

©Copyright 2023

Xingfan Huang

Computational methods for analyzing high-dimensional datasets
derived from molecular profiling of biological systems

Xingfan Huang

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

William Noble

David Beier

Program Authorized to Offer Degree:

Paul G. Allen School of Computer Science & Engineering

University of Washington

Abstract

Computational methods for analyzing high-dimensional datasets
derived from molecular profiling of biological systems

Xingfan Huang

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

Embryonic development is the process by which a single fertilized egg gives rise to all cells in a complex multicellular organism. Recent advances in single-cell sequencing techniques allow researchers to generate genomic profiles of large numbers of single cells and study metazoan development and complex tissues at unprecedented resolution. The objectives of my research projects are to develop computational methods to analyze high-dimensional genomics datasets, and direct those methods towards understanding metazoan embryonic development and complex tissues, in both normal and disease related genetic backgrounds.

In this thesis, we introduce three projects: 1) *Drosophila* Embryonic Atlas Project (DEAP), a continuous, single-cell atlas of chromatin accessibility and gene expression that spans *Drosophila* embryogenesis; 2) Mouse Mutant Cell Atlas (MMCA), a single-cell atlas of gene expression of mouse embryos in both normal and developmental disease related genetic backgrounds at the end of organogenesis; and 3) Macaque Brain Atlas, a multi-omic single-cell atlas of adult macaque brain samples to study regional heterogeneity and regulatory grammar of diverse brain cell types.

We introduce several novel computational methods: 1) A neural network to infer the precise time in development of individual nuclei and understand continuous gene expression and regulatory dynamics in *Drosophila* embryogenesis; 2) a novel statistic lochNESS to

quantify sample specific effects in single-cell data based on k -NN graphs, allowing systematic comparison of molecular profiles from normal and disease datasets to investigate effects of genetic perturbations; and 3) an extension of lochNESS for multiple comparisons to identify brain region specific transcriptional programs in the Macaque Brain Atlas. With these novel methods, combined with existing computational tools and packages, we derive biological insights from the various single-cell atlases and add to our understanding of the various biological systems.

TABLE OF CONTENTS

	Page
List of Figures	iv
Chapter 1: Introduction	1
1.1 Topics in this dissertation	1
1.2 Single-cell sequencing to build organism-scale atlases	4
1.3 Temporal molecular dynamics in metazoan development	6
1.4 Effects of genetic mutations in metazoan development	12
1.5 Dissecting molecular heterogeneity and regulatory information in complex brain tissues	17
Chapter 2: The continuum of drosophila embryonic development at single cell resolution	22
2.1 Introduction	24
2.2 Results	25
2.2.1 Predicting the absolute age of individual nuclei	28
2.2.2 Annotation and inference of diversifying developmental trajectories	29
2.2.3 Tracing dynamic gene modules across development	35
2.2.4 Nominating stage- and cell type-specific TF regulators	39
2.3 Discussion	46
2.4 Materials and methods	48
2.4.1 Sci-RNA-seq methods	52
2.4.2 Sci-ATAC-seq methods	55
2.4.3 Data analysis	60
2.5 Supplementary materials	67
2.5.1 Supplementary notes	67
2.5.2 Supplementary tables	69
2.5.3 Supplementary figures	72

Chapter 3:	Single cell, whole embryo phenotyping of pleiotropic disorders of mammalian development	84
3.1	Introduction	86
3.2	Results	88
3.2.1	Single-cell RNA-seq of 101 mouse embryos	88
3.2.2	Mutant-specific differences in cell type composition	91
3.2.3	LochNESS analysis reveals differences in transcriptional state within cell type trajectories	95
3.2.4	Identification of mutant-specific and mutant-shared effects	99
3.2.5	Global developmental defects in Sox9 regulatory mutant	103
3.3	Discussion	106
3.4	Materials and methods	109
3.5	Supplementary materials	123
3.5.1	Supplementary notes	123
3.5.2	Supplementary figures	123
Chapter 4:	A single-cell multi-omic atlas spanning the adult rhesus macaque brain	140
4.1	Introduction	141
4.2	Results	143
4.2.1	molecular taxonomy of cell types across the primate brain	143
4.2.2	Regional variation in cell subtype composition	146
4.2.3	Joint analysis of single-nucleus transcriptomic and epigenomic data	151
4.2.4	The gene regulatory landscape of the rhesus macaque brain	154
4.3	Discussion	165
4.4	Materials and methods	167
4.5	Supplementary Materials	190
4.5.1	Supplementary Tables	190
4.5.2	Supplementary Figures	194
Chapter 5:	Discussion	213
5.1	Limitations and future directions	214
5.1.1	Limitations and future directions of DEAP	215
5.1.2	Limitations and future directions of MMCA	217
5.1.3	Limitations and future directions of MBA	220

5.2	General thoughts and directions on computational genomics	222
5.2.1	Validation and technology development	223
5.2.2	Representation and visualization	225
5.2.3	Integration and models	229
5.2.4	AI and Genomics	232

LIST OF FIGURES

Figure Number	Page
2.1 Figure 1: Single cell profiling of chromatin accessibility and gene expression throughout <i>Drosophila</i> embryogenesis.	27
2.2 Figure 2: Inferring developmental age from cellular state.	30
2.3 Figure 3: Annotation of diversifying developmental trajectories. . .	36
2.4 Figure 4: Dynamic regulation of mesoderm-specific gene modules.	40
2.5 Figure 5: Integration of scRNA and scATAC data	45
2.6 Figure S1: QC/stats of the sci-ATAC-seq and sci-RNA-seq datasets.	73
2.7 Figure S2: Data filtering, evaluating batch and technical effects . .	74
2.8 Figure S3: Examples of time-dependent genes	75
2.9 Figure S4: A neural network-based model	76
2.10 Figure S5: Model identifies “contaminating” older nuclei in the earliest collection window.	77
2.11 Figure S6: Distribution of cells and cell states through inferred time.	78
2.12 Figure S7: Neuronal marker genes and spatial analysis.	79
2.13 Figure S8: Parameter selection and ATAC mesoderm clustering. . .	80
2.14 Figure S9: Motif activity and linking of scATAC and scRNA-derived clusters	81
2.15 Figure S10: Expression validation of TFs by in situ hybridization .	82
2.16 Figure S11: Additional analyses of interest.	83
3.1 Figure 1. Single-cell RNA-seq of 103 whole mouse embryos staged at E13.5	92
3.2 Figure 2. Cell composition changes for individual mutants across developmental trajectories	96
3.3 Figure 3. LochNESS analysis identifies mutant related changes . . .	100
3.4 Figure 4. Apparent stalling and redirection of mesenchyme differentiation in the Sox9 regulatory INV mutant.	107
3.5 Figure S1. Images of mouse embryos	124

3.6	Figure S2. Integrating cells derived from embryos of multiple genetic backgrounds to a single, wildtype-based “reference embedding”	125
3.7	Figure S3. Annotation of sub-trajectories from wildtype E13.5 embryos	126
3.8	Figure S4. <i>NNLS</i> correlation between developmental trajectories of MOCA and MMCA	127
3.9	Figure S5. Cell composition for individual embryos across trajectories	128
3.10	Figure S6. Multiple retinal trajectories are diminished in <i>Ttc21b</i> KO mice.	129
3.11	Figure S7. Co-embedding cells from selected trajectories from ZRS limb enhancer KO and FVB wildtype	130
3.12	Figure S8. Quantitative analysis of lochNESS distributions and analysis of <i>Gli2</i> KO in the roof plate and floor plate trajectories	131
3.13	Figure S9. Systematic screening of lochNESS distributions identifies altered epithelial sub-trajectories in the <i>Tbx3</i> TAD Boundary KO mutant	132
3.14	Figure S10. Similarity scores reveal mutant-shared and mutant-specific effects	133
3.15	Figure S11. Misregulation of <i>Sox9</i> and <i>Kcnj2</i> in the <i>Sox9</i> regulatory INV mutant	134
3.16	Figure S13. Stalling of regulatory INV cells in the undifferentiated mesenchyme	135
3.17	Figure S14. Density plots of the UMAP co-embedding of G4 samples	136
3.18	Figure S15. Density plots of the UMAP co-embedding of FVB samples	137
3.19	Figure S16. Density plots of the UMAP co-embedding of C57BL/6 samples	138
3.20	Figure S17. Density and marker gene expression plots of UMAP co-embeddings of C57BL/6 mice in the limb mesenchyme trajectory	139
4.1	Fig. 1. Experimental setup and summary of the Macaque Brain Atlas snRNA-seq dataset	147
4.2	Fig. 2. Cell subtype distribution and variation across the brain	152
4.3	Fig. 3. Generation of the Macaque Brain Atlas sci-ATAC-seq dataset and identification of cell classes	155

4.4	Fig. 4. Enrichment of transcription factor binding site motifs in candidate regulatory elements	159
4.5	Fig. 5. The landscape of <i>cis</i> -regulatory interactions in the Macaque Brain Atlas	162
4.6	Fig. 6. Enrichment of heritable disease-relevant sites among candidate regulatory elements.	195
4.7	Fig. S1. Schematic depicting snRNA-seq quality control pipeline .	196
4.8	Fig. S2. Identification and removal of exogenous nuclei	197
4.9	Fig. S3. Comparison of UMAP projections before batch and after batch correction	198
4.10	Fig. S4. Marker genes for each cell class	199
4.11	Fig. S5. Unsupervised clustering of cell classes and brain regions .	200
4.12	Fig. S6. Comparison of UMAP embeddings of nuclei derived from samples of different biological sex and brain hemisphere	201
4.13	Fig. S7. Cell-class specific UMAP projections colored and labeled according to identified cell subtypes.	202
4.14	Fig. S8. Cell subtype concordance with human cortical, human brain vascular cell and macaque hippocampus atlas	203
4.15	Fig. S9. GABAergic neuron subtype markers	204
4.16	Fig. S10. Schematic of extended lochNESS and additional lochNESS examples	205
4.17	Fig. S11. Schematic depicting snATAC-seq quality control pipeline	206
4.18	Fig. S12. Assessment of snRNA-seq/snATAC-seq integration quality.	207
4.19	Fig. S13. Integration of snRNA-seq and snATAC-seq data in cell classes	208
4.20	Fig. S14. Correlation of TF expression and motif enrichments . . .	209
4.21	Fig. S15. Concordance between GLUE regulatory scores and logistic regression standardized effect sizes	210
4.22	Fig. S16. Validation of marker peaks	211
4.23	Fig. S17. Number of cCREs interacting with a gene	212

ACKNOWLEDGMENTS

On the path to completing this PhD, I am deeply grateful for the many people who have supported me. First and foremost, I would like to thank my PhD advisor Jay Shendure for his support and mentorship. Jay is an amazing scientist and a great mentor, making it possible for me to work on interesting science and receive valuable training during my time in the lab. Jay is also an amazing person and colleague. He is always passionate and thoughtful, and is sincere and kind to everyone around him. It has been a pleasure to know him from all these aspects. I would like to thank my committee members, William Noble, David Beier, Neda Bagheri, and Luis Ceze, for their insightful feedback on my talks and dissertation, which helped me improve my work. Additionally, I would like to thank my earlier academic mentors Ashu Sabharwal and Erez Aiden for introducing me to the world of research during my undergrad and influencing my research interests and career trajectory.

The PhD projects would not have been possible without the support from people in both CSE and GS at UW, and collaborators from other institutes. I would like to thank Junyue Cao, Andrew Hill and Darren Cusanovich for the mentorship in the early years in the lab, and Chengxiang Qiu and Diego Calderon for being analysis buddies in the later years. I would like to thank all other members of the Shendure lab, especially Wei Chen, Wei Yang, Xiaoyi Li, Greg Findley, Lea Starita, Vikram Argawal, Charlie Lee, Molly Gasperini, Sanjay Srivatsan, Silvia Domcke, Anh Leith, Ronnie Blecher, Jacob Tome, Tony Li, Flo Chardon, Riza Daza, Beth Martin, Ruolan Qiu, Eva Nichols, Hanna Liao, Chase Suiter, Connor Kubo, David Lee, Jenny Nathans, Megan Taylor, Sereno Lopez-Darwin and Cooper Marshall for

the many wonderful scientific discussions, non-scientific conversations, as well as time spent together in the lab and on the volleyball courts. I would like to thank my comp bio buddies in CSE, Nick Nuechterlein, Lee Organick, Jacob Schreiber, Wei Qiu and Xinming Tu. I would like to thank Malte Spielmann, Jana Henck, Eileen Furlong, Stefano Secchia, Noah Snyder-Mackler and Kenny Chiou for the wonderful collaborations. I would like to thank Stephen Williams for being the most supportive internship mentor at 10x Genomics. I would also like to thank anyone else who provided administrative support during my PhD.

I would like to thank my friends for being with me through the PhD, which was complicated by the pandemic years. I would like to thank Sinan Xie, Wenjun Wu, Qisheng Li (and Sweet Potato the Corgi), Sang Lin (and Pipi and Doudou the Cats), HoiYan Wu, Mengdi Sun, Xiaoshi Quan (and Popcorn the Goldendoodle), Yuanyi Xu, Jasmine Li, Lianhui Qin, Beibin Li, Jiayin Qu (and Hulu and Sydney the Cats) and Yiran Zhao for exploring the food and mountains of the Pacific Northwest together. I would like to thank my friends Yuqi Tang, Xinyu Feng, Amber Tong, Xiaotong Xi, Shumao Zhang, Shiqi Li, Fangzhou Yu, Yiling Hu and Jing Feng for the many conversations throughout the years. I would like to thank my cat Gwen Yimiao Huang for being a constant source of comfort though the late nights and early meetings.

Finally, I want to express my deepest gratitude to my family, especially my parents, Hong Tang and Feijun Huang, for supporting me since the very beginning.

Chapter 1

INTRODUCTION

Embryonic development is one of the most fascinating and complex biological processes that occur in nature. The transformation of a single fertilized egg cell into a complex multicellular organism with multiple tissues and organs is quite remarkable. This process is a critical step in defining the emergence of complex life and has been of significant interest to scientists. Recent technological advancements in the field of single-cell sequencing have revolutionized the study of embryonic development. With the ability to generate genomic profiles of large numbers of single cells, researchers are now able to study the mechanisms of metazoan development with unprecedented resolution. This has opened new avenues for understanding how complex tissues and organs arise during embryonic development.

The objectives of my research projects in this field are to develop computational methods to analyze high-dimensional genomics datasets and apply these methods towards understanding the mechanisms of embryonic development and the cellular heterogeneity of complex tissues. This includes investigating both normal and disease-related genetic backgrounds, which are critical for understanding the underlying causes of developmental disorders.

1.1 Topics in this dissertation

In this dissertation, we aim to showcase the power of cutting-edge computational methods in the study of embryonic development by introducing three major projects that focus on

different model organisms and tissues. Each of these projects presents its own unique set of challenges and opportunities for the application of advanced computational tools.

The first project, the *Drosophila* Embryonic Atlas Project (DEAP), focuses on the study of embryonic development in the fruit fly *Drosophila melanogaster*, which is a widely used model for researching mechanisms of embryonic development due to its rapid development process, ease of manipulation and shared developmental mechanisms with other species. Utilizing single-cell combinatorial indexing (sci-) techniques, we aim to generate a comprehensive atlas of the gene expression and chromatin accessibility of cells from different stages of *Drosophila* embryonic development. The complex experimental design of this project, profiling embryos from overlapping time windows spanning *Drosophila* embryogenesis, provides a rich source of data for the development and application of new computational tools. More specifically, we apply sci-ATAC-seq [1] and sci-RNA-seq [2], to *Drosophila* nuclei from eleven tightly staged, overlapping windows of embryogenesis, profiling chromatin accessibility in almost one million, and gene expression in half a million, single nuclei. Leveraging the asynchrony of embryos within each collection, we train a neural network (NN) model on the genomics data to infer the precise age of each nucleus. We use a k -NN approach to connect cell states from adjacent time stages and construct a branching tree structure corresponding to lineage specification events in *Drosophila* embryogenesis. Our analyses result in continuous views of molecular and cellular transitions throughout the first 20 hours of fly embryonic development, where we can zoom into specific lineages and time windows of interest.

The second project, the Mutant Mouse Cell Atlas (MMCA), focuses on dissecting the phenotypic changes of mouse mutants through analysis of a comprehensive atlas of the transcriptome of cells from whole mouse embryos with various genetic perturbations. With MMCA, we set out to establish single-cell RNA sequencing (scRNA-seq) at the “whole embryo” scale as a tool for unbiased phenotyping of mouse genetic models. Mouse models represent an important tool to study human disease, particularly developmental disorders.

We apply sci-RNA-seq to profile 104 embryos of 23 different mouse mutants and 4 different wildtype backgrounds at 13.5 days after fertilization (embryonic day 13.5 or E13.5) in a single experiment. Towards evaluating the sensitivity of this approach, the severity of the mouse mutants investigated range from established multisystem disorders to single enhancer knockouts. The resulting MMCA consists of transcriptional profiles for over 1.7 million mouse cells. We develop an analytical framework for molecular phenotyping of cell type composition and trajectory changes, gene expression alterations and developmental phenotypes. The analyses allow us to identify mutant-specific cell type composition changes, as well as more subtle shifts in cell distributions. Intriguingly, some specific changes are shared by distinct mutants, suggesting that the molecular and cellular correlates of both gross and subtle anatomical phenotypes might be recovered by further scaling of this approach. Overall, we hope to establish whole embryo single-cell profiling as a powerful tool to systematically investigate developmental disorders.

The third project, the Macaque Brain Atlas (MBA), aims to generate a comprehensive atlas of cells from different regions of the rhesus macaque brain, providing a valuable resource for the study of brain function in a non-human primate model. Understanding the functioning of the primate brain is essential for comprehending human cognition, behavior, and disease. In this study, we have created a comprehensive, single-cell, multimodal molecular atlas of the macaque brain that covers 30 regions of the adult brain. We have analyzed over 2.58 million transcriptomes and 1.59 million epigenomes from single nuclei. Our computational methods revealed differences in cell type composition and variations in gene expression signatures across different regions of the brain. Moreover, we have identified 1.19 million novel candidate regulatory elements from the chromatin accessibility profiles, allowing us to investigate the cis-regulatory grammar and neurological disease risks in a cell type-specific way. Finally, we study the cell type-specific regulatory grammar by integrating the paired multi-omics datasets and identifying links between genes and regulatory regions. Overall, this open resource provides a multi-omics atlas that can aid in the exploration of

the evolution of the human brain and identifying potential targets for disease intervention.

In conclusion, these three major projects represent exciting opportunities for the development and application of new computational tools for the study of embryonic development and complex tissues. By analyzing complex datasets from different model organisms and tissues, we aim to gain a deeper understanding of the underlying mechanisms of embryonic development and contribute to the advancement of medical knowledge.

All chapters in this dissertation are modified from published (**Chapter 2**) or under revision work (**Chapters 3 & 4**). To provide context and motivation for the upcoming chapters, I will introduce prior work related to each project. This will help to establish the foundation upon which these projects are built and further motivate their importance.

1.2 Single-cell sequencing to build organism-scale atlases

In recent years, the emergence of single-cell sequencing has provided researchers a powerful experimental tool to generate molecular profiles of large numbers of cells. Depending on the experimental setup, single-cell sequencing can profile various elements of the genome in a high-throughput fashion. A commonly performed experiment is single-cell RNA sequencing (scRNA-seq), which profiles the messenger RNA contents in individual cells [3]. The resulting dataset can be summarized in a high-dimensional count matrix, where rows correspond to genes, columns correspond to individual cells and matrix elements are read counts that correspond to quantity of gene expression in a cell. Another type of experiment with increasing popularity is single-cell ATAC sequencing (scATAC-seq), which profiles the DNA contents of open chromatin [4]. Open chromatin regions usually contain important regulatory elements in the genome, including genes, promoters, and enhancers. The resulting dataset can also be summarized in a high-dimensional matrix, where rows correspond

to genomic regions, columns correspond to individual cells and matrix elements are binary values indicating whether the genomic region is open in a cell. Other modalities that can be characterized with single-cell sequencing include spatial transcriptomics [5], methylation state [6], chromatin organization [7] and combinations of modalities [8].

By applying single-cell sequencing to adult animal tissue samples, researchers can generate organism-scale cell atlases. The Tabula Muris Consortium is an example of using scRNA-seq to report a compendium of cell types in the mouse [9]. Taking 20 organs and tissues from 7 mice, the authors profiled a total of over 100,000 transcriptomes in 3-week-old mice of the C57BL/6JN background strain. Single-cell sequencing was performed on around half of the cells with a FACS-based cell capture, while the other half went through a microfluidic droplet-based system. Each organ was annotated separately to identify organ specific cell types. Through linear dimensionality reduction with PCA followed by graph-based clustering, the cells were grouped into clusters based on transcriptional similarity and assigned a cell type label based on expression of known marker genes. Similarly, a global analysis was conducted on all cells captured through FACS, yielding 54 cell clusters, 25 of which contained cells from multiple organs. The dataset establishes scRNA-seq as a tool to comprehensively capture different types of cells in a complex organism, including discovery of novel cell types that have not been identified with previous experimental techniques. This annotated dataset also allows researchers to identify novel markers for established cell types, compare gene expression of the same cell type across tissues, compare different scRNA-seq capture technologies and conduct cell type-specific transcription factor analysis.

Recently, advancements in single-cell technologies, reduction in sequencing costs, and increase in computational power have made generating and analyzing large-scale atlases more accessible to researchers. Single-cell sequencing is now applied to more complex samples with the aim of understanding complex biological systems and results in more complex datasets, as in each of the projects in the dissertation. Such complexity can arise from the addition

of a temporal component, as in the case of DEAP, various disease-related genotypes, as in the case of MMCA, or multiple tissues of origin, as in the case of MBA. These datasets present unique computational challenges, and thus the development of computational tools is essential to fill gaps in our efforts to gain specific biological insights from each dataset.

1.3 Temporal molecular dynamics in metazoan development

In DEAP, one of the main focuses is to gain a continuous view of molecular changes along trajectories in embryonic development. The biological system of our interest, embryonic development, is a highly dynamic process in which cells follow a complex branching lineage structure. In the embryonic development process, cells proliferate and differentiate into new cell types. These processes are continuous, rapid, and robust. For example, in the three days of mouse gastrulation, cells from one layer give rise to multiple major germ layers; and in the five days of mouse organogenesis, cells from the major germ layers give rise to all major organs. Although embryonic development has been intensively studied for over the past century, we are still far from a comprehensive understanding of the molecular events in the lineage specification processes in embryonic development.

In contrast with cells in adult tissues, cells in embryonic development are on continuous trajectories instead of in distinct states. A challenge in studying a continuum with single-cell datasets is the added time axis on top of a high-dimensional matrix, which requires specialized computational methods to place single cells on a continuum based on their molecular profile. Monocle 2 tackles this issue by inferring a “pseudotime” for each single cell in a dynamic population [10]. Monocle 2 uses reverse graph embedding (RGE) [11] to learn complex single-cell trajectory structures in an unsupervised fashion and study gene regulation underlying cell fate decisions along trajectories. RGE is a machine learning algorithm that finds a low dimensional embedding of high dimensional single-cell expression datasets and learns the

structure of the principal graph in the embedded space. Monocle 2 implements DDRTree, a scalable version of RGE, to generate a tree-like principal graph. The user can define a root, from which pseudotime can be calculated for each cell based on its distance from the root along the learned tree. Monocle2 also implements the BEAM algorithm [12] to identify branch-specific expression patterns. The accuracy of Monocle 2 is established by applying the algorithm to differentiating skeletal myoblasts and differentiating blood cells and showing that the algorithm captures both known trajectories and novel branches with specific gene expression programs. Monocle 2 is useful for systems with simple branching structures but may be insufficient for larger datasets with more complicated underlying lineage structure.

Another challenge to study cells in embryonic development is scale, *i.e.* the number of cells that needs to be profiled to capture all the lineage specification events. To tackle the issue of scale, previous work has applied single-cell combinatorial indexing (sci-) with three rounds of split-pool barcoding to increase the number of cells profiled in a single experiment [2]. The resulting technique, sci-RNA-seq3, was applied to mouse embryos in organogenesis to profile transcriptional profiles of around 2 million cells from 61 embryos across 5 time points spaced by 24 hours (E9.5 to E13.5) in a single experiment. The resulting Mouse Organogenesis Cell Atlas (MOCA) dataset is extensive, with 38 main cell types and over 500 subtypes resolved through global dimensionality reduction and clustering. The power of the scale of MOCA is shown by focused analysis of gene expression dynamics in the highly specialized AER cell subtype, which includes merely 1,237 cells or 0.06% of the dataset. These types of cells and dynamics would not be seen using other single-cell sequencing platforms given the complexity of mouse embryos at the end of organogenesis.

As in MOCA, to study the temporal dynamics in development with single-cell sequencing data, experiments are usually set up so that individual samples are obtained at discrete time points. Usually, the time stamp for each cell is provided by the time point at which the individual sample is collected in the lab, and cells within each time point are first analyzed

separately. As a result, computational methods have been specifically designed for time series in metazoan development to infer the structure and mechanisms underlying cell fate transitions. Accompanying the massive MOCA dataset, a computational package Monocle 3 [2] was developed to project cells into low dimensional space with UMAP [13], group transcriptionally similar cells with graph-based clustering and form “supergroups” by merging similar groups of cells. The “supergroups” are eventually placed along developmental trajectories, and the branching and convergence points are identified. Monocle 3 organized 1.5 million high quality MOCA cells into 12 distinct main trajectories and 56 sub trajectories with strong temporal components in the embedded space. A focused analysis of trajectories related to myogenesis identified potential regulators of myogenesis.

While Monocle 3 is successful at temporal ordering of a population of cells, the method outputs somewhat complicated network structures. While this structure accurately represents the dataset, it makes it difficult to focus on lineage specification events and the underlying molecular mechanisms. For datasets with developmental temporal resolution, it can be useful to add simplifying assumptions and force the reconstructed graph to be a branching tree, as cell types are mostly diversifying along the developmental time axis. URD is a diffusion-based algorithm that aims to reconstruct cell type specifying trajectories using single-cell transcriptomes from 38,731 cells in tightly staged zebrafish embryos in early embryogenesis [14]. URD is a multi-step algorithm: first, a k -NN graph is built based on transcriptional similarity and edge weights are saved as random walk transition probabilities; then, a pseudotime is assigned to each cell based on distance or number of random walks from a user defined root; finally, trajectories are constructed from user defined tips back to the root determined by random walks through cells backwards in pseudotime. The branching tree is eventually constructed by joining trajectories that pass through the same cells and visualized by placing cells on a branching tree structure in embedded space through a force directed layout. The developmental tree constructed allows for easy identification of gene modules that are active in each branch and differential expression at branching points. Fur-

thermore, inspection of the sequence of branching points in the branching tree may provide insight on the order of occurrence of lineage specification events. URD is a generalizable tree construction and pseudotime calculation method for scRNA-seq datasets with a temporal axis. However, treating each cell as an individual entity in the tree is computationally intensive, and may not scale well for the increasingly large developmental cell atlases.

Another method uses cell state, i.e. a group of transcriptionally similar cells in a time stage, as the basic entity in the developmental tree [15]. The authors profiled transcriptomes of 136,966 cells from 10 developmental stages across the first day of life (up to early organogenesis) of the frog with scRNA-seq. An explicit tree-like structure where branching points indicate emergence of a new cell state is created by linking cell clusters between neighboring stages using a k -NN based algorithm. Briefly, cells from neighboring developmental stages are co-embedded and for each cell from the later developmental stage k -NN cells are found in the earlier developmental stage. A connection is made between cell clusters using the consensus of per cell k -NN connections in the cluster. The resulting tree consists of 259 cell states or nodes across 10 developmental stages and largely recapitulates known lineage relationships. Each branch or lineage can be analyzed at single-cell resolution and show lineage specific gene expression dynamics as in previous methods. Interestingly, by building such compact trees for different species, researchers can start to align and contrast lineage specification events and identify gene homologues across species.

These papers along with many other comprehensive surveys of cellular diversity at profiled timepoints have shed light on the gene expression changes and dynamics of regulatory element activity during the *in vivo* emergence of each cell type in many organisms [1, 2, 14, 15, 16, 17, 18, 19, 20]. However, nearly all such “whole embryo” atlases of embryogenesis remain limited by sampling density, *i.e.* the number of discrete time points at which individual embryos are harvested and cells/nuclei collected. Given the rapidity with which molecular and cellular programs unfold, this limits the resolution at which regulatory transitions can

be characterized. For example, in the mouse, typically there are 6-24 hours between sampled embryonic time points, gaps within which massive molecular and morphological changes may take place [2, 19].

In most organisms, it can be quite hard to obtain many individual samples separated by small time steps to recover more continuous dynamics. In a study attempting to build a near continuous map for mouse gastrulation, the authors sampled about 33,000 single cells from 153 gastrulating mouse embryos with scRNA-seq [21]. The embryos were tightly sampled from the gastrulation period (E6.5 to E8.25). Taking advantage of the asynchronous development between embryos of the same litter, the authors could assign a timestamp to each embryo based on morphology and pseudobulk embryo transcriptional similarity. Each embryo was then assigned to one of 13 time groups with 0.2 day or 0.1 day separations spanning E6.5 to E8.25. The dataset was presented in a network flow model, where each node is a metacell (MC, or group of transcriptional similar cells), each column is a time group, and each edge is a connection between MCs in neighboring time groups. Connections are determined based on an optimal transport approach as the problem can be written as a min-cost max-flow optimization. Briefly, the cost function considers the inherent links between cells of different time groups within MCs and the transition probabilities between MCs in adjacent time windows based on transcriptional similarity in embedded space. The cost minimization problem is further constrained by the balance between influx and outflux of cells for each MC from and to adjacent time groups with adjustments based on cell type and time dependent growth rates. Similar to branches in previously described developmental trees, each flow corresponds to a lineage that can be traced back and can be used to study lineage-specific transcriptional dynamics. Overall, large amounts of work were required to collect sufficient samples to construct this nearly continuous model of mouse gastrulation, which spans 2 days of the 21-day mouse embryogenesis process.

To construct a continuous representation of embryogenesis *in vivo*, we would ideally

sample embryos continuously across the whole embryogenesis process. Although hardly possible with most model organisms, it is potentially possible in *Drosophila melanogaster*, where collections of timed and yet somewhat asynchronous embryos are easy to obtain, such that in theory one can achieve relatively high temporal resolution. *Drosophila* could therefore serve as a test case to develop a framework for the inference of continuous gene regulatory dynamics and cellular trajectories as embryogenesis progresses. As *Drosophila* is a preeminent model organism whose study has enabled many advances in the biological and biomedical sciences, obtaining a single-cell atlas of *Drosophila* is also an important goal alone. This includes its embryonic development, where its use in conjunction with powerful genetic tools has transformed our understanding of the mechanisms by which developmental complexity is achieved, in addition to uncovering many general principles of both genetic and epigenetic gene regulation.

Although synchronized egg laying yields embryos of known and somewhat homogenous developmental stages, the staging is based on morphological indicators that often span several hours such that even a staged two-hour collection will contain embryos at different “micro-stages”. In DEAP, we leverage this temporal variability within each collection time window, together with overlaps between adjacent time windows, to obtain a dynamic continuum. We develop computational methods to infer the developmental time from molecular profiles, construct the cell lineage structures and underlying molecular dynamics. In addition, in contrast to previous unimodal developmental atlases, we applied both sci-RNA-seq and sci-ATAC-seq to the same groups of embryos to study gene expression dynamics. With integrative analysis, we can characterize gene regulatory changes and cellular transitions across all major lineages of *Drosophila* embryonic development simultaneously.

1.4 Effects of genetic mutations in metazoan development

In MMCA, we hope to establish single-cell RNA sequencing at the whole embryo scale as a tool for unbiased phenotyping of mouse genetic models. Mouse models represent a critical tool to study human diseases, particularly developmental disorders. There has been a tremendous amount of effort put into investigating mutations correlated with human diseases in model organisms, particularly in mice. In an early study that disrupts *Tal1*, a frequently activated gene in T-cell leukemias that encodes a candidate regulator in hematopoietic development, the authors showed that *Tal1* is essential for embryonic blood formation in mouse embryogenesis [22]. Loss of *Tal1* was shown to be lethal between E8.5 and E10.5. Other phenotypes of *Tal1*^{-/-} embryos were determined through visual inspection and comparison with wildtype embryos. Major alterations include loss of blood in yolk sac vessels, smaller embryo size and absence of nucleated blood cells in traverse sections through the heart and placenta, while most other developmental events like formation of somites and the neural tube proceed as normal. Due to natural variation between embryos, observations on multiple embryos would be needed to draw concrete conclusions, requiring extensive manual work. It was also difficult to probe the molecular differences caused by genetic disruptions with technologies at that time.

Recently, the limit of introducing mutations has been lowered significantly through new technologies like the CRISPR-Cas9 system [23], but an unbiased whole organism analysis of potential phenotypes caused by mutations remains a challenge. A fundamental challenge to study pleiotropic developmental disorders *in vivo* is the lack of technologies with sufficient throughput and resolution to obtain a global view of the molecular states and trajectories of a rapidly diversifying and expanding number of cell types during embryogenesis. As described in the previous section, single-cell sequencing is a high throughput method capable of capturing diverse and detailed cell types in a complex organism and has been proven to

be widely successful in studying mouse embryonic development. While most studies focus on embryogenesis in normal (wildtype) mouse embryos, the effects of single mutations can be studied alongside these wildtype atlases [19]. The authors first constructed a molecular map of cellular diversification in mouse gastrulation by profiling 116,312 single-cell transcriptomes from embryos collected at 9 time points spanning E6.5 to E8.5. After establishing this single-cell map of gastrulation, the authors created *Tal1*^{-/-} chimeric embryos by injecting *Tal1*^{-/-}tdTomato⁺ embryonic stem cells into blastocytes and used the established map as a platform to dissect the effects of this genetic mutation. Wildtype (tdTomato⁺) and *Tal1*^{-/-}(*Tal1*^{-/-}tdTomato⁺) cells were sorted at E8.5, and their transcriptomes were profiled with scRNA-seq. These cells were then mapped back to the previously established wildtype atlas of gastrulation. The new wildtype cells were largely consistent with the cells in the wildtype atlas in terms of observed cell types, while *Tal1*^{-/-} cells did not map to the blood lineages. This lack of cell type in the embedded space is strong evidence that *Tal1* is essential for the first wave of blood development. By comparing the transcriptomes of *Tal1*^{-/-} cells mapping to the endothelial cell types to wildtype cells in the same cell types, the authors identified a downregulation of blood development associated genes and suggest a disruption to the second hematopoietic wave. Interestingly, focusing on the endothelial group EC3, it is shown that EC3-mapped *Tal1*^{-/-} cells are transcriptionally similar to EC3-mapped *Tal1*^{-/-} cells with expression of endothelial genes, but have an additional upregulation of cardiac-related genes, indicating involvement in mesodermal programs when unable to proceed in blood development.

These observations are consistent with previous studies and show the power of single-cell sequencing for phenotyping whole embryos with mutations. However, *Tal1*^{-/-} is a later lethal mutation with strong phenotypic changes that is even obvious under visual inspection. Researchers are also interested in understanding the power of single-cell sequencing in phenotyping mutants causing more subtle changes in embryogenesis. In a follow up study to the single-cell map of mouse gastrulation, the authors reanalyzed the somatic and paraxial

mesoderm, NMPs and spinal cord cells from the wildtype atlas, identifying multiple trajectories that underlie somitogenesis based on a “chorus” of expression dynamics of known gastrulation, NMP signature, mesoderm development and somatic mesoderm related genes. Primitive streak cells at E6.5 give rise to anterior somatic tissues, posterior somatic tissues and NMPs E8.5 in three separate waves. *T* is an established NMP marker co-expressed with *Sox2* and *Nkx1-2*. Interestingly, the authors observed differential expression of *T* between the early anterior and posterior somatic trajectories within the wildtype atlas. While the homozygous *T* mutant mouse model is shown to be embryonic lethal due to arrest in axis elongation and defects in neural tube, allantois and notochord, changes versus wildtype at E8.5 have not been dissected in detail. The authors apply single-cell profiling to *T*^{-/-} chimeric embryos and additional wildtype cells at E8.5 and map the cells back to the reference wildtype atlas to acquire cell type labels. Differential cell type abundance testing showed that in *T*^{-/-} chimeric embryos, the intermediate and somatic mesoderm cells are significantly reduced and NMP cells are increased. To understand when the changes occur along trajectories, the authors generated additional single-cell data for *T*^{-/-} chimeric embryos and wildtype embryos at E7.5. The cells from *T*^{-/-} chimeric embryos tend to contribute less to the posterior somite trajectory, contribute more to the NMP trajectory and remain stable in the anterior somite trajectory. Overall, the single-cell data revealed that in the absence of a regulatory gene, cells are diverted from one trajectory to another. This further established scRNA-seq as a whole embryo phenotyping tool in mouse embryogenesis, with the resolution to identify changes in cell fate decisions after genetic disruption.

Compared to these previous studies, MMCA is a more comprehensive dataset in terms of the number of mutants included. We were able to profile large amounts of cells from a variety of mouse mutants and different background strains in the same experiment using sci-RNA-seq³. We included mouse mutants of various severities and different levels of prior understanding, ranging from pleiotropic effects of multisystem disorders, mouse models of developmental disorders and unpublished loci that have not been studied before, to deletions

of TAD boundaries in the vicinity of key regulatory genes in development. In a single experiment, we profiled over 1.9 million cells from 104 embryos of 23 different mouse mutants and four different wildtype backgrounds at embryonic stage E13.5. Due to the unique experimental setup and scale of the resulting dataset, we needed computational methods to dissect cell type composition changes and transcriptional profile shifts in many genotype, despite having limited replicates. While this is a new challenge for computational tool development, we can take inspiration from many existing tools previously developed for various other purposes.

One of our goals in analyzing MMCA is to understand the cell type composition changes related to each genotype. While it is straightforward to calculate ratios between cell type compositions of mutant embryos versus wildtype embryos, it is challenging to assess the significance of each composition change due to limited sample sizes in each group. One of the recent models that detect cell type compositional changes from single-cell datasets with low sample sizes is scCODA [24]. A challenge in detecting cell type composition changes is that the biological change in the proportion of one cell type leads to inherent opposite changes in other cell types, leading to inflation in differential detection if not controlled for. To account for this inherent bias, the authors introduce a Bayesian model that is sensitive to changes due to disease or other stimuli while reliably controlling for false discoveries. The scCODA framework accounts for the uncertainty in cell type proportions by modeling counts as a hierarchical Dirichlet-Multinomial distribution, and accounts for the negative correlation bias by modeling all cell type proportions together. Then the model uses a Logit-normal prior with a log-link function to estimate effects of covariates on cell type proportions. The model also automatically selects an appropriate cell type as reference, so that detected changes can be interpreted in relation to this reference. In MMCA, we used a more straightforward regression-based model to detect differential cell type proportions and were conservative in interpreting the proposed changes.

Another important goal in our comparative analysis is to identify differences in single-

cell transcriptional profiles between mutant and wildtype cells in specific cell types, which would lead to differential abundance between mutant and wildtype in certain groups of cells. Compared to methods detecting cell type composition changes in pre-defined cell types, we would like to develop a method that can detect such changes in a label-agnostic fashion. First, we took inspiration from an established metric in geospatial analysis called Moran’s I, which captures spatial autocorrelation and is commonly used to examine the degree of clustering or dispersion across a geographical region [25]. Moran’s I takes on values between -1 and +1, with values towards -1 indicating dispersion, 0 indicating random distribution and +1 indicating clustering. In the context of single-cell sequencing and spatial transcriptomics, Moran’s I can be used to identify genes or features that exhibit spatial autocorrelation in embeddings or spatial slides [2, 26]. In the context of MMCA, we can potentially lift over the concept of Moran’s I to compare two samples in single-cell datasets by calculating Moran’s I for the mutant label (1 = mutant cell, 0 = wildtype cell) for cells in a co-embedding of mutant and wildtype cells. If mutant and wildtype cells are similar, they would be randomly mixed in the co-embedding and thus the Moran’s I value for the co-embedding would be close to 0. In contrast, if mutant and wildtype cells are somewhat dissimilar, they would be clustered separately in the co-embedding and thus the Moran’s I value for the co-embedding would be close to +1. However, Moran’s I essentially summarizes a co-embedding as a single value and is oversimplifying for our purpose.

Another type of method that can be ported over is batch mixing metrics. In the context of assessing whether a batch correction algorithm successfully integrates two batches of single-cell sequencing data, a metric may be used to quantify the amount of mixing of the two datasets before and after batch correction. For example, LISI (Local Inverse Simpson Index) is a metric used by the batch correcting algorithm Harmony to measure the degree of batch mixing by counting the effective number of batches represented in each cell’s k -NNs [27]. LISI ranges from 1 to 2, with the value of 1 indicating batches not mixing (all cells in the k -NNs are from the same batch) and 2 indicating well mixing in the cell’s k -NNs. However,

batch mixing metrics like LISI are insufficient for our purposes as they would not differentiate between enrichments and depletions and are not robust on imbalanced datasets.

While developing our own analysis framework for MMCA, researchers developed a similar package in parallel called Milo [28], a statistical method to test for differential abundance between two conditions in single-cell datasets. Instead of performing differential abundance testing in discrete clusters as in previous methods, Milo assigns cells to partially overlapping neighborhoods based on a k -NN graph, in which cell counts are modelled with negative binomial GLM and hypothesis testing is to identify neighborhoods with differential abundance of cell groups. The GLMs account for batch effects and variability in cell numbers across samples and makes Milo robust to complex experimental designs. When tested on several simulated and real datasets, Milo could identify perturbation effects that were obscured by clustering results while controlling for false discovery rates. In MMCA, we developed a similar framework. We calculate a score termed lochNESS for each cell based on the relative abundance of cells from different genotypes in the cell's k -NNs. Given the scores, we can later fit regression models to identify genes with expression associated with the scores and identify marker genes for cells in a cell type agnostic fashion.

1.5 Dissecting molecular heterogeneity and regulatory information in complex brain tissues

The study of complex brain tissues has been a long-standing challenge in the field of neuroscience, as the intricate relationships between cells, their functions, and their interactions are crucial for the proper functioning of the brain. Due to the complexity and importance of the brain, as single-cell sequencing and spatial transcriptomics techniques emerged, the brain has been the natural first step in testing such techniques. The generation of brain atlases in many different organisms has revolutionized the study of the brain, enabling re-

searchers to generate high-resolution genomic profiles of large numbers of individual cells from the brain. These brain atlases have rapidly increased in depth and scope as single-cell sequencing techniques improve.

In an earlier single-cell transcriptomic atlas of the mouse brain, Tasic and colleagues analyzed around 24,000 cells from the layers of the primary visual cortex and the anterior lateral motor cortex of the mouse neocortex [29]. Using deep single-cell transcriptional profiling via SMART-Seq v4, the authors were able to define 133 cell types organized into a taxonomy. The authors used the single-cell dataset to study cell type-specific, region-specific and layer-specific gene expression, observing that GABAergic neurons were shared in both areas profiled but glutamatergic neurons were mostly specific to one of the two areas. The authors also combined retrograde labelling, which traces neural connections from the synapse of the cell body, allowing for identification of connections between neurons in a terminal tissue and the target neurons in the brain. Glutamatergic neuron types were resolved using a combination of projection patterns from retrograde labelling, gene expression markers from scRNA-seq, and layer information from dissection. Overall, the atlas did not include a large number of cells, but with detailed dissection and deep sequencing, the authors were able to resolve cell type identities of very detailed cell types and characterize projection patterns, expression markers and layers/regions.

In a more recently published adult mouse brain epigenomic atlas, Li and colleagues profiled over 800,000 individual nuclei from 45 brain regions spanning the mouse cerebrum using sci-ATAC-seq [30]. The resulting dataset was used to identify 43 subclasses of cells, which were then further clustered into 160 cell types. Leveraging the large amount of brain regions profiled, the authors found high regional specificity in most glutamatergic neuron types, some GABAergic neuron types and a small set of non-neuronal cell types. The atlas characterized 491,818 candidate cis-regulatory elements (cCREs), which were differentially accessible in the different cell types. Then the authors used a computational package Cicero

[31] to link cCREs with target genes based on co-accessibility in cells, revealing various gene regulatory programs indifferent cell types. Finally, the authors used the human orthologs of cell type-specific mouse cCREs to identify potential genetic variance that are associated with neurological traits.

While mouse brain samples are the most accessible mammalian brain tissue and commonly used, non-human primate (NHP) models are important for several reasons in neuroscience research. NHPs are the closest living relatives to humans, sharing over 90% of our genetic makeup. This genetic similarity makes NHPs a valuable model for studying human disease and testing new drugs and therapies before they are used in human clinical trials. On the other hand, while rodent models have been used extensively in preclinical research, there are many important differences between rodents and humans in terms of anatomy, physiology, and behavior. Therefore, the use of the macaque model in the Macaque Brain Atlas (MBA) provides several advantages over traditional rodent models, as non-human primates are more closely related to humans, both anatomically and physiologically, providing a more accurate representation of human brain function. Overall, NHP models are important for advancing our understanding of human neurobiology and diseases, and for developing new treatments and therapies that benefit human health.

In recent years, many NHP brain atlases have been generated. One of the more recent examples is a single-cell transcriptomics atlas of the adult macaque hippocampus [32]. The authors profiled over 200,000 cells and identified 34 cell populations comprising all expected cell types in the hippocampus. This dataset helped identify novel cell type markers (e.g. *HMGB2* as a novel IPC marker) and because the dataset included samples from both young and middle aged adults, made it possible to study adult neurogenesis in the primate hippocampus.

Previous single-cell atlases of the NHP brain have largely focused on a limited number

of cell types or regions, lacking comprehensive representation of the entire brain. This has limited the ability to study the heterogeneity between different cell types and cells in different brain regions. The MBA aims to fill this gap by generating a comprehensive, high-resolution map of the brain at the single-cell level by profiling from a wide range of macaque brain regions. Additionally, previous single-cell atlases generally focus on a single modality, gene expression or chromatin accessibility, while MBA is a multi-omic atlas. Thus, the MBA provides a more comprehensive view of the molecular landscape of the NHP brain, allowing detailed analysis of the heterogeneity between cell types and brain regions, and enabling the creation of cell type-specific regulatory maps.

In the analysis of the MBA, since we performed sci-RNA-seq and sci-ATAC-seq on the same samples, an important aspect was the integration of the single-cell RNA and ATAC datasets. This task more generally is the integration of “unpaired” multi-modal single-cell datasets. The modalities correspond to various genomics layers, such as gene expression from scRNA-seq and chromatin accessibility from scATAC-seq, and “unpaired” refers to different omics layers not profiled in the same set of cells, but rather in independent samples from the same cell population. Previous methods that focus on this task either relied on preprocessing to match the features (generating a gene activity matrix by aggregating ATAC reads near a gene’s promoter region) or construction of a computationally expensive cell-cell similarity graph [33, 34]. The applicability of such methods to MBA, which profiles a large number of cells but has relatively shallow coverage per cell, is quite limited. A recent ML-based method scGLUE [35] was capable of integrating the 4.2 million cell MBA. The authors of scGLUE implemented a Graph Linked Unified Embedding (GLUE) model for unpaired single-cell multi-omics datasets. The model relies on guidance from a knowledge graph, which is automatically generated from the dataset. In the case of integrating RNA and ATAC datasets, the nodes of the knowledge graph are RNA genes and ATAC peaks, while the links are established based on the genomic distance between genes and peaks. The model takes the knowledge graph and single-cell datasets as input and generates feature embeddings

and cell embeddings as output, which can be used in downstream analysis to build regulatory maps.

Another important aspect of the analysis of MBA was gene regulatory inference, linking RNA genes to nearby ATAC peaks with regulatory function (cis-regulatory elements, CREs). In scGLUE [35], the authors demonstrated how the feature embeddings can be used for regulatory inference. Specifically, the cosine similarity between the feature embeddings of RNA genes and ATAC peaks was used to establish links between genes and potential CREs. Apart from similarity-based metrics, another type of method used for regulatory inference is regression. In a multi-omic dataset or an integrated dataset, regression models predicting gene expression using chromatin accessibility profiles, or vice versa, can identify cases where open-ness of candidate CREs are correlated with expression of genes and establish regulatory links. A recent preprint demonstrates the use of Poisson regression to generate cell type-specific gene regulation maps in immune cell types with multi-omics datasets from RNA & ATAC co-assays [36]. In this study, the authors aim to construct cell type-specific enhancer gene maps, with the overarching goal of understanding causal variants and their target genes from GWAS. The authors developed a computational method SCENT to build such enhancer gene maps based on multi-omic single-cell RNA & ATAC datasets. SCENT was applied to 9 multiome datasets, creating cell type-specific enhancer gene maps. The authors then characterize these maps by multiple systematic comparisons and showing specific examples. In the MBA, we elected to use an ensemble model, combining links identified with the similarity between scGLUE feature embeddings and results from a logistic regression model. This model identifies a high confidence set of CREs, and when further integrated with differential accessibility results, helped us construct cell type-specific regulatory maps.

Chapter 2

THE CONTINUUM OF DROSOPHILA EMBRYONIC DEVELOPMENT AT SINGLE CELL RESOLUTION

This Chapter is adopted from published work with minimum changes:

D., Calderon*, R., Blecher-Gonen*, **X.**, **Huang***, S., Secchia*, J. Kentro, R. M. Daza, B. Martin, A. Dulja, C. Schaub, C. Trapnell, E. Larschan, K. M. O'Connor-Giles, E. E. M., Furlong#, and J., Shendure#, “The continuum of drosophila embryonic development at single cell resolution”, Science, 2022.

In collaboration with the Furlong lab, Ronnie started the project to perform multi-omic profiling of *Drosophila* embryos, extending Darren’s previous sci-ATAC-seq atlas of *Drosophila* embryogenesis. During lab meetings, I was impressed by the scope of the experiments, and it was a pleasant surprise when Diego and I were invited to join the analysis efforts in September 2019. In the early days, we spent a lot of time setting up the processing pipelines. To expedite the work, we divided the tasks such that Diego focused on sci-RNA-seq and I on sci-ATAC-seq data. The sci-ATAC-seq processing pipeline was relatively new at the time and took a while to set up on the updated cluster system. The sci-ATAC-seq downstream analysis pipeline was also challenging due to the dataset’s size, which exceeded R’s memory limits even when reading in half of the data. As a result, most of the resource-intensive analysis was completed in chunks and then merged. Our analysis gained traction when Diego proposed training a regression model to infer a time stamp for each single cell. These time stamps enabled the later analyses I focused on, such as annotating sci-ATAC-seq clusters in each time window, building lineage trees, and conducting a detailed analysis of the mesoderm system. Although waking up at 8AM PST was challenging, our project moved

much faster once we began regular meetings with Eileen and Stefano on Zoom. At some point during these meetings, we started using the acronym DEAP. The revision process was also quite a journey, with some lengthy reviews and unusually tight deadlines. With Ronnie back in Israel and new collaborators from Brown joining, our calls included people from three different countries and four different time zones. Despite the challenges, Diego and I were able to reprocess all the data, create a more complex neural network-based time inference model, and add lots of computational validation and new analyses. I was able to squeeze in some spatial mapping results and create an interactive webpage for data exploration at the last minute. I remember the day before the revision submission deadline was a work party where Diego and I generated all the supplementary figures. In 2022, I presented our work at the *Drosophila* Research Conference, which was a fantastic experience with thousands of researchers who share the appreciation for *Drosophila* as the best model organism.

More formally, the author contributions are listed in the manuscript as follows: S.S. and A.D. performed embryo collections and carried out embryo fixation and nuclei isolation for sci-ATAC-seq. B.M. and R.B. performed the sci-RNA-seq experiments. R.D. and R.B. performed the sci-ATAC-seq experiments. D.C., X.H., S.S., and R.B. carried out the formal analysis. D.C., R.B., and X.H. wrote the first draft of the manuscript, which was then finalized with guidance from E.E.F. and J.S. and input from all authors. S.S. performed data analysis, results interpretation and participated in figure generation and manuscript writing. C.S. performed in situ hybridization experiments. J.K. aided with additional analyses and fine scale neuronal cell type annotation supervised by E.L and K.M.O.-G. C.T. contributed to conceptualizing the time modeling and comparisons to pseudotime, and revised the manuscript. E.E.F. and J.S. conceptualized, supervised and funded the project.

Abstract *Drosophila melanogaster* is a powerful, long-standing model for metazoan development and gene regulation. We profiled chromatin accessibility in almost one million, and gene expression in half a million, nuclei from overlapping windows spanning the entirety

of embryogenesis. Leveraging developmental asynchronicity within embryo collections, we applied deep neural networks to infer the age of each nucleus, resulting in continuous, multi-modal views of molecular and cellular transitions in absolute time. We identify cell lineages, infer their developmental relationships, and link dynamic changes in enhancer usage, transcription factor (TF) expression and the accessibility of TFs' cognate motifs. With these data, the dynamics of enhancer usage and gene expression can be explored within and across lineages at the scale of minutes, including for precise transitions like zygotic genome activation.

2.1 Introduction

Single cell technologies are a powerful new means of studying metazoan development, shedding light on the emergence of cellular diversity and the dynamics of gene regulation. However, nearly all such atlases of embryogenesis are limited in terms of the number of discrete timepoints and cells sampled per timepoint. Given the rapidity with which molecular and cellular programs unfold, this limits the resolution at which regulatory transitions can be characterized.

To more completely represent development, embryos would ideally be sampled continuously. Although impractical for most model organisms, it is feasible in *Drosophila*, where collections of timed and yet somewhat asynchronous embryos are easy to obtain, such that in principle one can achieve arbitrarily high temporal resolution. This sharply contrasts with mouse, for which there are typically 6-24 hrs between sampled timepoints, gaps within which massive molecular and morphological changes take place [2, 19, 21, 37]. Although sampling gaps can be computationally filled via the continuum of cell states represented in single embryos [21, 38], the asynchronous ages of *Drosophila* embryos within staged collections presents an opportunity for more bona fide continuity, *e.g.* with seconds or minutes separating the developmental ages of consecutive embryos, rather than hours or days. More-

over, as *D. melanogaster* is a preeminent model organism that has yielded many discoveries and general principles of metazoan development and gene regulation, obtaining a single cell atlas of *Drosophila* embryogenesis is an important goal in itself.

2.2 Results

We set out to measure chromatin accessibility and gene expression from individual nuclei spanning a continuum of *Drosophila melanogaster* embryogenesis. Staged embryos were collected in 11 overlapping time windows, collectively 0-20 hrs, covering the entirety of embryogenesis at 25°C. Overlapping 2 hr collections were used to capture the rapid transitions during early stages, followed by overlapping 4 hr collections from 3 hrs onwards (**Fig. 2.1A**). From each collection, samples were split and separately processed for ATAC-seq or RNA-seq. Although we hereafter refer to cells, all data were generated from nuclei. Single cell profiling was conducted using three-level combinatorial indexing (sci-ATAC-seq3 and sci-RNA-seq3) with minor modifications [16, 2].

Sci-ATAC-seq3 and sci-RNA-seq3 libraries were sequenced to generate 30 billion and 6.8 billion raw reads, respectively (**Fig. 2.6**). After deduplication and application of quality filters, we obtained chromatin accessibility profiles for 976,460 cells (scATAC: median 5,206 non-duplicate reads per cell) and gene expression profiles for 547,805 cells (scRNA: median 399 unique molecular identifiers (UMIs) and 274 genes detected per cell). Although our scRNA data has fewer UMIs per nucleus than previously obtained from *Drosophila* embryos [39], we profiled many more nuclei spanning many more stages of embryogenesis, and complement this with scATAC with a high number of unique reads per nucleus. Given the small size of the *Drosophila* embryo, such deep ‘shotgun cellular coverage’ should effectively sample all tissue types during embryogenesis. The data did not appear to be confounded by batch effects (**Fig. 2.7A-G**).

For both data modalities, integrating and visualizing single cell profiles across all time-points resulted in branching structures going from early to late stages, consistent with increasing complexity (**Fig. 2.1B-C**). From the scATAC data, we identified 110,185 regions exhibiting accessibility at some point during embryogenesis. Collectively, these candidate regulatory elements cover 30.4 Mb (22%) of *Drosophila* euchromatin (dm6) and include 85% of known embryonic enhancers, based on overlap with nearly 5,000 curated enhancers confirmed in transgenic embryos (**Fig. 2.1D**) [40, 41, 42]. This, together with the high coverage of both bulk DHS peaks (87%) and scATAC-derived peaks (98%) from 2-12 hrs [1, 43] supports the comprehensiveness of this compendium. Similar results were obtained computing overlaps on a per-base rather than per-element basis (**Fig. 2.7H**). We additionally uncovered over 40,000 distal accessible regions not identified in these previous studies (**Fig. 2.1D**), which are enriched for enhancer-associated histone marks, suggesting that they are new developmental enhancers (**Fig. 2.7I**). The compendium also recovered 94% of 8,008 extensively validated mesodermal *cis*-Regulatory Modules [44], and 96% of nearly one million ChIP-defined binding sites across 233 TFs [45]) (Fig. 2.7J).

In exploring these data, we identified thousands of genomic regions and transcripts whose accessibility and expression levels, respectively, were strongly correlated with the progression of developmental time (**Fig. 2.1E-F**). Interestingly, not all of these correlations were cell type-specific (**Fig. 2.8**). The presence of such time-dependent elements/transcripts suggests that a dynamic process is unfolding across development, at least some aspects of which are cell-type specific, while other aspects appear general to germ layers or the entire organism. We reasoned that we could leverage these correlations to build a model to predict absolute developmental age of any given nucleus with greater temporal resolution than our 2-4 hr collection windows.

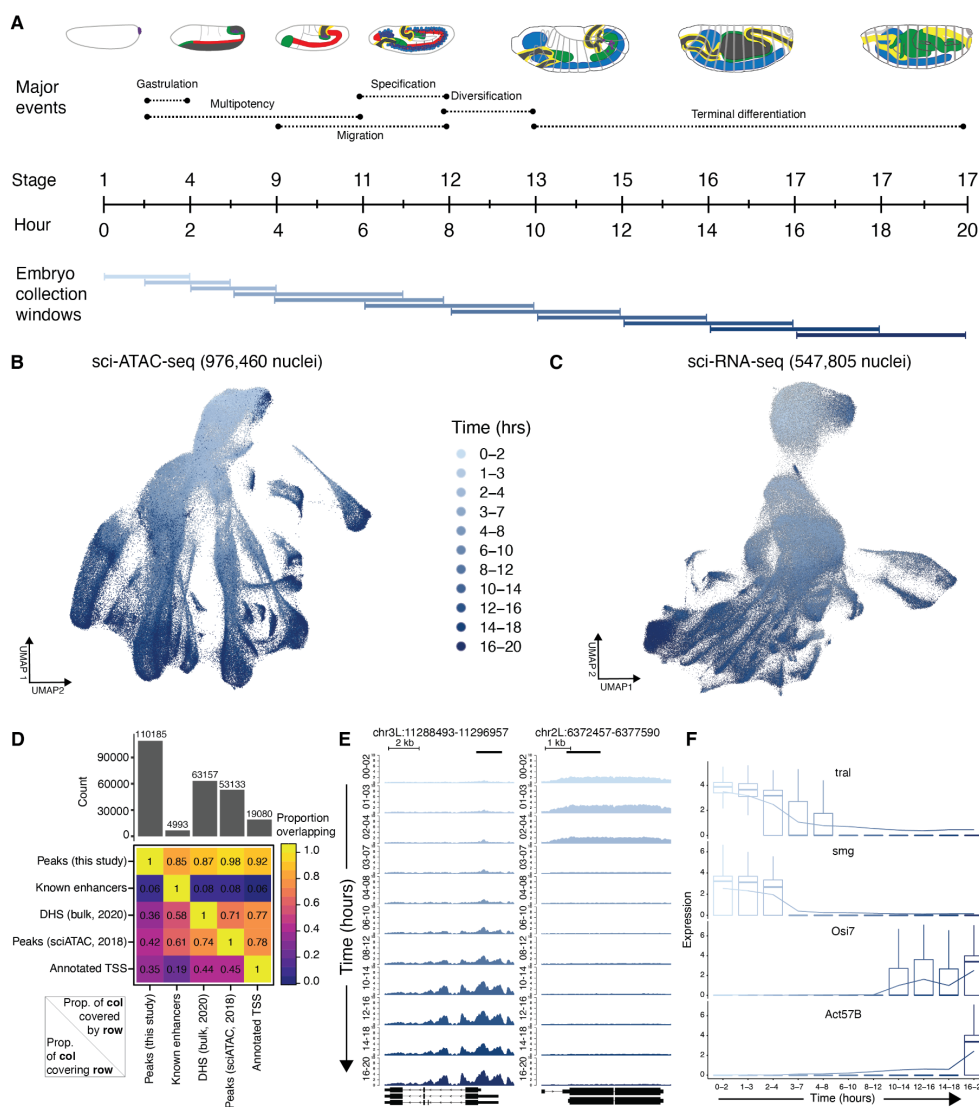


Figure 2.1: Figure 1: Single cell profiling of chromatin accessibility and gene expression throughout *Drosophila* embryogenesis. (A) Eleven overlapping collection windows that collectively cover embryogenesis. (B) UMAP visualization of cell-x-peak matrix of evenly time-sampled sci-ATAC-seq nuclei that passed QC. (C) Same as panel B but for sci-RNA-seq. (D) Heatmap showing proportion of our scATAC peaks overlapping ~5,000 curated enhancers [41, 42, 40], bulk DHS peaks from 2-12 hrs [43], scATAC peaks from 2-12 hrs [1] or annotated transcriptional start sites (TSS) [46]. (E) Chromatin accessibility, normalized by counts per million reads, across representative regions exhibiting time dependence across 11 collection windows. (F) Gene expression of representative genes exhibiting time dependence across 11 collection windows. Read counts were normalized, multiplied by a scale-factor, log-transformed after the addition of a pseudocount, and averaged across all cells within each window.

2.2.1 Predicting the absolute age of individual nuclei

In these data, the precise developmental age of each sampled nucleus is unknown, only the 2-4 hr collection window from which it derived. To estimate the age of each nucleus with greater precision, we fit a series of models using either the scATAC or scRNA data as input, and predicting the center hour of the collection window from which any given nucleus was obtained (**Fig. 2.2A**). Specifically, we split a subset of each dataset, evenly subsampled with respect to time, into 11 partitions, 10 of which were used as training data to fit either a lasso linear (LL) model or a neural net (NN)-based model with 10-fold cross-validation across various test parameters. After selecting the highest performing parameterization, the NN-based models markedly outperformed LL models, for both data types, in predicting the developmental age of nuclei within the held-out 11th partition (for NN vs. LL: Mean-Squared-Error (MSE) \rightarrow ATAC=5.26 vs 8.8. RNA=2.54 vs 4.72; proportion correct \rightarrow ATAC=0.67 vs 0.53, RNA=0.87 vs 0.65). We therefore moved forward with NN-based nuclear age predictions for the remainder of this study (**Fig. 2.2B**; **Fig. 2.9**).

To further assess accuracy, we applied the scRNA-derived models to a bulk RNA-seq time-course of staged embryos in 2-hr intervals [47], and found high concordance between predicted and actual developmental age (**Fig. 2.2C**). The scATAC-derived models were similarly able to order a time-course of bulk DNase-seq data from either whole embryos or specific FACS-purified lineages [43] (**Fig. 2.2D**). To assess predicted ages at much finer time scales (minutes rather than hours), we focused on genes whose expression is activated at specific nuclear cycles during zygotic genome activation (ZGA) [48]. Genes turning on during ZGA were dynamically upregulated in association with predicted nuclear ages (scRNA-based; 5 min increments), while maternal and silent genes were not (**Fig. 2.2E**). Early dynamically accessible enhancers and promoters could similarly be predicted (scATAC-based; 1 min increments) (**Fig. 2.2F**), opening in the same order as previously observed by bulk ATAC-seq of hand-picked embryos at 3 minute intervals (**Fig. 2.2G**) [49]. To further illustrate the

value of this framework, we note pseudobulk profiles corresponding to collection windows lead to piecewise expression dynamics (**Fig. 2.2H**). In contrast, pseudobulk profiles based on model-predicted ages yield more continuous dynamics (**Fig. 2.2I**).

Although there are similarities between the goal of our approach and the concept of ‘pseudotime’ [50], a key advantage of inferred age is that, both in training and prediction, cells are anchored to absolute time, which enables more interpretable ordering of cellular processes, as well as their ‘synchronization’ across lineages. One concern is that contamination with embryos whose developmental age falls outside the collection window will have exaggerated confounding effects on early time-points, as older embryos contain vastly more nuclei. Consistent with this, our model predicted that 2.8% of the $\sim 80,000$ scRNA-profiled cells from 0-2 hrs were at least 4 hours in developmental age. These ‘older’ cells represent the overwhelming majority of a discrete cluster in UMAP space (**Fig. 2.10A**). Similar contamination is also observed with scATAC profiles from this early time window (12.7% of $\sim 20,000$ cells; **Fig. 2.10B-D**). Clustering and visualizing only the cells inferred to be 0-2 hrs in age eliminates this developmentally advanced cluster (**Fig. 2.10E**).

2.2.2 Annotation and inference of diversifying developmental trajectories

To systematically track the emergence and diversification of developmental trajectories, we used inferred ages to separately process and cluster cells from a series of 2-hr non-overlapping time windows. Clusters were then annotated by leveraging stage-matched information on gene expression from thousands of *in-situ* hybridizations spanning embryogenesis, as well as extensive enhancer activity data [1, 51, 52] (**Fig. 2.3A-B**).

Of note, the last few hours of the time course had reduced numbers of inferred cells (*e.g.* after 18 hrs, 61% fewer than would be expected under uniform sampling) and fewer identified clusters (**Fig. 2.11A**). We suspect that this may be due to ‘edge effects’ of the model, as

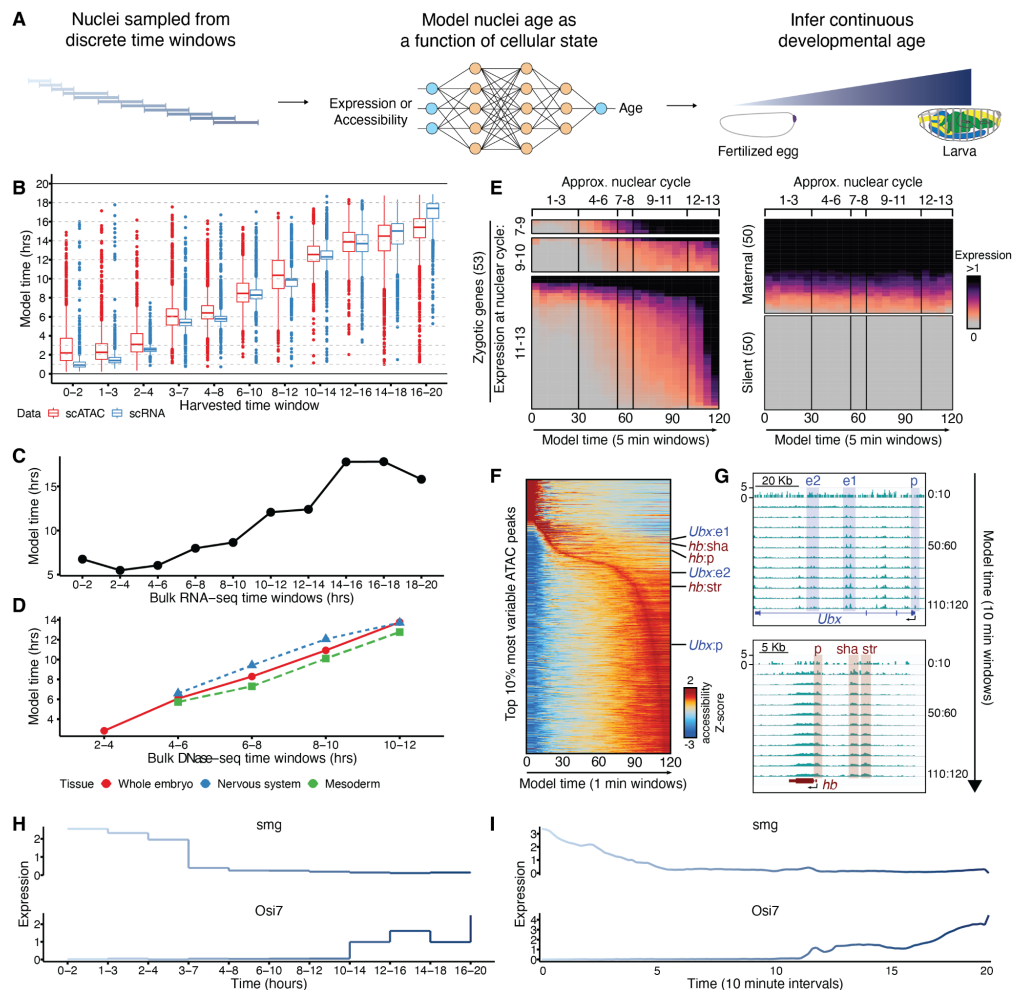


Figure 2: Inferring developmental age from cellular state. (A) We fit a NN-based model that uses either gene expression or chromatin accessibility to predict the center hour of the time window from which each nucleus was sampled. (B) NN model predicted-developmental ages (y-axis) of test set nuclei, equally sampled from discrete time windows (x-axis) and not included in model training. (C) NN model predicted-developmental ages (y-axis) of bulk RNA-seq samples [47] collected from 2-hr windows (x-axis). (D) NN model predicted-developmental ages (y-axis) of bulk DNase-seq samples from either whole embryo or purified tissues collected from 2-hr windows (x-axis). (E) Expression of zygotic (left), maternal (top right) or silent (bottom right) genes in nuclei from predicted-age windows in 5-min increments across 0-2 hrs of development. (F) Accessibility of most variable scATAC peaks from predicted-age windows in 5-min increments across 0-2 hrs of development. Labels indicate regions illustrated in panel G. (G) Examples of *cis*-regulatory regions known to exhibit dynamic accessibility in early embryos [49]. (H,I) Examples of time-associated genes, with expression values averaged across all nuclei from indicated collection windows (H) or from predicted-age windows in 10-min increments (I).

we also observe reduced numbers of inferred cells for the first several hours, although there they have less impact because the data from early timepoints lacks extensive structure. For this reason, we excluded cells with an inferred age of >18 hours from this set of analyses.

Here we use ‘cell state’ to mean an annotated cluster at a given time window. Altogether, we identified 171 cell states in sci-ATAC-seq data and 268 in sci-RNA-seq data across the 9 time windows, each of which received one of 38 cell type annotations for ATAC or one of 54 cell type annotations for RNA (**Table S1-S2**; **Fig. 2.3A-B**). Across time windows, we identified an average of 109 marker genes and 2469 marker accessible regions per cluster (**Table S3-S4**).

