and C.T.), the National Institutes of Health (grant UM1HG011531 to J.S.), Alex's Lemonade Stand's Crazy 8 Initiative (to J.S.) and the Bonita and David Brewer Fellowship (C.Q.). Work at the E.O. Lawrence Berkeley National Laboratory was supported by U.S. National Institutes of Health (NIH) grants to L.A.P. and A.V. (UM1HG009421 and R01HG003988) and performed under U.S. Department of Energy Contract DE-AC02-05CH11231, University of California. J.S. is an Investigator of the Howard Hughes Medical Institute. M.S. is a DZHK principal investigator and is supported by grants from the Deutsche Forschungsgemeinschaft (DFG) (SP1532/3-1, SP1532/4-1, and SP1532/5-1) and the Deutsches Zentrum für Luft- und Raumfahrt (DLR 01GM1925). D.R.B was supported by R01HD36404 from NICHD. J.C. is supported by the National Institutes of Health (grant DP2 HG012522-01 and RM1HG011014) and the Rockefeller University. M.H.A. and work at the German Mouse Clinic was supported by the German Federal Ministry of Education and Research (Infrafrontier grant 01KX1012); German Center for Diabetes Research (DZD).

Competing Financial Interests Statement

J.S. is a SAB member, consultant and/or co-founder of Cajal Neuroscience, Guardant Health, Maze Therapeutics, Camp4 Therapeutics, Phase Genomics, Adaptive Biotechnologies, Scale Biosciences, Pacific Biosciences, Prime Medicine, and Sixth Street Capital. All other authors declare no competing interests.

4.9 Figures

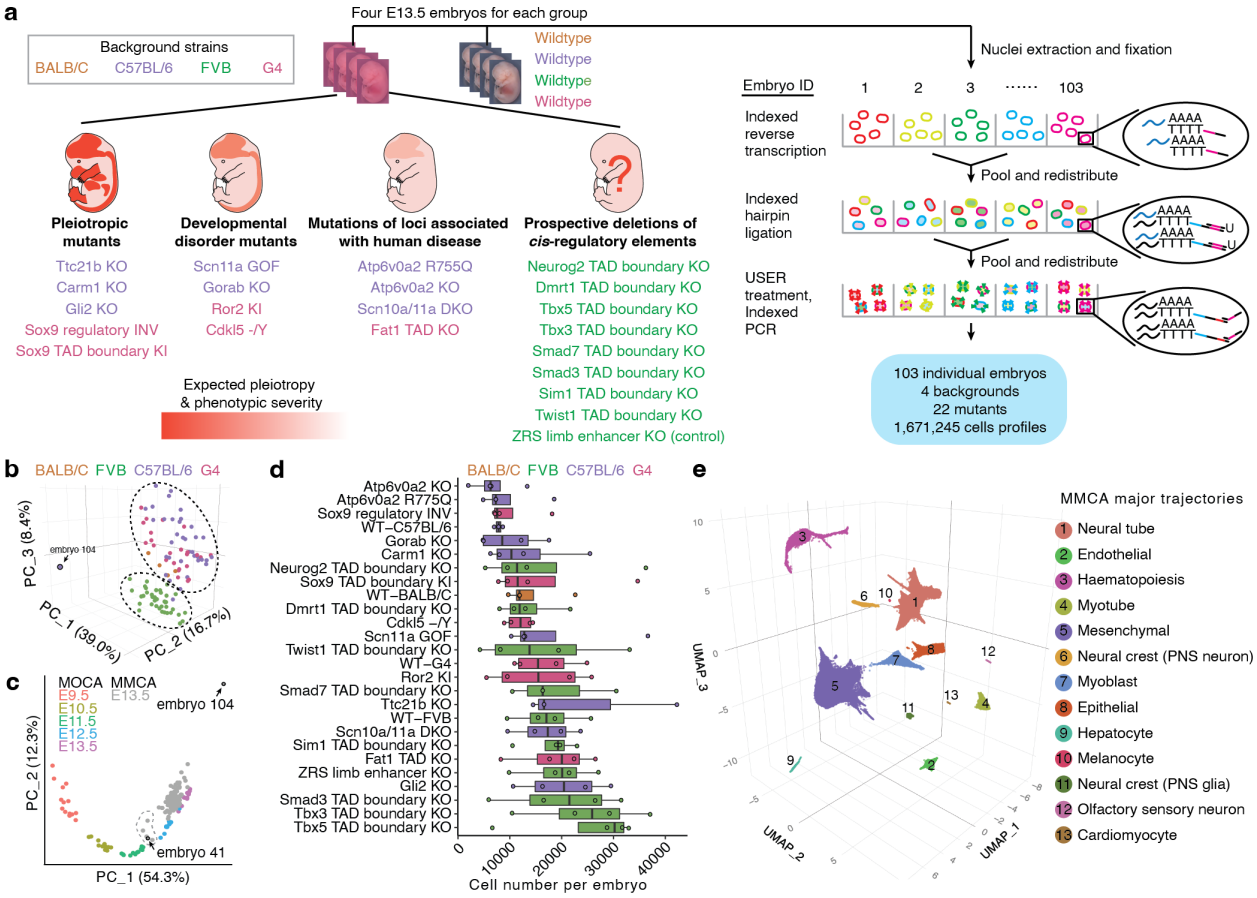

Figure 4.1: **Figure 1. Single cell transcriptional profiling of 103 whole mouse embryos staged at E13.5. a,** Categories of mutants (left) analyzed by whole embryo profiling with sci-RNA-seq3 (right). **b,** Embeddings of pseudobulk RNA-seq profiles of MMCA embryos in PCA space with visualization of top three PCs. Embryos are colored by background strain. **c,** Embeddings of pseudobulk RNA-seq profiles of MOCA[15] and MMCA embryos in PCA space defined solely by MOCA, with MMCA embryos (gray) projected onto it. Top two PCs are visualized. colored points correspond to MOCA embryos of different stages (E9.5-E13.5), and gray points to MMCA embryos (E13.5). **d,** Number of cells profiled per embryo for each strain. Genotypes are listed by median cell number in ascending order. **e,** 3D UMAP visualization of wildtype subset of MMCA dataset (215,575 cells from 15 embryos). Cells colored by major trajectory annotation.

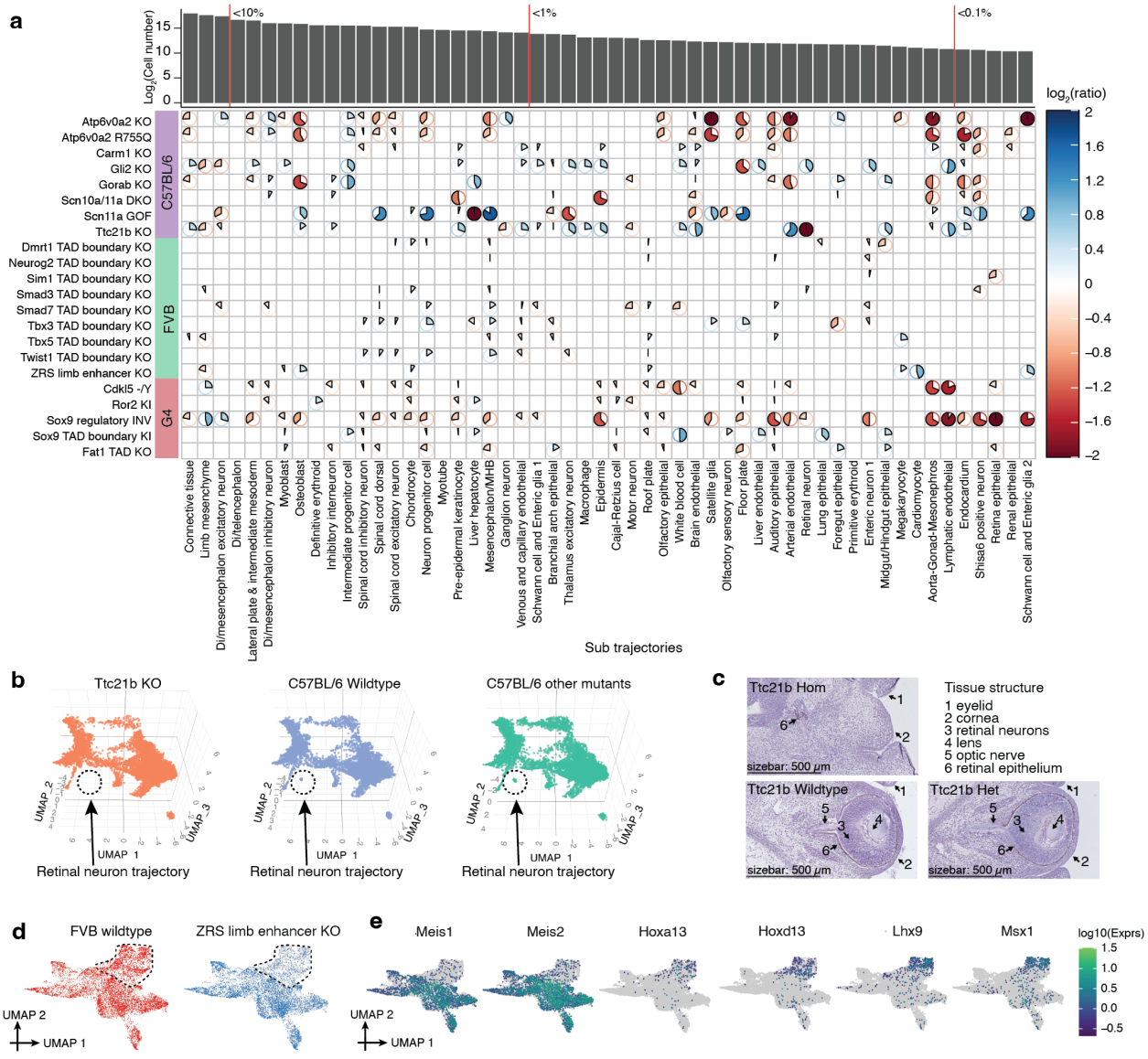

Figure 4.2: **Figure 2. Cell composition changes for individual mutants across developmental trajectories. a,** Heatmap shows log2-transformed ratios of cell proportions between each mutant type (y-axis) vs. pooled reference across individual sub-trajectories (x-axis). Pie color and direction correspond to log2-transformed ratio, while pie size and color intensity correspond to scale of log2-transformed ratio. **b,** 3D UMAP visualization of neural tube trajectory, highlighting cells from either *Ttc21b* KO (left), C57BL/6 wildtype (middle), or other C57BL/6 mutants (right). **c,** H&E staining of the developing eye of homozygous (Hom), heterozygous (Het) and wildtype *Ttc21b* E13.5 embryos. Structures are lost exclusively in homozygous mutants. **d,** UMAP visualization of co-embedded cells of limb mesenchyme trajectory from ZRS limb enhancer KO and FVB wildtype. **e,** Same UMAP as in panel d, colored by expression of marker genes which appear specific to proximal (*Meis1*, *Meis2*) or distal (*Hoxa13*, *Hoxd13*, *Lhx9*, *Msx1*) limb development.

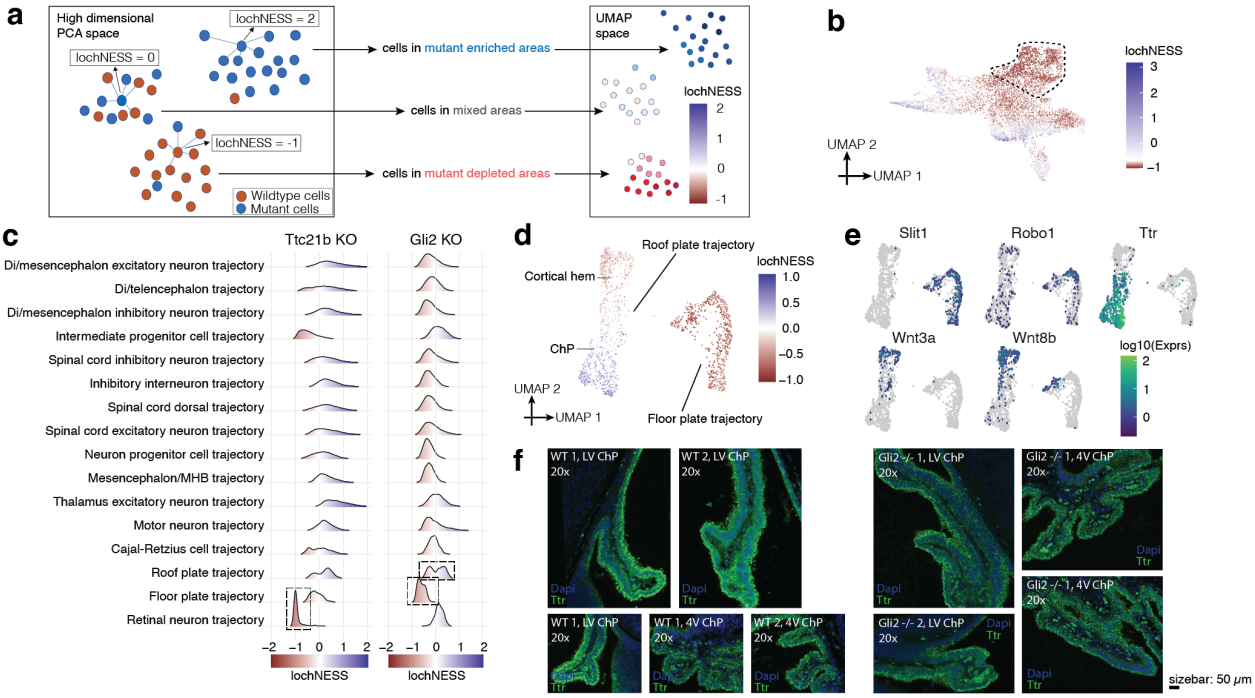

Figure 4.3: **Figure 3. LochNESS highlights mutant-related changes. a,** Schematic of lochNESS calculation and visualization. **b,** UMAP of limb mesenchyme trajectory from ZRS limb enhancer KO and wildtype cells, colored by lochNESS. Color scale centered at the median. Cells corresponding to a subset of ZRS limb enhancer KO cells with more extreme loss in **Fig. 2d** are circled. **c,** Distribution of lochNESS in the neural tube sub-trajectories of *Ttc21b* KO and *Gli2* KO mutants. Dashed boxes highlight shifted distributions of retinal neuron sub-trajectory of *Ttc21b* KO mutant and floor and roof plate sub-trajectories of *Gli2* KO mutant. **d,** UMAP of co-embedded cells of floor plate and roof plate sub-trajectories from *Gli2* KO mutant and pooled wildtype, colored by lochNESS. ChP, choroid plexus. **e,** same as panel d, but colored by selected marker gene expression. **f,** Immunofluorescence staining of *Ttr* (choroid plexus marker) in brain regions (LV = lateral ventricle, 4V = 4th ventricle) in sections from wildtype and *Gli2* KO mutant (**Methods**).

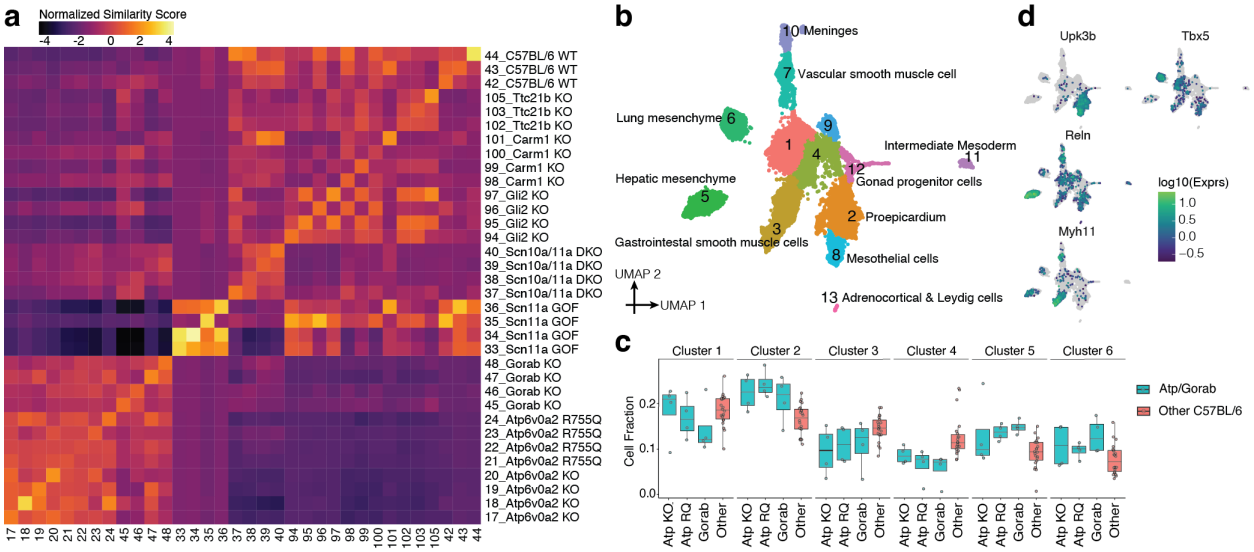

Figure 4.4: **Figure 4. Similarity scores identify mutant-shared and mutant-specific effects. a,** Heatmap showing similarity scores between individual C57BL/6 embryos in mesenchymal trajectory. Rows and columns grouped by genotype and labeled by embryo id and genotype. **b,** UMAP of lateral plate and intermediate mesoderm sub-trajectory for mutants from C57BL/6 background strain, colored and labeled by subcluster and detailed cell type. **c,** Boxplots showing composition of top 6 subclusters for individual *Atp6v0a2* KO, *Atp6v0a2* R755Q and *Gorab* KO embryos (blue, n=4 each genotype) and other C57BL/6 embryos (red, n=23). **d,** same as panel b, but colored by log-transformed expression of selected marker genes.

