2021

Understanding Gene Regulation In Development And Differentiation Using Single Cell Multi-Omics

Qin Zhu
University of Pennsylvania

Understanding Gene Regulation In Development And Differentiation Using Single Cell Multi-Omics

Abstract

Transcriptional regulation is a major determinant of tissue-specific gene expression during development. My thesis research leverages powerful single-cell approaches to address this fundamental question in two developmental systems, C. elegans embryogenesis and mouse embryonic hematopoiesis. I have also developed much-needed computational algorithms for single-cell data analysis and exploration. C. elegans is an animal with few cells, but a striking diversity of cell types. In this thesis, I characterize the molecular basis for their specification by analyzing the transcriptomes of 86,024 single embryonic cells. I identified 502 terminal and pre-terminal cell types, mapping most single cell transcriptomes to their exact position in C. elegans' invariant lineage. Using these annotations, I find that: 1) the correlation between a cell's lineage and its transcriptome increases from mid to late gastrulation, then falls dramatically as cells in the nervous system and pharynx adopt their terminal fates; 2) multilineage priming contributes to the differentiation of sister cells at dozens of lineage branches; and 3) most distinct lineages that produce the same anatomical cell type converge to a homogenous transcriptomic state. Next, I studied the development of hematopoietic stem cells (HSCs). All HSCs come from a specialized type of endothelial cells in the major arteries of the embryo called hemogenic endothelium (HE). To examine the cellular and molecular transitions underlying the formation of HSCs, we profiled nearly 40,000 rare single cells from the caudal arteries of embryonic day 9.5 (E9.5) to E11.5 mouse embryos using single-cell RNA-Seq and single-cell ATAC-Seq. I identified a continuous developmental trajectory from endothelial cells to early precursors of HSCs, and several critical transitional cell types during this process. The intermediate stage most proximal to HE, which we termed pre-HE, is characterized by increased accessibility of chromatin enriched for SOX, FOX, GATA, and SMAD binding motifs. I also identified a developmental bottleneck separates pre-HE from HE, and RUNX1 dosage regulates the efficiency of the pre-HE to HE transition. A distal enhancer of Runx1 shows high accessibility in pre-HE cells at the bottleneck, but loses accessibility thereafter. Once cells pass the bottleneck, they follow distinct developmental trajectories leading to an initial wave of lympho-myeloid-biased progenitors, followed by precursors of HSCs. During the course of both projects, I have developed novel computational methods for analyzing single-cell multi-omics data, including VERSE, PIVOT and VisCello. Together, these tools constitute a comprehensive single cell data analysis suite that facilitates the discovery of novel biological mechanisms.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Genomics & Computational Biology

First Advisor

Kai Tan

Second Advisor

Junhyong Kim

Subject Categories

Bioinformatics

UNDERSTANDING GENE REGULATION IN DEVELOPMENT AND DIFFERENTIATION USING SINGLE CELL MULTI-OMICS

Qin Zhu

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Co-Supervisor of Dissertation

Kai Tan

Kai Tan

Professor of Pediatrics

Co-Supervisor of Dissertation

Junhyong Kim

Professor of Biology

Graduate Group Chairperson

Brian Gregory, Associate Professor of Biology

Dissertation Committee

Klaus Kaestner, Professor of Genetics

Nancy A. Speck, Professor of Cell and Developmental Biology

Kun Zhang, Professor of Bioengineering, University of California San Diego

UNDERSTANDING GENE REGULATION IN DEVELOPMENT AND DIFFERENTIATION USING SINGLE CELL MULTI-OMICS

ACKNOWLEDGMENT

First and foremost, I want to thank my Ph.D. mentors Professor Kai Tan and Professor Junhyong Kim for their continuous support and mentorship. Kai and Junhyong have always been there for me and helped me grow as an independent scientist. Kai guided me into the field of systems biology, taught me critical thinking and writing, and provided me many opportunities and resources. Junhyong kindled my passion for single cell biology and bioinformatics when I was a Master's student and has taught me numerous things about life and science since then. I am forever indebted to both of them for all their guidance and support.

My thesis work builds on successful collaborations. I want to thank Professor Nancy Speck for introducing me to the field of HSC development and guiding me through our many collaboration projects. Her enthusiasm for science inspired me a lot and over the years I have learned so much from her. I also want to thank Professor John Murray and Professor Bob Waterston for introducing me into the world of worms. Through our collaboration, John and Bob have turned me from a "worm rookie" into a worm enthusiast. I am grateful for my collaborator Professor Christopher Lengner, who taught me a lot about intestine development and colon cancer. I want to thank members of my collaborating labs, who worked closely with me on various projects: Ning Li, Jonathan Packer, Joanna Tober, Laura Bennett, Priya Sivaramakrishnan, Melanie Mumau, Yan Li and Elizabeth Howell.

I would also like to take this opportunity to thank my thesis committee members: Brian Gregory, Nancy Speck, Klaus Kaestner and Kun Zhang, for their valuable feedbacks and suggestions during my committee meeting and whenever I needed them.

The work presented in this thesis would not have been possible without the contribution from my lab mates. I would like to specifically thank Peng Gao, Changya Chen, Yasin Uzun, Wenbao Yu and Bing He from the Tan Lab, and Stephen Fisher, Youtao Lu, Hannah Dueck, Derek Stefanik, Erik Nordgren, Sarah Middleton and Mugdha Khaladkar from the Kim Lab. I am extremely grateful to all current and former lab members of the Tan Lab and the Kim Lab for creating a productive, collaborative, and pleasant lab environment.

I would like to thank the Blavatnik Family Foundation for selecting me as a 2020 recipient of the Blavatnik Family Fellowship in Biomedical Research. I would also like to thank the GCB program for providing a wonderful, supportive environment that allowed me to thrive. I want to specifically thank Maureen Kirsch for being such a wonderful coordinator.

Many of my friends have made my graduate school life exciting and colorful. I want to thank Changya Chen, Youtao Lu, Wenbao Yu, Xiang Yu, Yuxuan Hu, Tao Peng, Jun Xiao, Peng Hu, Qi Qiu, Jian Chang and Yang Zhu for playing table tennis, badminton, and soccer with me over the years. I also want to thank my friends and fellow students in GCB, especially Chi-Yun Wu, Zhen Miao, Samantha Klasfeld, Michelle Lee, Blair Zhang, Katherine Huang and Zilu Zhou, who brought lots of joy to my life at Penn.

Finally, I want to thank my girlfriend, my parents, and my grandparents for their unwavering and unconditional love. I would not have reached today’s milestone without their love and support.

ABSTRACT

UNDERSTANDING GENE REGULATION IN DEVELOPMENT AND DIFFERENTIATION USING SINGLE CELL MULTI-OMICS

Qin Zhu

Kai Tan and Junhyong Kim

Transcriptional regulation is a major determinant of tissue-specific gene expression during development. My thesis research leverages powerful single-cell approaches to address this fundamental question in two developmental systems, *C. elegans* embryogenesis and mouse embryonic hematopoiesis. I have also developed much-needed computational algorithms for single-cell data analysis and exploration.

C. elegans is an animal with few cells, but a striking diversity of cell types. In this thesis, I characterize the molecular basis for their specification by analyzing the transcriptomes of 86,024 single embryonic cells. I identified 502 terminal and pre-terminal cell types, mapping most single cell transcriptomes to their exact position in *C. elegans'* invariant lineage. Using these annotations, I find that: 1) the correlation between a cell's lineage and its transcriptome increases from mid to late gastrulation, then falls dramatically as cells in the nervous system and pharynx adopt their terminal fates; 2) multilineage priming contributes to the differentiation of sister cells at dozens of lineage branches; and 3) most distinct lineages that produce the same anatomical cell type converge to a homogenous transcriptomic state.

Next, I studied the development of hematopoietic stem cells (HSCs). All HSCs come from a specialized type of endothelial cells in the major arteries of the embryo called hemogenic endothelium (HE). To examine the cellular and molecular transitions underlying the formation of HSCs, we profiled nearly 40,000 rare single cells from the caudal arteries of embryonic day 9.5 (E9.5) to E11.5 mouse embryos using single-cell RNA-Seq and single-cell ATAC-Seq. I identified a continuous developmental trajectory from endothelial cells to early precursors of HSCs, and several critical transitional cell types during this process. The intermediate stage most proximal to HE, which we termed pre-HE, is characterized by increased accessibility of chromatin enriched for SOX, FOX, GATA, and SMAD binding motifs. I also identified a developmental bottleneck separates pre-HE from HE, and RUNX1 dosage regulates the efficiency of the pre-HE to HE transition. A distal enhancer of *Runx1* shows high accessibility in pre-HE cells at the bottleneck, but loses accessibility thereafter. Once cells pass the bottleneck, they follow distinct developmental trajectories leading to an initial wave of lympho-myeloid-biased progenitors, followed by precursors of HSCs.

During the course of both projects, I have developed novel computational methods for analyzing single-cell multi-omics data, including VERSE, PIVOT and VisCello. Together, these tools constitute a comprehensive single cell data analysis suite that facilitates the discovery of novel biological mechanisms.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 INTRODUCTION

During development, a fertilized egg undergoes repeated cell divisions to produce an embryo that contains distinct cell types. This sequence of cell divisions is called the organism's cell lineage. Each cell in the lineage expresses a different set of genes in various quantities (the cell's transcriptome), thus directing cells to differentiate into specific cell types. It is not yet fully understood how cells control its gene expression during differentiation, and how cells interact with each other to form complex tissue structures.

Historically, researchers have studied gene expression during development through analysis of pooled population of cells from different developmental time points. These "bulk" methods were able to capture global expression changes over time and upon perturbation, but often fail to address the heterogeneous change and response of each individual cell type. Therefore, it is extremely difficult to study the development of rare cell types such as hematopoietic stem cells (HSCs). The emergence of single cell technology made it possible to simultaneously profile the molecular state of almost every cell in a multi-cellular organism. Unlike bulk methods, single-cell approach does not require purification of cell population or synchronization of the developing organism. Instead, by sampling cells across tissues and developmental stages, methods like single cell RNA sequencing (scRNA-Seq) are able to capture transcriptomes of various cell types and differentiation states (*1, 2*), bringing unprecedented resolution to the study of development. Recent technologies such as single cell ATAC (Assay for Transposase Accessible Chromatin) sequencing (scATAC-Seq) (*3, 4*) can identify active DNA elements that likely promote or inhibit gene expression, thus facilitating mechanistic understanding of gene regulatory program. Furthermore, multiplexed cytometric imaging techniques such as CODEX (CO-Detection by indEXing) (*5*) imaging made it possible to visualize the spatial distribution of different cell types in the tissue microenvironment and investigate cell-cell interactions. Such interactions play critical roles in cell fate specification and tissue patterning during normal development, and mediate pathological processes such as immune cell infiltration and tumor metastasis in cancer.

This technological revolution provides exciting opportunities to study development, but also poses challenges in data analysis and visualization. For developmental biology, it is particularly important to model the differentiation trajectory and characterize the underlying gene regulatory network. Novel methods are required to integrate different data modalities, such as scRNA-Seq and scATAC-Seq data, to gain mechanistic understanding of cell differentiation. The rapid growth of single cell data also requires robust software environment for easy and fast data exploration to bridge the gap between data generation and biological discovery.

Development and differentiation

C. elegans embryogenesis

Caenorhabditis elegans (*C. elegans*) is a simple, transparent organism with only 558 somatic cells upon hatching and 959 somatic cells in its adult hermaphroditic form. Unlike many complex organisms, the cell lineage of *C. elegans* is invariant across different individuals and has been fully resolved (*6, 7*), making it an ideal organism to study cellular differentiation and organismal development. *C. elegans* is the first multicellular organism whose genome gets completely sequenced, revealing a small genome size of 97Mb and about 19,000 genes in total (*8*). The self-fertilization capability of the hermaphrodites enables easy genetic screening to recover recessive mutants and identify genes critical to the developmental process. Cross-fertilization could also happen between males and hermaphrodites and will produce over 1000 fertilized eggs. Fertilization initiates the embryogenesis process, which takes about 14 hours to transform a single zygote into a moving L1 larva.

Embryogenesis of *C. elegans* can be divided into several stages, including fertilization, proliferation, gastrulation, morphogenesis, elongation, quickening and hatching (Figure 1.1) (*9*). Upon fertilization, anterior-posterior axis is established based on the entry position of the sperm. The sperm pronucleus is pushed to the nearest end of the elongated oocyte, making it the posterior pole (*10*). In the first 150 min post fertilization, a series of asymmetric division happens to establish a set of founder cells, including AB, MS, E, C and D (Figure 1.2). Each founder cell undergoes subsequent divisions to give rise to cells that will differentiate into various cell types. Gastrulation happens in conjunction with the proliferation of the founder cell lineages, where the AB and C lineages produce ectoderm and pharynx, MS descendants form mesoderm and pharynx, and E descendants form endoderm. Towards the end of gastrulation, the majority of cells start terminal differentiation and organize into various tissues, while the rest of cells undergo programmed cell death. Elongation occurs in parallel with tissue morphogenesis, and the embryo grows about three-fold to adopt its long worm-like shape. The embryo can be seen moving inside the egg about 650 minutes post-fertilization, and hatches with about 600 cells 14 hours post-fertilization.

The simplicity of the organism makes it possible to experimentally perturb the developmental process to gain insight into cellular differentiation and tissue development. For example, individual cells in the early developing embryo can be destroyed via laser ablation to reveal its developmental potential (*7, 11, 12*). Using this approach, it has been shown that every single cell in the early embryo is indispensable for normal development, and their fate is determined either autonomously or conditionally. Autonomous fate specification means the intrinsic cellular factors determine the fate of the cell and its progenies, and no extrinsic signaling is required. The determination of P1 lineage is autonomous as it can generate the posterior part of the embryo without the presence of AB (*11*). After P1 divides, one of its daughter cells, EMS, also have the capacity to produce pharyngeal tissues in isolation. The maternal protein SKN-1 is likely a key intrinsic factor that specifies the fate of EMS. Hermaphrodites with *skn-1* mutation produce embryos without pharyngeal mesoderm and endoderm, but with extra skin and muscle cells (*13*). When EMS cell divides, the specification of the E cell lineage (endoderm) is conditional, as it requires interaction with the P2 blastomere through Wnt signaling (*14, 15*). Without contact with P2, which expresses Wnt ligands *mom*, EMS cell will produce two MS cells but no intestine will be formed (*15*). Similarly, specification of the ABp cell fates depends on the contact between ABp and P2 through Notch signaling (*16*). The Notch receptor GLP-1 protein can be found in both ABa and ABp cells, but only ABp contacts P2 which has the Notch ligand APX-1. This signaling asymmetry cause ABp to adopt a different fate from its sister, thus establishing the dorsal-ventral axis of the embryo.

In recent years, omics technologies have enabled profiling of hundreds and thousands of gene expression patterns in the *C. elegans* embryo, allowing systematic characterization of gene regulatory pathways. For example, using fluorescent reporters, Murray *et al.* traced expression of 127 genes across cell lineages and developmental time points (*17*). They identified several genes downstream of the Wnt signaling pathway, which interact with the transcription factor POP-1 to generate diverse lineage-specific expression patterns. One of the earliest scRNA-Seq technologies, CEL-Seq, was first applied to study *C. elegans* embryogenesis (*18*). By comparing the gene expression profile of daughter cells with the mother cell, the researchers found transcription factors are highly enriched among daughter-cell-expressed genes, suggesting their critical roles in early lineage specification. Similarly, using scRNA-Seq, Tintori *et al.* profiled gene expression in the early embryo up to the 16-cell stage (*19*). They observed a global similarity in gene expression between AB descendants, except for a few genes including known Notch targets, *hlh-27*, *ref-1*, and *tbx-38*. Therefore, Notch signaling may be one of the key discriminating factors that drives different fate choice of early AB lineage cells.

Despite all these efforts, it largely remains a mystery how the fate of every single cell is robustly specified throughout embryonic development. Comprehensive gene expression profiling of single cells during *C. elegans* embryogenesis could provide the first step towards mechanistic understanding of this process. In Chapter 2, I describe our efforts to construct a lineage-resolved molecular atlas of the *C. elegans* embryo, which includes 86,024 single-cell gene expression profiles covering 87% of the embryonic lineages. Using this dataset, we modeled the differentiation trajectories of the cells and identified developmental patterns that are prevalent across embryonic lineages.

Embryonic origin of blood

Development of the blood system is a highly conserved process across vertebrates, and has been extensively studied in chicken, zebrafish, and mouse (*20*). The mouse hematopoiesis system shares a lot of similarity with that of human and therefore has been used as a model system for studying mammalian blood development. During early embryogenesis, there are multiple waves of blood cell progenitor formation (*21*). The first wave occurs around embryonic day (E) 7 in the extra-embryonic yolk sac of mouse, where a group of mesodermal progenitor cells differentiate into erythrocyte, megakaryocyte, and macrophage progenitors (*22-24*). These cells are transiently produced to support fetal development and will not last until adulthood, thus were termed the “primitive wave” of blood formation. The second wave originates from the vasculature of yolk sac around E8.25 (*25*). A subset of the endothelial cells in the blood vessel undergoes “endothelial to hematopoietic transition” (EHT), and form erythro-myeloid progenitors (EMPs) (*26, 27*). The second wave also involves generation of lymphoid progenitors from the yolk sac, dorsal aorta, vitelline arteries, and umbilical arteries, which will differentiate into T cells and B cells (*28-30*).

Hematopoietic stem cells (HSCs) first emerge from the aorta-gonad-mesonephros (AGM) region between E10.5 to E11.5 during the third “definitive wave” of blood formation (*31, 32*). These cells have both self-renew capability and the potential to differentiate into all blood cell types. Formation of HSCs involves multiple differentiation steps and several intermediate cell types (*20, 33*). Around E9.5, a subset of endothelial cells in the dorsal aorta is specified as hemogenic endothelial (HE) cells, which can be distinguished by the expression of the transcription factor Runx1 (*34*). The HE cells undergo EHT to give rise to hematopoietic stem and progenitor cells (HSPCs) that accumulate in intra-arterial clusters (IAC) (*35*). Limiting dilution analysis revealed that at E10.5, fewer than 10% of the several hundred IAC cells are HSC precursors, or pre-HSCs (*36, 37*). Pre-HSCs cannot engraft adult mice directly, but can mature *in vivo* or *ex vivo* into HSCs that can engraft (*38*). All pre-HSCs at E10.5 lack the pan-hematopoietic marker CD45, and are called type I pre-HSCs (*38*). At E11.5, the IACs contain type I pre-HSCs, CD45+ type II pre-HSCs that have matured from type I pre-HSCs, and ~1 HSC (*36*). CD45+ cells are also found in the IACs at E10.5, but these cells cannot be matured into HSCs and hence are not pre-HSCs (*38*). The lineage relationship between the CD45+ IAC cells that appear at E10.5 and the CD45+ type II pre-HSCs at E11.5 is unknown. Pioneering scRNA-seq analyses identified type I and II pre-HSCs within IACs (*39, 40*), but the overall composition of the IACs was not described due to the small number of cells that were previously analyzed. The pre-HSCs eventually detaches from the blood vessel and migrates to fetal liver, where they undergo rapid expansion and further maturation before colonizing the bone marrow to support adult hematopoiesis (*33*).

Formation of pre-HSCs from the arterial endothelium is regulated by multiple transcription factors and signaling pathways. The transcription factor *Runx1* is a critical regulator of the definitive waves of hematopoiesis (*33, 41, 42*). *Runx1*-deficient embryos dies by E12.5 with severe anemia due to the lack of blood production from the last two waves (*33*). Conditional knock-out of *Runx1* in vascular endothelial cells shows *Runx1* is essential for the generation of pre-HSCs and formation of the IACs (*41*). But *Runx1* is not required to maintain the hematopoietic cell identity as conditional knock-out of *Runx1* in the hematopoietic cells does not cause embryonic lethality (*41*). Previous studies show *Runx1* functions in HE to recruit hematopoietic regulators, *Tal1* and *Fli1* to induce the transition towards hematopoietic fate (*43*). Direct targets of *Runx1* include *Spi1* and *Gfi1* (*33, 44*), both are essential for EHT and are used in a transduction cocktail along with *Runx1* and *Fosb* to reprogram endothelial cells to hematopoietic cells *in vitro* (*45*). Besides *Runx1*, *Gata2* is another transcription factor essential for EHT (*46*). Haploinsufficiency of *Gata2* results in severe reduction of HSC production in the AGM region (*47*). *Gata2* expression is induced by Notch1 signaling, which is transiently required for the specification of HE (*48*). The Notch ligand, *Dll4* and *Jag1* are both expressed in the dorsal aorta. When in contact with *Dll4*, Notch activity is upregulated and promotes the arterial fate of endothelium (*49*). When bound by *Jag1*, Notch activity is restricted, permitting the induction of hematopoietic program (*49*). Other signaling pathways involved in HE specification and EHT includes retinoid acid signaling (*50*), BMP signaling (*51*), cytokine signaling (*52*) and fluid shear stress (*53*). However, due to the technical difficulties in isolating the rare cell population involved in EHT, little is known about the interplay between these pathways and the transcriptional regulatory network downstream of the signaling pathways.

In Chapter 3, I describe our effort to map a continuous developmental trajectory from endothelial cells to early precursors of HSCs by analyzing ~40,000 rare single cells from early developing mouse embryos. Through joint analysis of scRNA-Seq and scATAC-Seq data, I identified a developmental bottleneck regulated by *Runx1* dosage, and a distal enhancer of *Runx1* with transient activity at the bottleneck. I also identified pathways with increased activities at the bottleneck. Once cells pass the bottleneck, they follow distinct developmental trajectories leading to an initial wave of lympho-myeloid-biased progenitors, followed by precursors of HSCs.

Single cell technologies

The past decade has witnessed rapid development of single cell technology. The following section reviews several key methodology innovations that enable measuring molecular phenotypes of cells in a high-throughput fashion.

Single cell RNA sequencing (scRNA-Seq)

Single cell RNA sequencing enables measurement of mRNA transcript abundance across thousands of cells simultaneously. Figure 1.3 shows an overview of the general procedure of scRNA-Seq. First, biological samples are harvested and are dissociated into a single cell suspension. Cells of interest are then loaded onto a single cell isolating device which captures each cell in e.g., a droplet. After cells are lysed, reverse transcription and cDNA amplification are performed to generate libraries for sequencing. Sequencing reads are analyzed to obtain mRNA abundance measure of each gene in every cell, which can be further analyzed to gain biological insights.

scRNA-Seq technology has evolved a lot over the past years and several major breakthroughs were made to improve the throughput and robustness of the method. For example, the utilization of microfluidic device enables sorting cells into separate oil droplets, beads, or wells (*54, 55*). Each of the cells were lysed and labeled with a unique molecular barcode, such that they can be pooled for library construction and sequencing. The Fluidigm® C1 system was one of the earliest microfluidic-based technology which allows capturing of 96 to 800 cells (*2*). Since then, technologies such as inDrop, Drop-Seq and Next GEM from 10x Genomics^{TM} were developed, enabling partitioning and labeling of thousands of single cells (*54-56*). A different strategy to increase the throughput is using combinatorial barcoding, which is implemented in the

SPLiT-Seq and sci-RNA-Seq protocols (*57-59*). The methods randomly distribute fixed cells into wells and label the cells with well-specific barcodes. With several rounds of random splitting, barcoding, and pooling, the overall complexity of the ligated barcodes will be high enough to uniquely label almost every single cell, allowing 1-2 million cells to be sequenced in a single experiment. Such approaches significantly reduce the cost for library preparation but may not robustly capture lowly expressed genes in the cell due to limitations in sequencing depth. For certain biological questions, medium cell number and high read coverage per cell may be preferred, as many key regulatory genes, such as transcription factors, are present in small amount in each cell but have global regulatory effects.

The limited materials in each single cell also require robust library preparation methods, such that technical variations do not distort the biological signal. To this end, linear amplification protocols such as *in vitro* transcription (IVT) have been developed (*18, 60, 61*), which avoids the uncontrolled scaling by exponential amplification methods. To make the counting of mRNA molecules even more accurate, a unique molecular identifier (UMI) can be attached to each molecule, such that after amplification the reads can be readily de-multiplicated (*62*). Technologies such as IVT and UMI have been incorporated into modern single cell platforms such as the 10x Genomics Chromium system, allowing robust quantification of single cell gene expression.

Single cell ATAC sequencing (scATAC-Seq) and single-cell multi-omics

Gene expression in development is regulated by both *cis*-acting DNA elements and *trans*-acting factors. ATAC-Seq uses the hyperactive transposase Tn5 to insert sequencing adaptors to accessible chromatin regions, enabling identification of active DNA regions and potential trans-factor binding sites (*63*) (Figure 1.4). The successful application of microfluidic device for scRNA-Seq inspired development of single-cell ATAC sequencing. Using the Fluidigm® C1 system, Buenrostro *et al.* developed a single cell ATAC-Seq protocol allowing simultaneous measurement of chromatin accessibility in hundreds of cells (*3*). Although the data are very sparse, it shows that scATAC was able to capture cell-type-specific epigenetic features and global chromosome compartments. Around the same time, Cusanovich *et al.* developed a different scATAC-Seq protocol using combinatorial indexing (*4*). Like sci-RNA-Seq, the sci-ATAC-Seq protocol uses two rounds of splitting and pooling of cells on a 96-well plate to introduce unique combination of barcodes for each single cell. The method is capable of profiling 1500 cells in a single run but have ~11-12% collision rate. Recently, scATAC-Seq using the 10x Chromium platform has gained much popularity due to its high throughput (thousands of cells) and low collision rate (0.8-4%). After nuclei suspension is made and incubated with Tn5 transposase, the Chromium device encapsulates each cell in a Gel bead-in EMulsion (GEM), where cell barcoding and linear amplification happens. The GEMs are then pooled and broken to release the barcoded DNA fragments for library construction and sequencing. A study by Satpathy *et al.* used this technology to profile more than 200,000 single cells from human blood and tumor microenvironment. They were able to reconstruct the trajectory of multiple immune cell lineages using this dataset, demonstrating that scATAC is a valuable tool to understand gene regulatory programs in development and differentiation (*64*).

Recently, several methods were developed to jointly profile mRNA level and chromatin accessibility in the same cell, such as sci-CAR (*65*), Paired-Seq (*66*), SNAREseq (*67*) and SHARE-seq (*68*). Besides these, other types of multi-omics methods were developed for joint assay of transcriptome and epitopes (*69*), chromosome conformation and methylomes (*70*), transcriptome and histone modification (*71*), and even three different modalities (*72, 73*). These assays show lots of promises for understanding hierarchical gene regulatory program during development and differentiation. For example, using SHARE-Seq, Ma *et al.* observed during skin development, domains of regulatory chromatin (DORCs) became accessible before many fate-specific genes are expressed, suggesting chromatin accessibility is predictive of cell fate decisions (*68*). This dynamic relationship between chromatin accessibility and gene expression, which they termed “chromatin potential”, can be quantified to estimate the time scale of fate commitment, and to facilitate the discovery of key lineage-determining genes.

Single cell imaging technologies

Although scRNA-Seq and scATAC-Seq provide quantitative readout of key molecular features of each cell, the sequencing-based technologies alone are unable to resolve the spatial location and dynamical changes of the cell. Such information is critical for understanding the interaction between the cell and its surrounding environment, which plays a pivotal role in cell fate specification during development and differentiation.

The development of reporter genes and microscopy technology has made it possible to visualize gene expression in living organism at single cell resolution (*74*). For example, using confocal microscopy and fluorescent reporter constructs, Murray *et al.* measured reporter expression of more than one hundred genes in the *C. elegans* embryo on a cell-by-cell basis (*17, 75*). With this cellular-resolution compendium of gene expression, the researchers identified interesting spatial expression patterns such as left-right-asymmetric gene expression, as well as temporal expression cascades that define cell fates in a sequential manner. The GFP-based single cell imaging technique has also been applied to other model organisms such as *Drosophila* (*76*), but has fundamental limitation in throughput due to overlapping fluorophore emission spectra.

Multiplexed *in situ* fluorescent imaging, such as CODEX (CO-Detection by indEXing) (*5*), DEI (DNA Exchange Imaging) (*77*) and t-CyCIF (Tissue-based cyclic immunofluorescence) (*78*), circumvents this limitation by repeated imaging of the same specimen over multiple cycles. For example, the CODEX technology (*5*) uses DNA-conjugated antibodies to stain the cells. In each cycle, three fluorophores tagged with complementary DNA sequence were introduced and bound to matched antibodies, allowing imaging of three proteins at the same time. The fluorophores were then washed away to start another cycle of imaging. With repeated cycling, CODEX is able to capture fluorescent images of up to 45 proteins for a single tissue section. Using this technology, researchers were able to identify spatial distribution of different cell types and infer cell-cell communication network in normal and diseased tissues (*5, 79*).

Besides measuring protein abundance at single cell level, direct imaging of single RNA molecules in the cell has been made possible through single-molecule fluorescence *in situ* hybridization (smFISH) (*80, 81*). Several variations of this technique have been developed to increase the throughput and robustness of the method, allowing thousands of mRNA species to be measured at cellular or sub-cellular resolution (*82-86*). For example, Long Cai's group developed seqFISH (sequential Fluorescence *In Situ* Hybridization) technology (*82, 87-89*), which uses sequential rounds of hybridization and fluorescent imaging to read out the temporal barcode for each mRNA transcript.

seqFISH enables *in situ* quantification of gene expression that preserves spatial gene expression pattern (*87*), as well as direct read out of lineage barcodes introduced using CRISPR/Cas9-based targeted mutagenesis (*88*). Another technology, multiplexed error-robust FISH (MERFISH) (*84*), also enables transcriptome-scale quantification of RNA species, and has been extended to DNA imaging to enable simultaneous capture of the 3D organization and transcription activity (*90*). Compared to scRNA-Seq, smFISH data contain additional information about cellular and sub-cellular localization of individual mRNA species. Therefore, smFISH offers critical insight into many developmental processes that are regulated by mRNA localization, such as embryonic patterning, cell fate specification, cell migration and synapse development of neurons.

Lineage tracing

A cell's developmental history could involve multiple rounds of cell division, cell migration and programmed cell death. During cell division, a parental cell gives rise to two daughter cells with same identity, or with completely different cell fates. Methods such as scRNA-Seq and scATAC-Seq capture snapshots of each cell's molecular state but cannot record the consecutive cell divisions which represents a cell's lineage history. The cell lineage history, although by nature is tree-like, may be reflected in the molecular state space as several disjoint clusters, bifurcating trajectories, or loops. Therefore, tracking a cell's lineage history and mapping it back to the state space could help researchers understand how cells traverse the molecular state space to reach its terminally differentiated state.

In the past, lineage tracing was done using microscopy on simple, transparent organisms such as ascidians (*91*) and *C. elegans* (*7*). Recent *in toto* confocal microscopy technique enables cell tracking in more complicated organisms such as zebrafish (*92*) and mouse (*93*). In addition, it is possible to label cells with dyes, radioactive tracers, or reporter genes at an early embryonic stage, and track the fate of its descendants (*94*). Although these methods were able to reconstruct partial or full cell lineage tree, they cannot be readily used in conjunction with single cell omics methods to link a cell's lineage with its molecular state.

With high-throughput sequencing, it is possible to use DNA or RNA sequences as lineage labels, which can be directly read out using single cell sequencing methods. For example, using naturally occurring mutations in mitochondria genome, Ludwig *et al.* was able to identify subclones in human cell lines and human colorectal cancer and simultaneously measure chromatin accessibility using scATAC-Seq (*95*). Genetic modification can also be introduced using recombination (*96*), transposition (*97*), and *in vivo* editing of DNA targets by CRISPR-Cas9 (*40, 88, 98, 99*). Through consecutive editing of a lineage-recording barcode, CRISPR-Cas9 based methods such as scGESTALT (*98*) and ScarTrace (*40*) enables tracking cell lineage history and measuring transcriptome at the same time. However, resolution of such methods is limited by the editing efficiency and barcode detection rate, thus not every cell lineage can be confidently resolved.

Computational methods for single cell analysis of development and differentiation

Identifying cell types and cell states

A multicellular organism is composed of various cell types that are functionally specialized and morphologically distinct. Although the definition of cell type is controversial, scRNA-Seq offers a unique way to relate a cell's identity with its transcriptomic profile, which is inherently associated with morphology and function (*100, 101*). Furthermore, scRNA-Seq enables detailed characterization of different molecular states of the same cell type as cell cycle, circadian rhythm, aging and disease often cause altered gene expression pattern (*102-104*). To robustly group and annotate cells as different cell types and cell states, numerous computational methods have been developed, which can be grouped into two major categories – the unsupervised clustering-based approach and the supervised reference-based approach.

Clustering algorithms have long been developed in the field of statistics and computer science. However, most of these methods cannot be readily applied to single cell data due to several important limitations. One of such limitations is the sparsity of the data, as current single cell technologies sometimes fail to capture transcripts of lowly expressed genes, leading to zeros in the data matrix (*105*). This "dropout" effect distorts the true gene expression pattern, leading to biases in the downstream analysis. To reduce the effect of dropouts, many groups have developed algorithms to impute the missing values (*106-110*). Others take a slightly different approach by explicitly modeling the dropout events. For example, the ZIFA algorithm uses zero-inflated factor analysis to impute the cell coordinates in a low dimensional latent space (*111*). A similar algorithm, ZINB-WaVE, uses zero-inflated negative binomial model on the count data, and observed tighter, biologically meaningful clusters (*112*).

Another challenge for single cell clustering is the curse of dimensionality, which refers to the instability of distance metrics in high dimensional space (*113*). Therefore, a common procedure prior to clustering analysis is to reduce the dimensionality of the data. Principal component analysis (PCA) is one of the most extensively used linear dimensionality reduction method. By successively maximizing the variance in each principal component (PC), the method projects cells into a low-dimensional space where many clustering methods, such as K-means clustering and density-based clustering, can be readily applied. Furthermore, each PC represents a linear combination of genes that contribute to the separation of the clusters, thus can be biologically interpreted. In one single-cell study on metastatic melanoma, the researchers observed different partition of cells along each PC, which suggests transcriptional heterogeneity associated with cell cycle, spatial context and drug resistance program (*114*). In another study on the development of early *C. elegans* embryo, iterative PCA was applied to identify distinct cell lineages arising from asymmetric divisions, where each PC corresponds to lineage-specific transcription programs (*19*).

Although linear methods like PCA enable visualization and clustering of high dimensional gene expression data, biological data are intrinsically nonlinear. In recent years, several nonlinear dimensionality reduction methods, such as t-distributed stochastic neighbor embedding (t-SNE) (*115*) and uniform manifold approximation and projection (UMAP) (*116, 117*) have gained much popularity. The t-SNE algorithm creates an embedding of cells that preserves the probability distribution of neighbors around each cell, such that cells close to each other in the high dimensional space have high probability being close together in the low dimensional embedding. With t-SNE, distinct cell types often appear as disjoint islands in the embedding, making it easy to identify rare cell types. For example, with t-SNE performed on ~7000 single cells from the airway epithelium, Montoro *et al.* discovered the presence of a rare CFTR-expressing cell type (*96*). These cells are found to be the primary source of the cystic fibrosis transmembrane conductance regulator CFTR, and their dysfunction leads to phenotypes that are characteristic of cystic fibrosis (*96*). The discovery was also made independently by another group using a non-linear, graph-based algorithm, SPRING, which uses force-directed layout to preserve the nearest-neighbor relationship of cells (*118, 119*). The UMAP algorithm was recently favored by many biologists, as it not only preserves local nearest neighbor structure like t-SNE, but also preserves the global distance relationship to some degree (*116, 117*). In this way, dissimilar cells are often segregated into clusters that are far away from each other, making it ideal to apply density-based clustering algorithms, such as DBSCAN (*120*), or graph-based community detection algorithms, such as the Louvain method (*121*).

To further improve accuracy and robustness of the clustering result, several methods explore different metrics of cell-cell similarity. For example, SIMLR uses multi-kernel learning to generate a similarity matrix with an approximate block-diagonal structure, which can be visualized with t-SNE and clustered using spectral method (*122*). Similarly, the SC3 algorithm performs K-means clustering with various distance metrics and performs consensus clustering on top of the K-means results (*123*). Other clustering approaches aim to improve the flexibility of single cell clustering by allow varying resolutions, such that cell types, subtypes and states can be hierarchically explored. To this end, several graph-based clustering methods have been developed, such as PAGA (*124*) and Toomanycells (*125*).

Currently, annotation of clusters heavily relies on cell-type marker genes curated from past literatures, and there is no gold standard for evaluating the accuracy of clustering and cell type annotation. This paradigm may soon change due to the influx of well-annotated single cell data, such as those from the Human Cell Atlas project (*126*) and the Human BioMolecular Atlas Program (HuBMAP) (*127*). For developmental biology, transcriptional atlases have also been created for multiple species (see Table 1.1 for a list of published datasets). Therefore, it is possible to train a cell type classifier using large-scale reference datasets and predict cell type labels for newly profiled single cells. Many methods have adapted classification algorithms such as random forest, elastic net and neural network for this purpose, which has been extensively reviewed by Pasquini *et al* (*128*). One limitation of such approaches is the accuracy of the classifier heavily relies on the quality of the reference data. If a rare cell type or a transient cell type is missing from the reference data, the algorithm may fail to identify it and may classify it into a close cell type. Therefore, for analysis described in this thesis, I mainly used unsupervised clustering to discover various cell types and cell states. In addition, I compared data from this study to published datasets using enrichment of cell-type signatures and projection method, which further validates our cell type annotations.

Modeling of differentiation trajectory, velocity, and potential

Single cell RNA-Seq provides a powerful way to investigate the sequential transcriptional regulatory events during differentiation. To recapitulate the temporal sequence of gene regulation, a trajectory inference algorithm can be applied to order cells based on their transcriptome similarity. Trajectory inference algorithms differ in their assumption about the temporal structure of the underlying process, and the resulting trajectory can be linear, cyclic, tree-like or of mixed types. For example, Wanderlust constructs K-nearest neighbor (KNN) graphs based on cosine similarity between cells and draws shortest path on the graph to obtain a single linear trajectory (*129*). The algorithm was applied to model the B cell development and revealed a rare B cell precursor marked by its unique pSTAT5 response to the cytokine IL-7. However, for complex developmental systems, algorithms that allow branching and cyclic trajectories may be preferred, as cell cycle and fate bifurcation occur recurrently throughout development. One of the earliest trajectory inference method uses the PQ-tree algorithm to model the time-series microarray data such as the cell cycle of bacteria (*130*), and was later adapted for single cell data (*131*). Many other graph-based or tree-based algorithms for modeling differentiation trajectory have been developed in recent years (*124, 132-134*) and have been extensively reviewed by Saelens *et al.* (*135*).

Like single cell clustering, trajectory inference methods often rely on dimensionality reduction techniques to obtain a low dimensional representation of the "manifold" of transcriptomic states. Besides the widely used PCA, t-SNE and UMAP, diffusion map (*136*) is another popular method for modeling differentiation as a diffusion-like dynamical process and has been successfully applied to complex developmental systems such as the differentiation hierarchy of HSCs (*137*). A flexible trajectory inference method, Slingshot (*138*), allows user to input custom dimensionality reduction results and use principal curve analysis to fit linear or bifurcating trajectories. Slingshot is one of the top ranked methods in terms of accuracy, stability and usability based on a recent comparison of trajectory inference methods (*135*). Therefore, it is used to reconstruct the developmental trajectory of HSCs in this thesis, as described in Chapter 3.

With the inferred trajectory, “pseudo-time” can be assigned for each cell to reflect its relative position along the differentiation trajectory. This enables differential expression analysis to derive temporally differentially expressed genes that may be responsible for cell fate specification. For example, one can apply the switchde algorithm to find switch-like genes along single-cell trajectories (*139*), or use the BEAM method to discover genes that are differentially activated at branching points (*132*).

Trajectory inference methods connect cells across developmental stages but cannot determine the direction and rate of differentiation. To model the velocity of transcriptome change during differentiation, the Velocyto algorithm exploits the RNA splicing dynamics (*140*). When a gene is up-regulated, transcription initiation produces large amount of unspliced mRNAs. The immature mRNAs undergo alternative splicing and degradation, which brings the system to a steady state. Conversely, when a gene’s expression is repressed, reduction of unspliced mRNA precedes the downregulation of spliced mRNA, leading to a fast drop of unspliced versus spliced mRNA ratio. The RNA dynamics can be modeled with a set of ordinary differential equations (ODEs), solving which gives estimation of the rate of transcriptome change, or “RNA velocity”, for each cell. Using the RNA velocity, one can extrapolate the future state of the cell, and project that onto a dimensionality reduction plot. The projected velocity field enables automatic identification of start and end point of differentiation and allows investigation of the cell fate choices at bifurcating points. With RNA velocity estimated for the developing mouse hippocampus, the authors identified radial glia cells as the root of the lineage tree of the hippocampus and observed fate biases before the trajectory branches into CA and granule fate. One limitation of Velocyto is that it assumes the induction and repression of gene expression last long enough to reach steady state, which is often not true for transient cell populations in development. Therefore, a recent method scVelo was developed to address this limitation by solving the full gene-wise transcriptional dynamics (*141*). The concept was further extended to account for the protein translation kinetics, such that protein velocity and acceleration can be estimated with joint profiling of proteins and RNAs (*142*).

Lastly, during development, cells may have multiple differentiation potentials before committing to a terminal fate. For example, scRNA-Seq and lineage tracing experiment reveals that the hematopoietic stem and progenitor cells (HSPCs) exhibit a continuum of transcriptome states with different fate biases (*143*). To computationally resolve the fate potential of early progenitor cells, several algorithms, such as FateID (*144*), Waddington-OT (*145*), and population balance analysis (*146*) have been developed. Weinreb *et al.* benchmarked these methods using the HSPC lineage tracing data and observed all three methods were able to resolve lineage potential for late-stage cells but performed poorly on early-stage progenitors (*143*). Nevertheless, these algorithms, combined with lineage tracing, provides important information about fate specification mechanisms in early progenitors. By comparing the progenitors with different fate biases, factors that drive fate choices can be derived and functionally tested.

Identifying transcriptional regulatory network

Transcriptional regulatory network (TRN) describes the regulatory relationship between genes and provides systematic understanding of transcriptional regulation. However, multiple challenges exist for the robust reconstruction of TRN with single cell data. For example, currently, the state-of-art TRN algorithms for bulk gene expression data has a precision of ~50% for prokaryotes, and performs poorly for eukaryotic organisms (*147*). The technical noise and high dropout rate of single cell data adds to the variance, so the performance of these bulk algorithms, if applied directly to single cell data, would be even worse.

Despite of these caveats, single cell data presents unique opportunities for accurate and robust inference of TRN. First, gene correlation analysis using single cell data yields much more accurate estimation of co-expression relationship compared to bulk methods, as the latter may suffer from Simpson's Paradox (*148*). With single cell clustering analysis, individual cell types can be identified and cell-type-specific TRNs can be constructed. Second, trajectory inference and velocity analysis give temporal ordering to the gene expression profiles, making causal inference of gene regulatory relationship possible. Finally, other single cell data modalities such as scATAC-Seq provide important information about active DNA regulatory elements, which can be used to infer the *cis*- and *trans*-regulatory relationship. Indeed, many recently developed single cell network analysis methods exploit these unique properties of single cell data and have been successfully applied to several developmental systems (*149-153*).

In one of the earliest single cell study of blood stem and progenitor cells, cell-type-specific TRN was constructed using expression profiles of 18 transcription factors with known important roles in hematopoiesis (*154*). The authors observed that although most regulatory connections remain stable across cell types, some show clear differences. For example, strong negative correlation between *Gfi1* and *Gata2* was only observed in HSCs, and the correlation between *Gata1* and *PU.1* was strongly dependent on the cell type, consistent with their switch-like function in controlling erythroid and myelomonocytic fates (*155, 156*). Several single cell TRN inference methods, such as LEAP (*149*), SINCERITIES (*150*), SCODE (*151*) and SCRIBE (*152*), explicitly uses time information in their model. SINCERITIES first computes the Kolmogorov-Smirnov (KS) distance to measure the differences in marginal gene expression distributions between two consecutive time points (*150*). Regulatory connections were then inferred using Granger causality, where the changes in transcription factor (TF) expression were used to predict changes in its target genes. SINCERITIES were benchmarked using time-stamped single cell data of monocytic THP-1 human myeloid leukemia cell differentiation and gave much better predictions compared to methods that do not use time information. To further improve the accuracy of TRN inference, several algorithms utilize TF motif enrichment in active regulatory DNA elements. For example, the SCENIC method (*157*) first infers TRN using expression-based methods, GRNBoost (*157*) and GENIE3 (*158*), and then refines the TRN using motif enrichment analysis on promoters of co-expressed genes. A recently developed method, CellOracle (*153*), uses scATAC-Seq data and motif enrichment analysis to assemble a “base” TRN, and then uses scRNA-Seq data to convert the base TRN into cell-type specific TRNs. CellOracle was used to map the network structure during hematopoiesis and correctly predicts the effect of *Gata1* knock-out on myeloid cell identity. Finally, it is worth noting that a newly emerged single cell technology, Perturb-Seq (*159, 160*), enables profiling of thousands of cells with genome-scale CRISPR perturbations, therefore presenting a new opportunity for causal TRN inference. So far, only a few TRN inference methods have been developed for this type of data (*159, 161*).

Integrative analysis of single cell transcriptome and epigenome

The fast development of single cell technology has enabled profiling of various types of molecular features at single cell resolution. The collection of multimodal single cell data can be performed separately using biological replicates, or parallelly on one sample with true multi-omics technologies (*65-68*). In the first case, both features and cells are different across modalities, while in the second case, various types of features are measured for the same set of cells. Statistical challenges for each of these cases are different. At the time of the study described in this thesis, only the first type of technology is commercially available. Therefore, I will focus on reviewing integrative analysis methods for separately collected scRNA and scATAC data. However, it is worth noting that several methods have been recently developed for integrative analysis of true multi-omics datasets (*162, 163*).

One important challenge of integrative analysis of transcriptome and epigenome data is that the correspondence between the features is unclear. For example, the association between chromatin accessibility and gene expression is not completely understood, thus there is no simple transformation that maps cells from gene expression space to chromatin accessibility space, and *vice versa*. However, such relationship can be learned if cells collected for both modalities have the same underlying distributions of molecular states, which can be achieved through collection of biological replicates. The MATCHER algorithm makes a simple assumption that single cell measurements are made from cells changing unidirectionally along a one-dimensional manifold, and for each data modality cells are sampled from the same population, process, and cell type (*164*). Given these assumptions, MATCHR performs manifold alignment to project cells of different data modalities onto a shared 1D pseudo-time space. The authors applied MATCHR to jointly analyze gene expression and DNA methylation changes during human induced pluripotent stem cell (iPSC) reprogramming and observed an overall trend that DNA methylation changes lag behind gene expression changes.

Several integration algorithms do not have strong assumption on the latent manifold structure, but make certain assumptions about feature correspondence. For example, LIGER leverages the well-established negative relationship between gene-body methylation and expression to integrate single cell methylome and transcriptome of mouse cortical cells (*165, 166*). Joint embedding of the two data modalities were learned using integrative nonnegative matrix factorization and revealed multiple cell populations that could not be identified using the methylation data alone. The Seurat v3 algorithm builds upon a gene activity matrix, which can be computed by summing the scATAC-Seq fragments in the gene body and promoter region (*167*). Canonical correlation analysis (CCA) is applied to the gene activity matrix and gene expression matrix to obtain a joint dimensionality reduction of both data modalities. Next, Mutual nearest neighbors (MNNs) are identified and used as “anchors” for computing a set of correction vectors. Subtracting these vectors from one dataset allows the two datasets to be merged and jointly analyzed. GLUER (*168*) and scDART (*169*) also requires a pre-defined gene activity matrix and utilizes the deep learning framework to learn a joint latent representation of the cells.

The technological development presents a unique opportunity for investigating the association between transcriptome and epigenome during development and differentiation. For example, Cao *et al.* (*65*) used a linear regression model to predict gene expression from chromatin accessibility data and found that accounting for accessibility at distal sites improved prediction by four-fold compared to using promoter accessibility alone. One limitation of such method is that it can only be applied to true multi-omics dataset. In Chapter 3, I developed a computational framework for joint analysis of paired scRNA-Seq and scATAC-Seq data. The method first learns feature correspondence from the data by matching differentially accessible peaks with differentially expressed genes, then uses Seurat v3 to obtain a joint embedding of scRNA and scATAC cells. Using paired meta-cells, I performed linear regression to identify enhancer-promoter links that are important for the endothelial to hematopoietic transition.

Figures

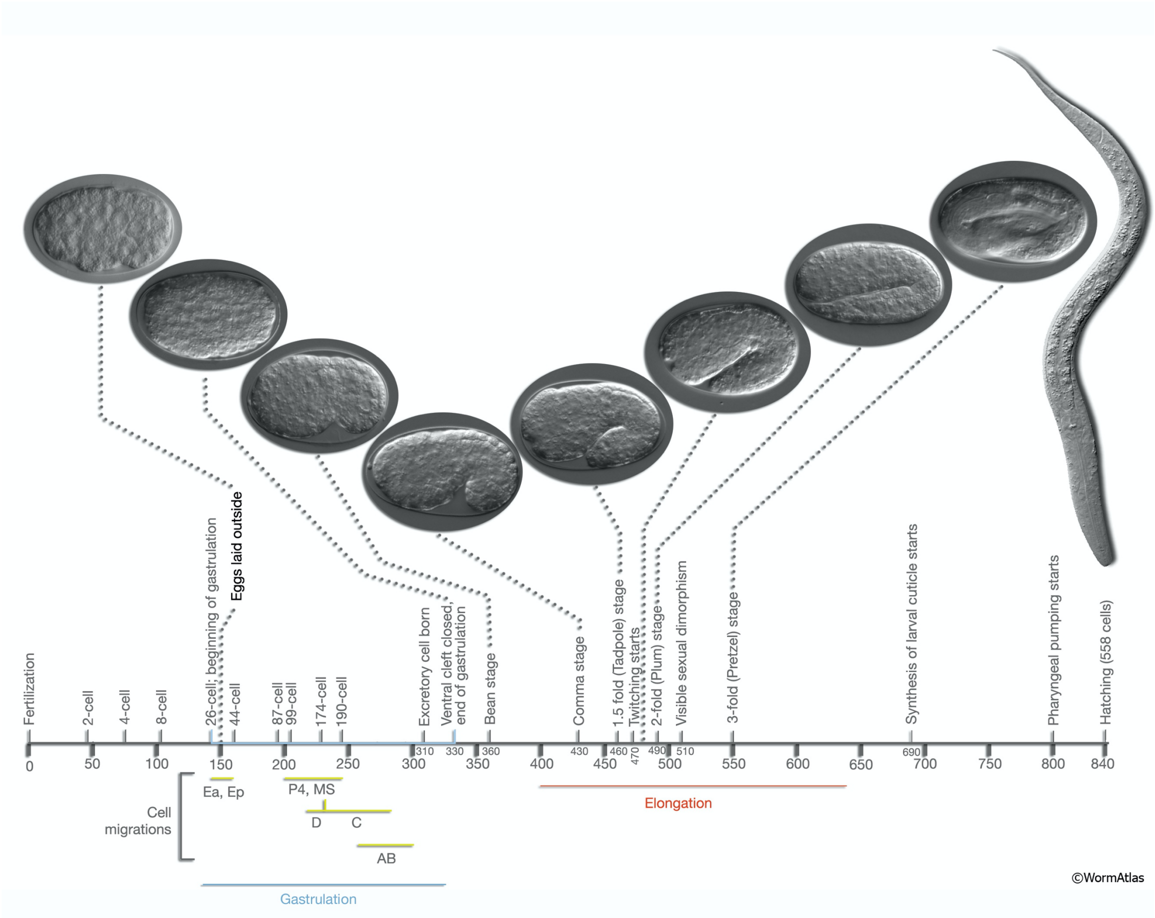


Figure 1.1 *C. elegans* embryogenesis.

Reprinted from WormAtlas (*9*). Horizontal axis shows approximate time in minutes after fertilization at 20-22°C. The stages, number of nuclei, marker events and DIC images of the embryos and larva are shown above the axis. Yellow bars indicate period of early cell migration. Blue bar shows gastrulation period (between 270 and 330 minutes). Red bar indicates elongation of the embryo between 400 and 640 minutes.

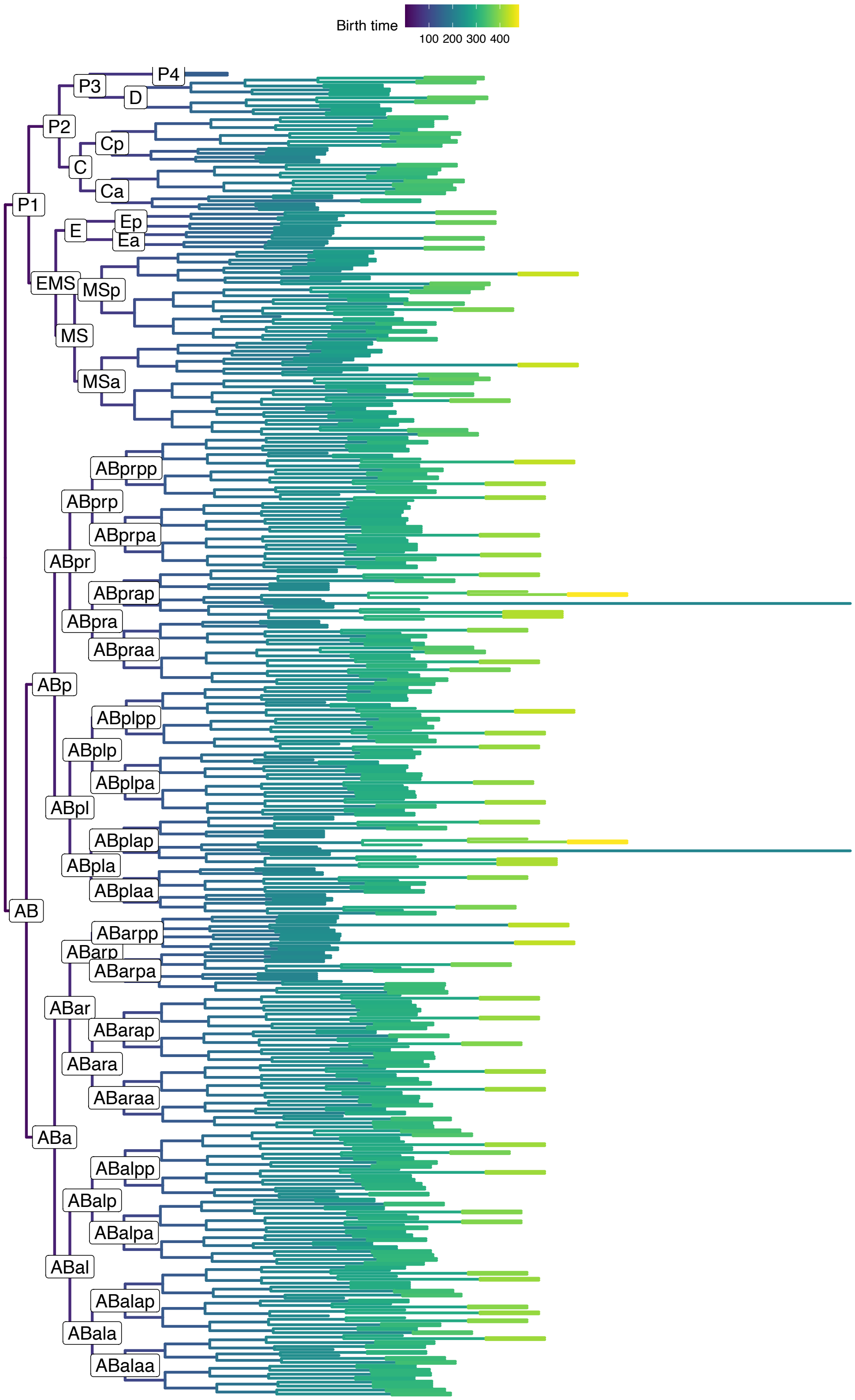

Birth time
100 200 300 400
P4
P3
D
P2
Cp
C
Ca
P1
Ep
E
Ea
EMS
MSp
MS
MSa
ABprpp
ABprp
ABprpa
ABpr
ABprap
ABpra
ABpraa
ABp
ABplpp
ABplp
ABplpa
ABpl
ABplap
ABpla
ABplaa
AB
ABarpp
ABarp
ABarpa
ABar
ABarap
ABara
ABaraa
ABa
ABalpp
ABalp
ABalpa
ABal
ABalap
ABala
ABalaa

Figure 1.2 Cell lineage tree of *C. elegans* embryogenesis.

Only cells with birth time before 500 mins are shown. Lineage tree reproduced from WormAtlas (*9*) and Sulston *et al.* (*7*) using VisCello.celegans (*170*). Length of each edge is proportional to the lifespan of the cell, and the color represents the birth time of the cell.

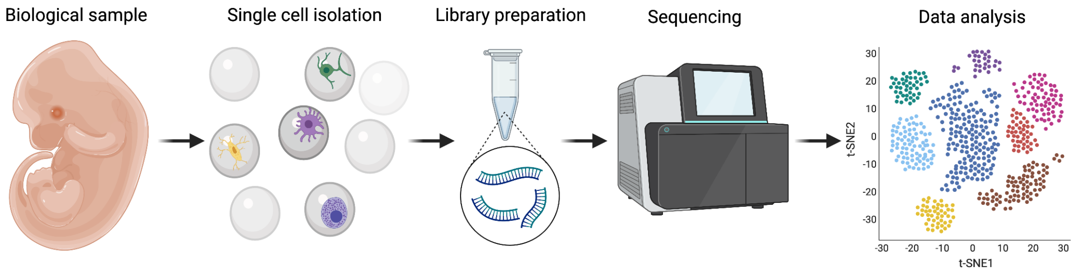


Figure 1.3 Single-cell RNA sequencing.

Created with BioRender.com. Biological samples are harvested and are dissociated into a single cell suspension. Single cells are then isolated with a cell isolating device. Once cells are isolated and lysed, sequencing library is prepared by reverse transcription and cDNA amplification. After sequencing, the reads are aligned and quantified for downstream data analysis.

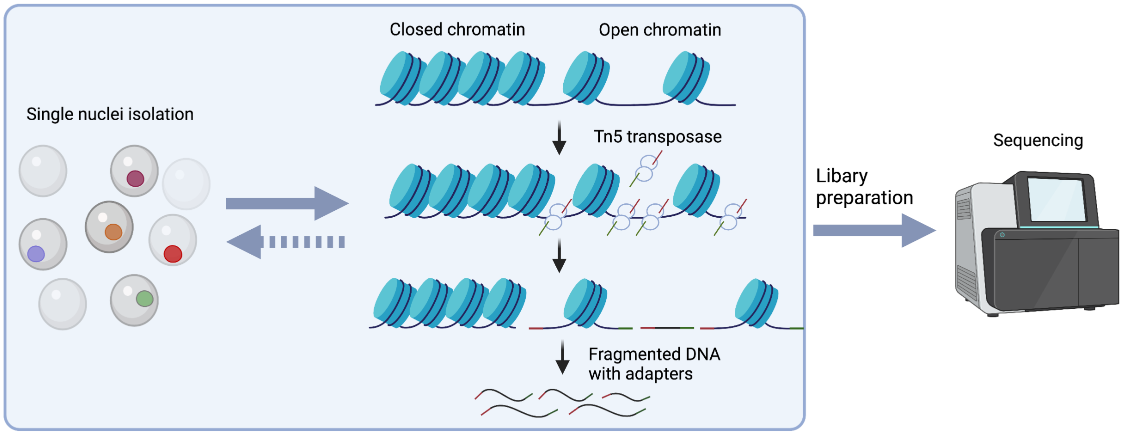


Figure 1.4 Single-cell ATAC sequencing.

Created with BioRender.com. Depending on the protocol, the order of single nuclei isolation and tagmentation can be reversed. Single nuclei isolation can be performed using methods such as microfluidics technology. Tagmentation involves using Tn5 transposase to insert sequencing adaptors to open chromatin regions. The resulting DNA fragments are then amplified and indexed for sequencing.

Tables

Table 1.1 List of large-scale single cell datasets for developmental biology.

SPECIES	DEV. STAGE	TECHNOLOGY	CELL#	REFERENCE
Mus musculus	E6.5-8.5	10x Genomics	116,312	(*171*)
Mus musculus	E8.5-E9.5	10x Genomics	22,264	(*172*)
Mus musculus	E9.5-13.5	sci-RNA-seq3	2,072,011	(*59*)
Drosophila melanogaster	Embryo	Drop-seq	7,975	(*173*)
Xenopus tropicalis	Embryo	InDrop	136,966	(*174*)
Danio rerio (Zebrafish)	Embryo	inDrop	92,000	(*97*)
Danio rerio (Zebrafish)	Embryo	Drop-seq	38,731	(*175*)
Hydra	Adult	Drop-Seq	25,000	(*176*)
Schmidtea mediterranea	Adult	Drop-seq	21,612	(*177*)
Spongilla lacustris	Juvenile	10x Genomics	39,552	(*178*)
Schmidtea mediterranea	Adult	Drop-seq	66,783	(*179*)
Ciona intestinalis	Embryo	10x Genomics	90,579	(*180*)
Caenorhabditis elegans	Embryo	10x scRNA	86,024	*(170), this thesis*

CHAPTER 2 A LINEAGE-RESOLVED MOLECULAR ATLAS OF *C. ELEGANS* EMBRYOGENESIS AT SINGLE-CELL RESOLUTION

The contents of this chapter have been previously published as:

Packer JS*, **Zhu Q***, Huynh C, Sivaramakrishnan P, Preston E, Dueck H, Stefanik D, Tan K, Trapnell C, Kim J, Waterston RH, Murray JI. A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science*. 2019 Sep 20;365(6459):eaax1971.

* These authors contributed equally

Introduction

To understand how cell fates are specified during development, it is essential to know the temporal sequence of gene expression in cells during their trajectories from early uncommitted precursors to differentiated terminal cell types. Gene expression patterns near branch points in these developmental trajectories can help identify candidate regulators of cell fate decisions (*181*). Single cell RNA sequencing (sc-RNA-seq) has made it possible to obtain comprehensive measurements of gene expression in whole animals (*58, 177, 179, 182-184*) and embryos (*19, 59, 97, 171, 173-175*). sc-RNA-seq profiling of multiple developmental stages in a time series can be particularly informative, as algorithms can use the data to reconstruct the developmental trajectories followed by specific cell types. However, confounding factors can generate misleading trajectories. For example, progenitor cell populations with distinct lineage origins may be conflated if their transcriptomes are too similar, and abrupt changes in gene expression can result in discontinuous trajectories. Thus, information from independent assays is necessary to conclusively validate an inferred trajectory as an accurate model of development.

Here, we comprehensively reconstruct and validate developmental trajectories for the embryo of the nematode worm *Caenorhabditis elegans*. *C. elegans* develops through a known and invariant cell lineage from the fertilized egg to an adult hermaphrodite with 959 somatic cells (*6, 7*), which creates the potential for a truly comprehensive understanding of its development. Using sc-RNA-seq, the known *C. elegans* lineage, and imaging of fluorescent reporter genes (*17, 185*), we produce a lineage-resolved single cell atlas of embryonic development that includes trajectories for most individual cells in the organism. Our atlas expands on previous studies of the earliest embryonic blastomeres (*18, 19*), covering 87% of embryonic lineage branches.

We use this dataset to quantitatively model the relationship between the cell lineage and the temporal dynamics of gene expression. We find that during gastrulation, lineage distance between cells is a strong predictor of transcriptome dissimilarity. The strength of this correlation increases from the middle to the end of gastrulation. After gastrulation, expression patterns of closely related cells diverge as they adopt their terminal cell fates. Body wall muscle, hypodermis, and the intestine are exceptions to this trend, as they are produced by semi-clonal lineage clades that maintain within-clade transcriptomic similarity. In the ectoderm, the final two rounds of cell division produce distinct neuron and glia cell types, which rapidly differentiate, often resulting in discontinuities in computational reconstructions of their developmental trajectories. In several cases, the transcriptomes of distant lineages converge as they adopt the same terminal cell fate, and at the same time diverge from their close relatives in the lineage.

Our ability to reconstruct these complex gene expression dynamics highlights both the utility of the known *C. elegans* lineage and the challenges that will be faced when trying to use single cell RNA sequencing to reconstruct the lineages of other organisms.

Results

Single-cell RNA-seq of *C. elegans* embryos

We sequenced the transcriptomes of single cells from *C. elegans* embryos with the 10x Genomics platform. We assayed loosely synchronized embryos enriched for pre-terminal cells as well as embryos that had been allowed to develop for ~300, ~400, and ~500 minutes after the first cleavage of the fertilized egg. We processed the datasets with the Monocle software package (*132*). After quality control, the final integrated dataset contained 86,024 single cells, representing a more than 60x oversampling of the 1,341 branches in the *C. elegans* embryonic lineage.

We estimated the embryo stage of each cell by comparing its expression profile with a high-resolution whole-embryo RNA-seq time series (*186*) (Supplemental Figure 2.1). We then visualized the data with the Uniform Manifold Approximation and Projection (UMAP) (*116, 187*) algorithm, which projects the data into a low-dimensional space and is well suited for data with complex branching structures (*187*). We found that trajectories in the UMAP projection reflect a smooth progression of embryo time (Figure 2.1A), with cells collected from later time points usually occupying more peripheral positions (Figure 2.1B). Unique transcripts per cell, as estimated with Unique Molecular Identifiers (UMIs), decreased with increasing embryo time throughout the period of embryonic cell division, consistent with decreasing physical cell size (Supplemental Figure 2.2). These observations suggest that UMAP trajectories corresponded to developmental progression and that embryo time estimates are a reasonable proxy for developmental stage for most cells. Approximately 75% of the cells recovered (64,384 cells) were from embryos spanning 210-510 minutes post first cleavage, corresponding to mid-gastrulation (~190 cell stage) to terminal differentiation (3-fold stage of development) (Figure 2.1C); however, cells were also recovered from earlier embryos (< 210 minutes, 9,886 cells), and later embryos (> 510 minutes, 11,754 cells).

We clustered cells in the UMAP using the Louvain algorithm (*121*) and annotated clusters with cell type identities using marker genes from the literature on *C. elegans* gene expression (*188*). Markers used for each annotation are listed in Table 2.1. The global UMAP arranges cells into a central group of progenitor cells and branches corresponding to eight major tissues (Figure 2.1A, Supplemental Figure 2.3): muscle/mesoderm, epidermis, pharynx, ciliated neurons, non-ciliated neurons, glia/excretory cells, intestine, and germline. While some individual cell types were identifiable in this global UMAP, many were not, especially progenitor lineages. To gain resolution, we hierarchically created separate UMAPs of each tissue (Supplemental Figures 2.4-2.13)**.** These "sub-UMAPs" better resolved specific cell types, allowing us to make extensive, fine-grained annotations.