The early stages of *Drosophila* embryogenesis, represented by our 0-2 hr time window, include 13 rapid nuclear divisions within a syncytium that generates 6,000 nuclei, regulated by maternal genes. At approximately 2 hrs and 20 minutes post-fertilization, cellularization occurs and the zygotic genome is activated [53], followed by gastrulation to generate the three germ layers. Our single cell data recapitulates these events, where the earliest time window (0-2 hrs) has two large clusters annotated as maternal or unknown. At 2-4 hrs, the maternal cluster is no longer present and instead, pole cells and anlage clusters appear. A dramatic expansion in the diversity of cell types follows across 6-10 hrs, matching expectations for when the major lineages in each germ layer are specified (**Fig. 2.3A-B**).

To follow the emergence and diversification of cell lineages, we systematically linked cell clusters across developmental time, applying similar methods as [15, 37] to co-embeddings of cells from adjacent, non-overlapping, inferred time windows (**Fig. 2.11B,C**). For cells of each state derived from the ‘child’ time window, we calculated the median proportion of nearest neighbors from the ‘parent’ window that were derived from each potential parental cell state, and treated this as the weight of the corresponding edge. The maximum edge weights greater than 0.2 were retained, resulting in acyclic, directed graphs, independently

generated from scRNA and scATAC data (**Fig. 2.3C-D**). Although these procedures were generated independently of our cell cluster annotations at each time window, they overwhelmingly yielded internally consistent results. For example, muscle clusters in one time window connect to muscle clusters in the next time window, and the same is true for other major lineages (*e.g.* CNS, PNS, etc.), as embryogenesis proceeds. We note that some paths seem to terminate prematurely, potentially due to drastic increases in cell number in later embryogenesis which were not matched by corresponding increases in our sampling or due to unknown technical or biological factors. More generally, because these are inferences based on cellular state rather than lineage tracing, they may be prone to certain kinds of error [37].

To illustrate the potential of these data to facilitate exploration of specific lineages at finer resolution, we reanalyzed 59,012 cells annotated as neuroectoderm using scRNA data from 6-18 hrs (**Fig. 2.3E; Fig. 2.12A**). This revealed 20 subclusters, including a large group of early cells corresponding to the brain primordium and neural progenitors that express regulators of neurogenesis such as *Notch (N)* and *Delta (Dl)*, and neuroblast temporal TFs such as *miranda (mira)* and *castor (cas)*. Two additional neural progenitor clusters correspond to sensory progenitors, while immature neurons express low levels of both neural progenitor and pan-synaptic genes, including *cacophony (cac)* and *synaptotagmin 1 (syt1)*. Mature neurons are marked by higher levels of pan- and subtype specific synaptic genes coupled with low/no expression of earlier developmental genes. Finally, midline cells, consisting of both neurons and glia cluster together, become evident at 6-8 hours; using the midline TF *single minded (sim)* and glial immunoglobulin family member *wrapper* as markers, we can follow them forward in time as they mature (**Fig. 2.12B**). We can also follow the maturation of sensory neural progenitors, marked by *shaven (sv)*, from 6 to 16 hrs (**Fig. 2.12B**).

To further explore neuronal diversity, we reclustered 6,703 mature neurons, revealing 11 neuronal subtypes which we manually curated (**Fig. 2.3F**). Among these, we identify four clearly separable sensory cell clusters. There are two types of *Drosophila* sensory neurons

based on dendritic morphology: Type I sensilla, which include both external sensory (ES) neurons and internal chordotonal (Ch) neurons, and Type II multidendritic (MD) neurons. We can clearly distinguish MD neurons based on expression of genes such as dendritic arbor reduction 1 (*dar1*), which promotes their characteristic branching dendrites, and the pseudouridine synthase RluA-1, which was recently identified as a marker of multidendritic neurons [54] (Fig. 2.3F-G). Consistent with their nociceptive role, this cluster also uniquely expresses the mechanical nociception degenerin/epithelial sodium channel subunits *picklepocket* (*ppk*) and *ppk26*. Mechanosensory ES neurons are specified by the TF *hamlet* (*ham*), which is uniquely expressed in the middle sensory cluster (**Fig. 2.3F-G**) [55]. Ch sensory neurons detect stretch and vibration, and are identified by expression of the mechanosensitive non-selective cation channel subunit *no mechanoreceptor potential C* (*nompC*) as well as fate-determinant *Rfx* and a number of as-yet uncharacterized genes unique to this cluster [56, 57] (Fig. 2.3F-G). The final sensory cluster likely corresponds to chordotonal glia based on the expression of glial markers, including *moody*, and *Cbl-associated protein* (*CAP*) and *nompA*, which promote the development and function of chordotonal glia, respectively (**Fig. 2.3F-G**). Based on vesicular neurotransmitter transporter expression, we also identify two clusters of central cholinergic neurons, a glutamatergic cluster that likely includes motor neurons, and monoaminergic neurons (**Fig. 2.3F-G**). Finally, peptidergic neurons cluster separately and were identified based on the expression of neuropeptides (*Ion transport peptide* (*ITP*)), enzymes involved in their synthesis (*amontillado* (*amon*)) and receptors (*Myosuppressin receptor 1* (*MsR1*)) (**Fig. 2.3F-G**).

We validated the expression of two genes, *complexin* (*cpx*) and *CG4328*, identified in our analysis as enriched in the monoaminergic cluster, which includes midline neurons, and an uncharacterized lncRNA *CR31451* as enriched in mature neurons (**Fig. 2.3H**). This neuronal subtype enrichment is unexpected for *cpx*, which encodes a presynaptic regulator of synaptic vesicle release, and may point to additional requirements for Complexin in midline monoaminergic neurons. In the course of exploring these fine neuronal subtypes, we also

made an unexpected finding regarding *elav*, a ‘classic’ marker gene for neurons. Specifically, we noticed lower level expression of *Elav* in clusters annotated as visceral muscle. Performing double fluorescent *in-situ* hybridization with a visceral muscle specific marker gene (*binou*) confirmed this surprising finding (**Fig. 2.12C**) and suggests a potential new role of this well-studied gene.

This deeper exploration of neuroectoderm, validating and extending years of research from many groups, illustrates the depth of information that can be obtained from these data. We additionally performed a more detailed annotation of non-myogenic mesoderm (**Supplementary Note 1**). A full exploration of all lineages represented in these data will require a community-wide effort, with tissue experts (as done here for neuronal diversity).

In addition to delineating developmental trajectories, these data can also capture spatial differences arising during developmental patterning. Previous bulk ATAC-seq on embryo halves showed variability in the accessibility of enhancers along the anterior-posterior (A-P) axis of the blastoderm embryo [58]. Using label transfer to map anterior or posterior identities from a previous blastoderm dataset [1] onto our 2-4 hr data, we computed a positional accessibility skew score for validated enhancers with strict A-P activity [58]. This indicates that accessibility of most A-P enhancers is skewed in the expected anterior or posterior cell group (**Fig. 2.12D**), recapitulating the bulk data [58]. Interestingly, we also identify differences among enhancers of the same gene. For example, in the *eve* locus, the stripe 1 enhancer has a much stronger skew for anterior accessibility compared to stripe 2, as also reported by [58]. Our single-cell data thus captures the biological variability in enhancer accessibility along the A-P axis, extending previous observations. We similarly could transfer labels from our sci-RNA-seq clusters to spatial coordinates from a Stereo-seq-based spatial study of *Drosophila* embryos at 14-16 hrs and 16-18 hrs of development [59]. Using the assigned annotations of tissues from the spatial study, we observe a correspondence with our cluster annotations, which again suggests the spatial-relevant variability present in these

data (**Fig. 2.12E**).

2.2.3 Tracing dynamic gene modules across development

To further leverage continuous views of unfolding trajectories, we next explored the gene regulatory modules active in germ layer-specific development. We focused on the mesoderm and its derivatives as a complex, well-characterized system that we and others have studied previously [43, 44, 60, 61]. For this, we selected all cells corresponding to mesoderm-derived cell states, collectively 51,672 (scRNA) and (scATAC) profiles across 4-20 hrs and 2-20 hrs of inferred developmental age, respectively (**Fig. 2.4A-B**).

Focusing first on RNA, we selected the top 2,000 most variable genes. After normalizing expression values to be comparable across time, we used dynamic time warp clustering to group genes into 4 clusters with distinct temporal regulation (**Fig. 2.4C**; **Fig. 2.13A**; **Table S6**). These clusters define broad successive waves of gene expression during mesoderm development (**Fig. 2.4D**), and notably exhibit similarly ordered waves of chromatin accessibility (**Fig. 2.13B,D**; **Supplementary Note 2**). Gene pathway enrichment suggests different functional roles for each cluster (**Fig. 2.13C**). Cluster 1 genes ($n = 571$) are highly expressed from the beginning of mesoderm development (directly after gastrulation; 4-9 hrs), are enriched for TFs ($p = 1.4 \times 10^{-6}$) and likely represent a mixture of genes involved in progenitor cells, mesoderm development and transcriptional activation (**Fig. 2.4D**; **Fig. 2.13C**). Cluster 2 genes ($n = 433$) peak around 9-11 hrs, during the subdivision of the mesoderm into different muscle primordia and their subsequent specification; this cluster is enriched for genes involved in mesoderm development, including myoblast fusion and myotube differentiation, while losing enrichment for stem cell/self-renewal terms (**Fig. 2.4D**; **Fig. 2.13C**). In contrast, cluster 3 genes ($n = 365$) initiate expression at around 10 hrs and steadily increase to the end of embryogenesis, while cluster 4 genes ($n = 631$) only switch on around 15 hrs, during muscle terminal differentiation. The last cluster

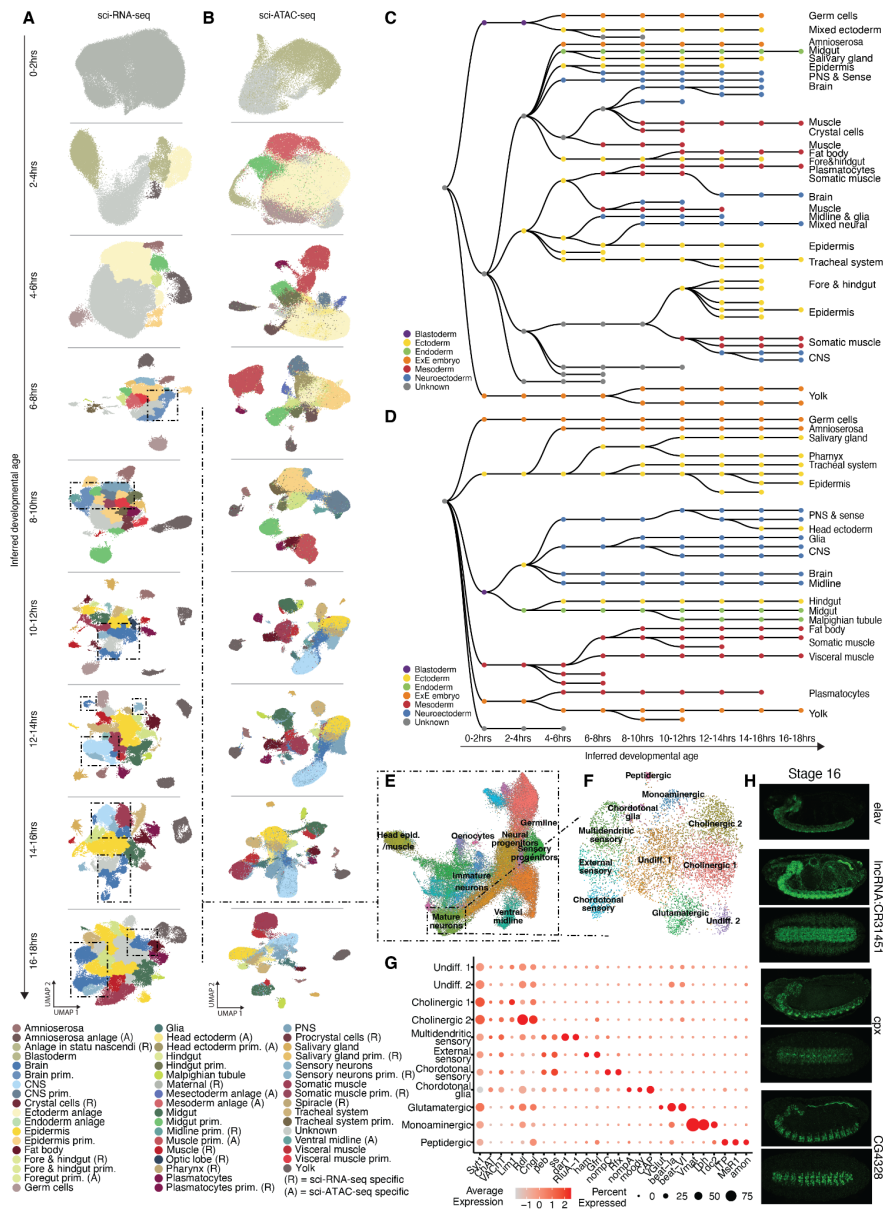


Figure 2.3: Figure 3: Annotation of diversifying developmental trajectories. (A) UMAP visualization of non-overlapping, inferred 2-hr time-windows for scRNA clusters colored by cell state annotation. Dashed boxes highlight neuroectodermal clusters. (B) Same as panel A, but for scATAC data. (C) ScRNA-based acyclic directed graph representation of clusters linked through non-overlapping time windows. (D) Same as panel C, but from scATAC data. (E) UMAP of scRNA data for ~60,000 annotated neuroectodermal cells, *i.e.* cell states highlighted in panel A with dashed boxes, colored by cluster. (F) UMAP of ~6,000 mature neurons, colored by cluster. (G) Dot plot showing marker gene expression for annotated clusters in panel F. (H) *In-situ* hybridization of stage 16 embryos, showing the expression of lncRNA *CR31451*, *cpx* and *CG4328* in the nervous system. A tissue marker (*elav*) is provided in the top panel. A lateral and ventral embryo view is shown for each gene.

lacks enrichment for TFs and rather includes genes involved in myofibril assembly and muscle assembly/maintenance, as well as essential contractile proteins for differentiated muscle (**Fig. 2.4D**; **Fig. 2.13C**). We validated the spatiotemporal expression of five poorly characterized genes by *in-situ* hybridization, confirming that they are indeed expressed in the mesoderm/muscle at the inferred time window (**Fig. 2.4E**).

The temporal and cell-type specific nature of these expression signatures for both the downstream effector molecules and their upstream regulators should provide the resolution to order genes into putative regulatory hierarchies. For example, several genes with essential roles in muscle differentiation, such as myosin heavy chain (*Mhc*), are present in clusters 3 and 4. *Mhc* protein plays a critical role in providing muscle-contractile force. Our scRNA data shows increasing *Mhc* expression along the muscle lineages in cells with later embryonic ages (**Fig. 2.4A,F**), matching the expression pattern of *Mhc*. Concomitantly, there is a gradual increase in open chromatin at characterized *Mhc* enhancers at later stages along multiple muscle trajectories (**Fig. 2.4G**).

Prior to the expression of *Mhc* and other muscle differentiation genes, we observe transient expression of mesoderm-associated TFs (cluster 2; **Fig. 2.4C**). One example is *Kahuli* (*Kah*), a TF associated with muscle development, which has peak expression at 10 hrs of embryogenesis (cluster 2; **Fig. 2.4C, D, F, G**). To investigate the relationship between open chromatin and gene expression, we computed ‘gene activity scores’, defined as the sum of sci-ATAC-seq reads in the gene body and 2 kb flanking the TSS. The gene activity scores for both *Mhc* and *Kah* recapitulate their sequential temporal patterns of expression, with *Kah*’s activity signature appearing earlier along the mesodermal trajectories compared to *Mhc* (**Fig. 2.4F-G**). To determine the extent to which we could map the exact ordering of accessibility and expression changes, we overlaid the scaled expression values and gene activity scores averaged across bins with equal numbers of cells (**Fig. 2.4G**). Intriguingly, for *Kah*, gene expression temporally follows the trajectory of the corresponding gene activity

score based on open chromatin, suggesting an ordering where first the gene body becomes accessible, followed by accumulating levels of the corresponding transcript; however, this was not the case for *Mhc*, for which expression and accessibility increased in tandem (**Fig. 2.4G**). Of note, Kah binds to several characterized *Mhc* enhancers near the gene’s promoter, as observed in bulk ChIP-seq data [45], suggesting a regulatory link between Kah and *Mhc* expression (**Fig. 2.4H**).

To extend this analysis more globally, we searched for TF motifs enriched in putative enhancers (mesoderm-specific scATAC peaks in the vicinity (1-10 kb upstream of TSS) of genes belonging to each of the four scRNA mesoderm expression clusters. This identified 458 TF motif-to-cluster enrichments (q -value $< 1 \times 10^{-3}$ and presence in $>1\%$ of target peaks; **Table S7**), corresponding with 152 unique TFs. Of these, 31 are TFs whose expression changes along mesoderm differentiation and are thus included in the expression-based clustering (**Table S7**). These 31 include many TFs essential for mesoderm development, including a number of direct target genes of the master regulator Twist (the functional ortholog of MyoD) at the beginning of mesoderm development (*e.g.* *hb*, *en*, *Ubx*, *pb*) and concordantly are expressed in the first temporal cluster. These factors have many functions, including setting up the segmentation of the mesoderm, regulating the expression of somatic muscle identity genes, establishing midgut constrictions in the visceral mesoderm and heart patterning. Other examples from the second and third temporal clusters are genes required for cell fate specification of somatic muscle founder cells (*Six4*, *ap*) and heart development (*tup*, *Lim3*).

We note that this approach may miss the contribution of important TFs that were not variably expressed in mesoderm. In particular, if a TF is variably expressed and has corresponding variability in motif activity, this TF is likely active. However, this does not imply that all expressed TFs are active (*e.g.* there may be co-activators or post-translational modifications that are required). This caveat notwithstanding, these analyses highlight the

potential for further discovery of co-regulated gene modules related to distinct germ layers or cell types.

2.2.4 Nominating stage- and cell type-specific TF regulators

We next asked if we could leverage the diversity of cell states across embryogenesis to infer which TFs drive specific programs of cell type differentiation. For this, we used all scATAC clusters at all timepoints (in contrast to the scRNA-focused cluster analysis above) and searched for differential enrichment of TF position weight matrices (PWMs) within each cluster’s open chromatin regions.

We first characterized enrichments across clusters from the 10-12 hr time window based on predicted time (**Fig. 2.5A**). Encouragingly, hierarchical clustering of the enrichment profiles of all associated PWMs grouped each cluster roughly by germ layer (this was also observed in other time windows; **Fig. 2.14A**). The non-myogenic mesoderm (fat body) and myogenic mesoderm (somatic muscle) cluster together (**Fig. 2.5A**). Open chromatin regions in the myogenic clusters are enriched in motifs for many TFs known to play a role in muscle development, including Mef2 and fork head (fkh) TFs. The myogenic clusters also appear close to two neuronal clusters (**Fig. 2.5A**), which is driven by shared motif enrichment with neuroectoderm and glial cells, in particular many C2H2 zinc finger TFs, including Btd, CG7368, Crol, Sr, and Dar1. Many of these factors have known roles in neuronal development (*e.g.* Dar1), while Stripe (Sr) is essential for muscle tendon cell fate and muscle attachment in the epidermis at late stages of embryogenesis [62].

As members of the same family of TFs typically recognize similar motif sequences (*e.g.* GATAe/ GATAd/pnr), it is often difficult from motif analysis alone to pinpoint the responsible TF. To address this, we leveraged our scRNA data to identify the most likely active TF, based on its expression within the clusters among all factors that share the same motif

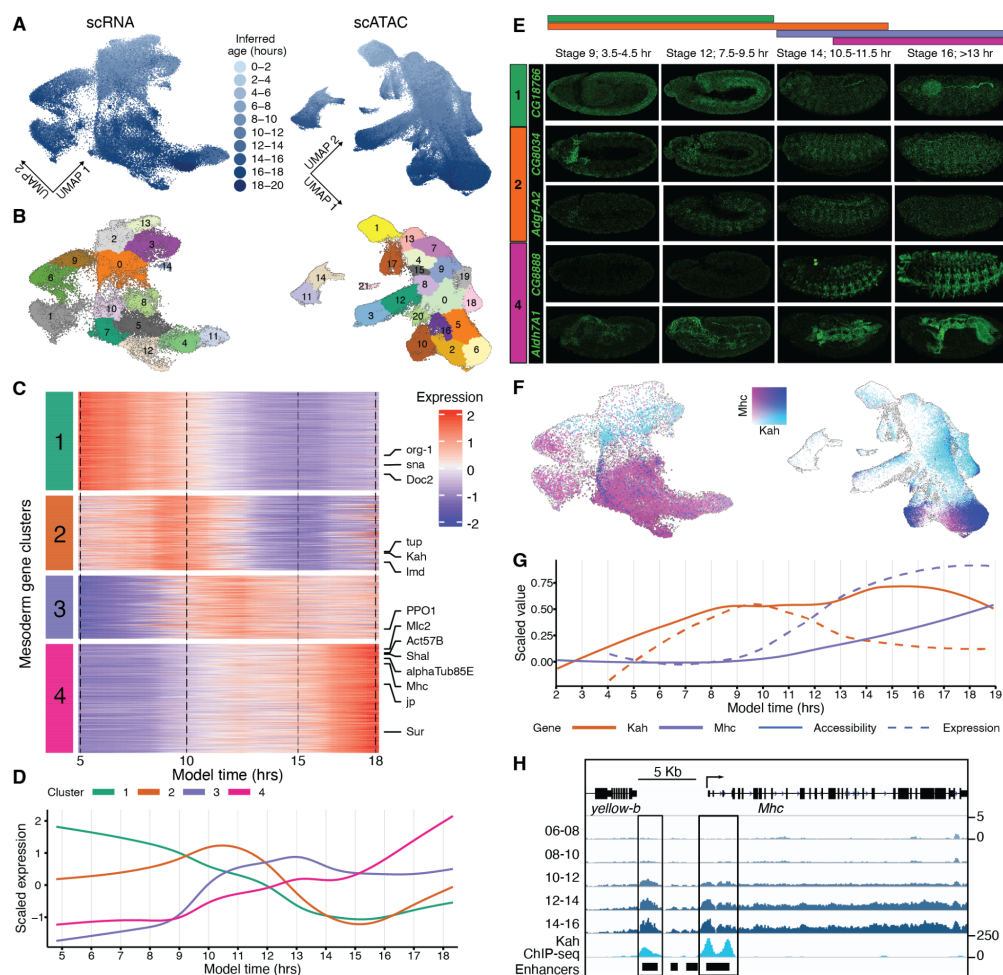


Figure 2.4: **Figure 4: Dynamic regulation of mesoderm-specific gene modules.** (A) UMAP of scRNA (left) or scATAC (right) data for all mesodermal cells, colored by inferred developmental age. (B) Same as panel A, but colored as reprocessed leiden-based clusters. (C) Normalized expression of mesoderm genes across inferred developmental time. (D) Average expression of the gene modules across inferred time. (E) *In-situ* hybridization experiments validating temporal expression of selected genes with predicted expression in mesoderm/muscle. (F) Same as panel A, but expression of *Kah* (cyan) and *Mhc* (purple) is overlaid. Points from cells that express both *Kah* and *Mhc* are colored dark blue. (G) Comparison of gene activity score (solid line) and gene expression (dashed line) over the continuum of inferred developmental age for *Kah* (cluster 2) and *Mhc* (cluster 3) in mesoderm-annotated cells. Gene activity scores and expression were binned into 100 equal partitions by inferred age, averaged, and scaled to 0-1 with min-max values. (H) Chromatin accessibility profile surrounding *Mhc* for pseudobulk mesoderm cells from 6-16 hrs inferred time in 2-hr increments, along with *Kah* ChIP-seq generated from 0-16 hr whole embryos [45]

binding pattern. First, we used a regression-based framework to integrate the scATAC and scRNA datasets and identify links between the different cell clusters [16, 2]. Specifically, we adopted a non-negative least square (NNLS) matrix factorization approach that decomposes expression data as a mixture of components derived from proximal gene activity scores generated from the scATAC data. Despite possible temporal differences between accessibility and expression, NNLS identifies stronger links between clusters from the same 2-hr window compared to adjacent 2-hr windows (**Fig. 2.14B**). We also inferred NNLS links in the opposite direction by decomposing proximal gene activity scores by gene expression associated with scRNA clusters. For each cluster of a given data type, the result of NNLS factorization is a mixture proportion of clusters from the other data type, where a higher value represents a stronger association between the scRNA and scATAC cluster (**Fig. 2.14C-F; Table S8**). This factor decomposition approach resulted in a strong linkage (NNLS-mixture coefficient of > 0.1) of 120 cell state clusters present in the same inferred time windows, with most of the strongly linked clusters being from 4-6 hrs onward. Upon manual inspection, many linked scATAC and scRNA clusters, which had been independently annotated, are from matching tissues. For example, from the 10-12 hr window, the epidermis cluster (cluster 0) in scATAC data was matched to the epidermis (cluster 3) in scRNA data. Altogether, of 21 ATAC clusters from the 10-12 hr window, 16 had a linked RNA annotation with a NNLS correlation value >0.1 , of which 14 were between comparable tissue annotations.

These integrated scRNA and scATAC clusters, which span 0-18 hrs of embryogenesis, enabled a more direct analysis of the role of specific TFs in different cell type's differentiation. We reasoned that active TFs should be more highly expressed in cell types for which they have a functional role, and their associated PWM should be more enriched or depleted in accessible regions when the TF is activating or repressing expression [16]. In line with this, correlation values between motif-associated accessibility and gene expression were shifted towards more positive values for TFs annotated (by GO) as activators, and more negative values for annotated repressors (**Fig. 2.5B; Table S9**), a trend also observed in human fetal

tissues [16]. This approach of linking TFs' cluster-specific expression and motif enrichments allowed us to nominate TFs as active at specific times in specific tissues (**Fig. 2.5C**). For example, this analysis predicts a specific role for Sage in salivary gland development, as the salivary gland is the only cell type exhibiting both high expression of the *sage* transcript and high accessibility of the Sage-associated PWM (**Fig. 2.5C**, top). This finding matches the essential role for *sage* in salivary gland development as determined by genetic loss-of-function analysis [63]. Similar predictions were made for GATAe in the midgut at 16-18 hours, and Awh in the epidermis at 14-16 hours (**Fig. 2.5C**, middle and bottom), matching the functional role for both TFs in midgut endoderm [64] and epidermis [65, 66] development, respectively.

To expand this analysis and systematically nominate TFs that potentially drive germ layer-specific differentiation programs, we fit a linear model that predicts a TF's motif associated chromatin changes from an estimated effect of an interaction term that includes the expression level of the TF in a specific germ layer and time window. Our model's effect estimates can identify TFs with specific motif activity in particular germ layers and suggest time windows from which a TF initiates its activity. For example, the model refined the role of Sage as becoming active in the ectoderm germ layer specifically from 10-12 hrs onward, and the activity of GATAe initiating in the endoderm from 08-10 hrs onward (**Fig. 2.5D**, top). Such a model encompassing germ layers across development time may also identify additional likely co-active TFs. For example, in addition to Sage, we found Fork head (Fkh) to be both co-expressed and co-active in the ectoderm, a TF reported to act together with Sage to activate salivary gland specific genes [67].

This analysis also generated additional interesting findings for other timepoints and germ layers (*e.g.* Fruitless (*fru*); **Supplementary Note 3; Fig. 2.5D**). Altogether, from eight high-level germ-layer associated tissue annotations and 316 TF motifs tested, we identified 1258 significant (BH-corrected p -value $< 1 \times 10^{-3}$) TF-to-tissue relationships, having both

associated expression and chromatin activity at one or more of the nine time windows assessed. We note that in time windows with fewer clusters, the association effect estimates are susceptible to outliers and should be interpreted with caution. Notwithstanding this caveat, these putative assignments represent an extensive resource for future studies (**Table S10**).

To demonstrate the potential of our approach to discover new putative roles for TFs, we selected four genes and validated whether they were expressed in the linked germ layer by fluorescent *in-situ* hybridization. While these genes were inferred to have effects in multiple germ layers, their function in either mesoderm (*CG5953*, *CG11617*) or neuroectodermal tissues (*Ets65A*, *CG12605*) was poorly characterized. We confirmed that these factors are in fact expressed in the tissue and time window predicted by our data (**Fig. 2.15**), suggesting potential roles for these TFs in mesoderm/neuronal development.

To complement the NNLS, we applied a recently developed tool, FigR [68] to further facilitate gene regulatory network (GRN) reconstruction. Because multi-omic ATAC-RNA data from the same cell is required for this task, we first integrated our two independent assays for all cells from 10-12 hours using canonical correlation analysis (CCA), identifying the most likely ATAC-RNA cell pairs using geodesic distance-based pairing [68] within the common CCA space. Using these pairs as input for GRN inference with FigR, we linked ATAC peaks to their target genes, based on peak-to-TSS accessibility correlation and then computed TF motif enrichments for the linked regions, which, together with the TF expression-accessibility correlation, allowed us to define hundreds of putative activators and repressors at this embryonic stage (**Fig. 2.16A**). Ranking the TFs by their regulation score (**Fig. 2.16B**) nominated many activators and repressors that we also identified in the NNLS analysis above, including *l(3)neo38*, *Lim3*, *lola*, *fkh*, and *fru* (**Fig. 2.5D**). Focusing on the targets of the regulatory networks across all cells at 10-12 hour, we found a large set of genes that appear to be extensively regulated (209 genes with >10 linked regulatory regions) (**Fig. 2.16C**). We then used the inferred TF activities to explore the factors acting on

these genes and their mode of regulation. For example, *tup*, a TF gene required for heart development, undergoes extensive self-regulation (highest motif-RNA correlation), besides being positively regulated by the pan-muscle TF Mef2 and repressed by Run and Opa (**Fig. 2.16D**). Another top ranking gene, *chinmo*, an essential neuronal TF, is activated by other nervous system TFs, such as Lim1 and Onecut, and is negatively regulated by Fru (**Fig. 2.16E**), which we also identified as a neuroectoderm-specific repressor in our NNLS-based analysis (**Fig. 2.5D; Supplementary Note 3**).

Finally, we sought to exploit the fine-grained resolution of inferred nuclear ages to explore the dynamics of an early pioneer TF, Zelda, in regulating chromatin opening followed by transcription during ZGA. We recovered the expression of a set of genes that are Zelda-dependent during ZGA [49], and for each gene aggregated accessibility at the linked Zelda-bound regions [69]) in intervals of one minute, across 0-3 hrs of embryogenesis (**Fig. 2.5E**). Clustering of gene expression identified two broad temporal clusters, a first group of early genes and a second group whose expression increases later, roughly after 1.5 hours of embryogenesis. Interestingly, while accessibility at the Zelda-bound regions linked to the early cluster seems to mirror the temporal expression, regions linked to the late expression gene cluster gain accessibility much earlier, almost as early as the first cluster, suggesting that Zelda is opening these regions for future activation (**Fig. 2.5F**). To verify if accessibility is indeed reflective of Zelda binding, we retrieved Zelda occupancy by nuclear cycle [69], which confirmed that more than 70% of regions in both temporal clusters are already occupied by Zelda at nuclear cycle 8-9, regardless of the associated gene expression (**Fig. 2.5G**). These results suggest that Zelda establishes chromatin accessibility at a large set of regulatory regions in the early embryo, independently of future gene expression, in agreement with its well known role as a pioneer factor. In some cases, Zelda possibly also functions as the activator of gene expression (cluster 1), while in others it retains a pioneering role and the gene's expression is induced by later TFs (cluster 2).

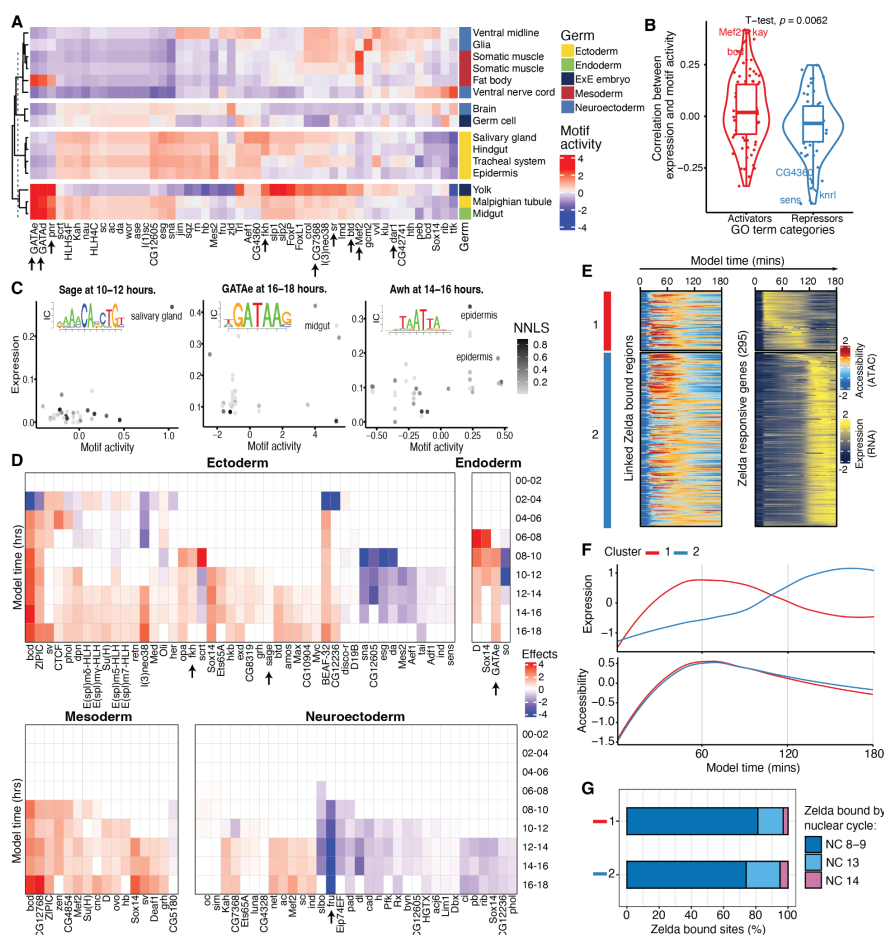


Figure 2.5: **Figure 5: Integration of scRNA and scATAC data to identify TFs with potential regulatory roles across differentiating tissues and developmental time.** (A) Heatmap with averaged chromatin accessibility differences associated with the 50 most variable TF-specific motifs from all cells in annotated ATAC-seq clusters from 10-12 hrs. Arrows indicate TFs discussed in the main text. (B) Correlation between expression and motif-associated accessibility grouped by expression activation or repression-associated GO categories. (C) Comparison of gene expression (y-axis) and motif-associated chromatin accessibility (x-axis) across NNLS-linked clusters for the TFs Sage (left), GATAe (middle), and Awh (right). Each TF's corresponding PWM is inset in each plot. (D) Heatmap of estimated effects of gene expression at predicting motif-associated chromatin accessibility changes through time in different germ layers. Displayed TFs had 3+ consecutive time windows with a significant sign-consistent effect. Arrows indicate TFs discussed in the main text. (E) Heatmap of expression at Zelda-responsive genes (right) and aggregated chromatin accessibility (left) at their Zelda-bound *cis*-regulatory regions. Values were averaged in 1 min windows. The colored bars at left indicate two temporal clusters of expression of Zelda-responsive genes. (F) Smoothed average expression and accessibility for the two Zelda gene-regions temporal clusters from panel E. (G) Proportion of accessible regions from panel E that are bound by Zelda in cluster 1 and 2 in ChIP-seq data from different nuclear cycles (NC).

2.3 Discussion

This continuum of *Drosophila* embryogenesis builds on our previous work generating sci-ATAC-seq from three non-overlapping time windows of embryogenesis [1] and complements other studies performed on specific tissues [70, 71, 72, 73, 74, 75] as well as scRNA from entire embryos at one specific stage [39] or on dissected tissues from adults [76]. Despite the growing use of single-cell assays to generate large-scale atlases, characterizing fine-scale dynamics of chromatin accessibility and gene expression across developmental time remains a challenge. The large number of cell types, and even greater number of cell states and branch points during embryogenesis requires extensive cell sampling at continuous stages to capture regulatory transitions, especially for rare cell types. This is very difficult, if not essentially impossible, to obtain in most model organisms.

Here, sampling embryo collections from overlapping 2-4 hr time windows, coupled with neural network-based inference of more precise nuclear ages, enabled continuous representation of *Drosophila* embryonic development. Other studies have attempted a similar ordering of embryos by developmental time over a two-day window of mouse development [21]). However, because only dozens rather than thousands of mouse embryos can practically be sampled, reliable inference at the scale of hours or minutes is challenging. Similarly, cell age was inferred in *C. elegans* using an independent time-series of bulk RNA-seq from whole-embryos [17]). But, relying on such whole embryo bulk data to predict developmental age in single cells risks inaccurate aging of rare or transient cell types, especially for more complex organisms.

Computationally, our neural network-based inference of developmental age bears some similarity to the concept of pseudotime [50]. As originally proposed, pseudotime aims to serve as “a quantitative measure of progress through a biological process”. Analogously, our inferred developmental age tracks the progression of nuclei through development. However,

the advantage of pairing an experimental design including overlapping yet tightly defined time windows with temporal ordering is that we can anchor inferred ages to fixed time-points, which can potentially lead to a more accurate representation of developmental age for complex cellular trajectories. Put another way, inferred ages are interpretable as units of absolute time that are synchronized across all tissue trajectories. With such a continuum of cellular states we can begin to infer cell type trajectories that more closely capture the continuous processes of cellular differentiation unfolding within a complex, developing multicellular organism.

There remain further possible improvements to our experimental framework. The alignment or anchoring to real time could be refined with sampling of more tightly staged windows. Multi-omic methods for characterizing multiple data types from the same nuclei may facilitate a joint model that can link paired gene expression and chromatin accessibility (and other modalities) to developmental age inference. There are cases where technical features of the data can lead to increased uncertainty of model predictions. For example, we found that cells annotated as germ cells, or from the first collection time window, or with low read count, were associated with greater prediction error (**Fig. 2.16F**). Moving forward, we suggest caution for interpreting findings solely based on inferred nuclear ages from clusters with these features.

The extensive scATAC data, with deep coverage across almost a million cells, likely captured the vast majority of regulatory elements active during embryonic development and provides a comprehensive resource of potential enhancers for almost any cell type in the embryo. In contrast, our scRNA data had relatively low unique reads per cell, and will likely miss some differentially expressed genes in specific cell types. As a result, some delicate analyses remain challenging. For example, we found transcriptional velocity estimates to be unstable with sparse scRNA data, although this issue was mitigated by constructing ‘metacells’ prior to velocity analysis (**Fig. 2.16G**), which may be useful for pursuing targeted

questions. In scATAC data, we were able to distinguish XX vs XY nuclei from the proportion of chrX-mapped reads (**Fig. 2.16H**); however, this was challenging for the scRNA data, again due to data sparsity. These shortcomings are to some degree compensated by the large number of cells profiled, as shown by our ability to recapitulate aspects of previously documented heterogeneity even for highly dynamic or restricted phenomena, *e.g.* zygotic gene activation (**Fig. 2.2E**).

Overall, this *Drosophila* embryonic atlas provides broad insights into the orchestration of cellular states during the most dynamic stages in the life-cycle of the organism. Our results represent a rich resource for understanding precise timepoints at which genes become active in distinct tissues as well as how chromatin is remodeled across time. The annotation of cell types within these data is an ongoing process and one that is much more challenging at early and mid-stages of embryogenesis as compared to late time-points or adults with differentiated tissues. A comprehensive annotation of embryonic cell states will require a collective effort from the *Drosophila* community. To support these ongoing efforts, we provide information on expression and peaks from all clusters (from **Fig. 2.3A-D**), in addition to all intermediate and raw data for further exploration (**Data availability**). Although larval stages remain insufficiently profiled, we hope that these data and methods, together with the recently released large-scale adult atlas [76], brings us closer to the community-wide goal of a multimodal *Drosophila* atlas spanning a continuum from zygote to adulthood.

2.4 Materials and methods

A detailed version of materials and methods is provided in the Supplementary Information. In brief, *Drosophila melanogaster* embryos were acquired for each of 11 collection windows and then each pool of embryos were divided, with each half being extracted and fixed for either sci-RNA-seq3 or sci-ATAC-seq3. Libraries were sequenced deeply, and the resulting reads mapped to dm6 and then processed with a uniform processing pipeline that included

QC filters for low read depth or high proportions of reads mapping to the mitochondria or ribosomal genes, and extensive doublet removal. Between the two data modalities, we obtained profiles for ~ 1.5 M nuclei, although unique read depth per nucleus was considerably lower for scRNA than scATAC data.

Using the center hour of the collection window, we used several machine learning approaches to fit a model that could infer the age of a nucleus with either gene expression or chromatin accessibility information. Both lasso linear regression and neural networks were fitted to the same training data, with a held-out subset used for model validation and comparison. Given its consistently superior performance, we then relied on specific parameterizations of neural network model-inferred ages to reposition nuclei in time. To zoom into fine scale timepoints, we binned data by small increments to explore the regulatory dynamics of zygotic gene activation. Then using 2 hr adjacent windows of cells, we computed clusters of similar cells and performed extensive manual review to annotate each cluster’s likely germ layer and cell type. We then used an iterative approach for constructing an acyclic tree of differentiation by identifying the likely precursor cluster for each cluster in a given time window.

Neuroectoderm was iteratively analyzed for deeper annotation of neuronal subtypes, while mesoderm was picked for analyses focused on identifying co-regulated genes and accessible regions, which were then subjected to ontology and TF motif enrichment analysis. To connect scATAC cell clusters with scRNA cell clusters we used a regression-based approach (NNLS). Such connections between ATAC and RNA clusters enabled a series of analyses such as correlating expression with motif accessibility, applying gene regulatory network analysis pipelines, etc.

Several additional analyses were performed. We used probabilistic label transfer to map likely cluster annotations from these data to spatial information from patterned DNA

nanoballs. We also found it is possible to infer the sex of cells from the proportion of chrX-mapped scATAC reads using a gaussian mixture model to classify cells. Although RNA velocity was challenging to apply to sparse scRNA data, it yielded more sensible results when subsets of cells were first aggregated to ‘metacells’.

The expression of several genes were verified by fluorescent *in situ* hybridization: specific neuronal genes active in identified clusters, unexpected co-activity of the *elav* with *binou*, genes active at specific mesoderm timepoints, and putative active TFs with less characterized roles in tissue development.

Raw data is available through the GEO, while additional scripts and intermediate files, including bigwigs and a custom R shiny app to visualize UMAPs, are available through our data sharing website.

Drosophila embryo collections

Drosophila melanogaster wild-type embryos from Canton S flies were collected as previously described [77, 1]. Briefly, embryos were collected on apple-agar plates in two-hour or four-hour windows, following three one-hour pre-lays to clear the females and synchronize the collections. After aging the embryos at 25°C to the specified time window (**Fig. 2.1A**), the embryos were collected from the plates, cleaned with water and dechorionated in 50% bleach for 2 min. Collected embryos for sci-ATAC-seq were fixed immediately, while embryos for sci-RNA-seq were flash-frozen and stored at -80°C until nuclei isolation. Embryo illustrations adapted from [78].

Nuclei isolation and fixation for sci-RNA-seq

Embryos were dissociated in an ice-cold hypotonic lysis buffer (20mM Hepes pH 7.5, 10mM KCl, 1.5mM MgCl₂, 1% Superase IN RNase inhibitor) with a Dounce homogenizer, then incubated on ice for 3 minutes, and filtered through a 40uM cell strainer into a tube that

contains 2ml of 1.5 mg/ml DSP or DSS cross linker (Thermo Fisher Scientific). Homogenizer and pestle were washed with an additional 0.5 ml hypotonic buffer, to increase nuclei yield, and the buffer filtered through the cell strainer and added to the cross-linking reaction. Nuclei were cross-linked on ice for 30 minutes, with gentle mixing every 5 minutes. After 30 minutes, 9 ml ice-cold methanol was added (to a final concentration of 80% methanol), and nuclei were further incubated on ice for 15 minutes. Nuclei were rehydrated with 15 ml nuclei purification buffer (NPB) (10mM Hepes pH 7.5, 90mM KCl, 40mM NaCl, 200 ng/ml BSA, 1% SuperaseIN), by adding the buffer drop-wise while gently mixing. Nuclei were collected by centrifugation, and separated from debris by sucrose cushion centrifugation (0.3M sucrose-PBS-triton over 1.4M sucrose-PBS-triton 2000 x g 20 min 4°C). Isolated nuclei were resuspended in 1ml of NPB + 50% glycerol, flash frozen and kept at -80°C until use.

Embryo fixation and nuclei isolation for sci-ATAC-seq

Embryos for sci-ATAC-seq were formaldehyde fixed in cross-linking solution (50mM Hepes, 1mM EDTA, 0.5mM EGTA, 100mM NaCl, pH 8, 1.8% formaldehyde v/v) with a heptane layer for 15 min with shaking at room temperature. Formaldehyde was quenched with PBS/glycine/triton solution (125 mM glycine + 0.1% Triton in PBS), after which the embryos were washed once in ice-cold PBS, dried and stored at -80°C in ~1 gram aliquots until further use. Embryo dissociation and nuclei isolation were performed as previously described, using a dounce homogenizer and a 22G needle (steps 1-10) [42]. Nuclei were then pelleted at 2,000g at 4°C, resuspended in nuclear freezing buffer (50mM Tris at pH 8.0, 25% glycerol, 5mM Mg(OAc)₂, 0.1mM EDTA, 5mM DTT, 1× protease inhibitor cocktail (Roche), 1:2,500 superasin (Ambion)) by trituration and flash frozen in liquid nitrogen.

In situ hybridization experiments

Whole-mount *Drosophila* embryo fluorescent *in situ* hybridization (FISH) was performed as previously described [79]. Briefly, overnight embryo collections spanning all stages of embryogenesis were fixed in 4% formaldehyde (Polysciences #18814-10) for 20 minutes, stained

by double fluorescent in situ hybridization, mounted in ProLong Gold Antifade reagent (Thermo Fisher Scientific # P36931) and imaged with a Zeiss LSM780 confocal microscope using a Plan Aplanachromat 20x/0.8 objective. Images were then visualized in Fiji [80].

2.4.1 Sci-RNA-seq methods

Library construction and sequencing

We performed sci-RNA-seq3 as previously described [2]. Several experiments were done and each one included samples from multiple time windows. Associated time-windows for nuclei batches were tracked through specific wells. Additionally, for a subset of the data we included mouse nuclei to serve as a control to determine the rate of cell doublets.

Read processing, nuclei filtering, and doublet removal

The read alignment and gene count matrix generation was performed as previously described [2]. With the single-cell gene count matrix cells with fewer than 250 UMIs, more than 10,000 UMIs, reads mapping to more than 7000 genes, or more than 10% of read counts mapping to ribosomal genes were excluded. Each nuclei was mapped to its original time window from which it was extracted by using the RT barcode.

We performed standard processing of the data split by experiments as recommended by Seurat v3 [33] documentation including `NormalizeData`, `FindVariableFeatures` (with method set to ‘vst’), `ScaleData`, `RunPCA`, `RunUMAP` (with `dims` set to 1:50 and `n.components` to 2), `FindNeighbors` (with `reduction` set to UMAP and `dims` set to 1:2), `FindClusters`. For detection and removal of doublet nuclei we relied on an iterative strategy described in [2]. We used a modified version of `DoubletFinder` [81] that could handle large data sets to document for each cell the proportion of neighbors that were simulated doublets. This version uses the first 30 principal components, along with `pN` of 0.2 and `pK` of 0.005. The number of nuclei

classified as doublets was based on estimates of the count of doublets (barnyard estimate multiplied by recovered nuclei) and were chosen to be the top rank nuclei with the greatest proportion of nuclei that were simulated doublets. Unfortunately, this process alone did not remove all putative doublets that may have been lost among different clusters of cells. Therefore, following clustering of the global dataset, we individually processed each cluster identified and sub-clustered the data. We then eliminated subclusters that were at least 15% classified as doublets. Following removal of all the initial nuclei identified as doublets and subclusters with a large proportion of doublets, we reprocessed the global dataset.

Dimensionality reduction, clustering, and identifying cluster-specific marker genes

The standard seurat processing pipeline, as described in the previous section, was performed on each non-overlapping inferred age window separately. We found that the default Seurat clustering resolution parameter did not capture the dynamics of the presence of different cell types across development. So for each time window we clustered the data with a variety of resolutions (from 0.1 to 1.5 in increments of 0.3) and then computed the within cluster sum of squares (WSS). Visualizing WSS across increasing resolutions we chose the resolution in which there was a visible plateau in the decrease of WSS. The selected set of resolutions are listed in **Table S11**. Finally we clustered the data for each inferred age time window using the per-window choice of resolution parameters. From these clusters we used Seurat’s FindMarkers to iteratively loop through all clusters and identify marker genes (with `only.pos=T`) which were used for cluster annotation. For the set of marker genes used for lineage annotation we excluded genes with ‘`log.FC < 0.25`’ or ‘`min.pct < 25`’.

Annotating cell types and tissues

For cluster annotation, we used the Berkeley Drosophila Genome Project (BDGP) database, which includes gene expression patterns of approximately 8600 genes in drosophila staged embryos as detected by *in situ* hybridization [51, 52]). The BDGP database gives a stage-

specific expression pattern (“term”) for each tested gene during embryogenesis. We used Fisher’s test to look for enrichment of BDGP gene expression terms in each cluster’s marker genes. Top ten terms per cluster were examined. To pick a specific term out of the top ten, we further examined the BDGP terms of the top 20 marker genes for each cluster.

Clustering genes

Inferred-time associated co-regulated genes specific to certain germ layers were determined with an unsupervised clustering approach. First, we subset the full seurat object that included cells from all time windows to only those cells annotated as a specific germ layer. For example, in analyses shown in **Fig. 2.4**, we focused on mesoderm. As described previously in the standard seurat processing pipeline, we scaled gene expression values and subsetted to the 5000 most highly variable genes. We constructed 100 bins of roughly equal numbers of cells across inferred time, and then smoothed expression values by computing the average expression of each gene in each window along with all expression values for cells in time proximal bins from a sliding window across inferred time. Once again we subsetted the genes to the top 2,000 that were highly variable across these smoothed time windows and then these variable expressed genes were scaled and centered to have mean=0 and sd=1. Finally, we performed dynamic time warp clustering with the ‘tsclust’ function from dtwclust v5.5.10 with type=’partitional’, distance=’dtw_basic’, and centroid=’pam’ on these scaled gene values to identify co-regulated genes (Sardá-Espinosa 2019). To choose the number of clusters (k) we performed clustering with all k values from 2 through 35 and then selected the earliest k at which several metrics of clustering metrics began to plateau (**Fig. 2.13**). We downloaded the Kah ChIP-seq data from the ENCODE portal [82] with the following identifier: ENCSR161YRO.

Gene pathway enrichment analysis

We were interested in relating various gene sets to known biological processes. For this task we used FlyEnrichr to perform gene set enrichment analyses [83, 84]. Gene sets were

uploaded to the FlyEnrichr server with the enrichR version 3.0 package for R available on CRAN. We restricted our results to enrichments in the “GO Biological Process 2018” database and the “RNAi Screens from GenomeRNAi 2017” database.

RNA velocity analysis

We randomly sampled 20,000 cells from 3 adjacent time windows (10-12 hr, 12-14 hr, and 14-16 hr) and then randomly grouped cells into 100 meta cell bins and aggregated all reads per meta cell bin. We then used the scVelo software package version 0.2.2 [85]) and the standard processing pipeline to estimate the RNA velocity graph, which we then visualized as a directional vector in the first two principal components space, thus preserving the linear interpretability. The script for performing this analysis is included in our data sharing page.

2.4.2 Sci-ATAC-seq methods

Library construction and sequencing

To create the sci-ATAC-seq3 libraries, we followed the protocol from [16]. As previously described, frozen fixed nuclei were thawed, re-permeabilized in Omni lysis buffer [86], and diluted in ATAC-resuspension buffer (RSB) buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl₂) supplemented with 0.1% Tween-20. We profiled 11 samples corresponding to developmental time windows in 2 experiments and one barnyard sample per experiment. For each time window, 50,000 cells were deposited across wells of a LoBind 96-well plate, 8 time windows per experiment (across 11 wells) in addition to a barnyard sample (across 8 wells) which is a mixture of mouse CH12-LX and human GM12878 cell lines. Re-permeabilized nuclei were tagmented with Tn5 enzyme at 55°C for 30 min then stop reaction buffer (40 mM EDTA with 1 mM Spermidine) were added afterwards and incubated at 37°C for 15 min. Pooled tagmented nuclei from each time window were pooled, pelleted and washed. Phosphorylation master mix [1X polynucleotide kinase (PNK) buffer, 1 mM rATP, T4 PNK]

was added to the washed tagmented nuclei and incubated at 37°C for 30 min. Next, ligation master mix [1X T7 ligase buffer, N5_splint, T7 DNA ligase enzyme] was added directly to the phosphorylation reaction followed by 384 distinct N5_oligos then incubated at 25°C for 1 hour. Stop reaction mixture was added to the ligation reaction and incubated at 37°C for 15 min. All wells were pooled then transferred into a 50-ml falcon tube, pelleted and washed with ATAC-RSB with 0.1% Tween-20. N7 ligation master mix [1X T7 ligase buffer, N7_splint, T7 DNA ligase] were added to the washed pellet and aliquoted into four 96-well LoBind plates. 384 distinct N7_oligos were added across four plates of N7 ligation and incubated at 25°C for 1 hour, then stop reaction mix were added for another 37°C for 15 min incubation. Afterwards, all wells were pooled and transferred to a clean 50-ml falcon tube then washed in ATAC-RSB with 0.1% Tween-20 before resuspending in Qiagen EB buffer. The ligated and washed nuclei were counted and aliquoted at 1000-3000 nuclei per well across four 96-well LoBind plates. Proteinase K and 1% SDS were added to the nuclei to reverse crosslink and incubated at 65°C for 16 hours. To determine the optimal cycle number, a test amplification was performed and monitored with SYBR green on a handful of wells of a reversed crosslink plate. The remaining plates were processed on the basis of the test PCR result. PCR amplifications were performed using NEBNext High Fidelity 2X PCR Master Mix, BSA [bovine serum albumin], indexed P5 oligo, and indexed P7 oligo. All wells were pooled and purified with Zymo Clean & Concentrate-5 and further purified with 1X AMPure bead to get rid of any remaining primers and adapter dimer. Purified libraries were quantified on an Agilent 4200 TapeStation System using D5000 reagents and screentape. Libraries were then diluted to 2 nM for sequencing using a custom recipe and primers on a NextSeq 500 to assess library complexity then further sequenced on an Illumina NovaSeq 6000 sequencer with custom sequencing recipe and primers.

Data processing for sci-ATAC-seq3

The sci-ATAC-seq3 raw reads were processed using the pipeline described in [16]. Reads were mapped to the dm6 reference genome. The non-duplicate fragments are used for peak

calling with MACS2 [87] in each sample, and then merged together with bedtools [88]. We generate sparse matrices counting reads falling into each 5 kb window in the genome for cells passing a sample-specific threshold for each sample. We also generated sparse matrices counting reads falling into the merged peak set and into gene bodies plus 2kb upstream regions (proximal gene activity matrices). The barnyard sample was mapped to the merged hg19-mm9 reference genome to estimate the collision (two cells receiving the same barcode by chance) rate in the experiment.

Dimensionality reduction and clustering

The downstream analysis steps also closely follow the pipeline described in [16]. We binarized the window-by-cell matrices for downstream analysis. We merge the binary matrices for time windows profiled in both experiments. We exclude peaks on sex chromosomes and peaks accessible in less than 1% of cells. We use latent semantic indexing with log-scaled term frequency to normalize the binary matrices. We use singular value decomposition on the normalized matrices to generate principal components (PCs). Retaining the 2nd through 50th PCs (discarding the 1st PC, which is generally correlated with read depth) and applying L2-normalization on the PCA matrix, we generate a low-dimensional representation for each time window. The normalized PCA matrices are used for Louvain clustering and UMAP (min.dist=0.3) as implemented in Seurat v3 [33]. For the Louvain clustering, we used resolution of 0.3 for the first round clustering and we varied the resolution parameter per time window for the final clustering after cluster specific peak calling. To select the clustering resolution parameter for the scATAC-seq clustering in each time window, we computed clusters with various clustering resolutions and selected the one at which the proportion of variance explained by the clustering plateaued (elbow method). We did not observe batch effects between the two experiments (**Fig. 2.7C**) in any developmental time window and thus did not run batch removal algorithms.

Doublet identification

The individual experiments had low estimated rates of doublets (**Fig. 2.7**). Based on the barnyard samples we expect 1.8% of doublets within our dataset (**Fig. 2.7**). We use a modified version of the scrublet algorithm [89] to calculate a per cell doublet score and set a threshold of the 95th percentile. For each developmental time window, cells with doublet scores above the threshold and clusters with over 25% cells of above the threshold are removed. The remaining cells for each developmental time window are re-embedded and re-clustered using the pipeline described above.

Cluster-specific peak calling

In order to generate a comprehensive set of peaks, we split the fragments files by cells in each Louvain cluster and call cluster-specific peaks with MACS2. The summit of each peak is extended to 150bp and then merged into a master peak set with bedtools. The peak-by-cell matrices are re-counted with the new peak set, and the cells for each developmental time window are re-embedded and re-clustered using the pipeline described above. In total, we identified 110,185 peaks (median length = 217 bp) that collectively cover 22% of the dm6 genome. We compared this peak set with sets of known elements, including annotated TSS sites (extended 2 kb upstream and 200 bp downstream), peaks identified in [1], curated set of known embryonic enhancers [42, 41, 40], and bulk DHS peaks [43]. For each pair of sets of elements, we calculate the portion of elements overlapping elements in the other set by 1 bp overlap, and vice versa. The peaks in this study overlap over 85% of each set of the known elements, while the known elements overlap less than 50% of the peaks in this study.

Global embedding

In order to visualize global trends of development based on the regulatory landscape, we randomly sampled 20,000 cells from each 2-hour developmental time window and 40,000 cells from each 4-hour developmental time window, merged the chromatin accessibility profiles and embedded the cells together using the pipeline described above.

Cell type annotation

To transfer cell type labels from the sci-ATAC-seq dataset in [1], we use the integration pipeline implemented in Signac V1 [90]) on five developmental time windows (2-4, 4-8, 6-10, 8-12, 10-14) that overlap one of the three time windows (2-4, 6-8, 10-12) profiled in [1]. We collapse the cell type subsets (*e.g.* CNS A, CNS B to CNS) in [1]) to a set of 37 cell type labels and count the number of cells in our dataset with each cell type label in each cluster. We assigned the most prominent cell type label to each cluster. For most clusters we could unambiguously assign a cell type label. We propagate the cell type labels to all developmental time windows through connections identified in the developmental tree described in the ‘Reconstruction of the developmental tree’ section. Since our study generated a much larger dataset, we expect to observe cell types that are not identified in [1]. Therefore, we also refined the transferred labels based on differentially accessible (DA) peaks and differentially expressed (DE) genes. DA peaks are calculated from the binary peak matrix with the FindMarkers() function as implemented in Seurat V3 with `test.use='LR'` and `logfc.threshold=0.1`. Peaks with a p-value less than 0.05 from the test are included in the foreground set and the remaining are included in the background set. The two sets of peaks in each sample are overlapped with a previously compiled embryonic enhancer database (CAD; [42, 41, 40]) of the closest matching time stage. We use a Fisher’s exact test to look for ‘terms’ in CAD that are enriched in the foreground set compared to the background set. We used terms with adjusted p-value less than 0.05 to refine the transferred labels. DE genes are calculated from the proximal gene activity matrix with the FindMarkers() function as implemented in Seurat V3. We look for ‘terms’ in the BDGP that are enriched with the same procedure as in sci-RNA-seq annotation.

Motif analysis

We calculate a per-cell motif activity score for known motifs with PWMs in the CisBP database (Weirauch et al. 2014) using chromVAR [91]). Within each cluster, chromVAR calculates a bias-corrected deviation of accessibility of each cell from the average accessibility

in all the cells. We also use Homer [92] to identify motifs enriched in the mesoderm-specific gene regions. We take the peaks open in over 2% of mesoderm cells and that overlap the 1kb-10kb region upstream of the TSS of the four clusters of genes in **Fig. 2.4C** as the target, peaks open in less than 2% of mesoderm cells and overlapping the 1kb-10kb region upstream of the TSS of non-mesoderm genes as the background, and run *findMotifsGenome* with the parameters *-cpg -size given* (**Table S7**).

2.4.3 Data analysis

Inferring developmental time

To estimate the developmental time stamp for each single cell, we trained lasso linear and neural net-based models using the window-by-cell read counts matrix for sci-ATAC-seq and the expression read count matrix for sci-RNA-seq to predict the midpoint of each developmental time window (*e.g.*, 7 for the 6-8 hours developmental time window). First, we equally subsampled cells for each time window to normalize the number of included cells per hour of collection time (*i.e.* 4 hour windows had twice as many nuclei as two hour windows). Following even subsampling, all genes and peaks that were constant values were removed. Then, we split the equally time subsampled data into 11 partitions of cells. The first 10 partitions were used for 10-fold cross validation to test many different model parameter choices (outlined below for each model type) and the final held out 11th partition was used as a test data set to evaluate the final models. To be clear, all model fitting described below was performed with the first 10 partitions of cells.

The lasso model was trained with `cv.glmnet()` as implemented in `glmnet` [93]. In this case, the only parameter being fit by `cv.glmnet()` is the strength of the lasso penalty. We use the trained model to infer the developmental time for all the cells with `predict()`, setting `s='lambda.min'`.

We used tensorflow v 2.6.1 (Abadi, Agarwal, et al. 2016; Abadi, Barham, et al. 2016) to fit all neural net models using a fully connected, feedforward 6-layer (4 hidden) neural network. For the first hidden layer we included variable l1 or l2 regularization with the ‘kernel_regularizer’ parameter, and in the last layer we optionally constrained the output to within 0-20 with either a sigmoid or tanh activation function. All hidden layers used relu activation functions. The ATAC and RNA model were very similar, and only differed in the number of units per hidden layer. For ATAC there were 10, 100, 60, 20, and for RNA there were 5, 100, 50, 20, 1 units per hidden layer. Moreover, the RNA model input was scaled with a ‘Normalization’ layer from tensorflow, whereas the ATAC model was not. Finally, we used ‘SparseTensor’ to encode the ATAC data.

For fitting these models we optionally used either the standard mean squared error from the center hour of the collection window (MSE) or a custom loss function based on MSE except the error is set to 0 if nuclei are placed within the correct collection window. We used 10-fold cross validation to estimate the generalization error of these models with a variety of parameterization choices for the L1 penalty [values=1, 0.1, 0.001, 0.0001, 0.00001, 0.000001, 0], L2 penalty [values=1, 0.1, 0.001, 0.0001, 0.00001, 0.000001, 0], activation function (linear, sigmoid, tanh), and loss (MSE, custom). We selected the two best performing models (one using MSE and the other using the custom loss) based on the best median MSE and the proportion of nuclei placed into the correct collection window across all the 10-folds.

We split the developmental time frame of 0-20 hours into 10 non overlapping 2 hour inferred time windows and reassign cells to each inferred time window based on their inferred times. The cells for each inferred time window are re-embedded and re-clustered using the pipeline described above.

Scripts for arranging the data and fitting the models can be found on our supplementary data sharing website.

Initially, when we tested these time inference models with bulk RNA/ATAC/DNase-seq libraries, the age estimates were outside the hour range of this experiment even though they produced the correct temporal ordering. We suspected that these large inferred time values were due to the differences in the read depth of the bulk libraries versus single-cell nuclei. To address this we simulated subsampling reads the median number of reads per nuclei (377 UMI for RNA; 5294 unique reads for ATAC) with the ‘`rmultinom`’ function from base R with the parameter ‘`prob`’ set to the read counts of the bulk library. To integrate over sampling error, from each bulk library we repeated the subsampling process 100 times and then averaged over the model-inferred predictions. This process of adjusting the read distribution of the bulk libraries to match the single-cell nuclei resulted in time predictions that were in the same scale as our time course experiment.

Inferring nuclei sex

To infer whether each nuclei was XX or XY we used the proportion of chrX-mapped sci-ATAC-seq reads as a summary statistic and identified two distinct populations **Fig. 2.16H**. After filtering out nuclei with 0 chrX-mapped reads or if they were in the tails of the distribution (prop. $X > 0.07$ and prop. $X < 0.22$), we fit a gaussian mixture model with the ‘`normalmixEM`’ function from the `mixtools` package version 1.2.0 [94]) and the parameter ‘`k`’=2. The gaussian mixture model was successfully able to separate these two populations of cells into a set with likely XX genotype (more proportion of chrX-mapped reads) compared to a set of cells that were likely XY genotype (less proportion of chrX-mapped reads). To verify that these classifications were accurate, we could validate that the XY genotyped cells were indeed male cells by verifying that they were enriched for reads mapping to the Y chromosome. Cells with a >95% probability of being classified as XX, based on this mixture model fit, only 8% of these nuclei had 0.05% or more proportion of Y-mapped reads. In contrast, for confidently predicted male cells (*i.e.* cells with >95% probability of being classified as XY), 42% had 0.05% or more Y-mapped reads. The full annotations of nuclei genotype are now included in the supplementary tables. Unfortunately,

this analysis was not possible with the sci-RNA-seq data likely due to dosage compensation and too few reads per nuclei.

Reconstruction of the developmental tree

To connect each cell state observed in a predicted time window with its most probable ancestor cell state from the previous predicted time window, we use the k -NN approach described in [95]. We took cells from neighboring predicted time windows and co-embedded them using the pipeline described above. For cells in the later inferred time window, we identify ten k -NN cells in the previous inferred time window based on 50 PCs. The edge weights connecting the clusters in each inferred time window to clusters in the previous inferred time windows are set to the percent of cells in each cluster that has majority k -NNs from a cluster in the previous inferred time window. Edge weights < 0.2 are removed. Branches with only one leaf and nodes in the 18-20 hours inferred time window are pruned for the developmental tree shown in **Fig. 2.3C-D**.

Spatial analysis

A tranche of spatial data was recently released including a study on *Drosophila* using SpaTial Enhanced REsolution Omics-sequencing (Stereo-seq) [59], which was collected from late-stage embryos and all stages of larvae. The late-stage embryos were from corresponding windows as our embryo samples 14-16 h and 16-18 h after egg laying. This technology is not single-cell based, but instead patterned DNA nanoballs are placed on a slide that then capture RNA transcripts which are then sequenced and associated back to the slide spatial position of the nanoball slide spatial position. With an anchor-based integration *FindTransferAnchors()* and *TransferData()* from Seurat v3 [33], we performed probabilistic label transfer to assign our cluster labels to each patterned nanoball's spatial location. Using the assigned annotations of tissues from the original study as reference above, we observe a correspondence with our cluster annotations (**Fig. 2.12E**).

Connecting cell states in the sci-ATAC-seq dataset to those in the sci-RNA-seq dataset

In order to identify matching cell states in the sci-ATAC-seq and sci-RNA-seq datasets, we implemented the non-negative least squares (NNLS) approach as described in [16, 2]. For the sci-RNA-seq dataset, we calculate an aggregate expression vector for each cluster in each time window by summing the log-transformed normalized UMI counts of all cells in that cluster. For the sci-ATAC-seq dataset we take the proximal gene activity matrix and calculate a similar aggregate activity vector for each cluster. Then we apply non-negative least squares (NNLS) regression to predict gene expression in a target cluster in the sci-RNA-seq dataset based on the gene activity of all clusters in the corresponding time window in the sci-ATAC-seq dataset. The resulting β matrix is denoted $\beta_{ra} \in R^{i \times j}$ where i and j are the number of RNA and ATAC clusters respectively. We then repeat the analysis predicting gene activity with gene expression to obtain $\beta_{ar} \in R^{j \times i}$. For each pair of clusters from the two datasets, we calculate a final beta value based on the element-wise multiplication of the two beta value matrices from NNLS: $\beta = 2(\beta_{ra} + 0.001)(\beta_{ar}^T + 0.001)$. Similar clusters, based on patterns of gene expression and the sum of ATAC-seq reads around a gene, in the two datasets have higher beta values.

Determining the relationship between TF expression and associated motif accessibility for NNLS-linked clusters

Relying on the NNLS-based links between clusters of ATAC and RNA data, we next set out to determine correlation between TF-associated motif accessibility and TF-associated gene expression. Presumably TFs with strong correlations are active, with positive correlations indicating TFs likely to be activators of gene expression and negative correlations indicating TFs likely to be repressors of gene expression. For each cluster we averaged the expression of each TF across all cells and averaged the associated motif accessibility score (as described previously) across all cells. Next for each RNA cluster we chose the best associated ATAC cluster as determined through NNLS analysis. For comparison we also performed the

vice versa association pairing each ATAC cluster with the top associated RNA cluster and the inferred correlations were similar (**Fig. 2.14G**). We then computed the spearman correlation and Pearson’s R correlation between motif accessibility score and gene expression for each TF with a known associated motif. Additionally, we computed these correlations for each gene at each two hour inferred time window. For the analysis visualized in **Fig. 2.5D** we fit a linear regression model with ‘lm()’ in R that predicts the TF motif-associated ATAC-seq activity (from chromVAR) from an interaction term including the expression of the related TF, the germ layer of the cluster, and time window. Prior to fitting this model, the TF expression values were scaled and normalized to have 0 mean and an sd of 1 with ‘scale()’. Additionally, we weighed the model observations by the NNLS correlation value representing the strength of the link between the ATAC and the RNA cluster. Intuitively, weighting the observations will increase the contribution from clusters that are strongly linked compared to clusters that have a weak link between ATAC and RNA.