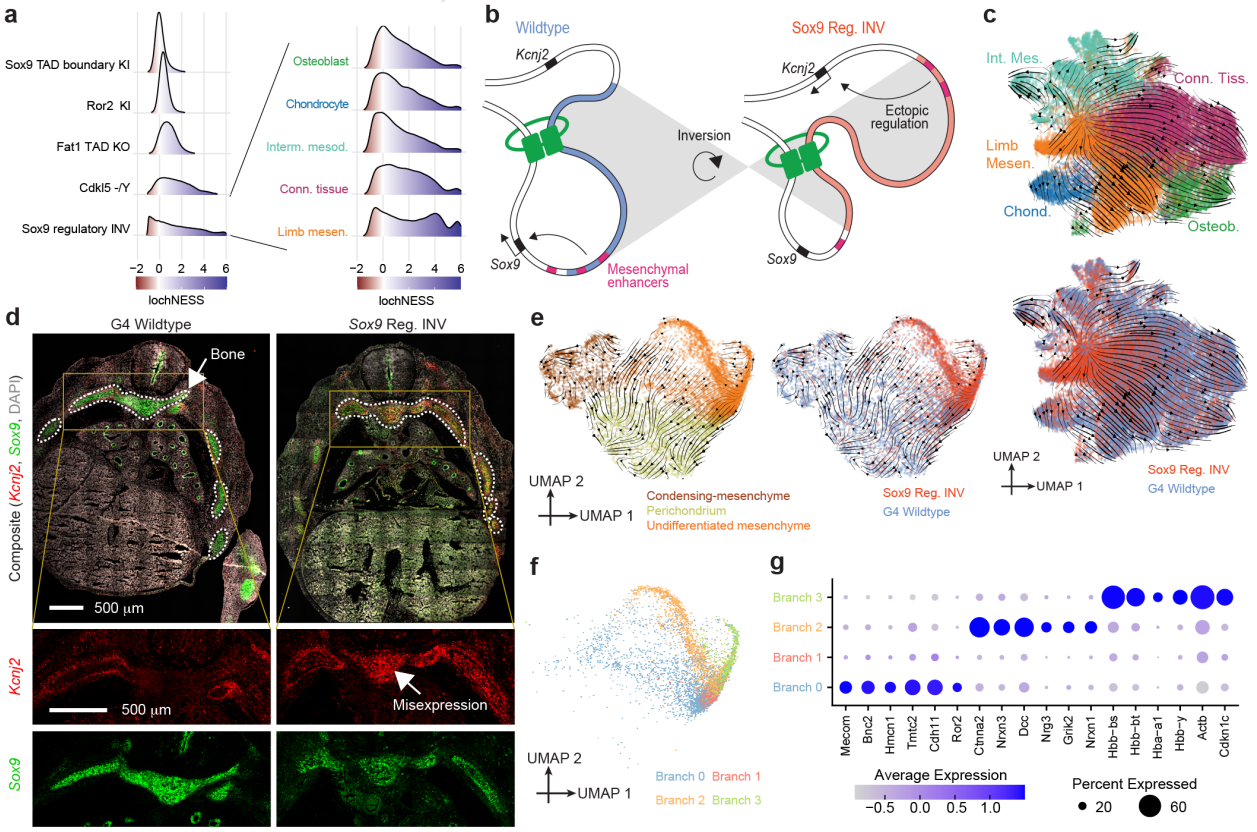

Figure 4.5: **Figure 5. Apparent stalling and redirection of mesenchyme differentiation in the Sox9 regulatory INV mutant. a,** LochNESS distributions for all G4 mutants in the mesenchymal trajectory (left) and *Sox9* regulatory INV mutant in mesenchymal sub-trajectories (right). **b,** Model of *Sox9* regulatory INV mutation depicting ectopic *Kcnj2* regulation through enhancer adoption. **c,** RNA velocity of mesenchymal G4 wildtype and *Sox9* regulatory INV cells colored by sub-trajectories (top) or genotype (bottom). **d,** *Sox9* regulatory INV heterozygous mutant and littermate wildtype RNA scope images (red: *Kcnj2*; green: *Sox9*), with insets below highlighting a region corresponding to developing bone (white circled area) **e,** RNA velocity of G4 wildtype and *Sox9* regulatory INV cells in the limb mesenchymal sub-trajectory labeled by annotation (left) or genotype (right). **f,** Same as penel e, but colored by branch number. **g,** Dot plot of top differentially expressed genes in the four branches shown in panel f.

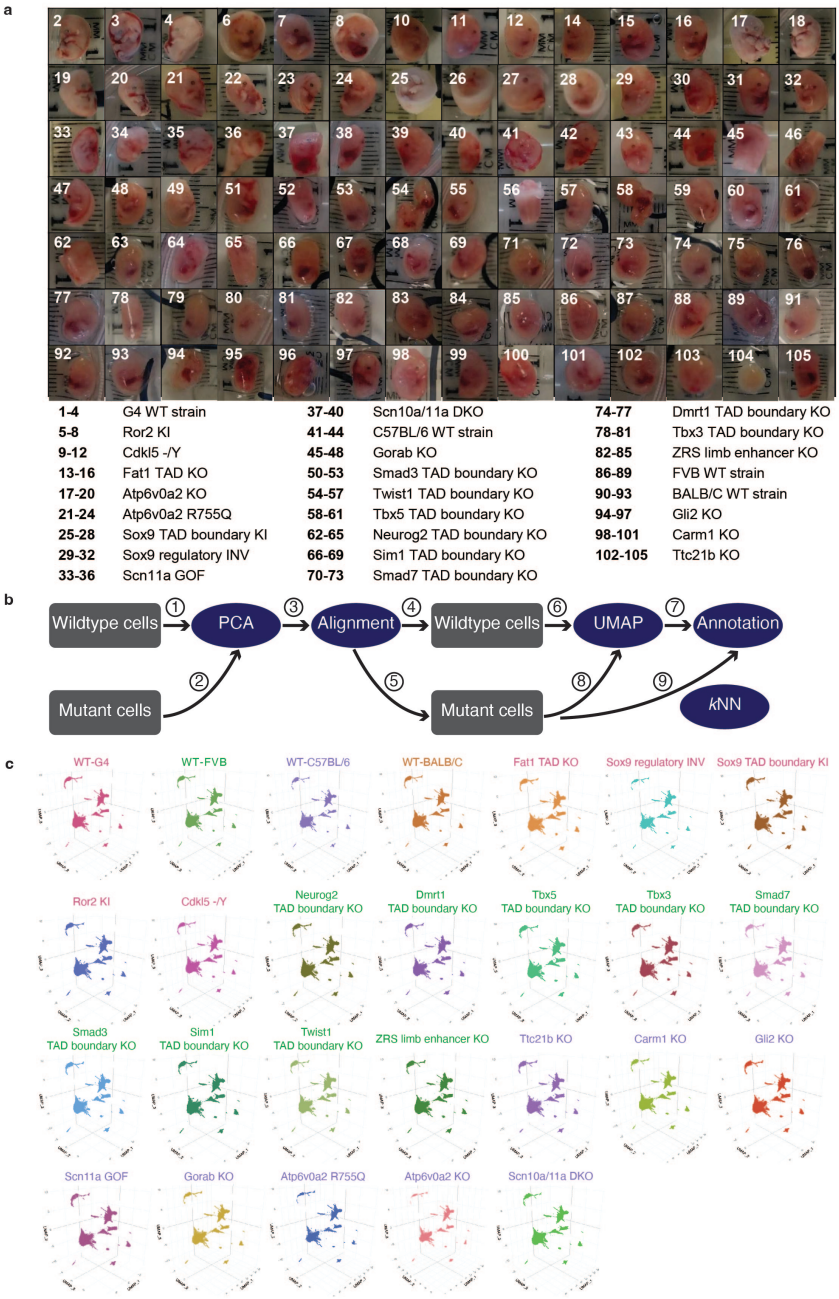

Figure 4.6: **Supplementary Figure 1. Images of mouse embryos and integrating cells derived from embryos of multiple genetic backgrounds to a single, wildtype-based "reference embedding". a,** 104 embryos (26 genotypes x 4 replicates) were staged at E13.5 and sent by five groups to a single site. **b,** Schematic of approach of dimensionality reduction and cell-type annotation. **c,** 3D UMAP visualizations of cells from each wildtype or mutant background within the shared "reference embedding" resulting from the aforedescribed procedures.

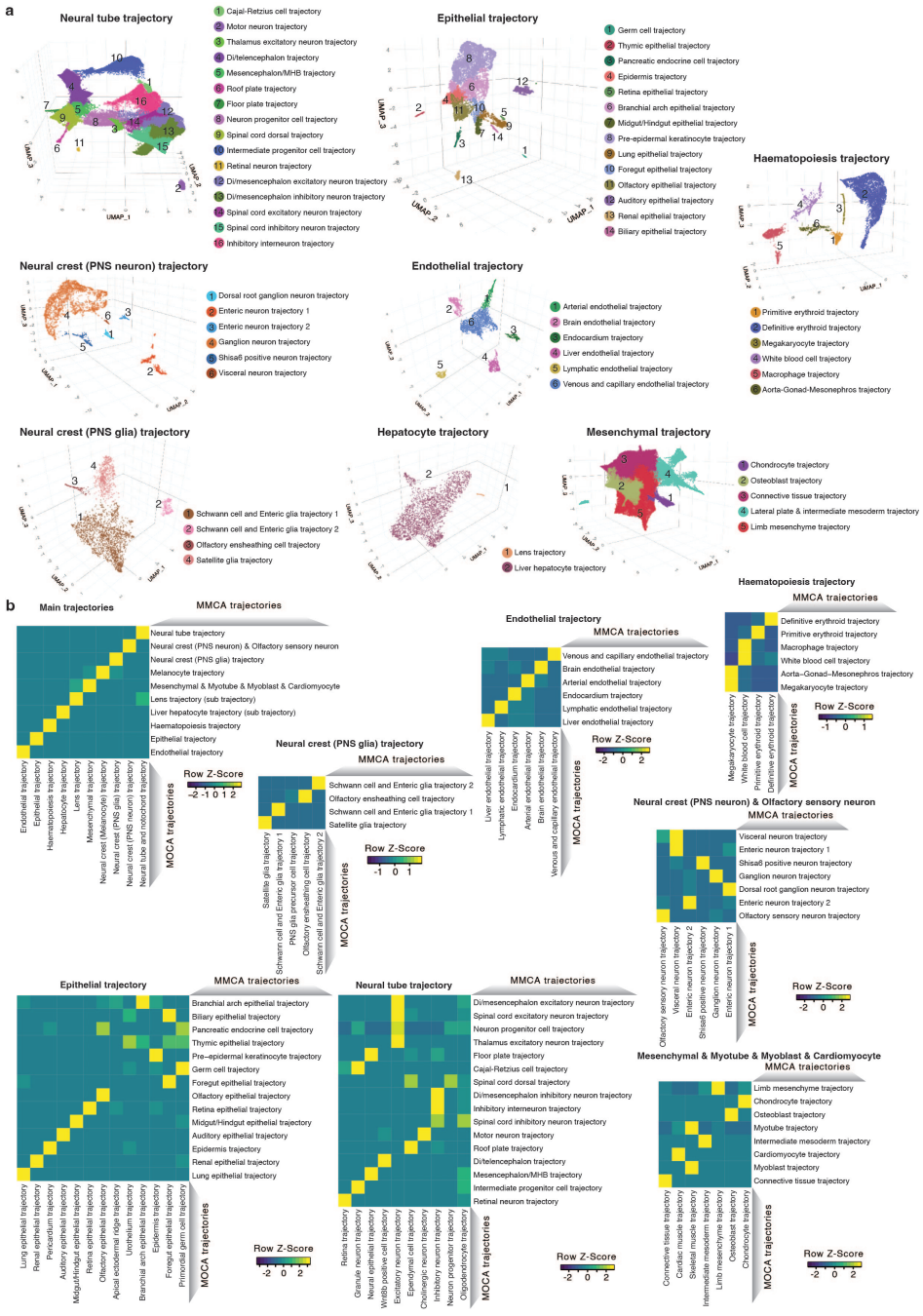

Figure 4.7: **Supplementary Figure 2. Annotation of sub-trajectories in data from wildtype E13.5 embryos and Correlated developmental major and sub-trajectories between MOCA (E9.5 - E13.5) and MMCA (E13.5 only) based on non-negative least-squares (NNLS) regression. a,** For 8 of these 13 major trajectories, iterative analysis identified the additional sub-trajectories shown here as 3D UMAP visualizations. **b,** Heat maps of the combined regression coefficients (row-scaled) between major or sub developmental trajectories from MMCA (rows) and corresponding developmental trajectories from the MOCA (columns).

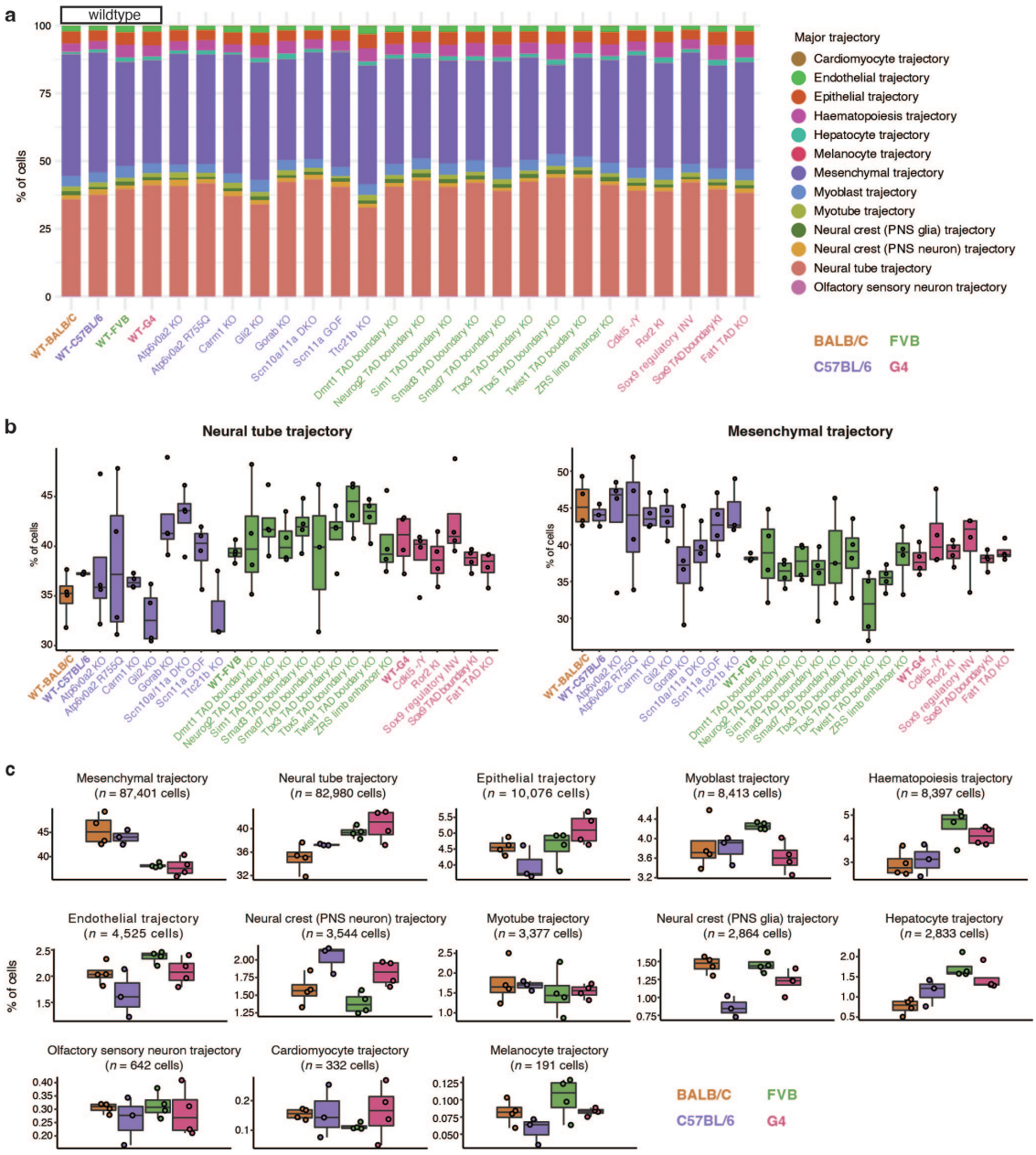

Figure 4.8: **Supplementary Figure 3. Cell composition for individual wildtype and mutant embryos across developmental trajectories. a,** Cell composition across 13 major trajectories of embryos from different wildtype or mutant strains. **b,** Boxplots of cell proportions falling into neural tube (left) or mesenchymal (right) trajectories for different wildtype or mutant strains. **c,** Boxplots of cell proportions falling into each of the 13 major trajectories for the four wildtype strains.

