A molecular atlas of C. elegans development at single-cell and single-lineage resolution

Jonathan Suresh Packer

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

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Cole Trapnell, Chair

Robert Waterston, Chair

Stephen Tapscott

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

A molecular atlas of C. elegans development at single-cell and single-lineage resolution

Jonathan Suresh Packer

Co-Chairs of the Supervisory Committee:
Associate Professor Cole Trapnell, Genome Sciences
Professor Robert Waterston, Genome Sciences

It takes many cell divisions to produce a complex, multicellular organism such as a human being. Every one of the trillions of cells in a human body was produced by the division of a parent cell, which in turn was produced by the division of its parent; and if one follows this lineage far back enough, one will reach the single zygote cell that is the common progenitor of all of the cells in the body. Collectively, the pattern of cell divisions that produce an organism is called its cell lineage. As cells divide in a developing organism, they also differentiate into specialized cell types. What cell type a cell will adopt—its "cell fate"—is restricted by its lineage. A cell that descends from an endoderm progenitor, for example, may differentiate into a liver cell or an intestine cell, but will not become a bone cell. This general principle, that different parts of an organism's cell lineage have different developmental potentials, has been known since the early 1800s. But our understanding of the molecular mechanisms that connect the cell lineage to the process of cell differentiation remains incomplete.

In this dissertation, I present a near-comprehensive atlas of gene expression in the embryonic cell lineage of the nematode *Caenorhabditis elegans*, the only animal for which the cell lineage is fully known. I describe the methods used to assemble this atlas from single cell RNA-seq data, which required finding the precise lineage identity of each assayed cell. Using the

atlas, I investigate the molecular mechanisms of cell fate commitment, finding that:

1. Multilineage priming is strikingly prevalent, contributing to the differentiation of over half of the cells in the lineage.

2. Distinct lineages that produce the same anatomical cell type tend to converge to a homogenous transcriptional state. This convergence is gradual for lineages that commit to their cell fate early in development, but can be abrupt for lineages that commit late.

3. A cell's lineage and its transcriptome are correlated, but this correlation is transient, peaking in late gastrulation and falling dramatically during terminal differentiation.

4. Developmental trajectories reconstructed from single cell RNA-seq data often do not accurately reflect the cell lineage, in large part due to the transcriptional convergence of similarly fated lineages. Supplementary datasets, e.g. from fluorescent reporter imaging, are necessary to accurately place single cell RNA-seq data in the context of the cell lineage.

This work provides an extensive resource to the *C. elegans* research community and an outline of the challenges that will need to be overcome in future studies of vertebrate cell lineages by single cell RNA-seq.

Table of Contents

List of Figures

List of Tables

Supplementary tables for this manuscript are provided as separate files, available online.

Supplementary tables for Chapter 2 are available at:

http://www.sciencemag.org/content/357/6352/661/suppl/DC1

Table S1. Summary of experiments.

Table S2. Summary statistics for cell type consensus expression profiles constructed in this study.

Table S3. Tissue-level consensus expression profiles.

Table S4. Cell type consensus expression profiles.

Table S5. Neuron cluster consensus expression profiles.

Table S6. Differential expression test results for the identification of tissue-enriched genes.

Table S7. Differential expression test results for the identification of cell type enriched genes.

Table S8. Differential expression test results for the identification of neuron cluster enriched genes.

Table S9. Differential expression test results for anterior vs. posterior body wall muscle.

Table S10. Differential expression test results for posterior vs. other intestine.

Table S11. Differential expression test results for amphid vs. phasmid sheath cells.

Table S12. Differential expression test results for the ASEL vs. ASER neuron.

Table S13. Differential expression test results for AWA vs. ASG neurons.

Table S14. List of genes used in heatmaps in Fig. 3F and Fig. 4C.

Supplementary tables for Chapter 3 are available at:

https://science.sciencemag.org/content/suppl/2019/09/04/science.aax1971.DC1

Table S1. Marker genes for terminal cell type annotations.

Table S2. Terminal cell type annotation statistics.

Table S3. List of EPiC movies used for lineage annotations.

Table S4. Marker genes for lineage annotations.

Table S5. Pre-terminal lineage annotation statistics.

Table S6. Map of *C. elegans* anatomical cells to annotations defined in this study.

Table S7. Gene expression profiles for terminal cell types.

Table S8. Gene expression profiles for annotated cell lineages.

Table S9. Differentially expressed genes between all pairs of sister lineages.

Table S10. Results from differential gene expression tests between all pairs of sister lineages.

Table S11. Results from differential gene expression tests between all pairs of parent vs. daughter lineages.

Table S12. Marker genes for cell types at the L2 stage.

Table S13. Annotation statistics for cell types at the L2 stage.

Table S14. Gene expression profiles for cell types at the L2 stage.

Table S15. Candidate terminal selectors of neuron types identified by differential gene expression analysis.

Table S16. Body wall muscle cells associated with each position-related group in Fig. 4C and 4D.

Acknowledgements

This work wasn't exactly what I set out to do when I came to graduate school, but it happened, and it turned out to be pretty cool in the end. I'm deeply grateful to all of the people who supported me along the way. Specifically, I'd like to thank:

- My advisors, Bob Waterston and Cole Trapnell.
    - Bob: thanks for getting me started on this project and setting me on this path. In a world where a lot of things are a mess, the orderliness and elegance the worm's development is a refreshing contrast. Except for that one AB lineage muscle cell. What's up with that?

    - Cole: as someone who is pessimistic and tends to overthink things, I really appreciate your attitude of "let's just try it and see what happens." Thanks for your guidance and for keeping my spirits up.

- My collaborators, especially Junyue Cao, Chau Hunyh, Qin Zhu, John Murray, and Junhyong Kim.
    - Junyue: I've always been in awe of you as an experimentalist, a scientist, and as a human being. I'm looking forward to reading about all of the cool projects you'll have to show us in the coming years.
    - Chau: thanks for prepping all the worms!
    - Qin: thanks for all your help on this project. It was a ton of work, but we did it!
    - John: thanks for your incredible dedication and thoroughness working on this project. Your lineage annotations were a heroic achievement. And even though I grumbled at the time, including every last cell we could find made the end-result that much more impactful and satisfying.
    - Junhyong: thanks for all of your advice in drafting the paper. I think your carefulness with language will help it better stand the test of time.

- My thesis committee members, Daniella Witten, Raphael Gottardo, and Stephen Tapscott; and my past scientific advisors, Jeffrey Reid and Yufeng Shen.

- Sanjay Srivatsan. You've been a true friend to me and I've learned so much from you. Your enthusiasm for science makes me feel like I've been doing something worthwhile for the last four years. Thank you so much, for everything.

Chapter 1: Introduction

The cell lineage is critical context for the study of gene regulation

One of the fundamental principles of biology is that cells are produced by the division of other cells. Every one of the trillions of cells in a human being was produced by the binary division of a progenitor cell, its "parent". Each progenitor cell itself was produced by the division of an earlier progenitor; and if one traces this lineage further, to earlier and earlier stages of development, one eventually reaches the single zygote cell from which all of the cells in the person ultimately originate. Collectively, the sequence of cell divisions that produce an organism, starting from the zygote, is referred to as its cell lineage.

As cells in a developing organism divide, the same DNA sequence is maintained in every daughter cell[1]; but the cells must also differentiate and adopt specialized functions. To accomplish this, cells express different subsets of the genes encoded in their DNA. Over time, differences in gene expression result in differences in protein production, which ultimately allow cells to differentiate into specialized cell types.

How does a cell "know" what genes it should express, and therefore what cell type it should adopt? When considering this question, one runs into a chicken-and-egg problem. The genes that a cell will express are determined by what regulatory proteins are currently present in the cell (as well as what signals it is receiving from its environment). But the current protein content of a cell is a result of what genes it expressed in the past; which is a result of what proteins it contained in the past; which is a result of what genes it expressed in the even more distant past; and so on. Furthermore, this simple model for how a cell's state depends on its past states is applicable across cell division events, as when a cell divides, each daughter cell inherits a portion of its parent's protein and RNA content. Thus, in order to understand why a cell expresses the genes that it does at one stage of development, one must first know what genes were expressed in each of its ancestors at previous stages of development. In other words, one must know the transcriptional history of the organism's cell lineage.

In this dissertation, I describe the results of a project to produce a transcriptional atlas of the cell lineage of the nematode *Caenorhabditis elegans*; how this atlas has provided new insights into mechanisms of gene regulation; and the implications of these findings for the study of vertebrate development. In this introduction, I will review the motivations for studying *C. elegans* and discuss previous studies of gene expression in this organism.

---

[1] As with most things in biology, exceptions to this "rule" do exist. For example, in an eclectic variety of species, including parasitic nematodes, hagfish, and bandicoots, somatic cell lineages eliminate large chunks of DNA that is maintained in the germline (*1*, *2*). Another example of a programmed rearrangement of the genome is V(D)J recombination in B and T cells (*3*).

C. elegans is the only animal with a fully known cell lineage

Nematodes have been studied as models of development since the late 1800s. Early studies found that every embryo of the species *Parascaris equorum*, a parasitic nematode that infects horses, developed through essentially the same cell lineage. The *P. equorum* lineage was mapped, comprehensively for the early embryo (*4*) and in broad strokes for later stages (*5*). Strikingly, the early cell lineages of other nematode species, such as *Caenorhabditis nigoni* (*6*) and *Metastrongylus elongatus* (*7*), were found to be almost identical to that of *P. equorum*. This conservation, together with the near-perfect reproducibility of *P. equorum* embryogenesis, suggested that the embryonic cell lineage of *P. equorum* was invariant between individuals.

Caenorhabditis elegans is a small, transparent nematode that develops rapidly and can be cultured on agar with a diet of bacteria. Like *P. equorum*, its cell lineage is invariant. Its cell number is limited, with an adult worm containing only 959 somatic cells. These properties led Ellsworth Dougherty to propose *C. elegans* as a model organism for genetic analysis in 1948 (*8*). Sydney Brenner, intrigued by the simplicity and structural consistency of its nervous system, obtained samples from Dougherty in 1963 and also advocated for its use as a model organism. In 1974, he published a landmark paper that characterized ethyl methanesulfonate induced mutants (*9*), drawing additional attention to *C. elegans* from the biological community. Subsequently, the cell lineage was mapped (*10–12*), a herculean task that was completed by Sulston *et al.* in 1983 (*12*); the neural connectome was mapped, another herculean task completed by White *et al.* in 1986 (*13*); and the genome was sequenced in 1998 (*14*).

To date, *C. elegans* is the only animal for which the cell lineage is fully known. This, in theory, allows for the process of cell differentiation to be studied "end-to-end", following the lineage of any particular cell all the way back to the first cell division. It allows cell fate decisions to be studied at the resolution of individual cell division events that are discrete and reproducible. Lastly, it allows *C. elegans* to serve as a ground truth for experimental and computational methods that seek to reconstruct the cell lineage of an organism.

Measuring gene expression in *C. elegans* with fluorescent reporters

Gene expression in *C. elegans* development has been studied extensively using fluorescent reporter assays. These assays use the promoter of a gene of interest to express either a fluorescent protein (a "promoter fusion"), or a fluorescent protein fused to the native protein (a "protein fusion") (*15*). Since *C. elegans* is a transparent animal, these reporter assays (in theory) allow one to identify the specific set of cells that express any given gene at any given stage of development. The *C. elegans* literature includes hundreds of such assays. Their results, i.e. which genes were observed to be expressed in which cells, have been compiled in an extensive online resource called WormBase (*16*).

Fluorescent reporter assays have several limitations.

1. Promoter fusion constructs will lack any *cis*-regulatory elements that exist in the native gene's introns or in distal intergenic sequence, potentially leading to incorrect expression patterns. Some protein fusion constructs include intronic sequence, but usually not distal intergenic sequence. This is not as severe of a problem as it would be in a vertebrate context, since the *C. elegans* genome is very compact, and most regulatory elements are within 1-2 kb of a gene's transcription start site. But many older studies do not include enough promoter sequence in their reporter constructs, e.g. using only 200-500 bp.

2. Many native proteins are degraded much faster than fluorescent proteins like GFP or mCherry. A promoter fusion reporter can therefore make it seem like a protein is continuously expressed when in fact the native gene and/or protein was only present for a limited period of time in development.

3. Fluorescent proteins take a significant amount of time to fold, causing observed protein expression to lag actual protein expression by ~30-45 minutes. This delay can be qualitatively significant in the rapidly developing *C. elegans* embryo.

4. Most studies do not quantify the fluorescent signal coming from their reporters. Furthermore, fluorescent signal can become saturated, making it hard to distinguish moderate expression from high expression.

5. The *C. elegans* intestine contains autofluorescent granules that can mask fluorescent signal from reporters of lowly-expressed genes (*17*).

6. Generating a line of worms that has a reporter construct stably integrated into the genome is a difficult and laborious process. Thus, reporter constructs are often expressed from extrachromosomal arrays. These arrays are heritable, but are unstable with respect to mitotic cell divisions: the copy number of the array will vary from cell to cell, and some cells will completely silence the array (*18*).

7. A well-designed fluorescent reporter construct will light up the cells that express a gene of interest, but one still has to determine the identity of those cells. In most studies, this is done simply by manual inspection, which can be a difficult and error-prone process. Additionally, many studies only examine cells that are of immediate interest, and fail to report expression observed in other tissues.

Murray *et al.* used methods that overcame several of these limitations when they created EPiC (*19*), a database that consists of 3D movies of fluorescent reporter constructs from over 200 different transcription factors. These movies show gene and protein expression in *C. elegans* embryos during a time period spanning from early to mid gastrulation.

Murray *et al.*'s reporter constructs were a mix of promoter and protein fusions. The promoter fusions included up to 5 kb of sequence upstream of the gene's transcription start site, while the protein fusions included all intragenic sequence and even larger stretches of upstream and downstream sequence. Thus, the constructs were likely to have included most of the relevant regulatory elements for each gene. The constructs were expressed from stably integrated, low copy number arrays. Expression from these reporters was quantified in each cell in the early *C. elegans* lineage using (mostly-)automated imaging processing software. This software had trouble distinguishing cells after the penultimate and final rounds of cell division in the embryo, so the gene expression profiles stop short of when many interesting cell fate decisions are made. Nevertheless, the uniform processing and single cell resolution of the data made EPiC a valuable resource that I used extensively in my work.

Measuring gene expression in *C. elegans* using bulk assays

Gene expression in *C. elegans* has also been profiled using microarrays and bulk RNA sequencing. The most notable datasets that have been published are time series that profile gene expression in whole *C. elegans* embryos, larva, and adults at different stages in development using bulk RNA-seq (*20*, *21*). Since they profile whole embryos/animals, not individual cells or cell types, these datasets on their own are of limited utility for studying gene regulation; but they are useful for getting a rough sense of when in development a gene is expressed.

A few studies have isolated specific *C. elegans* cell types using fluorescence activated cell sorting (FACS) and a reporter construct to sort for cells that express a marker gene than is known to be specific to the cell type of interest. Gene expression in the recovered cells is then profiled using microarrays (*22*) or bulk RNA-seq (*23*, *24*). This approach provided some insights into cell-type specific gene regulation, but has been limited by the need to create a reporter construct and perform a separate experiment for every cell type one wishes to profile. It also requires one to have a marker gene that is completely specific to the cell type of interest. For many cell types, there are no known marker genes with complete specificity. In some cases, one can achieve specificity by sorting cells that co-express two marker genes, but making worm lines that express multiple reporter constructs can be a hassle.

Measuring gene expression in *C. elegans* using single cell RNA-seq

Single cell RNA sequencing ("sc-RNA-seq") refers to a family of experimental methods that aim to quantify the abundance of mRNA molecules independently in individual cells from a

biological sample. From this data, one can retrospectively determine the cell type of each cell in the sample, making sc-RNA-seq an ideal assay for exploring gene expression in complex tissues, organs, or even whole organisms.

Prior to the work described in this dissertation, only two studies had applied sc-RNA-seq to *C. elegans*. In 2012, Hashimshony *et al.* (*25*) used manual dissection to isolate selected single cells from early *C. elegans* embryos at the 1- to 8-cell stages. They validated that sc-RNA-seq recovered distinct transcriptional profiles for different cells in the early *C. elegans* lineage, and that these profiles were consistent with results from fluorescent reporters.

In 2016, Tintori *et al.* (*26*) performed a deeper analysis of the early *C. elegans* embryo. Like Hashimshony *et al. (25)*, they used manual dissection to isolate single cells; however, their dataset was more comprehensive, covering every cell in the early *C. elegans* lineage up to the 16-cell stage. They determined the lineage identity of each cell in their dataset using marker genes from the literature and single molecule imaging assays.

The main conclusion of Tintori *et al. (26)* was that while most cell lineages were distinguished by the expression of many different genes, cells within the AB lineage (one of two lineages that produce ectoderm in *C. elegans*) were distinguished by the expression of only a handful of genes. Nevertheless, cells within the AB lineage are known to have different developmental potentials (i.e. they produce different sets of cell types), even at the 16-cell stage. Thus, the results of Tintori *et al.* (*26*) highlight the fact that cells that are distinct from a developmental perspective can have similar global transcriptomic profiles.

Both Hashimshony *et al. (25)* and Tintori *et al.* (*26*) used protocols that required them to manually dissect each embryo and place each single cell into its own reaction tube. This laborious approach limited them to examining only the early embryo, where there are very few cells and each cell is large enough to be manually manipulated. Fortunately, in recent years, sc-RNA-seq technologies have significantly improved. Droplet-based protocols such as Drop-seq and 10X Chromium use microfluidics to automate the isolation of single cells (*27*, *28*). The sci-RNA-seq protocol, described in Chapter 2, uses a clever approach that avoids the need to isolate single cells entirely (*29*, *30*). These protocols make it possible to profile thousands or even millions of cells at low cost, sufficient to cover any stage of *C. elegans* development.

The input material to each of the high-throughput sc-RNA-seq protocols is a suspension of single cells (i.e. the cells can be in the same tube, but they should not be attached to each other). To generate this suspension, one must use enzymatic digestion and/or mechanical force to dissociate one's biological sample. The dissociation procedure usually needs to be optimized for the specific type of sample one is working with, and even when optimized, requires deft hands. My work was made possible by Chau Hunyh, a scientist in the Waterston lab here at the University of Washington, who developed protocols for dissociating *C. elegans* embryos and larvae into single cell suspensions (described in the methods sections of Chapters 2 and 3).

Outline of this dissertation

In the following chapters, I describe my efforts to use data from large-scale sc-RNA-seq experiments to assemble a transcriptional atlas of the *C. elegans* cell lineage.

- In Chapter 2, I describe a proof-of-concept study in which I demonstrate that "shotgun" single cell RNA sequencing of whole *C. elegans* larva provides data that, with careful analysis, is sufficient to reconstruct the gene expression profiles of specific cells in the *C. elegans* anatomy, even down to the resolution of single anatomical cells.

- In Chapter 3, I describe my main project, generating a transcriptional atlas of the *C. elegans* cell lineage. Using sc-RNA-seq data from *C. elegans* embryos, I reconstruct gene expression trajectories for nearly every cell in the organism, spanning multiple developmental stages from gastrulation to terminal differentiation. Using this atlas, I investigate the transcriptional dynamics associated with individual cell division events in the lineage, with a focus on those that produce two daughter cells of different types, a simple model of a cell fate decision. I also construct statistical models that show that the extent to which gene expression differs between two cells correlates with the cells' distance from each other in the lineage tree.

- In Chapter 4, I present closing remarks, discussing how this work should inform future studies of both *C. elegans* and vertebrate development.

This work spans two fields of research that until now have not substantially overlapped: the study of *C. elegans* biology, and the development and use of computational methods for sc-RNA-seq data analysis. For readers who are not familiar with these fields, I have written four appendices.

- Appendix A provides a glossary of terms related to *C. elegans* anatomy.
- Appendix B describes the nomenclature used in the study of the *C. elegans* cell lineage.
- Appendix C describes the methods used in a "typical" workflow for analyzing sc-RNA-seq data.
- Appendix D discusses challenges associated the quality control of sc-RNA-seq data.

Chapter 2: Comprehensive single-cell transcriptional profiling of a multicellular organism

Introductory note

The main text of this chapter is adapted with minimal modification from the publication:

> Cao, Packer *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* 357, 661-667 (2017).
> doi: 10.1126/science.aam8940.

This project was a collaboration with Junyue Cao, a member of Jay Shendure's lab. Junyue was developing a new single cell RNA sequencing protocol called sci-RNA-seq, that allows one to assay large numbers of cells at low cost. Junyue was seeking to publish this protocol as soon as possible, and needed a compelling biological system with which to demonstrate its potential. At the time, I had been analyzing data from small-scale single cell RNA-seq profiling of *C. elegans* embryos. My initial results were promising, so Junyue decided to demonstrate sci-RNA-seq using a large-scale *C. elegans* experiment. In this experiment, Junyue and Chau Hunyh, a scientist in Bob Waterston's lab, dissociated whole worms at the L2 larval stage into single cells, and then profiled their transcriptomes using sci-RNA-seq. For this proof-of-concept experiment, we chose to profile larva instead of embryos, hoping that the cell types of differentiated larval cells would be easier to identify than those of embryonic cells.

My role in this project was to analyze the resulting data and demonstrate that this whole-organism, "shotgun" sc-RNA-seq approach could reconstruct gene expression profiles for all of the major cell types in the worm, as well as many rare cell types. In some cases, I found clusters of cells that corresponded to individual anatomical cells in the worm.

For unclear reasons, sci-RNA-seq did not work as well when applied to *C. elegans* embryos (rather than larva), so for future projects we used 10X Chromium, a different sc-RNA-seq technology. Still, the larval stage data was a useful point of comparison in my later analysis of *C. elegans* embryogenesis. I also learned several lessons about sc-RNA-seq data analysis from this project:

1. Computational methods that aim to find clusters of related cells in sc-RNA-seq data, such as the t-SNE algorithm, often fail to resolve closely related cell types if there are a large number of very distinct cell types in the dataset. To get around this, one can re-run the analysis for specific subsets of cells, e.g. just cells that were identified as neurons. In this experiment, reprocessing just the neurons identified many more neuronal cell types than the global analysis of all of the cells.

2. Many cells in the *C. elegans* anatomy are not uniquely identified by any single known marker gene. Identifying which clusters of cells in a sc-RNA-seq experiment correspond to which cell types can require one to carefully examine overlaps between the expression patterns of several marker genes from the literature, and be clever in one's usage of process-of-elimination.

3. There are two major technical issues that can confound sc-RNA-seq data analysis: doublets (instances where sequence reads from two cells receive the same barcode, making them look like one cell in the data), and background contamination (leakage of RNA from damaged cells that contaminates the whole sample). Both issues need to be addressed in order to accurately reconstruct the gene expression profiles of each cell type in an experiment. I discuss these problems in more detail in **Appendix D**.

   For the project described in this chapter, we were rushing to publish and I didn't have time to fully address these issues: while I did do some manual filtering of doublets, I did not do any correction for background contamination. I later developed methods to address these problems for the project described in Chapter 3, and also used them to clean up the data from this project (also described in Chapter 3).

Abstract

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

Main text

Individual cells are the natural unit of form and function in biological systems. However, conventional methods for profiling the molecular content of biological samples mask cellular heterogeneity, likely present even in ostensibly homogenous tissues (*31*). Recently, profiling the transcriptome of individual cells has emerged as a powerful strategy for resolving such heterogeneity. The expression levels of mRNA species are linked to cellular function, and therefore can be used to classify cell types (*28*, *32–39*) and to order cell states (*40*). Although methods for single cell RNA-seq have proliferated, they rely on the isolation of individual cells within physical compartments (*27*, *28*, *41–47*). Consequently, preparing single cell RNA-seq libraries with these methods can be expensive, the cost scaling linearly with the numbers of cells processed (*48*, *49*).

We recently developed combinatorial indexing, a method using split-pool barcoding of nucleic acids to uniquely label a large number of single molecules or single cells. Single-molecule combinatorial indexing can be used for haplotype-resolved genome sequencing and *de novo* genome assembly (*50*, *51*), while single-cell combinatorial indexing ("sci") can be used to profile chromatin accessibility (sci-ATAC-seq) (*52*), genome sequence (sci-DNA-seq) (*53*), genome-wide chromosome conformation (sci-Hi-C) (*54*), and DNA methylation (sci-MET) (*55*) in large numbers of single cells.

In this work, we developed a combinatorial indexing method to uniquely label the transcriptomes of large numbers of single cells or nuclei (sci-RNA-seq). We then applied sci-RNA-seq to deeply profile single cell transcriptomes in the nematode *C. elegans* at the L2 stage. *C. elegans* is the only multicellular organism for which all cells and cell types are defined, as is its entire developmental lineage (*10*, *12*). However, despite its modest cell count (*e.g.* 762 somatic cells per L2 larva), our knowledge of the molecular state of each cell and cell type remains fragmentary. We therefore saw an opportunity to generate a powerful resource for nematode biologists as well as for the single cell genomics community.

Overview of sci-RNA-seq

In its current form, sci-RNA-seq relies on the following steps (**Fig. 1A**): (1) Cells are fixed and permeabilized with methanol (alternatively, cells are lysed and nuclei recovered), and then split across 96- or 384-well plates. (2) A first molecular index is introduced to the mRNA of cells within each well with *in situ* reverse transcription (RT) incorporating a barcode-bearing, well-specific polyT primer containing unique molecular identifiers (UMI). (3) All cells are pooled and redistributed by fluorescence activated cell sorting (FACS) to 96- or 384-well plates in limiting numbers (*e.g.* 10-100 per well). Cells are gated on the basis of DAPI (4',6-diamidino-2-phenylindole) staining to discriminate single cells from doublets during sorting. (4) Second strand synthesis, transposition with Tn5 transposase, lysis, and PCR

amplification are performed. The PCR primers target the barcoded poly(T) primer on one end, and the Tn5 adaptor insertion on the other end, such that resulting PCR amplicons preferentially capture the 3' ends of transcripts. These primers introduce a second barcode, specific to each well of the PCR plate. (5) Amplicons are pooled and subjected to massively parallel sequencing, resulting in 3'-tag digital gene expression profiles, with each read associated with two barcodes corresponding to the first and second rounds of cellular indexing (**Fig. 1B**). In a variant of the method described below, we introduce a third round of cellular indexing during Tn5 transposition of double-stranded cDNA.

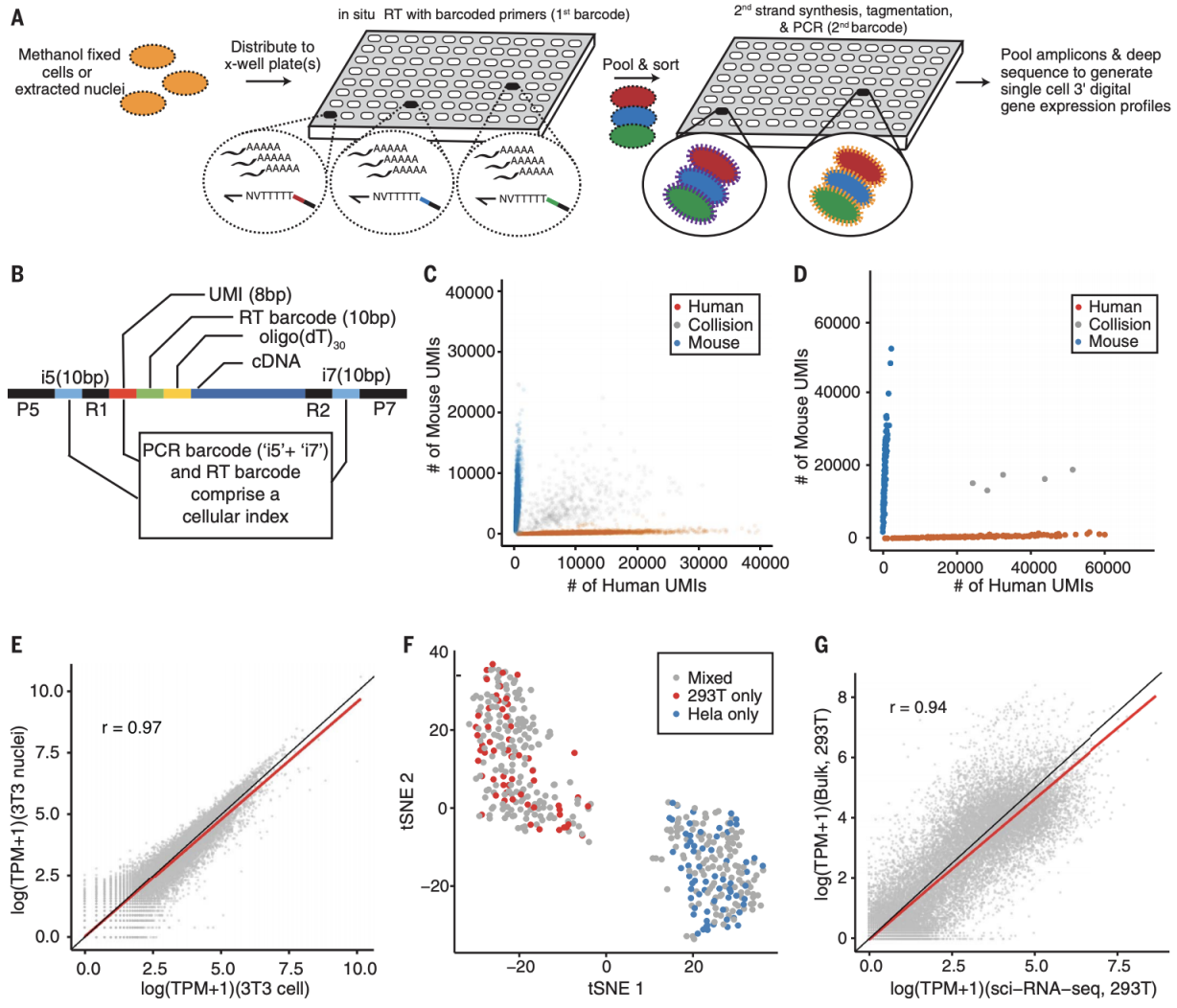


Fig. 1. sci-RNA-seq enables multiplex single-cell transcriptome profiling. (**A**) Schematic of the sci-RNA-seq workflow. AAAAA, polyadenosine tail; NVTTTTT, polythymidine primer. (**B**) Schematic of sci-RNA-seq library amplicons for Illumina sequencing. bp, base pairs; R, annealing sites for Illumina sequencing primers; P, Illumina P5 or P7 adaptor sequence. (**C**) Scatter plot of unique molecular identifier (UMI) counts from human and mouse cells, determined by 384 × 384 sci-RNA-seq. Blue, inferred mouse cells (n = 5953). Red, inferred human cells (n = 3967). Gray, collisions (n = 884). (*Legend continued on the following page*)

Fig. 1 (continued). (**D**) Scatter plot of UMI counts from human and mouse cells, determined by 96 × 96 sci-RNA-seq with an optimized protocol. Blue, inferred mouse cells (n = 129). Red, inferred human cells (n = 160). Gray, collisions (n = 5). In (C) and (D), only cells originating from wells containing mixed human and mouse cells are shown. (**E**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of NIH/3T3 cells (n = 238) and nuclei (n = 124). (**F**) t-SNE plot of cells originating in wells containing HEK293T (red; n = 60), HeLa S3 (blue; n = 69), or a mixture (gray; n = 321). (**G**) Correlation between gene expression measurements from aggregated sci-RNA-seq data and bulk RNA-seq data obtained using a related protocol (29). In (E) and (G), the red line is the linear regression, and the black line is y = x.

Most cells pass through a unique combination of wells, resulting in a unique combination of barcodes for each cell that tags its transcripts. The rate of two or more cells receiving the same combination of barcodes can be tuned by adjusting how many cells are distributed to the second set of wells (*52*). Increasing the number of barcodes used during each round of indexing leads boosts the number of cells that can be profiled while reducing the effective cost per cell (**Fig. S1**). Additional levels of indexing can potentially offer even greater complexity and lower costs. Multiple samples (*e.g.* different cell populations, tissues, individuals, time-points, perturbations, replicates, etc.) can be concurrently processed within one experiment, using different subsets of wells for each sample during the first round of indexing.

Scalability of sci-RNA-seq

We tested 262 sci-RNA-seq conditions with mammalian cells, optimizing the protocol and reaction conditions. We demonstrate scalability with 384 x 384 well sci-RNA-seq. During the first round of indexing, half of 384 wells contained pure populations of either human (HEK293T or HeLa S3) or mouse (NIH/3T3) cells, and the other half mixed human and mouse cells (**Table S1**). After barcoded RT, cells were pooled and then sorted to a new 384 well plate for the second round of barcoding and deep sequencing of pooled PCR amplicons. We recovered 15,997 single cell transcriptomes and readily assigned cells as human or mouse (**Fig. 1C**).

Optimization of sci-RNA-seq and application to nuclei

We performed optimized 96 x 96 well sci-RNA-seq on five cell populations, each present in distinct subsets of wells during the first round of barcoding (**Table S1**): HEK293T cells (8 wells); HeLa S3 cells (8 wells); an intraspecies mixture of HEK293T and HeLa S3 cells (32 wells); and interspecies mixtures of HEK293T and NIH/3T3 cells (24 wells) or nuclei (24 wells). We deeply sequenced the resulting library (~250,000 reads per cell; ~210,000 reads per nucleus; ~88% duplication rate), profiling 744 single cell and 175 single nucleus transcriptomes.

Transcriptomes in the 24 wells containing an interspecies mixture of human and mouse cells overwhelmingly mapped to the genome of one species or the other (289 of 294 cells), with only 5 ‘collisions’ (where collisions likely represent coincidental passage through the same wells

by two or more cells) (**Fig. 1D**). Excluding collisions, we observed an average of 24,454 UMIs (5,604 genes) per human cell and 17,665 UMIs (4,065 genes) per mouse cell, with 1.9% and 3.3% of reads per cell mapping to the incorrect species.

Transcriptomes originating in the 24 wells containing an interspecies mixture of human and mouse nuclei also overwhelmingly mapped to the genome of one species or the other (172 of 175 nuclei), with only 3 collisions (**Fig. S2A**). Excluding collisions, we observed an average of 32,951 UMIs (5,737 genes) per human nucleus and 20,123 UMIs (4,107 genes) per mouse nucleus (**Figs. S2B-C**), with 2.2% and 1.9% of reads per cell mapping to the incorrect species. The greater UMI counts in nuclei are potentially due to the higher amounts of mRNA in cells resulting in a reduced RT efficiency per molecule. Consistent with this, optimizing the number of cells per RT reaction increased UMI counts per cell (*56*).

Estimates of gene expression from the aggregated transcriptomes of nuclei versus cells were well correlated (Pearson: 0.96 for HEK293T, 0.97 for NIH/3T3; **Figs. 1E**, **S2D**). From cells, 81% of reads mapped to the expected strand of genic regions (47% exonic, 34% intronic), and 19% to intergenic regions or the unexpected strand of genic regions. From nuclei, 84% of reads mapped to the expected strand of genic regions (35% exonic, 49% intronic) and 16% to intergenic regions or the unexpected strand of genic regions, similar to previous studies (*57*). Whereas exonic reads show an expected enrichment at the 3' ends of gene bodies, intronic reads do not, and may be the result of poly(dT) priming from poly(dA) tracts in heterogeneous nuclear RNA (**Fig. S3**).

Transcriptomes originating in the 48 wells containing pure or an intraspecies mixture of HEK293T and HeLa S3 cells were readily separated into two clusters by t-stochastic neighbor embedding (t-SNE) (**Figs. 1F, S4**). Estimates of gene expression from the aggregated transcriptomes of all identified HEK293T cells versus a related bulk RNA-seq workflow (Tn5-RNA-seq (*58*)) without methanol fixation were well correlated (Pearson: 0.94, **Fig. 1G**).

Robustness of sci-RNA-seq

After optimizing the number of cells per RT reaction, we fixed a mixture of HEK293T and NIH/3T3 cells, and performed 16 x 84 well sci-RNA-seq (**Table S1**) (*56*). We recovered 185 human cells and 109 mouse cells with 22 collisions (**Fig. 2A**). At ~240,000 reads per cell (73% duplication rate), we observed an average of 49,043 UMIs (7,563 genes) per human cell and 36,737 UMIs (6,263 genes) per mouse cell (**Figs. 2B, S5A**), with 0.9% and 1.2% of reads per cell mapping to the incorrect species. Although this and the previous experiment were performed two months apart on independently grown and fixed cells, the aggregated transcriptomes were well correlated (Pearson: 0.98 for HEK293T, 0.98 for NIH/3T3; **Figs. 2C, S5B**).

We stored a portion of the methanol-fixed mixture of HEK293T and NIH/3T3 cells at -80C for 4 days and repeated sci-RNA-seq (**Table S1**). At ~200,000 reads per cell (73% duplication rate), we observed an average of 30,024 UMIs (5,965 genes) per human cell and

21,393 UMIs (4,503 genes) per mouse cell, with comparable purity (**Fig. S5C**). The aggregated transcriptomes of the fixed-fresh vs. fixed-frozen cells were well correlated (Pearson: 0.99 for HEK293T cells, 0.98 for NIH/3T3 cells; **Figs. 2D**, **S5D**).

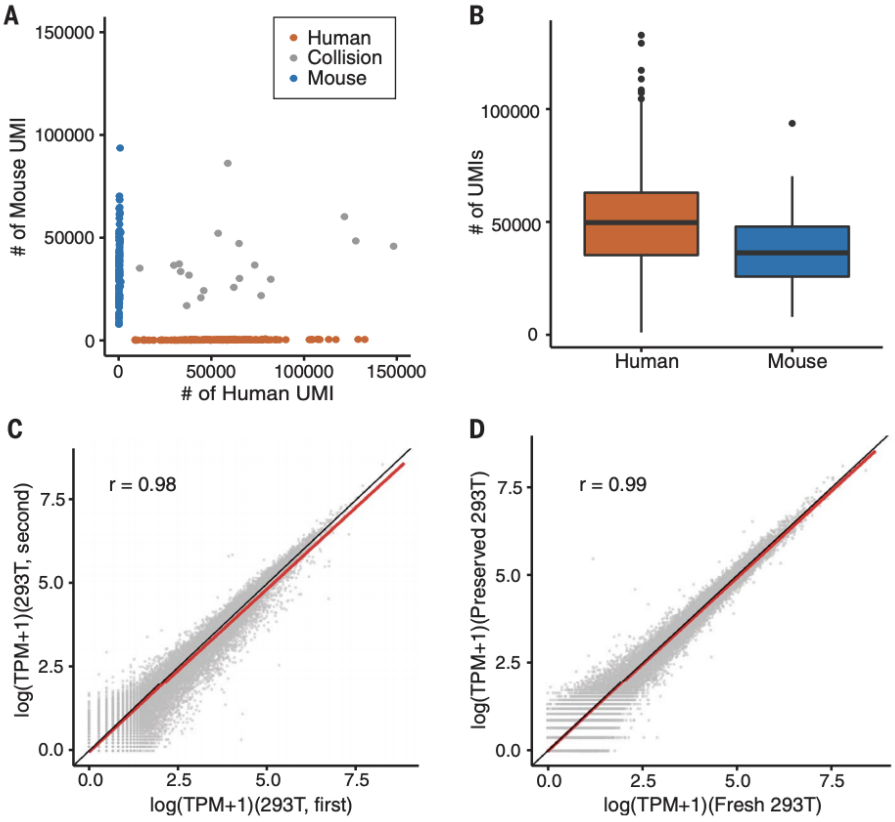


Fig. 2. sci-RNA-seq shows robust gene expression measurements. (**A**) Scatter plot of UMI counts from human and mouse cells, determined by a 16 × 84 sci-RNA-seq experiment on mixed HEK293T and NIH/3T3 cells (**Table S1**). Blue, inferred mouse cells (n = 109). Red, inferred human cells (n = 168). Gray, collisions (n = 19). (**B**) Box plots showing the number of UMIs detected per cell (thick horizontal lines, medians; upper and lower box edges, first and third quartiles, respectively; whiskers, 1.5 times the interquartile range; circles, outliers). (**C**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles from two experiments performed 2 months apart on independently grown and fixed cells. (**D**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of fixed-fresh and fixed-frozen cells. In (C) and (D), the red line is the linear regression, and the black line is y = x.

sci-RNA-seq with three levels of indexing

Two-level combinatorial indexing enables routine profiling of ~10^4 single cells per experiment. We tested an additional level of indexing during Tn5 transposition of double-stranded cDNA (*52*). We performed 16 x 6 x 16 well sci-RNA-seq on mixed HEK293T and NIH/3T3 cells after methanol fixation. After RT with 16 barcodes and second strand

synthesis, cells were pooled and distributed to 6 wells for tagmentation with indexed Tn5 (6 barcodes), then pooled again and sorted to 16 wells for PCR with indexed primers. At ~20,000 reads per cell (51% duplication rate), we recovered 119 human and 62 mouse cells with 5 collisions (**Fig. S6A**). The aggregated transcriptomes of three-level vs. two-level sci-RNA-seq were well correlated (Pearson: 0.96 for HEK293T, 0.94 for NIH/3T3; **Fig. S6B-C**). Downsampling to 15,000 reads per cell, three-level indexing recovered fewer UMIs per cell than two-level indexing (3-level: on average, 6,033 for HEK293T, 3,640 for NIH/3T3; 2-level: 9,942 for HEK293T, 8,611 for NIH/3T3; **Fig. S6D-G**), possibly due to lower efficiency of indexed vs. unindexed Tn5. This limitation notwithstanding, three-level combinatorial indexing has the potential to enable routine profiling of $>10^6$ single cells per experiment (**Fig. S6H**; (*56*)).

Single cell RNA profiling of C. elegans

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

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

Identifying cell types

Semi-supervised clustering analysis segregated the cells into 29 distinct groups, the largest containing 13,205 (31.4%) and the smallest only 131 (0.3%) cells (**Fig. 3A**). Somatic cell types comprised 37,734 cells. We identified genes that were expressed specifically in a single cluster, and by comparing those genes to expression patterns reported in the literature, assigned the clusters to cell types (**Figs. S15-23**). Twenty-six cell types were represented in the 29

clusters: 19 represented exactly one literature-defined cell type, 7 contained multiple distinct cell types, 2 contained cells of a specific cell type but had abnormally low UMI counts, and 1 could not be readily assigned. Neurons, which were present in 7 clusters in the global analysis, were independently reclustered, initially revealing 10 major neuronal subtypes.

Intestine cells were not represented in any cluster. Intestine cells comprise 2.5% of the somatic cells but are polyploid in *C. elegans* larvae (*59*) and also autofluorescent in the DAPI channel used to measure DNA content (*17*). We speculated they may have been excluded by how we gated on DNA content. We therefore performed a second 384 x 144 well *C. elegans* experiment, collecting all cells including polyploid cells on the basis of DAPI fluorescence (96 wells), or gating to enrich for polyploid cells (48 wells). Intestine cells were present (as compared with their absence in the previous experiment) and 2-fold enriched in wells gated for polyploidy. This experiment yielded 7,325 cells (UMI counts per cell for protein-coding genes ≥ 200), of which 6,335 were somatic and 511 intestine cells (**Fig. S8A**).

Gene expression patterns in hypodermal cells suggested that the worm cells from the second *C. elegans* experiment were more tightly synchronized, overlapping but not identical in developmental timing to the first experiment (**Fig. S8B-F**). *C. elegans* larvae feature pervasive oscillations in gene expression within each larval stage (*60*), making it difficult to distinguish biological variation from batch effects. However, the aggregated transcriptomes of human HEK293T cells from these same experiments were well correlated (Pearson: 0.97) and not readily separated by tSNE (**Fig. S9**). This suggests that the variation observed is primarily due to differences in the developmental timing or preparation of the *C. elegans* larvae and cells, rather than technical variation in the sci-RNA-seq protocol. Regardless of its source, to minimize confounding by this variation, we only included the intestine cells from the second *C. elegans* experiment in subsequent analyses, with all other cell types being represented by the first experiment only.

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

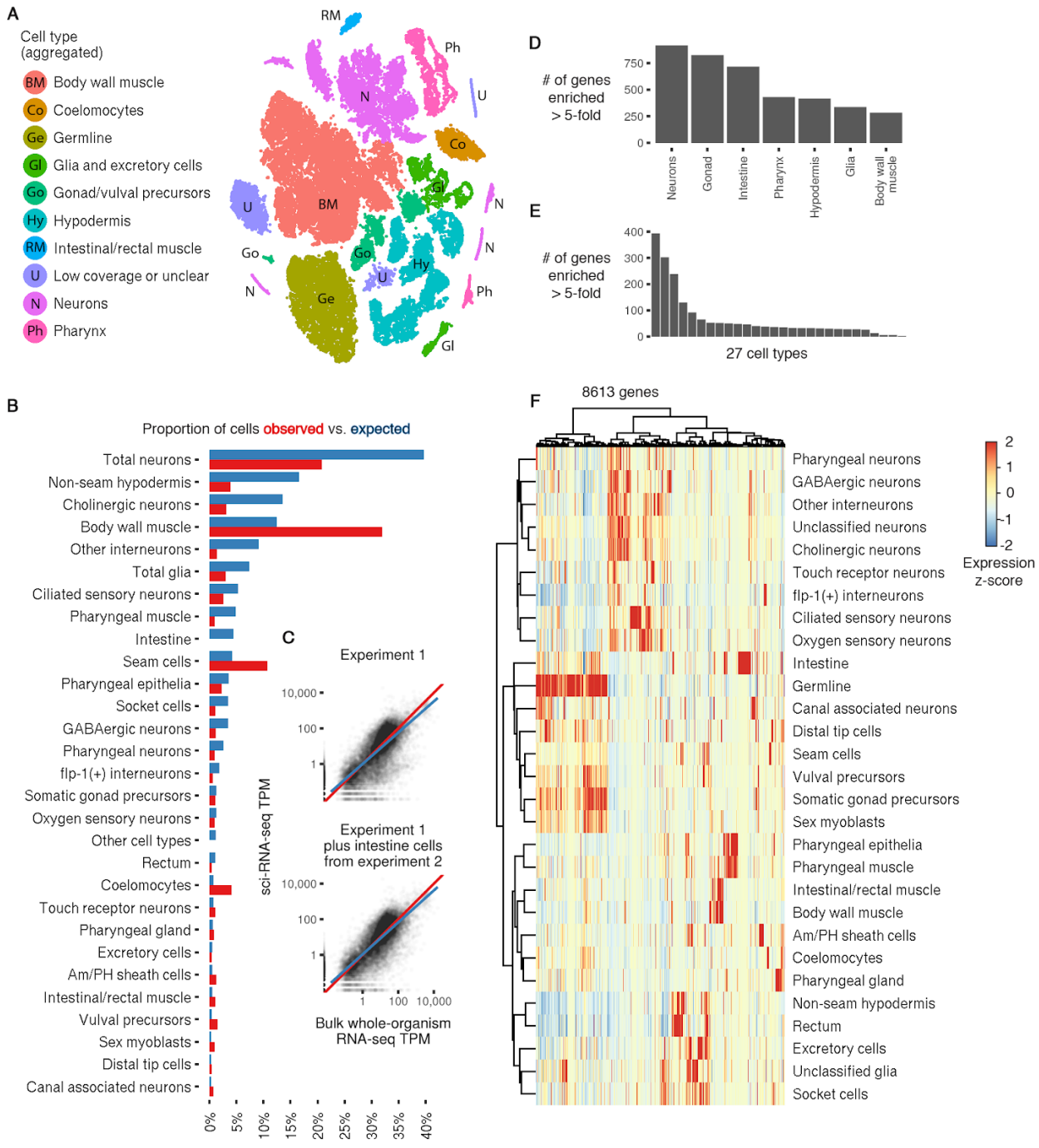


Fig. 3. A single sci-RNA-seq experiment highlights the single-cell transcriptomes of the C. elegans larva. (**A**) t-SNE visualization of the high-level cell types identified. (**B**) Bar graph showing the percentage of somatic cells profiled in the first sci-RNA-seq C. elegans experiment that could be identified as belonging to each cell type (red), compared with the percentage of cells from that type expected in an L2 C. elegans individual (blue). (**C**) Scatter plots showing the log-scaled transcripts per million (TPM) values of genes in the aggregation of all sci-RNA-seq reads (x axis) or in bulk RNA-seq (y axis; geometric mean of three experiments). Red line, y = x; blue line, linear regression. The top plot includes only the first sci-RNA-seq experiment. The bottom plot also includes intestine cells from the second sci-RNA-seq experiment. (*Legend continued on the following page*)

Fig. 3 (continued). (**D**) Number of genes that are at least five times as highly expressed in a specific tissue as in the second-highest expressing tissue, excluding genes for which the differential expression between the first- and second-highest expressing tissues is not significant ($q > 0.05$). (**E**) Same as (D), except comparing cell types instead of tissues. (**F**) Heat map showing the relative expression of genes in consensus transcriptomes for each cell type, estimated by sci-RNA-seq. Genes are included if they have a size factor–normalized mean expression of >0.05 in at least one cell type (8613 genes in total). The raw expression data (UMI count matrix) is log-transformed, column-centered, and scaled (using the R function scale), and the resulting values are clamped to the interval (–2, 2). GABA, g-aminobutyric acid.