Gene regulatory network analysis

To integrate our transcriptome and chromatin accessibility data, we first created subsampled Seurat objects (5,000 cells) for cells at 10-12 hr and log normalized the RNA counts and ATAC gene activities. We identified variable features for each object using Seurat ‘SelectIntegrationFeatures’ function and used the features to perform Canonical Correlation Analysis (CCA) with Seurat RunCCA function, in order to create a common embedding for our independently-assayed ATAC and RNA modalities. In CCA space, we used FigR ‘pairCells’ function to identify pairs of ATAC-RNA cells by geodesic distance-based pairing [68]. Gene expression and peak accessibility counts from paired cells were used as input for FigR ‘runGenePeakcorr’ in order to identify significant (p -value < 0.05) peak-gene associations. The ‘runGenePeakcorr’ function was slightly modified to accept the *Drosophila melanogaster* dm6 genome (Bioconductor library BSgenome.Dmelanogaster.UCSC.dm6) as a valid input. At this point, the DORC scores and the original RNA counts were smoothed with FigR function ‘smoothScoresNN’, and fed together with the original ATAC peak counts

and the significant peak-gene associations into FigR function ‘runFigRGRN’, for inference of the gene regulatory networks. The ‘runFigRGRN’ was slightly modified to accept *Drosophila melanogaster* CisBP PWMs as a valid input. For generating the plots presented in **Figure S11**, we ranked TFs by average regulation score using the FigR function ‘rankDrivers’, while TFs regulating individual genes were retrieved with the function ‘plotDrivers’. Domains of regulatory chromatin (DORCs) were identified with FigR function ‘dorcJPlot’ with ‘cutoff’ = 10 (at least 10 significantly linked peaks to be called a DORC) (**Fig. 2.16C**).

2.5 Supplementary materials

2.5.1 Supplementary notes

Supplementary Note 1: Diversification of non-myogenic mesoderm

We examined the diversification of the non-myogenic mesoderm trajectories in more detail . The non-myogenic mesoderm includes the fat body and haemocyte lineages, which in the embryo includes plasmatocytes and crystal cells. Previous genetic studies demonstrated that the fat body develops from the trunk mesoderm (Hoshizaki et al. 1994), similar to the somatic muscle, while the embryonic haemocytes originate from the head mesoderm (de Velasco et al. 2006). In agreement with this, and without any prior knowledge, the scATAC-derived graphs suggest that the fat body shares a developmental origin with somatic muscle, whereas the haemocytes originate from a separate mesodermal trajectory.

To explore these trajectories at finer resolution, we isolated and re-clustered all cells annotated as plasmatocytes, fat body and crystal cells, using scRNA data from 6-18 hrs . This revealed 17 subclusters, with a large group of early cells (right, clusters 1, 4, 13, 6, 7), that diversifies at 8-12 hrs. The early clusters are defined by the expression of several transcription factors (TFs) including Ultrabithorax (*Ubx*), restricted to the earlier stages, and homothorax (*hth*) and Zn finger homeodomain 1 (*zfh1*), expressed from early to late stages in development . All branches express common haemocyte and fat body marker genes such as the collagenase *Col4a1*, the procollagen lysyl hydroxylase *Plod* and the extracellular matrix gene *Tig*. These branches correspond to sub-trajectories for plasmatocytes (clusters 8, 9, 12, 14), fat body (clusters 0, 3, 10, 11, 16), and additional two muscle clusters (cluster 2, 5). Finally, the crystal cells (a relatively rare cell type, expressing the marker genes *Lozenge*, *PPO1*, *PPO2*) are clearly separated and disconnected from the rest (cluster 15), suggesting

that they may have a different developmental origin or, alternatively, that they deviate from a common precursor population earlier than 6 hrs.

Plasmatocytes represent $\sim 90\%$ of the embryonic haemocytes, and are phagocytic cells involved in clearing apoptotic cells (Wood and Jacinto 2007), which requires the expression of the scavenger receptors *crq* and *drpr* (Franc et al. 1999; Manaka et al. 2004). We observe both genes, in addition to other scavenger receptors (*e.g.* *Nimrod C4* and *NimB4*, specific to later embryonic stages), expressed along the plasmatocyte sub-trajectory. The second function of plasmatocytes is the secretion of extracellular matrix (ECM) proteins, which is also evident from our single cell trajectory revealing dynamic expression of a large repertoire of ECM proteins, such as Papilin (*Ppn*), Peroxidasin (*Pxn*), Glutactin (*Glt*), Tiggrin (*Tig*), basement membrane-associated SPARC, laminin A (*lanA*), and two collagenase IV molecules - *Col4a1* and Viking (*vkq*). The second major branch is the fat body, which is thought to have similar roles to the mammalian liver and adipose tissues, in addition to having an essential role in immune responses, as the main source of antimicrobial peptides (AMPs) (Vlisidou and Wood 2015). We identified 289 genes significantly associated with specific subclusters that highlight the dynamic differentiation of the fat body. For example, the expression of genes involved in lipid transport (*e.g.* *apolpp* and *apoltp* (Ugrankar et al. 2019; Rodríguez-Vázquez et al. 2015)), increases with the maturation of the fat body cells.

Supplementary Note 2: Tracing dynamic gene modules across development

The analysis steps associated with mesoderm gene clustering of **Fig. 2.4** were motivated to identify the TFs driving mesoderm-specific expression variability and determine whether we were capable of ordering putative regulatory events (*e.g.* *Kah/Mhc*). Of course, there are different directions to approach this or related questions. Therefore, we performed a complementary analysis by clustering accessible regions variable in the mesoderm tissue (**Fig. 2.13D**). The clustering approach was the same as in **Fig. 2.4** but using variable

chromatin accessibility regions from the set of mesoderm annotated nuclei. This analysis identified a similar set of 4 clusters of accessibility elements that are likely involved with regulatory patterning linked to mesoderm differentiation. Additionally, we performed a motif enrichment analysis on these peaks and found 483 significant associations between a peak group and TF (q -value $< 1 \times 10^{-3}$ and a match present in at least 1% of target peaks) from 192 unique TFs. Tying this back to the mesoderm expression analysis, we found that 40 of these TFs were in the set of genes that were variably expressed in the mesoderm tissue. These significant TF associations and the clustered peaks are available from our supplementary data website and **Table S12-S13**.

Supplementary Note 3: Nomination of fruitless as a repressive regulator in neuroectoderm

In the neuroectoderm, Fruitless (Fru) is associated with decreasing chromatin accessibility; *i.e.* increasing *fru* expression led to chromatin regions containing the Fru motif to become less accessible (**Fig. 2.5D**, bottom). Intriguingly, Fruitless is associated with specifying the molecular determinants of male courtship behavior by ‘masculinizing’ specific neurons [96, 97, 98]. However, only recently have studies shown that Fru likely acts through the repression of chromatin accessibility at specific regulatory elements [99, 100, 101].

2.5.2 Supplementary tables

Additional Data Table S1 (separate file)

RNA cell type annotations. For each cluster in each time window of the RNA-seq data we provide an automated cell type or tissue annotation, a manual cell type annotation that incorporates information from marker genes and germ layer.

Additional Data Table S2 (separate file)

ATAC cell type annotations. For each cluster in each time window of the ATAC-

seq data we provide an automated cell type annotation based on integration with previously annotated datasets, a refined cell type annotation that incorporates information from marker regions, marker genes and germ layer, and lineage information based on connecting the clusters from neighboring time windows.

Additional Data Table S3 (separate file)

RNA marker genes. Selected marker genes for each cluster in each time window of the RNA-seq data identified with a differential accessibility test with high statistical significance

Additional Data Table S4 (separate file)

ATAC marker accessible regions. Selected marker regions for each cluster in each time window of the ATAC-seq data identified with a differential accessibility test with high statistical significance.

Additional Data Table S5 (separate file)

Marker genes that differentiate each identified cluster in the neuronal data subset. Selected marker genes for each sci-RNA-seq cell state from the neuroectodermal vignette in Fig. 2.3, identified with a differential accessibility test with high statistical significance. Marker genes for the clustering of the full neuroectodermal set of cells are indicated by the ‘Clustering’ field set to ‘Broad’, whereas marker genes for the clustering of just the mature neuron annotated set of cells are indicated by the ‘Clustering’ field set to ‘Fine’.

Additional Data Table S6 (separate file)

Variably expressed mesoderm genes and their cluster assignment. The set of variably expressed genes identified in the mesoderm-annotated cell types which were then grouped into clusters by their patterns of temporal co-regulation.

Additional Data Table S7 (separate file)

TF motifs enriched in accessible regions near genes from mesoderm gene clusters. Results from the home-based motif enrichment in accessible regions that were proximal to the variably expressed mesoderm genes. These values are separated into the different mesoderm cluster groups.

Additional Data Table S8 (separate file)

NNLS-based links between RNA and ATAC clusters. NNLS coefficients for each pair of clusters of the ATAC-seq data and the RNA-seq data separated by the different 2 hr time windows.

Additional Data Table S9 (separate file)

Correlations between motif and expression. For each of the TFs with a motif in cisBP we computed the Pearson correlation between the expression of the TF and associated motif accessibility score. These values were computed at every time window.

Additional Data Table S10 (separate file)

Model effect estimates between motif and expression in germ layers by time. Linear regression effect estimates comparing TF expression with its corresponding motif accessibility score with a model that includes an interaction term that includes germ layer, time window.

Additional Data Table S11 (separate file)

UMAP clustering resolutions per time window for RNA and ATAC. For each time window we performed clustering of the data with multiple resolution parameter values then selected the parameter value at which the variability explained of the resulting clusters plateaued.

Additional Data Table S12 (separate file)

Variably accessible mesoderm peaks and their cluster assignment. The set of variably expressed accessible regions identified in the mesoderm-annotated cell types which were then grouped into clusters by their patterns of temporal co-regulation.

Additional Data Table S13 (separate file)

TF motifs enriched in mesoderm accessible region clusters. Results from the home-based motif enrichment in the accessible regions that were variable in mesoderm-annotated cells. These values are separated into the different mesoderm cluster groups.

2.5.3 Supplementary figures

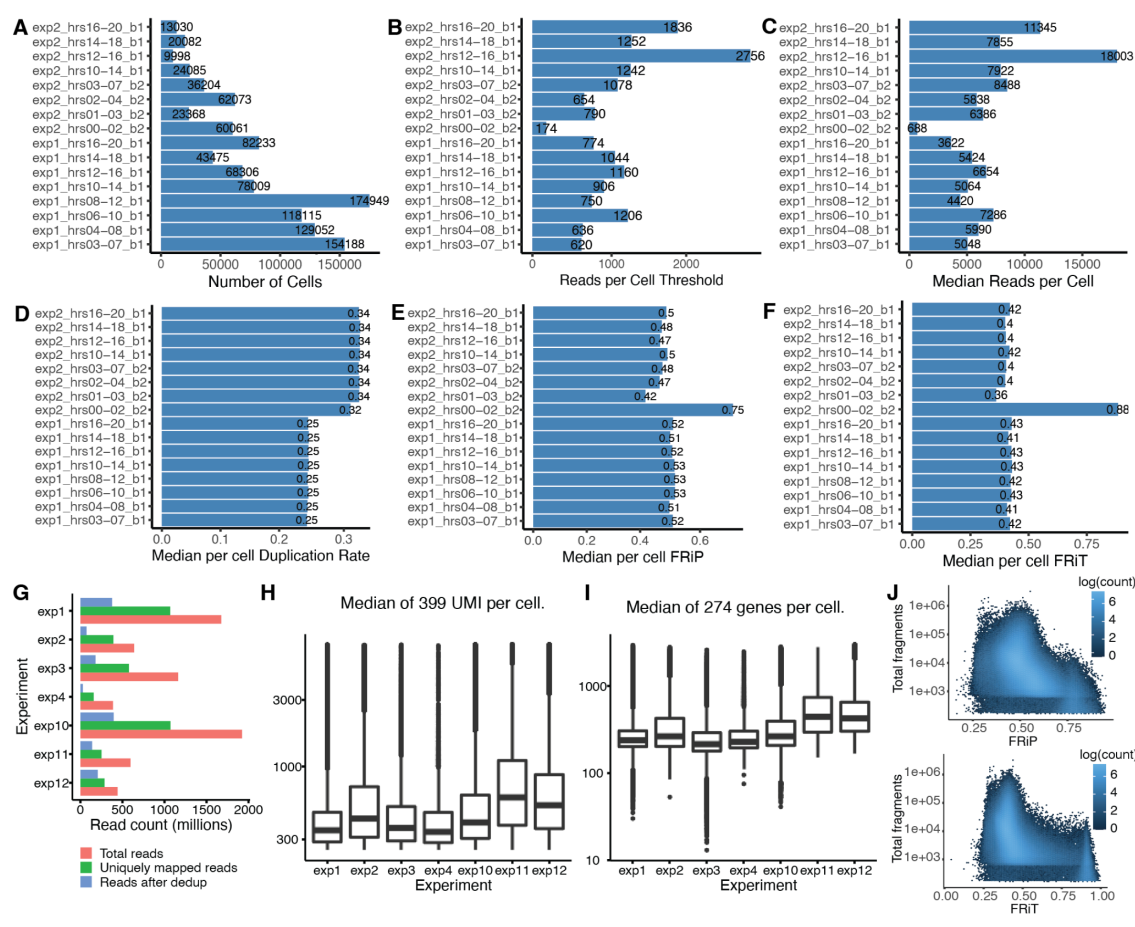


Figure 2.6: **Figure S1: QC/stats of the sci-ATAC-seq and sci-RNA-seq datasets.** (A) Barplot showing number of cells passing threshold in each sample in the sci-ATAC-seq dataset. Samples are labeled by the experiment (exp1/2), developmental hours of the nuclei collection, and the batch in which the nuclei were collected in (b1/2). (B) Barplot showing reads per cell threshold for each sample in the sci-ATAC-seq dataset. (C) Barplot showing the median reads per cell for each sample in the sci-ATAC-seq dataset. (D) Barplot showing median duplication rate per cell in each sample in the sci-ATAC-seq dataset. (E) Barplot showing median fraction of reads in peaks (FRiP) per cell in each sample in the sci-ATAC-seq dataset. (F) Barplot showing median fraction of reads in TSS sites (FRiT) per cell in each sample in the sci-ATAC-seq dataset. (G) Barplot showing the total number of reads, uniquely mapped reads and reads after deduplication in each experiment in the sci-RNA-seq dataset. (H) Boxplot showing the number of UMIs per cell in each experiment in the sci-RNA-seq dataset. (I) Boxplot showing the number of genes per cell in each experiment in the sci-RNA-seq dataset. (J) Hex bin plot showing distribution of nuclei FRiP (top) or FRiT (bottom) scores by total unique fragments.

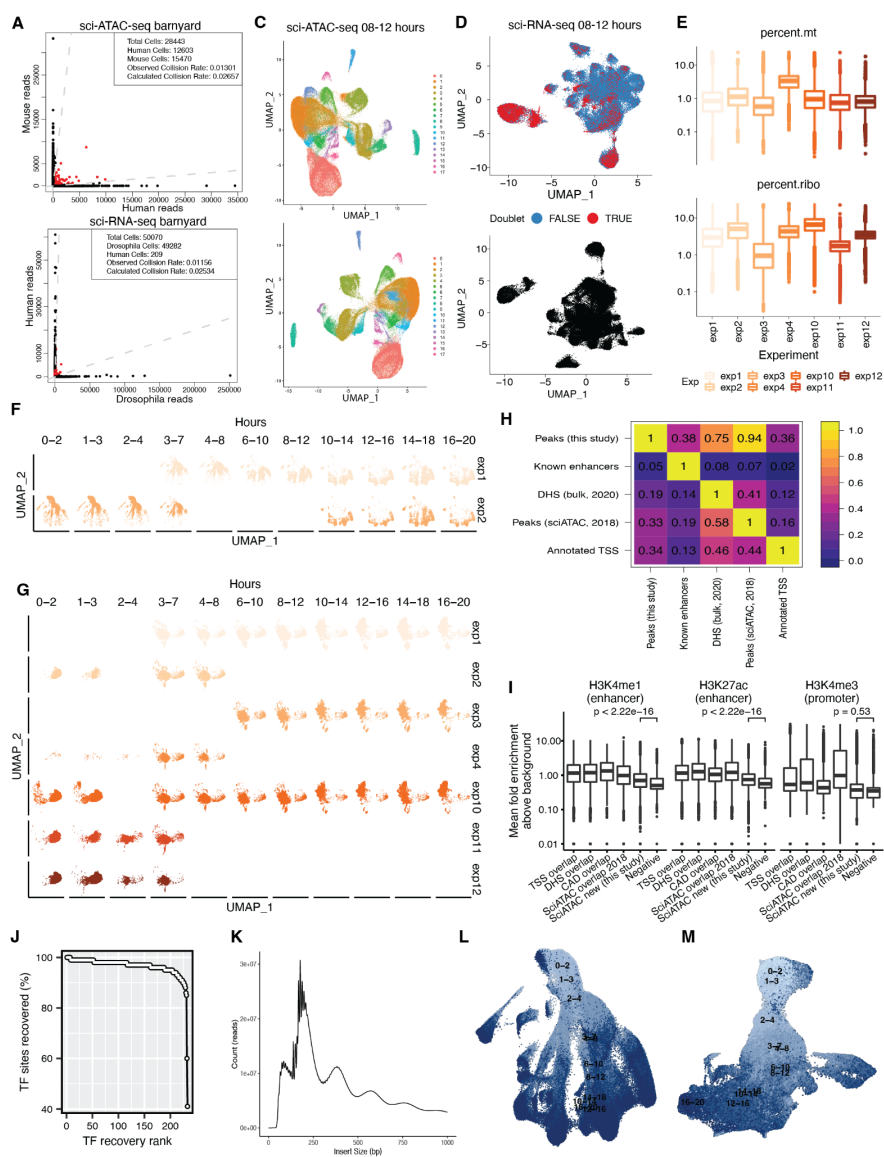


Figure 2.7: **Figure S2: Data filtering, evaluating batch and technical effects.** (A) Scatterplot showing estimated doublet rates from a human-mouse barnyard experiment for the scATAC data. (B) Scatterplot showing estimated doublet rates from a human-drosophila barnyard experiment for the scRNA data. (C) UMAPs for the scATAC data from the 8-12 hr time window before (top) and after (bottom) doublet removal. (D) UMAPs for the scRNA cells from the 8-12 hr time window before (top) and after (bottom) doublet removal. (E) Barplots of percentage of scRNA reads from ribosomal genes or mitochondria per experiment. (F) Global scATAC UMAP faceted by time window and experiment. (G) Global scRNA UMAPs faceted by time window and experiment. (H) Same as Fig. 2.1D but displaying percent overlap of bases between two sets of peaks. (J) Proportion of TF ChIP-seq binding sites that overlaps with peaks from our data for the full set of 233 TFs included in modERN. (K) Insert size distribution across all experiments. (L) Same as Fig. 2.1B but including labels of the time window placed at the median of all nuclei collected from this time window. (M) Same as L but for RNA-seq data.

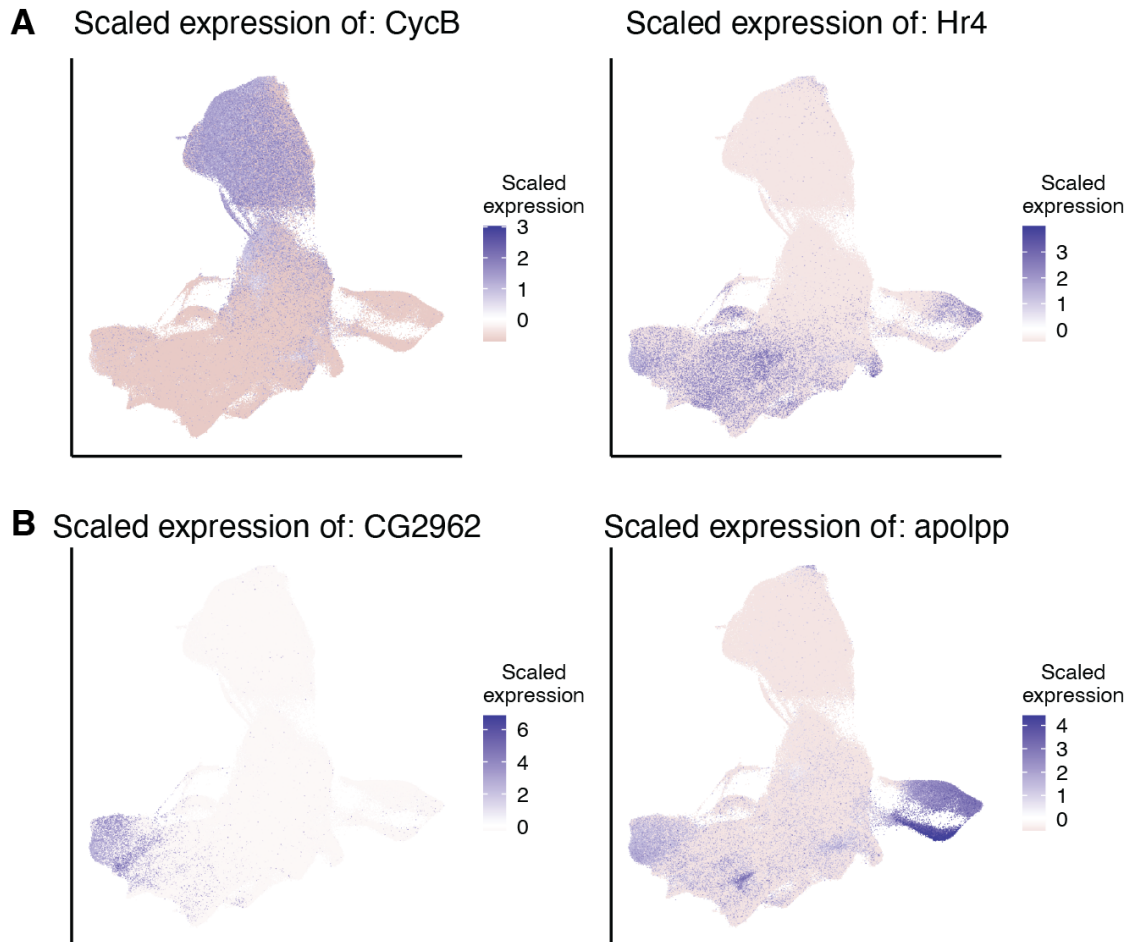


Figure 2.8: **Figure S3: Examples of time-dependent genes, some of which appear universal, others of which are specific to certain clusters.** (A) Two examples of genes that are associated with general expression in cells from early (left) or late (right) time windows. Scaled expression values overlaid on a global UMAP of scRNA data. This is similar to the UMAP shown in **Fig. 1C**, but here with all cells included instead of an evenly time subsampled dataset. (B) Similar to panel A, except that these time-dependent genes tend to be expressed in later time windows in restricted subsets of UMAP clusters.

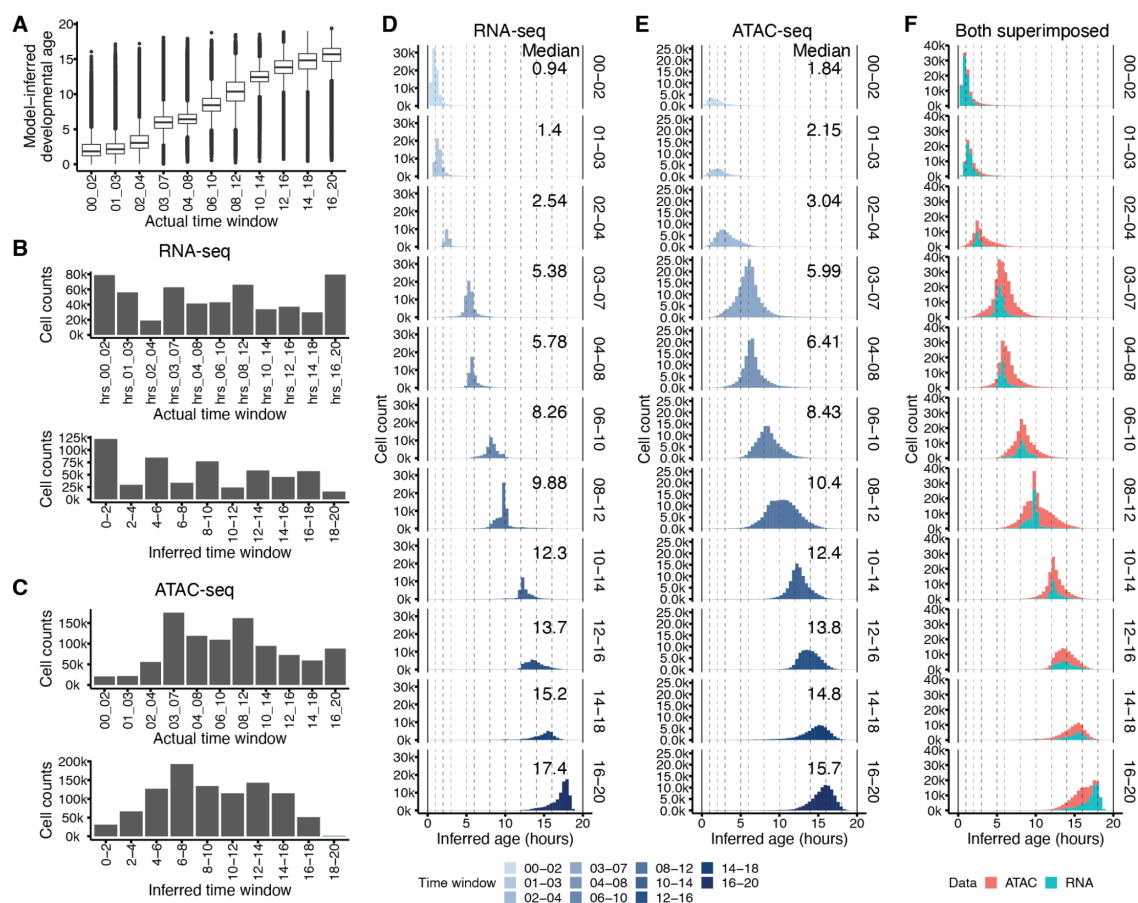


Figure 2.9: **Figure S4: A neural network-based model for inferring cellular developmental ages.** (A) The distribution of model-inferred developmental age for nuclei from the scATAC model across actual time windows (x-axis). (B) Counts of cells in scRNA data per actual 2-hr time window (top) and equivalent set of inferred 2-hr time windows (bottom). (C) Same as panel B except for counts of scATAC nuclei from actual vs. inferred time windows. (D) Distribution of inferred ages for cells stratified by the actual time windows from which the cells were collected. Dotted lines correspond to centers of each time window. The median age of cells per actual time window are listed to the right. (E) Same as panel D but for scATAC data. (F) Similar as E and D but superimposing model-inferred ages from both data types.

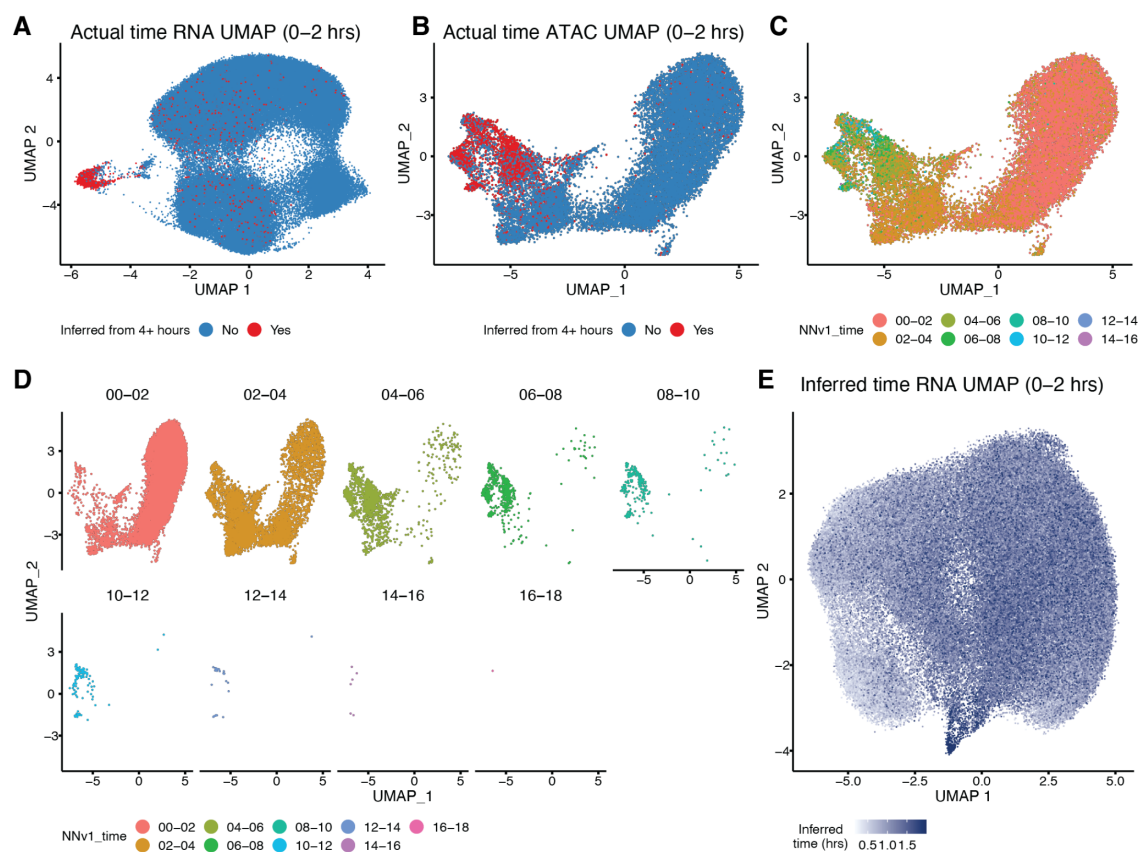


Figure 2.10: **Figure S5: Model of developmental ages identifies “contaminating” older nuclei in the earliest collection window.** (A) UMAP of scRNA data of cells from the 0-2 hr actual time window. Cells inferred to be over 4+ hours in developmental age by the model are highlighted in red. (B) Same as panel A but visualizing the scATAC data. (C) Same as panel B but cells are colored by model-inferred time. (D) Same as panel D but cells are split by model-inferred time. (E) UMAP of scRNA data of cells inferred to be up to 2 hrs in developmental age.

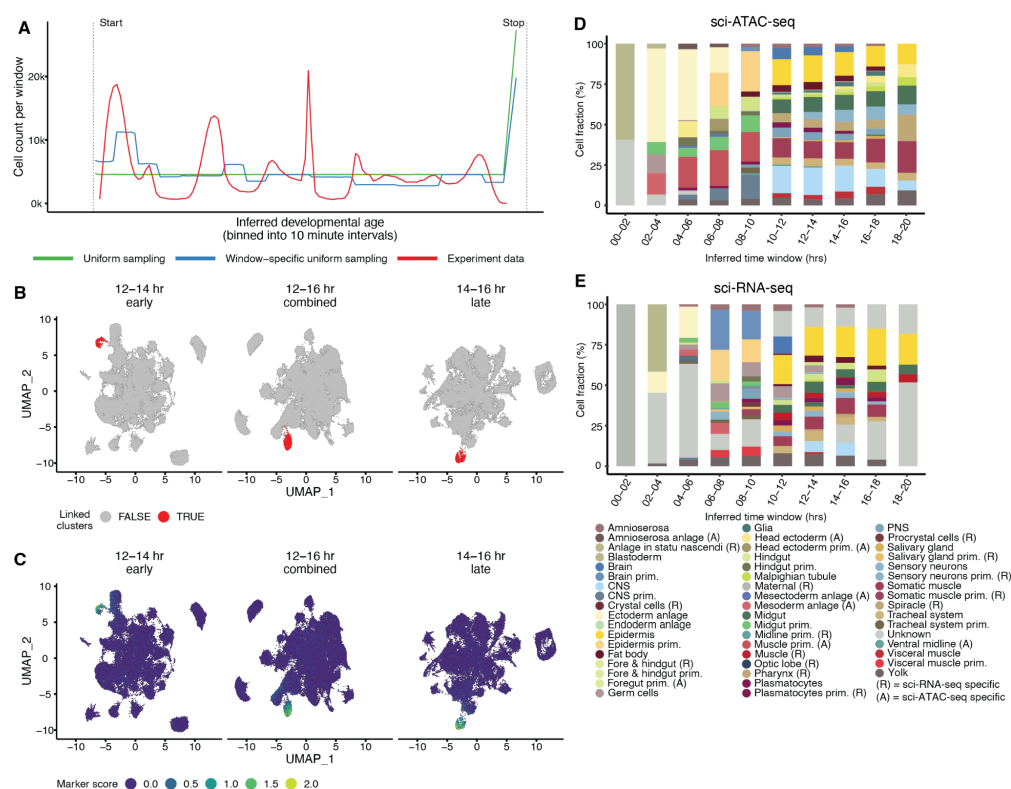


Figure 2.11: Figure S6: Distribution of cells and cell states through inferred time. (A) Cells were binned into 10 minute intervals based on their inferred developmental ages. The counts of cells per window are displayed on the y-axis. Counts of cells from the actual distribution of inferred ages for all nuclei are highlighted in red. We wanted to compare the distribution of inferred ages to what might be expected under different potential cell sampling strategies. A uniform sampling of all 547,805 cells across 10 min bins is visualized in green. Alternatively, uniformly sampling cells from actual time windows into 10 min bins is visualized in blue. This strategy accounts for different numbers of cells that were collected from different actual time windows. (B) Illustration of the process of linking clusters from adjacent time windows. Clusters eventually found to be linked are highlighted in red in adjacent windows (left, right panels), in the co-embedded UMAP space that includes nuclei from both time windows the linked cells are mapped to the same cluster. (C) Same UMAP embedding as panel B, however cells are colored by their marker gene score defined by marker genes identified from just the 12-14 hr time window linked clustered. (D) Cell type proportions in each 2 hour inferred time window of the scATAC data. Cell type colors match those in **Fig. 2.3A-B**. Cell types annotated as unknown are excluded from the bar graph. (E) Same as panel D but for the scRNA data.

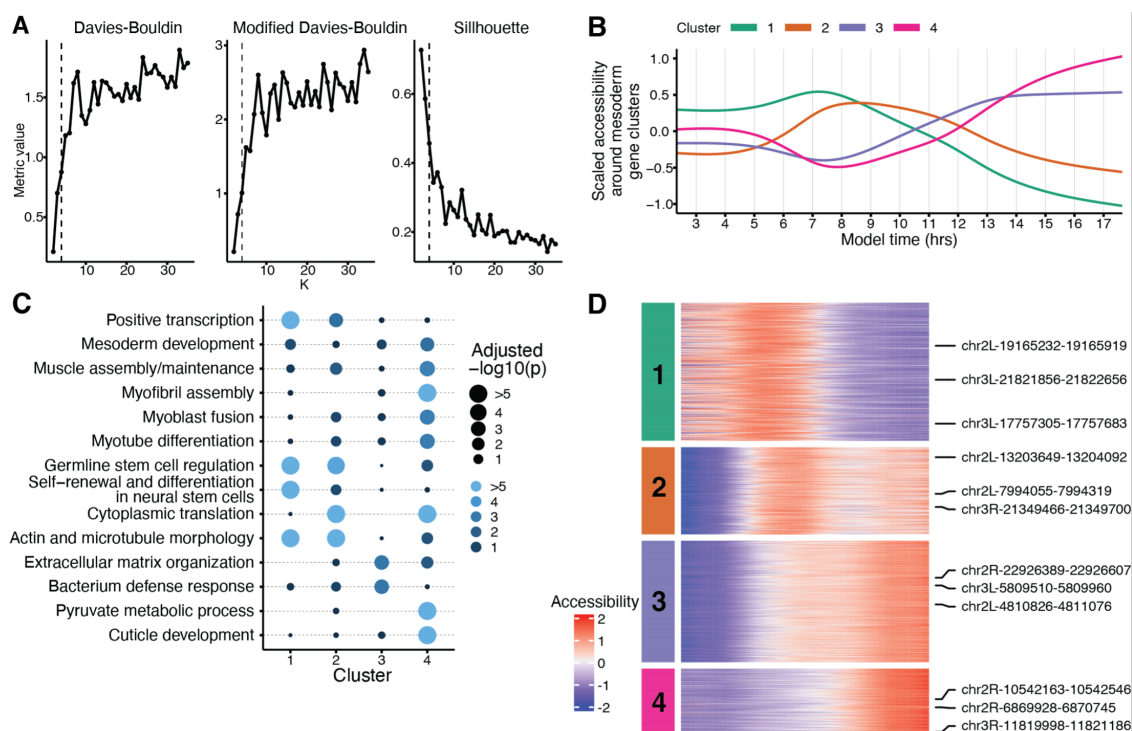


Figure 2.13: **Figure S8: Choosing the number of clusters for clustering of variable mesoderm genes, and additional ATAC mesoderm clustering.** (A) Visualization of several metrics for evaluating different choices of k for clustering. The dotted line indicates our choice of $k=4$ for clustering mesoderm genes. (B) Scaled accessibility around genes that were included in the four mesoderm clusters. (C) Results from gene pathway enrichment of genes in the mesoderm-upregulated gene cluster. (D) Heatmap of clustered mesoderm accessibility regions using a similar approach as in Fig. 2.4C.

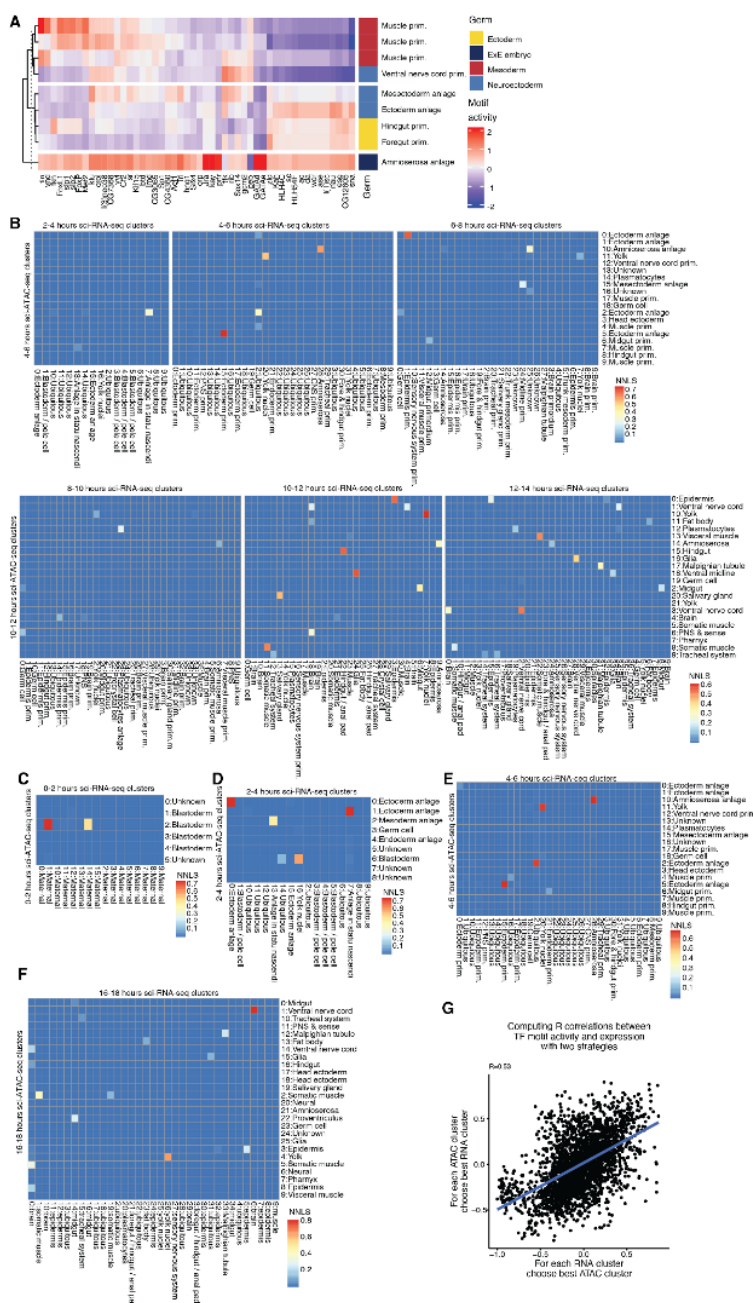


Figure 2.14: **Figure S9: Motif activity and linking of scATAC and scRNA-derived clusters.** (A) Same as **Fig. 2.5A**, except for the 6-8 hr time window. (B) NNLs results for between scATAC annotations and connections to an earlier 2 hr window (left), the matching hour window (middle) and a later 2 hr window of scRNA annotations. (C-F) Heatmaps visualizing the NNLs analyses derived coefficients linking scATAC (rows) and scRNA (columns) clusters in several 2 hour inferred time windows. (G) Comparing the Pearson's R correlation values computed between motif-associated chromatin accessibility and TF expression using two strategies of linking ATAC and RNA clusters. Blue line indicates a linear regression fit.

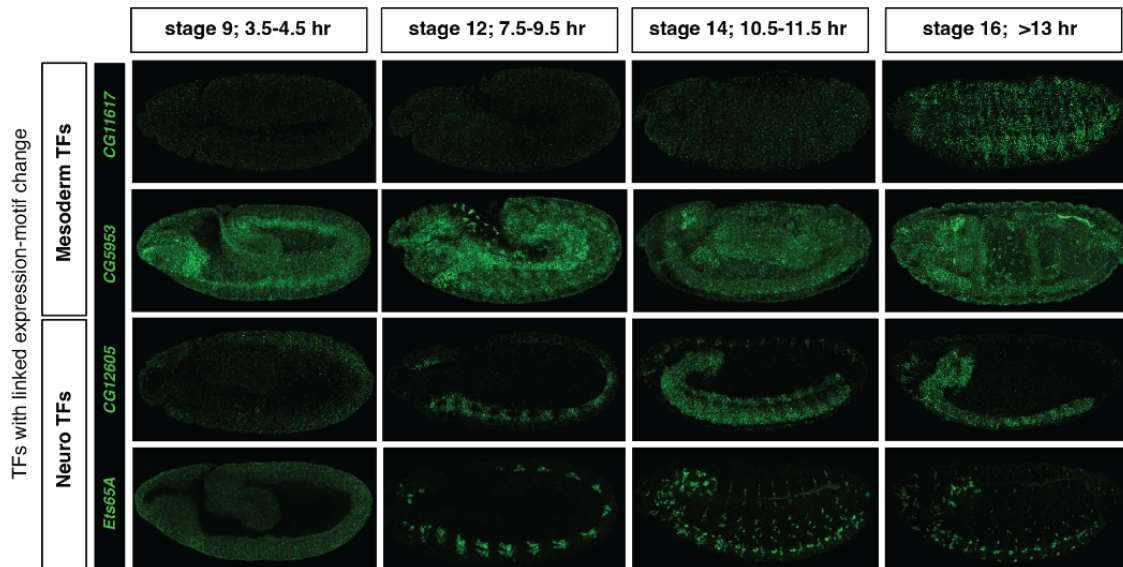


Figure 2.15: **Figure S10: Expression validation of putative mesodermal and neuroectodermal TFs by in situ hybridization.** *In-situ* hybridization demonstrates the expected temporal and tissue expression of four TFs in the mesoderm / muscle (CG11617, CG5953) or neuroectoderm (CG12605, Ets65A).

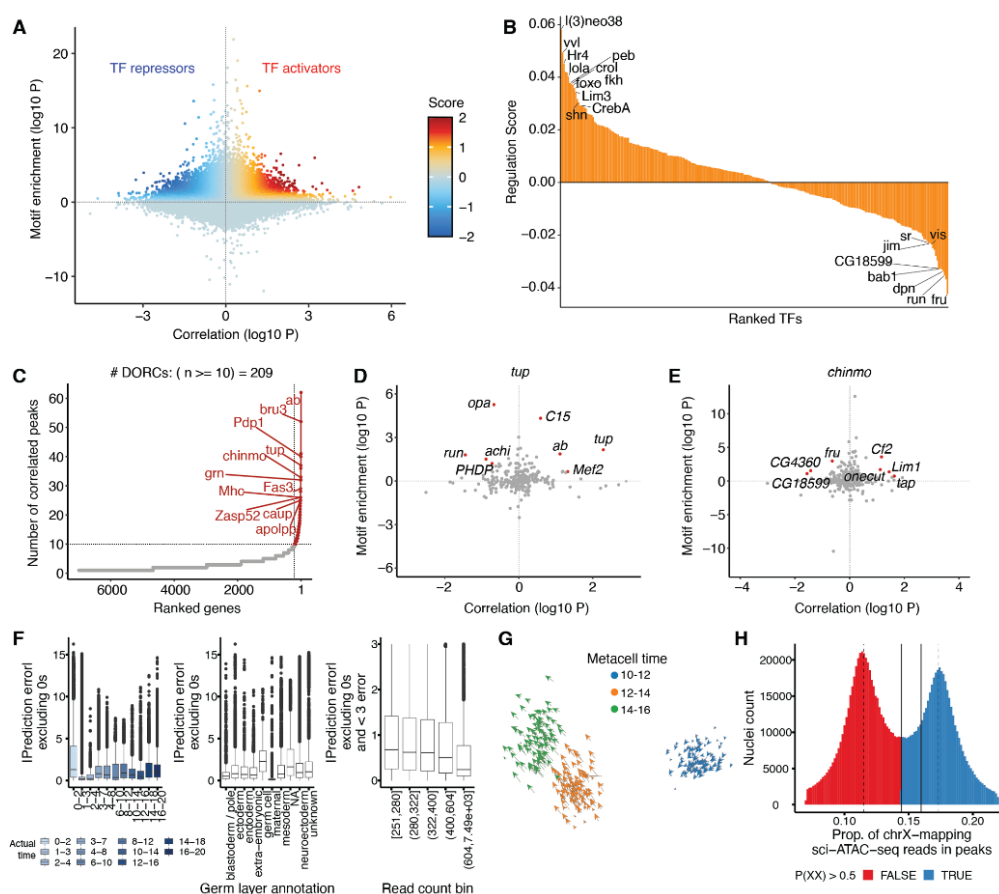


Figure 2.16: **Figure S11: Additional analyses of interest.** (A) Scatterplot of all TF-regulatory domains associations identified by FigR, colored by the regulation score. A positive score indicates putative TF activators, a negative score indicates repressors. (B) TFs ranked by their average regulation score (across all associated regulatory regions), putative TF activators (left-skewed) and repressors (right-skewed) are highlighted. (C) Genes ranked by the number of significant peak-gene links identified by FigR. Highlighted in red are genes with a high number of peak-gene links ($n \geq 10$), which are referred to as domains of regulatory chromatin (DORCs). (D) Candidate TF regulators of *tup*. Highlighted are TFs with an absolute regulation score > 0.5 , all other TFs are colored in gray. (E) Same as panel D, but showing the candidate TF regulators of *chinmo*. (F) The error of the model-inferred age prediction is visualized as a boxplot stratified by experimental time window (left), germ layer (middle), or read depth (right). (G) The velocity vector estimated on ‘metacells’, which consisted of a random subsample of cells per window that were pseudobulked prior to velocity analysis. (H) Distribution of the proportion of chrX-mapped reads from sci-ATAC data. Cells are classified as XX or XY with a gaussian mixture model.

Chapter 3

SINGLE CELL, WHOLE EMBRYO PHENOTYPING OF PLEIOTROPIC DISORDERS OF MAMMALIAN DEVELOPMENT

This Chapter is adopted from in revision work with minimum changes:

X., Huang*, J., Henck*, C., Qiu*, V.K.A., Sreenivasan, S., Balachandran, R., Behncke, W., Chan, A., Despang, D. E., Dickel, N., Haag, R., Hägerling, N., Hansmeier, F., Hennig, C., Marshall, S., Rajderkar, A., Ringel, M., Robson, L., Saunders, S. R., Srivatsan, S., Ulferts, L., Wittler, Y., Zhu, V. M., Kalscheuer, D., Ibrahim, I., Kurth, U., Kornak, D. R., Beier, A., Visel, L. A., Pennacchio, C., Trapnell, J., Cao#, M., Spielmann#, and J., Shendure#, “Single cell, whole embryo phenotyping of pleiotropic disorders of mammalian development”, bioRxiv, 2022.

In May 2020, Jun invited CX and me to participate in a project involving a “mutant mouse atlas with 1.7 million cells”. Jun had initially described the analysis as “straight-forward”, but the reality proved to be more challenging than we anticipated. Fortunately (for me), CX spent a lot of time perfecting the annotations and regressing out potential technical biases, resulting in a clean and fully annotated dataset to work with. At first, I tried to search for differential expression, but we soon realized that the phenotypic changes associated with many of the mutants would be subtle, requiring new analysis methods to detect. I then shifted my focus to detecting cell distribution shifts between mutant and wildtype within cell types. Initially, I attempted to apply Moran’s I for this purpose, which provided a summary of each pairwise embedding as a statistic ranging from zero to one as mutant and wildtype cells transitioned from mixed to segregated. After taking inspiration

from k -NN graph construction in Moran's I and some of Jun's suggestions, I developed a novel statistic called "mutant score" to measure the over-representation of mutant cells in each cell's neighborhood. It took many iterations to determine the most effective normalization, vectorization, and visualization of mutant scores, as well as substantial research to find the best examples for further validation. The part of the project that I am most proud of is the story we were able to construct around the *Gli2* KO mutant. Jana and I spent a lot of time researching and delving into the literature on mouse roof plate development. By examining both marker genes and spatial mapping results, we were able to identify changes in the choroid plexus as the mutant's likely phenotype. This hypothesis allowed us to design and carry out validation experiments, which ultimately confirmed our findings. This aspect of the project demonstrated the power of combining careful analysis with a deep understanding of developmental biology to gain insights into complex biological systems. The project meetings were (again) scheduled at 8AM PST, so I would usually be half-awake with a cat sitting in front of the camera while learning about mouse development. When we shared the manuscript with our collaborators, their most urgent comment was that we needed a better name for mutant scores. Jay jokingly suggested the acronym "lochNESS," and we ended up all liking it.

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. 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. 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, as well as for advancing our general understanding of mammalian biology. However, it has long been suspected that conventional approaches for phenotyping are insufficiently sensitive to detect subtle defects throughout the developing mouse. Here we set out to establish single cell RNA sequencing (sc-RNA-seq) of the whole embryo as a scalable platform for the systematic molecular and cellular phenotyping of mouse genetic models. We applied combinatorial indexing-based sc-RNA-seq to profile 101 embryos of 26 genotypes at embryonic stage E13.5, altogether profiling gene expression in over 1.6M nuclei. The 26 genotypes include 22 mouse mutants representing a range of anticipated severities, from established multisystem disorders to deletions of individual enhancers, as well as the 4 wildtype backgrounds on which these mutants reside. We developed and applied several analytical frameworks for detecting differences in composition and/or gene expression across 52 cell types or trajectories. Some mutants exhibited changes in dozens of trajectories (*e.g.*, the pleiotropic consequences of altering the *Sox9* regulatory landscape) whereas others showed phenotypes affecting specific subsets of cells. We also identify differences between widely used wildtype strains, compare phenotyping of gain vs. loss of function mutants, and characterise deletions of topological associating domain (TAD) boundaries. Intriguingly, even among these 22 mutants, some changes are shared by heretofore unrelated models, 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.

3.1 Introduction

For over 100 years, the laboratory mouse (*Mus musculus*) has served as the quintessential animal model for studying both common and rare human diseases [102, 103, 104]. 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 [105].

In the first decades of the field, mouse genetics relied on spontaneous or induced mutations resulting in visible physical defects that could then be mapped. However, gene-targeting techniques subsequently paved the way for “reverse genetics”, *i.e.* analysing the phenotypic effects of intentionally engineered mutations. Through systematic efforts such as the International Knockout Mouse Consortium, knockout models are now available for thousands of genes [106]). Furthermore, with the emergence of CRISPR/Cas genome editing [107, 108], it is increasingly practical to delete individual regulatory elements or otherwise modify the *cis*-regulatory landscape, and to then study the *in vivo* consequences of these alterations [109, 110].

Phenotyping has also grown more sophisticated. Conventional investigations of developmental syndromes typically focus on one organ system at a specific stage of development, *e.g.* combining expression analyses, histology, and imaging to investigate a visible malformation [111, 112, 102]. However, pleiotropy is a pervasive phenomenon in mammalian development, and focusing on one aspect of a phenotype may come at the expense of detecting or characterising others, particularly if they are subtle or masked by lethality. The concept of the Mouse Clinic, in which a given model is subjected to a battery of standardised tests, reflects a more systematic approach [113]. However, such clinics are expensive and time-consuming to conduct in practice. Furthermore, many kinds of phenotypes detected through such tests (*e.g.*, behavioural, electrophysiological) may require years of additional work to link to their molecular and cellular correlates. It is also the case that knockouts of even highly conserved coding or regulatory sequences frequently result in no detectable abnormality or only minor transcriptional changes [114, 115, 116]). 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.

The recent emergence of single cell molecular profiling technologies (*e.g.*, sc-RNA-seq) offer a potential path to overcome this barrier. As a first step, we and others have extensively applied sc-RNA-seq to profile wildtype mouse development at the scale of the whole embryo [2, 117, 118, 19, 21, 37]. Applying sc-RNA-seq to mouse mutants, several groups have successfully unravelled how specific mutations affect transcriptional networks and lead to altered cell fate decisions in individual organs [119, 120, 121, 122]. However, there is still no clear framework for analysing such data at the scale of the whole embryo, nor for how such data from multiple mutants might be combined to better understand the molecular and cellular basis of classic phenomena like pleiotropy.

Here we set out to establish sc-RNA-seq of whole embryos as a scalable framework for the systematic molecular and cellular phenotyping of mouse genetic models. We profiled 101 embryos of 22 different mouse mutants and 4 wildtype backgrounds at E13.5. The resulting mouse mutant cell atlas (MMCA) includes over 1.6M sc-RNA-seq profiles. To analyze these data, we develop and apply new strategies for detecting differences in composition and/or gene expression across 52 cell types or trajectories spanning the whole mid-gestational embryo.

3.2 Results

3.2.1 Single-cell RNA-seq of 101 mouse embryos

We collected a total of 103 mouse embryos, including 22 different mutants and four wildtype (WT) strains (C57BL/6J, G4, FVB, and BALB/C) at embryonic stage E13.5, and generally four replicates per strain (**Fig. 3.1a**). The mouse mutants were chosen to represent a spectrum of phenotypes ranging from very severe pleiotropic developmental disorders (*e.g.*, *Sox9*, which we expected to affect many organ systems) to knockouts of individual, noncoding regulatory elements (many of which we expected to result in, at best, subtle defects).

We grouped the 22 mutants, all homozygous, into four rough categories (**Supplementary Table 1**): 1) pleiotropic mutants, representing knockouts of developmental genes expressed in multiple organs (*Ttc21b* KO, *Carm1* KO, *Gli2* KO), as well as two mutations of the *Sox9* regulatory landscape suspected to have pleiotropic effects, both of which effectively result in the introduction of a boundary element between endogenous *Sox9* enhancers and the *Sox9* promoter (*Sox9* TAD boundary KI; *Sox9* regulatory INV) [123, 124, 125, 126]. 2) developmental disorder mutants, intended to model specific human diseases (*Scn11a* GOF, *Ror2* KI, *Gorab* KO, *Cdkl5* -/Y) [127, 128, 129], 3) mutations of loci associated with human disease (*Scn10a/Scn11a* DKO, *Atp6v0a2* KO, *Atp6v0a2* R755Q, *Fat1TAD* KO) [130, 131]. 4) prospective deletions of cis-regulatory elements, including of TAD boundaries in the vicinity of developmental transcription factors including *Smad3*, *Twist1*, *Tbx5*, *Neurog2*, *Sim1*, *Smad7*, *Dmrt1*, *Tbx3*, and *Twist1* [132], and, as a positive control, the ZRS distal enhancer (Zone of polarizing activity Regulatory Sequence) which regulates sonic hedgehog (SHH) expression and results in absent distal limb structures [133]).

The 103 flash-frozen embryos (26 genotypes x 4 replicates; one embryo was lost in transport), all staged at E13.5, were sent by five groups to a single site, where they were subjected to sci-RNA-seq3 as previously described [2]. After removing potential doublets, we profiled 1,671,245 nuclei altogether (16,226 +/- 9,289 per embryo; 64,279 +/- 18,530 per strain; median UMI count of 843 per cell and median genes detected of 534 at 75% duplication rate).

Applying principal components analysis (PCA) to “pseudobulk” profiles of the 103 embryos resulted in two roughly clustered groups corresponding to genetic background (**Fig. 3.1b**). In particular, wildtype and mutant FVB embryos clustered separately from C57BL/6J, G4, and BALB/C embryos. However, embryos corresponding to individual mutants did not cluster separately, suggesting that none were affected with severe, global aberrations and highlighting the inadequacy of bulk RNA-seq for detecting mutant-specific effects. A single

outlier embryo (#104) was aberrant with respect to cell recovery ($n = 1,047$) as well as appearance (**Fig. 3.5**).

We next sought to validate the staging of these embryos, leveraging our previous mouse organogenesis cell atlas (MOCA), which spans E9.5 to E13.5 [2]. PCA of pseudobulk profiles of 61 wildtype embryos from MOCA resulted in a first component (PC1) that was strongly correlated with developmental age (**Fig. 3.1c**). Projecting pseudobulk profiles of the 103 MMCA embryos to this embedding resulted in the vast majority of MMCA embryos clustering with E13.5 embryos from MOCA along PC1, consistent with accurate staging. However, five embryos from MMCA appeared closer to E11.5 or E12.5 embryos from MOCA. Four of these were retained as their delay might be explained by their mutant genotype, while one from a wildtype background (C57BL/6; #41) was designated as 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. 3.1d**).

To facilitate an integrated analysis, we sought to project cells from all genotypes to a wildtype derived “reference embedding” (**Fig. 3.6; Methods**). We first applied principal components (PC) dimensionality reduction to cells from wildtype genotypes only ($n = 215,575$; 13.2% of dataset). We then projected cells from mutant genotypes to this embedding, followed by alignment on the combined data to mitigate the effects of technical factors. Next, we applied the UMAP algorithm to the aligned principal components of wildtype cells, followed by Louvain clustering and manual annotation of the resulting major trajectories and sub-trajectories based on marker gene expression. Finally, we projected mutant cells into this UMAP space and assigned them major trajectory and sub-trajectory labels via a k -nearest neighbour (k -NN) heuristic.

Altogether, we identified 13 major trajectories, 8 of which could be further stratified

into 59 sub-trajectories (**Fig. 3.1e**; **Fig. 3.7**; **Supplementary Table 2**). These were generally consistent with our annotations of MOCA, albeit with some corrections as we have described elsewhere [37, 2], as well as greater granularity for some cell types that is likely a consequence of the deeper sampling of E13.5 cells in these new data (**Fig. 3.1f**; **Fig. 3.8**). For example, what we had previously annotated as the excitatory neuron trajectory could be further stratified into a di/mesencephalon (*Slc17a6+*, *Barhl1+*, *Shox2+*), thalamus (*Ntnng1+*, *Gbx2+*) and spinal cord (*Ebf1+*, *Ebf3+*) sub-trajectories, while skeletal muscle could be further stratified into myoblast (*Pax7+*) and myotube (*Myh3+*, *Myog+*) sub-trajectories.

3.2.2 Mutant-specific differences in cell type composition

Analogous to how there are many assays for phenotyping a mouse, there are many computational strategies that one might adopt in order to investigate mutant-specific differences in these embryo-scale sc-RNA-seq data. Here we pursued three main 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 the 13 major trajectories across the 4 wildtype and 22 mutant strains. For the most part, these proportions were consistent across genotypes (**Fig. 3.9a**). However, some mutants exhibited substantial differences. For example, compared to the C57BL/6 wildtype, the proportion of cells falling in the neural tube trajectory decreased from 37.3% to 33.7% and 32.6% in the *Gli2* KO and *Ttc21b* KO mice, respectively, while the proportion of cells falling in the mesenchymal trajectory decreased from 44.1% to 37.1% in the *Gorab* KO mice. These changes are broadly consistent with the gross pheno-

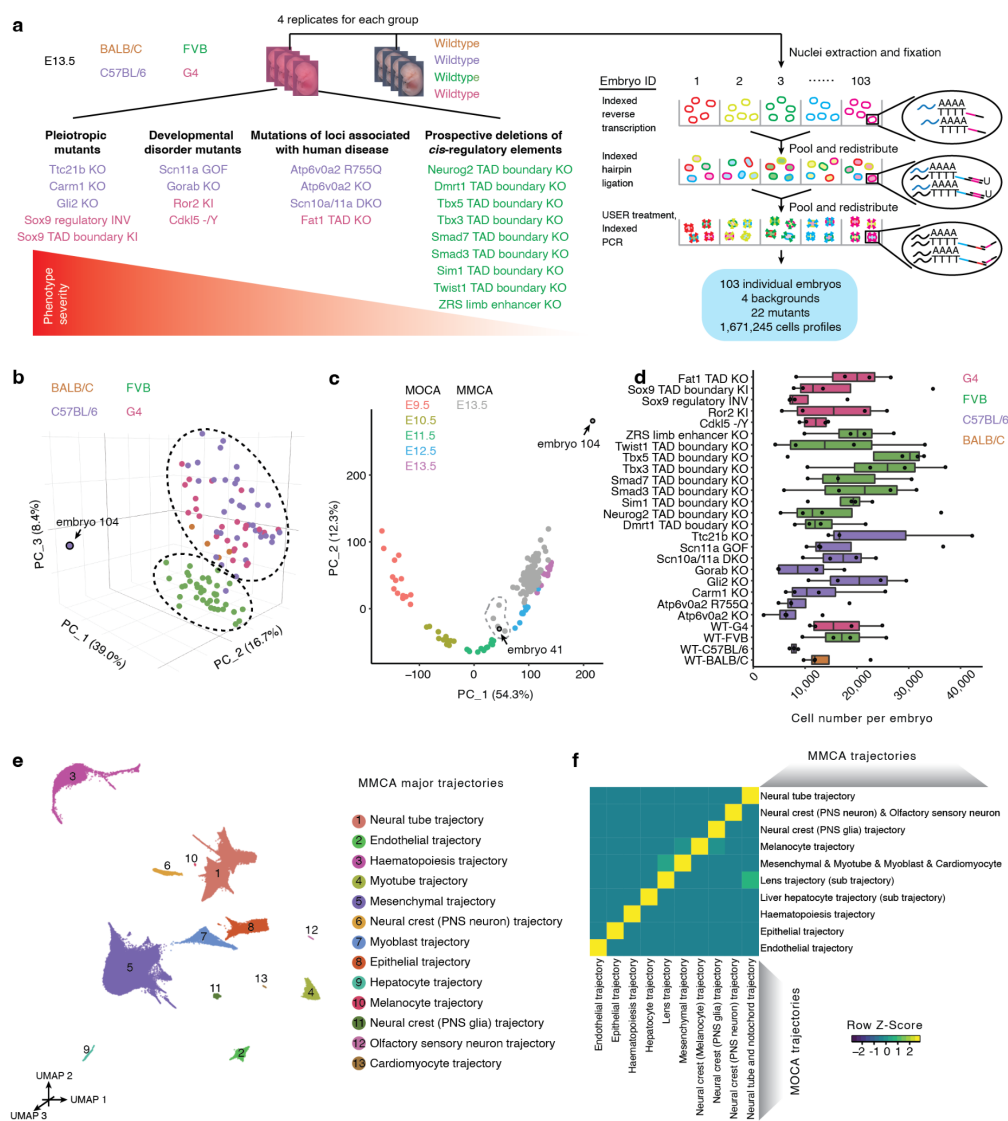


Figure 3.1: Figure 1. Single-cell RNA-seq of 103 whole mouse embryos staged at E13.5. **a**, We applied sci-RNA-seq3 to profile 1.6M single cell transcriptomes from 103 individual E13.5 embryos, derived from 22 mutants and four wildtype strains, in one experiment. **b**, Embeddings of pseudobulk RNA-seq profiles of MMCA mouse embryos in PCA space with visualisation of top three PCs. **c**, Embeddings of pseudobulk RNA-seq profiles of MOCA and MMCA mouse embryos in PCA space defined solely by MOCA, with MMCA embryos (gray) projected onto it. The dashed line highlights five MMCA embryos which are colocalized with E11.5 or E12.5 embryos from MOCA: *Scn11a* GOF (#33, #34, #36), *Carm1* KO (#101), and C57BL/6 wildtype (#41). **d**, The number of cells profiled per embryo for each strain. **e**, 3D UMAP visualisation of wildtype subset of MMCA dataset (215,575 cells from 15 wildtype E13.5 embryos). Cells are colored by major trajectory annotation. **f**, Heatmap of the combined non-negative least-squares (NNLS) regression coefficients (row-scaled) between developmental trajectories from MMCA (rows) and MOCA (columns).

types associated with these mutations [134, 135, 124, 129], but are caveated by substantial interindividual heterogeneity within each genotype (**Fig. 3.9b**). Also of note, we observe differences in major trajectory composition between the four wildtype strains. For example, relative to BALB/C and C57BL/6, the FVB and G4 wildtype mice consistently had substantially lower proportions of cells in the mesenchymal trajectory and higher proportions of cells in the neural tube trajectory (**Fig. 3.9c**).

To increase resolution, we sought to investigate compositional differences at the level of sub-trajectories. For each combination of background (C57BL/6, FVB, G4) and sub-trajectory ($n = 54$), we performed a regression analysis to identify instances where a particular mutation was nominally predictive of the proportion of cells falling in that sub-trajectory (uncorrected p -value < 0.05 ; beta-binomial regression; **Methods**). Across the 22 mutants, this analysis highlighted 300 nominally significant changes (**Fig. 3.2a**; **Supplementary Table 3**). Due to the limited number of replicate embryos per wildtype and mutant strain, our power to detect changes is limited, particularly in the smaller trajectories. Nevertheless, several patterns were clear:

First, it is evident that *Atp6v0a2* KO and *Atp6v0a2* R755Q, two distinct mutants of the same gene [130], are assigned very similar patterns by this analysis, both with respect to which sub-trajectories are nominally significant as well as the direction and magnitude of changes (first two rows of **Fig. 3.2a**). Although perhaps expected, the consistency supports the validity of this analytical approach.

Second, the mutants varied considerably with respect to the number of sub-trajectories that were nominally significant for compositional differences. At the higher extreme, the proportions of cells falling in 30 of 54 sub-trajectories were nominally altered by the *Sox9* regulatory INV mutation, consistent with the wide-ranging roles of Sox9 in development [136, 137]. On the other hand, other mutants, such as the TAD boundary knockouts, exhibited

comparatively few changes, consistent with the paucity of gross phenotypes in such mutants [116]. Nonetheless, all TAD boundary knockouts did show some changes, including specific ones, *e.g.* the lung epithelial and liver hepatocyte trajectories were decreased in the *Dmrt1* and *Tbx3* TAD boundary KOs, respectively, but not in other TAD boundary knockouts. At the lower extreme, the *Sim1* TAD boundary KO exhibited just two altered sub-trajectories.

Third, some sub-trajectories exhibited altered proportions in many mutants (*e.g.* the mesencephalon/MHB trajectory in 12 mutants) while others were changed only in a few (*e.g.* the definitive erythroid trajectory in *Ror2* KI only). In some cases, such patterns were “block-like” by background strain (*e.g.* all B6 mutants exhibited gains in endothelial cells and losses in endocardium). Although particular sub-trajectories might be vulnerable to disruption in a strain-specific way, it is also possible that this is a technical artefact (*e.g.* if the four wildtype replicates that we profiled for a given strain were atypical).

There were a few extreme examples, *e.g.* where a sub-trajectory appeared to be fully lost in a specific mutant. For example, *Ttc21b*, which encodes a ciliary protein and whose knockout is associated with brain, bone and eye phenotypes [138, 139, 124], exhibited a dramatic reduction in the proportion of cells in the retinal neuron trajectory ($\log_2(\text{ratio}) = -6.69$; unadjusted p -value = 0.028; beta-binomial regression) (**Fig. 3.2b**), as well as the lens ($\log_2(\text{ratio}) = -2.64$) and retina epithelium ($\log_2(\text{ratio}) = -2.32$) trajectories (**Fig. 3.10**). Validating this finding, the developing eye appears diminished in the homozygous *Ttc21b* mutant at E11.5 embryos compared to the wildtype or heterozygous mutant (**Fig. 3.2c**).

However, most changes were relatively subtle. For example, the ZRS limb enhancer KO is a well-studied mutant which shows a loss of the distal limb structure at birth [133]. This analytical framework highlighted eight sub-trajectories whose proportions were nominally altered in the ZRS limb enhancer KO, most of which were mesenchymal. However, although the most extreme, the reduction in limb mesenchymal cells was only about 30% ($\log_2(\text{ratio})$

= -0.49; unadjusted p -value = 6.32e-3; beta-binomial regression). To assess whether further subpopulations of the limb mesenchyme were more substantially changed, we performed co-embedding of limb mesenchyme cells from the ZRS limb enhancer KO and the FVB wildtype. Indeed, a subpopulation of the limb mesenchyme was much more markedly affected (**Fig. 3.2d**; **Fig. 3.11a**), and this subpopulation specifically expressed markers of the distal mesenchyme of the early embryonic limb bud, such as *Hoxa13* and *Hoxd13* (**Fig. 3.2e**) [140]. Of note, we did not observe such heterogeneity when we examined the seven other sub-trajectories whose proportions were nominally altered in the ZRS limb enhancer KO (**Fig. 3.11b**), consistent with the specificity of this phenotype.

3.2.3 *LochNESS analysis reveals differences in transcriptional state within cell type trajectories*

Given that most of the mutants that we studied did not exhibit macroscopic anatomical defects or otherwise severe phenotypes at E13.5, we next sought to develop a more sensitive approach for detecting deviations in transcriptional programs within cell type trajectories. Specifically, we developed “lochNESS” (local cellular heuristic Neighbourhood Enrichment Specificity Score), score that is calculated based on the “neighbourhood” of each cell in a sub-trajectory co-embedding of a given mutant (all replicates) vs. a pooled wildtype (all replicates of all backgrounds) (**Fig. 3.3a**; **Methods**; although developed independently, this approach is similar to recent work by Dann and colleagues [28]). Briefly, we took the aligned PC features of each sub-trajectory, as described above, and found k -NNs for each cell, excluding cells from the same mutant replicate from consideration. For each mutant cell, we then computed the ratio of the observed vs. expected number of mutant cells in its neighbourhood, with expectation simply based on the overall ratio of mutant vs. wildtype cells in co-embedding. In the scenario where mutant and wildtype cells are fully mixed, the resulting ratio should be close to 1. The final lochNESS was defined as the ratio minus 1, equivalent to the fold change of mutant cell composition.