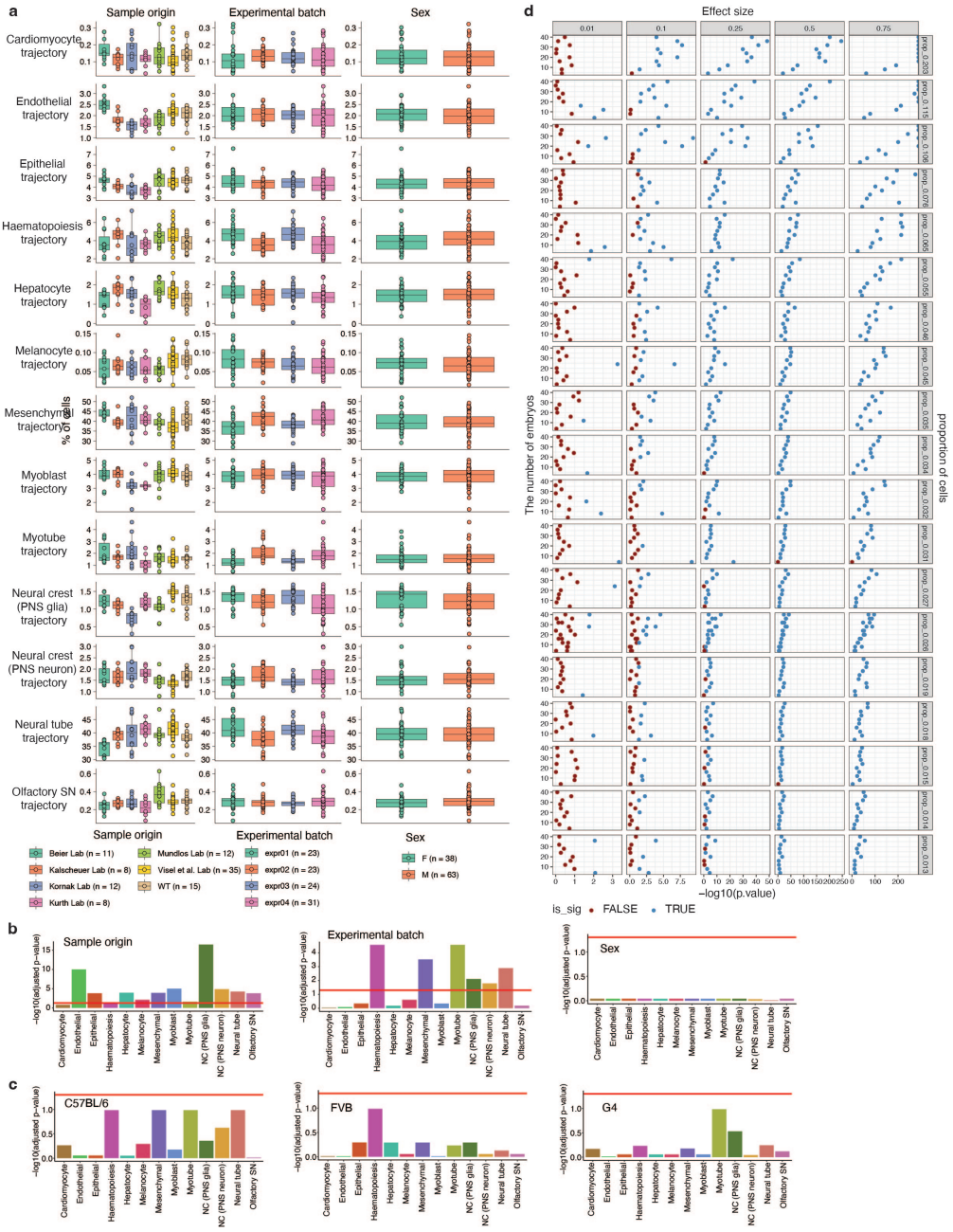

Figure 4.9: **Supplementary Figure 4. Cell composition for individual wildtype and mutant embryos across developmental trajectories, from different technical or biological groups and Simulation-based estimation of the number of replicates required to detect cell proportion changes. a,** Boxplots of cell proportions falling into each of the 13 major trajectories from different sample origins (left), experimental batch (middle), or sex (right). **b,** ANOVA was performed on cell proportions falling into each of the 13 major trajectories from different sample origins, experimental batch, or sex. **c,** ANOVA was performed on cell proportions falling into each of the 13 major trajectories from different experimental batches after subsetting samples from C57BL/6, FVB, or G4. **d,** Simulation-based estimation of the number of replicates required to detect cell proportion changes.

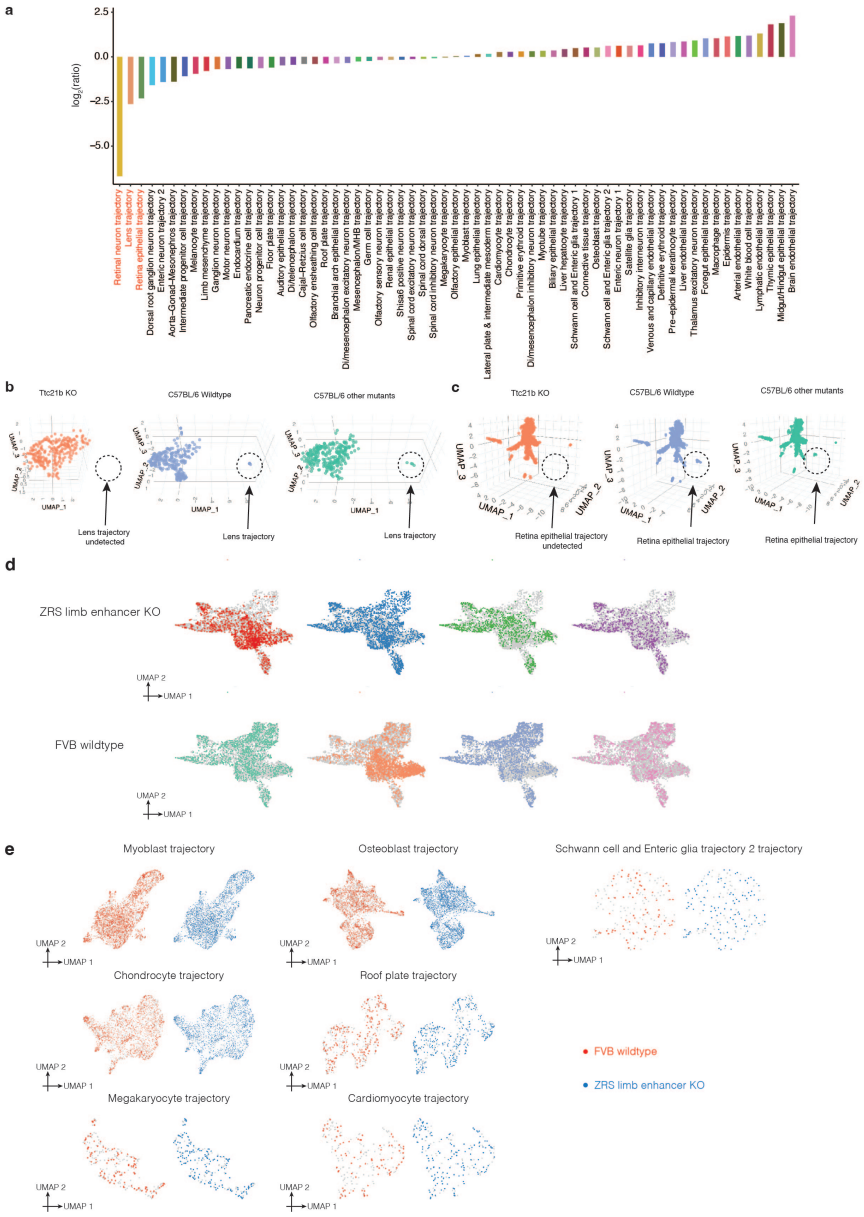

Figure 4.10: **Supplementary Figure 5. Multiple retinal trajectories are diminished in Ttc21b KO mice. a,** The log2 transformed ratio of the cell proportions of each sub-trajectory, comparing *Ttc21b* KO and C57BL/6 wildtype embryos, are shown. **b,** 3D UMAP visualization of the hepatocyte major trajectory, highlighting cells from either the *Ttc21b* KO, C57BL/6 wildtype, or other mutants on the C57BL/6 background. **c,** 3D UMAP visualization of the epithelial major trajectory, highlighting cells from either the *Ttc21b* KO, C57BL/6 wildtype, or other mutants on the C57BL/6 background. **d,** UMAP visualization of co-embedded cells of limb mesenchyme trajectory from the ZRS limb enhancer KO and FVB wildtype. **e,** UMAP visualization of co-embedded cells of various sub-trajectories from the ZRS limb enhancer KO and FVB wildtype.

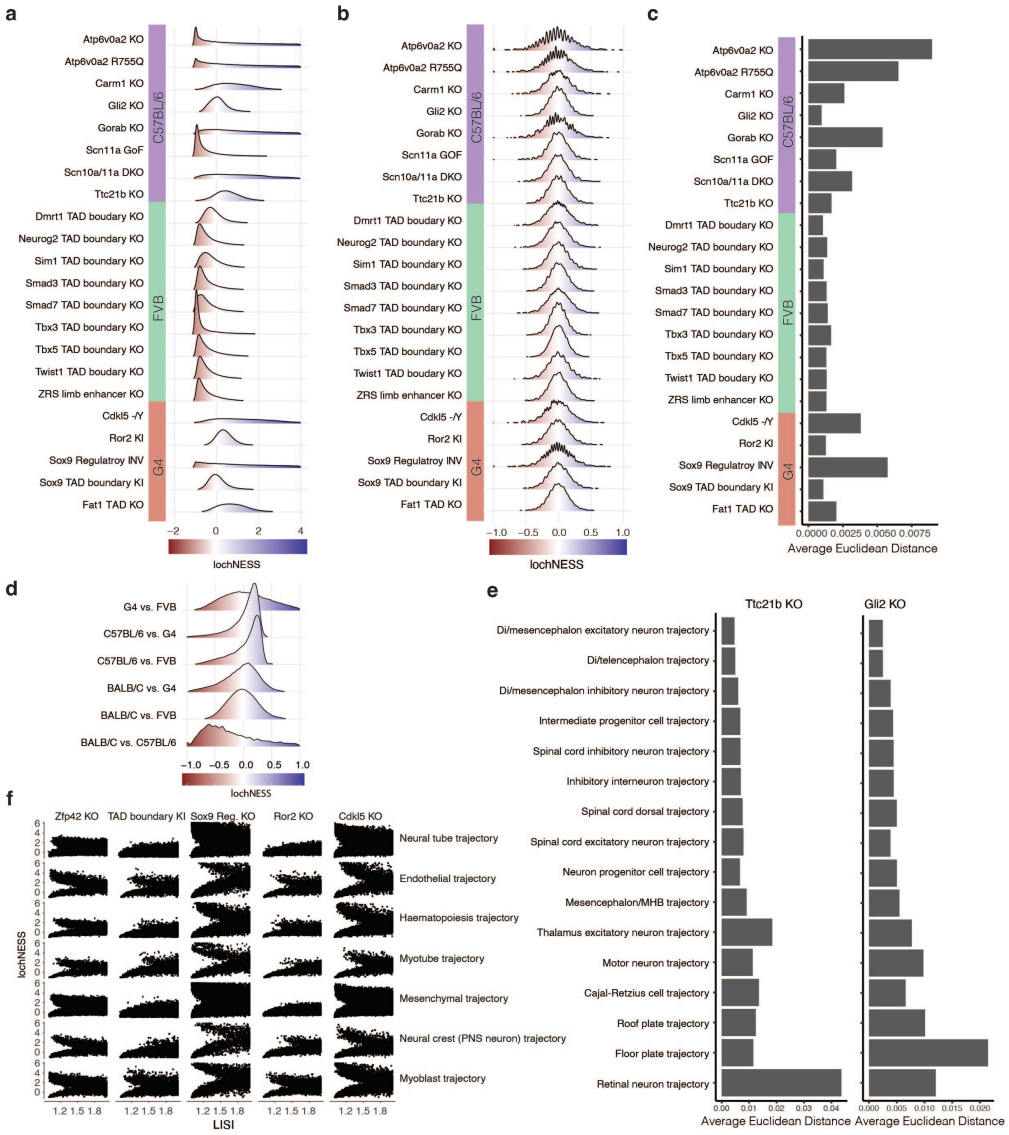

Figure 4.11: **Supplementary Figure 6. Quantitative analysis of lochNESS distributions. a,** Distribution of lochNESS across all 64 sub-trajectories in each mutant. **b,** Distribution of lochNESS in all cells of each mutant under random permutation of mutant labels. **c,** Barplot showing the average euclidean distance between lochNESS vs. lochNESS under permutation across all cells within a mutant. **d,** Estimated density graphs of lochNESS shows distribution of lochNESS in wildtype comparisons. **e,** Barplots showing the average euclidean distance between lochNESS and lochNESS under permutation, across all cells in neural tube sub-trajectories of the Ttc21b KO and Gli2 KO mutants. **f,** Scaterplots showing the concordance of lochNESS and LISI of cells from the G4 mutants in various main trajectories. More extreme lochNESS (indicating separation between mutant and wildtype) is associated with LISI scores approaching one (indicating non-mixing).

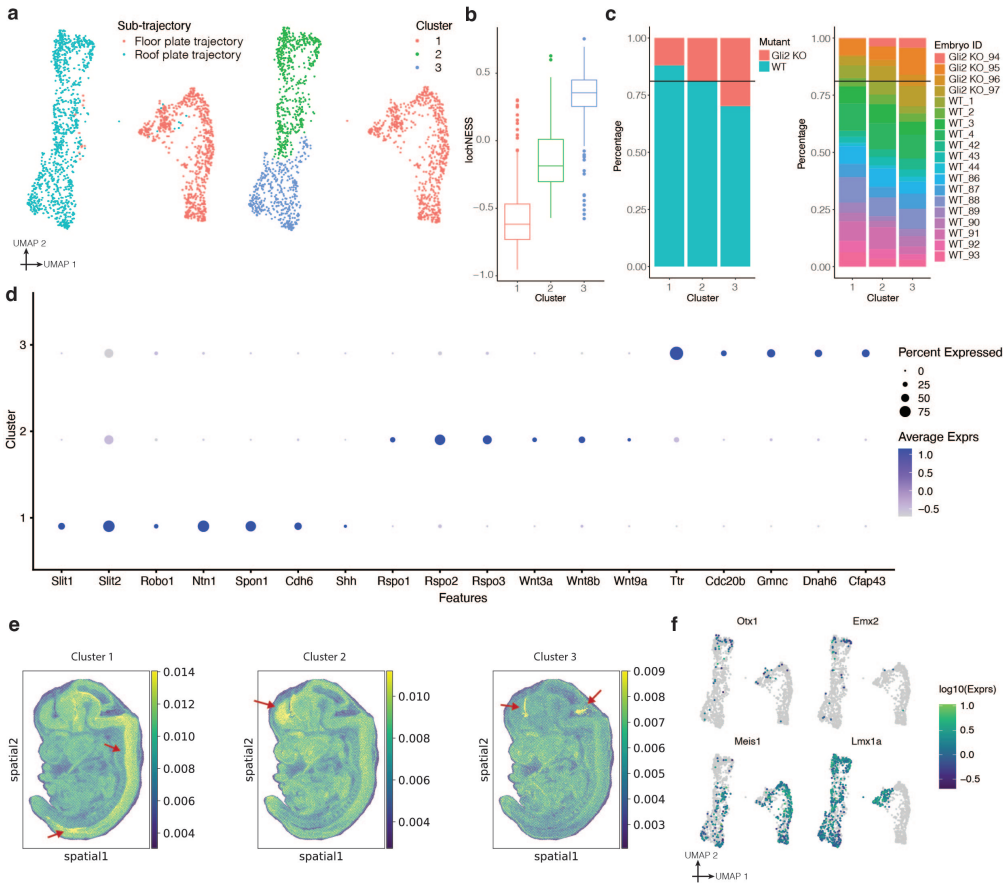

Figure 4.12: **Supplementary Figure 7. Analysis of Gli2 KO in the roof plate and floor plate trajectories. a,** UMAP visualization of co-embedded cells of the floor plate and roof plate sub-trajectories from the *Gli2* KO mutant and pooled wildtype. **b,** Boxplot showing the lochNESS distribution in each cluster shown on the right of panel a. **c,** Barplots showing the cell composition of each cluster shown on the right of panel a, split by mutant vs. wildtype (left) or individual embryo (right). **d,** Dotplot summarizing the expression of and percent of cells expressing selected marker genes in each cluster shown on the right of panel a. **e,** Tangram-inferred locations of each cluster shown on the right of panel a. **f,** UMAP visualization of co-embedded cells of the floor plate and roof plate sub-trajectories from the *Gli2* KO mutant and pooled wildtype, colored by expression of marker genes.