Transcriptional programs can be readily distinguished within single cell transcriptome datasets at shallow sequencing depths (*61*). Thus, despite being able to distinguish many distinct cell types in the worm, our molecular definition for each would be incomplete. However, we observed that half of all *C. elegans* protein-coding genes were expressed in at least 100 cells in the full dataset, and 66% of protein-coding genes in at least 20 cells. This compares favorably with the estimates of expressed genes at the L2 stage from whole animal RNA-seq (69%) (*21*). The "whole worm" expression profile derived by aggregating all sci-RNA-seq reads correlated well with whole animal bulk RNA-seq (*21*) for L2 *C. elegans* (**Fig. 3C**; Spearman: 0.796 with cells from the first experiment only, 0.824 including intestine cells from the second experiment). Furthermore, 3,925 genes were enriched in a single tissue (differential expression at least five-fold greater than the 2nd-highest expressing tissue; **Fig. 3D**, **Table S6**), and 1,939 genes were enriched for expression in a single cell type (**Fig. 3E**, **Table S7**). Thus, despite the fact that sci-RNA-seq captures a minority of transcripts in each cell, our 'oversampling' of the cellular composition of the organism enables us to construct representative expression profiles for individual cell types (**Fig. 3F**).

Neuronal cell types

Because the transcripts of tissue or cell type clusters suggested subclasses within groups (**Fig. 4A**), we examined expression within several tissues in more detail. We confirmed and extended findings that anterior and posterior body wall muscle have distinct expression patterns (**Fig. S10A-B**, **Table S9**, (*62*)), and also observed distinct expression patterns for posterior vs. other intestine cells (**Fig. S10C-D**, **Table S10**) and amphid vs. phasmid sheath cells (**Fig. S10E-F**, **Table S11**). Gene expression patterns were particularly diverse in neuronal cell types.

By morphological criteria, the 302 neurons of worm are classified into 118 distinct types (*63*) and from the database of reporter transgene expression patterns, most of these are postulated to have unique molecular signatures (*64*). Our initial re-clustering of neuronal cells divided them into 10 broad classes (**Fig. 4A**). Most classes of neurons were represented by several small but highly distinct clusters in the t-SNE plot. Further analysis of cluster-specific gene expression showed that many clusters corresponded to highly specific subsets of neurons in the L2 worm (**Fig. 4B**, **Table S7**). Three clusters corresponded to sets of four neurons in an individual worm, 8

clusters corresponded to a single pair of neurons (AFD, ASG, ASK, AWA, BAG, CAN, RIA, and RIC), and 3 clusters corresponded to exactly one neuron (ASEL, ASER, and DVA). Hierarchical clustering analysis showed that of the most of 917 genes highly enriched in neurons, compared to other tissues, were expressed in only a minority of neuronal clusters (**Fig. 4C**). 73% of neuron-enriched genes had no more than 10 neuron clusters (out of 40 total) in which they were expressed at ≥ 10% of the level of the highest-expressed cluster. 155 genes were highly enriched in a single neuron cluster relative to all others (**Fig. 4D**, **Table S8**).

Expression of marker genes, such as *gcy-3* and *gcy-6,* were key in identifying two neuronal clusters as left ASE (ASEL) and right ASE (ASER) gustatory neurons, respectively (**Fig. 4E**). These neurons have asymmetry in gene expression (*65*), and we observe 44 genes to be differentially expressed (**Fig. 4F**, **Table S12**, fold difference > 3, FDR < 5%). mRNA from these neurons has previously been profiled with co-immunoprecipitation of RNA and a transgenic poly(A)-binding protein expressed specifically in ASEL or ASER, followed by microarray analysis (*66*). The differentially expressed genes we observe are consistent with this study (**Fig. S11**), highlighting the ability of sci-RNA-seq to facilitate the analysis of cell types as rare as a single cell per individual.

Two neuronal clusters correspond to sister cells, the AWA and ASG neurons, (**Fig. 4G**), which arise from the same parental cell in the last round of *C. elegans* embryonic cell divisions. Their differentiation has previously been used as a model for the study of the regulation of cell fate decisions (*67*). In our data, 136 genes were differentially expressed between these two cell types (**Fig. 4H**, **Table S13**, fold difference > 3, FDR < 5%). The divergent transcriptomes of the AWA and ASG neurons, along with the left and right ASE neurons, highlight the potential of cells that are extremely closely related in morphology and developmental lineage to feature distinct programs of gene regulation.

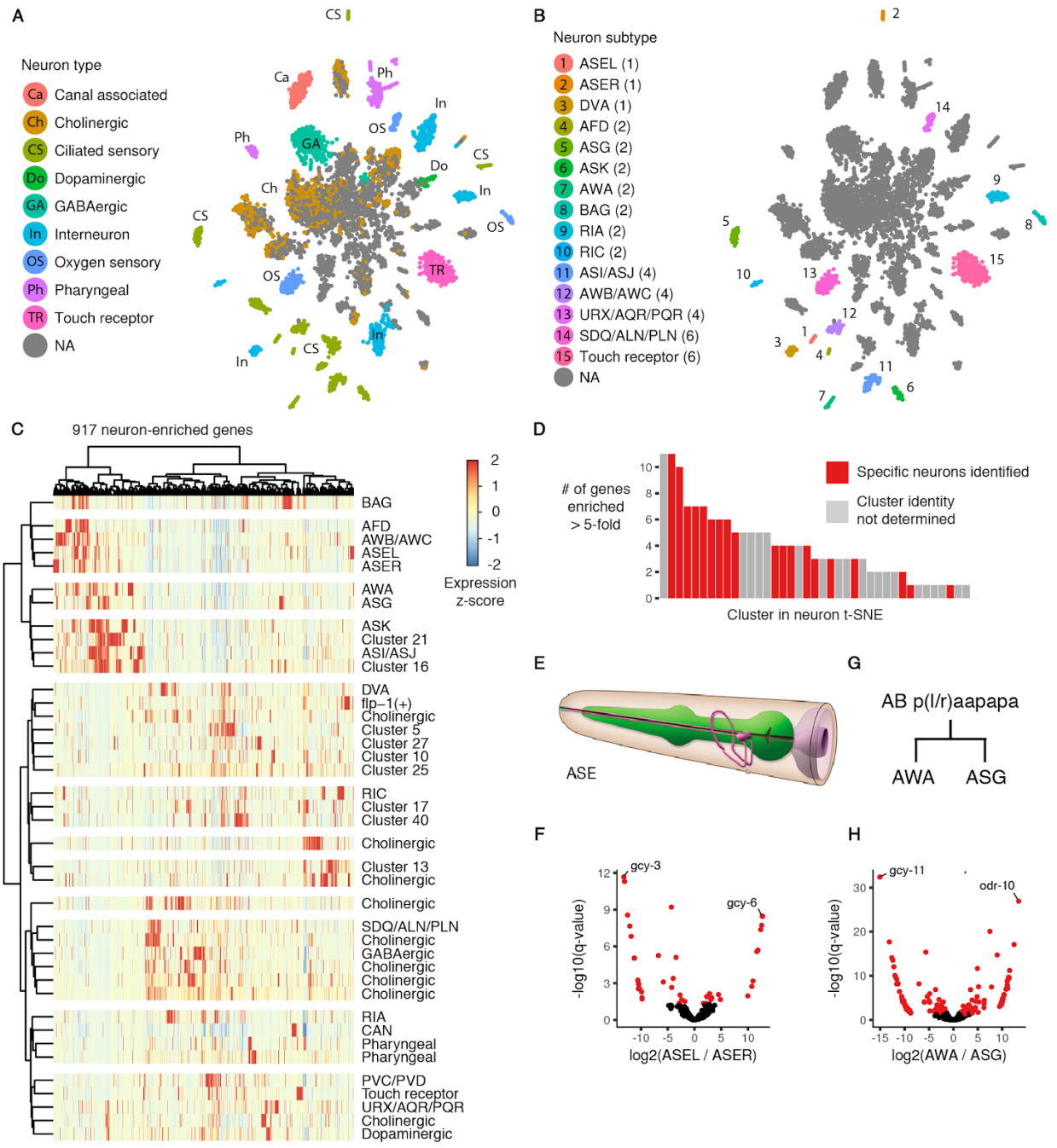


Fig. 4. sci-RNA-seq reveals the transcriptomes of fine-grained anatomical classes of *C. elegans* neurons. (**A**) t-SNE visualization of high-level neuronal subtypes. Cells identified as neurons from the t-SNE clustering shown in **Fig. 3A** were reclustered with t-SNE. NA, not assigned. (**B**) Clusters in the neuron t-SNE that can be identified as corresponding to one, two, or four specific neurons in an individual *C. elegans* larva. The number of neurons of each type is shown in parentheses. (*Legend continued on the following page*)

Fig. 4 (continued) (**C**) Heat map showing the relative expression of high-neuronal-expression genes across 40 neuron clusters identified by t-SNE and density peak clustering. Genes are included if their expression in the aggregate transcriptome of all neurons in our data is more than five times that of their expression in any other tissue, excluding cases where the differential expression is not significant ($q > 0.05$). (**D**) Distribution for each neuron cluster of the number of genes in that cluster whose expression is more than five times that in the second-highest expressing neuron cluster (q for differential expression < 0.05). (**E**) Cartoon illustrating the position of the left and right ASE neurons (pink) relative to the pharynx (green). [From www.wormatlas.org] (**F**) Volcano plot showing differentially expressed genes between the left and right ASE neurons. Points in red correspond to genes that are differentially expressed ($q < 0.05$) with more than a threefold difference between the higher- and lower-expressing neuron(s). (**G**) The left AWA and ASG neurons arise from the embryonic cell AB plaapapa; the right AWA and ASG neurons arise from AB praapapa. (**H**) Volcano plot showing differentially expressed genes between the AWA and ASG neurons.

Integration with transcription factor binding sites

We hypothesized that correlating transcription factor (TF) binding patterns—profiled in ChIP-seq experiments from the modENCODE (*68*) and modERN (*69*) consortia—with cell type gene expression profiles could give insights into the regulatory programs underlying the gene expression profiles. For each of 27 cell types, we constructed regularized regression models to predict each gene's expression as a function of the TF ChIP peaks present in its promoter (**Fig. 5**). We restricted a cell type's model to those TFs that were detectably expressed within it (>10 transcripts per million (TPM)), increasing the proportion of TF-to-cell-type associations that are likely to reflect causal gene regulation. Our regression analysis predicted gene expression by selecting numerous regulators critical for development or proper function-specific cell types, including *hlh-1* and *unc-120* in body wall muscle (*70*), *pha-4* in pharyngeal cell types (*71*), *hlh-8* (CeTwist) in sex myoblasts (*72*), *blmp-1* and *nhr-25* in hypodermis (*73*, *74*), *elt-2* in the intestine (*75*), and *xnd-1* in the germline (*76*, *77*).

The regression identified several putative novel regulators of cell-type specific expression. For example, *fkh-8,* which is expressed specifically in ciliated sensory neurons (our data and reporter construct from (*78*)) was predictive of their gene expression program (**Fig. S12**). The uncharacterized TF *F49E8.2* is expressed specifically in the germline and associated with germline gene expression (**Fig. S12**). F49E8.2 is an ortholog of the human gene "E2F-associated phosphoprotein" (EAPP) (*79*), and F49E8.2 ChIP-seq peaks co-localize with germline-specific EFL-1 peaks (ortholog of E2F, data from (*80*)) more often than could be expected due to chance (**Fig. S13A, B**, χ^2-test, $p = 2.8 \times 10^{-21}$), suggesting that these proteins may physically interact. The hypodermis-associated TFs *blmp-1* and *nhr-25* were also associated with gene expression in socket cells, excretory cells, and rectal cells. *nhr-25* is expressed 4.5-fold higher in socket cells than in seam cells (560 vs. 124 TPM) and 8.7-fold more than in the non-seam hypodermis (560 vs. 64 TPM), suggesting a role in glial development.

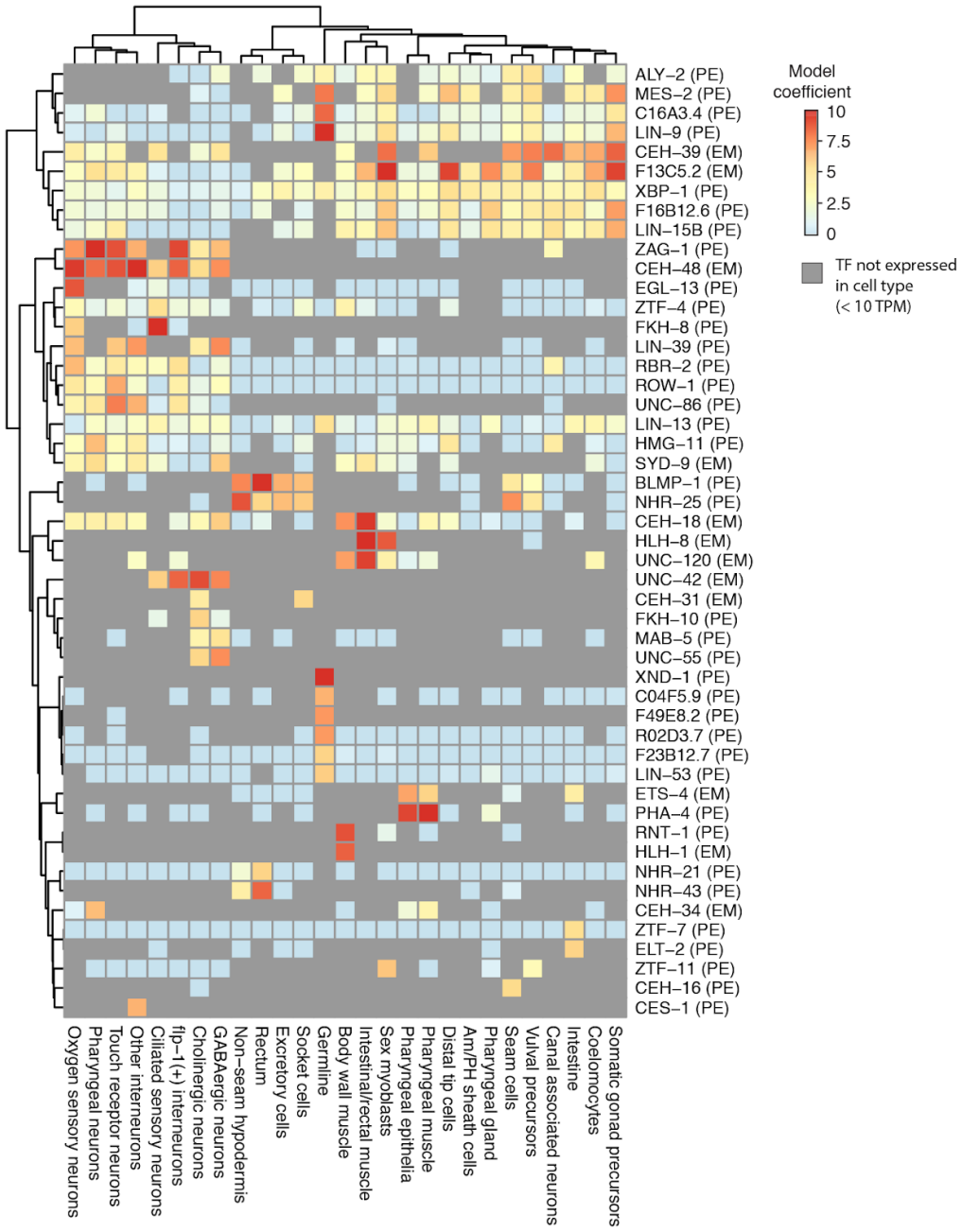


Fig. 5. Cell type–specific expression profiles from sci-RNA-seq enable the deconvolution of whole-animal transcription factor ChIP-seq data. For each of 27 cell types, a regularized regression model was fit to predict log-transformed gene expression levels in that cell type on the basis of ChIP-seq peaks in gene promoters (see **Methods**). The ChIP-seq data were generated by the modENCODE (*68*) and modERN (*81*) consortia, profiling transcription factor binding in whole C. elegans animals. “EM” next to a TF label indicates that the ChIP-seq data for the TF are from an embryonic stage; “PE” indicates that the data are from a postembryonic stage. Colors in the heat map show the extent to which having a ChIP-seq peak for a given TF in a gene promoter correlates with increased expression in a given cell type. Peaks in “HOT” regions (see **Methods**) are excluded. Gray cells in the heat map correspond to cases where a TF is not expressed in a cell type (<10 TPM), in which case ChIP-seq data for that TF are not considered by the regression model.

Discussion

Our method for single cell RNA-seq combinatorial indexing of cells or nuclei can be applied to profile the transcriptomes of tens-of-thousands of single cells per experiment through a library construction completed by a single individual in two days at a cost of $0.03-$0.20 per cell. sci-RNA-seq is compatible with cell fixation, which can minimize perturbations to cell state or RNA integrity before or during processing and facilitates the concurrent processing of multiple samples within a single experiment, potentially reducing batch effects relative to platforms requiring serial processing, an area of concern for the single cell RNA-seq field (*82*). Given that the second barcode is introduced after flow sorting, it is also possible to associate wells on the PCR plate with FACS-defined subpopulations. sci-RNA-seq is also compatible nuclei, which may be important for tissues for which unbiased cell disaggregation protocols are not well established (possibly most tissues). Lastly, sci-RNA-seq is scalable. We demonstrate up to 576 x 960 indexing, which enabled the generation of ~4 x 10^4 single cell transcriptomes in one experiment. However, processing of more cells with sub-linear cost scaling is possible by using more barcoded RT and PCR primers (*e.g.* 1,536 x 1,536 combinatorial indexing) and/or introducing additional rounds of indexing. With 384 x 384 x 384 combinatorial indexing, one can hypothetically profile the transcriptomes of over 10 million cells per experiment.

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

sci-RNA-seq expands the repertoire of single cell molecular phenotypes that can be resolved by combinatorial indexing (*52–55*). Provided that multiple aspects of cellular biology can be concurrently barcoded, combinatorial indexing may also facilitate the scalable generation of 'joint' single cell molecular profiles (*e.g.* RNA-seq and ATAC-seq from each of many single cells). We also envision that large-scale, integrated profiling of the molecular states and lineage histories (*83*) of single cells in other organisms will begin to give shape to "global views" of their developmental biology.

Materials and Methods

Mammalian cell culture

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

Generation of whole *C. elegans* cell suspensions

A *C. elegans* strain, (RW12139 stIs11435(unc-120::H1-Wcherry; unc-119(+)); unc-119(tm4063)) carrying an integrated Punc-120::mCherry gene in a wild type background was used in all experiments. A synchronized L2 population was obtained by two cycles of bleaching gravid adults to isolate fertilized eggs allowing the eggs to hatch in the absence of food to generate a population of starved L1 animals. Around 150,000 L1 larvae were plated on each 100 mm petri plate seeded with NA22 bacteria and incubated at 24°C for 15 hr to produce early L2 larvae. Dissociated cells were recovered following a published protocol (*84*) with modification. Specifically, L2 stage worms were collected by adding 10 ml sterile ddH2O to each plate. The collected L2s were pelleted by centrifugation at 1300 g for 1 min. The larval pellet was washed five times with sterile ddH2O to remove bacteria. The resulting pellet was transferred to a 1.6 ml microcentrifuge tube. Around 40 µl of the final compact pellet was used for each cell dissociation experiment. The worm pellet was treated with 250 µl of SDS-DTT solution (20 nM HEPES pH8, 0.25% SDS, 200 mM DTT, 3% sucrose) for 4 min. Immediately after SDS-DTT treatment, egg buffer (118 mM NaCl. 48 mM KCl. 3 mM CaCl2. 3 mM MgCl2. 5 mM HEPES (pH 7.2) was added to the SDS-DTT treated worms. Worms were pelleted at 500 g for 1 min, then washed 5 times with egg buffer). Pelleted SDS-DTT treated worms were digested with 200 µl of 15 mg/ml pronase (Sigma-Aldrich, St. Louis, MO) for 20 min. The treated worms were broken up to release cells by aspirating up and down through 21G1 ¼ needle. When sufficient single cells were observed the reaction was stopped by adding 900 µl L-15 medium containing 10% fetal bovine serum. Cells were separated from worm debris by centrifuging the pronase-treated worms at 150 g for 5 min at 4°C. The supernatant was transferred to 1.6 ml microcentrifuge tube and centrifuged at 500 g for 5 min at 4°C. The cell pellet was washed twice with egg-buffer containing 1% BSA.

Sample processing

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

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

For sci-RNA-seq on nuclei, 5M cells were combined and lysed using 1 mL ice-cold lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2 and 0.1% IGEPAL CA-630 from (*85*)), modified to also include 1% SUPERase In and 1% BSA). The isolated nuclei were then pelleted, washed twice with 1 mL ice-cold 1X PBS containing 1% DEPC, twice with 500 µL cold lysis buffer, once with 500 µL cold lysis buffer without IGEPAL CA-630, and then resuspended in lysis buffer without IGEPAL CA-630 at a final concentration of 5000 nuclei/µL. For all washes, nuclei were pelleted through centrifugation at 300xg for 3 min. at 4°C).

For cell-mixing experiments, trypsinized cells were counted and the appropriate number of cells from each cell line were combined prior to fixation or lysis. Fixed cells or nuclei were then distributed into 96- or 384-well plates (**Table S1**). For each well, 1,000-10,000 cells or nuclei (2 µL) were mixed with 1 µl of 25 µM anchored oligo-dT primer (5′-ACGACGCTCTTCCGATCTNNNNNNNN[10bp index]TTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3′, where “N” is any base and “V” is either “A”, “C” or “G”; IDT) and 0.25 µL 10 mM dNTP mix (Thermo), denatured at 55°C for 5 min and immediately placed on ice. 1.75 µL of first-strand reaction mix, containing 1 µL 5X Superscript IV First-Strand Buffer (Invitrogen), 0.25 µl 100 mM DTT (Invitrogen), 0.25 µl SuperScript IV reverse transcriptase (200 U/µl, Invitrogen), 0.25 µL RNaseOUT Recombinant Ribonuclease Inhibitor (Invitrogen), was then added to each well. Of note, the RT efficiency was affected by the number of cells (or nuclei) per reaction and too many cells (>4,000) per reaction resulted in lower reaction efficiency and higher impurity. For optimized efficiency, we use 2,000 mammalian cells or 5,000 mammalian nuclei per well for RT reaction. Reverse transcription was carried out by incubating plates at 55°C for 10 min, and was stopped by adding 5 µl 2X stop solution (40 mM EDTA, 1 mM spermidine) to each well. All cells (or nuclei) were then pooled, stained with 4',6-diamidino-2-phenylindole (DAPI, Invitrogen) at a final concentration of 3 µM, and sorted at varying numbers of cells/nuclei per well (depending on experiment; **Table S1**) into 5 uL buffer EB using a FACSAria III cell sorter (BD). Cells are gated based on DAPI stain such that singlets are discriminated from doublets and sorted into the each well. 0.5 µl mRNA Second Strand Synthesis buffer (NEB) and 0.25 µl mRNA Second Strand Synthesis enzyme (NEB) were then added to each well, and second strand synthesis was carried out at 16°C for 150 min. The reaction was then terminated by incubation at 75°C for 20 min.

Tagmentation was carried out on double-stranded cDNA using the Nextera DNA Sample Preparation kit (Illumina). Each well was mixed with 5 ng Human Genomic DNA (Promega), as carrier to avoid over-tagmentation and reduce losses during purification, 5 µL Nextera TD buffer

(Illumina) and 0.5 μL TDE1 enyzme (Illumina), and then incubated at 55°C for 5 min to carry out tagmentation. Note that because the PCR primers used to amplify libraries are specific to the RT products, tagmented carrier genomic DNA are not appreciably amplified or sequenced. The reaction was then stopped by adding 12 μL DNA binding buffer (Zymo) and incubating at room temperature for 5 min. Each well was then purified using 36 uL AMPure XP beads (Beckman Coulter), eluted in 16 μL of buffer EB (Qiagen), then transferred to a fresh multi-well plate.

For PCR reactions, each well was mixed with 2μL of 10 μM P5 primer (5′-AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTC TTCCGATCT-3′; IDT), 2 μL of 10 μM P7 primer (5′-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3′; IDT), and 20 μL NEBNext High-Fidelity 2X PCR Master Mix (NEB). Amplification was carried out using the following program: 72°C for 5 min, 98°C for 30 sec, 18-22 cycles of (98°C for 10 sec, 66°C for 30 sec, 72°C for 1 min) and a final 72°C for 5 min. After PCR, samples were pooled and purified using 0.8 volumes of AMPure XP beads. Library concentrations were determined by Qubit (Invitrogen) and the libraries were visualized by electrophoresis on a 6% TBE-PAGE gel. Libraries were sequenced on the NextSeq 500 platform (Illumina) using a V2 75 cycle kit (Read 1: 18 cycles, Read 2: 52 cycles, Index 1: 10 cycles, Index 2: 10 cycles).

sci-RNA-seq with three-level indexing

Cells were harvested and processed for reverse transcription following the same procedure as sci-RNA-seq with two-level indexing. After reverse transcription, each well was mixed with 0.66 μL second strand synthesis buffer (NEB), 0.33 μL second strand synthesis enzyme (NEB), and incubated at 16°C for 2 hours. Cells from all wells were pooled and distributed to a new 96 well plate (4.5 μL per well). 5 μL Nextera TD buffer (Illumina) and 0.5 μL indexed TDE1 enzyme (Illumina) were added to each well. Tagmentation was performed at 55°C for 10 min and stopped by adding 5 μl 2X stop solution (40 mM EDTA, 1 mM spermidine) to each well. All cells (or nuclei) were then pooled, stained with 4',6-diamidino-2-phenylindole (DAPI, Invitrogen) at a final concentration of 3 μM, and sorted at varying numbers of cells/nuclei per well (depending on experiment; see **Table S1**) into 5 μL buffer (4.6 μL EB buffer, 0.2 μL 1% SDS, 0.2 μL BSA (NEB)) using a FACSAria III cell sorter (BD). Cells are gated based on DAPI stain such that singlets are discriminated from doublets and sorted into the each well. After sorting, each well was mixed with 1 μL of 10 μM P7 primer (5′-CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG-3′, IDT) and incubated at 55°C for 15 min. Then each well was added with 1 μL 10% Tween-20, 1 μL nuclease-free water, 1μL of 10 μM indexed P5 primer (5′-AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTC TTCCGATCT-3′; IDT), and 10 μL NEBNext High-Fidelity 2X PCR Master Mix (NEB).

Amplification program and following steps were the same with sci-RNA-seq with two-level indexing.

Read alignments and construction of gene expression matrix

Base calls were converted to fastq format and demultiplexed using Illumina's bcl2fastq/ 2.16.0.10 tolerating one mismatched base in barcodes (edit distance (ED) < 2). Data were processed with GNU Parallel (*86*). Demultiplexed reads were then adaptor clipped using trim_galore/0.4.1 with default settings. Trimmed reads were mapped to the human reference genome (hg19), mouse reference genome (mm10), C.elegans reference genome (PRJNA13758) or a chimeric reference genome of hg19, mm10 and PRJNA13758, using STAR/v 2.5.2b (*87*) with default settings and gene annotations (GENCODE V19 for human; GENCODE VM11 for mouse, WormBase PRJNA13758.WS253.canonical_gene set for C.elegans). Uniquely mapping reads were extracted, and duplicates were removed using the unique molecular identifier (UMI) sequence (ED < 2, including insertions and deletions), reverse transcription (RT) index, and read 2 end-coordinate (i.e. reads with identical UMI, RT index, and tagmentation site were considered duplicates). Finally, mapped reads were split into constituent cellular indices by further demultiplexing reads using the RT index (ED < 2, including insertions and deletions). For mixed-species experiment, the percentage of uniquely mapping reads for genomes of each species was calculated. Cells with over 85% of UMIs assigned to one species were regarded as species-specific cells, with the remaining cells classified as mixed cells or "collisions". The collision rate was calculated as twice the ratio of mixed cells (as we are blind to collisions involving cells of the same species). For gene body coverage analysis of exonic reads, the split human and mouse single cell SAM files were concatenated and exonic reads were selected and analyzed using RSEQC/2.6.1, using BED annotation files downloaded from the UCSC Golden Path. For read position analysis for intronic reads, the split human and mouse single cell SAM files were concatenated and intronic reads were selected; the fractional position of each intronic read along the genomic distance between the TSS and transcript terminus was calculated, and these values used to generate a density plot.

To generate digital expression matrices, we calculated the number of strand-specific UMIs for each cell mapping to the exonic and intronic regions of each gene with python HTseq package (*88*). Generally, fewer than 3% of total UMIs strand-specifically mapped to multiple genes. For multi-mapped reads, reads were assigned to the closest gene, except in cases where another intersected gene fell within 100 bp to the end of the closest gene, in which case the read was discarded. For most analyses we included both expected-strand intronic and exonic UMIs in per-gene single-cell expression matrices.

For sci-RNA-seq with three-level indexing, reads were analyzed with the same procedure, except that RT index was combined with Tn5 index, and thus the mapped reads were

split into constituent cellular indices by demultiplexing reads using both the RT index and Tn5 index (ED < 2, including insertions and deletions).

t-SNE visualization of HEK293T cells and HeLa S3 cells

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

Genotyping of single HeLa cells by 3' tag sequences

HeLa S3 cell identity was verified on the basis of homozygous alleles not present in the hg19 assembly, using a callset derived from (*89*). Single-cell BAM files (with cellular indices encoded in the "read_id" field) were concatenated, and then processed as follows using a python wrapper of the samtools API (i.e. pysam). For each homozygous alternate SNV overlapping with a GENCODE V19 defined gene (n = 865,417) in the HeLa S3 variant callset, we computed the fraction of matching (i.e. HeLa S3 specific) alleles, and computed this value for all cells where at least 1 read containing a polymorphic site. We then re-plotted in R the tSNE visualization shown in **Fig. S4B**, now colored by the relative fraction of homozygous alternate alleles called for each cell.

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

To compare aggregated sci-RNA-seq single cell transcriptomes with bulk RNA-seq, we performed bulk RNA-seq using a modified protocol (*58*). In brief, 500 ng total RNA extracted from three biological replicate HEK293T samples (extraction using RNeasy kit (Qiagen)) with the RNeasy kit (Qiagen) were used for reverse transcription following the standard SuperScript II protocol. 500 ng total RNA (in 9 µL water) was mixed with 2 µL 25 uM oligo-dT(VN) (5′-ACGACGCTCTTCCGATCTNNNNNNNN[10bp index]TTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN-3′, where "N" is any base and "V" is either "A", "C" or "G"; IDT) and 1 µL 10 mM dNTPs, then incubated at 65°C for 5 min Following incubation, 8 µL reaction mix (4 µL 5X Superscript II First-Strand Buffer, 2 µl 100 mM DTT, 1 µl SuperScript II reverse transcriptase, 1 µL RnaseOUT) was added. Reactions were

incubated at 42°C for 50 min and terminated at 70°C for 15 min. For second strand synthesis, 2 μL RT product was mixed with 6.5 μL water, 1 μL mRNA Second Strand Synthesis buffer (NEB) and 0.25 μl mRNA Second Strand Synthesis enzyme (NEB). Second strand synthesis was carried out at 16°C for 150 min, followed by 75°C for 20 min. Tagmentation was carried out by adding 10 μL Nextera TD buffer, 1 μL Nextera Tn5 enyzme and incubating at 55°C for 5 min. Tagmented cDNA was purified using a Clean & Concentrator^{TM}-100 kit (Zymo) and eluted in 16 uL buffer EB. PCR, purification, and quantification were then performed as detailed above.

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

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

Both *C. elegans* sci-RNA-seq experiments were processed identically except as noted. A digital gene expression matrix was constructed from the raw sequencing data as described above. Cells with UMI count for protein-coding genes < 100 (experiment 1) or < 200 (experiment 2; higher threshold to compensate for slightly more leakage between cells) were excluded from the analysis. The dimensionality of this matrix was reduced first with PCA (40 components) and then with t-SNE, giving a two-dimensional representation of the data. This t-SNE was performed using the implementation in Monocle version 2.3.5(*90*). Similar to the approach in (*91*), cells in this two-dimensional representation were clustered using the density peak algorithm (*92*) as implemented in Monocle 2.3.5. Genes specific to each cluster were identified and compared to microscopy-based expression profiles reported in the literature (**Table S2**, **Figs. S15-23**), allowing the distinct cell types represented in each cluster to be identified. Based on these results, in experiment 1, we manually merged two clusters that both corresponded to body wall muscle, and manually split two clusters that included hypodermis, somatic gonad cells, and glia. Seven clusters exclusively contained neurons. We identified neuronal subtypes applying PCA, t-SNE, and density peak clustering to this subset of cells using the same approach as for the global cluster analysis.

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

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

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

Genes with expression patterns highly enriched in a single tissue were identified as follows. For each gene (excluding those expressed in fewer than 10 cells), the tissue in which it is expressed highest and the tissue in which it is expressed second-highest (relative to other tissues) are enumerated. The gene is considered enriched in the highest expressing tissue if it is both expressed at a >5-fold greater level than in the second-highest expressing tissue and the differential expression of this gene between the highest and second-highest expressing tissues is non-zero at a false detection rate of < 5%. The differential expression tests are performed with the differentialGeneTest function of Monocle 2 (*90*). The false detection rates are computed based on the tests for all genes, not just the genes with a given highest/second-highest expressing tissue. Genes with expression patterns enriched in a single cell type or a single neuron cluster were identified using the same method (i.e. comparing the highest and second-highest expressing cell type instead of tissue).

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

Integration of sci-RNA-seq expression profiles and modENCODE/modERN ChIP-seq data

Transcription factor (TF) ChIP-seq datasets were downloaded from the ENCODE data portal. The ChIP-seq data included experiments conducted on whole embryos or whole larvae at

different developmental stages. ChIP peaks for the same TF were merged if they overlapped and were either both from an embryonic stage experiment or both from a post-embryonic stage experiment. If a TF had both embryonic and post-embryonic data available, only the post-embryonic data was used.

A ChIP-seq peak was considered to be associated with a gene if: 1) the peak summit was within 2 kb of the canonical transcription start site (TSS) for the gene, 2) the distance from the peak summit to the second closest TSS (regardless of strand) was at least 50% greater than the distance to the closest TSS, and 3) the peak overlapped peaks for < 20% of assayed TFs from the same broad developmental stage (embryonic or post-embryonic). This excludes so-called "HOT regions" which are likely to reflect either non-sequence-specific TF binding or an artifact of the ChIP-seq assay (*93*).

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

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

To identify pairs of TFs that have co-localized binding patterns more often than could be expected by chance (**Fig. S13**), peaks were first clustered by recursively merging those with summits within 150 bp of each other. This analysis was limited to TFs with ChIP-seq data from post-embryonic worms, and also included germline-specific ChIP-seq for EFL-1 and DPL-1 produced by (*80*). Peak clusters that contained peaks for >20% of the TFs ("HOT regions") were excluded from further analysis. Peak clusters were associated with genes using the same criteria as used for individual peaks (described above, treating the midpoint of the cluster's genomic interval as the "summit" of the cluster). Peak clusters that could not be associated with a gene were excluded from further analysis. From the remaining peak clusters, a matrix was constructed where the rows are identifiers for each peak cluster and the columns are binary variables with value 1 if the cluster includes at least one peak for a given TF, 0 otherwise.

This matrix was used as input to the Graphical LASSO (*94*), an algorithm which provides robust estimates of partial correlations between a set of random variables given a limited number of observations and under the assumption that most variables are conditionally independent from another. In this context, the partial correlation between two columns of the input matrix is equal to the correlation of the events ">0 peaks for TF 1 are present in this peak cluster" and ">0 peaks for TF 2 are present in this peak cluster", conditioned on the presence or absence of peaks for all other TFs. The Graphical LASSO was applied to either the full matrix (**Fig. S13D**) or the subset of rows in the matrix that corresponded to peak clusters in the promoters of gonad-enriched genes (**Fig. S13A**) or neuron-enriched genes (**Fig. S13C**). From the partial correlations outputted by each Graphical LASSO, we constructed a network where the nodes are TFs (columns in the matrix) and undirected edges connect each pair of TFs for which the partial correlation in either direction (TF 1 → TF 2 or TF 2 → TF 1) is > 0.01.

The Graphical LASSO model requires a regularization parameter to be set by the user, with increasing values. We set this parameter to the smallest value that satisfied the requirement that the probability that a non-zero partial correlation in the output is in fact zero in the "true" model—the false detection rate—is less than 5%. To find a mapping between regularization parameter values and the false detection rate, we constructed a null model by shuffling the values of the input matrix in a manner that preserves both row and column sums, using the CurveBall algorithm (*95*). In a shuffled matrix, all non-zero partial correlations reported by the Graphical LASSO are false detections. We therefore estimate the false detection rate of a given regularization parameter value to be equal to the mean number of non-zero partial correlations reported by the Graphical LASSO for shuffled matrices (averaging over 50 shuffles) divided by the number of non-zero partial correlations reported by the Graphical LASSO on the unshuffled input data.

Cost estimation

Using the 576 x 960 sci-RNA-seq experiment as an example, reagent costs are largely enzyme-driven and include SuperScript IV reverse transcriptase ($934), second strand synthesis mix ($750), Nextera Tn5 enzyme ($5,000), NEBnext master mix ($1,150), FACS sorting ($250) and other reagents and plates ($250). If we sort 60 cells per well (assuming recovery rate is 100%) for 960 wells (5% collision rate), then the reagent cost of library preparation is around $0.14 per cell (expected yield of around 55,000 cells). However, it is worth noting that simply increasing the number of cells sorted per well decreases costs (e.g. sorting 150 cells to each well would yield around 140,000 cells at a cost of $0.05 per cell), but also results in an increased collision rate (12%). Alternatively, by increasing to 1,536 barcodes during the first (RT-based) round of indexing, we can sort up to 320 cells per well at a 10% collision rate, thereby reducing the cost per cell to less than $0.025 per cell. Straightforward reductions in reaction volumes and/or in-house enzyme production at all steps may also lead to further reductions in costs, as

would additional rounds of molecular indexing. For example, with 384 x 384 x 384 combinatorial indexing, we can potentially uniquely barcode the transcriptomes of around 12 million cells at a 10% collision rate, corresponding to >200-fold increase in detection capacity relative to the 576 x 960 experiment, without much increase in reagent costs.

Supplementary Figures

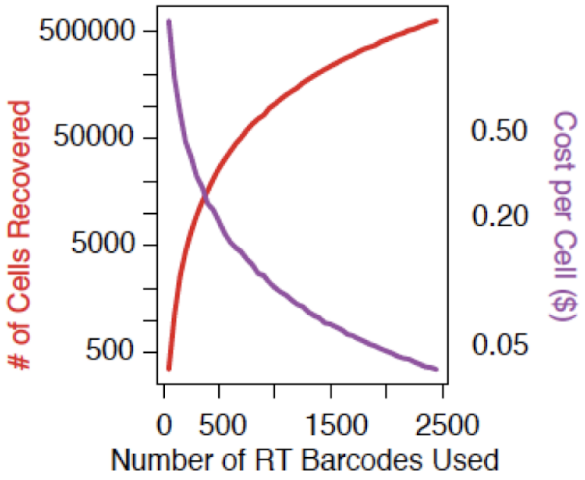


Fig. S1. Combinatorial indexing with increasing numbers of reverse transcription (RT) barcodes enables sublinear scaling of cost per cell. Plot assumes two-level indexing and estimates how detection capacity (i.e. the number of cells detected in a sci-RNA-seq experiment, red) and cost per cell (blue) vary as a function of the number of RT barcodes used, assuming a collision rate of 5%.

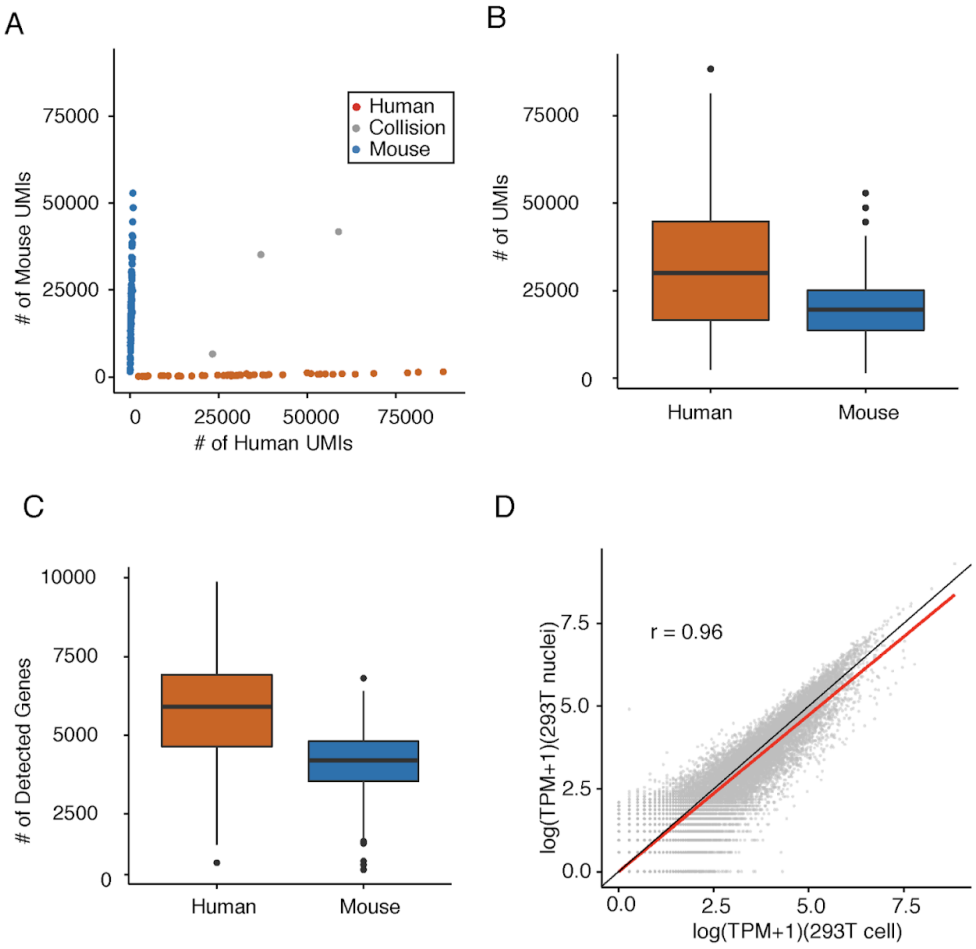


Fig. S2. sci-RNA-seq is compatible with isolated nuclei as starting material. (**A**) Scatter plot of unique human and mouse nuclei UMI counts from a 96 x 96 sci-RNA-seq experiment. This experiment included different cell populations (**Table S1**), but only cells originating from a mixture of human (HEK293T) and mouse (NIH/3T3) nuclei are plotted here. Inferred mouse cells (n = 124) are colored in blue; inferred human cells (n = 48) are colored in red, and "collisions" (n = 3) are colored in grey. (**B, C**) Boxplots showing the number of UMIs (B) and genes (C) detected per cell in nuclear sci-RNA-seq experiments. (**D**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of HEK293T cells (n = 328) vs. HEK293T nuclei (n = 48), together with a linear regression line (red) and y=x line (black).