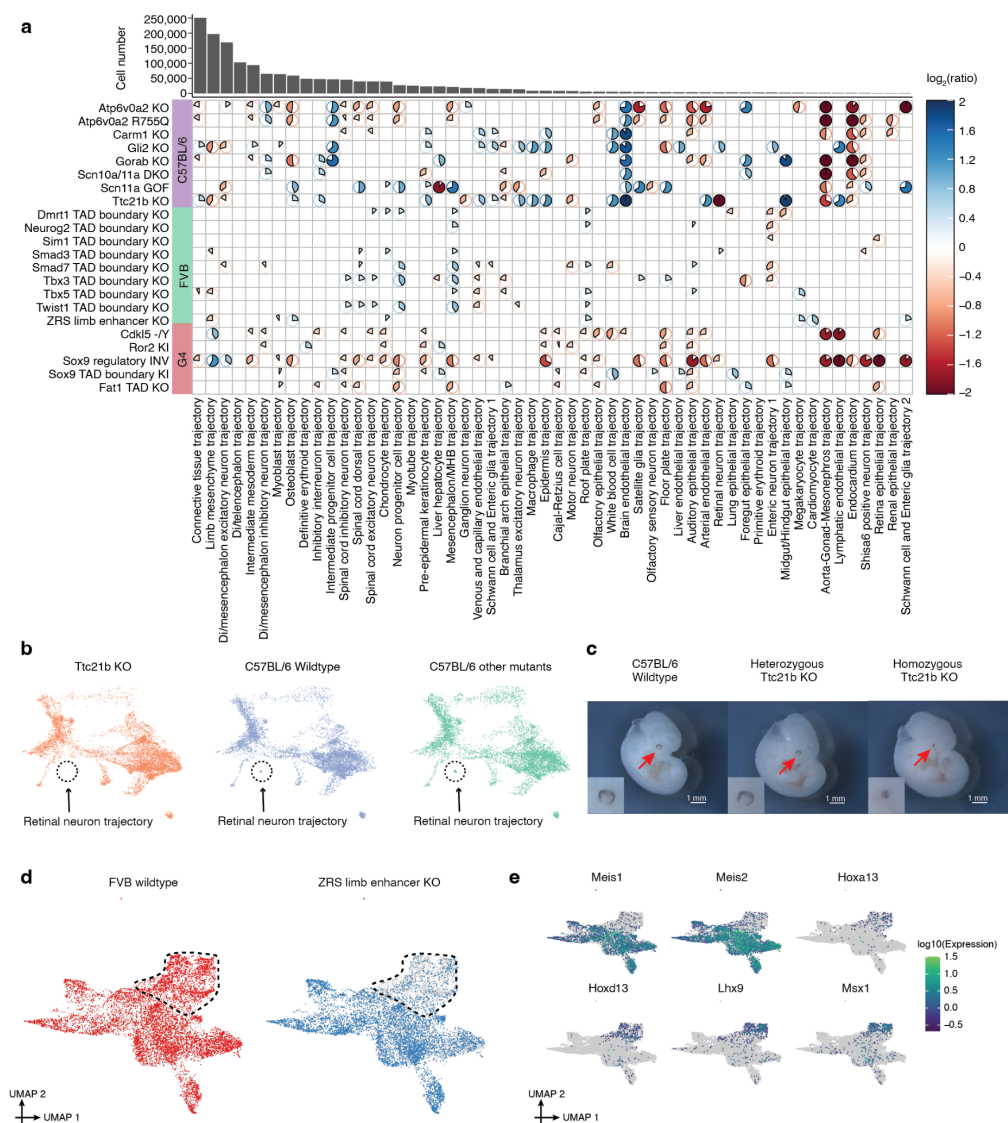


Figure 3.2: Figure 2. Cell composition changes for individual mutants across developmental trajectories. **a**, Heatmap shows \log_2 transformed ratios of the cell proportions between each mutant type (y -axis) and its corresponding wildtype background, across individual sub-trajectories (x -axis). **b**, 3D UMAP visualisation of the neural tube trajectory, highlighting cells from either the *Ttc21b* KO (left), C57BL/6 wildtype (middle), or other mutants on the C57BL/6 background (right). The three plots were randomly downsampled to the same number of cells ($n = 8,749$ cells). **c**, Homozygous *Ttc21b* KO mice embryo (E11.5) showed abnormal eye development. **d**, UMAP visualisations of co-embedded cells of limb mesenchyme trajectory from the ZRS limb enhancer KO and FVB wildtype, highlighting cells from either FVB wildtype (left) or ZRS limb enhancer KO (right). The subset of cells in this co-embedding exhibiting more extreme loss in the ZRS limb enhancer KO is highlighted. **e**, The same UMAP as in panel **d**, colored by gene expression of marker genes which appear specific to proximal limb development (*Meis1*, *Meis2*) [141, 142]) and distal limb development (*Hoxa13*, *Hoxd13*, *Lhx9*, *Mx1*) [140, 143, 144].

Visualisation of lochNESS in the embedded space highlights areas with enrichment or depletion of mutant cells. For example, returning to the previously discussed ZRS limb enhancer KO mice, we observed markedly low lochNESS in a portion of the limb mesenchymal trajectory corresponding to the distal limb (**Fig. 3.3b**; **Fig. 3.2d**). This highlights the value of the lochNESS framework, as within the sub-trajectory (limb mesenchyme), an effect could be detected and also assigned to a subset of cells in a label-agnostic fashion.

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 (**Fig. 3.3c**). 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. Of note, we confirmed that repeating the calculation of lochNESS after random permutation of mutant and wildtype labels resulted in bell-shaped distributions centred around zero (**Fig. 3.12a**). As such, the deviance of lochNESS can be summarised as the average euclidean distance between lochNESS vs lochNESS under permutation (**Fig. 3.12b**).

We next examined lochNESS within each mutant of each sub-trajectory to identify system-specific phenotypes. For example, consistent with results shown above, we observed low lochNESS within the retinal neuron sub-trajectory in the *Ttc21b* KO (**Fig. 3.3d**; **Fig. 3.12c**). We also observed a strong shift towards low scores for the floor plate sub-trajectory in the *Gli2* KO, and interestingly, a more subtle change in lochNESS distribution for the roof plate trajectory, which is forming opposite to the floor plate along the D-V axis of the developing neural tube (**Fig. 3.3d**; **Fig. 3.12c**). To explore this further, we extracted and reanalyzed cells corresponding to the floor plate and roof plate. Within the floor plate, *Gli2* KO cells consistently exhibited low lochNESS (**Fig. 3.3e**). However, there were only a handful of differentially expressed genes between wildtype and mutant cells, and no significantly enriched pathways within that set. For example, genes like *Robo1* and *Slit1*, both involved

in neuronal axon guidance, are specifically expressed in the floor plate relative to the roof plate (**Fig. 3.3f**; **Fig. 3.12g**), 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 and the previous demonstration that *Gli2* knockouts fail to induce a floor plate [145]).

Less expectedly, this focused analysis also revealed two subpopulations of roof plate cells, one depleted and the other enriched for *Gli2* KO cells (**Fig. 3.3e**; **Fig. 3.12d-f**). To annotate these subpopulations, we examined genes whose expression was predictive of lochNESS via regression (**Methods**). The mutant-enriched group of roof plate cells was marked by *Ttr*, a marker for choroid plexus and dorsal roof plate development [146], as well as genes associated with the development of cilia (*e.g.* *Cdc20b*, *Gmnc*, *Dnah6* and *Cfap43*), while the mutant-depleted group was marked by Wnt signaling-related genes including *Rspo1/2/3* and *Wnt3a/8b/9a* (**Fig. 3.3f**; **Fig. 3.12g**; **Supplementary Table 3** [147, 148, 149, 150]. It has been shown that ventrally-expressed *Gli2* plays a central role in dorsal-ventral patterning of the neural tube by antagonising Wnt/Bmp signalling from the dorsally-located roof plate [145]. Our results are consistent with this, and also define two subpopulations of roof plate cells on which *Gli2* KO appears to have differential effects. 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 the lochNESS strategy.

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 (**Fig. 3.13a**). **Such**

deviations, summarised as the average euclidean distances between lochNESS and lochNESS under permutation, are visualised in **Fig. 3.13b**. 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 (**Fig. 3.13c-d; Supplementary Table 4**) [151, 152]. 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* signalling genes that regulate *Tbx3* during lung development (*Bmp1/4*), and distal airway markers *Sox9* and *Id2* [153, 152, 154].

3.2.4 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 [155]. A corollary of pleiotropy is that there are also specific traits that appear to be influenced by multiple, unrelated genes. For the most part, the characterization of the sharing of phenotypic features between multiple Mendelian disorders has remained coarse. For example, many disorders share macrocephaly as a feature, but it remains largely unexplored whether the molecular and cellular basis for macrocephaly is shared between them, unique to each, or somewhere in between.

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 strain, we computed k -NNs as in **Fig. 3.3a**, and then calculated the observed vs. expected ratio of each genotype among a cell’s k -NNs. The “similarity

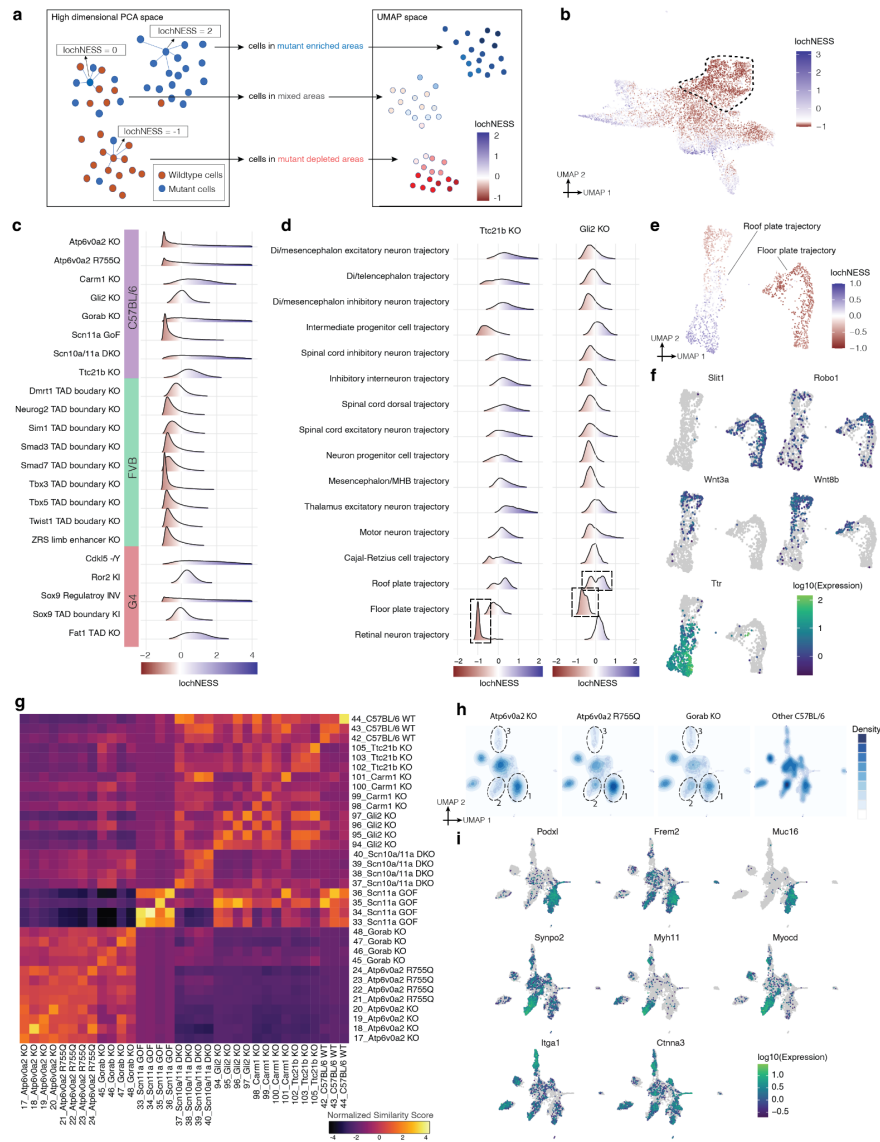


Figure 3.3: Figure 3. LochNESS analysis identifies mutant related changes.

a, Schematic of lochNESS calculation and visualisation. **b**, UMAP visualisation of co-embedded cells of limb mesenchyme trajectory from the ZRS limb enhancer KO and FVB wildtype, colored by lochNESS. Dashed lines highlight same area as in Fig. 3.2d. **c**, Distribution of lochNESS across all cells in each mutant. **d**, Distribution of lochNESS in the neural tube sub-trajectories of the *Ttc21b* KO and *Gli2* KO mutants. Dashed boxes highlight the sub-trajectories with shifted distributions. **e**, UMAP visualisation of co-embedded cells of the floor plate and roof plate sub-trajectories from the *Gli2* KO mutant and pooled wildtype, colored by lochNESS. **f**, same as in panel e, but colored by expression of selected marker genes. **g**, Heatmap showing similarity scores between individual C57BL/6 embryos in the mesenchymal trajectory. **h**, UMAPs showing the co-embedding of the intermediate mesoderm sub-trajectory for mutants from the C57BL/6 background strain. Dashed circles highlight clusters of cells where *Atp6v0a2* KO, *Atp6v0a2* R755Q and *Gorab* KO mice exhibit enrichment (cluster 1) or depletion (clusters 2 & 3). **i**, same as in panel h, but colored by expression of marker genes of the clusters highlighted in panel h.

score” between one genotype vs. all others is defined as the mean of these ratios across cells of the genotype. 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. 3.3g**; **Fig. 3.14a-b**).

The *Scn11a* GOF mutant exhibited the most extreme similarity scores, in terms of both similarity between replicates and dissimilarity with other genotypes (**Fig. 3.3g**; **Fig. 3.14a**). 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 [127]. 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. 3.1c**), we speculated that its extreme similarity scores might be attributable to developmental delay of the *Scn11a* GOF mutant at the scale of the whole embryo. 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 (**Fig. 3.14d**). 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; **Fig. 3.14e**). As such, the apparently unique signature of *Scn11a* GOF cells might be attributable to these embryos simply being earlier in development, suggesting a more global role for sodium ion channels not only for neuronal function but also early development and cell fate determination [156]).

Incorrect staging is formally possible, but unlikely because the embryos derived from three independent litters.

In sharp contrast with the relative uniqueness of the *Scn11a* GOF mutant, we also observed that the similarity scores between three mutants – *Atp6v0a2* KO, *Atp6v0a2* R755Q and *Gorab* KO – was 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. 3.3g; Fig. 3.14a,c,f**). Such sharing is perhaps expected between the *Atp6v0a2* KO and *Atp6v0a2* R755Q mutants, as they involve the same gene. 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 [130, 129]. However, only the ATP6V0A2-related disorder displays a prominent CNS phenotype, consistent with the changes in the neural tube trajectory seen only in both *Atp6v0a2* models (**Fig. 3.14a,c,f**).

In order to explore phenotypic sharing between these genotypically distinct mutants at greater granularity, we co-embedded cells of the intermediate mesoderm sub-trajectory from C57BL/6 strains. We identified three subclusters of intermediate mesoderm where *Atp6v0a2* KO, *Atp6v0a2* R755Q and *Gorab* KO mice are similarly distributed compared to other C57BL/6 genotypes (**Fig. 3.3h,i**). In particular, cluster 1 is enriched for cells from *Atp6v0a2* KO, *Atp6v0a2* R755Q and *Gorab* KO mice and is marked by genes related to epithelial-to-mesenchymal transition, cell-cell adhesion and migration, such as *Podxl*, *Frem2* and *Muc16* [157, 158, 159]. Clusters 2 and 3 are depleted in cells from *Atp6v0a2* KO, *Atp6v0a2* R755Q and *Gorab* KO mice and are marked by muscular development related genes like *Synpo2*, *Myh11* and *Myocd* (cluster 2), and cell-cell adhesion related genes like *Itga1* and *Ctnna3* (cluster 3) [160, 161, 162, 163, 164].

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

3.2.5 Global developmental defects in *Sox9* regulatory mutant

About half of the mutants profiled in this study 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 the mesenchymal trajectory (**Fig. 3.3c; Fig. 3.4a**). **The *Sox9* locus encodes a pleiotropic transcription factor that plays a central role during the** development of the skeleton, the brain, in sex determination as well as several other tissues during embryogenesis, orchestrated by a complex regulatory landscape [165, 166, 167, 168, 169, 170, 171, 172, 173]. This particular mutant features an inversion of a 1Mb region upstream of *Sox9* that includes several distal enhancers and a TAD boundary, essentially relocating these elements into a TAD with *Kcnj2*, which encodes a potassium channel (**Fig. 3.4b**) [123, 174, 175]. Consistent with the heterozygous and homozygous *Sox9* knockout, the homozygous *Sox9* regulatory INV is perinatally lethal, with extensive 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 was previously shown to lead to pronounced misexpression of *Kcnj2* in the digit anlagen in a wildtype *Sox9* pattern [123]. 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 lochNESS distribution for *Sox9* regulatory INV were consistently observed, but the limb mesenchyme and connective tis-

sue were particularly enriched for cells with extremely high lochNESS (**Fig. 3.4a**, right). Of relevance, 2 of the 3 major enhancers (E250 and E195) known to drive *Sox9*-mediated chondrogenesis in mesenchymal stem cells are located within the inverted region (**Fig. 3.4b**) [167]. Cell type composition analysis (**Fig. 3.2a**) showed that *Sox9* regulatory INV mutants harbor considerably larger numbers of cells classified as limb mesenchyme, at the expense of osteoblasts, intermediate mesoderm, chondrocytes and connective tissue trajectory. This shift can also be seen in a UMAP embedding (**Fig. 3.4c**), a topic that 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; **Fig. 3.15a**), 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. 3.4d**; **Fig. 3.15b**).

Since the inverted *Sox9* regulatory region also hosts multiple enhancers active in other tissues (*e.g.* E161-lung; E239-cerebral cortex) [167], we wondered whether these patterns were also seen in other tissues. Indeed, both sc-RNA-seq expression analysis 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 in our single cell data, RNAscope quantification showed reductions in *Sox9* expression in the telencephalon and lung as well (**Fig. 3.15**). Taken together, these data suggest marked changes in mesenchyme

due to reductions of *Sox9* expression (presumably due to separation from key enhancers), together with broader increases in *Kcnj2* expression (presumably due to the appropriation of *Sox9* enhancers).

To explore the apparent effects of the *Sox9* regulatory inversion on mesenchyme in more detail, in particular the apparent accumulation of limb mesenchyme (**Fig. 3.4c**), we reanalyzed mutant and wildtype cells from the limb mesenchyme sub-trajectory on their own, which revealed subsets corresponding to condensing mesenchyme, perichondrium, and undifferentiated mesenchyme (**Fig. ??a,b**). This analysis further revealed that the vast majority of limb mesenchyme “accumulation” in mutant embryos was due to a large proportion of cells that appear delayed or stalled in an undifferentiated or stem-like state, rather than an accumulation of more advanced limb mesenchyme (**Fig. 3.4c**, bottom panels; **Fig. ??a**). Of note, because the annotation of “limb mesenchyme” for this sub-trajectory was propagated forward from earlier stages of development during the creation of MOCA, we cannot rule out that other, non-limb mesenchymal populations contribute to this expanded, undifferentiated pool in the *Sox9* regulatory INV embryos as well.

Inspection of density plots and RNA velocity suggested that wildtype undifferentiated mesenchymal cells (a subset of cells annotated as limb mesenchyme in **Fig. 3.4c**) are poised to undergo differentiation into diverse subtypes (**Fig. 3.4c**; **Fig. ??a**). In sharp contrast, undifferentiated mesenchymal cells from *Sox9* regulatory INV embryos accumulate at the “source” of differentiation, and also appear to acquire a distinct state (high density region in bottom right sub-panel of **Fig. 3.4c**). This accumulation is even more apparent in integrated views of the limb mesenchyme sub-trajectory, where we observe two distinct branches, each heavily enriched for *Sox9* regulatory INV mutant cells, within undifferentiated mesenchyme (**Fig. 3.4e**; **Fig. 3.16a**).

To investigate these two branches further, we performed sub-clustering of *Sox9* regulatory

INV undifferentiated mesenchyme cells, followed by differential expression analysis (**Fig. 3.4f,g**). Interestingly, the most differentially expressed genes in “branch 1” were neuronal, *e.g.* several neurexins and neuregulin 3, an observation that was supported by single-sample gene set enrichment analysis (ssGSEA) [176], which further highlighted KRAS and other signalling pathways (**Fig. 3.4g; Fig. 3.16b,c**). Of note, mesenchymal stem cells can be differentiated to neuronal states *in vitro* [177]. Although further investigation is necessary, we note that cells contributing to “branch 0” as well as the *neuronal-trending* “branch 1” are present in wildtype embryos, albeit at much reduced frequencies compared to the *Sox9* regulatory INV mutant (**Fig. 3.16a, left**).

In sum, consistent with what is known about the role of *Sox9* as a driver gene in cartilage and skeletal development, our data reveals a redirection in the differentiation of osteoblast, chondrocytes and other derivatives of the undifferentiated mesenchyme in the *Sox9* regulatory INV mutant. Among mutants on the G4 background, the observed pattern is specific to the *Sox9* regulatory INV mutant (**Fig. 3.17**). Remarkably however, when we examine the entire dataset (**Fig. 3.17-3.19**), we observe a similar accumulation of undifferentiated mesenchymal cells in the *Atp6v0a2* KO, *Atp6v0a2* R755Q, and *Gorab* KO mutants, indicating sharing of this sub-phenotype amongst 4 of 22 mutants examined (**Fig. 3.19-3.20**). This observation further illustrates the potential for systematic, whole embryo analysis to reveal sharing of molecular and cellular sub-phenotypes across pleiotropic developmental mutants in unexpected ways.

3.3 Discussion

In this study, we set out to establish whole embryo sc-RNA-seq as a new paradigm for the systematic, scalable phenotyping of mouse developmental mutants. In one experiment, we generated ~1.6M single cell transcriptomes from just over 100 E13.5 embryos corresponding to 22 mutant genotypes and 4 wildtype strains. To investigate the resulting dataset, we

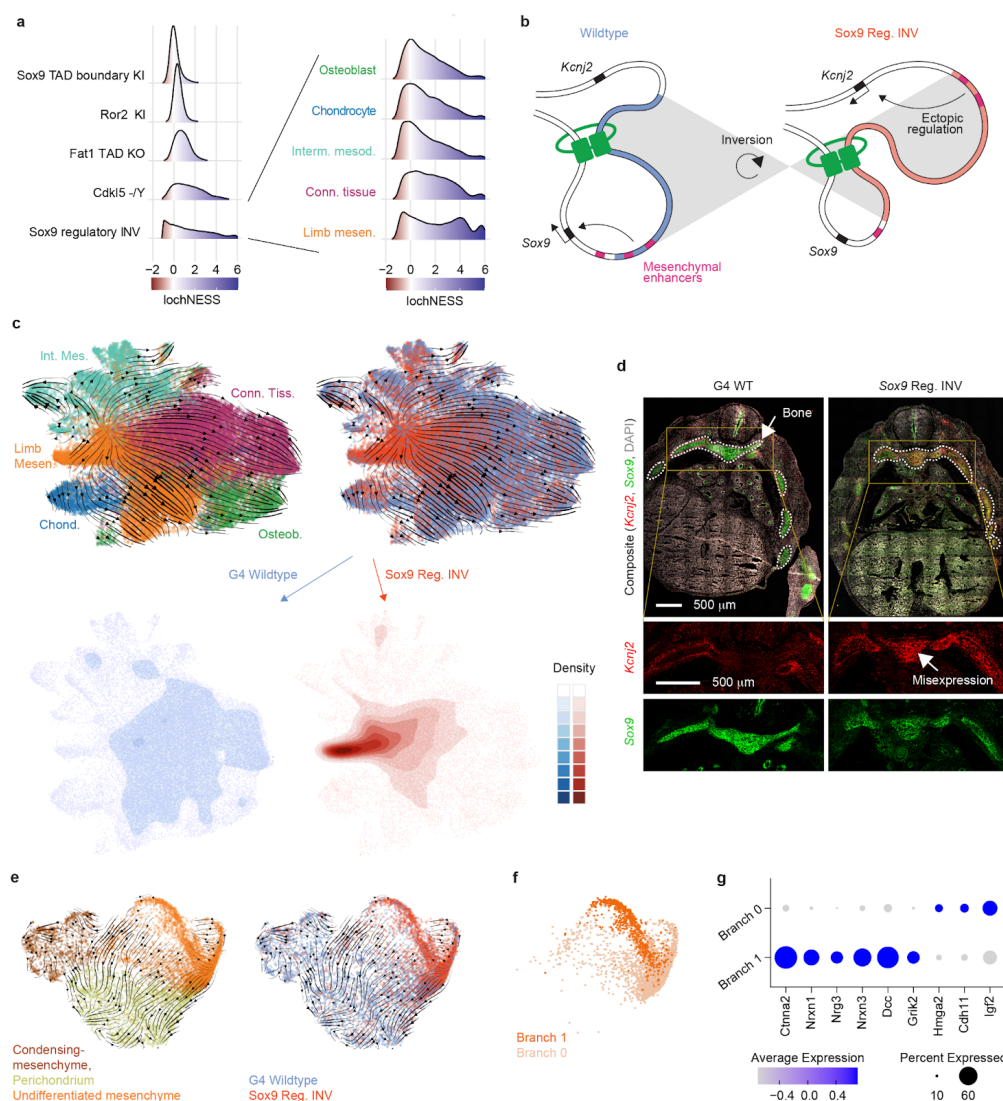


Figure 3.4: Figure 4. Apparent stalling and redirection of mesenchyme differentiation in the *Sox9* regulatory INV mutant. **a**, LochNESS distributions for all G4 mutants in the mesenchymal trajectory (left) and the *Sox9* regulatory INV mutant in mesenchymal sub-trajectories (right). **b**, Model of *Sox9* regulatory INV mutation depicting ectopic *Kcnj2* expression due to adoption of chondrogenesis and osteogenesis specific enhancers. **c**, top: RNA velocity on UMAP embedding of mesenchymal G4 wildtype and *Sox9* regulatory INV cells labelled by annotation (left) or sample (right). bottom: 2D density plots of the same UMAP embedding for G4 wildtype (left) and *Sox9* regulatory INV cells (right). **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 on UMAP embedding of G4 wildtype and *Sox9* regulatory INV cells in the limb mesenchymal trajectory labelled by annotation (left) or sample (right). **f**, UMAP embedding of *Sox9* regulatory INV cells in the undifferentiated mesenchyme, visualised in the same embedding as in panel **e**. **g**, Dot plot of the top six (where available) significantly differentially expressed genes between the two branches.

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”). We also evaluated how a range of gross phenotypic severities manifest at the molecular and cellular levels, and show how global analysis can in some cases reveal molecular and cellular phenotypes that may be missed by conventional phenotyping. Such “*in silico* developmental biology”, wherein global profiles of developmental mutants are subjected to systematic, outcome-agnostic computational analyses, may complement and guide conventional phenotyping, which can be impractical to scale to all physiological systems even for a single mutant.

We emphasise that the concurrent analysis of many mutants proved essential to the contextualization of particular observations, *i.e.* to understand how specific or non-specific any apparent deviation really was, against a background of dozens of genotypes and over 100 embryos. This aspect of the study 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” (*e.g.* the stalling of undifferentiated mesenchyme in 4 of 22 mutants examined).

Our mouse mutant cell atlas (MMCA) has limitations. First, we only profiled 4 replicates per mutant at a single developmental time point. We can’t exclude that some subtle effects were missed that might have been captured through profiling of a larger number of replicate embryos. Second, we profiled only ~15,000 cells per embryo, which is only a small fraction of the millions of cells that are present in E13.5 embryos, which may also have limited sensitivity. A counterweight to these limitations is that 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 even rare cell types, *e.g.* in the retina (*Ttc21b* KO) and roof plate (*Gli2* KO).

Third, although we performed more detailed *in silico* analyses of selected mutants and phenotypes, 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). Even for these 22 mutants, but also looking to the future, we anticipate the community input and domain expertise will be essential to extract full value from these data, including the development of additional analytical strategies. 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/).

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 [178]. Although over 7,000 lines have already been analysed, thousands more still await phenotyping, and even what phenotyping has been done is not necessarily comprehensive [179]. 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, to advance our understanding of the molecular and cellular basis of human developmental disorders, to decompose pleiotropy, and to shed light on the function(s) of mammalian genes.

3.4 Materials and methods

Data reporting

No statistical methods were used to predetermine sample size. Embryos used in experiments were randomised 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 [180]. 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 [2]. 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 μ m) 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 500 μ l NSB and split into 2 tubes, each containing 250 μ l 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.

Library preparation and sequencing

The library preparation was performed previously described [2, 181]. In short, 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 analysed 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 [2] with the following 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* [182] with default settings. Downstream sequence processing and cell-x-gene expression count matrix generation were similar to sci-RNA-seq [20] 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* [183] 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 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 [184]. 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 [89] 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* [33] 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 neighbourhood graph of observations (k.param = 50) followed by clustering the cells into sub-groups using the Louvain algorithm. For UMAP visualisation, 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* [2]. 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 visualisation 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 [2], 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 normalised by dividing its estimated size factor, and then the data were log2-transformed after adding a pseudocount followed by performing the PCA. The normalisation 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 normalised 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. 3.1e**). 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 neighbour (k -NN) heuristic. Specifically, for each mutant-derived cell, we identified its 15 nearest neighbour wildtype-derived cells in UMAP space and then assigned the major trajectory with the maximum frequency within that set of 15 neighbours 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 normalised 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 normalised 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, *i.e.* 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 normalised 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) \sim genotype”. Of note, embryos from the four different mouse strain backgrounds were analysed independently.

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 neighbourhood. 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 neighbourhood 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. 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.

Identifying lochNESS associated gene expression changes

To identify gene expression changes associated with mutant enriched or depleted areas, we find differentially expressed genes through fitting a regression model for each gene accounting for lochNESS. We use the *fit_models()* function implemented in *monocle/v3* with lochNESS

as the *model_formula_str*. This essentially fits a generalized linear model for each gene: $\log(y_i) = \beta_0 + \beta_n * x_n$, where y_i is the gene expression of $gene_i$, β_n captures the effect of the lochNESS x_n on expression of $gene_i$ and β_0 is the intercept. For each $gene_i$, we test if β_i is significantly different from zero using a Wald test and after testing all genes, we adjust the p-values using the Benjamini and Hochberg procedure to account for multiple hypotheses testing. We identify the genes that have adjusted p-value < 0.05 and large positive β_i values as associated with mutant enriched areas, and those with large negative β_i values as associated with mutant depleted areas.

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:

$$similarityscore_{cell_n, embryo_j} = \frac{\#of\ cells\ from\ embryo_i\ in\ k\ NN\ of\ cell_n}{k} / \frac{\#of\ cells\ from\ embryo_j\ in\ dataset}{N}$$

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 $embryo_i$. The embryo similarity score matrix can be visualised in a square heatmap where rows and/or columns are hierarchically clustered.

Identifying and quantifying developmental delay

To identify potential mutant related developmental delay, we integrate MMCA with MOCA. We consider a mutant and its corresponding wildtype in a sub trajectory as a dataset. We take the cells from E11.5-E13.5 with similar annotations from MOCA and co-embed with the MMCA cells. We take the raw counts from both datasets, normalise, and

process the data together without explicit batch correction as both datasets were generated with sci-RNA-seq3 and were similar in dataset quality. We visualise the co-embedded data in 3D UMAP space and check for developmental delay in the mutant cells (*i.e.* mutant cells embedded closer to early MOCA cells compared to wildtype cells). To quantify the amount of developmental delay, we find k -NNs in MOCA for each cell in MMCA and calculate $timescore = \frac{\sum_{n=1}^k T_n}{k}$, where T_n is the developmental time of MOCA $cell_n$ in the k -NNs of the MMCA cell. Afterwards, we test if the average time scores of mutant cells are significantly different from that of wildtype cells using a student's t-test.

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 $5\ \mu\text{m}$ 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 five μm sections of the mouse embryos. RNAscope probes were purchased by ACD and designed as described by [185]. 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 labelling 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 analysed using the image processing software Fiji [80]. Each organ of interest mRNA signal was counted in a defined area (1 x 1 mm²) with an n=6 per condition. Statistics were calculated using student t-Test and evaluated (- p > 0,05 = non-significant, p < 0,05 - ≥ 0,01 = * , p < 0,01 - ≥ 0,001= ** - p < 0,001= ***).

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-normalised after which PCA was performed with default parameters on top 2000 highly variable genes selected using the “vst” method. Nearest neighbours 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 (**Fig. ??b**). 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 visualization and RNA velocity analysis

Using *Seurat/v4.0.6*, the raw counts were log-normalised, 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 normalised using *scv.pp.filter_and_normalize*, with *min_shared_counts*=20 and *n_top_genes*=2000. Means and variances between 30 nearest neighbours 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 [176]. The *msigdb* 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 [186, 187]. *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. 3.4f**) using the two sample Wilcoxon test (*wilcox.test*) with default parameters and adjusted for multiple comparisons using Bonferroni correction.

3.5 *Supplementary materials*

3.5.1 *Supplementary notes*

Detailed explanation of Fig. 3.2a

In Fig. 3.2a, only sub-trajectories with a mean number of cells across individual embryos of less than ten were excluded from this analysis, leaving 54 (columns). Only those combinations of mutant and sub-trajectory which were nominally significant in the regression analysis are shown (see text and **Methods**; uncorrected p -value < 0.05 ; beta-binomial regression). For calculating the displayed ratios, cell counts from replicates were merged. The pie color and direction correspond to whether the log₂ transformed ratio is above 0 (blue, clockwise) or below 0 (red, anticlockwise), while the pie size and colour intensity correspond to the scale of log₂ transformed ratio. A handful of log₂ transformed ratios with > 2 (or < -2) were manually set to 2 (or -2) for a better visualisation. The number of cells assigned to each developmental trajectory in the overall dataset is shown above the heatmap.

3.5.2 *Supplementary figures*



Figure 3.5: **Figure S1. Images of mouse embryos.** 104 embryos (26 genotypes x 4 replicates) were staged at E13.5 and sent by five groups to a single site. #49 was accidentally skipped in our numbering systems. Embryo #70 was lost in transport. Pictures of embryos #1, #5, #9, #13 and #91 were not taken, but the embryos were included in the sci-RNA-seq3 experiment. As discussed in the text, embryos #41 and #104 were labelled as outliers based on computational analyses and their data discarded, while data from the remaining 101 embryos were retained and analysed further. Of note, in addition to the computational analyses suggesting that embryo #104 was an outlier, it was also relatively small in size upon visualisation.

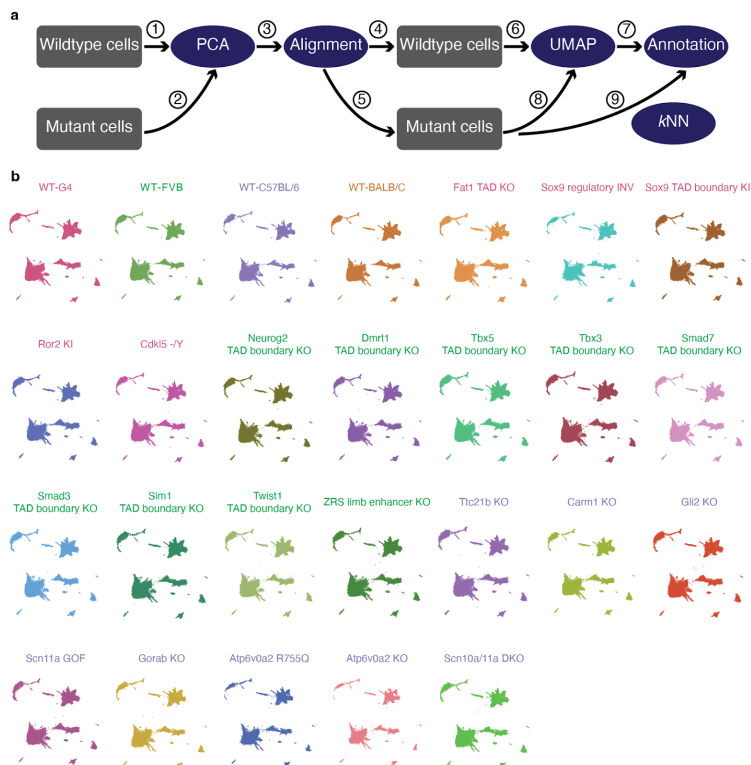


Figure 3.6: **Figure S2. Integrating cells derived from embryos of multiple genetic backgrounds to a single, wildtype-based “reference embedding”.** **a**, Schematic of approach. We first applied principal components-based dimensionality reduction to cells from wildtype genotypes only (①), then projected cells from the mutant embryos to this PCA embedding (②). Next, we applied the *align_cds* function in *Monocle/v3*, with the MT%, Ribo%, and log-transformed total UMIs of each cell as covariates (③). We then split wildtype and mutant cells (④ & ⑤), and applied the UMAP algorithm to wildtype cells only using their “aligned” PC features (⑥), followed by Louvain clustering and manual annotation of individual clusters based on marker gene expression to identify major trajectories, and then iterative clustering and annotation to identity and annotate sub-trajectories (⑦). Finally, cells from mutant embryos were projected to this wildtype-based UMAP embedding using aligned PC features (⑧). Major trajectory and sub-trajectory labels were assigned to mutant cells via a k -nearest neighbour (k -NN) heuristic (⑨). **b**, 3D UMAP visualisations of cells from each wildtype or mutant background within the shared “reference embedding” resulting from the aforescribed procedures.

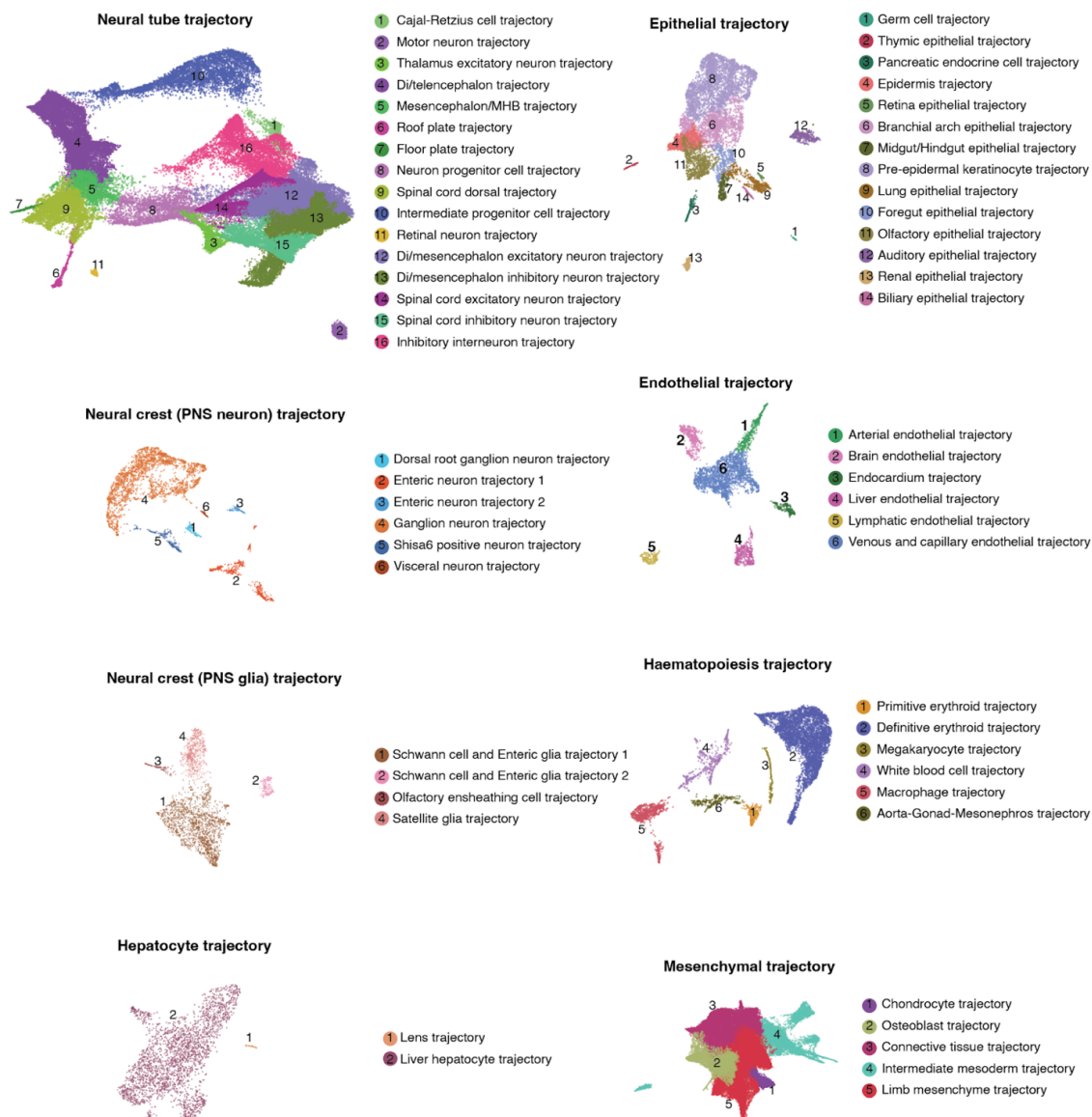


Figure 3.7: **Figure S3. Annotation of sub-trajectories in data from wildtype E13.5 embryos.** From 215,517 single cell profiles of wildtype E13.5 embryos of four strains in MMCA, we annotated 13 major trajectories. For 8 of these 13 major trajectories, iterative analysis identified the additional sub-trajectories shown here as 3D UMAP visualisations. Cells are colored by sub-trajectory annotations. PNS: peripheral nervous system. MHB: midbrain-hindbrain boundary. Di: Diencephalon.

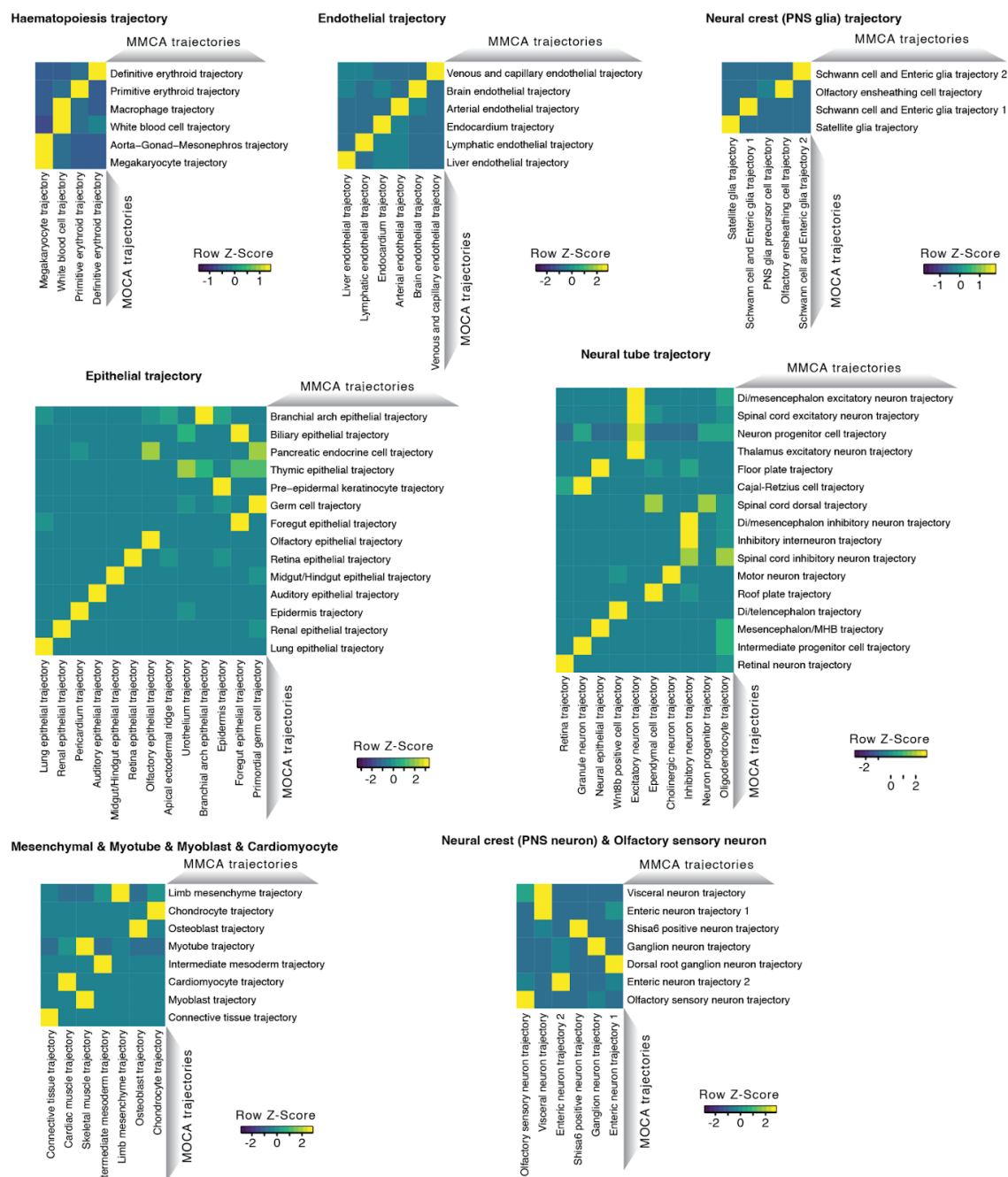


Figure 3.8: Figure S4. Correlated developmental sub-trajectories between MOCA (E9.5 - E13.5) and MMCA (E13.5 only) based on non-negative least-squares (*NNLS*) regression. Similar to Fig. 3.1f, shown here are heat maps of the combined β values (row-scaled) between developmental sub-trajectories from MMCA (rows) and developmental sub-trajectories from the MOCA (columns), within each major trajectory. PNS: peripheral nervous system. MHB: midbrain-hindbrain boundary. Di: Di-encephalon.

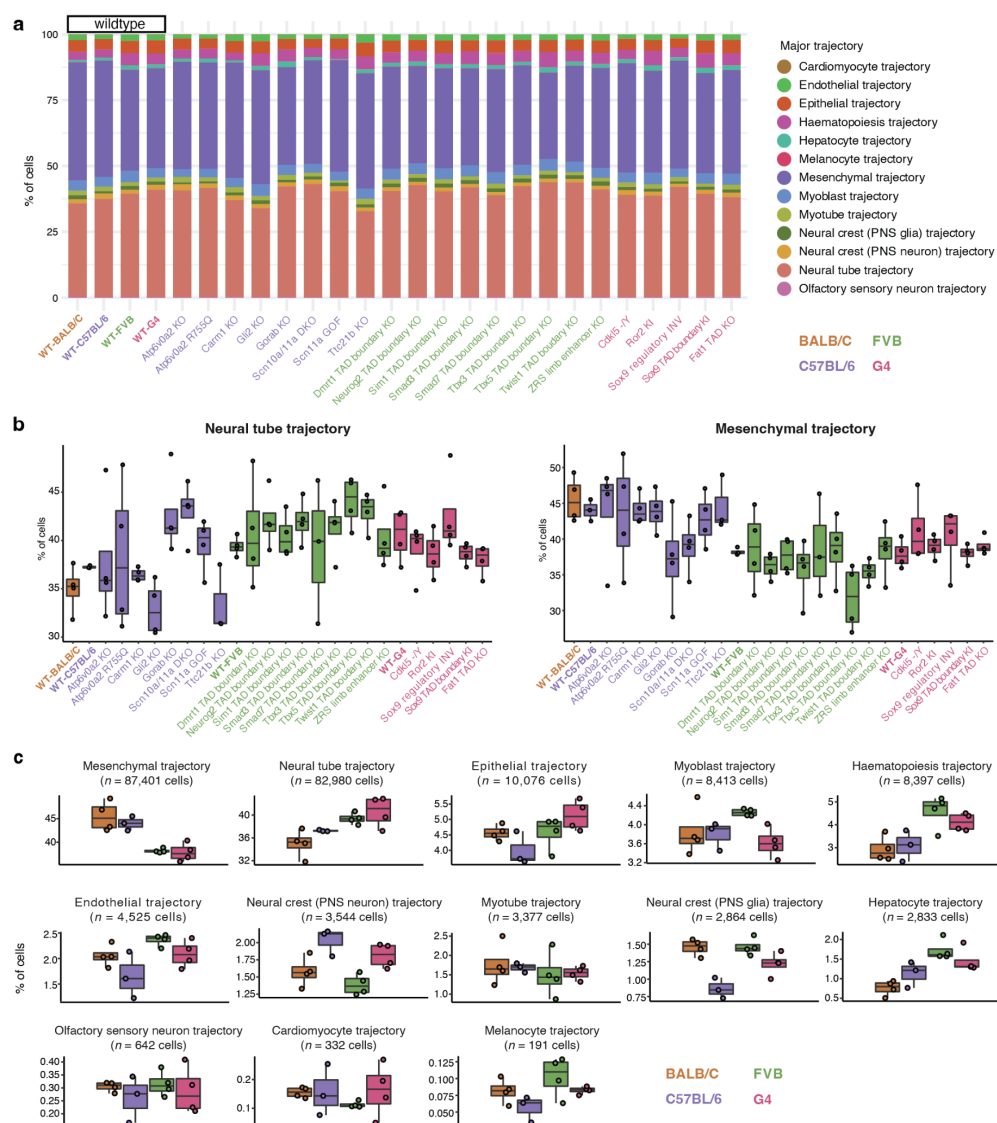


Figure 3.9: Figure S5. Cell composition for individual wildtype and mutant embryos across developmental trajectories. **a**, Cell composition across 13 major trajectories of embryos from different wildtype or mutant strains. Cells from all replicates for each strain were pooled for this visualisation. **b**, Boxplots of cell proportions falling into neural tube (left) or mesenchymal (right) trajectories for different wildtype or mutant strains. Each point corresponds to an individual embryo. **c**, Boxplots of cell proportions falling into each of the 13 major trajectories for the four wildtype strains. Each point corresponds to an individual embryo. The total number of cells from each major trajectory profiled from wildtype embryos is also listed. PNS: peripheral nervous system.

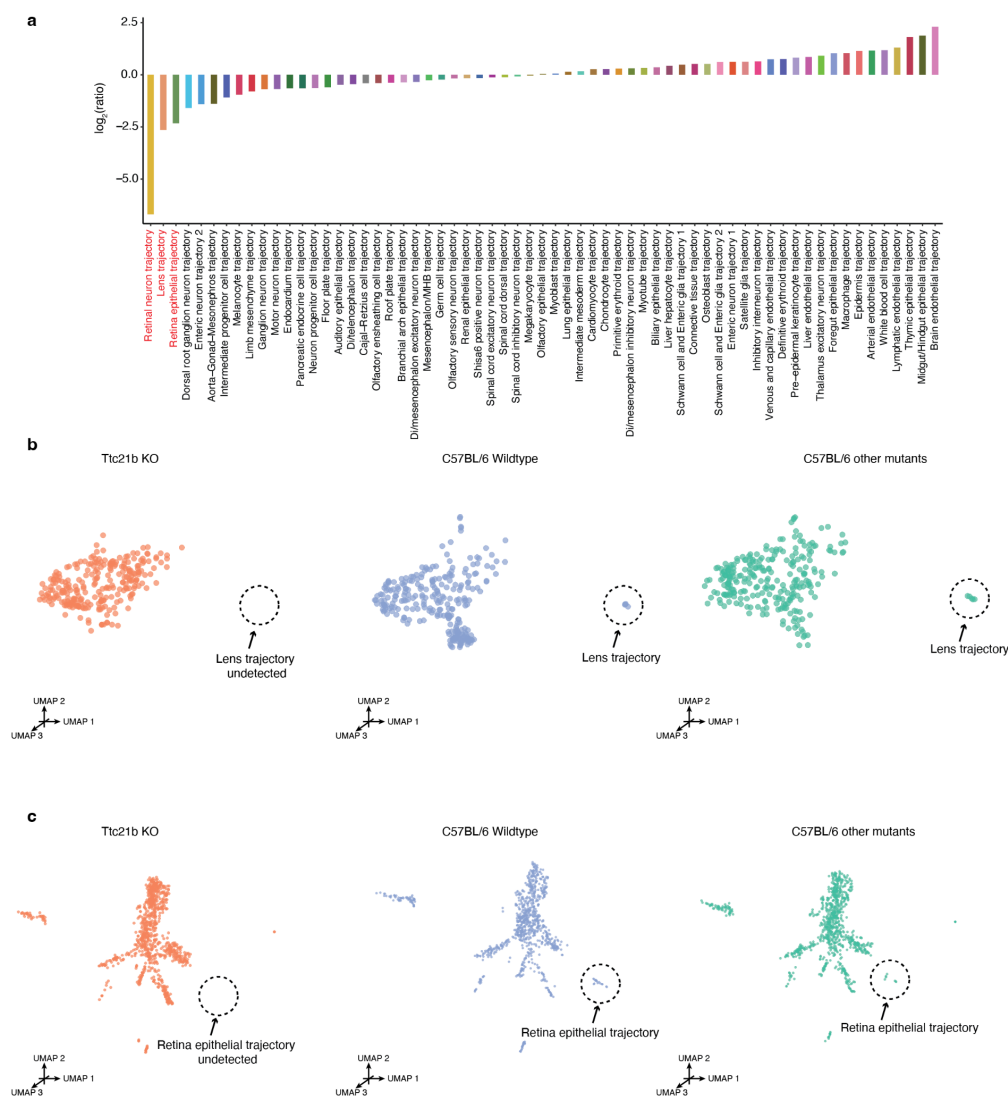


Figure 3.10: Figure S6. Multiple retinal trajectories are diminished in *Ttc21b* KO mice. **a**, The log₂ transformed ratio of the cell proportions of each sub-trajectory, comparing *Ttc21b* KO and C57BL/6 wildtype embryos, are shown. Although reductions in the retina epithelial and lens trajectories were excluded from the regression analysis due to their low numbers, they were, together with the retinal neuron trajectory, the most extreme in magnitude. **b**, 3D UMAP visualisation of the hepatocyte major trajectory, highlighting cells from either the *Ttc21b* KO (left), C57BL/6 wildtype (middle), or other mutants on the C57BL/6 background (right). The three plots were randomly downsampled to the same number of cells ($n = 264$ cells). **c**, 3D UMAP visualisation of the epithelial major trajectory, highlighting cells from either the *Ttc21b* KO (left), C57BL/6 wildtype (middle), or other mutants on the C57BL/6 background (right). The three plots were randomly downsampled to the same number of cells ($n = 937$ cells).

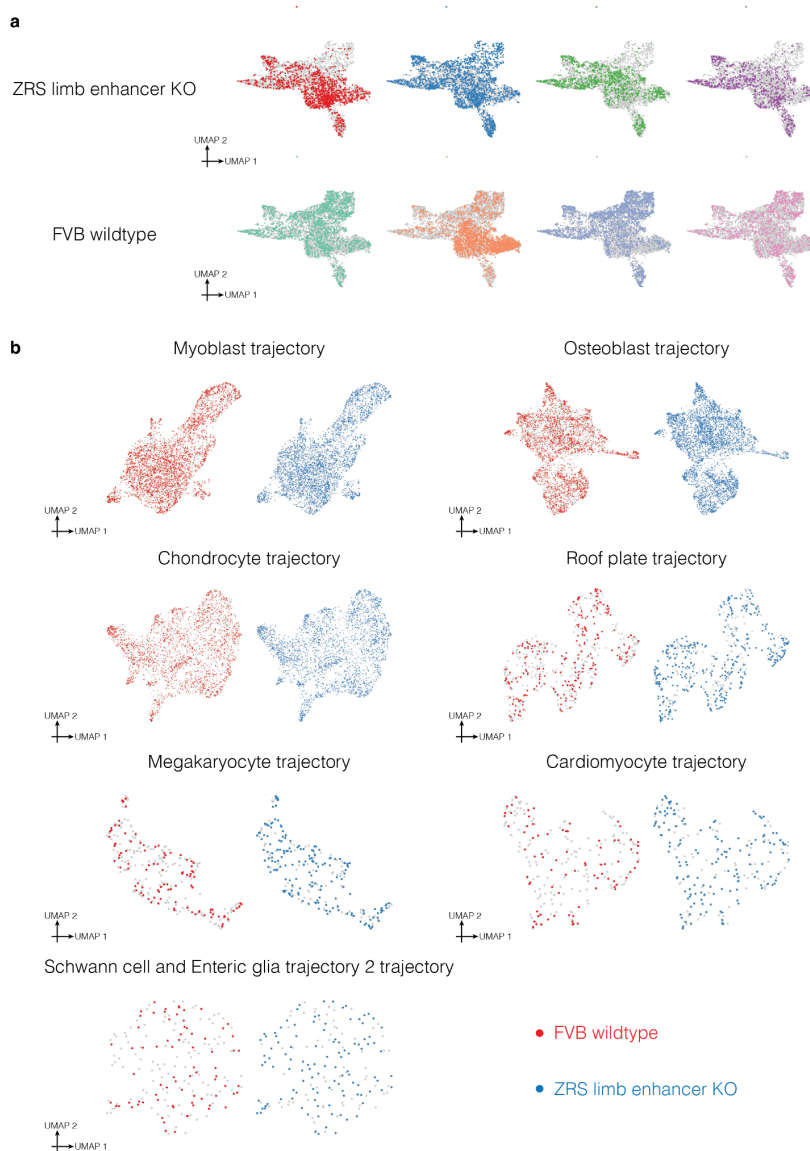


Figure 3.11: **Figure S7. Co-embedding cells from nominally altered trajectories from ZRS limb enhancer KO and FVB wildtype.** **a**, UMAP visualisation of co-embedded cells of limb mesenchyme trajectory from the ZRS limb enhancer KO and FVB wildtype. The same UMAP is shown eight times, highlighting cells from either ZRS limb enhancer KO (top row) or FVB wildtype (bottom row), and breaking out the four individual replicates for each strain. **b**, UMAP visualisation of co-embedded cells of various sub-trajectories from the ZRS limb enhancer KO and FVB wildtype. The same UMAP is shown twice for each, highlighting cells from either FVB wildtype (left) or ZRS limb enhancer KO (right). These are the seven sub-trajectories in which, in addition to limb mesenchyme, we detected nominally significant differences in cell type proportions for the ZRS limb enhancer KO.

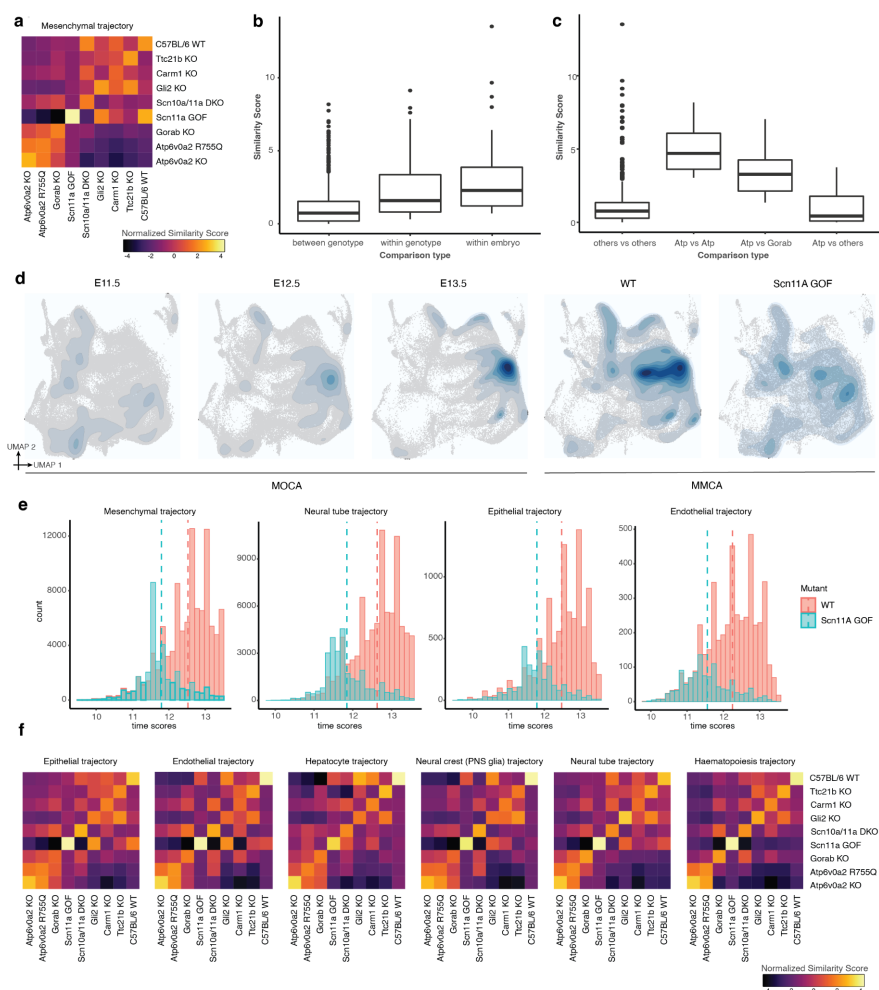


Figure 3.12: Figure S8. Quantitative analysis of lochNESS distributions and analysis of *Gli2* KO in the roof plate and floor plate trajectories. **a**, Distribution of lochNESS in all cells of each mutant under random permutation of mutant labels. **b**, Barplot showing the average euclidean distance between lochNESS vs. lochNESS under permutation across all cells within a mutant. **c**, 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. **d**, UMAP visualisation of co-embedded cells of the floor plate and roof plate sub-trajectories from the *Gli2* KO mutant and pooled wildtype, colored by sub-trajectory (left) or cluster number (right). **e**, Boxplot showing the lochNESS distribution in each cluster shown on the right of panel **d**. **f**, Barplots showing the cell composition of each cluster shown on the right of panel **d**, split by mutant vs. wildtype (left) or individual embryo (right), with a reference line at the overall wildtype cell proportion. **g**, Dotplot summarising the expression of and percent of cells expressing selected marker genes in each cluster shown on the right of panel **d**.

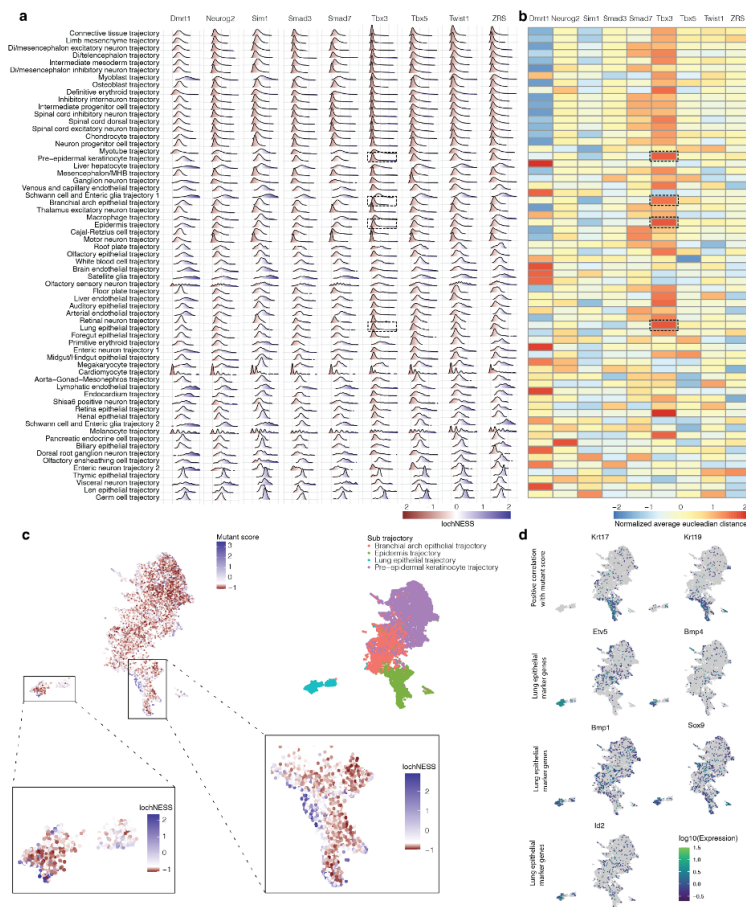


Figure 3.13: Figure S9. Systematic screening of lochNESS distributions identifies altered epithelial sub-trajectories in the *Tbx3* TAD Boundary KO mutant. **a**, Distribution of lochNESS in each sub-trajectory of the mutants in the FVB background strain, all of which are TAD boundary KOs. Dashed boxes in the sixth column highlight the most deviated epithelial sub-trajectories in the *Tbx3* TAD Boundary KO mutant. **b**, Row-normalised heatmap showing the average euclidean distance between lochNESS and lochNESS under permutation in each sub-trajectory for the same mutants shown in panel **a**, centred and scaled by row. Dashed boxes in the sixth column again highlight the most deviated epithelial sub-trajectories in the *Tbx3* TAD Boundary KO mutant. **c**, UMAP showing co-embedding of *Tbx3* TAD Boundary KO and pooled wildtype cells in the pre-epidermal keratinocyte, epidermis, branchial arch, and lung epithelial sub trajectories, colored by lochNESS (top left) [with blown up insets showing lochNESS in lung epithelial (bottom left) and epidermis (bottom right) sub-trajectories], or by sub-trajectory identity (right). LochNESS colour scale is centred at the median of lochNESS. **d**, same as in panel **c**, but colored by expression of selected mutant related genes and marker genes.

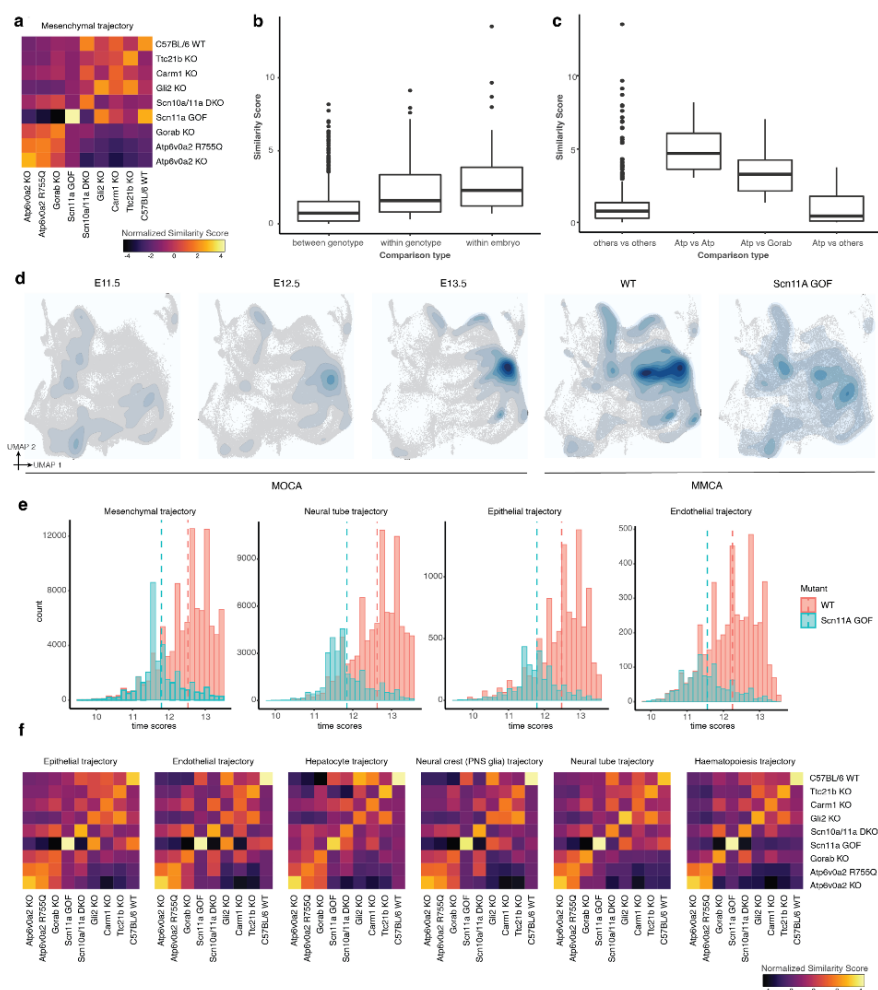


Figure 3.14: **Figure S10. Similarity scores reveal mutant-shared and mutant-specific effects.**

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 (left), between embryos of the same genotype (middle), and within the same embryos (right) for C57BL/6 genotypes in the mesenchymal trajectory. **c**, Boxplot showing the similarity scores of comparisons between *Atp6v0a2* KO vs. *Atp6v0a2* R755Q (left), *Atp6v0a2* KO or *Atp6v0a2* R755Q vs. *Gorab* KO (middle), *Atp6v0a2* KO or *Atp6v0a2* R755Q vs. other C57BL/6 genotypes. **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), with cell density and distributions overlaid. **e**, Barplots showing the distribution of “time scores” for *Scn11a* GOF cells and pooled wildtype cells in the mesenchyme, neural tube, endothelial and epithelial main trajectories, with reference lines at the mean value of time scores. **f**, Heatmaps showing similarity scores between C57BL/6 genotypes in selected main trajectories.

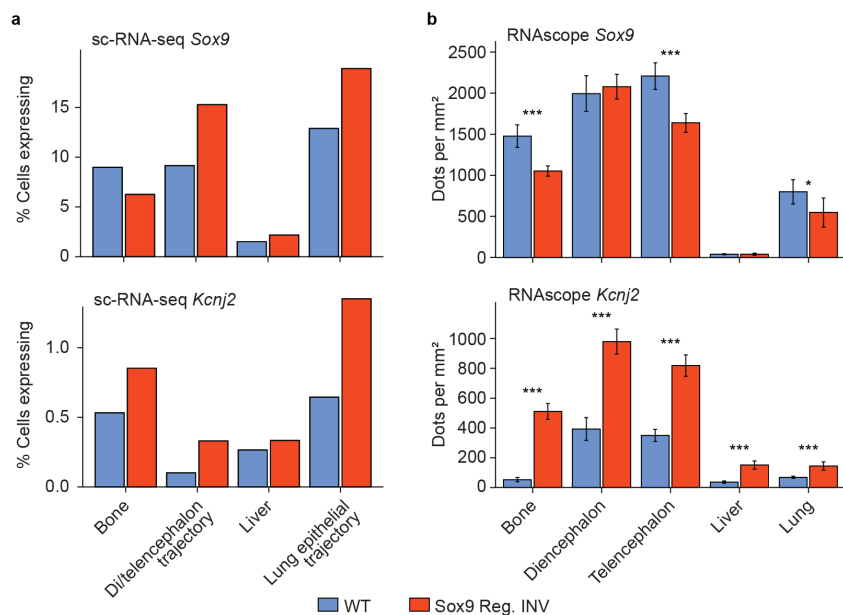


Figure 3.15: **Figure S11. Misregulation of *Sox9* and *Kcnj2* in the *Sox9* regulatory INV mutant.** **a**, Quantification of *Sox9* (top row) and *Kcnj2* (bottom row) expression in sc-RNA-seq data in the wildtype (blue) and *Sox9* regulatory INV (red) genotypes in selected trajectories. For “bone” and “liver”, multiple sub-trajectories were pooled to match the tissue labels in the RNAscope data in panel **b**. Specifically, “bone” refers to cells from chondrocyte, osteoblast, and limb mesenchyme trajectories, whereas “liver” refers to cells from the liver endothelial and liver hepatocyte trajectories. **b**, Quantification of *Sox9* and *Kcnj2* expression based on RNAscope images of heterozygous E13.5 wildtype and *Sox9* regulatory INV mutant embryos (images not shown; available upon request). The mRNA signal was counted in a defined area (1 x 1 mm²), n=6 each condition. Statistics were calculated using student t-test and evaluated the following: p > 0.05 = non-significant; p < 0,05 - ≥ 0.01 = *; p < 0,01 - ≥ 0.001= **; p < 0.001= ***.