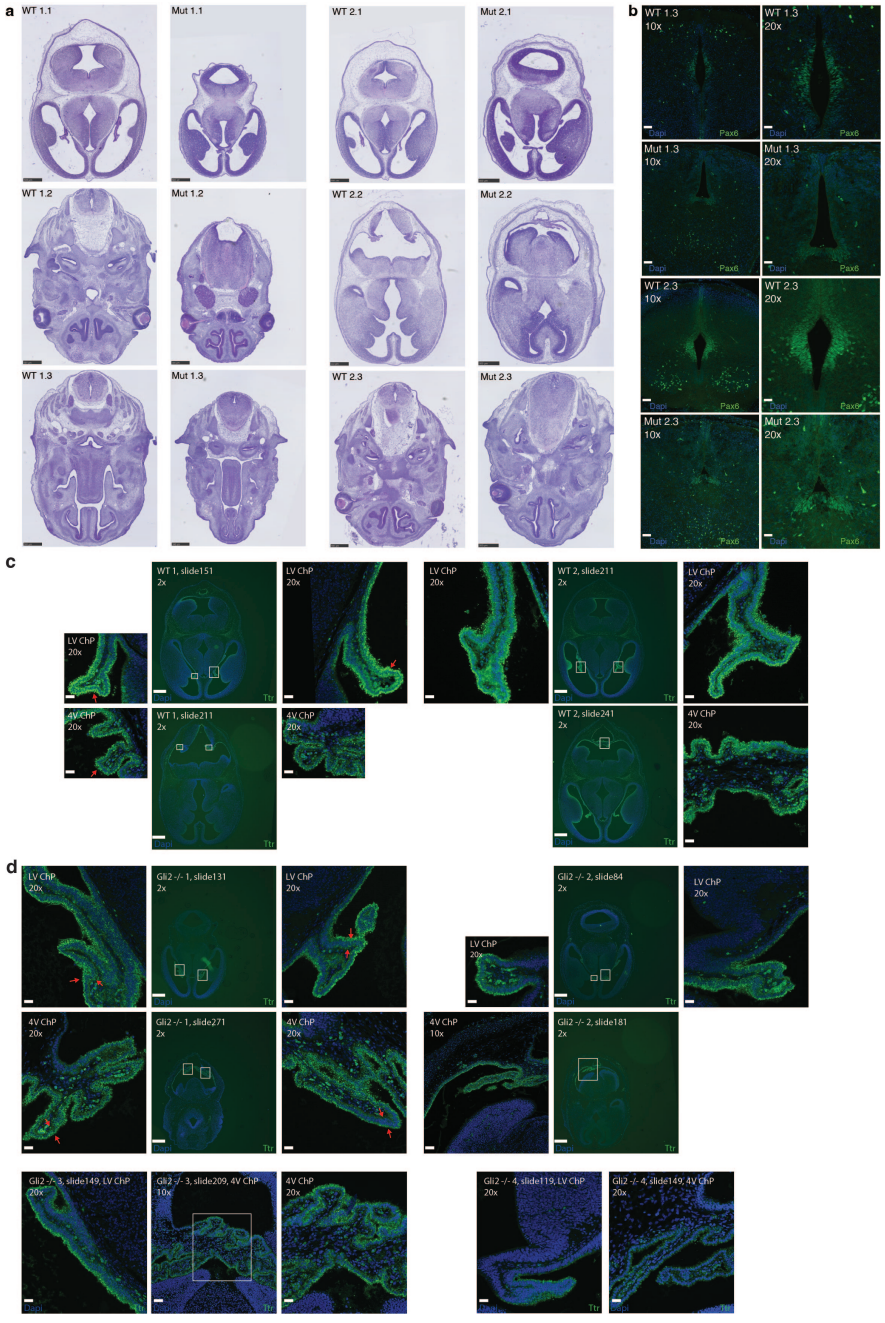

Figure 4.13: **Supplementary Figure 8. Morphological phenotype of Gli2 -/- mutants and Ttr staining in wildtype mice and Gli2 -/- mutants. a,** H&E staining of two mutant and two wild type E13.5 embryos in cranial-caudal (1-3) order within the head. **b,** Neural tube marker *Pax6* staining of the developing neural tube in consecutive sections 1.3 and 2.3 to visualize the structure of the neural tube formation in wildtype and mutant. *Ttr* staining of the developing brain regions (LV = lateral ventricle, 4V = 4th ventricle, ChP = choroid plexus) in sections of **c,** wildtype and **d,** *Gli2* -/- mutants.

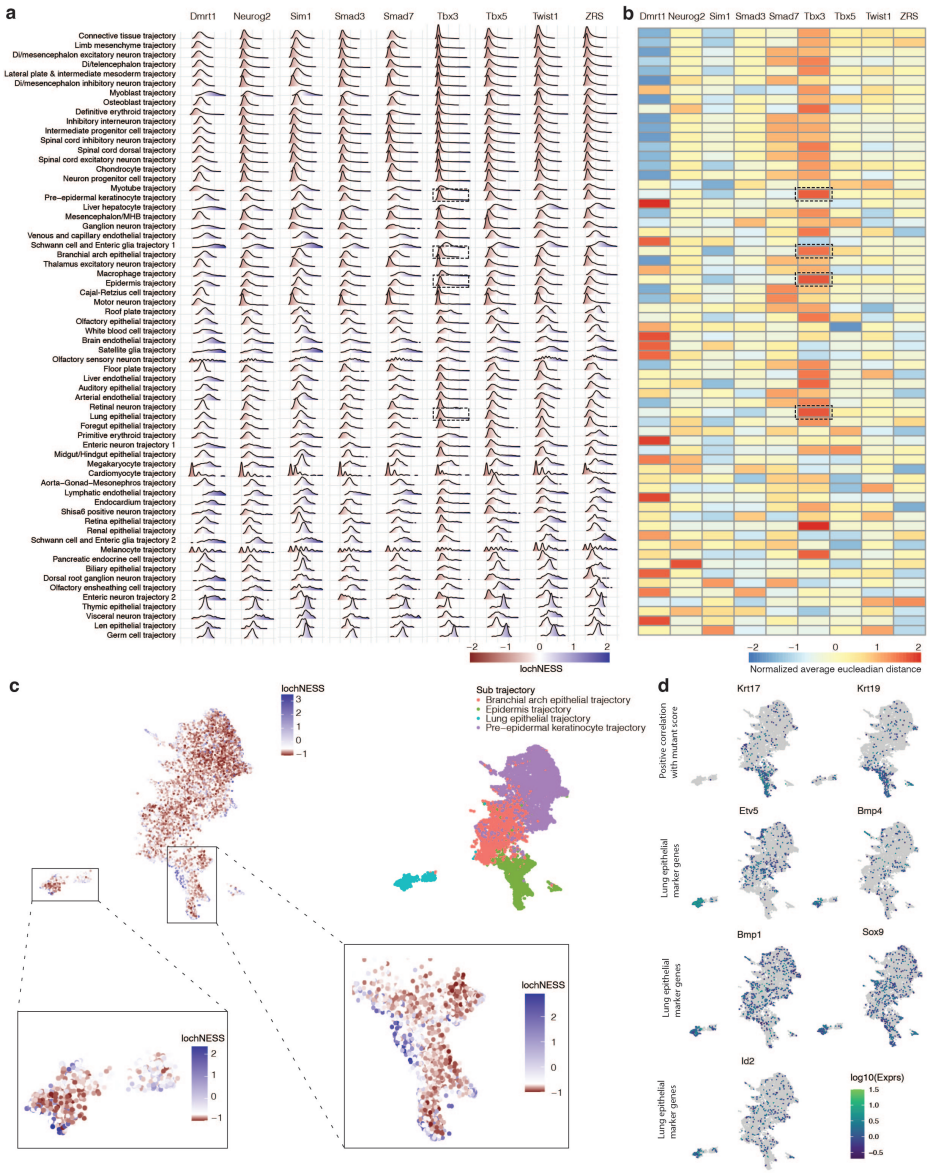

Figure 4.14: **Supplementary Figure 9. Systematic screening of lochNESS distributions identifies altered epithelial sub-trajectories in the Tbx3 TAD Boundary KO mutant. a,** Distribution of lochNESS in each sub-trajectory of the mutants in the FVB background strain, all of which are TAD boundary KOs. **b,** Row normalized heatmap showing the average euclidean distance between lochNESS and lochNESS under permutation in each sub-trajectory for the same mutants shown in panel a. **c,** UMAP showing coembedding of *Tbx3* TAD Boundary KO and pooled wildtype cells in the pre-epidermal keratinocyte, epidermis, branchial arch, and lung epithelial sub-trajectories. **d,** same as in panel c, but colored by expression of selected mutant related genes and marker genes.

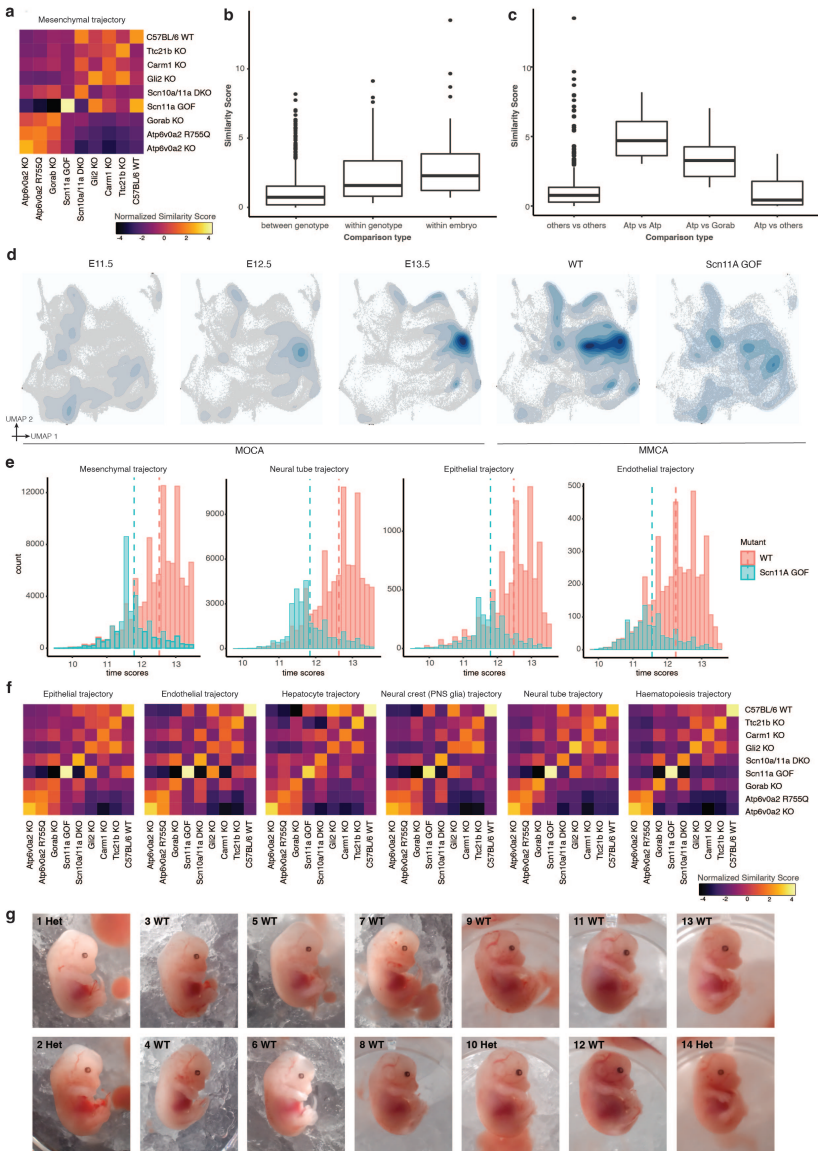

Figure 4.15: **Supplementary Figure 10. Similarity scores reveal mutant-shared and mutant-specific effects. a,** Heatmap showing similarity scores between C57BL/6 genotypes in the mesenchymal trajectory. **b,** Boxplot showing the similarity scores of comparisons between embryos of different genotypes, between embryos of the same genotype, and within the same embryos for C57BL/6 genotypes in the mesenchymal trajectory. **c,** Boxplot showing the similarity scores of comparisons between *Atp6v0a2* KO vs. *Atp6v0a2* R755Q, *Atp6v0a2* KO or *Atp6v0a2* R755Q vs. *Gorab* KO, *Atp6v0a2* KO or *Atp6v0a2* R755Q vs. other C57BL/6 genotypes, in the mesenchymal trajectory. **d,** UMAPs showing co-embedding of *Scn11a* GOF cells with pooled wildtype cells and E11.5-E13.5 MOCA cells, in the neural tube trajectory, split by mutant (MMCA) and time point (MOCA). **e,** Barplots showing the distribution of "time scores" for *Scn11a* GOF cells and pooled wildtype cells in selected main trajectories. **f,** Heatmaps showing similarity scores between C57BL/6 genotypes in selected main trajectories. **g,** *Scn11a* mutant and wildtype morphology comparison.

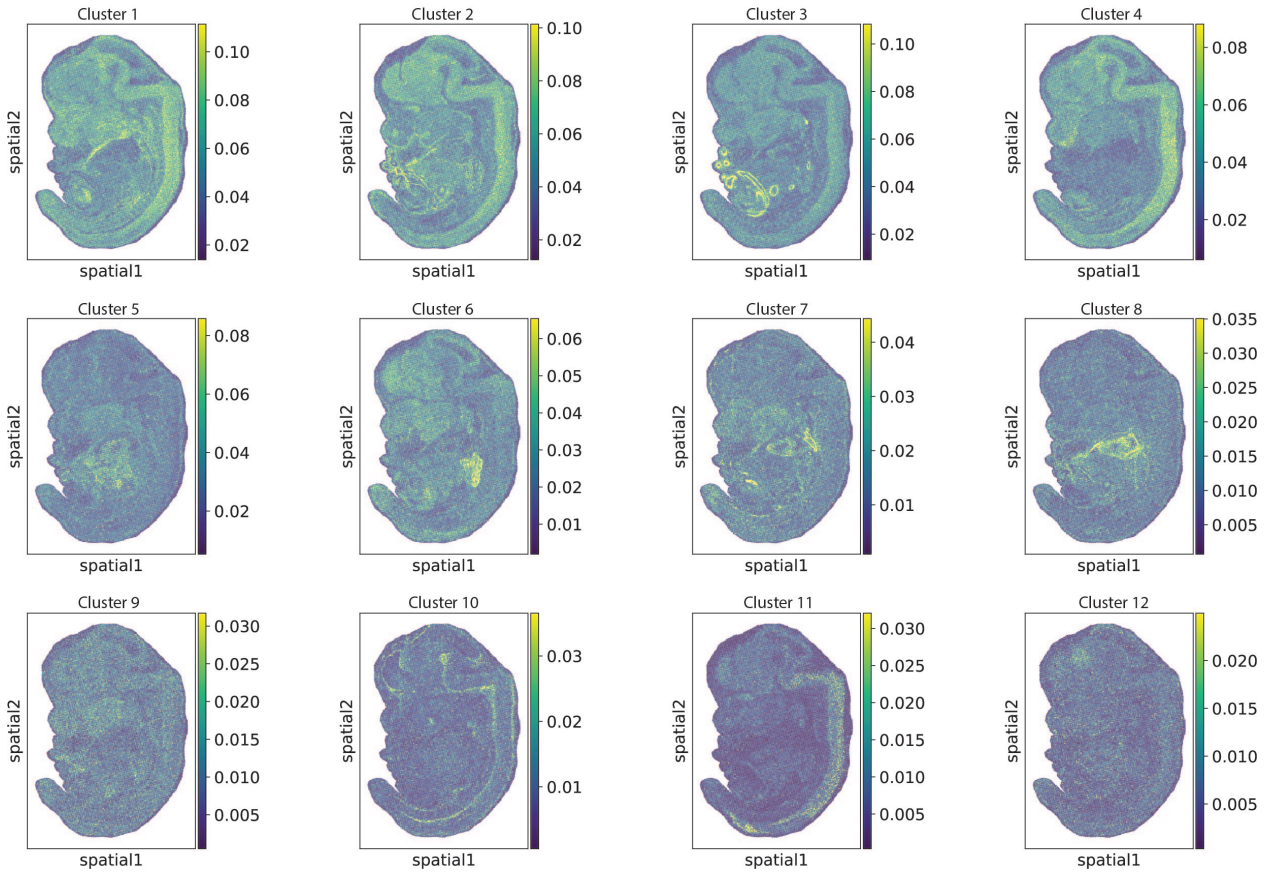

Figure 4.16: **Supplementary Figure 11. Spatial mapping of lateral plate and intermediate mesoderm sub-clusters.** Spatial mapping results by Tangram showing the most likely physical location of the cells from each cluster in the lateral plate and intermediate mesoderm sub-trajectory on a sagittal mouse section. Top 12 sub-clusters are shown. The color scale is set from 1st percentile to 99th percentile.