A combination of marker genes, lineage assignments, and developmental time allowed us to locate 112 specific terminal anatomical cell types, including every lineage input to body wall muscle, every distinct subtype of pharyngeal muscle (pm1-2, pm3-5, pm6, pm7, and pm8) and hypodermis (hyp1-2, hyp3, hyp4-6, hyp7, hyp8-11, seam, and P cells), and every non-neuronal cell type in the mesoderm. We identified 69 of 82 non-pharyngeal neuron types and 9 of 12 glial cell types present in the embryo. We could not identify 12 of 14 pharyngeal neuron types. A cluster corresponding to the most differentiated pm3-5 pharyngeal muscle cells had a low level of expression of neuron-specific genes, suggesting that we failed to dissociate the neurons that innervate these muscles in late embryos.

We successfully annotated 93% of cells in our dataset with a cell type (for terminal cells) or a cell lineage (for progenitor cells, discussed below) (Figure 2.1D). The number of cells annotated for each cell type was variable but roughly fit the expectation on the basis of the number of cells of that type present in a single embryo (Figure 2.1E, r = 0.64, p = 2.4e-13, t test).

Mapping single cells to known *C. elegans* cell lineage tree

The structure of the global and single-tissue UMAPs was dominated by trajectories of terminal cell differentiation. We hypothesized that closely related lineages could be better resolved by separately analyzing progenitor cells prior to terminal differentiation. Thus, we ran UMAP with only cells with embryo time <= 150, 250, or 300 minutes and found branching patterns that reflect lineage identities (Figure 2.2, Supplemental Figures 2.14-2.16). Intestine and germline cells commit to their terminal fates very early and have very divergent expression that distorts the projections, so they were removed and analyzed separately (Supplemental Figures 2.7, 2.12). The 300-minute UMAP contained several large quasi-connected groups corresponding approximately to major founding lineages, roughly organized by the major fates produced by each founder cell lineage (MS muscle, MS pharynx, C/D muscle and AB-derived lineages that produce either pharynx, neurons/glia, or hypodermis). We were able to resolve additional details by recursively making sub-UMAP projections of these cell subsets.

To annotate progenitor lineages, we exploited lineage marker genes from the literature and the EPiC database, which contains single cell resolution expression profiles extracted by cell tracking software from confocal movies of *C. elegans* embryos expressing fluorescent reporters (*17*). In addition to the 180 previously described patterns (*17, 189*), we have collected movies for 71 additional genes, increasing the total number of patterns in EPiC to 251 genes. We annotated branches with lineage identities between the 28-cell and 350-cell stages by finding genes that were differentially expressed both between sister lineages in the EPiC data and between branches of the sub-UMAP trajectories in a concordant manner (Figure 2.2, Supplemental Figures 2.14-2.16)**.** For example, expression of *ceh-51* is restricted to the MS (mesoderm-producing) lineage (*190*), allowing us to label the single group of *ceh-51(+)* cells in 150-minute UMAP as part of the MS lineage (Figure 2.2A, B). Within this lineage, we used expression of *pha-4* to annotate the anterior granddaughters of MS (MSaa and MSpa) and *hnd-1* to annotate the posterior granddaughters (MSap and MSpp) (Figure 2.2C). We applied this same logic iteratively across the different UMAPs and lineage marker genes to annotate each branch with its lineage identity.

In most cases, branches in the progenitor lineage UMAPs corresponded directly to sister cells in the lineage (Figure 2.2D, E), but some branches were unclear or misleading, and marker gene expression was critical to annotate lineages correctly. For example, ABpxpaaaa and ABpxpaapa are cousin lineages, but appear to branch as sisters in the UMAP trajectory, and the same is true for their sisters (ABpxpaaap and ABpxpaapp) (Figure 2.2D)**.** In other cases, such as the ABpxppap lineage (Figure 2.2D), marker gene combinations were required to annotate lineages that were not contiguous with their parent or sister lineages in the UMAP. These misleading branches demonstrate the importance of having independent expression or lineage data to correctly interpret trajectories visualized in low-dimensional embeddings of sc-RNA-seq data.

To complete our annotations, we used UMAPs of selected subsets of cells with embryo time <= 350 or 400 minutes to reconstruct trajectories leading from the grandparents and parents of terminal cells to their terminal descendants (Supplemental Figure 2.17). Most terminal cell types were thus identified by two methods: first using marker genes for the differentiated cell type, and second by following UMAP trajectories from the cell’s progenitors. Notably, in all cases, the cell type predictions of these two mostly independent methods were concordant.

In total, we annotated 502 distinct cell lineages. Most lineage annotations correspond to a symmetric pair of cells, with the exception of some terminal cell types in which 3-18 cells converge to a homogenous transcriptomic state and could not be further resolved. Our annotations account for 1,068 out of 1,228 individual branches in the *C. elegans* embryonic lineage (Supplemental Figure 2.18), excluding the 113 branches that lead to programmed cell death. Combined with the dataset of Tintori *et al.* (*19*), which profiles the 1- to 16-cell stages, we now have a near-complete molecular atlas of *C. elegans* embryogenesis.

The lineages included in our atlas partially overlap with the Tintori *et al.* dataset (*19*) at the 16-cell stage. Gene expression profiles for lineages annotated in both datasets were concordant (Supplemental Figure 2.19). Additionally, gene expression profiles for terminal cells in our data were concordant with previously published microarray data (*191*) (Supplemental Figure 2.20).

Bifurcating cell fates and multi-lineage priming

Developmental trajectories in which a parent cell divides to produce two terminal daughter cells of different cell types are a basic type of cell fate decision. Bifurcations like these are common in neuronal lineages in *C. elegans*, such as those that produce ciliated neurons. To examine the molecular basis for such developmental decisions, we used recursive UMAP projections of ciliated neurons (Figure 2.3A) to identify developmental trajectories for all but one of the 22 ciliated neuron types and their parents, missing only the PHA phasmid neurons. The distinction between neuroblasts and terminal neurons was supported by embryo time estimates consistent with terminal cell division times (*192*), by the expression patterns of cell cycle associated genes and transcription factors (Figure 2.3B), and by the structure of the UMAP projection. A 3D version of the UMAP featured better continuity for several trajectories, including those connecting the ASG-AWA, ADF-AWB, and ASJ-AUA neuroblasts with their daughter cells, as well as the branching of the laterally asymmetric left and right ASE neurons (Supplemental Figure 2.21).

To identify potential regulators of cell fate decisions, we identified genes that were differentially expressed between the branches of each bifurcating ciliated neuron lineage. The lineage of the ASE, ASJ, and AUA neurons (spanning embryo time ~215-650 minutes) serves as a representative example (Figure 2.3C). About 3-4 TFs are specific to each terminal neuron type in this lineage (Figure 2.3D). Similar numbers of branch-specific TFs were observed for other lineage bifurcations (Supplemental Figure 2.22). Beyond these simple cases, we also found several TFs that were expressed in a parent cell and had expression selectively maintained in one daughter but not the other. For example, the TFs *ceh-36/37/43/45*, *ham-1*, and *hlh-3* are all co-expressed within single ASE-ASJ-AUA neuroblast cells. *ceh-36/37* and *hlh-3* expression was maintained in only one daughter of this neuroblast, the ASE parent, while *ceh-43/45* and *ham-1* expression was maintained only in the other daughter, the ASJ-AUA neuroblast (Supplemental Figure 2.23).

This pattern, where a progenitor cell co-expresses genes specific to each of its daughters, has been termed “multilineage priming” and has been observed in several organisms and developmental contexts (*174, 193-198*). Our transcriptomic atlas of the *C. elegans* cell lineage allows us to provide an unbiased quantification of the prevalence of multilineage priming throughout the organism’s ectoderm and mesoderm (we lack sufficient resolution in our annotations of the endoderm, which produces only one cell type, the intestine). There are 172 instances in which we have data for a parent cell and both of its distinct daughters. Of these, 52% exhibit multilineage priming. Multilineage priming events are distributed throughout several generations of both the ectoderm and mesoderm (Supplemental Figure 2.24), demonstrating that it is a common and pervasive mechanism of gene regulation. The expression patterns of many TFs involved in multilineage priming, e.g. *hlh-3* (Supplemental Figure 2.23D), are confirmed by the movies in EPiC (*17*).

Transcription factors that are both required for neuron type specification and have expression maintained throughout the lifetime of the neuron are referred to as “terminal selectors” (*198*). To identify potential terminal selectors, we looked for transcription factors that were 1) expressed in a neuron type but not its sister in the embryo and 2) expressed in the same neuron type at the L2 stage. This analysis replicated 23 known neuron-TF associations (*198*) and identified 116 novel associations. Other known associations may have been missed due to the extreme sparsity of the L2 stage data, and the fact that many terminal selectors are expressed at low levels in fully differentiated neurons, or are expressed in both daughters of a terminal division. In cases where a neuron’s sister undergoes programmed cell death, we looked for TFs that are both enriched in the terminal cell’s most recent ancestor that has a surviving sister cell (compared to that sister), and also have expression maintained throughout the lifespan of the terminal neuron. This revealed novel associations, including *ceh-6* for AVH, *ceh-8* for RIA, *unc-62* for RIC, and *lin-11* for RIC and RIM, in which the putative terminal selector TF is expressed in a neuroblast before the terminal cell is produced, suggesting that these lineages commit to a cell fate early.

Only two neurons (ASE and AWC) are known to have left-right asymmetric gene expression (*199, 200*). For both neuron types, the lineages of the left and right neurons diverge in the early embryo at the 4-cell stage (< 50 minutes). Asymmetric gene expression in our data, however, emerges only much later in embryogenesis. The transcriptomes of ASEL and ASER diverged in our UMAP at ~650-700 minutes, with *lim-6* expressed specifically in the ASEL branch, consistent with previous studies (*201, 202*). AWC left/right asymmetry occurs stochastically, with one neuron becoming “AWC-ON” and the other becoming “AWC-OFF” (*200*). We identified a small cluster in the UMAP with embryo time >700 minutes as AWC-ON based on *srt-28* expression (Figure 2.3A) (*203*). AWC-OFF is putatively part of the main AWC trajectory. No evidence of left/right asymmetry was observed in neurons besides ASE and AWC.

Transcriptional convergence of co-fated lineages

While most bilaterally symmetric cells were not distinguishable by UMAP (as expected), several cell types with >2-fold symmetry are produced by multiple non-symmetric lineage inputs. These lineage inputs tended to cluster separately in our progenitor cell UMAPs, while in our late-cell tissue UMAPs, we saw almost no evidence of heterogeneity within the terminal cell types that they produce. This difference suggested that the transcriptomes of these co-fated lineages were converging during differentiation.

One example of apparent molecular convergence of cells from distinct lineages was the IL1-IL2 neuroblasts. The six IL1 and six IL2 neurons are produced by three symmetric pairs of neuroblast lineages. Each neuroblast pair produces a pair of bilaterally symmetric IL1 neurons, and likewise a pair of IL2 neurons. A UMAP of IL1/2 neurons and progenitors revealed trajectories for these neuroblasts that converge gradually over their lifespan (Figure 2.4A). The transcription factor *ast-1* was transiently expressed at extremely high levels (>10,000 TPM) during this process, suggesting that it might play a role in homogenizing the IL1/2 neuroblast transcriptomes (Figure 2.4B). Correspondingly, expression of genes differentially expressed between the input lineages decreased over time, while expression of genes specific to terminal neurons increased (Figure 2.4C-D). We observed similar lineage convergence via continuous gene expression trajectories for other cell types, including hypodermis (Supplemental Figure 2.8), head body wall muscle (Supplemental Figure 2.17), and GLR cells (Supplemental Figure 2.17).

Like the IL1/2 neurons, IL socket glia (ILso) are produced by three symmetric pairs of lineages. In contrast to the examples discussed above, trajectories formed by the ILso progenitors and their terminal descendants were discontinuous in UMAP space (Supplemental Figure 2.25). Discontinuous trajectories were also observed for several other cell types from multiple tissues, including other glia, several neuron types, the excretory gland, coelomocytes, and somatic gonad precursors (Z1/Z4) (Supplemental Figure 2.25). Several lines of evidence suggest that these discontinuities reflect sudden changes in the transcriptome rather than technical artifacts of sc-RNA-seq or UMAP. Discontinuous trajectories had more genes differentially expressed between the parent and daughter cells than continuous trajectories (Supplemental Figure 2.26). Almost all discontinuous trajectories were observed in lineages where a parent cell gives rise to two daughters of different broadly-defined cell types, e.g. a glia and a non-glial cell, or a ciliated neuron and a non-ciliated neuron (Supplemental Figure 2.26). These discontinuities were seen in both the global and the tissue-specific UMAPs, and with different UMAP parameters. Finally, for most discontinuous trajectories, cells had a continuous distribution of embryo times (Supplemental Figure 2.27). However, a few trajectories, such as that of the BAG neuron, had gaps in the embryo time distribution indicative of potential sampling bias.

Body wall muscle (BWM) was exceptional in that lineage-related heterogeneity persisted throughout differentiation. BWM is produced by multiple distinct lineages (C, D, MS) and occupies a wide range of positions along the anterior-posterior (A-P) axis of the animal. A UMAP of BWM cells identified distinct trajectories for the 1st row of head BWM vs. all other BWM (Figure 2.4E). The non-1st-row trajectory was formed by input trajectories that corresponded to lineages and progressed in parallel along the temporal axis. Using marker genes that are expressed in domains along the A-P axis (*17, 204-206*), we divided BWM cells in the UMAP into six "bands" (Figure 2.4E) and identified the specific anatomical cells present in each band (Figure 2.4F). We found that the Jensen-Shannon (JS) distance, a measure of transcriptome difference, between the transcriptomes of posterior BWM (C lineage) vs. both the 1st and 2nd rows of BWM (D/MS lineage) did not decrease over time (Figure 2.4G), indicating that BWM heterogeneity persists throughout differentiation.

Temporal dynamics of the lineage-transcriptome relationship

The presence of discontinuities between progenitor cells and terminal cells in the UMAP projections suggested that the terminal division could mark a shift from lineage-correlated to fate-correlated gene expression. We asked how well the distance between two cells in the lineage predicts the difference between their transcriptomes (as defined with the JS distance). We focused on the AB lineage, which produces mostly ectoderm and accounts for ~70% of the terminal cells in the embryo. The AB lineage undergoes roughly synchronized cell divisions, allowing us to group cells by generation. For example, we refer to the 32 cells produced by 5 divisions of AB as "AB5" and so on.

In AB5 (early/mid-gastrulation; 50-cell stage), the earliest stage where our lineage annotations were near-complete, sister cells were more similar than distant relatives, but the difference was not large (Figure 2.5A). In AB6 (mid-gastrulation; 100-cell stage) and AB7 (late gastrulation; 200-cell stage), the transcriptomes of sister cells become more similar than in AB5, while those of distant relatives become more divergent, resulting in a strong correlation between transcriptome distance and lineage distance. In AB8 (350-cell stage), most epidermal cells exit the cell cycle and begin terminal differentiation, while neuron/glia progenitors continue for 1-2 more cell divisions. AB8 thus features a bimodal distribution of transcriptome JS distances: terminal epidermal cells become highly distinct from neuron/glia progenitors, but cells within each group are more similar (Supplemental Figure 2.28). Finally, most neuron/glia progenitors in AB8 produce two terminal daughters in AB9 that have distinct cell fates and a much weaker lineage-transcriptome correlation than in earlier generations.

Together, these statistics suggest that progenitor cells develop strong expression signatures of their lineage identity, and that these signatures are rapidly lost or overshadowed by new expression at the time of the terminal division. An analysis of cells from the mesoderm (MS lineage) replicated the trends observed in the ectoderm (Supplemental Figure 2.29A).

To summarize the strength of the lineage-transcriptome correlation in a cell generation as a single number, we developed a statistic analogous to the concept of pseudo-R^2 in generalized linear regression models. Consistent with the above analysis, we find that the extent to which lineage predicts the transcriptome increases throughout gastrulation, peaks at 55% in AB7, and then falls to 18% after terminal differentiation in AB9 (Figure 2.5B). Next, we asked how much of the total pseudo-R^2 for one cell generation was attributable to gene expression signatures associated with each preceding cell generation. For cells in AB5-8, the largest contributor to pseudo-R^2 was the identity of their ancestor in the AB3 generation (Supplemental Figure 2.30). This is interesting because many of the clades formed at AB3 share a broadly-defined tissue fate. For example, the clade founded by the cell ABala produces only neurons and glia, while the clade founded by the cell ABarp produces mostly (but not exclusively) epidermal cells. The second largest lineage signal was from the identity of a cell's parent in the preceding generation (i.e. the tendency of sister cells to be more similar than cousins). Thus, both broad and fine-grained structure in the lineage contribute towards shaping the transcriptome.

To investigate the potential regulatory mechanisms that differentiate sister cells, we identified transcription factors (TFs) that distinguish each cell in AB5-9 from its sister. The median number of these “lineage signature TFs” per cell increased over time, ranging from 1.5 in AB5 to 14 in AB9 (Figure 2.5C). A substantial number of lineage signature TFs (~40-50%) had expression selectively maintained in only one of a cell's two daughters (Figure 2.5D). In other words, TFs that distinguish a cell from its sister in one generation are frequently re-used to distinguish that cell's daughters from each other. Sister cells are also differentiated by the expression of new TFs not present in their parents. The proportion of lineage signature TFs that are newly expressed ranged from 33-61% and increased over time in AB6-9 (Figure 2.5E). Temporal dynamics of lineage signature TFs were similar in the mesoderm (Supplemental Figure 2.29).

Taken together, these results highlight the incremental nature of cell fate decisions: every terminal cell is the result of a series of lineage bifurcations, each of which, on average, involves multiple differentially expressed TFs.

Global patterns of gene expression and transcriptome specialization

Hierarchical clustering of expression levels in all annotated lineages and cell types provides a global view of expression dynamics for all genes in our dataset. A heatmap of pre-terminal lineage expression profiles (Supplemental Figure 2.31) does not reveal large clusters of genes specific to specific lineages, other than one cluster of genes specific to the early C and D lineages. Similarly, most marker genes used for lineage annotation are not part of large clusters of co-expressed genes. The clusters that do form are composed of early tissue-specific genes. The lack of cluster structure in the heatmap suggests that differential fates for tissue sub-lineages are specified by relatively small sets of genes. By contrast, a heatmap of terminal cell type expression profiles (Supplemental Figure 2.32) has more obvious structure. Cells in each major tissue express ~500-1500 tissue-enriched genes. There is little reuse of tissue-enriched genes between tissues other than hypodermis, which shares many genes with glia and intestine. Neuron subtypes and other specialized cells (such as the hmc or M cell) are typically distinguished from other cells within their tissue by expression of <20-300 genes. Finally, there are substantial temporal changes in expression, especially in muscle and hypodermis.

We observed substantial variation between cells in the Gini coefficient, which measures how unequally different genes are expressed in a given cell type (Supplemental Figure 2.33A). Hypodermis, seam cells, and the pharyngeal gland express small sets of cell type specific genes at very high levels (high Gini coefficient), while the intestine and germline feature diverse gene expression patterns (low Gini coefficient). In several cell types, such as the pharyngeal gland, increases in Gini coefficient over time coincide with decreases in the number of TFs expressed per cell (Supplemental Figure 2.33B). Families of TFs also exhibit differential expression patterns over time and across lineages. Nuclear hormone receptors (NHRs) are on average activated later in development than other TF families, such as Forkhead and Homeodomain TFs (Supplemental Figure 2.33C). Hypodermis and intestine express many distinct NHRs, while expression of Sox family TFs is largely restricted to neurons, glia and pharynx (Supplemental Figure 2.33D).

Discussion

The cells of *C. elegans* are limited in number and invariant in lineage and cell fate, making it feasible to conduct comprehensive, whole-organism investigations. Yet within this limited repertoire of cells exists an impressive diversity of cell types, which work together to produce complex anatomical structures and behaviors. This study and our previous work (*17, 58*) have shed light on the molecular basis for the specification of these cell types, but are only the first step toward a comprehensive understanding of the molecular basis of development. We hope that this resource will help guide future projects in the *C. elegans* community.

In contrast to developmental sc-RNA-seq datasets from other species, this dataset links gene expression trajectories to the exact cell lineages they correspond to, allowing steps in the process of differentiation to be associated with specific cell division events. Thus, our data provide a quantitative portrait of Waddington's landscape (*207*) for a whole organism. The abruptness of many cell fate decisions in *C. elegans*, with many distinct terminal cell types becoming distinguished only in the final embryonic cell division, contrasts, however, with the smooth landscape in Waddington's illustrations and warrants further investigation.

We observe convergence of gene expression patterns in many instances where distinct cell lineages produce identical or related cell types. Data from a recent atlas of mouse organogenesis (*59*) suggests that this phenomenon is also prevalent in vertebrates. For example, myocytes in the mouse atlas are produced by two convergent trajectories, and excitatory neurons are produced by several trajectories.

Our analysis highlights two important challenges that will be faced by efforts to reconstruct the cell lineages of other organisms using single cell RNA-seq. First is the difficulty of accurately connecting developmental trajectories that start after the convergence of lineages with similar cell fates to trajectories that span earlier stages of development. A naive interpretation of the UMAP projection of the full dataset (Figure 2.1A) could lead to inferred trajectories that are inconsistent with the correct lineage (for example, incorrectly concluding that hypodermis and seam cells are produced from a common ancestor that previously diverged from the progenitors of neurons). Second is the difficulty of constructing continuous trajectories for lineages that undergo abrupt changes in gene expression. In our data, progenitor cells that give rise to glia, excretory cells, and non-ciliated neurons were more often than not disconnected to their terminal daughters in UMAP space (Supplemental Figure 2.25, Supplemental Figure 2.26), reflecting the fact that many of these lineages only commit to a terminal fate after their final cell division.

Due to these challenges, we anticipate that constructing end-to-end trajectories of vertebrate organogenesis will require single cell RNA-seq to be integrated with experimental lineage tracing methods (*208*). It will also require improved computational methods that can model heterogeneity among poorly-differentiated progenitor cells and highly-differentiated cell types in an integrated manner.

Between this study, our previous study of the L2 stage (*58*), and earlier studies of the 1 to 16-cell stage embryos (*18, 19*), a large portion of the early *C. elegans* life-cycle has now been profiled by single cell transcriptomics. However, more datasets will be needed to complete missing stages, including other larval stages and the adult soma and germline. In the future, single cell profiling of different strains or species will be a useful approach to examine the evolution of cell types and their expression programs. All of these datasets will ideally be integrated into a single visualization platform, such as VisCello (*170*), to allow full tracking of cell trajectories from fertilization through the end of life. A greater challenge will be to discover the precise mechanisms that produce transcriptomic outputs. Single cell transcriptome analysis of mutants will likely need to be integrated with new single cell multi-omic technologies (*209*) to bring mechanistic studies to a whole-organism scale.

Materials and Methods

Sample preparation

To obtain a broad range of embryo ages, including early stages, roughly synchronized *C. elegans* adults (N2 strain) were obtained by releasing embryos with standard hypochlorite treatment and letting the L1 larvae hatch and undergo growth arrest on unseeded plates. Starved L1s were transferred to NGM plates seeded with *E. coli* OP50 bacteria. Embryos were released from these synchronized young adults using hypochlorite treatment followed by three washes with L15-10 media. To generate cell suspensions, embryos were then treated with 0.5 mg/ml chitinase at room temperature until the shells were dissolved (30-40 minutes at ~22 °C) followed by dissociation of the cells using a 3 ml syringe fitted with a 21 gauge 1¼ inch needle until >80% of embryos were disrupted. The cell suspension was then passed through a 10 μm filter, washed in phosphate buffered saline (PBS) and finally resuspended in PBS. An estimated 14,000 cells were loaded immediately onto a 10x Chromium instrument. The trypan blue negative viable cell count was estimated using a hemocytometer and was >84% for all samples.

To sample later stages more deeply, more tightly synchronized embryo populations (used for the 300-minute, 400-minute, and 500-minute time series shown in Figure 2.1B) were obtained through two cycles of bleaching adult worms (strain VC2010, a strain derived from N2 that has been completely sequenced). On the first round of synchronization, populations of mixed stage embryos recovered by hypochlorite treatment of mixed populations were hatched overnight in egg buffer (118 mM NaCl, 48 mM KCl, 3 mM CaCl2, 3 mM MgCl2, 5 mM HEPES pH 7.2) with gentle shaking. The hatched L1s were plated onto 150 mm peptone rich NGM plates seeded with *E. coli* NA22 at no more than 100,000 worms per plate. When worms reached the adult stage, the number of embryos inside the adults was monitored until most had about 4 embryos on each gonad arm. The adult worms were collected and treated with hypochlorite to release embryos. The embryos were again allowed to hatch in the absence of food at 20 °C for 12 hours yielding a more tightly synchronized population of L1 worms. Around 250,000 L1 larvae were plated onto four 100 mm petri plates seeded with NA22 bacteria and allowed to develop at 20 °C. As the worms reached the young adult stage, the population was closely monitored. When about 20-30% of the adults had a single embryo in either arm of the gonad, worms were subjected to hypochlorite treatment. The time hypochlorite was added to the worms was considered t = 0 (see Warner *et al.* (*210*) for typical age distributions). The capture time was taken as when the cells were loaded onto the 10x Chromium instrument. The embryos were allowed to develop in egg buffer until one hour prior to capture time. The embryos were collected by centrifugation, resuspended in 0.5 ml egg buffer and 1 ml chitinase (1 U/ml), and transferred to 30 mm petri dishes. The degradation of eggshell was monitored; after ~20 min (when about half the eggs had lost the shell), the suspension was transferred to a 15 ml falcon tube and centrifuged at 200 g for 5 min. The chitinase solution was aspirated; a solution of 200 ul pronase (15 mg/ml) together with 0.5 ml egg buffer was added to the embryo pellet. The vitelline membrane was disrupted and the cells released by repeated passage through 21 gauge 1¼ inch needle attached to a 1 ml syringe. When sufficient single cells were observed, the reaction was stopped by adding 1 ml of egg buffer containing 1% BSA. Cells were separated from intact embryos by centrifuging the pronase treated embryos at 150 g for 5 min at 4 °C. The supernatant was transferred to a 1.5 ml microcentrifuge tube and centrifuged at 500 g for 5 min at 4 °C. The cell pellet was washed twice with egg-buffer containing 1% BSA.

Single cell capture and library preparation followed 10x Genomics published protocols. For each channel, 14,000 *C. elegans* cells were mixed with reverse transcriptase reaction solution and loaded immediately onto the capture chip to minimize the time that the cells spent in the reverse transcription cocktail. The exception was the first 500 minute sample, when three channels were loaded with 14,000, 4,666, and 1,555 cells respectively.

Read mapping and gene expression quantification

The single cell RNA-seq data was processed using the 10x Genomics CellRanger pipeline. Reads were mapped to the *C. elegans* reference transcriptome from WormBase (*188*), version WS260. We noticed that many 3' UTR annotations in the reference transcriptome were too short, causing genic reads to be called as intergenic, affecting gene expression quantification. To address this, we also mapped reads to modified versions of the WS260 transcriptome in which all 3' UTRs were extended by either 100, 200, 300, 400, or 500 bp (these 3' UTR extensions were cut short if the extended UTR would overlap with a downstream gene).

We then defined a set of criteria that specified for each gene whether it was beneficial to extend the 3' UTR for that gene, and if so, by how much. For each gene, we counted the number of reads across the entire dataset mapped to that gene for each version of the reference. We computed the ratio of the read counts from the 500 bp 3' UTR extended reference to the baseline reference. If this ratio was < 1.2, or if the total read count for the gene in the 500 bp 3’ UTR extended reference was < 20, we used the baseline 3’ UTR annotation for that gene. Otherwise, we used the shortest 3’ UTR extension (100, 200, 300, 400, or 500 bp) that gave at least 90% of the read count gain that was given by the 500 bp 3’ UTR extension.

We repeated this process with reads from our previous study on L2 worms (*58*). If a gene met our criteria for extending the 3’ UTR based on embryo reads, we used the extension length determined by the embryo reads. If a gene did not meet our criteria for extending the 3’ UTR based on embryo reads but did meet the criteria based on L2 stage reads, we used the extension length determined by the L2 stage reads. After deciding on how much to extend each gene’s 3’ UTR, we made a final reference transcriptome incorporating all of the per-gene 3’ UTR extension lengths. We then used this final reference transcriptome as input to the CellRanger pipeline to generate gene-by-cell UMI count matrices.

Criteria for distinguishing cells from empty droplets

The default barcode filtering algorithm in the 10x CellRanger pipeline can fail for experiments where the cells profiled are highly variable in size, resulting in a non-normal distribution of UMIs per cell. This is the case for our data. The total volume of the *C. elegans* embryo remains constant as cells divide within it, making cells of later generations smaller than those from earlier generations. Additionally, some cell types are more prone to damage and mRNA leakage than others. Neurons in particular usually have lower UMI counts than other cell types. To account for these factors, we manually set UMI count thresholds to distinguish cell barcodes from empty droplet barcodes on a sample-by-sample basis, based on the knee plots reported by CellRanger. The UMI count thresholds ranged for 700-1100.

While performing downstream analyses, we noticed that several neuronal, glial, rectal, and excretory cell types were missing from our data. We discovered that this was due to cells with extra low UMI counts (< 700 UMIs) being excluded by our UMI count thresholds. Lowering the UMI count threshold for all cells, however, would include low-quality, potentially damaged cells for other cell types where the average UMIs/cell is higher. To integrate the low-UMI count cells, we:

1. made a set of all cells with UMI count >= 500 (vs. the previous threshold of 700)
2. ran UMAP dimensionality reduction (described below) on this set of cells
3. identified clusters of cells corresponding to neurons (using the pan-neuronal marker genes *sbt-1* and *egl-21*) or glia, rectal, and excretory cells (using a variety of markers, see Table 2.1)
4. made new UMAPs from just neurons, just glia and excretory cells, or just rectal cells
5. filtered putative doublets (i.e. cells also expressing markers of non-neuronal cell types in the neuron UMAP, or cells also expressing markers of non-glia/hypodermal cell types in the glia UMAP)
6. made whitelists of the remaining cells

These whitelisted low-UMI count cells were then included when generating the final tissue UMAPs presented in this paper (Figure 2.3A**,** Supplemental Figures 2.9-2.11, 2.13)**.** They are not included in the original global UMAP (Figure 2.1A)**.**

Dimensionality reduction

For each dimensionality reduction (both for the global analysis of all cells and the tissue specific analyses), the first step was to perform PCA and adjust the PCA results to correct for batch effects. We performed PCA on the size-factor corrected, log transformed expression matrix, typically with 50-100 PCs depending on the dataset.

For batch effect correction, we noted that the predominant source of batch effects in our data appeared to be background contamination where RNA from lysed or damaged cells enters droplets in the 10x sc-RNA-seq apparatus that contain intact cells, causing each cell to receive reads from exogenous RNA. For each experimental sample, we computed the gene expression distribution of this background RNA by summing the read counts for cell barcodes that had < 50 UMIs, i.e. empty droplets. We transformed the background RNA count vector for each sample as if it were the count vector for a cell, and projected this vector into the PCA space computed from real cells. We then computed the dot product of each real cell PCA coordinate vector with each sample's background vector, calling this the "background loading" of a given cell for a given sample (each cell actually comes from exactly one sample, but computing each cell's loading for each sample's background made the next step mathematically/computationally simpler). Next, we fit a linear regression model, real cell PCA coordinate matrix ~ cell background loadings, and called its residuals the "background corrected PCA matrix." This background correction method is similar to, but developed independently of, a recently published method (*211*).

We found that the UMAP (*116, 187*) algorithm, which provides a way to project the data into a low-dimensional space, better maintains the topology of the dataset compared to the commonly used t-SNE algorithm. In our dataset, UMAP often creates long, continuous trajectories, while t-SNE clusters distinct cell types but does not clearly show the relationships between them. UMAP and t-SNE have been compared in the context of sc-RNA-seq by Becht *et al.* (*187*), but this paper focuses on the empirical performance of the algorithms and does not explain precisely how and why the mathematical differences between the algorithms underlie their qualitatively different results. We chose UMAP over t-SNE based on our subjective evaluation of how the two algorithms' results compared to our expectations given the known *C. elegans* lineage.

We reduced the dimensionality of the background corrected PCA matrix to 2 or 3 dimensions using UMAP, using the wrapper function for this algorithm provided by the Monocle software package, version 3 alpha (the reduceDimension function). The UMAP parameters were: metric = "cosine", min_dist = 0.1, n_neighbors = 20.

Lastly, cells in the UMAP space were clustered using the Louvain algorithm (*121*). The Louvain algorithm is one of several algorithms that group nodes in a weighted, undirected graph into clusters in a way that seeks to maximize a statistic called "modularity." Modularity is essentially the difference between the total edge weight between nodes assigned to the same cluster and the expectation of the total within-cluster edge weight if all edges were randomized. Exact optimization of modularity is computationally intractable for large graphs, so the Louvain algorithm uses a heuristic. In the context of our study, the graph used for the Louvain algorithm is a *k*-nearest neighbor graph (k = 20) constructed from cell coordinates in UMAP space.

Doublet identification

We used two complementary methods to identify doublets. The first method involved identifying clusters of doublets in iterated UMAP projections of the data on the basis of co-expression of high-confidence cell type specific marker genes, reported in WormBase (*188*), for >1 cell type (e.g. a cluster expressing the muscle markers *myo-3* and *pat-10* along with the neuron markers *egl-21* and *sbt-1* was considered a muscle-neuron doublet cluster). We applied this simple approach to a global UMAP of all cells and iterated UMAPs of tissues / related groups of cells from the global UMAP (e.g. muscle, intestine, ciliated neurons, *etc.*).

The second approach involved logistic regression models, one for each broadly-defined terminal cell type (e.g. body wall muscle, intestine, ciliated neurons, non-ciliated neurons, *etc.*), that predict whether a cell is part of that cell type or not. We fit one such model for each broadly-defined cell type and used the models to score each cell for the probability of it being a member of each broadly-defined cell type. Cells that had >= 2 cell types with a >= 20% predicted probability of the cell being a member of that cell type were considered doublets. Clusters in the UMAP projections that were enriched for cells considered doublets by these regression models were manually examined, and in some cases manually filtered.

Due to the abundance of cell type specific marker genes, we estimate that we were able to filter out almost all terminal cell type doublets. Residual expression of genes from one cell type in a cluster corresponding to another cell type appears to be driven by background RNA contamination, not doublets. Our approach is less likely to catch doublets between progenitor cells that do not yet express marker genes of differentiated terminal cell types. For earlier-stage embryos however, the cell dissociation protocol works more reliably than for late-stage embryos, so we expect the doublet rate to be close to the reported rate for the 10x Genomics Chromium platform, which is low (~4.5% given ~9k cells loaded per lane).

While performing downstream analyses, we noticed that a few cell types were missing from our data, including rectal epithelial and gland cells, the excretory duct and pore, and the T cell. These were erroneously excluded by our doublet filter due to co-expressing genes that were enriched in two or more tissues (e.g. co-expressing hypodermis-enriched genes with pharynx-enriched genes). We used marker genes to identify these cells in a non-doublet-filtered global UMAP, whitelisted them, and included them in the appropriate tissue UMAPs (Figure 2.3A, Supplemental Figures 2.9-2.11, 2.13). These cells are not included in the global UMAP (Figure 2.1A).

Embryo time estimation

For each cell, we estimated the age of the embryo that the cell came from (“embryo time”) based on Pearson correlation of its transcriptome with bulk RNA-seq time series data from Hashimshony *et al*. (*18*). Their data show that the majority of genes that change expression over time in any given lineage are not lineage specific. Thus, we first defined a list of genes with time-dependent expression patterns, requiring an auto-correlation greater than 0.6 and standard deviation greater than 1.5 across bulk RNA-seq time points (units = log TPM). Pearson correlation was then computed between log-scaled single cell and bulk data using only the time-dependent genes. We observed for non-multiplet cells, the Pearson correlation across time shows a strong peak pattern

(Supplemental Figure 2.1A). Thus, by fitting a Loess regression curve and finding its maximal point, we were able to assign each cell with its most correlated bulk time point.

Embryo times estimated based on data from Hashimshony *et al. (18)* approximately agree with embryo collection times from our experimental design (Supplemental Figure 2.1B), and also have a strong correlation with embryo times estimated based on data from Boeck *et al. (212)* (Supplemental Figure 2.1C). To further validate our embryo time estimates, we computed for each anatomical cell in the *C. elegans* embryonic lineage the 5th percentile of the embryo times for the set of sc-RNA-seq cells that we annotated as corresponding to that anatomical cell. This effectively estimates the birth time of the anatomical cell. These cell birth time estimates correlated well with cell birth time estimates derived from live imaging (*192*) (Supplemental Figure 2.1D).

In the Waterston lab samples, embryos were incubated for a specific amount of time after hypochlorite treatment. However, each sample has some outlier cells with abnormally low embryo time estimates, i.e. lower than the incubation time. There are several biological and technical factors that could produce these outlier cells. The developmental rate of *C. elegans* embryos can vary by over 2-fold depending on temperature, and may also be influenced by differences in crowding, hypoxia, or the effects of hypochlorite and chitinase treatment. Consistent with this, embryo times estimated using data from Boeck *et al* (*212*), which was collected using methods more similar to those used in this study, were systematically later than embryo times estimated using data from Hashimshony *et al.* (*18*) (Supplemental Figure 2.1C). Alternatively, some cells may have embryo time estimates that are lower than the true developmental age of the embryo they came from. Sparsity in the single cell data contributes to noise in the estimates. Finally, the most extreme outlier embryo time estimates in each sample are for germline cells. The germline maintains expression of many genes that turn off during early embryogenesis in all other cells. This causes embryo time estimates based on correlation to bulk RNA-seq to be inaccurate for this cell type.

Per-cell background correction and filtering

Our method for correcting for background RNA contamination, described in the section above titled “Dimensionality reduction”, works solely on the level of PCA coordinates and does not change the underlying gene-by-cell expression matrices. We used a separate background correction method to adjust these gene expression matrices on a per-cell basis for purposes of making plots of gene expression.

Our per-cell background correction method relies on a panel of cell-type specific marker genes that are assumed, based on the literature (and confirmed empirically in our data), to be specific to either hypodermis (including seam and P cells) or body wall muscle (BWM). The hypodermis-specific genes were: *sqt-3*, *dpy-17*, *dpy-14*, *dpy-10*, *dpy-7*, *dpy-2*, *dpy-3*, *bus-8*, *wrt-2*, and *noah-1*. The BWM-specific genes were: *pat-10*, *mlc-3*, *cpn-3*, *clik-1*, *ost-1*, *mlc-1*, *mlc-2*, *tni-1*, *ttn-1*, *unc-15*, and *myo-3*.

The gene expression distribution for the background contamination of each biological sample was estimated by aggregating the reads for cell barcodes that had < 50 UMIs, which were assumed to correspond to empty droplets in the 10x sc-RNA-seq apparatus. The expression level of each gene in the panel was computed for each sample’s background, measured in transcripts per million (TPM). Similarly, the expression level of each gene in the panel was computed for each cell, also measured in TPM. The background fraction of a cell was estimated as the sum of the expression of panel genes in the cell divided by the sum of the expression of panel genes in the background distribution for the sample that cell came from. For cells annotated as hypodermis, glia, or potential progenitors of those cell types, hypodermis-specific genes from the panel were excluded from the computation. Likewise, for cells annotated as body wall muscle, intestinal/rectal muscle, or a non-pharyngeal mesoderm cell type, as well as progenitors of those cell types, BWM-specific genes from the panel were excluded from the computation. For all other cells, all genes from the panel were used.

The median estimated background fraction across all cells in the dataset was 17.7%. Putatively damaged cells with an estimated background fraction >= 75% (8.3% of all cells, see Supplemental Figure 2.34A) were filtered entirely from all subsequent plots and analyses. For the remaining cells, the cells' gene expression profiles were corrected to subtract the contribution from background. A cell's raw gene expression vector (UMI counts) was converted to transcripts per million by dividing each entry by the sum and multiplying by one million. The background-corrected TPM value for each gene was computed according to the formula:

```
background-corrected TPM =

    max(raw TPM - background fraction * background TPM, 0)
```

where background TPM is the expression of the given gene in the background distribution for the biological sample that the cell came from. The background-corrected TPM values were then rescaled to sum to 1,000,000 and then converted back to (pseudo-)counts based on the total UMI count of the cell. Fractional count values were rounded probabilistically (i.e. a value of 2.7 was rounded to 3.0 with a 70% chance and to 2.0 with a 30% chance).

After background correction, cells with low background fractions and cells with high background fractions have near-identical average gene expression profiles (Supplemental Figure 2.34B). This indicates that non-background gene expression observed in high background cells is not systematically biased compared to low background cells.

Differential expression analysis for Figure 2.3D and Supplemental Figure 2.22

We included four classes of transcription factors (TFs) in the heatmaps of Figure 2.3D and Supplemental Figure 2.22. Both figures consider differential expression of TFs between different ciliated neuron lineages. For the division of a parent neuroblast into two daughter cells, the four TF classes of interest were:

1. TFs enriched in one daughter vs. the parent and vs. the other daughter
2. TFs depleted in one daughter vs. the parent and vs. the other daughter
3. TFs enriched in the parent vs. both daughters and vs. other neuroblasts of the same cell generation
4. TFs enriched in parent vs. other neuroblasts of the same cell generation; and in both daughters vs. other terminal cells

We considered a TF “enriched” in cell set A vs. cell set B if the expression in A was at least 3-fold higher than in B; and if the difference in expression was statistically significant with q-value < 0.01. We considered a TF “depleted” in cell set A vs. cell set B if it was “enriched” in B vs. A. q-values were computed using the Monocle (version 3 alpha) function “differentialGeneTest”. Differential expression tests were performed for all genes, not just TFs-the non-TF results were discarded, but this was done to produce more conservative q-values compared to considering only TF DE tests. Cells with embryo time >650 minutes were excluded from all comparisons. Due to limited figure space, some TFs that matched the criteria of the four TF classes but had low absolute expression levels were excluded from the figure heatmaps.

Derivation of lineage specific and terminal cell type specific genes for Figure 2.4D

Lineage specific genes were derived by one vs. rest differential expression analysis on the three input branches based on Louvain clustering results and annotations from Supplemental Figure 2.15, using “sSeq” (*213*), as implemented in the cellrangerRkit package. Genes associated with IL1/IL2 terminal cell types were derived by comparing IL1/IL2 cells to all other ciliated neurons in Figure 2.3A. For each of the gene sets, the average TPM across all genes in the set was computed for cells from each of the three input branches, binned in 30-minute intervals up to 390 minutes, where the branches can no longer be distinguished from each other in the UMAP. Values in each heatmap were linearly rescaled to be within the range of 0 to 1.

Pseudo-R^2 statistic

For each anatomical cell annotated in our dataset, we compute an aggregate gene expression profile from all of the sc-RNA-seq cells that we annotated as corresponding to that anatomical cell. This procedure is described in above section titled, “Computing aggregate gene expression profiles for cell types and lineages.” The result is that each anatomical cell is associated with a vector of relative gene expression values. We refer to this vector as the anatomical cell’s “transcriptome.”

In Figure 2.5B and Supplemental Figure 2.29B, we seek to estimate the extent to which the transcriptomes of cells in a given generation of the AB or MS lineages are predicted by the lineage. To do this, we have defined a statistic that measures how much more similar, on average, are the transcriptomes of sister cells compared to random pairs of cells. Specifically, we compute:

$$1 - \frac{\begin{array}{l}\texttt{average Jensen-Shannon divergence between the transcriptomes}\\ \texttt{of pairs of sister cells in the cell generation}\end{array}}{\begin{array}{l}\texttt{average Jensen-Shannon divergence between the transcriptomes}\\ \texttt{of random pairs of cells in the cell generation}\end{array}}$$

In the main text and figures, we refer to our statistic as a pseudo-R^2 statistic. The so-called pseudo-R^2 statistics are a family of statistics that have been proposed in the context of generalized linear regression models (*214*) and aim to have similar properties to the coefficient of determination, R^2, that is commonly used in the analysis of ordinary linear regression models. Similarly, the statistic we have defined aims to have similar properties to R^2, despite not being mathematically comparable to it in a rigorous sense. Below, we discuss the similarities between our pseudo-R^2 statistic and R^2.

One of several equivalent definitions of R^2 for an ordinary linear regression model is:

$$1 - \frac{\text{mean squared error of the regression model's predictions}}{\text{overall variance of the response variable}}$$

This formula for R^2 and our formula for pseudo-R^2 are both expressed in terms of a fraction subtracted from one. The numerator in our formula for pseudo-R^2, which we defined in terms of the Jensen-Shannon divergence, can be re-expressed as the average prediction error of a certain regression model, analogous to the numerator of regular R^2.

Specifically, the numerator in our pseudo-R^2 is equivalent to the average prediction error of a model that:

1. seeks to predict a cell's transcriptome based on the identity of its parent.
2. measures the deviation between its predicted transcriptome and the observed transcriptome for a cell using the Kullback-Leibler (KL) divergence.

Methods used in Supplemental Figure 2.30

In Supplemental Figure 2.30, we estimate the extent to which the ability of lineage to predict the transcriptome in a given cell generation, "generation N", is a consequence of gene expression signatures associated with each of the preceding cell generations 1 to N-1. We compute the overall ability of the lineage to predict the transcriptome in generation N using the pseudo-R^2 statistic described in the previous section. To compute the contribution of the parent generation N-1 to the total pseudo-R^2 for generation N, we use the formula:

$$\frac{\texttt{(average JS divergence between cells that share a grandparent -}\ \texttt{average JS divergence between sisters)}}{\texttt{average JS divergence between random pairs of cells}}$$

This formula evaluates how much more similar are cells that share a parent (i.e. sisters) than cells that share a grandparent (i.e. cousins or sisters), and scales this relative to the average dissimilarity of random pairs of cells in the same generation.

Generalizing this formula, we estimate the contribution of the generation N - M as:

$$\frac{\texttt{(average JS divergence between cells with lineage distance <= M+1 -}\ \texttt{average JS divergence between cells with lineage distance <= M)}}{\texttt{average JS divergence between random pairs of cells}}$$

where the lineage distance between two cells is the number of cell divisions since their most recent common ancestor (1 for sisters, 2 for cousins, etc.).

Using this formula, the sum of the contributions of each ancestor generation 1 to N-1 simplifies to:

$$\frac{\texttt{(average JS divergence between cells with lineage distance <= N-1 -}\ \texttt{average JS divergence between cells with lineage distance <= 1)}}{\texttt{average JS divergence between random pairs of cells}}$$

All cells in generation N have lineage distance <= N-1, so the first term in the numerator is equal to the average JS divergence between random pairs of cells (same as the denominator). Furthermore, the only cells with lineage distance <= 1 are sisters. Making these substitutions, we get:

$$\frac{\text{(average JS divergence between random pairs of cells - average JS divergence between sisters)}}{\text{average JS divergence between random pairs of cells}}$$

Which simplifies to our original statistic for total pseudo-R^2:

$$1 - \frac{\text{average JS divergence between sister cells}}{\text{average JS divergence between random pairs of cells}}$$

This equivalence is a consequence of the following:

1. When tasked to predict the transcriptomes of two sister cells, a model that predicts a cell's transcriptome based on the identity of its parent effectively guesses the midpoint of the two sister cells' transcriptomes.

2. Therefore, if one measures the deviation between the model's predictions and the observed transcriptomes using KL divergence, then the mean prediction error of the model, when applied to pairs of sister cells, is simply the average KL divergence between each cell's transcriptome and the midpoint of it and its sister's transcriptomes.

3. By the definition of Jensen-Shannon (JS) divergence, this is the same as the average JS divergence between each pair of sister cells' transcriptomes, which is the numerator used in our pseudo-R^2.

The denominator of our formula for pseudo-R^2, the average JS divergence between the transcriptomes of random pairs of cells, is a measure of the overall variability in the transcriptomic data. This is analogous to the denominator of regular R^2, which is also a measure of the overall variability (i.e. the variance) of the response variable in an ordinary linear regression model.

Thus, both the numerator and the denominator in our formula for pseudo-R^2 are qualitatively similar measurements to the numerator and denominator of regular R^2.

Computing the adjusted Gini coefficient for Supplemental Figure 2.33A

The Gini coefficient is biased by sample size (*215*). Therefore, to adjust for total UMI count differences between cells, we first downsampled the data from each cell to a total of 500 UMIs (the minimum UMI count across all cells) using a multinomial distribution, with probability equal to each gene's UMI count divided by the total UMI count of the cell. We then computed Gini coefficients for each cell using the downsampled data, and used the z-score of the adjusted Gini coefficients to compare transcriptome inequality across cells.

Comparison of data from this study to public single cell dataset

Due to technical limitations, we have data from relatively few cells prior to the 28-cell stage. Therefore, we compared single cell RNA-seq profiles of cells from the 16-cell stage collected by Tintori *et al.* (*19*) to their corresponding lineages or immediate descendants in our dataset (Supplemental Figure 2.19). We downloaded normalized expression data (measured in reads per kilobase of transcript per million mapped reads, RPKM) from Tintori *et al.* (*19*) and computed average log2 normalized expression levels for each of their annotated lineages. We then applied the same log2 transformation on our normalized gene expression data, and measured pairwise similarity between the expression vectors for each lineage using Pearson correlation. To enrich for lineage-specific signals, we computed correlation using gene sets that had been selected by Tintori *et al. (19)* using an iterative PCA approach. Gene sets 7, 9, and 10 in supplemental document S1 of Tintori *et al.* (*19*) were used to discriminate 16-cell stage lineages. Set 8 was excluded because most germline (P4) specific genes are also differentially expressed over time throughout the whole embryo and thus confound time with lineage. Intersecting genes from sets 7, 9, and 10 with genes detected in our data, we obtained a list of 593 genes that we then used to generate the correlation matrix shown in Supplemental Figure 2.19A. Hierarchical clustering was performed on the correlation matrix using the pheatmap package with default parameters (*216*).

To demonstrate that our data are consistent with Tintori *et al.* (*19*) at the level of single cells, we repeated their PCA analysis and projected 16- and 28-cell stage cells from our dataset onto the PCA space derived from their dataset (Supplemental Figure 2.19B-E). The distribution and orientation of lineages in the PCA space was similar for our and their data. For example, the PCA in the top sub-panel of Supplemental Figure

2.19B was computed using all 16-stage cells from Tintori *et al.* (*19*) and 421 genes (intersection of Set 7 and expressed genes in our data). In the bottom sub-panel of Supplemental Figure 2.19B, we projected 292 cells from the 16- and 28-cell stages from our dataset using the loading matrix derived from the PCA of the Tintori *et al.* data (*19*). Germline (P4, Z2/Z3) and endoderm lineage (Ex, Exx) cells from our data are located at the left and right-hand sections of the PCA projection respectively, consistent with the pattern observed with cells from Tintori *et al (19)*.

Spencer *et al.* (*191*) used microarrays to profile the transcriptomes of *C. elegans* cell types obtained by fluorescent activated cell sorting. For each cell type they profiled, they derived a set of genes that are enriched in that cell type compared to all other cells. We used these “signature” gene sets to validate our cell type annotation and the robustness of our data. First, we downloaded signature gene sets from cell types profiled at the embryonic stage from https://www.vanderbilt.edu/wormdoc/wormmap/Enriched_genes.html. We then used the AUCell package (*217*) to check for enrichment of Spencer *et al.* (*191*) signature genes in single cells from our dataset. For each cell, AUCell ranks genes by expression level and computes a recovery curve for each gene set. It then uses “Area Under the Curve” (AUC) as a measure of enrichment of the gene set.

We found most Spencer *et al.* (*191*) signature genes have strong enrichment in the corresponding cell types in our data (Supplemental Figure 2.20). Due to the method by which the Spencer *et al. (191)* signature genes were derived—comparing one cell type to all other cells—most of the genes are tissue-specific, not cell-type specific, so enrichment was in some cases also observed in a set of several related cell types in our data.

Spencer *et al.* (*191*) signature genes for pharyngeal muscle were unusual in that they were enriched in intestine cells from our dataset. Examining the pharyngeal muscle gene set, we noticed it contains *elt-2* and *elt-7*, which are known to be endoderm specific (*218*). Checking this gene set against expression patterns from Warner *et al.*, 2019 (*210*), we found that 18 out of the top 20 genes are intestine specific/enriched. Therefore, we concluded the pharyngeal muscle signature list is problematic and dropped the comparison from Supplemental Figure 2.20.

Figures

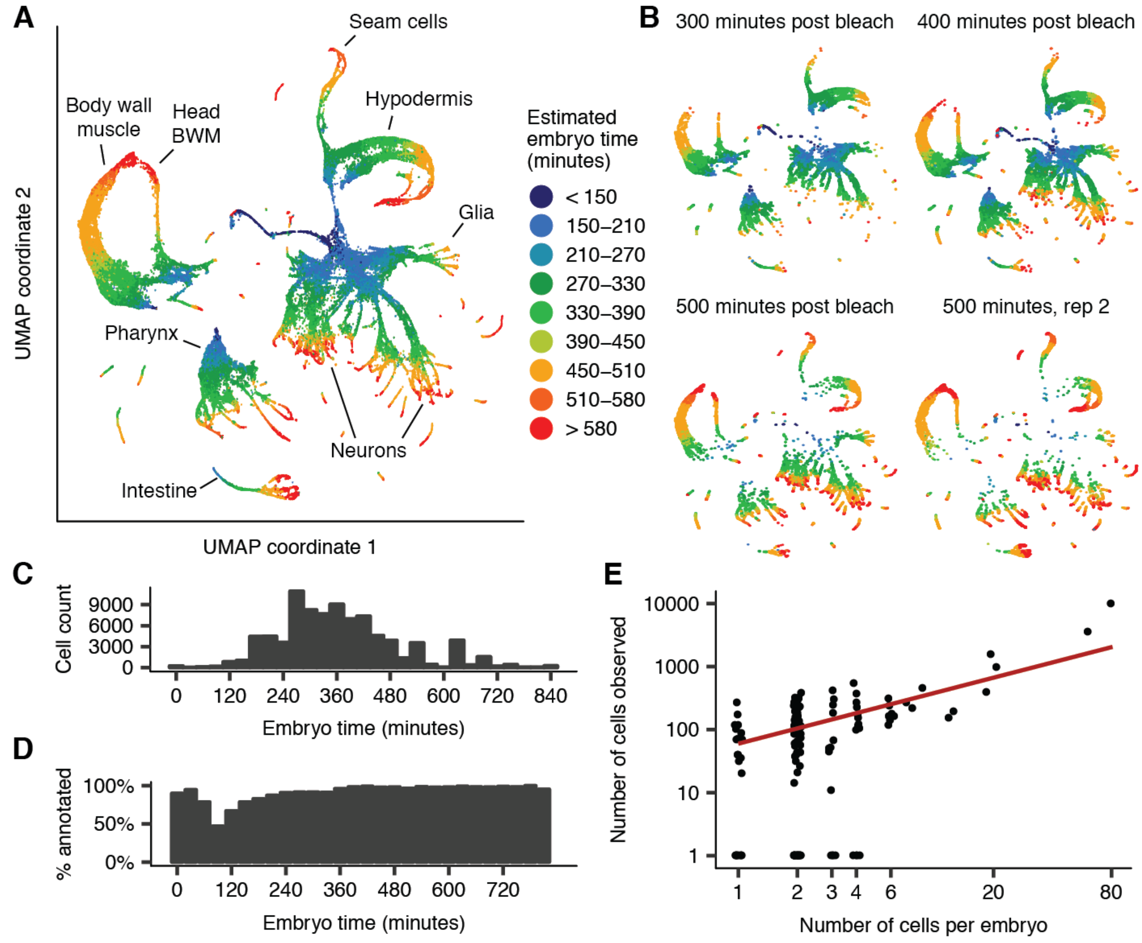


Figure 2.1 UMAP projection shows tissues and developmental trajectories in *C. elegans* embryogenesis.

(A) UMAP projection of the 81,286 cells from our sc-RNA-seq dataset that passed our initial QC. This UMAP does not include 4,738 additional cells that were initially filtered, but were later whitelisted and included in downstream analyses. Color indicates the age of the embryo that a cell came from, estimated from correlation to a whole-embryo RNA-seq time series (*186*) and measured in minutes after an embryo's first cell cleavage. **(B)** Positions of cells from four samples of synchronized embryos on the UMAP plot. **(C)** Histogram of estimated embryo time for all cells in the dataset. **(D)** Bar plot showing for bins of embryo time, the percentage of cells in that embryo time bin that we were able to assign to a terminal cell type or pre-terminal lineage. **(E)** Scatter plot showing correlation of the number of cells of a given anatomical cell class in a single embryo (X axis, log scale) with the number of cells recovered in our data (Y axis, log scale). Each point corresponds to a cell class. Only cells with estimated embryo time >= 390 minutes are included in the counts (many earlier cells are still dividing). Red line is a linear fit, excluding points with y = 0.

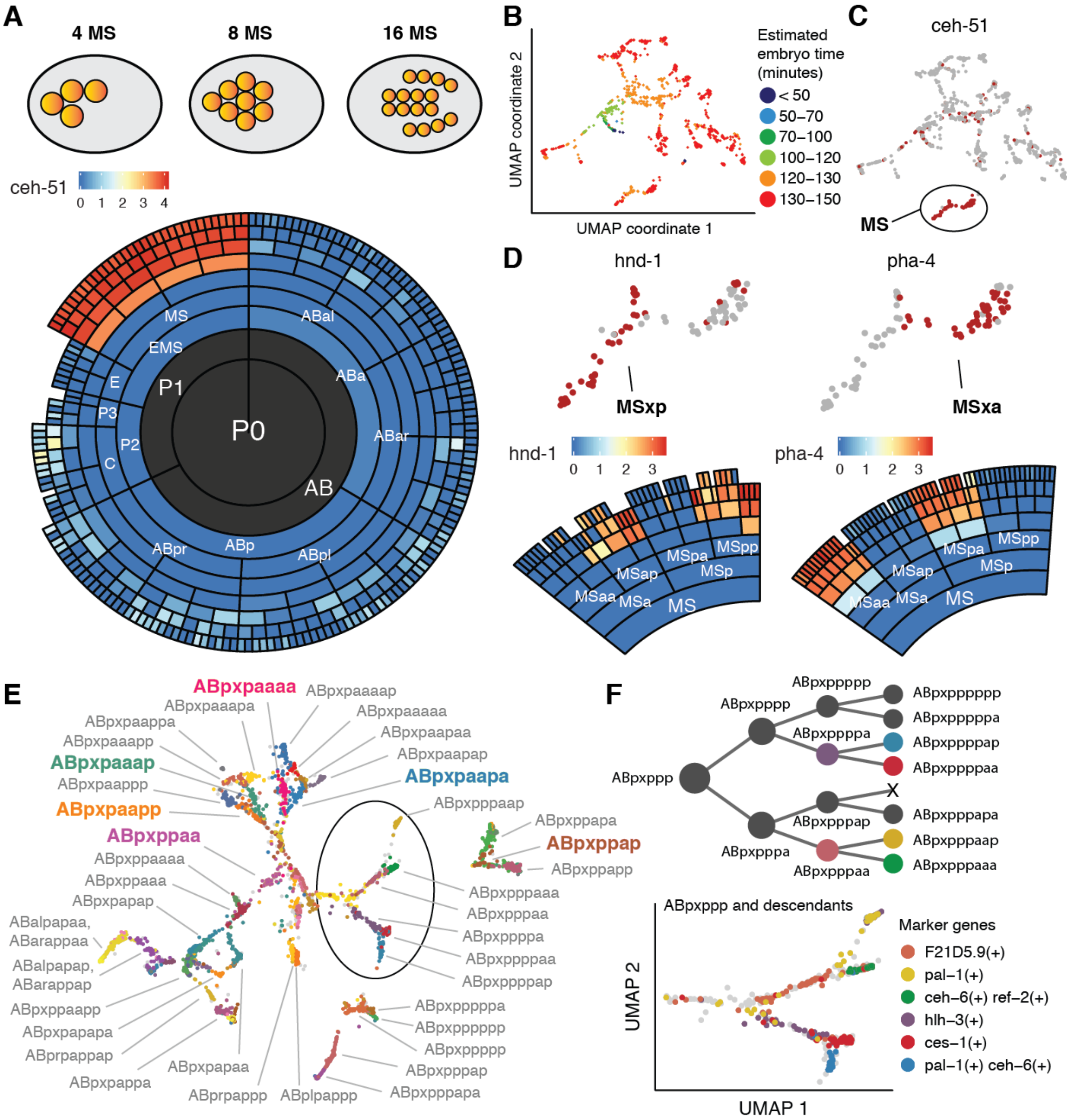


Figure 2.2 Annotation of the early lineage.

(A) Diagram showing the position of early mesoderm (MS lineage) cells marked by expression of *ceh-51*. The lineage radiograph shows the average fluorescent intensity (log10 scaled) of a CEH-51::GFP protein fusion measured by live imaging. The inner rings show the generation of the founder cells, AB (which produces almost exclusively ectoderm and pharynx), MS (mesoderm and pharynx), C (muscle and ectoderm) and P3, which gives rise to P4 (germline) and D (muscle). Daughter cells are named by their relative positions at mitosis (e.g. ABa is the anterior daughter of AB, ABal is left daughter of ABa). **(B)** UMAP projection of 926 early-stage cells (estimated embryo time <= 150 minutes), colored by embryo time. E lineage and germline cells are excluded and shown separately in Supplemental Figure 2.7 and Supplemental Figure 2.12, as they differentiate early compared to other lineages. **(C)** Same UMAP as (B), colored by *ceh-51* expression (red indicates cells with >0 UMIs for *ceh-51*). **(D)** Expression of *hnd-1* and *pha-4* measured by sc-RNA-seq (UMAP) and live imaging of GFP protein fusions (radiograph). **(E)** Cropped section of a UMAP of 8,083 neuron/glia/rectal progenitor cells with embryo time <= 250 minutes (Supplemental Figure 2.15). This plot shows the section of that UMAP that corresponds to the 3,233 cells from the ABpxp ectodermal lineage ("ABpxp" is short-hand for two symmetric lineages, ABplp and ABprp). Colored bold annotations highlight specific lineages that are discussed in the text. **(F)** Lineage tree for the ABpxppp sub-lineage, highlighting cells that are present in the circled section of (E). The (co-)expression pattern of marker genes identifies branches in the UMAP that correspond to specific ABpxppp descendants. Additional ABpxppp descendants not shown in this panel are annotated in (E), below the circled section.

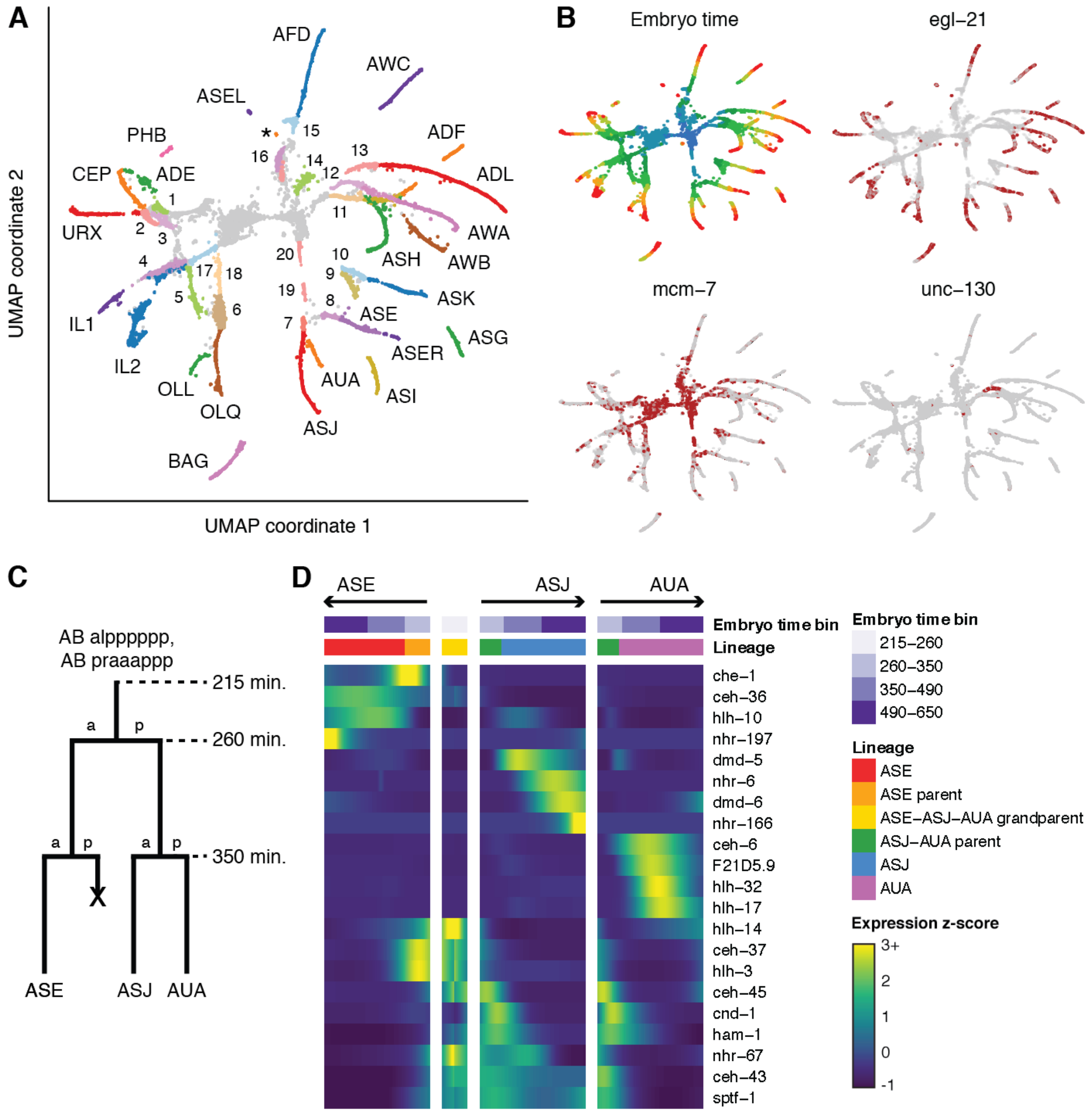


Figure 2.3 Developmental trajectories of ciliated neurons.