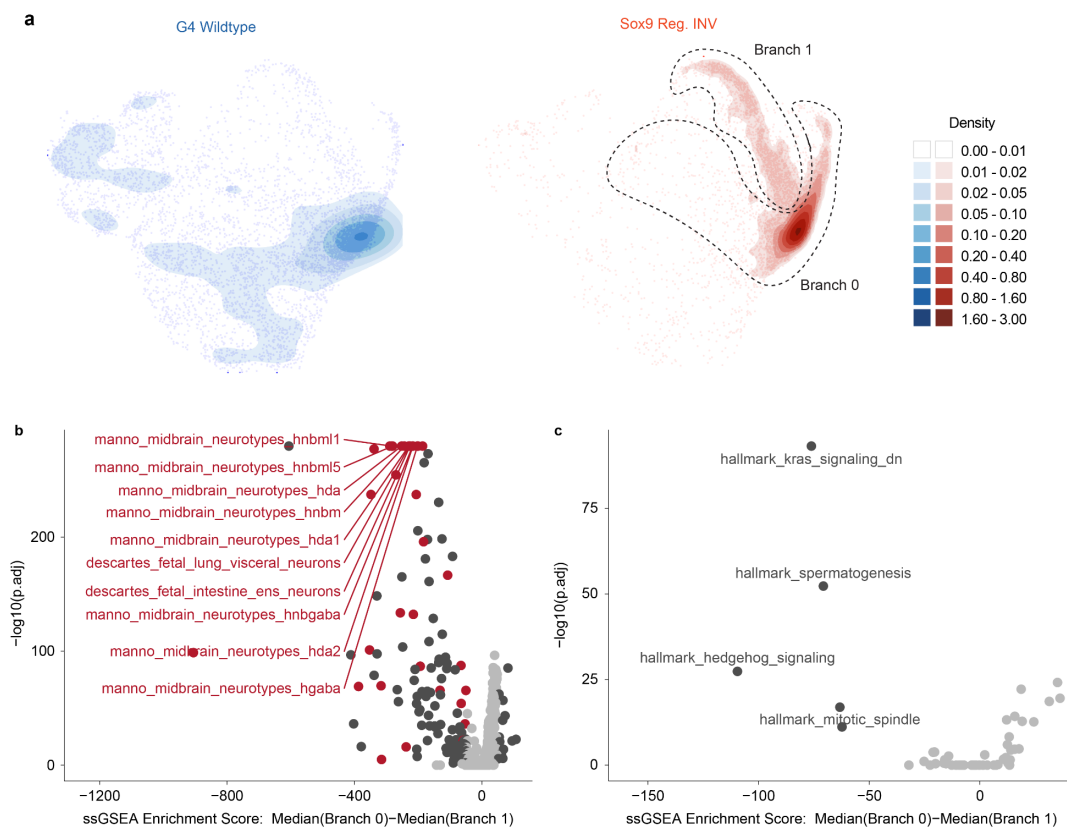


Figure 3.16: **Figure S13. Stalling of regulatory INV cells in the undifferentiated mesenchyme and gene set enrichment analysis on these cells.** **a**, Density for UMAP embedding of G4 wildtype and Sox9 regulatory INV cells in the limb mesenchymal trajectory (same embedding as **Fig. 3.4e**). Dotted black lines demarcate the two branches of the undifferentiated mesenchyme, based on the sub-clustering shown in **Fig. 3.4f**. **b,c**, Comparison of the ssGSEA [176] scores between the two branches of undifferentiated mesenchyme for Sox9 regulatory INV cells for (a) cell type signature (C8) and (b) Hallmark gene sets. Gene sets that are both significantly different between the two branches and that have a difference in median ssGSEA scores greater than 50 are highlighted in dark grey, and the ten most significantly different gene sets are also labelled. In panel **b**, all significantly different gene sets with names containing "neuro" are highlighted in red.

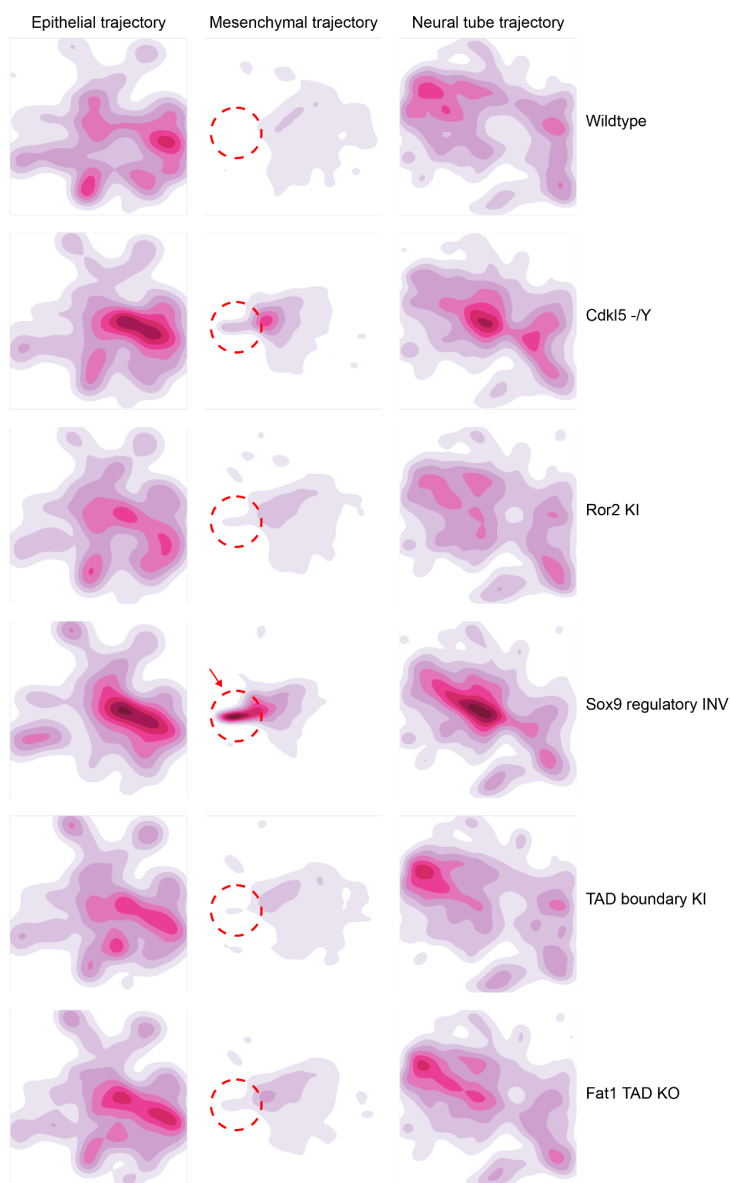


Figure 3.17: **Figure S14. Density plots of the UMAP co-embedding of G4 samples.** We focus on the epithelial, mesenchymal and neural tube main trajectories, which are the three largest. The densities are corrected for the total number of cells. The colour scale is kept consistent across mutants (rows), but varied across the trajectories (columns). Arrow points to the accumulation of cells in the *Sox9* regulatory INV mutant. Dotted circles demarcate the location of cellular accumulation in *Sox9* regulatory INV mutant in the same embedding across all the other mutants.

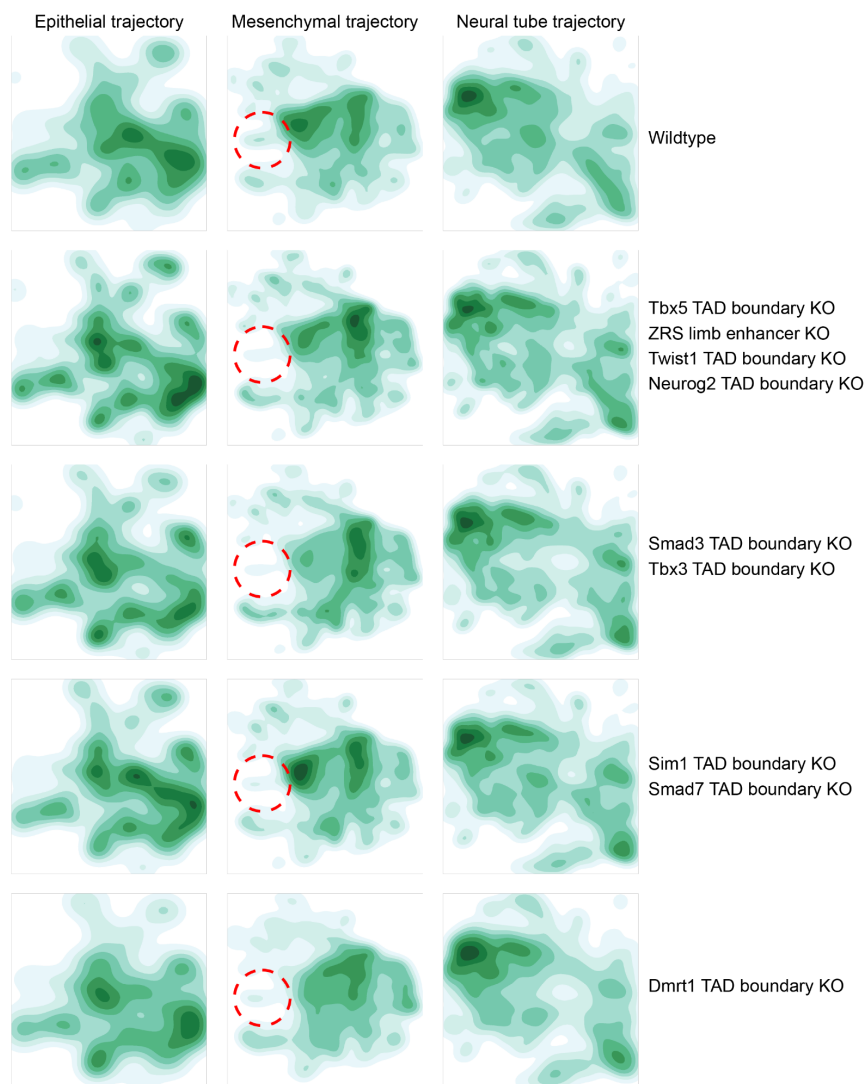


Figure 3.18: **Figure S15. Density plots of the UMAP co-embedding of FVB samples.** We focus on the epithelial, mesenchymal and neural tube main trajectories, which are the three largest. The same embedding as in Fig. 3.17 was used. Mutants with visually similar UMAP embeddings were combined for presentation. The densities are corrected for the total number of cells. The colour scale is kept consistent across mutants (rows), but varied across the trajectories (columns). Dotted circles demarcate the location of cellular accumulation in *Sox9* regulatory INV mutant in the same embedding across all the other mutants.

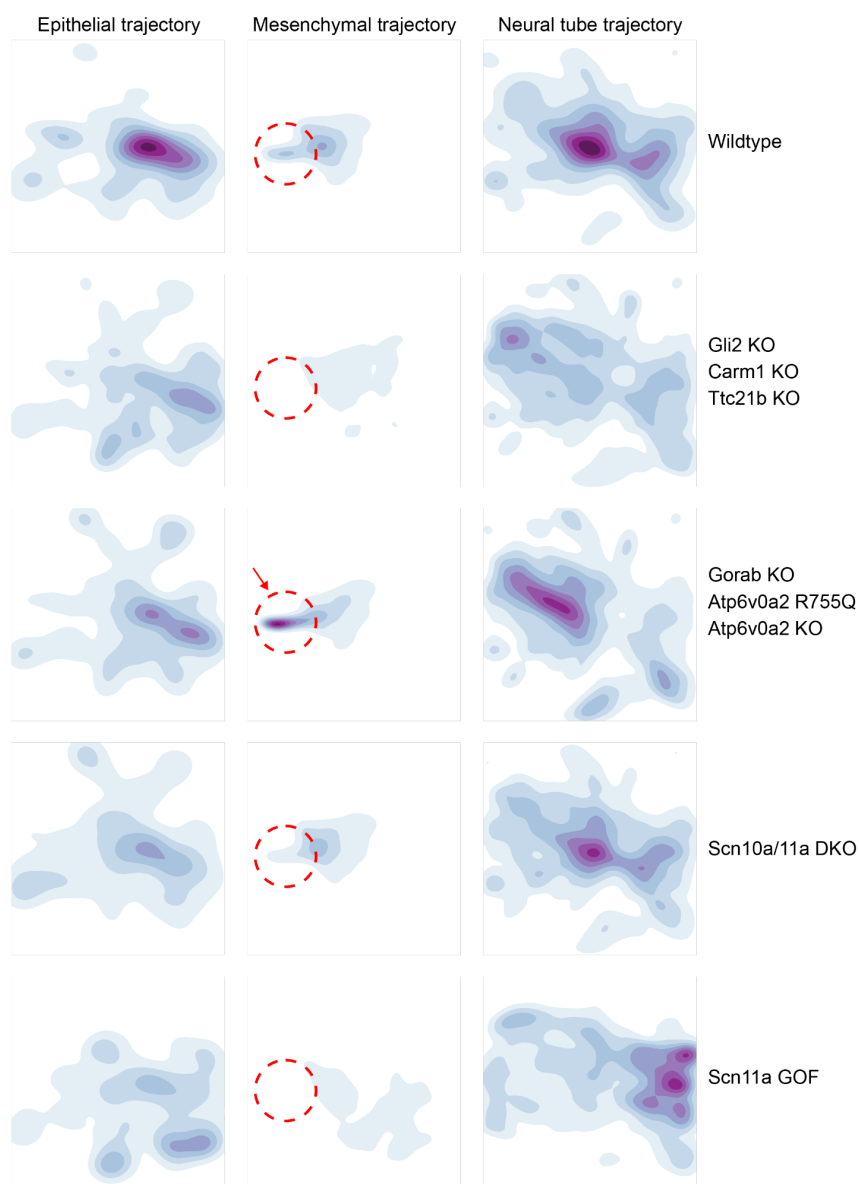


Figure 3.19: **Figure S16. Density plots of the UMAP co-embedding of C57BL/6 samples.** We focus on the epithelial, mesenchymal and neural tube main trajectories, which are the three largest. The same embedding as in Fig. 3.17 was used. Mutants with visually similar UMAP embeddings were combined for presentation. The densities are corrected for the total number of cells. The colour scale is kept consistent across mutants (rows), but varied across the trajectories (columns). Dotted circles demarcate the location of cellular accumulation in *Sox9* regulatory INV mutant in the same embedding across all the other mutants. Arrow highlights a similar accumulation of cells in the *Gorab* KO, *Atp6v0a2* R755Q, and *Atp6v0a2KO* mutants.

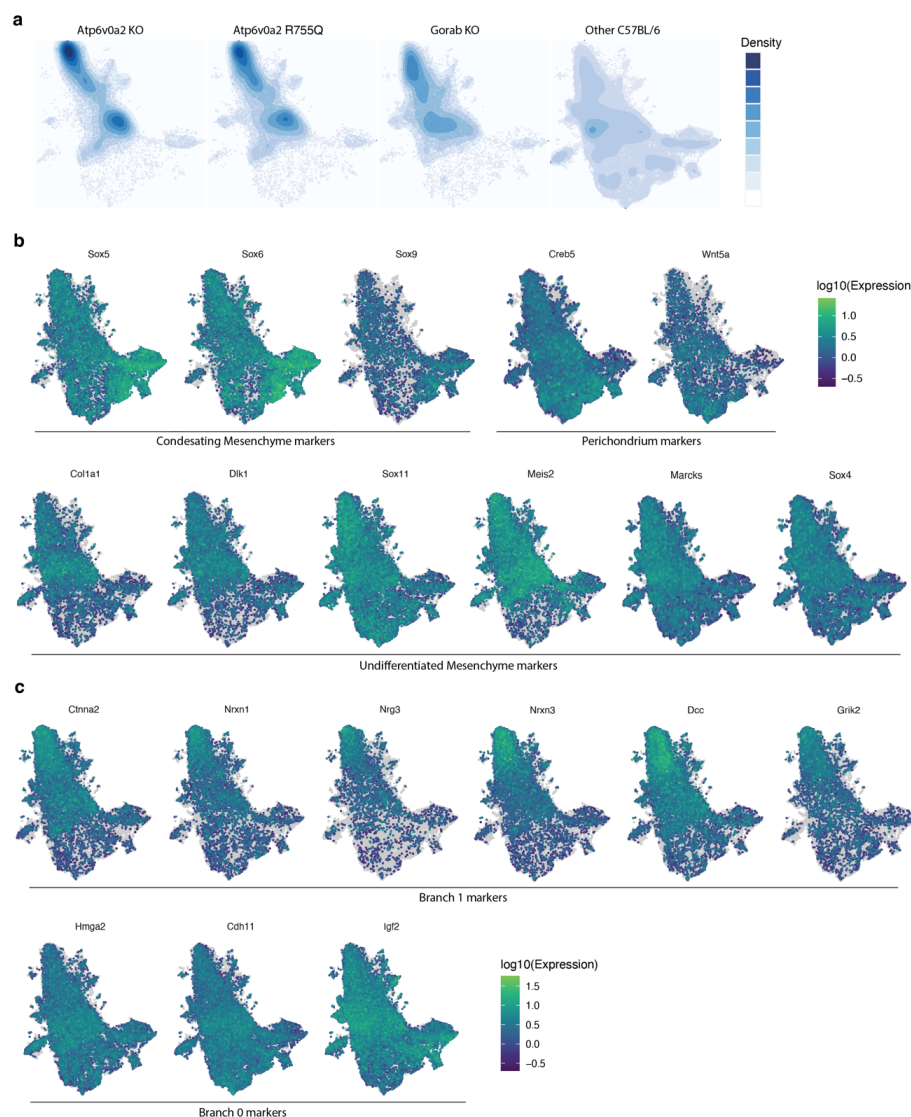


Figure 3.20: **Figure S17. Density and marker gene expression plots of UMAP co-embeddings of wildtype and mutant samples from C57BL/6 mouse background in the limb mesenchyme trajectory.** **a**, UMAPs showing the co-embeddings of the limb mesenchyme trajectory for wildtype and mutant genotypes from the C57BL/6 background strain, with cell density and distributions overlaid. **b**, same as in panel **a**, but colored by expression of limb mesenchyme sub-cluster marker genes. The accumulation of cells in the *Gorab* KO, *Atp6v0a2* R755Q, and *Atp6v0a2* KO mutants express markers of undifferentiated mesenchyme. **c**, same as in panel **a**, but colored by expression of significantly differentially expressed genes between the two branches of Sox9 regulatory INV undifferentiated mesenchyme cells as shown in Fig. 3.4g.

Chapter 4

A SINGLE-CELL MULTI-OMIC ATLAS SPANNING THE ADULT RHESUS MACAQUE BRAIN

This Chapter is adopted from in revision work with minimum changes:

K. L., Chiou*, **X.**, **Huang***, M. O., Bohlen, S., Tremblay, D. R., O’Day, C. H., Spurrell, A. A., Gogate, T. M., Zintel, Cayo Biobank Research Unit, M. G., Andrews, M. I., Martínez, L. M., Starita, M. J., Montague, M. L., Platt#, J., Shendure#, N., Snyder-Mackler#, “A single-cell multi-omic atlas spanning the adult rhesus macaque brain”, bioRxiv, 2022.

I had been aware of a massive macaque brain dataset for some time due to emails asking for advice on analyzing large sci-ATAC-seq datasets, but it wasn’t until April 2022 that I had the chance to speak with Kenny and Noah about it. At that point, they had already planned to submit to a journal as part of a BICCN package, even before the paper was written. The timing was perfect for me, as I was just wrapping up the revision on DEAP and was waiting for reviewers to return feedback on MMCA. I was able to quickly jump on board and adapt some of my previous analyses to this new dataset. Just as in MMCA, we were interested in detecting changes in cell composition and cell distributions, but this time between multiple brain regions. There were only minor challenges in extending lochNESS from pairwise comparisons to multiple regions, and the analysis helped identify marker genes for region-specific cell subtypes without clustering. Collaborating with Kenny, Noah, and others was incredibly smooth, and I learned a lot about neuroscience. I also provided input on important analysis decisions, such as using scGLUE for integration, NNLS for cell type label comparison, and the ensemble model for regulatory inference. It was a truly amazing experience analyzing a massive dataset with a small team and writing a paper in less than

six months.

More formally, the author contributions are listed in the manuscript as follows: N.S.M., J.S., M.L.P., M.J.M., L.M.S., K.L.C., and X.H. conceived the study. K.L.C., M.J.M., and N.S.M. collected samples, with logistical support from CBRU and MIM. M.O.B. performed neuroanatomical dissections, assisted by K.L.C., T.M.Z., M.J.M., and N.S.M.. K.L.C., D.R.O., C.H.S., and T.M.Z. performed lab work. A.A.G. managed data and provided bioinformatic support. K.L.C., X.H., T.M.Z., and N.S.M. analyzed the data, with input from M.O.B., S.T., M.G.A., L.M.S., M.J.M., M.L.P., and J.S. K.L.C., X.H., M.O.B., S.T., M.J.M., M.L.P., J.S., and N.S.M. wrote the paper. All authors edited and approved the manuscript.

Abstract Cataloging the diverse cellular architecture of the primate brain is crucial for understanding cognition, behavior and disease in humans. Here, we generated a brain-wide single-cell multimodal molecular atlas of the rhesus macaque brain. Altogether, we profiled 2.58M transcriptomes and 1.59M epigenomes from single nuclei sampled from 30 regions across the adult brain. Cell composition differed extensively across the brain, revealing cellular signatures of region-specific functions. We also identified 1.19M candidate regulatory elements, many novel, allowing us to explore the landscape of *cis*-regulatory grammar and neurological disease risk in a cell-type-specific manner. Together, this multi-omic atlas provides an open resource for investigating the evolution of the human brain and identifying novel targets for disease interventions.

4.1 Introduction

The cellular and molecular origins of complex human thought and behavior remain largely a mystery. Historically, proposed explanations have centered on the large relative size [188, 189, 190], high cell numbers [191], or the large cortical surface area and thickness [192] of

the human brain. These explanations in isolation, however, fail to explain the many uniquely human faculties, nor do they explain the extreme variety and complexity of impairments that accompany human neurodevelopmental, neuropsychiatric, and neurodegenerative disorders [193]. The human brain is composed of myriad cell types and this cellular heterogeneity contributes to our cognitive and behavioral complexity [194, 195]. Supporting this hypothesis is the observation that the number of distinct cell types in the brain is positively correlated with behavioral complexity across vertebrates [196]. In recent decades, it has been proposed that certain aspects of higher human cognition are supported by specific cell types such as von Economo neurons [197] and “mirror neurons” [198], which have been hypothesized to support intuition and empathy, respectively. These propositions, however, remain largely untested due to gaps in our understanding of the cellular landscape of the human brain and, crucially, differences in cell-type composition and regional heterogeneity among the brains of humans, nonhuman primates, and other animals.

In recent years, the application of rapidly developing single-cell technologies to the brain has begun to address these gaps. Single-cell molecular surveys of targeted regions of the mouse and human brain, for example, have revealed specialized species-specific cell types—e.g., rosehip neurons in humans [199]—and regional biases in cell-type distribution and function (e.g., [200]). Such atlases are yielding unprecedented cross-species insights into the cellular architecture supporting the structure and function of the brain [201, 202]), but the general paucity of comparative nonhuman primate brain atlases has left a conspicuous gap [203]. Moreover, much effort has focused on single molecular modalities (e.g., transcriptomics), typically in only one or a few regions, leaving a lacuna in our understanding of the molecular mechanisms underlying cell function across much of the primate brain.

Here, we generated a 4.2 million cell (combined) transcriptomic and epigenomic atlas across the brain of the rhesus macaque (*Macaca mulatta*), the most widely used nonhuman primate model organism for studies of human perception, cognition, aging, and neurolog-

ical disease [204]). These single-cell profiles derive from 30 distinct brain regions that collectively represent major cortical, subcortical, and cerebellar areas involved in sensory, cognitive, emotional, and motor functions. Many of these regions are also implicated in one or more clinically relevant neurological disorders. By integrating measures of gene expression and chromatin accessibility, we discover molecular signatures that define cell types across the macaque brain, characterize their distribution and molecular function across disparate anatomical regions, and nominate sets of *cis*-regulatory regions that likely contribute to mature cell fate and function across the brain.

4.2 Results

4.2.1 molecular taxonomy of cell types across the primate brain

We generated single-nucleus RNA sequencing (snRNA-seq) data from 30 distinct regions across the cortex, subcortex, cerebellum, and brainstem ($N=5$ animals, 3 female) using sci-RNA-seq3 combinatorial indexing [20, 2] (**Fig. 4.1A, table S1**). With the original sci-RNA-seq3 protocol [2], we generated 1,008,204 single-nucleus transcriptomes from 110 age-, sex-, and hemisphere-matched samples representing 28 brain regions of 10 year-old (mid-adult aged) macaques ($N=3$ animals; 2 female). Over the course of the study, we implemented improvements in nuclei isolation and preservation [205] which increased nuclear transcriptome recovery by $\sim 60\%$ (median unique molecular indices [UMIs], before=202, after=320) and, consequently, the number of nuclei passing our UMI threshold. With the improved protocol, we generated an additional 1,702,081 single-nucleus transcriptomes from the right hemisphere of two animals, the vast majority ($N=1,579,908$) of which were sampled from 27 brain regions of a single 10 year-old female macaque. Altogether, after applying quality control filters (**Methods, Fig. 4.7,4.8**), we recovered transcriptome profiles for 2,583,967 nuclei (median UMI per cell=265, median genes expressed per cell=221, **table S2**).

Controlling for batch effects across sequencing runs (**Methods, Fig. 4.9**), we jointly clustered single cell profiles across all sampled brain regions to identify 17 molecularly distinct cell types, which we refer to as “cell classes” (**Fig. 4.1B-C**). Based on established cell markers (**Fig. 4.10, table S3**), we annotated these 17 cell classes as either: (i) neuronal cells, including cortical glutamatergic neurons (*CAMK2A*), cortical GABAergic neurons (*GAD1, GAD2*), basket cells (*GRID2, SORCS3*), other cerebellar neurons (primarily granule cells; *GRM4*), medium spiny neurons (*DACH1, PPP1R1B, BCL11B*), serotonergic neurons (*TPH2*), dopaminergic neurons (*TH, DBH*); or (ii) non-neuronal cells, including microglia (*DOCK2*), oligodendrocyte precursor cells (OPCs; *VCAN*), astrocytes (*ALDH1A1, GFAP*), oligodendrocytes (*MOG, MBP*), vascular cells (*CFH*), and ependymal cells (*FOXJ1*). Our broad survey also captured four rare, possibly novel cell populations that to our knowledge, have not yet been identified in other studies: three *RBFOX3*⁺ (NeuN⁺) neuron-like populations (marker genes: *APOA2*, $N=7,055$ cells; *F5*, $N=880$; *KIR2DL1/2*, $N=84$) and one *RBFOX3*⁻ microglia-like population (marker gene: *KIR3DL1/2*⁺, $N=44$ cells, also *P2RY12*⁺/*PTPRC*⁺/*ENTPD1*⁺). Given their rarity, we removed these four cell populations from downstream analyses. Hierarchical clustering of cell classes by the top 50 principal components of gene expression largely recapitulated broad ontogenetic relationships, with most neuronal classes clustering together (dopaminergic neurons being the exception) and the two mesoderm-derived classes (microglia and vascular cells) clustering together (**Fig. 4.11A**).

By sampling across a broad range of anatomical regions within the same individuals, we were able to characterize cellular composition across 30 distinct brain regions—to our knowledge, the most regionally expansive primate single-cell brain atlas to date (**Fig. 4.1D-E**). The distribution of major cell classes were balanced between sexes and hemispheres (**Fig. 4.12**), but differed extensively across regions, reflecting the cellular makeup underlying region-specific functions (**Fig. 4.1E**). Unsupervised hierarchical clustering of brain regions according to cell-class composition for the most part conformed to broader anatomical catego-

rizations, with regions of the cortex, subcortex, brainstem, and cerebellum usually grouping together (**Fig. 4.11B**), which was also the case when clustering regions based on the top 50 principal components of gene expression (**Fig. 4.11B**). Two of these four broad regional classes were comprised primarily of a single cell class: in the cortex ($N=16$ regions, **table S4**), glutamatergic neurons were the most abundant cell type (mean=63.7% of all cells per sample) and outnumbered GABAergic neurons by almost four-fold (**Fig. 4.1E**; mean=17.4%), while the cerebellum ($N=2$ regions) was composed almost entirely of cerebellar neurons (mean=85.1%). In contrast, the subcortex ($N=8$ regions) and brainstem ($N=4$ regions), were more heterogeneous with respect to their cellular composition, with samples from these regions containing roughly equal proportions of glutamatergic neurons (mean_{subcortex}=25.1%; mean_{brainstem}=25.5%), GABAergic neurons (mean_{subcortex}=20.2%; mean_{brainstem}=23.0%), and oligodendrocytes (mean_{subcortex}=18.5%; mean_{brainstem}=25.5%). We further subdivided the cortical and subcortical samples into “region subclasses” based on neuroanatomical groups (**table S1**), in which there was more limited variation in cellular composition (**Fig. 4.1E**). For instance, in the subcortex, medium spiny neurons (MSN) comprised around half of the cells in the basal ganglia (nucleus accumbens [NAc] mean=44.7%; caudate nucleus [CN] mean=60.0% MSNs), while the thalamus was enriched for GABAergic neurons (lateral geniculate nucleus [LGN] mean=55.7%; mediodorsal thalamic nucleus [mdTN] mean=43.8%; ventrolateral thalamic nucleus [vlTN] mean=28.6%).

Our broad survey also captured two rarer, but important, cell classes: dopaminergic and serotonergic neurons. These two neurons collectively represented less than 0.3% of all profiled cells (dopaminergic=0.14%; serotonergic=0.12% of all cells) and 0.5% of all neurons (dopaminergic=0.19%; serotonergic=0.17% of all neurons), suggesting that targeted approaches that enrich for these cells (e.g., [206, 207]) are necessary to identify transcriptional variation among subtypes. Dopaminergic neurons, which are found primarily in the substantia nigra pars compacta at low frequency (1.1% of cells sampled in the midbrain vs. mean 0.1% in other sampled regions), are involved in a range of important processes,

including voluntary movement, reinforcement learning, and addiction, and their loss is a neuropathological hallmark of Parkinson’s disease [208]. We found that serotonergic neurons were most abundant in the brainstem (mean 0.35% in the 4 brainstem regions vs. mean 0.09% in other sampled regions), where they play a major role in sleep, mood, and appetite, and are key targets of pharmacological therapies for major depressive disorder in humans [209].

4.2.2 Regional variation in cell subtype composition

To characterize heterogeneity within cell classes, we partitioned the dataset and repeated preprocessing and clustering separately for each of the 17 cell classes. Collectively, we identified 112 distinct clusters (**Fig. 4.13, table S5**) that captured neuronal and non-neuronal diversity across the primate brain (**Fig. 4.2A**). We refer to clusters at this level as “cell subtypes”. We identified extensive heterogeneity in glutamatergic (39 subtypes) and GABAergic (20 subtypes) neurons primarily found in the cortex and some regions of the subcortex (e.g., hippocampus, thalamus), while neurons derived from other non-cortical brain regions (e.g., cerebellum, striatum) were transcriptionally distinct and relatively homogeneous within those regions (**Fig. 4.2A**). This is due in part to the large number of specialized neurons present in some of these regions, including granule and Purkinje cells in the cerebellum, and medium spiny neurons in the basal ganglia (**table S5**).

Our systematic approach also allowed us to characterize and compare the regional cellular distribution of non-neuronal subtypes, including those of glia, which have not often been the focus on most single cell atlases to date (**Fig. 4.2A, table S6**). Overall, we identified 6 astrocyte, 2 microglial, 7 oligodendrocyte, and 6 vascular cell subtypes, the latter including endothelial cells, smooth muscle cells, pericytes, and both perivascular and meningeal fibroblasts (**Fig. 4.13, table S5**) (A. C. [210]). We compared cell subtypes to published datasets using a NNLS approach [2], and found broad correspondence with subtypes observed in hu-

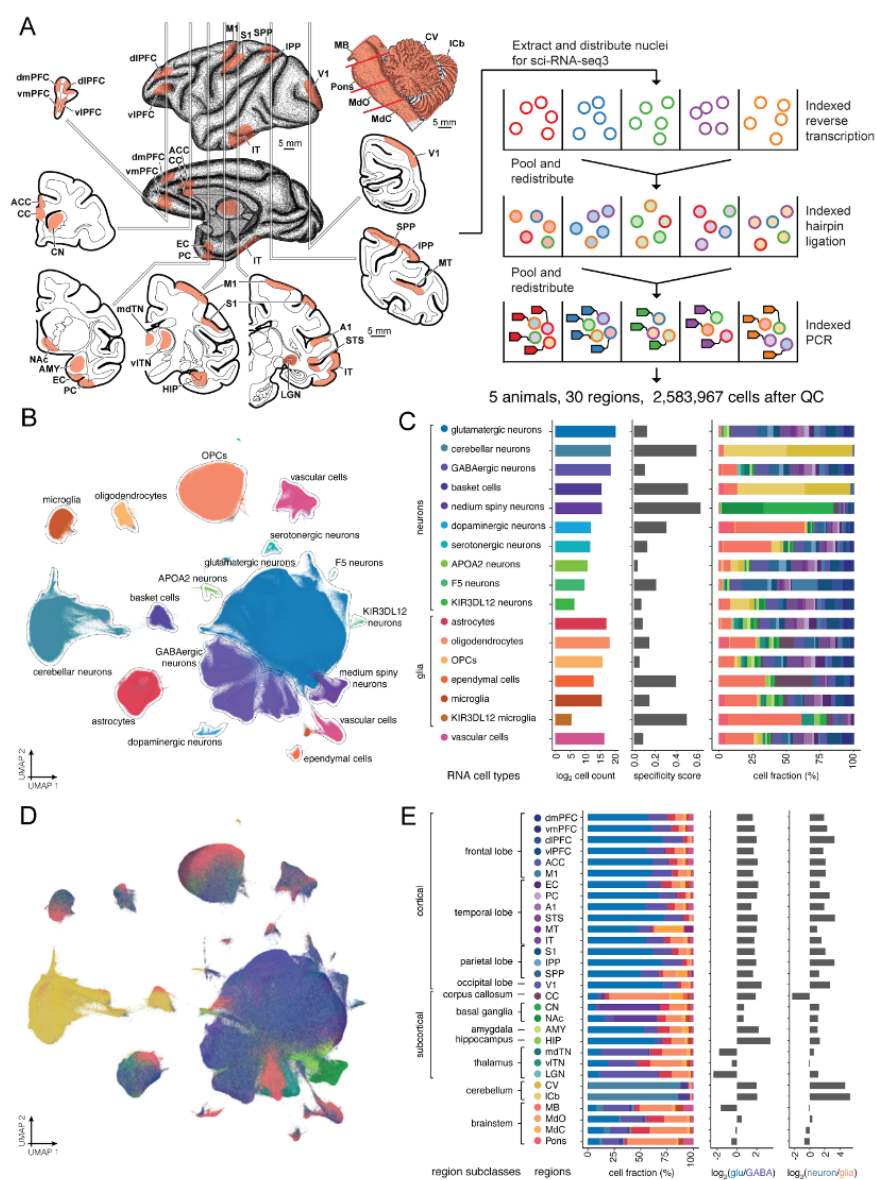


Figure 4.1: **Fig. 1. Experimental setup and summary of the Macaque Brain Atlas snRNA-seq dataset.** **A**, Schematic of biopsied brain regions for sci-RNA-seq3 experiment. A full list of sampled regions is provided in **table S1**. **B**, UMAP visualization of all snRNA-seq profiled cells colored by cell type (with color code shown in panel C). **C**, Barplots showing the \log_2 transformed cell counts (left), regional specificity score (middle) and regional composition (right, with color code shown in panel E) of each cell type. **D**, UMAP visualization of all snRNA-seq cells colored by cell type (with color code shown in panel E). **E**, Barplots showing the cell type composition (left, with color code shown in panel C), \log_2 transformed ratio of glutamatergic neurons and GABAergic neurons (middle) and \log_2 transformed ratio of neurons and glial cells (right) of each region. Regions are organized by the regional subclass to which they belong.

man cortical [201], human brain vascular (A. C. [210], and macaque hippocampal atlases [32]) (Fig. 4.14A-E).

To identify cell subtypes that were specific or biased towards a single region or set of regions, we calculated a measure of “regional specificity” using the Jensen-Shannon divergence statistic (**Methods**, [211, 30]). Overall, glial subtypes were more evenly distributed across all regions compared to neuronal subtypes (**Fig. 4.2A**). This is reflected in lower Jensen-Shannon specificity scores for glial subtypes (mean=0.20; median=0.15; range=[0.04,0.81]) compared to cortical neurons (mean=0.31; median=0.18; range=[0.08,0.89]). A number of cell subtypes, both neuronal and non-neuronal, were highly region-specific. For instance, oligodendrocyte subtype 8, the rarest oligodendrocyte subtype ($N=3,439$ cells; 1.5% of oligodendrocytes) overwhelmingly derived from the highly myelinated corpus callosum (93.0% of these cells; **Fig. 4.2A, table S6**). Among cortical neurons, GABAergic interneuron subtypes generally exhibited a lower median regional specificities than to glutamatergic neuron subtypes, although there were a number of interneuron subtypes specific to the thalamus (cluster 6) or brainstem (clusters 3 and 16), discussed below.

Given that the regional specificity of excitatory neuronal subtypes has been explored in depth in other studies (e.g., [212, 213], we focus here instead on populations that are vital for neuronal signal transduction but for which cellular diversity has not previously been explored across the macaque brain. Specifically, we concentrated on the regional diversity of interneurons, because they are important components of long-range circuitry and have been characterized in a few regions across mice, monkeys, and humans (e.g., [199, 202, 201, 214], allowing us to both benchmark our atlas but also extend current knowledge to understudied regions. We also examine regional distribution among astrocytes, which are crucial for maintaining neuronal homeostasis [215] and are implicated in neurological disorders [216], but have been relatively understudied at the single-cell level.

We pursued three main approaches to dissect the regional heterogeneity within interneuron and astrocyte subtypes, discussed in further detail below: 1) quantification of cell subtype composition to identify nuanced differences in detailed regions within the cortex; 2) identification of regionally specific gene expression programs by analyzing region specific subtypes of interneurons; and 3) in the case of minimal region specific subtypes, leveraging a recently developed statistic to identify region-specific gene expression patterns in astrocyte subtypes in a cell subtype-agnostic fashion.

Within specific regions of the cortex, cell subtype composition differences become more subtle and require focused quantification. As a first approach, for every sufficiently abundant interneuron and astrocyte subtype in the cortex (>100 cells), we calculated the \log_2 transformed ratio of cell subtype composition in a region, compared to the average composition of that subtype across all cortical regions (**Fig. 4.2B**). Within the five most abundant interneuron subtypes, we note general balance across all cortical regions, but also observe a relative enrichment of cluster 2 (*PVALB+*) in the occipital lobe (primary visual cortex [V1]) with depletion in regions within the temporal lobe, and depletion of cluster 5 (*ADARB/PAX6+*) in V1. In the superior temporal sulcus (STS) and middle temporal visual area (MT), there is a strong depletion of astrocyte subtype 3 (*LUZP2/GPC5+*) but an enrichment of subtype 6 (*KCNIP4/RBFOX1+*).

Interneurons are the primary drivers of inhibitory control through the release of GABA (γ -aminobutyric acid) and thus strongly impact neural circuitry. Inappropriate development of GABAergic interneurons and subsequent loss of inhibitory regulation contributes to disorders of neurodevelopment, including epilepsy and autism [217, 218]. Despite their importance, the molecular identities and distribution of interneuron subtypes across the adult primate brain remain relatively unknown outside of a few regions [202, 214, 219]. Our snRNA-seq sample captured 371,548 GABAergic interneurons corresponding to 20 subtypes. As a second approach, we focused on gene markers of the region specific interneuron subtypes.

Eleven interneuronal subtypes were primarily found in the cortex and could be assigned to four primary interneuron groups that are conserved between mouse and human brains [214], marked by *SST*, *PVALB*, *VIP*, and *LAMP5* expression (**Fig. 4.15**). Compared to the cortex, the brainstem and thalamus had a unique distribution of interneuron subtypes (**Fig. 4.2C**). Thalamic interneurons, which use feed-forward inhibition to relay and tune visual responses to thalamocortical neurons, expressed high levels of *NTNG1* and *RNF220* (**Fig. 4.2D**), which is indicative of long-range interneurons in the first-order relay nuclei of the thalamus ([220]. Sampling across the striatum, which is a critical part of the reward pathway and the largest part of the basal ganglia, a recent single-cell study identified a molecularly unique primate interneuron [214], which was most similar to our GABAergic cluster 18 and represented 15% of interneurons in the CN (**Fig. 4.2A**).

Astrocytes, the second most abundant non-neuronal cell type in our dataset, are multifaceted support cells of the brain that perform a variety of tasks related to neuronal homeostasis. These tasks can vary across brain regions [215] and astrocyte dysfunction has been linked to neurological diseases, including Alzheimer’s disease [221]. Given these regional differences, we examined whether astrocyte subtypes exhibited regional biases in macaque, similar to what has been observed in the mouse brain [222]). However, while astrocyte subtypes were widely distributed across multiple regions, the cell clusters did not correspond neatly to regions of origin, making claims about inter-region differences in cell composition difficult to systematically analyze across the many regions profiled. To address this complexity, as a third approach, we adapted our recently developed statistic, lochNESS [223], to quantitatively measure regional enrichment in each cell’s “neighborhood” of transcriptionally similar cells. Briefly, for each cell, we tally the number of cells from each brain region in its neighborhood and calculate a focal regional enrichment score (**Fig. 4.16A, Methods**). We illustrate the utility of this approach by calculating the lochNESS score on astrocytes at the level of brain region subclasses. Each cell had 11 lochNESS scores calculated, one for each region subclass, with each such score quantifying the enrichment of the given region subclass

in a cell’s transcriptional vicinity. We then identified the most enriched region subclass in a cell’s neighborhood and examined the regional heterogeneity agnostic to the cluster-assigned subtype labels (**Fig. 4.2E**). We also extended the lochNESS to identify genes whose expression can be predicted by lochNESS scores for given regions. To do so, we modeled the lochNESS score of each region in each cell as a function of gene expression with generalized linear regression (**Methods, table S7**). The resulting set of genes with significant positive associations with a region’s lochNESS score have higher expression in, and are putatively markers for, cell subtypes in that region.

Using this approach, we identified markers for astrocytes in specific regions (e.g., *TCAF2* and *FRK* in the occipital lobe) and in combinations of regions (e.g., *PGD* in the brainstem, basal ganglia, and thalamus), that we would not have identified if we focused solely on discrete, computationally-defined clusters (**Fig. 4.2F**). This strategy thus facilitates the identification of more complex region-specific gene expression patterns. For example, *EMID1*, which is a marker for a subpopulation of astrocyte-like NG2 cells [224], is more highly expressed in astrocytes in the cortex but not in the thalamus, brainstem, or cerebellum. In contrast, *ADAP2*, which is involved in protection from RNA virus infections [225], is highly specific to a subset of astrocytes found in the thalamus (**Fig. 4.2F**). LochNESS can thus provide a more nuanced approach to identifying regionally-biased cell subtypes and gene expression than conventional clustering. While we focused on astrocytes in this example, lochNESS could be iteratively applied to regions within a subclass in each cell class, e.g. for all glutamatergic neurons across all cortical regions or oligodendrocytes across all subcortical regions (**Fig. 4.16B-D**).

4.2.3 Joint analysis of single-nucleus transcriptomic and epigenomic data

To complement our transcriptomic dataset and identify key regulatory genomic regions in brain cells, we applied sci-ATAC-seq3 [226, 16]) to profile single-nucleus ATAC sequencing

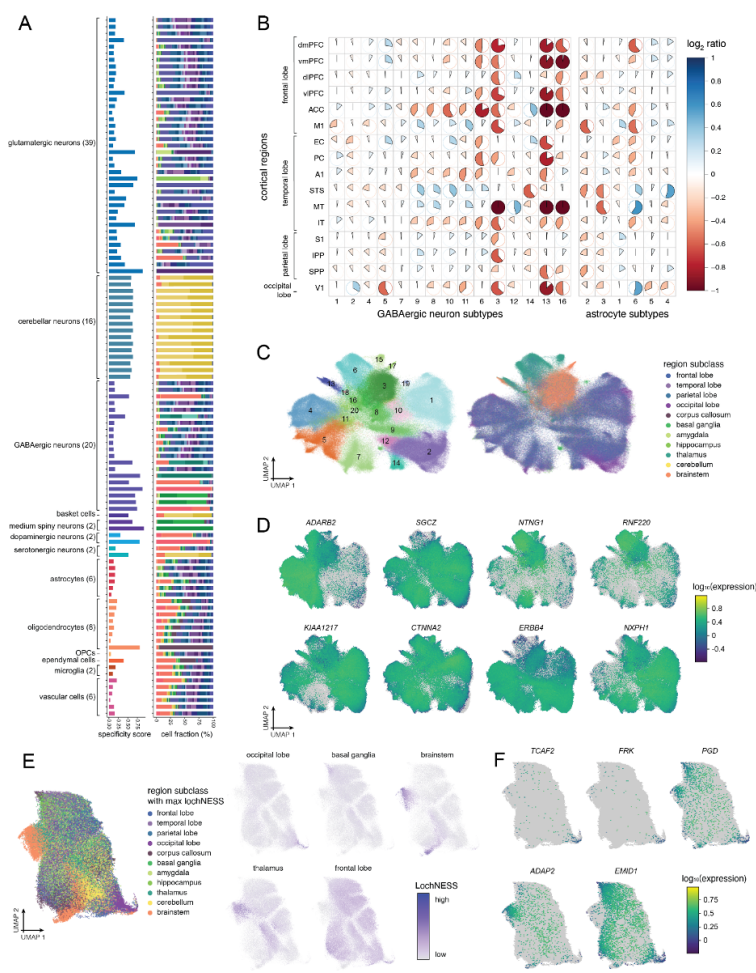


Figure 4.2: Fig. 2. Cell subtype distribution and variation across the brain. A, Barplots showing the region specificity score (i.e., Jensen-Shannon divergence statistic) and composition for cell subtypes (with color code shown in Fig. 4.1E). **B**, Heatmap showing scaled \log_2 ratios of GABAergic neuron and astrocyte subtype compositions within cortical region, compared to the average across all regions. Cell subtypes with at least 100 cells profiled are shown in the order of abundance (x-axis, left to right) in the cortical regions organized by region subclasses (y-axis). The color and direction of each pie corresponds to relative enrichment (blue, clockwise) and depletion (red, anti-clockwise) of a cell subtype in a region. \log_2 ratios were capped at positive and negative 2 prior to scaling. **C**, UMAP visualizations of GABAergic neurons colored by cell subtype (left) and regional subclass (right). **D**, UMAP visualizations of GABAergic neurons colored by cell subtype marker gene expression. **E**, UMAP visualization of astrocytes colored by the region with the highest lochNESS, indicating enrichment of a region subclass in the cell's transcriptional vicinity. LochNESS distribution in a few example regions are highlighted in separate panels as examples. **F**, UMAP visualizations of astrocytes colored by lochNESS-derived region-related marker genes.

(snATAC-seq) epigenomes from nearly all of the brain regions represented in our snRNA-seq dataset. To maximize comparability among datasets, we used 110 of the same age-, sex-, and hemisphere-matched tissue samples (representing the same three animals) profiled in our snRNA-seq dataset. To ensure that the snRNA-seq and snATAC-seq datasets captured the same heterogeneous populations of cells, we homogenized tissue samples on dry ice prior to separately preparing separate nuclei isolations for each library type (**Methods**). Together, the snATAC-seq samples represented 28 of the 30 regions (n=3 animals; midbrain [MB] and MT snATAC-seq data were not generated). After quality control (**Methods, Fig. 4.17**), the total number of nuclei profiled was 1,587,880 and ranged from 5,100 (in the closed medulla [MdC]) to 114,410 (in the inferior temporal cortex [IT]) nuclei per region (median=63,739 nuclei per region). We called peaks on a per-sample basis and combined them across all samples based on genomic overlap, resulting in (after filtering) a combined set of 1,192,873 candidate *cis*-regulatory elements (cCREs) spanning 24.4% (725 Mb) of the genome.

We first applied UMAP dimensionality reduction and Leiden clustering to the batch-corrected epigenomic data (**Fig. 4.3A**) and identified 42 clusters which, based on promoter accessibility, could be assigned to most major cell classes found across the brain (**Fig. 4.3B**). However, given that unsupervised approaches to cell-type identification are consistently more sensitive using single cell/nucleus RNA-seq data [33], we drew from our transcriptionally defined cell annotations in order to assign cell labels to our snATAC-seq nuclei. To integrate the datasets, we used the graph-linked unified embedding (GLUE) approach [35] and generated a unified transcriptomic and epigenomic embedding of 4,171,847 nuclei (**Fig. 4.3C-D**). Subsequent cell-type predictions based on our multimodal integration assigned the majority of snATAC-seq nuclei to a cell class (73.7% with confidence ≥ 0.95 ; **Fig. 4.18**), and captured all of the major cell classes (**Fig. 4.3D**) with the exception of serotonergic and dopaminergic neurons, which are relatively rare and fairly specific to the MB (which as noted above was not sampled in our snATAC-seq data). The regional distribution of cell classes captured from snATAC-seq and snRNA-seq data were highly concordant, both within regions (**Fig.**

4.3E) and overall (**Fig. 4.3F**), which demonstrates that our homogenization and nuclei isolation protocols captured the same heterogeneous populations of cells in the same regions across both modalities.

4.2.4 *The gene regulatory landscape of the rhesus macaque brain*

We leveraged the scRNA-based cell class annotations (**Fig. 4.3G**) to explore heterogeneity in cell type-specific gene regulation across the brain. To do so, we partitioned all unique snATAC-seq reads by predicted cell class (**Fig. 4.3G**), then called peaks separately for each partition using a similar peak-calling approach to that used for the overall dataset, thereby generating an inventory of putative cCREs derived from each cell class in isolation (**Methods**). Across 11 cell classes with snATAC-seq-assigned nuclei, we identified an average of 210,572 peaks per cell class, ranging from 99,323 in microglia to 425,738 in cortical GABAergic neurons (**Fig. 4.4A**). On average for any given cell class, these peaks covered 7.7% of the genome and 28.8% were found >2 Kb from the nearest gene or promoter (**Fig. 4.4A**).

Transcription factor regulatory networks

Multi-modal integration of cell specific snATAC-seq and snRNA-seq data allowed us to examine the *cis*- and *trans*-regulatory links between chromatin accessibility and gene expression within individual cell types. We first examined putative *trans*-regulatory factors within cell classes and subtypes. Transcription factors (TFs) are key *trans*-regulatory proteins that control cell differentiation and function during neurodevelopment [227, 228, 229, 230, 231, 232, 233]. The extremely high cell-type specificity of some nuclear TFs have also made them useful targets for identifying and enriching rarer cell types prior to single-cell sequencing (PsychENCODE [234, 235, 236]).

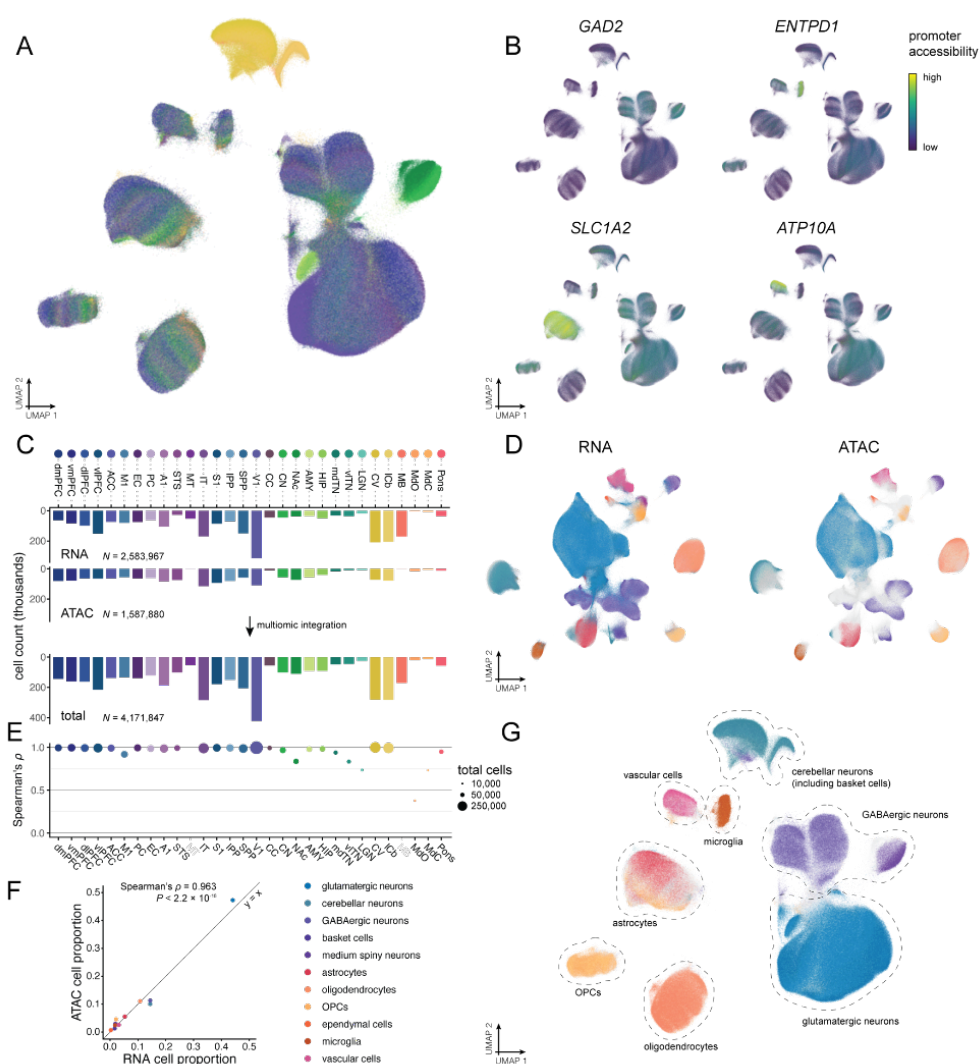


Figure 4.3: Fig. 3. Generation of the Macaque Brain Atlas sci-ATAC-seq dataset and identification of cell classes. **A**, UMAP visualization of all snATAC-seq cells colored by brain region (with color code shown in C). **B**, UMAP visualizations of promoter accessibility scores of cell markers (*GAD2*: GABAergic neurons, *ENTPD1*: microglia, *SLC1A2*: astrocytes, *ATP10A*: vascular cells) reveal high specificity. **C**, Barplots showing nuclei counts by brain region of the snRNA-seq, snATAC-seq, and integrated datasets. **D**, UMAP visualizations of integrated multimodal data, with cell classes colored separately for (left) snRNA-seq and (right) snATAC-seq nuclei (with color code shown in panel F). **E**, Spearman's rank correlation coefficients showing the correlation between cell-class proportions in the snRNA-seq and snATAC-seq datasets within each region (representing data generated from the same homogenized sample). **F**, Scatterplot showing the correlation between cell-class proportions in the overall snRNA-seq and snATAC-seq datasets (combined across brain regions). **G**, Integration-derived cell-class annotations visualized over the same snATAC-seq UMAP visualization shown in panel A (with color code shown in panel F).

To identify candidate *trans*-acting regulatory networks in each cell class, we carried out TF binding motif enrichment analysis on each set of cell-class-specific peaks, defined as the subset of a cell class’s cCREs that did not overlap with any peaks called in other cell classes (**Methods, Fig. 4.4A**). Cell-class-specific cCREs were highly enriched for many TF binding motifs that are likely involved in cell-specific gene regulation (**Fig. 4.4B, table S8**), including many motifs previously implicated (**Fig. 4.4C**). For instance, microglial cCREs contained 6.6-fold more binding sites of the nuclear TF SPI1 (also known as PU.1) than expected by chance ($P_{adj}=1.22\times 10^{-284}$; **Fig. 4.4B-C**). In addition to such canonical examples, we identified numerous motifs that distinguish relatively similar cell classes. For instance, the TF binding motif for NFE2, from the NRF TF family, was most enriched (odds ratio [OR] > 2) in cCREs in both medium spiny neurons (OR=3.07, $P_{adj}=1.77\times 10^{-87}$) and basket cells (OR=2.34, $P_{adj}=8.76\times 10^{-7}$), while the binding motif for NEUROD1 was most enriched (OR>2) in cCREs of basket cells (OR=2.04, $P_{adj}=3.53\times 10^{-29}$), where this TF is necessary for basket cell terminal differentiation and, consequently, axon growth and inhibitory circuit formation [237].

We also characterized TF binding motif enrichment at the cell subtype level. To do this, we extended our multimodal integration and label-transferring approach to each cell class independently by tabulating the reads per-cell falling within cell-class-specific cCREs described above for all cells of a given cell class (**Fig. 4.3D**). We then integrated the data with corresponding snRNA-seq data of the same cell class using GLUE (**Methods**). The resulting integrated embeddings for each cell class were then used as the basis for predicting cell subtypes, which we carried out on all snATAC-seq cells within each class (**Fig. 4.19**).

Since cell subtypes are preselected to already share broadly similar chromatin accessibility profiles, identifying peaks that are specific to a single subtype—similar to our approach at the cell-class level—was not feasible and left most cell subtypes with no or very few unique peaks to analyze. As an alternative strategy, we carried out differential accessibility analyses

among cell subtypes to identify peaks that were predictive of each individual cell subtype within a given cell class (**Methods**). We then identified TF binding motifs enriched in highly differentially accessible regions within cell subtypes (**table S9**). For example, we observed numerous TF binding motifs ($N=433$, $P_{adj} < 0.05$) that were enriched within highly accessible peaks in Purkinje cells, a GABAergic neuron type of the cerebellum that is implicated in autism spectrum disorders (ASD). In our snRNA-seq dataset, of all tested diseases [238], genes associated with autism (DOID:12849) were overrepresented (Fisher’s exact test, $OR=10.2$, $P_{adj}=8.52\times 10^{-16}$) among the top 100 Purkinje-cell marker genes, including *RORA* (fold-change [FC]=331.9), *AUTS2* (FC=43.1), and *SHANK2* (FC=13.8) (**table S10**). Correspondingly, we found that TF motifs enriched in differentially accessible peaks included RORA, four members of the EGR family (EGR1–EGR4), and CTCF (EGR1, EGR3, and CTCF were among the top 5 TF motifs ranked by OR; RORA ranked 182nd). RORA is a regulator of circadian rhythm that exhibits decreased expression in ASD brains and may play a role in ASD pathogenesis [239, 240]. EGR-family TFs have been implicated in the disruption of human-specific developmental programs in autism [241]. CTCF is an insulator protein that regulates chromatin structure and may play a critical role in maintaining dendrite structure in Purkinje cells [242] and is also a risk gene for ASD [243].

Given that families of TFs have similar binding motifs, it is often difficult to identify the specific TF in a given family that is responsible for enrichment in cell type-specific cCREs. To identify the most likely TF, we therefore employed our recently-developed approach [16, 244]) that uses the computationally paired snRNA-seq and snATAC-seq data. In brief, this approach relies on the assumption that TFs will be highly expressed in cell types where they play a key role, while their associated motif should be enriched (or depleted) in that cell’s cCREs, indicating TF activation (or repression). Overall, we compared the accessibility of 369 TF binding motifs and their corresponding gene’s expression across the cell classes in four region subclasses, with 189 TFs showing positive Pearson’s correlation between gene expression and accessibility of the cognate motif, and 180 showing negative correlation (**Fig.**

4.20A, table S11). Among the TFs with largest positive or negative Pearson’s correlation values were strong cell-class-specific activators and repressors (**Fig. 4.4D, Fig. 4.20B**). For instance, *SPI1*, which has been identified as a candidate gene for Alzheimer’s disease via various functional genetics approaches [245], shows a strong activating effect with high expression of the *SPI1* gene and high accessibility for the SPI1 binding motif in microglia. In contrast, *NFATC2* has a repressing effect in microglia and vascular cells, as shown by high expression of the *NFATC2* gene associated with lower *NFATC2* motif binding in those cell types. We also found evidence for a clear distinction between neurons and non-neuronal cells at two TFs, with *ELF1* functioning as a non-neuronal specific activator and *NEUROD2* as a neuron-specific activator. Additionally, we note that *FLI1*, an activator in vascular and microglia cell types, and *ELF1* have motif sequences similar to *SPI1* (**Fig. 4.20C**), but their activating effects impact a broader set of cell types.

The cis-regulatory landscape of brain cell variation

We next sought to characterize *cis*-regulatory interactions between cCREs and proximate genes in the rhesus macaque brain. We used two complementary analyses to scan for interactions using our integrated multimodal dataset. First, we used the regulatory inference framework of GLUE [35], which leverages the unified feature embedding (i.e., joint integration of snRNA-seq genes and snATAC-seq peaks in a common data space) generated during GLUE integration to assess similarity between peaks and genes. Putative regulatory interactions are defined as a high cosine similarity between peak and gene feature embeddings in the unified data space, with statistical significance assessed by permutation ([35]). Second, we used a metacell-based approach to aggregate snRNA-seq transcriptomes and snATAC-seq epigenomes into multimodal metacells based on *k*-means clustering of the unified cell embeddings, then used logistic regression to model the relationship between gene expression and chromatin accessibility within a given metacell [246]. In contrast to the GLUE regulatory score, the logistic regression analysis enabled us to differentiate between positive and

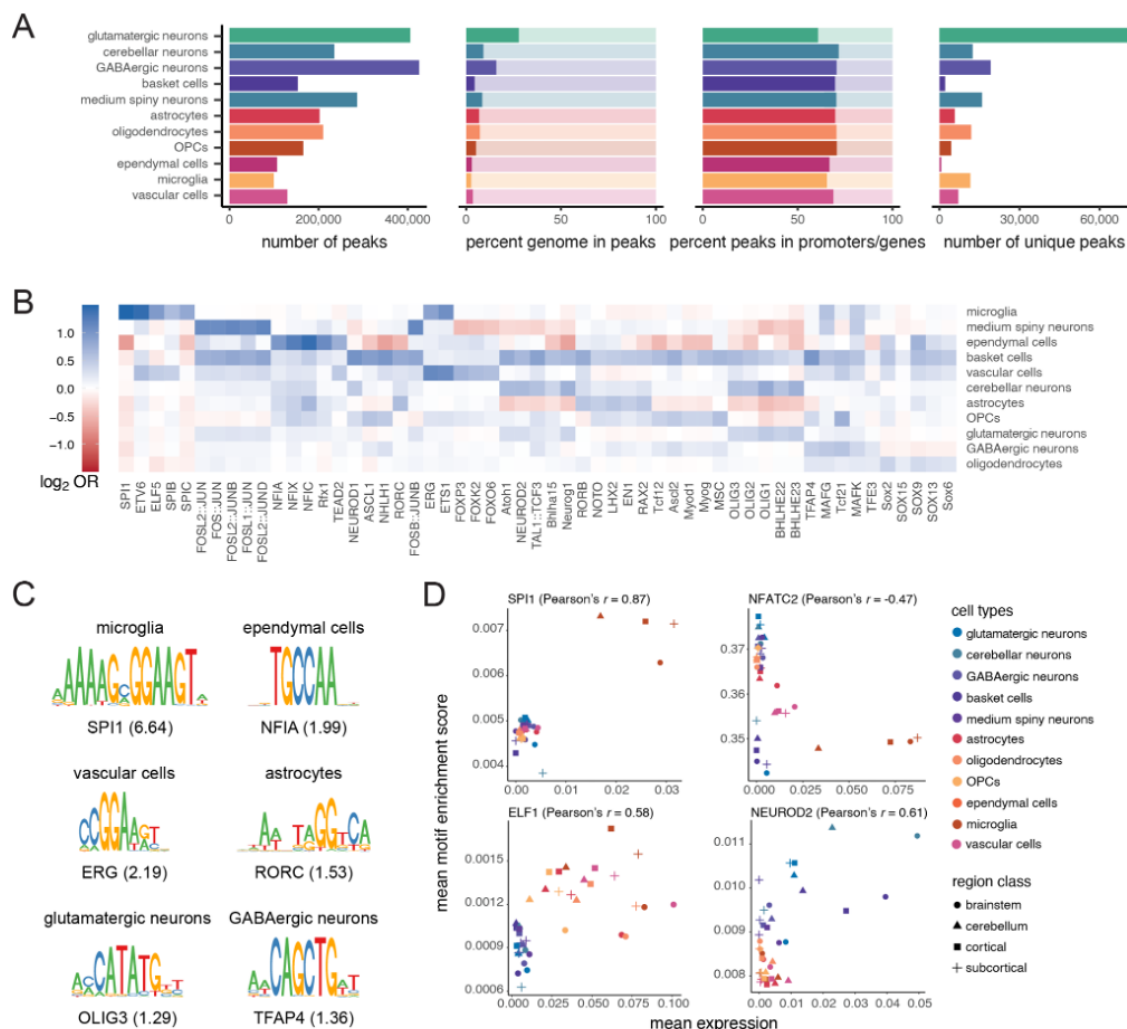


Figure 4.4: **Fig. 4. Enrichment of transcription factor binding site motifs in candidate regulatory elements.** **A**, Barplots showing summary statistics for peak sets called separately on reads derived from cells assigned to each of 11 cell classes. **B**, Heatmap showing enrichment (\log_2 odds ratio) [OR] of transcription factor (TF) binding motifs among cell classes. The top five most-enriched nonredundant TF motifs (all $P_{adj} < 0.05$) are shown per cell class, ordered from left to right by increasing P_{adj} . \log_2 OR color ranges are capped at ± 1.5 . **C**, Position weight matrices of the most-enriched TF motifs for six example cell classes. Odds ratios are shown in parentheses. **D**, Scatterplots showing correlation between snATAC-seq accessibility of TF binding motifs and snRNA-seq gene expression of corresponding TF genes within cell classes in regional classes for four example TFs.

negative regulatory interactions between peaks and genes. We considered peak-gene pairs to be putatively regulatory if $P_{adj} < 0.05$ for both analyses (**Fig. 4.5A**, **Fig. 4.21**). For each cell class, we also scanned for differentially accessible peaks using both a regularized logistic regression and a t -test, testing accessibility in a given cell class against accessibility in all other cell classes. We consider cCREs with differentially high accessibility (regularized LR coefficient > 0 , \log_2 fold-change > 0 , and t -test $P_{adj} < 0.05$, **Fig. 4.22**) to be candidate regulators of cell-type-specific genes (**Fig. 4.5A**).

We focused our analysis on the 6,000 most variable genes in our snRNA-seq dataset and tested all snATAC-seq peaks that fell within 150 Kb of the gene promoter (defined as TSS extended 2 Kb upstream). In total, we tested 223,752 peak-gene pairs (151,083 unique peaks, 5,765 unique genes), of which 142,324 peak-gene pairs (63.6%) met our criteria for being considered candidate *cis*-regulatory interactions (**table S12**). 128,741 peaks (85.2%) that we evaluated were cCREs for at least one gene and 4,811 genes (83.5%) that we evaluated had at least one cCRE.

Of all peak-gene pairs, 132,805 (93.3%) involved a peak that was highly differentially accessible in at least one cell class, thereby fulfilling our criteria for being considered candidate *cis* molecular interactions regulating cell-type-specific markers. cCREs were highly differentially accessible in a maximum of 7 cell classes, with 37% exclusive to a single cell class and 88% highly differentially accessible in 1–3 cell classes.

The vast majority (133,496, or 93.8%) of candidate regulatory interactions were positively associated (i.e., had a positive effect size in the metacell logistic regression)—this held true whether peaks were upstream (13,650/14,575, or 93.7%), downstream (116,939/124,592, or 93.9%), or overlapping (2,907/3,157, or 92.1%) the gene’s transcription start site (TSS). For peak-gene pairs where the peak was upstream of the TSS, the GLUE regulatory scores were highest (indicating high similarity between peak and gene feature embeddings) when peaks

were in closer proximity to the TSS (**Fig. 4.5B**). For peaks downstream of the TSS, GLUE regulatory scores remained high across all distances, with only a modest decrease farther from the TSS (**Fig. 4.5B**). This result was particularly striking for peaks that had significant, mainly positive, associations between accessibility and gene expression, likely reflecting (*i*) higher global accessibility across the gene body resulting from higher expression of the gene (as opposed to distal regulation) and/or, (*ii*) methodological limitations of using a single gene-wide TSS (i.e., the most upstream TSS of all isoforms), thereby ignoring variation in TSS positioning among isoforms, which likely vary in their usage across tissues and contexts [247].

Using the cell-class-specific gene expression and cCRE peak sets, we repeated our integration, regulatory inference, and differential accessibility workflows on each cell class individually. We tested a mean of 72,914 peak-gene pairs (range: 45,539–114,200) per cell class and identified a mean of 11,442 peak-gene pairs (range: 881–41,966) showing evidence of regulatory interactions (**Fig. 4.23** and **table S13**).

To illustrate how these maps of putative interactions might be useful to investigate the regulatory landscape at the level of an individual locus, we focused on the myelin basic protein (*MBP*) gene (**Fig. 4.5C**), which encodes one of the most abundant proteins in central nervous system myelin [248, 249], has a range of splice isoforms [250], and is a canonical marker of oligodendrocytes. *MBP* is located on chromosome 18 (positions 2,932,531–3,086,873) on the rhesus macaque (*Mmul_10*) genome and has 8 annotated mRNA isoforms (Ensembl). In humans, classic *MBP* isoform 3 (18.5 kDa) predominates in adult myelin [250].