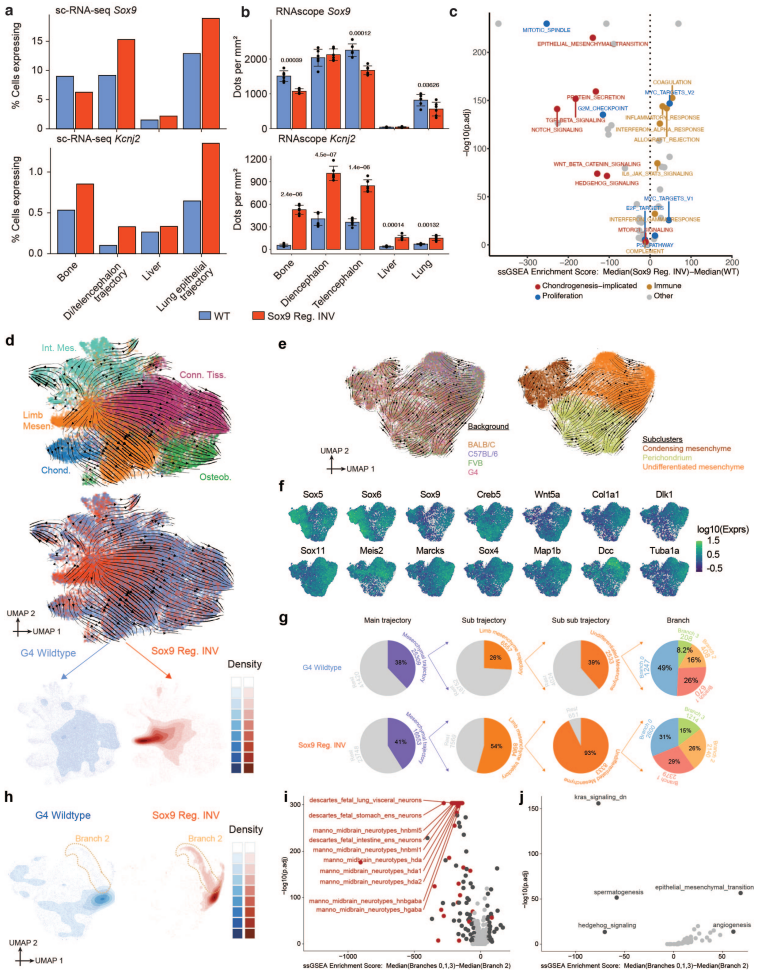

Figure 4.17: **Supplementary Figure 12. Misregulation of Sox9 and Kcnj2, and stalling of cells in the undifferentiated mesenchyme in the Sox9 regulatory INV mutant. a,** Quantification of *Sox9* and *Kcnj2* expression (scRNA-seq) in selected trajectories. For "bone" and "liver", multiple sub-trajectories were pooled to match the tissue labels in the RNAscope data in panel b. **b,** Quantification of *Sox9* and *Kcnj2* expression based on RNAscope images of heterozygous E13.5 wildtype and *Sox9* regulatory INV mutant embryos. **c,** Gene set enrichment analysis on bone cells. **d,** RNA velocity of mesenchymal G4 wildtype and *Sox9* regulatory INV cells labeled by sub-trajectories or genotype and the corresponding 2D density plots. **e,** Sub-clustering of the limb mesenchyme sub-trajectory based on cells from pooled wildtype. **f,** Marker gene expression used to annotate limb mesenchyme sub-clusters. **g,** Proportion and the number of cells at different levels of clustering, leading up to the four branches of the undifferentiated mesenchyme. **h,** Density plots for UMAP embedding of G4 wildtype and *Sox9* regulatory INV cells in the limb mesenchymal trajectory. Comparison of the ssGSEA scores between the two branches of undifferentiated mesenchyme for *Sox9* regulatory INV cells for (**i**) cell type signature (C8) and (**j**) Hallmark gene sets.

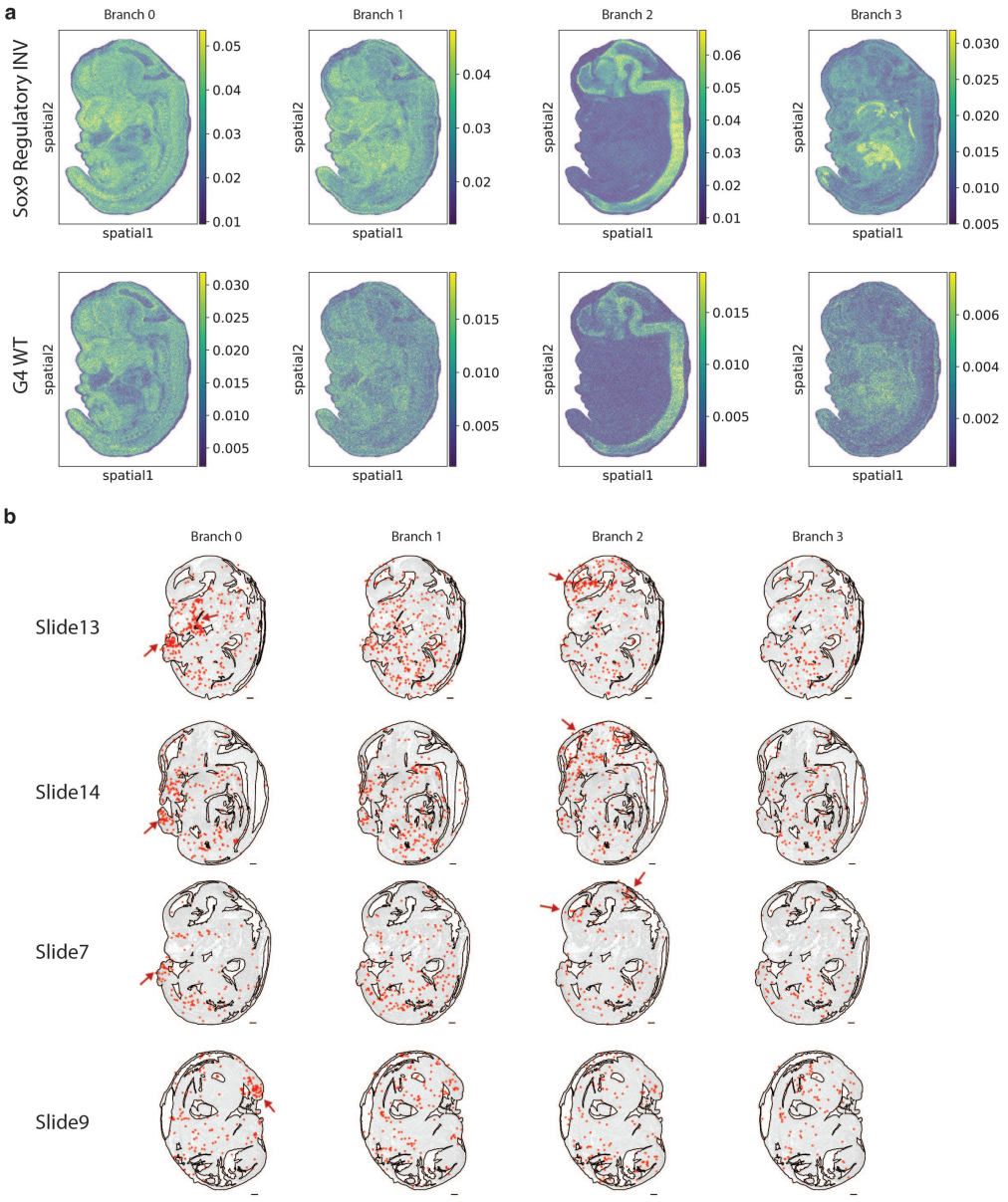

Figure 4.18: **Supplementary Figure 13. Spatial mapping of the cells of undifferentiated mesenchyme onto the Stereo-seq dataset and integration with the sci-space dataset. a,** Tangram inferred locations of cells from each branch shown in **Fig. 5f**, split by mutant (top) and wildtype (bottom) identity. **b,** Cells from the *Sox9* regulatory INV mutant assigned to the undifferentiated mesenchyme were integrated with a spatial transcriptomics dataset on mid-gestational mice (E14.5), generated via the sci-space method[267]. We find the nearest neighbor of each *Sox9* regulatory INV mutant cell in sci-space data in the integrated co-embedding, and plot the location of the neighboring sci-space cell where it is known (red dots).

Chapter 5
DISCUSSION AND FUTURE DIRECTION

Although the data and strategies presented in this thesis provide valuable insights into the cellular trajectories and key regulators during mammalian embryogenesis, they also have some shortcomings.

- The scRNA-seq data from different stages of development were generated from different studies using different technologies (gastrulation from Pijuan-Sala *et al.*, organogenesis and fetal development from Cao *et al.* and Qiu *et al.*). Batch correction is essential for integrating single-cell RNA-seq data sets, but it can be problematic because it may fail to remove all batch effects or it may remove some biological signal from the data.

- The early-stage scRNA-seq datasets were generated by pooling embryos together to obtain a sufficient number of cells. In contrast, the later-stage datasets were generated by profiling a single sample for each time point. Both approaches may have lost variation between replicates, making the results less robust.

- We have estimated the total number of cells in a mouse embryo at different timepoints in **Chapter 3**. These estimates suggest that our sampling strategy (*i.e.* the number of cells profiled at each timepoint) for the later stages may not have been sufficient to characterize the rare cell states and trajectories.

- The trajectory reconstruction strategy used in **Chapter 2** is not easily generalizable to more complex systems because it divides cells into different time points. This means that cells at a single time point may not be sufficient to identify all possible cell

states. Additionally, the strategy identifies trajectories between cell states in a global embedding, which may not be appropriate for cell states identified by subclustering.

- The trajectory reconstruction strategy used in **Chapter 3** did not account for the cellular heterogeneity and temporal heterogeneity within each single cell state node. Additionally, the strategy manually split cells into different subsystems, which may have resulted in the loss of trajectories that cross between systems (*e.g.* cross-germ layers).

The abrupt changes in gene expression that we observed in some tissues after birth (P0), including the liver, fat, and lung, have raised our interest in the developmental process from birth to adulthood. The development of a mouse after birth is a highly complex and coordinated process. It involves the interaction of many different genes and environmental factors. After birth, a mouse's brain, senses, muscles, immune system, and endocrine system continue to develop rapidly. This development includes the formation of new synapses and the refinement of neural circuits. The senses of sight, hearing, taste, smell, and touch also continue to develop, allowing the mouse to interact with its environment and learn about the world around it. The mouse gains control over its muscles and learns how to move around. It also starts to develop coordinated movements, such as walking and running. The immune system develops, allowing the mouse to fight off infections and disease. The endocrine system starts to function, producing hormones that regulate growth, development, and metabolism. To better understand how mice develop during this period, my colleagues and I have started collecting and staging embryos from birth to adulthood. We have adapted our sci-RNA-seq3 technology to work with larger samples, such as adult mice, allowing us to systematically profile the transcriptome of single cells from embryos collected at a series of timepoints after birth. This is particularly interesting as it allows us to study how the mouse embryo interacts with its environment during development. We could even extend this strategy to later stages, up to death, resulting in a complete developmental timeline from fertilization (E0) to death

(D0).

In a recent review paper, Domcke *et al.* argue that the current approach to cell classification is inadequate [212]. They point out that the traditional approach, which is based on atlases or periodic tables, does not consider the developmental relationships between cell types. As a result, we lack a systematic way to organize and interpret the vast amounts of data that have been generated about single cells. This makes it difficult to discover new cell types and to study their development and functions. To address this problem, Domcke *et al.* propose a new approach based on a consensus ontogeny. A consensus ontogeny is a tree-like structure that represents the developmental relationships between all cell types in a given organism. This approach would provide a universal, stable, and extendable framework for the precise communication of scientific knowledge about cell types. The transcriptome profiles we are creating from samples collected from E0 to P0 (or even D0) provide an opportunity to make this perspective a reality. By systematically profiling the transcriptome of single cells from different developmental stages, we can create a comprehensive map of the developmental relationships between cell types.

Most of the trajectories described in this thesis were reconstructed computationally using scRNA-seq data. The key regulators were identified based on the differential gene expression between cell states along the trajectory. However, the true function of these regulators, especially transcription factors (TFs), is still not well-established, as expression does not always equal function. Therefore, data with chromatin accessibility from samples collected from a series of timepoints during mouse embryogenesis is essential to identify the true functional regulators. In Calderon *et al.*'s study in flies [6], they did not perform scRNA-seq and scATAC-seq on the same cells, but they were still able to explore the dynamics of enhancer usage and gene expression within and across lineages at the scale of minutes. Similarly, Domcke *et al.* integrated scRNA-seq and scATAC-seq with data collected from the paired organs of fetal humans, identifying the relationship between TF expression and

activity across organs [120]. Our next step is to collect and profile the chromatin accessibility of single cells from the corresponding timepoints during mouse organogenesis, and then integrate these data with scRNA-seq to identify the key functional regulators.

The sampling strategy based on profiling single-cell transcriptomes is an indirect way to identify developmental trajectories. It would be better to combine it with other direct strategies, such as cell lineage tracing. In the past two years, new technologies have been developed that may overcome their original limitations. For example, new methods for genome editing are being developed for cell lineage tracing that can create more stable barcodes that can be tracked for longer periods of time [213, 214]. Additionally, new light-sheet microscopes are being developed that are capable of imaging larger and more complex embryos [278, 279, 280]. Choi *et al.* developed DNA Ticker Tape, a DNA-based memory device that can record events in sequence [214]. It overcomes the limitations of previous systems by recording multiple signals simultaneously and capturing the order of events. Maizels *et al.* developed sci-FATE2, an optimized metabolic labeling method for studying the dynamics of cell fate determination, by which they profiled approximately 45,000 embryonic stem cells differentiating into multiple neural tube identities [281]. These advanced technologies, together with single-cell transcriptome data and other omics data, provide new insights into our understanding of mammalian development.

The advent of novel technologies, particularly single-cell-based methods, has opened up new avenues for investigating the intricacies of embryo development, shedding light on cell identity and cell fate during embryogenesis. The massive amount of single-cell data generated by various studies, some of which may overlap in terms of coverage and even contradict each other on their findings, has shifted our research focus from data generation to data interpretation. Integrating heterogeneous data and systematically extracting valuable and accurate information from this vast trove of data has become a critical next step. Geneformer, for instance, is a revolutionary deep learning model that has transformed the analysis of various

biological data [215]. Trained on a vast collection of single-cell transcriptomes, Geneformer boasts an exceptional ability to unravel intricate gene regulatory networks, even in situations with limited data. By fine-tuning Geneformer for specific tasks, researchers have unlocked its potential to predict chromatin and network dynamics, and even pinpoint promising therapeutic targets for cardiomyopathy. This work, along with other innovative computational methods powered by deep learning models, pave the way for future advancements in the field.

BIBLIOGRAPHY

[1] Domingos Henrique et al. “Neuromesodermal progenitors and the making of the spinal cord”. en. In: *Development* 142.17 (Sept. 2015), pp. 2864–2875.

[2] J E Sulston et al. “The embryonic cell lineage of the nematode Caenorhabditis elegans”. en. In: *Dev. Biol.* 100.1 (Nov. 1983), pp. 64–119.

[3] Aaron McKenna et al. “Whole-organism lineage tracing by combinatorial and cumulative genome editing”. en. In: *Science* 353.6298 (July 2016), aaf7907.

[4] Katie McDole et al. “In Toto Imaging and Reconstruction of Post-Implantation Mouse Development at the Single-Cell Level”. en. In: *Cell* 175.3 (Oct. 2018), 859–876.e33.

[5] Jonathan S Packer et al. “A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution”. en. In: *Science* 365.6459 (Sept. 2019).

[6] Diego Calderon et al. “The continuum of Drosophila embryonic development at single-cell resolution”. en. In: *Science* 377.6606 (Aug. 2022), eabn5800.

[7] Jeffrey A Farrell et al. “Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis”. en. In: *Science* 360.6392 (June 2018).

[8] Daniel E Wagner et al. “Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo”. en. In: *Science* 360.6392 (June 2018), pp. 981–987.

[9] James A Briggs et al. “The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution”. en. In: *Science* 360.6392 (June 2018).

[10] Lauren M Saunders et al. “Deep molecular, cellular and temporal phenotyping of developmental perturbations at whole organism scale”. en. Aug. 2022.

[11] Merlin Lange et al. “Zebrahub – Multimodal Zebrafish Developmental Atlas Reveals the State-Transition Dynamics of Late-Vertebrate Pluripotent Axial Progenitors”. en. June 2023.

[12] Abhinav Sur et al. “Single-cell analysis of shared signatures and transcriptional diversity during zebrafish development”. en. Apr. 2023.

[13] Blanca Pijuan-Sala et al. “A single-cell molecular map of mouse gastrulation and early organogenesis”. en. In: *Nature* 566.7745 (Feb. 2019), pp. 490–495.

[14] Markus Mittnenzweig et al. “A single-embryo, single-cell time-resolved model for mouse gastrulation”. en. In: *Cell* 184.11 (May 2021), 2825–2842.e22.

[15] Junyue Cao et al. “The single-cell transcriptional landscape of mammalian organogenesis”. en. In: *Nature* 566.7745 (Feb. 2019), pp. 496–502.