(A) UMAP of 10,740 ciliated neurons and precursors. Colors correspond to cell identity. Text labels indicate terminal cell types. Numbers 1-16 indicate parents of **1** ADE-ADA, **2** CEP-URX **3** PHB-HSN **4** IL1 **5** OLL **6** OLQ **7** ASJ-AUA **8** ASE **9** ASI **10** ASK **11** ADF-AWB **12** ASG-AWA **13** ADL **14** ASH-RIB **15** AFD-RMD **16** AWC-SAA (purple) and BAG-SMD (red)**. 4-6**, **8-10**, and **13** are listed as parents of only one cell type as the sister cells die. Numbers 17-20 indicate grandparents of **17** IL1 (= IL2 parent) **18** OLQ-URY **19**, **20** ASE-ASJ-AUA. Differentiated PHA was not conclusively identified but may co-cluster with PHB. The parent of PHA is not present in this UMAP, but was located separately within the area annotated as “rectal cells” in the UMAP in Supplemental Figure 2.3. The tiny cluster labeled with an asterisk (*****) is putatively AWC-ON on the basis of *srt-28* expression. **(B)** UMAP plot colored by embryo time (colors matched to Figure 2.1A) and gene expression (red indicates >0 reads for the listed gene). *egl-21* codes for an enzyme that is essential for processing neuropeptides (*219*). Its expression is used as a proxy for the onset of neuron differentiation. *mcm-7* codes for a DNA replication licensing factor. Loss of *mcm-7* expression in each UMAP trajectory approximately marks the boundary between neuroblasts and terminal cells. *unc-130* is known to be expressed in the ASG-AWA neuroblast but neither terminal cell (*220*). **(C)** Cartoon illustrating the lineage of the ASE, ASJ, and AUA neurons. **(D)** Heatmap showing patterns of differential transcription factor expression associated with branches in the ASE-ASJ-AUA lineage. Expression values are log-transformed, then centered and scaled by standard deviation for each row (gene).

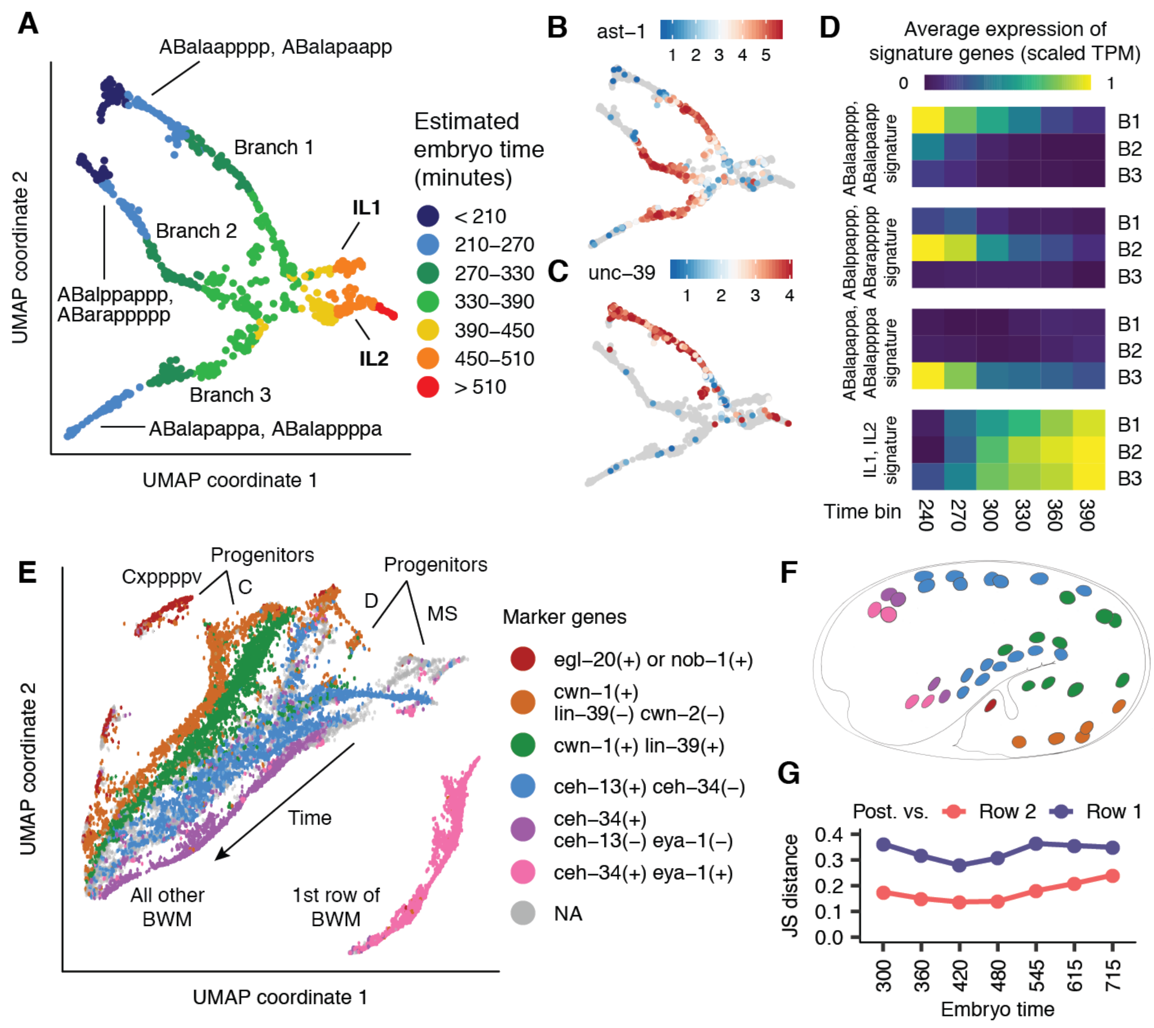


Figure 2.4 Full vs. incomplete convergence of lineages producing common cell types.

(A) UMAP of 854 IL1/2 neurons and progenitors colored by estimated embryo time (cells selected on the basis of annotations in Figure 2.3A and Supplemental Figure 2.15). **(B)** IL1/2 UMAP colored by *ast-1* expression level (log2 size-factor normalized UMI counts). **(C)** IL1/2 UMAP colored by expression of *unc-39*, a gene specific to branch 1. **(D)** Heatmap showing the average expression level of lineage specific and terminal cell type specific genes over time for each of the 3 branches. **(E)** Supplemental Figure 2.5A shows a UMAP of body wall muscle and mesoderm cells. This panel is a zoomed-in view of that UMAP, including only 17,520 BWM cells, which are grouped into “bands” based on marker gene expression patterns (here, a cell is considered to express a gene if it or >= 2 of 5 of its nearest neighbors have >0 reads for the gene). **(F)** Physical positions of cells in each BWM band (colors matched to panel E) in the embryo at 430 minutes. Adapted from Fig. 8B of (*7*). **(G)** Transcriptome Jensen-Shannon distance for posterior (orange+green bands in panel E) BWM vs. row 2 (blue band) or row 1 (pink band) head BWM over time. Heterogeneity between BWM subsets persists throughout development and may reflect functional differences.

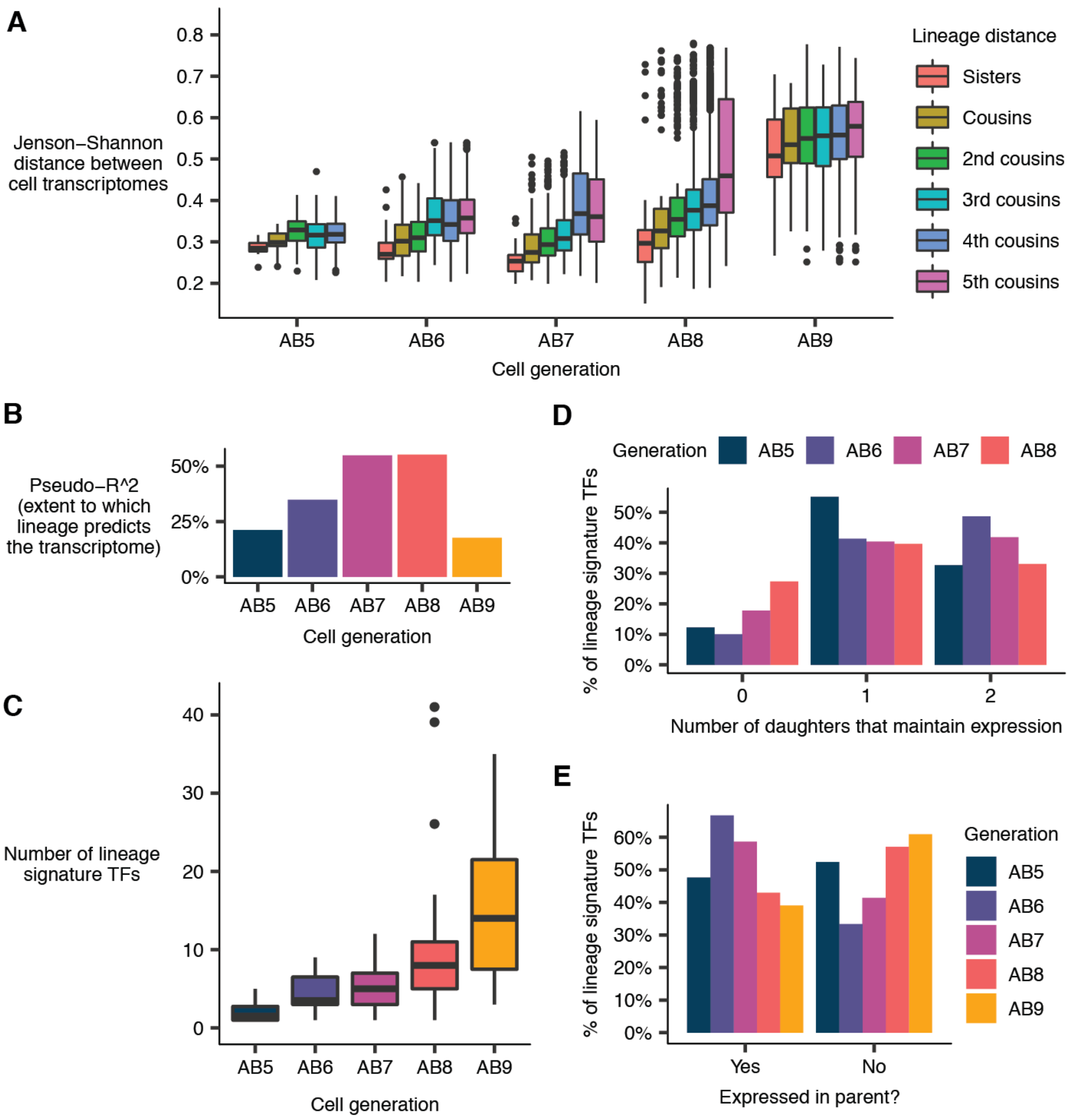


Figure 2.5 Correlation between cell lineage and the transcriptome in the ectoderm.

(A) Jensen-Shannon (JS) distance between the transcriptomes of pairs of ectodermal cells (AB lineage), faceted by cell generation and lineage distance. AB5 refers to the cell generation produced by 5 divisions of the AB founder cell, and likewise for generations AB6-9. The “transcriptome” of a given anatomical cell is defined as the average gene expression profile of all sc-RNA-seq cells annotated as that anatomical cell. Pairs of bilaterally symmetric cells are excluded from the statistics. **(B)** Estimates of the extent to which lineage predicts the transcriptome in AB5-9. **(C)** Distribution of the number of “lineage signature transcription factors”—TFs that distinguish a cell from its sister—for all cells in AB5-9. The outlier points in AB8 are instances where a terminal epidermal cell is a sister of a neuroblast. **(D)** Proportion of lineage signature transcription factors for a cell in a given generation that have expression maintained in 0, 1, or 2 of the cell’s daughters in the subsequent generation. **(E)** Proportion of lineage signature TFs for which expression in a given cell was maintained from the cell’s parent vs. newly activated after the parent’s division.

Tables

Table 2.1 Marker genes for terminal cell type annotations.

This table lists the marker genes that were used to annotate sc-RNA-seq cells with their corresponding cell types. Expression patterns for marker genes were retrieved from Wormbase (*188*) (https://wormbase.org) and EPiC (*17*) (http://epic.gs.washington.edu/Epic2/). The UMAP column lists which UMAP the cell type was located in. These UMAPs are shown in Supplemental Figures 2.4-S2.13 and can be explored in the VisCello application (https://cello.shinyapps.io/celegans_explorer/). For marker genes that have expression profiles in EPiC (*17*) (http://epic.gs.washington.edu/Epic2/), we used protein fusion datasets when available, and only used promoter fusion datasets when protein fusions were unavailable.

Cell type	UMAP	Marker genes
BWM_far_posterior	Muscle and mesoderm	hlh-1, myo-3, egl-20
BWM_posterior	Muscle and mesoderm	hlh-1, myo-3, cwn-1
BWM_anterior	Muscle and mesoderm	hlh-1, myo-3, ceh-13
BWM_head_row_2	Muscle and mesoderm	hlh-1, myo-3, ceh-34
BWM_head_row_1	Muscle and mesoderm	hlh-1, myo-3, ceh-34, eya-1
Coelomocyte	Muscle and mesoderm	cup-4, lgc-26, let-381
GLR	Muscle and mesoderm	unc-30, let-381, sfrp-1
hmc	Muscle and mesoderm	hlh-8, sfrp-1, glb-26, dmd-4
M_cell	Muscle and mesoderm	hlh-8, pal-1
mu_int_mu_anal	Muscle and mesoderm	hlh-8, mls-1, dsc-1, exp-1, mig-1, unc-62
mu_sph	Muscle and mesoderm	hlh-8, mls-1, dsc-1
Z1_Z4	Muscle and mesoderm	ehn-3, unc-39
g1A	Pharynx	hlh-6, phat-2, phat-5, lys-8
g1P	Pharynx	hlh-6, phat-2, phat-1, lys-8
g2	Pharynx	hlh-6, ceh-6, irx-1, dmd-4, gly-15

mc1	Pharynx	ttx-1, pax-1, agr-1, ceh-45, ceh-2
mc2	Pharynx	ttx-1, pax-1, agr-1
mc3	Pharynx	ttx-1, pax-1, agr-1, irx-1
Pharyngeal_intestinal_valve	Pharynx	ttx-1, lec-8, cwn-2, unc-62, fos-1
pm1_pm2	Pharynx	tnc-2, tnt-4, tni-4, inx-20, eyg-1
pm3_pm4_pm5	Pharynx	tnc-2, tnt-4, tni-4, ceh-22
pm6	Pharynx	tnc-2, ser-2, elt-4, W05B10.4
pm7	Pharynx	tnc-2, tni-4, ceh-22, spp-7, W05B10.4
pm8	Pharynx	ref-1, aff-1, pax-1, inx-20, unc-129
Anterior_arcade_cell	Pharynx	inx-12
Posterior_arcade_cell	Pharynx	inx-12, let-23
hyp1_hyp2	Pharynx	mlt-11, mlt-8, slt-1, nhr-25, nhr-67
MC	Pharynx	ceh-19, nhr-239, glr-8
Intestine_anterior	Intestine	ZC204.12, cpr-1, ceh-37
Intestine_middle_and_posterior	Intestine	irg-7, pal-1 cpr-1 and ceh-37 in subset
Intestine_far_posterior	Intestine	irg-7, faah-1, pbo-4, psa-3
hyp4_hyp5_hyp6	Hypodermis and seam cells	elt-1, elt-3, slt-1, vab-3, unc-130, egl-17, ceh-32 in subset
hyp7_AB_lineage	Hypodermis and seam cells	elt-1, elt-3, unc-62, vab-3, unc-130, tbx-2 ceh-13 in subset
hyp7_C_lineage	Hypodermis and seam cells	elt-1, elt-3, tbx-8, tbx-9 lin-39 in subset
Tail_hypodermis	Hypodermis and seam cells	elt-1, elt-3, lin-44, vab-7
P_cell	Hypodermis and seam cells	elt-1, elt-3, pax-3, plx-2, lin-39, mab-5
Seam_cell	Hypodermis and seam cells	bus-4, bus-8, bus-12, ceh-16, rnt-1, elt-6
G2_and_W_blasts	Hypodermis and seam cells	lin-12, ahr-1
AMsh	Glia and excretory cells	aff-1, kcc-3, nas-31, pros-1, F52E1.2, F16F9.3
ADEsh	Glia and excretory cells	aff-1, unc-62

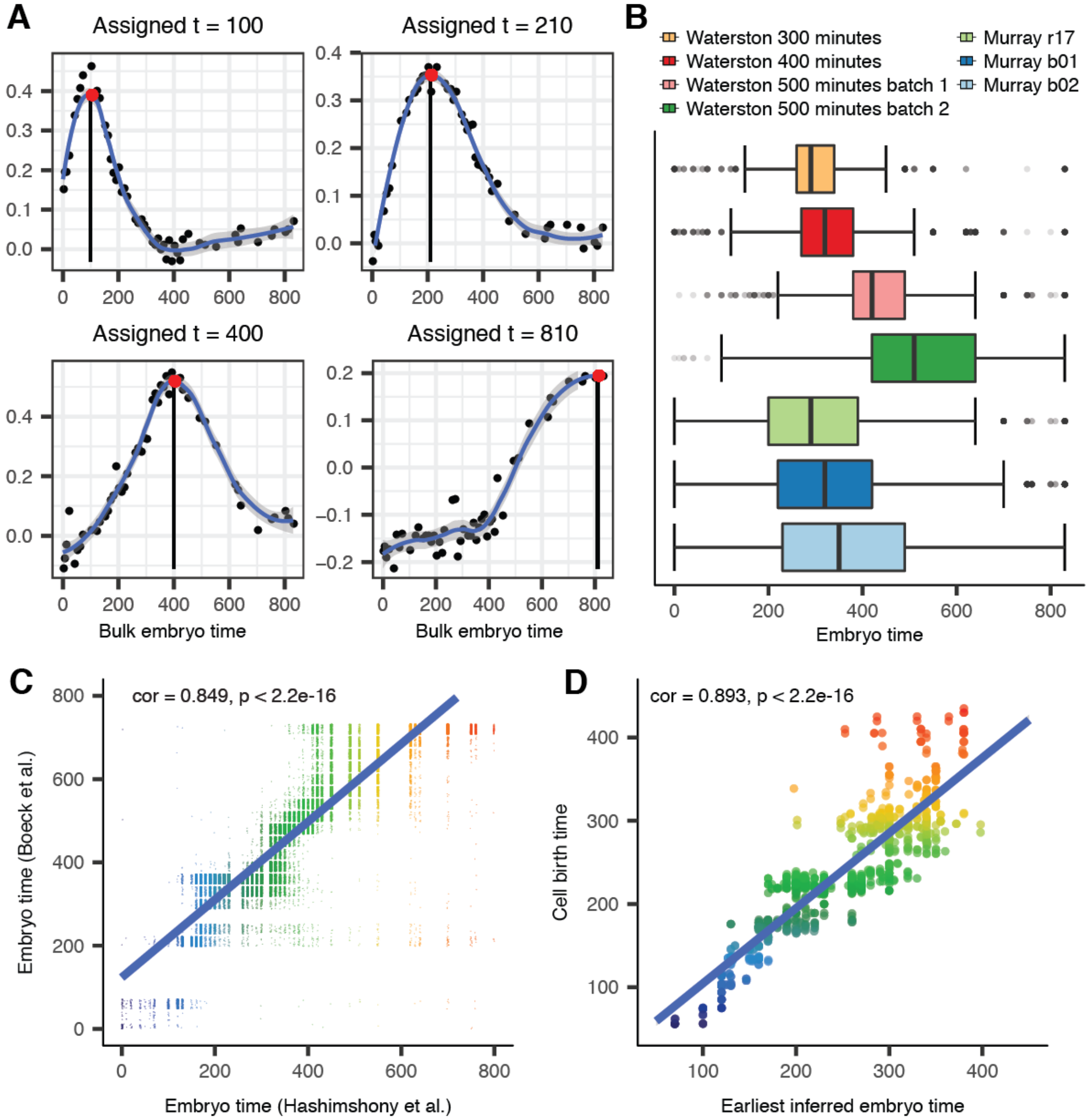
CEPsh	Glia and excretory cells	aff-1, kcc-3, aqp-7, K09F5.6, mltn-13, K08D12.4
ILsh_OLLsh_OLQsh	Glia and excretory cells	aff-1, kcc-3
AMso	Glia and excretory cells	grd-15, grl-12
CEPso	Glia and excretory cells	mls-2, inx-12, inx-13
ILso	Glia and excretory cells	grl-18, wrt-6
hyp3	Glia and excretory cells	nhr-25, ceh-32, slt-1, sym-1
Excretory_cell	Glia and excretory cells	pros-1, ceh-37, ceh-6, hlh-11
Excretory_gland	Glia and excretory cells	lim-6, ser-2, aat-1
Excretory_duct_and_pore	Glia and excretory cells	irx-1, ceh-37, grl-2, let-23
XXX	Glia and excretory cells	eak-3, sdf-9, eak-6
Possibly_hyp1V	Glia and excretory cells	mlt-8, qua-1, nhr-25
Possibly_ant_arc_V	Glia and excretory cells	See note.
Germline	Early embryo, germline, and rectum	glh-1, pgl-1, nos-1
ADF	Ciliated neurons	ceh-19, cat-4, bas-1
ADL	Ciliated neurons	K04D7.6, xbx-9, F15A4.5
AFD	Ciliated neurons	gcy-8, gcy-18, gcy-23, dac-1, ttx-1
ASE	Ciliated neurons	che-1, ceh-36 ASEL: gcy-6, gcy-14, lim-6 ASER: gcy-5, gcy-22
ASG	Ciliated neurons	gcy-11, capa-1
ASH	Ciliated neurons	osm-10, R102.2, deg-1, M04B2.6, unc-42
ASI	Ciliated neurons	ins-6, cng-2
ASJ	Ciliated neurons	ssu-1, trx-1, nhr-6, sptf-1
ASK	Ciliated neurons	F09E8.8, pax-2, C47D2.1
AWA	Ciliated neurons	odr-7, nhr-216, ocr-1
AWB	Ciliated neurons	srd-23, odr-1, daf-11, sox-2
AWC	Ciliated neurons	ceh-36, odr-1, daf-11, sox-2
ADE	Ciliated neurons	dat-1, cat-2, tba-9, pdf-1, unc-62, cwn-2, ceh-13
CEP	Ciliated neurons	dat-1, cat-2, tba-9, nhr-67, nhr-67
URX	Ciliated neurons	gcy-32, gcy-35, gcy-36, gcy-37

BAG	Ciliated neurons	gcy-9, gcy-31, gcy-33
IL1	Ciliated neurons	flp-3, agr-1, sox-2
IL2	Ciliated neurons	tba-6, klp-6, cil-7, agr-1, sox-2
OLL	Ciliated neurons	sox-2, tbx-2
OLQ	Ciliated neurons	ocr-4, dyla-1, dhc-3, pcrg-1
PHB_and_possibly_PHA	Ciliated neurons	osm-10, R102.2, gpa-6, cog-1 low expression of ceh-14, srb-6, srh-74
AUA	Non-ciliated neurons and Ciliated neurons	ceh-6, dop-1, flr-4, che-7
AIA	Non-ciliated neurons	ttx-3, mgl-1, flp-2, ins-1
AIB	Non-ciliated neurons	snet-1, aptf-1, odr-2, glr-2
AIM	Non-ciliated neurons	snet-1, flp-22, mbr-1, mls-2, mod-5, unc-86, inx-19
AIN	Non-ciliated neurons	ttx-3, mgl-1, K07C5.9, ast-1
AIY	Non-ciliated neurons	ttx-3, ceh-10, F17C11.2, flp-9, glc-3, bus-18, ser-2, nlp-15
AIZ	Non-ciliated neurons	ser-2, unc-86, eat-4, acc-2
ALA	Non-ciliated neurons	flp-24, ceh-17, des-2, deg-3, snf-11, flp-13
ALN	Non-ciliated neurons	unc-86, lad-2, gcy-35
ALM_PLM	Non-ciliated neurons	mec-17, mec-3, mec-7, unc-86
ALM_BDU	Non-ciliated neurons	mec-17, mec-3, mec-7, unc-86, unc-62
AVA	Non-ciliated neurons	acc-1, fax-1, unc-42, unc-3, flp-18, acr-16, acr-15 rig-3, gpa-14, glr-1, glr-2, nmr-2, unc-3
AVB	Non-ciliated neurons	fax-1, unc-42, unc-3, ceh-31, pdf-1
AVD	Non-ciliated neurons	unc-42, unc-3, unc-17, rig-5, glr-1, glr-2, nmr-2
AVE	Non-ciliated neurons	fax-1, unc-42, unc-3, glr-1, glr-2, glr-5
AVG	Non-ciliated neurons	lite-1, glr-1, glr-2, nmr-2, lin-11, ast-1, odr-2, F59E11.7, unc-62

AVH	Non-ciliated neurons	unc-42, lin-11, hlh-34, ceh-6, flp-12, pdf-1
AVJ	Non-ciliated neurons	unc-42, lin-11, glr-1
AVK	Non-ciliated neurons	flp-1, fax-1, unc-42, sox-2, glr-5
AVL	Non-ciliated neurons	unc-25, unc-46, unc-47, lim-6, ceh-27, alr-1
CAN	Non-ciliated neurons	pks-1, ceh-10, ace-3, acy-2
DA	Non-ciliated neurons	gbb-1, gbb-2, unc-3, unc-4, unc-17, unc-62, mab-9
DB	Non-ciliated neurons	gbb-1, gbb-2, unc-3, unc-4, unc-17, unc-62, mab-9, vab-7, ceh-6
DD	Non-ciliated neurons	unc-25, unc-30, unc-46, unc-47, unc-62, snf-11
DVA	Non-ciliated neurons	lin-44, nob-1, fax-1, nlp-12, twk-16, lin-11
DVC	Non-ciliated neurons	ceh-63, hlh-14, hlh-13, egl-20
FLP	Non-ciliated neurons	unc-86, mec-7, mec-3, unc-62
I5	Non-ciliated neurons	unc-4, ceh-34, tbx-2, flp-4, flp-13, unc-7
PLM	Non-ciliated neurons	mec-17, mec-3, mec-7, unc-86, egl-5
PVP	Non-ciliated neurons	mbr-1, nlp-7, unc-30, lin-11, pdf-1, glb-17
PVQ_and_possibly_PVC	Non-ciliated neurons	lin-11, vab-15, ceh-43, glr-1 nlp-17, C35B1.7, F26A10.1, Y43F8B.20, acr-23, irx-1
PVR	Non-ciliated neurons	hlh-14, unc-86, egl-20
PVT	Non-ciliated neurons	gpa-2, mec-1, zig-5, vab-15, dop-5, pdf-1, lim-6, unc-6
RIA	Non-ciliated neurons	glr-3, glr-6
RIB	Non-ciliated neurons	aptf-1, glr-4, ser-4, sto-3, unc-29
RIC	Non-ciliated neurons	tbh-1, tdc-1, glr-5
RID	Non-ciliated neurons	unc-3, ceh-10, lim-4, pdf-1
RIH	Non-ciliated neurons	unc-86, unc-130, nhr-67, slt-1, rig-4
RIM	Non-ciliated neurons	tdc-1, cex-1, glr-1, nmr-2
RIS	Non-ciliated neurons	flp-11, unc-25, aptf-1, lim-6

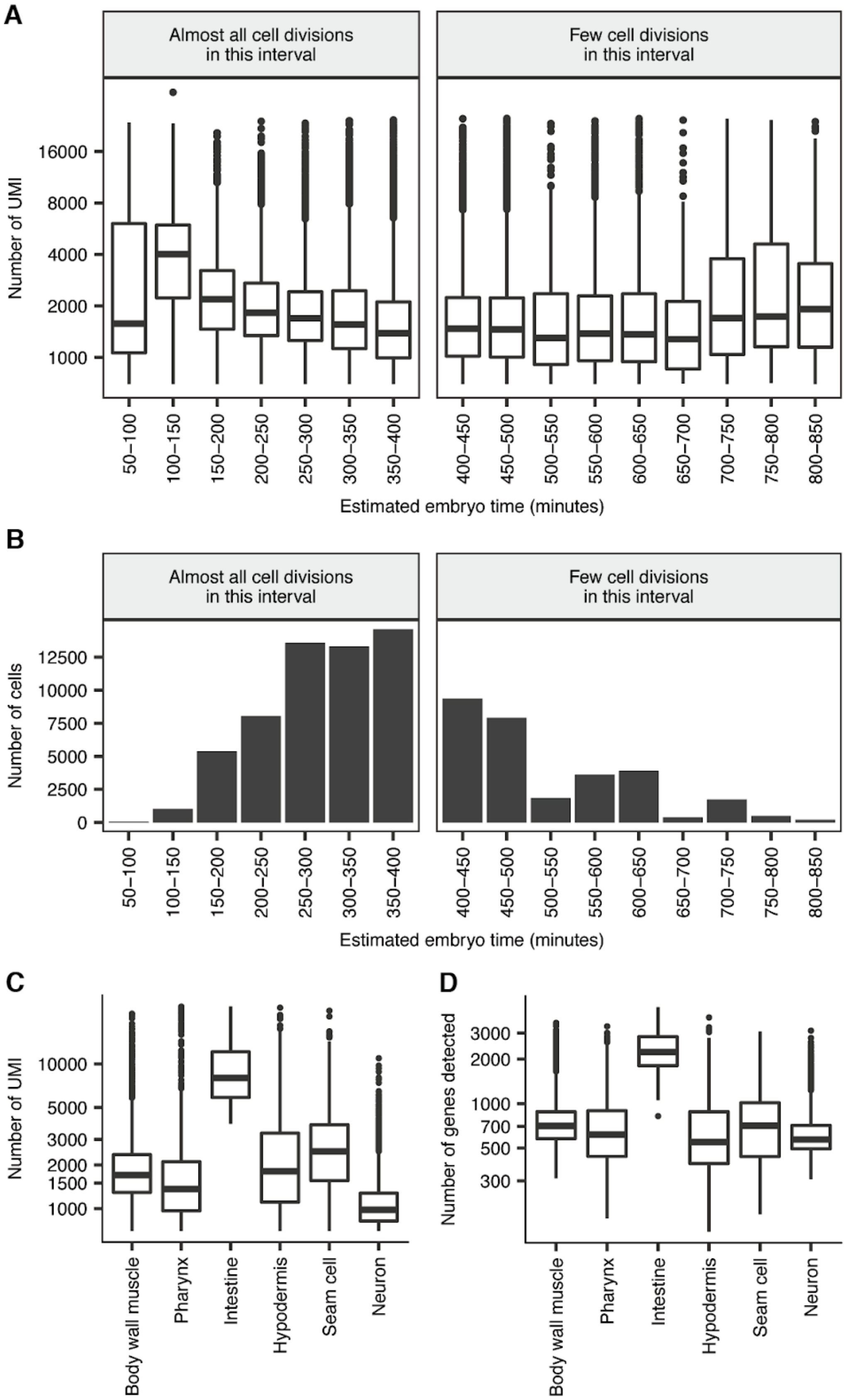
RIV	Non-ciliated neurons	odr-2, lim-4, ceh-75, ast-1
RMD	Non-ciliated neurons	lad-2, acc-1, unc-42, glr-1, glr-4, glr-5, ceh-6, mgl-1, unc-7, ast-1
RME	Non-ciliated neurons	ceh-32, unc-25, unc-46, snf-11, sox-2, ser-2
SIA	Non-ciliated neurons	ceh-17, ceh-24, unc-42, vab-8, ser-6, lim-4
SIB	Non-ciliated neurons	fax-1, ceh-24, unc-42, vab-8, tmc-1, glr-5
SMB	Non-ciliated neurons	sox-3, ceh-24, lim-4, vab-8, unc-42
SMD	Non-ciliated neurons	lad-2, acc-1, flp-22, odr-2, unc-42, glr-1, glr-5
URB_and_possibly_URA	Non-ciliated neurons	unc-86, sox-2, glr-8
T	Time 350min hypodermis + glia	psa-3, ceh-16, tlp-1, php-3, elt-1
Excretory_duct	Duct and pore	irx-1, ceh-37, grl-2, aff-1
Excretory_pore_G1	Duct and pore	irx-1, ceh-37, grl-2, lack of aff-1
B	Rectal cells	ceh-6, ref-2, mab-9, ceh-27
F_U	Rectal cells	egl-38, egl-20, mom-2
K_Kprime	Rectal cells	pha-4, pal-1, egl-38
Y	Rectal cells	ceh-6, ref-2, mom-2, nhr-25, lack of cnd-1
B_F_K_Kp_U_Y	Rectal cells	ceh-6, ref-2, mab-9, daf-6
rect_D	Rectal cells	pha-4, pal-1, dve-1
Rectal_gland	Rectal cells	pha-4, pal-1, dve-1, nac-2, elt-3, tat-4

Supplemental Figures



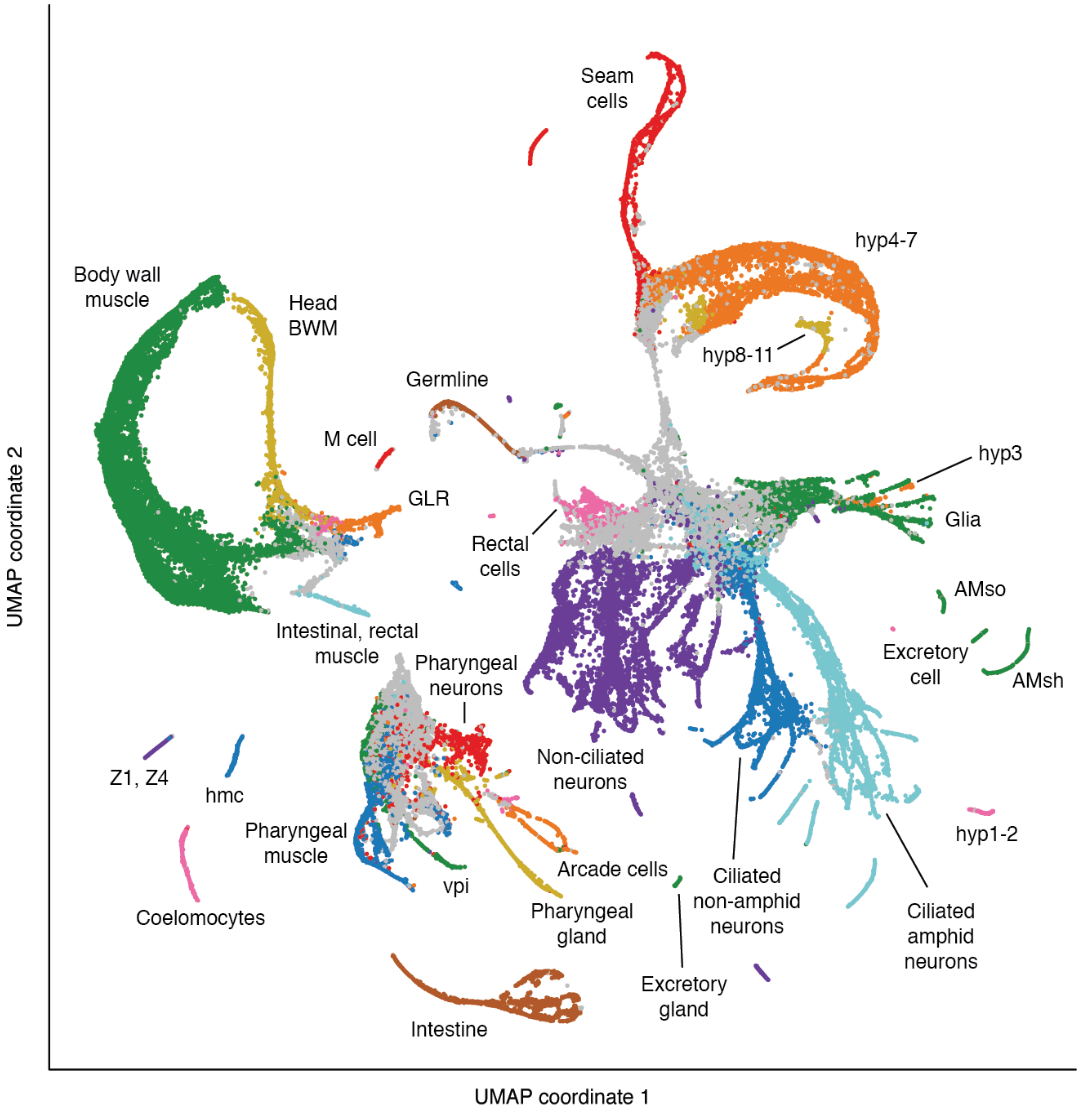
Supplemental Figure 2.1 Method for estimating the age of the embryo that a sc-RNA-seq cell came from.

Embryo times are measured in minutes post first cleavage. **(A)** Embryo times are estimated based on Pearson correlation of a single cell's transcriptome to a bulk RNA-seq time series (see Methods). Pointwise estimates of the correlation to each time point are smoothed using a Loess regression. **(B)** Distribution of estimated embryo times for each biological sample. The average embryo time estimate in the Waterston lab sample correlates with the real time duration that the embryos were incubated. Each sample contains some outlier cells with abnormally low embryo times. Potential biological and technical causes for the presence of these outlier cells are discussed in the Methods. **(C)** Correlation of embryo time estimates based on Hashimshony *et al.* (*186*) to an alternate set of embryo time estimates based on Boeck *et al.* (*212*). Estimates based on Hashimshony *et al.* (*186*) were used for all downstream analyses. **(D)** Correlation between cell birth times estimated based on our lineage annotations (x-axis) with cell birth times computed based on automated analysis of imaging data (y-axis) (*192*).


A
Almost all cell divisions in this interval
Few cell divisions in this interval
Number of UMI
16000
8000
4000
2000
1000
50–100
100–150
150–200
200–250
250–300
300–350
350–400
400–450
450–500
500–550
550–600
600–650
650–700
700–750
750–800
800–850
Estimated embryo time (minutes)
B
Almost all cell divisions in this interval
Few cell divisions in this interval
Number of cells
12500
10000
7500
5000
2500
0
50–100
100–150
150–200
200–250
250–300
300–350
350–400
400–450
450–500
500–550
550–600
600–650
650–700
700–750
750–800
800–850
Estimated embryo time (minutes)
C
Number of UMI
10000
5000
3000
2000
1500
1000
Body wall muscle
Pharynx
Intestine
Hypodermis
Seam cell
Neuron
D
Number of genes detected
3000
2000
1000
700
500
300
Body wall muscle
Pharynx
Intestine
Hypodermis
Seam cell
Neuron

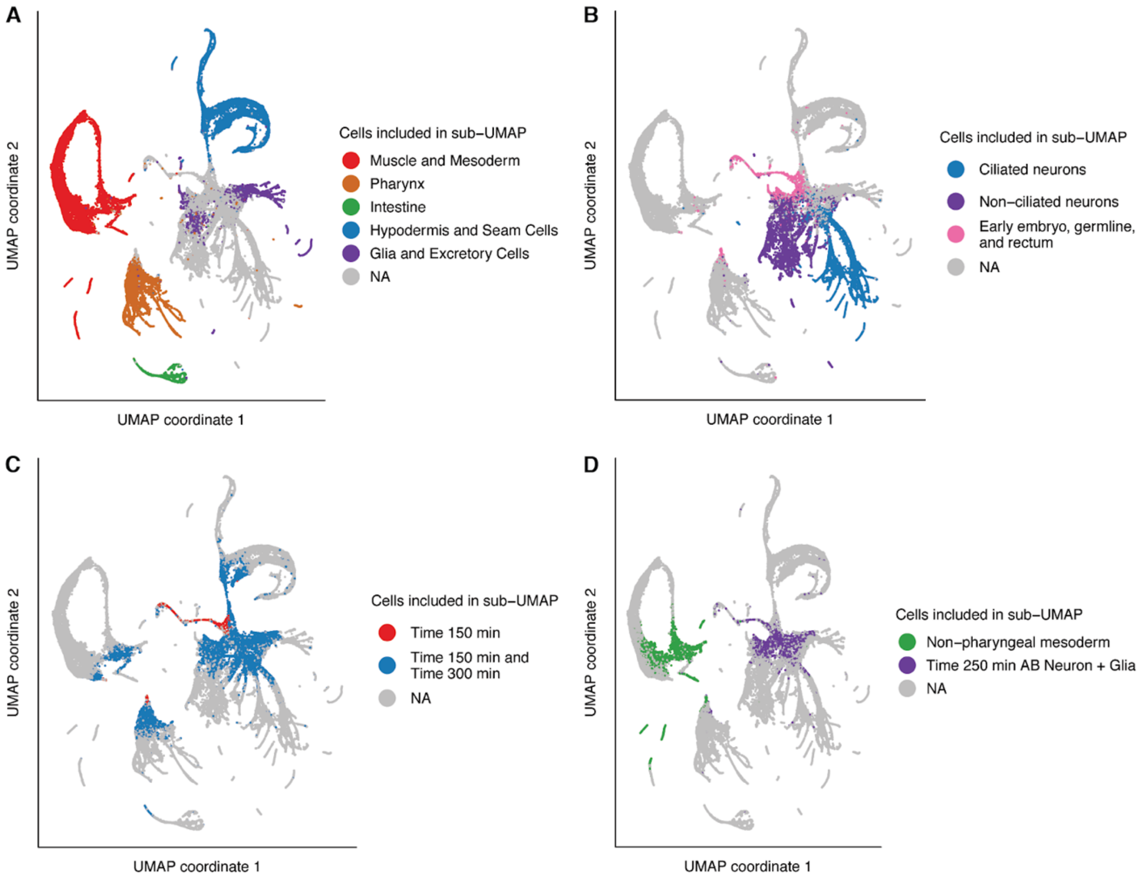
Supplemental Figure 2.2 UMIs recovered per cell decreases with embryo age.

All Y-axes are log scaled. **(A)** Distributions of number of UMIs recovered per cell, binned by estimated embryo age. Median UMIs per cell decreases until ~400 minutes, after which almost all cell division has stopped. Comparing each embryo time bin on the X-axis to the subsequent bin, e.g. comparing 100-150 minutes to 150-200 minutes, the decrease in median UMIs per cell is statistically significant for each step from 100-400 minutes (Wilcoxon rank sum tests, all p-values < 2.2e-16). Note that our quality control procedures exclude cells with < 700 UMIs (or < 500 UMIs for neurons), causing the decrease in UMIs/cell to be understated, as the proportion of cells falling below the cutoff is greater for later stage embryos. **(B)** Number of cells included in each time bin from panel A. **(C** and **D)** Number of UMIs and genes detected for cells with embryo time in the range of 390-650 minutes, by tissue.



Supplemental Figure 2.3 Cell type annotations for the global UMAP of 81,286 cells.

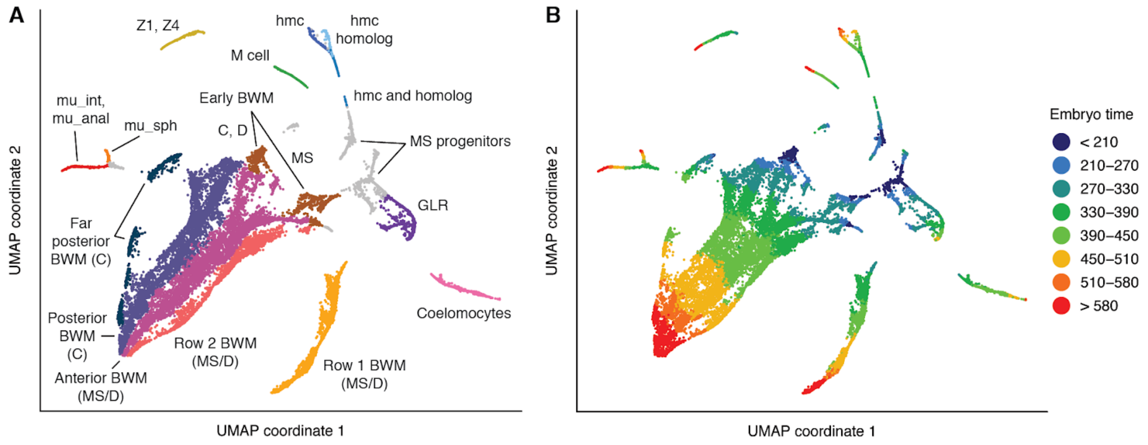
This plot shows more cell type annotations for the global UMAP from Figure 2.1A. This UMAP does not include 4,738 additional cells that were initially filtered, but were later whitelisted and included in downstream analyses (see **Materials and Methods**). For fine-grained annotations of cell types in each major tissue, see Figure 2.3A and Supplemental Figures 2.5-2.13. For fine-grained annotations of progenitor cell lineages, see Supplemental Figures 2.14-2.17.



Supplemental Figure 2.4 Cells included in each sub-UMAP.

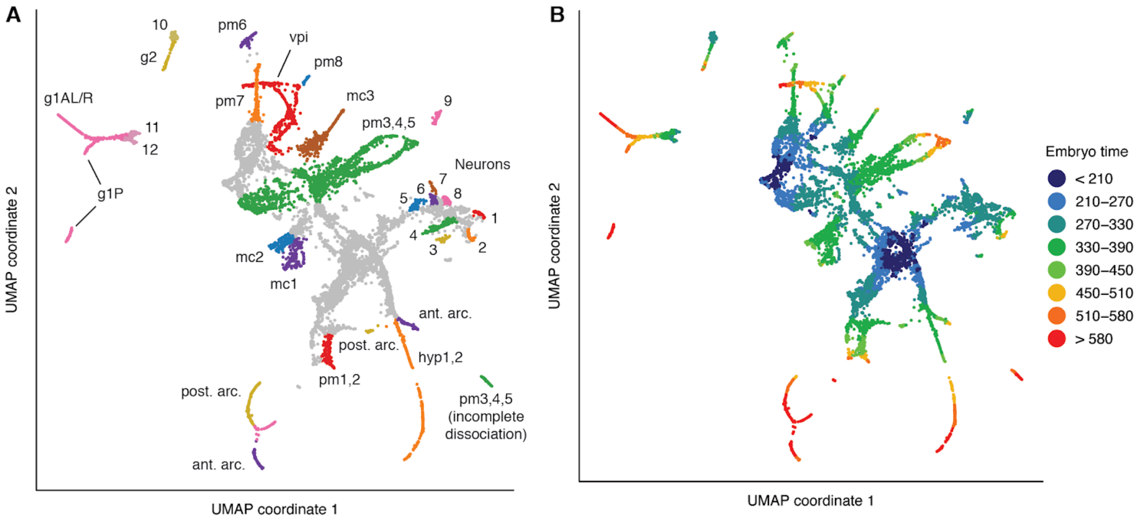
Plots show which cells from the global UMAP (Supplemental Figure 2.3) are included in each sub-UMAP (Supplemental Figures 2.5-2.17), including UMAPs aimed at visualizing terminal cell types **(A, B)** and UMAPs focused aimed at visualizing progenitor lineages **(C, D)**. Note that the actual assignment of cells to sub-UMAPs was performed based on a 3D version of the global UMAP (not shown). In **(C)**, all cells included in the Time 150 min. sub-UMAP are also included in the Time 300 min. sub-UMAP.

Note: The figures below show UMAPs of muscle and the non-pharyngeal mesoderm (Supplemental Figure 2.5), pharynx (Supplemental Figure 2.6), intestine (Supplemental Figure 2.7), hypodermis and seam cells (Supplemental Figure 2.8), glia and excretory cells (Supplemental Figure 2.9), non-ciliated neurons (Supplemental Figure 2.10), touch receptor neurons (Supplemental Figure 2.11), germline (Supplemental Figure 2.12), and rectum (Supplemental Figure 2.13). A UMAP of ciliated neurons is shown in the main text (Figure 2.3A). UMAPs focused on annotating progenitor lineages are shown in Supplemental Figures 2.14-2.17.



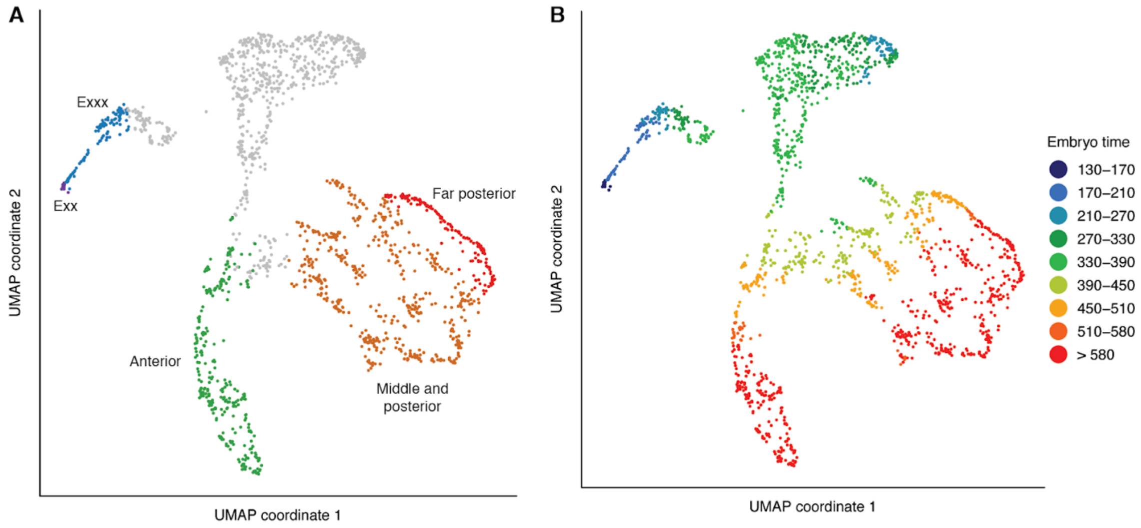
Supplemental Figure 2.5 UMAP of 22,371 body wall muscle and non-pharyngeal mesoderm cells.

(A) Labels indicate cell types. See Table 2.1 for marker genes used to annotate cell types. MS, C, and D indicate cell lineages. Abbreviations: BWM = body wall muscle, mu_int = intestinal muscle, mu_anal = anal depressor muscle, mu_sph = anal sphincter muscle, hmc = head mesodermal cell. **(B)** Colors show estimated embryo times (minutes post first cleavage) for each cell.



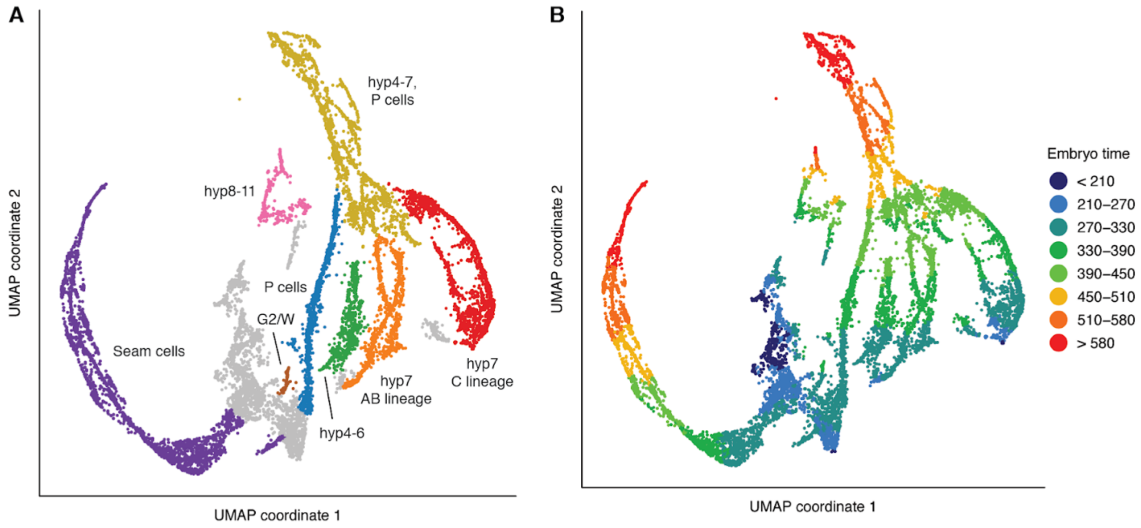
Supplemental Figure 2.6 UMAP of 10,784 pharyngeal cells.

(A) Labels indicate cell types. Abbreviations: pm = pharyngeal muscle, mc = pharyngeal marginal cell, g1A/g1P/g2 = pharyngeal gland, vpi = pharyngeal-intestinal valve, hyp = hypodermis, ant. arc. = anterior arcade cells, post. arc. = posterior arcade cells. Anterior and posterior arcades from late embryos converge in the UMAP to a common transcriptomic profile (pink cells at the bottom of the plot). Numeric labels indicate: **1** parent of NSM **2** MC **3** parent of MI and pm1DR **4** grandparent of I2 **5** parent of M1 **6** parent of M2 and M3 **7** parent of M5 and I6 **8** parent of I1 **9** parent of M4 **10** parent of g2 **11** parent of g1P and I3 **12** parent of g1A. **(B)** Colors show estimated embryo times (minutes post first cleavage) for each cell.



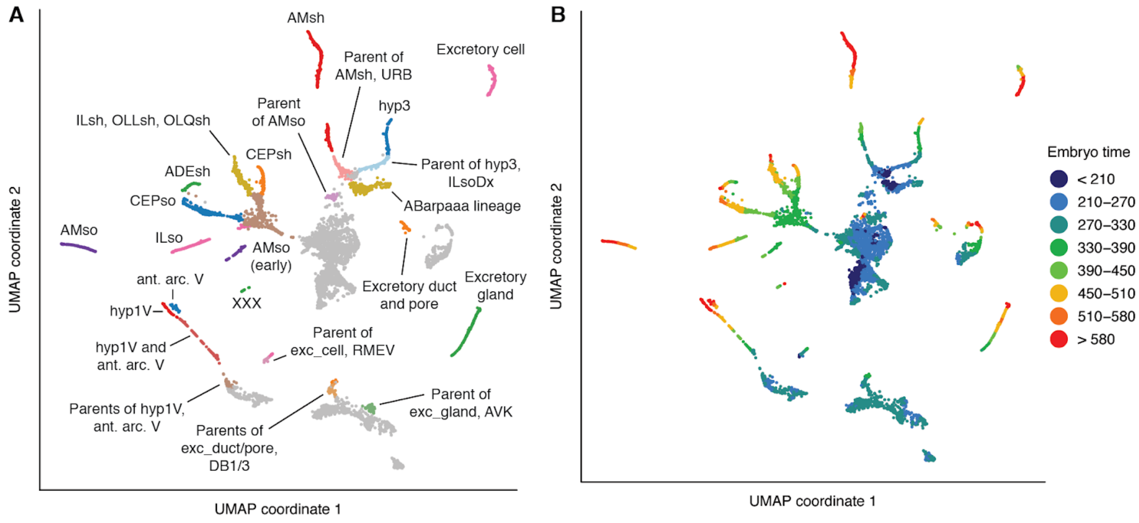
Supplemental Figure 2.7 UMAP of 1,734 intestine cells.

(A) Labels indicate subsets of intestine cells and their relative position on the anterior-posterior axis. See Table 2.1 for marker genes used to annotate cell types. **(B)** Colors show estimated embryo times (minutes post first cleavage) for each cell.



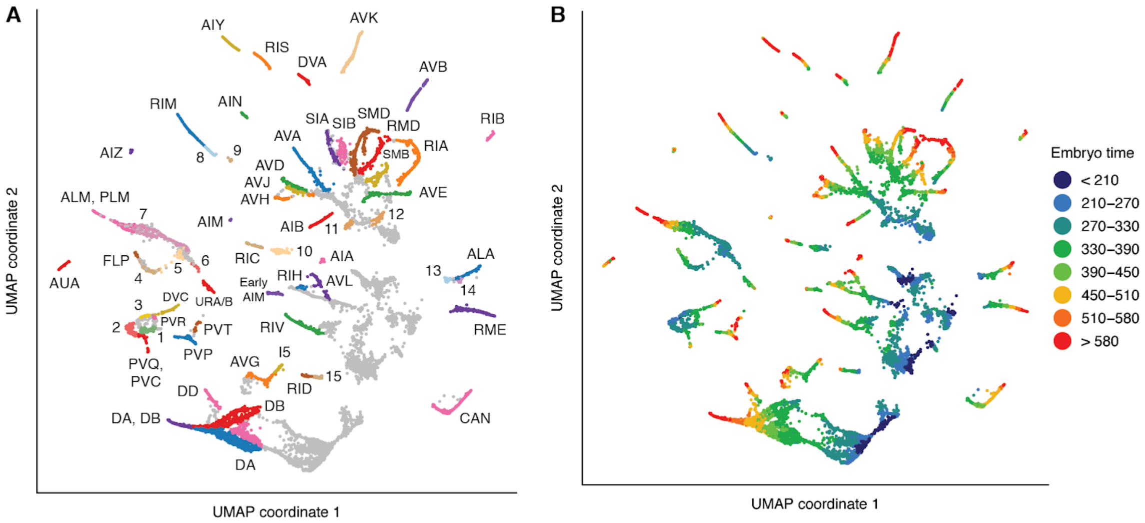
Supplemental Figure 2.8 UMAP of 12,254 hypodermis and seam cells.

(A) Labels indicate cell types. See Table 2.1 for marker genes used to annotate cell types. hyp1-3 are not included here. hyp1-2 appear in the pharynx UMAP (Supplemental Figure 2.6), and hyp3 appears in the glia UMAP (Supplemental Figure 2.9), consistent with their cell lineage (hyp1-2 are sisters/cousins of arcade cells, and hyp3 are sisters of ILsoDx). **(B)** Colors show estimated embryo times (minutes post first cleavage) for each cell.



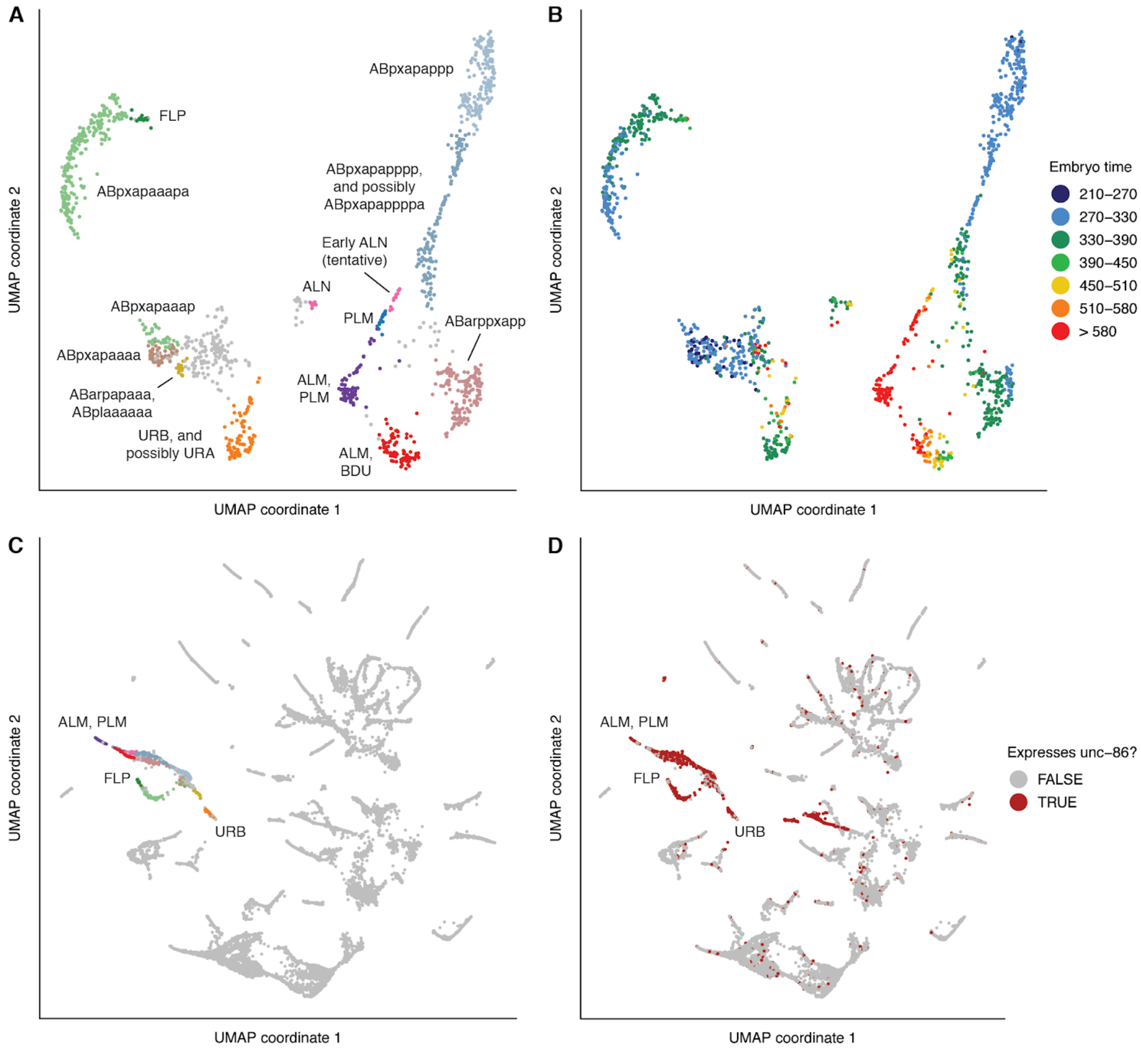
Supplemental Figure 2.9 UMAP of 7,512 glia, excretory cells, and progenitors.

(A) Labels indicate cell types. Some non-glial/excretory cells are also included in the UMAP, such as neuron/glia/rectal progenitors. **(B)** Colors show estimated embryo times (minutes post first cleavage) for each cell.



Supplemental Figure 2.10 UMAP of 14,728 non-ciliated neurons and progenitors.

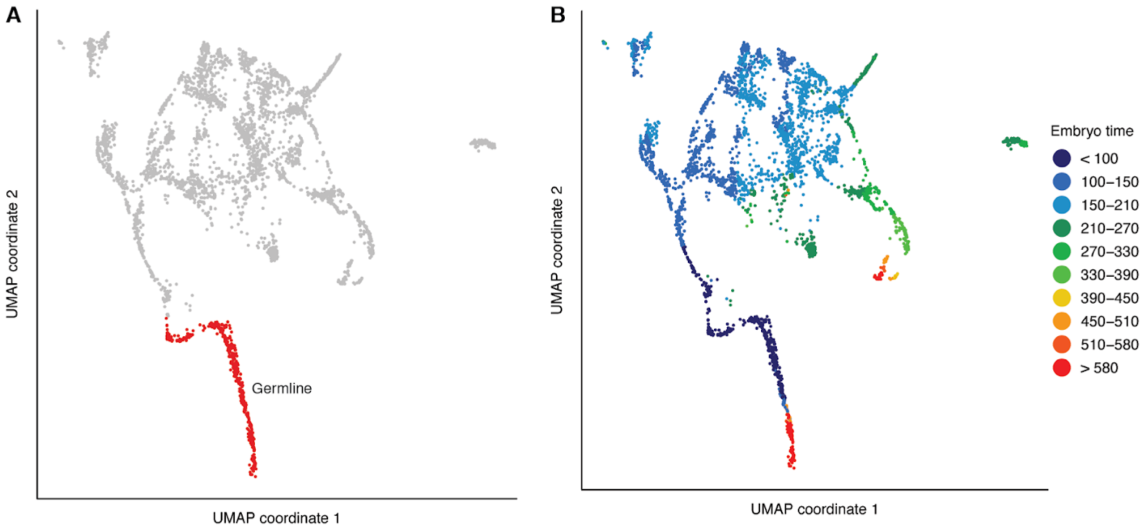
For a UMAP of ciliated neurons, see Figure 2.3A. **(A)** Text labels indicate terminal cell types. Numeric labels indicate: **1** PVC-LUA neuroblast **2** parent of PVQ **3** parent of DVC **4** FLP-AIZ neuroblast **5** FLP-AIZ-RMG neuroblast **6** parent of URADx **7** progenitors of ALM, BDU, PLM, and ALN (see Supplemental Figure 2.11 for a UMAP of the touch receptor lineages) **8** parent of RIM **9** AVG-RIR neuroblast **10** parent of RIC **11** parent of AVH **12** parent of RIA **13** ALA-RMED neuroblast **14** RMED, early after parent's division **15** parent of RID. **(B)** Colors show estimated embryo times (minutes post first cleavage) for each cell.



Supplemental Figure 2.11 UMAP of 1,300 touch receptor neurons, URB neurons, and progenitors.

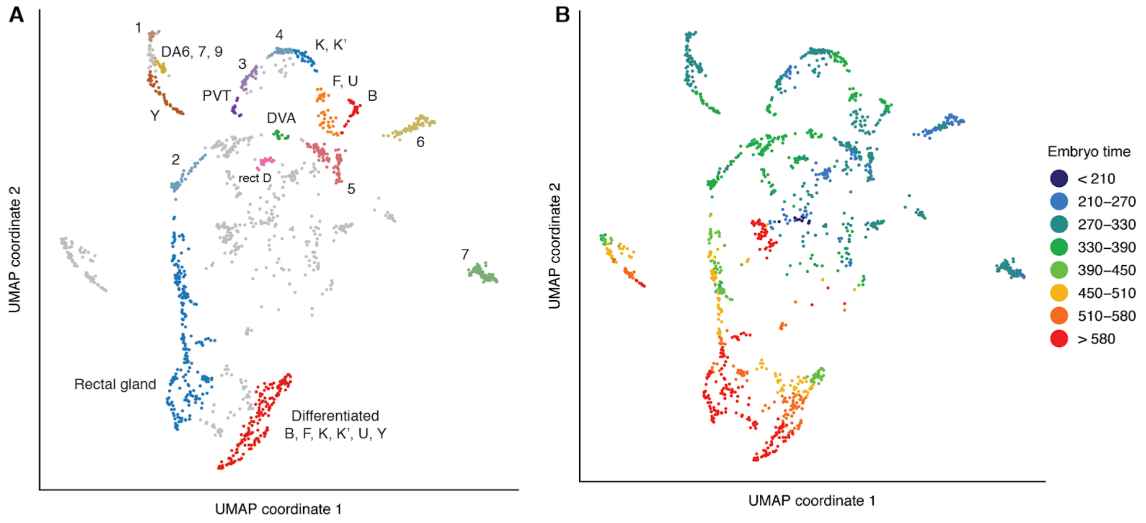
URB neurons are included because they cluster near the touch receptors in the UMAP of all non-ciliated neurons (Supplemental Figure 2.10). This is in part due to high *unc-86* expression. **(A)** Labels indicate cell type (for terminal cells) or lineage (for progenitors). **(B)** Colors show estimated embryo times (minutes post first cleavage) for each cell. **(C)** Location of cells shown in panel A on the UMAP of all non-ciliated neurons from

Supplemental Figure 2.10. **(D)** Expression pattern of *unc-86* on the UMAP of all non-ciliated neurons. Both touch receptor lineages and URB express high levels of *unc-86*.



Supplemental Figure 2.12 UMAP of 3,476 early embryo, germline, and rectal cells.