In our global peak set (all cells), 94 peaks fell within 150 Kb of the *MBP* promoter and were included in our analysis. Of these peaks, 83 (88.3%) were identified as candidate regulators of *MBP* (crMBP), with 38 crMBPs (45.8%) positively associated with *MBP* expression. Of all crMBPs, only one was not located within the *MBP* gene boundaries—it was, however,

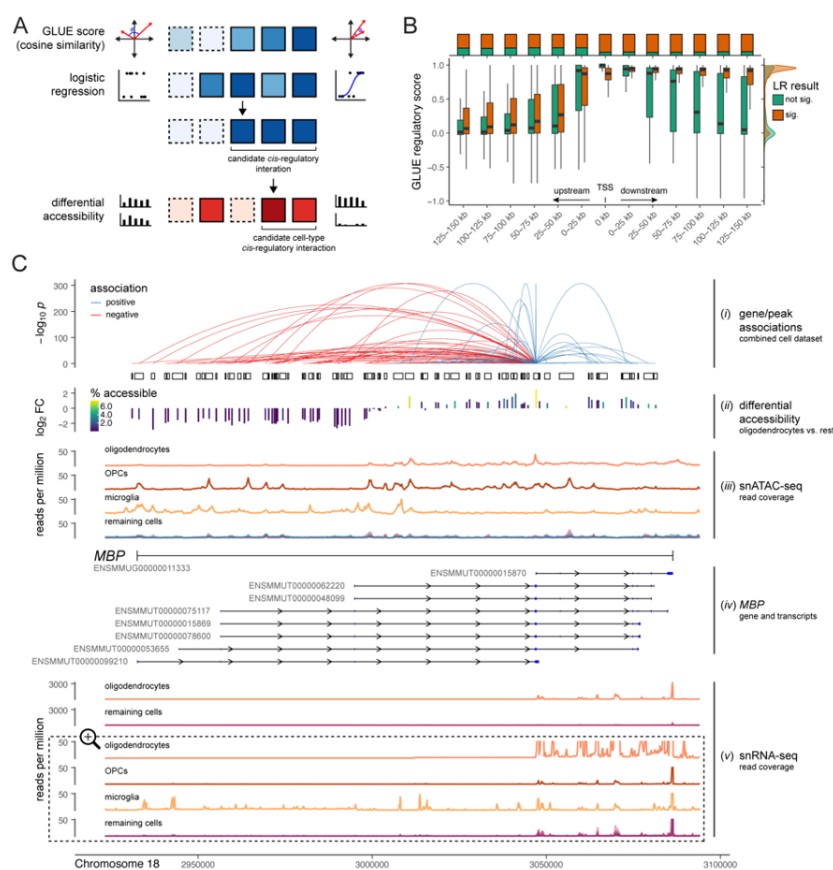


Figure 4.5: **Fig. 5. The landscape of *cis*-regulatory interactions in the Macaque Brain Atlas.** **A**, Schematic outlining criteria for identification of candidate *cis*-regulatory elements (cCREs). Squares represent peak:gene pairs, darker colors symbolize stronger evidence for a given measure, and solid borders represent statistically significant measures. **B**, Distribution of gene-peak GLUE regulatory scores binned according to the minimum signed distance (left: upstream, right: downstream) between peaks and gene transcription start sites (TSS). Distributions are shown separately according to whether the gene:peak pair also exhibited a significant association ($P_{adj} < 0.05$) based on the metacell-based logistic regression analysis. **C**, Candidate regulatory elements are shown in relation to, from top to bottom, (i) the strength of and direction of inferred regulatory links connecting peaks to *MBP* expression (based on metacell logistic-regression analysis). The height of links represents the strength ($-\log_{10} P$ values) of evidence for regulatory connections and the color symbolizes the direction of the relationship; (ii) the differential accessibility ($-\log_2$ fold change) of peaks in oligodendrocytes relative to all other cell classes; (iii) the distribution of normalized snATAC-seq reads by cell class; (iv) gene and transcript boundaries of *MBP* and its known isoforms in the rhesus macaque genome, with exons shown in blue; (v) the distribution of normalized snRNA-seq reads by cell class. Oligodendrocyte reads are shown in relation to all other cell classes on the upper portion of the plot. On the bottom portion, the y axis is magnified 60x and cropped to highlight more subtle differences among cell classes.

located less than 2 Kb upstream within the likely promoter region.

In accordance with the well-known status of *MBP* as an oligodendrocyte marker, we found that *MBP* was differentially expressed in oligodendrocytes, with detected expression in 80.9% of cells and 1,434-fold higher expression than all other cells averaged together. Fine-grained inspection of normalized read distributions from oligodendrocyte nuclei revealed the highest densities of snRNA-seq reads corresponding to the polyadenylation site (position 3,086,373) and snATAC-seq reads corresponding to the TSS (position 3,046,976) of a single transcript, ENSMMUT00000015870, indicating that it is likely the dominant *MBP* isoform expressed in adult macaque oligodendrocytes.

By examining the genomic-distance relationships between crMBPs and the dominant *MBP* transcript in adult oligodendrocytes, we found that all 16 crMBPs that either overlapped or were downstream of the isoform's TSS were positively associated with *MBP* expression. Among the 67 crMBPs that were located upstream of the TSS, 22 (32.8%) were positively associated with *MBP* expression while 45 (67.2%) were negatively associated. Several of these negatively associated crMBPs corresponded with sci-ATAC-seq3 peaks in other cell types, particularly oligodendrocyte precursor cells (OPCs) and microglia (**Fig. 4.5C**). However, the accessibility landscape of OPCs is overall more similar to that of oligodendrocytes across the region upstream of the TSS of the dominant isoform, with greater accessibility at most peaks except for that of the promoter of the dominant isoform (**Fig. 4.5C**). As OPCs play a critical role in myelinogenesis by giving rise to oligodendrocytes [251], these crMBPs likely serve as critical markers of the OPC-oligodendrocyte transition, during which the expression of this gene, and this isoform in particular, is massively upregulated.

Enrichment of disease heritability among candidate regulatory elements

Lastly, we used our cCREs to identify cell-type-associated regulatory networks that may drive polygenic disease risk. We tested for enrichment of disease trait heritability using the linkage disequilibrium score regression (LDSC) tool [252, 253], after lifting over macaque cCREs to human genome coordinates [211]. We tested a total of 53 phenotypes relevant to neurological diseases, disorders, syndromes, behaviors, or other traits (**table S14**), and examined enrichment among cell-class cCREs called separately in each of 11 cell classes.

Our results broadly recapitulated several known roles of cell classes in neurological disease (**Fig. 4.6** and **table S15**). For example, sites associated with cardioembolic stroke (OR=32.2) or ischemic stroke (OR=9.2) were enriched ($P_{adj} < 0.05$) only in vascular cells, which play a crucial role in forming and maintaining the blood-brain barrier [254]. We also found that Alzheimer’s disease-associated sites were enriched only in microglia—a result replicated using loci from three independent genome-wide association studies (GWAS) (OR range: 13.9–15.0)—consistent with the prominent role of microglia proliferation and activation in Alzheimer’s disease [255].

Across all cell classes, basket cells were enriched for the greatest number ($N=37$) of GWAS phenotypes, including disorders such as schizophrenia (OR range: 5.9–6.2), bipolar disorder (OR range: 5.6–6.2), and major depressive disorder (OR range: 5.1–5.3), and, most strongly, epilepsy (OR=9.0)—a disease that basket cells have been connected to in animal models and some genetically-linked human forms of the disease [256].

Other notable results included the enrichment of multiple-sclerosis-associated sites among open regions in microglia (OR=46.6), highlighting the outsized role of these immune cells in the etiology of multiple sclerosis and as a putative therapeutic target [257, 258]. In multiple sclerosis, disease-associated microglia alter their transcriptional profiles and may

contribute to neuroinflammatory processes underpinning this autoimmune disorder [258]. We also found enrichment of Parkinson’s disease-associated sites among open regions in the glial OPC, oligodendrocyte, and astrocyte cell classes (OR range: 7.0–8.4). In Parkinson’s disease, glial cells may play a major role in the progressive degeneration of dopaminergic neurons [259], a classic hallmark of Parkinson’s disease, or in alterations to glutamatergic neurotransmission [260].

Finally, we found that heritable sites associated with attention deficit/hyperactivity disorder (ADHD) in our analysis were enriched only among open regions of medium spiny neurons. While the magnitude of the enrichment was relatively mild (OR=2.6, P_{adj} =0.031), genetic variants associated with ADHD have been historically difficult to identify, with the first risk loci only recently reported [261]. Medium spiny neurons have been linked to behavioral hyperactivity and disrupted attention via activation of astrocyte-mediated synaptogenesis [262]. Our results therefore suggest that medium spiny neurons may be a promising target for future ADHD-related study.

4.3 Discussion

Understanding the cellular architecture of the adult primate brain is crucial both for understanding the evolution of human cognition and behavior as well as for identifying mechanisms underlying neurological disorders. In service of these goals, we used snRNA-seq and snATAC-seq to derive a molecular atlas spanning the adult rhesus macaque brain, comprising data from over 4 million cells profiled from 30 brain regions. Based on our multimodal molecular data, we identified 112 distinct molecular cell types or subtypes and characterized their distribution across the macaque brain, adding to the growing number of primate single-cell molecular brain atlases [202, 214]. The data are freely available (NeMO archive, nemo:dat-rtmm5q2) and will serve as a rich resource for the neuroscience and neurogenomics communities.

In generating a multi-region transcriptomic and epigenomic atlas of the most widely used nonhuman primate in neuroscience, we: (i) identified all of the major brain cell classes and many cell types that have been previously reported (**Fig. 4.1, Fig. 4.2**); (ii) quantified regional distribution of cell types and subtypes within individuals, which allowed us to identify compositional differences in samples collected at the same time and from the same animals (**Fig. 4.2**); (iii) identified rare and regionally specific cell types (e.g., Purkinje cells), which may facilitate the development of molecular tools such as cell type-specific viral vectors that, in combination with new technologies such as CellREADR (Cell access through RNA sensing by Endogenous ADAR) [263] may enable precise targeting of cell types based on their unique patterns of chromatin accessibility and gene expression; (iv) characterized multiple *trans*- and *cis*-regulatory mechanisms that differentiate cell classes and subtypes (**Fig. 4.4A-D, Fig. 5**); and (v) identified numerous associations between genetic risk for neurological disorders and the epigenomic states of specific cell types (**Fig. 4.6**).

This single-cell atlas of the adult primate is notably generated from samples collected from healthy adults. The paired nature of the dataset, with regions sampled from the same individual brains, avoids many of the inter-individual variables (e.g., genotype, environment) that can impact neurological development and function. The atlas may thus be a valuable resource for characterizing molecular features that play a role in myriad neurological disorders. The relatively few unique individuals sampled also represents a limitation of the current study—we currently know very little about how brains of healthy individuals differ in cell composition and function and what that confers for disease susceptibility and/or progression. Given continuing improvements in cost and throughput of single-cell sequencing, characterizing multi-region cellular variation across many healthy individuals is becoming not only a possibility, but also an emerging priority for the field.

To our knowledge, these data represent the largest and most comprehensive multimodal molecular atlas in a primate to date and provide a resource for exploring how the heteroge-

neous molecular and cellular composition of the brain gives rise to the behavioral complexity of primates including humans. We anticipate that these data will also provide a critical and much-needed molecular and neurobiological map of complex human-relevant social behavior and disease, as well as an extensive substrate for comparative analyses across animal brains.

All code for this project is available through the following GitHub repositories: <https://github.com/bbi-lab/bbi-dmux> (sci-RNA-seq3 demultiplexing), <https://github.com/bbi-lab/bbi-sciatac-demux> (sci-ATAC-seq3 demultiplexing), <https://github.com/bbi-lab/bbi-sci> (sci-RNA-seq3 preprocessing up to count matrix generation), <https://github.com/bbi-lab/bbi-sciatac-analyze> (sci-ATAC-seq3 preprocessing up to count matrix generation), and <https://github.com/CayoBiobankResearchUnit/macaque-brain-atlas> (remainder of the analyses).

4.4 Materials and methods

The detailed materials and methods are available in the supplementary materials. Briefly, we collected fresh-frozen brains from 5 adult rhesus macaques that were part of the free-ranging Caribbean Primate Research Center research colony on Cayo Santiago. We focused our atlas on 30 anatomically defined regions that are associated with key cognitive, behavioral, and disease traits. To allow for the profiling of multiple genomic modalities from the same representative cell populations, we pulverized all samples on dry ice to homogenize and divide tissue for single nucleus sequencing. We generated single-nucleus RNA-seq data from 2,583,967 nuclei spanning a total of 30 unique regions from both hemispheres of the brain, and paired those data with single-nucleus ATAC-seq data from 1,587,880 28 regions across 28 unique regions. These data were generated using sci-RNAseq3 [2] and sci-ATACseq3 [16]) combinatorial indexing. Single-nucleus libraries were deeply sequenced and processed using a uniform protocol that included extensive QC filters (**Fig. 4.7,4.8**).

Using Leiden-clustering on snRNA-seq nuclei [2], we identified 17 primary cell classes and then iteratively clustered each cell class for deeper annotation of cell subtypes. Whenever external data were available, we validated our cell classifications using a non-negative least squares (NNLS) approach [2] to identify correlations between cell subtypes and annotated labels in reference datasets. We then identified marker genes for each cell class and subtype, characterized the regional distribution and expression of each cell class and subtype across the brain, and identified cell-specific enrichment of disease-associated genes.

To connect snATAC-seq profiles to snRNA-seq nuclei, we used the GLUE integration approach [35], which allowed us to annotate all snATAC nuclei based on the cell classes and subtypes identified in our snRNA-seq data. These connections allowed us to carry out a range of analyses, including TF binding site enrichment, linking TF enrichment to and TF expression within cell types, and identifying cell-specific regulatory links between candidate cis-regulatory elements (cCREs) and nearby genes. Lastly, following coordinate liftover between the primate and human genomes, we used LDSC [252, 253] to quantify enrichment of neurological disease-associated variants in cell class biased cCREs.

Raw sequencing data and the annotated count matrices are available through NeMO (RRID:SCR_002001), protocols for data generation are on protocols.io (DOI:10.17504/protocols.io.9yih7ue and DOI:10.17504/protocols.io.be8mjhu6); and scripts to process samples and recreate all analyses are available on GitHub (**Data and Materials Availability**).

Study population and sample collection

All animals sampled in this study are rhesus macaques (*Macaca mulatta*) from the semi-free-ranging colony on the island of Cayo Santiago, Puerto Rico. Maintained by the Caribbean Primate Research Center (CPRC) within the University of Puerto Rico, the Cayo Santiago macaque colony has been largely continuously studied since its founding in 1938 [264]. All present-day macaques are descended from an initial founder population of 409

animals and have since maintained an outbred population structure despite generations of isolation [265]. Apart from being provisioned with commercial feed and occasionally subject to capture-and-release sampling, the macaques otherwise live in naturalistic conditions, subject to minimal intervention and manipulation, as approved by IACUC. The study used animals that needed to be removed from Cayo Santiago [266] and were immediately euthanized. Standardized tissue collection and sample archiving was coordinated by the Cayo Biobank Research Unit (CBRU), which provided the brain samples used in this study [267, 268].

Procedures for necropsy, brain removal, and dissection followed those previously described for this population [268] and are briefly outlined here. Following veterinary euthanasia, brains were perfused with sterile saline, removed from the cranium, and hemisected into left and right hemispheres using a long single-edge razor blade. After sectioning off the cerebellum/brainstem from each hemisphere, the cerebral hemispheres were placed on custom molds (designed either for left or right hemispheres) and coronally sectioned into 11 roughly 5-mm-thick blocks, numbered in order rostral to caudal. All 12 blocks (with the cerebellum/brainstem considered block 12) were then sealed in Whirl-pak bags, flash-frozen in liquid nitrogen vapor, and archived in ultralow -80°C freezers. The interval between euthanasia and permanent storage of frozen tissue averaged 51 minutes, with a standard deviation of 5.8.

All procedures were performed in accordance with the NIH Guide for the Care and Use of Laboratory Animals and were approved by the Institutional Animal Care and Use Committee at the University of Puerto Rico (protocol #338300). Five macaques were included in this study (**table S2**). The vast majority of the data derived from four 10-year-old macaques, which are considered middle-aged adults in this population [269, 268].

Region selection and biopsy

Frozen brain blocks were placed on a dissection tray over dry ice in order to keep tissue frozen during biopsy collection. Individual blocks were then moved from the dry ice to a tray sitting on wet ice, allowing for tissues to be acutely warmed to the point that biopsies could be taken from targeted structures. Biopsies were made using a cutting spoon (Fine Science Tools, Inc., cat. #10360-13). Dissected brain regions are listed in **table S1** and approximate locations for biopsy are illustrated in **Fig. 4.1A**. For a given structure, attempts were made to minimize inclusion of off-target surrounding tissues (e.g., white matter underlying a targeted gray matter structure). Below, we document the most common block numbers where structures were located. Due to interindividual differences and/or variation in sectioning, regions of interest were sometimes identified and dissected from adjacent blocks based on neuroanatomical landmarks. Alternate block numbers are therefore also documented below.

The most anterior block sampled for this study (block 2) contained gray matter for the dorsomedial (dmPFC), ventromedial (vmPFC), dorsolateral (dlPFC), and ventrolateral prefrontal cortices (vlPFC). dmPFC and vmPFC were defined as being on the medial side of block 2. The dmPFC biopsy was pulled from the gray matter in the ~top half of the medial edge of the block. A space along the medial edge was left to separate dmPFC from vmPFC. The vmPFC biopsy was pulled from the medial ventral half of the tissue block. Biopsies of dlPFC came from the cortical tissues surrounding the dorsal lateral portion of the block that included the superior and inferior portions of the principal sulcus. Samples from vlPFC came from the ventral and lateral portion of the block. As was the case on the medial side, a portion of the cortex was left between each lateral biopsy to avoid overlap (**Fig. 4.1A**).

Block 3 (sometimes 4) contained biopsies for the anterior cingulate cortex (ACC), corpus callosum (CC), and head of the caudate nucleus (CN). The biopsy for ACC was the gray matter sitting between the CC, which is ventral to ACC and the cingulate sulcus (cs), which sits dorsal to the cingulate gyrus. CC was defined as the white matter track sitting ventral to the ACC and medial to the lateral ventricle. The CN was the gray matter sitting ventrolateral

to the lateral ventricle and surrounded on all other sides by white matter. The CN was the only biopsy in the second block that was scooped out of the block face to minimize inclusion of any white matter sitting anteriorly past the CN within the block (**Fig. 4.1A**).

Block 5 (sometimes 4 or 6) contained the amygdala (AMY), entorhinal cortex (EC), perirhinal cortex (PC), and nucleus accumbens (NAc). The NAc is located ventral to the caudate, internal capsule, and putamen (Pu). Furthermore, in fresh-frozen tissue, there was a slightly darker color to the NAc. The tissue making up the NAc was scooped out of the block face. Similarly, the AMY was identified as ventral to the Pu, medial to the ventral portion of the claustrum, and dorsal to the EC. The AMY was also scooped out to minimize the inadvertent collection of neurons within the hippocampus (HIP). Finally, the EC and PC were collected, the delineation between the two was the rhinal fissure (**Fig. 4.1A**).

Blocks 5–6 (sometimes 4 or 7) contained tissue that were biopsied to represent cortical regions primary motor cortex (M1), primary somatosensory cortex (S1), primary auditory cortex (A1), superior temporal cortex (STS), and inferior temporal cortex (IT). Subcortical structures that were biopsied included mediodorsal thalamic nucleus (mdTN), ventrolateral thalamic nucleus (vlTN), lateral geniculate nucleus (LGN), and hippocampus (HIP). The delineation between M1 and S1 was the central sulcus and were taken from the approximate central third of the lateral portion of each respective gyrus. Within a case, attempts were made to biopsy from approximately the same putative mototopic and somatotopic regions. A1 biopsies were taken from the dorsal portion of the superior temporal gyrus which is within the ls (i.e., inferior operculum). The gray matter forming the STS sits ventral to the superior temporal gyrus and dorsal to the inferior temporal gyrus. IT was defined as the gray matter forming the lateral portion of the inferior temporal cortex. mdTN sits bilaterally on midline, within the thalamus. It is bound by ventricles dorsally, laterally by the centrolateral thalamic nucleus and ventrally by the centromedial thalamic nucleus. vlTN is bound by the centrolateral thalamic nucleus medially, body of the caudate nucleus (CN)

dorsally, and the reticular thalamic nucleus laterally. Biopsies for mdTN were taken from the central and central medial portions of the nucleus, while vITN biopsies were taken from the central portion of the nucleus. In both cases, this was in an effort to avoid inclusion of other thalamic nuclei. The LGN is a 6 layered structure that is easily observed on the coronal face of fresh-frozen slabs. When observed, the biopsy was scooped out. Like the LGN, the HIP was defined by its classic cytoarchitectonic features within the medial temporal lobe. For biopsies, efforts were made to not include EC, which sits ventral and ventromedial to HIP (**Fig. 4.1A**).

Block 7 (sometimes 6 or 8) contained tissues representing the superior posterior parietal (SPP), inferior posterior parietal (IPP), and area MT (MT). SPP biopsies were from the gray matter of the superior lobule. The intraparietal sulcus sits between SPP and IPP. Therefore, IPP biopsies were taken from the gray matter of the second, more lateral lobule. Finally, area MT was defined by the gray matter of the insular cortex, bound on its medial edge by white matter of the extreme external capsule and laterally by the superior and inferior operculum divided by the superior temporal sulcus (**Fig. 4.1A**) [270, 271].

The final cerebral block, block 11, contained the visual cortex. Biopsies from primary visual cortex (V1) were taken from the dorsolateral surface gray matter above the external calcarine sulcus (**Fig. 4.1A**).

The hemisected cerebellum/brainstem block was dissected as follows. First, the cerebellum was dissected off and the cerebellar vermis (CV) was separated from the lateral cerebellar cortex (lCb). Next, the remaining brainstem was dissected such that the midbrain (MB) block was separated by making a cut from just behind the inferior colliculus to the top of the basilar pons. Next, the pons was separated from the medulla by making a cut from the stria medullaris (approximate center of the fourth ventricle) to the base of the pons. A final cut at the base of the fourth ventricle to separate the open medulla (MdO) from the closed

medulla (MdC) (**Fig. 4.1A**).

To allow for the profiling of multiple genomic modalities from the same representative cell populations, we pulverized all biopsies on dry ice to homogenize and divide tissue for downstream experiments. We followed the tissue pulverization procedures described by [16] to achieve a powder consistency on a sterile aluminum foil work surface. Once sufficiently pulverized, we stirred the sample thoroughly, then divided the sample using the folded edge of foil as a funnel into new 1.5 ml pre-chilled and pre-labeled microcentrifuge tubes. Foil and tubes were set on aluminum trays or tube racks set on dry ice to keep powdered tissue frozen throughout this process. We divided samples into roughly a 2:1 ratio given the expected efficiencies/yields for single-nucleus RNA-seq and single-nucleus ATAC-seq protocols, respectively. Pulverized tissue was stored at -80°C up until processing for downstream library preparation procedures.

snRNA-seq data generation

To profile single-nucleus gene expression, we performed single-nucleus RNA-seq (snRNA-seq) using the three-level single-cell combinatorial indexing RNA-seq (sci-RNA-seq3) approach [2], which is the improved version of the original sci-RNA-seq protocol [20].

For two out of the three experimental batches in our dataset, we used a protocol closely adhering to the sci-RNA-seq3 protocol described by [2]. For the third batch, we used the improved protocol (“tiny sci”) described by [205]. Sample order was randomized between the first two batches, and within the third batch, to minimize batch effects and other technical artifacts. For the first two batches, we slightly modified the protocol described by [2] for a different tissue type and smaller input amounts. Briefly, we added 50 μl of cell lysis buffer to pulverized tissue in a 1.5 ml microcentrifuge tube, then homogenized the tissue using 5–10 strokes with a disposable RNase-free plastic pestle (Fisherbrand, cat. #12-141-364). We then added another 950 μl of cell lysis buffer, mixed by pipette, then transferred

the suspension through a 70 μm cell strainer (pluriSelect cat. #43-10070-70) into a 15 ml conical tube containing 5 ml ice-cold 4% paraformaldehyde. Nuclei were fixed in 4% paraformaldehyde for 15 min with occasional mixing, washed once in 1 ml ice-cold nuclei wash buffer, then suspended in 200 μl nuclei wash buffer. Nuclei were counted by mixing with 1 μM of YOYO-1 iodide (ThermoFisher cat. #Y3601) using a Countess II FL automated cell counter (Life Technologies), divided into tubes in 100 μl aliquots, then flash-frozen in liquid nitrogen.

For nuclei fixed with paraformaldehyde, library construction was similar to the sci-RNA-seq3 method from [2] with minor modifications including the substitution of Quick Ligase (NEB) for 10 minutes at 25°C for the second index step, instead of T4 DNA ligase (NEB) for 180 minutes at 16°C. For tagmentation, we used N7 adaptor-loaded Tn5 from QB3 MacroLab at the University of California Berkeley in tagmentation buffer (2X TD) as previously described in [86]: 20 mM Tris-HCl, pH 7.5, 10 mM MgCl₂, 20% (vol/vol) dimethylformamide (DMF). Libraries were sequenced on a NextSeq or NovaSeq platform (Illumina) (read 1: 34 cycles, read 2: 100 cycles, index 1: 10 cycles, index 2: 10 cycles).

For the DSP/MeOH nuclei isolations and library construction based on [205], we used hypotonic lysis buffer solution B (with BSA) for small volume tiny sci-RNA-seq3 nuclei isolation methods. For sci-RNA-seq3 library construction, we loaded \sim 20,000 nuclei per index 1 reverse transcriptase (RT) well in a 384 RT-well experiment with mouse and human brain added as separate quality control nuclei and nuclei from cell lines HEK293T (RRID:CVCL_0063) and NIH/3T3 (RRID:CVCL_0594) combined as barnyard controls per RT plate. Nuclei from all RT plates were pooled and redistributed to ligation plates for the second index as previously published; after the addition of the second index, nuclei were again re-pooled for their final distribution of 4,000 nuclei per well prior to second strand synthesis, protease digestion, tagmentation and PCR all on this final third index plate.

snRNA-seq pre-processing

snRNA-seq sequencing reads were processed into a gene-by-nucleus expression matrix of unique molecular index (UMI) counts following the methods described by [2]. We used largely an identical pipeline which, briefly, (1) converts base calls to fastq files with bcl2fastq/v.2.20 (RRID:SCR_015058) (Illumina), (2) removes adapter sequences using Trim Galore/v.0.6.7 (RRID:SCR_011847) [Krueger], (3) aligns trimmed reads to a reference genome with STAR/v.2.7.6 (RRID:SCR_004463) [183], (4) extracts mapped reads, (5) removes duplicates, and (6) generates UMI counts for exonic and intronic regions of each gene, tabulated according to the unique three-level barcode design in sci-RNA-seq3. We used the rhesus macaque reference genome (Mmul_10) [272] and annotation, obtained from Ensembl (version 101) (RRID:SCR_002344). We extended the 3' UTR annotations of genes and transcripts by 500 bp to avoid misclassifying genic reads as intergenic. The remainder of our pipeline followed the procedures described by [2]. After generating the count matrix, we removed all nuclei with UMI counts < 100 .

For each sample, we imported gene-by-nucleus count matrices into the AnnData/v.0.8.0 (RRID:SCR_018209) [273] framework, then ran Scrublet/v.0.2.3 (RRID:SCR_018098) [89] (expected_doublet_rate=0.05) to calculate doublet scores. We marked nuclei as doublets if they had Scrublet doublet scores > 0.20 . For each sample, we additionally marked nuclei as doublets using per-sample thresholds determined by Scrublet and adjusted by eye as necessary in order to separate bimodal peaks visualized on the Scrublet doublet score histogram (**Fig. 4.7**).

To further identify potential doublet nuclei, we employed an iterative clustering strategy [37] implemented with Scanpy/v.1.9.1 (RRID:SCR_018139) [274]. First, we combined all nuclei into a single AnnData object and filtered nuclei to those with UMI ≥ 100 , number of expressed genes $< 2,500$, and a percentage of reads mapping to the mitochondrial genome $< 5\%$. We then removed all non-autosomal genes, genes located on unplaced scaffolds, and

unexpressed genes. Next, we normalized the data to the total UMI per nucleus, logarithmized the data, and subsetted the data to the 10,000 most variable genes. For each cell, we regressed out total UMI counts per nucleus, then mean-centered and scaled the data. The dimensionality of the data was then reduced by PCA (50 components).

To further reduce the dimensionality, we ran a UMAP (using `umap-learn/v.0.5.2` (RRID:SCR_018217) analysis (McInnes and Healy 2018) with `BBKNN/v.1.5.1` (RRID:SCR_022807) [275] to simultaneously correct for batch differences. For the `BBKNN` integration, we set `neighbors_within_batch=10` (given three batches, tantamount to UMAP `n_neighbors=30`), used the cosine distance metric, and used the `PyNNDescent/v.0.5.6` algorithm (RRID:SCR_022806) [276]. We then ran UMAP using the settings `min_dist=0`, `spread=1.0`, and `n_components=10` to facilitate clustering (<https://umap-learn.readthedocs.io/en/latest/clustering.html>). For data visualization only (not clustering), we ran a similar `BBKNN/UMAP` pipeline with `neighbors_within_batch=5` (for three batches, tantamount to UMAP `n_neighbors=15`), `min_dist=0.25`, `spread=1.0`, and `n_components=2`. To cluster the data, we exported and imported the 10-dimensional UMAP matrix into `Monocle3/v.1.2.9` (RRID:SCR_018685) in `R/v.4.0.2` (RRID:SCR_001905), then implemented the Leiden-clustering workflow in `Monocle3` with a relatively high-resolution setting (`resolution=1×10-4`). For each cluster, we then calculated the mean `Scrublet` doublet score and marked all clusters with a mean `Scrublet` doublet score > 0.15 as doublet clusters (**Fig. 4.7**).

After identifying doublets as described above, we removed all marked doublets and repeated the normalization, dimensionality, and clustering procedures almost exactly as described above, with the only difference being a coarser cluster resolution setting in `Monocle3` (`resolution=1×10-5`). We confirmed adequate removal of doublet cells by observing the clean separation of distinct cell types and the absence of clusters expressing obviously ambiguous marker gene profiles (**Fig. 4.7**).

Removal of sci-RNA-seq cell contamination

During the course of cell-type identification (see following section), we observed the presence of two distinct clusters of cells (**Fig. 4.8A**) with expression profiles resembling embryonic progenitors (marker genes, unknown cluster 1: *ASPM*, *CENPE*, *CENPF*, *MKI67*; unknown cluster 2: *COL1A1*, *COL1A2*, *FN1*, *VIM*), an unusual finding in adult primate brain samples. Because these were present in relatively large proportions in some samples ($\sim 25\%$)—but at low levels overall (2.2%)—and because our sci-RNA-seq experiments included control samples of exogenous (i.e., non-macaque brain) origin (specifically, a fetal mouse brain positive control and a “barnyard” sample consisting of mixed human HEK293T and mouse NIH/3T3 cells), we tested for the presence of contaminating nuclei of exogenous origin. We identified and removed contaminating cells as follows.

Because the only non-macaque samples included in all experiments were the control samples of either human or mouse origin, we used BBSplit/v.38.38 (RRID:SCR_016965) [277] to assign reads to the macaque, human, or mouse genomes. BBSplit is a competitive aligner that maps to several references simultaneously, assigning reads to the genome with the best unambiguous match. We used the following reference assemblies from Ensembl version 101: Mmul_10 (macaque), GRCh38.p13 (human), and GRCm38.p6 (mouse). After indexing the three references simultaneously using BBSplit, we aligned 10 million randomly sampled unique (de-duplicated) reads for each sample using default settings in BBSplit, which partitioned reads assigned to each genome into separate fastq files. Unmapped and ambiguous reads were directed to additional fastq files that were not used. Using a similar demultiplexing workflow to the sci-RNA-seq3 preprocessing pipeline, we tabulated reads-per-cell for each of the three genomes and calculated summary statistics.

After filtering to only cells with ≥ 10 unambiguously assigned reads by BBSplit, we observed that discernible fractions of exogenous reads (reads unambiguously assigned to human or mouse) were specific to certain barcodes from the first round of sci-RNA-seq barcoding

(reverse transcription), indicating that a low level of cross-well contamination of cells or barcoded primers likely occurred at this stage (**Fig. 4.8C**). We also observed that, after filtering to cells passing all previous quality control filters, our clustering and annotation workflow had partitioned exogenous cells into two clusters corresponding to human and mouse cells respectively, with no discernible exogenous contamination in other annotated cell types (**Fig. 4.8B**). After removing the entirety of the two exogenous clusters from the dataset ($N=58,443$ cells), we re-examined the distribution of exogenous read fractions across reverse-transcription barcodes and confirmed that human and mouse cells were effectively removed (**Fig. 4.8C**).

snRNA-seq cell-type and cell-subtype identification

To identify cell types, we visualized the expression of canonical marker genes (**table S3**) on normalized, log-transformed gene expression data using Scanpy. Most clusters were readily assigned to well-characterized cell types in this manner. To aid in the classification of more nuanced cell types, we determined top marker genes using logistic-regression and *t*-test marker-gene methods implemented via the ‘rank_genes_groups’ function in Scanpy. For each discrete cell type, we ran marker gene tests by testing gene expression in a given cell type against gene expression in all other cells in our dataset.

Based on canonical markers and data-derived marker genes, we identified 17 parent cell types (not including the two cells of exogenous origin, see section above), which we refer to as cell classes. In all but two cases, our parent cell types corresponded with partitions identified through our clustering using Monocle3 (q -value threshold=0.05). In two cases, we considered clusters assigned to the same partition to be discrete parent cell types because they exhibited clear separation in our global analysis while clearly expressing canonical markers of known cell types (dopaminergic and serotonergic neurons; **table S3**), yet did not effectively segregate when their assigned partition (the partition also including GABAergic neurons) was analyzed separately.

To identify cell subtypes, we partitioned the data by cell class and reanalyzed each data partition individually. For each cell-class-specific analysis, we repeated a preprocessing, dimensionality reduction, and clustering analysis that largely followed the pipeline described above for our global analysis, with the following exceptions. After normalizing and log-transforming the data, we identified the 2,000 most variable genes for each given cell type and subset the data to those highly variable genes. Because we observed that differences in total UMI among batches resulted in artifactual clusters being identified downstream (even after batch-correction with BBKNN, a problem we did not observe in our global analysis), we regressed out total UMI counts per nucleus separately for each batch. We then combined residual values from all batches before mean-centering and scaling for PCA and UMAP analysis. For Leiden clustering, we used the same resolution parameter (resolution= 1×10^{-5}) for most cell types, but in four cases defaulted to partitions identified using Monocle3 (q -value threshold=0.05) after observing small clusters with unusually high UMI. We considered clusters/partitions identified in this manner to be cell subtypes. As with our global (all cell classes combined) analysis, for each cell subtype we identified top marker genes using logistic-regression and t -test marker gene methods implemented in Scanpy. Additionally we used a non-negative least squares (NNLS) approach [2] to identify correlations between cell subtypes and annotated labels in reference datasets (**table S5, Fig. 4.14**).

Additionally, we scanned for gene-disease associations that were enriched among the top 100 marker genes for each cell subtype. We used gene-disease associations from the DISEASES database (RRID:SCR_015664) [238] and used Fisher's exact test to identify overrepresented disease associations among the top 100 marker genes for a given cell subtype, using all macaque genes in our analysis as the background (**table S10**).

Cell composition and regional heterogeneity analysis

To assess the specificity of cell classes and/or subtypes, we calculated the Jensen-Shannon divergence statistic using the 'JSD' function from the philentropy package (RRID:SCR_022805)

(Drost 2018) in R. We calculated the Jensen-Shannon divergence by comparing, for a given cell class or cell subtype, the cell type’s count distribution across brain regions to the count distribution (combining all cell types per region) of the entire dataset combined [30]). To measure regional heterogeneity within cell types, we extended our recently developed statistic, lochNESS [223], to quantitatively measure enrichment of each region subclass or region within each cell’s neighborhood. For each cell type, we define lochNESS of $cell_n$ for $region_m$ as:

$$lochNESS_{cell_n, region_m} = \frac{\#of\ cells\ from\ region_m\ in\ k\ NN\ of\ cell_n}{k} / \frac{\#of\ cells\ from\ region_m\ in\ cell\ type}{N}$$

where N is the total number of cells in the the cell type and k is the number of nearest neighbors for $cell_n$.

For each cell type the calculation results in a cell x region matrix, where each row can be separately visualized. For a summarizing visualization, each cell can be colored by the region with the largest lochNESS. Additionally, when we focus on a subset of regions (e.g. just the cortical regions), we calculate a normalized lochNESS that is comparable across the regions of interest:

$$lochNESS^*_{cell_n, region_m} = lochNESS_{cell_n, region_m} / \sum_{m=1}^M lochNESS_{cell_n, region_m}$$

where M is the number of regions or region subclasses of interest.

To identify genes that are expressed with regional bias, we fit a regression model for each gene to identify regions with significant non-zero correlation with gene expression as implemented in Monocle3 [2]. The model for each cell type is:

$$\log(\text{expression}) = \beta_0 + \beta_1 * lochNESS_{region_1} + \beta_2 * lochNESS_{region_2} + \dots + \beta_m * lochNESS_{region_m}$$

where β_0 is the intercept and $lochNESS_{region_i}$ is a vector of lochNESS across all cells in the cell type.

Hierarchical clustering of cells and regions

We used Scanpy to cluster cell classes and brain regions based on the top 50 princi-

pal components (PCs) of gene expression. Because of our use of BBKNN for batch correction in our main workflow, our PCA was not actually corrected for batch. To rectify this, we first used the `harmony/v.0.0.5` (RRID:SCR_022798) implementation of Harmony (RRID:SCR_022206) [27] to generate a batch-corrected PCA matrix (convergence after 2 generations). We then used the Scanpy ‘`dendrogram`’ function to perform hierarchical clustering using the batch-corrected PCA embedding. To visualize uncertainty, we performed 1,000 bootstrap iterations in which we resampled cells randomly with replacement and computed new dendrograms. We then used the ‘`DensiTree`’ function [278] implemented in the `phangorn/v.2.6.3` (RRID:SCR_017302) [279] R package to visualize trees. We performed this procedure using both cell class and brain region as labels (**Fig. 4.11A-B**).

For brain regions, we also performed hierarchical clustering using the cell proportion (cell class×brain region) matrix. We used the ‘`hclust`’ function in R to cluster using the “complete” method based on Euclidean distances. To again visualize uncertainty, we resampled all cells in our dataset 1,000× with replacement, then repeated calculation of cell class proportions and hierarchical clustering. We visualized the final tree with ‘`DensiTree`’ (**Fig. 4.11B**).

snATAC-seq data generation

To profile single-nucleus chromatin accessibility, we performed single-nucleus ATAC-seq (snATAC-seq) using the three-level single-cell combinatorial indexing ATAC-seq (sci-ATAC-seq3) approach [16], which is the improved version of the original sci-ATAC-seq protocol [226]. We followed the protocol of [16], with slight modifications. Briefly, we added 50 μl of Omni-ATAC lysis buffer to pulverized tissue and homogenized the tissue with 5–10 strokes with a disposable RNase-free plastic pestle (Fisherbrand, cat. #12-141-364). We then added another 950 μl of Omni-ATAC lysis buffer, mixed by pipette, incubated on ice for 3 min, then transferred the suspension to a new 15 ml conical tube containing 5 ml ATAC-RSB with 0.1% Tween-20. We then pelleted the nuclei, removed the supernatant, and resuspended the

pellet in 1 ml of 1x DPBS. We then transferred the suspension through a 70 μm cell strainer (pluriSelect cat. #43-10070-70) into a 15 ml conical tube containing 4 ml of 1x DPBS and 140 μl of 37% formaldehyde (final concentration 1% formaldehyde). We then incubated the nuclei for 10 min with occasional mixing. The fixation was then quenched with 250 μl 2.5 M glycine, incubated for 5 min at room temperature, then incubated for another 15 min on ice. We then pelleted the nuclei, removed the supernatant, and resuspended the pellet in 2 ml freezing buffer. Nuclei were counted by mixing with 1 μM of YOYO-1 iodide (ThermoFisher cat. #Y3601) using a Countess II FL automated cell counter (Life Technologies), divided into tubes in 50 μl aliquots, then flash-frozen in liquid nitrogen

Frozen fixed nuclei were prepared for the sci-ATAC-seq3 library similar to [16]. Omni-ATAC lysis buffer (10 mM NaCl, 3 mM MgCl_2 , 10 mM Tris-HCl pH 7.4, 0.1% IGEPAL CA-630, 0.1% Tween 20 and 0.01% Digitonin) was used to permeabilize fixed nuclei before diluting samples with ATAC-RSB (10 mM NaCl, 3 mM MgCl_2 , 10 mM Tris-HCl pH 7.4) supplemented with 0.1% Tween-20. Approximately 200,000 nuclei per sample was spread across 4 wells for tagmentation as previously described. Barnyard control for each set of experiments included mouse cell line (CH12-LX; RRID:CVCL_0211) and human pancreas as a quality control tissue.

Our combined snATAC-seq dataset encompasses data prepared using five sci-ATAC-seq3 experimental runs (i.e., library preparation/sequencing batches). Sample order was randomized between batches to ensure balance of brain regions, sex, and hemispheres between runs and to minimize batch effects.

snATAC-seq pre-processing

snATAC-seq sequencing reads were processed into a peak-by-nucleus count matrix following the methods described by [16]. We followed largely an identical pipeline which, briefly, (1) converts base calls to fastq files with bcl2fastq/v.2.20 (Illumina), (2) removes adapter se-

quences using Trimmomatic/v.0.39 (RRID:SCR_011848) (Bolger, Lohse, and Usadel 2014), (3) aligns trimmed reads to a reference genome with bowtie2/v.2.4.1 (RRID:SCR_016368) (Langmead and Salzberg 2012), (4) calculates nonduplicate fragment endpoints for each cell, (5) calls peaks using MACS2/v.2.2.7.1 (RRID:SCR_013291) [87, 280] and merges peaks across samples to create a merged BED file, (6) tabulates reads from merged peaks and annotated TSSs (± 1 Kb around each TSS) for quality control (QC), (7) separate cell barcodes from background barcodes by fitting a mixture of two negative binomials (noise vs. signal), and (8) assembles sparse matrix tabulating reads per cell barcode falling within the master set of peaks and within gene bodies extended by 2 Kb upstream. We used the rhesus macaque reference genome (Mmul_10) [272] and annotation, obtained from Ensembl (version 101), and merged peaks across all samples (encompassing five library preparation and sequencing batches) to create a global set of peaks. After binarizing UMI counts, we filtered the peak-by-nucleus matrix to include only nuclei with ≥ 1000 binarized UMI, less than 100,000 binarized UMI, and $\geq 30\%$ fraction of reads in peaks (FRIP) (**Fig. 4.17**).

We identified and removed doublets using a similar iterative clustering approach to that described for our single-nucleus RNA dataset (**Fig. 4.17**). Briefly, we ran Scrublet/v.0.2.3 [89] on each sample individually and marked doublets using both a universal threshold (Scrublet doublet score > 0.20) and a per-sample threshold determined by Scrublet and checked and adjusted (if necessary) by eye. We then performed a similar preprocessing, dimensionality reduction, and clustering pipeline to identify clusters with relatively high Scrublet doublet scores (mean Scrublet doublet score > 0.15). We finally removed all nuclei marked as doublets based on the described criteria before concatenating all singlet nuclei across all samples together.

Our snATAC-seq preprocessing, dimensionality reduction, and clustering pipeline likewise tracked closely to our single-nucleus RNA-seq analysis, with minor modifications to accommodate best practices for ATAC-seq data. Briefly, we filtered the data to remove

peaks that were not accessible in a minimum of 5 cells as well as peaks that were located on non-autosomal or unplaced scaffolds in the macaque genome. We then filtered the data to the top 100,000 variable features. We performed latent semantic analysis (LSI) on the resulting peak-by-cell matrix to reduce the dimensionality of the data. We performed term frequency/inverse document frequency (TF-IDF) normalization followed by singular value decomposition (SVD) following previously described procedures [16] to reduce the data to 50 PCA dimensions. L2 normalization was then performed on the last 49 principal components, thereby excluding the first principal component, which tends to capture read depth [16]. TF-IDF, SVD, and L2-normalization procedures were implemented using scikit-learn/v.1.0.2 (RRID:SCR_002577) [281]. The L2-normalized PCA matrix was then reduced further and clustered using an identical BBKNN/UMAP/Monocle3 approach to that used for our single-nucleus RNA-seq data. Doublet-derived clusters were also marked for removal using an identical threshold (mean Scrublet doublet score > 0.15).

After marking and removing doublets from our data, we repeated our preprocessing, dimensionality reduction, and clustering pipeline. After observing clear separation of distinct cell classes, we used MUON/v.0.1.2 (RRID:SCR_022804) [282] to calculate promoter accessibility scores by tabulating binarized UMI counts within the region 2,000 bp upstream of a transcriptional start site (TSS). Because at the time of this analysis MUON did not factor in DNA strand information, we ran the function ‘count_fragments_features’ separately for + and – strand genes, using the “upstream_bp” or “downstream_bp” arguments as necessary to tabulate counts in the correct upstream region (extending from the TSS to [TSS – 2,000 bp] or [TSS + 2,000 bp], respectively) (<https://github.com/scverse/muon/issues/59>). We used Scanpy to normalize and visualize resulting promoter accessibility scores (**Fig. 4.3B**). We provisionally classified nuclei based on promoter accessibility scores of known marker genes.

Integration of snRNA-seq and snATAC-seq data

We used Graph Linked Unified Embedding (GLUE) implemented in `scglue/v.0.2.3` (RRID:SCR_022803) [35] to integrate our snRNA-seq and snATAC-seq datasets. To run `scglue`, we followed pre-processing procedures in Scanpy recommended by the `scglue` authors for both our snRNA-seq and snATAC-seq data, after filtering out doublets as described above. For snRNA-seq data, we identified the top 2,000 most variable genes, then normalized, log-transformed, and scaled the data using default parameters in Scanpy. We then reduced the dimensionality of the data to the top 100 principal components using PCA, based on the top 2,000 variable genes and the automatic SVD solver selected by Scanpy. For snATAC-seq data, we used the LSI implementation in `scglue` to reduce the data to the top 100 principal components, with the number of power iterations set to 15.

We then used `scglue` to compute a prior guidance graph and propagated highly variable snRNA-seq features (genes) to identify highly variable snATAC-seq features (peaks) based on the guidance graph. We then built and trained the GLUE integration model using the PCA and LSI embeddings, respectively, as the first encoding transformation, modeling raw counts of both snRNA-seq and snATAC-seq data using the negative binomial model, using the batch-correction option to correct for sequencing batches, and using the previously computed prior guidance graph as input. As all nuclei from this study were included (totaling over 4 million nuclei), this analysis was particularly computationally demanding. We performed this analysis on a machine with 1.5 TB RAM, accelerated by 4 Tesla V100 (NVIDIA) GPUs.

After training a GLUE model, we validated effective integration by calculating integration consistency scores using `scglue` (**Fig. 4.18A**). We then calculated integrated cell and feature embeddings for both snRNA-seq and snATAC-seq data using `scglue`. After projecting all cells to a unified embedding, we performed UMAP dimensionality reduction using the same procedures as described previously, with one exception. Because the unified GLUE embedding was already batch-corrected, we computed the neighborhood graph using the Scanpy ‘neighbors’ function rather than BBKNN, with `n_neighbors=15`.

To transfer cell-class labels from our snRNA-seq data to our snATAC-seq data, we used the ‘transfer_labels’ function in scglue, which computes shared nearest neighbors between reference (snRNA-seq) and query (snATAC-seq) nuclei, weighted by the Jaccard index. Jaccard indices are then normalized per query nucleus to form a mapping matrix, which is then multiplied by one-hot-encoded reference labels. The reference label with the highest score is then assigned as the predicted cell type, with the highest score retained as the confidence score. For label transfer, because a subset of our snRNA-seq data was derived from samples that were unprofiled in our snATAC-seq data, we limited our reference RNA-seq dataset to only those nuclei deriving from samples profiled in both snRNA-seq and snATAC-seq experiments. We then retained 100,000 nuclei from withheld (unmatched) snRNA-seq samples as a query dataset to evaluate label transfer accuracy. For snATAC-seq label transfer, we used all snATAC-seq nuclei as a query dataset. We used previously assigned parent cell types for our snRNA-seq dataset as reference labels. For our snATAC-seq query nuclei, we retained all predicted cell-class labels with a label transfer confidence score ≥ 0.95 . At this threshold, the error rate in our evaluation dataset was 0.43% (**Fig. 4.18B**).

Identification of candidate *cis*-regulatory elements

To scan for candidate *cis*-regulatory elements underlying differential expression among brain cells, we used two complementary approaches. First, we used the integrative GLUE regulatory inference approach implemented in scglue/v.2.0.3 [35], which calculates regulatory scores based on the cosine similarities between multi-omics data features in an integrated space. Second, we used a metacell approach to construct multi-omic samples (determined via *k*-means clustering in integrated space) with aggregated (pseudobulk) gene expression and chromatin accessibility counts, which we then modeled using logistic regression. Finally, we calculated differentially accessible peaks using a similar workflow to our snRNA-seq marker gene analysis.

To calculate GLUE regulatory scores, we performed a second integration of our snRNA-

seq and snATAC-seq datasets, following an identical pipeline except including the top 6,000 most variable genes (rather than 2,000). This allowed us to identify putative gene:peak regulatory connections and to generate an integrated feature embedding for a greater number of genes and genomic regions. We constructed a window graph between inferred promoters—which we calculated as the region from the strand-specific transcription start site extended upstream 2,000 bp—and peaks using the ‘window_graph’ function, with the window size set to 150 Kb and a distance-decaying weight, as recommended by the scglue authors. We then used the previously computed window graph and feature embeddings to perform the regulatory inference analysis using the ‘regulatory_inference’ function, with the alternative hypothesis set to “greater” in order to perform a one-sided test.

In order to determine the directionality of putative regulatory relationships, we used a second approach based on metacell identification and logistic regression (**Fig. 4.5A**). We use the ‘get_metacells’ function to generate multi-omic (snRNA-seq/snATAC-seq) metacells based on k -means clustering of their integrated cell embeddings. As our snRNA-seq dataset included 2,583,967 single-cell transcriptomes, we set k (n_meta) to 10,335 in order to target a mean of roughly 250 RNA transcriptomes per metacell. After identifying metacells in this manner, we summed (pseudobulked) gene expression per metacell. For each gene:peak pair tested in our GLUE regulatory inference, we then performed a logistic regression modeling accessibility of each individual snATAC-seq cell in a given metacell (1: open, 0: closed) as a function of \log_2 CPM-normalized gene expression for that snATAC-seq cell’s respective metacell. Logistic regressions were performed in R/v.4.0.2.

We considered candidate *cis*-regulatory relationships to be gene:peak pairs for which FDR-adjusted $P < 0.05$ for both the GLUE regulatory inference and metacell-based logistic regression tests. We classified candidate *cis*-regulatory relationships as positive or negative relationships based on the sign of their logistic regression coefficients (β values) (**Fig. 4.21**).

For all peaks, we also tested for marker peaks (peaks with differential accessibility) using logistic-regression and *t*-test marker-gene methods implemented via the ‘rank_genes_groups’ function in Scanpy. Similar to our snRNA-seq marker genes analysis, we ran marker peak tests by testing chromatin accessibility in a given cell type against accessibility in all other cells in our dataset. Additionally, to validate marker peaks, we used a second logistic-regression approach implemented via the ‘FindMarkers’ function in Seurat/v.4.1.1 (RRID:SCR_016341) [283]. In contrast to the logistic regression in Scanpy, the Seurat implementation is not a regularized procedure and is thus able to control for latent variables and to calculate *P* values. To reduce computational burden, we ran ‘FindMarkers’ on a dataset with 1,000 cells per cell class. As we found that output statistics (regularized logistic regression coefficient in Scanpy and log fold-change in Seurat) were highly concordant (**Fig. 4.22**), we report Scanpy results here as they included all possible cells. We considered peaks to be differentially accessible if the regularized logistic regression coefficient > 0 , the log fold change > 0 , and the *t*-test $P_{adj} < 0.05$.

snATAC-seq cell subtype analysis

To mitigate peak-calling biases while allowing us to probe more nuanced regulatory variation within cell populations, we called a new set of cell-class-specific peaks for each cell type with assigned cells, skipping rarer cell types for which no snATAC-seq nuclei passed our prediction threshold above.

Following scglue integration and assignment of snATAC-seq cells to cell classes, we created cell-class-specific pseudobulk epigenomes by aggregating all nonduplicate fragment endpoints for each cell class. These cell-class-level ATAC-seq data were then used for peak-calling using MACS3/v.3.0.0a6 [87, 280], with the same peak calling parameters that we used for each sample and batch described in the “snATAC-seq pre-processing” section above (‘-g 2.7e9 -call-summits -nomodel’). For each cell class, we repeated steps from our snATAC-seq data generation pipeline to tabulate reads from newly called peaks and to assemble sparse count

matrices matrix tabulating reads per cell barcode falling within the master set of peaks and within gene bodies extended by 2 Kb upstream. We then imported peak-by-nucleus count matrices into the AnnData/v.0.8.0 [273] framework.

To assign cell subtypes for our snATAC-seq data (**Fig. 4.19**), we repeated preprocessing, data integration, label transfer, and regulatory inference procedures described above on each cell class individually. In contrast to our global joint analysis, we only included snRNA-seq nuclei deriving from samples that were profiled in both snRNA-seq and snATAC-seq experiments, and used the top 6,000 most variable genes in our snRNA-seq analysis, and used the snATAC-seq peak sets specific to each cell type. The remainder of our preprocessing and data integration procedures followed the same pipeline described previously for our global integration analysis. For label transfer, we also followed largely the same procedures as for our global label transfer pipeline. We did not, however, use a label transfer confidence score threshold under the assumption that snATAC-seq nuclei would, on average, be assigned to the correct cell subtype and, if incorrect, would be assigned to a closely related cell subtype (i.e., a neighboring subtype in the integrated multidimensional cell space).

For metacell-based regulatory inference, we varied the settings for k based on dataset size in order to target a mean of 50 transcriptomes per metacell.

Transcription factor binding site enrichment

For enrichment analyses at the cell-class level, we focused on peaks that were deemed accessible in one and only one cell class, which we called “cell-class-unique peaks”. We identified these peaks using BedTools/v.2.30.0 (‘intersect -v’) (RRID:SCR_006646) [88] to find all peaks in a cell class that did not overlap with any peak called in another cell class. The number of peaks identified in this manner ranged from 655 (in ependymal cells) to 71,049 (in glutamatergic neurons). We tested for enrichment of TF binding motifs in cell-class-unique peaks compared to the background of the rhesus macaque genome while controlling

for GC content, implemented in the monaLisa/v.1.3.1 (RRID:SCR_022802) [284] in R/v.4.1.0 (**table S8**). We used the JASPAR 2018 (RRID:SCR_003030) non-redundant vertebrate core position weight matrices [285].

At the cell-subtype level, we tested for enrichment using the top differentially accessible peaks among subtypes of the same cell class, excluding peaks with regularized logistic regression coefficients < 0 (**table S9**). We retained the top first percentile of marker peaks, ranked according to their regularized logistic regression coefficients.

Disease heritability enrichment

We calculated enrichment of disease-associated variants in cell-class-specific accessible chromatin regions using linkage disequilibrium score regression, LDSC (RRID:SCR_022801) [252, 253] (Fig. 4.6 and **table S15**). Because the trait-associated loci are annotated in the human genome, we converted all peaks (at the combined level as well as each individual cell-class level) from the rhesus macaque genome coordinates to GRCh37 using UCSC's liftOver/v.302 (RRID:SCR_018160) tool [286]. We followed the standard pipeline using the 1000 Genomes baseline model and precomputed .sumstats files. A list of phenotypes tested can be found in **table S14**.

4.5 Supplementary Materials

4.5.1 Supplementary Tables

Table S1. List of brain regions sampled in this study. Cell counts for both snRNA-seq and snATAC-seq are final counts after all quality control filters have been applied.

Table S2. Sample metadata for this study. Animal IDs and social group IDs follow the identifiers assigned by the Caribbean Primate Research Center (CPRC) for the Cayo Santiago macaque colony. Brain regions follow the abbreviations in **table S1**. Hemispheres with <NA> values indicate that samples are from midline structures and were sampled as a single sample from either or both hemisphere(s). Library IDs and batch IDs for both sci-RNA-seq3 and sci-ATAC-seq3 are internal identifiers in use by the Brotman Baty Institute Advanced Technology (BAT) Laboratory. Cell counts for both snRNA-seq and snATAC-seq are final counts after all quality control filters have been applied.

Table S3. Cell class statistics and marker genes. Counts for each cell class are included for both the snRNA-seq and snATAC-seq datasets. The comma-separated list of top 10 marker genes for each cell class is included, ranked in descending order by regularized logistic regression coefficient.

Table S4. Cell-class compositions across sampled brain regions in snRNA-seq dataset. Relative proportions are for cell classes (rows) found in each brain region (columns) in the snRNA-seq dataset.

Table S5. Cell subtype statistics and marker genes. Counts for each cell subtype are included for both the snRNA-seq and snATAC-seq datasets. The comma-separated list of top 5 marker genes for each cell subtype is included, ranked in descending order by regularized logistic regression coefficient (testing expression of each subtype compared to expression in other subtypes within a given cell-class partition). In some cases, cell subtypes were annotated based on non-negative least-squares (NNLS) regression comparisons with external datasets. When applicable, cell-type labels and DOIs are included for NNLS matches.

Table S6. Cell-subtype compositions across sampled brain regions in snRNA-seq dataset. Relative proportions are for cell subtypes (rows) found in each brain region (columns) in the

snRNA-seq dataset.

Table S7. Genes associated with lochNESS at the region subclass level in the astrocytes. Generalized linear models were fit to each variable gene in the astrocytes. For each gene and region subclass, an estimate and adjusted p-value is calculated based on the model. Gene and region subclass pairs with significant non-zero estimates ($P_{adj} < 0.05$) are included in the table.

Table S8. Enrichment of transcription factor binding motifs among cell-class-specific peaks. Enrichments were performed on peaks that were called separately for each cell class and that showed no genomic overlap with peaks called using other cell classes. Enrichment was performed using monaLisa on JASPAR transcription factor (TF) binding profiles, using the rhesus macaque genome as the background. TF binding motifs with $P_{adj} < 0.05$ for a given cell class are included in this table.

Table S9. Enrichment of transcription factor binding motifs among cell-subtype differentially accessible peaks. Enrichments were on peaks that were called separately for each cell class and that were classified as differentially accessible marker peaks, defined as being in the top percentile (1%) of peaks (ranked by regularized logistic regression coefficient) with regularized logistic regression coefficient > 0 . Enrichment was performed using monaLisa on JASPAR transcription factor (TF) binding profiles, using the rhesus macaque genome as the background. The top 20 TF binding motifs with $P_{adj} < 0.05$ (ranked by P_{adj}) for a given cell subtype are included in this table.

Table S10. Enrichment of disease-associated genes among cell-subtype marker genes. Enrichment of disease-associated genes was performed using disease association from the DISEASES database. Enrichment of disease associations among the top 100 marker genes (t -test $P_{adj} < 0.05$; ranked by regularized logistic regression coefficient) for each cell subtype

were performed using Fisher’s exact test.

Table S11. Correlation between transcription factor expression and motif enrichment. A total of 369 JASPAR TFs with a corresponding gene in the snRNA-seq dataset were included in the analysis. The mean expression and mean motif enrichment of each TF were calculated for each cell type and region class. Correlation between gene expression and motif enrichment across cell type and region classes was quantified with the Pearson’s correlation coefficient.

Table S12. Candidate *cis*-regulatory interactions between peaks and genes (global). Regulatory inference results using peaks called on the global (all cells combined) dataset. Peak-gene pairs were considered significant if $P_{adj} < 0.05$ for both regulatory inference methods (GLUE regulatory inference and metacell-based logistic regression). Because of the large number of identified candidate regulatory interactions, only results with $P_{adj} < 0.005$ for both statistics are included in this table.

Table S13. Candidate *cis*-regulatory interactions between peaks and genes (cell classes). Regulatory inference results using peaks called on the dataset partitioned by cell class. Peak-gene pairs were considered significant if $P_{adj} < 0.05$ for both regulatory inference methods (GLUE regulatory inference and metacell-based logistic regression). Because of the large number of identified candidate regulatory interactions across cell classes, only results with $P_{adj} < 0.001$ for both statistics are included in this table.

Table S14. List of phenotypes included in linkage disequilibrium score (LDSC) enrichment analysis. Genome-wide association study (GWAS) citations and digital object identifiers (DOIs) are included.

Table S15. Linkage disequilibrium score (LDSC) enrichment results. LDSC enrichment

was performed on peaks called separately for each cell class. All results with $P_{adj} < 0.05$ are included in this table.

4.5.2 Supplementary Figures

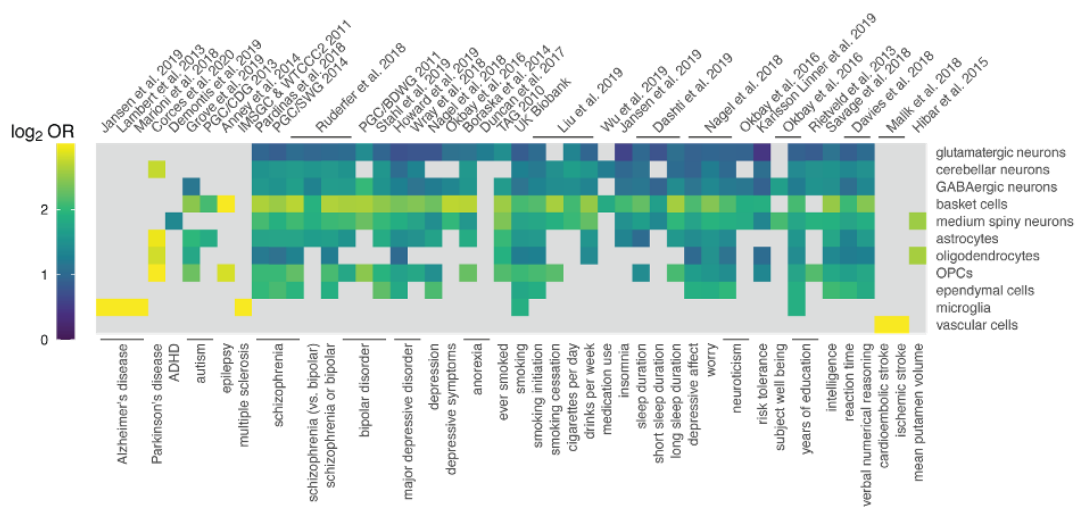


Figure 4.6: **Fig. 6. Enrichment of heritable disease-relevant sites among candidate regulatory elements.** The heatmap displays heritability enrichment (\log_2 odds ratio [OR]) of diseases among cell-class snATAC-seq peaks for tested diseases, syndromes, and phenotypes. Only results passing a threshold of $P_{adj} < 0.05$ are shown. The \log_2 OR color range is capped at 3.0.

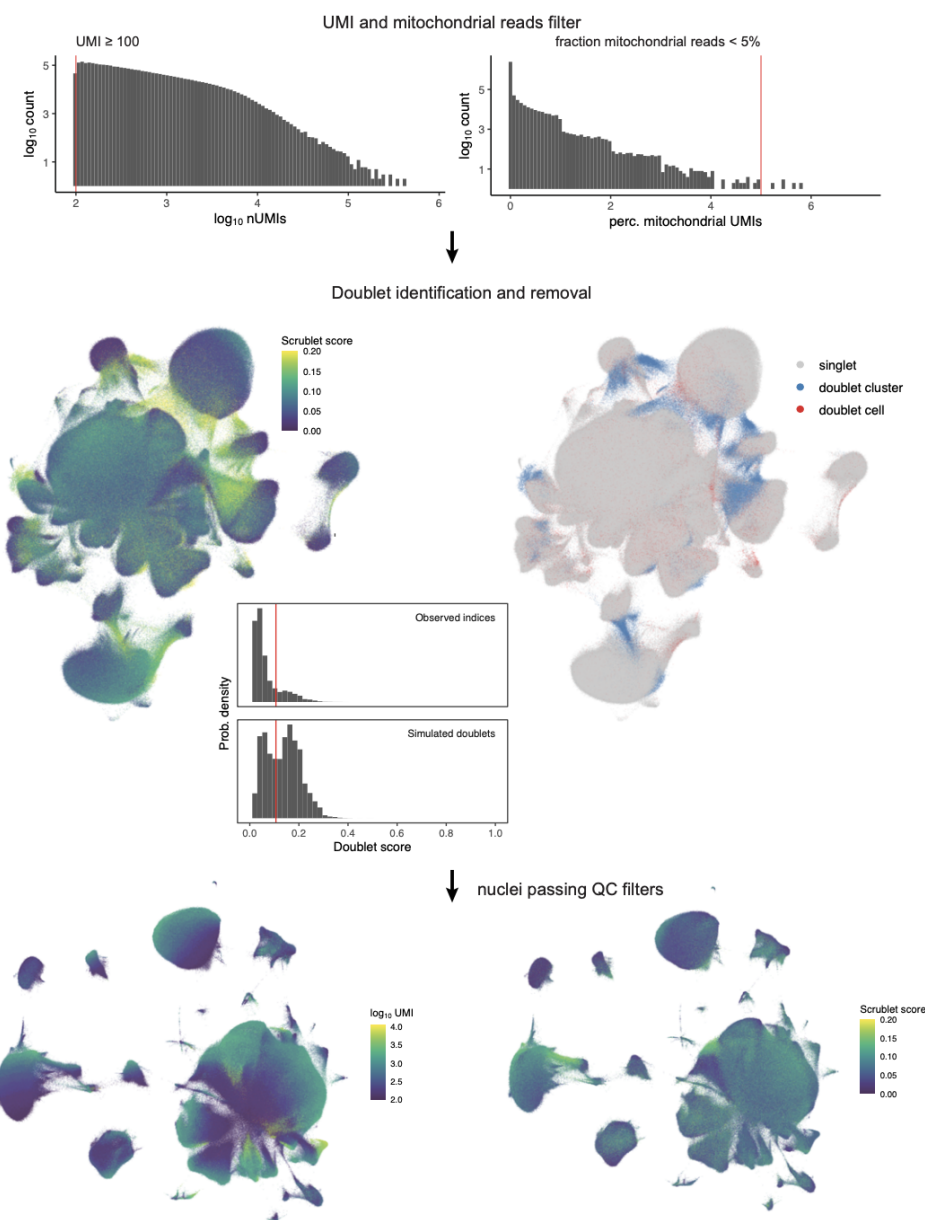


Figure 4.7: **Fig. S1. Schematic depicting snRNA-seq quality control pipeline.** **A**, Nuclei (combinatorial indices) with fewer than 100 UMIs and greater than 5% reads mapping to the mitochondrial genome were removed. **B**, Scrublet k -nearest-neighbor (kNN) doublet scores were calculated per-sample and doublets with scores > 0.20 were marked (using automated Scrublet thresholds with manual adjustment) but not removed. All nuclei, including doublets, were then jointly preprocessed and clustered. Clusters with mean doublet scores > 0.15 were then removed along with previously marked doublets. **C**, UMI counts and Scrublet doublet-detection scores visualized on the post-quality-control dataset.

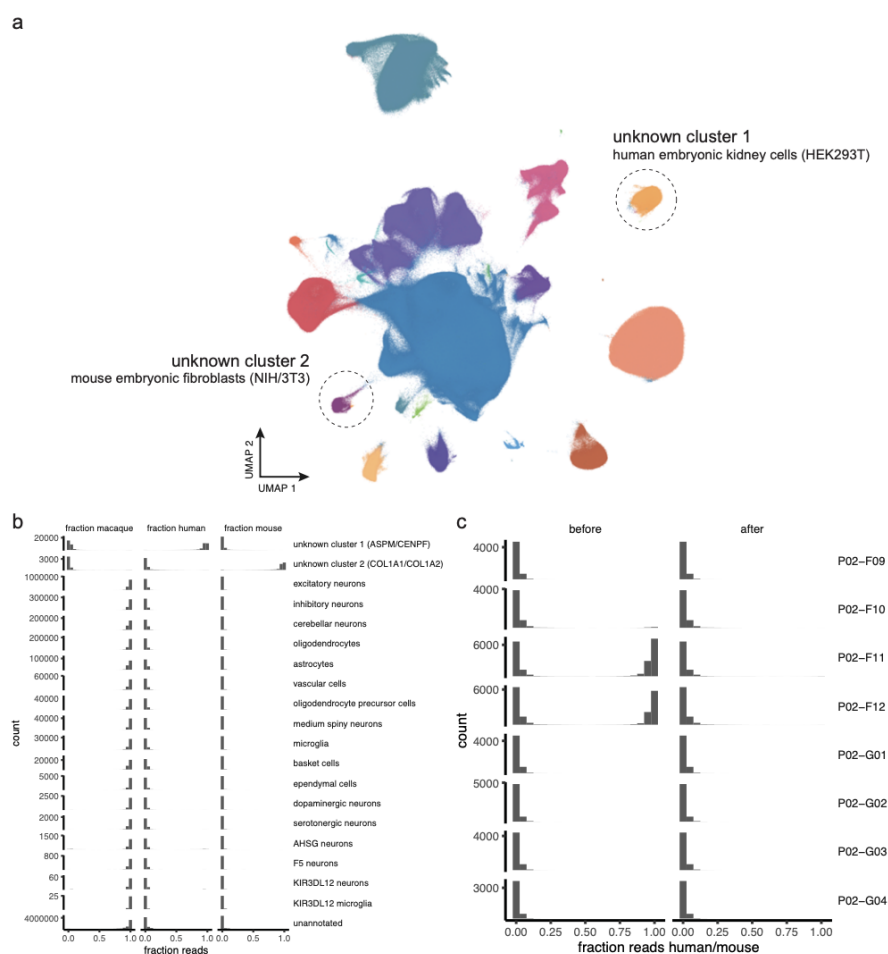


Figure 4.8: Fig. S2. Identification and removal of exogenous nuclei. **A**, Two anomalous clusters were identified during the course of cell-type annotation had marker gene profiles (unknown cluster 1 : ASPM, CENPE, CENPF, MKI67; unknown cluster 2 : COL1A1, COL1A2, FN1, VIM) characteristic of stem cell progenitors. **B**, Using the BBSplit multi-genome mapping strategy, reads were assigned to either the rhesus macaque, human, or mouse genomes. Histograms showed that exogenous (human or mouse) reads were specific to the two anomalous clusters and identified them as human-derived (unknown cluster 1) and mouse-derived (unknown cluster 2), respectively. **C**, Histograms of exogenous read fractions reveal that exogenous reads were specific to particular reversetranscription (RT) barcodes. The 8 barcodes shown (named according to plate number and position in 96-well plate) were assigned to equal aliquots of a single tissue sample, the right SPP from individual 2C0. Reads associated with two barcodes (P02-F11 and P02-F12) showed clear evidence of contamination (notably, a human-mouse barnyard control sample was loaded in adjacent wells P02-G11 and P02-G12). After the two anomalous clusters were removed from the entire dataset, these two barcodes no longer showed discernible evidence of exogenous contamination, indicating that human-and mouse-derived nuclei had been effectively.