[16] Yichi Xu et al. “A single-cell transcriptome atlas profiles early organogenesis in human embryos”. en. In: *Nat. Cell Biol.* 25.4 (Apr. 2023), pp. 604–615.

[17] Bo Zeng et al. “The single-cell and spatial transcriptional landscape of human gastrulation and early brain development”. en. In: *Cell Stem Cell* 30.6 (June 2023), 851–866.e7.

[18] Junyue Cao et al. “A human cell atlas of fetal gene expression”. en. In: *Science* 370.6518 (Nov. 2020).

[19] Richard C V Tyser et al. “Characterization of a common progenitor pool of the epicardium and myocardium”. en. In: *Science* 371.6533 (Mar. 2021).

[20] Gioele La Manno et al. “Molecular architecture of the developing mouse brain”. en. In: *Nature* 596.7870 (Aug. 2021), pp. 92–96.

[21] Leland McInnes, John Healy, and James Melville. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”. In: (Feb. 2018). eprint: 1802.03426.

[22] Tara Chari and Lior Pachter. “The Specious Art of Single-Cell Genomics”. en. Dec. 2022.

[23] F Alexander Wolf et al. “PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells”. en. In: *Genome Biol.* 20.1 (Mar. 2019), p. 59.

[24] Gioele La Manno et al. “RNA velocity of single cells”. en. In: *Nature* 560.7719 (Aug. 2018), pp. 494–498.

[25] Geoffrey Schiebinger et al. “Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming”. en. In: *Cell* 176.4 (Feb. 2019), 928–943.e22.

[26] Dominik Klein et al. “Mapping cells through time and space with moscot”. en. May 2023.

[27] Samuel L Wolock, Romain Lopez, and Allon M Klein. “Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data”. en. In: *Cell Syst* 8.4 (Apr. 2019), 281–291.e9.

[28] Chee-Huat Linus Eng et al. “Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH”. en. In: *Nature* 568.7751 (Apr. 2019), pp. 235–239.

[29] Kok Hao Chen et al. “RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells”. en. In: *Science* 348.6233 (Apr. 2015), aaa6090.

[30] Samuel G Rodriques et al. “Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution”. en. In: *Science* 363.6434 (Mar. 2019), pp. 1463–1467.

[31] Marlon Stoeckius et al. “Simultaneous epitope and transcriptome measurement in single cells”. en. In: *Nat. Methods* 14.9 (Sept. 2017), pp. 865–868.

[32] Ao Chen et al. “Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays”. en. In: *Cell* 185.10 (May 2022), 1777–1792.e21.

[33] Guangdun Peng et al. “Molecular architecture of lineage allocation and tissue organization in early mouse embryo”. en. In: *Nature* 572.7770 (Aug. 2019), pp. 528–532.

[34] Vitalii Kleshchevnikov et al. “Cell2location maps fine-grained cell types in spatial transcriptomics”. en. In: *Nat. Biotechnol.* 40.5 (May 2022), pp. 661–671.

[35] Aaron M Newman et al. “Determining cell type abundance and expression from bulk tissues with digital cytometry”. en. In: *Nat. Biotechnol.* 37.7 (July 2019), pp. 773–782.

[36] Mor Nitzan et al. “Gene expression cartography”. en. In: *Nature* 576.7785 (Dec. 2019), pp. 132–137.

[37] Tommaso Biancalani et al. “Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram”. en. In: *Nat. Methods* 18.11 (Nov. 2021), pp. 1352–1362.

[38] Ricard Argelaguet et al. “Multi-omics profiling of mouse gastrulation at single-cell resolution”. en. In: *Nature* 576.7787 (Dec. 2019), pp. 487–491.

[39] Blanca Pijuan-Sala et al. “Single-cell chromatin accessibility maps reveal regulatory programs driving early mouse organogenesis”. en. In: *Nat. Cell Biol.* 22.4 (Apr. 2020), pp. 487–497.

[40] Ioannis Sarropoulos et al. “Developmental and evolutionary dynamics of cis-regulatory elements in mouse cerebellar cells”. en. In: *Science* 373.6558 (Aug. 2021).

[41] Michael W Dorrity et al. “Proteostasis governs differential temperature sensitivity across embryonic cell types”. en. Aug. 2022.

[42] Emma Dann et al. “Differential abundance testing on single-cell data using k-nearest neighbor graphs”. en. In: *Nat. Biotechnol.* 40.2 (Feb. 2022), pp. 245–253.

[43] Daniel B Burkhardt et al. “Quantifying the effect of experimental perturbations at single-cell resolution”. en. In: *Nat. Biotechnol.* 39.5 (May 2021), pp. 619–629.

[44] Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. “scGen predicts single-cell perturbation responses”. en. In: *Nat. Methods* 16.8 (Aug. 2019), pp. 715–721.

[45] Michael A Skinnider et al. “Cell type prioritization in single-cell data”. en. In: *Nat. Biotechnol.* 39.1 (Jan. 2021), pp. 30–34.

[46] Yakir A Reshef et al. “Co-varying neighborhood analysis identifies cell populations associated with phenotypes of interest from single-cell transcriptomics”. en. In: *Nat. Biotechnol.* 40.3 (Mar. 2022), pp. 355–363.

[47] Stefan Peidli et al. “scPerturb: Information Resource for Harmonized Single-Cell Perturbation Data”. en. Aug. 2022.

[48] Margarida Cardoso-Moreira et al. “Gene expression across mammalian organ development”. en. In: *Nature* 571.7766 (July 2019), pp. 505–509.

[49] Susanne C van den Brink et al. “Single-cell and spatial transcriptomics reveal somitogenesis in gastruloids”. en. In: *Nature* 582.7812 (June 2020), pp. 405–409.

[50] Gianluca Amadei et al. “Embryo model completes gastrulation to neurulation and organogenesis”. en. In: *Nature* 610.7930 (Oct. 2022), pp. 143–153.

[51] Tim Stuart et al. “Comprehensive Integration of Single-Cell Data”. en. In: *Cell* 177.7 (June 2019), 1888–1902.e21.

[52] Krzysztof Polański et al. “BBKNN: fast batch alignment of single cell transcriptomes”. en. In: *Bioinformatics* 36.3 (Feb. 2020), pp. 964–965.

[53] Junyue Cao et al. “Comprehensive single-cell transcriptional profiling of a multicellular organism”. en. In: *Science* 357.6352 (Aug. 2017), pp. 661–667.

[54] Shangli Cheng et al. “Single-Cell RNA-Seq Reveals Cellular Heterogeneity of Pluripotency Transition and X Chromosome Dynamics during Early Mouse Development”. en. In: *Cell Rep.* 26.10 (Mar. 2019), 2593–2607.e3.

[55] Hisham Mohammed et al. “Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions during Mouse Early Gastrulation”. en. In: *Cell Rep.* 20.5 (Aug. 2017), pp. 1215–1228.

[56] Cole Trapnell et al. “The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells”. en. In: *Nat. Biotechnol.* 32.4 (Apr. 2014), pp. 381–386.

[57] Daniel E Wagner and Allon M Klein. “Lineage tracing meets single-cell omics: opportunities and challenges”. en. In: *Nat. Rev. Genet.* 21.7 (July 2020), pp. 410–427.

[58] Beth K Martin et al. “Optimized single-nucleus transcriptional profiling by combinatorial indexing”. en. In: *Nat. Protoc.* (Oct. 2022).

[59] Marysia Placzek and James Briscoe. “The floor plate: multiple cells, multiple signals”. en. In: *Nat. Rev. Neurosci.* 6.3 (Mar. 2005), pp. 230–240.

[60] K Dale et al. “Differential patterning of ventral midline cells by axial mesoderm is regulated by BMP7 and chordin”. en. In: *Development* 126.2 (Jan. 1999), pp. 397–408.

[61] M Sameer Rana et al. “Tbx1 coordinates addition of posterior second heart field progenitor cells to the arterial and venous poles of the heart”. en. In: *Circ. Res.* 115.9 (Oct. 2014), pp. 790–799.

[62] Chen-Leng Cai et al. “Isl1 identifies a cardiac progenitor population that proliferates prior to differentiation and contributes a majority of cells to the heart”. en. In: *Dev. Cell* 5.6 (Dec. 2003), pp. 877–889.

[63] Franziska Herrmann et al. “Tbx5 overexpression favors a first heart field lineage in murine embryonic stem cells and in Xenopus laevis embryos”. en. In: *Dev. Dyn.* 240.12 (Dec. 2011), pp. 2634–2645.

[64] Daniela Später et al. “A HCN4+ cardiomyogenic progenitor derived from the first heart field and human pluripotent stem cells”. en. In: *Nat. Cell Biol.* 15.9 (Sept. 2013), pp. 1098–1106.

[65] D G Wilkinson et al. “Segmental expression of Hox-2 homoeobox-containing genes in the developing mouse hindbrain”. en. In: *Nature* 341.6241 (Oct. 1989), pp. 405–409.

[66] C B Moens et al. “Equivalence in the genetic control of hindbrain segmentation in fish and mouse”. en. In: *Development* 125.3 (Feb. 1998), pp. 381–391.

[67] Lisa Maves, William Jackman, and Charles B Kimmel. “FGF3 and FGF8 mediate a rhombomere 4 signaling activity in the zebrafish hindbrain”. en. In: *Development* 129.16 (Aug. 2002), pp. 3825–3837.

[68] M Studer et al. “Genetic interactions between Hoxa1 and Hoxb1 reveal new roles in regulation of early hindbrain patterning”. en. In: *Development* 125.6 (Mar. 1998), pp. 1025–1036.

[69] B A Parr et al. “Mouse Wnt genes exhibit discrete domains of expression in the early embryonic CNS and limb buds”. en. In: *Development* 119.1 (Sept. 1993), pp. 247–261.

[70] M Sander et al. “Ventral neural patterning by Nkx homeobox genes: Nkx6.1 controls somatic motor neuron and ventral interneuron fates”. en. In: *Genes Dev.* 14.17 (Sept. 2000), pp. 2134–2139.

[71] Lu Teng et al. “Requirement for Foxd3 in the maintenance of neural crest progenitors”. en. In: *Development* 135.9 (May 2008), pp. 1615–1624.

[72] Alok Javali et al. “Co-expression of Tbx6 and Sox2 identifies a novel transient neuromesoderm progenitor cell state”. en. In: *Development* 144.24 (Dec. 2017), pp. 4522–4529.

[73] Ramkumar Sambasivan and Benjamin Steventon. “Neuromesodermal Progenitors: A Basis for Robust Axial Patterning in Development and Evolution”. en. In: *Front Cell Dev Biol* 8 (2020), p. 607516.

[74] V Wilson et al. “The T gene is necessary for normal mesodermal morphogenetic cell movements during gastrulation”. en. In: *Development* 121.3 (Mar. 1995), pp. 877–886.

[75] Hiromi Hirata et al. “Instability of Hes7 protein is crucial for the somite segmentation clock”. en. In: *Nat. Genet.* 36.7 (July 2004), pp. 750–754.

[76] Hanna Berger, Andreas Wodarz, and Annette Borchers. “PTK7 Faces the Wnt in Development and Disease”. en. In: *Front Cell Dev Biol* 5 (Apr. 2017), p. 31.

[77] Wim B M de Lau, Berend Snel, and Hans C Clevers. “The R-spondin protein family”. en. In: *Genome Biol.* 13.3 (2012), p. 242.

[78] T Shintani et al. “Neurons as well as astrocytes express proteoglycan-type protein tyrosine phosphatase zeta/RPTPbeta: analysis of mice in which the PTPzeta/RPTPbeta gene was replaced with the LacZ gene”. en. In: *Neurosci. Lett.* 247.2-3 (May 1998), pp. 135–138.

[79] Takeshi Sakurai. “The role of NrCAM in neural development and disorders–beyond a simple glue in the brain”. en. In: *Mol. Cell. Neurosci.* 49.3 (Mar. 2012), pp. 351–363.

[80] Changyong Tang et al. “Neural Stem Cells Behave as a Functional Niche for the Maturation of Newborn Neurons through the Secretion of PTN”. en. In: *Neuron* 101.1 (Jan. 2019), 32–44.e6.

[81] Shoichiro Tani et al. “Understanding paraxial mesoderm development and sclerotome specification for skeletal repair”. en. In: *Exp. Mol. Med.* 52.8 (Aug. 2020), pp. 1166–1177.

[82] Aida Rodrigo Albors, Pamela A Halley, and Kate G Storey. *Lineage tracing of axial progenitors using Nkx1-2CreERT2 mice defines their trunk and tail contributions.* 2018.

[83] Maxime Bouchard. “Transcriptional control of kidney development”. en. In: *Differentiation* 72.7 (Sept. 2004), pp. 295–306.

[84] Jaime A Rivera-Pérez and Anna-Katerina Hadjantonakis. “The Dynamics of Morphogenesis in the Early Mouse Embryo”. en. In: *Cold Spring Harb. Perspect. Biol.* 7.11 (June 2014).

[85] Sophie Balmer, Sonja Nowotschin, and Anna-Katerina Hadjantonakis. "Notochord morphogenesis in mice: Current understanding & open questions". en. In: *Dev. Dyn.* 245.5 (May 2016), pp. 547–557.

[86] J M Wells and D A Melton. "Vertebrate endoderm development". en. In: *Annu. Rev. Cell Dev. Biol.* 15 (1999), pp. 393–410.

[87] Kenzo Ivanovitch, Susana Temiño, and Miguel Torres. "Live imaging of heart tube development in mouse reveals alternating phases of cardiac differentiation and morphogenesis". en. In: *Elife* 6 (Dec. 2017).

[88] Sonja Nowotschin and Anna-Katerina Hadjantonakis. "Guts and gastrulation: Emergence and convergence of endoderm in the mouse embryo". en. In: *Curr. Top. Dev. Biol.* 136 (2020), pp. 429–454.

[89] Volker Bergen et al. "Generalizing RNA velocity to transient cell states through dynamical modeling". en. In: *Nat. Biotechnol.* 38.12 (Dec. 2020), pp. 1408–1414.

[90] Andreas Sagner et al. "A shared transcriptional code orchestrates temporal patterning of the central nervous system". en. In: *PLoS Biol.* 19.11 (Nov. 2021), e3001450.

[91] Rahul Satija et al. "Spatial reconstruction of single-cell gene expression data". en. In: *Nat. Biotechnol.* 33.5 (May 2015), pp. 495–502.

[92] Nikos Karaiskos et al. "The embryo at single-cell transcriptome resolution". en. In: *Science* 358.6360 (Oct. 2017), pp. 194–199.

[93] P P Tam and R R Behringer. "Mouse gastrulation: the formation of a mammalian body plan". en. In: *Mech. Dev.* 68.1-2 (Nov. 1997), pp. 3–25.

[94] Samuel A Lambert et al. "The Human Transcription Factors". en. In: *Cell* 172.4 (Feb. 2018), pp. 650–665.

[95] Hitoshi Niwa. "The principles that govern transcription factor network functions in stem cells". en. In: *Development* 145.6 (Mar. 2018).

[96] Hui Hu et al. “AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors”. en. In: *Nucleic Acids Res.* 47.D1 (Jan. 2019), pp. D33–D38.

[97] M Blum et al. “Gastrulation in the mouse: the role of the homeobox gene goosecoid”. en. In: *Cell* 69.7 (June 1992), pp. 1097–1106.

[98] Timothy J Nelson et al. “SRF-dependent gene expression in isolated cardiomyocytes: regulation of genes involved in cardiac hypertrophy”. en. In: *J. Mol. Cell. Cardiol.* 39.3 (Sept. 2005), pp. 479–489.

[99] Joseph M Miano et al. “Restricted inactivation of serum response factor to the cardiovascular system”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 101.49 (Dec. 2004), pp. 17132–17137.

[100] R P Harvey. “NK-2 homeobox genes and heart development”. en. In: *Dev. Biol.* 178.2 (Sept. 1996), pp. 203–216.

[101] Stefan C Materna et al. “Cardiovascular development and survival require Mef2c function in the myocardial but not the endothelial lineage”. en. In: *Dev. Biol.* 445.2 (Jan. 2019), pp. 170–177.

[102] Manvendra K Singh et al. “Gata4 and Gata5 cooperatively regulate cardiac myocyte proliferation in mice”. en. In: *J. Biol. Chem.* 285.3 (Jan. 2010), pp. 1765–1772.

[103] Anja Beckers et al. “The mouse homeobox gene Noto regulates node morphogenesis, notochordal ciliogenesis, and left right patterning”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 104.40 (Oct. 2007), pp. 15765–15770.

[104] B G Herrmann and A Kispert. “The T genes in embryogenesis”. en. In: *Trends Genet.* 10.8 (Aug. 1994), pp. 280–286.

[105] Marica Zizic Mitrecic et al. “The mouse gene Noto is expressed in the tail bud and essential for its morphogenesis”. en. In: *Cells Tissues Organs* 192.2 (Feb. 2010), pp. 85–92.

[106] Martin Cheung and James Briscoe. "Neural crest development is regulated by the transcription factor Sox9". en. In: *Development* 130.23 (Dec. 2003), pp. 5681–5693.