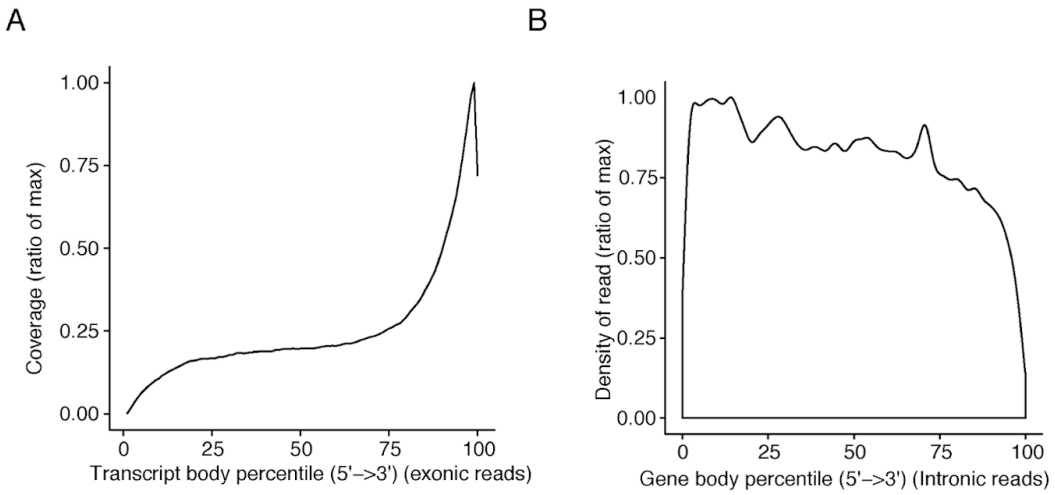


Fig. S3. Positional bias of exonic and intronic sci-RNA-seq reads. (**A**) Density plot showing that as expected, sci-RNA-seq reads mapping to exons are strongly biased to originate near the 3' ends of transcripts (intronic regions excluded from percentile scaling). (**B**) Density plot showing that in contrast, sci-RNA-seq reads mapping to introns do not exhibit 3' bias (intronic regions included in percentile scaling). Y-axis is scaled to the ratio of max.

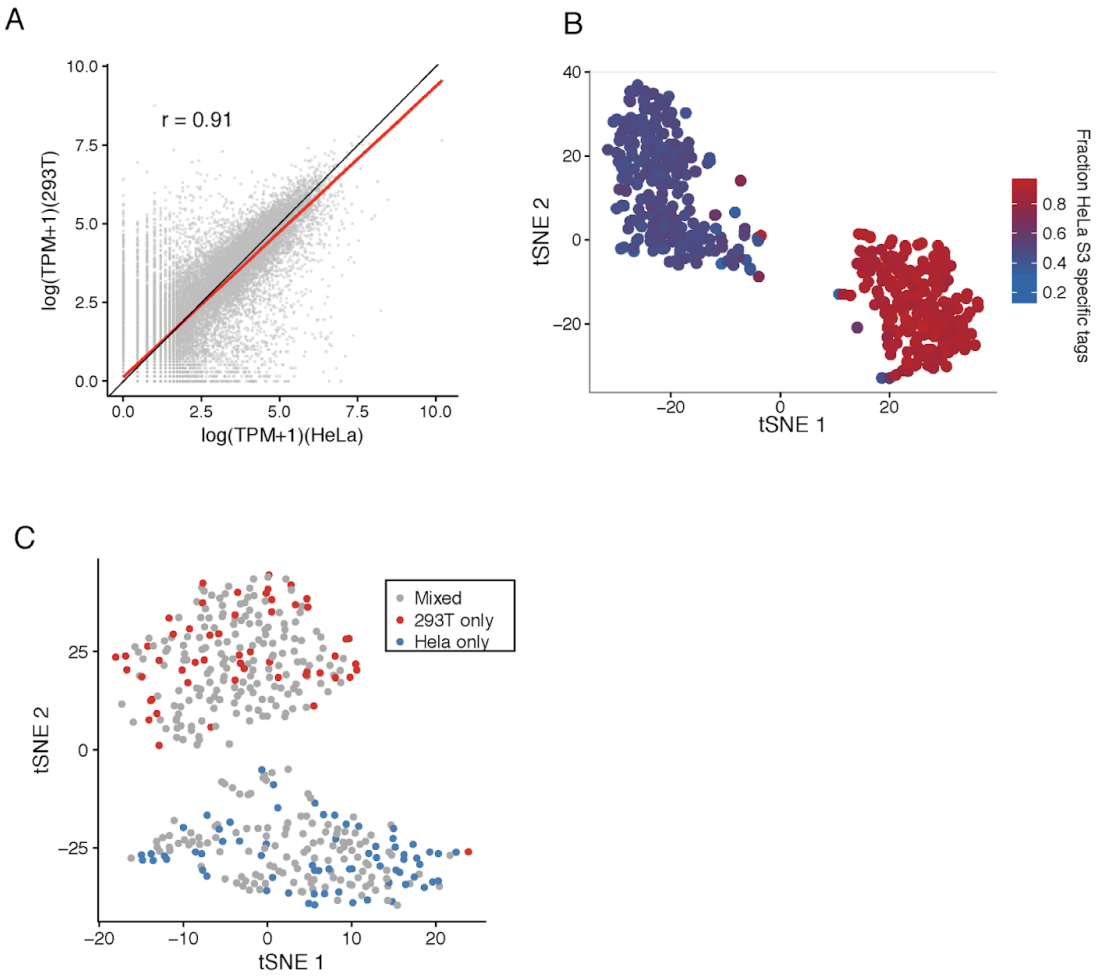


Fig. S4. Quality control for sci-RNA-seq on mixed populations of HeLa S3 and HEK293T cells. (**A**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of HeLa S3 vs. HEK293T cells, together with a linear regression line (red) and y=x line (black). (**B**) tSNE plot (as in **Fig. 1F**), with cells colored by fraction of reads harboring HeLa S3 specific SNVs (single nucleotide variants) relative to hg19 assembly. (**C**) tSNE using digital gene expression matrices constructed from only intronic reads. Cells are colored by the population from which they derived, with pure HEK293T in red, pure HeLa S3 in blue, and mixed cells in grey.

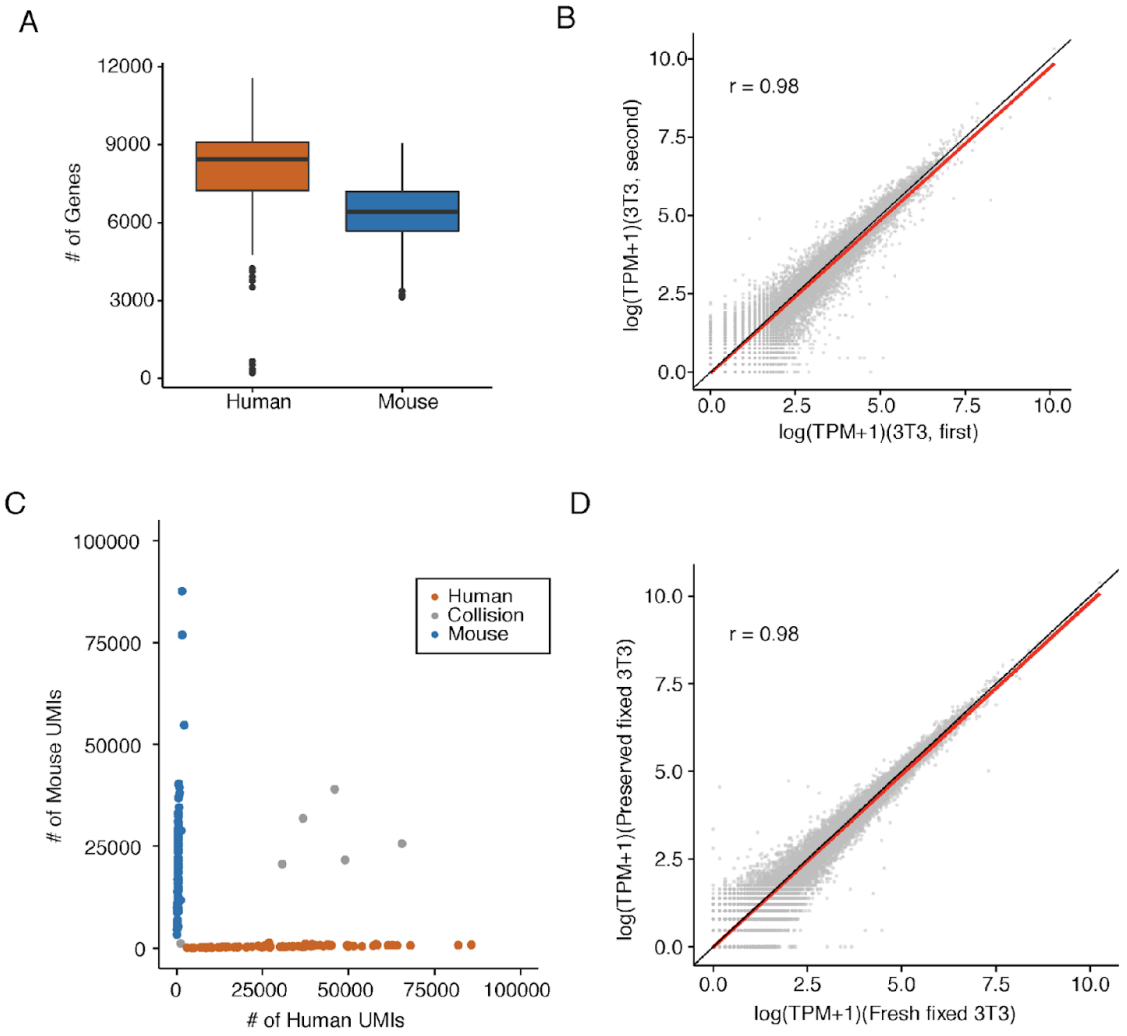


Fig. S5. sci-RNA-seq shows robust gene expression measurements. (**A**) Boxplots showing the number of genes detected per cell in a 16 x 84 well sci-RNA-seq experiment. (**B**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of NIH/3T3 cells from two sci-RNA-seq experiments, performed two months apart and on independently grown and fixed cells, together with a linear regression line (red) and y=x line (black). (**C**) Scatter plot of unique human and mouse UMI counts from a 16 x 84 sci-RNA-seq experiment on mixed HEK293T and NIH/3T3 cells after methanol fixation and freezing at -80°C for 4 days. Inferred mouse cells (n = 90) are colored in blue; inferred human cells (n = 89) are colored in red, and "collisions" (n = 6) are colored in grey. (**D**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of fixed-fresh vs. fixed-frozen NIH/3T3 cells, together with a linear regression line (red) and y=x line (black).

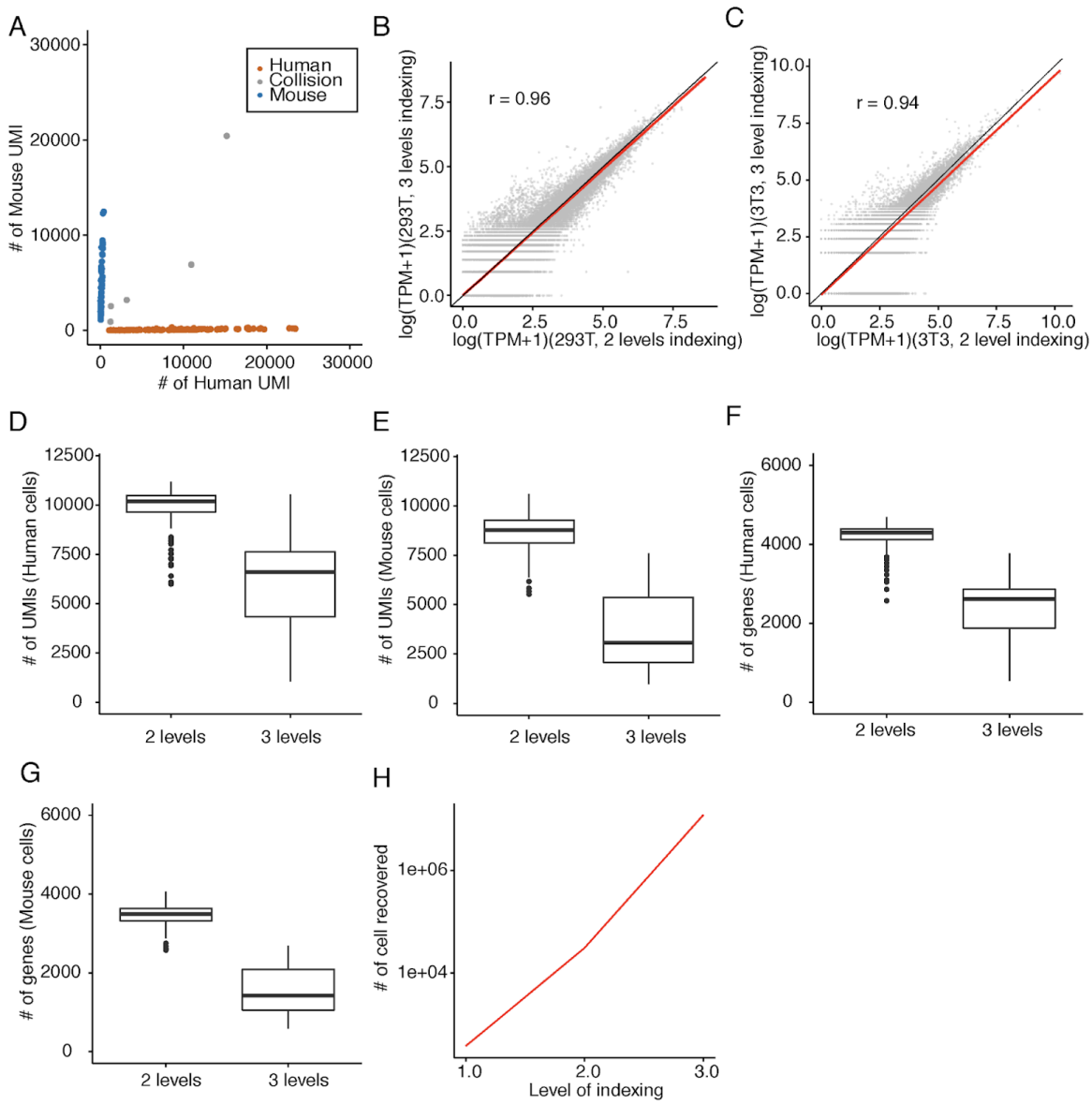


Fig. S6. Representative result from sci-RNA-seq with 3-level indexing. (**A**) Scatter plot of unique human and mouse UMI counts from a 16 x 6 x 16 sci-RNA-seq experiment on mixed HEK293T and NIH/3T3 cells. Inferred mouse cells (n = 62) are colored in blue; inferred human cells (n = 119) are colored in red, and "collisions" (n = 5) are colored in grey. (**B**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of HEK293T cells with 2-level vs. 3-level indexing, together with a linear regression line (red) and y=x line (black). (**C**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of NIH/3T3 cells in sci-RNA-seq with 2-level vs. 3-level indexing, together with a linear regression line (red) and y=x line (black). (**D, E**) Boxplots showing the number of UMIs detected per HEK293T cell (D) and NIH/3T3 cell (E) in sci-RNA-seq with 2-level or 3-level indexing, sampling 15,000 total reads per cell. (**F, G**) Boxplots showing the number of genes detected per HEK293T cell (F) and NIH/3T3 cell (G) in sci-RNA-seq with 2-level or 3-level indexing, sampling 15,000 total reads per cell. (**H**) Plot illustrating how estimated detection capacity (i.e. the number of cells detected in a sci-RNA-seq experiment, red) varies as a function of number of rounds of indexing used, assuming a collision rate of 10% and 384 indexes at each level.

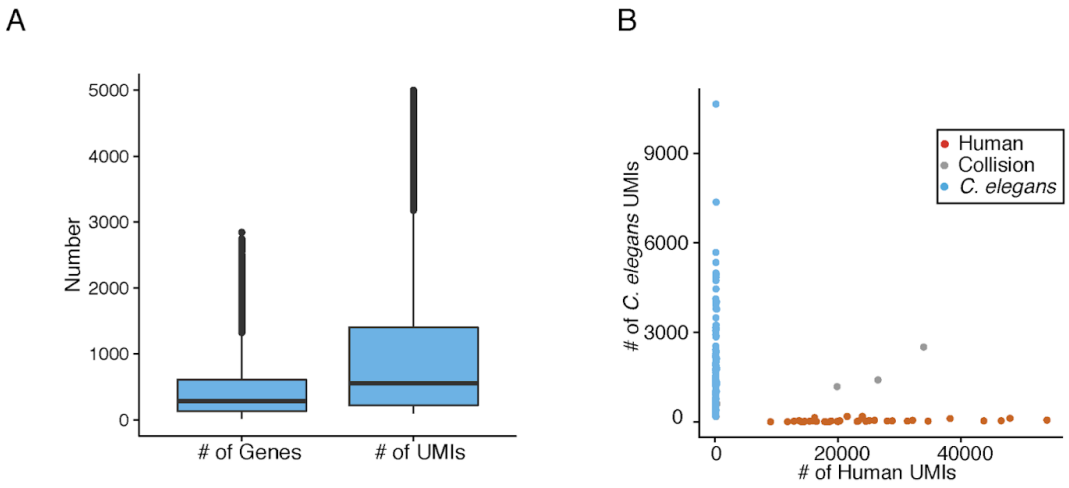


Fig. S7. Quality control metrics for *C. elegans* sci-RNA-seq experiments. (**A**) Distribution of number of protein-coding genes and UMI counts (mapping to protein-coding genes) detected per *C. elegans* cell. (**B**) Scatter plot of unique UMI counts per cell from a sci-RNA-seq experiment performed on mixture of HEK293T (human) and *C. elegans* cells.

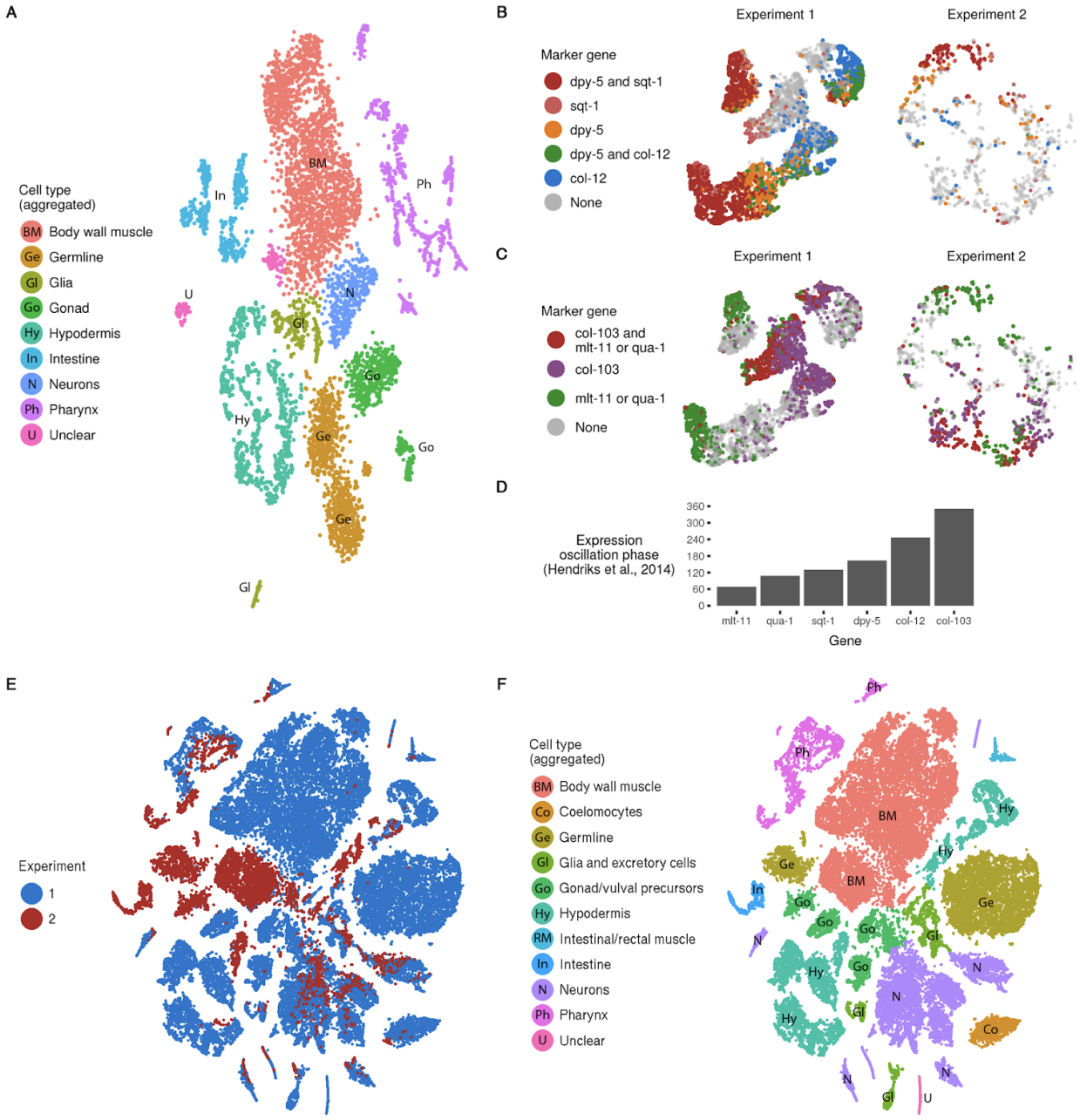


Fig. S8. A second *C. elegans* sci-RNA-seq experiment recovers intestine cells. (**A**) t-SNE visualization of cells from the second *C. elegans* experiment, which included all cells (96 wells) or only cells with high DAPI stain (48 wells). 511 intestine cells were successfully recovered. (**B**) Expression of the cuticle collagens *dpy-5*, *sqt-1*, and *col-12* in cells from experiments 1 and 2. t-SNE coordinates for cells are the same as in **Fig. 3A** (for experiment 1) and (A) (for experiment 2), but only hypodermal cells are shown. *dpy-5* and *sqt-1* are expressed during the synthesis of new cuticle preceding each larval molt, while *col-12* is expressed during molting and ecdysis (*96*). (**C**) Expression of the signaling gene *qua-1*, the protease inhibitor *mlt-11*, and the collagen *col-103*, in experiments 1 and 2. *qua-1* and *mlt-11* are expressed at the initiation of new cuticle synthesis (*97*). *col-103* is expressed in the intermolt, after ecdysis but before new cuticle synthesis begins (*60*). (*Legend continued on the following page*)

Fig. S8 (continued). Taken together with (B), the expression patterns suggest that the worms in experiment 1 spanned a range of developmental sub-stages from late L2 to around the L3 molt, while worms from experiment 2 had greater synchrony and were mostly from the early L2 stage. (**D**) Phase of the molting-cycle associated gene expression oscillations of selected genes, as reported by (*60*). The values are modulo 360, i.e. 360 is the same as 0 and equidistant from 90 and 270. (**E, F**) t-SNE visualizations of cells from both *C. elegans* experiments processed together.

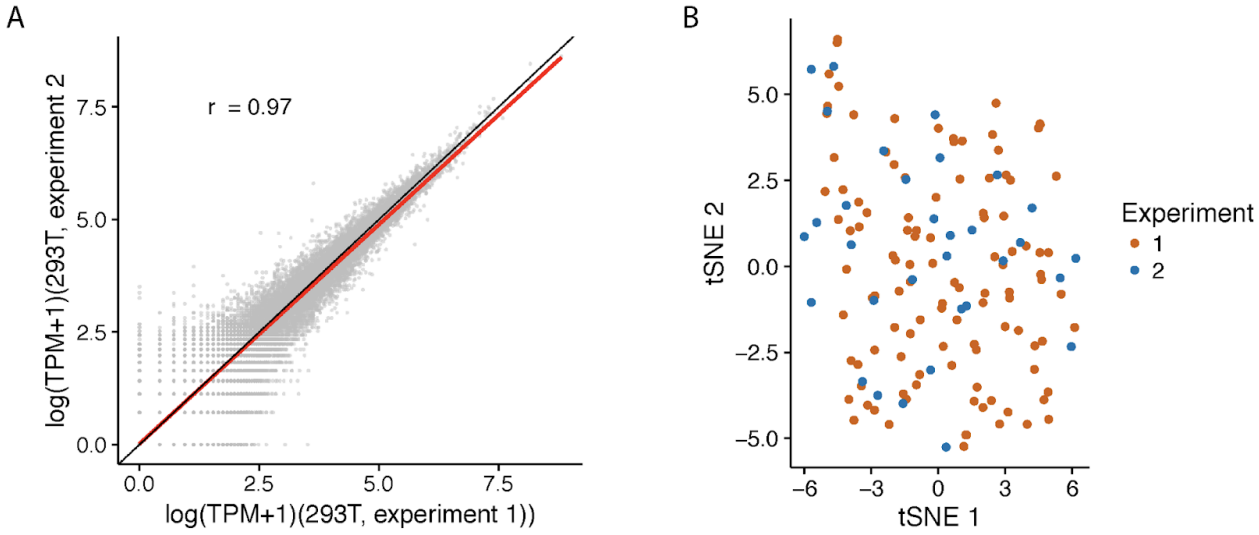


Fig. S9. Evaluation of technical variance between the two *C. elegans* experiments. (**A**) Correlation between gene expression measurements in aggregated sci-RNA-seq profiles of HEK293T cells spiked in with in the first *C. elegans* experiment (n = 32) vs. the second experiment (n = 111), together with a linear regression line (red) and y=x line (black). (**B**) t-SNE clustering of HEK293T cells recovered from the two experiments. Cells are colored by the experiment from which they derived.

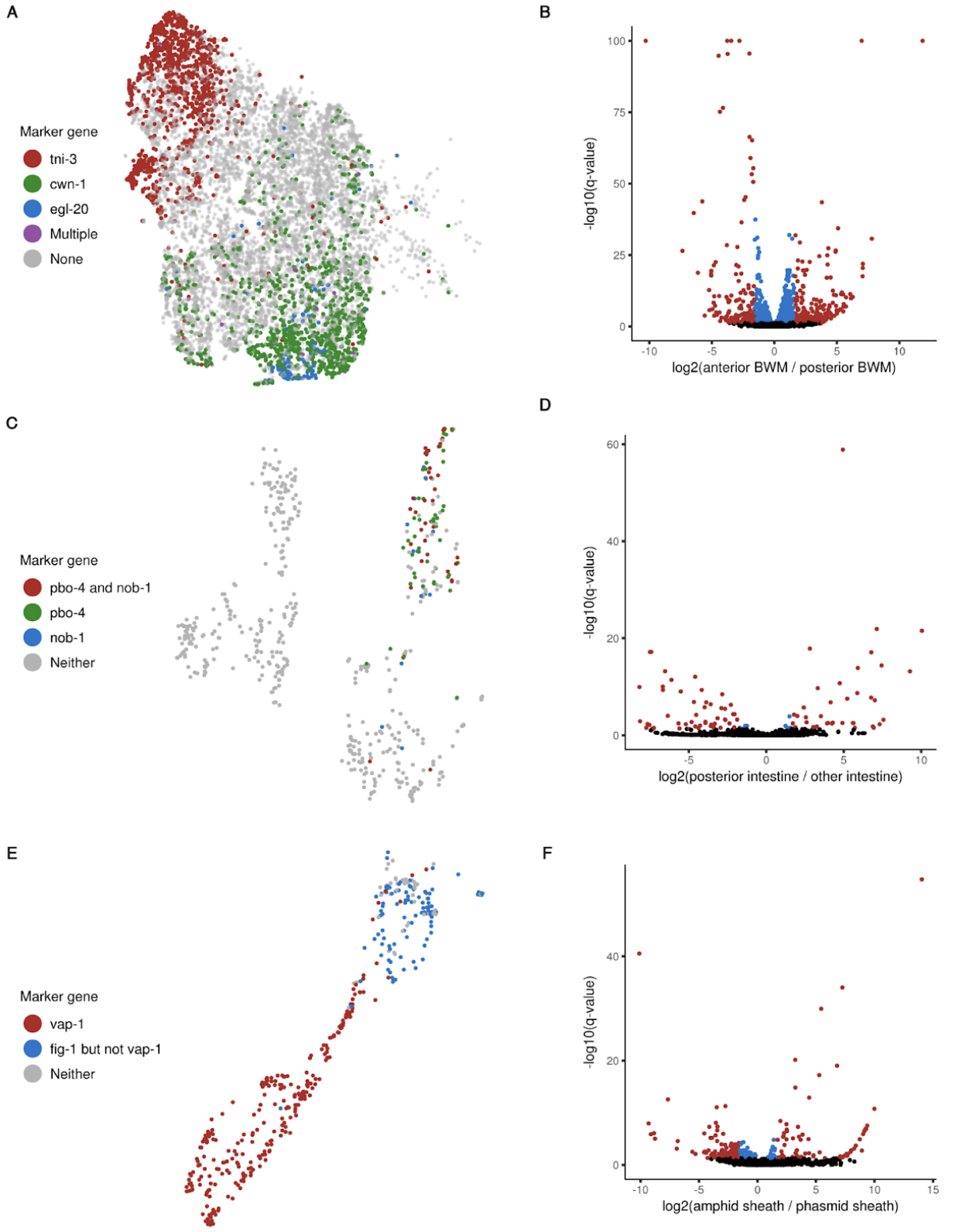


Fig. S10. sci-RNA-seq reveals genes differentially expressed between anterior and posterior cells for three cell types. (*Legend continued on the following page*)

Fig. S10 (continued). (**A**) Expression of anterior/posterior marker genes in body wall muscle cells. Cell t-SNE coordinates are the same as in **Fig. 3A**, except only BWM cells are shown. *tni-3* (red) is specific to the head (*62*), while *cwn-1* (green) and *egl-20* (blue) are specific to the posterior and tail respectively (*98*). (**B**) Volcano plot showing genes differentially expressed between anterior [*tni-3*(+)] and posterior [*cwn-1*(+) or *egl-20*(+)] body wall muscle. -log10 q-values (y-axis) are capped at 100. Genes with differential expression q-value < 0.05 are colored red if the fold difference in expression is >3, blue otherwise. (**C**) Expression of posterior marker genes in intestine cells. Cell t-SNE coordinates are the same as in **Fig. S10A**, except only intestine cells are shown. *pbo-4* and *nob-1* are specific to the posterior (*19*, *99*). (**D**) Volcano plot showing genes differentially expressed between posterior [*pbo-4*(+) or *nob-1*(+)] intestine and other intestine. Colors are the same as in (B). (**E**) Expression of amphid/phasmid (anterior/posterior) marker genes in amphid/phasmid sheath cells. Cell t-SNE coordinates are the same as in **Fig. 3A**, except only amphid/phasmid sheath cells are shown. *fig-1* is expressed in both amphid and phasmid sheath cells, while *vap-1* is specific to the amphid sheath cells (*100*, *101*). (**F**) Volcano plot showing genes differentially expressed between amphid [*vap-1*(+)] and phasmid [*fig-1*(+) *vap-1*(-)] sheath cells. Colors are the same as in (B).

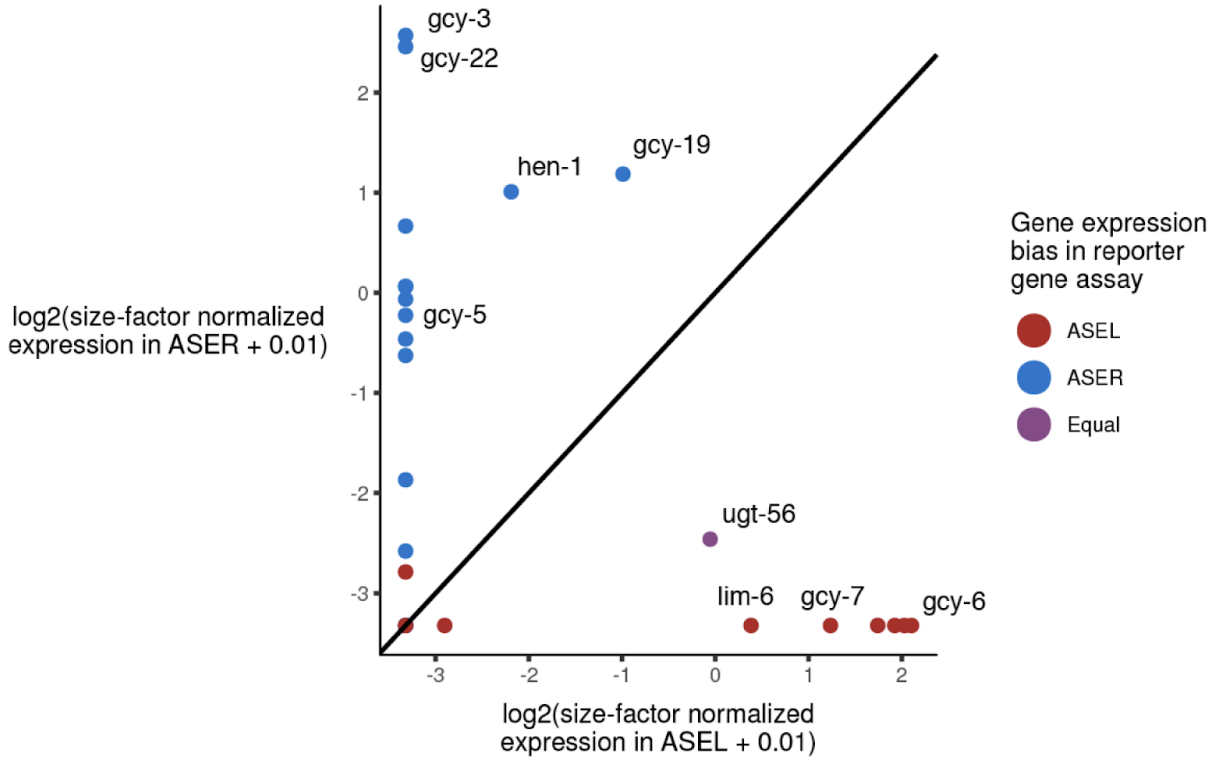


Fig. S11. sci-RNA-seq expression profiles for the ASEL and ASER neurons are consistent with reporter gene assays for asymmetric gene expression. Points represent genes which were tested for asymmetric expression between the ASEL and ASER neurons in promoter-fusion reporter gene assays, as reported by (*66*). Point colors show the expression bias observed in the reporter gene assay for a given gene. The x-axis and y-axis show the log-transformed, size-factor normalized mean number of unique molecular identifiers observed for a given gene per ASEL and ASER cell respectively in the sci-RNA-seq data.

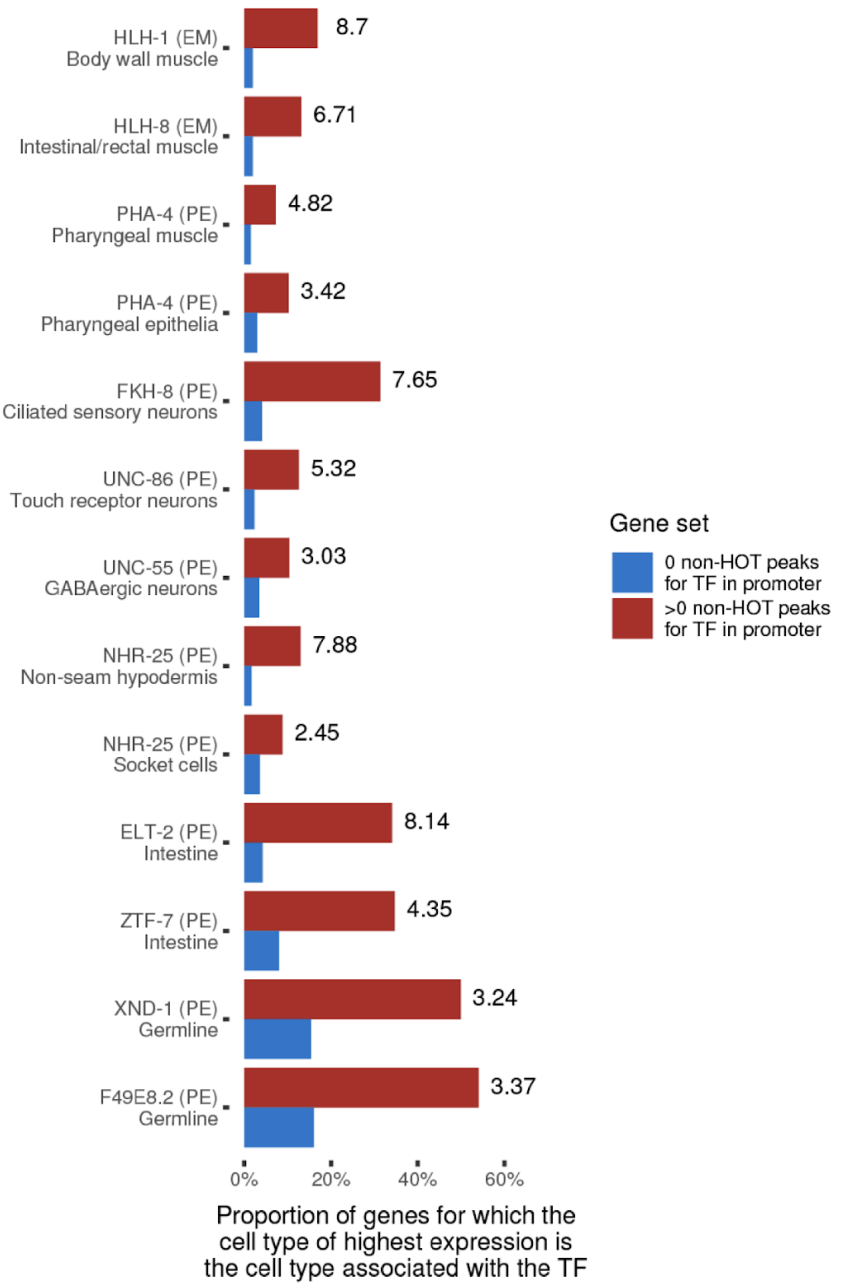


Fig. S12. Transcription factor ChIP-seq peaks predict cell type enriched gene expression. For many TF-to-cell-type associations, the presence of a ChIP-seq peak for the TF in the promoter of a given gene substantially increases the likelihood of the associated cell type being the cell type in which the gene is most highly expressed. Red bars show this probability for genes with at least one peak for the listed TF in their promoter; blue bars show the probability for genes with no peak for the TF in their promoter. Numbers next to the red bars show the ratio of the probabilities for genes with >0 vs. 0 peaks for the TF in their promoter. The associations here are selected examples, each having a positive coefficient in **Fig. 5**. A "PE" following a TF name indicates that the ChIP-seq dataset(s) for that TF are from post-embryonic worms; "EM" indicates that they are from embryos. "HOT region" peaks, defined as those which overlap peaks >20% of all TFs assayed in the same broad developmental stage (embryonic or post-embryonic), are excluded from the analysis.

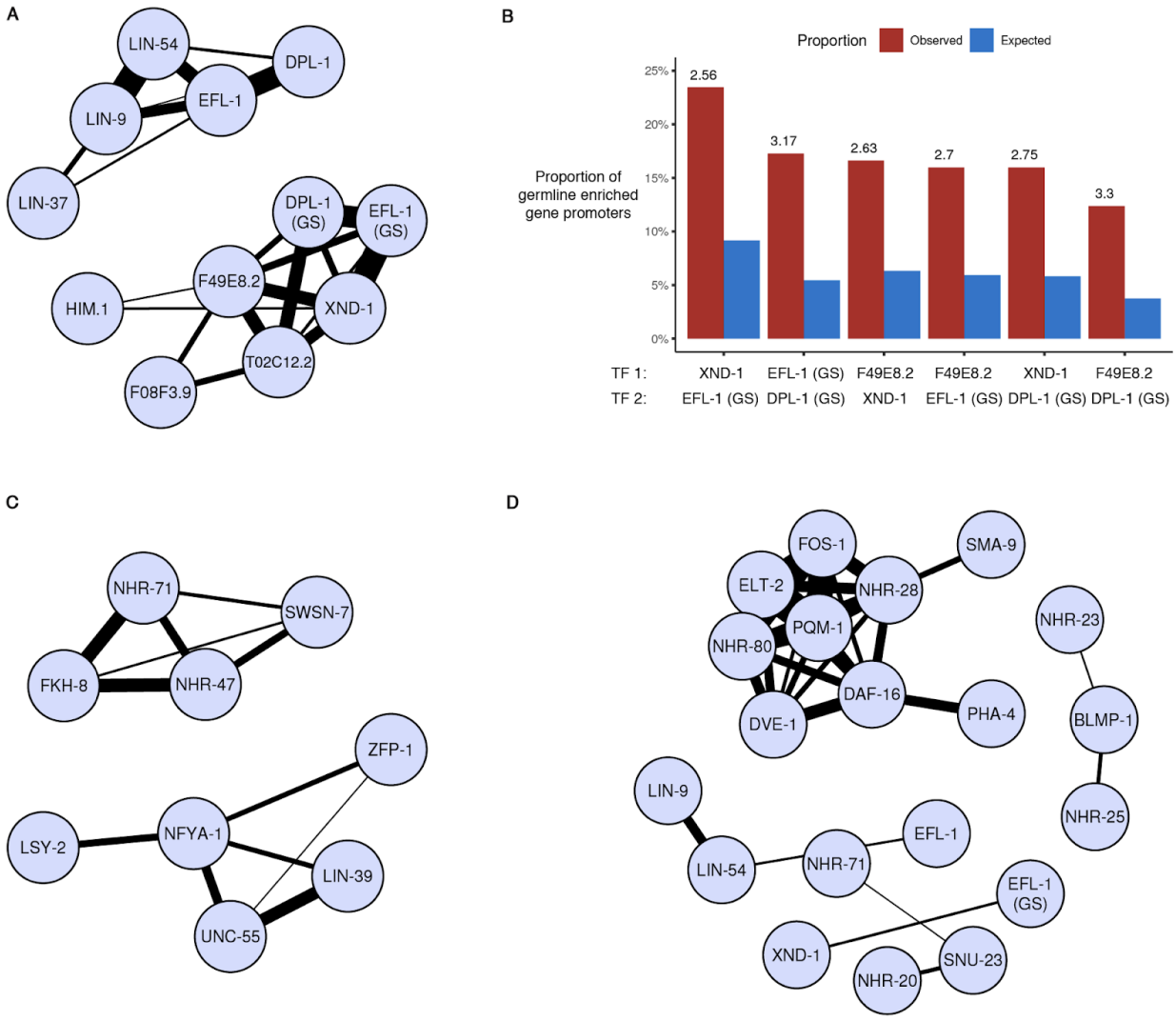


Fig. S13. Transcription factor ChIP-seq peaks have distinct co-localization patterns in the promoters of genes with tissue-enriched expression patterns. (**A, C, D**) A Graphical LASSO model (see **Methods**) is used to find pairs of transcription factors which have overlapping ChIP-seq peaks more often than could be expected by chance, in the context of (A) the promoters of genes with gonad-enriched expression (>5-fold greater in gonad than in any other tissue), (C) the promoters of genes with neuron-enriched expression, or (D) the promoters of all genes. All TF ChIP-seq in this analysis is from post-embryonic stages. EFL-1 (GS) and DPL-1 (GS) refer to peaks from germline-specific ChIP-seq datasets from (*80*). EFL-1, DPL-1, LIN-9, LIN-37, and LIN-54 are members of the DRM complex (*C. elegans* ortholog of the mammalian DREAM complex), which activates a subset of genes in the germline while repressing them in soma (*80*, *102–104*). (**B**) The observed proportion of germline-enriched genes (those with germline expression >5-fold higher than in any other cell type) that have peaks for both listed TFs in their promoter (in red), compared to the proportion that would be expected if the TF binding patterns were independent conditional on being in a germline-enriched gene promoter (in blue). The numbers above each red bar is the ratio of observed / expected. The conditioning of these statistics on the context of being in a germline-enriched gene promoter rules out the possibility that the co-localizations observed in (A) are simply due to each TF independently being associated with germline-specific genes.

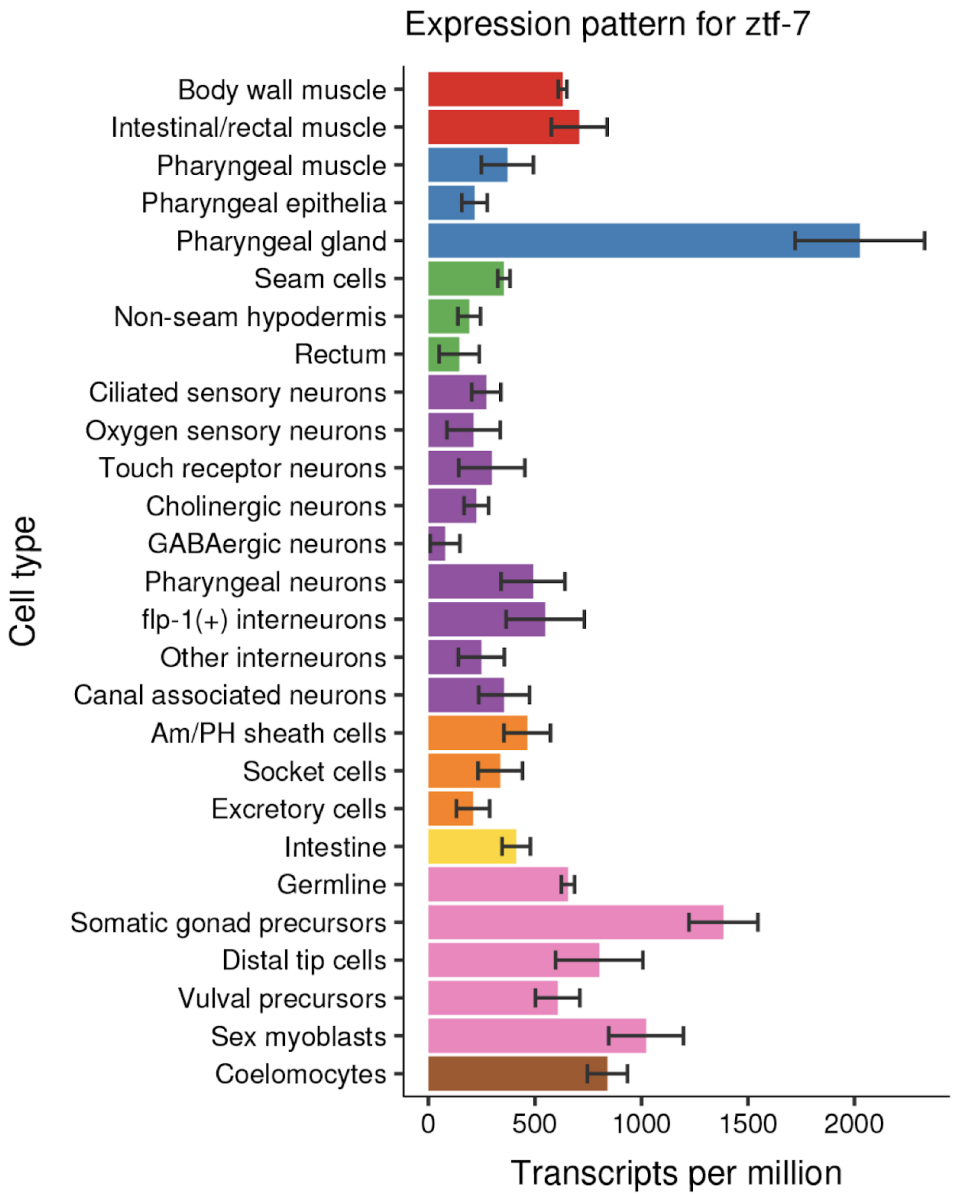


Fig. S14. Example of a "gene expression report" image, with the full set hosted on GExplore. For a given gene, mean expression values are shown for each of 27 cell types. Black bars indicate the 95% confidence interval. All gene profiles are viewable at: http://genome.sfu.ca/gexplore/gexplore_search_tissues.html.