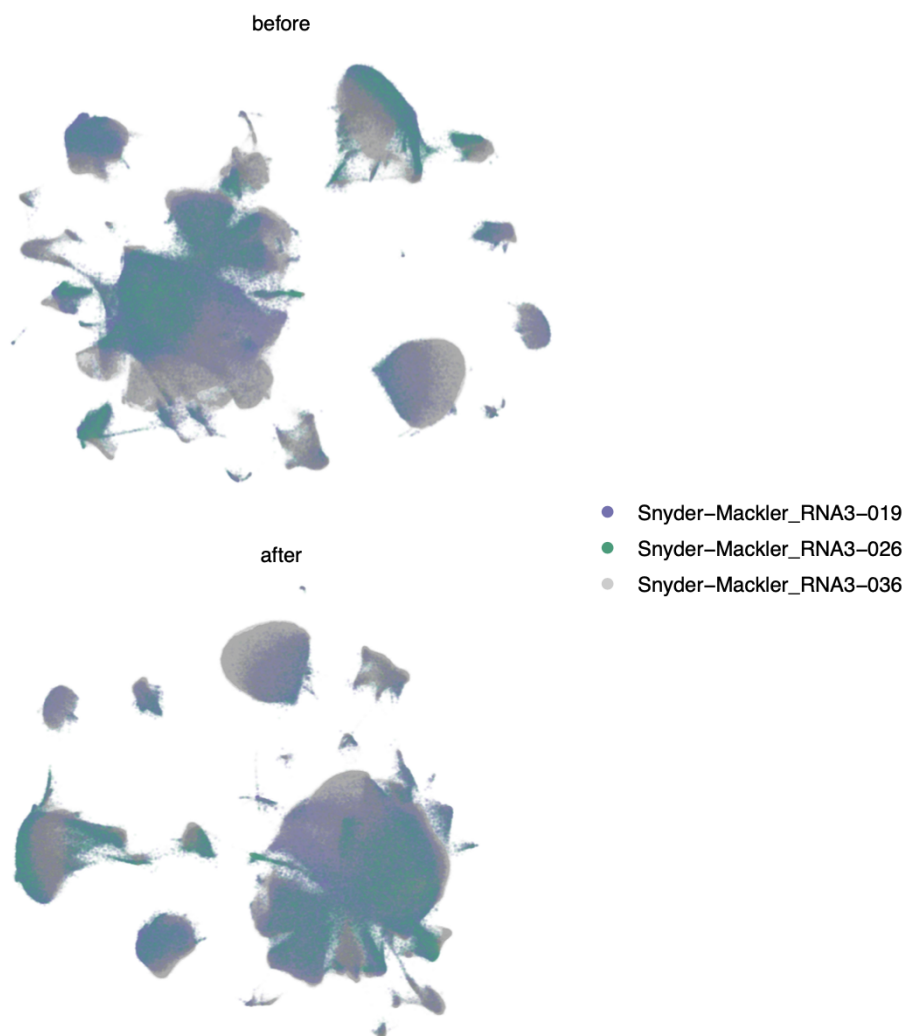


Figure 4.9: **Fig. S3. Comparison of UMAP projections before batch and after batch correction.**

The UMAP projection prior to batch correction was generated using the Scanpy ‘neighbors’ function to build a neighborhood-graph while the UMAP projection with batch correction used BBKNN in place of ‘neighbors’. Colors highlight the three library-preparation/sequencing batches, with the third batch shown in a lighter gray color with increased transparency due to the higher nuclei numbers from this batch implementing protocol improvements.

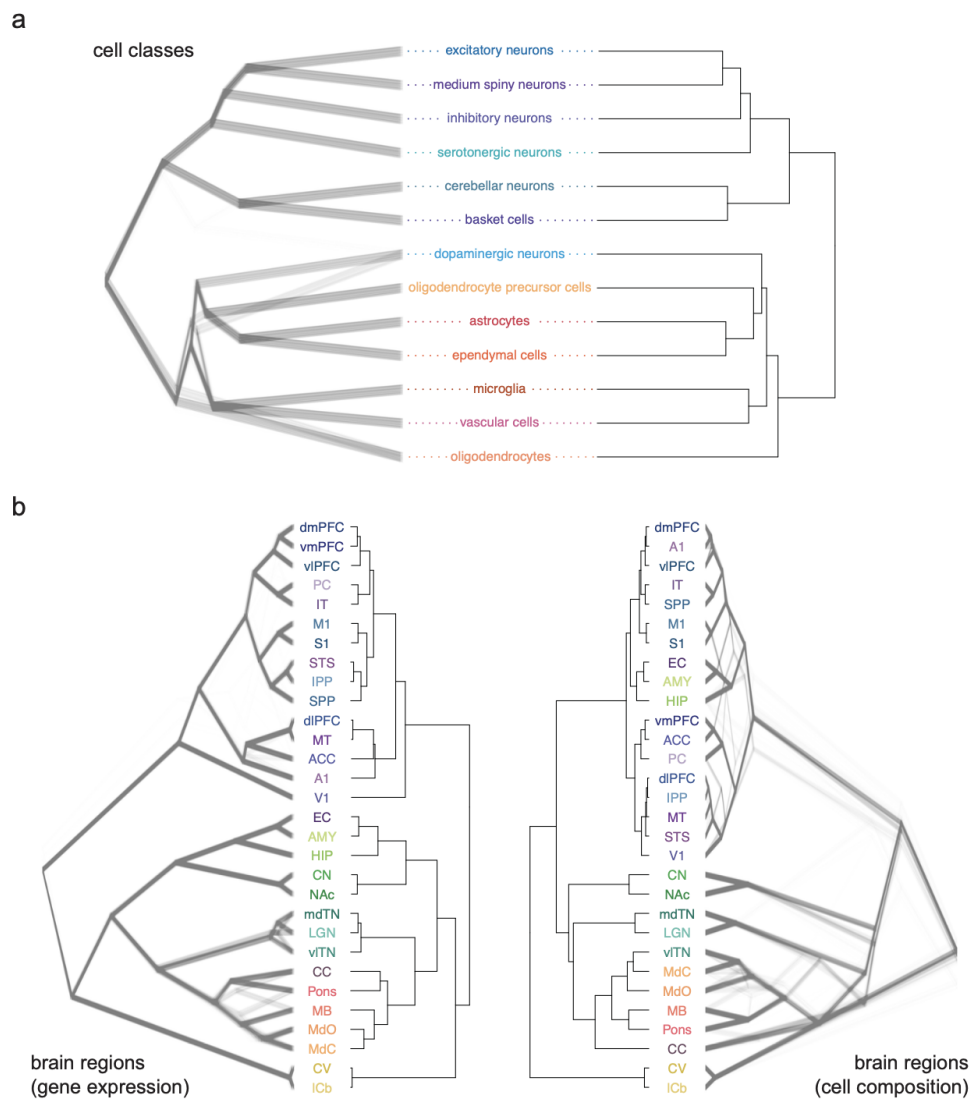


Figure 4.11: **Fig. S5. Unsupervised clustering of cell classes and brain regions.** **A**, Dendrograms showing unsupervised hierarchical clustering of cell classes by the top 50 principal components of gene expression. The consensus tree is shown on the right, opposite an uncertainty tree derived from 1,000 bootstrap replicates. **B**, Dendrograms showing unsupervised clustering of brain regions by, left, the top 50 principal components of gene expression and, right, relative proportions of cell classes. Consensus trees are shown opposite uncertainty trees which were also each derived from 1,000 bootstrap replicates.

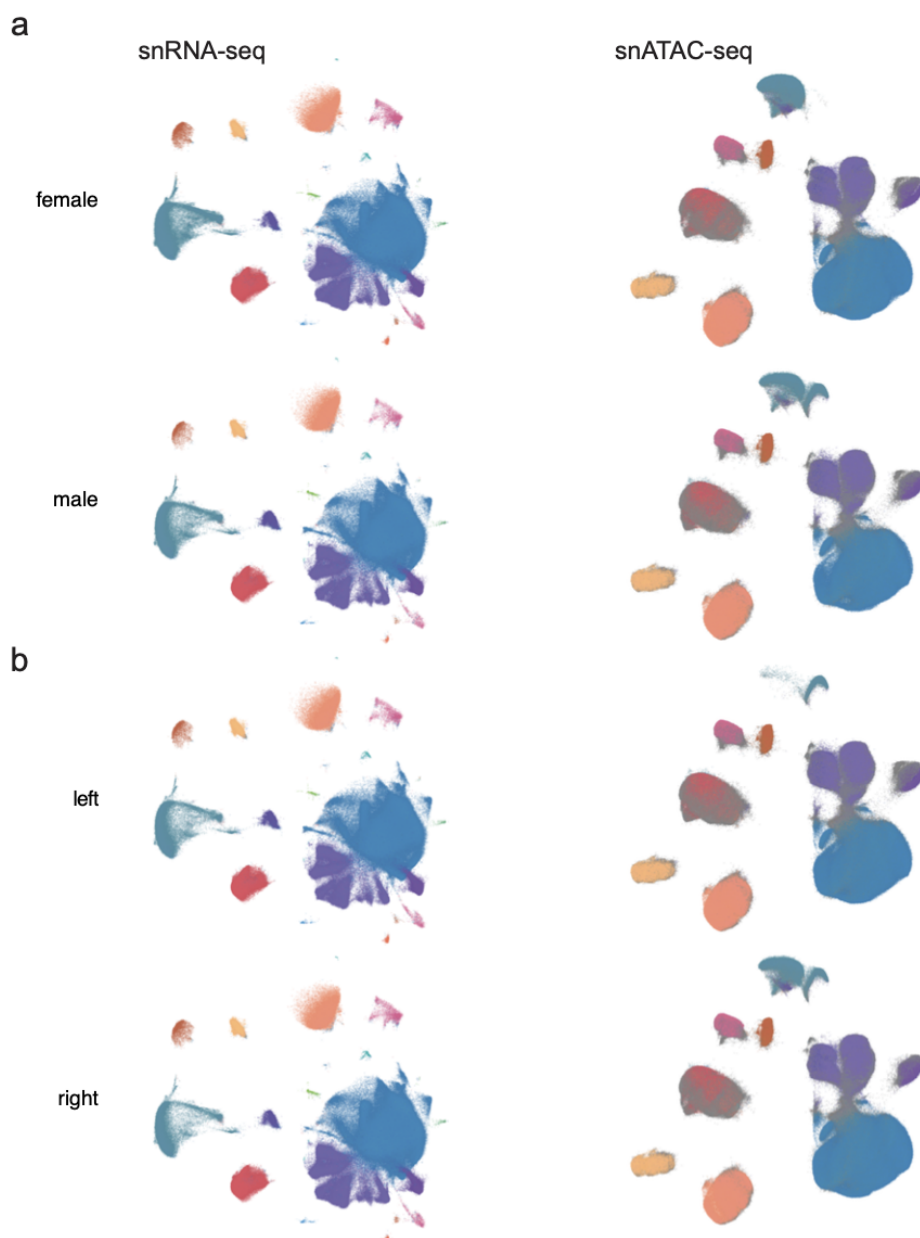


Figure 4.12: **Fig. S6. Comparison of UMAP embeddings of nuclei derived from samples of different A, biological sex and B, brain hemisphere.** snRNA-seq data are shown on the left and snATAC-seq data are shown on the right. For the snATAC-seq dataset, nuclei lacking cell-class assignments are shown in gray. All other colors follow the color scale in Fig. 4.1B and Fig. 4.2G. For hemisphere comparisons, nuclei from the cerebellar vermis (Vrm) and midbrain (MB) are not shown because the structures are located on the midline and were sampled from either or both hemisphere(s) depending on where they were located.

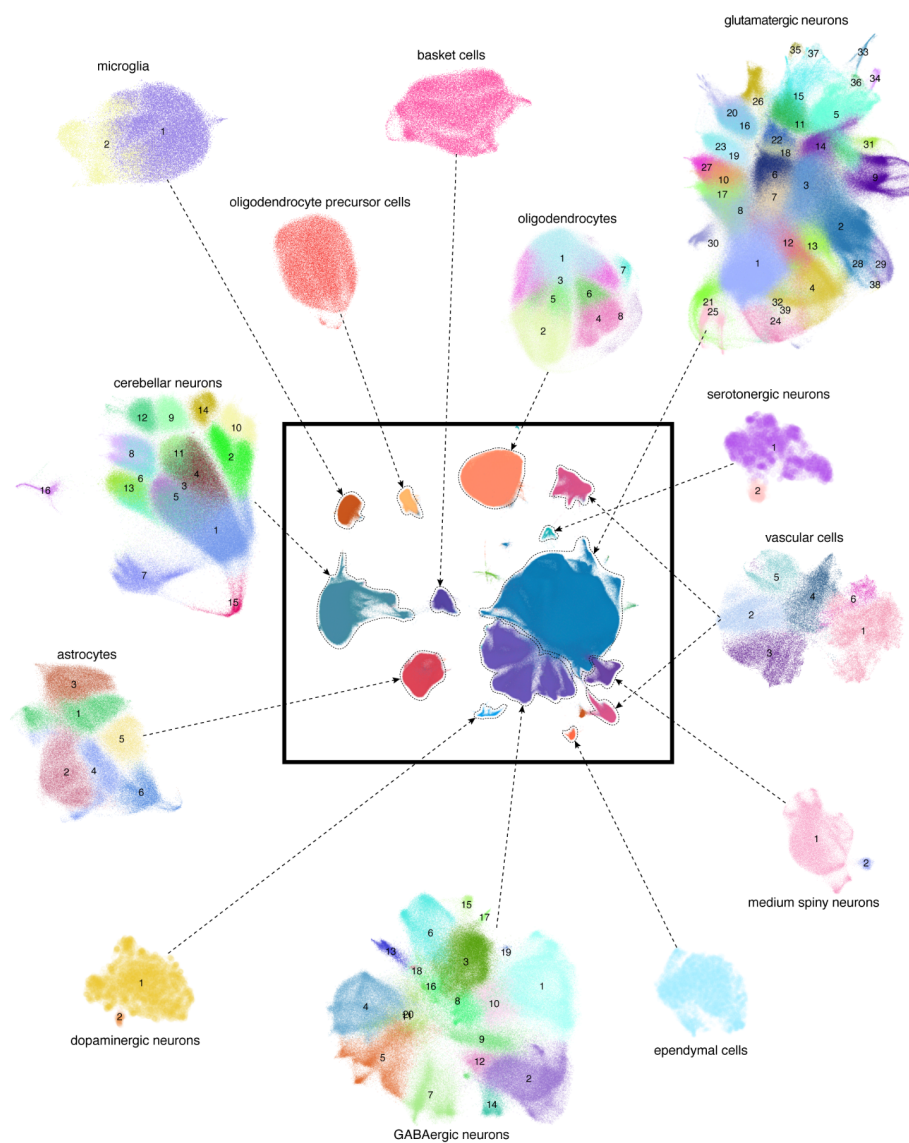


Figure 4.13: **Fig. S7. Cell-class specific UMAP projections colored and labeled according to identified cell subtypes.** To identify cell subtypes, the dataset was partitioned by cell class and preprocessing, clustering, and annotation steps were repeated on each partition separately. Cell subtype colors were generated separately for each cell class partition using the randomcoloR/v.1.1.0.1 package in R.

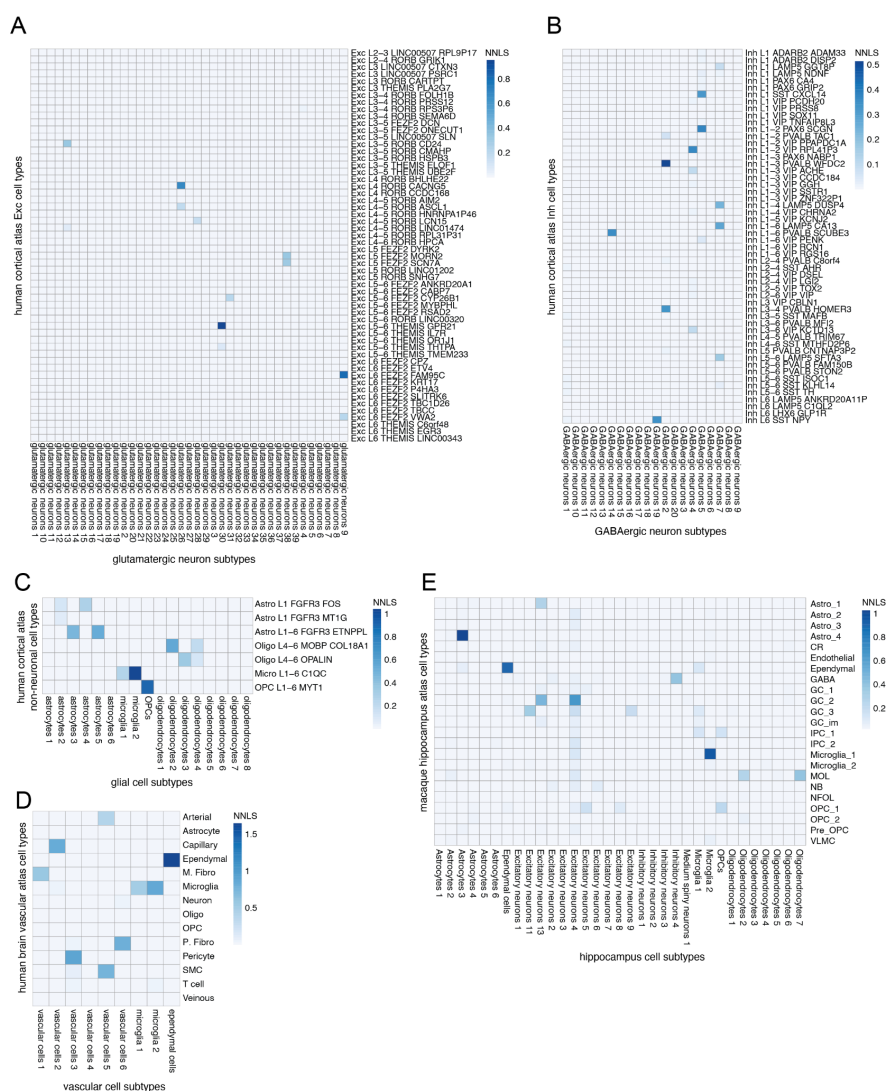


Figure 4.14: **Fig. S8. Cell subtype concordance with human cortical, human brain vascular cell and macaque hippocampus atlas.** **A-C**, Correlations between cell subtypes and annotated labels in a cortical human brain atlas. Combined β values from bi-directional non-negative least squares (NNLS) regression are shown for adult macaque cell subtypes (x axis) and cortical cell types annotated in the Allen human cortex dataset comprising several cortical brain regions (y axis). Glutamatergic neuron, GABAergic neuron and glial subtypes are shown in three separate panels. **D**, Correlations between cell subtypes and annotated labels in a human brain vascular cell atlas. Combined β values from bi-directional NNLS regression are shown for adult macaque cell subtypes (vascular, myeloid, and ependymal cells, x axis) and reference cell types in the vascular cell atlas (y axis). **E**, Correlations between cell subtypes and annotated labels in a macaque hippocampus cell atlas. Combined β values from bidirectional NNLS regression are shown for adult macaque glutamatergic neuron, GABAergic neuron and glial cell subtypes that are sufficiently abundant in the hippocampus ($N > 100$, x axis) and reference cell types in the macaque hippocampus cell atlas (y axis).

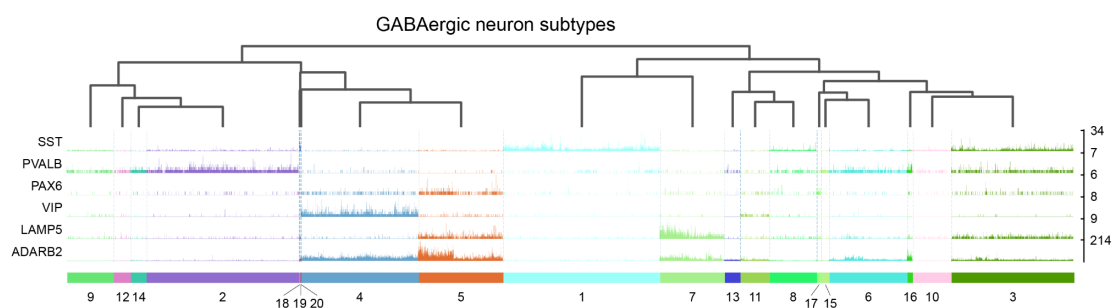


Figure 4.15: **Fig. S9. GABAergic neuron subtype markers** "Tracks plot" showing snRNA-seq read counts for cells assigned to each of 20 GABAergic neuron subtypes for six known GABAergic neuron markers. Dendrogram is based on hierarchical clustering of the top 50 principal components of gene expression in the GABAergic neuron class dataset partition.

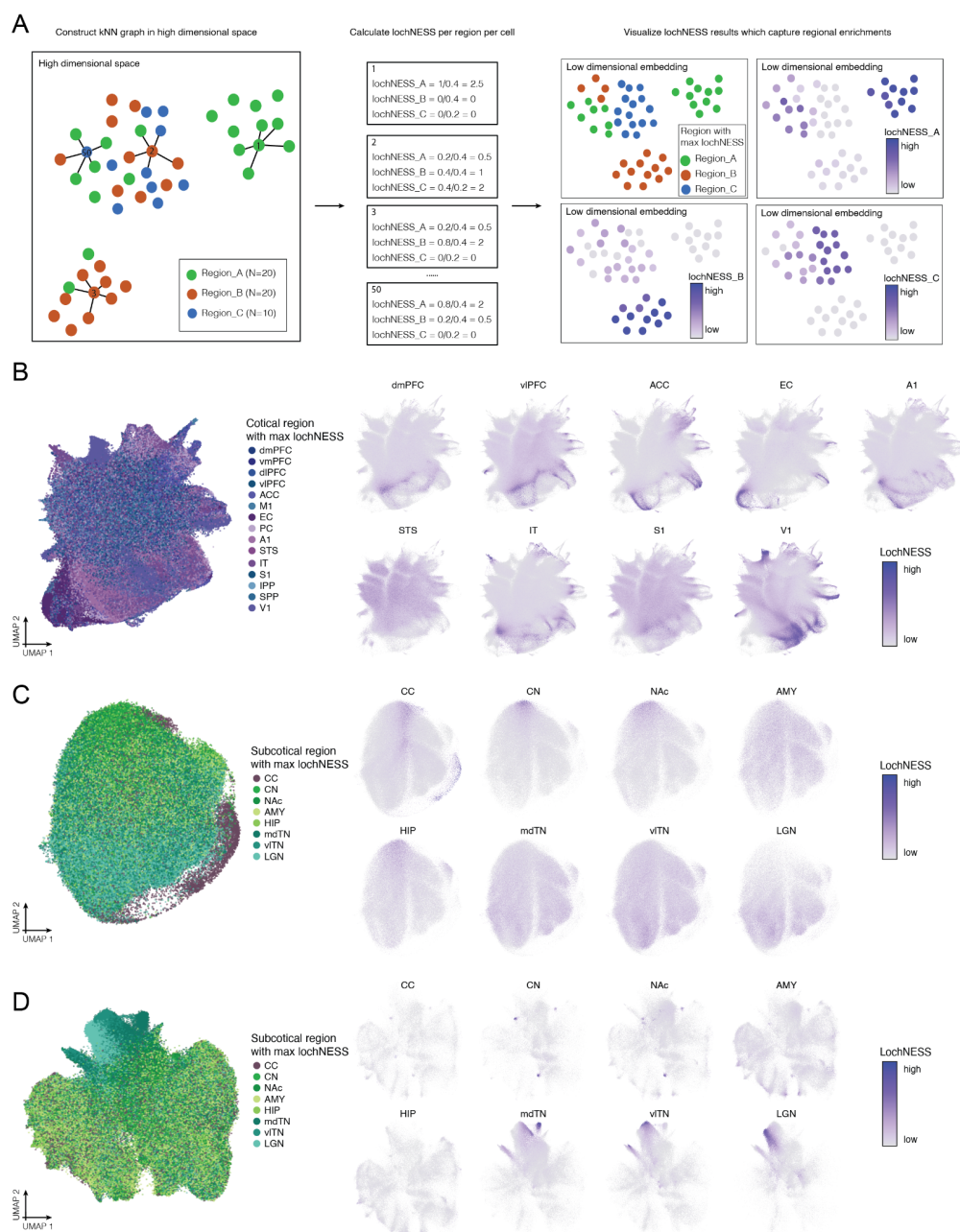


Figure 4.16: **Fig. S10. Schematic of extended lochNESS and additional lochNESS examples.** **A**, Schematic of the lochNESS analysis. **B**, UMAP visualizations of glutamatergic neurons colored by the cortical region with the highest lochNESS (left). LochNESS distributions in a subset of cortical regions are shown in separate panels (right). **C**, UMAP visualizations of oligodendrocytes colored by the subcortical region with the highest lochNESS (left). LochNESS distribution in subcortical regions are shown in separate panels (right). **D**, UMAP visualizations of GABAergic neurons colored by the subcortical region with the highest lochNESS (left). LochNESS distribution in subcortical regions are shown in separate panels (right).

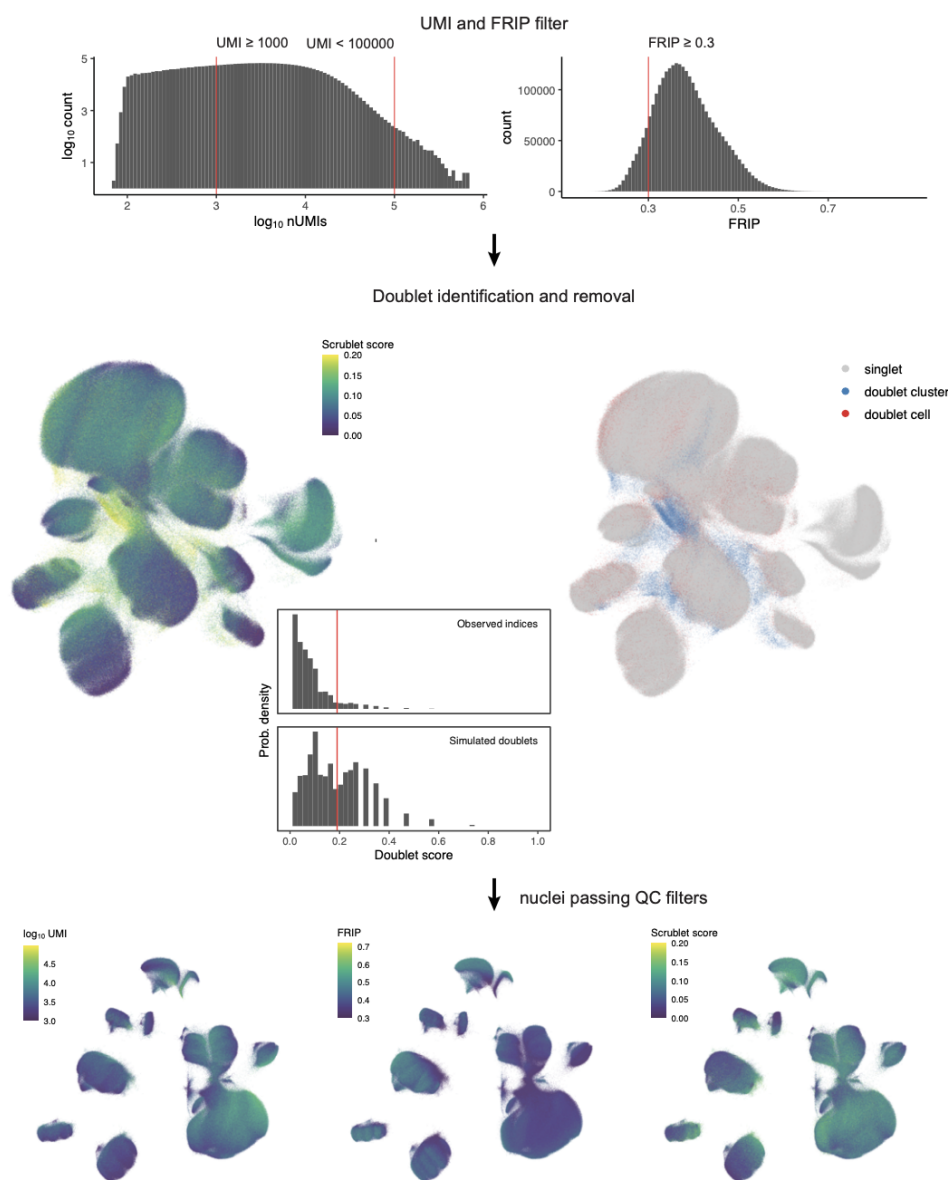


Figure 4.17: **Fig. S11. Schematic depicting snATAC-seq quality control pipeline.** **A**, Nuclei (combinatorial indices) with fewer than 100 or greater than 100,000 UMIs were removed, as were nuclei with fractions of reads in peaks (FRIP) < 0.3 . **B**, Scrublet k-nearest-neighbor (kNN) doublet scores were calculated per-sample and doublets with scores > 0.20 were marked (using automated Scrublet thresholds with manual adjustment) but not removed. All nuclei, including doublets, were then jointly preprocessed and clustered. Clusters with mean doublet scores > 0.15 were then removed along with previously marked doublets. **C**, UMI counts, FRIP, and Scrublet doublet-detection scores visualized on the post-quality-control dataset.

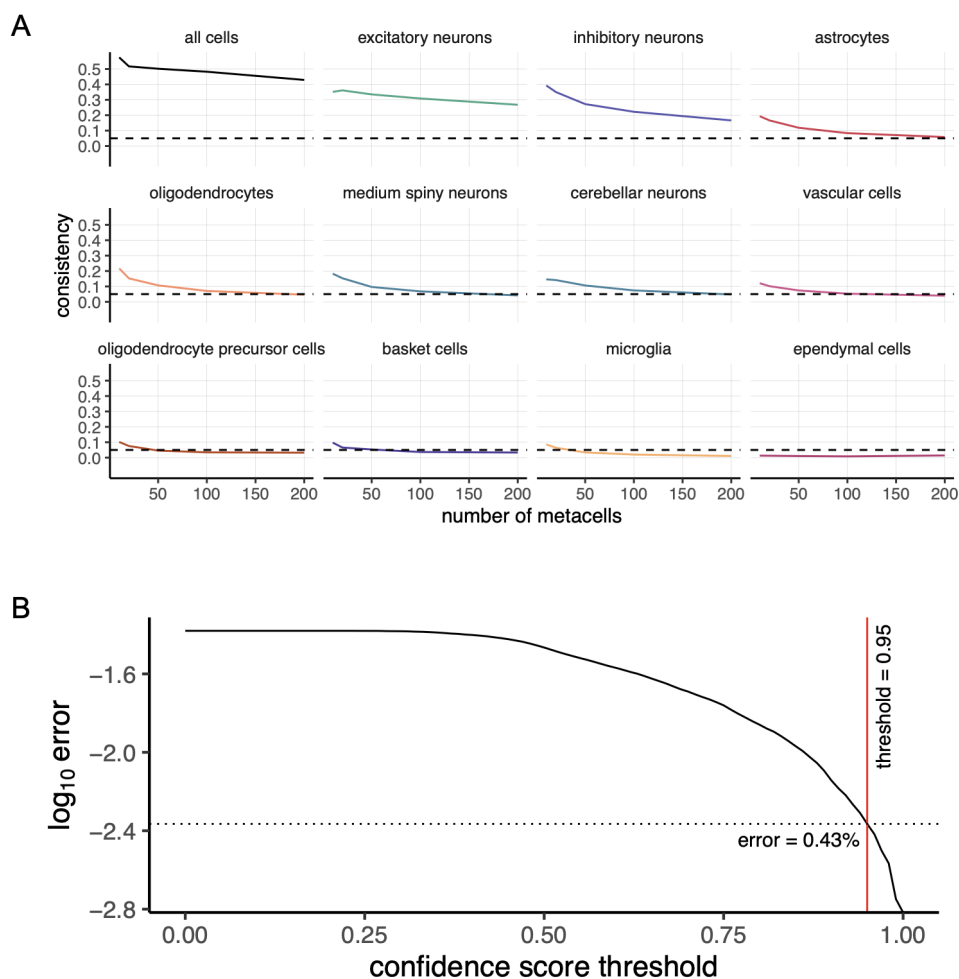


Figure 4.18: **Fig. S12. Assessment of snRNA-seq/snATAC-seq integration quality.** **A**, Integration consistency scores calculated by grouping neighboring cells into "metacells" and computing correlations were calculated using glue and are plotted here. These plots include both integrations performed at the dataset-wide level ("all cells") and at the cell-class-specific level. Integrations are considered more reliable the higher the curve is. We observed that larger cell classes tended to have higher integration consistency scores. **B**, Cell-class prediction accuracy was calculated using an evaluation dataset of 100,000 snRNA-seq cells not used in the reference dataset. A range of confidence score thresholds were then tested. At a confidence threshold of 0.95 (the chosen threshold), the prediction error (percentage of incorrectly predicted cell-class labels) was 0.43%.

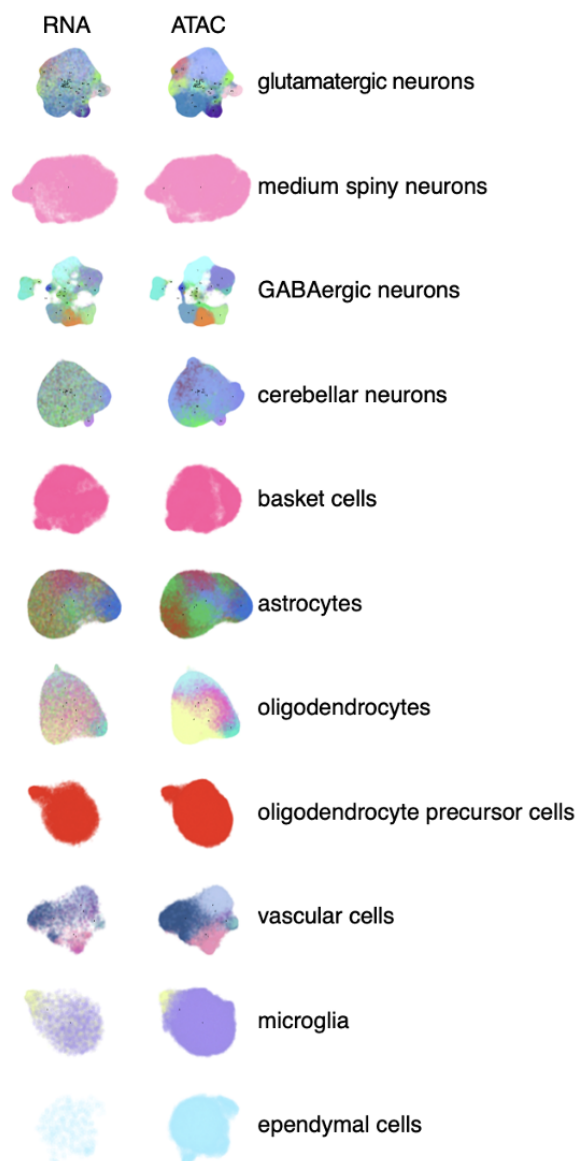


Figure 4.19: **Fig. S13. Integration of snRNA-seq and snATAC-seq data in cell classes.** UMAP embeddings of snRNA-seq and snATAC-seq data integrated separately across cell classes. Cells are colored according to annotated or predicted cell subtypes and match the colors in **Fig. 4.13**

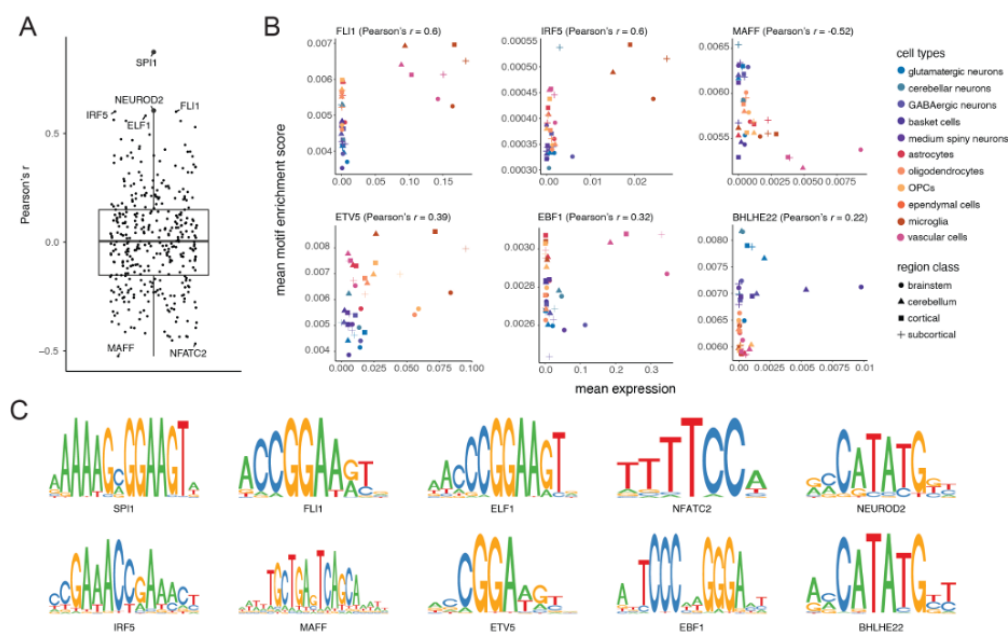


Figure 4.20: **Fig. S14. Correlation of TF expression and motif enrichments.** **A**, Boxplot showing distribution of Pearson's correlation coefficients with TFs with largest coefficient values labeled. **B**, Scatterplots showing correlation between snATAC-seq accessibility of TF binding motifs and snRNA-seq gene expression of corresponding TF genes within cell classes in regional classes for six TFs (in addition to examples in Fig. 4.4D). These TFs with cell-class-specific activating and repressing effects were selected either by systematically screening for large coefficient values (top row) or manual inspection (bottom row). **C**, Position weight matrices of the ten TF motifs shown in the TF expression and motif enrichment correlation analysis.

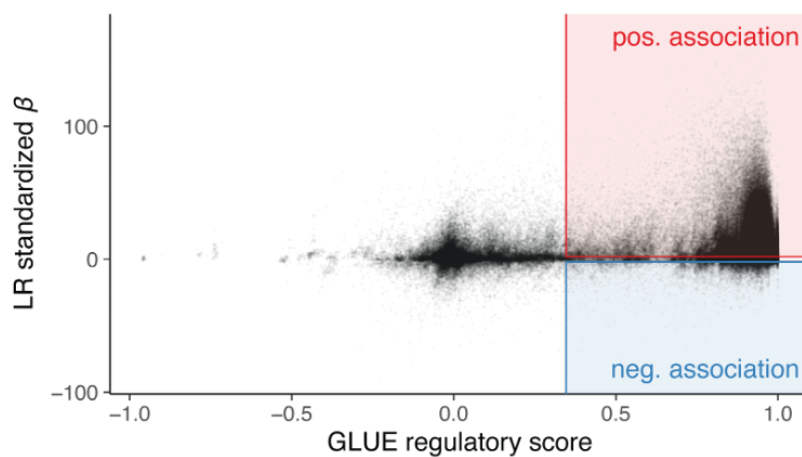


Figure 4.21: **Fig. S15. Concordance between GLUE regulatory scores and logistic regression standardized effect sizes.** Concordance between GLUE regulatory scores—which measure the cosine similarity between genes and peaks in the integrated multidimensional embedding—and logistic regression standardized effect sizes (standardized β), calculated based on metacells. Shaded areas encompass peak-gene pairs that are identified as candidate regulatory interactions ($P_{adj} < 0.05$ for both statistics), with the color depicting the direction of the association based on the sign of the logistic regression β estimate.

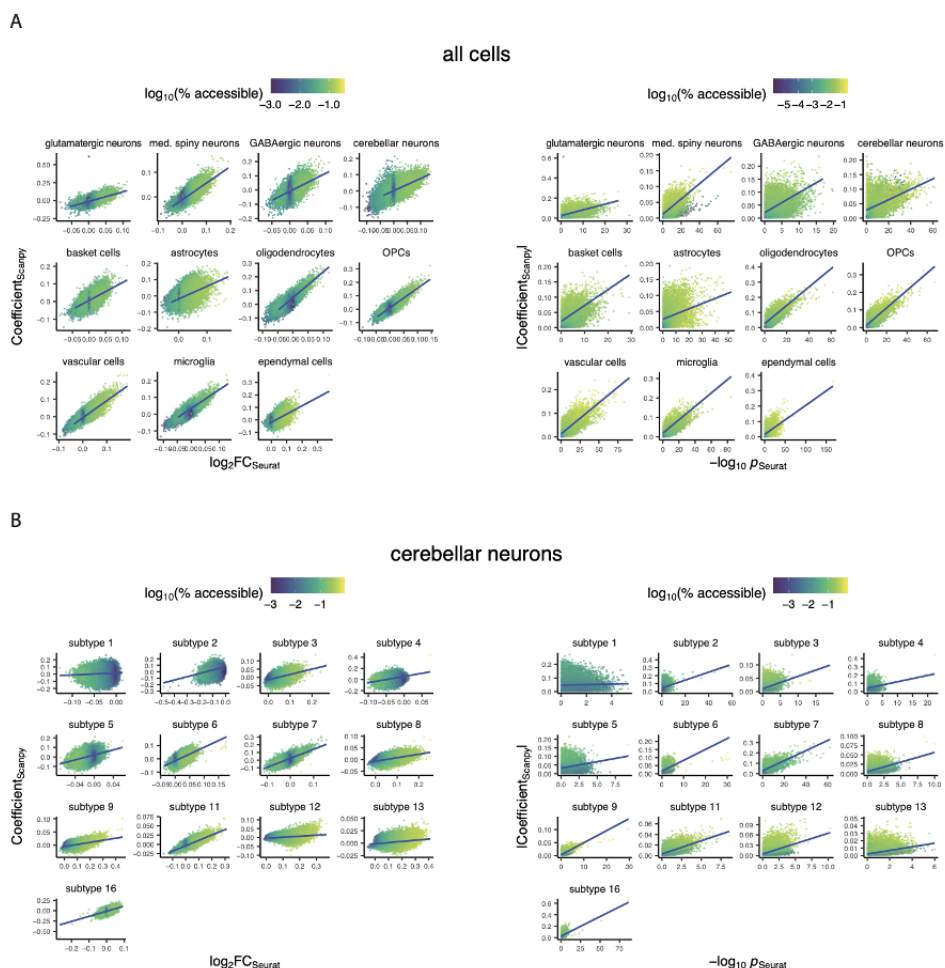


Figure 4.22: **Fig. S16. Validation of marker peaks.** Due to the large size of the data, the relatively computationally tractable regularized logistic regression method implemented in Scanpy was used to calculate differential accessibility across peaks. The regularized logistic regression, however, is unable to calculate P values or to control for covariates such as UMI counts. We therefore subsampled our dataset (1,000 cells per cell class or subtype) and repeated the analysis using the Seurat implementation (‘FindMarkers’) of logistic regression, which allowed us to control for UMI counts and to calculate P values. We confirmed that results were similar, both at the **A**, class, and **B**, subtype levels (a representative subtype-level analysis is shown here for the cerebellar neuron data partition).

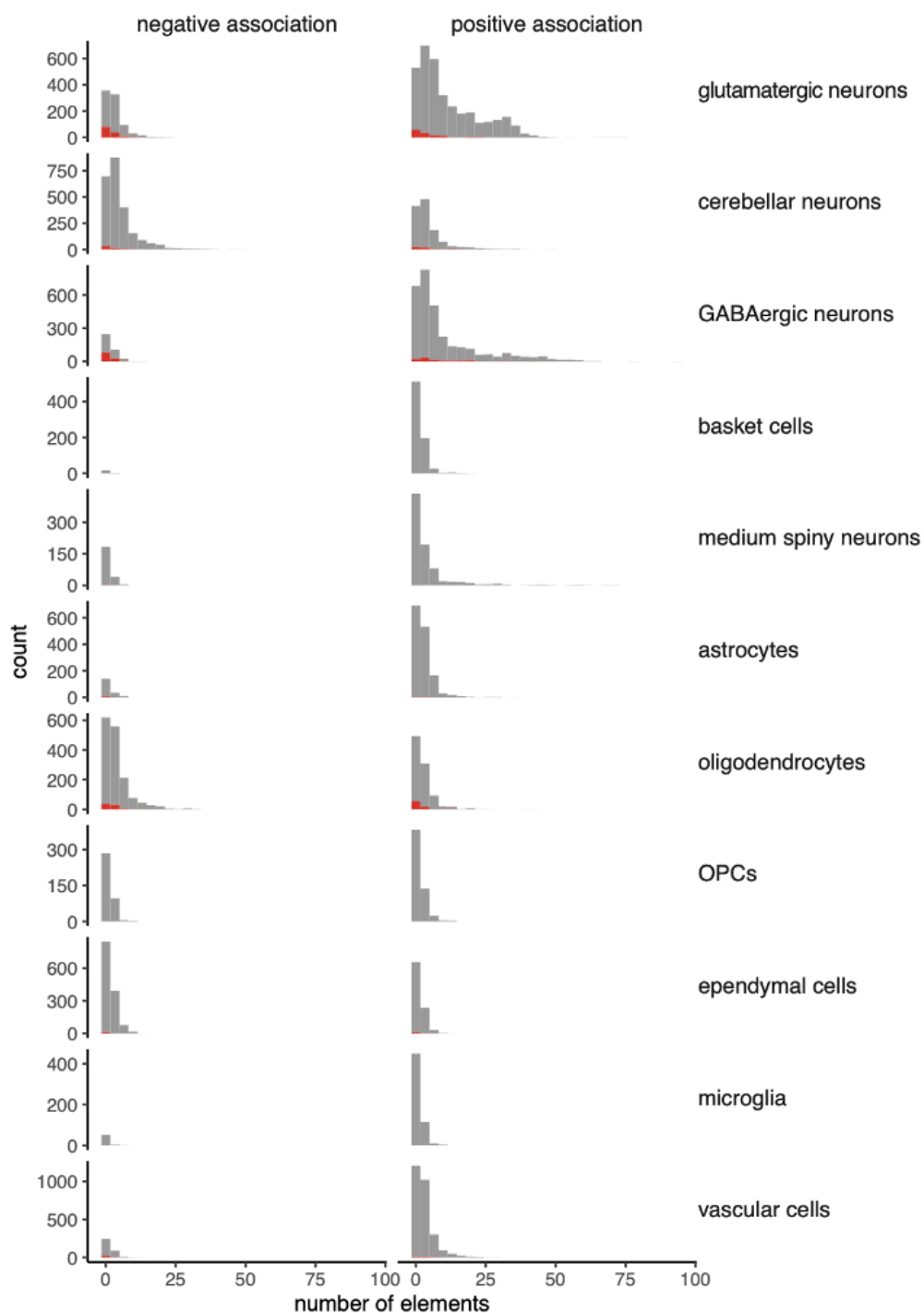


Figure 4.23: **Fig. S17. Number of cCREs interacting with a gene.** Histograms showing the number of cCREs (peaks) identified as interacting with a gene for each of 11 cell classes. Interactions are grouped separately into positive and negative interactions. Genes having both positively and negatively associated cCREs are shaded in red.

Chapter 5

DISCUSSION

In this dissertation, we introduced three major projects, the *Drosophila* Embryonic Atlas Project (DEAP), the Mutant Mouse Cell Atlas (MMCA), and the Macaque Brain Atlas (MBA), aimed at improving our understanding of embryonic development and complex tissues. These projects produced high-dimensional single-cell datasets of developing embryos and tissues and are challenging to analyze due to the complexity of the experimental design, which includes multiple time points, different genotypes, and different tissue regions. My work focused on developing novel computational tools as part of these projects.

In DEAP, the focus is on gaining a continuous view of molecular changes along trajectories in embryonic development. The cellular processes of embryonic development, including cell proliferation and differentiation, are highly dynamic and follow complex branching lineage structures. Despite previous research efforts, there is still much to learn about the molecular events underlying lineage specification in embryonic development of fruit flies. To this end, we generate single-cell transcriptomic and epigenomic datasets to provide a comprehensive view of the molecular landscape of *Drosophila* embryonic development. The major computational advances include a neural network for inferring the developmental age of each cell based on molecular profiles, and systematic construction of lineage trees using a k -NN approach.

MMCA aims to establish scRNA-seq as a tool for systematic phenotyping. We use scRNA-seq to generate a comprehensive single-cell transcriptomic atlas of the E13.5 mouse embryos from wildtype mice of different background strains and various disease relevant

mutant mice. My work focused on developing a computational framework for comparative single-cell analysis. The first part of this framework focuses on cell type composition analysis, where we use beta-binomial regression to identify statistically significant cell sub-trajectory proportion changes in mutants. Then we introduce lochNESS, a novel statistic based on k -NN graphs that quantifies differential abundance of mutant cells. Finally, we extend lochNESS to calculate similarity scores, providing a distance or similarity metric between embryos or genotypes.

Finally, the Macaque Brain Atlas (MBA) aims to generate a comprehensive map of cell types and regions in the macaque brain. This is of particular importance because non-human primates, such as the macaque, serve as critical models for studying the human brain. In MBA, the computational efforts focused on dissecting the regional heterogeneity of cell types and integrative analysis to understand cell type-specific gene regulation. To understand regional heterogeneity of cell types, we are extending the lochNESS statistic developed in MMCA to a multiple comparison case, calculating regional lochNESS statistics for each brain region in each cell type. Leveraging the regional lochNESS statistics, we can identify regional specific cell type markers with a regression model. For integration of the RNA-seq and ATAC-seq datasets, we applied a recently developed model scGLUE. For regulatory inference, we use an ensemble mode, combining results from cosine similarity between scGLUE feature embeddings and results from a regression model to propose a set of high confidence candidate cis-regulatory elements.

5.1 Limitations and future directions

In conclusion, the three projects described in this dissertation are designed to provide a comprehensive understanding of embryonic development and complex tissues. I believe that the datasets, analyses and computational tools in these projects will provide valuable re-

sources for researchers studying development, disease and neuroscience, and will further our understanding of the molecular basis of metazoan development and complex brain tissues.

However, each of the projects, especially the computational aspects, have certain limitations and potential for future projects. In this final chapter, I would like to discuss the limitations and future directions of each project.

5.1.1 Limitations and future directions of DEAP

In DEAP, we utilized a neural network (NN) as our central computational model for inferring the precise developmental time for single cells based on molecular profiles. However, we acknowledge that we did not provide a systematic comparison between our NN model and existing methods, such as pseudotime analysis, which could be considered in future studies. Additionally, we did not explore the use of more complex NN model architectures or other parameter adjustments, such as dropout, which could further improve the accuracy of the model. One of the considerations when developing the time inference model was the tradeoff between interpretability and precision. Earlier, we also implemented a Lasso linear model, which allowed us to identify features of importance and gain insight into the regulatory mechanisms driving development. However, it was not as accurate as the NN model. Furthermore, we believe that our NN model has the potential to be extended for staging external datasets and for cross-species prediction of developmental time, which could be of great value in comparative developmental biology studies.

Another computational aspect in DEAP was lineage tree construction. Currently we are binning cells into two-hour time windows and connecting cell types in neighboring time windows using a k -NN approach. In constructing our lineage trees, we can go beyond the two-hour time windows for a more continuous view. While binning cells into two-hour time

windows was a practical approach for lineage tree construction, it may not capture more subtle changes that occur over shorter time frames. To obtain a more continuous view, we could potentially use a sliding window approach, where the window size is smaller than two hours. This would enable us to capture more fine-grained changes in the transcriptomic profiles of cells over time, which could improve the accuracy of the lineage tree. We can also go beyond the simple k -NN model and identify branching points more precisely. While the k -NN model is a useful way to connect cells across time points, it may not always accurately capture the complexity of cell fate decisions that occur during development. To identify branching points more precisely, we could consider more sophisticated models that consider the topology of the data, such as graph-based methods or diffusion maps. These approaches could enable us to better identify key transition points between different cell states and lineages. Additionally, we can further utilize the lineage trees and propose sets of TFs at branching points. By identifying key branching points in the lineage tree, we can pinpoint the transcription factors (TFs) that are likely to be important for driving cell fate decisions. We can use gene expression data from the cells that branch at a particular point to identify TFs that are differentially expressed between the two lineages.

While DEAP is a comprehensive atlas containing both gene expression and chromatin accessibility information spanning embryogenesis, our integrative analysis was limited to the cell type level. Future work utilizing the dataset can focus on integrative analysis at the single-cell level, both within the dataset and with external datasets. Within the dataset, we can work on integration of the RNA and ATAC data to better understand gene regulation and chromatin accessibility across different cell types and time points. This can lead to the creation of cell type-specific and time point-specific regulatory maps, which can be useful for identifying key regulators of development. With external datasets, future work can focus on integrating other modalities, such as spatial datasets or 3D reconstructed spatial data [287]. This can provide additional information about the spatial organization of cells during development and help us understand how gene expression and chromatin accessibility

patterns are related to cell positions.

Given a more integrated dataset, analysis can focus on specific time windows or specific tissues exhibiting large changes, such as early development, the transition from embryo to larva, or the transition from larva to adult. This can provide insights into key developmental processes and help us understand how gene expression and chromatin accessibility patterns change over a longer period of time. Future experimental work can focus on generating new data, such as mutant data or spatial data. This can provide additional information about the genetic and epigenetic factors that contribute to development and help us test hypotheses about the roles of specific genes and regulatory elements.

Overall, it is important to note that our current analysis relies on the modalities that we captured (expression and chromatin accessibility) and may miss other information like spatial coordinates. Therefore, it is important to understand that each representation is built on some level of simplification or abstraction and conveys a somewhat summarized view of the actual continuous process. Future work can focus on developing more comprehensive representations that capture additional aspects of development, such as spatial organization or regulatory networks.

5.1.2 Limitations and future directions of MMCA

In MMCA, one of the major challenges was to tease apart signal and noise. We faced this challenge due to the limited statistical power to detect mutant-related cell type composition and transcription changes, which was primarily due to the small number of replicates per genotype. With a small number of replicates, it becomes difficult to distinguish true biological signals from random noise. Future experiments should consider increasing the number of replicates per genotype, which can help increase the statistical power, thus allowing for more

accurate identification of true biological signals. Many of the statistical analyses in MMCA rely on the multiple replicates in each mutant and wildtype. However, it is important to note that there is always natural variation between individuals even within the same genotype or cell type. When evaluating differential abundance of cell types or mutant types, it is important to utilize the statistical power from multiple replicates to separate the true effects from natural noise. This can be achieved by using statistical tests that consider the natural variation between individuals, such as linear mixed models, or by calculating effect sizes and performing power analyses to ensure that the effects observed are statistically significant and biologically meaningful.

In MMCA, we introduced a novel statistic called lochNESS, which was used to identify cells with differential abundance between mutant and wildtype in k -NNs. While the performance of lochNESS was validated within the context of MMCA, we did not perform a systematic comparison of lochNESS with other available methods for differential abundance testing, such as LISI [27] and Milo [28]. To further validate and benchmark the performance of lochNESS, future work can consider developing simulations with known cell types and abundances and introducing different levels of noise and variability to assess the robustness and accuracy of the method. Methods like lochNESS may be of broad interest given the increasingly available disease vs. wildtype datasets and Perturb-seq experiments. By developing rigorous benchmarks and comparing lochNESS to other methods, we can better understand the strengths and limitations of the approach and identify potential applications in other experimental contexts beyond MMCA. This will ultimately contribute to the development of more robust and accurate methods for differential abundance testing in single-cell data analysis.

To inform validation experiments such as *in situ* hybridizations, spatial mapping was a crucial tool for following up on some of our analysis in MMCA. However, both the spatial transcriptomics datasets used as a reference and the spatial mapping methods still have

room for improvement. One major challenge with spatial transcriptomics techniques is the tradeoff between region-of-interest size and imaging time/cost and number of features. Most methods do not offer single-cell resolution and/or might need additional processing like cell segmentation. Additionally, spatial mapping methods can be difficult to evaluate, as there is limited ground-truth to rely on. Future work can focus on improving the resolution and accuracy of spatial transcriptomics datasets, as well as developing more robust and versatile spatial mapping methods.

Another more general challenge in single-cell genomics analysis that we noticed during analysis of MMCA was the difficulty in accurately determining the significance of gene expression changes due to the high levels of noise and low sample sizes in the dataset. Current methods for identifying differential gene expression often rely on cluster-based differential expression testing. Consequently, the sensitivity and specificity of detecting gene expression changes are dependent on the clustering results or grouping, which can be arbitrary in granularity. Cluster-based differential expression testing can also lead to “double-dipping” or circular reasoning, where gene expression patterns that are randomly distributed across cells are associated with clusters simply because clusters were defined on the dataset. Future work can focus on developing alternative statistical methods that can account for the high levels of noise and low sample sizes in single-cell genomics data, as well as identify marker genes while controlling for false discovery.

When we were trying to validate the results of MMCA using experimental methods, we found that some of the computationally identified phenotypes were too subtle to be reliably detected with conventional methods like *in situ* hybridization. To overcome this limitation, future validation experiments may need to use more advanced techniques such as Stereo-seq [288] or sci-Space [289], which can provide higher resolution information about gene expression and spatial localization. In addition, generating time course datasets that capture the dynamic changes in gene expression during development or in response to perturbations

may also be useful in validating computational predictions in MMCA.

Furthermore, additional datasets on the same mutant mice could provide more information about the phenotypes and allow for the generation of better hypotheses about the observed changes. For example, generating other single-cell data types such as scATAC-seq or Hi-C data could provide insight into the chromatin state and higher-order chromatin organization that underlies the observed changes in gene expression.

5.1.3 Limitations and future directions of MBA

In the analysis of MBA (and almost every other single-cell dataset), one of the first processing steps is to cluster the dataset and identify cell types. However, although clustering is an essential step in identifying different cell types and their specific molecular profiles, determining the appropriate number of clusters and the distance metric to use for clustering is not always clear-cut. The choice of clustering algorithms and clustering parameters is often arbitrary and can have a significant impact on the resulting clusters and the biological interpretation of the data. Moreover, in many cases, it can be difficult to determine whether the clusters identified represent biologically meaningful cell types or simply reflect technical variation or noise in the data. This can make it challenging to draw meaningful conclusions about the underlying biology of the system being studied. To address this issue, various strategies have been proposed for evaluating and validating clustering results. These include assessing the stability of clustering across different parameter choices, comparing clustering results to known cell types or biological markers, and using independent validation datasets to confirm the robustness of the clustering. In summary, it is important to be aware of the arbitrary nature of clustering parameters and the challenges associated with identifying biologically meaningful cell types when defining cell types in single-cell datasets. Careful evaluation and validation of clustering results can help ensure that the resulting clusters accurately reflect the underlying biology of the system being studied.

MBA is one of the largest single-cell datasets to date, with over 4 million cells across two data modalities. Such large datasets pose significant computational challenges that researchers must consider when designing processing pipelines. In previous years, most single-cell genomics analysis was performed using R, which was the dominant programming language for genomics analysis software. However, as the complexity and dimensions of single-cell datasets have increased, the memory limitations and processing speed of R may not be sufficient to handle the data. As a result, more researchers are turning to Python as an alternative to R for processing single-cell data. Python offers many advantages for large-scale data analysis, including its ability to handle large datasets efficiently and its compatibility with distributed computing frameworks and machine learning libraries. However, this shift to Python also presents a challenge for researchers who must port their existing analysis tools and pipelines to the new language. In addition to programming language considerations, there are other computational challenges associated with processing large single-cell datasets. For example, many algorithms used in single-cell analysis, such as clustering and dimensionality reduction, scale poorly with increasing dataset size. To address these challenges, researchers might need to explore new approaches, such as distributed computing and hardware accelerated deep learning, to enable faster and more efficient processing and analysis of large-scale single-cell datasets.

Integrating multi-omics data from single cells can provide a more comprehensive understanding of the molecular processes underlying cell behavior and function. In MBA, we used scGLUE to integrate the RNA and ATAC data modalities and obtained both feature embeddings and cell embeddings. These embeddings can be used to identify relationships between genes and regulatory regions in a cell type-specific manner. To achieve this, we developed an ensemble model that combined a feature correlation approach and a regression-based approach. However, building cell type-specific regulatory maps using single-cell multi-omics data is still a challenging task, and there is much ongoing research in this area. One of the limitations is the scarcity of high-quality multi-omics data that can be used to train

and validate models. Another challenge is the computational complexity of integrating and analyzing large-scale multi-omics datasets, especially unpaired datasets like our own. In addition, there is a need to develop standardized methods and benchmarks for evaluating the performance of cell type-specific regulatory mapping methods. Despite these challenges, the integration of multi-omics data has the potential to uncover novel molecular mechanisms underlying cell behavior and function, and it is an area of with opportunities for new methods and frameworks.

One limitation we encountered in MBA was the lack of novel biological insight from our data analysis. While we were able to identify novel cell types and regulatory elements, further experiments could use retrograde labeling techniques to understand connectivity between cell types. Retrograde labeling is a technique that can be used to study neural connectivity, which involves injecting a tracer molecule into a downstream target region and then allowing it to be transported back up to the upstream neurons that project to that region. By studying these upstream neurons with single-cell sequencing, researchers can gain insight into the circuitry and connections within the brain. Additionally, integrating datasets from other species with MBA could provide a broader context for understanding the evolution and conservation of these cell types and regulatory elements. By comparing the single-cell data from different species like rodents and human, researchers can gain insight into the similarities and differences in cellular diversity and gene expression patterns between species, which can inform our understanding of the evolution and development of different brain cell types.

5.2 General thoughts and directions on computational genomics

Next, thinking beyond the scope of our projects, I would like to start a more general discussion on computational methods in single-cell sequencing, while contemplating the general

directions of the field.

With the emergence of new experimental technologies, researchers can create atlases to profile the position, lineage, chromatin accessibility, gene expression and other modalities of individual cells in metazoan development and complex tissues. As researchers create more comprehensive descriptive single-cell atlases in systems and organisms of their interest, the scientific community can start working towards integrating and representing such datasets in informative and effective ways. Assuming in the future there exists full observatory data of the development of a complex model organism, we will be able to work towards a computational representation of the development process. In addition, as introducing changes to regulatory elements and coding sequences in a high throughput fashion becomes increasingly feasible, biological studies are at a turning point where we can move beyond descriptive assays and generate functionally meaningful models. A conceivable yet elusive computational model of embryogenesis would be able to not only fully describe the state of the system at any time along the development process, but also predict the change in the system when perturbations are introduced. Such computational models become more feasible with the exponential increase in sequencing power and computational power, providing larger amounts of data to learn from and more resources to train and store complex models on. Overall, we believe that our datasets and computational tools will contribute to this larger effort to provide a comprehensive understanding of embryogenesis in complex organisms. In the following sections I would like to discuss these ideas in more detail.

5.2.1 Validation and technology development

As the complexity of biological systems becomes better understood, it becomes clear that existing single-cell analysis methods are not always sufficient to capture the full scope of biological diversity. For example, current methods may be limited in their ability to capture

rare cell types or low-abundance transcripts, or may not be able to capture the modality of interest. There is consequently a need for new experimental techniques that can address these limitations and provide more comprehensive data. One promising area of development is the integration of multiple genomics modalities into a single experimental pipeline. For example, methods are being developed that can measure both the transcriptome and epigenome of individual cells simultaneously, allowing for a more comprehensive understanding of the regulation of gene expression. Similarly, methods that can measure both the transcriptome and proteome of individual cells are also being developed, which could provide insights into post-transcriptional regulation and protein-protein interactions. Additionally, methods to capture such information in a spatial context are becoming increasingly important. The development of such techniques will require advances in sample preparation, sequencing technology, and use of chemistry, but could ultimately lead to a much deeper understanding of the molecular mechanisms underlying biological processes. Additionally, there is a need for more sophisticated computational methods to analyze and interpret these data, as the complexity of the datasets generated by single-cell sequencing continues to grow.

Furthermore, it is essential to have methods in place to validate the accuracy and reliability of the data and analysis results of single-cell datasets. In the case of single-cell sequencing, this can be particularly challenging, as the datasets can be noisy and subject to technical artifacts. The conventional experimental approach to validation is to compare the results of single-cell analysis to other complementary techniques, such as fluorescence microscopy or *in situ* hybridization. Another approach is to use synthetic data or “ground truth” datasets to test the accuracy of different computational methods. However, current validations may not be able to fully establish the data-derived findings in single-cell datasets, given the complexity of some analyses. Developing such validation methods will be critical for ensuring that the results of single-cell analysis are robust and reproducible and can be used to drive further discoveries in the field.

5.2.2 Representation and visualization

The most common representation of single-cell datasets is a low dimensional embedding from the UMAP or t-SNE algorithm. These embeddings are useful for visualizing high-dimensional data in two or three dimensions, but they can introduce biases and result in non-representative visualizations. For example, UMAP and t-SNE embeddings can be affected by the choice of hyperparameters, and they can be influenced by the presence of batch effects and outliers in the data. In a recent preprint discussing the biases in results from applying t-SNE and UMAP to single-cell sequencing data, the authors pointed out that these commonly used dimensionality reduction methods do not preserve distances in high dimensional space well in the low dimensional embedding and suggested alternative approaches [290]. On the other hand, it is also important to note that preserving distances is not the main goal of such dimensionality reduction methods and it would be unfair to evaluate t-SNE and UMAP embeddings based on distance-preserving criteria. The value of t-SNE and UMAP embedding have been proved by the fact that cells of the similar type tend to cluster together in embedded space. In general, it is not futile to generate these low dimensional representations, but we should be careful in drawing conclusions based on coordinates in embedded space. After visual inspection of low dimensional embeddings and observing a potential effect (e.g. less expression of a gene in an area, less abundance of a mutant in an area), one should use valid metrics to quantify and test for the significance of the observed effect.

For developmental biology, there is even more room to think about the representation and visualization of the underlying dynamic and robust biological process. In the past, lineage trees and the Waddington landscape have been used to represent developmental processes. One example of a successful model for representing developmental processes is the Sulston lineage tree of *C. elegans*, which describes the entire developmental lineage

of each cell in the nematode [291]. However, constructing lineage trees for more complex systems can be challenging, as they are variable, large, and opaque. Additionally, certain cell types in more complex organisms can be difficult to trace due to their mobility or the lack of unique markers. Therefore, developing new experimental techniques and computational methods for lineage tracing will be essential for constructing comprehensive lineage trees for variable systems. Another way to conceptualize differentiation and development is through the Waddington landscape, a theoretical framework that models cells as progressing through a multi-dimensional “landscape”. It envisions the differentiation process as a ball rolling down a landscape of hills and valleys, where each valley represents a stable cell type, and the hilltops represent unstable cell states. The ball rolling down the hills represents the cell moving towards a stable cell state through the process of differentiation. It may be important to expand this concept and think of the landscape as dynamic landscape (i.e. the Waddington landscape is not static but can change over time). The landscape can be altered by various factors such as genetic mutations, environmental cues, and stochastic events, which can cause the ball to roll in different directions and end up in different valleys. It may also be important to think about the landscape with concepts of “velocity” and “acceleration”, which can be used to describe the speed and rate of change of the ball rolling down the landscape. In the context of cellular differentiation, velocity can refer to the rate at which cells move towards their final differentiated state, while acceleration refers to the rate at which the velocity changes over time. For example, a sudden change in environmental cues or genetic mutations may accelerate the rate of differentiation or change the direction of differentiation, leading to a different final cell state. It may also be interesting to compare the landscape with gradient descent, as both involve the concept of a landscape with multiple possible states and a process that leads towards a desired state. In gradient descent, the goal is to minimize a cost function by iteratively adjusting the parameters to move towards the minimum of the function, and the direction of movement is determined by the gradient of the cost function, whereas in the Waddington landscape, the direction is determined by the topology of the landscape and the position of the cell.

In a single-cell project, we often describe the resulting dataset as an atlas or map. It is interesting to think about in what capacity can these datasets be used as maps. Thinking in the context of digital maps (like Google Maps), we can draw analogies to various functions as follows:

- Search functions: In a single-cell map, search functions can be used to find specific cell types or genes of interest. This can help researchers to quickly locate and analyze the relevant data points in a large dataset.
- Zooming in and out: A single-cell map can be visualized at different resolutions or layers of information. Researchers can zoom in to view individual cells and zoom out to see the larger picture of the tissue or organ context (which is more straightforward in spatially resolved datasets). Additionally, the map can be layered with additional information such as gene expression levels, epigenetic marks, and cell lineage information, which can provide deeper insights into the cellular landscape.
- Reference mapping: In a single-cell map, reference mapping can be used to align a query dataset to a common reference. This can help researchers compare their dataset with other datasets and facilitate data sharing and integration. An illustration of the function can be found in Azimuth [292], a web-based platform that leverages existing single-cell datasets as references to automate the analysis of new datasets.
- Landmarks and annotations: Landmarks and annotations can be used to mark important features or special points in the dataset. For example, a researcher can annotate the center of a specific cell type in the dataset, which can help others quickly learn about the typical molecular signatures of that specific cell type (maybe with a understanding of the variance within the cell type) and identify relevant/similar data points.
- Path finding: In a single-cell map, path finding can be used to identify the develop-

mental or differentiation trajectories of different cell types. This can help researchers to understand the lineage relationships between different cell types and to reconstruct the developmental trajectories of tissues and organs. Additionally, path finding can also be used to identify potential cell type transitions or trajectories that are associated with specific biological processes or diseases.

Another important aspect after creating large single-cell datasets is interactive visualization, data sharing and data access. We hope to find effective ways to share our datasets and allow the community of developmental biologists and neuroscientists to utilize our datasets and computational tools. Interactive visualization was especially useful in MMCA where we presented the embeddings in three-dimensional space. This added to the challenge of presentation as points can be obscured by other dots in a 3-dimensional view presented on paper or a screen. Our proposed solution is to create an interactive web app that allows users to explore our dataset. The main interactive component was a 3-dimensional embedding of our dataset, and the user will be able to choose a subset of data to visualize (one trajectory in one mutant) and color the cells by any metadata column (e.g. sub trajectory or MT%) or by expression of a user-selected gene of interest. Since our dataset is naturally comparative by the experimental setup, we implemented side by side 3 dimensional views with synched cameras. Overall, such interactive tools would allow researchers to investigate any groups of cells from a mutant and genes of interest to gain insights from our dataset. An example platform towards facilitating data sharing and collaboration in genomics is the Chan Zuckerberg CELLxGENE Discover platform <https://cellxgene.cziscience.com/>. By curating high-quality published data and presenting each dataset in a uniform way, the platform enables researchers to easily access and explore various datasets. The streamlined data upload process offered by the platform also saves time and resources for researchers who want to share their data alongside their publication. Looking forward, the sharing of single-cell datasets from different biological systems and perturbations will allow for more

comprehensive and meaningful comparisons, ultimately leading to a deeper understanding of biological processes.