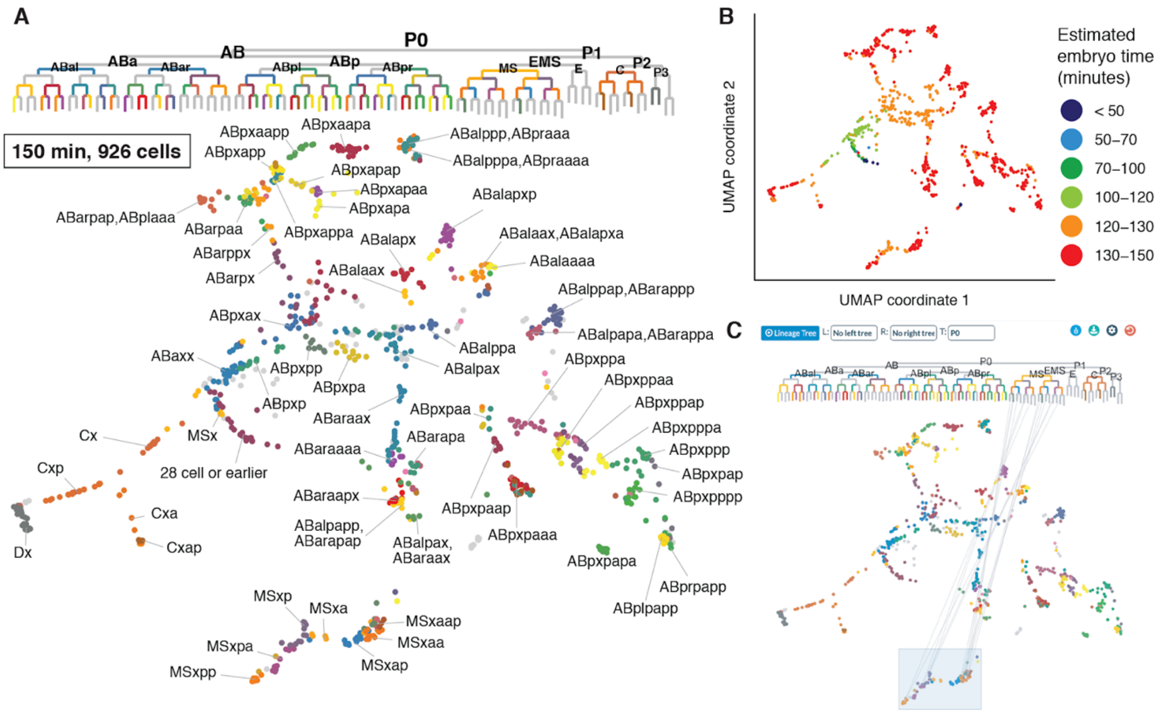
This UMAP was used only for its trajectory of germline development (500 cells). Other lineages that are included in this UMAP were better resolved in other UMAPs, shown below. **(A)** Germline cells highlighted in red. **(B)** Colors show estimated embryo times (minutes post first cleavage) for each cell. These estimates, which are based on correlation to a whole-embryo bulk RNA-seq time series, are inaccurate for germline cells, as genes that follow the same temporal dynamics for all somatic cells often have different expression dynamics in the germline.



Supplemental Figure 2.13 UMAP of 1,598 rectal cells and progenitors.

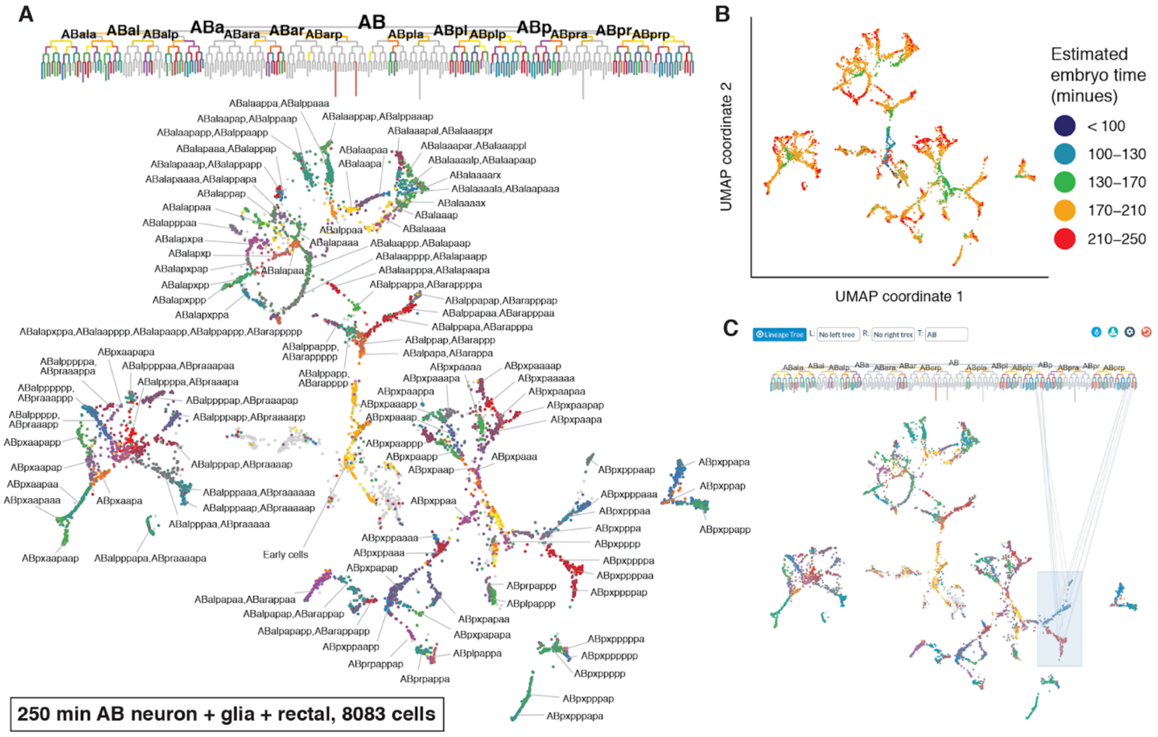
(A) Text labels indicate terminal cell types. Numeric labels indicate: **1** parents of (Y and DA7) and (DA6 and DA9). **2** parent of PVP and rect_V **3** parent of PVT and rect_D **4** parent of K and K' **5** parents of (B and DVA) and (F and U) **6** Parent of the tail spike cells and hyp10 **7** Parent of PHsh and hyp8/9. **(B)** colors show estimated embryo times (minutes post first cleavage) for each cell. The cluster of cells from late embryos (>580 minutes) in the center of the UMAP are AMsh (glia, not rectal cells) that were included in this UMAP by mistake.

Note: Supplemental Figures 2.14-2.17 show a representative subset of the UMAPs that were used to annotate progenitor lineages. Several additional UMAPs can be visualized in VisCello (https://cello.shinyapps.io/celegans_explorer/).



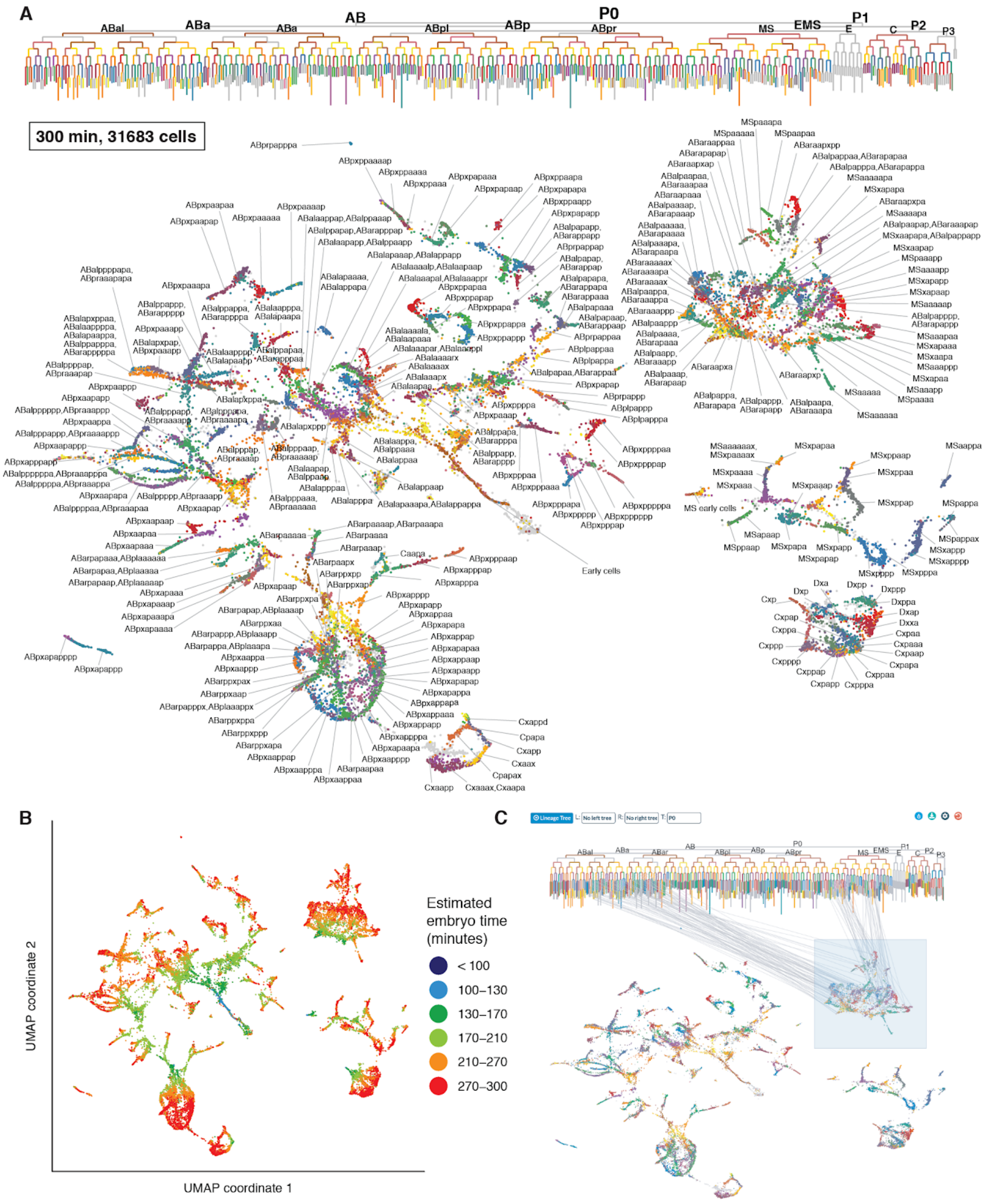
Supplemental Figure 2.14 UMAP and detailed annotation of 926 cells from embryos < 150 minutes post first cleavage.

E lineage and germline cells are excluded from the UMAP and were analyzed separately (Supplemental Figure 2.7 and Supplemental Figure 2.12). **(A)** Detailed labeling of lineages, co-visualized with the lineage tree. **(B)** Colors show estimated embryo times (minutes post first cleavage) for each cell. **(C)** Screenshot of an interactive co-visualization implemented in VisCello (https://cello.shinyapps.io/celegans_explorer/), highlighting the connection between MS lineage clusters and corresponding leaves in the lineage tree.



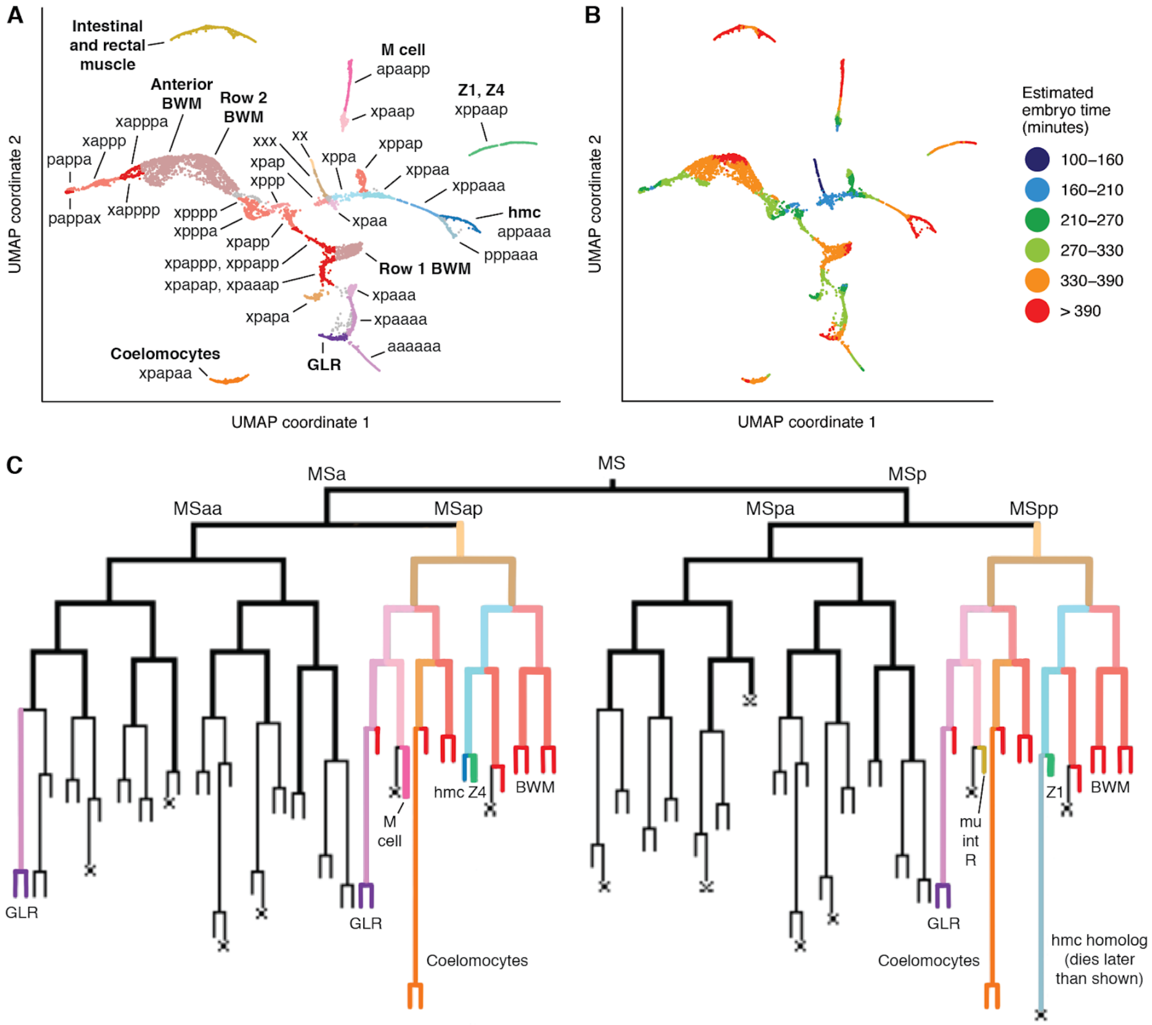
Supplemental Figure 2.15 UMAP and detailed annotation of 8,083 AB lineage neuron/glia/rectal progenitor cells from embryos < 250 minutes post first cleavage.

This UMAP includes only AB lineage cells that give rise to neurons, glia, and rectal cells. **(A)** Detailed labeling of lineages, co-visualized with the AB lineage tree. **(B)** Colors show estimated embryo times (minutes post first cleavage) for each cell. **(C)** Screenshot of an interactive co-visualization implemented in VisCello (https://cello.shinyapps.io/celegans_explorer/), highlighting the connection between ABpxppp lineage clusters and corresponding leaves in the lineage tree.



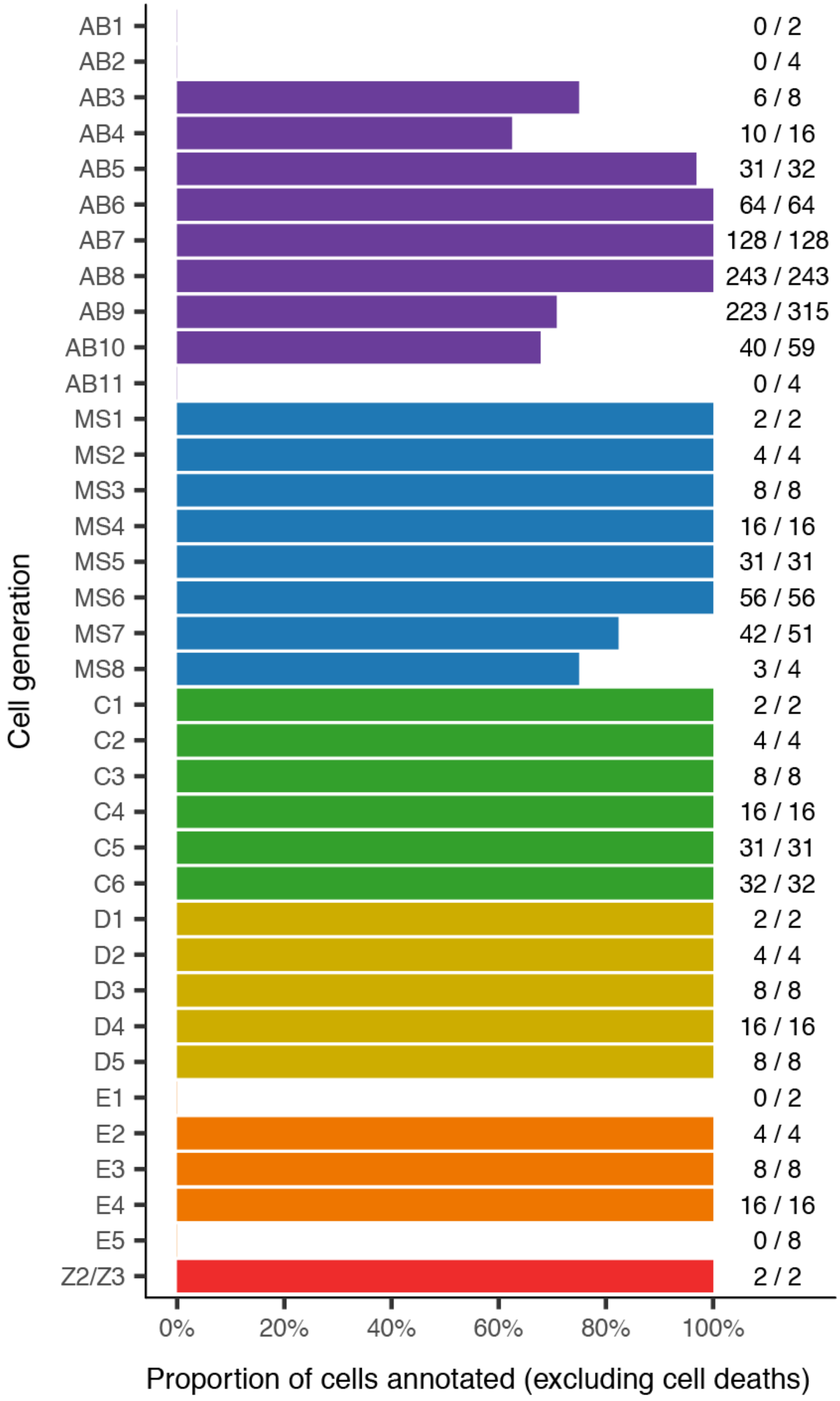
Supplemental Figure 2.16 UMAP and detailed annotation of 31,683 cells from embryos < 300 minutes post first cleavage.

E lineage and germline cells are excluded from the UMAPs and were analyzed separately (Supplemental Figure 2.7 and Supplemental Figure 2.12). **(A)** Detailed labeling of lineages, co-visualized with the lineage tree. **(B)** Colors show estimated embryo times (minutes post first cleavage) for each cell. **(C)** Screenshot of an interactive co-visualization implemented in VisCello (https://cello.shinyapps.io/celegans_explorer/), highlighting the connection between the pharynx cluster in the UMAP and the corresponding leaves in the lineage tree. All cells in the pharynx cluster are annotated as descendants of the ABalp, ABara and MS lineages, consistent with previous observations that pharyngeal cells only arise from these lineages.



Supplemental Figure 2.17 UMAP of 8,233 non-pharyngeal mesoderm cells, focused on the early lineage.

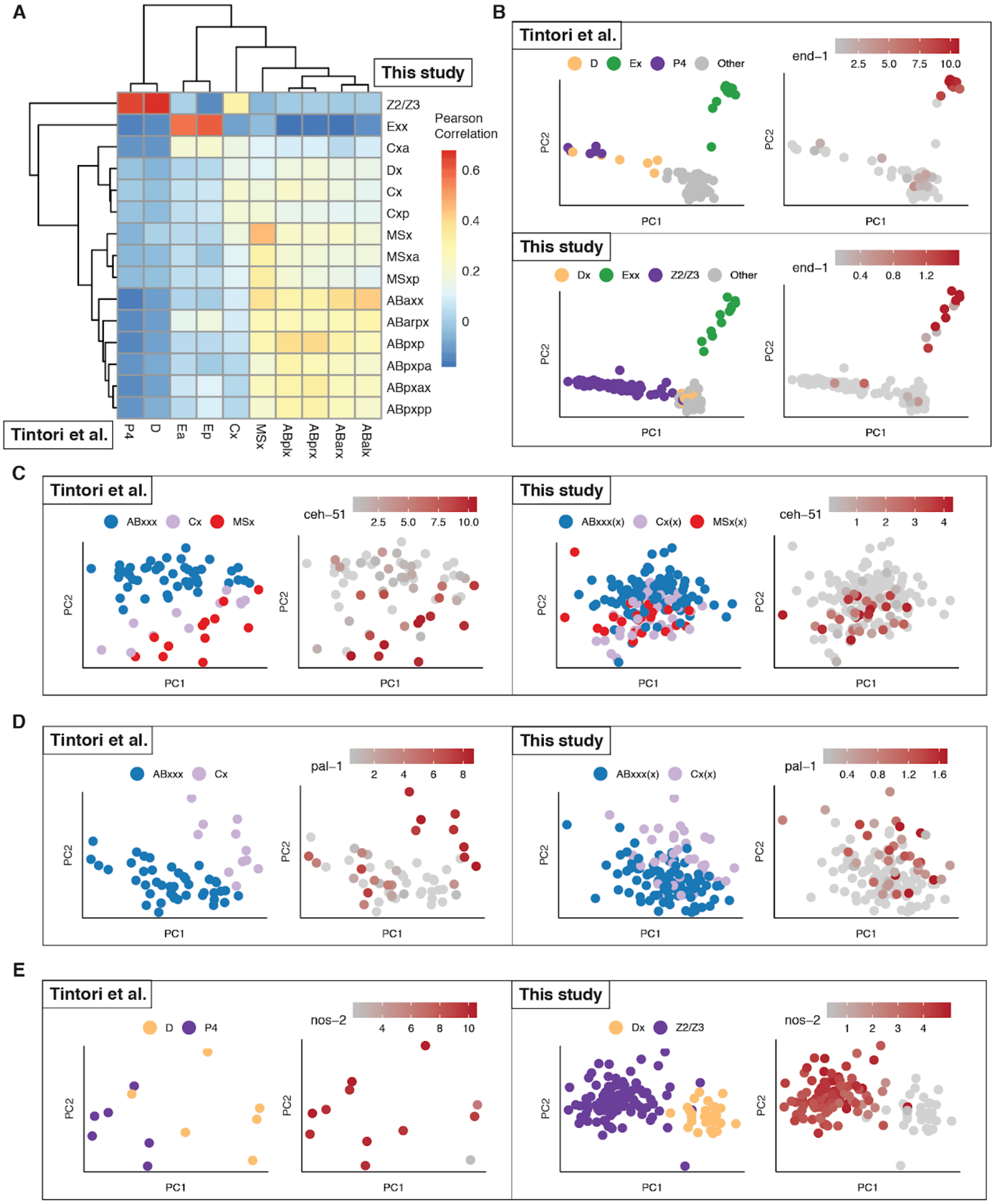
This UMAP includes the same cells as the muscle and mesoderm UMAP (Supplemental Figure 2.5), but excludes putative C and D lineage body wall muscle, MS lineage body wall muscle with estimated embryo time >400 minutes (post first cleavage), and coelomocytes with embryo time >400 minutes. This UMAP serves as a representative example of a set of several UMAPs used to connect terminal cells to their immediate progenitors. Additional UMAPs can be viewed in VisCello. **(A)** Text labels indicate MS lineages (i.e. “xppa” = MSxppa). Bold text labels indicate cell types. MSxppapx was not conclusively identified, but is presumed to be included in the head BWM cluster. **(B)** Estimated embryo time for each cell. **(C)** diagram of the MS lineage. Colored sub-lineages match the colors of cell groups in panel (A).


Cell generation
AB1
AB2
AB3
AB4
AB5
AB6
AB7
AB8
AB9
AB10
AB11
MS1
MS2
MS3
MS4
MS5
MS6
MS7
MS8
C1
C2
C3
C4
C5
C6
D1
D2
D3
D4
D5
E1
E2
E3
E4
E5
Z2/Z3
0 / 2
0 / 4
6 / 8
10 / 16
31 / 32
64 / 64
128 / 128
243 / 243
223 / 315
40 / 59
0 / 4
2 / 2
4 / 4
8 / 8
16 / 16
31 / 31
56 / 56
42 / 51
3 / 4
2 / 2
4 / 4
8 / 8
16 / 16
31 / 31
32 / 32
2 / 2
4 / 4
8 / 8
16 / 16
8 / 8
0 / 2
4 / 4
8 / 8
16 / 16
0 / 8
2 / 2
0%
20%
40%
60%
80%
100%
Proportion of cells annotated (excluding cell deaths)

Supplemental Figure 2.18 Summary of lineage annotations.

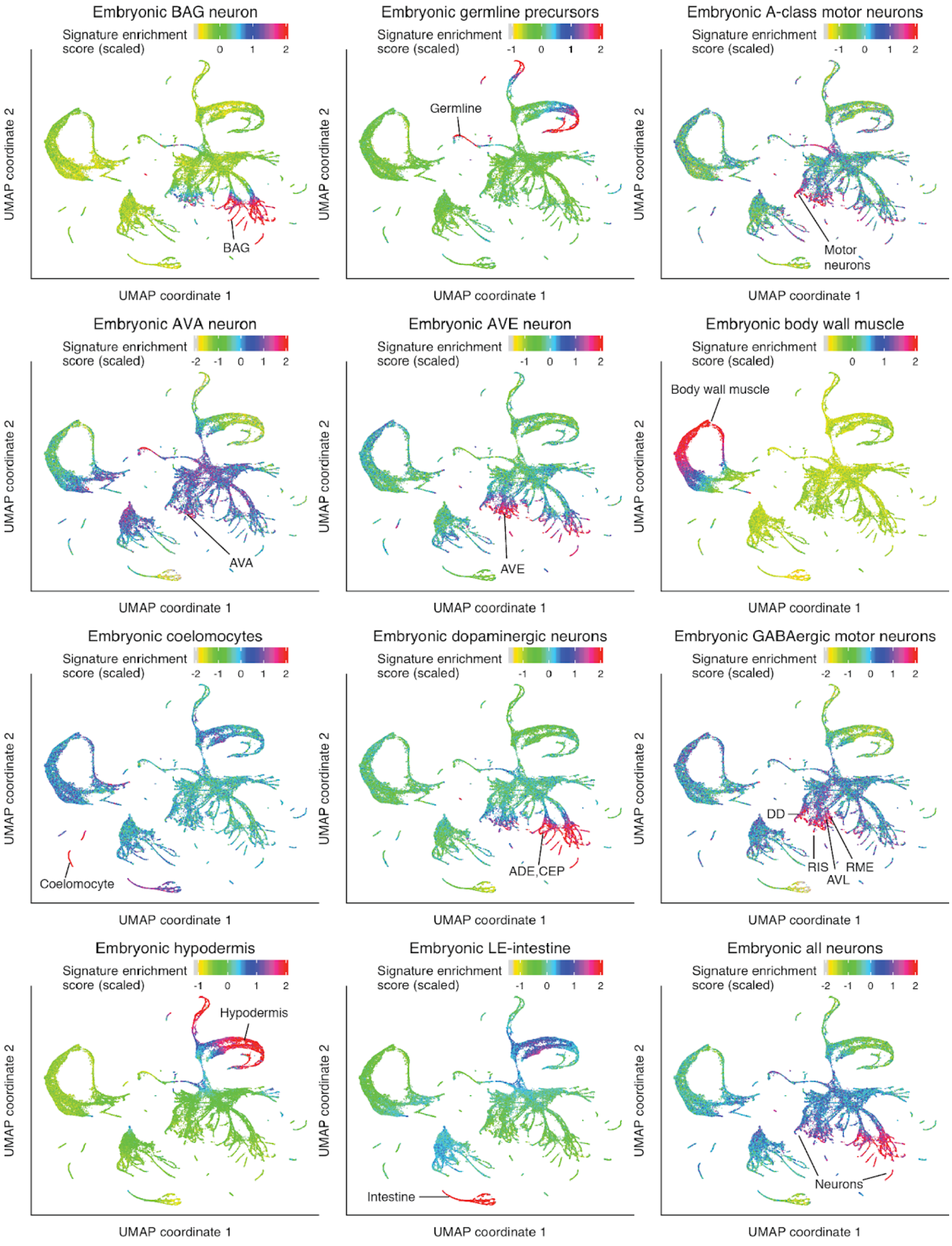
Each row corresponds to a subset of cells in the *C. elegans* embryonic cell lineage. Row labels consist of one or two letters, which identify a broad lineage (AB, MS, C, D, or E), and a number, which specifies the number of cell divisions since the founding cell of the broad lineage. For example, “AB5” refers to the 32 cells produced by 5 divisions of the AB founder cell, and “C2” refers to the 4 cells produced by 2 divisions of the C founder cell. The founder cells themselves are not included in the plot. The label “Z2/Z3” is an exception to the nomenclature and refers to the two germline lineages, Z2 and Z3.

Bar lengths indicate the percent of cells within the specific lineage and cell generation specified by the row label that are included in our annotations of our single cell RNA-seq dataset. Lineages that undergo programmed cell death are excluded from the statistics. Numbers to the right of the bars indicate the absolute number of lineages annotated and the total number of lineages present within a particular cell generation.



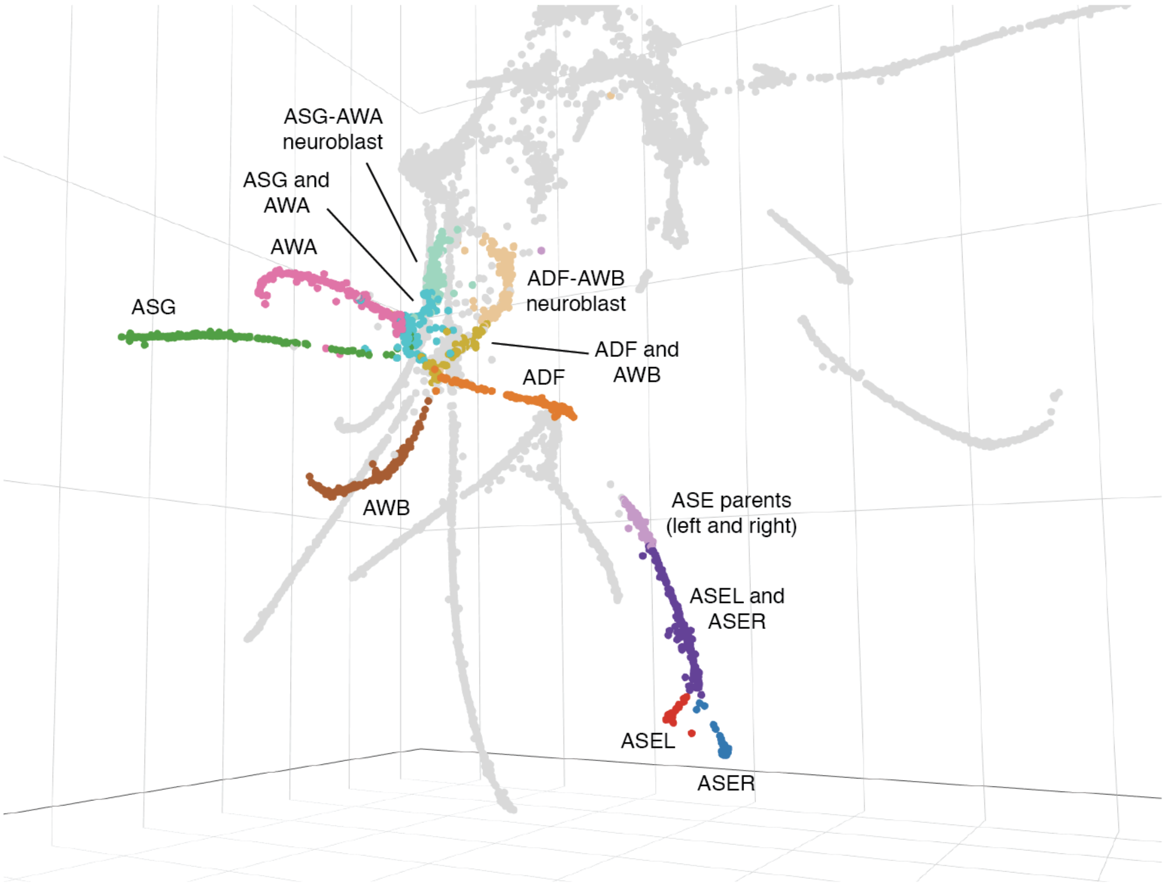
Supplemental Figure 2.19 Comparison of data from this study to data from Tintori *et al.*, 2016 (*19*).

Tintori *et al.* (*19*) profiled the transcriptomes of single cells from the *C. elegans* 1- to 16-cell stages. **(A)** Heatmap showing Pearson correlations between the log2-scaled gene expression profiles of 16-cell stage cells from Tintori *et al.* (*19*) vs. 16- and 28-cell stage cells from this study. Correlation was computed using informative genes selected by an iterative PCA approach used by Tintori *et al.* (*19*) (see **Materials and Methods**). **(B-E)** First sub-panel shows a PCA projection computed using 16-cell stage cells from Tintori *et al.* (*19*), reproducing their original analysis. Second sub-panel shows a projection of 16- and 28-cell stage cells from this study into the same PCA space. Each PCA uses a different set of informative genes, as originally defined by Tintori *et al.* (*19*), to discriminate particular lineages (see **Materials and Methods**). For each PCA, the gene expression level of a selected lineage-specific marker gene was plotted. Gene expression is measured in log2 RPKM for data from Tintori *et al.* (*19*), and log2 size-factor normalized UMI counts for data from this study.


Embryonic BAG neuron
Signature enrichment score (scaled)
0 1 2
BAG
UMAP coordinate 1
UMAP coordinate 2
Embryonic germline precursors
Signature enrichment score (scaled)
-1 0 1 2
Germline
Embryonic A-class motor neurons
Signature enrichment score (scaled)
-1 0 1 2
Motor neurons
Embryonic AVA neuron
Signature enrichment score (scaled)
-2 -1 0 1 2
AVA
Embryonic AVE neuron
Signature enrichment score (scaled)
-1 0 1 2
AVE
Embryonic body wall muscle
Signature enrichment score (scaled)
0 1 2
Body wall muscle
Embryonic coelomocytes
Signature enrichment score (scaled)
-2 -1 0 1 2
Coelomocyte
Embryonic dopaminergic neurons
Signature enrichment score (scaled)
-1 0 1 2
ADE,CEP
Embryonic GABAergic motor neurons
Signature enrichment score (scaled)
-2 -1 0 1 2
DD
RIS
RME
AVL
Embryonic hypodermis
Signature enrichment score (scaled)
-1 0 1 2
Hypodermis
Embryonic LE-intestine
Signature enrichment score (scaled)
-1 0 1 2
Intestine
Embryonic all neurons
Signature enrichment score (scaled)
-2 -1 0 1 2
Neurons

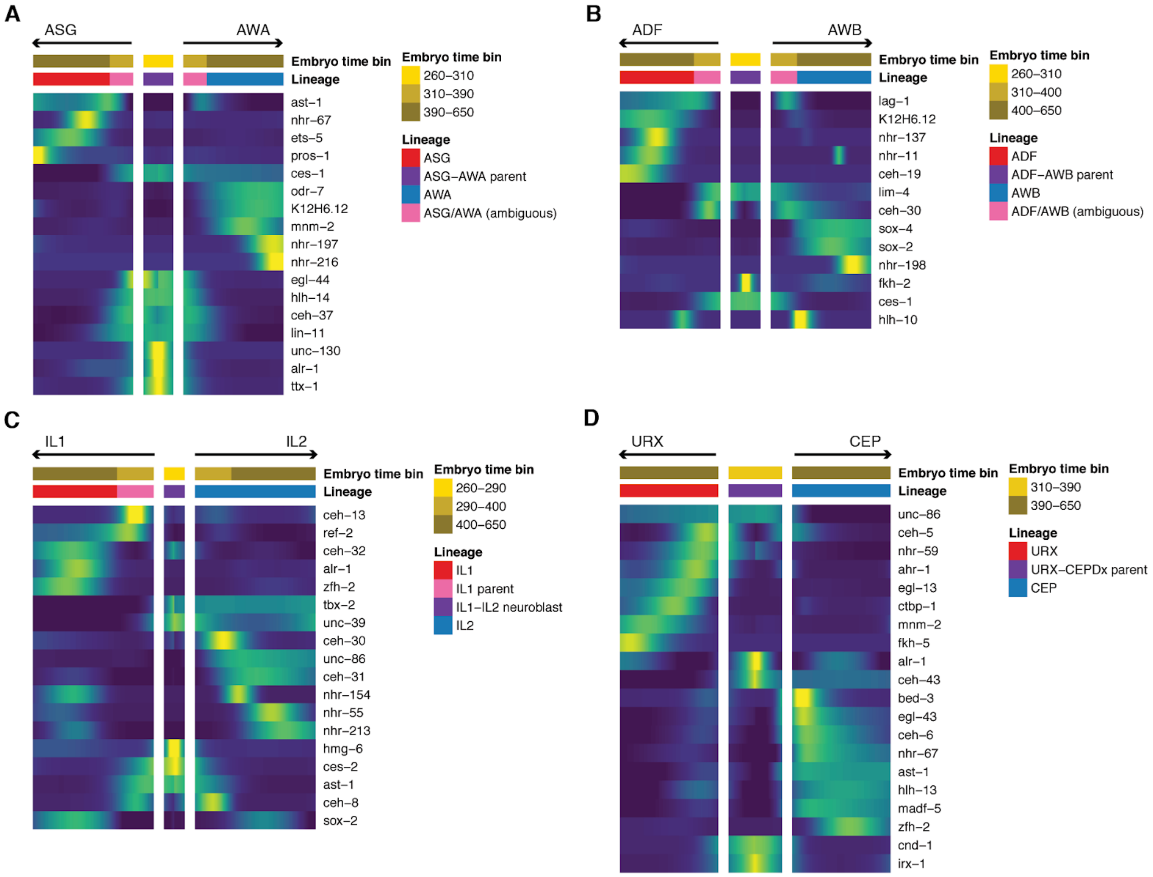
Supplemental Figure 2.20 Comparison of data from this study to microarray data from Spencer *et al.*, 2011 (*191*).

Each panel shows a global UMAP of cells from this study, colored by a score that measures the extent to which each single-cell transcriptome is enriched for genes from a particular gene set reported by Spencer *et al.* (*191*). Signature gene sets from Spencer *et al.* (*191*) were downloaded from https://www.vanderbilt.edu/wormdoc/wormmap/Enriched_genes.html. Each signature gene set corresponds to genes that are enriched in a particular embryonic cell type compared to all other cells in the Spencer *et al.* microarray data (*191*). Signature genes are therefore mostly tissue-specific, rather than cell-type specific. Gene set enrichment scores were computed using the AUCell package (*157*). Comparison with pharyngeal muscle was dropped because most of the signature genes reported in Spencer *et al.* (*191*) for this cell type are intestine specific, as confirmed by a third dataset (*191*). See **Materials and Methods** for more details.



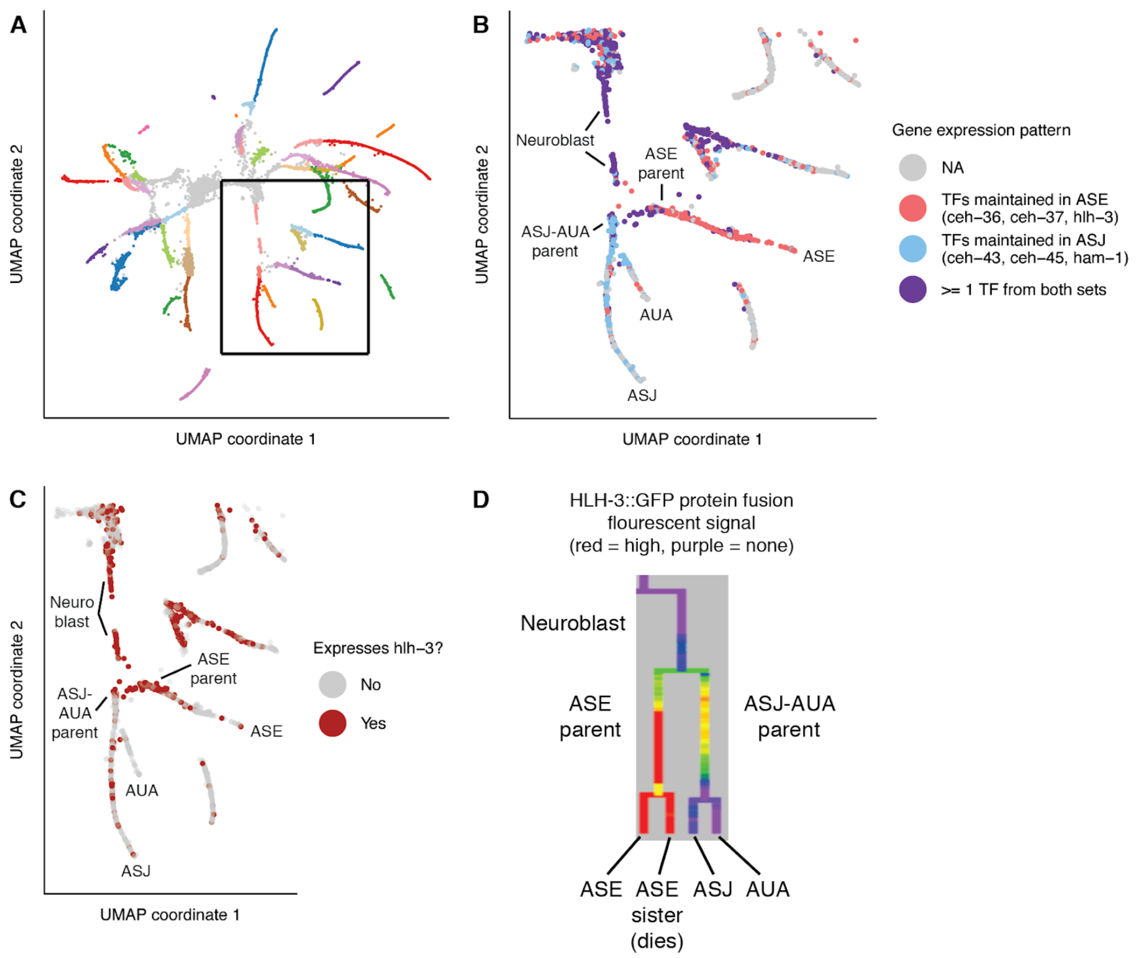
Supplemental Figure 2.21 Ciliated neuron developmental trajectories are more continuous in a 3D UMAP.

This plot is a 2D screenshot of part of a 3D UMAP of ciliated neuron cells, oriented to show specific lineage relationships. The cells are the same as in Figure 2.3A; the only difference is projecting into 3D instead of 2D. Developmental trajectories connecting the ASG-AWA and ADF-AWB neuroblasts to their respective daughter cells are continuous in this UMAP space, as is the branching trajectory of the left and right ASE neurons (ASEL and ASER). In the ASG-AWA and ADF-AWB trajectories, there are sections that appear before the branch points in the UMAP, but based on our embryo time estimates are likely to be terminal cells and not the parent neuroblasts. These sections may contain both daughter cells of each trajectory after their birth but before they differentiate. Cells in the “ADF and AWB” section co-express in the same cells the marker genes *lag-1*, which persists only in ADF, and *lim-4*, which persists only in AWB; however, their estimated embryo times span ~100 minutes after the parent cells’ division time. Note that the grey, unannotated cells below the ADF trajectory are behind the ADF cells in 3D space, as are the grey cells overlapping the AWB trajectory.



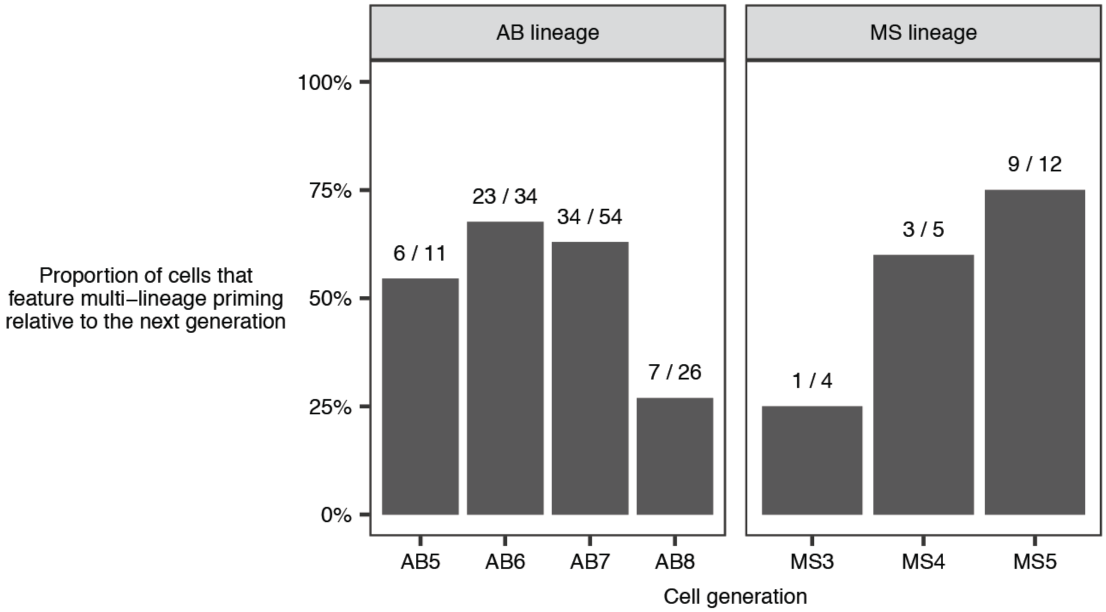
Supplemental Figure 2.22 Differentially expressed transcription factors associated with ciliated neuron lineage branches.

Heatmaps showing patterns of differential transcription factor expression associated with branches in **(A)** the ASG-AWA lineage, **(B)** the ADF-AWB lineage, **(C)** the IL1-IL2 lineage, and **(D)** the URX-CEPDx lineage. A heatmap for the ASE-ASJ-AUA lineage is shown in Figure 2.3D. Expression values are log-transformed, then centered and scaled by standard deviation for each row (gene). In each of the ASG-AWA and ADF-AWB lineages, there is a set of cells that are before the branch point of the trajectory in UMAP space (see Supplemental Figure 2.21), but based on embryo time estimates and marker gene expression patterns, are likely to be terminal cells. In the ADF-AWB lineage, these cells co-express *lag-1*, which is selectively retained in ADF, and *lim-4*, which is selectively retained in AWB, suggesting that this cell set may include undifferentiated, terminal ADF and AWB cells.



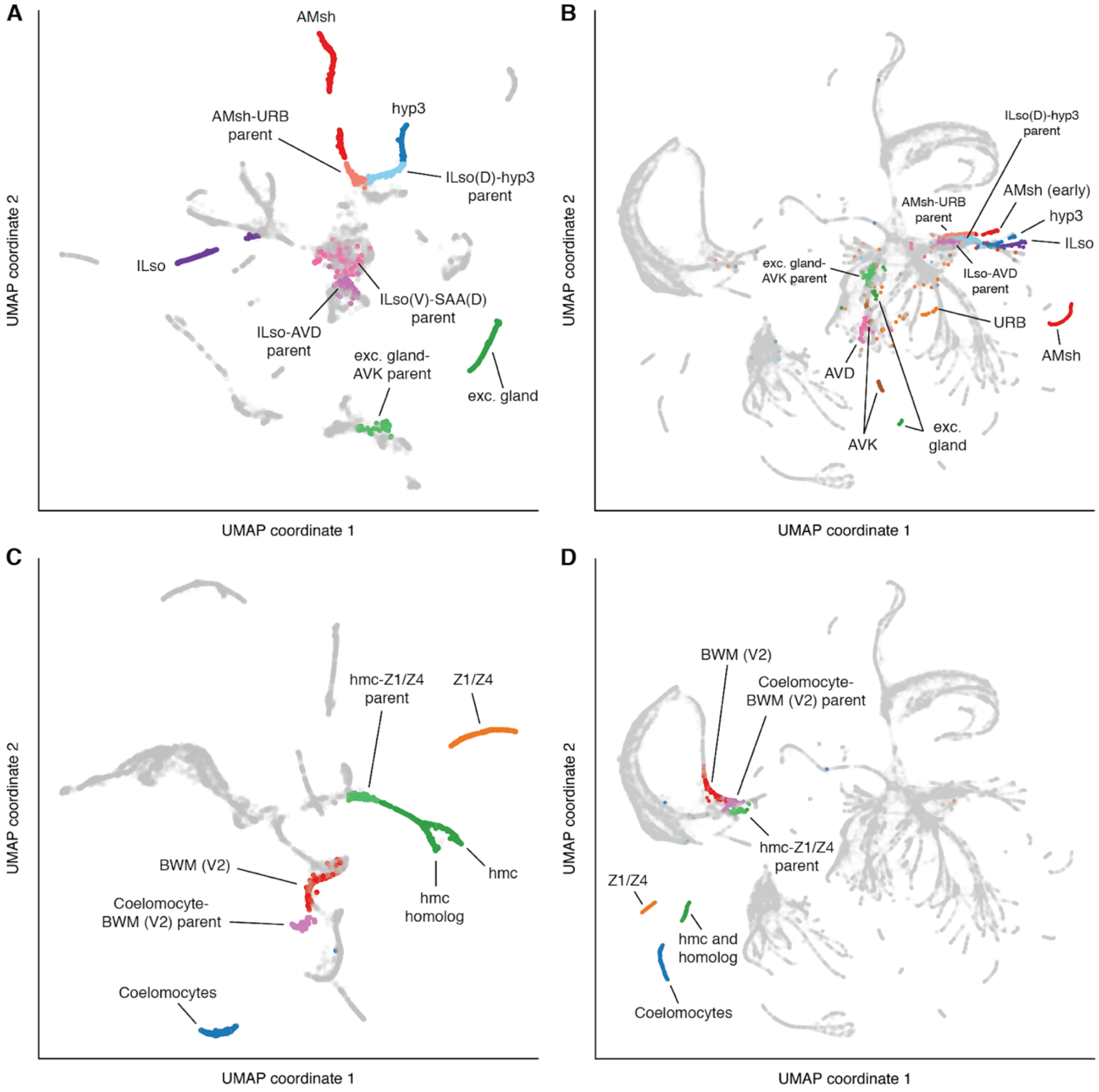
Supplemental Figure 2.23 Multilineage priming in the ASE-ASJ-AUA lineage.

(A) Section of the ciliated neuron UMAP from Figure 2.3A that is shown in panels B and C. This section includes the trajectory of the lineage that produces the ASE, ASJ, and AUA neurons (ABalpppppp/ABpraaappp). **(B)** Expression patterns for transcription factors that are expressed in the ASE-ASJ-AUA neuroblast and selectively maintained in only one of its daughters. Red and blue points indicate cells that express >= 1 TF for which expression is maintained only in the ASE lineage (red) or only the ASJ lineage (blue). Purple points indicate cells that express >= 1 TF from both sets. **(C)** Expression pattern of *hlh-3*, which is expressed in the ASE-ASJ-AUA neuroblast and maintained in the ASE parent but not the ASJ-AUA parent. **(D)** Fluorescent signal from a HLH-3::GFP protein fusion from EPiC (*17*) (series 20160301_hlh-3_OP650_L2). Red indicates high signal, yellow/green indicate medium signal, blue indicates low signal, and purple indicates no signal. Due to translation and the folding time of GFP, the fluorescent signal has a time lag compared to the RNA expression in panel C. The presence of signal in the ASJ-AUA parent indicates that HLH-3 protein does not undergo asymmetric localization during cell division; instead, it is simply maintained in the ASE lineage and allowed to degrade in the ASJ-AUA lineage.



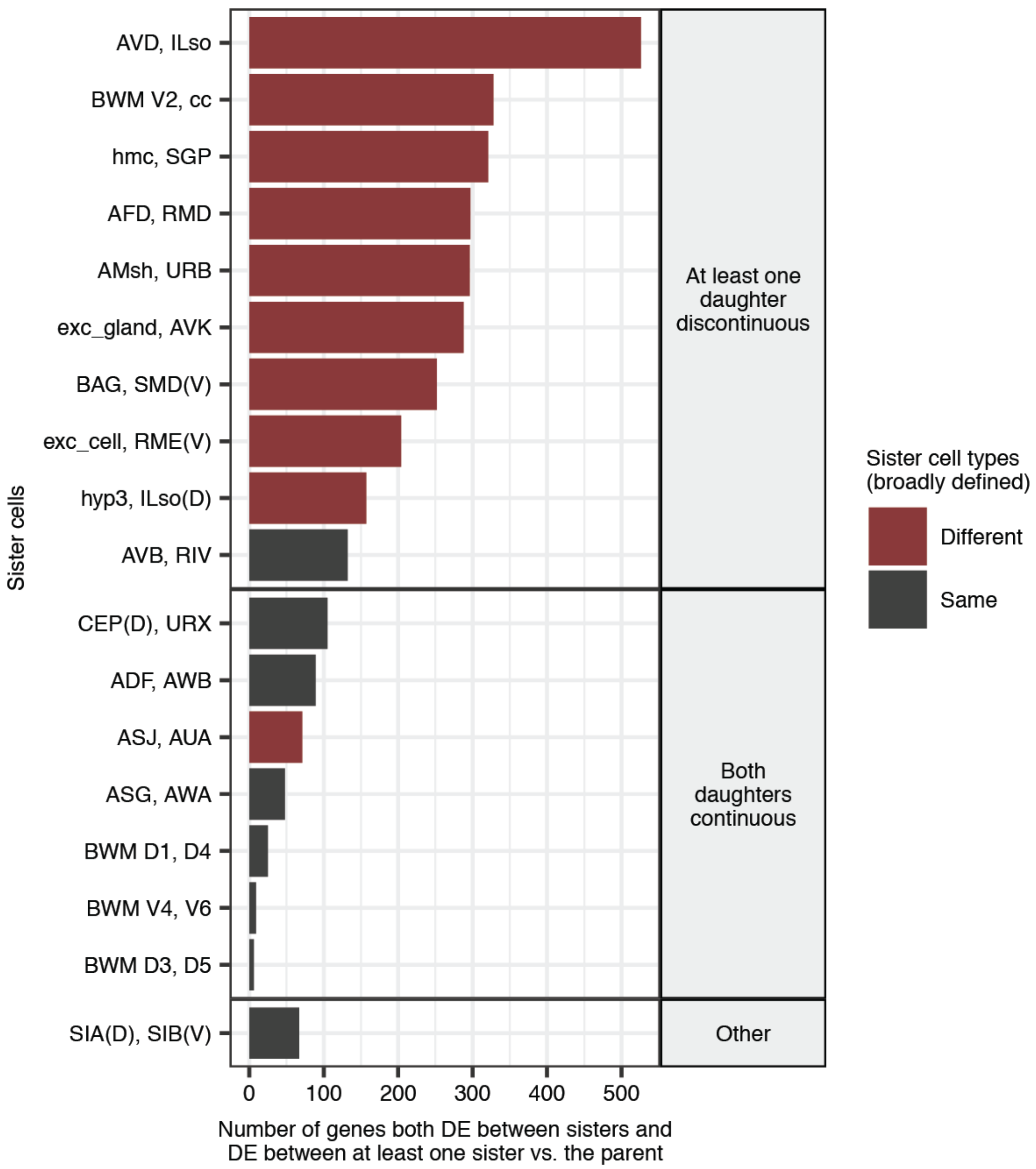
Supplemental Figure 2.24 Prevalence of multilineage priming in *C. elegans*.

X-axis shows different cell generations of the ectoderm (AB lineage) and mesoderm (MS lineage). “AB5” refers to the generation produced by 5 divisions of the AB founder cell, and likewise for AB6-8 and MS3-5. Y-axis shows the proportion of lineages in a given generation that co-express at least one transcription factor (TF) that has expression selectively maintained in one daughter, and at least one TF that has expression selectively maintained in the other daughter (e.g. TF A expressed in parent and daughter 1, TF B expressed in parent and daughter 2). Lineages that satisfy these criteria are considered to exhibit “multilineage priming.” Text labels above each bar indicate the absolute number of lineages in each generation that exhibit multilineage priming (numerator) and the total number of lineages included in the analysis (denominator). Lineages that do not have exactly two, transcriptomically distinct daughters annotated in our dataset are excluded from the statistics. Cell generations that are not shown in this plot were excluded due to having a sample size of <= 3 lineages.



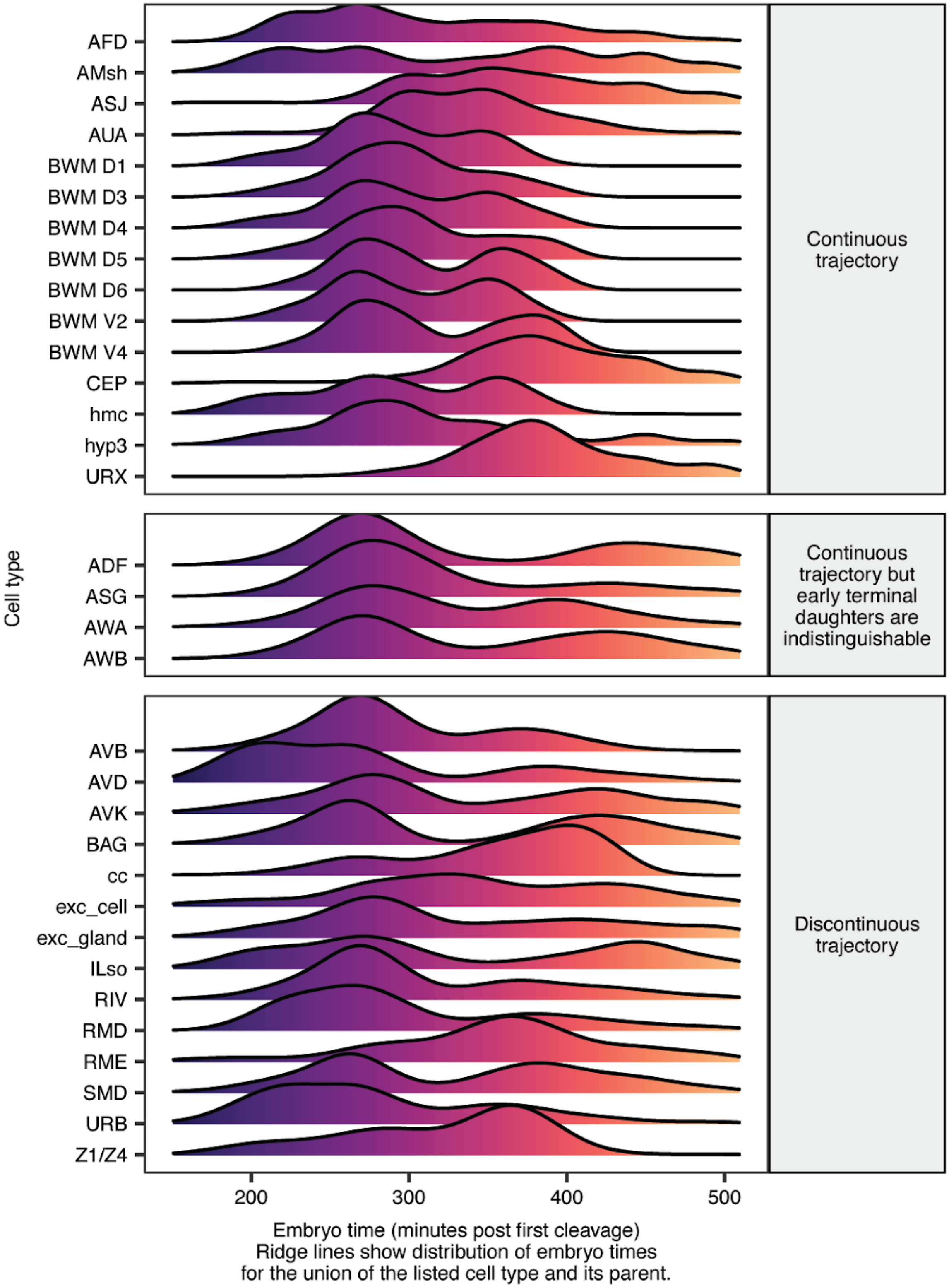
Supplemental Figure 2.25 Examples of lineages that form discontinuous trajectories in UMAP space.

(A) UMAP of 7,512 glia, excretory cells, and progenitors (same as Supplemental Figure 2.9). ILso glia are formed by three input lineages. Two input lineages, the ILso-AVD parent and the ILso(D)-hyp3 parent, form discontinuous trajectories with terminal ILso. Some early terminal ILso cells are likely to be unannotated, so it is not clear if there is a continuous or discontinuous trajectory with the third input lineage, the ILso(V)-SAA(D) parent. **(B)** Global UMAP of 81,286 cells (same as Figure 2.1A). Annotated cell populations are the same as in panel A, plus additional neuron types. The AVD, AVK, and URB neurons are sisters of glia/excretory cells, but form discontinuous trajectories with their parents. **(C)** UMAP of 8,233 non-pharyngeal mesoderm cells (same as Supplemental Figure 2.17). Coelomocytes and Z1/Z4 (the somatic gonad precursors) form discontinuous trajectories with their parents. **(D)** Global UMAP, same as panel B. Annotated cell populations are the same as in panel C.



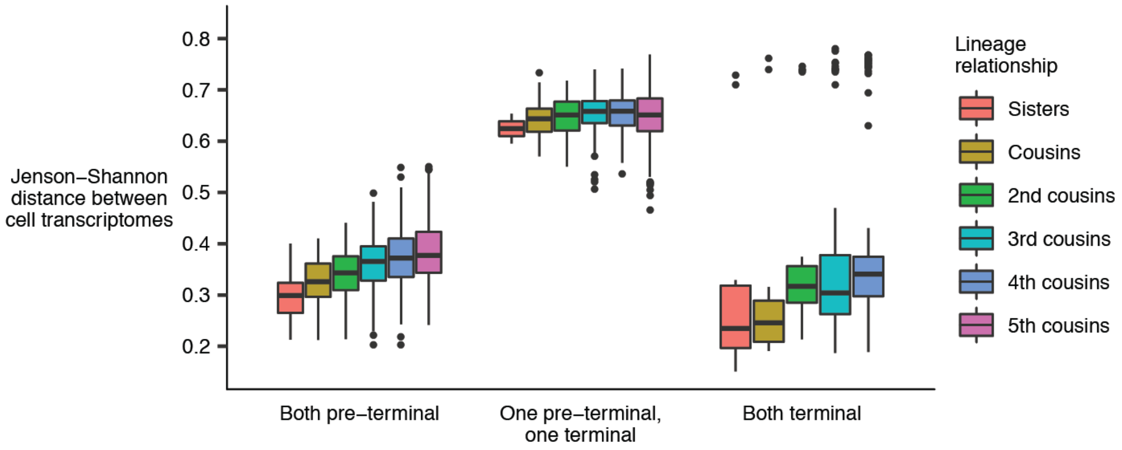
Supplemental Figure 2.26 Counts of differentially expressed genes for lineages that form continuous vs. discontinuous trajectories in UMAP space.

Each row (y-axis) corresponds to a pair of terminal sister cells in the ectoderm (AB lineage, generations 9 and 10) or mesoderm (MS lineage, generation 6). Bar length (x-axis) indicates the number of genes that are both differentially expressed (fold difference > 3, q-value < 0.1) between the sister cells and also differentially expressed (same thresholds) between at least one of the sisters and their parent. Genes that satisfy these criteria are genes that are changing over time in a lineage-specific manner (and therefore exclude broadly expressed genes). Before performing differential expression analysis, the sc-RNA-seq cells that correspond to each of the listed anatomical cells and their parent were downsampled to ensure that each comparison had approximately the same statistical power. Rows are grouped based on whether or not the developmental trajectories formed by the sister cells and their parent in UMAP space were discontinuous for at least one sister. Trajectories were considered discontinuous only if the discontinuity was present in both the global UMAP (Figure 2.1A, Supplemental Figure 2.3) and the relevant tissue UMAP (Figure 2.3A, Supplemental Figures 2.9-2.10, 2.17). Rows are colored to indicate whether or not the sister cells share the same broadly-defined cell type. For example, ASG and AWA, two ciliated neurons, are considered to have the same broadly-defined cell type, while AFD and RMD, a ciliated and non-ciliated neuron respectively, are considered to have different broadly-defined cell types.


AFD
AMsh
ASJ
AUA
BWM D1
BWM D3
BWM D4
BWM D5
BWM D6
BWM V2
BWM V4
CEP
hmc
hyp3
URX
Continuous trajectory
ADF
ASG
AWA
AWB
Continuous trajectory but early terminal daughters are indistinguishable
AVB
AVD
AVK
BAG
cc
exc_cell
exc_gland
ILso
RIV
RMD
RME
SMD
URB
Z1/Z4
Discontinuous trajectory
Cell type
200
300
400
500
Embryo time (minutes post first cleavage)
Ridge lines show distribution of embryo times
for the union of the listed cell type and its parent.