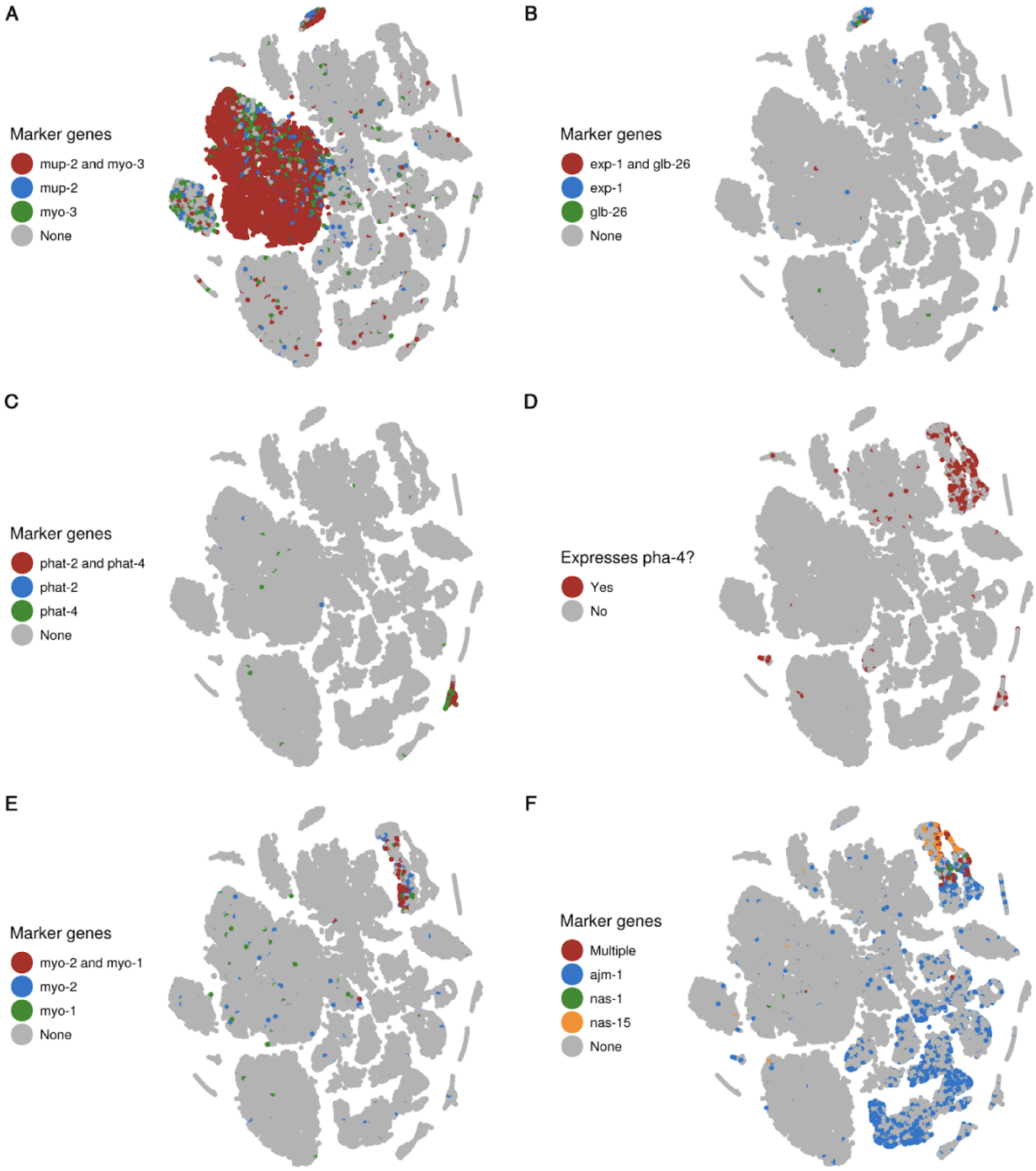


Fig. S15. Expression patterns of marker genes for body wall muscle, intestinal/rectal muscle, and pharynx. (**A**) *mup-2* (troponin T) and *myo-3* (myosin heavy chain A) expression identifies body wall muscle and intestinal/rectal muscle cells (*105*). The cluster to the left of the large muscle cluster are low UMI-count cells that we believe to be damaged body wall muscle cells. They were excluded from downstream analysis. (**B**) *exp-1* and *glb-26* expression distinguishes intestinal/rectal muscle cells from body wall muscle (*106*, *107*). (**C**) *phat-2* and *phat-4* expression identifies pharyngeal gland cells (*108*). (**D**) *pha-4* expression identifies a cluster (top right) of non-gland pharyngeal cells (*71*). The small *pha-4*(+) cluster on the left are distal tip cells (see **Fig. S18B**). (*Legend continued on the following page*)

Fig. S15 (continued). (**E**) *myo-1* and *myo-2* expression identifies pharyngeal muscle cells (*109*). For the purpose of constructing consensus expression profiles, cells in this t-SNE cluster were considered pharyngeal muscle if they expressed at least two of *myo-1*, *myo-2*, *myo-5*, *tnt-4*, *mlc-1* or *mlc-2*. (**F**) *ajm-1*, *nas-1*, and *nas-15* expression identifies non-muscle epithelial cells in the pharyngeal t-SNE cluster. *ajm-1* is expressed in all epithelial cells, while *nas-1* and *nas-15* are specific to the pharynx (*110*, *111*). For the purpose of constructing consensus expression profiles, cells in the pharyngeal muscle/epithelial t-SNE cluster were considered to be epithelial if they do not express any of the markers listed in (E) and expressed at least one of *ajm-1*, *sma-1*, *nas-1*, *nas-15*, or *ifa-1*.

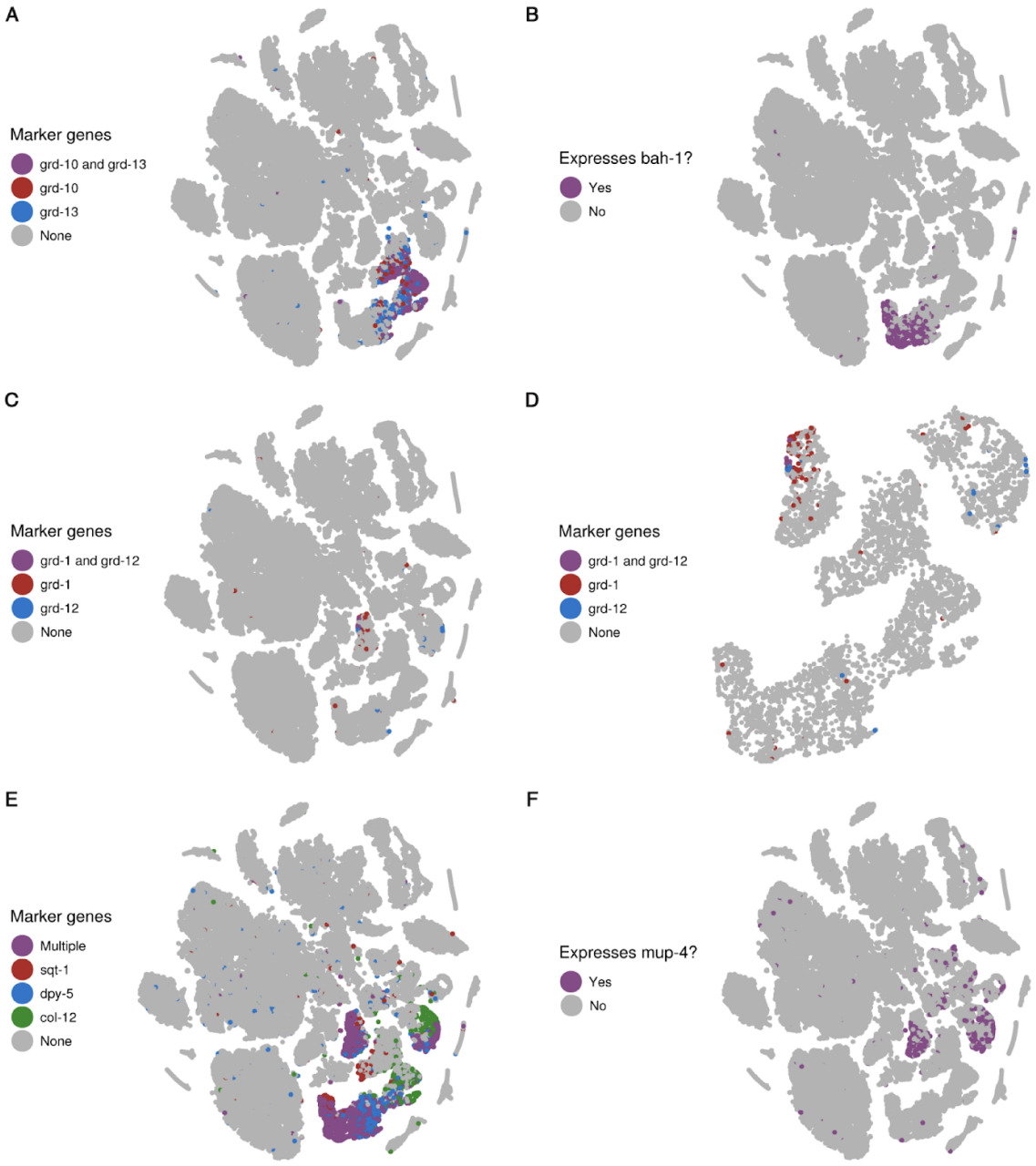


Fig. S16. Expression patterns for marker genes for hypodermis and the rectum. (**A**) *grd-10* and *grd-13* expression identifies seam cells (*112*). (**B**) *bah-1* expression identifies additional seam cells (*113*) and shows that the t-SNE cluster with *grd-10/13* expression is likely to be entirely seam cells. This cluster also expresses seam cell

specific transcription factors including *ceh-18* and *nhr-73*. (**C, D**) *grd-1* and *grd-12* expression identifies rectal cells. *grd-1* is expressed in the rectal gland cells (*114*), while *grd-12* is expressed in the B and Y rectal epithelial cells (*112*) (D) is a zoomed-in view of the hypodermal cell clusters in (C). (**E**) Expression of the cuticle collagen genes *sqt-1*, *dpy-5*, *col-12* identify hypodermal cells (*115*), including two clusters of non-seam hypodermal cells. We were unable to clearly identify the anatomical differences between the cells in the two non-seam hypodermal clusters. (**F**) Expression of *mup-4* is exclusive to non-seam hypodermis and glia, consistent with previous reports of its expression in the circumferential rings of the cuticle (*116*).

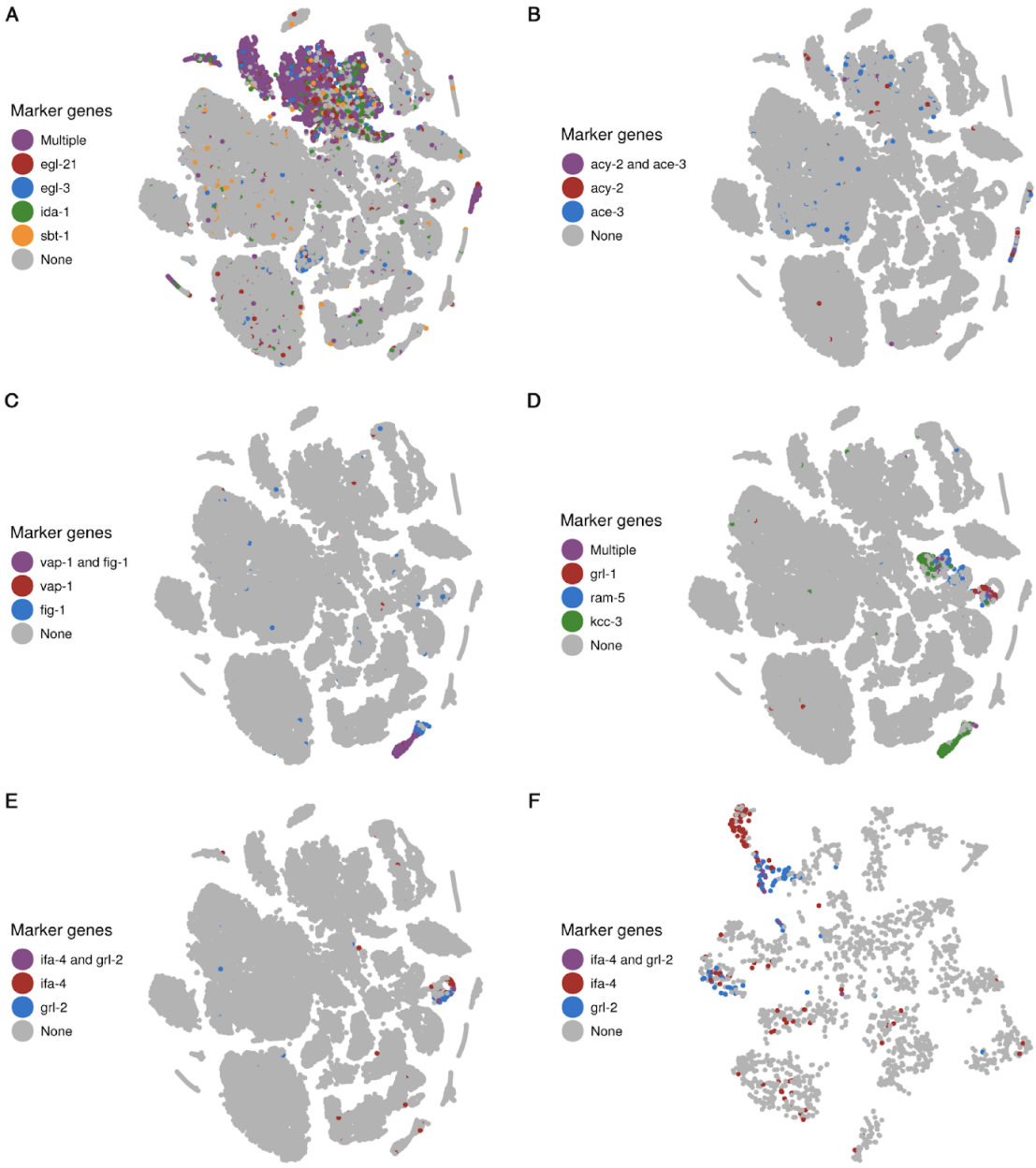


Fig. S17. Expression patterns of marker genes for neurons, glia, and excretory cells. (**A**) Expression of *egl-21*, *egl-3*, *ida-1*, and *sbt-1* identifies neuronal cells (*117–120*). (**B**) The canal associated neurons do not express the marker genes listed in (A), but are identified by their expression of *acy-2* and *ace-3 (121, 122)*. (**C**) Expression of *vap-1* and *fig-1* identifies the amphid and phasmid sheath cells (*100*). (**D**) Expression of *grl-1* and *ram-5* identifies

socket cells (*112*, *123*). Expression of *kcc-3* outside the amphid/phasmid sheath cell cluster identifies additional sheath cells (*124*). For the purpose of constructing consensus expression profiles, cells in the non-amphid/phasmid sheath glial t-SNE clusters were considered to be socket cells if they were not identified to be excretory cells, expressed at least one of *grl-1*, *grd-15*, *daf-6*, or *ram-5*, and did not express *kcc-3*. (**E**) Expression of *ifa-4* and *grl-2* identifies excretory cells (*112*, *125*). (**F**) *ifa-4*(+) and *grl-2*(+) cells cluster together in a t-SNE of only cells from the glial/excretory cell clusters. We suspect that the *ifa-4*(+) cluster at the top corresponds to the excretory canal cell, while the *grl-2*(+) cluster corresponds to the excretory duct, pore, and/or gland cells.

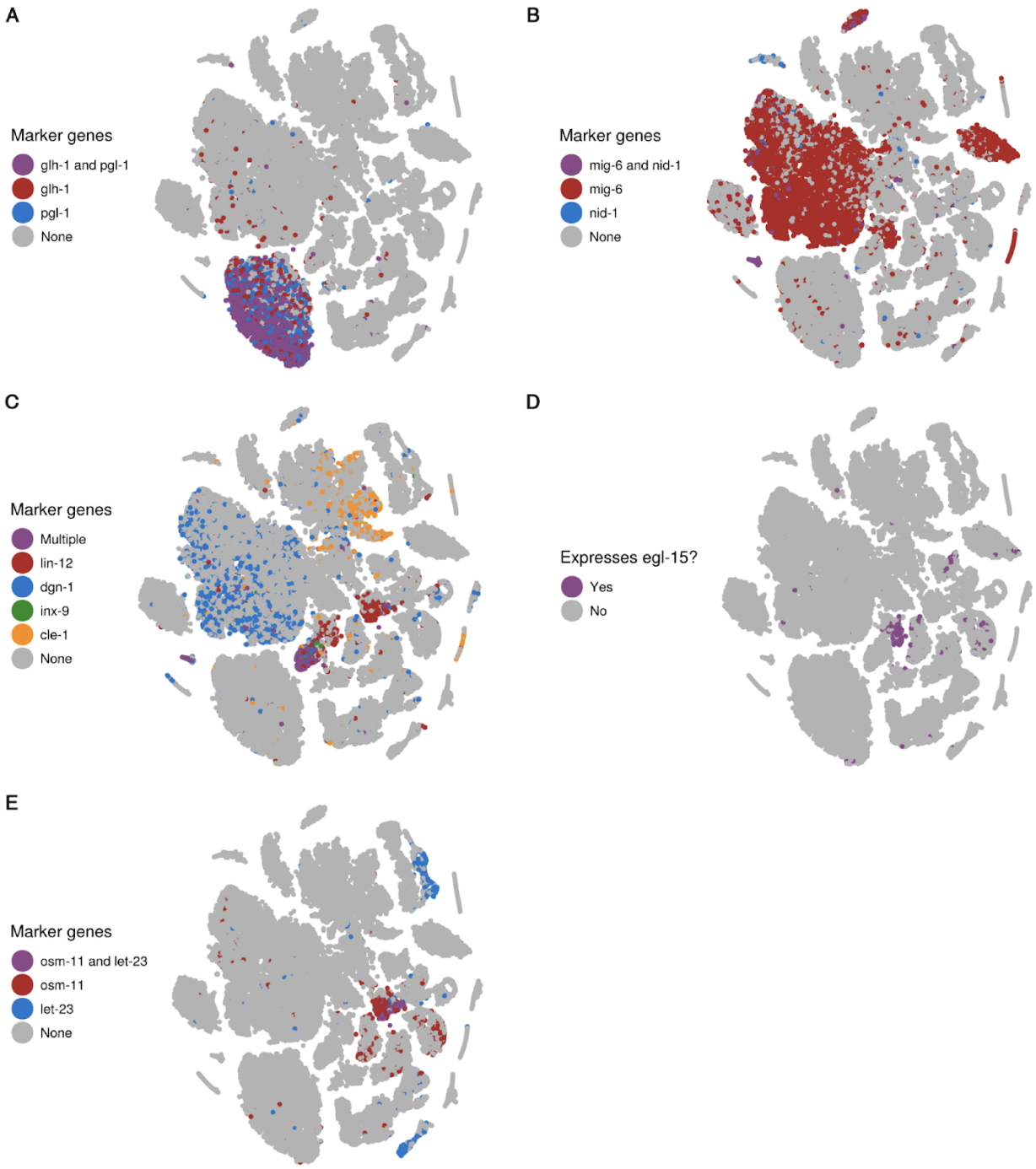


Fig. S18. Expression of marker genes for the germline, somatic gonad, and other sex-related tissues. (**A**) Expression of *glh-1* and *pgl-1* identifies germline cells (*126*, *127*). (**B**) Co-expression of *mig-6* and *nid-1* identifies the distal tip cells of the somatic gonad (small purple cluster on the lower left; (*128*, *129*)). (**C**) Co-expression of at least two of *lin-12*, *dgn-1*, *inx-9*, and *cle-1* identifies the somatic gonad precursor cells (*130–133*). (**D**) Expression of *egl-15* identifies sex myoblasts (*134*). (**E**) Expression of *osm-11* and *let-23* identifies vulval precursor cells (*135*, *136*).

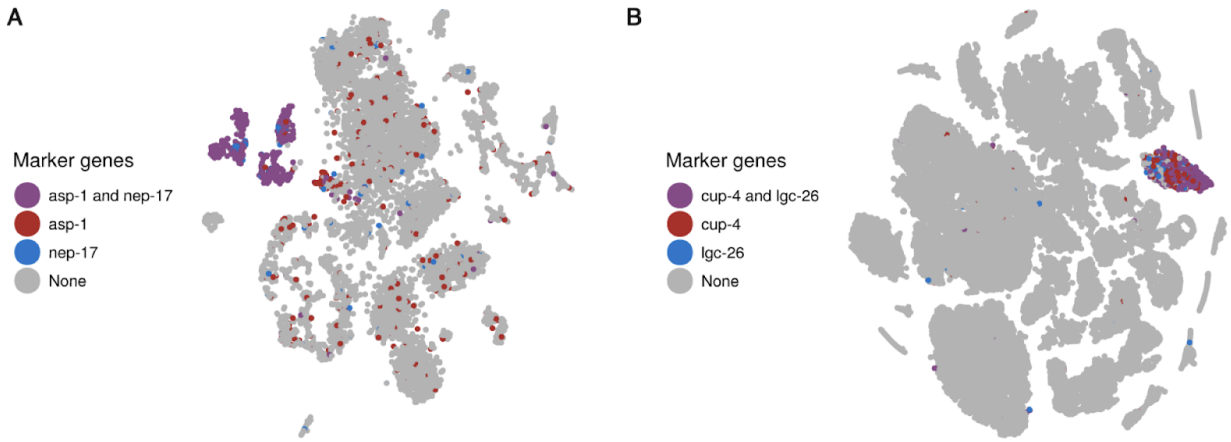


Fig. S19. Expression of marker genes for the intestine and coelomocytes. (**A**) Expression of *asp-1* and *nep-17* identifies intestine cells from the second *C. elegans* experiment (experiment 7 in **Table S1**). (*137*, *138*). (**B**) Expression of *cup-4* and *lgc-26* identifies coelomocytes (*139*).

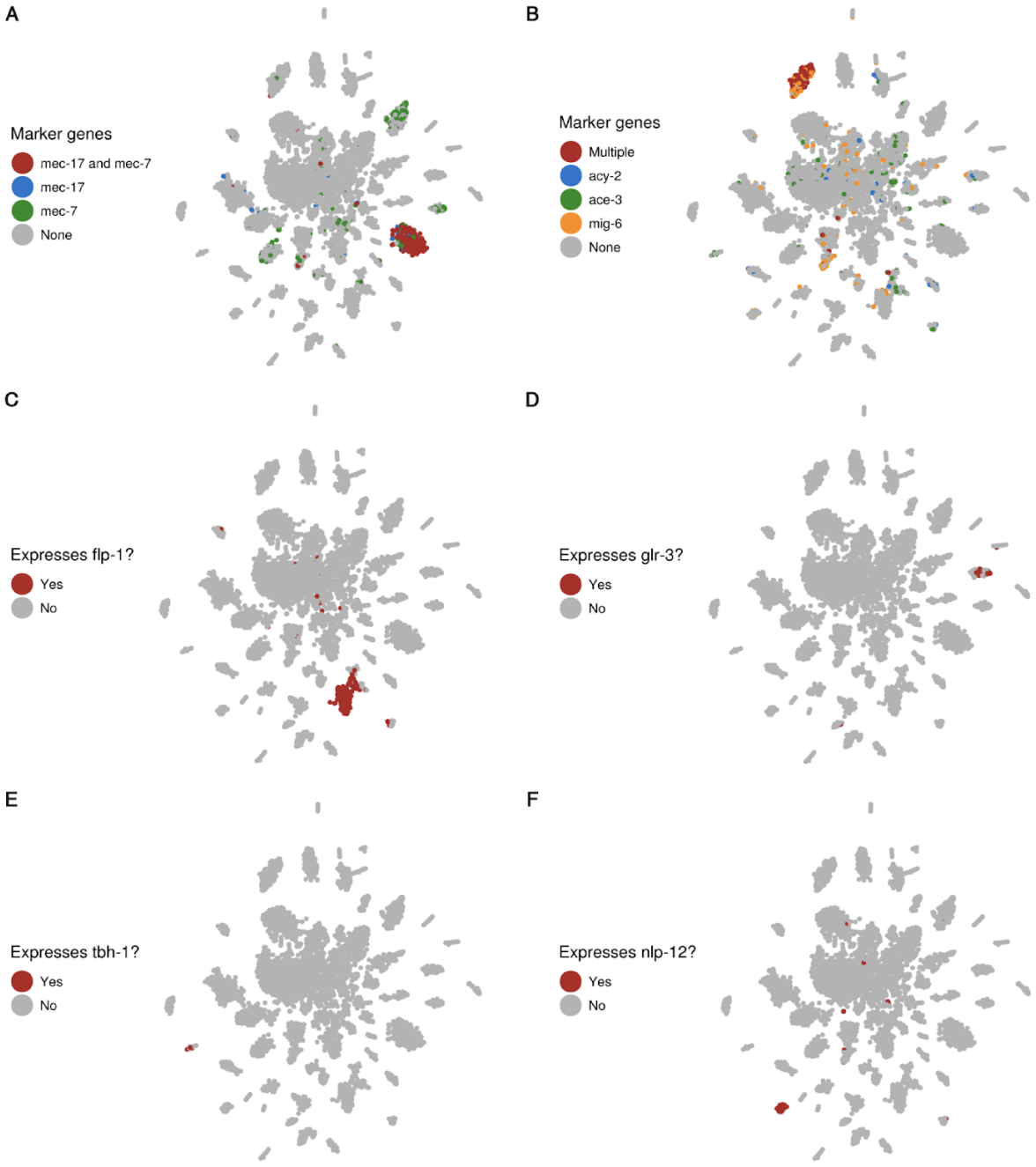


Fig. S20. Expression patterns of marker genes for touch receptor neurons and interneuron subtypes. t-SNE plots shown are from a clustering of just neuronal cells (identified in **Fig. S17A,B**). (**A**) Expression of *mec-17* and *mec-7* identifies touch receptor neurons (*140*). (**B**) Expression of *acy-2* and *ace-3* identifies canal associated neurons (*121*, *122*). The canal associated neurons are also the only neuron class that expresses *mig-6 (141)*. (**C**) *flp-1* expression identifies interneurons of the anatomical classes AVK, AVA, AVE, RIG, RMG, AIY, AIA (*142*). *flp-1* has also been reported to be expressed in the M5 pharyngeal motor neuron. (**D**) *glr-3* is expressed exclusively in the RIA interneurons (*143*). (**E**) Among neurons, *tbh-1* is expressed exclusively in the RIC interneurons (*144*). (**F**) *nlp-12* expression identifies the DVA tail interneuron (*145*).

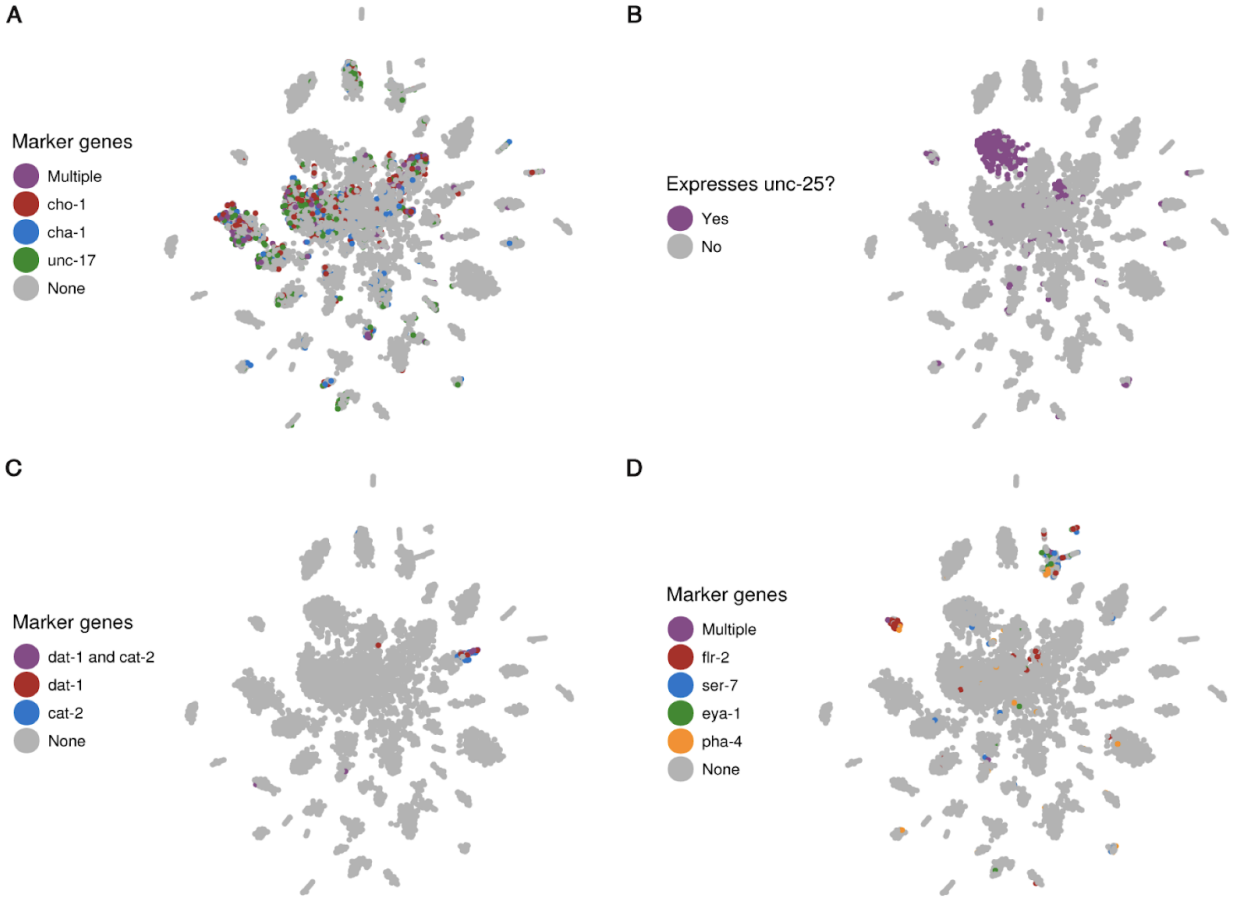


Fig. S21. Expression of marker genes for cholinergic, GABAergic, dopaminergic, and pharyngeal neurons. t-SNE plots shown are from a clustering of just neuronal cells (identified in **Fig. S17A,B**). (**A**) Expression of *cho-1*, *cha-1*, and *unc-17* identifies cholinergic neurons (*146*). For the purpose of constructing consensus expression profiles, neuronal cells were identified as cholinergic if they were not part of a t-SNE cluster identified as any other neuronal subtype and they expressed at least one of *cho-1*, *cha-1*, *unc-17*, *acr-15*, or *acr-18*. (**B**) *unc-25* expression identifies GABAergic neurons (*147*). (**C**) Expression of *dat-1* and *cat-2* identifies dopaminergic neurons (*148*, *149*). (**D**) While no single marker is both highly expressed and specific to pharyngeal neurons, the expression patterns of *flr-2*, *ser-7*, *eya-1*, and *pha-4* together identify two clusters as highly likely to correspond to pharyngeal neurons (*71*, *150–152*).

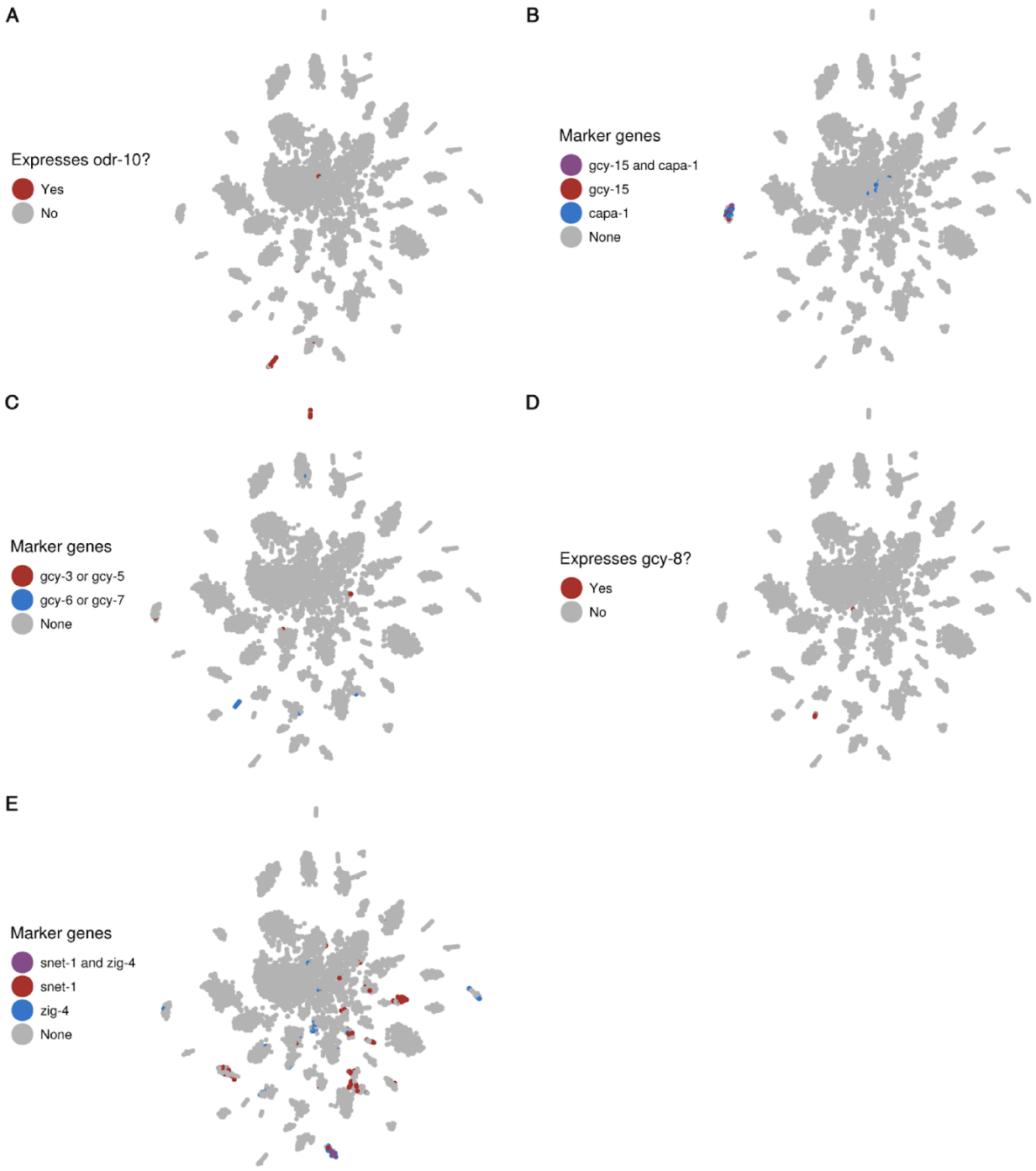


Fig. S22. Expression patterns of marker genes for the AWA, ASG, ASE, AFD, and ASK neurons. t-SNE plots shown are from a clustering of just neuronal cells (identified in **Fig. S17A,B**). (**A**) *odr-10* expression identifies the AWA neurons (*153*). (**B**) *gcy-15* expression identifies the ASG neurons (*154*). *capa-1* has also been reported to be expressed in two specific but unidentified pairs of neurons in the head (*155*); in our data it is expressed predominantly in the same cluster as *gcy-15*. (**C**) Expression of *gcy-3* and *gcy-5* identifies the ASER neuron, while expression of *gcy-6* and *gcy-7* identifies the ASEL neuron (*65*, *66*). (**D**) *gcy-8* expression identifies the AFD neurons (*156*). (**E**) Co-expression of *snet-1* and *zig-4* identifies the ASK neurons (*157*, *158*).

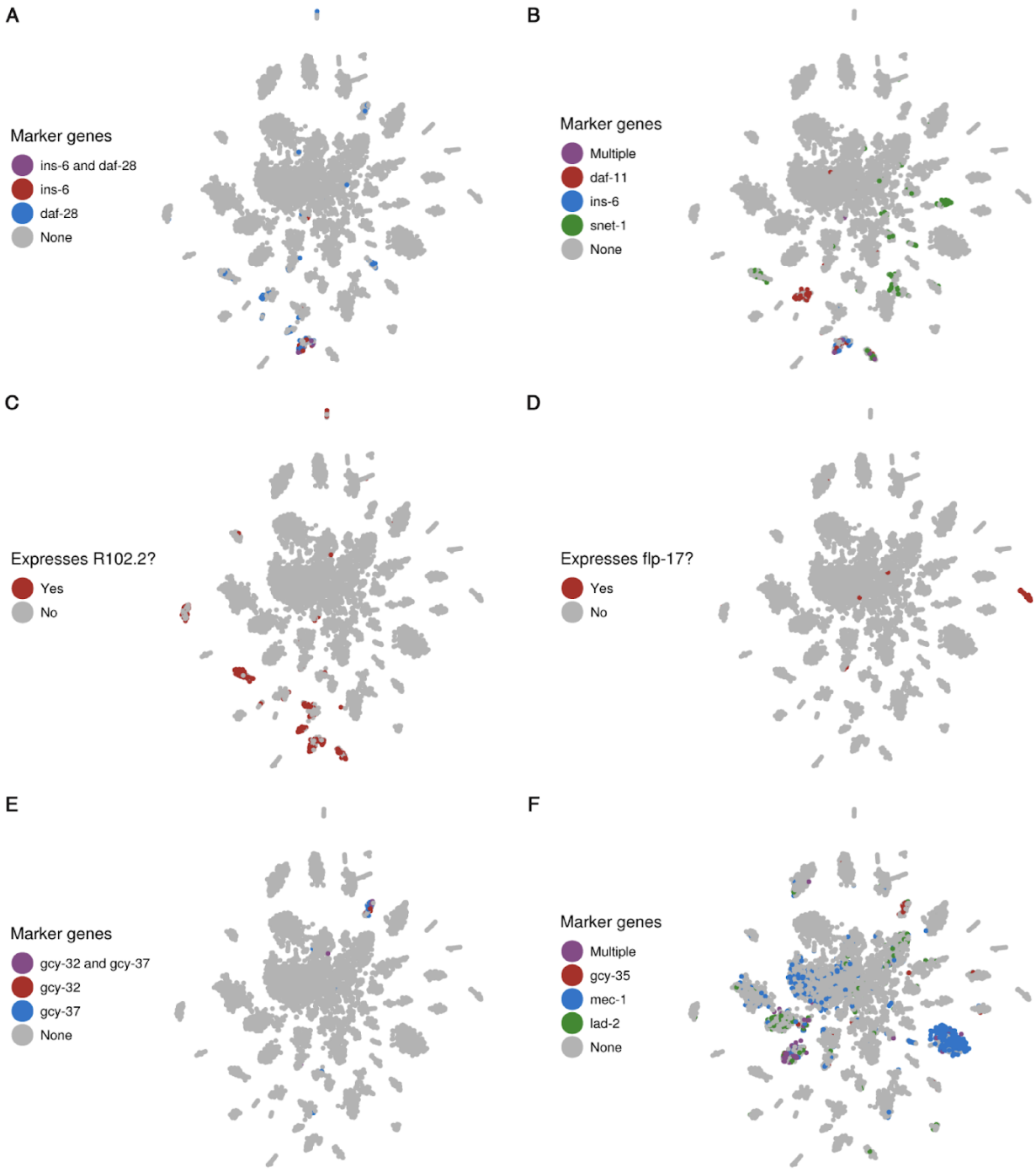


Fig. S23. Expression patterns of marker genes for ASI/ASJ, AWB/AWC, BAG, URX, SDQ, and other ciliated sensory neurons. t-SNE plots shown are from a clustering of just neuronal cells (identified in **Fig. S17A,B**). (**A**) Expression of *ins-6* and *daf-28* identifies a neuron cluster that consists of the ASI and ASJ neurons (*159*, *160*). (**B**) Based on reported expression patterns, a neuron cluster that expresses *daf-11* but not *ins-6* or *snet-1* can only correspond to the AWB and/or AWC neurons (*157*, *159*, *161*). (**C**) Beyond those identified in **Fig. S22,** and (A) of this figure, three additional neuron clusters express *R102.2*. Based on the expression patterns reported by (*162*), these clusters correspond to the ciliated sensory neurons classes ADF, ASH, PHA, and/or PHB. We could not precisely identify them however. (*Legend continued on the following page*)

Fig. S23 (continued). For the purpose of constructing consensus expression profiles, neuronal cells were considered ciliated sensory neurons if they either were part of a cluster that was identified as a ciliated sensory neuron class or were part of a cluster that could not be conclusively identified but expressed high levels of *R102.2*, *dyf-2*, *che-3*, or *nphp-4*. (**D**) *flp-17* expression identifies the BAG neurons (*142*). (**E**) Expression of *gcy-32* and *gcy-37* identifies a neuron cluster that consists of the URX, AQR, and PQR neurons (*163*, *164*). (**F**) Among neurons, *gcy-35* is expressed in the URX, AQR, PQR, SDQ, ALN, PLN O_2-sensory neurons, as well as the AVM and BDU neurons (*164*). *mec-1* was reported to be expressed in the touch receptor neurons, SDQ/ALN/PLN O_2-sensory neurons, and PVT neurons (*165*). *lad-2* was reported to be expressed in the SDQ/ALN/PLN O_2-sensory neurons and some sublateral motor neurons (*166*). Based on these expression patterns, a neuron cluster enriched for expression of all three of these genes is likely to correspond to the SDQ/ALN/PLN O_2-sensory neurons.

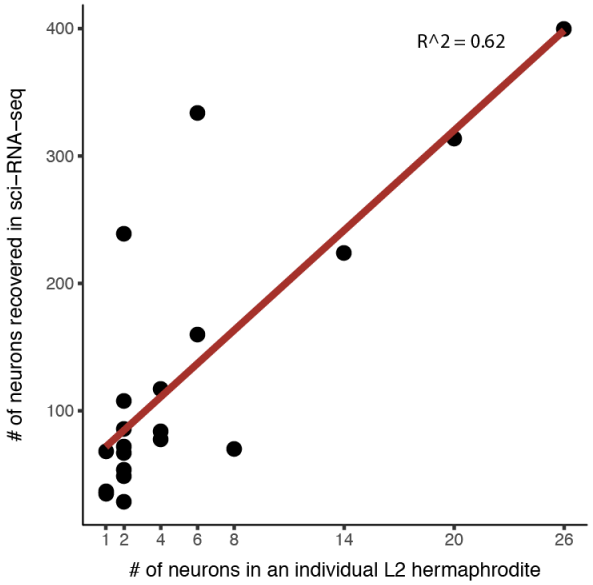


Fig. S24. Recovery rates of neuron types in sci-RNA-seq. The observed number of cells identified in sci-RNA-seq for a given neuron type (y axis) is compared to the number of neurons of that type in an individual L2 hermaphrodite *C. elegans* (x-axis). The plot includes all specific neuron types that we were able to identify, excluding cholinergic neurons, which were not limited to distinct t-SNE clusters and therefore may be under-counted as we only considered a cell cholinergic if we observed expression of at least one cholinergic marker gene (see **Fig. S21**). The neuron types included in the plot are: ASEL, ASER, DVA, AFD, ASG, ASK, AWA, BAG, CAN, RIA, RIC, ASI/ASJ, AWB/AWC, URX/AQR/PQR, SDQ/ALN/PLN, touch receptor neurons (ALM/PLM/AVM/PVM), dopaminergic neurons (CEP/ADE/PDE), *flp-1*(+) neurons (excluding the pharyngeal neuron M5), pharyngeal neurons, and GABAergic neurons.

Supplementary Tables

Tables S1-S14 are available online at:
www.sciencemag.org/content/357/6352/661/suppl/DC1

Data Availability

The raw data have been deposited with the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) under accession code GSE98561.

Project Acknowledgements

We thank members of the Shendure, Trapnell, and Waterston laboratories for helpful discussions and feedback, particularly A. Hill, V. Agarwal, M. Gasperini, L. Starita, Y. Yin, and B. Martin; S. Zimmerman and C. Berg for helpful technique suggestions; the modERN consortium for allowing us to use their ChIP-seq data; D. Prunkard and L. Gitari in the Pathology Flow Cytometry Core Facility for their exceptional assistance in flow sorting; the T. Reh laboratory for sharing the NIH/3T3 cell line; and H. Hutter for adding our tissue-specific expression profiles to gExplore. HeLa S3 cells were used as part of this study. Henrietta Lacks, and the HeLa cell line that was established from her tumor cells in 1951, have made significant contributions to scientific progress and advances in human health. We are grateful to Henrietta Lacks, now deceased, and to her surviving family members for their contributions to biomedical research. This work was funded by grants from the NIH (DP1HG007811 and R01HG006283 to J.S., U41HG007355 and R01GM072675 to R.H.W, and DP2 HD088158 to C.T.), the Paul G. Allen Family Foundation (to J.S.), the W. M. Keck Foundation (to C.T. and J.S.), the Dale F. Frey Award for Breakthrough Scientists (to C.T.), the Alfred P. Sloan Foundation Research Fellowship (to C.T.), and the William Gates III Endowed Chair in Biomedical Sciences (to R.H.W.). D.A.C. was supported in part by T32HL007828 from the National Heart, Lung, and Blood Institute. J.S. is an investigator of the Howard Hughes Medical Institute. F.J.S. declares competing financial interests in the form of stock ownership and paid employment by Illumina. One or more embodiments of one or more patents and patent applications filed by Illumina may encompass the methods, reagents, and data disclosed in this manuscript.

Chapter 3: A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single cell resolution

Abstract

C. elegans is an animal with few cells, but a striking diversity of cell types. Here, we characterize the molecular basis for their specification by profiling the transcriptomes of 86,024 single embryonic cells. We identify 502 terminal and pre-terminal cell types, mapping most single cell transcriptomes to their exact position in *C. elegans*' invariant lineage. Using these annotations, we 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.

Acknowledgements

The main text of this chapter is adapted with minimal modification from the publication:

> Packer, Zhu *et al.* A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single cell resolution. *Science*, accepted manuscript in press. doi: 10.1126/science.aax1971.