5.2.3 *Integration and models*

As more single-cell datasets are generated by different labs with different protocols on different systems, an important consideration is integration. Integration of single-cell datasets is crucial for achieving a more holistic understanding of cellular function. Integration can occur within the same data modality, where multiple experimental batches, species, techniques, or sequencing depths can be combined to increase statistical power and improve the quality of the resulting dataset. By combining multiple datasets, we can better capture the heterogeneity in the biological system, and we can also identify subtle differences in cell states that may not be observable in individual datasets. Integration can also occur across different data modalities, such as transcriptomics, epigenomics, proteomics, and imaging data. This is particularly important since no single data modality can fully capture the complexity of a biological system. By integrating multiple modalities, we can obtain a more complete picture of cellular function and the interplay between different levels of cellular organization, including gene regulation, transcription factor binding, cellular morphology, and so on. More specifically for example, the integration of transcriptomics and epigenomics data can reveal the relationship between gene expression and chromatin modifications, which can shed light on the regulatory mechanisms underlying cellular function. Integration of spatial or imaging data can provide spatial context and enable the identification of cellular subpopulations and cell types. Based on my personal experience, I have found that machine learning or deep learning-based methods such as scGLUE [35] and Tangram [293] hold the most promise for integrating diverse genomics datasets. These approaches leverage the power of neural networks to identify patterns and relationships within large multi-modal datasets. Overall, integration of single-cell datasets is essential for advancing our understanding of cellular

function and can lead to the development of more accurate and comprehensive models of biological systems.

It is interesting to contemplate the broader possibility of creating a global embedding onto which all cells can be mapped. A global embedding would enable us to compare and contrast cells from different experimental conditions, species, and tissue types in a standardized way, allowing for a full understanding of cellular function. However, it is important to consider the limitations of such an approach. One potential limitation is the technical variation between different experiments, such as differences in sequencing depth or library preparation methods, which may limit the ability to create a single global embedding. Additionally, the biological variation between different organisms or modalities may simply be too large to consolidate and would require separate embeddings. Furthermore, it may also be important to consider the computational limitations of embedding large numbers of cells from multiple datasets. The size of the resulting dataset, as well as the complexity of the biological and technical factors that contribute to the variation between samples, may make it challenging to create a single, unified embedding. Despite these challenges, the progress in developing methods for integrating multiple single-cell datasets and the progress in profiling more cells with more modalities are promising and we can expect more unified, multimodal, and integrative datasets in the future.

As such comprehensive datasets emerge, it will be important to start thinking about the data-derived models that can effectively extract biological insights while accounting for technical and biological confounding factors. This requires an interdisciplinary approach that involves biologists, statisticians, and computational scientists working together to develop and validate these models. Generally, we would be moving from descriptive models or building atlases to predictive models. As the field of single-cell genomics continues to mature, there is a shift towards developing predictive models that can provide a mechanistic understanding of the underlying biology. This involves moving beyond simply describing the

observed patterns in the data towards understanding the rules that govern these patterns. This requires the development of computational models that can simulate and predict cellular behavior in response to perturbations or environmental changes. At the more extreme end, there is a possibility of distilling a basic set of rules that can explain the fundamental processes underlying development and cellular behavior. Similar to the Game of Life, a system invented by mathematician John Horton Conway in 1970 and an example of a system that is built on simple rules but exhibits complex behaviors such as self-replication, gliders, and oscillators, development could have an underlying set of simple rules that are capable of generating complex patterns and behaviors. While building such a model may be an ambitious or even infeasible goal for a complex system like mammalian development, simplified models for specific subsets of cells or time windows seem to be within reach given current datasets. In the middle ground between fully descriptive datasets and rule-guided models, there are many levels of abstractions and simplifications that can be applied. For example, the aforementioned models of lineage trees and the Waddington landscape are useful tools for visualizing and conceptualizing cellular differentiation and development, with necessary simplifications of the complex reality of these processes. As more comprehensive datasets become available, it becomes possible to develop increasingly sophisticated models that can capture the nuances of these processes while remaining tractable for analysis. Eventually, we can imagine the field of single-cell genomics going towards developing a digital model of development that can accurately capture the dynamics of cellular behavior from the earliest stages of embryogenesis to the mature organism in a mammalian system. This would probably require the integration of multiple data modalities, including genomics, epigenomics, proteomics, and imaging data, into a coherent computational framework. Such a model would not only describe the state of the system at any time during development but also predict the changes that occur when perturbations are introduced. My perspective is that artificial intelligence has great potential to contribute to the development of data-driven digital models for complex biological processes. As sequencing and computational power continue to grow exponentially, we will have increasing amounts of data and hardware resources to

train models on, making this goal more achievable.

When developing computational tools and models, it is important to keep the focus on biology and ensure that the research is addressing relevant biological questions. There is a growing interest in using genomics techniques to study complex diseases such as cancer, neurodegenerative diseases, and developmental disorders. By understanding the molecular mechanisms underlying these diseases, researchers can identify new therapeutic targets and develop better treatments. For example, genomics can be used to identify genes that are mutated in cancer cells or to study gene expression changes in neurodegenerative disorders. This knowledge can then be used to develop targeted therapies that are tailored to individual patients based on their genomic profile. In addition, genomics can also be used to study the genetic basis of complex traits such as susceptibility to diseases, which can lead to a better understanding of the underlying biology of these traits.

Finally, it is important to acknowledge that genomics is a highly interdisciplinary field that requires researchers to have expertise in multiple areas. Genomics, especially data analysis, involves the knowledge of a wide range of techniques, including sequencing, computational biology, statistics, machine learning, and bioinformatics. Properly using tools from other areas such as statistics and machine learning, and integrating best practices from software engineering and benchmarking, can help ensure that computational genomics research is conducted in a rigorous and reproducible manner.

5.2.4 *AI and Genomics*

The use of artificial intelligence (AI), machine learning (ML), and deep learning (DL) methods is becoming increasingly prevalent in the analysis of genomics data. These tools are being applied to various aspects of computational genomics data, including data process-

ing, analysis, and modeling. Within our projects, a few places where I found ML models helpful were the NN for staging cells in developmental time, spatial mapping to combine single-cell transcriptomics and spatial transcriptomics datasets and integrating RNA and ATAC datasets with graph-based models. Going forward, there are many places where I think AI methods would be helpful. First is the construction of lineage trees, where machine learning methods can be used to build models to infer lineage relationships from single-cell data without relying on pseudobulked datasets. The second is comparative analysis, which involves comparing single-cell data from different experimental conditions, such as healthy vs. diseased tissue, or wildtype vs. knockout cells. Machine learning can be used to identify differences in gene expression, cell type composition, and other features between these conditions. Contrastive learning is a specific type of machine learning that can be used to learn representations of data that emphasize differences between samples and could be useful in this case. Third is integration (especially with imaging data) and reference mapping of single-cell datasets. Machine learning can be used to integrate datasets and build reference maps that can help us understand the spatial organization of cells and their interactions. Fourth is sequence-based models for understanding regulation. Machine learning can be used to build models that can learn the rules governing gene regulation from DNA sequence data. This can help us understand how genetic variation affects gene expression and is especially promising given the recent success of large language models. Finally, there are genetics models for understanding causal relationships, where machine learning can be used to analyze large-scale genetic data to identify causal relationships between genetic variation and disease. This can help us identify new therapeutic targets and develop personalized medicine strategies.

It is also important to think about the challenges of applying AI tools to genomics datasets. Compared to fields where AI has been largely successful, like computer vision and natural language process, there is limited amounts of labeled data available in genomics. Unlike image datasets such as ImageNet, where millions of labeled images are available,

genomics datasets are often expensive to generate and hard to uniformly label. It can be difficult to accurately label genomic data due to the complexity of the biological systems being studied. There is also the challenge of genomics dataset being non-uniform. Biological images and sequencing data come in all sorts of platforms and formats, making it difficult to create a unified dataset for training and testing AI models. Additionally, different types of genomics data (such as RNA-seq, ATAC-seq and ChIP-seq) have different characteristics, making it challenging to apply the same AI model to different types of data. There are many experimental biases that can affect genomics data, such as differences in sequencing depth, variation in sample quality, or differences in lab protocols. These biases can be difficult to control for and may affect the performance of AI models trained on the data. Another challenge is benchmarking and establishing ground truth in genomics. Benchmarking is a crucial step in developing and evaluating AI models, but establishing ground truth in genomics data can be difficult. For example, while photo labels can be easily verified by humans, labels for cell types or genetic interactions can be more subjective and difficult to establish. Overall, addressing these challenges will require interdisciplinary collaboration between genomics researchers, computer scientists, and statisticians, as well as a better understanding of the biological systems being studied.

Apart from specific research projects, AI has the potential to make a significant impact in other areas of scientific work. For example, ChatGPT could be utilized for scientific writing or as a tool for scientific brainstorming. Given the following prompt: “In the next ten years, how can AI help genomics research”, ChatGPT (GPT3.5) produced the following answer:

AI has already made significant contributions to genomics research, and its potential to revolutionize this field is only growing. In the next ten years, AI is likely to help genomics research in the following ways:

1. Accelerating Genome Analysis: AI can help automate and accelerate genome analysis, making it faster and more accurate. By automating routine tasks, such as identifying and annotating genetic

variants, AI can allow researchers to focus on more complex analysis and interpretation.

2. Predicting Disease Risk: AI can be used to predict an individual's risk of developing a particular disease based on their genetic makeup. This can help clinicians tailor preventive measures and treatment plans to individual patients, improving overall health outcomes.

3. Drug Discovery: AI can help researchers identify new drug targets and develop more effective treatments for diseases. By analyzing vast amounts of genomic and biological data, AI can identify new drug targets and predict the effectiveness of potential treatments.

4. Personalized Medicine: AI can help tailor treatments to individual patients based on their genetic makeup, lifestyle, and other factors. This can lead to more personalized and effective treatments that improve patient outcomes.

5. Enhancing Research Collaboration: AI can help facilitate collaboration among researchers by providing tools for data sharing, analysis, and visualization. This can help accelerate scientific discovery and foster innovation in the field.

Overall, AI has the potential to revolutionize genomics research by accelerating analysis, predicting disease risk, facilitating drug discovery, enabling personalized medicine, and enhancing research collaboration.

And given the following prompt: "In the next ten years, how can AI help single-cell genomics analysis", ChatGPT produced the following answer:

Single-cell genomics is a rapidly growing field that aims to understand the genetic makeup and behavior of individual cells. AI can play a significant role in advancing single-cell genomics research by providing powerful tools for data analysis and interpretation. Here are a few ways AI could help in the next ten years:

1. Improved Data Analysis: Single-cell genomics generates vast amounts of data that require advanced computational methods for analysis. AI techniques such as machine learning and deep learning can help automate the analysis of single-cell genomic data, enabling researchers to identify patterns and associations that would be difficult to detect manually.

2. Cell Classification: AI can help classify cells based on their genomic profiles. For example,

deep learning algorithms can be trained to identify cell types and subtypes based on gene expression data. This could help researchers better understand cell differentiation and function.

3. Biomarker Discovery: AI can help identify novel biomarkers that can be used to diagnose diseases or monitor disease progression. By analyzing single-cell genomic data, AI algorithms can identify gene expression patterns that are associated with specific diseases, enabling the development of targeted therapies.

4. Personalized Medicine: AI can help identify the genetic variations that contribute to disease and enable personalized medicine. By analyzing single-cell genomic data, AI algorithms can identify specific genetic mutations or gene expression patterns that are associated with disease. This information could be used to develop personalized treatment plans that target the underlying genetic causes of disease.

Overall, AI has the potential to revolutionize single-cell genomics research by providing powerful tools for data analysis and interpretation. As the field continues to grow, we can expect to see more innovative applications of AI in single-cell genomics analysis.

Finally, after providing ChatGPT with the text from the title and abstract of the dissertation, I asked: “What would be a good ending sentence for this dissertation” and received the following response:

In summary, the integration of cutting-edge computational methods with experimental techniques has the potential to revolutionize our understanding of embryonic development and complex tissues, paving the way for new discoveries and advancements in the field of developmental biology and beyond.

BIBLIOGRAPHY

- [1] Cusanovich Darren et al. “The Cis-Regulatory Dynamics of Embryonic Development at Single-Cell Resolution”. In: *Nature* 555 (7697 2018), pp. 538–542.
- [2] Cao Junyue et al. “The Single-Cell Transcriptional Landscape of Mammalian Organogenesis”. In: *Nature* 566 (7745 2019), pp. 496–502.
- [3] Fuchou Tang et al. “RNA-Seq analysis to capture the transcriptome landscape of a single cell”. en. In: *Nat. Protoc.* 5.3 (Feb. 2010), pp. 516–535.
- [4] Jason D Buenrostro et al. “Single-cell chromatin accessibility reveals principles of regulatory variation”. en. In: *Nature* 523.7561 (July 2015), pp. 486–490.
- [5] Rodriques Samuel et al. “Slide-Seq: A Scalable Technology for Measuring Genome-Wide Expression at High Spatial Resolution”. In: *Science* 363 (6434 2019), pp. 1463–1467.
- [6] Smallwood Sébastien et al. “Single-Cell Genome-Wide Bisulfite Sequencing for Assessing Epigenetic Heterogeneity”. In: *Nature Methods* 11 (8 2014), pp. 817–820.
- [7] Ramani Vijay et al. “Massively Multiplex Single-Cell Hi-C”. In: *Nature Methods* 14 (3 2017), pp. 263–266.
- [8] Clark Stephen et al. “ScNMT-Seq Enables Joint Profiling of Chromatin Accessibility DNA Methylation and Transcription in Single Cells”. In: *Nature Communications* 9 (1 2018), p. 781.
- [9] Tabula Muris Consortium et al. “Single-Cell Transcriptomics of 20 Mouse Organs Creates a Tabula Muris”. In: *Nature* 562 (2018), pp. 367–372.
- [10] Qiu Xiaojie et al. “Reversed Graph Embedding Resolves Complex Single-Cell Trajectories”. In: *Nature Methods* 14 (10 2017), pp. 979–982.

- [11] Qi Mao et al. “Principal Graph and Structure Learning Based on Reversed Graph Embedding”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 39.11 (Nov. 2017), pp. 2227–2241.
- [12] Qiu Xiaojie et al. “Single-Cell mRNA Quantification and Differential Analysis with Census”. In: *Nature Methods* 14 (3 2017), pp. 309–315.
- [13] McInnes Leland, Healy John, and Melville James. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”. In: *ArXiv* (2018).
- [14] Farrell Jeffrey et al. “Single-Cell Reconstruction of Developmental Trajectories during Zebrafish Embryogenesis”. In: *Science* 360 (6392 2018).
- [15] Briggs James et al. “The Dynamics of Gene Expression in Vertebrate Embryogenesis at Single-Cell Resolution”. In: *Science* 360 (6392 2018).
- [16] Domcke et al. “A Human Cell Atlas of Fetal Chromatin Accessibility”. In: *Science* 370 (6518 2020). DOI: 10.1126/science.aba7612.
- [17] Packer Jonathan et al. “A Lineage-Resolved Molecular Atlas of *C. Elegans* Embryogenesis at Single-Cell Resolution”. In: *Science* 365 (6459 2019).
- [18] Wagner Daniel et al. “Single-Cell Mapping of Gene Expression Landscapes and Lineage in the Zebrafish Embryo”. In: *Science* 360 (6392 2018), pp. 981–987.
- [19] Pijuan-Sala et al. “A Single-Cell Molecular Map of Mouse Gastrulation and Early Organogenesis”. In: *Nature* 566 (7745 2019), pp. 490–495.
- [20] Cao et al. “Comprehensive Single-Cell Transcriptional Profiling of a Multicellular Organism”. In: *Science* 357 (6352 2017), pp. 661–667.
- [21] Mittnenzweig Markus et al. “A Single-Embryo, Single-Cell Time-Resolved Model for Mouse Gastrulation”. In: *Cell* 184 (11 2021), pp. 2825–2842.
- [22] Shivdasani R, Mayer E, and Orkin S. “Absence of Blood Formation in Mice Lacking the T-Cell Leukaemia Oncoprotein Tal-1/SCL”. In: *Nature* 373 (6513 1995), pp. 432–434.

- [23] F Ann Ran et al. “Genome engineering using the CRISPR-Cas9 system”. en. In: *Nat. Protoc.* 8.11 (Nov. 2013), pp. 2281–2308.
- [24] Büttner M et al. “ScCODA Is a Bayesian Model for Compositional Single-Cell Data Analysis”. In: *Nature Communications* 12 (1 2021), p. 6876.
- [25] Andrew Cliff and Keith Ord. “Testing for spatial autocorrelation among regression residuals”. en. In: *Geogr. Anal.* 4.3 (July 1972), pp. 215–332.
- [26] Detomaso David and Yosef Nir. “Hotspot Identifies Informative Gene Modules across Modalities of Single-Cell Genomics”. In: *Cell Systems* 12 (5 2021), pp. 446–456.
- [27] Korsunsky Ilya et al. “Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony”. In: *Nature Methods* 16 (12 2019), pp. 1289–1296.
- [28] Dann Emma et al. “Differential Abundance Testing on Single-Cell Data Using k-Nearest Neighbor Graphs”. In: *Nature Biotechnology* (Sept. 2021). DOI: 10.1038/s41587-021-01033-z.
- [29] Tasic Bosiljka et al. “Shared and Distinct Transcriptomic Cell Types across Neocortical Areas”. In: *Nature* 563 (7729 2018), pp. 72–78.
- [30] Li Yang et al. “An Atlas of Gene Regulatory Elements in Adult Mouse Cerebrum”. In: *Nature* 598 (7879 2021), pp. 129–136.
- [31] Hannah A Pliner et al. “Cicero Predicts cis-Regulatory DNA Interactions from Single-Cell Chromatin Accessibility Data”. en. In: *Mol. Cell* 71.5 (Sept. 2018), 858–871.e8.
- [32] Hao et al. “Single-Cell Transcriptomics of Adult Macaque Hippocampus Reveals Neural Precursor Cell Populations”. In: *Nature Neuroscience* 25 (6 2022), pp. 805–817.
- [33] Stuart Tim et al. “Comprehensive Integration of Single-Cell Data”. In: *Cell* 177 (7 2019), pp. 1888–1902.
- [34] Demetci Pinar et al. “SCOT: Single-Cell Multi-Omics Alignment with Optimal Transport”. In: *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 29 (1 2022), pp. 3–18.

- [35] Cao Zhi-Jie and Gao Ge. “Multi-Omics Single-Cell Data Integration and Regulatory Inference with Graph-Linked Embedding”. In: *Nature Biotechnology* 40 (10 2022), pp. 1458–1466.
- [36] Sakaue Saori et al. “Tissue-Specific Enhancer-Gene Maps from Multimodal Single-Cell Data Identify Causal Disease Alleles”. In: *BioRxiv* (2022). DOI: 10.1101/2022.10.27.22281574.
- [37] Qiu Chengxiang et al. “Systematic Reconstruction of Cellular Trajectories across Mouse Embryogenesis”. In: *Nature Genetics* 54 (3 2022), pp. 328–341.
- [38] Manno La et al. “Molecular Architecture of the Developing Mouse Brain”. In: *Nature* 596 (2021), pp. 92–96.
- [39] Karaiskos Nikos et al. “The Drosophila Embryo at Single-Cell Transcriptome Resolution”. In: *Science* 358 (6360 2017), pp. 194–199.
- [40] Rivera John et al. “REDfly: The Transcriptional Regulatory Element Database for Drosophila”. In: *Nucleic Acids Research* 47 (D1 2019), pp. D828–834.
- [41] Kvon Evgeny et al. “Genome-Scale Functional Characterization of Drosophila Developmental Enhancers in Vivo”. In: *Nature* 512 (7512 2014), pp. 91–95.
- [42] Bonn et al. “Tissue-Specific Analysis of Chromatin State Identifies Temporal Signatures of Enhancer Activity during Embryonic Development”. In: *Nature Genetics* 44 (2 2012), pp. 148–156.
- [43] Reddington James et al. “Lineage-Resolved Enhancer and Promoter Usage during a Time Course of Embryogenesis”. In: *Developmental Cell* 55 (5 2020), pp. 648–664.
- [44] Zinzen Robert et al. “Combinatorial Binding Predicts Spatio-Temporal Cis-Regulatory Activity”. In: *Nature* 462 (7269 2009), pp. 65–70.
- [45] Kudron Michelle et al. “The ModERN Resource: Genome-Wide Binding Profiles for Hundreds of Drosophila and Caenorhabditis Elegans Transcription Factors”. In: *Genetics* 208 (3 2018), pp. 937–949.

- [46] Howe Kevin et al. “Ensembl 2021”. In: *Nucleic Acids Research* 49 (D1 2021), pp. D884–D891.
- [47] Graveley Brenton et al. “The Developmental Transcriptome of *Drosophila Melanogaster*”. In: *Nature* 471 (7339 2010), pp. 473–479.
- [48] Kwasnieski Jamie et al. “Early Genome Activation in Is Extensive with an Initial Tendency for Aborted Transcripts and Retained Introns”. In: *Genome Research* 29 (7 2019), pp. 1188–1197.
- [49] Blythe Shelby and Wieschaus Eric. “Establishment and Maintenance of Heritable Chromatin Structure during Early *Drosophila* Embryogenesis”. In: *Cell* 160 (6 2015), p. 5. DOI: 10.7554/eLife.20148.
- [50] Trapnell Cole et al. “The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells”. In: *Nature Biotechnology* 32 (4 2014), pp. 381–386.
- [51] Hammonds Ann et al. “Spatial Expression of Transcription Factors in *Drosophila* Embryonic Organ Development”. In: *Genome Biology* 14 (12 2013), R140.
- [52] Tomancak Pavel et al. “Global Analysis of Patterns of Gene Expression during *Drosophila* Embryogenesis”. In: *Genome Biology* 8 (7 2007), R145.
- [53] Laver John et al. “Chapter Two -Regulation and Function of Maternal Gene Products During the Maternal-to-Zygotic Transition in *Drosophila*”. In: *In Current Topics in Developmental Biology* 113 (2015), pp. 43–84.
- [54] Song Wan, Ressler Susanne, and Tracey W. Daniel. “Loss of Pseudouridine Synthases in the RluA Family Causes Hypersensitive Nociception in *Drosophila*”. In: *G3* 10 (12 2020), pp. 4425–4438.
- [55] Moore Adrian and Yeh Lily. “Hamlet, a Binary Genetic Switch between Single-and Multiple-Dendrite Neuron Morphology”. In: *Science* 297 (5585 Jan. 2002), pp. 1355–1358.

- [56] Dubruille Raphaelle et al. “Drosophila Regulatory Factor X Is Necessary for Ciliated Sensory Neuron Differentiation”. In: *Development* 129 (23 2002), pp. 5487–5498.
- [57] Walker R, Willingham A, and Zuker C. “A Drosophila Mechanosensory Transduction Channel”. In: *Science* 287 (5461 2000), pp. 2229–2234.
- [58] Haines Jenna and Eisen Michael. “Patterns of Chromatin Accessibility along the Anterior-Posterior Axis in the Early Drosophila Embryo”. In: *PLoS Genetics* 14 (5 2018), e1007367.
- [59] Wang Mingyue et al. “High-Resolution 3D Spatiotemporal Transcriptomic Maps of Developing Drosophila Embryos and Larvae”. In: *Developmental Cell* 158 (6 2014), pp. 1431–1443. DOI: 10.1016/j.devcel.2022.04.006.Weirauch.
- [60] Junion Guillaume et al. “A Transcription Factor Collective Defines Cardiac Cell Fate and Reflects Lineage History”. In: *Cell* 148 (3 2012), pp. 473–486.
- [61] Secchia Stefano et al. “Simultaneous Cellular and Molecular Phenotyping of Embryonic Mutants Using Single-Cell Regulatory Trajectories”. In: *Developmental Cell* 57 (4 2022), pp. 496–511.
- [62] Frommer G et al. “Epidermal Egr-like Zinc Finger Protein of Drosophila Participates in Myotube Guidance”. In: *The EMBO Journal* 15 (7 1996), pp. 1642–1649.
- [63] Fox Rebecca et al. “Organ-Specific Gene Expression: The bHLH Protein Sage Provides Tissue Specificity to Drosophila FoxA”. In: *Development* 140 (10 2013), pp. 2160–2171.
- [64] Reuter R. “The Gene Serpent Has Homeotic Properties and Specifies Endoderm versus Ectoderm within the Drosophila Gut”. In: *Development* 120 (5 1994), pp. 1123–1135.
- [65] Curtiss J and Heilig J. “Arrowhead Encodes a LIM Homeodomain Protein That Distinguishes Subsets of Drosophila Imaginal Cells”. In: *Developmental Biology* 190 (1 1997), pp. 129–141.

- [66] Preger-Ben Noon et al. “Evolved Repression Overcomes Enhancer Robustness”. In: *Developmental Cell* 39 (5 2016), pp. 572–584.
- [67] Myat M and Andrew D. “Fork Head Prevents Apoptosis and Promotes Cell Shape Change during Formation of the Drosophila Salivary Glands”. In: *Development* 127 (2000), pp. 4217–4226.
- [68] Kartha Vinay et al. “Functional Inference of Gene Regulation Using Single-Cell Multi-Omics”. In: (2021). DOI: 10.1101/2021.07.28.453784.
- [69] Harrison Melissa et al. “Zelda Binding in the Early Drosophila Melanogaster Embryo Marks Regions Subsequently Activated at the Maternal-to-Zygotic Transition”. In: *PLoS Genetics* 7 (10 2011), e1002266.
- [70] Özel Mehmet et al. “Neuronal Diversity and Convergence in a Visual System Developmental Atlas”. In: *Nature* 589 (7840 2021), pp. 88–95.
- [71] Bageritz Josephine et al. “Gene Expression Atlas of a Developing Tissue by Single Cell Expression Correlation Analysis”. In: *Nature Methods* 16 (8 2019), pp. 750–756.
- [72] Mclaughlin Colleen et al. “Single-Cell Transcriptomes of Developing and Adult Olfactory Receptor Neurons in Drosophila”. In: (Feb. 2021), p. 10. DOI: 10.7554/eLife.63856.
- [73] Shi Zhen et al. “Single-Cyst Transcriptome Analysis of Drosophila Male Germline Stem Cell Lineage”. In: *Development* 147 (8 2020). DOI: 10.1242/dev.184259.
- [74] Everetts Nicholas et al. “Single-Cell Transcriptomics of the Drosophila Wing Disc Reveals Instructive Epithelium-to-Myoblast Interactions”. In: (Mar. 2021), p. 10. DOI: 10.7554/eLife.61276.
- [75] Hung et al. “A Cell Atlas of the Adult Drosophila Midgut”. In: *Proceedings of the National Academy of Sciences of the United States of America* 117 (3 2020), pp. 1514–1523.

- [76] Li Hongjie et al. “Fly Cell Atlas: A Single-Nucleus Transcriptomic Atlas of the Adult Fruit Fly”. In: *Science* 375 (6584 2022), p. 2432.
- [77] Sandmann et al. “ChIP-on-Chip Protocol for Genome-Wide Analysis of Transcription Factor Binding in *Drosophila Melanogaster* Embryos”. In: *Nature Protocols* 1 (6 2006), pp. 2839–2855.
- [78] Hartenstein Volker. “Atlas of *Drosophila* Development”. In: 328 (1993).
- [79] Schor Ignacio et al. “Non-Coding RNA Expression, Function, and Variation during *Drosophila* Embryogenesis”. In: *Current Biology: CB* 28 (22 2018), pp. 3547–3561.
- [80] Schindelin Johannes et al. “Fiji: An Open-Source Platform for Biological-Image Analysis”. In: *Nature Methods* 9 (7 2012), pp. 676–682.
- [81] Christopher S McGinnis, Lyndsay M Murrow, and Zev J Gartner. “DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors”. en. In: *Cell Syst* 8.4 (Apr. 2019), 329–337.e4.
- [82] Davis Carrie et al. “The Encyclopedia of DNA Elements (ENCODE): Data Portal Update”. In: *Nucleic Acids Research* 46 (D1 2018), pp. D794–801.
- [83] Chen Edward et al. “Enrichr: Interactive and Collaborative HTML5 Gene List Enrichment Analysis Tool”. In: *BMC Bioinformatics* 14 (Apr. 2013), p. 128.
- [84] Kuleshov Maxim et al. “modEnrichr: A Suite of Gene Set Enrichment Analysis Tools for Model Organisms”. In: *Nucleic Acids Research* 47 (W1), W183–W190.
- [85] Bergen et al. “Generalizing RNA Velocity to Transient Cell States through Dynamical Modeling”. In: *Nature Biotechnology* 38 (12 2020), pp. 1408–1414.
- [86] Corces M et al. “An Improved ATAC-Seq Protocol Reduces Background and Enables Interrogation of Frozen Tissues”. In: *Nature Methods* 14 (10 2017), pp. 959–962.
- [87] Zhang Yong et al. “Model-Based Analysis of ChIP-Seq (MACS)”. In: *Genome Biology* 9 (9 2008), R137.

- [88] Quinlan Aaron and Hall Ira. “BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features”. In: *Bioinformatics* 26 (6 2010), pp. 841–842.
- [89] Wolock Samuel, Lopez Romain, and Klein Allon. “Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data”. In: *Cell Systems* 8 (4 2019), pp. 281–291.
- [90] Stuart Tim et al. “Multimodal Single-Cell Chromatin Analysis with Signac”. In: (2020). DOI: 10.1101/2020.11.09.373613.
- [91] Schep A et al. “chromVAR: Inferring Transcription-Factor-Associated Accessibility from Single-Cell Epigenomic Data”. In: *Nature Methods* 14 (10 2017), pp. 975–978.
- [92] Heinz Sven et al. “Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities”. In: *Molecular Cell* 38 (4 2010), pp. 576–589.
- [93] Friedman Jerome, Hastie Trevor, and Tibshirani Rob. “Regularization Paths for Generalized Linear Models via Coordinate Descent”. In: *Journal of Statistical Software* 33 (1 2010), pp. 1–22.
- [94] Benaglia Tatiana et al. “Mixtools: An R Package for Analyzing Mixture Models”. In: *Journal of Statistical Software* 32 (2010), pp. 1–29.
- [95] Qiu Chengxiang et al. “Systematic Reconstruction of the Cellular Trajectories of Mammalian Embryogenesis”. In: (2021). DOI: 10.1101/2021.06.08.447626.
- [96] Demir Ebru and Dickson Barry. “Fruitless Splicing Specifies Male Courtship Behavior in *Drosophila*”. In: *Cell* 121 (5 2005), pp. 785–794.
- [97] Manoli Devanand et al. “Male-Specific Fruitless Specifies the Neural Substrates of *Drosophila* Courtship Behaviour”. In: *Nature* 436 (7049 2005), pp. 395–400.
- [98] Kimura Ken-Ichi et al. “Fruitless Specifies Sexually Dimorphic Neural Circuitry in the *Drosophila* Brain”. In: *Nature* 438 (7065 2005), pp. 229–233.

- [99] Brovkina Margarita et al. “Fruitless Decommissions Regulatory Elements to Implement Cell-Type-Specific Neuronal Masculinization”. In: *PLoS Genetics* 17 (2 2021), e1009338.
- [100] Sato Kosei and Yamamoto Daisuke. “The Mode of Action of Fruitless: Is It an Easy Matter to Switch the Sex?” In: *Genes, Brain, and Behavior* 19 (2 2020), e12606.
- [101] Ito Hiroki et al. “Fruitless Represses robo1 Transcription to Shape Male-Specific Neural Morphology and Behavior in *Drosophila*”. In: *Current Biology: CB* 26 (12 2016), pp. 1532–1542.
- [102] Rosenthal Nadia and Brown Steve. “The Mouse Ascending: Perspectives for Human-Disease Models”. In: *Nature Cell Biology* 9 (9 2007), pp. 993–999.
- [103] Gurumurthy Channabasavaiah and Lloyd Kevin. “Generating Mouse Models for Biomedical Research: Technological Advances”. In: *Disease Models & Mechanisms* 12 (1 2019). DOI: 10.1242/dmm.029462.
- [104] Steensma David et al. “Abbie Lathrop, the ‘Mouse Woman of Granby’: Rodent Fancier and Accidental Genetics Pioneer”. In: *Mayo Clinic Proceedings. Mayo Clinic* 85 (11 2010), e83.
- [105] Silver Molecular Biology Lee M, Dr Professor, and Silver Lee. In: *Mouse Genetics: Concepts and Applications* (1995).
- [106] Ringwald Martin et al. “The IKMC Web Portal: A Central Point of Entry to Data and Resources from the International Knockout Mouse Consortium”. In: *Nucleic Acids Research* 39 (Database 2011), pp. D849–D855.
- [107] Jinek Martin et al. “A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity”. In: *Science* 337 (6096 2012), pp. 816–821.
- [108] Singh et al. “A Mouse Geneticist’s Practical Guide to CRISPR Applications”. In: *Genetics* 199 (1 2015), pp. 1–15.

- [109] Lupiáñez Darío et al. “Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions”. In: *Cell* 161 (5 2015), pp. 1012–1025.
- [110] Spielmann et al. “Structural Variation in the 3D Genome”. In: *Nature Reviews Genetics* (2018).
- [111] Otto F et al. “Cbfa1, a Candidate Gene for Cleidocranial Dysplasia Syndrome, Is Essential for Osteoblast Differentiation and Bone Development”. In: *Cell* 89 (5 1997), pp. 765–771.
- [112] Scambler Peter. “Faculty Opinions Recommendation of Mechanism for the Learning Deficits in a Mouse Model of Neurofibromatosis Type 1”. In: (2002). DOI: 10.3410/f.1003995.44355.
- [113] Gailus-Durner et al. “Introducing the German Mouse Clinic: Open Access Platform for Standardized Phenotyping”. In: *Nature Methods* 2 (6 2005), pp. 403–404.
- [114] Dickel Diane et al. “Ultraconserved Enhancers Are Required for Normal Development”. In: *Cell* 172 (3 2018), pp. 491–499.
- [115] Flöttmann Ricarda et al. “Noncoding Copy-Number Variations Are Associated with Congenital Limb Malformation”. In: *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 20 (6 2018), pp. 599–607.
- [116] Osterwalder Marco et al. “Enhancer Redundancy Provides Phenotypic Robustness in Mammalian Development”. In: *Nature* 554 (7691 2018), pp. 239–243.
- [117] Cheng Shangli et al. “Single-Cell RNA-Seq Reveals Cellular Heterogeneity of Pluripotency Transition and X Chromosome Dynamics during Early Mouse Development”. In: *Cell Reports* (2019). DOI: 10.1016/j.celrep.2019.02.031.
- [118] Mohammed Hisham et al. “Single-Cell Landscape of Transcriptional Heterogeneity and Cell Fate Decisions during Mouse Early Gastrulation”. In: *Cell Reports* 20 (5 2017), pp. 1215–1228.

- [119] Zhou Yingyue et al. “Human and Mouse Single-Nucleus Transcriptomics Reveal TREM2-Dependent and TREM2-Independent Cellular Responses in Alzheimer’s Disease”. In: *Nature Medicine* 26 (1 2020), pp. 131–142.
- [120] Soysa T et al. “Single-Cell Analysis of Cardiogenesis Reveals Basis for Organ-Level Developmental Defects”. In: *Nature* 572 (7767 2019), pp. 120–124.
- [121] Zhong Jixing et al. “Single-Cell Brain Atlas of Parkinson’s Disease Mouse Model”. In: *Journal of Genetics and Genomics = Yi Chuan Xue Bao* 48 (4 2021), pp. 277–288.
- [122] Li Xiuying et al. “Single Cell RNA Sequencing Identifies IGFBP5 and QKI as Ciliated Epithelial Cell Genes Associated with Severe COPD”. In: *Respiratory Research* 22 (1 2021), p. 100.
- [123] Despang Alexandra et al. “Functional Dissection of the Sox9-Kcnj2 Locus Identifies Nonessential and Instructive Roles of TAD Architecture”. In: *Nature Genetics* 51 (8 2019), pp. 1263–1271.
- [124] Stottmann R et al. “Ttc21b Is Required to Restrict Sonic Hedgehog Activity in the Developing Mouse Forebrain”. In: *Developmental Biology* 335 (1 2009), pp. 166–178.
- [125] Yadav Neelu et al. “Specific Protein Methylation Defects and Gene Expression Perturbations in Coactivator-Associated Arginine Methyltransferase 1-Deficient Mice”. In: *Proceedings of the National Academy of Sciences of the United States of America* 100 (11 2003), pp. 6464–6468.
- [126] Mo R et al. “Specific and Redundant Functions of Gli2 and Gli3 Zinc Finger Genes in Skeletal Patterning and Development”. In: *Development* 124 (1 1997), pp. 113–123.
- [127] Leipold Enrico et al. “A de Novo Gain-of-Function Mutation in SCN11A Causes Loss of Pain Perception”. In: *Nature Genetics* (2013). DOI: 10.1038/ng.2767.

- [128] Schwabe Georg et al. “Ror2 Knockout Mouse as a Model for the Developmental Pathology of Autosomal Recessive Robinow Syndrome”. In: *Developmental Dynamics: An Official Publication of the American Association of Anatomists* 229 (2 2004), pp. 400–410.
- [129] Chan Wing 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”. In: *PLOS Genetics* (2018). DOI: 10.1371/journal.pgen.1007242.
- [130] Fischer Björn et al. “Further Characterization of ATP6V0A2-Related Autosomal Recessive Cutis Laxa”. In: *Human Genetics* 131 (11 2012), pp. 1761–1773.
- [131] Ringel Alessa et al. “Promoter Repression and 3D-Restructuring Resolves Divergent Developmental Gene Expression in TADs”. In: *SSRN Electronic Journal* (). DOI: 10.2139/ssrn.3947354.
- [132] Rajderkar Sudha et al. “Topologically Associating Domain Boundaries Are Commonly Required for Normal Genome Function”. In: (). DOI: 10.1101/2021.05.06.443037.
- [133] Kvon Evgeny et al. “Progressive Loss of Function in a Limb Enhancer during Snake Evolution”. In: *Cell* 167 (3 2016), pp. 633–642. DOI: org/10.1016/j.cell.2016.09.028.---.2016b.
- [134] Jacob J and Briscoe J. “Gli Proteins and the Control of Spinal-Cord Patterning”. In: *EMBO Reports* 4 (8 2003). DOI: 10.1038/sj.embor.embor896.
- [135] Nolte Mark et al. “Functional Analysis of Limb Transcriptional Enhancers in the Mouse”. In: *Evolution & Development* 16 (4 1978), pp. 207–223. DOI: 10.1016/b978-0-12-507850-4.x5001-x.
- [136] Jo Alice et al. “The Versatile Functions of Sox9 in Development, Stem Cells, and Human Diseases”. In: *Genes & Diseases* 1 (2 2014), pp. 149–161.

- [137] Gordon C et al. “Long-Range Regulation at the SOX9 Locus in Development and Disease”. In: *Journal of Medical Genetics* 46 (10 2009), pp. 649–656.
- [138] Tran Pamela et al. “THM1 Negatively Modulates Mouse Sonic Hedgehog Signal Transduction and Affects Retrograde Intraflagellar Transport in Cilia”. In: *Nature Genetics* 40 (4 2008), pp. 403–410.
- [139] Davis Erica et al. “TTC21B Contributes Both Causal and Modifying Alleles across the Ciliopathy Spectrum”. In: *Nature Genetics* 43 (3 2011), pp. 189–196.
- [140] Sheth Rushikesh et al. “Distal Limb Patterning Requires Modulation of Cis-Regulatory Activities by HOX13”. In: *Cell Reports* 17 (11 2016), pp. 2913–2926.
- [141] Capdevila J et al. “Control of Vertebrate Limb Outgrowth by the Proximal Factor Meis2 and Distal Antagonism of BMPs by Gremlin”. In: *Molecular Cell* 4 (5 1999), pp. 839–849.
- [142] Mercader Nadia et al. “Ectopic Meis1 Expression in the Mouse Limb Bud Alters P-D Patterning in a Pbx1-Independent Manner”. In: *The International Journal of Developmental Biology* 53 (8-10 2009), pp. 1483–1494.
- [143] Reginelli A et al. “Digit Tip Regeneration Correlates with Regions of Msx1 (Hox 7) Expression in Fetal and Newborn Mice”. In: *Development* 121 (4 1995), pp. 1065–1076.
- [144] Tzchori et al. “LIM Homeobox Transcription Factors Integrate Signaling Events That Control Three-Dimensional Limb Patterning and Growth”. In: *Development* 136 (8 2009), pp. 1375–1385.
- [145] Matise M et al. “Gli2 Is Required for Induction of Floor Plate and Adjacent Cells, but Not Most Ventral Neurons in the Mouse Central Nervous System”. In: *Development* 125 (15 1998), pp. 2759–2770.

- [146] Broom Emma et al. “The Roof Plate Boundary Is a Bi-Directional Organiser of Dorsal Neural Tube and Choroid Plexus Development”. In: *Development* 139 (22 2012), pp. 4261–4270.
- [147] Revinski Diego et al. “CDC20B Is Required for Deuterosome-Mediated Centriole Production in Multiciliated Cells”. In: *Nature Communications* 9 (1 2018), p. 4668.
- [148] Terré Berta et al. “Defects in Efferent Duct Multiciliogenesis Underlie Male Infertility in GEMC1-, MCIDAS-or CCNO-Deficient Mice”. In: *Development* 146 (8 2019). DOI: 10.1242/dev.162628.
- [149] Li You et al. “DNAH6 and Its Interactions with PCD Genes in Heterotaxy and Primary Ciliary Dyskinesia”. In: *PLoS Genetics* 12 (2 2016), e1005821.
- [150] Coutton Charles et al. “Mutations in CFAP43 and CFAP44 Cause Male Infertility and Flagellum Defects in Trypanosoma and Human”. In: *Nature Communications* 9 (1 2018), p. 686.
- [151] Ichijo Ryo et al. “Essential Roles of Tbx3 in Embryonic Skin Development during Epidermal Stratification”. In: *Genes to Cells: Devoted to Molecular & Cellular Mechanisms* 22 (3 2017), pp. 284–292.
- [152] Khan Saif et al. “The Roles and Regulation of TBX3 in Development and Disease”. In: *Gene* 726 (Feb. 2020), p. 144223.
- [153] Zhang Zhen et al. “Transcription Factor Etv5 Is Essential for the Maintenance of Alveolar Type II Cells”. In: *Proceedings of the National Academy of Sciences of the United States of America* 114 (15 2017), pp. 3903–3908.
- [154] Herriges Michael and Morrisey Edward. “Lung Development: Orchestrating the Generation and Regeneration of a Complex Organ”. In: *Development* 141 (3 2014), pp. 502–513.
- [155] Paaby Annalise and Rockman Matthew. “The Many Faces of Pleiotropy”. In: *Trends in Genetics: TIG* 29 (2 2013), pp. 66–73.

- [156] Zeidler Maximilian et al. “NOCICEPTRA: Gene and microRNA Signatures and Their Trajectories Characterizing Human iPSC-Derived Nociceptor Maturation”. In: *Advancement of Science* 8 (21 2021), e2102354.
- [157] Timmer John et al. “Tissue Morphogenesis and Vascular Stability Require the Frem2 Protein, Product of the Mouse Myelencephalic Blebs Gene”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102 (33 2005), pp. 11746–11750.
- [158] Larrucea Susana et al. “Expression of Podocalyxin Enhances the Adherence, Migration, and Intercellular Communication of Cells”. In: *Experimental Cell Research* 314 (10 2008), pp. 2004–2015.
- [159] Ponnusamy Moorthy et al. “Emerging Role of Mucins in Epithelial to Mesenchymal Transition”. In: *Current Cancer Drug Targets* 13 (9 2013), pp. 945–956.
- [160] Weins A et al. “Differentiation and Stress-Dependent Nuclear Cytoplasmic Redistribution of Myopodin, a Novel Actin-Bundling Protein”. In: *The Journal of Cell Biology* 155 (3 2001), pp. 393–404.
- [161] Kwartler Callie et al. “Overexpression of Smooth Muscle Myosin Heavy Chain Leads to Activation of the Unfolded Protein Response and Autophagic Turnover of Thick Filament-Associated Proteins in Vascular Smooth Muscle Cells”. In: *The Journal of Biological Chemistry* 289 (1920), pp. 14075–14088.
- [162] Wang Zhigao et al. “Myocardin Is a Master Regulator of Smooth Muscle Gene Expression”. In: *Proceedings of the National Academy of Sciences of the United States of America* 100 (12 2003), pp. 7129–7134.
- [163] Gharibi Armen et al. “ITGA1 Is a Pre-Malignant Biomarker That Promotes Therapy Resistance and Metastatic Potential in Pancreatic Cancer”. In: *Scientific Reports* 7 (1 2017), p. 10060.

- [164] Fanjul-Fernández Miriam et al. “Cell-cell Adhesion Genes CTNNA2 and CTNNA3 Are Tumour Suppressors Frequently Mutated in Laryngeal Carcinomas”. In: *Nature Communications* (2013). DOI: 10.1038/ncomms3531.
- [165] Mead Timothy et al. “A Far-Upstream (-70 Kb) Enhancer Mediates Sox9 Auto-Regulation in Somatic Tissues during Development and Adult Regeneration”. In: *Nucleic Acids Research* 41 (8 2013), pp. 4459–4469.
- [166] Long Hannah et al. “Loss of Extreme Long-Range Enhancers in Human Neural Crest Drives a Craniofacial Disorder”. In: *Cell Stem Cell* 27 (5 2020), pp. 765–783.
- [167] Yao Baojin et al. “The SOX9 Upstream Region Prone to Chromosomal Aberrations Causing Campomelic Dysplasia Contains Multiple Cartilage Enhancers”. In: *Nucleic Acids Research* 43 (11 2015), pp. 5394–5408.
- [168] Nagakura Ryotaro et al. “Switching of Sox9 Expression during Musculoskeletal System Development”. In: *Scientific Reports* 10 (1 2020), p. 8425.
- [169] Scott Charlotte et al. “SOX9 Induces and Maintains Neural Stem Cells”. In: *Nature Neuroscience* 13 (10 2010), pp. 1181–1189.
- [170] Kawaguchi Yoshiya. “Sox9 and Programming of Liver and Pancreatic Progenitors”. In: *The Journal of Clinical Investigation* 123 (5 2013), pp. 1881–1886.
- [171] Kumar Sanjeev et al. “Sox9 Activation Highlights a Cellular Pathway of Renal Repair in the Acutely Injured Mammalian Kidney”. In: *Cell Reports* 12 (8 2015), pp. 1325–1338.
- [172] Tanimizu Naoki et al. “Progressive Induction of Hepatocyte Progenitor Cells in Chronically Injured Liver”. In: *Scientific Reports* 7 (Jan. 2017), p. 39990.
- [173] Rockich Briana et al. “Sox9 Plays Multiple Roles in the Lung Epithelium during Branching Morphogenesis”. In: *Proceedings of the National Academy of Sciences of the United States of America* 110 (47 2013), E4456–E4464.

- [174] Wagner T et al. “Autosomal Sex Reversal and Campomelic Dysplasia Are Caused by Mutations in and around the SRY-Related Gene SOX9”. In: *Cell* 79 (6 1994), pp. 1111–1120.
- [175] Bi W et al. “Haploinsufficiency of Sox9 Results in Defective Cartilage Primordia and Premature Skeletal Mineralization”. In: *Proceedings of the National Academy of Sciences of the United States of America* 98 (12 2001), pp. 6698–6703.
- [176] Borcharding Nicholas et al. “Mapping the Immune Environment in Clear Cell Renal Carcinoma by Single-Cell Genomics”. In: *Communications Biology* 4 (1 2021), p. 122.
- [177] Hernández Rosa et al. “Differentiation of Human Mesenchymal Stem Cells towards Neuronal Lineage: Clinical Trials in Nervous System Disorders”. In: *Biomolecules & Therapeutics* 28 (1 2020), pp. 34–44.
- [178] Dickinson Mary et al. “High-Throughput Discovery of Novel Developmental Phenotypes”. In: *Nature* 537 (7621 2016), pp. 508–514.
- [179] Brown Steve. “Advances in Mouse Genetics for the Study of Human Disease”. In: *Human Molecular Genetics* (2021). DOI: 10.1093/hmg/ddab153.
- [180] Kraft Katerina et al. “Deletions, Inversions, Duplications: Engineering of Structural Variants Using CRISPR/Cas in Mice”. In: *Cell Reports* 10 (5 2015), pp. 833–839.
- [181] Cao Junyue. “Sci-RNA-seq3 v1”. In: *Protocols.io. ZappyLab, Inc* (2019). DOI: 10.17504/protocols.io.9yih7ue.
- [182] Renaud Gabriel et al. “deML: Robust Demultiplexing of Illumina Sequences Using a Likelihood-Based Approach”. In: *Bioinformatics* 31 (5 2015), pp. 770–772.
- [183] Dobin Alexander et al. “STAR: Ultrafast Universal RNA-Seq Aligner”. In: *Bioinformatics* 29 (1 2013), pp. 15–21.
- [184] Anders Simon, Pyl Paul, and Huber Wolfgang. “HTSeq—a Python Framework to Work with High-Throughput Sequencing Data”. In: *Bioinformatics* 31 (2 2015), pp. 166–169.

- [185] Wang Fay et al. “RNAscope: A Novel in Situ RNA Analysis Platform for Formalin-Fixed, Paraffin-Embedded Tissues”. In: *The Journal of Molecular Diagnostics: JMD* 14 (1 2012), pp. 22–29.
- [186] Subramanian Aravind et al. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles”. In: *Proceedings of the National Academy of Sciences of the United States of America* 102 (43 2005), pp. 15545–15550.
- [187] Liberzon Arthur et al. “The Molecular Signatures Database (MSigDB) Hallmark Gene Set Collection”. In: *Cell Systems* 1 (6 2015), pp. 417–425.
- [188] Dunbar R and Shultz Susanne. “Evolution in the Social Brain”. In: *Science* 317 (5843 2007), pp. 1344–1347.
- [189] Navarrete Ana, Van Schaik Carel, and Isler Karin. “Energetics and the Evolution of Human Brain Size”. In: *Nature* 480 (7375 2011), pp. 91–93.
- [190] Darwin Charles. “The Descent of Man, and Selection in Relation to Sex”. In: (1871).
- [191] Herculano-Houzel et al. “Cellular Scaling Rules for Primate Brains”. In: *Proceedings of the National Academy of Sciences of the United States of America* 104 (9 2007), pp. 3562–3567.
- [192] Grasby Katrina et al. “The Genetic Architecture of the Human Cerebral Cortex”. In: *Science* 367 (6484 2020), p. 6690.
- [193] Zeng Jia et al. “Divergent Whole-Genome Methylation Maps of Human and Chimpanzee Brains Reveal Epigenetic Basis of Human Regulatory Evolution”. In: *American Journal of Human Genetics* 91 (3 2012), pp. 455–465.
- [194] Preuss Todd. “Human Brain Evolution: From Gene Discovery to Phenotype Discovery”. In: *Proceedings of the National Academy of Sciences of the United States of America* 109 (1 June 2012), pp. 10709–10716.

- [195] Cajal S. “The Croonian Lecture: La Fine Structure Des Centres Nerveux”. In: *Proceedings of the Royal Society of London* 55 (1894), pp. 444–468.
- [196] Striedter Georg. “Principles of Brain Evolution”. In: (2005).
- [197] Allman John et al. “Intuition and Autism: A Possible Role for Von Economo Neurons”. In: *Trends in Cognitive Sciences* 9 (8 2005), pp. 367–373.
- [198] Rizzolatti Giacomo, Fabbri-Destro Maddalena, and Cattaneo Luigi. “Mirror Neurons and Their Clinical Relevance”. In: *Nature Clinical Practice. Neurology* 5 (1 2009), pp. 24–34.
- [199] Boldog et al. “Transcriptomic and Morphophysiological Evidence for a Specialized Human Cortical GABAergic Cell Type”. In: *Nature Neuroscience* 21 (9 2018), pp. 1185–1195.
- [200] Nowakowski Tomasz et al. “Spatiotemporal Gene Expression Trajectories Reveal Developmental Hierarchies of the Human Cortex”. In: *Science* 358 (6368 2017), pp. 1318–1323.
- [201] Hodge Rebecca et al. “Conserved Cell Types with Divergent Features in Human versus Mouse Cortex”. In: *Nature* 573 (7772 2019), pp. 61–68.
- [202] Bakken Trygve et al. “Comparative Cellular Analysis of Motor Cortex in Human, Marmoset and Mouse”. In: *Nature* 598 (7879 2021), pp. 111–119.
- [203] Ecker Joseph et al. In: *The BRAIN Initiative Cell Census Consortium: Lessons Learned toward Generating a Comprehensive Brain Cell Atlas* 96 (2017), pp. 542–557.
- [204] Gibbs Richard et al. “Evolutionary and biomedical insights from the rhesus macaque genome”. In: *Science* 316 (5822 2007), pp. 222–234.
- [205] Martin Beth et al. “An Optimized Protocol for Single Cell Transcriptional Profiling by Combinatorial Indexing”. In: (2021).

- [206] Kamath Tushar et al. “Single-Cell Genomic Profiling of Human Dopamine Neurons Identifies a Population That Selectively Degenerates in Parkinson’s Disease”. In: *Nature Neuroscience* 25 (5 2022), pp. 588–595.
- [207] Ren Jing et al. “Single-Cell Transcriptomes and Whole-Brain Projections of Serotonin Neurons in the Mouse Dorsal and Median Raphe Nuclei”. In: 8 (Oct. 2019), e49424.
- [208] Poewe Werner et al. “Parkinson Disease”. In: *Nature Reviews. Disease Primers* 3 (Mar. 2017), p. 17013.
- [209] Fakhoury Marc. “Revisiting the Serotonin Hypothesis: Implications for Major Depressive Disorders”. In: *Molecular Neurobiology* 53 (5 2016), pp. 2778–2786.
- [210] Yang Andrew et al. “A Human Brain Vascular Atlas Reveals Diverse Mediators of Alzheimer’s Risk”. In: *Nature* 603 (7903 2022), pp. 885–892.
- [211] Cusanovich Darren et al. “A Single-Cell Atlas of in Vivo Mammalian Chromatin Accessibility”. In: *Cell* 174 (5 2018), pp. 1309–1324.
- [212] Berg Jim et al. “Human Neocortical Expansion Involves Glutamatergic Neuron Diversification”. In: *Nature* 598 (7879 2021), pp. 151–158.
- [213] Trevino Alexandro et al. “Chromatin and Gene-Regulatory Dynamics of the Developing Human Cerebral Cortex at Single-Cell Resolution”. In: *Cell* 184 (2019), pp. 5053–5069.
- [214] Krienen Fenna et al. “Innovations Present in the Primate Interneuron Repertoire”. In: *Nature* 586 (7828 2020), pp. 262–269.
- [215] Barres Ben. “The Mystery and Magic of Glia: A Perspective on Their Roles in Health and Disease”. In: *Neuron* 60 (3 2008), pp. 430–440.
- [216] Valori Chiara et al. “Astrocytes: Emerging Therapeutic Targets in Neurological Disorders”. In: *Trends in Molecular Medicine* 25 (9 2019), pp. 750–759.

- [217] Wiebe Shane et al. “Inhibitory Interneurons Mediate Autism-Associated Behaviors via 4E-BP2”. In: *Proceedings of the National Academy of Sciences of the United States of America* 116 (36 2019), pp. 18060–18067.
- [218] Miyoshi Goichi et al. “FoxG1 Regulates the Formation of Cortical GABAergic Circuit during an Early Postnatal Critical Period Resulting in Autism Spectrum Disorder-like Phenotypes”. In: *Nature Communications* 12 (1 2021), p. 3773.
- [219] Bakken Trygve et al. “Single-Cell and Single-Nucleus RNA-Seq Uncovers Shared and Distinct Axes of Variation in Dorsal LGN Neurons in Mice, Non-Human Primates, and Humans”. In: *eLife* 10 (Sept. 2021), e64875.
- [220] Zhang Qi et al. “Netrin-G1 Regulates Fear-like and Anxiety-like Behaviors in Dissociable Neural Circuits”. In: *Scientific Reports* 6 (June 2016), p. 28750.
- [221] Mathys Hansruedi et al. “Single-Cell Transcriptomic Analysis of Alzheimer’s Disease”. In: *Nature* 570 (7761 2019), pp. 332–337.
- [222] Batiuk Mykhailo et al. “Identification of Region-Specific Astrocyte Subtypes at Single Cell Resolution”. In: *Nature Communications* 11 (1 2020), p. 1220.
- [223] Huang Xingfan et al. “Single Cell, Whole Embryo Phenotyping of Pleiotropic Disorders of Mammalian Development”. In: *bioRxiv* (Aug. 2022).
- [224] Kirdajova Denisa et al. “Transient Astrocyte-like NG2 Glia Subpopulation Emerges Solely Following Permanent Brain Ischemia”. In: *Glia* 69 (11 2021), pp. 2658–2681.
- [225] Shu et al. “ADAP2 Is an Interferon Stimulated Gene That Restricts RNA Virus Entry”. In: *PLoS Pathogens* 11 (9 2015), e1005150.
- [226] Cusanovich Darren et al. “Multiplex Single Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing”. In: *Science* 348 (2015), pp. 910–914.

- [227] Ishibashi M et al. “Targeted Disruption of Mammalian Hairy and Enhancer of Split Homolog-1 (HES-1) Leads to up-Regulation of Neural Helix-Loop-Helix Factors, Premature Neurogenesis, and Severe Neural Tube Defects”. In: *Genes & Development* 9 (24 1995), pp. 3136–3148.
- [228] Collignon J et al. “A Comparison of the Properties of Sox-3 with Sry and Two Related Genes, Sox-1 and Sox-2”. In: *Development* 122 (2 1996), pp. 509–520.
- [229] Gaiano N, Nye J, and Fishell G. “Radial Glial Identity Is Promoted by Notch1 Signaling in the Murine Forebrain”. In: *Neuron* 26 (2 2000), pp. 395–404.
- [230] Palma Veronica and Ruiz I Altaba Ariel. “Hedgehog-GLI Signaling Regulates the Behavior of Cells with Stem Cell Properties in the Developing Neocortex”. In: *Development* 131 (2 2004), pp. 337–345.
- [231] Kitamura Y et al. “Alteration of Transcription Factors NF- κ B and STAT1 in Alzheimer’s Disease Brains”. In: *Neuroscience Letters* 237 (1 1997), pp. 17–20.
- [232] Citron Bruce et al. “Transcription Factor Sp1 Dysregulation in Alzheimer’s Disease”. In: *Journal of Neuroscience Research* 86 (11 2008), pp. 2499–2504.
- [233] Tiwari Prafulla, Chandra Rishi, and Pal. “The Potential Role of Neuroinflammation and Transcription Factors in Parkinson Disease”. In: *Dialogues in Clinical Neuroscience* 19 (1 2017), pp. 71–80.
- [234] Consortium Psychencode et al. “The PsychENCODE Project”. In: *Nature Neuroscience* 18 (12 2015), pp. 1707–1712.
- [235] Prater Katherine et al. “Transcriptomic Profiling of Myeloid Cells in Alzheimer’s Disease Brain Illustrates Heterogeneity of Microglia Endolysosomal Subtypes”. In: (2021). DOI: 10.1101/2021.10.25.465802.
- [236] Gerrits Emma et al. “Neurovascular Dysfunction in GRN-Associated Frontotemporal Dementia Identified by Single-Nucleus RNA Sequencing of Human Cerebral Cortex”. In: *Nature Neuroscience* 25 (8 2022), pp. 1034–1048.

- [237] Pieper Alexander et al. “NeuroD2 Controls Inhibitory Circuit Formation in the Molecular Layer of the Cerebellum”. In: *Scientific Reports* 9 (1 2019), p. 1448.
- [238] Pletscher-Frankild et al. “DISEASES: Text Mining and Data Integration of Disease-Gene Associations”. In: *Methods* 74 (Mar. 2015), pp. 83–89.
- [239] Nguyen et al. “Global Methylation Profiling of Lymphoblastoid Cell Lines Reveals Epigenetic Contributions to Autism Spectrum Disorders and a Novel Autism Candidate Gene, RORA, Whose Protein Product Is Reduced in Autistic Brain”. In: *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology* 24 (8 2010), pp. 3036–3051.
- [240] Sayad Arezou et al. “Retinoic Acid-Related Orphan Receptor Alpha (RORA) Variants Are Associated with Autism Spectrum Disorder”. In: *Metabolic Brain Disease* 32 (5 2017), pp. 1595–1601.
- [241] Liu Xiling et al. “Disruption of an Evolutionarily Novel Synaptic Expression Pattern in Autism”. In: *PLoS Biology* 14 (9 2016), e1002558.
- [242] Hirayama Teruyoshi et al. “CTCF Loss Induces Giant Lamellar Bodies in Purkinje Cell Dendrites”. In: (2022). DOI: [10.21203/rs.3.rs-1970261/v1](https://doi.org/10.21203/rs.3.rs-1970261/v1).
- [243] Chang Jonathan et al. “Genotype to Phenotype Relationships in Autism Spectrum Disorders”. In: *Nature Neuroscience* 18 (2 2015), pp. 191–198.
- [244] Calderon Diego et al. “The Continuum of Drosophila Embryonic Development at Single-Cell Resolution”. In: *Science* 377 (6606 2022), p. 5800.
- [245] Efthymiou Anastasia and Goate Alison. “Late Onset Alzheimer’s Disease Genetics Implicates Microglial Pathways in Disease Risk”. In: *Molecular Neurodegeneration* 12 (1 2017), p. 43.
- [246] Fang Rongxin et al. “Comprehensive Analysis of Single Cell ATAC-Seq Data with SnapATAC”. In: *Nature Communications* 12 (1 2021), p. 1337.

- [247] Leung Szi et al. “Full-Length Transcript Sequencing of Human and Mouse Cerebral Cortex Identifies Widespread Isoform Diversity and Alternative Splicing”. In: *Cell Reports* 37 (7 2021), p. 110022.
- [248] Boggs J. “Myelin Basic Protein: A Multifunctional Protein”. In: *Cellular and Molecular Life Sciences: CMLS* 63 (17 2006), pp. 1945–1961.
- [249] Harauz George, Ladizhansky Vladimir, and Boggs Joan. “Structural Polymorphism and Multifunctionality of Myelin Basic Protein”. In: *Biochemistry* 48 (34 2009), pp. 8094–8104.
- [250] Harauz George and Boggs Joan. “Myelin Management by the 18.5-kDa and 21.5-kDa Classic Myelin Basic Protein Isoforms”. In: *Journal of Neurochemistry* 125 (3 2013), pp. 334–361.
- [251] Dietz Karen et al. “Targeting Human Oligodendrocyte Progenitors for Myelin Repair”. In: *Experimental Neurology* 283 (2016), pp. 489–500.
- [252] Bulik-Sullivan Brendan et al. “LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies”. In: *Nature Genetics* 47 (3 2015), pp. 291–295.
- [253] Finucane Hilary et al. “Partitioning Heritability by Functional Annotation Using Genome-Wide Association Summary Statistics”. In: *Nature Genetics* 47 (11 2015), pp. 1228–1235.
- [254] Yang et al. “Neuroinflammatory Mechanisms of Blood-Brain Barrier Damage in Ischemic Stroke”. In: *American Journal of Physiology* 316 (2 2019), pp. C135–C153.
- [255] Bohlen Christopher et al. “Microglia in Brain Development, Homeostasis, and Neurodegeneration”. In: *Annual Review of Genetics* 53 (Dec. 2019), pp. 263–288.
- [256] Jiang X, Lachance M, and Rossignol E. “Involvement of Cortical Fast-Spiking Parvalbumin-Positive Basket Cells in Epilepsy”. In: *Progress in Brain Research* 226 (June 2016), pp. 81–126.

- [257] Kamma Emily et al. “Central Nervous System Macrophages in Progressive Multiple Sclerosis: Relationship to Neurodegeneration and Therapeutics”. In: *Journal of Neuroinflammation* 19 (1 2022), p. 45.
- [258] Voet Sofie, Prinz Marco, and Van Loo Geert. “Microglia in Central Nervous System Inflammation and Multiple Sclerosis Pathology”. In: *Trends in Molecular Medicine* 25 (2 2019), pp. 112–123.
- [259] Domingues Ana et al. “Glial Cells in Parkinson’s Disease: Protective or Deleterious?” In: *Cellular and Molecular Life Sciences: CMLS* 77 (24 2020), pp. 5171–5188.
- [260] Iovino L, Tremblay M, and Civiero L. “Glutamate-Induced Excitotoxicity in Parkinson’s Disease: The Role of Glial Cells”. In: *Journal of Pharmacological Sciences* 144 (3 2020), pp. 151–164.
- [261] Demontis et al. “Discovery of the First Genome-Wide Significant Risk Loci for Attention Deficit/hyperactivity Disorder”. In: *Nature Genetics* 51 (1 2019), pp. 63–75.
- [262] Nagai Jun et al. “Hyperactivity with Disrupted Attention by Activation of an Astrocyte Synaptogenic Cue”. In: *Cell* 177 (5 2019), pp. 1280–1292.
- [263] Qian Yongjun et al. “Programmable RNA Sensing for Cell Monitoring and Manipulation”. In: *bioRxiv* (2022). DOI: 10.1101/2022.05.25.493141.
- [264] Kessler Matthew and Rawlins Richard. “A 75-Year Pictorial History of the Cayo Santiago Rhesus Monkey Colony”. In: *American Journal of Primatology* 78 (1 2016), pp. 6–43.
- [265] Widdig Anja et al. “Low Incidence of Inbreeding in a Long-Lived Primate Population Isolated for 75 Years”. In: *Behavioral Ecology and Sociobiology* 71 (1 2017), p. 18.
- [266] Hernandez-Pacheco et al. “Managing the Cayo Santiago Rhesus Macaque Population: The Role of Density”. In: *American Journal of Primatology* 78 (1 2016), pp. 167–181.
- [267] Testard Camille et al. “Social Connections Predict Brain Structure in a Multidimensional Free-Ranging Primate Society”. In: *Science Advances* 8 (15 2022), p. 5794.

- [268] Chiou Kenneth et al. “Multi-Region Transcriptomic Profiling of the Primate Brain Reveals Signatures of Aging and the Social Environment”. In: *Nature Neuroscience* (2022).
- [269] Chiou Kenneth et al. “Rhesus Macaques as a Tractable Physiological Model of Human Ageing”. In: *Philosophical Transactions of the Royal Society of London. Series B* 375 (1811), p. 20190612.
- [270] Desimone R and Ungerleider L. “Multiple Visual Areas in the Caudal Superior Temporal Sulcus of the Macaque”. In: *The Journal of Comparative Neurology* 248 (2 1986), pp. 164–189.
- [271] Born Richard and Bradley David. “Structure and Function of Visual Area MT”. In: *Annual Review of Neuroscience* 28 (2005), pp. 157–189.
- [272] Warren Wesley et al. “Sequence Diversity Analyses of an Improved Rhesus Macaque Genome Enhance Its Biomedical Utility”. In: *Science* 370 (6523 2020), p. 6617.
- [273] Virshup Isaac et al. “Anndata: Annotated Data”. In: (2021). DOI: 10.1101/2021.12.16.473007.
- [274] Wolf F et al. “SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis”. In: *Genome Biology* 19 (1 2018), p. 15.
- [275] Polański et al. “BBKNN: Fast Batch Alignment of Single Cell Transcriptomes”. In: *Bioinformatics* 36 (3 2020), pp. 964–965.
- [276] Dong W, Moses C, and Li K. “Efficient K-Nearest Neighbor Graph Construction for Generic Similarity Measures”. In: *Proceedings of the 20th International Conference on World Wide Web* (2011), pp. 577–586.
- [277] Bushnell Brian. “BBMap: A Fast, Accurate, Splice-Aware Aligner”. In: (2014).
- [278] Bouckaert Remco. “DensiTree: Making Sense of Sets of Phylogenetic Trees”. In: *Bioinformatics* 26 (10 2010), pp. 1372–1373.

- [279] Schliep Klaus and Peter. “Phangorn: Phylogenetic Analysis in R”. In: *Bioinformatics* 27 (4 2011), pp. 592–593.
- [280] Gaspar John. “Improved Peak-Calling with MACS2”. In: (2018). DOI: 10.1101/496521.
- [281] Pedregosa Fabian et al. “Scikit-Learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.
- [282] Bredikhin Danila, Kats Ilia, and Stegle Oliver. “MUON: Multimodal Omics Analysis Framework”. In: *Genome Biology* 23 (1 2022), p. 42.
- [283] Satija et al. “Spatial Reconstruction of Single-Cell Gene Expression Data”. In: *Nature Biotechnology* 33 (5 2015), pp. 495–502.
- [284] Machlab Dania et al. “monaLisa: An R/Bioconductor Package for Identifying Regulatory Motifs”. In: *Bioinformatics* 38 (9 2022), pp. 2624–2625.
- [285] Khan Aziz et al. “JASPAR 2018: Update of the Open-Access Database of Transcription Factor Binding Profiles and Its Web Framework”. In: *Nucleic Acids Research* 46 (D1 2018), pp. D260–D266.
- [286] Hinrichs A et al. “The UCSC Genome Browser Database: Update”. In: *Nucleic Acids Research* 34 (Database 2006), pp. D590–D598.
- [287] Wang Mingyue et al. “High-Resolution 3D Spatiotemporal Transcriptomic Maps of Developing Drosophila Embryos and Larvae”. In: *Developmental Cell* 57 (10 2022), pp. 1271–1283.
- [288] 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.
- [289] Sanjay R Srivatsan et al. “Embryo-scale, single-cell spatial transcriptomics”. en. In: *Science* 373.6550 (July 2021), pp. 111–117.
- [290] Chari Tara, Banerjee Joeyta, and Pachter Lior. “The Specious Art of Single-Cell Genomics”. In: *BioRxiv* (2021). DOI: 10.1101/2021.08.25.457696.

- [291] 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.
- [292] Yuhao Hao et al. “Integrated analysis of multimodal single-cell data”. en. In: *Cell* 184.13 (June 2021), 3573–3587.e29.
- [293] 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.