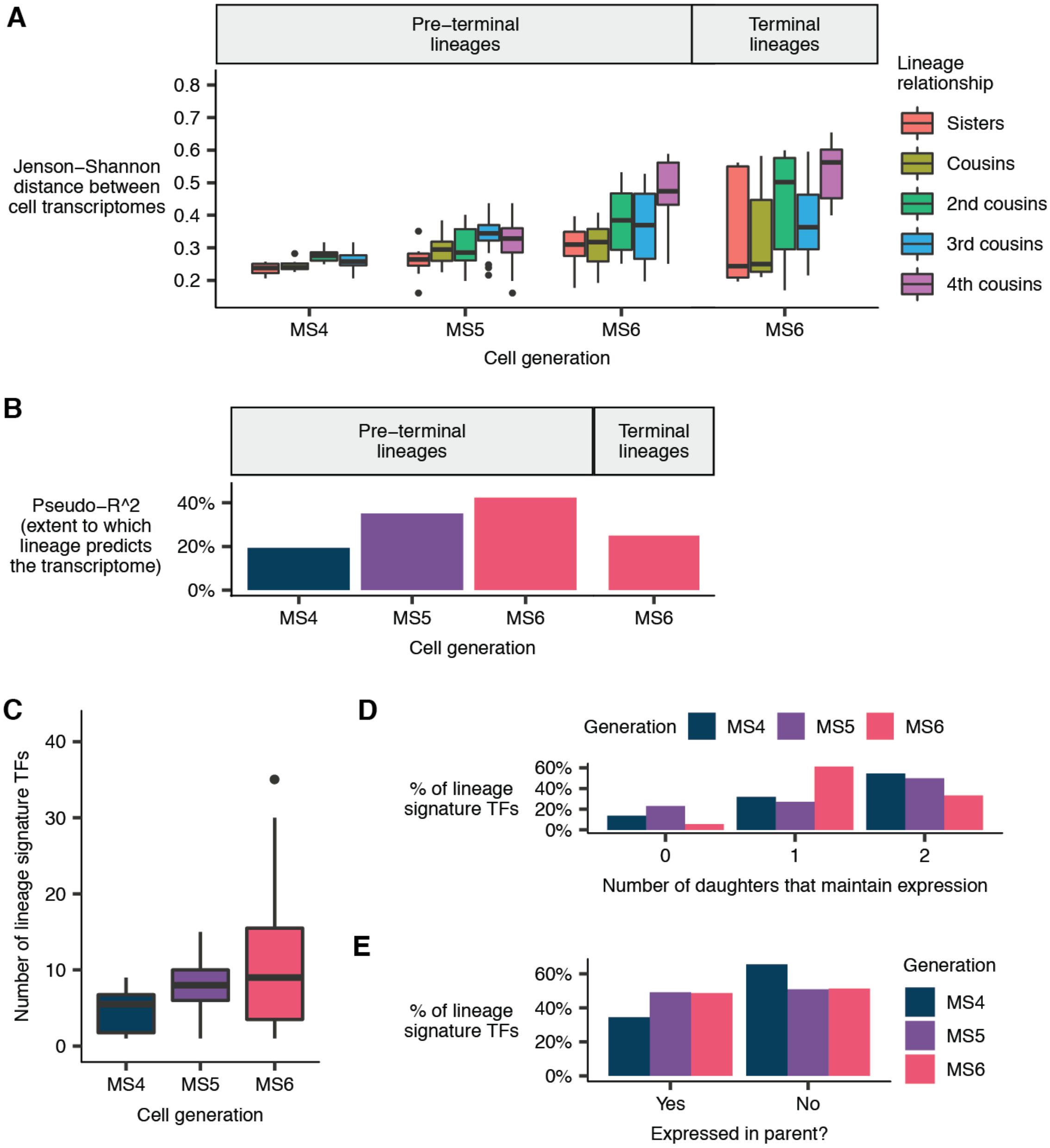
Supplemental Figure 2.27 Embryo time distributions for trajectories included in Supplemental Figure 2.26.

Ridge plot shows the distribution of estimated embryo times (minutes post first cleavage) for all of the sc-RNA-seq cells annotated as one of the terminal cells listed in Supplemental Figure 2.26, or its parent. For example, the ridge line for the row labeled AFD has the distribution of embryo times for all sc-RNA-seq cells annotated as either AFD (lineage = ABalpppapav/ABpraaaapav) or the AFD-RMD parent (lineage = ABalpppapa/ABpraaaapa). Rows are grouped based on whether or not the listed terminal cell forms a discontinuous trajectory with its parent in UMAP space. Trajectories were considered discontinuous only if the discontinuity was present in both the global UMAP (Figure 2.1A, Supplemental Figure 2.3) and the relevant tissue UMAP (Figure 2.3A, Supplemental Figures 2.9-2.10, 2.17).



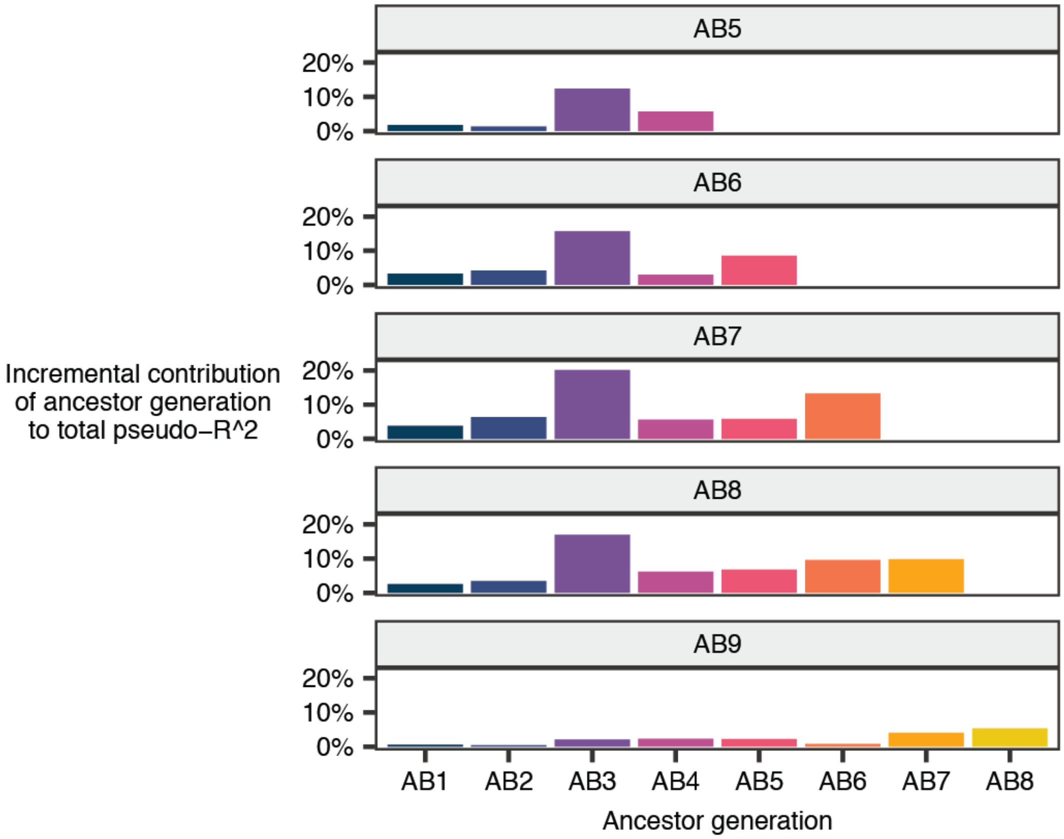
Supplemental Figure 2.28 Lineage distance vs. transcriptome distance in AB generation 8.

Jensen-Shannon (JS) distance between the transcriptomes of pairs of cells in AB8, the generation produced by 8 cell divisions since the AB founder cell. Data is faceted by lineage distance and by whether the pair consists of two pre-terminal cells, one pre-terminal and one terminal cell, or two terminal cells. Most terminal epidermal cells in the AB lineage are produced in AB8, while most terminal neurons, glia, and pharyngeal cells are produced in the subsequent generation, AB9. The terminal epidermal cells in AB8 exit the cell cycle and begin to differentiate, resulting in a large transcriptome distance between them and neuron/glia/pharynx progenitor cells that remain in the cell cycle.



Supplemental Figure 2.29 Correlation between cell lineage and the transcriptome in the mesoderm.

(A) Jensen-Shannon (JS) distance between the transcriptomes of pairs of mesoderm cells (MS lineage), faceted by cell generation and lineage distance. MS4 refers to the cell generation produced by 4 divisions of the mesoderm founder cell (MS), and likewise for generations MS5-6. The “transcriptome” of a given anatomical cell is defined as the average gene expression profile of all sc-RNA-seq cells annotated as that anatomical cell. Pairs of bilaterally symmetric cells are excluded from the statistics. The MS6 generation contains both terminal cells and pre-terminal cells that are still dividing. The data for MS6 in the plot is faceted to separate these, comparing only pairs of pre-terminal cells (left panel) or only pairs of terminal cells (right panel). **(B)** Estimates of the extent to which lineage explains the transcriptome in MS4-6, using a pseudo-R^2 statistic (see **Materials and Methods**). **(C)** Distribution of the number of “lineage signature transcription factors”—TFs that distinguish a cell from its sister—for cells in MS4-6. **(D)** Proportion of lineage signature transcription factors for a cell in a given generation that have expression maintained in 0, 1, or 2 of the cell’s daughters in the subsequent generation. **(E)** Proportion of lineage signature TFs for which expression in a given cell was maintained from the cell’s parent vs. newly activated after the parent’s division.

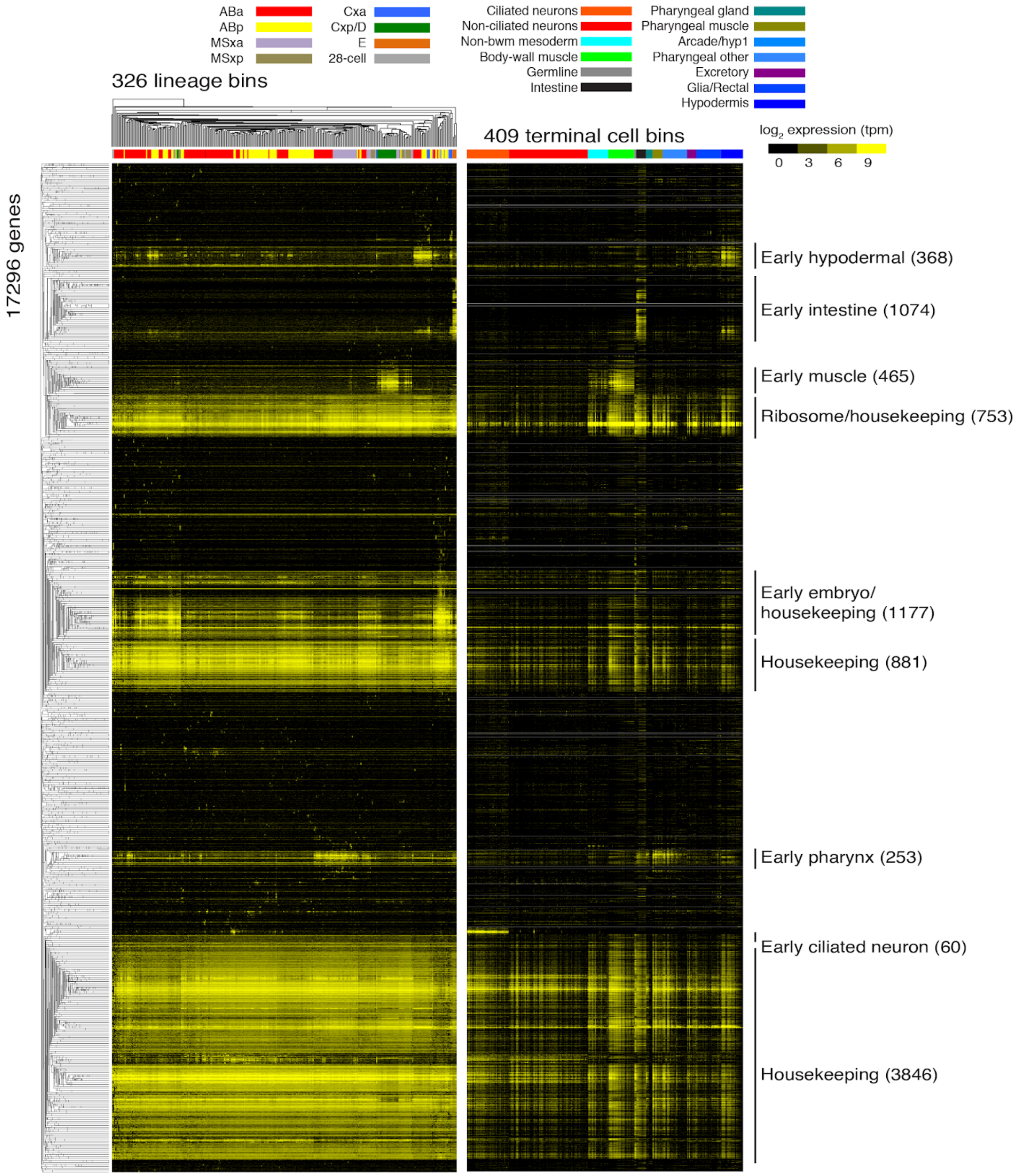


Supplemental Figure 2.30 Both recent and distant ancestry contribute to the ability of the lineage to predict a cell's transcriptome.

In Figure 2.5B, we used a pseudo-R^2 statistic to estimate the extent to which lineage predicts the transcriptomes of cells within a given generation. Specifically, our pseudo-R^2 statistic computes how much more similar are the transcriptomes of sister cells than those of random pairs of cells (see methods section titled "Pseudo-R^2 statistic").

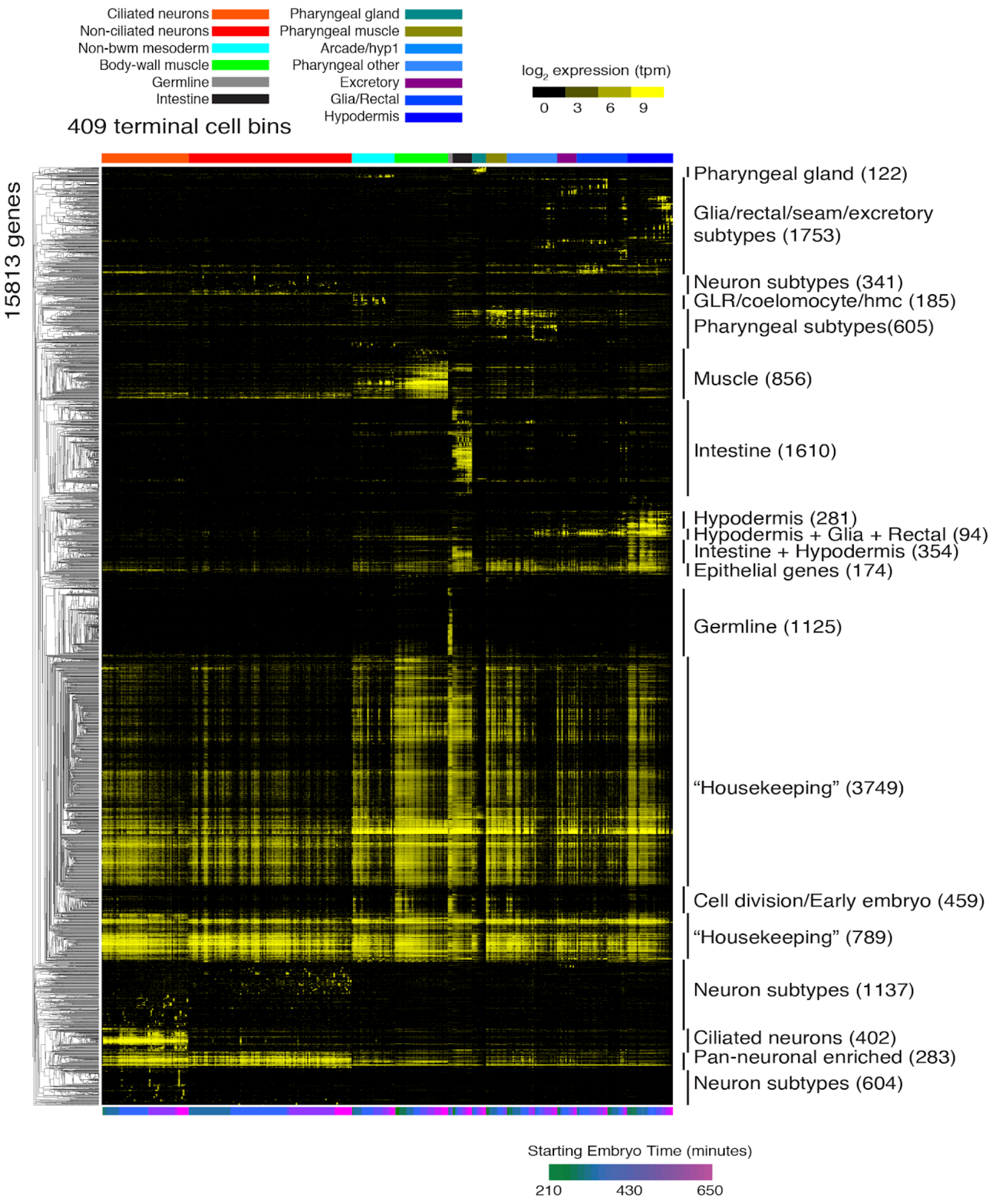
Here, we estimate how much of the similarity of sisters is specifically due to gene expression signatures associated with their parent, and how much is due to gene expression signatures associated with more distant ancestors. We describe how these estimates are computed in the methods section titled “Methods used in Supplemental Figure 2.30”.

Each panel in the figure corresponds to a generation of the AB lineage. Each bar on the x-axis corresponds to one of the generations that precede it. For example, AB5 is preceded by the generations AB4, AB3, AB2, and AB1. The height of each bar represents the contribution of gene expression signatures associated with that specific ancestor generation to the ability of the lineage to predict the transcriptome in the descendant generation. The sum of the heights of all bars in a panel is equal to the total pseudo-R^2 for the descendant generation (Figure 2.5B).



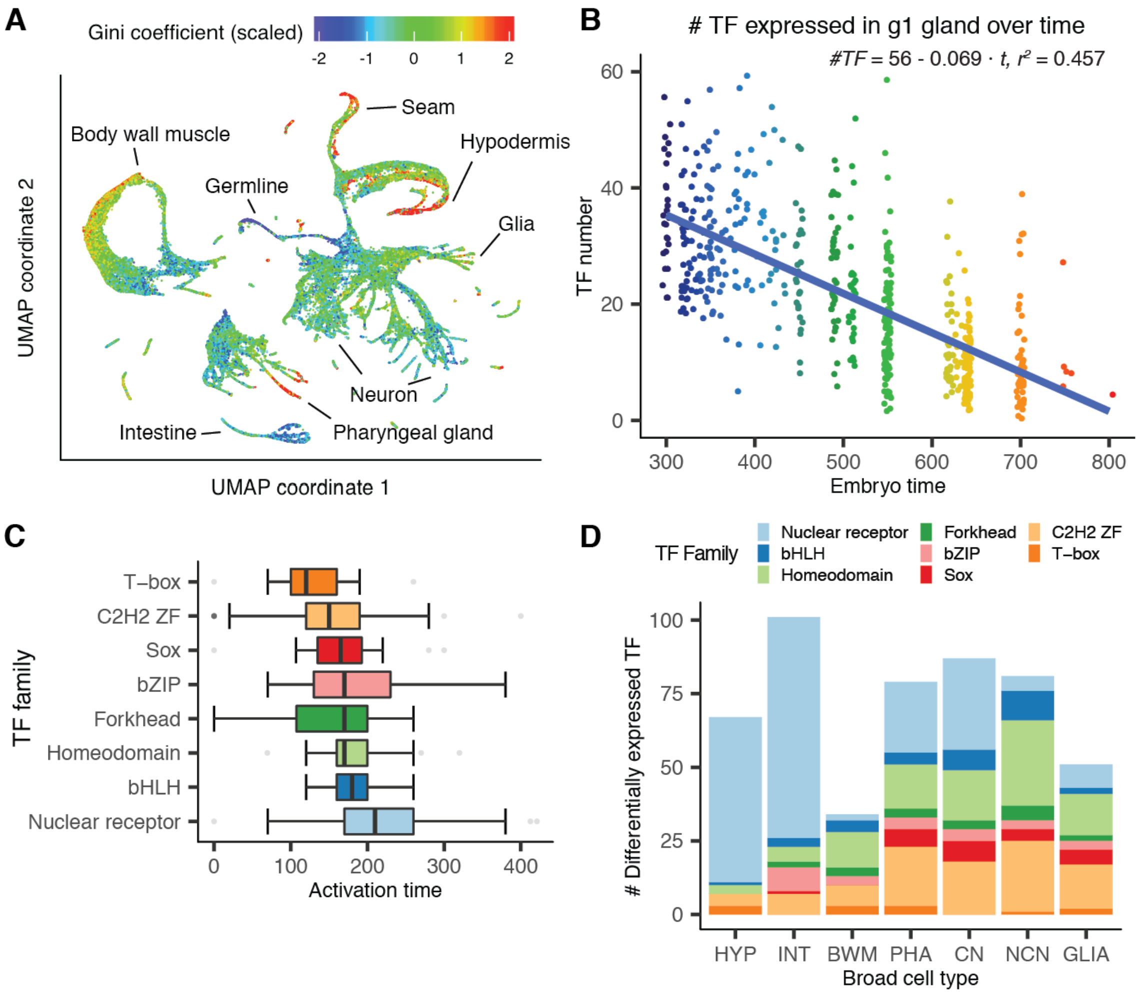
Supplemental Figure 2.31 Hierarchical clustering of progenitor lineage transcriptomes.

This heatmap shows the $\log_2$ expression ($\log_2$ transcripts per million) of all genes (rows) that are expressed in at least one pre-terminal lineage (columns). Genes and lineages are ordered by hierarchical clustering. The right panel shows the expression values in terminal cell bins, with genes (rows) ordered by the clustering as generated from the pre-terminal lineages and terminal cell bins (columns) ordered as in Supplemental Figure 2.32.



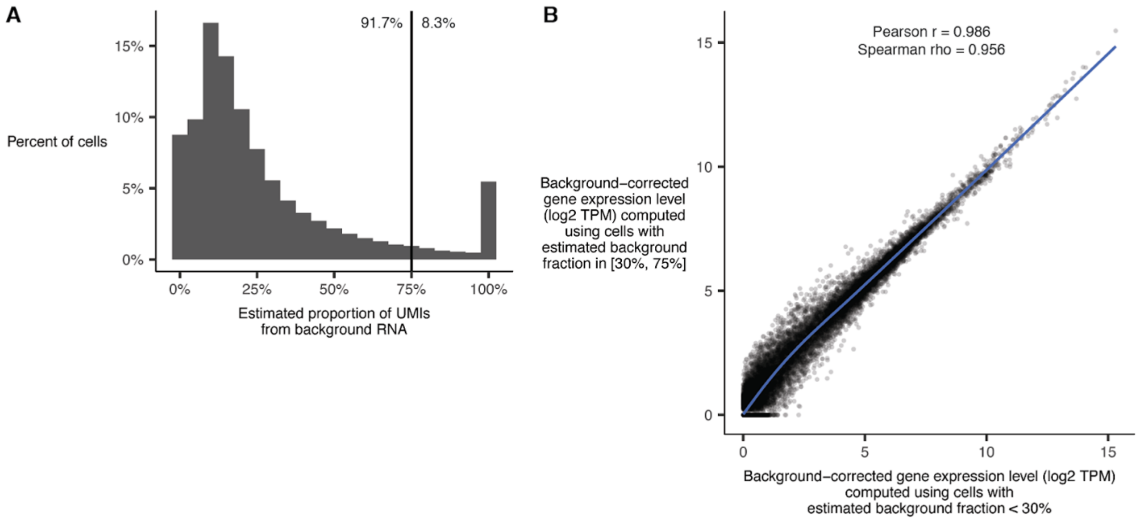
Supplemental Figure 2.32 Hierarchical clustering identifies signatures of tissue and cell type differentiation.

This heatmap shows the $\log_2$ expression ($\log_2$ transcripts per million) of all genes (rows) that are expressed in at least one terminal cell bin (columns). Genes are ordered by hierarchical clustering, and cell bins are ordered by tissues (colored as in the legend), and within tissues by the beginning of the time bin in minutes (early to late). Gene clusters are labeled by sites of predominant expression. Numbers in parentheses are the number of genes in that cluster.



Supplemental Figure 2.33 Transcriptome specialization and transcription factor usage across cell types and time.

(A) A global UMAP with 81,286 cells colored by the Gini coefficient of their gene expression vector, adjusted to correct for sample size bias and scaled by converting to z-scores. High Gini coefficients indicate that a small set of genes produces a large fraction of cell mRNA content. **(B)** Number of TF expressed in g1 gland over time. Equation shows linear regression result. Points are colored by estimated embryo time. **(C)** Box plot showing TF activation times—the embryo time when a TF first becomes expressed—grouped by TF family. For each TF, its activation time is defined as the 5th percentile of the estimated embryo time values for cells that express that TF. TF family annotations are taken from the CIS-BP database (*221*). Families that have fewer than 10 members detected in the current dataset were excluded from this plot. **(D)** Number of differentially expressed TFs and TF family composition across broad cell types.



Supplemental Figure 2.34 Distribution of estimates for the proportion of UMIs in a cell that come from background RNA.

(A) The process for making the estimates is described in the methods section “Per-cell background correction and filtering”. Due to the sparsity of the single cell data, the estimates are noisy. Numbers to the left and right of the vertical line indicate the proportion of cells with estimated background fraction < or >= 75%. Cells with background fraction >= 75% are filtered from all downstream analyses. **(B)** After per-cell background correction, cells with low and high background fractions have near-identical average gene expression profiles. Plot shows average gene expression profiles (measured in transcripts per million) computed from non-head body wall muscle cells divided into two groups: cells with estimated background fraction < 30% (x axis) and cells with background fraction in the range [30%, 75%].

CHAPTER 3 DEVELOPMENTAL TRAJECTORY OF PRE-HEMATOPOIETIC STEM CELL FORMATION FROM ENDOTHELIUM

The contents of this chapter have been previously published as:

Zhu Q*, Gao P*, Tober J*, Bennett L, Chen C, Uzun Y, Li Y, Howell ED, Mumau M, Yu W, He B. Developmental trajectory of pre-hematopoietic stem cell formation from endothelium. *Blood*. 2020 May 11.

* These authors contributed equally.

Introduction

Hematopoietic ontogeny involves multiple “waves” in which HSPCs with different potentials differentiate from HE cells. HE cells in the yolk sac (YS) differentiate into committed erythro-myeloid (EMP) and lymphoid progenitors, and the caudal arteries produce lymphoid progenitors and pre-HSCs (*222, 223*). YS hematopoiesis can be recapitulated in embryonic stem (ES) cell cultures, where the molecular events are well-described (*224, 225*). Groundbreaking studies described the transcriptomes of HE and pre-HSCs in the major caudal artery, the dorsal aorta, at single cell resolution (*39, 226-228*). However, these analyses did not examine the distribution or chromatin landscapes of cells along the trajectory, or the heterogeneity of cells in the intra-arterial clusters (IACs), due to the limited number of cells sequenced. To gain insights into the molecular mechanisms mediating the differentiation of arterial E cells into IACs we performed single-cell RNA sequencing (scRNA-Seq) and single-cell assay for transposase-accessible chromatin sequencing (scATAC-Seq). Our data reveal a continuous trajectory from E to IAC cells, previously undefined transitional cell populations along the trajectory, the pathways and transcription factors active in these cells, and describe the molecular heterogeneity of IAC cells.

Results

scRNA-Seq reveals a continuous trajectory from endothelial cells to IAC cells

Our strategy was to analyze all cells along the trajectory in a single sample to determine their distribution between different transcriptional states, and combine that with analyses of purified sub-populations to make accurate cell assignments and obtain additional coverage of rare cells (Figure 3.1A). We captured the entire trajectory by purifying a population containing all E, HE, and IAC cells (E+HE+IAC) from E9.5 and E10.5 embryos using a combination of endothelial markers (Supplemental Figure 3.1A). We also purified subpopulations of HE and E cells from E9.5 and E10.5 embryos based on expression of green fluorescent protein (GFP) from the *Runx1* locus(*229*) (Supplemental Figure 3.1B). We confirmed that only HE cells were capable of producing hematopoietic cells *ex vivo* (Supplemental Figure 3.1C). Kit^{hi} IAC cells were excluded in the sorts, therefore HE and E cells were negligibly contaminated with HSPCs (Supplemental Figure 3.1D). We purified IAC cells from E10.5 and E11.5 embryos using antibodies recognizing endothelial markers and Kit, E9.5 yolk sac EMPs (E9.5 YS-EMP), and E14.5 fetal liver HSCs (FL-HSCs) (Supplemental Figure 3.2).

Summary statistics for collected cell populations are shown in Figure 3.1B and Table 3.1. We used uniform manifold approximation and projection (UMAP) to reduce the data dimension (*117*). After filtering out non-endothelial and non-hematopoietic cells (Supplemental Figure 3.3A) and reducing batch effect using an “informative feature selection” method (Supplemental Figure 3.4, Materials and Methods), UMAP of the combined datasets shows a continuous trajectory from E to IAC cells (Figure 3.1C,D, Supplemental Figure 3.3). E14.5 FL-HSCs are disconnected from this trajectory, therefore, are more distantly related (Figure 3.1C).

Two streams of *Efbn2*$^{+}$ E cells in the UMAP converge to form a stem leading to HE and IACs (Figure 3.1E). Analyses of E10.5 E+HE+IAC cells manually separated into VU arteries and DA demonstrated that VU cells contribute to one of these streams and DA to both streams (Figure 3.1F). The E+HE+IAC samples, which demonstrate the distribution of cells at various stages, show that at E9.5, IAC cells constitute only 0.5% of the E+HE+IAC population, but at E10.5 the fraction of IAC cells expands 7 fold, representing 3.5% of the population, consistent with histological analyses showing increased numbers of IACs between these two embryonic stages (*230, 231*) (Figure 3.1D,G).

Unsupervised clustering identified 7 distinct populations in the combined dataset, and separated the two streams of E cells into distinct clusters (Supplemental Figure 3.5A-C). One cluster, containing only DA E cells, expresses high levels of Wnt target genes (Wnt^{hi} E) (Supplemental Figure 3.5D). The second cluster, Wnt^{lo} E containing both DA and UV cells, expresses lower levels of Wnt target genes. Wnt^{hi} E and Wnt^{lo} E could be further subdivided into arterial E (AE) and venous E (VE) by computing an arterial/venous score based on sets of AE and VE specific genes (*232*) (Figure 3.2A-D). Pseudo-time-ordered Wnt^{hi} and Wnt^{lo} E cells prior to the point where the AE score exceeded the VE score were defined as VE, and after that point were defined as AE. Wnt^{hi} AE and Wnt^{lo} AE then converge to form a distinct cluster determined by both UMAP and another method PHATE (*233*), that we termed conflux AE (Figure 3.2A, Supplemental Figure 3.5E). The confluence of transcriptomes in conflux AE is driven by the loss of Wnt$^{hi/lo}$ AE-specific gene expression and increased levels of transcripts from later stage-specific genes (Figure 3.2D,E). For example, expression of Wnt target genes *Foxq1* and *Nkd1* in Wnt^{hi} AE cells*,* and *Tmem255a* in Wnt^{lo} AE cells are down-regulated in conflux AE (Figure 3.2D). Cell cycle is also significantly inhibited in conflux AE (Supplemental Figure 3.5F,G), while Notch signaling is elevated, seen by increased expression of the Notch ligand *Dll4* and transcription factor *Hey2* (Figure 3.2D; Supplemental Figure 3.3C). Additional pathways activated in conflux AE include those regulating cell shape and motility (“elastin fiber formation””, “platelet adhesion to exposed collagen", "gap junction assembly") and processes important in hematopoietic cells ("MAPK signaling for integrins") (Figure 3.2F).

Runx1 regulates progression through a developmental bottleneck between pre-HE and HE

Conflux AE gives rise to HE and IAC cells, which are characterized by high levels of *Runx1* and *Gfi1*, and IAC by expression of the pan-hematopoietic marker gene *Ptprc* (encoding CD45) (Figure 3.3B; Supplemental Figure 3.3C). Between conflux AE and HE is a distinct cluster of endothelial cells that we named pre-HE. UMAP and pseudotime trajectories of E10.5 E+HE+IAC reveal an accumulation of pre-HE cells, suggesting a bottleneck between pre-HE and HE (Figure 3.3C) that is prominent at E10.5 although not at E9.5 (Figure 3.1D). *Gfi1*, a direct RUNX1 target that participates in extinguishing endothelial fate (*234*), shows elevated expression immediately after cells pass through the bottleneck and become HE, while high levels of *Sox17,* the Notch target *Hey2,* and the arterial marker *Cd44* are found in pre-bottleneck populations including conflux E and pre-HE (Figure 3.3B). To provide further evidence for the bottleneck, we utilized Velocyto and scVelo, which infer directionality of differentiation by modeling dynamics of unspliced versus spliced RNAs when a gene is up or down-regulated (*141, 235*). Velocyto and scVelo showed a marked decrease in RNA velocity in pre-HE cells, suggesting a differentiation barrier restricting their progression towards HE (Figure 3.3C, Supplemental Figure 3.6). Once pre-HE cells transit to HE, however, they smoothly differentiate to IAC cells. Several pathways known to promote *Runx1* expression and HSPC formation are upregulated in pre-HE, including Notch, tumor necrosis factor, fluid shear stress, cytokine signaling, and synthesis of eicosanoids, vitamins, and sterols (*52, 53, 236-241*), suggesting these pathways are important in pre-HE (Figure 3.3D, Supplemental Figure 3.7). Once cells transition to HE, RUNX1 plays a predominant role.

Runx1 expression is upregulated in approximately 7% of pre-HE cells suggesting that RUNX1 levels regulate passage through the bottleneck (Figure 3.3B, Supplemental Figure 3.6B). We tested this hypothesis using several approaches. First, we compared the distribution of cells between conflux AE, pre-HE, HE, and IAC in *Runx1*$^{+/-}$ and *Runx1*$^{+/+}$ littermates by scRNA-Seq. We observed a 68% reduction in the proportion of HE and IAC cells in E10.5 *Runx1*$^{+/-}$ compared to *Runx1*$^{+/+}$ embryos, and a commensurate 56% increase in pre-HE, consistent with the hypothesis that RUNX1 levels regulate transit through the bottleneck (Figure 3.3E). We also performed the reciprocal experiment; ectopically expressing RUNX1 in all endothelial cells by activating a conditional *Runx1* cDNA in the *Rosa26* locus (cR1) using an endothelial-specific tamoxifen-inducible Cre driven from the vascular endothelial cadherin (*Cdh5*) regulatory sequences (Cre) (*242*). We previously showed that ectopic expression of RUNX1 in all endothelial cells in Cre;cR1/+ embryos increased the frequency of functional HE cells compared to control embryos (cR1/+) (*242*). scRNA-Seq analysis demonstrates these results from an increase in the proportion of HE cells and a proportionate decrease in pre-HE cells (Figure 3.3F), confirming that RUNX1 levels regulate the number of pre-HE cells that transit through the bottleneck to become HE. Second, we determined whether RUNX1 haploinsufficiency reduced the number of phenotypic HE cells by confocal microscopy. SOX17 is expressed in AE cells and promotes HE specification, whereas HE cells are RUNX1^{+}SOX17$^{low/-}$ (*243, 244*). The ratio of RUNX1^{+}SOX17$^{low/-}$ HE cells versus RUNX1^{-}SOX17^{+} AE cells in the dorsal aorta was significantly lower in E9.5

Runx1$^{+/-}$ embryos compared to *Runx1*$^{+/+}$ embryos (Supplemental Figure 3.8). Finally, we measured the frequency of functional HE cells within a purified population of CD44^{+} E+HE+IAC cells (Supplemental Figure 3.2G), which are enriched for conflux AE, pre-HE, HE, and IAC (*245*) (Figure 3.3B). RUNX1 haploinsufficiency reduced the frequency of functional HE cells by 77% (Figure 3.3G), consistent with the observed reduction in the scRNA-Seq experiment (Figure 3.3E). Together these data confirm that RUNX1 level regulates the number of pre-HE cells that transit through the bottleneck to become HE cells.

scATAC-Seq identifies putative Runx1 enhancers and transcription factor motifs that gain accessibility in pre-HE

To identify signals that may activate *Runx1* expression in pre-HE, we performed paired scRNA-Seq and scATAC-Seq on E10.5 CD44^{+} E+HE+IAC cells to identify *Runx1* enhancers and the stages they are accessible. High quality open chromatin profiles were obtained for 1670 cells, covering various cell types from E to IAC (Supplemental Figure 3.9). The joint embedding of scRNA-Seq and scATAC introduced a gap between pre-HE and IAC cells on the UMAP. This results from the developmental bottleneck around E10.5 that causes an underrepresentation of HE cells connecting pre-HE and IAC in some samples. (Figure 3.4A). We devised a computational approach that matches scATAC-Seq clusters with scRNA-Seq clusters (Figure 3.4A, Supplemental Figure 3.10), and subsequently linked enhancers with their target promoters (Figure 3.4B). Accuracy of our method was benchmarked using known hematopoietic and endothelial enhancers (Figure 3.4C, D). We applied chromVar (*246*) to assess differential transcription factor

(TF) binding patterns along the EHT trajectory (Figure 3.4E). Results show strong correlation with the TF expression patterns and are consistent with pathway analyses from scRNA-Seq data. For example, strong TCF/LEF binding activity was detected in Wnt^{hi} E that abruptly decreased in conflux AE (Figure 3.4E, F). SOX and FOX binding sites are mostly open in conflux AE and pre-HE. Binding sites for a large group of TFs had increased accessibility beginning at the pre-HE stage, including HES1, GATA, SMAD, and TFs such as MECOM, EGR1, and YY1 that regulate HSC homeostasis (*247-249*), the latter group suggesting that an HSC-specific transcriptional program may initiate at the pre-HE stage.

Runx1 contains two promoters, an upstream P1 promoter that is first utilized in committed HSPCs, and a more proximal P2 promoter that is active in HE and HSPCs (*250*). Consistent with this, P1 first becomes accessible in IAC cells, whereas P2 is accessible in all endothelial cells including pre-HE (Figure 3.5A), which may permit or contribute to the stochastic *Runx1* expression observed in a subset of endothelial cells (Figure 3.3B). Using our computational approach, we predicted 27 enhancer-promoter (E-P) interactions, which recapitulate 11 out of 22 previously identified E-Ps based on chromosome conformation capture assays (*251, 252*) (Figure 3.5A, also see Materials and Methods). All of the predicted enhancers exhibit higher co-accessibility with P1 compared to P2, therefore only E-Ps to P1 are indicated. A significance plot of the predicted E-Ps reveals several enhancers whose chromatin openness is significantly correlated with *Runx1* expression, including the *Runx1* +23 enhancer (Figure 3.5B). Several of the predicted enhancers exhibit stage-specific co-accessibility with the P1 promoter (Figure 3.5C). Interestingly, one candidate enhancer located 371 kb upstream of *Runx1* P1 was accessible only in pre-HE and IACs, and not in other endothelial cell populations (Figure 3.5A, C). This candidate enhancer was previously shown by circular chromosome conformation capture sequencing to interact with the +23 enhancer and P1 in a hematopoietic progenitor cell line (*251*). The -371 enhancer drove expression of a reporter gene in the intermediate cell mass and posterior blood island of zebrafish embryos, both of which are sites of hematopoietic ontogeny (*251*). The scATAC-Seq signal encompassing the -371 enhancer begins to increase in conflux AE cells and reaches a maximum in pre-HE cells (Figure 3.5A, D). This change in accessibility in pre-HE coincides with the activation of *Runx1* expression in a subset of pre-HE cells (Figure 3.5D). However, unlike the +23 enhancer, the chromatin accessibility of the -371 enhancer subsequently decreases in IAC cells and is no longer open in FL-HSCs (Figure 3.5A). The candidate -371 enhancer contains GATA, STAT, and JUN motifs, indicating that GATA2 and cytokine and/or inflammatory signaling may contribute to the opening of this enhancer in pre-HE (Figure 3.5A). An independent co-expression analysis based on the scRNA-Seq data reveals that these factors form a co-expression gene module that precedes and correlates with *Runx1* expression (Figure 3.5E), suggesting they may cooperatively regulate *Runx1* expression. Notably, neither the -371 nor the +23 enhancers contain SOX motifs, which are recognized by a repressor of Runx1 expression, Sox17 (*244*). Other TF motifs enriched in the 27 called *Runx1* enhancers include ETS, FOX, SOX, KLF/SP, RUNX, and SMAD, which are recognized by TFs with well-documented roles in HSPC formation (*253-255*).

Two waves of HSPCs form in the IACs

We also examined the transition of HE to IAC cells and the composition of IAC cells. Principal component analysis (PCA) depicts a sharp U-turn as HE differentiates into IAC cells, reflective of a marked decrease in AE gene expression and activation of hematopoietic genes (Figure 3.6A). For example, the AE-specific gene *Gja5* is primarily expressed on the HE side of the trajectory, while expression of *Spn* (encoding CD43), *Ptprc* (encoding CD45) and the Rho GTPase *Rac2* rapidly increases in IAC cells (Figure 3.6B, Supplemental Figure 3.11A, B). Transient expression of the chromatin remodeling protein *Nupr1* occurs at the U-turn, while *Hey1* and *Sox17* transcripts significantly diminish as IAC cells mature (Figure 3.6B, Supplemental Figure 3.11B).

IACs contain pre-HSCs that cannot engraft adult mice directly, but can mature *in vivo* or *ex vivo* into adult-repopulating HSCs(*36, 38, 256*). Pre-HSCs are classified as type I or II based on CD45 expression; type I are CD45$^-$, and the more mature type II are CD45$^+$ (*38*). E10.5 IACs contain only type I pre-HSCs, whereas E11.5 IACs contain both type I and II pre-HSCs(*38*). Additionally, multiple progenitors with lymphoid, myeloid, lympho-myeloid, or multi-lineage potential emerge prior to or contemporaneously with pre-HSCs (*222*), at least a subset of which are CD45$^+$ (*226, 257, 258*). We compared E10.5 CD45$^+$IAC cells that contain HSC-independent progenitors and lack pre-HSCs to E11.5 CD45$^+$CD27$^+$CD144$^+$ IAC cells enriched for type II pre-HSCs (E11.5 pre-HSCs) (*38, 259*) to determine their developmental relationship (Supplemental Figure 3.2C,D). The two populations bifurcate in the third principal component of PCA plots; specifically, the majority of E10.5 CD45$^+$ IAC cells occupy one end of the PC3 axis, and E11.5 pre-HSCs reside on the other end (Figure 3.6C, D). E11.5 pre-HSCs demonstrated a high correspondence with previously published data (*39*) (Supplemental Figure 3.12A). We determined the fraction of pre-HSCs in E10.5 and E11.5 IAC cells using a K-nearest-neighbor classifier. About 2% of E10.5 IAC cells were found to be molecularly similar to E11.5 type II pre-HSCs; this fraction of pre-HSCs increases to 67% in E11.5 IAC cells

(Figure 3.6D), consistent with previous limiting dilution assay results demonstrating an increase in functional pre-HSCs between E10.5 and E11.5 (*36*). To determine the fate bias of cells from earlier stages, we used Palantir and FateID (*144, 260*). T-SNE plot generated by Palantir analysis shows a bifurcation pattern similar to the PCA result (Supplemental Figure 3.11C). Distribution of the fate probabilities suggests that compared to E9.5 HE, E10.5 HE has higher probability of choosing pre-HSC fate, and E11.5 IACs contain more pre-HSC-like cells than E10.5 IACs (one-sided Kolmogorov-Smirnov test, Supplemental Figure 3.11D, E).

The 974 genes more highly expressed in E11.5 pre-HSCs compared to E10.5 CD45$^+$ IAC cells include known markers of pre-HSCs and/or HSCs (*Eya2, Procr, Cd27,* and *Mecom*) (*39, 247, 259, 261, 262*), while the 877 genes up-regulated in E10.5 CD45$^+$ IAC cells include proliferation related genes (*Myc*) and lympho-myeloid associated genes (*Il7r, Fcer1g*) (Figure 3.6E, F). Among the differentially expressed genes, some transcription factors, such as *Myc*, *Klf2*, *Smad7*, *Mecom*, *Meis2* and *Nfix*, are expressed in HE cells, and show strong bifurcation in expression as cells become IAC cells (Supplemental Figure 3.11F). Pathway analysis suggests E11.5 pre-HSCs gain stem-cell specific features such as “OCT4, SOX2, NANOG represses genes related to differentiation”, while pathways associated with E10.5 CD45$^+$ IAC cells are associated with cell cycle and/or related to a specific hematopoietic lineage, such as “TCF dependent signaling in response to WNT” (Figure 3.6G). Interestingly, E11.5 pre-HSCs, although sampled 1 day later in development compared to E10.5 CD45$^+$ IAC cells, retain many pathways from E/HE stages, such as “Signaling by BMP” and “eNOS activation”, suggesting a relatively slow shutdown of the E/HE program in pre-HSCs. In contrast, subsets of E10.5 CD45$^+$ IAC cells show lineage-specific differentiation bias; *Il7r* is up-regulated in 26%, and *Gata1* in 3% of E10.5 CD45$^+$ IAC cells (Figure 3.6F).

Previous scRNA-Seq studies identified committed progenitors in E10.5 and E11.5 IACs but concluded they were contaminating erythro-myeloid progenitors (EMPs), likely originating from the yolk sac, that had been circulating in the blood and became attached to the IACs (*39, 226*). We addressed the possibility that the E10.5 CD45$^+$ IAC cells we profiled are contaminating YS EMPs. Direct comparison shows E10.5 CD45$^+$ IAC cells and E9.5 YS-EMP are molecularly and functionally distinct (Figure 3.6H, Supplemental Figure 3.12C). E10.5 CD45$^+$ IAC cells contained progenitors of macrophages and granulocytes/monocytes, but very few erythroid or megakaryocytic progenitors compared to E9.5 YS-EMPs (Figure 3.6H). E10.5 CD45$^+$ IACs have potent lymphoid potential; limiting dilution assays revealed a high frequency of cells (1:6) capable of producing B cells following culture on OP9 stromal cells or T cells on OP9 expressing the Notch ligand delta-like 1 (Figure 3.6I, J). E10.5 CD45$^-$ IAC cells also contained progenitors with lymphoid and myeloid potential, although their frequency was lower than in the E10.5 CD45$^+$ IAC population. In summary, E10.5 CD45$^+$ IAC cells represent a distinct wave of lympho-myeloid-biased progenitors in IACs that appear prior to E11.5 type II pre-HSCs.

Discussion

Our single cell analyses provide new insights into the process by which endothelial cells differentiate into pre-HSCs. First, we define a precursor of HE we have named pre-HE, in which multiple pathways known to regulate HSPC formation appear to act. Also, through trajectory analyses and genetic perturbation experiments we identified a bottleneck separating pre-HE from HE, indicative of a developmental barrier that must be overcome at that transition. It is long known that embryonic hematopoiesis is exquisitely sensitive to *Runx1* dosage, as reduced *Runx1* dosage decreases the number of HE cells, IACs and committed hematopoietic progenitors in the embryo (*263-265*). Our scRNA-Seq analyses show that the deficits caused by reduced *Runx1* dosage are caused, at least in part, by the inefficient transition of pre-HE to HE cells. The molecular underpinnings of the bottleneck at the pre-HE to HE transition are not known. One possibility is that *Runx1* expression may be actively repressed in the majority of pre-HE cells by TFs such as *Sox17* (*244, 266*), which is highly expressed in pre-HE, and binding sites for which are accessible in pre-HE. A requirement for chromatin remodeling may also be a limiting factor in pre-HE, as multiple epigenetic regulatory proteins have been shown to affect *Runx1* expression in HE, some of which may act at the pre-HE to HE transition (*267*).

Prior to HE, *Runx1* is expressed at low levels in a subset of endothelial cells, consistent with the chromatin accessibility of the P2 promoter and of several *Runx1* enhancers in endothelial cells. *Runx1* expression in endothelial cells appears to be stochastic; it then becomes elevated in a subset of pre-HE cells, and is uniformly high in HE and IACs. The mechanism by which *Runx1* expression is activated in a subset of pre-HE cells is not known, but our experiments provide some clues. scATAC-Seq revealed that a distal enhancer in *Runx1* (-371), previously validated in zebrafish transgenic embryos and conserved in mammals (*251*), first becomes accessible in pre-HE. Highly conserved TF motifs in the -371 enhancer include GATA, STAT, and JUN, implying that TFs that bind these motifs may play a role in opening the enhancer in pre-

HE. *Gata2* expression is activated in a pulsatile manner in endothelial cells in the DA (*268*), which may contribute to the stochastic expression of *Runx1* in arterial endothelial cells. STAT and JUN motifs are recognized by TFs that are effectors of inflammatory signaling pathways, including type I and II interferons, and tumor necrosis factor, all of which promote HSPC formation from arterial endothelium (*52, 269, 270*). Hence signaling pathways known to promote later *Runx1* expression in HE could potentially initiate *Runx1* expression in a subset of pre-HE cells by activating the candidate -371 pre-HE enhancer. At later stages, in IACs and FL-HSCs, multiple additional enhancers, including the +23 enhancer gain accessibility and interact with the P1 promoter to further elevate *Runx1* expression.

A second important concept gleaned from our data is that the IACs contain at least two distinct HSPC subtypes, committed lympho-myeloid-biased progenitors and pre-HSCs, that can be distinguished molecularly. These appear sequentially, with CD45$^+$ lympho-myeloid-biased progenitors preceding the formation of type II pre-HSCs. The mechanisms underlying the generation of these two types of HSPCs is of great interest. It is not known, for example, if they independently differentiate from an equivalent population of immature IAC cells. Alternatively, they may be derived from distinct populations of HE cells. Our cell fate analysis suggests that E10.5 HE is more likely to assume pre-HSC fate than E9.5 HE, which is consistent with the observation that E11.5 IACs contain more pre-HSCs than E10.5 IACs. The bifurcation of fate may be partially driven by early differential expression of transcription factors specific to pre-HSCs or lympho-myeloid-biased progenitors. The lympho-myeloid-biased progenitors are more developmentally “mature” compared to the type II pre-HSCs, suggesting that they are more driven towards terminal differentiation. A similar population of lympho-myeloid restricted progenitors that originates in the yolk sac colonizes the FL and thymus prior to HSCs (*271, 272*). Lympho-myeloid-biased progenitors in the arterial IACs may serve a similar function.

The earlier emergence of lympho-myeloid-biased progenitors in the arteries may have implications for ongoing efforts to generate pre-HSCs from ES cells. The acquisition of lymphoid potential is often used as a surrogate for pre-HSC formation. However, it is possible that conditions favoring the production of this earlier population of committed lympho-myeloid progenitors may be suboptimal for the later formation of pre-HSCs. If this is the case, then inhibiting the differentiation of lympho-myeloid progenitors in ES cell cultures may improve pre-HSC production ex vivo.

Materials and Methods

Animal husbandry

B6C3F1/J 3-week old female mice were purchased from Jackson Laboratories (Stock no: 100010). Females were injected with 5 IU pregnant mare serum gonadotropin and 48 hours later with 5 IU human chorionic gonadotropin (Sigma), then immediately paired overnight with C57BL6/J male mice. Runx1:GFP (*Runx1^{tm4Dow}*) (*229*) homozygous male mice were mated to super-ovulated B6C3F1/J 3-week old female mice to generate embryos for purification of E and HE cells. Female B6C3F1/J mice were mated with male B6129SF1/J mice for isolating fetal liver HSCs. Ectopic RUNX1 expression in endothelial cells in Tg(Cdh5-cre/ERT2)1Rha embryos (*273*) that contained an activatable *Runx1* cDNA in the *Rosa26* locus was described previously (*242*). *Runx1$^{+/-}$* mice (*Runx1^{tm1Spe}*) were described previously (*264*). The morning post mating is considered embryonic day (E) 0.5. E9.5-E11.5 embryos were accurately staged at the time of harvest by counting somites. Embryos that showed abnormal development were discarded. Mice were handled according to protocols approved by the University of Pennsylvania's Institutional Animal Care and Use Committee and housed in a specific-pathogen-free facility.

Embryo dissection and FACS

Yolk sacs were removed from embryos, and vitelline vessels were retained with the embryonic portion. The head, cardiac and pulmonary regions, liver, digestive tube, tail and limb buds were removed. The remaining portion containing the aorta-gonad-mesonephros (AGM) region, portions of somite, umbilical and vitelline vessels were collected. E9.5 and E10.5 yolk sacs were collected for isolation of EMPs. Tissues were dissected in phosphate buffered saline (PBS)/10% Fetal Bovine Serum (FBS) and Penicillin/Streptomycin (Sigma), followed by dissociation in 0.125 Collagenase (Sigma) for 1 hour. Tissues were washed and filtered through a 40-micron filter and resuspended in antibody solution. Cells were sorted on either BD Influx, MoFlow Astrios EQ (Beckman), BD Jazz, or BD Aria, all equipped with a 100-micron nozzle, and run at a pressure of 17 psi with flow rates less than 4000 events/second. Sorted cells for functional assays were collected in PBS/20% FBS/25mM HEPES. For scRNA-Seq and scATAC-Seq, cells were collected in IMDM/20% FBS in low-retention microcentrifuge tubes (Denville).

scRNA-Seq

Sorted cells were immediately processed for library preparation using the 10x Genomics Chromium Single Cell 3′ Reagent Kit v2. Libraries were quantified using the dsDNA High-Sensitivity (HS) Assay Kit (Invitrogen) on a Qubit fluorometer and the qPCR-based KAPA assay (Kapa Biosystems). Library quality assessment was performed on the Agilent 2100 Bioanalyzer in combination with the Agilent High Sensitivity DNA kit. Indexed libraries were pooled and sequenced on an Illumina HiSeq 4000 or NextSeq 550 using paired-end 26 × 98 bp read length.

scATAC-Seq

Sorted cells were centrifuged at 300x *g* for 5 min at 4 °C and resuspended in 50 µL of 1x PBS + 0.04% BSA. 45 µL of supernatant was carefully discarded, and 45 µL of chilled lysis buffer was added and mixed by pipetting gently. After incubation for 5 min on ice, 50 µL of chilled wash buffer was added without mixing. The mix was centrifuged at 500x *g* for 5 min at 4 °C, and 95 µL of supernatant was discarded. 45 µL of chilled diluted nuclei buffer was added without mixing, and the mix was centrifuged at 500x *g* for 5 min at 4 °C. The nuclei pellet was resuspended in 7 µL of chilled diluted nuclei buffer. 2 µL of nuclei suspension was used to determine the cell concentration by a Countess II cell counter (Invitrogen), and the remaining 5 µL of nuclei suspension was processed for library preparation using the Chromium Single Cell ATAC Reagent Kits protocol. Libraries were quantified using the dsDNA HS Assay Kit on a Qubit fluorometer and the qPCR-based KAPA assay. Library quality assessment was performed using the Agilent

2100 Bioanalyzer with the Agilent High Sensitivity DNA kit. Indexed libraries were pooled and sequenced on an Illumina NextSeq 550 using paired-end 50 × 50 bp read length.

OP9 co-culture assays

FACS sorted cells were plated in limiting dilutions on OP9 (ATCC) or OP9-delta-like 1 stromal cells in 96-well plates containing Minimum Essential Medium Eagle - alpha modification (alpha MEM), 20% FBS (Hyclone, Gibco) and Pen/Strep. 5 ng/mL Flt3L and 10 ng/mL IL-7 were added to the medium for OP9 co-cultures. The medium for the OP9-DL1 co-cultures was supplemented with 5 ng/mL Flt3L and 1 ng/mL IL-7. Co-cultures were conducted for 10-13 days and subsequently, flow cytometry was performed on a LSR-II (BD). The flow cytometry antibody panel for OP9 co-cultures included the hematopoietic markers Mac1, Gr1, CD19, B220, and CD45, while the OP9-DL1 co-cultures were analyzed for CD45, CD90, and CD25. Limiting range was determined using the extreme limiting dilution analysis (ELDA) software analysis tool (*274*).

Methylcellulose assays

To enumerate erythroid, myeloid, and megakaryocyte progenitors, sorted cells were cultured in M3434 (StemCell Technologies) for 7 days. Colonies were scored based on morphological criteria.

Hemogenic endothelial (HE) cell assay

Sorted cells were plated in limiting dilutions with OP9 stromal cells for 8-10 days in alpha MEM containing 10ng/mL of IL-3, IL-7, Flt3, and SCF. Cells were analyzed by flow cytometry for hematopoietic markers (B220, CD19, Mac1, Gr1, and CD45), and ELDA software analysis tool (*274*) was used to determine the frequency of HE.

Whole-mount immunofluorescence and confocal microscopy

Embryos were prepared as previously described (*37*). Embryos were stained with rabbit anti-mouse/human RUNX1/AML1+RUNX2+RUNX3 and rabbit anti-mouse/human Sox17 at a working concentration of 1:500. Secondary antibodies used were goat anti-rabbit Alexa Fluor 488 (1:1000, Abcam ab150077, against Runx1) and goat anti-rabbit Alexa Fluor 647 (1:500, Abcam ab 150083, against Sox17). Images were acquired on a Zeiss LSM 880 AxioObserver inverted microscope equipped to detect 488, 561, and 633nm wavelengths. Images were analyzed using Fiji software (*275*).

scRNA-Seq data analysis

Data pre-processing and filtering of non-endothelial and non-hematopoietic cells

Raw sequencing reads were first pre-processed with 10x Genomics Cell Ranger pipeline and aligned to the mouse mm10 reference genome. An initial filtering was performed on the raw gene-barcode matrix output by the Cell Ranger *cellranger count* function, removing barcodes that have less than 1000 transcripts (quantified by unique molecular identifier (UMI)) and 1000 expressed genes ("expressed" means that there is at least 1 transcript from the gene in the cell). Barcodes that pass this filter were considered as cells and were fed into downstream dimensionality reduction and clustering analysis. In the global UMAP with 37,766 cells combined from all datasets, we noticed several contaminant cell types, including mesenchymal-like cells that express high levels of collagen, erythroid progenitors, and *Lyve1*$^{+}$ endothelial cells that likely have a lymphatic or YS origin (Supplemental Figure 3.3A). Since these contaminant cell types are not directly associated with EHT, we removed them from our downstream analyses, thereby obtaining a UMAP exclusively with endothelial cells and hematopoietic cells (Supplemental Figure 3.3B).

Unsupervised clustering on the cleaned global UMAP reveals a clear separation between IACs and other hematopoietic progenitors. For example, Haptoglobin (*Hp*) is highly expressed in most YS EMPs, but has almost zero expression in IAC cells (Supplemental Figure 3.3B). Genes such as *Gata1* were found to be expressed in subset of EMP and a few IAC cells (Supplemental Figure 3.3B). We found a *Bnip3*^{hi} population in the E10.5 CD44^{+} E+HE+IAC samples, and a group of low-quality endothelial cells marked by low UMI counts per cell (Supplemental Figure 3.3B). After filtering out these cells, we obtained a final UMAP with 23,081 cells representing the EHT trajectory (Supplemental Figure 3.3C). We ran Louvain clustering on this global UMAP and assigned cell types based on differentially expressed genes (Supplemental Figure 3.3C). The cell distributions of each dataset post cleaning are shown in Figure 2C.

Feature selection, dimensionality reduction, and unsupervised clustering

Gene-barcode UMI count matrix combined from all datasets was first processed with a standard pipeline utilizing the Monocle 3 package (*276*). An initial variable expressed gene (VEG) selection was performed on the size-factor corrected, log2 transformed expression matrix using the feature dispersion table output by Monocle *estimateDispersions* and *dispersionTable* functions. The *estimateDispersions* function models how a gene's variance is related to its average expression. We tested various cutoffs for dispersion, and found a relatively consistent pattern in the resulting UMAP. We found relaxing the cutoff to *dispersion_empirical/dispersion_fit > 0.5* improves the clustering result, especially in detecting rare cell types/states. This procedure is similar to choosing x number of top variable genes in the *Seurat* pipeline, where x is an arbitrary number selected by the user. Furthermore, we require a gene to be expressed in at least more than 1 transcript in a minimum of 10 cells in order to be used as a VEG for dimensionality reduction. To produce a low dimensional embedding of the data, principal component analysis (PCA) was performed on the VEG-cell matrix, and the top PCs were used as features for the UMAP algorithm. UMAP was computed using the *umap* function in the uwot R package, with "cosine" distance metric, 20 nearest neighbors, and the rest of the parameters utilized were default. Louvain clustering was run on the K-nearest neighbor graph (K = 20) constructed from cell embeddings on the UMAP. Additionally, we ran PHATE (*233*) on the same set of EHT cells with the default parameter setting, and obtained a similar trajectory as in the UMAP (Supplemental Figure 3.5E).

We noticed that VEGs selected using Monocle 3 or Seurat contained genes that are cell-type specific, as well as genes associated with cell cycle and batch differences. Some highly expressed house-keeping genes are also called as VEGs, likely due to variation in sequencing depth across batches. The UMAP produced by the Monocle 3 or Seurat pipeline is globally reflective of cell type, but locally affected by batch and cell cycle, causing some clusters to be less representative of underlying cell states (Supplemental Figure 3.4D).

To select features that are most reflective of cell type transitions during EHT and less affected by batch or cell cycle difference, we devised a feature selection procedure called informative feature (IFF) selection. IFF takes an initial clustering generated by other single-cell clustering methods such as Monocle 3, Seurat and SC3 (*277-279*) and then computes the expressed (non-zero) fraction of each gene for each cluster. Genes that are detected in too few cells (e.g., less than 10% in every cluster) are filtered out. To determine the "inequality" of the gene's expression across clusters, we calculated Gini coefficient on the per-cluster-expressed-fraction vector. The distribution of Gini coefficients shows a clear peak on the left (Supplemental Figure 3.4A). Genes in the peak are highly enriched for housekeeping and cell cycle functions, while genes in the right long tail are strongly enriched for cell type specific ontologies (Supplemental Figure 3.4B, C). This allows us to separate the majority of "cell-type-informative-features" from "ubiquitous features". We found that IFF significantly improves the clustering result by mitigating the batch effect (Supplemental Figure 3.4D) and identifies underlying cell subtypes and states (Supplemental Figure 3.4F-H). The method is also robust to initial clustering parameter choice (Supplemental Figure 3.4E).

The IFF selection procedure is conceptually similar to the dpFeature selection introduced in the Monocle 2 package, which requires an initial clustering that is most reflective of cell type and selects cell-type-specific features by differential gene expression test. However, unlike the dpFeature, IFF selection is more permissive, as it does not require a gene to be significantly differentially expressed in one cell type to be selected. Additionally, genes expressed in a subset of clusters with a relatively weak but specific pattern are informative of cell type segregation, and a high Gini coefficient enables them to be selected as IFFs. Since Gini coefficient computation is based on the per-cluster-expressed-fraction rather than normalized expression, this procedure is also insensitive to various modeling assumptions underlying different single cell data normalization methods.

Differential expression analysis

We ran differential expression analysis using the “sSeq” algorithm implemented in the cellrangerRkit package and used FDR < 0.05 and log2 fold change >1 to call differentially expressed genes (DEGs).

Pathway enrichment analysis

To directly compute a per-cell enrichment score for each pathway in the Reactome database (*280*), we used an approach based on the AUCell package (*157*). We slightly modified the standard AUCell pipeline; instead of using all genes for ranking, we initially removed the majority of housekeeping genes using the IFF selection method described above, thereby retaining genes that are mostly cell-type specific (top 25% of genes ranked by Gini coefficient). To derive pathways that are differentially active along the EHT trajectory and between pre-HSC and lympho-myeloid-biased progenitors, we subsequently performed stage-wise Student’s *t* test on the enrichment score (q-value <=

0.01). For Figure 3.2F, 3.6G, Supplemental Figure 3.5D, and Supplemental Figure 3.7, we removed pathways with fewer than 5 genes. Redundant pathways were removed if the Jaccard index (*number of shared genes/number of all genes)* for the pair of pathways was greater than 0.1 and the pathway had a higher q-value. For Supplemental Figure 3.4B, C, we computed gene ontology (GO) enrichment using the ClusterProfiler package (*281*), q-value cutoff of 0.05 and ontology type “Biological Process” (BP).

Pseudotime assignment

We applied Slingshot, one of the best performing trajectory inference methods based on a benchmark study of 45 methods (*135*), to the cleaned data, as described above. Slingshot infers trajectory by fitting a principal curve along a user-selected low dimensional embedding of the data and assigns each cell a pseudotime based on its projection onto the curve. We used the UMAP in Figure 1C, excluding FL-HSC, as the input to the Slingshot algorithm. The starting cluster was set to the “E9.5 E” population, and the terminal cluster was set to the “IAC” population. Computed pseudotime was used for ordering cells along heatmaps in Figure 1G, 2F, and Supplemental Figure 3.7. For Figure 3.2B, C and F, Slingshot was re-run with cells in the Figure 3.2B UMAP, which is a subset of the Figure 3.2A UMAP containing only Wnt^{hi} E, Wnt^{lo} E, Conflux AE and pre-HE. Cells in heatmaps of Figure 3.6E, G and Supplemental Figure 3.11B were ordered based on the PC score, rather than Slingshot-assigned pseudotime.

Single cell RNA velocity analysis

We applied two methods, Velocyto and scVelo to estimate cell velocity in EHT. We used the "*velocyto run10x*" command with mm10 reference genome to quantify spliced and unspliced mRNAs. The output loom file was analyzed using the "velocyto.R" package and the scVelo python package. For the E10.5 E+HE+IAC dataset shown in Figure 3.3, we sequenced ~56k reads/cell, and 13.3% UMIs contained unspliced intronic sequences. Velocyto analysis allows estimation of RNA velocity of single cells by distinguishing between unspliced and spliced mRNAs, which is predictive of the rate of transcriptome change along the EHT trajectory. We used the "*gene.relative.velocity.estimates*" function with *fit.quantile* = 0.05, *deltaT = 1*, *kCells = 20* to calculate RNA velocity and subsequently, visualized the velocity vector field in the UMAP using the "*show.velocity.on.embedding.cor*" function with 20-cell neighborhood and 80 grid points along each UMAP axis. Compared with the steady-state model used in Velocyto, scVelo implements a more sophisticated dynamic model that models the full splicing kinetics. We ran scVelo with its default parameter setting and plotted its predicted velocity on the same UMAP as Velocyto.

Fate probability analysis

We applied Palantir and FateID to determine the probabilities of HE cells becoming pre-HSCs and lympho-myeloid-biased progenitors (*144, 260*). Input to both methods are log2 transformed normalized UMI count matrix, filtered with genes selected by the IFF method. Results with default parameter settings are plotted on the T-SNE embedding by Palantir as shown in Supplemental Figure 3.11D.