[107] Mamoru Ishii et al. "Combined deficiencies of Msx1 and Msx2 cause impaired patterning and survival of the cranial neural crest". en. In: *Development* 132.22 (Nov. 2005), pp. 4937–4950.

[108] Celeste Tribulo et al. "Regulation of Msx genes by a Bmp gradient is essential for neural crest specification". en. In: *Development* 130.26 (Dec. 2003), pp. 6441–6452.

[109] B J Martinsen and M Bronner-Fraser. "Neural crest specification regulated by the helix-loop-helix repressor Id2". en. In: *Science* 281.5379 (Aug. 1998), pp. 988–991.

[110] Daniel J Garry. "Etv2 IS A MASTER REGULATOR OF HEMATOENDOTHELIAL LINEAGES". en. In: *Trans. Am. Clin. Climatol. Assoc.* 127 (2016), pp. 212–223.

[111] Irina Elcheva et al. "Direct induction of haematoendothelial programs in human pluripotent stem cells by transcriptional regulators". en. In: *Nat. Commun.* 5 (July 2014), p. 4372.

[112] Emma de Pater et al. "Gata2 is required for HSC generation and survival". en. In: *J. Exp. Med.* 210.13 (Dec. 2013), pp. 2843–2850.

[113] Peter Laslo et al. "Multilineage transcriptional priming and determination of alternate hematopoietic cell fates". en. In: *Cell* 126.4 (Aug. 2006), pp. 755–766.

[114] Kaoru Mitsui et al. "The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells". en. In: *Cell* 113.5 (May 2003), pp. 631–642.

[115] Nadine Schrode et al. "GATA6 levels modulate primitive endoderm cell fate choice and timing in the mouse blastocyst". en. In: *Dev. Cell* 29.4 (May 2014), pp. 454–467.

[116] J Nichols et al. "Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4". en. In: *Cell* 95.3 (Oct. 1998), pp. 379–391.

[117] Guang Jin Pan et al. “Stem cell pluripotency and transcription factor Oct4”. en. In: *Cell Res.* 12.5-6 (Dec. 2002), pp. 321–329.

[118] Kok-Siong Chen et al. “The convergent roles of the nuclear factor I transcription factors in development and cancer”. en. In: *Cancer Lett.* 410 (Dec. 2017), pp. 124–138.

[119] A Z Chaudhry, G E Lyons, and R M Gronostajski. “Expression patterns of the four nuclear factor I genes during mouse embryogenesis indicate a potential role in development”. en. In: *Dev. Dyn.* 208.3 (Mar. 1997), pp. 313–325.

[120] Silvia Domcke et al. “A human cell atlas of fetal chromatin accessibility”. en. In: *Science* 370.6518 (Nov. 2020).

[121] Darren A Cusanovich et al. “The cis-regulatory dynamics of embryonic development at single-cell resolution”. en. In: *Nature* 555.7697 (Mar. 2018), pp. 538–542.

[122] Sven Heinz et al. “Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities”. en. In: *Mol. Cell* 38.4 (May 2010), pp. 576–589.

[123] E Bonnafe et al. “The transcription factor RFX3 directs nodal cilium development and left-right asymmetry specification”. en. In: *Mol. Cell. Biol.* 24.10 (May 2004), pp. 4417–4427.

[124] K Hemavathy, S I Ashraf, and Y T Ip. “Snail/slug family of repressors: slowly going into the fast lane of development and cancer”. en. In: *Gene* 257.1 (Oct. 2000), pp. 1–12.

[125] E A Carver et al. “The mouse snail gene encodes a key regulator of the epithelial-mesenchymal transition”. en. In: *Mol. Cell. Biol.* 21.23 (Dec. 2001), pp. 8184–8188.

[126] Detlev Arendt et al. “The origin and evolution of cell types”. en. In: *Nat. Rev. Genet.* 17.12 (Dec. 2016), pp. 744–757.

[127] Zhenyuan Yu et al. “Single-Cell Transcriptomic Map of the Human and Mouse Bladders”. en. In: *J. Am. Soc. Nephrol.* 30.11 (Nov. 2019), pp. 2159–2176.

[128] Cindy Fukazawa et al. “poky/chuk/ikk1 is required for differentiation of the zebrafish embryonic epidermis”. en. In: *Dev. Biol.* 346.2 (Oct. 2010), pp. 272–283.

[129] E M De Roberts et al. “Goosecoid and the organizer”. en. In: *Dev. Suppl.* (1992), pp. 167–171.

[130] Guilherme Costa et al. “SOX7 regulates the expression of VE-cadherin in the haemogenic endothelium at the onset of haematopoietic development”. en. In: *Development* 139.9 (May 2012), pp. 1587–1598.

[131] Y Takabatake, T Takabatake, and K Takeshima. “Conserved and divergent expression of T-box genes Tbx2-Tbx5 in Xenopus”. en. In: *Mech. Dev.* 91.1-2 (Mar. 2000), pp. 433–437.

[132] Francisco Barrionuevo et al. “Sox9 is required for invagination of the otic placode in mice”. en. In: *Dev. Biol.* 317.1 (May 2008), pp. 213–224.

[133] Wangjun Wu et al. “The role of Six1 in the genesis of muscle cell and skeletal muscle development”. en. In: *Int. J. Biol. Sci.* 10.9 (Sept. 2014), pp. 983–989.

[134] Woo Jun Shim et al. “Conserved Epigenetic Regulatory Logic Infers Genes Governing Cell Identity”. en. In: *Cell Syst* 11.6 (Dec. 2020), 625–639.e13.

[135] Sarah Bowling et al. “An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells”. en. In: *Cell* 181.6 (June 2020), 1410–1422.e27.

[136] Reza Kalhor et al. “Developmental barcoding of whole mouse via homing CRISPR”. en. In: *Science* 361.6405 (Aug. 2018).

[137] Michelle M Chan et al. “Molecular recording of mammalian embryogenesis”. en. In: *Nature* 570.7759 (June 2019), pp. 77–82.

[138] Gabriel Renaud et al. “deML: robust demultiplexing of Illumina sequences using a likelihood-based approach”. en. In: *Bioinformatics* 31.5 (Mar. 2015), pp. 770–772.

[139] Alexander Dobin et al. “STAR: ultrafast universal RNA-seq aligner”. en. In: *Bioinformatics* 29.1 (Jan. 2013), pp. 15–21.

[140] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. “HTSeq–a Python framework to work with high-throughput sequencing data”. en. In: *Bioinformatics* 31.2 (Jan. 2015), pp. 166–169.

[141] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. “SCANPY: large-scale single-cell gene expression data analysis”. en. In: *Genome Biol.* 19.1 (Feb. 2018), p. 15.

[142] Xiaojie Qiu et al. “Reversed graph embedding resolves complex single-cell trajectories”. en. In: *Nat. Methods* 14.10 (Oct. 2017), pp. 979–982.

[143] Andrew D Yates et al. “Ensembl 2020”. en. In: *Nucleic Acids Res.* 48.D1 (Jan. 2020), pp. D682–D688.

[144] Alexander J Tarashansky et al. “Mapping single-cell atlases throughout Metazoa unravels cell type evolution”. en. In: *Elife* 10 (May 2021).

[145] Aziz Khan et al. “JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework”. en. In: *Nucleic Acids Res.* 46.D1 (Jan. 2018), pp. D260–D266.

[146] Samantha A Morris et al. “Dissecting engineered cell types and enhancing cell fate conversion via CellNet”. en. In: *Cell* 158.4 (Aug. 2014), pp. 889–902.

[147] Eduardo Puelles et al. “Otx2 regulates the extent, identity and fate of neuronal progenitor domains in the ventral midbrain”. en. In: *Development* 131.9 (May 2004), pp. 2037–2048.

[148] Hidekiyo Harada, Tatsuya Sato, and Harukazu Nakamura. “Fgf8 signaling for development of the midbrain and hindbrain”. en. In: *Dev. Growth Differ.* 58.5 (June 2016), pp. 437–445.

[149] Holly C Gibbs et al. “Midbrain-Hindbrain Boundary Morphogenesis: At the Intersection of Wnt and Fgf Signaling”. en. In: *Front. Neuroanat.* 11 (Aug. 2017), p. 64.

[150] Kendra Sturgeon et al. “Cdx1 refines positional identity of the vertebrate hindbrain by directly repressing Mafb expression”. en. In: *Development* 138.1 (Jan. 2011), pp. 65–74.

[151] Hugo J Parker, Irina Pushel, and Robb Krumlauf. “Coupling the roles of Hox genes to regulatory networks patterning cranial neural crest”. en. In: *Dev. Biol.* 444 Suppl 1 (Dec. 2018), S67–S78.

[152] Chengxiang Qiu et al. “Systematic reconstruction of cellular trajectories across mouse embryogenesis”. en. In: *Nat. Genet.* 54.3 (Mar. 2022), pp. 328–341.

[153] Stephen A Murray et al. “Mouse gestation length is genetically determined”. en. In: *PLoS One* 5.8 (Aug. 2010), e12418.

[154] Eric H Davidson, David R McClay, and Leroy Hood. “Regulatory gene networks and the properties of the developmental process”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 100.4 (Feb. 2003), pp. 1475–1480.

[155] Eric H Davidson et al. “A genomic regulatory network for development”. en. In: *Science* 295.5560 (Mar. 2002), pp. 1669–1678.

[156] Bernd Boehm et al. “A landmark-free morphometric staging system for the mouse limb bud”. en. In: *Development* 138.6 (Mar. 2011), pp. 1227–1234.

[157] Marco Musy et al. “A quantitative method for staging mouse embryos based on limb morphometry”. en. In: *Development* 145.7 (Apr. 2018).

[158] Elena Tzouanacou et al. “Redefining the progression of lineage segregations during mammalian embryogenesis by clonal analysis”. en. In: *Dev. Cell* 17.3 (Sept. 2009), pp. 365–376.

[159] Isabel Olivera-Martinez et al. "Loss of FGF-dependent mesoderm identity and rise of endogenous retinoid signalling determine cessation of body axis elongation". en. In: *PLoS Biol.* 10.10 (Oct. 2012), e1001415.

[160] Carolina Guibentif et al. *Diverse Routes toward Early Somites in the Mouse Embryo.* 2021.

[161] Dorothee Mugele et al. *Genetic approaches in mice demonstrate that neuro-mesodermal progenitors express T/Brachyury but not Sox2.*

[162] Simne Langton and Lorraine J Gudas. "CYP26A1 knockout embryonic stem cells exhibit reduced differentiation and growth arrest in response to retinoic acid". en. In: *Dev. Biol.* 315.2 (Mar. 2008), pp. 331–354.

[163] Mina Gouti et al. "A Gene Regulatory Network Balances Neural and Mesoderm Specification during Vertebrate Trunk Development". en. In: *Dev. Cell* 41.3 (May 2017), 243–261.e7.

[164] André Dias et al. "A Tgfbr1/Snai1-dependent developmental module at the core of vertebrate axial elongation". en. In: *Elife* 9 (June 2020).

[165] S J Kinder et al. "The organizer of the mouse gastrula is composed of a dynamic population of progenitor cells for the axial mesoderm". en. In: *Development* 128.18 (Sept. 2001), pp. 3623–3634.

[166] Claudio D Stern. *Initial patterning of the central nervous system: How many organizers?* 2001.

[167] A C Foley, I Skromne, and C D Stern. "Reconciling different models of forebrain induction and patterning: a dual role for the hypoblast". en. In: *Development* 127.17 (Sept. 2000), pp. 3839–3854.

[168] Yojiro Yamanaka et al. "Live imaging and genetic analysis of mouse notochord formation reveals regional morphogenetic mechanisms". en. In: *Dev. Cell* 13.6 (Dec. 2007), pp. 884–896.

[169] Dennis Schifferl et al. “A 37 kb region upstream of brachyury comprising a notochord enhancer is essential for notochord and tail development”. en. In: *Development* 148.23 (Dec. 2021).

[170] Ashley E E Bruce and Rudolf Winklbauer. “Brachyury in the gastrula of basal vertebrates”. en. In: *Mech. Dev.* 163 (Sept. 2020), p. 103625.

[171] J B Singer et al. *Drosophila brachyenteron regulates gene activity and morphogenesis in the gut*. 1996.

[172] A Woollard and J Hodgkin. “The caenorhabditis elegans fate-determining gene mab-9 encodes a T-box protein required to pattern the posterior hindgut”. en. In: *Genes Dev.* 14.5 (Mar. 2000), pp. 596–603.

[173] Daisy A Robinton et al. “The Lin28/let-7 Pathway Regulates the Mammalian Caudal Body Axis Elongation Program”. en. In: *Dev. Cell* 48.3 (Feb. 2019), 396–405.e3.

[174] Eric G Moss and Lingjuan Tang. “Conservation of the heterochronic regulator Lin-28, its developmental expression and microRNA complementary sites”. en. In: *Dev. Biol.* 258.2 (June 2003), pp. 432–442.

[175] Frank Costantini and Reena Shakya. “GDNF/Ret signaling and the development of the kidney”. en. In: *Bioessays* 28.2 (Feb. 2006), pp. 117–127.

[176] Arindam Majumdar et al. “Wnt11 and Ret/Gdnf pathways cooperate in regulating ureteric branching during metanephric kidney development”. en. In: *Development* 130.14 (July 2003), pp. 3175–3185.

[177] Bree A Rumballe et al. “Nephron formation adopts a novel spatial topology at cessation of nephrogenesis”. en. In: *Dev. Biol.* 360.1 (Dec. 2011), pp. 110–122.

[178] Shunsuke Yuri et al. “In Vitro Propagation and Branching Morphogenesis from Single Ureteric Bud Cells”. en. In: *Stem Cell Reports* 8.2 (Feb. 2017), pp. 401–416.

[179] Adrian S Woolf and Jamie A Davies. “Cell biology of ureter development”. en. In: *J. Am. Soc. Nephrol.* 24.1 (Jan. 2013), pp. 19–25.

[180] Andrew Ransick et al. “Single-Cell Profiling Reveals Sex, Lineage, and Regional Diversity in the Mouse Kidney”. en. In: *Dev. Cell* 51.3 (Nov. 2019), 399–413.e7.

[181] Karin D Prummel, Susan Nieuwenhuize, and Christian Mosimann. “The lateral plate mesoderm”. en. In: *Development* 147.12 (June 2020).

[182] Lu Han et al. “Single cell transcriptomics identifies a signaling network coordinating endoderm and mesoderm diversification during foregut organogenesis”. en. In: *Nat. Commun.* 11.1 (Aug. 2020), pp. 1–16.

[183] L Ariza et al. “Coelomic epithelium-derived cells in visceral morphogenesis”. In: *Dev. Dyn.* 245.3 (Mar. 2016).

[184] Irene Delgado et al. “GATA4 loss in the septum transversum mesenchyme promotes liver fibrosis in mice”. en. In: *Hepatology* 59.6 (June 2014), pp. 2358–2370.

[185] Chenura D Jayewickreme and Ramesh A Shivdasani. “Control of stomach smooth muscle development and intestinal rotation by transcription factor BARX1”. en. In: *Dev. Biol.* 405.1 (Sept. 2015), pp. 21–32.

[186] Lázaro Centanin and Joachim Wittbrodt. “Retinal neurogenesis”. en. In: *Development* 141.2 (Jan. 2014), pp. 241–244.

[187] Brian S Clark et al. “Single-Cell RNA-Seq Analysis of Retinal Development Identifies NFI Factors as Regulating Mitotic Exit and Late-Born Cell Specification”. en. In: *Neuron* 102.6 (June 2019), 1111–1126.e5.

[188] Connie Cepko. “Intrinsically different retinal progenitor cells produce specific types of progeny”. en. In: *Nat. Rev. Neurosci.* 15.9 (Sept. 2014), pp. 615–627.

[189] Karthik Shekhar et al. “Diversification of multipotential postmitotic mouse retinal ganglion cell precursors into discrete types”. en. In: *Elife* 11 (Feb. 2022).

[190] Robert F Hevner. “From radial glia to pyramidal-projection neuron: transcription factor cascades in cerebral cortex development”. en. In: *Mol. Neurobiol.* 33.1 (Feb. 2006), pp. 33–50.

[191] Kei Hori and Mikio Hoshino. "GABAergic neuron specification in the spinal cord, the cerebellum, and the cochlear nucleus". en. In: *Neural Plast.* 2012 (June 2012), p. 921732.

[192] Emma R Broom et al. "The roof plate boundary is a bi-directional organiser of dorsal neural tube and choroid plexus development". en. In: *Development* 139.22 (Nov. 2012), pp. 4261–4270.

[193] Marc R Freeman. "Specification and morphogenesis of astrocytes". en. In: *Science* 330.6005 (Nov. 2010), pp. 774–778.

[194] Andreas Sagner and James Briscoe. *Establishing neuronal diversity in the spinal cord: a time and a place*. 2019.

[195] T M Jessell. "Neuronal specification in the spinal cord: inductive signals and transcriptional codes". en. In: *Nat. Rev. Genet.* 1.1 (Oct. 2000), pp. 20–29.

[196] Zhigang Xue et al. "Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing". en. In: *Nature* 500.7464 (Aug. 2013), pp. 593–597.