This project was a collaboration with the labs of John Murray and Junhyong Kim at the University of Pennsylvania. Priya Sivaramakrishnan, Elicia Preston, and Hannah Dueck from the Murray lab and Derek Stefanik from the Kim lab produced about half of the data, with the other half being produced by Chau Hunyh (from Bob Waterston's lab at the University of Washington). I led the analysis and drafting of the manuscript. Qin Zhu, a graduate student in the Murray lab, was a co-first author on the paper. Qin made several important contributions to the analysis, including developing an algorithm for estimating the age of the embryo that each sc-RNA-seq cell came from (described below). Qin also developed the web application "VisCello," which allows one to visualize and explore the data from the project. John Murray and I annotated the sc-RNA-seq cells with their corresponding cell types and lineage identities. John annotated all pre-terminal cells (which constitute most of the lineage) and I annotated the terminal cells.

Main Text

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 (*167*). Single cell RNA sequencing (sc-RNA-seq) has made it possible to obtain comprehensive measurements of gene expression in whole animals (*29*, *168–172*) and embryos (*26*, *30*, *173–177*). 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 (*10*, *12*), 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 (*19*, *178*), 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 (*25*, *26*), 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.

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 (*179*). 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 (*20*) (**Fig. S1**). We then visualized the data with the Uniform Manifold Approximation and Projection (UMAP) (*180*, *181*) algorithm, which projects the data into a low-dimensional space and is well suited for data with complex branching structures (*181*). We found that trajectories in the UMAP projection reflect a smooth progression of embryo time (**Fig. 1A**, also see **Supplemental Note 1** for a brief discussion of the term "trajectory"), with cells collected from later time points usually occupying more peripheral positions (**Fig. 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 (**Fig. S2**). 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) (**Fig. 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 (*182*) and annotated clusters with cell type identities using marker genes from the literature on *C. elegans* gene expression (*16*). Markers used for each annotation are listed in **Table S1**. The global UMAP arranges cells into a central group of progenitor cells and branches corresponding to eight major tissues (**Figs. 1A, S3**): 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 (**Figs. S4-S13**). 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

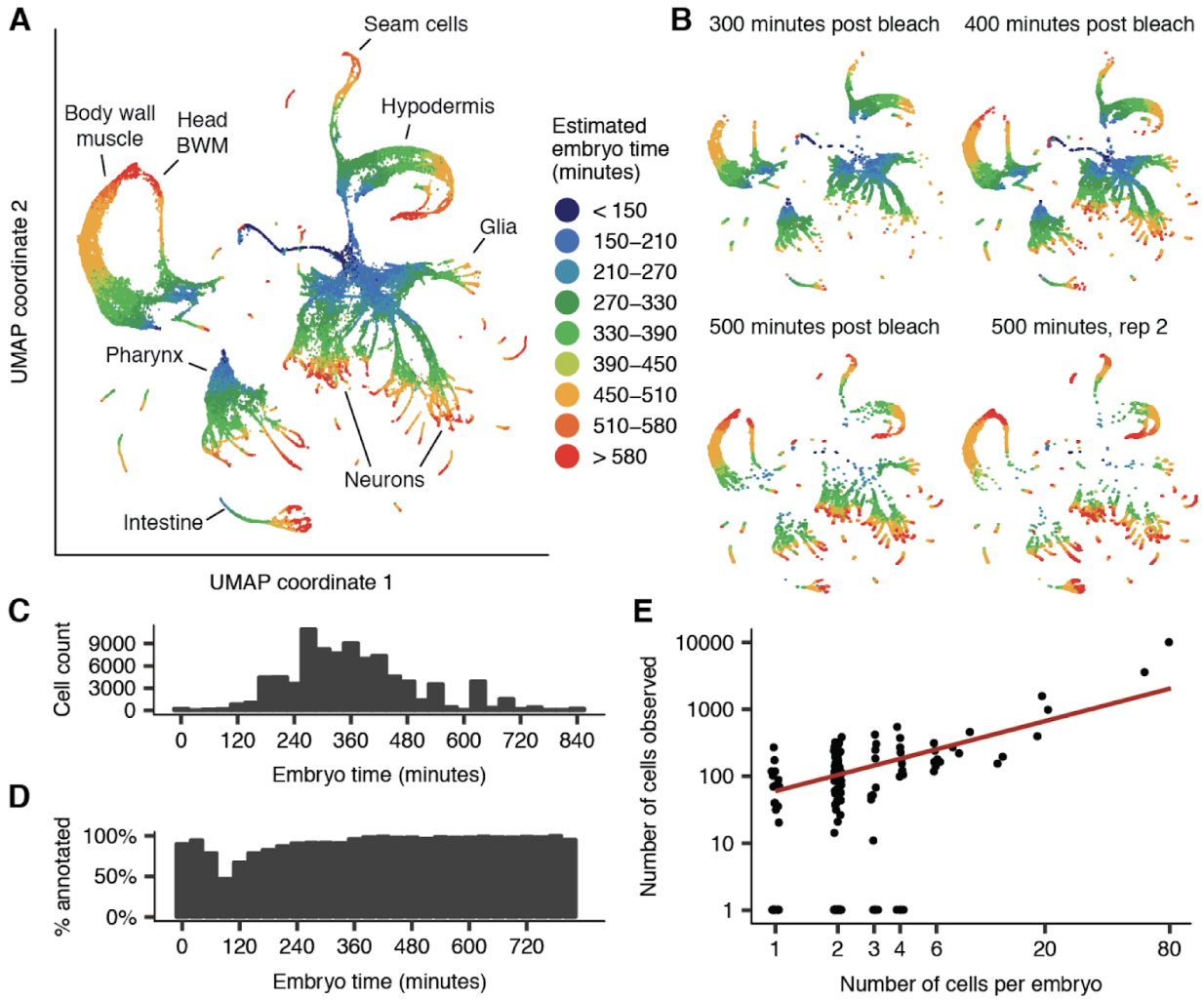


Fig. 1. UMAP projection shows tissues and developmental trajectories in *C. elegans* embryogenesis. (**A**) UMAP projection of the 81,286 cells from our sc-RNA-seq dataset that passed our initial QC. This UMAP does not include 4,738 additional cells that were initially filtered, but were later whitelisted 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 (*20*) 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.

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 (**Table S2**). 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) (**Fig. 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 (**Fig. 1E**, $r = 0.64$, p = 2.4e-13, *t* test).

Annotation of progenitor lineages

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 (**Figs. 2, S14-16**). 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 (**Figs. S7, S12**). 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 (*19*). In addition to the 180 previously described patterns (*19*, *183*), we have collected movies for 71 additional genes, increasing the total number of patterns in EPiC to 251 genes (**Table S3**). 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 (**Figs. 2, S14-16**, **Tables S4-S5).** For example, expression of *ceh-51* is restricted to the MS (mesoderm-producing) lineage (*184*), allowing us to label the single group of *ceh-51(+)* cells in 150-minute UMAP as part of the MS lineage (**Fig. 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) (**Fig. 2C**). We applied this same logic iteratively across the different UMAPs and lineage marker genes to annotate each branch with its lineage identity (**Table S4**).

In most cases, branches in the progenitor lineage UMAPs corresponded directly to sister cells in the lineage (**Fig. 2D, E**), but some branches were unclear or misleading, and marker gene expression was critical to annotate lineages correctly. For example, ABpxpaaaa and ABpxpaapa

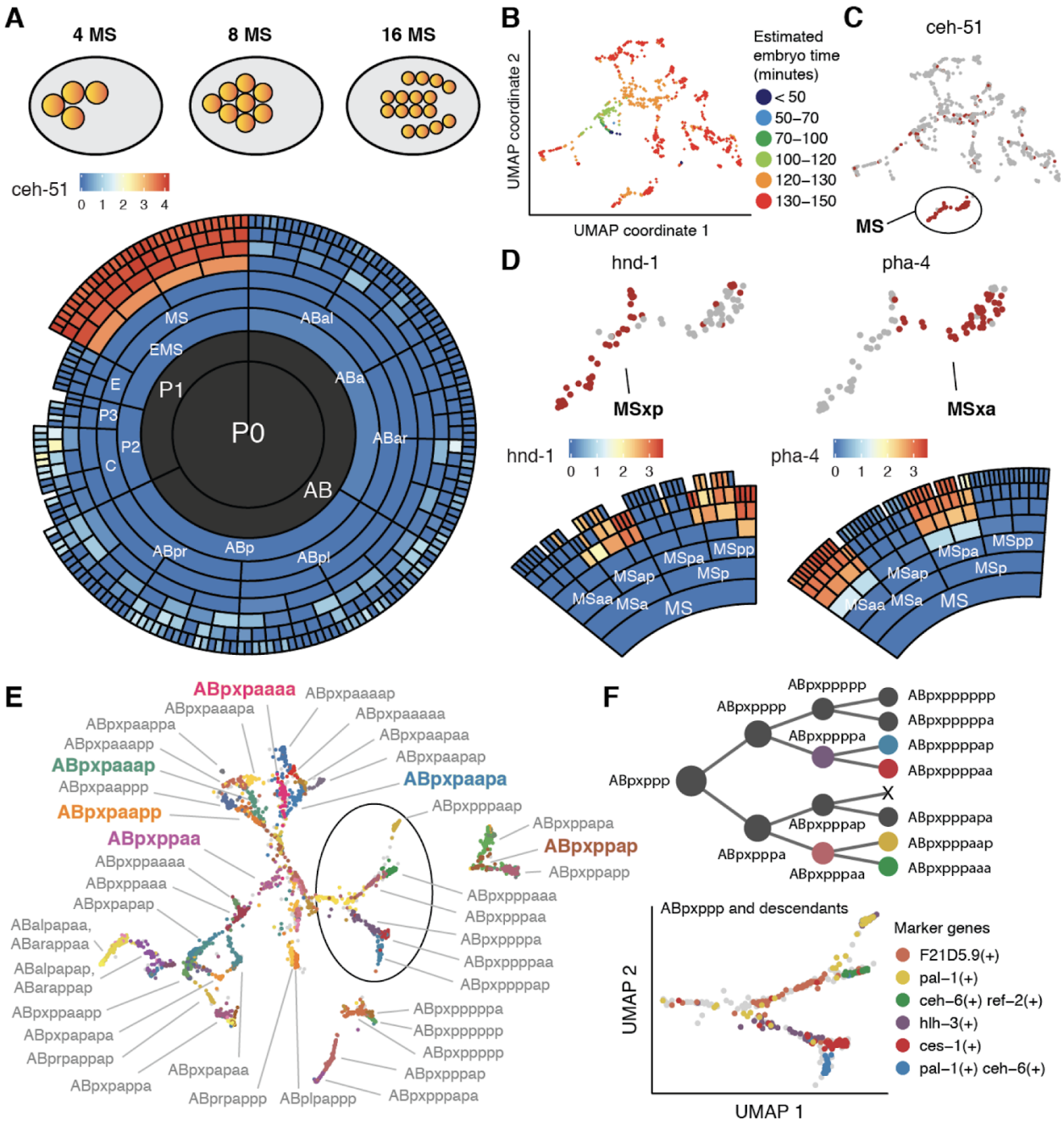


Fig. 2. Annotation of the early lineage. (**A**) Diagram showing the position of early mesoderm (MS lineage) cells marked by expression of *ceh-51*. The lineage radiograph shows the average fluorescent intensity (log10 scaled) of a CEH-51::GFP protein fusion measured by live imaging. The inner rings 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 **Figs S7 and S12**, 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*). (*Legend continued on the following page*)

Fig. 2 (continued). (**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 (**Fig. S15**). 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.

are cousin lineages, but appear to branch as sisters in the UMAP trajectory, and the same is true for their sisters (ABpxpaaap and ABpxpaapp) (**Fig. 2D**)**.** In other cases, such as the ABpxppap lineage (**Fig. 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 (**Fig. S17**). 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.

A near-complete atlas of the embryonic transcriptome

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 (**Fig. S18**), excluding the 113 branches that lead to programmed cell death. Combined with the dataset of Tintori *et al. (26)*, 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 (*26*) at the 16-cell stage. Gene expression profiles for lineages annotated in both datasets were concordant (**Fig. S19**). Additionally, gene expression profiles for terminal cells in our data were concordant with previously published microarray data (*22*) (**Fig. S20**).

In **Table S6**, we provide a map of anatomical cell names to annotations defined in this study. In **Tables S7-8**, we provide aggregate gene expression profiles for each terminal cell type (binned by embryo time) and each cell lineage annotation. We use bootstrap resampling to estimate a confidence interval for the expression level of each gene in each cell type. In **Tables S9-11**, we provide lists of differentially expressed genes between all pairs of sister lineages and

all pairs of parent vs. daughter lineages within our annotations. Lastly, we systematically re-annotated our previous data from the L2 stage (*29*), identifying 118 cell types (over twice as many as reported in the initial publication). **Tables S12-14** list marker genes, annotation statistics, and expression profiles for the L2 data.

Bifurcating cell fates and multilineage 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 (**Fig. 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 (*185*), by the expression patterns of cell cycle associated genes and transcription factors (**Fig. 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 (**Fig. S21**).

To identify potential regulators of cell fate decisions, we identified genes that were differentially expressed between the branches of each bifurcating ciliated neuron lineage (**Table S9**). The lineage of the ASE, ASJ, and AUA neurons (spanning embryo time ~215-650 minutes) serves as a representative example (**Fig. 3C**). About 3-4 TFs are specific to each terminal neuron type in this lineage (**Fig. 3D**). Similar numbers of branch-specific TFs were observed for other lineage bifurcations (**Fig. S22**). 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 (**Fig. S23**).

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*, *186–190*). 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 (**Fig. S24**), 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* (**Fig. S23D**), are confirmed by the movies in EPiC (*19*).

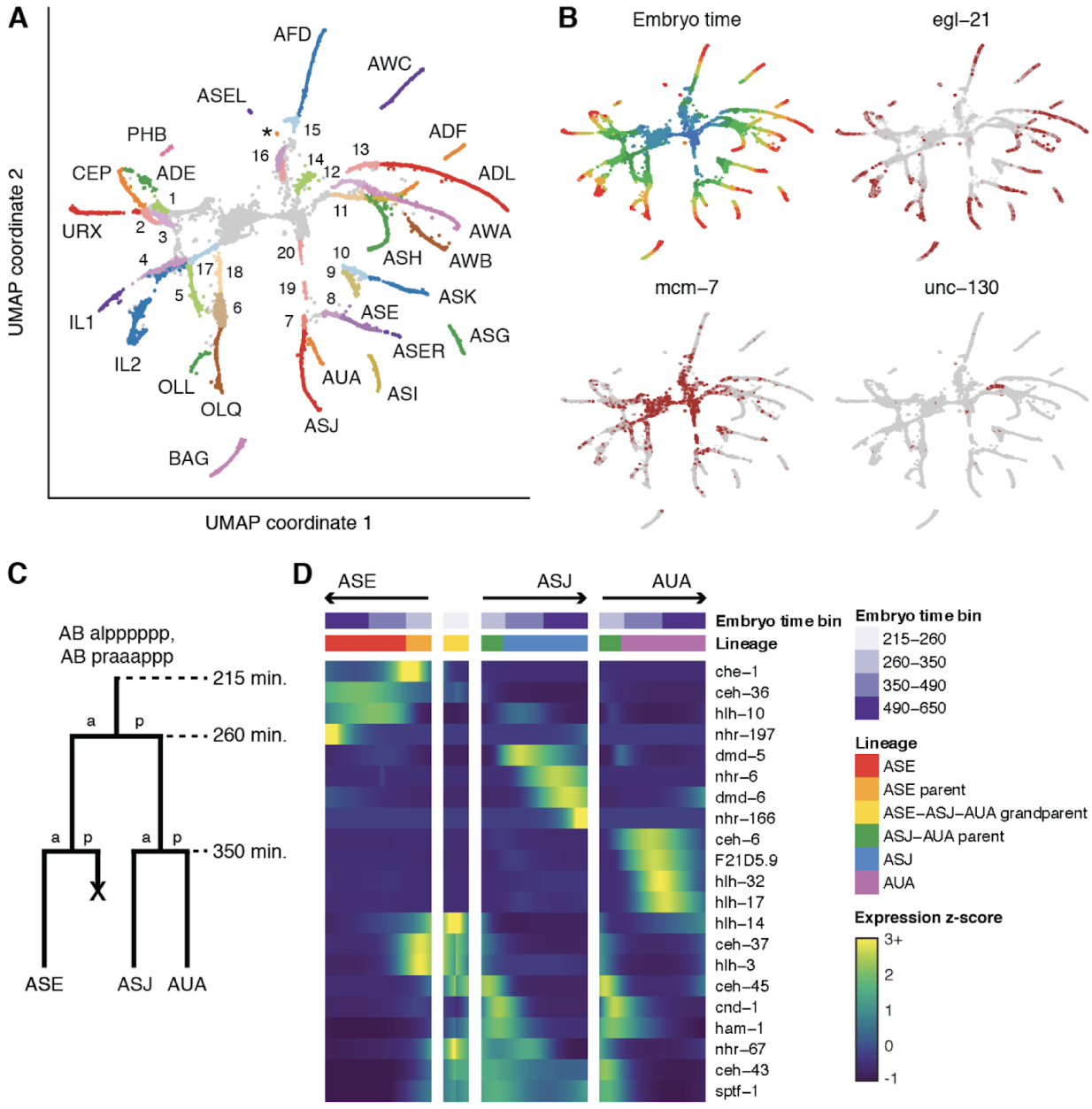


Fig. 3. Developmental trajectories of ciliated neurons. (**A**) UMAP of 10,740 ciliated neurons and precursors. Colors correspond to cell identity. Text labels indicate terminal cell types. Numbers 1-16 indicate parents of **1** ADE-ADA, **2** CEP-URX **3** PHB-HSN **4** IL1 **5** OLL **6** OLQ **7** 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. (*Legend continued on the following page*)

Fig. 3 (continued). 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 **Fig. S3**. 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 **Fig. 1A**) and gene expression (red indicates >0 reads for the listed gene). *egl-21* codes for an enzyme that is essential for processing neuropeptides (*191*). 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 (*192*). (**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).

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" (*193*). 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 (*193*) and identified 116 novel associations (**Table S15**). 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 (*163*, *194*). 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 (*195*, *196*). AWC left/right asymmetry occurs stochastically, with one neuron becoming "AWC-ON" and the other becoming "AWC-OFF" (*194*). We identified a small cluster in the UMAP with embryo time >700 minutes as AWC-ON based on *srt-28* expression (**Fig. 3A**) (*197*). 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 (**Fig. 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 (**Fig. 4B**). Correspondingly, expression of genes differentially expressed between the input lineages decreased over time, while expression of genes specific to terminal neurons increased (**Fig. 4C-D**). We observed similar lineage convergence via continuous gene expression trajectories for other cell types, including hypodermis (**Fig. S8**), head body wall muscle (**Fig. S17**), and GLR cells (**Fig. S17**).

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 (**Fig. S25**). 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) (**Fig. S25**). 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 (**Fig. S26**). 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 (**Fig. S26**). 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 (**Fig. S27**). 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

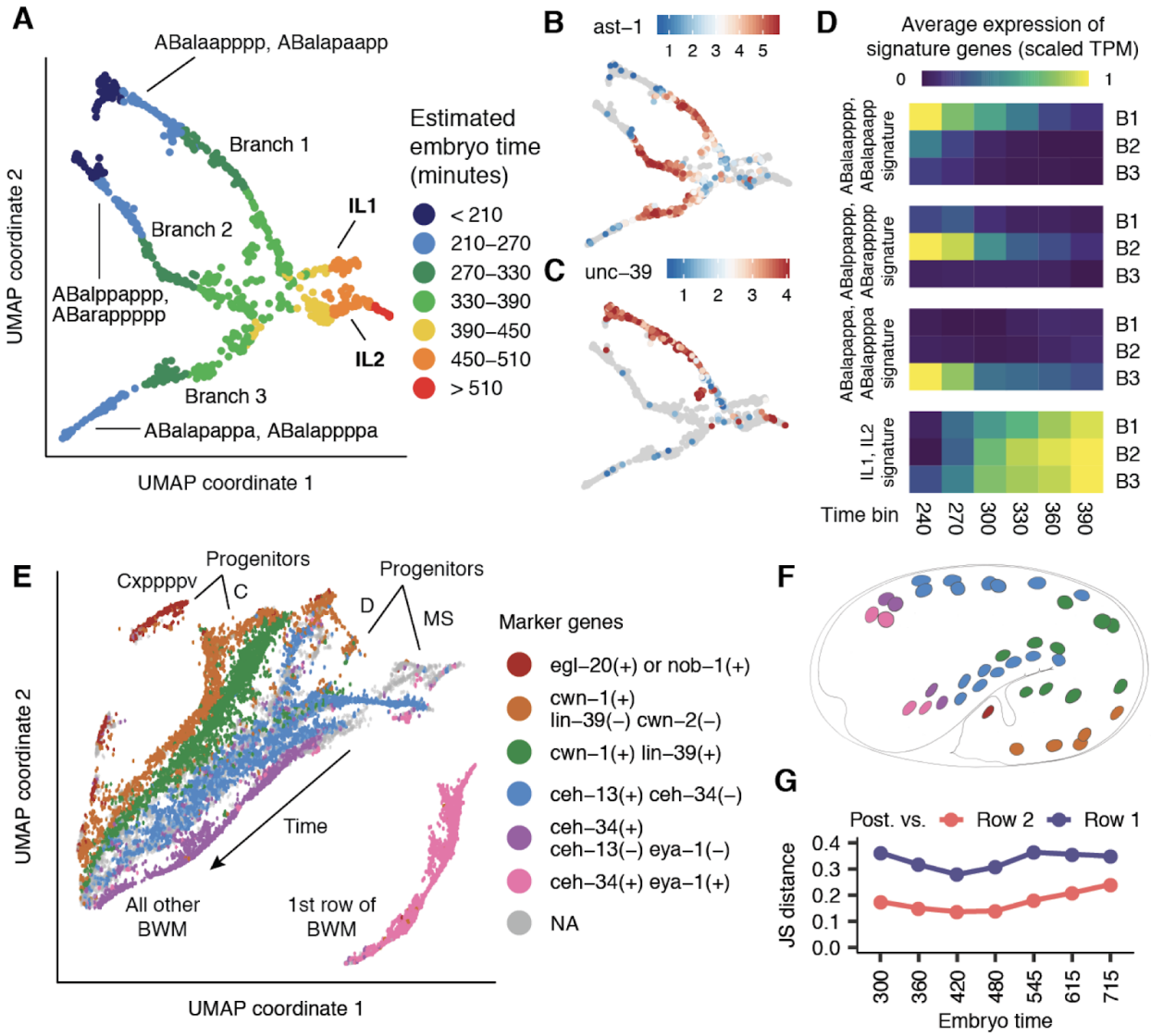


Fig 4. Full vs. incomplete convergence of lineages producing common cell types. (**A**) UMAP of 854 IL1/2 neurons and progenitors colored by estimated embryo time (cells selected on the basis of annotations in **Figs. 3A and S15**). (**B**) IL1/2 UMAP colored by *ast-1* expression level (log2 size-factor 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**) **Fig. S5A** 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 (*12*). (**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.

BWM (**Fig. 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 (*19*, *98*, *198*, *199*), we divided BWM cells in the UMAP into six "bands" (**Fig. 4E**) and identified the specific anatomical cells present in each band (**Fig. 4F**, **Table S16**). 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 (**Fig. 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 (**Fig. 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 (**Fig. S28**). 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 (**Fig. S29A**).

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 (see **Methods**). 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 (**Fig. 5B**). Next, we asked how much of the total pseudo-R^2 for one cell generation was attributable to gene

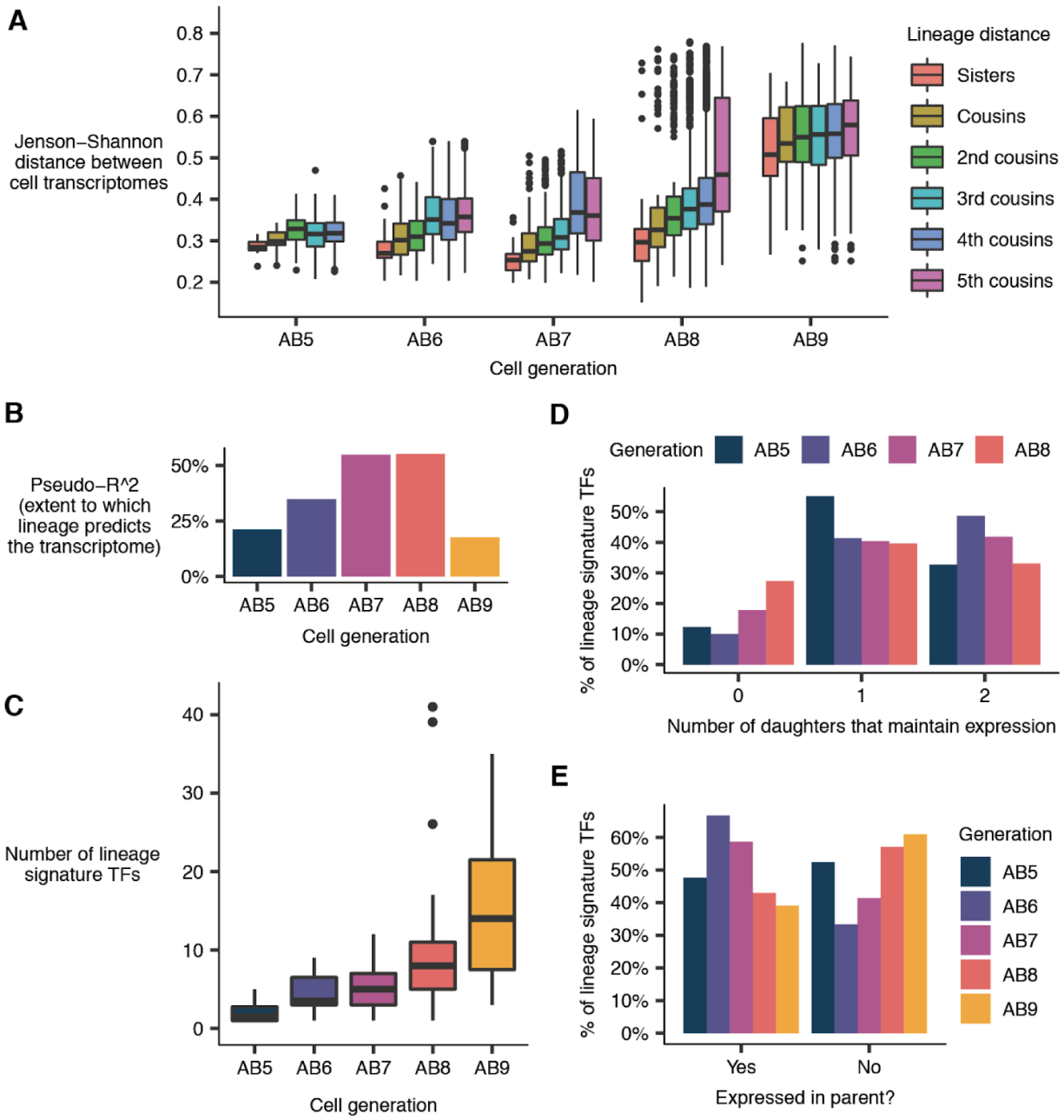


Fig 5. Correlation between cell lineage and the transcriptome in the ectoderm. (**A**) Jensen-Shannon (JS) distance between the transcriptomes of pairs of ectodermal cells (AB lineage), faceted by cell generation and lineage distance. AB5 refers to the cell generation produced by 5 divisions of the AB 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.

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 (**Fig. S30**). 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 (**Fig. 5C**). A substantial number of lineage signature TFs (~40-50%) had expression selectively maintained in only one of a cell's two daughters (**Fig. 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 (**Fig. 5E**). Temporal dynamics of lineage signature TFs were similar in the mesoderm (**Fig. S29**).

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 (**Tables S7-8**) provides a global view of expression dynamics for all genes in our dataset. A heatmap of pre-terminal lineage expression profiles (**Fig. S31**) 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 (**Fig. S32**) 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 (**Fig. S33A**). 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 (**Fig. S33B**). 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 (**Fig. S33C**). Hypodermis and intestine express many distinct NHRs, while expression of Sox family TFs is largely restricted to neurons, glia and pharynx (**Fig. S33D**).

An RShiny app to explore and extend on our analysis

We developed VisCello to distribute single cell analyses and provide interactive visualizations (**Fig. S34**). It is available as a web app (https://cello.shinyapps.io/celegans/) and can also be installed as an R package (https://github.com/qinzhu/VisCello.celegans). VisCello hosts dimensionality reductions (e.g. UMAPs), cell annotations, and marker gene tables for the different subsets of the data described in this manuscript. 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 can also be used to host and disseminate other single cell datasets, including data from the *C. elegans* 1-16 cell stage (*26*) and L2 stage (*29*) (https://github.com/qinzhu/VisCello).

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 (*19*, *29*) 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 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 (*30*) 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 (**Fig. 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 (**Figs. S25-26**), 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 (*200*). 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 (*29*), and earlier studies of the 1 to 16-cell stage embryos (*25*, *26*), 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, 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 (*201*) 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 degrees 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 **Fig. 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. (24)* 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 guage 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, 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 (*29*). 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.

Our final reference transcriptome is available as **Additional Data File 1**. It should be suitable for future studies of *C. elegans* embryos or early larva. However, neither the embryonic or L2 datasets contain reads from developed germline cells (sperm and oocytes) or developed somatic gonad cells (e.g. spermatheca), so our reference will not properly extend 3' UTRs for genes specific to these cell types.

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 S1**)
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 (**Figs. 3A, S9-11, S13).** They are not included in the original global UMAP (**Fig. 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 (*202*).

We found that the UMAP (*180*, *181*) 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.* (*181*), 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 (*182*). 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.

For more details on the UMAP and Louvain algorithms, we refer the reader to these web resources:

- A talk by Leland McInnes describing the intuition behind UMAP: https://www.youtube.com/watch?v=nq6iPZVUxZU.
- The section in the UMAP paper on arxiv (*180*) titled "A Computational View of UMAP".
- An explanation of the Louvain algorithm on Quora: https://www.quora.com/Is-there-a-simple-explanation-of-the-Louvain-Method-of-community-detection.

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 (*16*), 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 (**Figs 3A, S9-11, S13**). These cells are not included in the global UMAP (**Fig. 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*. (*20*). 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 (**Fig. S1A**). 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. (20)* approximately agree with embryo collection times from our experimental design (**Fig. S1B**), and also have a strong correlation with embryo times estimated based on data from Boeck *et al. (21)* (**Fig. S1C**). 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 (*185*) (**Fig. S1D**).

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.* (*21*), 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.* (*20*) (**Fig. S1C**). 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 **Fig. S35A**) 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 corrected TPM values were then rescaled to once again sum to 1,000,000 and then converted back into (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 (**Fig. S35B**). This indicates that non-background gene expression observed in high background cells is not systematically biased compared to low background cells.

Computing aggregate gene expression profiles for cell types and lineages

To compute the aggregate gene expression profile for a cell type (**Table S7**) or lineage (**Table S8**), we (1) subsetted the whole-dataset gene-by-cell gene expression matrix to include just the cells annotated as the given cell type or lineage; (2) divided each column by the corresponding cell's size factor (a statistic computed by the Monocle software package equal to the cell's total UMI count divided by the geometric mean of all cells' total UMI counts); (3) took the mean of each row (gene); and (4) rescaled the resulting vector to sum to 1,000,000. This results in a gene expression vector measured in transcripts per million (TPM).

We performed these computations using the original gene expression matrix, not corrected for background RNA contamination. After computing the aggregate gene expression vector for a cell type or lineage, we then corrected it for background contamination using the same method described in the previous section ("Per-cell background correction and filtering"), treating the aggregate vector as if it were the expression vector for a single cell. Compared to the alternative option of correcting each cell for background first and then computing the aggregate profile, aggregating first then correcting makes the estimate of the background fraction more robust due to an increased sample size (number of reads).

Even after background correction, we noticed some residual, aberrant expression of genes that should not be expressed in a given cell type. In several cases, this aberrant expression was due to just one or two outlier cells within a given UMAP cluster. We suspected these outlier cells included both doublets missed by our filtering procedure and "pseudo-doublets", consisting of a real cell plus debris from another cell. In order to reduce the impact of these outliers, in **Tables S7-8** we report a "robust" estimate of the mean expression for a gene in a given cell type or lineage (in addition to the "raw" estimate). This robust estimate excludes the highest-expressing cell and the lowest-expressing cell for a given gene in a given cell type / lineage before computing the mean expression.

The impact of outliers is greater for cell types represented by only a small number of cells in our dataset, and for cell types that have a low average number of UMIs per cell. Estimates of

mean gene expression values are therefore less precise for these cell types. To estimate the variance of our mean gene expression statistics, we used bootstrap resampling: for a cell type with N cells in our dataset, we randomly sampled, with replacement, N cells from that set and computed mean expression statistics from that sample. We repeated this process for 1,000 iterations for each cell type and computed bootstrap confidence intervals from the resulting distribution of mean estimates. If one must make a statement that a gene is "expressed" in a given cell type or lineage, we recommend using the criterion that the lower bound of the 95% bootstrap confidence interval is >0 TPM.

Differential expression analysis for Fig. 3D and Fig. S22

We included four classes of transcription factors (TFs) in the heatmaps of **Fig. 3D** and **Fig. S22**. 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 Fig. 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 **Fig. S15**, using "sSeq" (*203*), 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 **Fig. 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 used in Fig. 5B and Fig. S29B

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 the 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 **Fig. 5B** and **Fig. S29B**, 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:

$$1 - \frac{\text{average Jensen-Shannon divergence between the transcriptomes of pairs of sister cells in the cell generation}}{\text{average Jensen-Shannon divergence between the transcriptomes of random pairs of cells in the cell generation}}$$

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 (*204*) 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 Kullback-Leibler (KL) divergence.

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.

Methods used in Fig. S30

In **Fig. S30**, 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:

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

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:

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

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:

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

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:

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

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

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

Definition of lineage signature transcription factors for Fig. 5C-E and Fig. S29C-E

For the analyses presented in **Fig. 5C-E** and **Fig. S29C-E**, we introduced a concept of "lineage signature transcription factors." For each anatomical cell that we annotated in our data, we defined the set of lineage signature transcription factors associated with that cell to be set of TFs that satisfy both of the following criteria:

1. The lower bound of the 95% bootstrap confidence interval for the average expression level of the TF in the cell, as reported in **Table S8**, is >0. In other words, the TF must be robustly expressed in the cell.
2. The TF must be expressed at least 5-fold higher in the cell compared to its sister in the lineage, with a differential expression q-value < 0.01.

Computing the adjusted Gini coefficient for Fig. S33A

The Gini coefficient is biased by sample size (*205*). 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 single cell RNA-seq data from Tintori *et al.*, 2016 (*26*)

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.* (*26*) to their corresponding lineages or immediate descendants in our dataset (**Fig. S19**). We downloaded normalized expression data (measured in reads per kilobase of transcript per million mapped reads, RPKM) from Tintori *et al.* (*26*) 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. (26)* using an iterative PCA approach. Gene sets 7, 9, and 10 in supplemental document S1 of Tintori *et al.* (*26*) 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 **Fig. S19A**. Hierarchical clustering was performed on the correlation matrix using the pheatmap package with default parameters (*206*).

To demonstrate that our data are consistent with Tintori *et al.* (*26*) 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 (**Fig. S19B-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 **Fig. S19B** was computed using all 16-stage cells from Tintori *et al.* (*26*) and 421 genes (intersection of Set 7 and expressed genes in our data). In the bottom sub-panel of **Fig. S19B**, we project 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 (*26*). 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 (26)*.

Comparison of data from this study to data from Spencer *et al.*, 2011 (*22*)

Spencer *et al.* (*22*) 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 (*207*) to check for enrichment of Spencer *et al.* (*22*) 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.* (*22*) signature genes have strong enrichment in the corresponding cell types in our data (**Fig. S20**). Due to the method by which the Spencer *et al. (22)* 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.* (*22*) 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 (*208*). Checking this gene set against expression patterns from Warner *et al.*, 2019 (*24*), 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 **Fig. S20**.

Supplementary Text

Supplemental Note 1

In the analysis of single cell RNA-seq data, the term "trajectory" is often used to refer to a group of cells that 1) represent a specific cell lineage or cell type, and 2) can be ordered in a way that reflects the cells' progression over time from one transcriptomic state to another. Algorithms for trajectory inference can construct such orderings by mapping high dimensional gene expression data to a low dimension and fitting a graph structure to the data points in the low dimension. The distance between a user-defined root vertex on the graph to the location of a cell on the graph is called the cell's "pseudotime," and for a particular path along the graph, the ordering of cells on that path by pseudotime defines a "trajectory" (the graph as a whole can be considered a "branched trajectory").

In this manuscript, we have not used any trajectory inference algorithm. Instead, we use our embryo time estimates for each cell, which are computed based on correlation of the cell's transcriptome with a high-resolution bulk RNA-seq time series (see **Methods** and **Fig. S1**), as a universal ordering for all cells in the dataset. We annotate cell types and lineages based on marker genes from the literature (**Table S1, S4**) and clustering in UMAPs of our data. Ordering cells with a common cell type or lineage annotation by embryo time defines a "trajectory". In most cases, trajectories defined in this way also form contiguous shapes in UMAPs, to which one could fit a graph structure to if desired. Our universal ordering of cells by embryo time is a more robust approach, however, since 1) it allows annotations from multiple UMAPs to be integrated to define a single "trajectory" per cell type/lineage, and 2) the approach still works even when cells of a common type/lineage are split into disparate, non-contiguous groups in a UMAP (which may occur due to abrupt changes in gene expression or due to technical factors, such as non-uniform sampling with respect to time).

Supplementary Figures

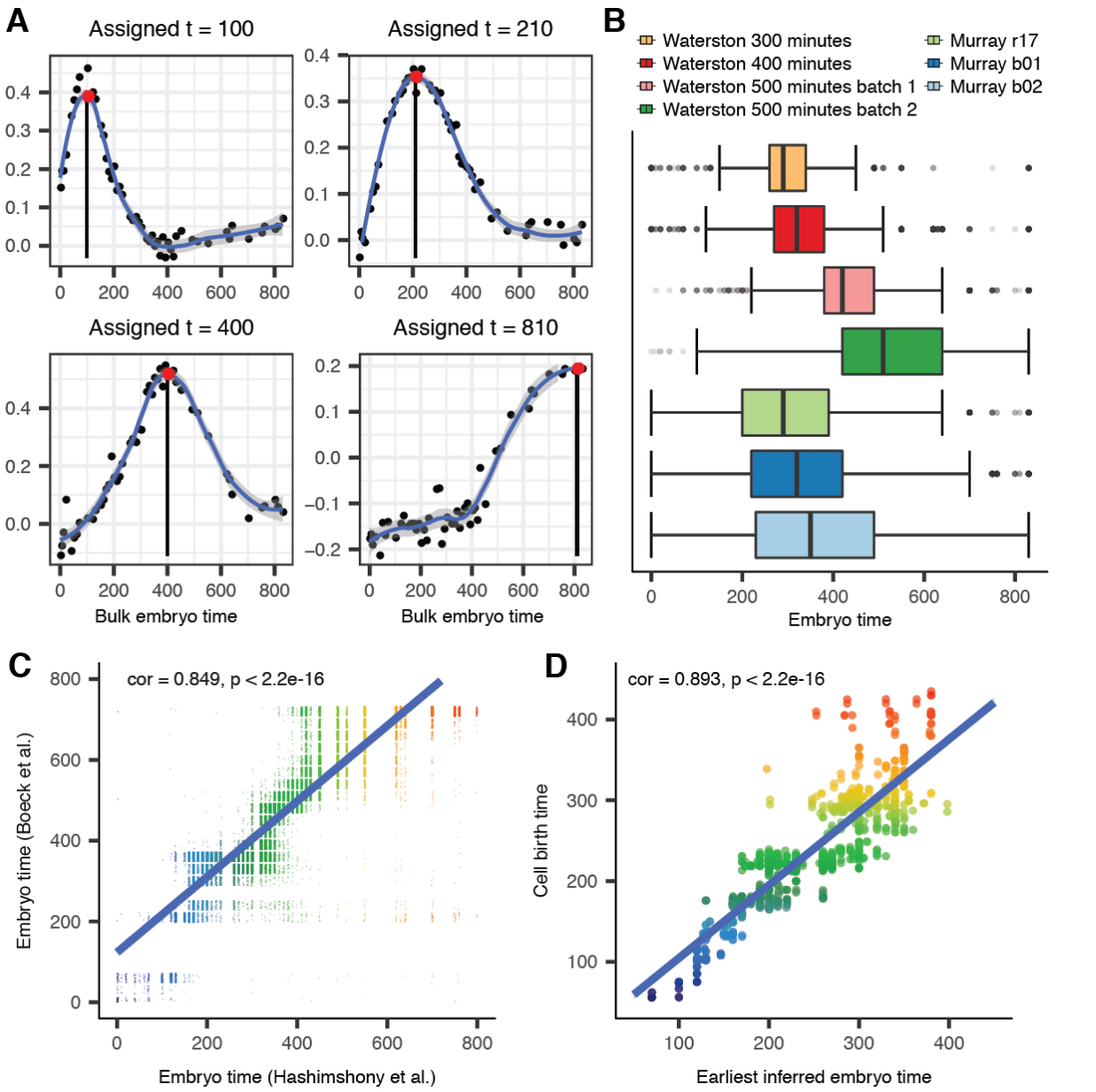


Fig. S1. Method for estimating the age of the embryo that a sc-RNA-seq cell came from. Embryo times are measured in minutes post first cleavage. (**A**) Embryo times are estimated based on Pearson correlation of a single cell's transcriptome to a bulk RNA-seq time series (see **Methods**). Pointwise 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.* (*20*) to an alternate set of embryo time estimates based on Boeck *et al*. (*21*). Estimates based on Hashimshony *et al.* 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) (*185*).

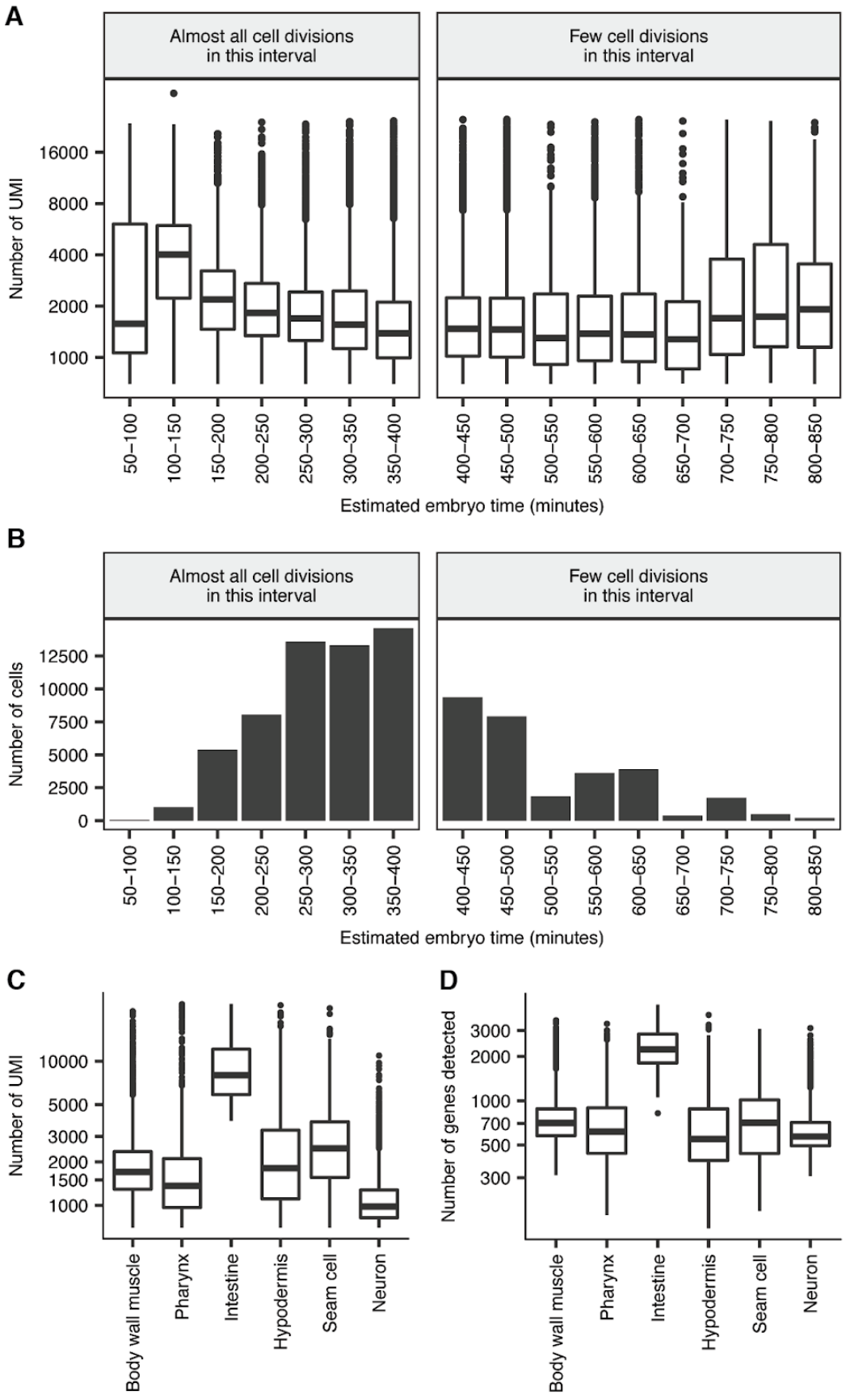

A
Almost all cell divisions in this interval
Few cell divisions in this interval
Number of UMI
16000
8000
4000
2000
1000
50–100
100–150
150–200
200–250
250–300
300–350
350–400
400–450
450–500
500–550
550–600
600–650
650–700
700–750
750–800
800–850
Estimated embryo time (minutes)
B
Almost all cell divisions in this interval
Few cell divisions in this interval
Number of cells
12500
10000
7500
5000
2500
0
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

Fig. S2. 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.