Comparison with published scRNA-Seq data

We compared our data to three published scRNA-Seq datasets. Zhou *et al*. (*39*) sequenced 181 cells including E, pre-HSC and HSC cells. Baron et al. (*226*) sequenced 1121 E, HE, EHT and IAC cells from E10 and E11 AGM using CEL-Seq. Mass *et al*. (*282*) sequenced ~90 E10.25 EMP cells. First, we performed PCA on our data using shared genes with the public data, then used the top 10 PCs to compute a UMAP using the *umap* function from the uwot package. Using the PCA loading matrix, we projected public data onto the same PCA space, then predicted UMAP embedding using *umap_transform* function with the previously computed UMAP model. For each projected cell, we mapped it to cell types annotated in this study by 3-nearest-neighbor classification. The final co-embedding for public data with our data are shown in Supplemental Figure 3.12.

scATAC-Seq data analysis

Data pre-processing and peak calling

scATAC-Seq reads were aligned to the mouse mm10 reference genome using the "*cellranger-atac mkfastq*" command. Peaks were called using MACS2 with the FDR cutoff of 0.10 and the following parameters: -q 0.10 --broad --broad-cutoff 0.10) (*283*). Quality control statistics of the data were generated using the scATAC-pro package (*284*) and are shown in Supplemental Figure 3.9. We implemented a custom PERL script to quantify the reads overlapping with peaks individually for each cell. The read is considered to overlap with the peak if at least half of the read overlaps. By quantifying the reads for each individual cell, we obtained a peak-barcode matrix. The peak-barcode matrix underwent an initial filtering, requiring a barcode to have at least 2,000 fragments, 1,000 detected peaks and at least 20 percent fragments in peak to be considered as a "cell". Additionally, for each peak, we computed the fraction of cells with non-zero value, and removed peaks that were detected in fewer than 1% of all cells. The final cleaned matrix contains 1670 cells and 150,427 peaks. We also computed a normalized data matrix by first log2 transforming the data (with 1 pseudocount added) and regressing out variance explained by total detected peaks per cell estimated with the "*lmFit*" function in the limma package (*285*). This normalized matrix was used for differential accessibility test and preliminary matching of scRNA-Seq and scATAC-Seq clusters.

scATAC-Seq clustering and differential accessibility analysis

Traditional feature selection methods designed for scRNA-Seq data do not work well on scATAC-Seq data, due to much greater sparsity and a binary data distribution (per genomic locus per cell, the expected read count is 0, 1 or 2). This makes it difficult to select the most informative features for clustering and cell identity mapping.

Using the IFF selection method described above, we were able to obtain a UMAP with cells separated into several distinct neighborhoods, which significantly improved the Louvain clustering quality (Supplemental Figure 3.4F-H). To identify differentially accessible peaks (DAPs), we first binarized data as either open (>1) or closed (0), then calculated the fraction of cells that have open states for each peak in each cluster. We ran one-vs-rest Chi-square test on the fractions and called cluster-specific peaks (DAPs) using FDR < 0.05 and absolute log2 fold change >1. Fold change is defined as the ratio of open fractions between the two groups.

We observed that DAPs and DEGs show strong co-enrichment patterns at genomic loci for certain pairs of scATAC-Seq and scRNA-Seq clusters (Supplemental Figure 3.10A). Many of the DAPs are located near promoters of the matched DEG, but some are much more distant. By computing a co-enrichment value, we established an initial mapping between DAPs and DEGs (Supplemental Figure 3.10A).

Seurat alignment and co-embedding of scATAC-Seq and scRNA-Seq cells

We used the Seurat alignment algorithm (*279*) to co-embed scATAC-Seq and scRNA-Seq cells onto a single UMAP shown in Figure 3.4A. We first matched each cell-type-specific DEGs with corresponding DAPs 200 kb up and downstream of the TSS, using the method described above, obtaining 12,768 links between 2,379 DEGs and 10,126 DAPs. We then summed up DAP fragments for each DEG, obtaining a gene-by-cell activity score matrix as the “gene activity matrix” for Seurat alignment. Transfer anchors were computed using the “*FindTransferAnchors*” function in Seurat, with dimensionality reduction method set to “cca” (canonical correlation analysis). scATAC-Seq cells were then transferred to scRNA-Seq reference using the “*TransferData*” function, using 15 nearest neighbors and PCA for computing the weighted correction vectors. Finally, scATAC-Seq and scRNA-Seq Seurat objects were merged using the “*merge.Seurat*” function, and joint UMAP was computed with top 20 PCs and 15 nearest neighbors. Cell type labels were transferred from scRNA to scATAC using Seurat. Contaminant cell types, including mesenchymal, *Lyve1*$^{+}$ E and *Bnip3*^{hi} E were removed from both scRNA-

Seq and scATAC-Seq data, and only cells involved in EHT were used to generate the UMAP shown in Figure 3.4A.

Inference of enhancer-promoter (E-P) links

The DAP-DEG matching procedure could link cell-type-specific peaks to nearby cell-type-specific genes. However, this association required both differential expression and differential chromatin accessibility to be significant, potentially missing E-P links with weaker signal. Inspired by the Seurat alignment algorithm and eQTL inference method, we used linear regression on matched scATAC-Seq and scRNA-Seq meta cells to find gene-distal peaks (defined as enhancers) that have a chromatin accessibility pattern significantly correlated with a gene's expression. First, for each scATAC-Seq cell, we paired it to its nearest scRNA-Seq neighbor in the joint UMAP, establishing links between 1,186 scATAC-Seq cells and 659 scRNA-Seq cells. Note that not all scRNA-Seq cells were paired with a scATAC-Seq cell and some scRNA-Seq cells were paired to multiple scATAC-Seq cells, but the paired scRNA-Seq cells were uniformly distributed along the EHT trajectory.

To overcome the sparsity in scRNA-Seq and scATAC-Seq data, we expanded the paired scATAC-Seq and scRNA-Seq neighbors to paired scATAC-Seq and scRNA-Seq neighborhoods by pooling counts from 10 nearest neighbors. We normalized the pooled expression and accessibility by regressing out per-meta-cell total counts from log2 transformed data, followed by z-score transformation. For each expressed gene, we ran linear regression with its pooled expression against pooled accessibility peaks 200kb upstream and downstream of its TSS in the paired scATAC-Seq meta-cell. Links with

Bonferroni corrected p-values < 0.01 and regression coefficients > 0.1 were considered significant and these called peaks are likely enhancers that contribute to the corresponding gene's expression. E-P links for Runx1 shown in Figure 3.5 were called separately, including additional peaks within 500kb upstream and downstream of Runx1 P1. For genes with multiple promoters (defined as regions 2,000 bp upstream and 500 bp downstream of each TSS), we ran a second regression using the promoter accessibility as dependent variable and each called peak as independent variable. This allowed us to link each called enhancer to a specific promoter. We computed "cis-regulatory-activity matrix" based on the called E-Ps, and observed consistent pattern between a gene's expression and its cis-regulatory-activity score (Supplemental Figure 3.10B).

TF activity assessment using chromVar

We assessed TF binding activity to enhancers with chromVar (*246*) using its default setting, but changed the default p.value cutoff in "*matchMotifs"* function to 0.1 / (2 * median(enhancer length)) to account for multiple testing. The input TF motifs were curated from the CIS-BP motif database (*286*). For each TF motif and each cell, a GC-bias and background-corrected deviation score was computed using the "*computeDeviations*" function, which represents the relative gain or loss of TF binding activity. Lastly, to identify TFs with stage-specific binding activity, we ran stage-wise Mann-Whitney U-test with the deviation scores, and considered those with FDR < 0.05 as significant.

Figures

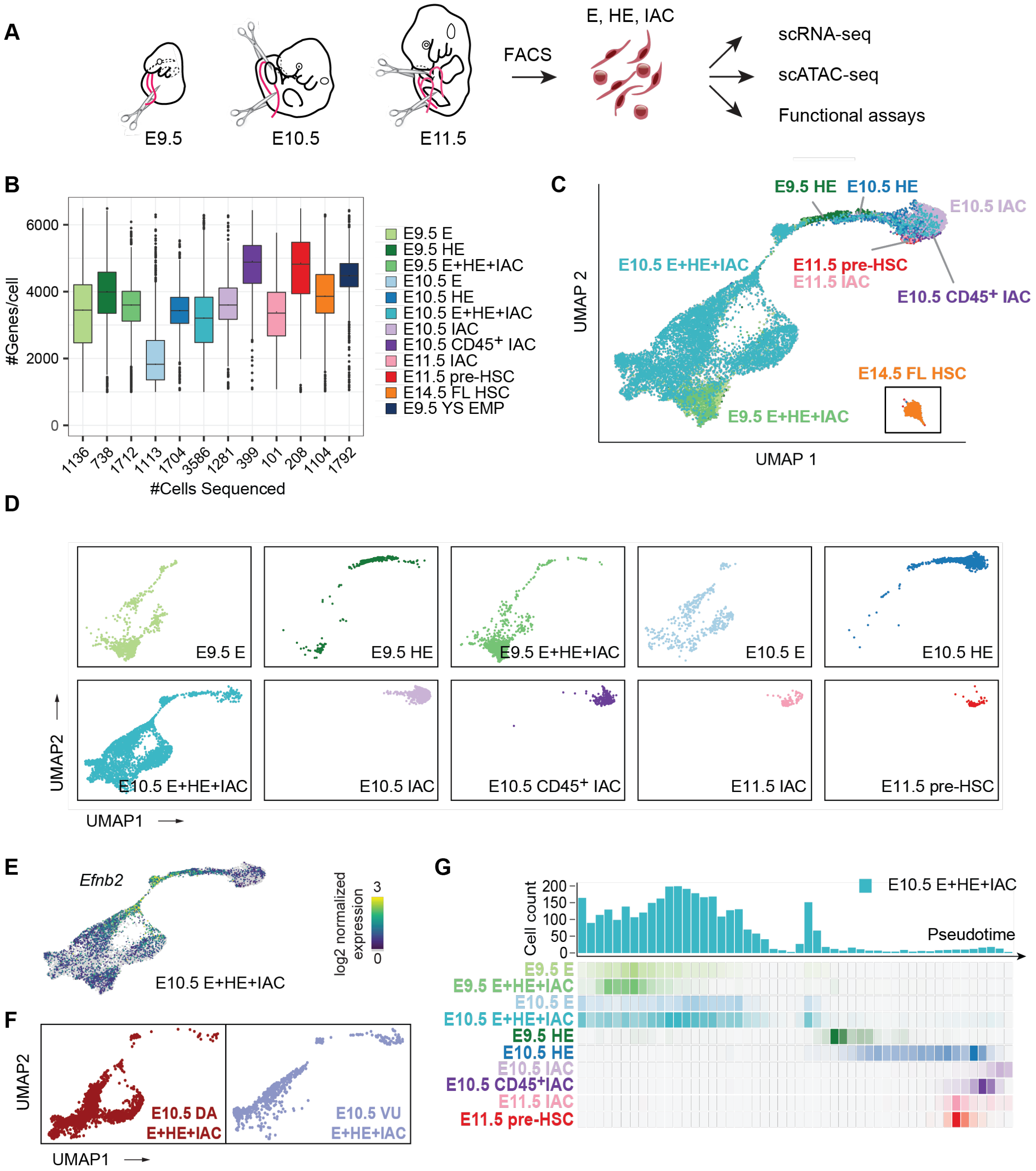


Figure 3.1 Experimental design, and overview of single cell RNA-Seq data.

(A) The caudal part of embryos were isolated (boundaries are illustrated with scissors), then organs and gut tube removed. Vitelline and umbilical arteries (VU) were isolated and included in the sample. The tissue was dissociated and cells were isolated by FACS, then analyzed by scRNA-Seq, scATAC-Seq, or in functional assays. All cell populations purified and sequenced are listed in Table 3.1, and sort plots are shown in Supplemental Figures 3.1 and 3.2. **(B)** The number of cells sequenced (x-axis) and genes per cell detected for representative samples. **(C)** UMAP of continuous EHT trajectory and FL-HSCs, with selected cell populations labeled. **(D)** Distribution of cells from each dataset in the UMAP reflecting EHT trajectory. **(E)** UMAP illustrating the two streams of E cells expressing high levels of the arterial marker *Efnb2* that converge to form the stem leading to HE and IACs. **(F)** E+HE+IAC cells separately purified from the vitelline and umbilical (VU) arteries, and from the dorsal aorta (DA) within the caudal half of the embryo, highlighted on the global UMAP plot. **(G)** Cell count along the pseudotime trajectory. Bar graph quantifies results from a single sort of E10.5 E+HE+IAC cells; heat maps below the graph show distribution of cells in all sorted cell populations.

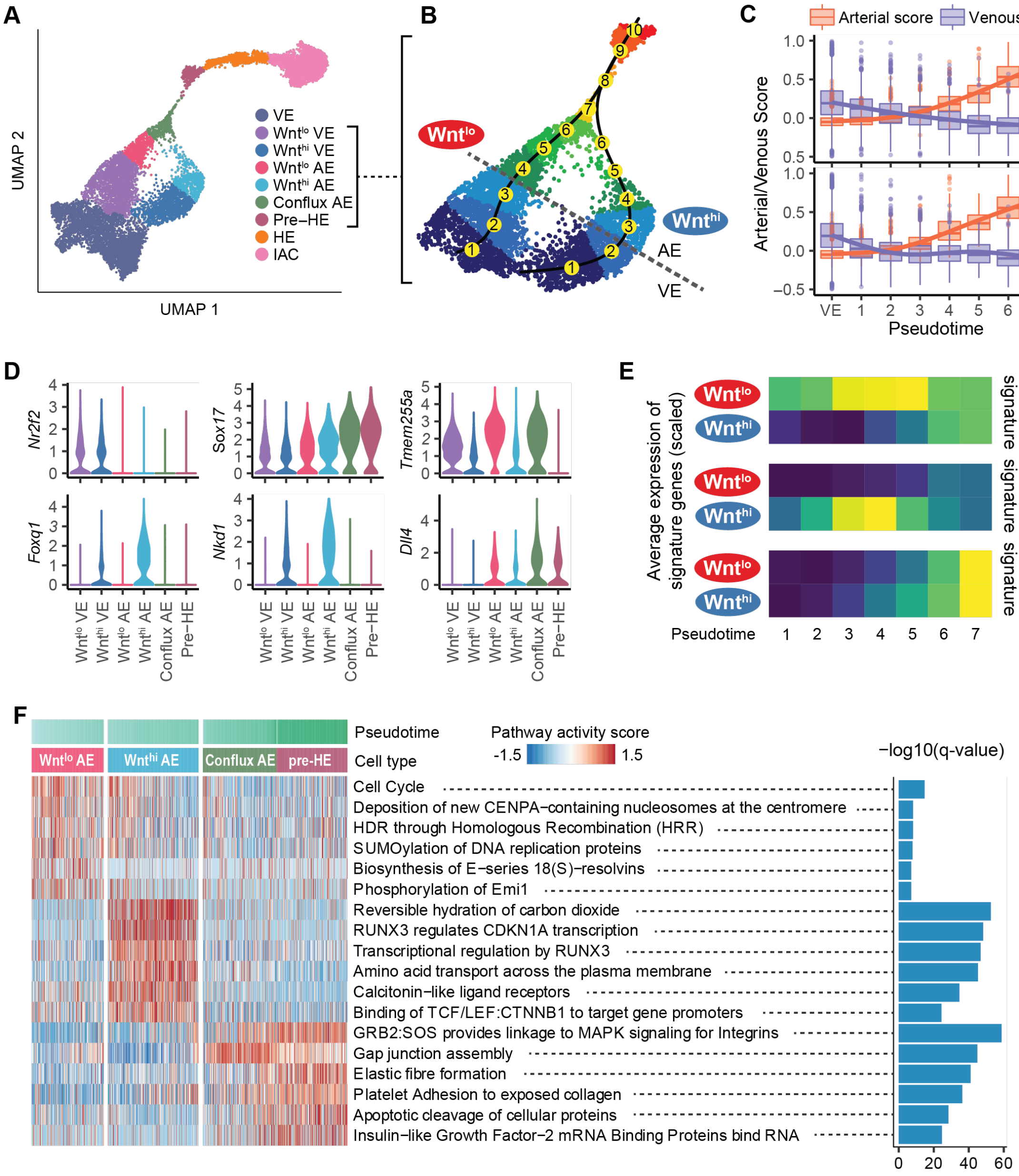


Figure 3.2 Two streams of endothelial cells converge before hemogenic endothelium.

(A) UMAP of EHT trajectory (from Figure 3.1C, with FL-HSC removed) showing the 7 clusters identified by Louvain clustering in Supplemental Figure 3.5A, with Wnt^{hi} E subdivided into Wnt^{hi} AE and Wnt^{hi} VE, plus Wnt^{lo} E subdivided into Wnt^{lo} AE and Wnt^{lo}

VE based on the arterial/venous score determined as shown in panels B and C. **(B)** Zoom-in UMAP highlighting the two streams of endothelial cells converging to conflux AE. Numbers in yellow circles represent pseudotime bins up to the point of convergence. The dotted gray line represents the boundary between AE and VE. **(C)** Arterial score vs venous score over pseudotime bins. Cluster VE from panel A is used as the first pseudotime bin. Curves are fitted for AE score and VE score of each branch using a generalized additive model. **(D)** Violin plots of expression of cluster specific genes, including venous marker *Nr2f2*, arterial marker *Sox17*, Wnt^{lo} AE specific gene *Tmem255a*, Wnt^{hi} AE specific gene *Foxq1* and *Nkd1*, and Notch ligand *Dll4*. **(E)** Average expression of Wnt^{lo} E, Wnt^{hi} E, and pre-HE-specific genes over pseudotime. Differentially expressed genes were derived by pairwise expression analysis between Wnt^{lo} E and Wnt^{hi} E. Pre-HE specific genes were derived by comparing pre-HE with Wnt^{lo} plus Wnt^{hi} E. **(F)** Heatmap showing stream-specific Reactome pathway activity over pseudotime. AUCell package(*157*) was used to compute a pathway activity score for each cell. One vs the rest Student's t-test was used to identify group-specific pathways and the top 6 most significant pathways were plotted.

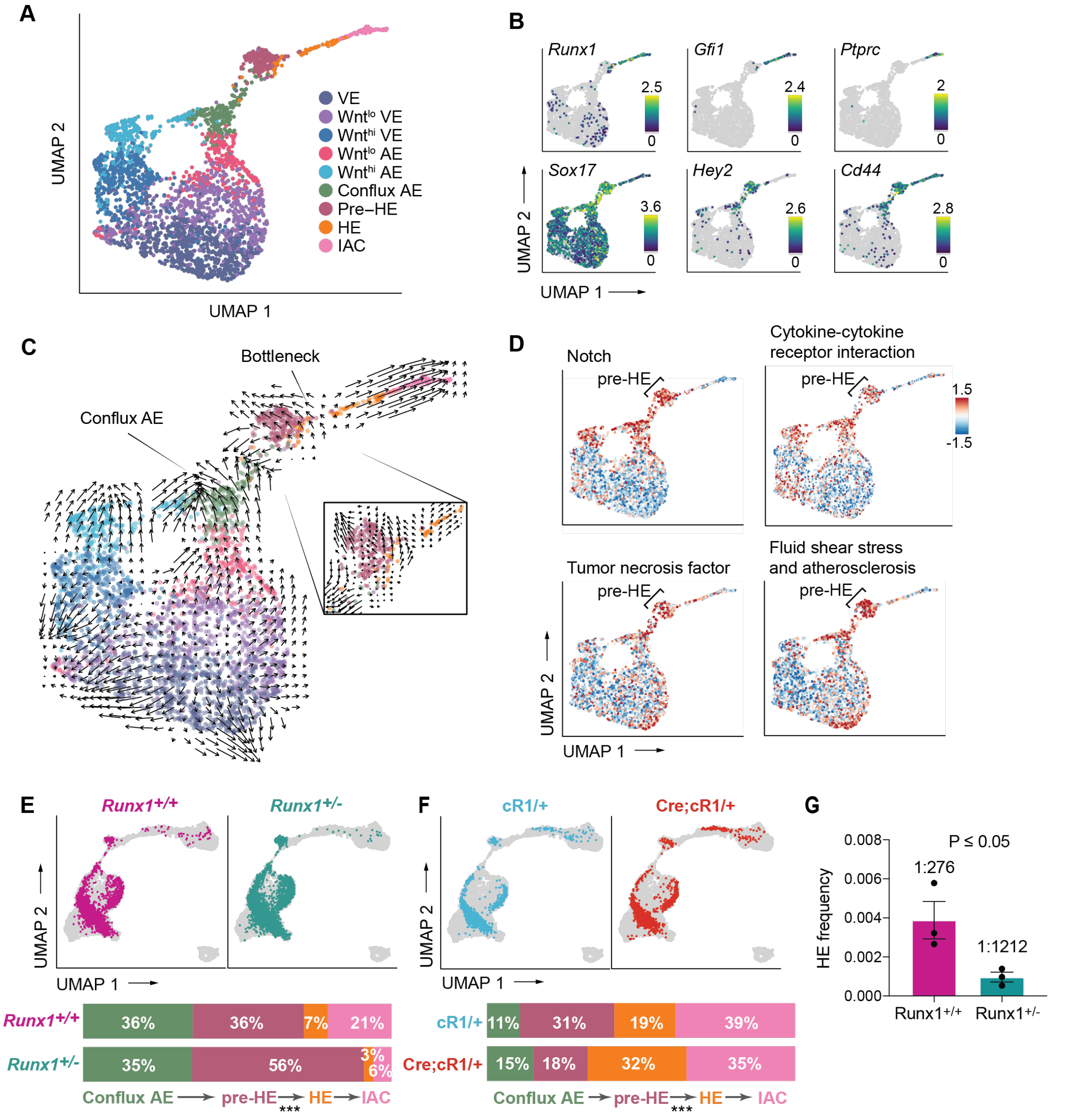


Figure 3.3 Developmental bottleneck between pre-HE and HE cells.

(A) UMAP of E10.5 E+HE+IAC cells showing 9 cell types from Figure 3.2A. **(B)** Expression of key markers of clusters, including *Hey2* in conflux AE and pre-HE, *Cd44* in conflux AE, pre-HE, HE and IACs, *Ptprc* in IACs, *Gfi1* and *Runx1* in HE and IACs, and high levels of *Sox17* in conflux AE and pre-HE, with downregulation in HE. Note *Runx1* is expressed at low levels in all subsets of endothelial cells. **(C)** Velocyto analysis revealing different differentiation dynamics along the EHT in E10.5 E+HE+IAC cells. To the right is a zoom-in velocity of pre-HE cells that have accumulated at the bottleneck between pre-HE and HE. **(D)** Activity of pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG) database, computed for each cell using the AUCell method (*157*). **(E)** UMAP of E+HE+IAC cells from E10.5 *Runx1*$^{+/+}$ and *Runx1*$^{+/-}$ littermates. Bars on the bottom depict the distribution of cells between conflux AE, pre-HE, combined HE, and IAC populations in E10.5 *Runx1*$^{+/+}$ and *Runx1*$^{+/-}$ littermates. *P*-values indicate significant differences in the distributions of cells in pre-HE and HE in *Runx1*$^{+/+}$ versus *Runx1*$^{+/-}$ samples based on proportion test. **(F)** UMAP of E+HE+IAC cells from E10.5 control embryos (cR1/+) and littermates ectopically expressing RUNX1 in all endothelial cells from the *Rosa26* locus (Cre;cR1/+) (*242*). Bars on the bottom as in panel E. **(G)** Limiting dilution assay to determine the frequency of HE in the CD44^{+} fraction of E+HE+IAC cells isolated from E10.5 embryos (see Supplemental Figure 3.2G for FACS plots). Shown are frequencies of cells that yielded hematopoietic cells (B220^{+}, CD19^{+}, Mac1^{+}, Gr1^{+}, and/or CD41 and CD45) *ex vivo.* Frequencies were calculated by ELDA(*274*). Data represent three independent cell purifications and limiting dilution assays (mean ± SD, unpaired two-tailed Student's *t*-test).

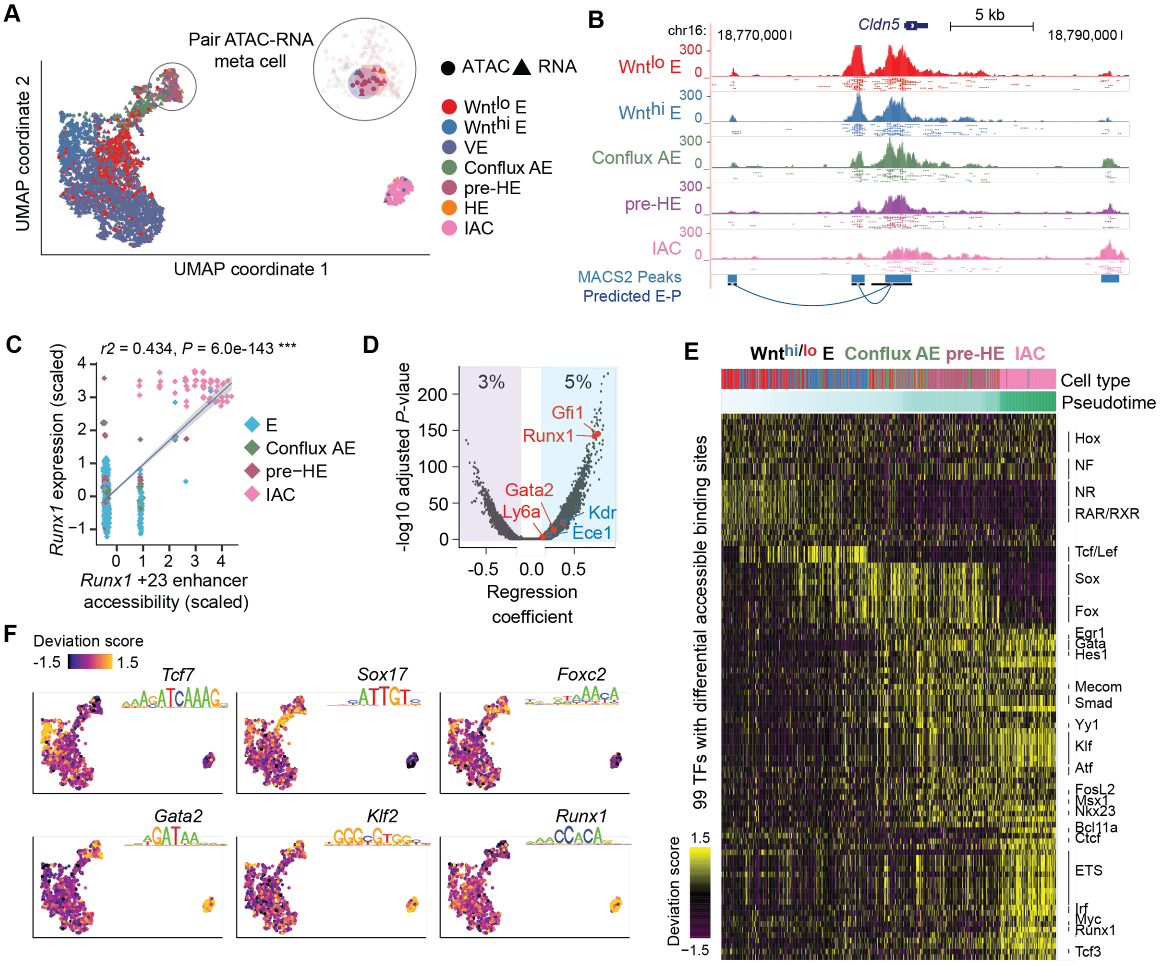


Figure 3.4 Joint scRNA-Seq and scATAC-Seq analysis of bottleneck populations.

(A) UMAP of 1637 cells from scRNA-Seq and 1186 cells from scATAC-Seq, aligned using Seurat algorithm with a custom defined gene-by-cell activity score matrix (see Materials and Methods). The number of HE cells was too few to be resolved by UMAP, and clustered with pre-HE. To gain enough statistical power for predicting E-P, we pooled reads from 10 nearest neighbors as “meta cells”, and paired scATAC meta cells to nearby scRNA meta cells. Additional details can be found in the Materials and Methods section. **(B)** UCSC genome browser tracks showing open chromatin signal of *Cldn5* promoter and its predicted enhancers. Dots below each aggregated signal track represent signal from 50 sampled cells of each type. **(C)** Linear regression shows high correlation between *Runx1* +23 enhancer chromatin accessibility and *Runx1* expression levels (z-score transformed). Each point represents a paired ATAC-RNA meta cell in panel A, with pooled RNA expression on the y-axis and pooled enhancer accessibility on the x-axis. **(D)** Prediction of enhancer-promoter interaction using linear regression. Predictions (points in blue shaded area, 5% of total candidate interactions) were made using $p < 0.01$ and regression coefficient > 0.1. We recapitulated the majority of known enhancer-promoter interactions (E-Ps) that function during EHT, with the *Runx1* +23 enhancer (*287*) and *Gfi1* enhancer (*288*) among the top predictions. **(E)** TF binding patterns among called scATAC-Seq peaks assessed using chromVar (*246*), which defines a deviation score reflecting the accessibility change at binding sites of each TF across all cells. Binding sites were determined using DNA motif scan on the called enhancers, which does not discriminate TFs in the same family with very similar motifs. Top significant TFs based on Mann-Whitney U test are plotted for each stage. **(F)** ChromVar deviation score for selected TF motifs plotted on the UMAP, showing specific binding pattern for Tcf7 in Wnt^{hi} E, Sox17 in conflux E, Foxc2 in pre-HE, Gata2 and Klf2 in both pre-HE and IAC. Runx1 binding sites are highly accessible post bottleneck, but also exhibit medium to high level of chromatin accessibility in some early-stage cells.

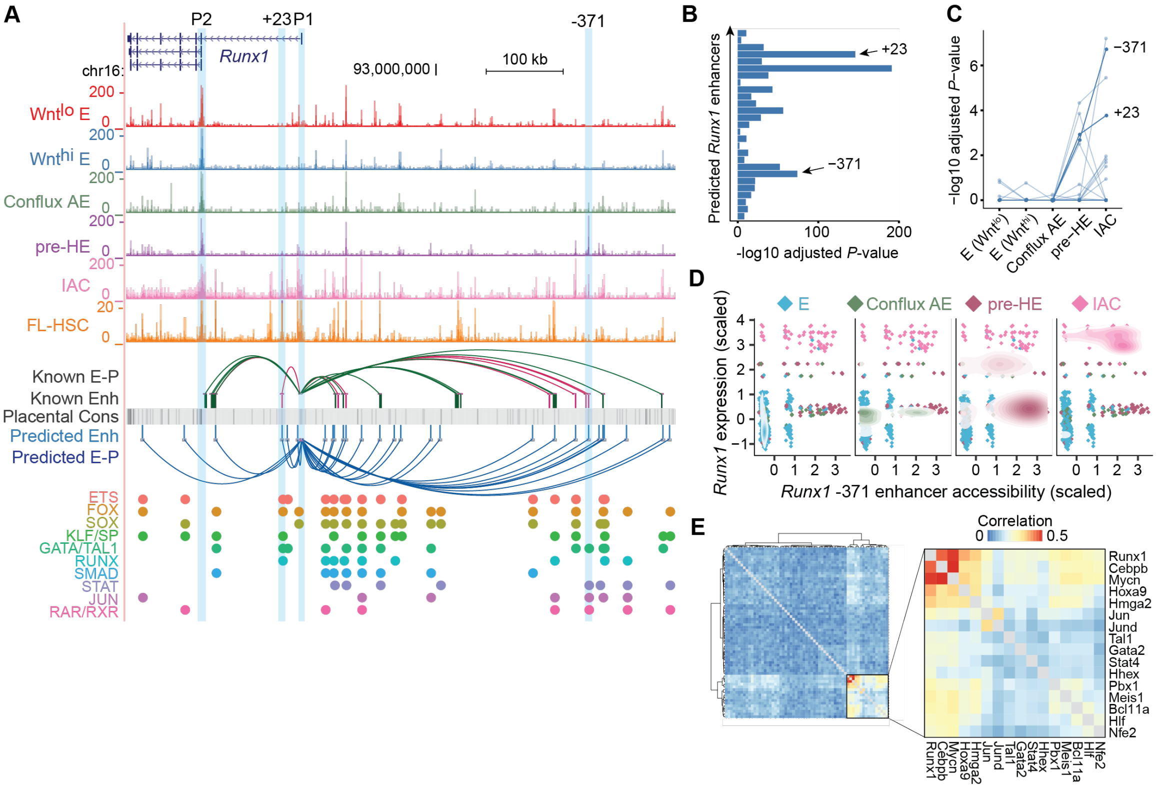


Figure 3.5 Developmental-stage-specific enhancers of *Runx1*.

(A) UCSC genome browser tracks showing open chromatin signal for each of the populations. Tracks from E to IAC are cumulative scATAC-Seq signals (per-base unique fragment coverage) normalized by the number of cells in that population. Tracks for FL-HSC are bulk ATAC-Seq data from Chen, C. *et al.* (*252*). Experimentally validated enhancers and E-Ps from Marsman *et al*. (*251*) are shown in magenta. Enhancers and E-P links from Chen, C. *et al.*(*252*) are shown in dark green. E-P links were inferred based on linear regression on paired scRNA-scATAC meta cells (see Materials and Methods). Placental mammal conservation by PhastCons score is shown as a grey track. For each of the inferred enhancers, we scanned for known motifs from CIS-BP database and grouped TFs from the same family having similar motifs. Motif hits of several previously reported early hematopoietic TFs are highlighted below the track. **(B)**

Distribution of linear regression P-values for predicted *Runx1* enhancers. Highly significant peaks include the validated +23 and -371 enhancers. The most significant peak is ~3.6 kb downstream of P1. **(C)** Co-accessibility of *Runx1* P1 promoter and its predicted linked enhancers in each cell type. P-values for co-accessibility in each cell type were computed using Fisher's exact test with multiple testing correction. **(D)** Stage-specific chromatin accessibility of *Runx1* -371 enhancer and *Runx1* expression levels (z-score transformed). Each point in the scatter plot represents a paired ATAC-RNA meta cell in Figure 3.5A, with pooled RNA expression on the y-axis and pooled enhancer accessibility on the x-axis. A 2-dimensional density plot is superimposed on the scatter plot. **(E)** Co-expression of transcription factors that have binding motifs at *Runx1* enhancers and whose expression precedes *Runx1*. Correlations were computed using gene expression matrix including conflux E, pre-HE and HE cells. TFs with Pearson correlation with Runx1 < 0.05 were removed. Hierarchical clustering was performed on the correlation matrix and a strong TF co-expression module was highlighted.

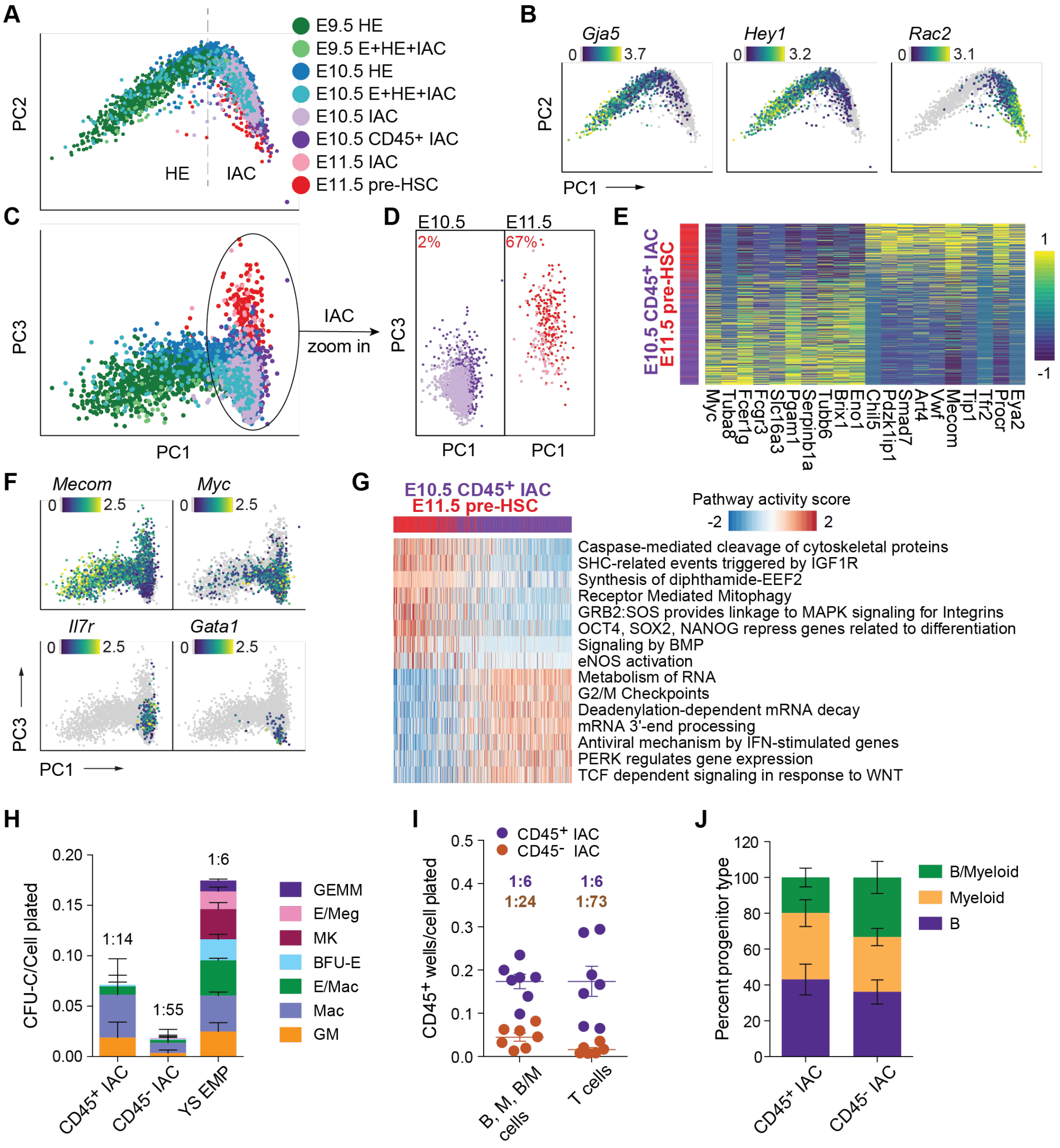


Figure 3.6 Two waves of CD45$^+$ HSPCs in IAC cells.

(A) PCA plot of a subset of data containing IACs, illustrating the trajectory of IAC differentiation from HE along the PC1 axis. **(B)** Expression of *Gja5*, *Hey1*, and *Rac2* illustrating the maturation of IAC cells along the trajectory. **(C)** PCA plot showing the separation of E10.5 and E11.5 IAC cells along the PC3 axis. **(D)** E10.5 and E11.5 IAC cells, E10.5 CD45$^+$ IAC cells, and E11.5 pre-HSCs plotted separately to visualize their relative distribution along the PC3 axis. A K-nearest-neighbor classifier (K = 3 with PC1-10 as feature input) was trained using E10.5 CD45$^+$ IAC cells and E11.5 pre-HSCs to determine the fraction of pre-HSCs (labeled in red) in E10.5 and E11.5 IAC cells. **(E)** Heatmap showing top differentially expressed genes in E10.5 CD45$^+$ IAC cells versus E11.5 pre-HSCs. **(F)** Preferential expression of *Mecom* in E11.5 pre-HSCs and IAC cells, versus *Myc, Il7r*, and *Gata1* in E10.5 CD45$^+$ IAC enriched for lympho-myeloid-biased progenitors and in E10.5 IAC cells. **(G)** Reactome pathway analysis comparing E11.5 pre-HSC and E10.5 CD45$^+$ IAC cells. Color indicates pathway activity score computed using the AUCell package (*157*). **(H)** Methylcellulose (colony forming unit-culture, CFU-C) assay performed in the presence of stem cell factor (SCF), interleukin 3 (IL3), IL6, and erythropoietin (EPO) to measure the frequency of committed erythroid and myeloid progenitors in E10.5 CD45$^+$ IAC cells, CD45$^-$ IAC cells, and E9.5 yolk sac EMPs. BFU-E, burst forming unit-erythroid; GM, granulocyte/macrophage; Mac, macrophage; MK, megakaryocyte; GEMM, granulocyte/erythroid/monocyte/megakaryocyte. Error bars; mean ± SD. Frequencies of total progenitors are indicated above the bars. n = 3 experiments. **(I)** Limiting dilution assays on OP9 stromal cells to determine the frequencies of progenitors in purified E10.5 CD45$^+$ IAC and E10.5 CD45$^-$ IAC cells yielding B (CD45$^+$CD19$^+$B220$^{mid/lo}$), myeloid (M) (Gr1$^+$Mac1$^+$or Gr1$^+$Mac1$^-$), and B+myeloid (B/M) cells in culture. Also shown are frequencies of progenitors in purified E10.5 CD45$^+$ IAC and E10.5 CD45$^-$ IAC cells that produced T cells (CD90$^+$ CD25$^+$) when cultured on OP9 cells expressing the Notch ligand delta like 1. Error bars; mean ± SD. Frequencies of all progenitors are indicated above the bars. n = 7 experiments. **(J)** Percentage of wells at the limiting cell dose containing B, M, or B/M cells from experiments in panel I. n = 8 experiments.

Tables

Table 3.1. Summary statistics for collected cell populations.

Dataset/ Cell type[1]	Surface Phenotype	# Embryos	Somite Pairs	# Cells Sequenced	Median # Expressed Genes	Median # UMIs
E9.5 E [1]	CD41$^-$ CD45$^-$ Kit$^-$ CD31$^+$ Runx1:GFP$^-$	60	22-27	1582	3454	15703
E9.5 E [2] [2]	Ter119$^-$ CD41$^-$ CD45$^-$ Kit$^-$ CD144$^+$ ESAM$^+$ Runx1:GFP$^-$	51	25-27	754	3398	12491
E9.5 HE [1]	CD41$^-$ CD45$^-$ Kit$^{lo/-}$ CD31$^+$ Runx1:GFP$^+$	60	22-27	182	3418	12554
E9.5 HE [2] [2]	Ter119$^-$ CD41$^-$ CD45$^-$ Kit$^{lo/-}$ CD144$^+$ ESAM$^+$ Runx1:GFP$^+$	51	25-27	686	4094	18223
E9.5 E+HE+IAC	Ter119$^-$ CD41$^{lo/-}$ Kit$^{lo/-}$ CD31$^+$ CD144$^+$ ESAM$^+$	60	23-26	2171	3598	17139
E9.5 EMP	CD41$^+$ Kit$^+$ CD16/32$^+$	30	25-29	1875	4476	25617
E10.5 E [1]	CD41$^-$ CD45$^-$ Kit$^-$ CD31$^+$ Runx1:GFP$^-$	43	ND[3]	1073	1675	5245
E10.5 E [2] [2]	Ter119$^-$ CD41$^-$ CD45$^-$ Kit$^-$ CD144$^+$	100	33-38	1017	2672	8131

	ESAM$^+$ Runx1:GFP$^-$					
E10.5 HE	Ter119$^-$ CD45$^-$ CD31$^+$ CD144$^+$ ESAM$^+$ Kit$^{lo/}$ CD41$^{lo/mid/-}$ Runx1:GFP$^+$	84	33-38	1795	3431	15518
E10.5 E+HE+IAC	Ter119$^-$ CD41$^{lo/-}$ CD31$^+$CD144$^+$ ESAM$^+$	60	33-36	4776	3208	11991
E10.5 IAC	Ter119$^-$ CD41$^{lo/-}$ CD31$^+$ CD144$^+$ ESAM$^+$ Kit$^+$	60	33-36	2288	3600	18246
E10.5 CD45$^+$ IAC	CD31$^+$ CD144$^+$ ESAM$^+$ Kit$^+$ CD45$^+$	18	33-37	602	4910	29978
E10.5 VU E+HE+IAC	Ter119$^-$ CD41$^{lo/-}$ CD31$^+$CD144$^+$ ESAM$^+$	46	35-37	1173	4059	17760
E10.5 DA E+HE+IAC	Ter119$^-$ CD41$^{lo/-}$ CD31$^+$CD144$^+$ ESAM$^+$	10	33-37	1951	3133	13707
E10.5 Runx1$^{+/+}$ E+HE+IAC [1]	Ter119$^-$ CD41$^{lo/-}$ CD31$^+$CD144$^+$ ESAM$^+$	7	34-37	963	4005	17965
E10.5 Runx1$^{+/-}$ E+HE+IAC [1]	Ter119$^-$ CD41$^{lo/-}$ CD31$^+$CD144$^+$ ESAM$^+$	11	34-37	1355	4320	20296
E10.5 Runx1$^{+/+}$ E+HE+IAC [2]	Ter119$^-$ CD41$^{lo/-}$ CD31$^+$CD144$^+$ ESAM$^+$	8	32-35	2449	3290	14659

E10.5 Runx1$^{+/-}$ E+HE+IAC [2]	Ter119^{-} CD41$^{lo/-}$ CD31^{+}CD144^{+} ESAM^{+}	10	32-35	3944	3396	14880
E10.5 CD44^{+} E+HE+IAC (paired scRNA) [1]	Ter119^{-} CD41$^{lo/-}$ CD31^{+}CD144^{+} ESAM^{+} CD44$^{lo/+}$	56	33-37	2317	3565	15307
E10.5 CD44^{+} E+HE+IAC (paired scATAC) [1]	Ter119^{-} CD41$^{lo/-}$ CD31^{+}CD144^{+} ESAM^{+} CD44$^{lo/+}$	56	33-37	2317	3565	15307
E10.5 CD44^{+} E+HE+IAC (cR1/+)[3]	Ter119^{-} CD41$^{lo/-}$ CD31^{+}CD144^{+} ESAM^{+} CD44$^{lo/+}$	42	30-34	3437	2505	8200
E10.5 CD44^{+} E+HE+IAC (Cre;cR1/+)[4]	Ter119^{-} CD41$^{lo/-}$ CD31^{+}CD144^{+} ESAM^{+} CD44$^{lo/+}$	31	30-34	4845	2741	9696
E11.5 IAC [1]	Ter119^{-} CD31^{+} CD144^{+} ESAM^{+} Kit^{+}	12	ND	1267	3520	16230
E11.5 IAC [2]	Ter119^{-} CD41$^{lo/-}$ CD31^{+} CD144^{+} ESAM^{+} Kit^{+}	71	45-48	299	3304	17188
E11.5 pre-HSC	CD144^{+} CD45^{+} CD27^{+}	41	ND	279	4854	29507
E11.5 FL-LMPP[5]	Ter119^{-} Nk1.1^{-} Gr1^{-} CD3e^{-} CD19^{-} B220^{-} F4/80^{-} Kit^{+}	90	ND	1860	3400	16624

	CD45$^+$CD135$^+$ IL7ra$^+$					
E14.5 FL-HSC	Ter119$^-$ Gr1$^-$ B220$^-$ CD3e$^-$ Sca1$^+$ Kit$^+$ CD48$^-$ CD150$^+$	22	ND	1108	3865	16500
Total		1134	-	44940	-	-

[1] Numbers in brackets indicate replicate samples.

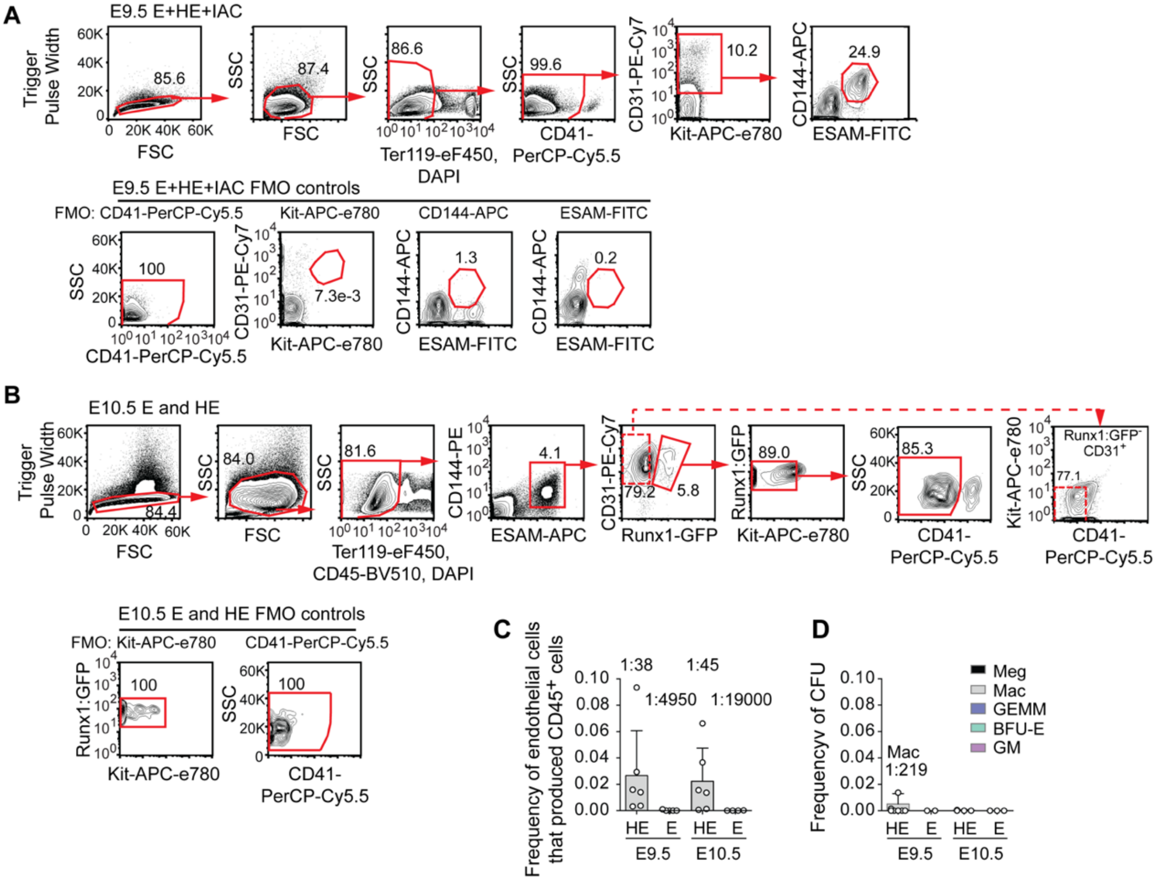
[2] Markers were adjusted on subsequent sorts to reduce the fraction of contaminating cells detected by scRNA-Seq.

[3] ND, not determined

[4] Ectopic RUNX1 expression in endothelial cells in Tg(*Cdh5-cre/ERT2*)1Rha embryos (Cre) (*273*) that contained an activatable *Runx1* cDNA in the *Rosa26* locus (cR1/+) (*242*). Cre was activated by injection of 1 mg tamoxifen into pregnant dams at E9.5.

[5] LMPP, lymphoid-primed multipotent progenitor.

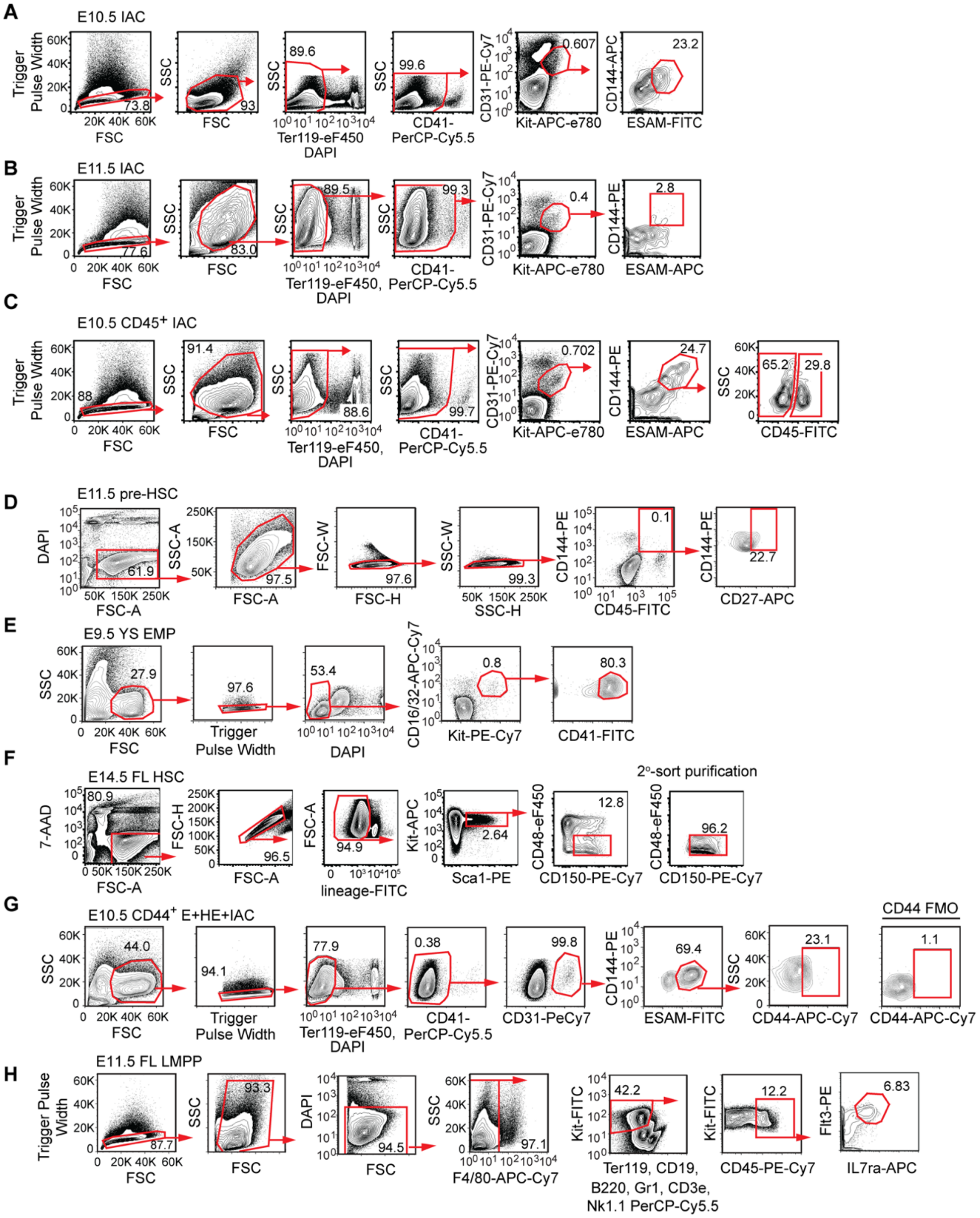
Supplemental Figures



Supplemental Figure 3.1 Purification of endothelial cell populations by FACS.

(A) Quantile contour FACS plots depicting gating strategy for isolating E+HE+IAC cells (plots are representative sorts from E9.5 embryos). Embryos were collected from B6C3F1 females mated to C57BL6/J males. Bulk endothelium and clusters were purified as Ter119^{-}CD41$^{mid/-}$CD31^{+}CD144^{+}ESAM^{+}cells. Kit was not used to exclude cells in order to capture IAC cells with the other endothelial cell populations. Fluor-minus-one (FMO) controls for CD41, Kit, CD144 (vascular endothelial cadherin), and ESAM are shown in the bottom four panels. Numbers on the x and y-axes are indicated on the first plot on the left, and unless changed are not depicted on plots to the right of the preceding plot.

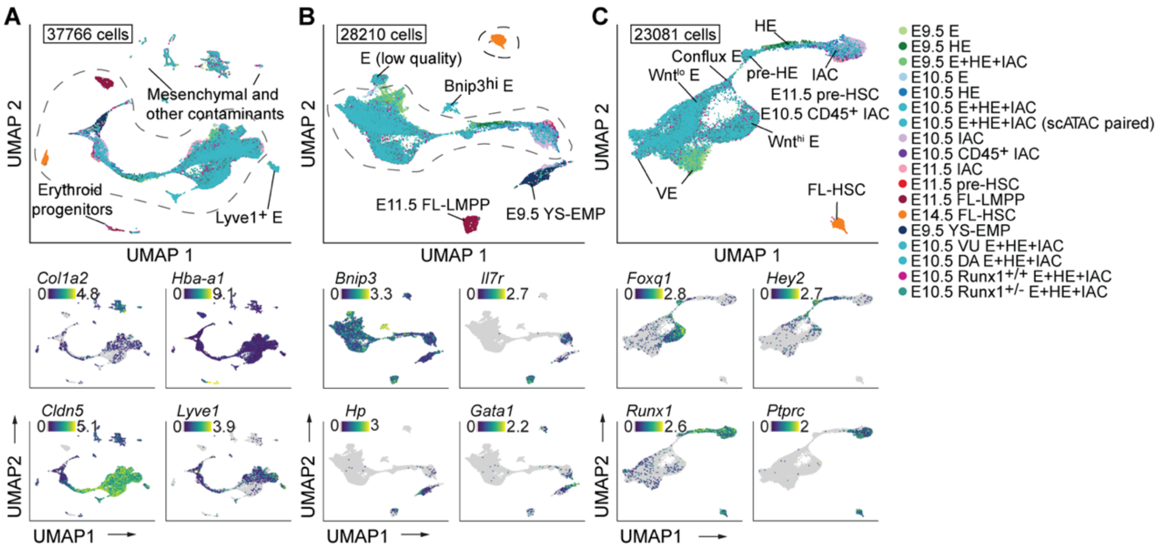
eF450, eFluor^{TM}-450; APC-e780, APC-eFluor 780; eF450, eFluor 450. **(B)** Quantile contour FACS plots depicting gating strategies for isolating endothelial cells (E) (Ter119^{-}CD41^{-}CD45^{-}CD31^{+}CD144^{+}ESAM^{+}Kit^{-}Runx1:GFP^{-}) and hemogenic endothelial cells (HE) (Ter119^{-}CD41^{-}CD45^{-}CD31^{+}CD144^{+}ESAM^{+}Kit$^{lo/-}$Runx1:GFP^{+}) from E10.5 *Runx1*-IRES-GFP embryos. Embryos were collected from *Runx1*-IRES-GFP male mice (*229*) mated to B6C3F1 females. FMO controls for Kit and CD41 are shown in the two panels below. **(C)** Frequency of purified endothelial cells from E9.5 and E10.5 embryos that gave rise to CD45^{+}cells when cultured in a limiting dilution assay on OP9 stromal cells. Frequencies are indicated on top of the bars. HE, Runx1:GFP^{+} endothelial cells, purified as described in panel B; E, Runx1:GFP^{-} endothelial cells. Error bars, mean ± standard deviation (SD); n = 5-6 experiments. Frequencies were calculated using ELDA software (*274*). **(D)** Colony forming units (CFU) representing frequency of contaminating committed HSPCs in purified HE and E populations. Meg, megakaryocyte; Mac, macrophage; GEMM, granulocyte/erythroid/monocyte/megakaryocyte; BFU-E, burst forming unit-erythroid; GM, granulocyte/monocyte. Error bars, mean ± SD; n = 2-3 experiments.



Supplemental Figure 3.2 IAC cell purification by FACS.

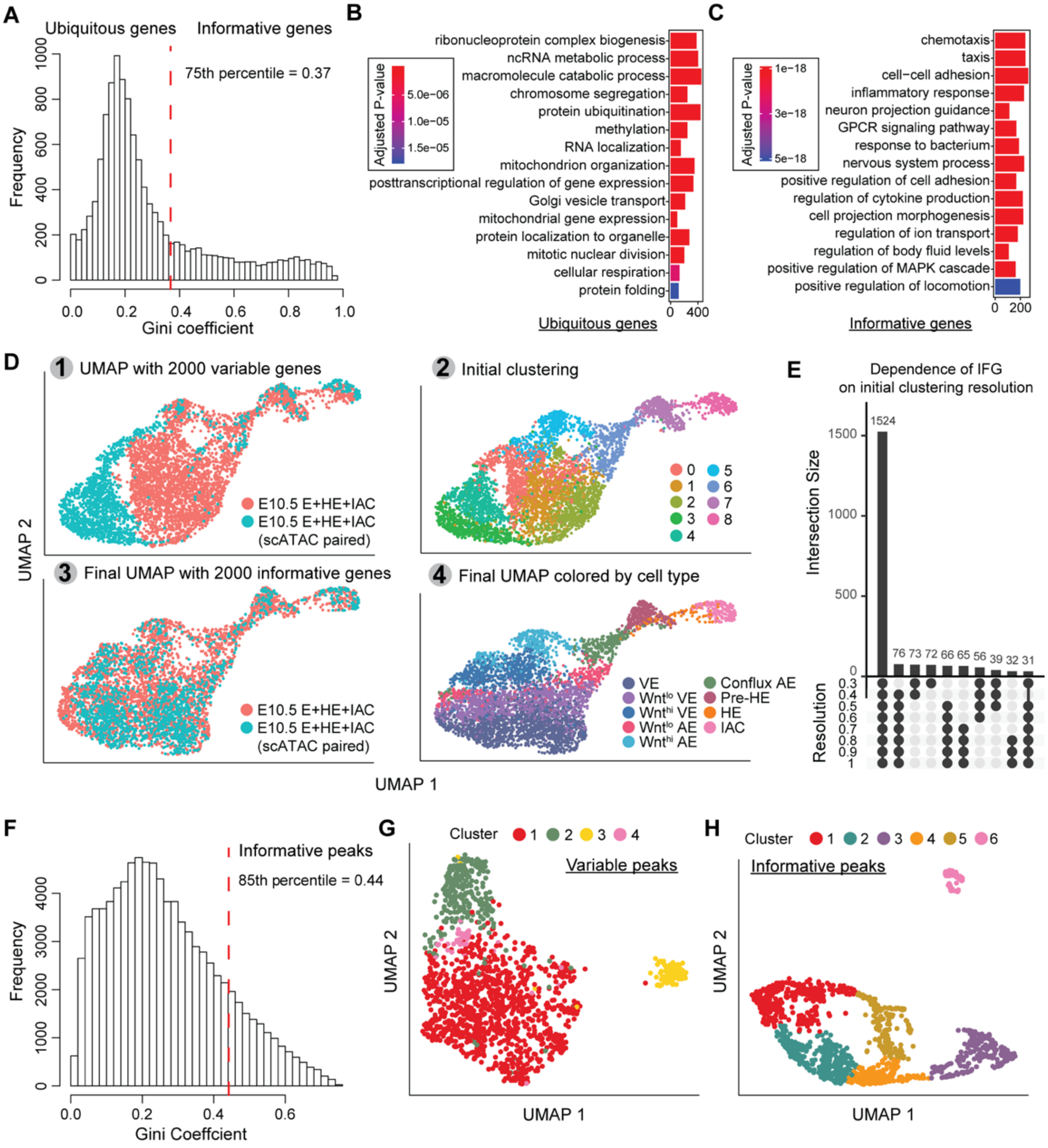
(A) Quantile contour FACS plots for purification of E10.5 IAC cells (Ter119$^-$CD41$^{med/-}$

$CD31^+Kit^+CD144^+ESAM^+$). **(B)** Quantile contour FACS plots for E11.5 IAC cells. **(C)** Quantile contour FACS plots for purification of E10.5 $CD45^+$ IAC cells containing lympho-myeloid-biased progenitors ($Ter119^-CD41^{med/-}CD31^+Kit^+CD144^+ESAM^+CD45^+$), and E10.5 $CD45^-$ IAC cells ($Ter119^-CD41^{med/-}CD31^+Kit^+CD144^+ESAM^+CD45^-$). **(D)** Quantile contour FACS plots from purification of E11.5 $CD144^+CD45^+CD27^+$ IAC cells containing type II pre-HSCs (E11.5 pre-HSC) **(E)** Quantile contour FACS plots from purification of E9.5 yolk sac EMPs ($CD16/32^+Kit^+CD41^{hi}$). **(F)** Quantile contour FACS plots from purification of E14.5 FL-HSCs [$lineage^-$ ($Ter119^-$ $Gr1^-$ $B220^-$ $CD3e^-$) $Kit^+Sca1^+CD48^-$ $CD150^+$]. **(G)** Quantile contour FACS plots from purification of E10.5 $CD44^+$ E+HE+IAC cells ($Ter119^-CD41^{med/-}CD31^+CD144^+ESAM^+CD44^+$). **(H)** Quantile contour FACS plots from purification of yolk sac-derived lymphoid-primed multipotent progenitors from the E11.5 FL (E11.5 FL-LMPP), which are included in the UMAP plots in Supplemental Figure 3.3A, B and Supplemental Figure 3.12C.



Supplemental Figure 3.3 Assignment of cell types, computational filtering of contaminant cells.

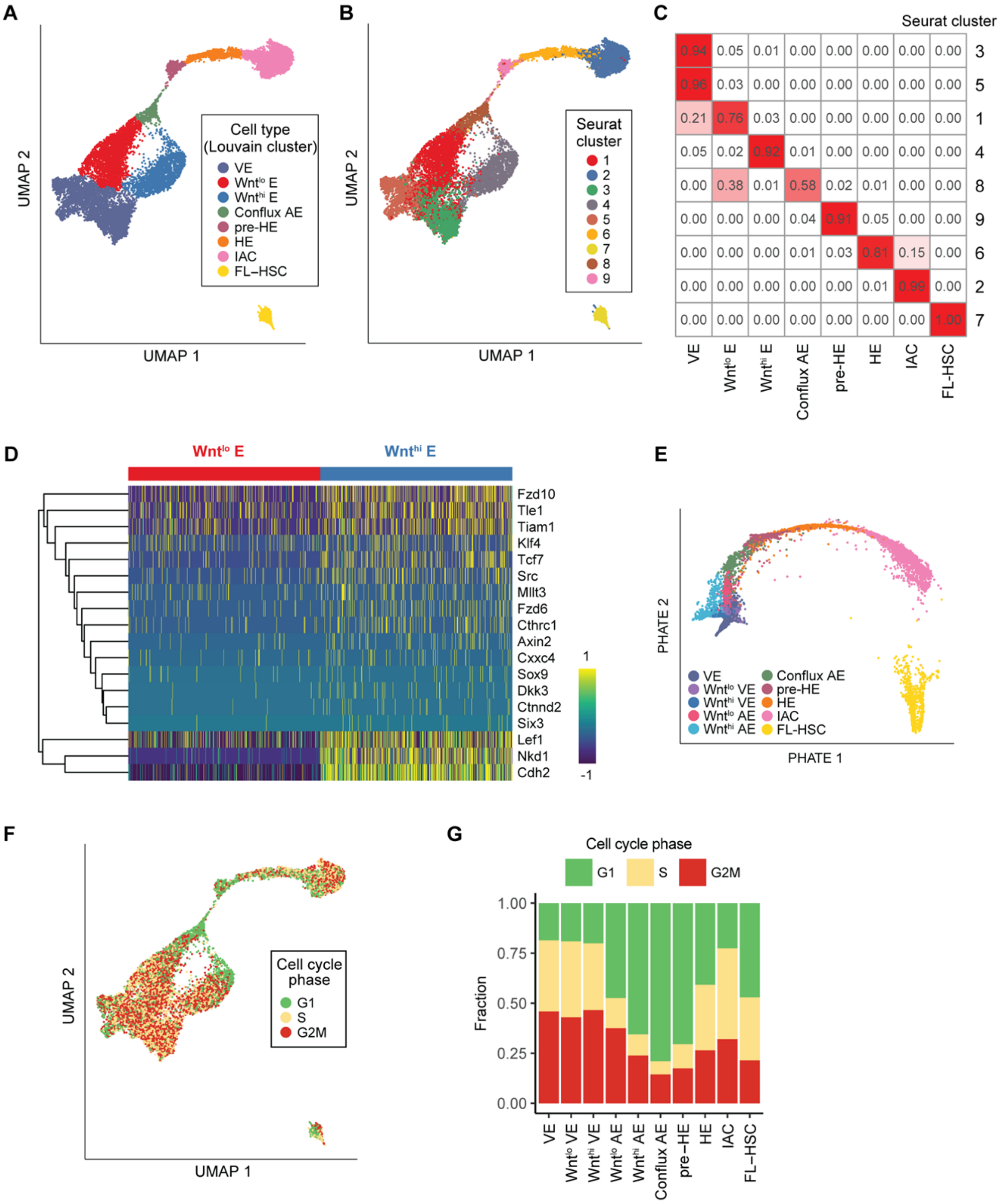
(A) Top: UMAP with all datasets. Labeled populations are contaminant cell types identified using Louvain clustering and differentially expressed genes (examples shown below). Populations circled with dashed line were used as input cells for a UMAP recomputed in panel B. Bottom: expression of representative top differentially expressed genes, including *Col1a2* (mesenchymal-fibroblast cells), *Hba-a1* (erythroid cells/progenitors), *Cldn5* (endothelial cells), *Lyve1* (combined with *Cldn5*, marks lymphatic endothelial or yolk sac endothelial cells). Clusters proximal to the mesenchymal-fibroblast cluster have distinct gene expression patterns, but are mostly non-endothelial or non-hematopoietic. **(B)** Top panel, UMAP with contaminant cells removed. Example DEGs shown below. E9.5 YS-EMP and E11.4 FL-LMPP are committed progenitors unrelated to pre-HSCs. Populations circled with a dashed line represent the continuous EHT trajectory that leads to E10.5 CD45$^+$ IAC cells and E11.5 pre-HSCs. Bottom panel, expression of representative top differentially expressed genes. *Bnip3* marks endothelial cells likely undergoing apoptosis, autophagy, or mitophagy (*289*); *Il7r* marks E11.5 FL-LMPP and a subset of IACs, and is low in E9.5 YS-EMPs; *Hp* (haptoglobin) is expressed in E11.5 FL-LMPP and E9.5 YS-EMP, at low levels in IAC cells; *Gata1* is expressed in the *Hp*$^{-}$ subset of E9.5 YS-EMPs, a subset of IAC cells, and at low levels in E11.5 FL-LMPPs. **(C)** Top panel, UMAP of continuous EHT trajectory and FL-HSCs, same as Figure 3.2A. Labeled clusters are cell types determined based on differentially expressed genes (examples shown below). Bottom panel, expression of representative top differentially expressed genes, including *Foxq1* in Wnt^{hi} E, *Hey2* in conflux AE and pre-HE, *Runx1* in HE and IACs, and *Ptprc* in IACs.