[197] Carmit Levy, Mehdi Khaled, and David E Fisher. "MITF: master regulator of melanocyte development and melanoma oncogene". en. In: *Trends Mol. Med.* 12.9 (Sept. 2006), pp. 406–414.

[198] Robert Nechanitzky et al. "Transcription factor EBF1 is essential for the maintenance of B cell identity and prevention of alternative fates in committed cells". en. In: *Nat. Immunol.* 14.8 (Aug. 2013), pp. 867–875.

[199] T Reya et al. "Wnt signaling regulates B lymphocyte proliferation through a LEF-1 dependent mechanism". en. In: *Immunity* 13.1 (July 2000), pp. 15–24.

[200] Gregory D Gregory et al. "FOG1 requires NuRD to promote hematopoiesis and maintain lineage fidelity within the megakaryocytic-erythroid compartment". en. In: *Blood* 115.11 (Mar. 2010), pp. 2156–2166.

[201] Hidefumi Iwashita et al. “Secreted cerberus1 as a marker for quantification of definitive endoderm differentiation of the pluripotent stem cells”. en. In: *PLoS One* 8.5 (May 2013), e64291.

[202] N Neckelmann et al. “cDNA sequence of a human skeletal muscle ADP/ATP translocator: lack of a leader peptide, divergence from a fibroblast translocator cDNA, and coevolution with mitochondrial DNA genes”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 84.21 (Nov. 1987), pp. 7580–7584.

[203] Ian A Simpson et al. “The facilitative glucose transporter GLUT3: 20 years of distinction”. en. In: *Am. J. Physiol. Endocrinol. Metab.* 295.2 (Aug. 2008), E242–53.

[204] Christine Hacker et al. “Transcriptional profiling identifies Id2 function in dendritic cell development”. en. In: *Nat. Immunol.* 4.4 (Apr. 2003), pp. 380–386.

[205] Huiyun Liang and Walter F Ward. “PGC-1alpha: a key regulator of energy metabolism”. en. In: *Adv. Physiol. Educ.* 30.4 (Dec. 2006), pp. 145–151.

[206] J Girard. “Metabolic adaptations to change of nutrition at birth”. en. In: *Biol. Neonate* 58 Suppl 1 (1990), pp. 3–15.

[207] Xingxing Kong et al. “IRF4 is a key thermogenic transcriptional partner of PGC-1α”. en. In: *Cell* 158.1 (July 2014), pp. 69–83.

[208] Leslie A Rowland et al. “Uncoupling Protein 1 and Sarcolipin Are Required to Maintain Optimal Thermogenesis, and Loss of Both Systems Compromises Survival of Mice under Cold Stress”. en. In: *J. Biol. Chem.* 290.19 (May 2015), pp. 12282–12289.

[209] R M Tribe et al. “Parturition and the perinatal period: can mode of delivery impact on the future health of the neonate?” en. In: *J. Physiol.* 596.23 (Dec. 2018), pp. 5709–5722.

[210] Nick Hopwood. “’Not birth, marriage or death, but gastrulation’: the life of a quotation in biology”. en. In: *Br. J. Hist. Sci.* 55.1 (Mar. 2022), pp. 1–26.

[211] Xingfan Huang et al. “Single cell, whole embryo phenotyping of pleiotropic disorders of mammalian development”. en. Aug. 2022.

[212] Silvia Domcke and Jay Shendure. “A reference cell tree will serve science better than a reference cell atlas”. en. In: *Cell* 186.6 (Mar. 2023), pp. 1103–1114.

[213] Dian Yang et al. “Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution”. en. In: *Cell* 185.11 (May 2022), 1905–1923.e25.

[214] Junhong Choi et al. “A time-resolved, multi-symbol molecular recorder via sequential genome editing”. en. In: *Nature* 608.7921 (Aug. 2022), pp. 98–107.

[215] Christina V Theodoris et al. “Transfer learning enables predictions in network biology”. en. In: *Nature* 618.7965 (June 2023), pp. 616–624.

[216] J E Sulston and H R Horvitz. “Post-embryonic cell lineages of the nematode, Caenorhabditis elegans”. en. In: *Dev. Biol.* 56.1 (Mar. 1977), pp. 110–156.

[217] Karen I Zeller et al. “An integrated database of genes responsive to the Myc oncogenic transcription factor: identification of direct genomic targets”. en. In: *Genome Biol.* 4.10 (Sept. 2003), R69.

[218] Mingzhu Jiang et al. “The emerging role of MEIS1 in cell proliferation and differentiation”. en. In: *Am. J. Physiol. Cell Physiol.* 320.3 (Mar. 2021), pp. C264–C269.

[219] Frederic Koch et al. “Antagonistic Activities of Sox2 and Brachyury Control the Fate Choice of Neuro-Mesodermal Progenitors”. en. In: *Dev. Cell* 42.5 (Sept. 2017), 514–526.e7.

[220] Brian Hie et al. “Geometric Sketching Compactly Summarizes the Single-Cell Transcriptomic Landscape”. en. In: *Cell Syst* 8.6 (June 2019), 483–493.e7.

[221] R A Coleman et al. “Expression of aquaporins in the renal connecting tubule”. en. In: *Am. J. Physiol. Renal Physiol.* 279.5 (Nov. 2000), F874–83.

[222] Fuguo Wu et al. “Single cell transcriptomics reveals lineage trajectory of retinal ganglion cells in wild-type and Atoh7-null retinas”. en. In: *Nat. Commun.* 12.1 (Mar. 2021), p. 1465.

[223] Chun Qiu et al. “Differential expression of TYRP1 in adult human retinal pigment epithelium and uveal melanoma cells”. en. In: *Oncol. Lett.* 11.4 (Apr. 2016), pp. 2379–2383.

[224] Stina H Mui et al. “Vax genes ventralize the embryonic eye”. en. In: *Genes Dev.* 19.10 (May 2005), pp. 1249–1259.

[225] Jie Wang, Amir Rattner, and Jeremy Nathans. “A transcriptome atlas of the mouse iris at single-cell resolution defines cell types and the genomic response to pupil dilation”. en. In: *Elife* 10 (Nov. 2021).

[226] Daisuke Kurokawa et al. “Regulation of Otx2 expression and its functions in mouse forebrain and midbrain”. en. In: *Development* 131.14 (July 2004), pp. 3319–3331.

[227] Fabrice Prin et al. “Hox proteins drive cell segregation and non-autonomous apical remodelling during hindbrain segmentation”. en. In: *Development* 141.7 (Apr. 2014), pp. 1492–1502.

[228] Laura E Mickelsen et al. “Single-cell transcriptomic analysis of the lateral hypothalamic area reveals molecularly distinct populations of inhibitory and excitatory neurons”. en. In: *Nat. Neurosci.* 22.4 (Apr. 2019), pp. 642–656.

[229] Asa Wallén-Mackenzie, Hanna Wootz, and Hillevi Englund. “Genetic inactivation of the vesicular glutamate transporter 2 (VGLUT2) in the mouse: what have we learnt about functional glutamatergic neurotransmission?” en. In: *Ups. J. Med. Sci.* 115.1 (Feb. 2010), pp. 11–20.

[230] Nadia Rosenthal and Steve Brown. “The mouse ascending: perspectives for human-disease models”. en. In: *Nat. Cell Biol.* 9.9 (Sept. 2007), pp. 993–999.

[231] Johannes Beckers, Wolfgang Wurst, and Martin Hrabé de Angelis. “Towards better mouse models: enhanced genotypes, systemic phenotyping and envirotype modelling”. en. In: *Nat. Rev. Genet.* 10.6 (June 2009), pp. 371–380.

[232] Sudha Rajderkar et al. “Topologically associating domain boundaries are required for normal genome function”. en. In: *Commun Biol* 6.1 (Apr. 2023), p. 435.

[233] Alexandra Despang et al. “Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture”. en. In: *Nat. Genet.* 51.8 (Aug. 2019), pp. 1263–1271.

[234] Professor Dr of Molecular Biology Lee M Silver and Lee M Silver. *Mouse Genetics: Concepts and Applications*. en. Oxford University Press on Demand, 1995.

[235] Martin Ringwald et al. “The IKMC web portal: a central point of entry to data and resources from the International Knockout Mouse Consortium”. en. In: *Nucleic Acids Res.* 39.Database issue (Jan. 2011), pp. D849–55.

[236] Martin Jinek et al. “A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity”. en. In: *Science* 337.6096 (Aug. 2012), pp. 816–821.

[237] Priti Singh, John C Schimenti, and Ewelina Bolcun-Filas. “A mouse geneticist’s practical guide to CRISPR applications”. en. In: *Genetics* 199.1 (Jan. 2015), pp. 1–15.

[238] Malte Spielmann, Darıo G Lupiáñez, and Stefan Mundlos. *Structural variation in the 3D genome*. 2018.

[239] Valérie Gailus-Durner et al. “Introducing the German Mouse Clinic: open access platform for standardized phenotyping”. en. In: *Nat. Methods* 2.6 (June 2005), pp. 403–404.

[240] Marco Osterwalder et al. “Enhancer redundancy provides phenotypic robustness in mammalian development”. en. In: *Nature* 554.7691 (Feb. 2018), pp. 239–243.

[241] Yingyue Zhou et al. "Human and mouse single-nucleus transcriptomics reveal TREM2-dependent and TREM2-independent cellular responses in Alzheimer's disease". en. In: *Nat. Med.* 26.1 (Jan. 2020), pp. 131–142.

[242] R W Stottmann et al. "Ttc21b is required to restrict sonic hedgehog activity in the developing mouse forebrain". en. In: *Dev. Biol.* 335.1 (Nov. 2009), pp. 166–178.

[243] Neelu Yadav et al. "Specific protein methylation defects and gene expression perturbations in coactivator-associated arginine methyltransferase 1-deficient mice". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 100.11 (May 2003), pp. 6464–6468.

[244] R Mo et al. "Specific and redundant functions of Gli2 and Gli3 zinc finger genes in skeletal patterning and development". en. In: *Development* 124.1 (Jan. 1997), pp. 113–123.

[245] Enrico Leipold et al. *A de novo gain-of-function mutation in SCN11A causes loss of pain perception*. 2013.

[246] Georg C Schwabe et al. "Ror2 knockout mouse as a model for the developmental pathology of autosomal recessive Robinow syndrome". en. In: *Dev. Dyn.* 229.2 (Feb. 2004), pp. 400–410.

[247] Wing Lee Chan et al. *Impaired proteoglycan glycosylation, elevated TGF-β signaling, and abnormal osteoblast differentiation as the basis for bone fragility in a mouse model for gerodermia osteodysplastica*. 2018.

[248] Björn Fischer et al. "Further characterization of ATP6V0A2-related autosomal recessive cutis laxa". en. In: *Hum. Genet.* 131.11 (Nov. 2012), pp. 1761–1773.

[249] Alessa R Ringel et al. "Repression and 3D-restructuring resolves regulatory conflicts in evolutionarily rearranged genomes". en. In: *Cell* 185.20 (Sept. 2022), 3689–3704.e21.

[250] Evgeny Z Kvon et al. "Progressive Loss of Function in a Limb Enhancer during Snake Evolution". en. In: *Cell* 167.3 (Oct. 2016), 633–642.e11.

[251] J Jacob and J Briscoe. “Gli proteins and the control of spinal-cord patterning”. In: *EMBO Rep.* 4.8 (Aug. 2003).

[252] Alice Jo et al. “The versatile functions of Sox9 in development, stem cells, and human diseases”. en. In: *Genes Dis* 1.2 (Dec. 2014), pp. 149–161.

[253] C T Gordon et al. “Long-range regulation at the SOX9 locus in development and disease”. en. In: *J. Med. Genet.* 46.10 (Oct. 2009), pp. 649–656.

[254] Pamela V Tran et al. “THM1 negatively modulates mouse sonic hedgehog signal transduction and affects retrograde intraflagellar transport in cilia”. en. In: *Nat. Genet.* 40.4 (Apr. 2008), pp. 403–410.

[255] Emma Dann et al. “Differential abundance testing on single-cell data using k-nearest neighbor graphs”. en. In: *Nat. Biotechnol.* (Sept. 2021).

[256] M P Matise et al. “Gli2 is required for induction of floor plate and adjacent cells, but not most ventral neurons in the mouse central nervous system”. en. In: *Development* 125.15 (Aug. 1998), pp. 2759–2770.

[257] Q Ding et al. “Diminished Sonic hedgehog signaling and lack of floor plate differentiation in Gli2 mutant mice”. en. In: *Development* 125.14 (July 1998), pp. 2533–2543.

[258] Ryo Ichijo et al. “Essential roles of Tbx3 in embryonic skin development during epidermal stratification”. en. In: *Genes Cells* 22.3 (Mar. 2017), pp. 284–292.

[259] Annalise B Paaby and Matthew V Rockman. “The many faces of pleiotropy”. en. In: *Trends Genet.* 29.2 (Feb. 2013), pp. 66–73.

[260] Baojin Yao et al. “The SOX9 upstream region prone to chromosomal aberrations causing campomelic dysplasia contains multiple cartilage enhancers”. en. In: *Nucleic Acids Res.* 43.11 (June 2015), pp. 5394–5408.

[261] Charlotte E Scott et al. “SOX9 induces and maintains neural stem cells”. en. In: *Nat. Neurosci.* 13.10 (Oct. 2010), pp. 1181–1189.

[262] T Wagner et al. “Autosomal sex reversal and campomelic dysplasia are caused by mutations in and around the SRY-related gene SOX9”. en. In: *Cell* 79.6 (Dec. 1994), pp. 1111–1120.

[263] George Coricor and Rosa Serra. “TGF-β regulates phosphorylation and stabilization of Sox9 protein in chondrocytes through p38 and Smad dependent mechanisms”. en. In: *Sci. Rep.* 6 (Dec. 2016), p. 38616.

[264] R Haller et al. “Notch1 signaling regulates chondrogenic lineage determination through Sox9 activation”. en. In: *Cell Death Differ.* 19.3 (Mar. 2012), pp. 461–469.

[265] Haruhiko Akiyama et al. “Interactions between Sox9 and beta-catenin control chondrocyte differentiation”. en. In: *Genes Dev.* 18.9 (May 2004), pp. 1072–1087.

[266] Kenta Hino et al. “Master regulator for chondrogenesis, Sox9, regulates transcriptional activation of the endoplasmic reticulum stress transducer BBF2H7/CREB3L2 in chondrocytes”. en. In: *J. Biol. Chem.* 289.20 (May 2014), pp. 13810–13820.

[267] Sanjay R Srivatsan et al. “Embryo-scale, single-cell spatial transcriptomics”. en. In: *Science* 373.6550 (July 2021), pp. 111–117.

[268] Rosa Hernández et al. “Differentiation of Human Mesenchymal Stem Cells towards Neuronal Lineage: Clinical Trials in Nervous System Disorders”. en. In: *Biomol. Ther.* 28.1 (Jan. 2020), pp. 34–44.

[269] Hongcheng Mai et al. “Whole-body cellular mapping in mouse using standard IgG antibodies”. en. In: *Nat. Biotechnol.* (July 2023).

[270] Mary E Dickinson et al. “High-throughput discovery of novel developmental phenotypes”. en. In: *Nature* 537.7621 (Sept. 2016), pp. 508–514.

[271] Katerina Kraft et al. “Deletions, Inversions, Duplications: Engineering of Structural Variants using CRISPR/Cas in Mice”. en. In: *Cell Rep.* 10.5 (Feb. 2015), pp. 833–839.

[272] Ilya Korsunsky et al. “Fast, sensitive and accurate integration of single-cell data with Harmony”. en. In: *Nat. Methods* 16.12 (Dec. 2019), pp. 1289–1296.

[273] Fay Wang et al. “RNAscope: a novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues”. en. In: *J. Mol. Diagn.* 14.1 (Jan. 2012), pp. 22–29.

[274] Johannes Schindelin et al. “Fiji: an open-source platform for biological-image analysis”. en. In: *Nat. Methods* 9.7 (June 2012), pp. 676–682.

[275] Nicholas Borcherding et al. “Mapping the immune environment in clear cell renal carcinoma by single-cell genomics”. en. In: *Commun Biol* 4.1 (Jan. 2021), p. 122.

[276] Aravind Subramanian et al. “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles”. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 102.43 (Oct. 2005), pp. 15545–15550.

[277] Arthur Liberzon et al. “The Molecular Signatures Database (MSigDB) hallmark gene set collection”. en. In: *Cell Syst* 1.6 (Dec. 2015), pp. 417–425.

[278] Hiroki R Ueda et al. “Whole-Brain Profiling of Cells and Circuits in Mammals by Tissue Clearing and Light-Sheet Microscopy”. en. In: *Neuron* 106.3 (May 2020), pp. 369–387.

[279] Alessandro Motta et al. “Dense connectomic reconstruction in layer 4 of the somatosensory cortex”. en. In: *Science* 366.6469 (Nov. 2019).

[280] Yinan Wan et al. “Single-Cell Reconstruction of Emerging Population Activity in an Entire Developing Circuit”. en. In: *Cell* 179.2 (Oct. 2019), 355–372.e23.

[281] Rory J Maizels, Daniel M Snell, and James Briscoe. “Deep dynamical modelling of developmental trajectories with temporal transcriptomics”. en. July 2023.