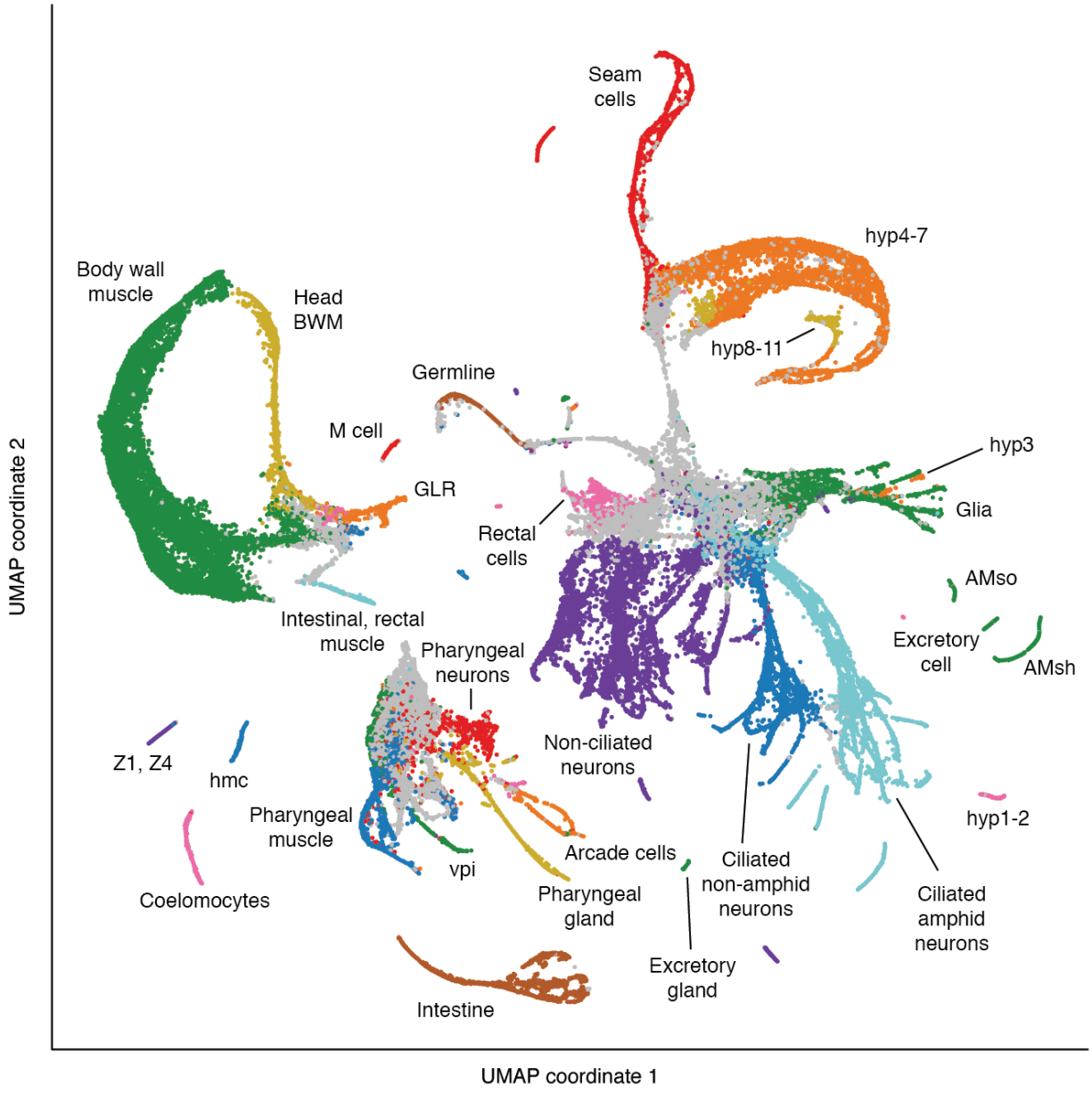


Fig. S3. Cell type annotations for the global UMAP of 81,286 cells. This plot shows more cell type annotations for the global UMAP from **Fig. 1A**. This UMAP does not include 4,738 additional cells that were initially filtered, but were later whitelisted and included in downstream analyses (see **Methods**). For fine-grained annotations of cell types in each major tissue, see **Figs. 3A and S5-13**. For fine-grained annotations of progenitor cell lineages, see **Figs. S14-17**.

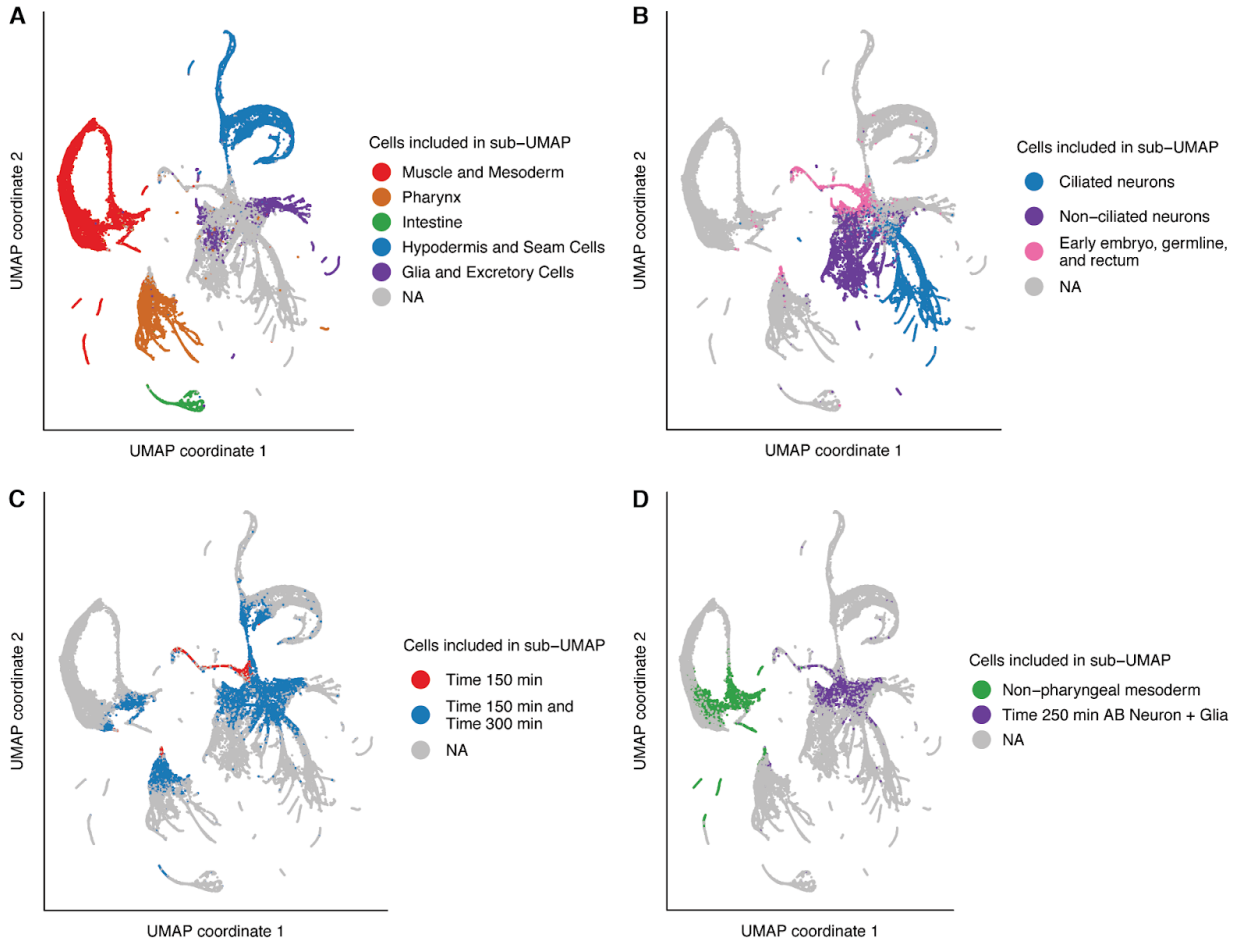


Fig. S4. **Cells included in each sub-UMAP.** Plots show which cells from the global UMAP (**Fig. S3**) are included in each sub-UMAP (**Figs. S5-17**), including UMAPs aimed at visualizing terminal cell types (**A, B**) and UMAPs focused aimed at visualizing progenitor lineages (**C, D**). Note that 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 (**Fig. S5**), pharynx (**Fig. S6**), intestine (**Fig. S7**), hypodermis and seam cells (**Fig. S8**), glia and excretory cells (**Fig. S9**), non-ciliated neurons (**Fig. S10**), touch receptor neurons (**Fig. S11**), germline (**Fig. S12**), and rectum (**Fig. S13**). A UMAP of ciliated neurons is shown in the main text (**Fig. 3A**). UMAPs focused on annotating progenitor lineages are shown in **Figs. S14-17**.

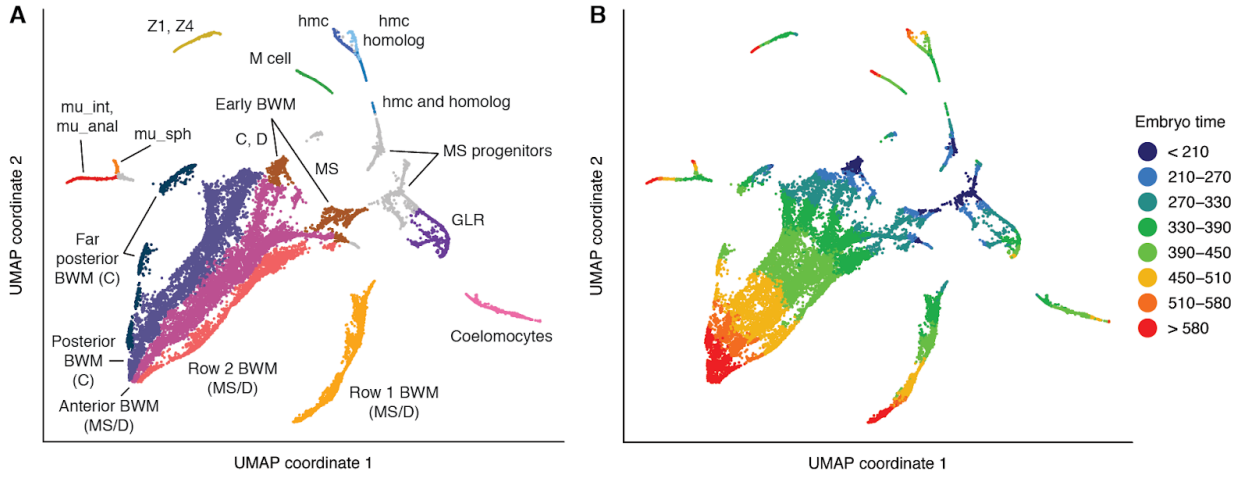


Fig. S5. UMAP of 22,371 body wall muscle and non-pharyngeal mesoderm cells. (**A**) Labels indicate cell types. See **Table S1** for marker genes used to annotate cell types. MS, C, and D indicate cell lineages. Abbreviations: BWM = body wall muscle, mu_int = intestinal muscle, mu_anal = anal depressor muscle, mu_sph = anal sphincter muscle, hmc = head mesodermal cell. (**B**) Colors show estimated embryo times (minutes post first cleavage) for each cell.

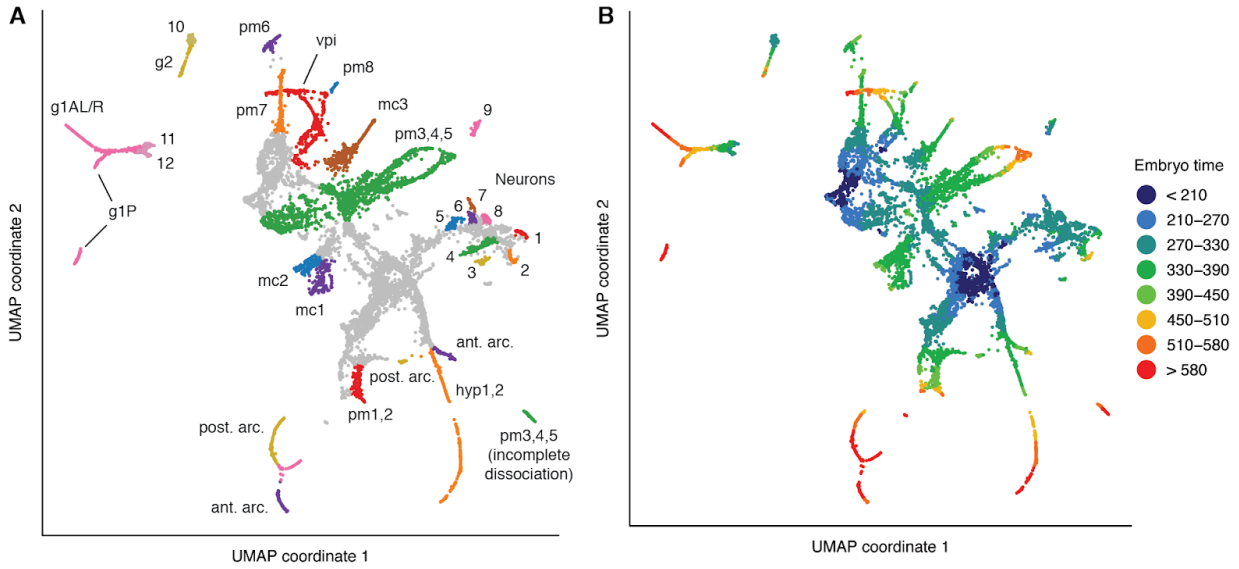


Fig. S6. UMAP of 10,784 pharyngeal cells. (**A**) Labels indicate cell types. See **Tables S1 and S4** for marker genes used to annotate cell types. Abbreviations: pm = pharyngeal muscle, mc = pharyngeal marginal cell, g1A/g1P/g2 = pharyngeal gland, vpi = pharyngeal-intestinal valve, hyp = hypodermis, 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.

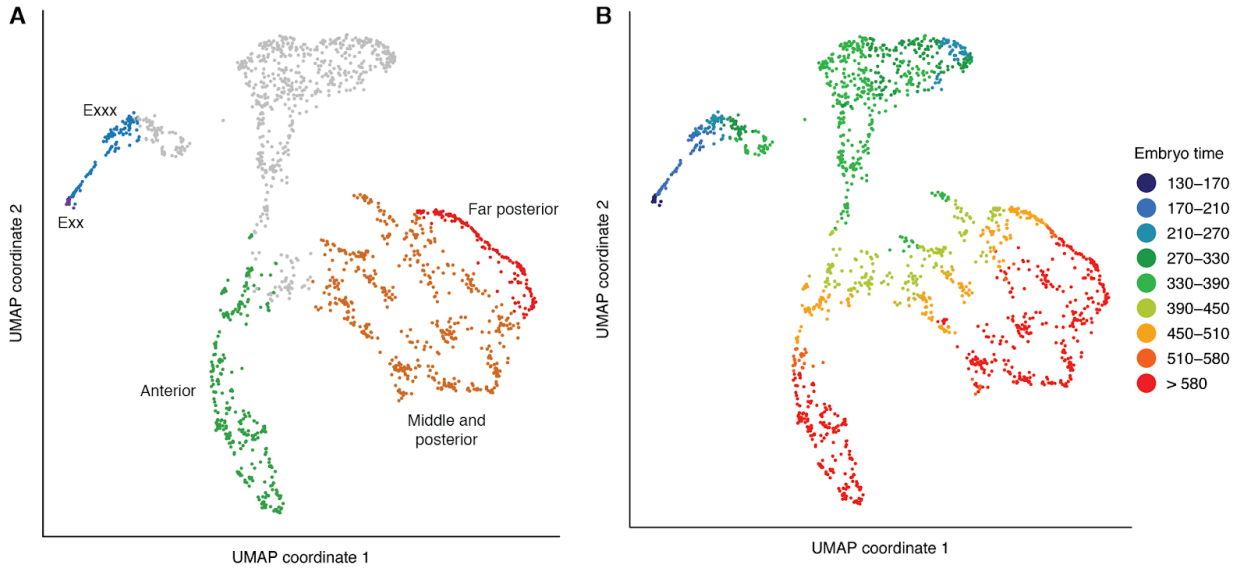


Fig. S7. UMAP of 1,734 intestine cells. (**A**) Labels indicate subsets of intestine cells and their relative position on the anterior-posterior axis. See **Table S1** for marker genes used to annotate cell types. (**B**) Colors show estimated embryo times (minutes post first cleavage) for each cell.

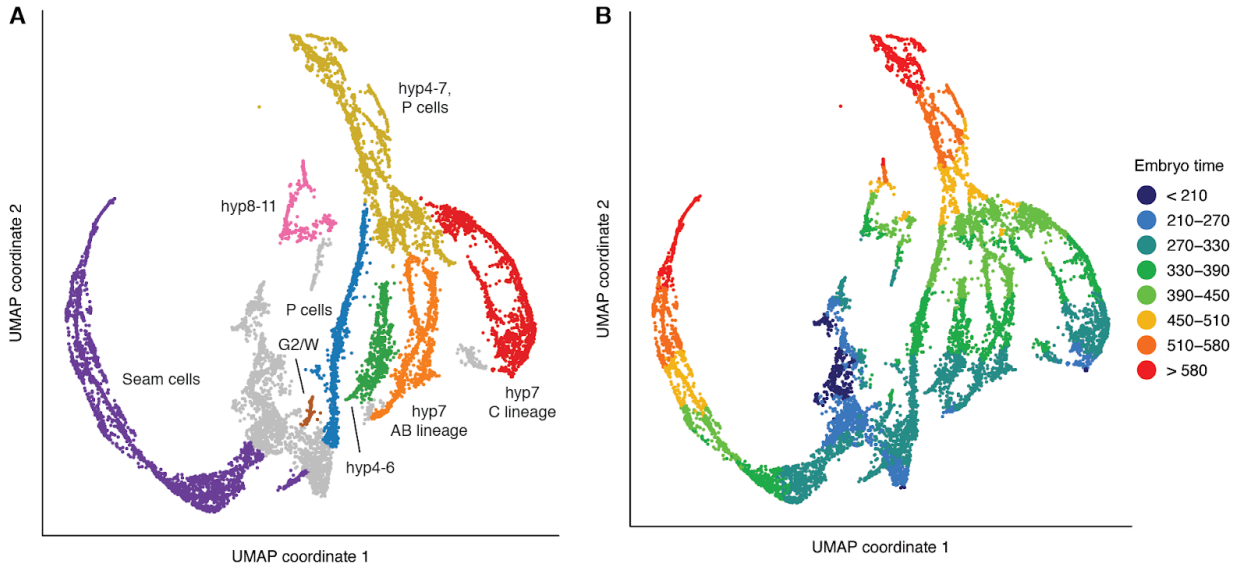


Fig. S8. UMAP of 12,254 hypodermis and seam cells. (**A**) Labels indicate cell types. See **Table S1** for marker genes used to annotate cell types. hyp1-3 are not included here. hyp1-2 appear in the pharynx UMAP (**Fig. S6**), and hyp3 appears in the glia UMAP (**Fig. S9**), consistent with their cell 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.

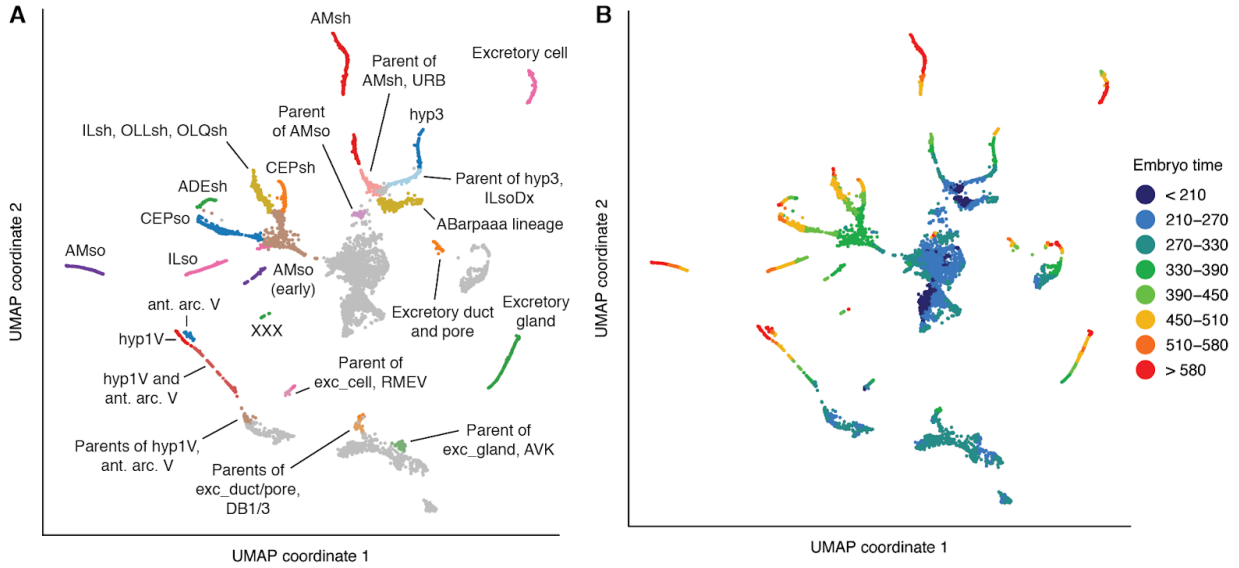


Fig. S9. UMAP of 7,512 glia, excretory cells, and progenitors. (**A**) Labels indicate cell types. See **Tables S1 and S4** for marker genes used to annotate cell types. Some non-glial/excretory cells are also included in the UMAP, such as neuron/glia/rectal progenitors. The annotations of hyp1V and anterior arcade V are very tentative—evidence is described in a note in **Table S1**. (**B**) Colors show estimated embryo times (minutes post first cleavage) for each cell.

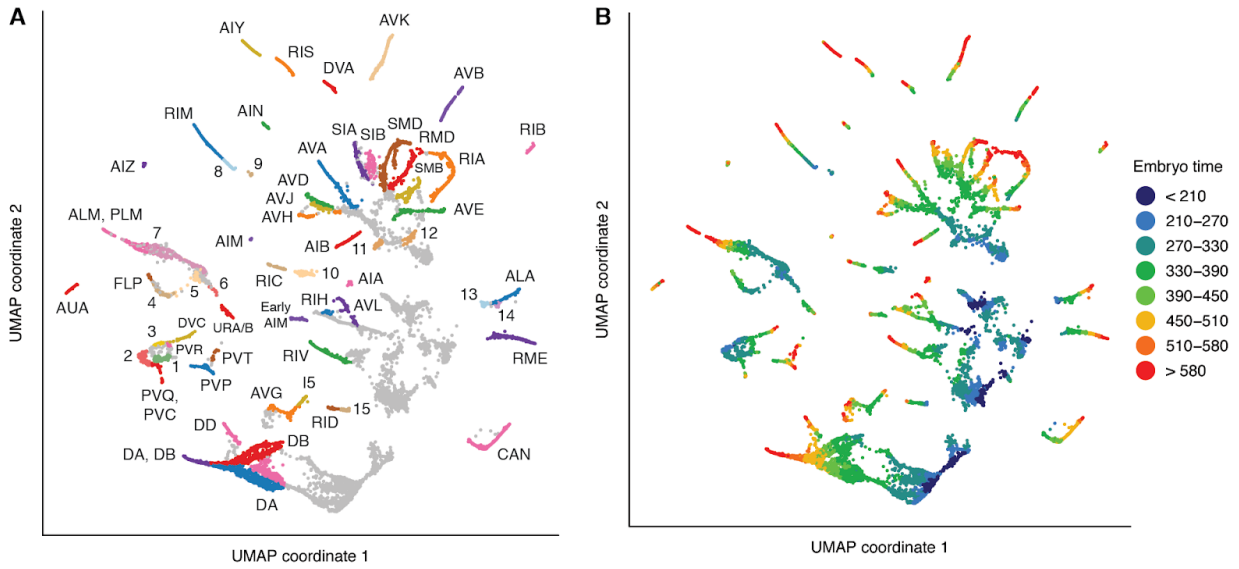


Fig. S10. UMAP of 14,728 non-ciliated neurons and progenitors. For a UMAP of ciliated neurons, see **Fig. 3A**. (**A**) Text labels indicate terminal cell types. Numeric labels indicate: **1** PVC-LUA neuroblast **2** parent of PVQ **3** parent of DVC **4** FLP-AIZ neuroblast **5** FLP-AIZ-RMG neuroblast **6** parent of URADx **7** progenitors of ALM, BDU, PLM, and ALN (see **Fig. S11** 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. See **Tables S1 and S4** for marker genes used to annotate cell types. (**B**) Colors show estimated embryo times (minutes post first cleavage) for each cell.

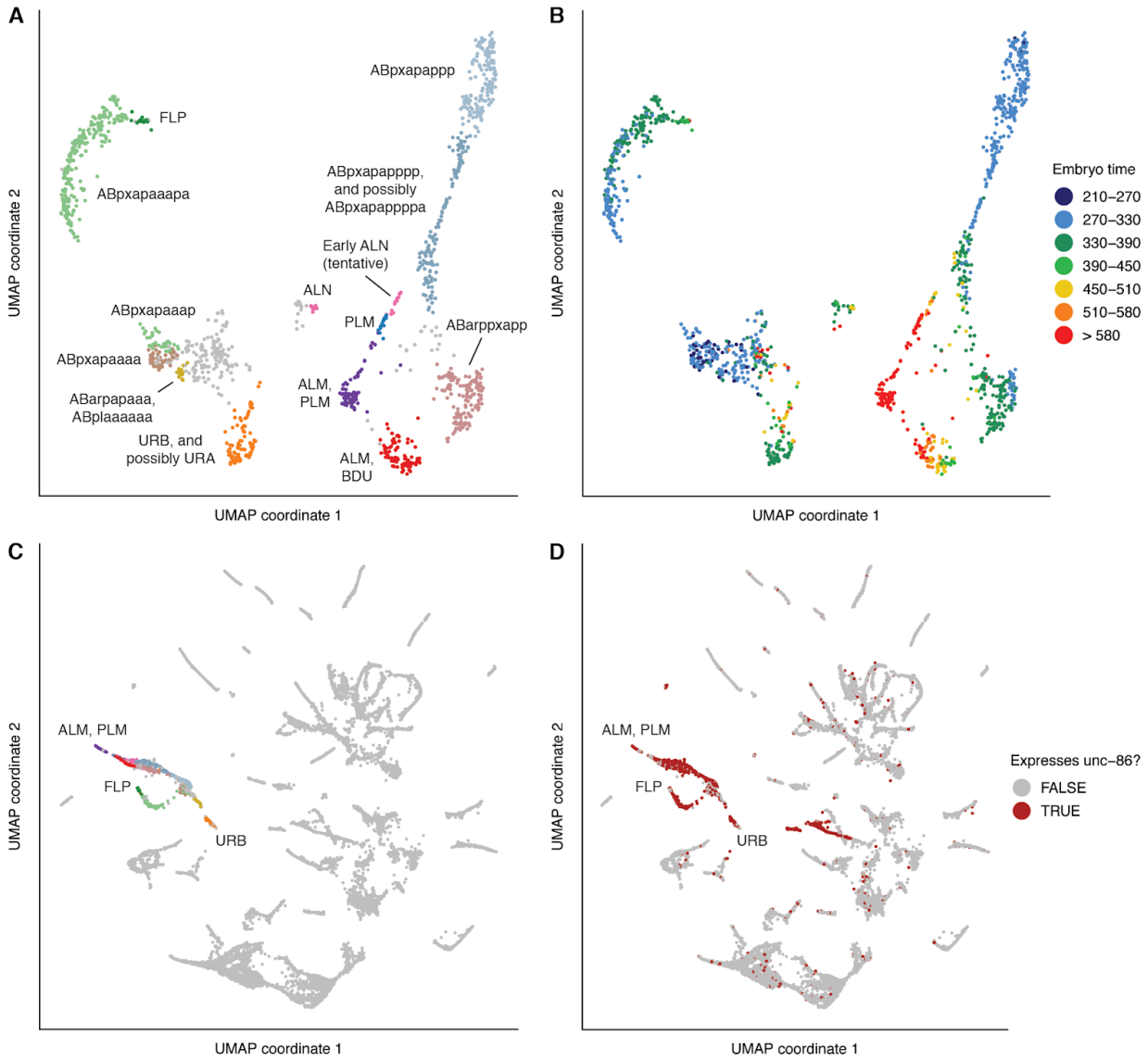


Fig. S11. UMAP of 1,300 touch receptor neurons, URB neurons, and progenitors. URB neurons are included because they cluster near the touch receptors in the UMAP of all non-ciliated neurons (**Fig. S10**). This is in part due to high *unc-86* expression. (**A**) Labels indicate cell type (for terminal 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 **Fig. S10**. (**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*.

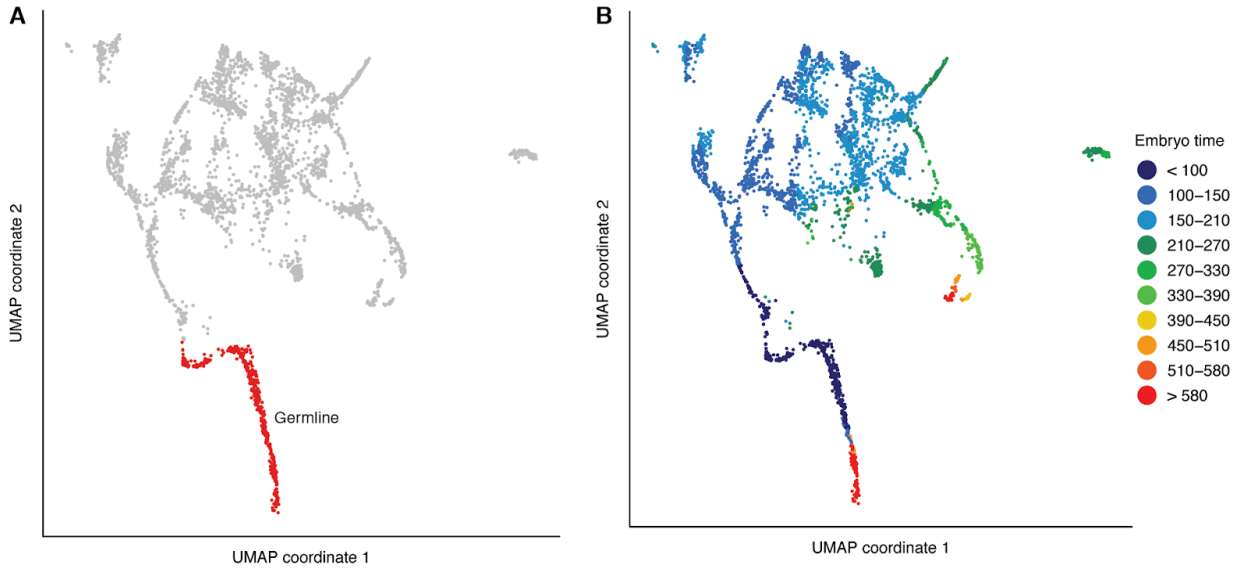


Fig. S12. UMAP of 3,476 early embryo, germline, and rectal cells. This UMAP was used only for its trajectory of germline development (500 cells). Other lineages that are included in this UMAP were better resolved in other UMAPs, shown below. (**A**) Germline cells highlighted in red. (**B**) Colors show 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.

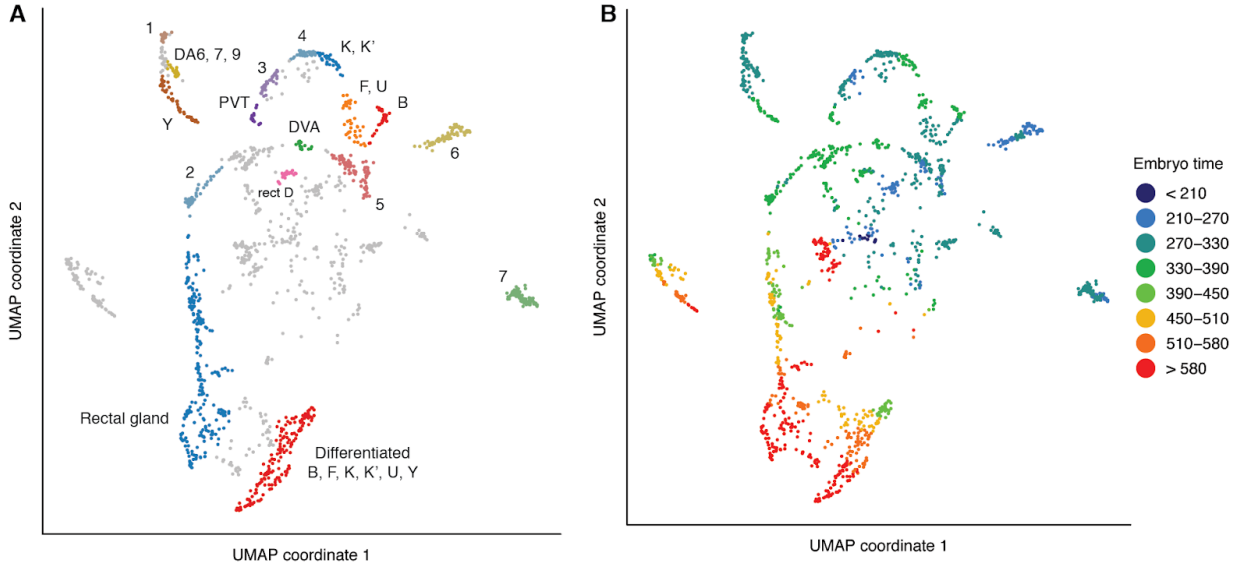


Fig. S13. UMAP of 1,598 rectal cells and progenitors. (**A**) Text labels indicate terminal cell types. Numeric labels indicate: **1** parents of (Y and DA7) and (DA6 and DA9). **2** parent of PVP and rect_V **3** parent of PVT and rect_D **4** parent of K and K' **5** parents of (B and DVA) and (F and 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: Figs S14-17 show a representative subset of the UMAPs that were used to annotate progenitor lineages. Several additional UMAPs can be visualized in VisCello. **Table S4** includes a list of which UMAP was used to annotate each lineage in our annotation set.

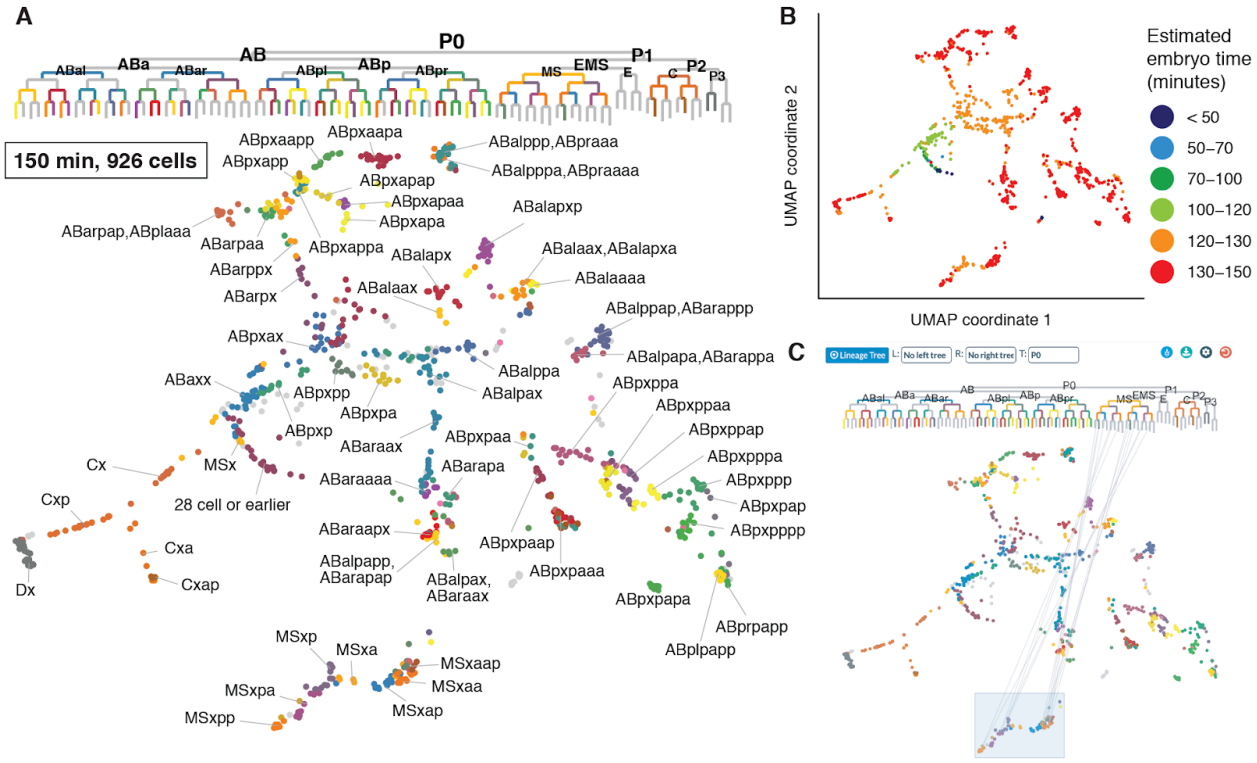


Fig. S14. UMAP and detailed annotation of 926 cells from embryos < 150 minutes post first cleavage. E lineage and germline cells are excluded from the UMAP and were analyzed separately (**Figs S7 and S12**). (**A**) Detailed labeling of lineages, co-visualized with the lineage tree. (**B**) Colors show estimated embryo times (minutes post first cleavage) for each cell. (**C**) Screenshot of an interactive co-visualization implemented in VisCello, highlighting the connection between MS lineage clusters and corresponding leaves in the lineage tree.

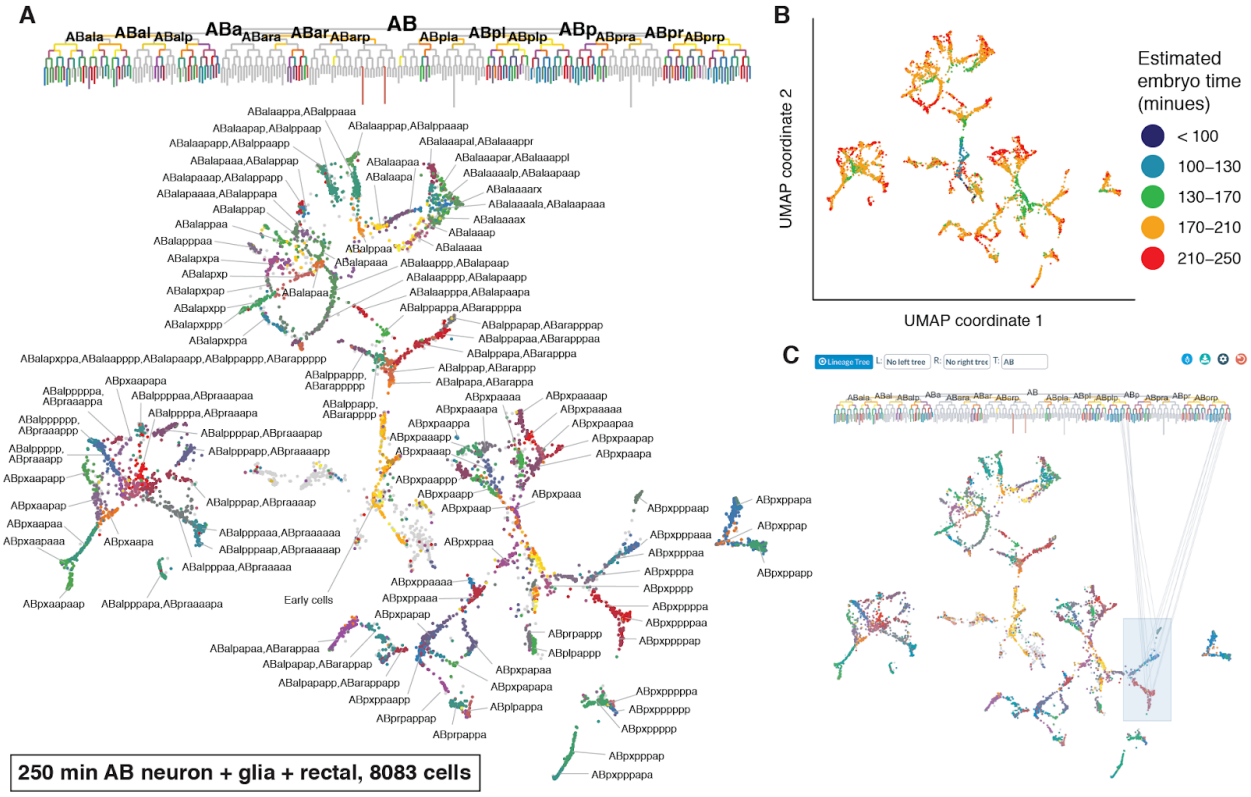


Fig. S15. UMAP and detailed annotation of 8,083 AB lineage neuron/glia/rectal progenitor cells from embryos < 250 minutes post first cleavage. This UMAP is derived from another UMAP that can be visualized in VisCello ("Time 250min lineage"). This UMAP includes only AB lineage cells that give rise to 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, highlighting the connection between ABpxppp lineage clusters and corresponding leaves in the lineage tree.

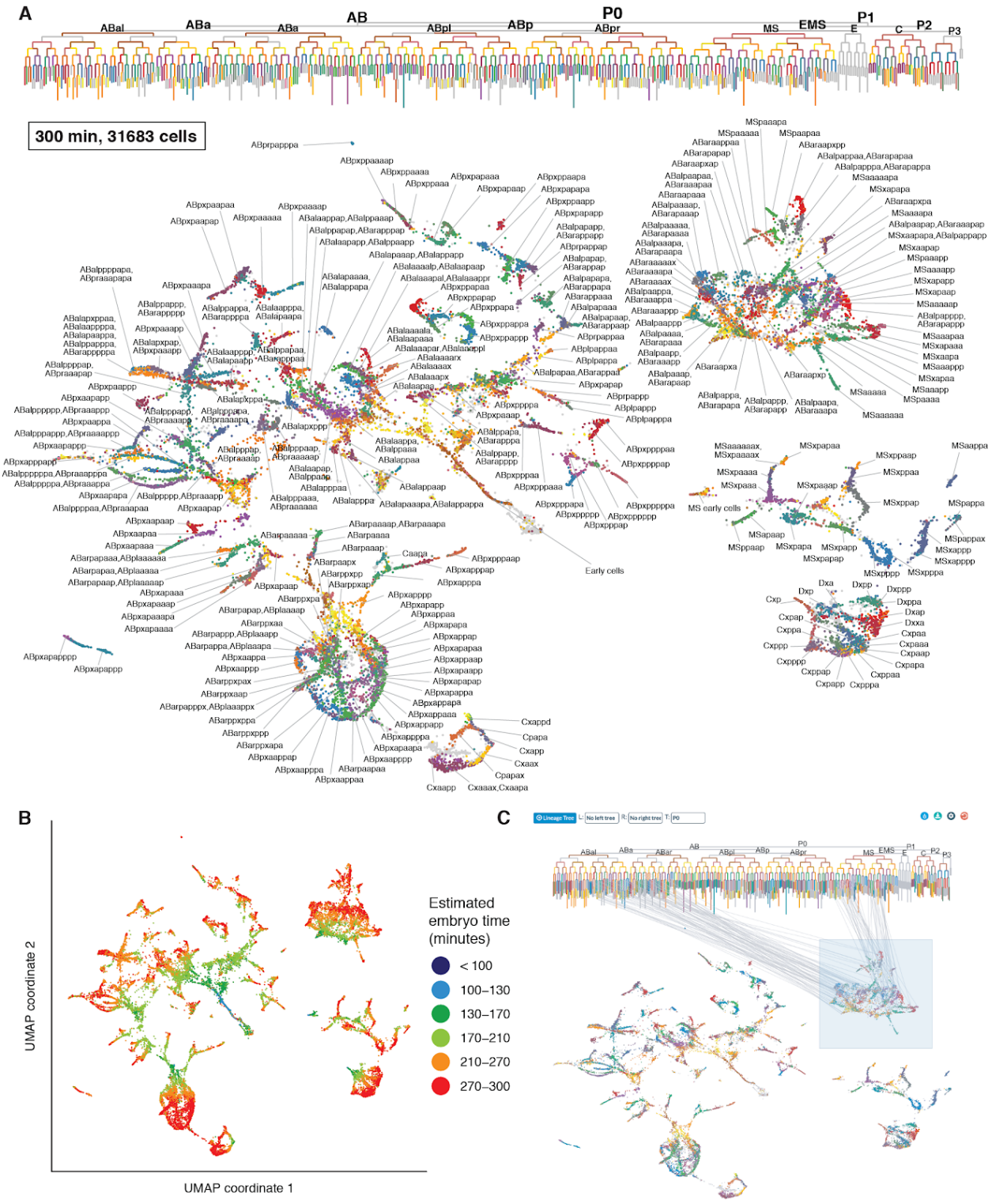


Fig. S16. UMAP and detailed annotation of 31,683 cells from embryos < 300 minutes post first cleavage. E lineage and germline cells are excluded from the UMAPs and were analyzed separately (**Figs S7 and S12**). (**A**) Detailed labeling of lineages, co-visualized with the lineage tree. (**B**) Colors show estimated embryo times (minutes post first cleavage) for each cell. (*Legend continued on the following page*)

Fig. S16 (continued). (**C**) Screenshot of an interactive co-visualization implemented in VisCello, 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.