Supplemental Figure 3.4 Informative feature selection.

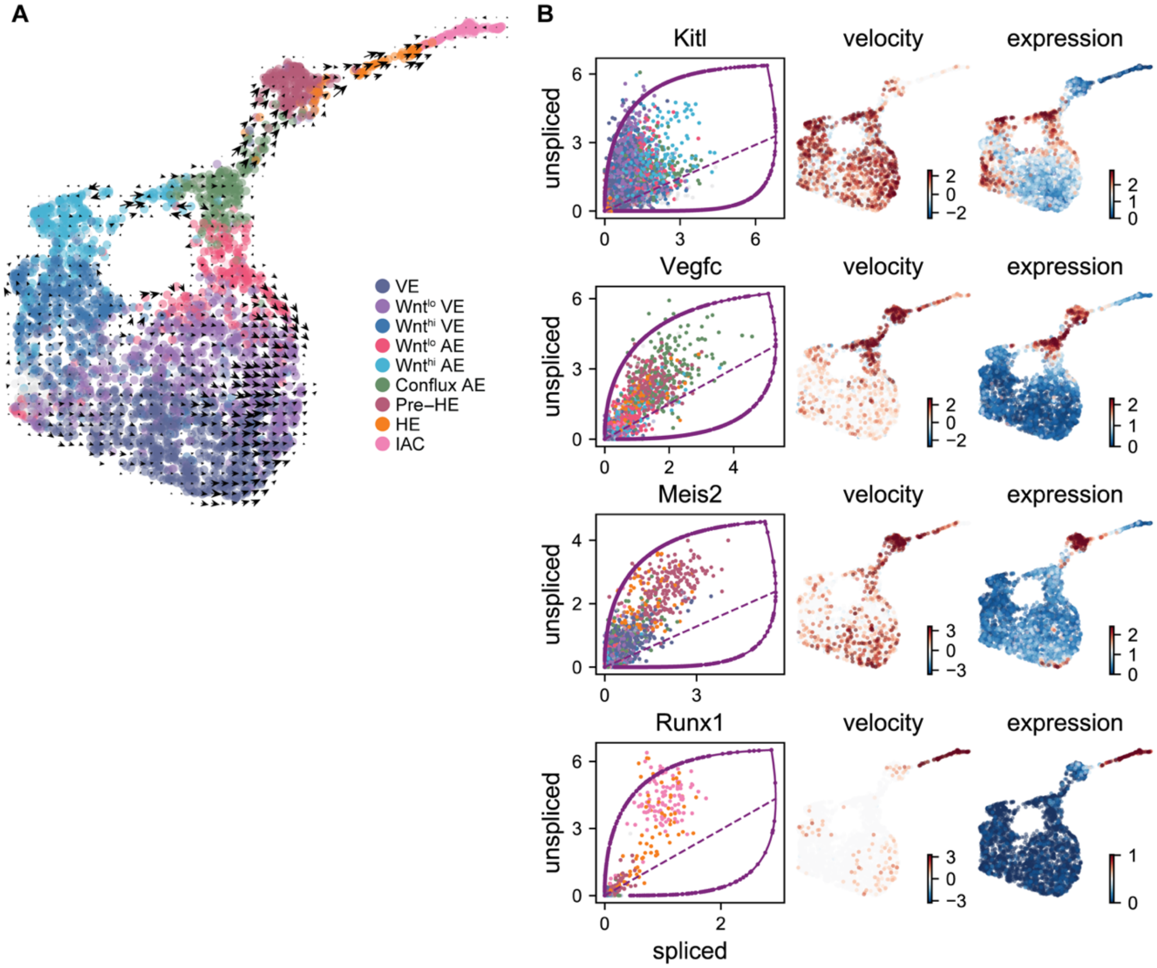
(A) Distribution of Gini coefficient for each expressed gene calculated with cells from Supplemental Figure 3.3A, using the approach described in Materials and Methods. Genes in the left peak have low Gini coefficients and ubiquitous expression patterns.

Genes in the right tail show unequal expression across clusters, and were chosen as IFFs for downstream dimensionality reduction analysis. **(B)** GO enrichment results for ubiquitous genes in panel A. Redundant terms were removed and top 15 enriched GO terms were plotted using clusterProfiler (*281*). Bar height represents gene counts for each enriched term. **(C)** GO enrichment results for IFFs in panel A. **(D)** IFF selection applied to two scRNA-Seq batches, E10.5 E+HE+IAC and E10.5 E+HE+IAC for paired scATAC, sampled under similar biological conditions. Panel 1 shows UMAP with top 2000 VEG from Seurat, 2 shows Seurat clusters, 3 shows UMAP with IFF selected based on Seurat clustering, 4 shows the UMAP colored by cell type. **(E)** Intersection of top 2000 IFFs selected across different Seurat clustering results with varied resolution. Intersections with less than 30 genes are not plotted. Seurat resolution parameters under which overlap is observed are indicated by a series of dots in a column below the x-axis. **(F)** Gini coefficient distribution of open chromatin peaks when applying IFF method to scATAC-Seq data. Peaks with Gini coefficient greater than the 85th percentile were selected as IFFs for analysis in panel F. **(G)** UMAP computed with EHT cells using variable features defined by Seurat (*279*). **(H)** UMAP computed using IFFs derived based on Gini coefficient distribution in panel F. Louvain clustering on the UMAP revealed finer endothelial cell subtypes compared to result in panel E.



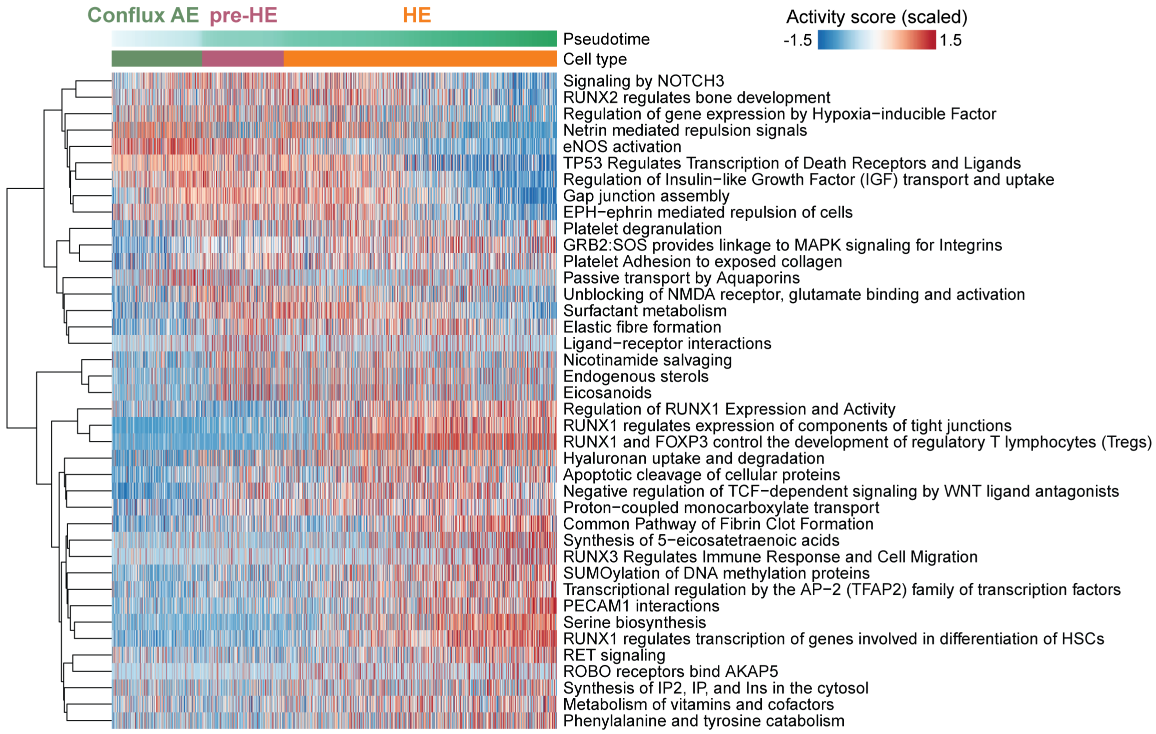
Supplemental Figure 3.5 Unsupervised clustering and dimension reduction on EHT cells.

(A) UMAP of continuous EHT trajectory and FL-HSCs colored by Louvain clustering result. **(B)** UMAP, same as in A, colored by Seurat clustering result. **(C)** Heatmap showing overlap of cells between Seurat clusters and Louvain clusters (number of cells in each Seurat cluster divided by number of cells in corresponding Louvain cluster). **(D)** Heatmap of expression of Wnt pathway genes (based on GO annotation) that are differentially expressed between Wnt^{lo} E and Wnt^{hi} E. **(E)** PHATE low dimensional embedding of EHT trajectory and FL-HSCs colored by cell type. **(F)** UMAP, same as in A, colored by Seurat predicted cell cycle phases. **(G)** Bar plot showing cell cycle phase composition for each cell type.



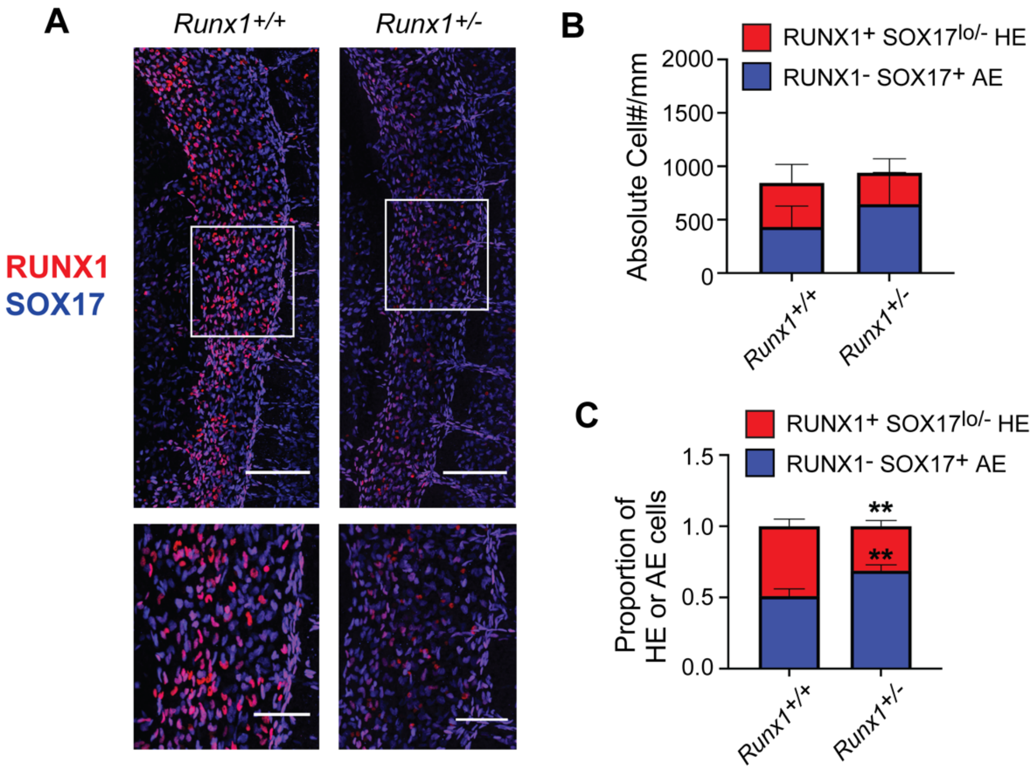
Supplemental Figure 3.6 RNA velocity estimate with scVelo dynamic model.

(A) scVelo estimated velocity embedded on the same UMAP as in Figure 3.3C. **(B)** Velocity estimation for genes induced at different stages of EHT, including *Kitl* in E, *Vegfc* in conflux AE, *Meis2* in pre-HE and *Runx1* in HE and subset of E cells. For each gene, the phase plot shows unspliced versus spliced transcript counts and the scVelo fitted model. The velocity and expression are plotted on the same UMAP in panel A.



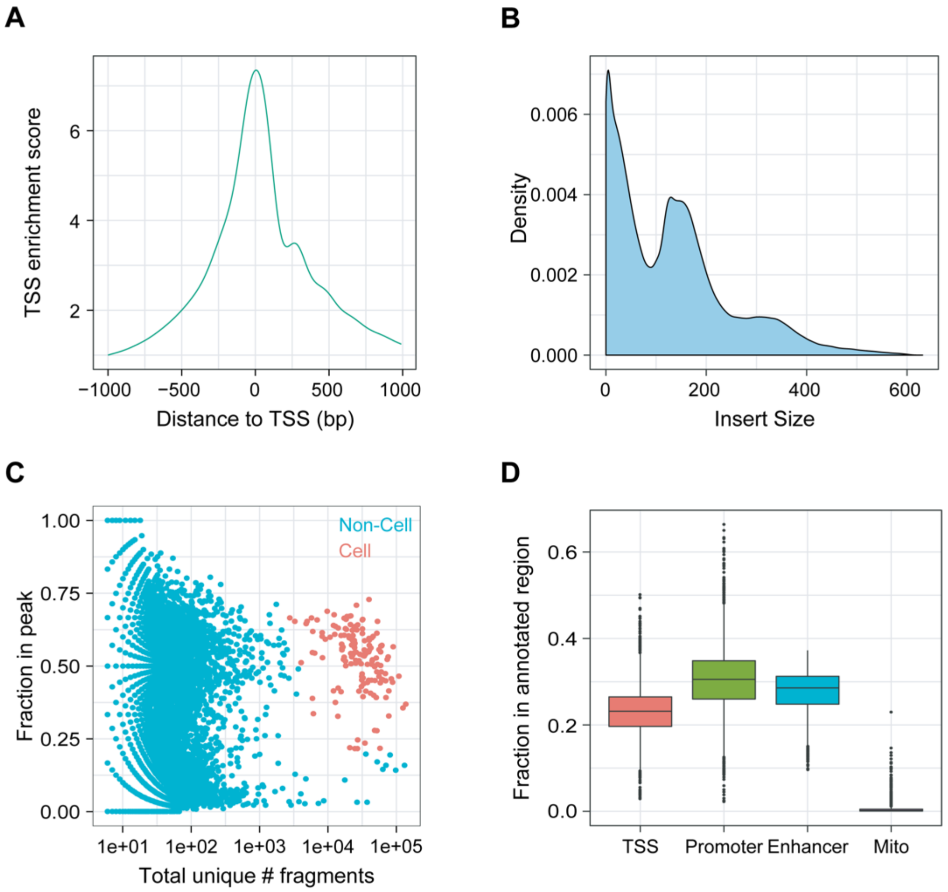
Supplemental Figure 3.7 Reactome pathway analysis for the transitions between conflux AE, pre-HE, and HE.

Top stage-wise differentially activated pathways were plotted. Rows were clustered using hierarchical clustering with Ward’s method.



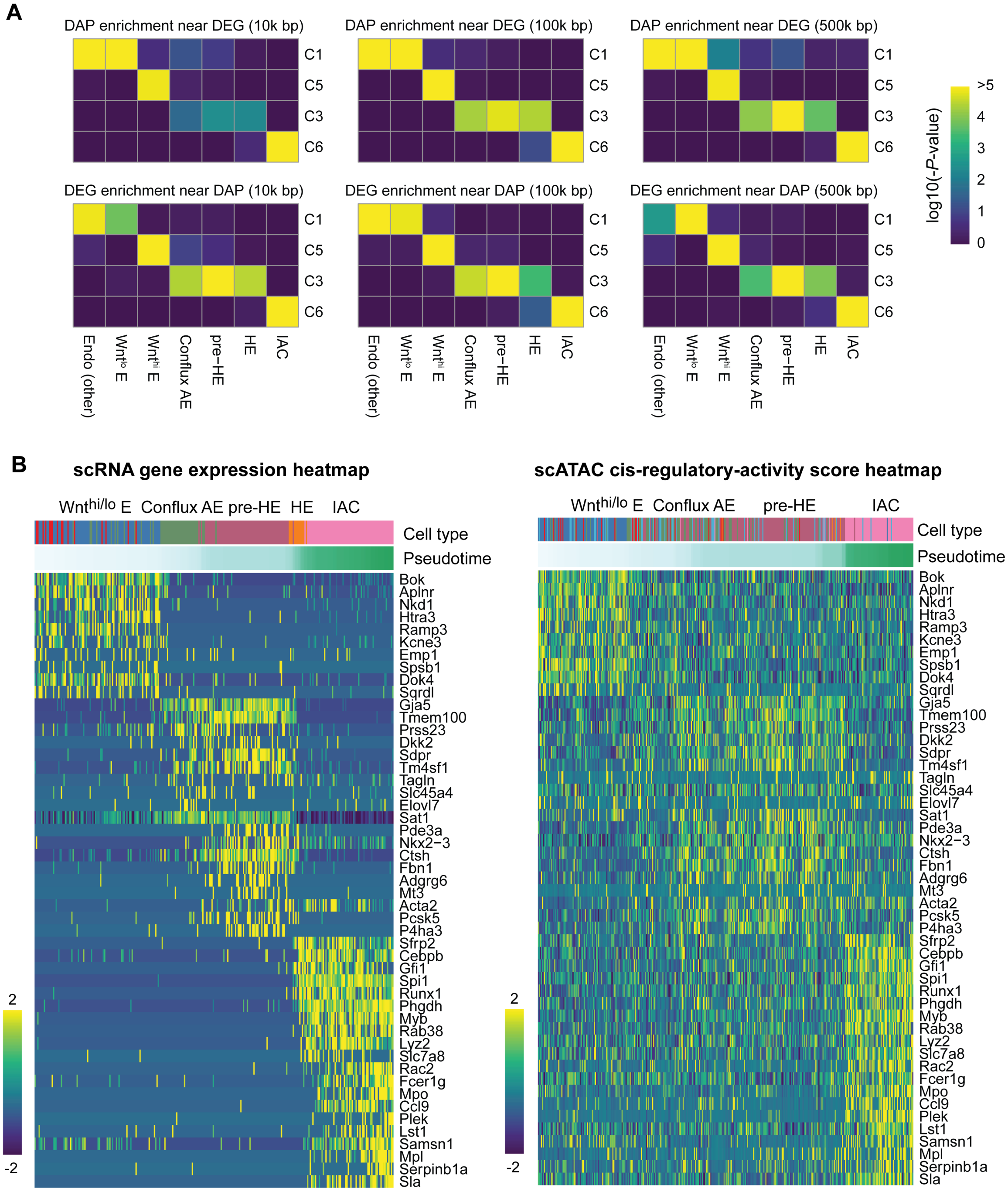
Supplemental Figure 3.8 RUNX1 haploinsufficiency reduces phenotypic hemogenic endothelial cells in the dorsal aorta.

(A) Confocal z-projection stacks of dorsal aortas with z-intervals of 2 μM of E9.5 embryos immunostained with antibodies against RUNX1 and SOX17. Scale bar: 100μM. Enlargement of outlined area shown below. Scale bar: 50μM. *Runx1*$^{+/+}$ n=6, *Runx1*$^{+/-}$ n=3 (23-27 somite pairs). Representative images are shown for each genotype. **(B)** Quantification of absolute number of AE cells (RUNX1^{+}SOX17^{-}) and HE cells (RUNX1^{+}SOX17$^{low/-}$) per mm of dorsal aorta. Two-way ANOVA and Tukey's test. **(C)** Proportion of HE and AE cells [(# of HE or AE)/(total # of AE+HE)] in E9.5 embryos. Two-way ANOVA and Tukey's test. **p<0.01.



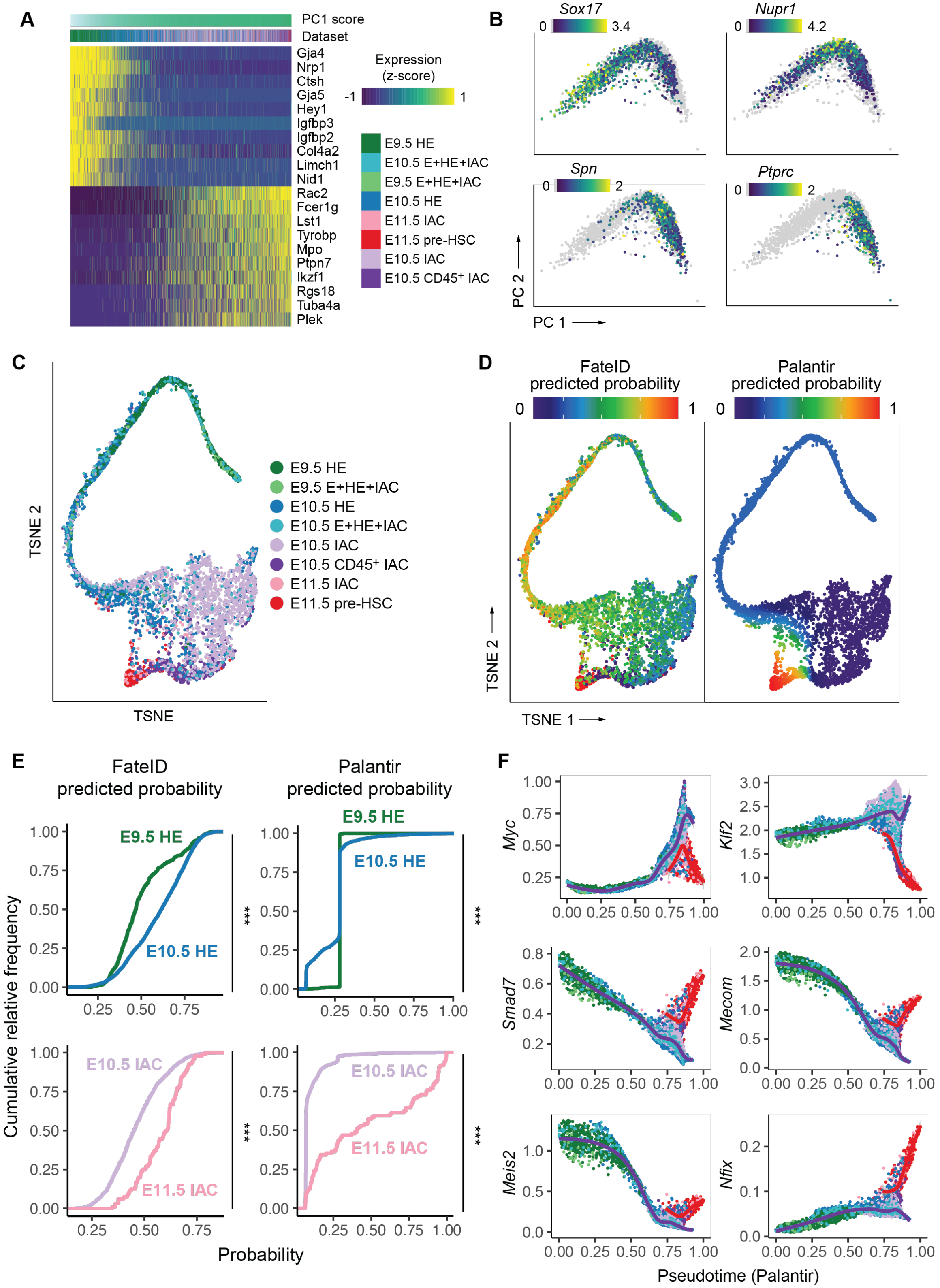
Supplemental Figure 3.9 Quality assessment of scATAC-Seq data.

Quality assessment was performed using the scATAC-pro software. **(A)** Enrichment profile of scATAC-seq reads around the transcription start site (TSS). Scores were calculated based on the aggregate read distribution centered on the TSSs, extending 1000 bp in both directions. **(B)** Distribution of ATAC-Seq insert size for all unique fragments. **(C)** Scatter plot (with points down-sampled) of the fraction of unique fragments in peaks versus total number of unique fragments per cell barcode, discriminating cells from non-cells. **(D)** Box plots of the fraction of unique fragments overlapping with annotated genomic regions from Ensemble regulatory build release 95. Mito, mitochondrial genome.



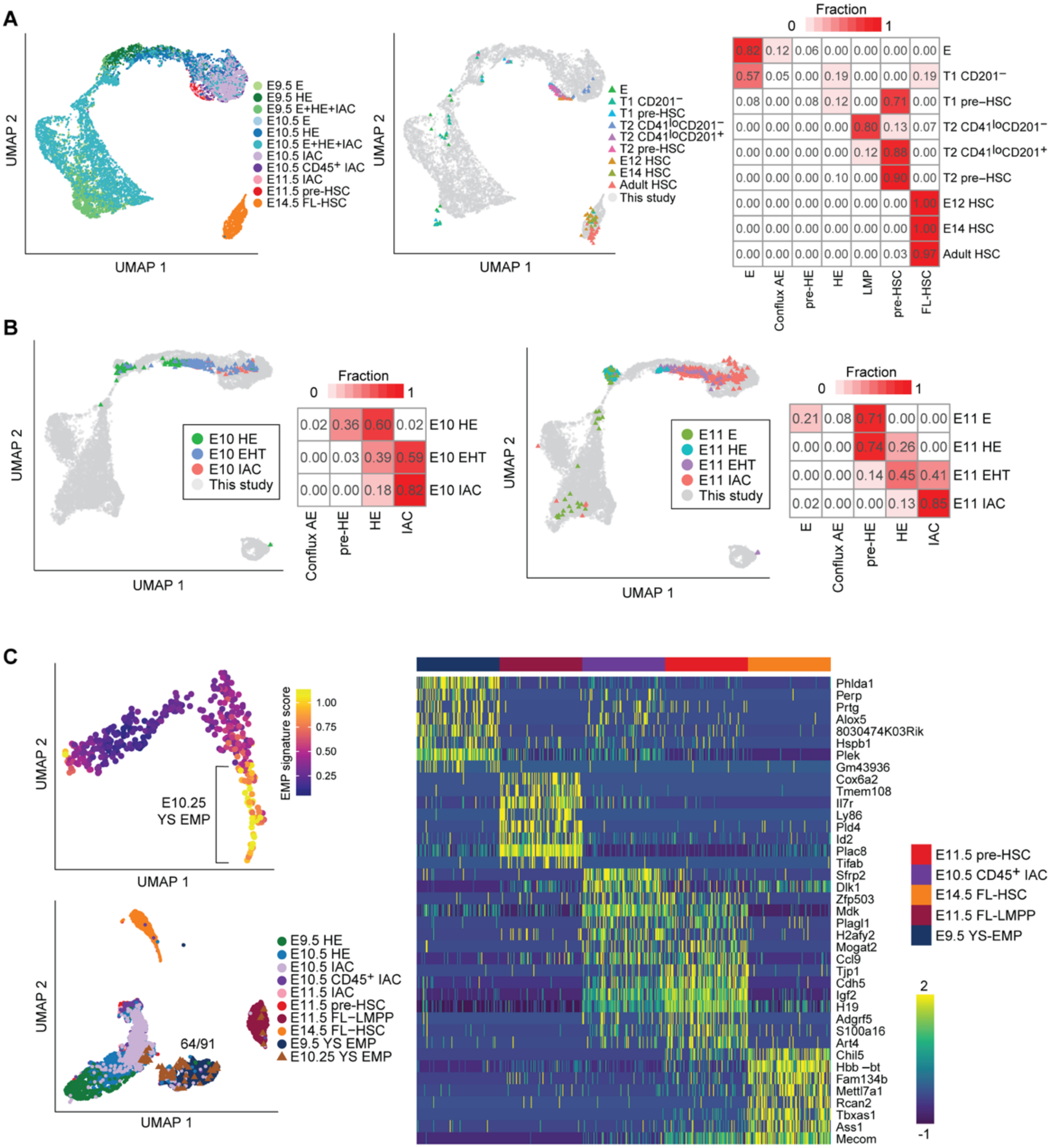
Supplemental Figure 3.10 Mapping between scRNA and scATAC data.

(A) Signature matching of scATAC-Seq and scRNA-Seq data. Differentially accessible peaks (DAPs) and differentially expressed genes (DEGs) were separately derived, and hypergeometric test was performed on DAPs and DEGs within predefined genomic windows around DEG/DAP (10k bp, 100k bp and 500k bp) for each pair of scATAC-Seq (labeled as C1-6) and scRNA-Seq clusters (labeled by cell type names). We did not obtain enough DAPs for C2 and C4 for this matching. **(B)** Heatmap of stage-wise differentially expressed genes and their corresponding cis-regulatory-activity score. Values are log2 transformed and scaled across cells.


A
PC1 score
Dataset
Gja4
Nrp1
Ctsh
Gja5
Hey1
Igfbp3
Igfbp2
Col4a2
Limch1
Nid1
Rac2
Fcer1g
Lst1
Tyrobp
Mpo
Ptpn7
Ikzf1
Rgs18
Tuba4a
Plek
Expression
(z-score)
-1
1
E9.5 HE
E10.5 E+HE+IAC
E9.5 E+HE+IAC
E10.5 HE
E11.5 IAC
E11.5 pre-HSC
E10.5 IAC
E10.5 CD45+ IAC
B
Sox17
0
3.4
Nupr1
0
4.2
Spn
0
2
Ptprc
0
2
PC 2
PC 1
C
TSNE 2
TSNE
E9.5 HE
E9.5 E+HE+IAC
E10.5 HE
E10.5 E+HE+IAC
E10.5 IAC
E10.5 CD45+ IAC
E11.5 IAC
E11.5 pre-HSC
D
FateID
predicted probability
0
1
Palantir
predicted probability
0
1
TSNE 2
TSNE 1
E
FateID
predicted probability
Palantir
predicted probability
Cumulative relative frequency
1.00
0.75
0.50
0.25
0.00
E9.5 HE
E10.5 HE
0.25
0.50
0.75
***
E9.5 HE
E10.5 HE
0.25
0.50
0.75
1.00
***
E10.5 IAC
E11.5 IAC
***
E10.5 IAC
E11.5 IAC
***
Probability
F
Myc
Klf2
Smad7
Mecom
Meis2
Nfix
1.00
0.75
0.50
0.25
3.0
2.5
2.0
1.5
1.0
0.8
0.6
0.4
0.2
2.0
1.5
1.0
0.5
0.0
1.0
0.5
0.0
0.2
0.1
0.0
0.00
0.25
0.50
0.75
1.00
Pseudotime (Palantir)

Supplemental Figure 3.11 Transition from HE to IAC and fate probability analysis.

(A) Differentially expressed genes in HE to IAC cell transition. Expression heatmap is plotted for top genes ranked by PC1 loading (PCA shown in Figure 3.6A). Cells are ordered based on PC1 score. Expression values are log2 transformed and z-score scaled. **(B)** Expression pattern of *Sox17*, *Nupr1*, *Spn* (*CD43*) and *Ptprc* (*CD45*) on the PC1 vs PC2 PCA plot. Values are log2 transformed normalized UMI counts. **(C)** T-SNE plot generated with the Palantir package (*260*), colored by cell population. **(D)** Fate probability of becoming pre-HSC versus lympho-myeloid-biased progenitors predicted by FateID and Palantir for each cell in the T-SNE plot. (**E**) Cumulative distribution of pre-HSC fate probability for each cell population. One-sided Kolmogorov–Smirnov test was carried out between E9.5 HE and E10.5 HE, E10.5 IAC and E11.5 IAC for probabilities computed with both methods. ***p<0.001. **(F)** Transcription factors expressed in HE that show bifurcation in expression along the trajectory. Expression is imputed with MAGIC (*290*). Gene expression trends are modeled with a generalized additive model, using cells with probability greater than 0.5 for each fate branch.



Supplemental Figure 3.12 Comparison with previously published scRNA-Seq data.

(A) Left panel shows UMAP of EHT trajectory computed with genes shared with Zhou *et al*. (*39*), with the same cells in Figure 3.1C. Middle panel shows projection of Zhou *et al*. (*39*) data to the same UMAP (see Methods). Right panel shows heatmap for the fraction of each cell population from Zhou *et al*. mapped to each of the cell types from this study.

(B) Projection of E10 (left panel) and E11 (right panel) cells from Baron *et al.* (*226*) to those from this study. The UMAP is slightly different from those in panel A, due to differences in shared genes, but each cell type has an almost identical relative position. Heatmap shows the fraction of each cell population from Baron *et al.* mapped to each of the cell types from this study. **(C)** Top panel: UMAP of EMP, macrophage precursors and macrophage single cell data highlighting the EMP signature score (average expression level of signature genes listed in Supplemental Table 2 from Mass *et al.* (*282*)). Bottom panel: UMAP of indicated populations from this study (circles). E9.5 EMPs form a separate cluster from IAC cells and other cell types. Cells with a high EMP signature score from Mass *et al.* (*282*) (brown triangles) were projected onto the same UMAP (see Materials and Methods). 64 out of 91 E10.25 EMP cells projected onto the E9.5 YS-EMP cluster, and none overlapped with E10.5 CD45$^+$ IAC cells. Yolk sac-derived lymphoid-primed multipotent progenitors purified from the E11.5 fetal liver (E11.5 FL-LMPP) are also separated from E10.5 CD45$^+$ IAC cells. Heatmap shows top differentially expressed genes of each cell population.

CHAPTER 4 SOFTWARE FOR SINGLE CELL TRANSCRIPTOMICS DATA ANALYSIS

The contents of this chapter have been previously published in:

Zhu Q, Fisher SA, Dueck H, Middleton S, Khaladkar M, Kim J. PIVOT: platform for interactive analysis and visualization of transcriptomics data. *BMC bioinformatics*. 2018 Dec;19(1):1-8.

Packer JS*, **Zhu Q***, Huynh C, Sivaramakrishnan P, Preston E, Dueck H, Stefanik D, Tan K, Trapnell C, Kim J, Waterston RH, Murray JI. A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution. *Science*. 2019 Sep 20;365(6459):eaax1971.

* These authors contributed equally.

Introduction

Technologies such as single cell RNA-sequencing measure gene expressions and present them as high-dimensional expression matrixes for downstream analyses. In recent years, many methods have been developed for the statistical analysis of transcriptomics data, such as edgeR (*291*) and DESeq (*292*) for differential expression testing, and monocle (*131*), Seurat (*293*), SC3 (*123*) and SCDE (*294*) for single cell RNA-Seq data analysis. Besides these, the Comprehensive R Archive Network (CRAN) (*295*) and Bioconductor (*296*) host various statistical packages addressing different aspects of transcriptomics study and provides recipes for a multitude of analysis workflows. Making use of these R analysis packages requires expertise in R and often custom scripts to integrate the results of different packages. In addition, many exploratory analyses of transcriptomics data involve repeated data manipulations such as normalizations, filtering, merging, etc., each step generating a derived dataset whose version and provenance must be tracked. Previous efforts to address these problems include designing standardized workflows (*297*), building a comprehensive package (*293*) or assembling pipelines into integrative platforms such as Galaxy (*298*). Designing workflows or using large packages still requires a significant amount of programming skills and it can be difficult to make various components compatible or applicable to specific datasets. Integrative platforms offer greater usability but trade off flexibility, functionality and efficiency due to limitations on data size, parameter choice and computing power. For example, the Galaxy platform is designed as discrete functional modules which require separate file inputs for different analysis. This design not only makes user-end file format conversion complicated and time-consuming, but also breaks the integrity of the analysis workflow, limiting the sharing of global parameters, filtering criteria and analysis results between modules. Tools such as RNASeqGUI (*299*), START (*300*), ASAP (*301*) and DEApp (*302*) provide an interactive graphical interface for a small number of packages. But these and other similar packages all adopt a rigid workflow design, have limited data provenance tracking, and none of the packages provide mechanisms for tracking, saving and sharing analysis results. Furthermore, many web-based applications require users to upload data to a server, which might be prohibited by HIPPA (Health Insurance Portability and Accountability Act of 1996) for clinical data analysis.

Here we developed PIVOT, an R-based platform for exploratory transcriptomics data analysis. We leverage the Shiny framework (*303*) to bridge open source R packages and JavaScript-based web applications, and to design a user-friendly graphical interface that is consistent across statistical packages. The Shiny framework translates user-driven events (e.g. pressing buttons) into R interpretable reactive data objects, and present results as dynamic web content. PIVOT incorporates four key features that assists user interactions, integrative analysis and provenance management:

- PIVOT directly integrates existing open-source packages by wrapping the packages with a uniform user-interface and visual output displays. The user interface replaces command line options of many packages with menus, sliders, and other option controls, while the visual outputs provide extra interactive features such as change of view, active objects, and other user selectable tools.
- PIVOT provides many tools to manipulate a dataset to derive new datasets including different ways to normalize a dataset, subset a dataset, etc. In particular, PIVOT supports manipulating the datasets using the results of an

analysis; for example, a user might use the results of differential gene expression analysis to select all gene satisfying some p-value filter. PIVOT implements a visual data management system, which allows users to create multiple data views and graphically display the linked relationship between data variants, allowing navigation through derived data objects and automated re-analysis.

- PIVOT dynamically bridges analysis packages to allow results from one package to be used as inputs to another. Thus, it provides a flexible framework for users to combine tools into customizable pipelines for various analysis purposes.
- PIVOT provides facilities to automatically generate reports, publication-quality figures, and reproducible computations. All analyses and data generated in an interactive session can be packaged as a single R object that can be shared to exactly reproduce any results.

Recently, we extended PIVOT to a new tool, VisCello, which supports analysis and hosting of large-scale single cell data. VisCello preserves key functions of PIVOT, incorporates several latest single cell analysis packages, and have multiple function and speed optimizations.

Results and Discussion

We describe the general workflow of PIVOT and demonstrate its versatile and practical use in the following sections. We also briefly describe VisCello, an extension of PIVOT which facilitates large-scale single cell data analysis.

Data input and transformations

Read counts obtained from RNA-Seq quantification tools such as HTSeq (*304*) or featureCounts (*305*) can be directly uploaded into PIVOT as text, csv or Excel files. Data generated using the 10x Genomics Cell Ranger pipeline can also be readily read in and processed by PIVOT. PIVOT automatically performs user selected data transformations including normalization, log transformation, or standardization. We have included multiple RNA-Seq data normalization methods including DESeq normalization (*306*), trimmed mean of M-values (TMM) (*307*), quantile normalization (*308*), RPKM/TPM (*309*), Census normalization (*181*), and Remove Unwanted Variation (RUVg) (*310*) (Table 4.1). If samples contain spike-in control mixes such as ERCC (*311*), PIVOT will also separately analyze the ERCC count distribution and allow users to normalize the data using the ERCC control. Existing methods can be customized by the user by setting detailed normalization parameters. For example, we implement a modification of the DESeq method by making the inclusion criterion a user set parameter, making it more applicable to sparse expression matrices such as single cell RNA-Seq data (*312*).

Users can upload experiment design information such as conditions and batches, which can be visualized as annotation attributes (e.g., color points/sidebars) or used as model specification variables for downstream analyses such as differential expression. PIVOT supports flexible operations to filter data for row and column subsets as well as for merging datasets, creating new derived datasets. Multiple summary statistics and quality control plots are automatically generated to help users identify possible outliers. Users can manually select samples for analysis, or specify statistical criteria on analysis results such as expression threshold, dropout rate cutoff, Cook’s distance or size factor range to remove unwanted features and samples.

Visual data management with data map

When analyzing large datasets, a common procedure is to first perform quality control to remove low quality elements, then normalize the data and finally generate different data subsets for various analysis purposes. Some analyses require filtering out genes with low expressions, while others are designed to be performed on a subset of the genes such as transcription factors. During secondary analyses, outliers may be detected requiring additional scrutiny. All these data manipulations generate a network of derived datasets from the original data and require a significant amount of effort to track. Failure to track the data lineage could affect the reproducibility and reliability of the study. Furthermore, an investigator might wish to repeat an analysis over a variety of derived datasets, which may be tedious and error-prone to carry out manually. To address this problem, we implemented a graphical data management system in PIVOT.

As the user generates derived datasets with various data manipulations, PIVOT records and presents the data provenance in an interactive tree graph, the "Data Map". As shown in Figure 4.1, each node in the data map represents a derived dataset and the edges contain information about the details of the derivation operation. Users can attach analysis results to the data nodes as interactive R markdown reports (*313*) and switch between different datasets or retrieve analysis reports by simply clicking the nodes. Upon switch to a new dataset selected from the Data Map, PIVOT automatically re-runs analyses and updates parameter choices when needed. Thus, a user can easily compare results of a workflow across derived datasets. The data map is generated with the visNetwork package (*314*) and can be directly edited, so that users can rename nodes, add notes, or delete data subsets and analysis reports that are no longer useful.

The full data history is also presented as downloadable tables with all sample and feature information as well as data manipulation details.

Comprehensive toolset for exploratory analysis

PIVOT is designed to aid exploratory analysis for both single cell and bulk RNA-Seq data, thus we have incorporated a large set of commonly used tools (see Table 4.1). PIVOT supports many visual data analytics including QC plots (number of detected genes, total read counts, dropout rates and estimated size factors; Figure 4.2A, data from (*100*)), transcriptome statistics plots (e.g., rank-frequency plots, mean-variability plots, etc; Figure 4.2B), and sample and feature correlation plots (e.g., heatmaps, smoothened scatter plots, etc.). All visual plots feature interactive options and a query function is provided which allows users to search for features sharing similar expression patterns with a target feature. PIVOT provides users extensive control over parameter choices. Each analysis module contains multiple visual controls allowing users to adjust parameters and obtain updated results on the fly.

Integrative analysis and interactive visualization

PIVOT transparently bridges multiple sequences of analyses to form customizable analysis pipelines. For example, with single cell data collected from heterogeneous tissues, a user can first perform PCA or t-SNE (*115*) (Figure 4.2C) to visualize the low dimensional embedding of the data. If there is clear clustering pattern, possibly originated from different cell types, the user can directly specify cell clusters by dragging selection boxes on the graph, or perform K-means or hierarchical clustering with the projection matrix. One can proceed to run DE or penalized LDA (*315*) to identify cluster-specific marker genes, which can then be used to filter the datasets for generating a heatmap showing distinctive expression pattern across cell types (Figure 4.2D). Within each determined cell type, a user may further apply the walk-trap community detection method (*316*) to identify densely connected network of cells, which are indicative of potential subpopulations.

As another example, for time-series data such as cells collected at different stages of development or differentiation, one can use diffusion pseudotime (DPT) (*317*), which reconstructs the lineage branching pattern based on the diffusion map algorithm (*136*), or Monocle (*131*), which implements an unsupervised algorithm for pseudo-temporal ordering of single cells (*130*). We have incorporated the latest Monocle 2 workflow in PIVOT, including cell state ordering, unsupervised cell clustering, gene clustering by pseudo-temporal expression pattern and cell trajectory analysis. Besides the DE method implemented in monocle, one can also run DESeq, edgeR, SCDE or the Mann-Whitney U test. A user can specify whether to perform basic DE analysis or a multi-factorial DE analysis with customized formula for complex experimental designs such as time-series or controlling for batch effects. Results are presented as dynamic tables including all essential statistics such as maximum likelihood estimation and confidence intervals. Each gene entry in the table can be clicked and visualized as violin plots or box plots, showing the actual expression level across conditions. Once DE results are obtained, the user can further explore the connections between DE genes and identify potential trans-differentiation factors as introduced in the Mogrify algorithm (*318*). PIVOT provides several extensions of functionality from the original Mogrify method. The network analysis module allows users to plot the log fold changes (LFC) of DE genes in a protein-protein interaction network obtained from the STRING database (Figure 4.3A) (*319*) or a directed regulatory network graph constructed from the Regnetwork repository (Figure 4.3B) (*320*). With scoring based on the p-value and log fold change, the graph can be filtered to only include top-rank genes, showing the regulatory "hot spot" of the network. PIVOT provides users with multiple options for defining the network influence score of transcription factors, and will produce lists of potential trans-differentiation factors based on the final ranking. As shown in Figure 4.3C, with the FANTOM5 expression data of fibroblasts and ES cells (*321*), PIVOT correctly reports OCT4 (POU5F1), NANOG and SOX2 as key factors for trans-differentiation (*322*). In addition to the DESeq results used by the original Mogrify algorithm, a user can choose to use SCDE or edgeR results to perform trans-differentiation analysis on single cell datasets.

Another useful feature of PIVOT is that it provides users multiple visualization options by exploiting the power of various plotting packages. For example, users can either generate publication-quality heatmap graphs (implemented in gplots package [43]), or interactively explore the heatmap with the heatmaply view (*323*). For principal component analysis, PIVOT uses three different packages to present the 2D and 3D projections. The plotly package (*324*) displays sample names and relevant information as mouse-over labels, while the ggbiplot (*325*) presents the loadings of each gene on the graph as vectors. The threejs package (*326*) fully utilizes the power of WebGL and outputs rotatable 3D projections. In the network analysis module, we utilize both igraph (*327*) and networkD3 (*328*) package to plot the transcription factor centered local network. The latter provides a force directed layout, which allows users to drag the nodes and visualize the physical simulation of the network response.

Reproducible research and complete provenance capture

PIVOT automatically records all data manipulations and analysis steps. Once an analysis has been performed, users will have the option of pasting related R markdown code to a shinyAce report editor (*329*), or download the report as either a pdf or interactive html document. All results and associated parameters will be captured and saved to the report along with user-provided comments. PIVOT states are automatically saved in cases of browser refresh, crash or user exit, and can also be manually exported, shared and loaded. Thus, all analyses performed in PIVOT are fully encapsulated and can be shared or disseminated as a single data+provenance object, allowing universally reproducible research.

VisCello: extending PIVOT for large-scale single cell data

We extended PIVOT to VisCello for distributing single cell analyses and providing interactive visualizations (Figure 4.4). VisCello can be installed as an R package (https://github.com/qinzhu/VisCello) or hosted as an interactive web app. Compared to PIVOT, VisCello had multiple optimizations, including the adoption of the sparse matrix format (*330*), which significantly improves the speed and reduces memory use. VisCello hosts dimensionality reductions (e.g. UMAPs), cell annotations, and marker gene tables for different subsets of the data. Users can visualize gene expression on UMAP or PCA plots, on a lineage tree diagram, or as box/violin plots grouped by cell type or lineage. The plots are interactive, allowing users to zoom in on subsets of cells, define new cell annotation groups, and run differential expression analysis and GO/KEGG enrichment with these newly defined groups. Program state can be downloaded and shared, facilitating collaboration. VisCello has been used to host and disseminate various single cell datasets (*170, 331, 332*).

Conclusions

We developed PIVOT and VisCello for easy, fast, and exploratory analysis of single cell transcriptomics data. Toward this goal we have automated the analysis procedures and data management, and we provide users with detailed explanations both in tooltips and user manuals. PIVOT and VisCello exploits the power of multiple plotting packages and gives users full control of key analysis and plotting parameters. Given user input that leads to function errors, PIVOT and VisCello will alert the user and provide corrective suggestions. Program states and reports can be shared between researchers to facilitate the discussion of expression analysis and future experimental design. Future versions of the software will continue to integrate popular transcriptome analysis routines as they are made available to the research community.

Figures

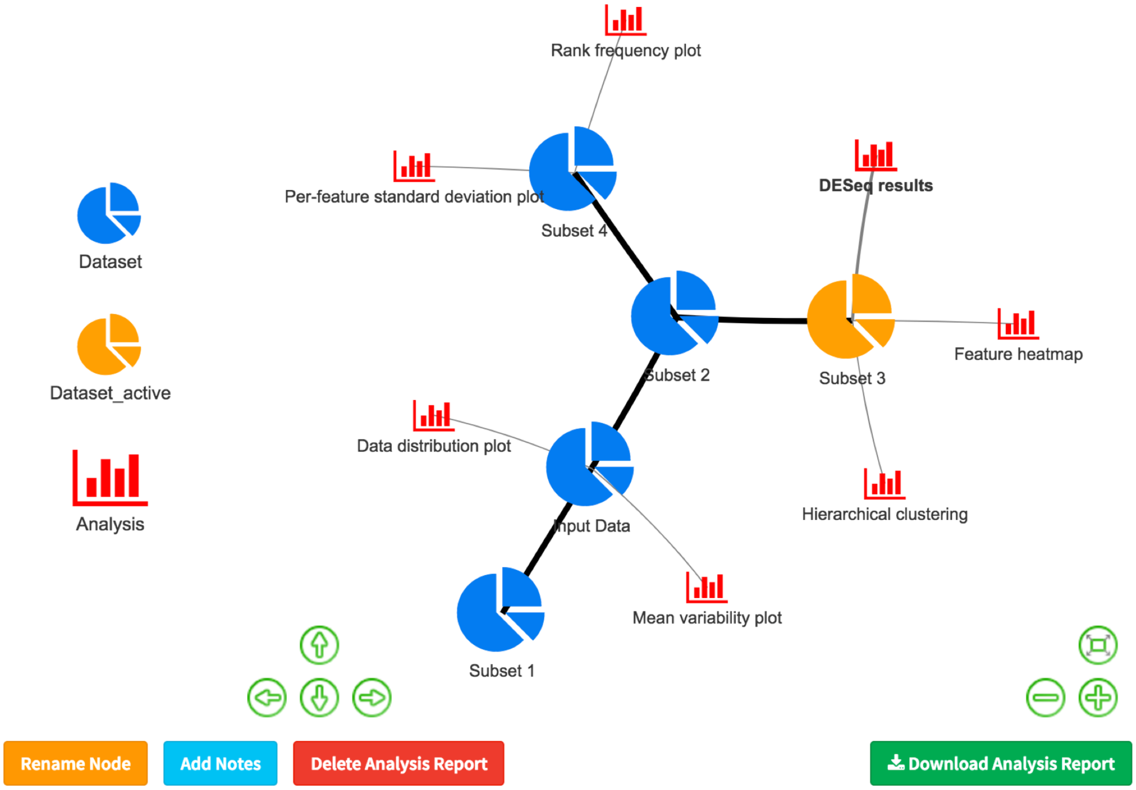
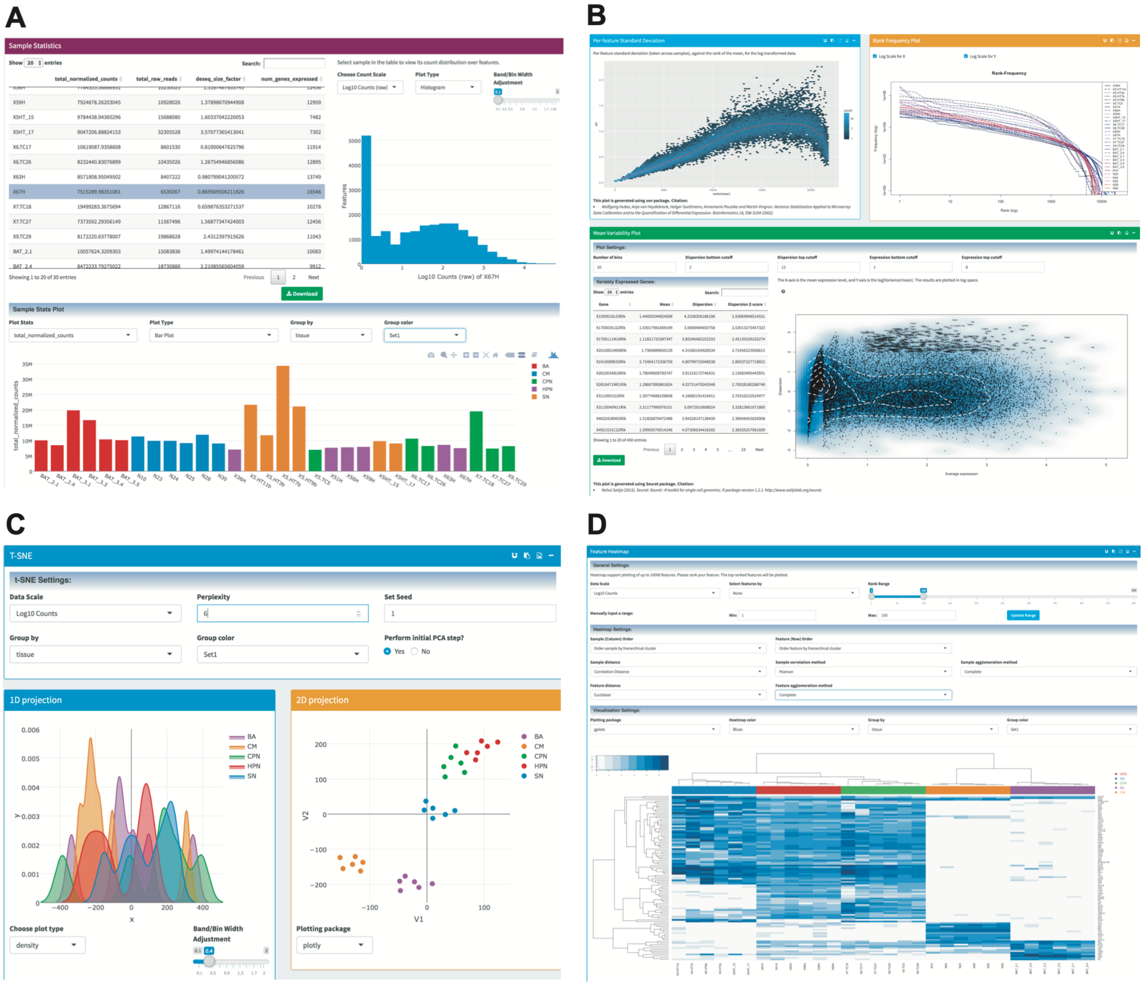


Figure 4.1 Data management with data map.

The map shows the history of the data change and the association between analysis and data nodes. Users can hover over edges to see operation details, or click nodes to get analysis reports or switch active subsets.



Sample Statistics

Show 20 entries Search:

	total_normalized_counts	total_raw_reads	deseq_size_factor	num_genes_expressed
X59H	7924678.26203045	10928026	1.37898670944908	12959
X5HT_15	9784438.94360296	15688080	1.60337042220053	7482
X5HT_17	9047206.88824153	32305528	3.57077365413041	7302
X6.TC17	10619087.9358608	8601530	0.81000647625796	11914
X6.TC26	8232440.83076899	10435026	1.26754946856086	12895
X63H	8571808.95049502	8407222	0.980799041200572	13749
X67H	7515289.98351061	6535067	0.869569506211826	16546
X7.TC18	19499283.3675694	12867116	0.659876353271537	10278
X7.TC27	7373592.29356149	11567496	1.56877347424003	12456
X9.TC29	8172220.63778007	19868628	2.4312397915626	11043
BAT_2.1	10057624.3209303	15083836	1.49974144178461	10083
BAT_2.4	8472233.79275022	18730886	2.21085565604059	9912

Showing 1 to 20 of 30 entries

Select sample in the table to view its count distribution over features.

Choose Count Scale: Log10 Counts (raw)

Plot Type: Histogram

Band/Bin Width Adjustment: 0.1

Sample Stats Plot

Plot Stats: total_normalized_counts

Plot Type: Bar Plot

Group by: tissue

Group color: Set1

T-SNE

t-SNE Settings:

Data Scale: Log10 Counts

Perplexity: 6

Set Seed: 1

Group by: tissue

Group color: Set1

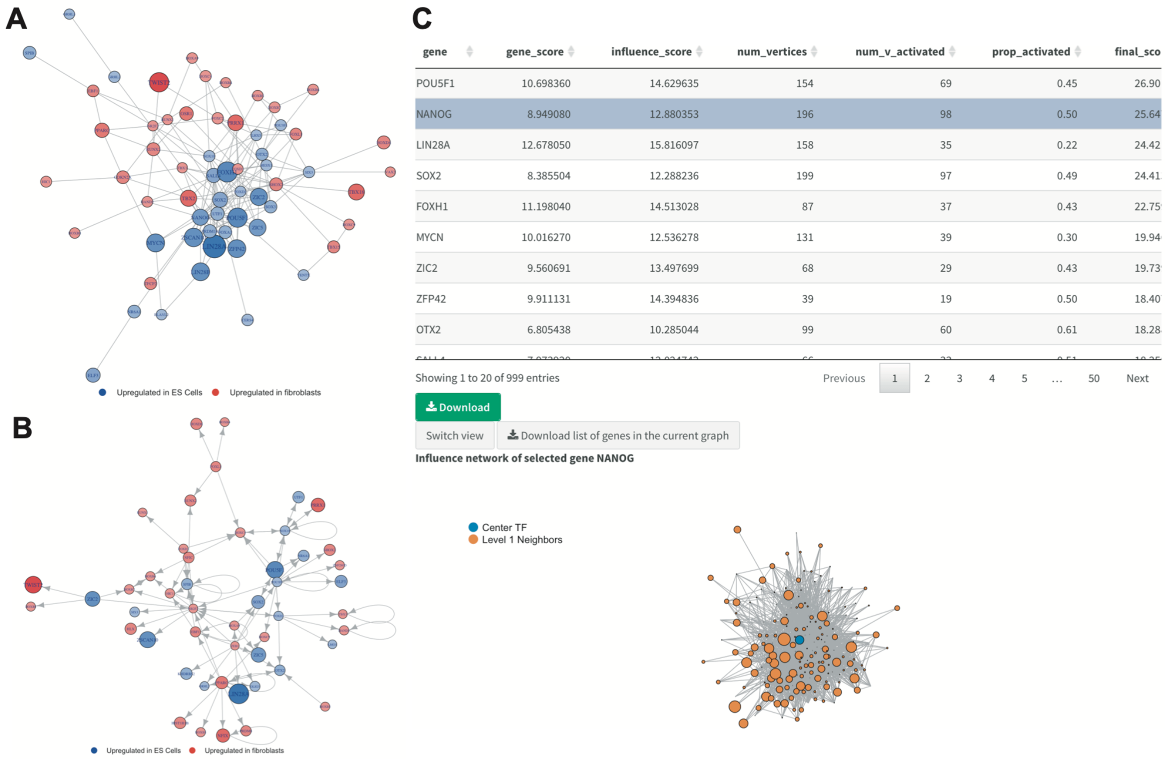
Perform initial PCA step? ● Yes ○ No

Choose plot type: density

Plotting package: plotly

Figure 4.2 Selected analysis modules in PIVOT.

(A) The table on the left lists basic sample statistics. The selected statistics are plotted below the table, and clicking a sample in the table will plot its count distribution. **(B)** Mean-Standard deviation plot (top left, with vsn package), rank frequency plot (top right) and mean variability plot (bottom, with Seurat package). **(C)** The t-SNE module plots 1D, 2D and 3D projections (3D not shown due to space). **(D)** Feature heatmap with the top 100 differentially expressed genes reported by DESeq2 likelihood ratio test.


A
Upregulated in ES Cells
Upregulated in fibroblasts
B
Upregulated in ES Cells
Upregulated in fibroblasts
C
Download
Switch view
Download list of genes in the current graph
Center TF
Level 1 Neighbors

gene	gene_score	influence_score	num_vertices	num_v_activated	prop_activated	final_sco
POU5F1	10.698360	14.629635	154	69	0.45	26.90
NANOG	8.949080	12.880353	196	98	0.50	25.64
LIN28A	12.678050	15.816097	158	35	0.22	24.42
SOX2	8.385504	12.288236	199	97	0.49	24.41
FOXH1	11.198040	14.513028	87	37	0.43	22.75
MYCN	10.016270	12.536278	131	39	0.30	19.94
ZIC2	9.560691	13.497699	68	29	0.43	19.73
ZFP42	9.911131	14.394836	39	19	0.50	18.40
OTX2	6.805438	10.285044	99	60	0.61	18.28

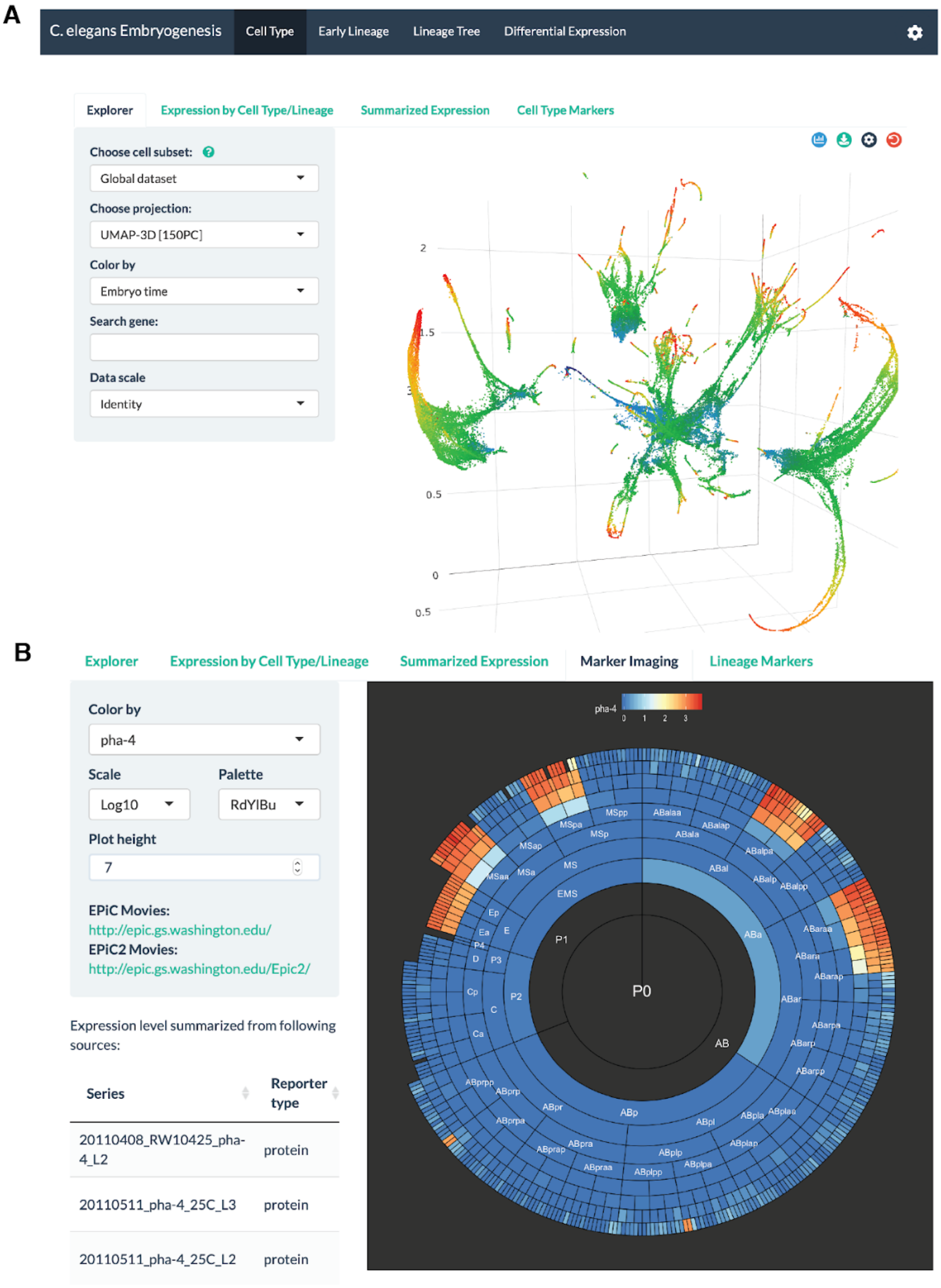
Showing 1 to 20 of 999 entries

Previous 1 2 3 4 5 … 50 Next

Influence network of selected gene NANOG

Figure 4.3 Network analysis for the identification of potential transdifferentiation factors.

(A, **B)** Graphs showing the connection between transcription factors differentially expressed between fibroblasts and ES cells. **A** is an undirected graph showing the protein-protein interaction relationship based on the STRING database, and **B** is constructed based on the Regnetwork repository, showing the regulatory relationship. The size of the nodes and the color gradient indicate the log fold change of the genes. The graphs have been zoomed in to only include the genes with large LFC and small *p*-value. **(C)** Predicted transdifferentiation factor lists based on the network score ranking. The table includes information such as the center transcription factor score, the total number of vertices in its direct neighborhood, and the number of activated neighbors with gene score above a user-specified threshold. Clicking entries on the table will plot the local neighborhood network centered on that TF.



Choose cell subset: Global dataset

Choose projection: UMAP-3D [150PC]

Color by: Embryo time

Search gene:

Data scale: Identity

Color by: pha-4

Scale: Log10

Palette: RdYlBu

Plot height: 7

EPiC Movies:
http://epic.gs.washington.edu/
EPiC2 Movies:
http://epic.gs.washington.edu/Epic2/

Expression level summarized from following sources:

Series	Reporter type
20110408_RW10425_pha-4_L2	protein
20110511_pha-4_25C_L3	protein
20110511_pha-4_25C_L2	protein

Figure 4.4 Screenshots of VisCello.