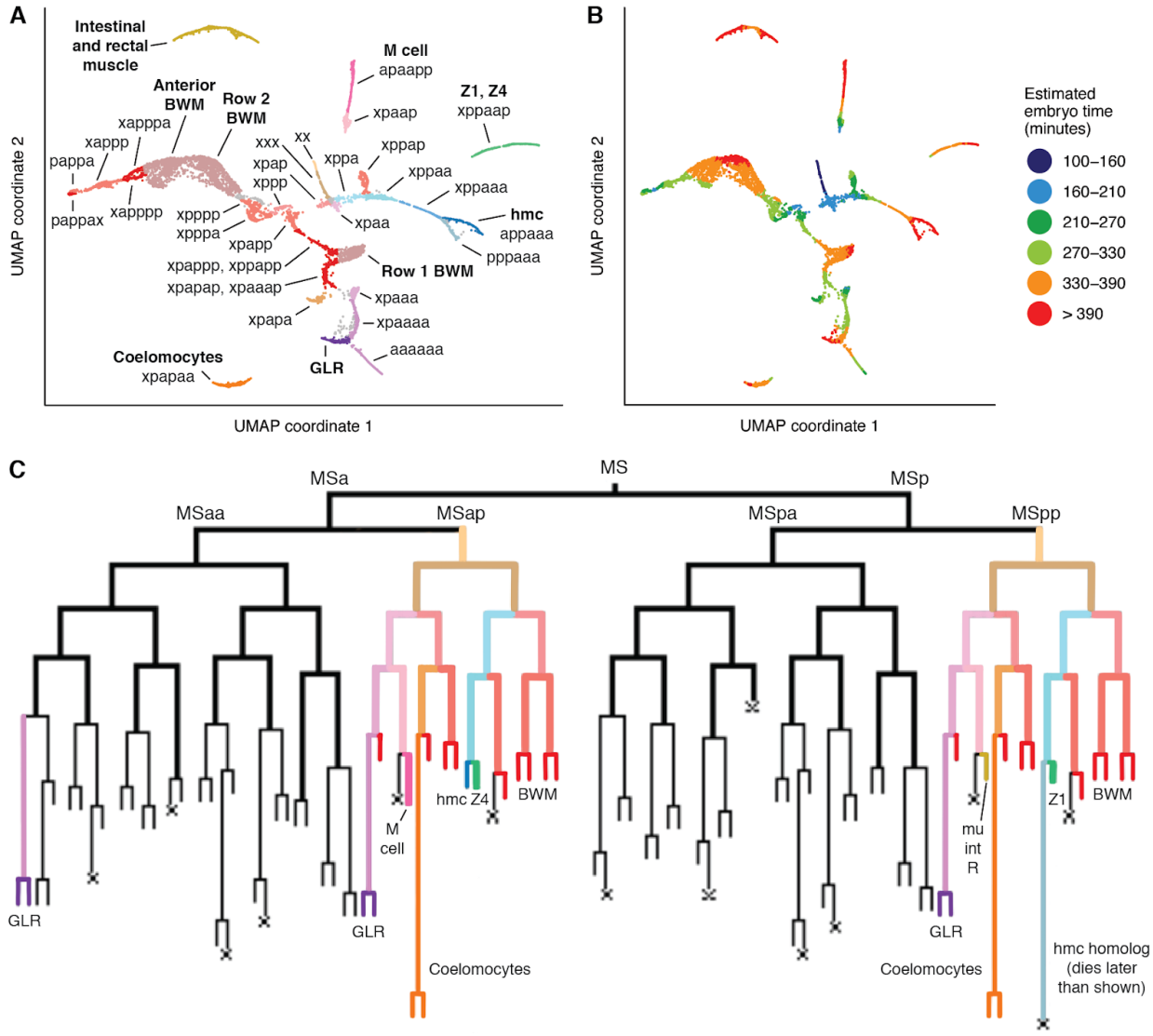


Fig. S17. **UMAP of 8,233 non-pharyngeal mesoderm cells, focused on the early lineage.** This UMAP includes the same cells as the muscle and mesoderm UMAP (**Fig. S5**), but excludes putative C and D lineage body wall muscle, MS lineage body wall muscle with estimated embryo time >400 minutes (post first 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).

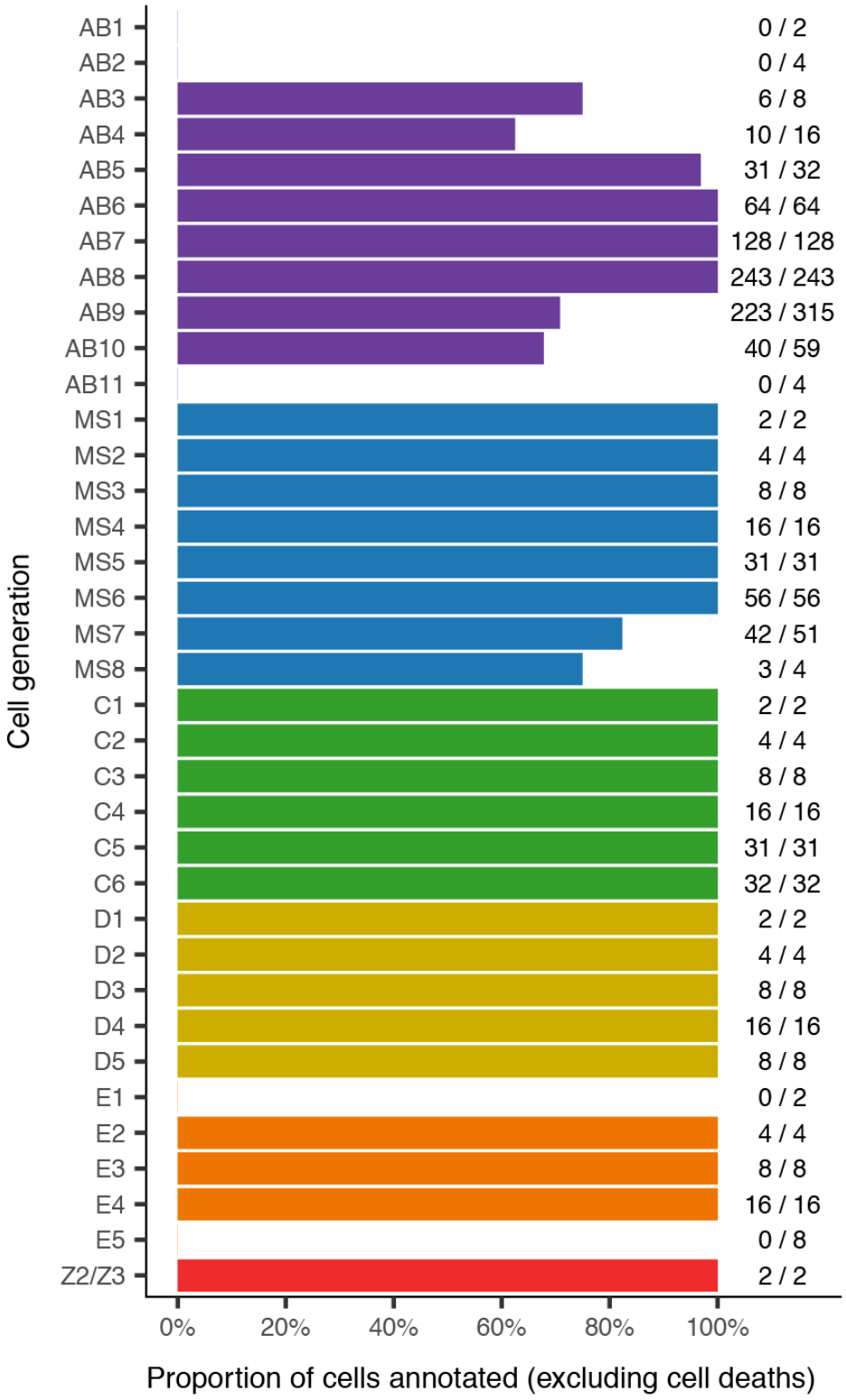

AB1
AB2
AB3
AB4
AB5
AB6
AB7
AB8
AB9
AB10
AB11
MS1
MS2
MS3
MS4
MS5
MS6
MS7
MS8
C1
C2
C3
C4
C5
C6
D1
D2
D3
D4
D5
E1
E2
E3
E4
E5
Z2/Z3
0 / 2
0 / 4
6 / 8
10 / 16
31 / 32
64 / 64
128 / 128
243 / 243
223 / 315
40 / 59
0 / 4
2 / 2
4 / 4
8 / 8
16 / 16
31 / 31
56 / 56
42 / 51
3 / 4
2 / 2
4 / 4
8 / 8
16 / 16
31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

Fig. S18. 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.

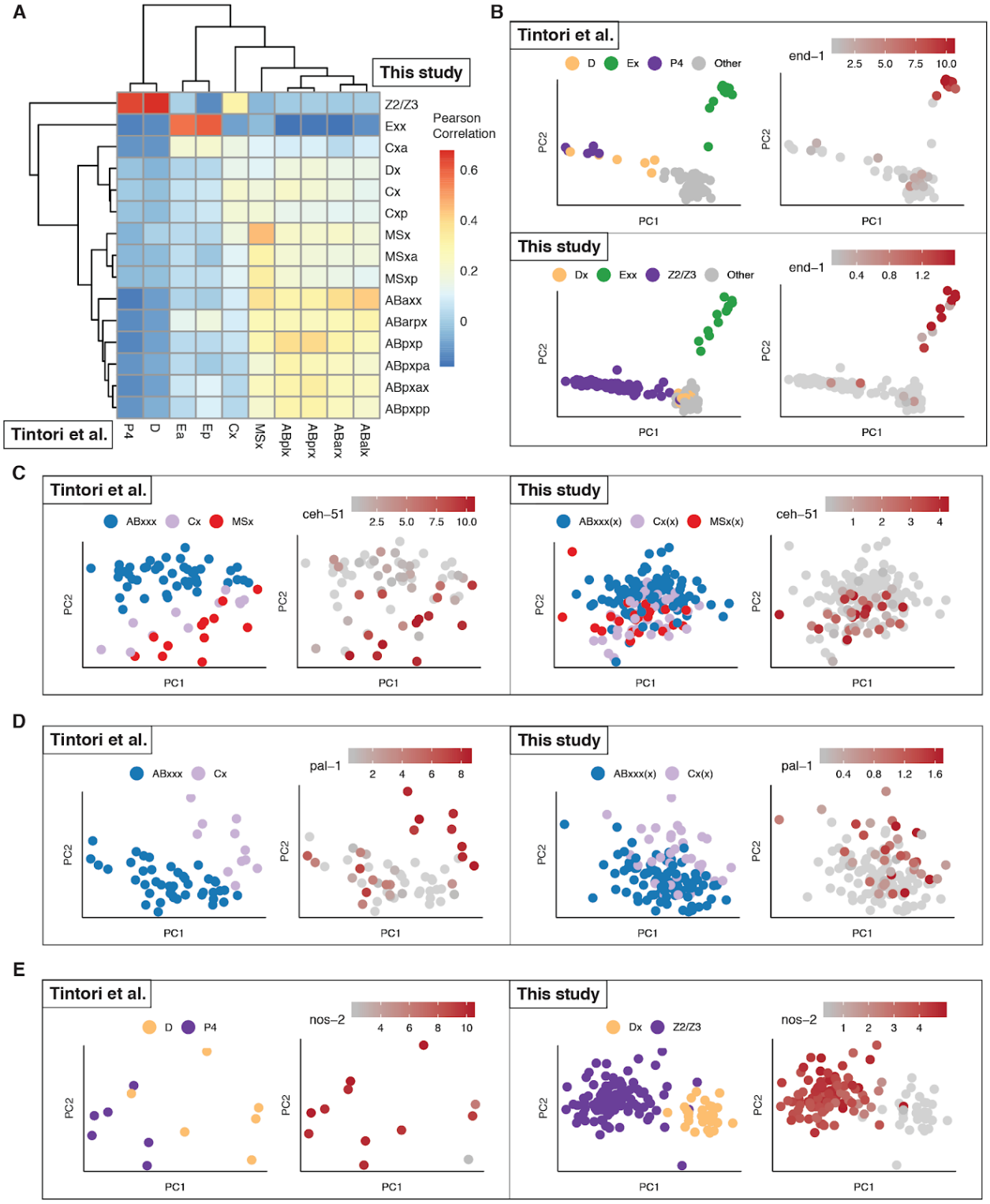

A
This study
Tintori et al.
Z2/Z3
Exx
Cxa
Dx
Cx
Cxp
MSx
MSxa
MSxp
ABaxx
ABarpx
ABpxp
ABpxpa
ABpxax
ABpxpp
P4
D
Ea
Ep
Cx
MSx
ABplx
ABprx
ABarx
ABalx
Pearson Correlation
0.6
0.4
0.2
0
B
Tintori et al.
D
Ex
P4
Other
end–1
2.5 5.0 7.5 10.0
PC2
PC1
This study
Dx
Exx
Z2/Z3
Other
end–1
0.4 0.8 1.2
C
Tintori et al.
ABxxx
Cx
MSx
ceh–51
2.5 5.0 7.5 10.0
This study
ABxxx(x)
Cx(x)
MSx(x)
ceh–51
1 2 3 4
D
Tintori et al.
ABxxx
Cx
pal–1
2 4 6 8
This study
ABxxx(x)
Cx(x)
pal–1
0.4 0.8 1.2 1.6
E
Tintori et al.
D
P4
nos–2
4 6 8 10
This study
Dx
Z2/Z3
nos–2
1 2 3 4

Fig. S19. Comparison of data from this study to data from Tintori *et al.*, 2016 (*26*). Tintori *et al.* (*26*) 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. (26)* 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. (26)* (see **Methods**). (**B-E**) First sub-panel shows a PCA projection computed using 16-cell stage cells from Tintori *et al. (26)*, 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. (26)*, to discriminate particular lineages (see **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. (26)*, and log2 size-factor normalized UMI counts for data from this study.

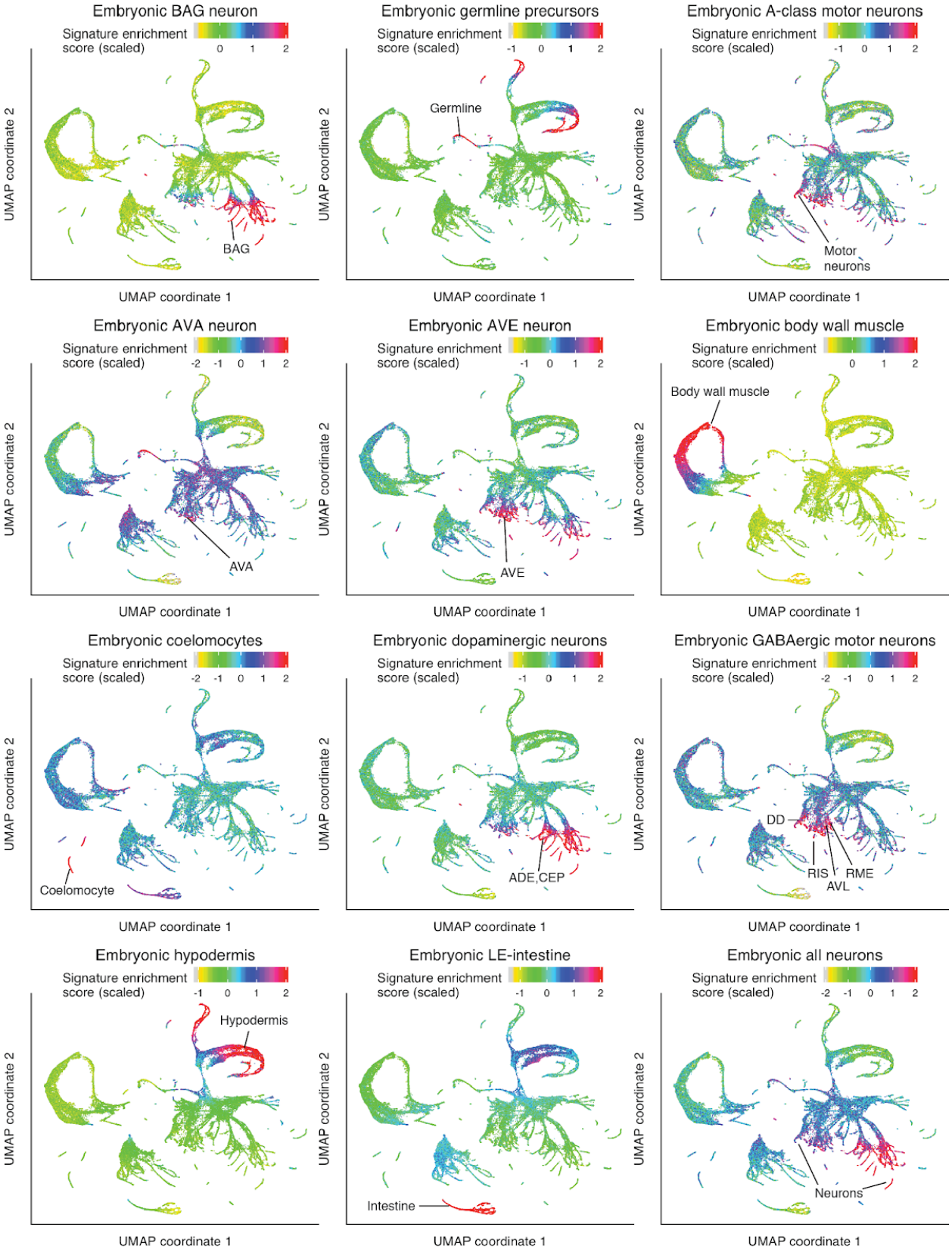

Embryonic BAG neuron
Signature enrichment score (scaled)
0 1 2
BAG
Embryonic germline precursors
Signature enrichment score (scaled)
-1 0 1 2
Germline
Embryonic A-class motor neurons
Signature enrichment score (scaled)
-1 0 1 2
Motor neurons
Embryonic AVA neuron
Signature enrichment score (scaled)
-2 -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
UMAP coordinate 1
UMAP coordinate 2

Fig. S20. Comparison of data from this study to microarray data from Spencer *et al.*, 2011 (*22*). 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. (22)*. Signature gene sets from Spencer *et al.* (*22*) 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. Signature genes are therefore mostly tissue-specific, rather than cell-type specific. Gene set enrichment scores were computed using the AUCell package (*207*). Comparison with pharyngeal muscle was dropped because most of the signature genes reported in Spencer *et al.* for this cell type are intestine specific, as confirmed by a third dataset (*24*). See **Methods** for more details.

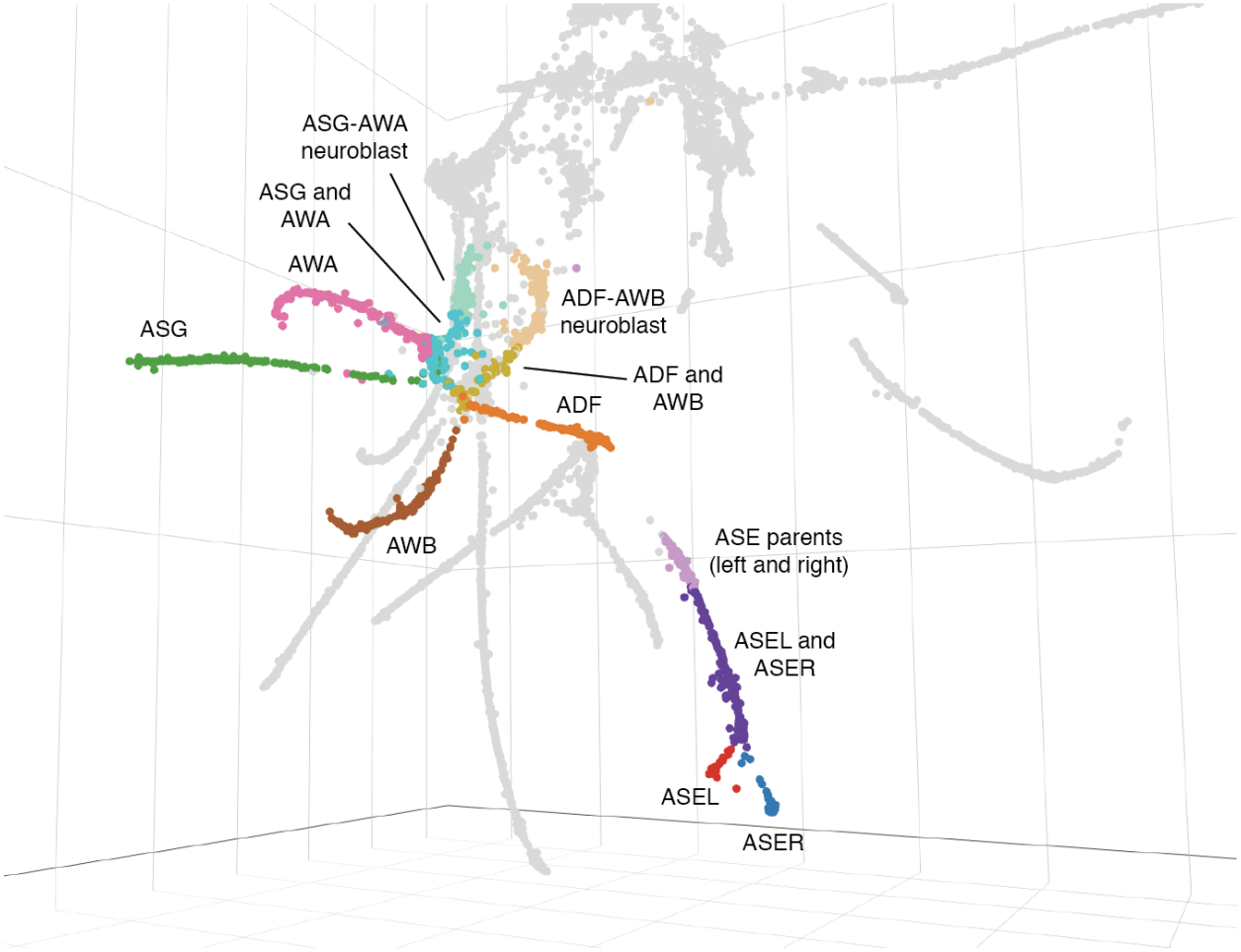


Fig. S21. Ciliated neuron developmental trajectories are more continuous in a 3D UMAP. This plot is a 2D screenshot of part of a 3D UMAP of ciliated neuron cells, oriented to show specific lineage relationships. The cells are the same as in **Fig. 3A**; the only difference is projecting into 3D instead 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.

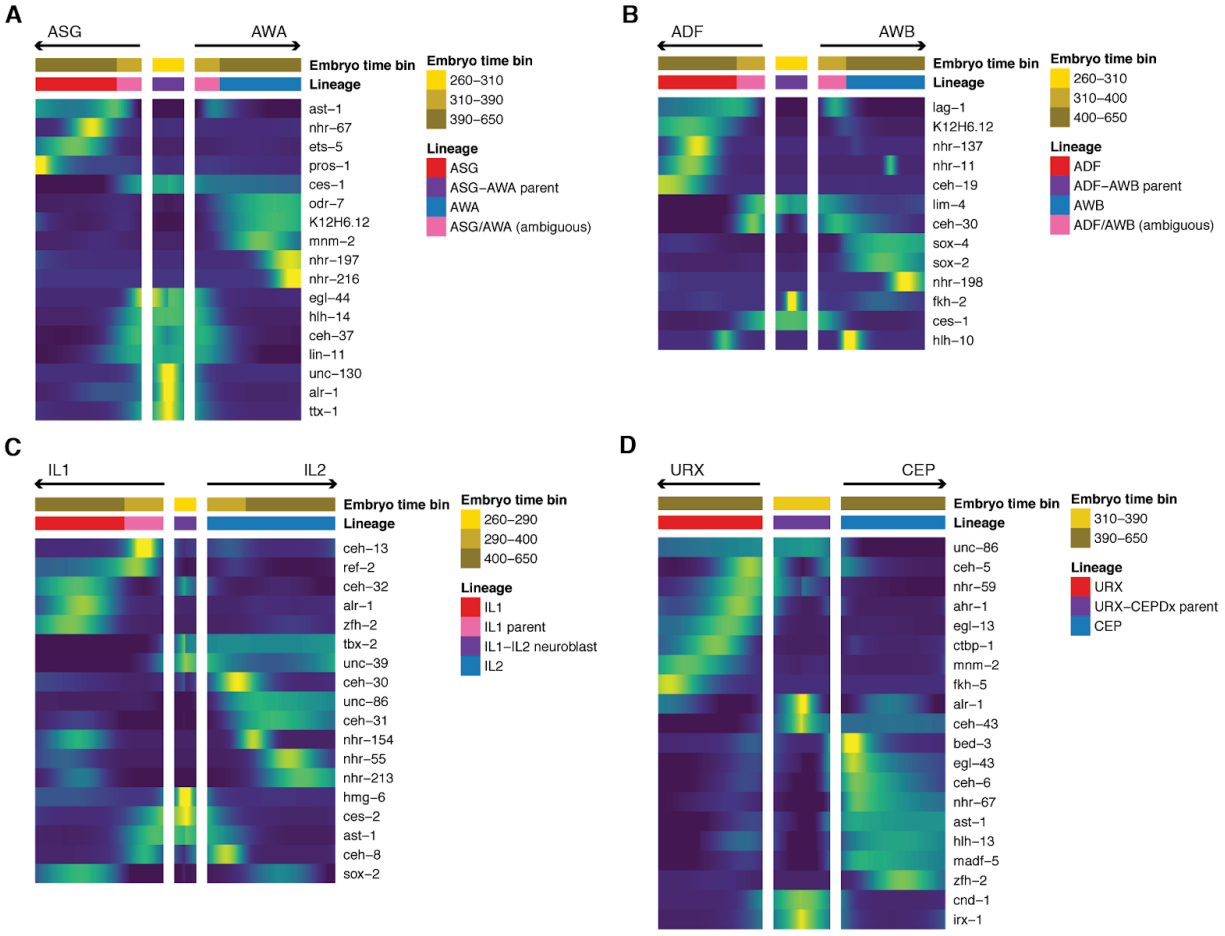


Fig. S22. Differentially expressed transcription factors associated with ciliated neuron lineage branches. Heatmaps showing patterns of differential transcription factor expression associated with branches in (**A**) the ASG-AWA lineage, (**B**) the ADF-AWB lineage, (**C**) the IL1-IL2 lineage, and (**D**) the URX-CEPDx lineage. A heatmap for the ASE-ASJ-AUA lineage is shown in **Fig. 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 **Fig. S21**), 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.

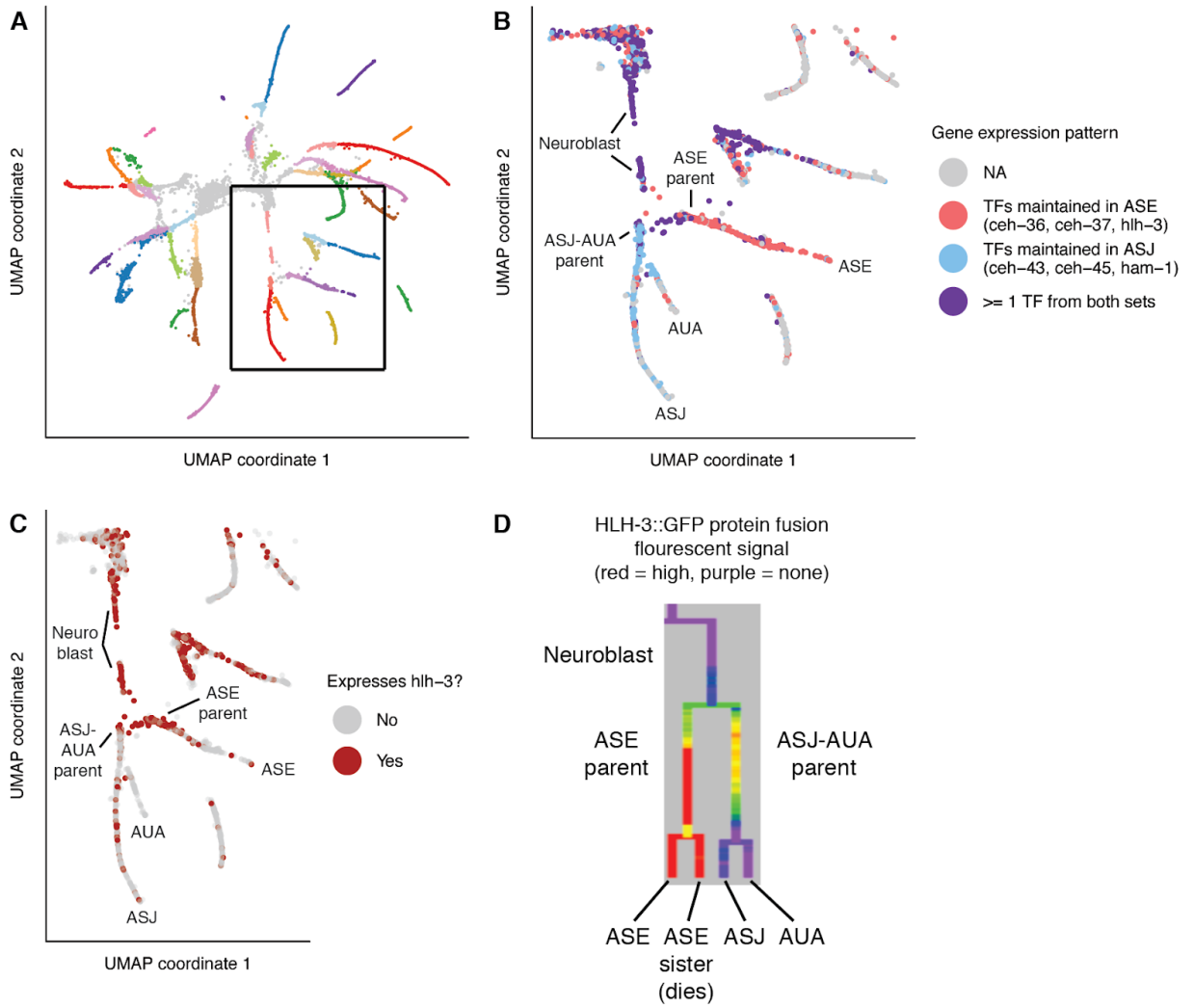


Fig. S23. Multilineage priming in the ASE-ASJ-AUA lineage. (**A**) Section of the ciliated neuron UMAP from **Fig. 3A** that is shown in panels B and C. This section includes the trajectory of the lineage that produces the ASE, ASJ, and AUA neurons (ABalpppppp/ABpraaappp). (**B**) Expression patterns 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 an HLH-3::GFP protein fusion (series 20160301_hlh-3_OP650_L2 in EPiC (*19*)). 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.

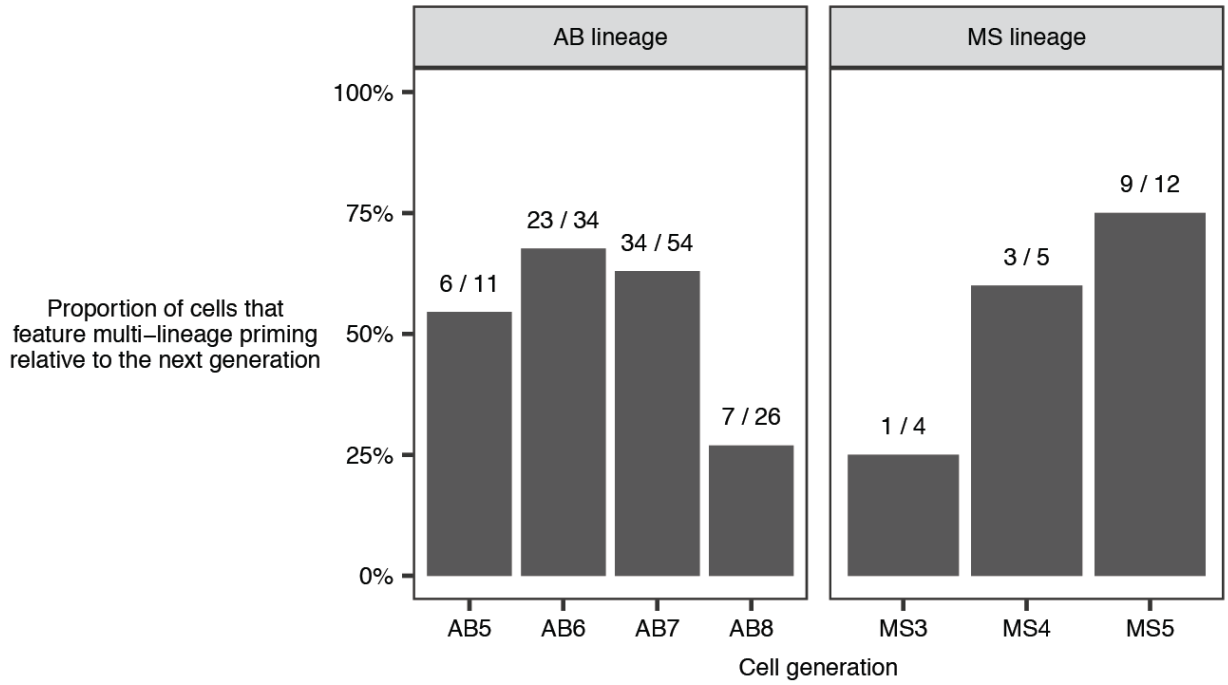


Fig. S24. Prevalence of multilineage priming in *C. elegans*. X-axis shows different cell generations of the ectoderm (AB lineage) and mesoderm (MS lineage). "AB5" refers to the generation produced by 5 divisions of the AB founder cell, and likewise for AB6-8 and MS3-5. Y-axis shows the proportion of 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). Gene expression levels are taken from **Table S8**. 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.

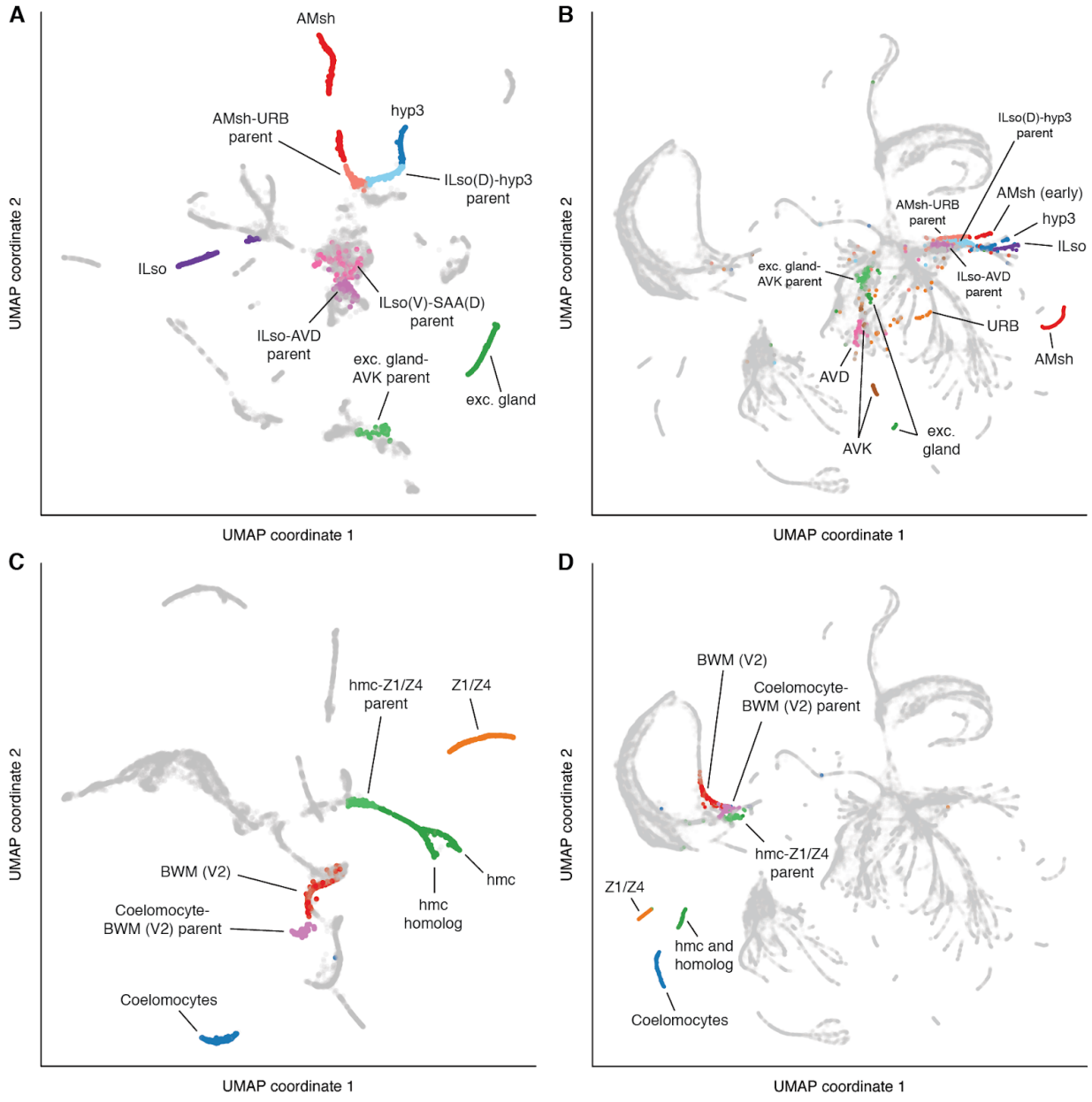


Fig. S25. Examples of lineages that form discontinuous trajectories in UMAP space. (**A**) UMAP of 7,512 glia, excretory cells, and progenitors (same as **Fig. S9**). ILso glia are formed by three input lineages. Two input lineages, the ILso-AVD parent and the ILso(D)-hyp3 parent, form discontinuous 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 **Fig. 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 **Fig. S17**). 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.

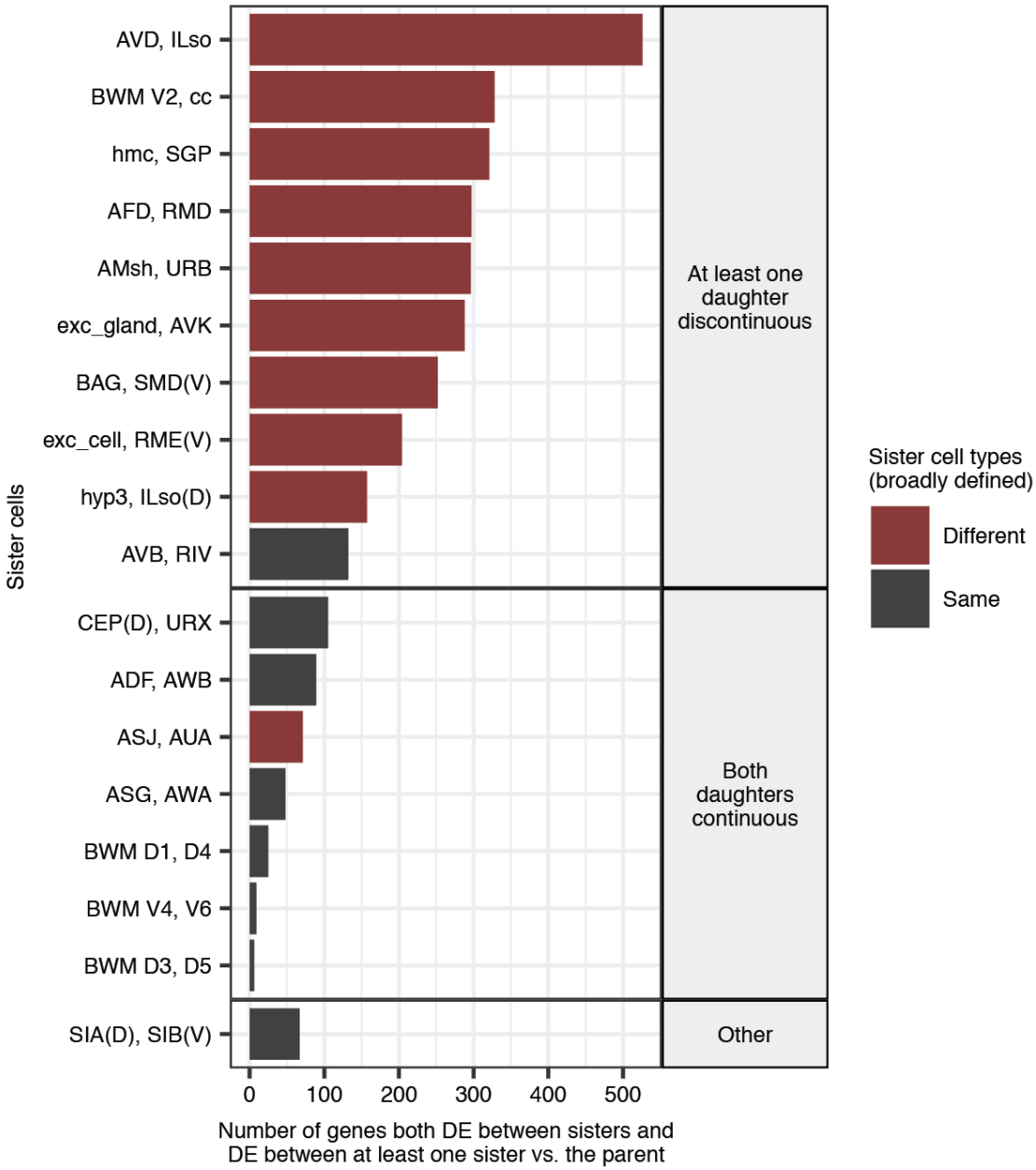


Fig. S26. Counts of differentially expressed genes for lineages that form continuous vs. discontinuous trajectories in UMAP space. Each row (y-axis) corresponds to a pair of terminal sister cells in the ectoderm (AB lineage, generations 9 and 10) or mesoderm (MS lineage, generation 6). Bar length (x-axis) 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 (**Figs. 1A, S3**) and the relevant tissue UMAP (**Figs. 3A, S9-10, S17**). (*Legend continued on the following page*)

Fig. S26 (continued). 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.

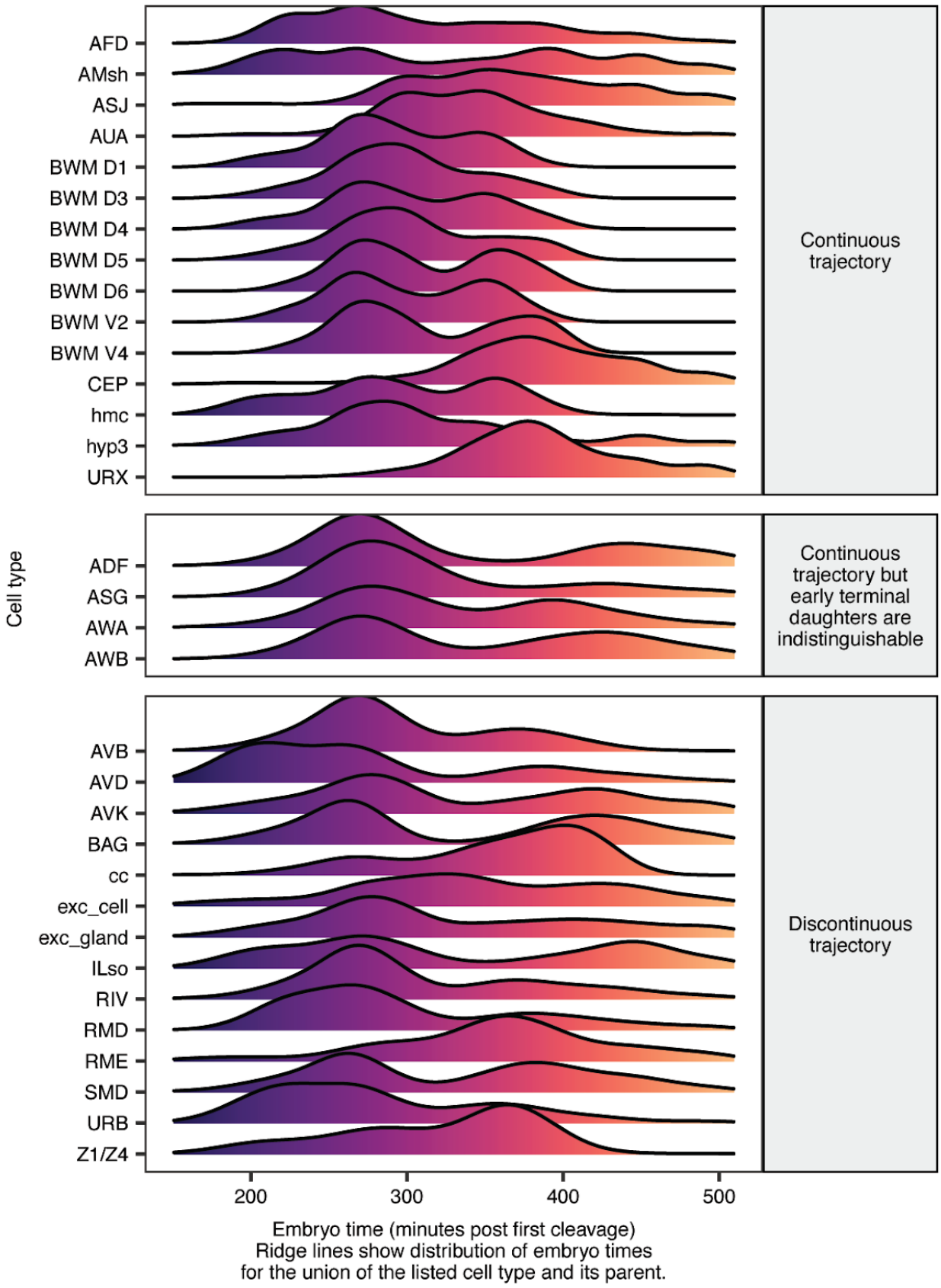

AFD
AMsh
ASJ
AUA
BWM D1
BWM D3
BWM D4
BWM D5
BWM D6
BWM V2
BWM V4
CEP
hmc
hyp3
URX
Continuous trajectory
ADF
ASG
AWA
AWB
Continuous trajectory but early terminal daughters are indistinguishable
AVB
AVD
AVK
BAG
cc
exc_cell
exc_gland
ILso
RIV
RMD
RME
SMD
URB
Z1/Z4
Discontinuous trajectory
Cell type
200
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.

Fig. S27. Embryo time distributions for trajectories included in Fig. S26. 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 **Fig. S26**, 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 (**Figs. 1A, S3**) and the relevant tissue UMAP (**Figs. 3A, S9-10, S17**).

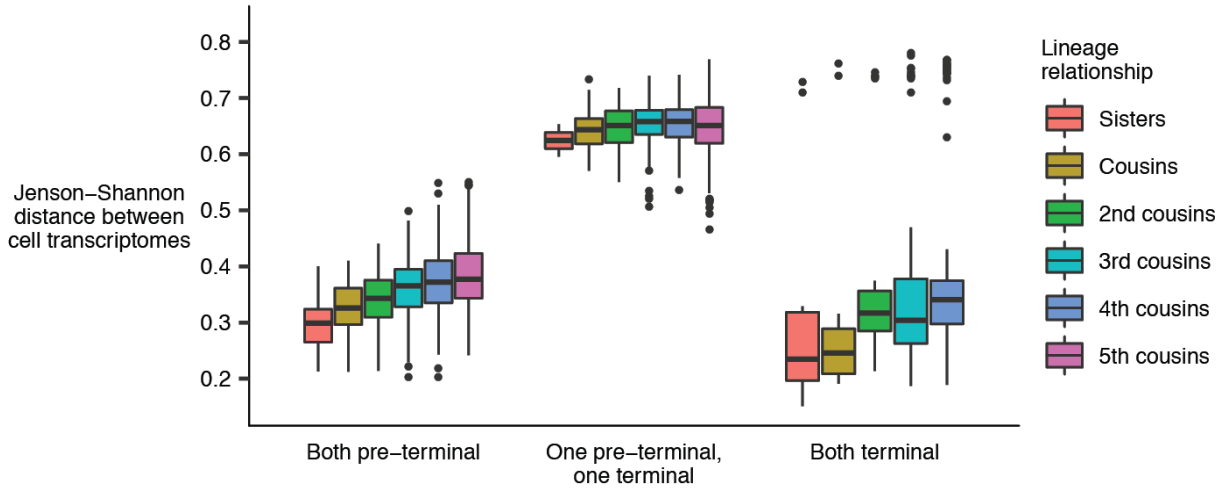


Fig. S28. Lineage distance vs. transcriptome distance in AB generation 8. Jensen-Shannon (JS) distance between the transcriptomes of pairs of cells in AB8, the generation produced by 8 cell divisions since the AB founder cell. Data is faceted by lineage distance and by whether the pair consists of two 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.