(A) Screenshot of the cell type explorer, which enables interactive visualization of 2D and 3D UMAPs and PCA plots for different subsets of the data. The view shown in the panel is a 3D UMAP for all cells colored by estimated embryo time. Users can overlay gene expression, cell type, number of expressed genes and other statistics on this plot. The cell type explorer also features box/violin plots for gene expression across cell types, lineages or time, summarized gene expression tables, and marker gene tables. **(B)** Screenshot of the early cell lineage explorer, which enables interactive visualization and comparison of the sc-RNA-seq data and summarized live imaging data. Panel shows a radiograph of average fluorescent intensity (log10 scaled) of *pha-4*, measured by live imaging.

Tables

Table 4.1. List of tools currently integrated/implemented in PIVOT.

PIVOT Modules	Tools Integrated
Normalization	DESeq, Modified DESeq, TMM, Upper quartile, CPM/RPKM/TPM, RUV, Spike-in regression, Census
Feature/Sample Filtering	List based, Expression based and Quality based filters
Basic Analysis Modules	Data distribution plots, Dispersion analysis, Rank-frequency plot, Spike-in analysis, Feature heatmap, etc.
Differential Expression	DESeq2, edgeR, SCDE, Monocle, Mann-Whitney U test
Clustering/Classification	Hierarchical, K-means, SC3, Community detection, Classification with caret, Cell state ordering with Monocle2/Diffusion pseudotime
Dimensionality Reduction	PCA, t-SNE, Metric/Non-Metric MDS, penalized LDA, Diffusion Map
Correlation Analysis	Pairwise scatter plots, Sample/feature correlation heatmap, Co-expression analysis
Gene Set Enrichment Analysis	KEGG pathway analysis, Gene ontology analysis
Network Analysis	STRING protein association network, Regnetwork visualization, Mogrify based transcription factor prediction
Other Utilities	Data map, Gene ID/Name conversion, BioMart gene annotation query, Venn diagram, Report generation, State saving

CHAPTER 5 CONCLUSION AND FUTURE DIRECTIONS

Conclusion

In this thesis I have presented two investigations of gene regulation in development and differentiation. The first study examines cell type transitions and gene expression change at whole-organism scale using the nematode *Caenorhabditis elegans*. The second study aims to understand a specific developmental process, endothelial to hematopoietic transition, which leads to the formation of pre-HSCs. For both studies, I have developed computational methods that facilitate sophisticated statistical analysis of single cell data.

In the *C. elegans* study described in Chapter 2, we profiled the transcriptome of >80,000 single cells across *C. elegans* developmental stages ranging from ~100-600 minutes post-fertilization. Using the scRNA-Seq data, a known *C. elegans* lineage tree, and imaging of fluorescent reporter genes, we generated a lineage-resolved single-cell atlas of embryonic development with 93% cells annotated with a cell type or a cell lineage. This atlas covers early events starting from initial diversification of founder lineages through the specification of terminal cell fates. It enables us to make several unique discoveries about the gene regulatory programs underly development and differentiation. First, by modeling the bifurcating differentiation in neuron development, we identified the widespread presence of 'multilineage priming', where the regulators of multiple alternative fates are co-expressed in a progenitor cell and were selectively inherited by its descendants. Second, by tracing the development of AB lineage, we found a time-dependent association between a cell's transcriptome with its lineage history and fate choice. Gene expression associated with lineage identity is rapidly lost at the time of the terminal division and is replaced by transcriptional program related to the terminal fate. Lastly, we identified several types of convergent development, where distinct lineages converge to a homogeneous transcriptomic state before differentiating to the same cell type.

In Chapter 3, we studied the process of pre-HSC formation in the mouse embryo. By profiling ~40,000 single cells from early mouse embryos, we were able to capture the entire developmental trajectory from arterial endothelial cells to lympho-myeloid biased progenitors and pre-HSCs. Using this dataset, we discovered an endothelial cell precursor of HE cells that we termed pre-HE, in which multiple signaling pathways known to be important for the specification of HE cells are active. Furthermore, through trajectory modeling and RNA velocity analysis, we identified a developmental bottleneck between pre-HE and HE cells. We observed *Runx1*, a key regulator of EHT (*33, 41*), is expressed in approximately 7% of pre-HE cells and hypothesized that its expression level regulates the cell passage through the bottleneck. We validated the hypothesis by performing scRNA-Seq on *Runx1*$^{+/-}$ and *Runx1*$^{+/+}$ littermates and observed a 68% reduction in the proportion of HE and IAC cells in *Runx1* haploinsufficient mice. To investigate the epigenetic mechanism regulating *Runx1* expression at the bottleneck, we performed paired scRNA-Seq and scATAC-Seq on E10.5 CD44^{+} E+HE+IAC cells and developed a computational pipeline for integrative analysis of these two data modalities. We identified a candidate enhancer 371 kb upstream of the *Runx1* P1 promoter that first becomes accessible in pre-HE cells and contains motifs of TFs that are downstream of signaling pathways active in pre-HE. We also found that after cells pass the bottleneck, they follow distinct developmental trajectories leading to an initial wave of lympho-myeloid-biased progenitors at E10.5, followed by precursors of HSCs at E11.5. The findings of this study will inform efforts to generate HSCs from human ESCs and iPSCs *in vitro*.

Interestingly, we observed recurrent gene regulatory patterns in both developmental systems. For example, both systems involve convergence in cellular differentiation that give rise to a common progenitor population. In *C. elegans* embryogenesis, transcriptome state of three sets of distinct cell lineages converges over time and give rise to IL1/IL2 neuroblasts (Figure 2.4A-D). Similarly, two streams of arterial endothelial cells with different Wnt signaling level converge to form conflux AE (Figure 3.2). This phenomenon, known as developmental homoplasy, has been reported for other developmental systems (*333*). The exact mechanism driving the convergence is unknown, but in both cases, we observed coordinated up-regulation of a set of transcription factors in the converging cells, suggesting TFs (and upstream signaling pathways) likely play a key role in homogenizing the transcriptome of distinct cell populations (Figure 2.4A-D and Figure 3.2).

In addition, we observed discontinuous trajectories for many cell types in the UMAP of *C. elegans* embryogenesis, suggesting sudden changes in the transcriptome. In contrast, a developmental bottleneck was found in the trajectory leading to HE, where most cells are blocked at the pre-HE stage. These observations suggest that rate of transcriptome changes may vary during differentiation. Therefore, analysis of the RNA velocity and acceleration will likely reveal important regulatory points in cellular differentiation.

Finally, we found during *C. elegans* embryogenesis, terminal cell fate is specified through a series of lineage bifurcations, each involving multiple differentially expressed

TFs (Figure 2.5C-E). Similarly, analysis of scATAC-Seq data reveals sequential activation of TFs during EHT (Figure 3.4E). Taken together, these results highlight the incremental nature of cell fate decisions and the coordination of the transcriptional regulatory program underlie development and differentiation.

Other contributions

Contributions to papers that are not discussed in this thesis are as follows:

Li N*, Zhu Q*, Ahn KJ, Pourfarhangi KE, Tan K, Lengner CJ. Single cell RNA-Seq and CODEX multiplexed imaging reveals novel macrophage-tumor cell interactions in colorectal cancer. *In preparation*.

Chen C*, Yu W*, Alikarami F, Qiu Q, Chen CH, Flournoy J, Gao P, Uzun Y, Fang L, Hu Y, Zhu Q, *et al.* Single-cell multi-omics reveals elevated plasticity and stem-cell-like blasts underlying the poor prognosis of *KMT2A*-rearranged leukemia. *Submitted*.

Yang Y, Mumau M, Tober J, Zhu Q, Bennet L, Hong C, Sung D, Keller T, Gao P, Shewale S, Chen M, Yang J, Chen X, Thomas SA, Tan K, Speck NA, Kahn ML. Endothelial MEKK3-KLF2/4 signaling integrates inflammatory and hemodynamic signals during definitive hematopoiesis. *In revision*.

Chen GM*, Chen C*, Das RK*, Gao P, Chen CH, Bandyopadhyay S, Ding YY, Uzun Y, Yu W, Zhu Q, Myers RM. Integrative bulk and single-cell profiling of pre-manufacture T-cell pop- ulations reveals factors mediating long-term persistence of CAR T-cell therapy. *Cancer Dis- covery*. 2021 Apr 5.

Gao P, Chen C, Howell ED, Li Y, Tober J, Uzun Y, He B, Gao L, Zhu Q, Siekmann AF, Speck NA. Transcriptional regulatory network controlling the ontogeny of hematopoietic stem cells. *Genes & Development*. 2020 July 1;34(13-14):950-964.

Yu W, Uzun Y, Zhu Q, Chen C, Tan K. scATAC-pro: a comprehensive workbench for single- cell chromatin accessibility sequencing data. *Genome biology*. 2020 April 20;21(1):94.

Bellissimo DC, Chen CH, Zhu Q, Bagga S, Lee CT, He B, Wertheim GB, Jordan M, Tan K, Worthen GS, Gilliland DG. Runx1 negatively regulates inflammatory cytokine production by neutrophils in response to Toll-like receptor signaling. *Blood advances*. 2020 Mar 24;4(6):1145-58.

Peng T, Zhu Q, Yin P, Tan K. SCRABBLE: single-cell RNA-seq imputation constrained by bulk RNA-seq data. *Genome biology*. 2019 Dec 1;20(1):88.

Tang AT, Sullivan KR, Hong CC, Goddard LM, Mahadevan A, Ren A, Pardo H, Peiper A, Griffin E, Tanes C, Mattei LM, Yang J, Li L, Mericko-Ishizuka P, Shen L, Hobson N, Girard R, Lightle R, Moore T, Shenkar R, Polster SP, Roedel CJ, Li N, Zhu Q, *et al.* Distinct cellular roles for PDCD10 define a gut-brain axis in cerebral cavernous malformation. *Science translational medicine*. 2019 Nov 27;11(520).

Ransick A*, Lindström NO*, Liu J*, Zhu Q, Guo JJ, Alvarado GF, Kim AD, Black HG, Kim J, McMahon AP. Single-cell profiling reveals sex, lineage, and regional diversity in the mouse kidney. *Developmental cell*. 2019 Nov 4;51(3):399-413.

Zhu Q, Fisher SA, Shallcross J, Kim J. VERSE: a versatile and efficient RNA-Seq read count- ing tool. *bioRxiv*. 2016 Jan 1:053306.

Future direction: Comparative analysis of embryogenesis across species

Comparative analysis of gene expression across species has enabled discovery of conserved gene expression patterns and coordinated evolution of regulators (*334*). With the increasing amount of single cell data for various developmental systems such as our *C. elegans* single cell atlas and those listed in Table 1.1, it is now possible to investigate evolutionary conserved gene regulatory mechanism at single cell resolution.

For example, using the single cell developmental datasets, one can test the “hourglass” model of development, which predicts that the most conserved developmental period of animal phyla is not the early and late embryonic stage, but a mid-embryonic period, or “phylotypic period” (*335*). The hourglass model is supported by many morphological evidence as well as molecular evidence (*336*), but has never been examined at the single cell resolution, where both cell type and embryonic time can be resolved. Using single cell analysis, one can test this model through a cell-centric approach and a gene-centric approach. In the first approach, homology between cell types across species can be established by comparing the gene expression profile. Then, differentiation trajectories of homologous cell types can be aligned through dynamic time warping (DTW), such that the transcriptome divergence across developmental timepoints can be compared. In the second gene-centric approach, one can perform TRN inference for each species and search for conserved TRN modules, and check if these modules are activated at specific developmental stage.

One potential caveat is that most published single cell data on development are from evolutionarily distant species, making it difficult to establish homology between cells from these species. The *Caenorhabditis* genus of nematodes includes ~50 species with almost identical cell lineage structure, and thus provides an ideal system for comparative study of embryogenesis. Furthermore, for one of these species, *C. elegans*, we have generated a lineage-resolved single cell atlas, which can be used as a reference dataset. Therefore, single cell sequencing on other closely related species, such as *C. remanei, C. brenneri* and *C. briggsae*, will likely enable a comprehensive survey of evolutionarily conserved gene regulatory principles during embryogenesis.

Future direction: Uncover gene regulatory mechanisms underlie EHT

Understanding how HSC form *in vivo* is essential for producing HSCs from other cell sources *ex vivo*. Our recent work expanded our knowledge about EHT, but the complete gene regulatory mechanisms underlie EHT has not been resolved.

For example, in Chapter 3, we showed using single cell analysis that a developmental bottleneck exists between pre-HE and HE, and *Runx1* dosage regulates cell passage through the bottleneck. We further showed, using scATAC analysis, that a distal enhancer 371 kb upstream of *Runx1* is transiently accessible in pre-HE cells. However, we do not yet know if the Runx1 -371 enhancer is required to activate *Runx1* expression at the bottleneck. Some evidence from zebrafish suggest that this enhancer is capable of driving reporter gene expression in the intermediate cell mass and posterior blood island of zebrafish embryos (*251*), but it remains to be shown that the enhancer directly regulates *Runx1* expression in pre-HE and HE cells. To answer this question, one can use the CRISPR/Cas9 system to delete this enhancer in mouse models and observe the effect on *Runx1* expression. Furthermore, to test which transcription factors bind to the enhancer to activate its activity, point mutations can be introduced to destroy the GATA/TAL1, STAT, JUN, and RAR/RXR motifs at the enhancer site. A significant reduction of enhancer activity in these mutants would suggest the requirement for the corresponding transcription factor. Finally, one can perform circular chromosome conformation capture sequencing (4C-seq) to identify the changes in enhancer-promoter looping during pre-HE to HE transition. It is possible that the -371 enhancer is required to initiate *Runx1* expression through a long-range interaction with *Runx1* promoter but is no longer required thereafter as other *Runx1* enhancers take over.

Another unsolved problem is what regulates the pre-HSC production from HE cells. Our single cell data suggest that at E10.5, most of the HE cells differentiate into lympho-myeloid progenitors, and only 2% of the IAC cells are pre-HSCs. However, at E11.5, the proportion shifts dramatically, where 67% of the IAC cells are found to be pre-HSCs. The fate probability analysis shown in Supplemental Figure 3.11 provides some clue, as we observe down-regulation of *Myc* and up-regulation of *Smad7*, *Mecom*, *Meis2* and *Nfix* in the trajectory leading to pre-HSC. *Myc* is a proto-oncogene which drives cell proliferation (*337*), and previous study have shown that cell cycle is slowed down considerably as pre-HSC mature (*338*). On the other hand, *Mecom* encodes the transcription factors *Evi1* and *Mds1-Evi1*, which are known to promote stemness and quiescence of adult HSCs (*339, 340*). These suggest that regulation of cell cycle may play an important role in the specification of the pre-HSC fate. Future experiments that perturb expression level of these transcription factors may reveal the underlying gene regulatory network controlling the fate choice of HE cells.

Future direction: *In silico* modeling of development and differentiation

The pace and volume of single cell data collection across a variety of developmental systems is exploding, giving the hope that we now have the data density to meaningfully apply latest machine learning technologies to model complex organismal development.

First, at the single-cell level, development involves cell differentiation where each cell adopts a series of molecular changes and differentiates into the terminal cell type, such as a neuron or a blood cell. Although significant progress has been made to discover molecular switches regulating differentiation, we do not yet know how the control fully works even for a small part of the process. For example, in this thesis I showed that the *Runx1* is a key regulator helping endothelial cells (blood vessel cells) overcome a developmental bottleneck to become hematopoietic stem cells (HSCs). Yet little is known about which genes activate *Runx1* expression in the first place. To decode the regulatory program of cell differentiation, a different approach from that described in previous section is to build an *in silico* cell differentiation model using deep learning. Deep learning has demonstrated great power in solving problems in dynamical domain such as autonomous driving (*341*). Self-driving cars uses information about a car's current state, such as speed, direction, and surroundings, to make decisions on the adjustments to speed and direction. Like self-driving cars, a deep neural network can take the cell's molecular state at each time point, as measured by single-cell sequencing, and make predictions about the cell's future state. The model can be trained with currently available 10^4-10^6 single cell data points along the differentiation trajectory, such that it can robustly predict the series of molecular changes happening inside a developing cell. One caveat of using deep neural network is that the learned model is presented as a black box and is hard to interpret (*342*). But recent advances such as SHAP (SHapley Additive exPlanations) (*343*) have made it possible to look inside the black box and examine which genes contribute most to the prediction result. Therefore, we might use the model as an *in silico* hypothesis generating platform to query the potential role of a gene during differentiation.

Second, development is a multi-cellular process involving different types of cells interacting with each other in a spatial context and organizing into intricate tissue structures. So far, efforts to re-create functional tissues *in vitro* by organoid culture systems have largely failed, as the organoids cannot recapitulate the cell type diversity and cell-cell interaction *in vivo* (*344*). Recently, multi-agent reinforcement learning has achieved great success in modeling multi-component systems such as real-time strategy games (*345*) and drone systems (*346*). Therefore, it might be possible to expand the single cell models into a multi-agent system to simulate cell-cell communications *in silico*. One of the best ground truth datasets for benchmarking is our embryogenesis atlas of *C. elegans*, which is a simple organism with only 558 cells at hatching and spatial location of every cell known. Building on top of this, it might be possible to use multi-agent learning to simulate the interactions between spatially adjacent cells and the development of a “virtual embryo” over time.

Finally, to make the models trained using data from model organisms applicable for understanding human developmental processes, transfer learning techniques can be applied. The Human Cell Atlas project has by far collected millions of cells from human post-mortem tissues. However, the majority of the cells come from adults. In other words, we only have information from the end point of human development. These existing data can be used as “anchor points” for establishing the connections between human and mouse. With these anchor points, a model trained in mouse can be adapted using transfer learning to make predictions for matched human developmental stages.

BIBLIOGRAPHY

1. J. Eberwine, J.-Y. Sul, T. Bartfai, J. Kim, The promise of single-cell sequencing. *Nature methods* **11**, 25-27 (2014).
2. A. A. Kolodziejczyk, J. K. Kim, V. Svensson, J. C. Marioni, S. A. Teichmann, The technology and biology of single-cell RNA sequencing. *Molecular cell* **58**, 610-620 (2015).
3. J. D. Buenrostro *et al.*, Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486-490 (2015).
4. D. A. Cusanovich *et al.*, Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910-914 (2015).
5. Y. Goltsev *et al.*, Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell* **174**, 968-981. e915 (2018).
6. J. E. Sulston, H. R. Horvitz, Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. *Developmental biology* **56**, 110-156 (1977).
7. J. E. Sulston, E. Schierenberg, J. G. White, J. N. Thomson, The embryonic cell lineage of the nematode Caenorhabditis elegans. *Developmental biology* **100**, 64-119 (1983).
8. C. e. S. Consortium*, Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-2018 (1998).
9. Z. Altun, L. Herndon, C. Crocker, R. Lints, D. Hall. (WormAtlas, 2002-2021).
10. B. Goldstein, S. N. Hird, Specification of the anteroposterior axis in Caenorhabditis elegans. *Development* **122**, 1467-1474 (1996).
11. J. R. Priess, J. N. Thomson, Cellular interactions in early C. elegans embryos. *Cell* **48**, 241-250 (1987).
12. C. I. Bargmann, L. Avery, Laser killing of cells in Caenorhabditis elegans. *Methods in cell biology* **48**, 225-250 (1995).
13. B. Bowerman, B. A. Eaton, J. R. Priess, skn-1, a maternally expressed gene required to specify the fate of ventral blastomeres in the early C. elegans embryo. *Cell* **68**, 1061-1075 (1992).
14. C. E. Rocheleau *et al.*, Wnt signaling and an APC-related gene specify endoderm in early C. elegans embryos. *Cell* **90**, 707-716 (1997).
15. C. J. Thorpe, A. Schlesinger, J. C. Carter, B. Bowerman, Wnt signaling polarizes an early C. elegans blastomere to distinguish endoderm from mesoderm. *Cell* **90**, 695-705 (1997).
16. C. C. Mello, B. W. Draper, J. R. Prless, The maternal genes apx-1 and glp-1 and establishment of dorsal-ventral polarity in the early C. elegans embryo. *Cell* **77**, 95-106 (1994).
17. J. I. Murray *et al.*, Multidimensional regulation of gene expression in the C. elegans embryo. *Genome research* **22**, 1282-1294 (2012).
18. T. Hashimshony, F. Wagner, N. Sher, I. Yanai, CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell reports* **2**, 666-673 (2012).
19. S. C. Tintori, E. O. Nishimura, P. Golden, J. D. Lieb, B. Goldstein, A transcriptional lineage of the early C. elegans embryo. *Developmental cell* **38**, 430-444 (2016).
20. G. Canu, C. Ruhrberg, First blood: the endothelial origins of hematopoietic progenitors. *Angiogenesis*, 1-13 (2021).
21. Y. Groner *et al.*, *RUNX proteins in development and cancer*. (Springer, 2017).

22. J. Palis, S. Robertson, M. Kennedy, C. Wall, G. Keller, Development of erythroid and myeloid progenitors in the yolk sac and embryo proper of the mouse. *Development* **126**, 5073-5084 (1999).
23. M.-j. Xu *et al.*, Evidence for the presence of murine primitive megakarycytopoiesis in the early yolk sac. *Blood, The Journal of the American Society of Hematology* **97**, 2016-2022 (2001).
24. J. Tober *et al.*, The megakaryocyte lineage originates from hemangioblast precursors and is an integral component both of primitive and of definitive hematopoiesis. *Blood* **109**, 1433-1441 (2007).
25. M. C. Yoder, Inducing definitive hematopoiesis in a dish. *Nature biotechnology* **32**, 539 (2014).
26. B. Kasaai *et al.*, Erythro-myeloid progenitors can differentiate from endothelial cells and modulate embryonic vascular remodeling. *Scientific reports* **7**, 1-12 (2017).
27. A. Plein, A. Fantin, L. Denti, J. W. Pollard, C. Ruhrberg, Erythro-myeloid progenitors contribute endothelial cells to blood vessels. *Nature* **562**, 223-228 (2018).
28. M. Yoshimoto *et al.*, Embryonic day 9 yolk sac and intra-embryonic hemogenic endothelium independently generate a B-1 and marginal zone progenitor lacking B-2 potential. *Proceedings of the National Academy of Sciences* **108**, 1468-1473 (2011).
29. M. Yoshimoto *et al.*, Autonomous murine T-cell progenitor production in the extra-embryonic yolk sac before HSC emergence. *Blood* **119**, 5706-5714 (2012).
30. C. Böiers *et al.*, Lymphomyeloid contribution of an immune-restricted progenitor emerging prior to definitive hematopoietic stem cells. *Cell stem cell* **13**, 535-548 (2013).
31. M. Tavian *et al.*, Aorta-associated CD34+ hematopoietic cells in the early human embryo. (1996).
32. E. Oberlin, B. El Hafny, L. Petit-Cocault, M. Souyri, Definitive human and mouse hematopoiesis originates from the embryonic endothelium: a new class of HSCs based on VE-cadherin expression. *International Journal of Developmental Biology* **54**, 1165-1173 (2010).
33. A. D. Yzaguirre, M. F. de Bruijn, N. A. Speck, The role of Runx1 in embryonic blood cell formation. *RUNX proteins in development and cancer*, 47-64 (2017).
34. G. Swiers, C. Rode, E. Azzoni, M. F. de Bruijn, A short history of hemogenic endothelium. *Blood Cells, Molecules, and Diseases* **51**, 206-212 (2013).
35. K. Kissa *et al.*, Live imaging of emerging hematopoietic stem cells and early thymus colonization. *Blood, The Journal of the American Society of Hematology* **111**, 1147-1156 (2008).
36. S. Rybtsov, A. Ivanovs, S. Zhao, A. Medvinsky, Concealed expansion of immature precursors underpins acute burst of adult HSC activity in foetal liver. *Development* **143**, 1284-1289 (2016).
37. T. Yokomizo *et al.*, Whole-mount three-dimensional imaging of internally localized immunostained cells within mouse embryos. *Nat Protoc* **7**, 421-431 (2012).
38. S. Rybtsov *et al.*, Hierarchical organization and early hematopoietic specification of the developing HSC lineage in the AGM region. *J Exp Med* **208**, 1305-1315 (2011).

39. F. Zhou *et al.*, Tracing haematopoietic stem cell formation at single-cell resolution. *Nature* **533**, 487-492 (2016).
40. A. Alemany, M. Florescu, C. S. Baron, J. Peterson-Maduro, A. Van Oudenaarden, Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108-112 (2018).
41. M. J. Chen, T. Yokomizo, B. M. Zeigler, E. Dzierzak, N. A. Speck, Runx1 is required for the endothelial to haematopoietic cell transition but not thereafter. *Nature* **457**, 887-891 (2009).
42. T. E. North, T. Stacy, C. J. Matheny, N. A. Speck, M. F. de Bruijn, Runx1 is expressed in adult mouse hematopoietic stem cells and differentiating myeloid and lymphoid cells, but not in maturing erythroid cells. *Stem cells* **22**, 158-168 (2004).
43. M. Lichtinger *et al.*, RUNX1 reshapes the epigenetic landscape at the onset of haematopoiesis. *The EMBO journal* **31**, 4318-4333 (2012).
44. C. Lancrin *et al.*, GFI1 and GFI1B control the loss of endothelial identity of hemogenic endothelium during hematopoietic commitment. *Blood, The Journal of the American Society of Hematology* **120**, 314-322 (2012).
45. R. Lis *et al.*, Conversion of adult endothelium to immunocompetent haematopoietic stem cells. *Nature* **545**, 439-445 (2017).
46. F.-Y. Tsai *et al.*, An early haematopoietic defect in mice lacking the transcription factor GATA-2. *Nature* **371**, 221-226 (1994).
47. K.-W. Ling *et al.*, GATA-2 plays two functionally distinct roles during the ontogeny of hematopoietic stem cells. *The Journal of experimental medicine* **200**, 871-882 (2004).
48. K. Ottersbach, Endothelial-to-haematopoietic transition: an update on the process of making blood. *Biochemical Society Transactions* **47**, 591-601 (2019).
49. L. Gama-Norton *et al.*, Notch signal strength controls cell fate in the haemogenic endothelium. *Nature communications* **6**, 1-12 (2015).
50. B. Chanda, A. Ditadi, N. N. Iscove, G. Keller, Retinoic acid signaling is essential for embryonic hematopoietic stem cell development. *Cell* **155**, 215-227 (2013).
51. C. Souilhol *et al.*, Inductive interactions mediated by interplay of asymmetric signalling underlie development of adult haematopoietic stem cells. *Nature communications* **7**, 1-13 (2016).
52. R. Espin-Palazon *et al.*, Proinflammatory signaling regulates hematopoietic stem cell emergence. *Cell* **159**, 1070-1085 (2014).
53. T. E. North *et al.*, Hematopoietic stem cell development is dependent on blood flow. *Cell* **137**, 736-748 (2009).
54. A. M. Klein *et al.*, Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201 (2015).
55. E. Z. Macosko *et al.*, Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202-1214 (2015).
56. G. X. Zheng *et al.*, Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 1-12 (2017).
57. A. B. Rosenberg *et al.*, Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176-182 (2018).
58. J. Cao *et al.*, Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661-667 (2017).
59. J. Cao *et al.*, The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502 (2019).

60. J. Eberwine *et al.*, Analysis of gene expression in single live neurons. *Proceedings of the National Academy of Sciences* **89**, 3010-3014 (1992).
61. J. Phillips, J. H. Eberwine, Antisense RNA amplification: a linear amplification method for analyzing the mRNA population from single living cells. *Methods* **10**, 283-288 (1996).
62. S. Islam *et al.*, Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature methods* **11**, 163 (2014).
63. J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. *Nature methods* **10**, 1213 (2013).
64. A. T. Satpathy *et al.*, Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature biotechnology* **37**, 925-936 (2019).
65. J. Cao *et al.*, Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380-1385 (2018).
66. C. Zhu *et al.*, An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nature structural & molecular biology* **26**, 1063-1070 (2019).
67. S. Chen, B. B. Lake, K. Zhang, High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature biotechnology* **37**, 1452-1457 (2019).
68. S. Ma *et al.*, Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116. e1120 (2020).
69. M. Stoeckius *et al.*, Simultaneous epitope and transcriptome measurement in single cells. *Nature methods* **14**, 865-868 (2017).
70. G. Li *et al.*, Joint profiling of DNA methylation and chromatin architecture in single cells. *Nature methods* **16**, 991-993 (2019).
71. C. Zhu *et al.*, Joint profiling of histone modifications and transcriptome in single cells from mouse brain. *Nature Methods* **18**, 283-292 (2021).
72. S. J. Clark *et al.*, scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nature communications* **9**, 1-9 (2018).
73. E. Swanson *et al.*, Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. *Elife* **10**, e63632 (2021).
74. G. A. Rutter, H. J. Kennedy, C. D. Wood, M. R. White, J. M. Tavaré, Real-time imaging of gene expression in single living cells. *Chemistry & biology* **5**, R285-R290 (1998).
75. J. I. Murray *et al.*, Automated analysis of embryonic gene expression with cellular resolution in C. elegans. *Nature methods* **5**, 703-709 (2008).
76. A.-S. Chiang *et al.*, Three-dimensional reconstruction of brain-wide wiring networks in Drosophila at single-cell resolution. *Current biology* **21**, 1-11 (2011).
77. Y. Wang *et al.*, Rapid sequential in situ multiplexing with DNA exchange imaging in neuronal cells and tissues. *Nano letters* **17**, 6131-6139 (2017).
78. J.-R. Lin *et al.*, Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. *Elife* **7**, (2018).
79. C. M. Schürch *et al.*, Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell* **182**, 1341-1359. e1319 (2020).

80. A. M. Femino, F. S. Fay, K. Fogarty, R. H. Singer, Visualization of single RNA transcripts in situ. *Science* **280**, 585-590 (1998).
81. A. Raj, P. Van Den Bogaard, S. A. Rifkin, A. Van Oudenaarden, S. Tyagi, Imaging individual mRNA molecules using multiple singly labeled probes. *Nature methods* **5**, 877-879 (2008).
82. E. Lubeck, A. F. Coskun, T. Zhiyentayev, M. Ahmad, L. Cai, Single-cell in situ RNA profiling by sequential hybridization. *Nature methods* **11**, 360 (2014).
83. J. H. Lee *et al.*, Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360-1363 (2014).
84. K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, X. Zhuang, Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, (2015).
85. L. Cai, Transcriptome-Scale Super-Resolved Imaging in Tissues by RNA SeqFISH. *European Journal of Human Genetics* **28**, 10 (2020).
86. G. Wang, J. R. Moffitt, X. Zhuang, Multiplexed imaging of high-density libraries of RNAs with MERFISH and expansion microscopy. *Scientific reports* **8**, 1-13 (2018).
87. S. Shah, E. Lubeck, W. Zhou, L. Cai, In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. *Neuron* **92**, 342-357 (2016).
88. K. L. Frieda *et al.*, Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107-111 (2017).
89. C.-H. L. Eng, S. Shah, J. Thomassie, L. Cai, Profiling the transcriptome with RNA SPOTs. *Nature methods* **14**, 1153-1155 (2017).
90. J.-H. Su, P. Zheng, S. S. Kinrot, B. Bintu, X. Zhuang, Genome-scale imaging of the 3D organization and transcriptional activity of chromatin. *Cell* **182**, 1641-1659. e1626 (2020).
91. E. G. Conklin. (Academy of Natural Sciences, 1905).
92. S. G. Megason, in *Zebrafish*. (Springer, 2009), pp. 317-332.
93. K. McDole *et al.*, In toto imaging and reconstruction of post-implantation mouse development at the single-cell level. *Cell* **175**, 859-876. e833 (2018).
94. K. Kretzschmar, F. M. Watt, Lineage tracing. *Cell* **148**, 33-45 (2012).
95. L. S. Ludwig *et al.*, Lineage tracing in humans enabled by mitochondrial mutations and single-cell genomics. *Cell* **176**, 1325-1339. e1322 (2019).
96. D. T. Montoro *et al.*, A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319-324 (2018).
97. D. E. Wagner *et al.*, Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981-987 (2018).
98. B. Raj *et al.*, Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nature biotechnology* **36**, 442-450 (2018).
99. B. Spanjaard *et al.*, Simultaneous lineage tracing and cell-type identification using CRISPR–Cas9-induced genetic scars. *Nature biotechnology* **36**, 469-473 (2018).
100. H. Dueck *et al.*, Deep sequencing reveals cell-type-specific patterns of single-cell transcriptome variation. *Genome biology* **16**, 1-17 (2015).
101. C. R. Cadwell *et al.*, Electrophysiological, transcriptomic and morphologic profiling of single neurons using Patch-seq. *Nature biotechnology* **34**, 199-203 (2016).
102. C. J. Hsiao *et al.*, Characterizing and inferring quantitative cell cycle phase in single-cell RNA-seq data analysis. *Genome research* **30**, 611-621 (2020).

103. S. S. Fonseca Costa, M. Robinson-Rechavi, J. A. Ripperger, Single-cell transcriptomics allows novel insights into aging and circadian processes. *Briefings in Functional Genomics* **19**, 343-349 (2020).
104. A. Saviano, N. C. Henderson, T. F. Baumert, Single-cell genomics and spatial transcriptomics: discovery of novel cell states and cellular interactions in liver physiology and disease biology. *Journal of hepatology*, (2020).
105. C. Ziegenhain *et al.*, Comparative analysis of single-cell RNA sequencing methods. *Molecular cell* **65**, 631-643. e634 (2017).
106. M. Huang *et al.*, SAVER: gene expression recovery for single-cell RNA sequencing. *Nature methods* **15**, 539-542 (2018).
107. W. Gong, I.-Y. Kwak, P. Pota, N. Koyano-Nakagawa, D. J. Garry, DrImpute: imputing dropout events in single cell RNA sequencing data. *BMC bioinformatics* **19**, 1-10 (2018).
108. W. V. Li, J. J. Li, An accurate and robust imputation method scImpute for single-cell RNA-seq data. *Nature communications* **9**, 1-9 (2018).
109. D. Talwar, A. Mongia, D. Sengupta, A. Majumdar, AutoImpute: Autoencoder based imputation of single-cell RNA-seq data. *Scientific reports* **8**, 1-11 (2018).
110. Y. Xu *et al.*, scIGANs: single-cell RNA-seq imputation using generative adversarial networks. *Nucleic acids research* **48**, e85-e85 (2020).
111. E. Pierson, C. Yau, ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome biology* **16**, 1-10 (2015).
112. D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, J.-P. Vert, A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature communications* **9**, 1-17 (2018).
113. K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, in *International conference on database theory*. (Springer, 1999), pp. 217-235.
114. I. Tirosh *et al.*, Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196 (2016).
115. L. Van der Maaten, G. Hinton, Visualizing data using t-SNE. *Journal of machine learning research* **9**, (2008).
116. L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, (2018).
117. E. Becht *et al.*, Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*, (2018).
118. L. W. Plasschaert *et al.*, A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377-381 (2018).
119. C. Weinreb, S. Wolock, A. M. Klein, SPRING: a kinetic interface for visualizing high dimensional single-cell expression data. *Bioinformatics* **34**, 1246-1248 (2018).
120. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, in *Kdd*. (1996), vol. 96, pp. 226-231.
121. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008**, P10008 (2008).
122. B. Wang, J. Zhu, E. Pierson, D. Ramazzotti, S. Batzoglou, Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature methods* **14**, 414-416 (2017).
123. V. Y. Kiselev *et al.*, SC3: consensus clustering of single-cell RNA-seq data. *Nature methods* **14**, 483-486 (2017).

124. F. A. Wolf *et al.*, PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology* **20**, 1-9 (2019).
125. G. W. Schwartz *et al.*, TooManyCells identifies and visualizes relationships of single-cell clades. *Nature methods* **17**, 405-413 (2020).
126. A. Regev *et al.*, Science forum: the human cell atlas. *Elife* **6**, e27041 (2017).
127. H. Consortium, The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187 (2019).
128. G. Pasquini, J. E. R. Arias, P. Schäfer, V. Busskamp, Automated methods for cell type annotation on scRNA-seq data. *Computational and Structural Biotechnology Journal*, (2021).
129. S. C. Bendall *et al.*, Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714-725 (2014).
130. P. M. Magwene, P. Lizardi, J. Kim, Reconstructing the temporal ordering of biological samples using microarray data. *Bioinformatics* **19**, 842-850 (2003).
131. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32**, 381 (2014).
132. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell trajectories. *Nature methods* **14**, 979 (2017).
133. J. Chen, A. Schlitzer, S. Chakarov, F. Ginhoux, M. Poidinger, Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. *Nature communications* **7**, 1-15 (2016).
134. J. D. Welch, A. J. Hartemink, J. F. Prins, SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome biology* **17**, 1-15 (2016).
135. W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods. *Nat Biotechnol* **37**, 547-554 (2019).
136. R. R. Coifman *et al.*, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the national academy of sciences* **102**, 7426-7431 (2005).
137. S. Nestorowa *et al.*, A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood, The Journal of the American Society of Hematology* **128**, e20-e31 (2016).
138. K. Street *et al.*, Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics* **19**, 1-16 (2018).
139. K. R. Campbell, C. Yau, switchde: inference of switch-like differential expression along single-cell trajectories. *Bioinformatics* **33**, 1241-1242 (2017).
140. G. La Manno *et al.*, RNA velocity of single cells. *Nature* **560**, 494-498 (2018).
141. V. Bergen, Lange, M., Peidli, S., Wolf, F.A., Fabian, F.J., Generalizing RNA velocity to transient cell states through dynamical modeling. *bioRxiv*. 2019 (https://doi.org/10.1101/820936).
142. G. Gorin, V. Svensson, L. Pachter, Protein velocity and acceleration from single-cell multiomics experiments. *Genome biology* **21**, 1-6 (2020).
143. C. Weinreb, A. Rodriguez-Fraticelli, F. D. Camargo, A. M. Klein, Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, (2020).
144. J. S. Herman, Sagar, D. Grun, FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nat Methods* **15**, 379-386 (2018).

145. G. Schiebinger *et al.*, Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* **176**, 928-943. e922 (2019).
146. C. Weinreb, S. Wolock, B. K. Tusi, M. Socolovsky, A. M. Klein, Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences* **115**, E2467-E2476 (2018).
147. D. Marbach *et al.*, Wisdom of crowds for robust gene network inference. *Nature methods* **9**, 796-804 (2012).
148. C. Trapnell, Defining cell types and states with single-cell genomics. *Genome research* **25**, 1491-1498 (2015).
149. A. T. Specht, J. Li, LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. *Bioinformatics* **33**, 764-766 (2017).
150. N. Papili Gao, S. M. Ud-Dean, O. Gandrillon, R. Gunawan, SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles. *Bioinformatics* **34**, 258-266 (2018).
151. H. Matsumoto *et al.*, SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics* **33**, 2314-2321 (2017).
152. X. Qiu *et al.*, Inferring causal gene regulatory networks from coupled single-cell expression dynamics using Scribe. *Cell systems* **10**, 265-274. e211 (2020).
153. K. Kamimoto, C. M. Hoffmann, S. A. Morris, CellOracle: Dissecting cell identity via network inference and in silico gene perturbation. *bioRxiv*, (2020).
154. V. Moignard *et al.*, Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nature cell biology* **15**, 363-372 (2013).
155. C. Nerlov, T. Graf, PU. 1 induces myeloid lineage commitment in multipotent hematopoietic progenitors. *Genes & development* **12**, 2403-2412 (1998).
156. S. Huang, Y.-P. Guo, G. May, T. Enver, Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Developmental biology* **305**, 695-713 (2007).
157. S. Aibar *et al.*, SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083-1086 (2017).
158. V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, P. Geurts, Inferring regulatory networks from expression data using tree-based methods. *PloS one* **5**, 1-10 (2010).
159. A. Dixit *et al.*, Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853-1866. e1817 (2016).
160. B. Adamson *et al.*, A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867-1882. e1821 (2016).
161. C. A. Jackson, D. M. Castro, G.-A. Saldi, R. Bonneau, D. Gresham, Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. *Elife* **9**, e51254 (2020).
162. R. Argelaguet *et al.*, MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology* **21**, 1-17 (2020).
163. Y. Hao *et al.*, Integrated analysis of multimodal single-cell data. *Cell*, (2021).

164. J. D. Welch, A. J. Hartemink, J. F. Prins, MATCHER: manifold alignment reveals correspondence between single cell transcriptome and epigenome dynamics. *Genome biology* **18**, 1-19 (2017).
165. J. D. Welch *et al.*, Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873-1887. e1817 (2019).
166. A. Mo *et al.*, Epigenomic signatures of neuronal diversity in the mammalian brain. *Neuron* **86**, 1369-1384 (2015).
167. T. Stuart *et al.*, Comprehensive integration of single-cell data. *Cell* **177**, 1888-1902. e1821 (2019).
168. T. Peng, G. Chen, K. Tan, GLUER: integrative analysis of single-cell omics and imaging data by deep neural network. *bioRxiv*, (2021).
169. Z. Zhang, C. Yang, X. Zhang, Learning latent embedding of multi-modal single cell data and cross-modality relationship simultaneously. *bioRxiv*, (2021).
170. J. S. Packer *et al.*, A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. *Science* **365**, (2019).
171. B. Pijuan-Sala *et al.*, A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490-495 (2019).
172. M. M. Chan *et al.*, Molecular recording of mammalian embryogenesis. *Nature* **570**, 77-82 (2019).
173. N. Karaiskos *et al.*, The Drosophila embryo at single-cell transcriptome resolution. *Science* **358**, 194-199 (2017).
174. J. A. Briggs *et al.*, The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, (2018).
175. J. A. Farrell *et al.*, Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, (2018).
176. S. Siebert *et al.*, Stem cell differentiation trajectories in Hydra resolved at single-cell resolution. *Science* **365**, (2019).
177. M. Plass *et al.*, Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science* **360**, (2018).
178. J. M. Musser *et al.*, Profiling cellular diversity in sponges informs animal cell type and nervous system evolution. *BioRxiv*, 758276 (2019).
179. C. T. Fincher, O. Wurtzel, T. de Hoog, K. M. Kravarik, P. W. Reddien, Cell type transcriptome atlas for the planarian Schmidtea mediterranea. *Science* **360**, (2018).
180. C. Cao *et al.*, Comprehensive single-cell transcriptome lineages of a proto-vertebrate. *Nature* **571**, 349-354 (2019).
181. X. Qiu *et al.*, Single-cell mRNA quantification and differential analysis with Census. *Nature methods* **14**, 309-315 (2017).
182. T. M. Consortium, Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367-372 (2018).
183. A. Zeisel *et al.*, Molecular architecture of the mouse nervous system. *Cell* **174**, 999-1014. e1022 (2018).
184. A. Sebé-Pedrós *et al.*, Cnidarian cell type diversity and regulation revealed by whole-organism single-cell RNA-Seq. *Cell* **173**, 1520-1534. e1520 (2018).
185. M. Sarov *et al.*, A genome-scale resource for in vivo tag-based protein function exploration in C. elegans. *Cell* **150**, 855-866 (2012).
186. T. Hashimshony, M. Feder, M. Levin, B. K. Hall, I. Yanai, Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer. *Nature* **519**, 219-222 (2015).

187. E. Becht *et al.*, Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology* **37**, 38-44 (2019).

188. R. Y. N. Lee *et al.*, WormBase 2017: molting into a new stage. *Nucleic acids research* **46**, D869-D874 (2018).

189. C. L. Araya *et al.*, Regulatory analysis of the C. elegans genome with spatiotemporal resolution. *Nature* **512**, 400-405 (2014).

190. G. Broitman-Maduro *et al.*, The NK-2 class homeodomain factor CEH-51 and the T-box factor TBX-35 have overlapping function in C. elegans mesoderm development. *Development* **136**, 2735-2746 (2009).

191. W. C. Spencer *et al.*, A spatial and temporal map of C. elegans gene expression. *Genome research* **21**, 325-341 (2011).

192. J. L. Richards, A. L. Zacharias, T. Walton, J. T. Burdick, J. I. Murray, A quantitative model of normal Caenorhabditis elegans embryogenesis and its disruption after stress. *Developmental biology* **374**, 12-23 (2013).

193. M. Hu *et al.*, Multilineage gene expression precedes commitment in the hemopoietic system. *Genes & development* **11**, 774-785 (1997).

194. P. Laslo *et al.*, Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell* **126**, 755-766 (2006).

195. M. Thomson *et al.*, Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. *Cell* **145**, 875-889 (2011).

196. E. W. Brunskill *et al.*, Single cell dissection of early kidney development: multilineage priming. *Development* **141**, 3093-3101 (2014).

197. W. Wang *et al.*, A single-cell transcriptional roadmap for cardiopharyngeal fate diversification. *Nature cell biology* **21**, 674-686 (2019).

198. O. Hobert, A map of terminal regulators of neuronal identity in Caenorhabditis elegans. *Wiley Interdisciplinary Reviews: Developmental Biology* **5**, 474-498 (2016).

199. S. Yu, L. Avery, E. Baude, D. L. Garbers, Guanylyl cyclase expression in specific sensory neurons: a new family of chemosensory receptors. *Proceedings of the National Academy of Sciences* **94**, 3384-3387 (1997).

200. E. R. Troemel, A. Sagasti, C. I. Bargmann, Lateral signaling mediated by axon contact and calcium entry regulates asymmetric odorant receptor expression in C. elegans. *Cell* **99**, 387-398 (1999).

201. O. Hobert, K. Tessmar, G. Ruvkun, The Caenorhabditis elegans lim-6 LIM homeobox gene regulates neurite outgrowth and function of particular GABAergic neurons. *Development* **126**, 1547-1562 (1999).

202. J. T. Pierce-Shimomura, S. Faumont, M. R. Gaston, B. J. Pearson, S. R. Lockery, The homeobox gene lim-6 is required for distinct chemosensory representations in C. elegans. *Nature* **410**, 694-698 (2001).

203. B. J. Lesch, C. I. Bargmann, The homeodomain protein hmbx-1 maintains asymmetric gene expression in adult C. elegans olfactory neurons. *Genes & development* **24**, 1802-1815 (2010).

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

205. K. Brunschwig *et al.*, Anterior organization of the Caenorhabditis elegans embryo by the labial-like Hox gene ceh-13. *Development* **126**, 1537-1546 (1999).

206. T. Hirose, B. D. Galvin, H. R. Horvitz, Six and Eya promote apoptosis through direct transcriptional activation of the proapoptotic BH3-only gene egl-1 in

Caenorhabditis elegans. *Proceedings of the National Academy of Sciences* **107**, 15479-15484 (2010).

207. C. H. Waddington, *The strategy of the genes.* (Routledge, 2014).
208. L. Kester, A. van Oudenaarden, Single-cell transcriptomics meets lineage tracing. *Cell stem cell* **23**, 166-179 (2018).
209. J. Packer, C. Trapnell, Single-cell multi-omics: an engine for new quantitative models of gene regulation. *Trends in Genetics* **34**, 653-665 (2018).
210. A. D. Warner, L. Gevirtzman, L. W. Hillier, B. Ewing, R. H. Waterston, The C. elegans embryonic transcriptome with tissue, time, and alternative splicing resolution. *Genome research* **29**, 1036-1045 (2019).
211. M. D. Young, S. Behjati, SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience* **9**, giaa151 (2020).
212. M. E. Boeck *et al.*, The time-resolved transcriptome of C. elegans. *Genome research* **26**, 1441-1450 (2016).
213. D. Yu, W. Huber, O. Vitek, Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. *Bioinformatics* **29**, 1275-1282 (2013).
214. J. Bruin, FAQ. What are pseudo-R-squareds. *UCLA: Academic Technology Services, Statistical Consulting Group. Retrieved*, (2006).
215. G. Deltas, The small-sample bias of the Gini coefficient: results and implications for empirical research. *Review of economics and statistics* **85**, 226-234 (2003).
216. K. Raivo, Pheatmap: pretty heatmaps. *R package version* **1**, (2019).
217. S. Aibar *et al.*, SCENIC: single-cell regulatory network inference and clustering. *Nature methods* **14**, 1083-1086 (2017).
218. E. M. Sommermann, K. R. Strohmaier, M. F. Maduro, J. H. Rothman, Endoderm development in Caenorhabditis elegans: the synergistic action of ELT-2 and-7 mediates the specification→ differentiation transition. *Developmental biology* **347**, 154-166 (2010).
219. S. J. Husson *et al.*, Impaired processing of FLP and NLP peptides in carboxypeptidase E (EGL‐21)‐deficient Caenorhabditis elegans as analyzed by mass spectrometry. *Journal of neurochemistry* **102**, 246-260 (2007).
220. T. R. Sarafi-Reinach, P. Sengupta, The forkhead domain gene unc-130 generates chemosensory neuron diversity in C. elegans. *Genes & Development* **14**, 2472-2485 (2000).
221. M. T. Weirauch *et al.*, Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443 (2014).
222. B. Hadland, M. Yoshimoto, Many layers of embryonic hematopoiesis: new insights into B-cell ontogeny and the origin of hematopoietic stem cells. *Exp Hematol* **60**, 1-9 (2018).
223. R. Auerbach, H. Huang, L. Lu, Hematopoietic stem cells in the mouse embryonic yolk sac. *Stem cells* **14**, 269-280 (1996).
224. S. Vanhee *et al.*, In vitro human embryonic stem cell hematopoiesis mimics MYB-independent yolk sac hematopoiesis. *Haematologica* **100**, 157-166 (2015).
225. D. K. Goode *et al.*, Dynamic Gene Regulatory Networks Drive Hematopoietic Specification and Differentiation. *Developmental cell* **36**, 572-587 (2016).
226. C. S. Baron *et al.*, Single-cell transcriptomics reveal the dynamic of haematopoietic stem cell production in the aorta. *Nature communications* **9**, 2517 (2018).

227. Y. Zeng *et al.*, Tracing the first hematopoietic stem cell generation in human embryo by single-cell RNA sequencing. *Cell Res*, (2019).
228. S. Hou *et al.*, Embryonic endothelial evolution towards first hematopoietic stem cells revealed by single-cell transcriptomic and functional analyses. *Cell Res*, (2020).
229. R. B. Lorsbach *et al.*, Role of Runx1 in adult hematopoiesis: analysis of Runx1-IRES-GFP knock-in mice reveals differential lineage expression. *Blood* **103**, 2522-2529 (2004).
230. J. A. Garcia-Porrero, I. E. Godin, F. Dieterlen-Lièvre, Potential intraembryonic hemogenic sites at pre-liver stages in the mouse. *Anat. Embryol.* **192**, 425-435 (1995).
231. T. Yokomizo, E. Dzierzak, Three-dimensional cartography of hematopoietic clusters in the vasculature of whole mouse embryos. *Development* **137**, 3651-3661 (2010).
232. N. G. dela Paz, P. A. D'Amore, Arterial versus venous endothelial cells. *Cell Tissue Res* **335**, 5-16 (2009).
233. K. R. Moon *et al.*, Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol* **37**, 1482-1492 (2019).
234. C. Lancrin *et al.*, GFI1 and GFI1B control the loss of endothelial identity of hemogenic endothelium during hematopoietic commitment. *Blood* **120**, 314-322 (2012).
235. G. La Manno *et al.*, RNA velocity of single cells. *Nature* **560**, 494-498 (2018).
236. M. Cortes *et al.*, Developmental Vitamin D Availability Impacts Hematopoietic Stem Cell Production. *Cell reports* **17**, 458-468 (2016).
237. Q. Gu *et al.*, AIBP-mediated cholesterol efflux instructs hematopoietic stem and progenitor cell fate. *Science* **363**, 1085-1088 (2019).
238. P. Li *et al.*, Epoxyeicosatrienoic acids enhance embryonic haematopoiesis and adult marrow engraftment. *Nature* **523**, 468-471 (2015).
239. T. E. North *et al.*, Prostaglandin E2 regulates vertebrate haematopoietic stem cell homeostasis. *Nature* **447**, 1007-1011 (2007).
240. A. D. Kim, D. L. Stachura, D. Traver, Cell signaling pathways involved in hematopoietic stem cell specification. *Exp Cell Res* **329**, 227-233 (2014).
241. L. Adamo *et al.*, Biomechanical forces promote embryonic haematopoiesis. *Nature* **459**, 1131-1135 (2009).
242. A. D. Yzaguirre, E. D. Howell, Y. Li, Z. Liu, N. A. Speck, Runx1 is sufficient for blood cell formation from non-hemogenic endothelial cells in vivo only during early embryogenesis. *Development* **145**, (2018).
243. R. L. Clarke *et al.*, The expression of Sox17 identifies and regulates haemogenic endothelium. *Nat Cell Biol* **15**, 502-510 (2013).
244. F. L. Bos, J. S. Hawkins, A. C. Zovein, Single-cell resolution of morphological changes in hemogenic endothelium. *Development* **142**, 2719-2724 (2015).
245. S. C. Wheatley, C. M. Isacke, P. H. Crossley, Restricted expression of the hyaluronan receptor, CD44, during postimplantation mouse embryogenesis suggests key roles in tissue formation and patterning. *Development* **119**, 295-306 (1993).
246. A. N. Schep, B. Wu, J. D. Buenrostro, W. J. Greenleaf, chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* **14**, 975-978 (2017).

247. S. Goyama *et al.*, Evi-1 is a critical regulator for hematopoietic stem cells and transformed leukemic cells. *Cell Stem Cell* **3**, 207-220 (2008).
248. I. M. Min *et al.*, The transcription factor EGR1 controls both the proliferation and localization of hematopoietic stem cells. *Cell Stem Cell* **2**, 380-391 (2008).
249. Z. Lu *et al.*, Polycomb Group Protein YY1 Is an Essential Regulator of Hematopoietic Stem Cell Quiescence. *Cell reports* **22**, 1545-1559 (2018).
250. P. Sroczynska, C. Lancrin, V. Kouskoff, G. Lacaud, The differential activities of Runx1 promoters define milestones during embryonic hematopoiesis. *Blood* **114**, 5279-5289 (2009).
251. J. Marsman, A. Thomas, M. Osato, J. M. O'Sullivan, J. A. Horsfield, A DNA Contact Map for the Mouse Runx1 Gene Identifies Novel Haematopoietic Enhancers. *Sci Rep* **7**, 13347 (2017).
252. C. Chen *et al.*, Spatial Genome Re-organization between Fetal and Adult Hematopoietic Stem Cells. *Cell reports* **29**, 4200-4211 e4207 (2019).
253. U. Blank, S. Karlsson, The role of Smad signaling in hematopoiesis and translational hematology. *Leukemia* **25**, 1379-1388 (2011).
254. S. Menegatti, M. de Kruijf, E. Garcia-Alegria, G. Lacaud, V. Kouskoff, Transcriptional control of blood cell emergence. *FEBS letters*, (2019).
255. J. Gilmour *et al.*, Robust hematopoietic specification requires the ubiquitous Sp1 and Sp3 transcription factors. *Epigenetics Chromatin* **12**, 33 (2019).
256. A. Kieusseian, P. Brunet de la Grange, O. Burlen-Defranoux, I. Godin, A. Cumano, Immature hematopoietic stem cells undergo maturation in the fetal liver. *Development* **139**, 3521-3530 (2012).
257. K. Ohmura *et al.*, Emergence of T, B, and myeloid lineage-committed as well as multipotent hemopoietic progenitors in the aorta-gonad-mesonephros region of day 10 fetuses of the mouse. *J Immunol* **163**, 4788-4795 (1999).
258. M. A. Inlay *et al.*, Identification of multipotent progenitors that emerge prior to hematopoietic stem cells in embryonic development. *Stem cell reports* **2**, 457-472 (2014).
259. Y. Li, L. Gao, B. Hadland, K. Tan, N. A. Speck, CD27 marks murine embryonic hematopoietic stem cells and type II prehematopoietic stem cells. *Blood* **130**, 372-376 (2017).
260. M. Setty *et al.*, Characterization of cell fate probabilities in single-cell data with Palantir. *Nat Biotechnol* **37**, 451-460 (2019).
261. E. C. Forsberg *et al.*, Differential expression of novel potential regulators in hematopoietic stem cells. *PLoS Genet* **1**, e28 (2005).
262. H. Iwasaki, F. Arai, Y. Kubota, M. Dahl, T. Suda, Endothelial protein C receptor-expressing hematopoietic stem cells reside in the perisinusoidal niche in fetal liver. *Blood* **116**, 544-553 (2010).
263. Z. Cai *et al.*, Haploinsufficiency of AML1/CBFA2 affects the embryonic generation of mouse hematopoietic stem cells. *Immunity* **13**, 423-431 (2000).
264. Q. Wang *et al.*, Disruption of the *Cbfa2* gene causes necrosis and hemorrhaging in the central nervous system and blocks definitive hematopoiesis. *Proc. Natl. Acad. Sci. USA* **93**, 3444-3449 (1996).
265. A. L. M. Lie *et al.*, Regulation of RUNX1 dosage is crucial for efficient blood formation from hemogenic endothelium. *Development* **145**, (2018).
266. C. O. Lizama *et al.*, Repression of arterial genes in hemogenic endothelium is sufficient for haematopoietic fate acquisition. *Nature communications* **6**, 7739 (2015).

267. D. M. Kasper, S. Nicoli, Epigenetic and Epitranscriptomic Factors Make a Mark on Hematopoietic Stem Cell Development. *Curr Stem Cell Rep* **4**, 22-32 (2018).
268. C. Eich *et al.*, In vivo single cell analysis reveals Gata2 dynamics in cells transitioning to hematopoietic fate. *J Exp Med* **215**, 233-248 (2018).
269. Y. Li *et al.*, Inflammatory signaling regulates embryonic hematopoietic stem and progenitor cell production. *Genes Dev* **28**, 2597-2612 (2014).
270. S. Sawamiphak, Z. Kontarakis, D. Y. Stainier, Interferon gamma signaling positively regulates hematopoietic stem cell emergence. *Developmental cell* **31**, 640-653 (2014).
271. C. Boiers *et al.*, Lymphomyeloid contribution of an immune-restricted progenitor emerging prior to definitive hematopoietic stem cells. *Cell Stem Cell* **13**, 535-548 (2013).
272. T. C. Luis *et al.*, Initial seeding of the embryonic thymus by immune-restricted lympho-myeloid progenitors. *Nat Immunol* **17**, 1424-1435 (2016).
273. I. Sorensen, R. H. Adams, A. Gossler, DLL1-mediated Notch activation regulates endothelial identity in mouse fetal arteries. *Blood* **113**, 5680-5688 (2009).
274. Y. Hu, G. K. Smyth, ELDA: extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. *Journal of immunological methods* **347**, 70-78 (2009).
275. J. Schindelin *et al.*, Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**, 676-682 (2012).
276. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**, 979-982 (2017).
277. J. Cao *et al.*, The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502 (2019).
278. V. Y. Kiselev *et al.*, SC3: consensus clustering of single-cell RNA-seq data. *Nat Methods* **14**, 483-486 (2017).
279. T. Stuart *et al.*, Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821 (2019).
280. A. Fabregat *et al.*, The Reactome pathway Knowledgebase. *Nucleic Acids Res* **44**, D481-487 (2016).
281. G. Yu, L. G. Wang, Y. Han, Q. Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287 (2012).
282. E. Mass *et al.*, Specification of tissue-resident macrophages during organogenesis. *Science* **353**, (2016).
283. Y. Zhang *et al.*, Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
284. W. Yu, Uzun, Y., Zhu, Q., Chen, C., Tan, K., scATAC-pro: a comprehensive workbench for single-cell chromatin accessibility sequencing data. *Genome Biol* **In press**, (2020).
285. M. E. Ritchie *et al.*, limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
286. M. T. Weirauch *et al.*, Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431-1443 (2014).
287. W. T. Nottingham *et al.*, Runx1-mediated hematopoietic stem-cell emergence is controlled by a Gata/Ets/SCL-regulated enhancer. *Blood* **110**, 4188-4197 (2007).
288. N. K. Wilson *et al.*, Gfi1 expression is controlled by five distinct regulatory regions spread over 100 kilobases, with Scl/Tal1, Gata2, PU.1, Erg, Meis1, and Runx1

acting as upstream regulators in early hematopoietic cells. *Mol Cell Biol* **30**, 3853-3863 (2010).

289. J. Zhang, P. A. Ney, Role of BNIP3 and NIX in cell death, autophagy, and mitophagy. *Cell Death Differ* **16**, 939-946 (2009).
290. D. van Dijk *et al.*, Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**, 716-729 e727 (2018).
291. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
292. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 1-21 (2014).
293. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data. *Nature biotechnology* **33**, 495-502 (2015).
294. P. V. Kharchenko, L. Silberstein, D. T. Scadden, Bayesian approach to single-cell differential expression analysis. *Nature methods* **11**, 740-742 (2014).
295. K. Hornik, The comprehensive R archive network. *Wiley interdisciplinary reviews: Computational statistics* **4**, 394-398 (2012).
296. W. Huber *et al.*, Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods* **12**, 115 (2015).
297. S. Anders *et al.*, Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature protocols* **8**, 1765 (2013).
298. J. Goecks, A. Nekrutenko, J. Taylor, Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology* **11**, 1-13 (2010).
299. F. Russo, C. Angelini, RNASeqGUI: a GUI for analysing RNA-Seq data. *Bioinformatics* **30**, 2514-2516 (2014).
300. J. W. Nelson, J. Sklenar, A. P. Barnes, J. Minnier, The START App: a web-based RNAseq analysis and visualization resource. *Bioinformatics* **33**, 447-449 (2017).
301. V. Gardeux, F. P. David, A. Shajkofci, P. C. Schwalie, B. Deplancke, ASAP: a web-based platform for the analysis and interactive visualization of single-cell RNA-seq data. *Bioinformatics* **33**, 3123-3125 (2017).
302. Y. Li, J. Andrade, DEApp: an interactive web interface for differential expression analysis of next generation sequence data. *Source code for biology and medicine* **12**, 1-4 (2017).
303. W. Chang *et al.*, shiny: Web Application Framework for R. *CRAN*, (2021).
304. S. Anders, P. T. Pyl, W. Huber, HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).
305. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).
306. S. Anders, W. Huber, Differential expression analysis for sequence count data. *Nature Precedings*, 1-1 (2010).
307. M. D. Robinson, A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology* **11**, 1-9 (2010).
308. M.-A. Dillies *et al.*, A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics* **14**, 671-683 (2013).

309. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**, 621-628 (2008).
310. D. Risso, J. Ngai, T. P. Speed, S. Dudoit, Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature biotechnology* **32**, 896-902 (2014).
311. A. Lemire *et al.*, Development of ERCC RNA spike-in control mixes. *Journal of biomolecular techniques: JBT* **22**, S46 (2011).
312. J. M. Spaethling *et al.*, Primary cell culture of live neurosurgically resected aged adult human brain cells and single cell transcriptomics. *Cell reports* **18**, 791-803 (2017).
313. Y. Xie, J. J. Allaire, G. Grolemund, *R markdown: The definitive guide*. (CRC Press, 2018).
314. B. Almende, B. Thieurmel, T. Robert. (CRAN, 2016).
315. D. M. Witten, R. Tibshirani, Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 753-772 (2011).
316. P. Pons, M. Latapy, in *International symposium on computer and information sciences*. (Springer, 2005), pp. 284-293.
317. L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, F. J. Theis, Diffusion pseudotime robustly reconstructs lineage branching. *Nature methods* **13**, 845 (2016).
318. O. J. Rackham *et al.*, A predictive computational framework for direct reprogramming between human cell types. *Nature genetics* **48**, 331 (2016).
319. D. Szklarczyk *et al.*, STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research* **43**, D447-D452 (2015).
320. Z.-P. Liu, C. Wu, H. Miao, H. Wu, RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* **2015**, (2015).
321. M. Lizio *et al.*, Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome biology* **16**, 1-14 (2015).
322. D. Huangfu *et al.*, Induction of pluripotent stem cells from primary human fibroblasts with only Oct4 and Sox2. *Nature biotechnology* **26**, 1269-1275 (2008).
323. T. Galili, A. O'Callaghan, J. Sidi, C. Sievert, heatmaply: an R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics* **34**, 1600-1602 (2018).
324. C. Sievert, *Interactive web-based data visualization with R, plotly, and shiny*. (CRC Press, 2020).
325. V. Q. Vu, ggbiplot: A ggplot2 based biplot. *R package* **342**, (2011).
326. B. Lewis, threejs: Interactive 3D Scatter Plots, Networks and Globes. *R package version 0.3* **1**, (2017).
327. G. Csardi, T. Nepusz, The igraph software package for complex network research. *InterJournal, complex systems* **1695**, 1-9 (2006).
328. J. Allaire *et al.*, Package 'networkD3'. *D3 JavaScript Network Graphs from R*, (2017).
329. L. TT. (CRAN, 2016).
330. D. Bates, M. Maechler, Matrix: sparse and dense matrix classes and methods. *R package version 0.999375-43, URL http://cran.r-project. org/package= Matrix*, (2010).

331. Q. Zhu *et al.*, Developmental trajectory of prehematopoietic stem cell formation from endothelium. *Blood, The Journal of the American Society of Hematology* **136**, 845-856 (2020).

332. Z. Miao *et al.*, Single cell regulatory landscape of the mouse kidney highlights cellular differentiation programs and disease targets. *Nature communications* **12**, 1-17 (2021).

333. A. Graham, Developmental homoplasy: convergence in cellular differentiation. *Journal of anatomy* **216**, 651-655 (2010).

334. M. B. Gerstein *et al.*, Comparative analysis of the transcriptome across distant species. *Nature* **512**, 445-448 (2014).

335. D. Duboule, Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Development* **1994**, 135-142 (1994).

336. N. Irie, S. Kuratani, The developmental hourglass model: a predictor of the basic body plan? *Development* **141**, 4649-4655 (2014).

337. C. Bouchard, P. Staller, M. Eilers, Control of cell proliferation by Myc. *Trends in cell biology* **8**, 202-206 (1998).

338. A. Batsivari *et al.*, Understanding hematopoietic stem cell development through functional correlation of their proliferative status with the intra-aortic cluster architecture. *Stem Cell Reports* **8**, 1549-1562 (2017).

339. K. Kataoka *et al.*, Evi1 is essential for hematopoietic stem cell self-renewal, and its expression marks hematopoietic cells with long-term multilineage repopulating activity. *Journal of Experimental Medicine* **208**, 2403-2416 (2011).

340. Y. Zhang *et al.*, PR-domain–containing Mds1-Evi1 is critical for long-term hematopoietic stem cell function. *Blood, The Journal of the American Society of Hematology* **118**, 3853-3861 (2011).

341. S. Grigorescu, B. Trasnea, T. Cocias, G. Macesanu, A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics* **37**, 362-386 (2020).

342. R. Shwartz-Ziv, N. Tishby, Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, (2017).

343. S. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, (2017).

344. J. He *et al.*, Organoid technology for tissue engineering. *Journal of molecular cell biology* **12**, 569-579 (2020).

345. W. J. Yun, S. Yi, J. Kim, Multi-agent deep reinforcement learning using attentive graph neural architectures for real-time strategy games. *arXiv preprint arXiv:2105.10211*, (2021).

346. H. X. Pham, H. M. La, D. Feil-Seifer, A. Nefian, Cooperative and distributed reinforcement learning of drones for field coverage. *arXiv preprint arXiv:1803.07250*, (2018).