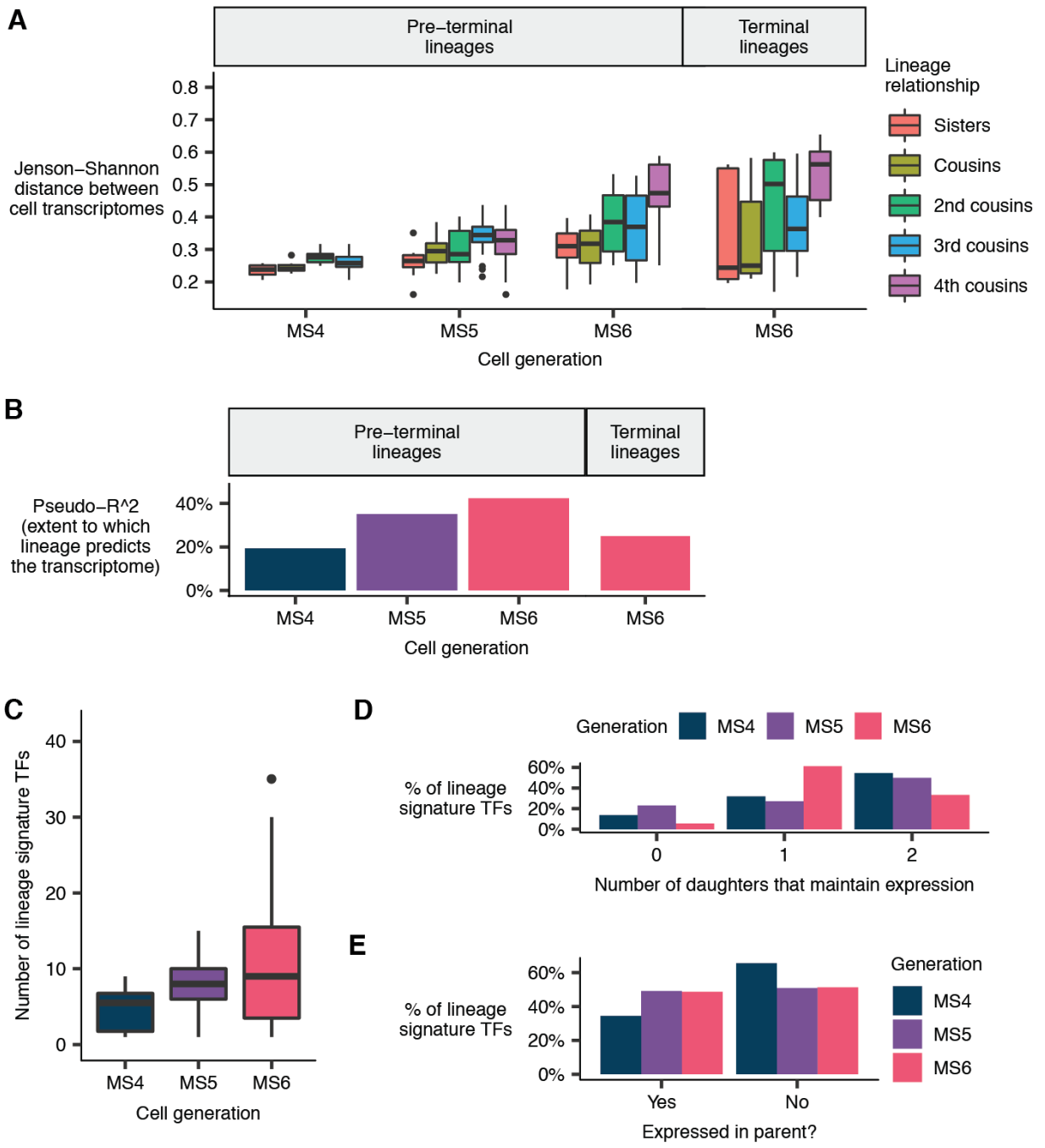


Fig. S29. Correlation between cell lineage and the transcriptome in the mesoderm. (**A**) Jensen-Shannon (JS) distance between the transcriptomes of pairs of mesoderm cells (MS lineage), faceted by cell generation and lineage distance. MS4 refers to the cell generation produced by 4 divisions of the 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 **Methods**). (*Legend continued on the following page*)

Fig. S29 (continued). (**C**) Distribution of the number of "lineage signature transcription factors"—TFs that distinguish a cell from its sister—for a 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.

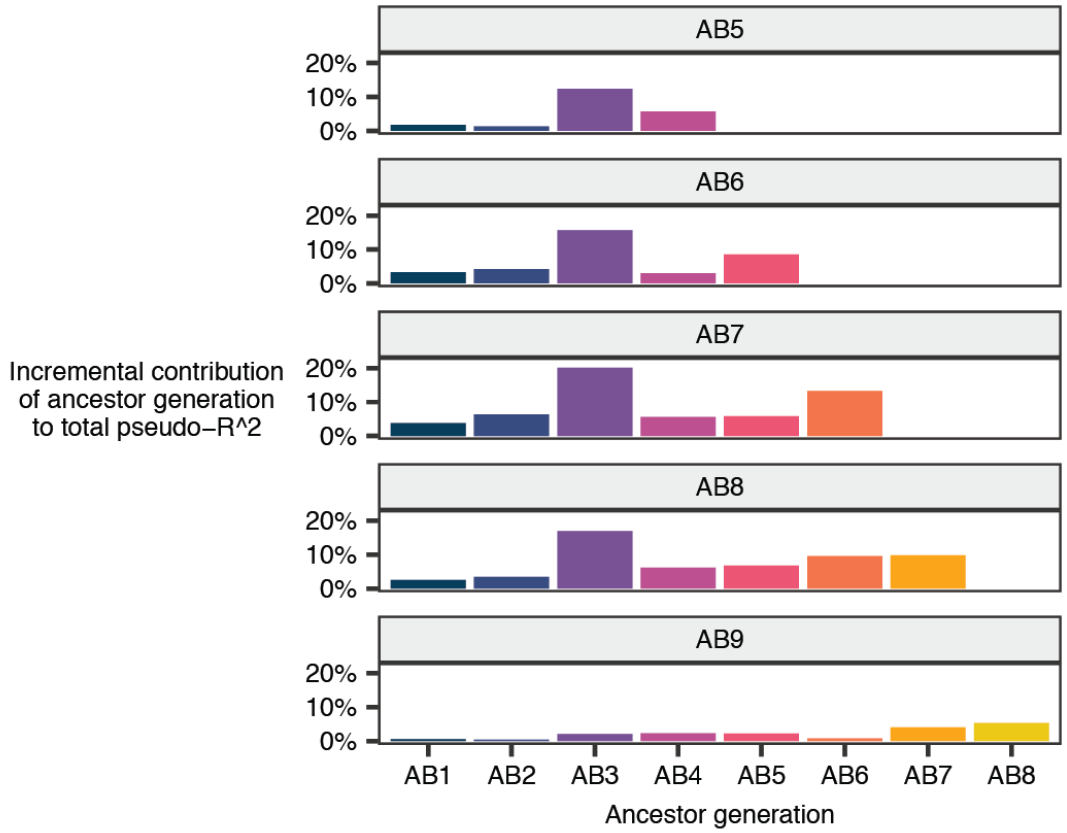


Fig. S30. Both recent and distant ancestry contribute to the ability of the lineage to predict a cell's transcriptome.

In **Fig. 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 used in Fig. 5B and Fig. S29B").

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 Fig. S30".

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 of the bars in a panel is equal to the total pseudo-R^2 for the descendant generation (**Fig. 5B**).

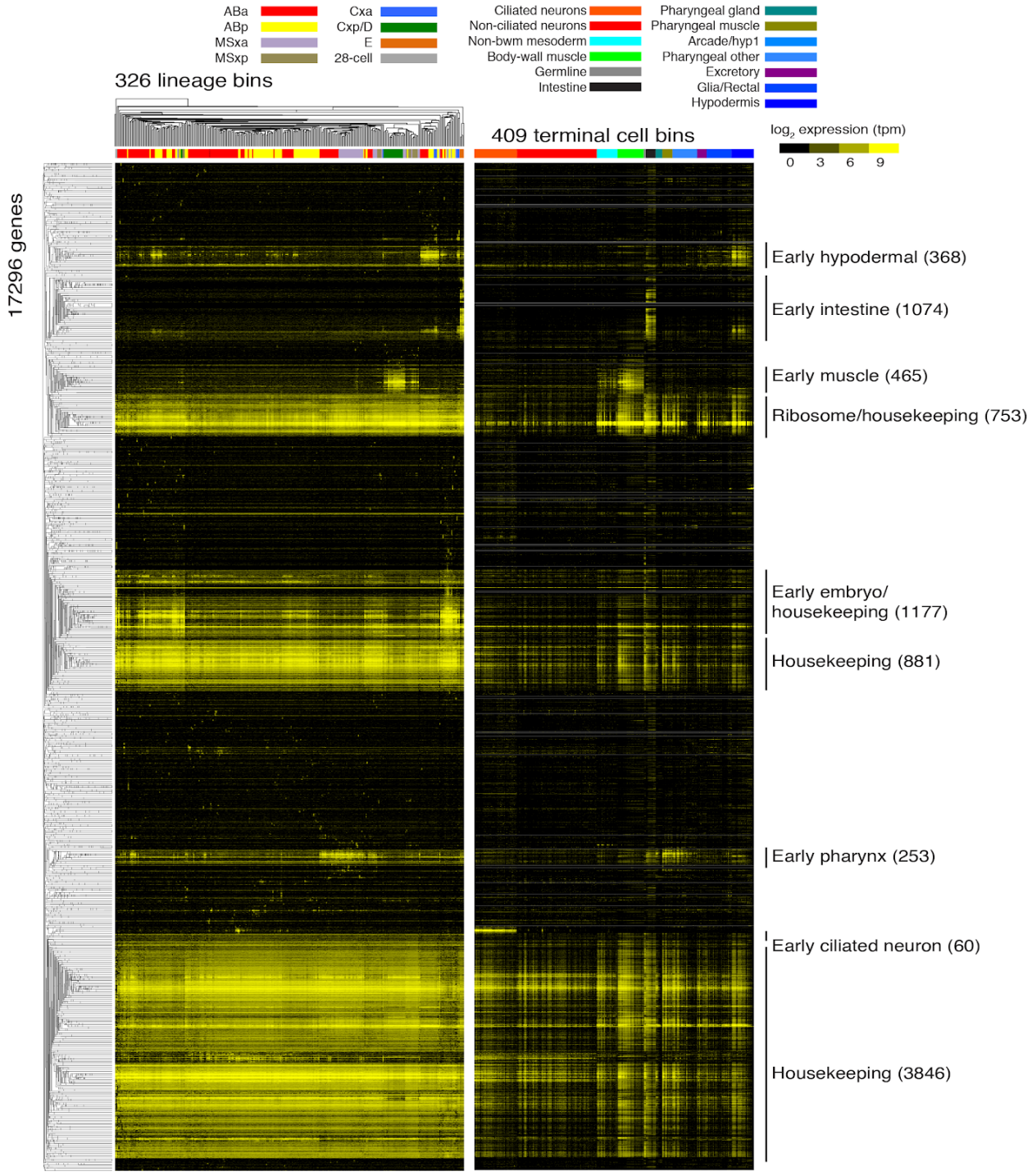


Fig. S31. Hierarchical clustering of progenitor lineage transcriptomes. This heatmap shows the $\log_2$ expression ($\log_2$ transcripts per million) of all genes (rows) that are expressed in at least one pre-terminal lineage (columns). Gene expression values are taken from **Table S8**. Genes and lineages are ordered by hierarchical clustering. The right panel shows the expression values in terminal cell bins (**Table S7**), with genes (rows) ordered by the clustering as generated from the pre-terminal lineages and terminal cell bins (columns) ordered as in **Fig. S32**.

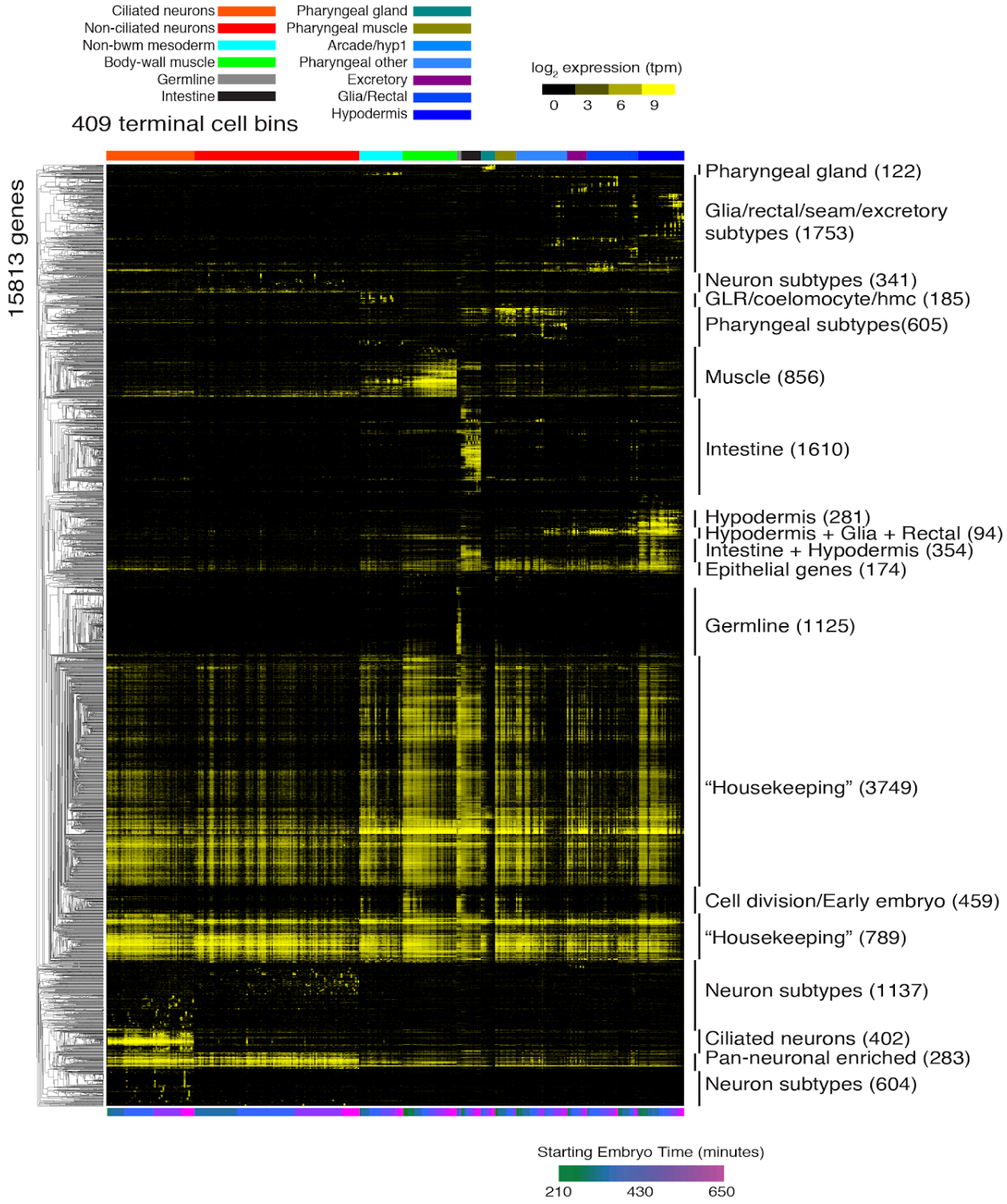


Fig. S32. Hierarchical clustering identifies signatures of tissue and cell type differentiation. This heatmap shows the $\log_2$ expression ($\log_2$ transcripts per million) of all genes (rows) that are expressed in at least one terminal cell bin (columns). Gene expression values are taken from **Table S7.** (*Legend continued on the following page*)

Fig. S32 (continued). 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.

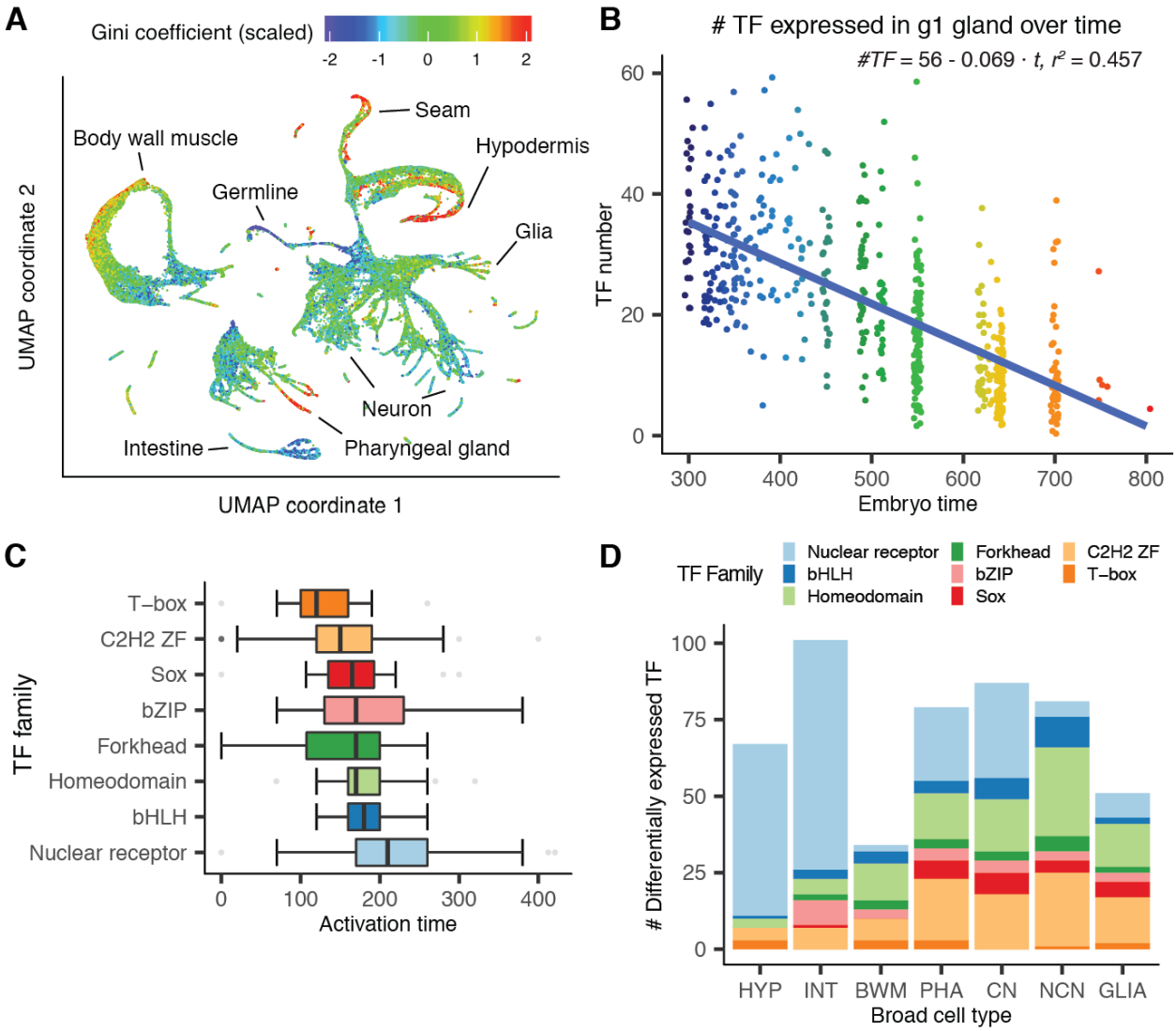
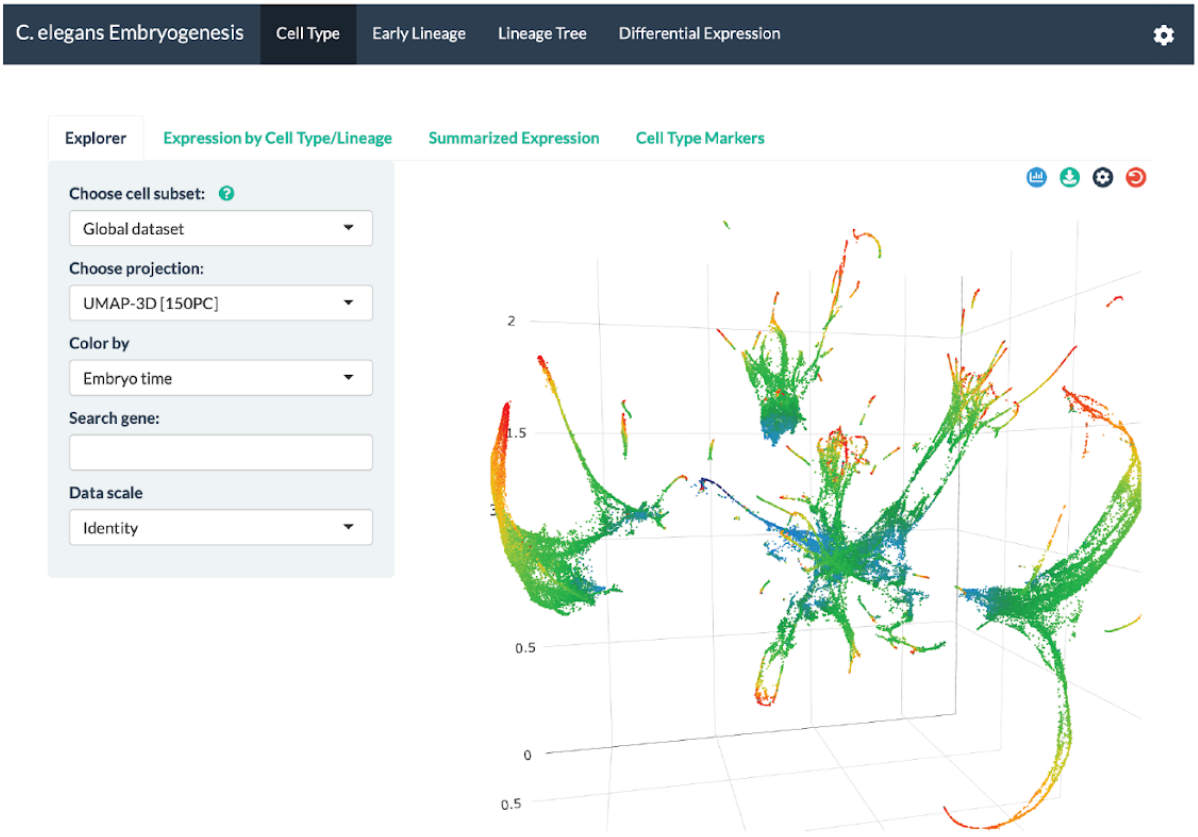


Fig S33. Transcriptome specialization and transcription factor usage across cell types and time. (**A**) A global UMAP with 81,286 cells colored by the Gini coefficient of their gene expression vector, adjusted to correct for sample size bias and scaled by converting to z-scores. High Gini coefficients 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 (*209*). 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.

A


C. elegans Embryogenesis
Cell Type
Early Lineage
Lineage Tree
Differential Expression
Explorer
Expression by Cell Type/Lineage
Summarized Expression
Cell Type Markers
Choose cell subset:
Global dataset
Choose projection:
UMAP-3D [150PC]
Color by
Embryo time
Search gene:
Data scale
Identity
2
1.5
0.5
0
0.5

B

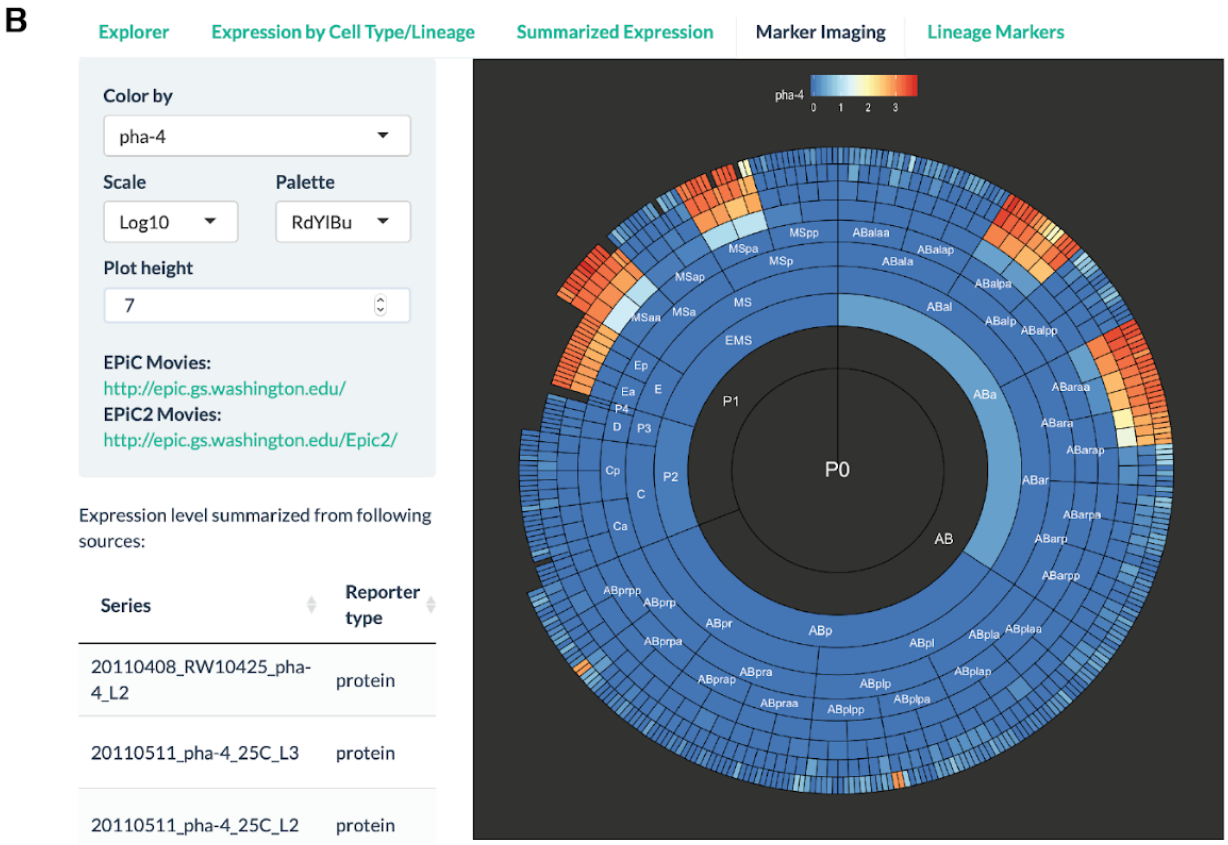

Explorer
Expression by Cell Type/Lineage
Summarized Expression
Marker Imaging
Lineage Markers
Color by
pha-4
Scale
Log10
Palette
RdYlBu
Plot height
7
EPiC Movies:
http://epic.gs.washington.edu/
EPiC2 Movies:
http://epic.gs.washington.edu/Epic2/
Expression level summarized from following sources:
Series
Reporter type
20110408_RW10425_pha-4_L2
protein
20110511_pha-4_25C_L3
protein
20110511_pha-4_25C_L2
protein
pha-4
0 1 2 3
P0
AB
P1
P2
P3
P4
MS
EMS
E
MSa
MSp
MSaa
MSpa
MSap
Ea
Ep
C
Ca
Cp
D
ABa
ABal
ABala
ABalp
ABar
ABara
ABarp
ABaraa
ABarpa
ABarap
ABarpp
ABalaa
ABalpa
ABalap
ABalpp
MSpp
ABp
ABpl
ABpla
ABplp
ABpra
ABprp
ABpraa
ABprpa
ABprap
ABprpp
ABplaa
ABplpa
ABplap
ABplpp
ABpr

Fig. S34. 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 plot 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.

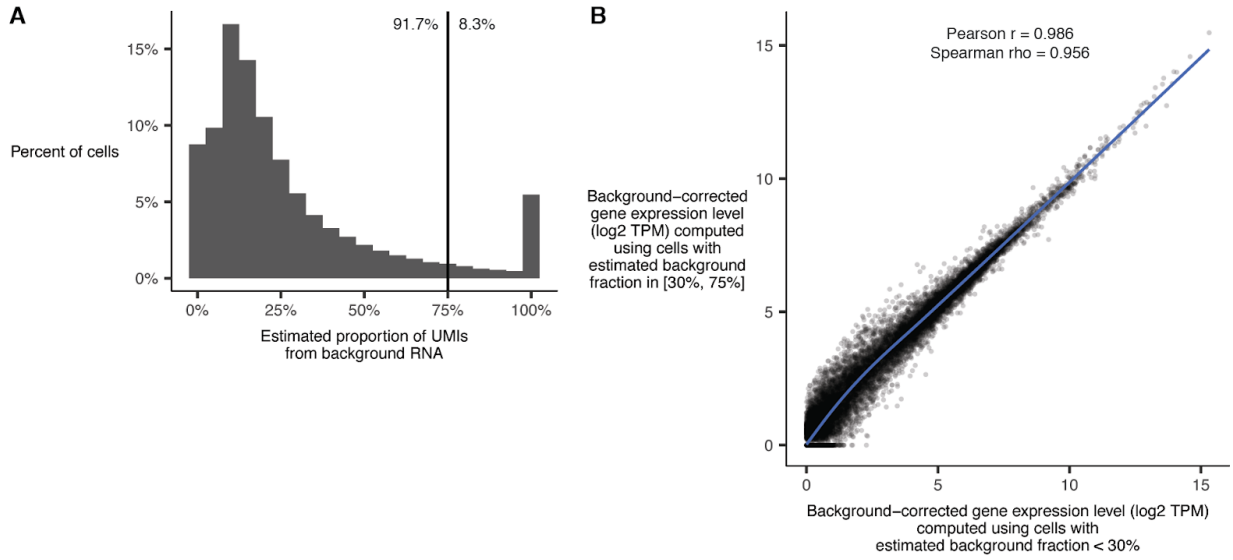


Fig. S35. Distribution of estimates for the proportion of UMIs in a cell that come from background RNA. (**A**) The process for making the estimates is described in the methods section "Per-cell background correction and filtering". Due to the sparsity of the single cell data, the estimates are noisy. 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%].

Supplementary Tables

Tables S1-S16 are available online at:
https://science.sciencemag.org/content/suppl/2019/09/04/science.aax1971.DC1

Data Availability

The raw data have been deposited with the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) under accession code GSE126954. Source code of VisCello (with *C. elegans* data) has been deposited at Github (https://github.com/qinzhu/VisCello.celegans) and Zenodo (*210*). Source code of VisCello for hosting other single cell data has been deposited at Github (https://github.com/qinzhu/VisCello) and Zenodo (*211*). Gene expression movies used in the annotation but not previously published have been deposited at Dryad (doi: 10.5061/dryad.7tg31p7).

Additional Acknowledgements

We thank members of the Murray, Waterston, and Kim labs, and Ben Lehner and Meera Sundaram for providing critical comments on the manuscript. We also thank A. Zacharias, D. Vafeados, M. Corson, R. Terrell, L. Gevirtzman, and P. Weisdepp for their contributions to the EPiC database. **Funding:** This work was funded by NIH grants U41HG007355 and R01GM072675 to RHW, and R35GM127093 and R21HD085201 to JIM. This work was also funded in part by Commonwealth of Pennsylvania Health Research Formula Funds and RM1HG010023 to JK, by U2C CA233285 to KT, by the William H. Gates Chair of Biomedical Sciences (RHW), and by the Allen Discovery Center for Lineage Tracing (JSP, CT). **Author contributions:** JP, CH, JK, RW, and JM conceived and designed the study; CH, PS, EP, HD, and DS performed the experiments; JP, QZ, RW, and JM did the analyses; CT, JK, RW, and JM supervised analyses; JK, JM, and KT supervised the development of VisCello; JP, QZ, JK, RW, and JM wrote the paper. **Competing interests:** The authors have no competing interests. This article was prepared while HD was employed by the University of Pennsylvania. The opinions expressed in this article are the authors' own and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States government.

Chapter 4: Conclusion

In this chapter, I will share a few brief thoughts on how my work may enable future research into *C. elegans* development, and inform future research into vertebrate development.

Identifying the transcription factors that regulate *C. elegans* cell fate decisions

The map of the *C. elegans* cell lineage created by Sulston *et al. (10, 12)* showed us what cell fate decisions are made at each cell division in the worm's development. The transcriptional atlas presented in this dissertation identifies genes whose expression distinguishes each pair of differently-fated sister cells. These resources provide foundations from which the *C. elegans* research community can develop mechanistic models of how each cell fate decision is made.

One step towards building such models will be to determine precisely which transcription factors (TFs) regulate each cell fate decision. Historically, identifying such TFs has been a laborious process. For example, one might use a forward genetic screen to identify mutants that have an abnormal developmental or behavioral phenotype (*9*); use linkage analysis to find the genes that those mutants affect (*212*); identify the subset of those genes that are predicted to be transcription factors (*213*); and then do further experiments to confirm that the candidate TFs regulate the particular cell type that one is interested in. This general approach has allowed researchers to identify many associations between TFs and cell types, e.g. finding TFs that are necessary and sufficient to specify the fates of different neuron types (*193*).

Having a *C. elegans* gene expression atlas should accelerate the process of regulatory TF discovery. It will facilitate "reverse" genetic screens, in which one first decides on a set of genes to study, and then characterizes the phenotype of mutations affecting those genes. TFs that are highly expressed in cells that have a particular cell fate but are not expressed in their sisters could be good candidates for reverse genetic analysis. As I describe in Chapter 3, there are usually many such differentially expressed TFs for any given pair of sister cells, so ideally one would use additional data to further narrow down the list of candidates. For example, one could use ChIP-seq (*81*, *183*) or CUT&RUN (*214*) assays to validate that a candidate TF binds to the promoters of genes that are expressed in the cell type of interest (see Chapter 2, Figure 5).

Regardless of how one discovers a candidate TF, it can be challenging to demonstrate that it has a causal role in regulating a particular cell type. One approach is to show that expression of a cell type specific marker gene is abolished in the context of a TF loss-of-function mutant (*215*, *216*). A limitation of this approach is that a TF may be important for the biological function of the cell type, but may not be necessary for the expression of that particular marker gene. Single mutant analyses will also fail to identify TFs that have redundant functions (*70*, *217*, *218*). Lastly, many mutations are lethal, making it hard to characterize their cell-type specific effects.

An alternative approach is to show that ectopic expression of a TF causes a cell to express marker genes specific to another cell type (*67*, *219*, *220*). This approach is informative if it works, but it will not work in the many cases where combinatorial interactions between multiple TFs are necessary to specify a cell fate (*221*).

Using single cell RNA-seq as a phenotypic readout could allow one to observe the effects of knocking out or ectopically expressing a TF in each cell type of the worm in a single assay. To reduce the number of irrelevant cells sequenced, one could use fluorescence-activated cell sorting to isolate a broad class of cells, e.g. all neurons, and then look for differential expression between wild-type and mutant worms for each cell type within that class.

Learning the "grammar" of *C. elegans* regulatory DNA

The ability of transcription factors to induce gene expression is mediated by their ability to bind to DNA in a sequence-specific manner. Ideally, a "mechanistic" model of gene regulation would include not just a map of which TFs regulate which cell types, but also maps of where those TFs bind in the genome, a model that explains why they bind there, and a model that explains how that TF binding relates to transcriptional output.

Constructing such a comprehensive model would require solving several open problems in computational molecular biology, and it is probably more feasible to attempt it with human cell lines than with *C. elegans*. A somewhat simpler problem to tackle is to try to predict cell type specific gene expression directly from DNA sequence, without requiring one to predict exactly which TFs bind to which sequences. To make a very loose analogy, to solve this sub-problem, one would have to learn the "words" and "grammar" that are used to compose regulatory DNA sequences in an organism, but one would not have to learn the "definition" of each "word" (i.e. which TF it corresponds to). I previously wrote a review paper that discusses this problem in the context of vertebrate gene regulation (*201*). I believe, however, that it may be possible to use the data from this dissertation to tackle the problem in *C. elegans*.

Convolutional neural networks offer a promising approach to predict gene expression from DNA sequence (*222*). I did some preliminary experiments using a convolutional neural network to predict *C. elegans* gene expression based on promoter sequences. For brevity, I'll omit the data and summarize the results: the model had mediocre performance for broadly expressed genes and very poor performance for cell type specific genes. This is not unexpected. Neural networks require large amounts of data to train, and there are only ~10,000 genes expressed at non-negligible levels in the *C. elegans* embryo; and of these, there may only be ~50-500 genes that are specific to or highly enriched in any particular cell type. So there simply weren't enough promoter sequences to learn the *C. elegans* regulatory grammar.

I believe this problem can be solved (at least partially) by taking advantage of the conservation of the developmental program between *C. elegans* and its evolutionary relatives. *C. briggsae*, a nematode whose most recent common ancestor with *C. elegans* lived ~60-110

million years ago (*223*) has an almost identical cell lineage to *C. elegans* (*224*). Single cell RNA-seq data from *C. briggsae* embryos (not shown) produces a UMAP plot that is very similar to the *C. elegans* UMAP from Chapter 3, suggesting that gene expression patterns are also highly conserved. Other species in the *Caenorhabditis* genus are more phenotypically distinct (for example *C. nigoni*, unlike *C. elegans* or *C. briggsae*, is a species with female and male sexes that occur approximately equally frequently (*225*)), but these species still share very similar developmental programs.

Therefore, I think it should be feasible to train a model that works under the assumption that a *C. briggsae* gene will have approximately the same expression pattern across cell types as its ortholog in *C. elegans*, and likewise for any other *Caenorhabditis* species. In other words, though the orthologous genes' promoter sequences have diverged over millions of years, they should maintain approximately the same regulatory information content. If this assumption is valid for most genes, one could train a neural network that predicts *C. elegans* gene expression based on the promoter sequences of gene orthologs in each other *Caenorhabditis* species. The amount of training data (i.e. promoter sequences) available to the neural network would scale linearly with the number of species included in the analysis. Hopefully, this large amount of training data would enable the model to learn a more complex regulatory grammar than the model that only used *C. elegans* promoter sequences.

There is no reason I couldn't have done this analysis myself, as all the necessary data is available. I just never got around to it.

Reconstructing the transcriptional histories of vertebrate cell lineages

Single cell RNA-seq has been used to reconstruct the gene expression trajectories followed by various vertebrate cell types in development (*30*, *174–177*, *226–229*). The analysis presented in Chapter 3 suggests that one should be cautious, however, before claiming that branches in these computationally reconstructed developmental trajectories correspond to lineage relationships between cell types. In some cases, distant lineages converge to similar or identical transcriptional states; in others, sister lineages differentiate extremely abruptly. Both of these scenarios can confound methods that infer developmental trajectories from sc-RNA-seq data alone. To make a stronger statement, trying to reconstruct an unknown cell lineage solely by applying computational tricks to transcriptomic data is probably a bad idea. I hope that one of the main impacts of my work will be to dissuade other researchers from attempting it.

To overcome the limitations of transcriptomic data analysis, a few labs have developed new experimental methods that use CRISPR-Cas9 genome editing to physically record lineage information in single cells and allow it to be read out through sequencing (*83*, *230*, *231*) or imaging (*232*) assays. Importantly, these assays can be coupled to single cell RNA-seq (*230*, *231*) or seqFISH (*232–234*), allowing one to profile gene expression in the same single cells.

Therefore, in theory, one could use these assays to reconstruct the transcriptional histories of vertebrate cell lineages.

There are still a few challenges that need to be overcome before CRISPR-based lineage tracing technologies can be used at scale. The experimental methods need to be improved to allow for more precise control over the rate and timing of the genome edits that record lineage information. The computational methods used to reconstruct lineage trees from the experimental data also need to be improved to better account for sequence biases in the CRISPR-Cas9 genome editing system. Lastly, computational methods and "best practices" need to be developed to reconcile lineage trees inferred from multiple replicate experiments, both with each other and with gene expression trajectories inferred from single cell RNA-seq data.

Most of the labs that work on these methods have been coordinating their efforts through the Allen Discovery Center for Lineage Tracing (http://www.allen-lineage.org/), and I expect much groundbreaking work to come out of this institution in the coming years.

Appendix A: Glossary of *C. elegans* hermaphrodite anatomy

C. elegans has an impressively complex anatomy for an organism with less than 1,000 somatic cells. In this appendix, I provide brief descriptions of each of the cell types in the worm, summarized from more detailed descriptions that are available at the website *WormAtlas (235)* (https://www.wormatlas.org/hermaphrodite/hermaphroditehomepage.htm).

C. elegans has two sexes, hermaphrodites and males. Males are very rare, accounting for only 0.1%-0.2% of individuals (*236*, *237*). Here, I will describe the anatomy of the hermaphrodite.

Arcade cells. A set of nine cells that form part of the worm's buccal (mouth) epithelium. Each arcade cell has a cell body near the anterior bulb of the pharynx and extends a thin cytoplasmic process towards the mouth. There are two groups of arcade cells: three anterior arcades and six posterior arcades. The processes of the anterior arcades merge to form a ring-shaped syncytium, and those of the posterior arcades form a second syncytium. Though not technically part of the pharynx, in terms of gene expression, arcades are more similar to pharyngeal cell types than those of any other tissue.

Anal depressor muscle. A single non-striated muscle cell that is located at the far posterior end of the worm. This cell expands and contracts the rectum to enable defecation.

Anal sphincter muscle. A single non-striated muscle cell that wraps around the junction between the intestine and the rectum. The cell helps squeeze the posterior intestine and push waste into the rectum to be defecated.

Body wall muscle. Analogous to skeletal muscle in other organisms, body wall muscle (BWM) is a striated muscle cell type that allows the worm to move around. The 95 BWM cells in the adult worm are arranged along the anterior-posterior axis in four quadrants (ventral left/right and dorsal left/right). Fourteen BWM cells are post-embryonic, produced by divisions of the M cell.

Coelomocytes. A macrophage-like cell type that has immune and scavenging functions. Unlike macrophages, they are not actively migratory. There are six coelomocytes in the adult worm, grouped into three pairs that located in the anterior, middle, and posterior of the worm respectively. Two coelomocytes are post-embryonic, produced by divisions of the M cell..

Distal tip cells. A set of two cells. Each distal tip cell forms a cap over one of the two arms of the worm's gonad, wrapping around the distal-most germ cells. The distal tip cells guide the elongation of the gonad arms, and promote mitosis and inhibit meiosis in the distal germ cells.

Excretory cell. Also referred to as the “excretory canal cell”. A single, large, H-shaped cell located near the posterior bulb of the pharynx. This cell helps regulate osmotic pressure and ion concentrations in the worm, analogous to the renal system in vertebrates. Two long “canals” grow from the cell body and span the entire length of the worm. Fluid flows through the canals to a junction beneath the excretory cell body. This junction connects to the excretory duct and pore cells, the latter of which forms an opening to the outside of the worm.

Excretory duct. A single cell that is part of the excretory system. It forms a duct that connects the excretory cell to the excretory pore.

Excretory gland. Two cells that are part of the excretory system and fuse to form a single binucleate cell. This cell connects to the excretory duct and pore and secretes various materials in membrane-bound vesicles.

Excretory pore. Also called the “excretory socket”. A single cell that is part of the excretory system. It forms a pore that allows fluid to flow out of the worm. In the embryo, a cell called G1 acts as the excretory pore. After hatching, the G1 cell becomes a neuroblast and a descendant of another cell (called G2) becomes the pore cell.

Germline (Z2/Z3). A set of two germline cells that are born in the early embryo. After hatching, they divide to produce eggs and sperm.

Glia. *C. elegans* glia provide structural support to ciliated sensory neurons and guide the extension of their dendrites. Unlike vertebrate glia, they do not form myelin, and are not found in association with non-ciliated neurons. There are two main *C. elegans* glial cell types. Sheath cells tend to have large cell bodies that completely envelop (ensheath) the dendrites of ciliated sensory neurons for most of their length. Socket cells tend to be smaller and wrap around the ends of the sensory cilia, forming pores that extend either into the cuticle or to the exterior of the worm. Like hypodermal cells, socket cells contribute to cuticle synthesis.

GLR cells. A set of six cells positioned in a ring around the middle of the pharynx. While the function of the GLR cells is not precisely known, they are thought to have a role in guiding muscle cells to their appropriate positions during development (*63*).

hmc (head mesodermal cell). A cell that is produced by the embryonic MS (mesoderm) lineage and migrates to the head of the worm, sitting above the back-end of the terminal bulb of the pharynx. The function of this cell is not known.

hmc homolog. A cell that is produced by a lineage that is bilaterally symmetric to the lineage that produces the head mesodermal cell (hmc). Unlike the hmc, it undergoes programmed cell death in late embryogenesis.

Hypodermis. The hypodermis lies beneath and produces the cuticle, the collagenous outer protective layer of the worm. Other specialized epithelial cell types, including the seam cells, P cells, arcade cells, pharyngeal epithelial cells, rectal epithelial cells, and socket cells also contribute to cuticle production. The *C. elegans* cell lineage produces 162 hypodermis cells (46 before hatching, and 116 after hatching). Most of these fuse to form ring-shaped, multinucleated syncytia. Six syncytia, called hyp1, hyp2, hyp3, hyp4, hyp5, and hyp6, form the cuticle around the mouth, head, and "neck". The cuticle around the main body is formed by a very large syncytium called hyp7, which contains 139 nuclei. The cuticle around the tail is produced by three single, non-syncytial cells (hyp8, hyp9, and hyp11) and one binucleate syncytia (hyp10).

Intestinal muscle. A pair of non-striated muscle cells that squeeze the posterior intestine and facilitate defecation.

Intestinal-rectal valve. A pair of cells that form a small passageway that connects the intestine to the rectum.

Intestine. The intestine (formed by 20 cells) digests the worm's food. Intestine cells are the only cell type produced by the endoderm lineage, and are positioned in bilaterally symmetric pairs along the anterior-posterior axis of the worm, starting just behind the pharynx and extending to the rectum. After hatching, and after each larval molt, most intestine cells undergo endoreduplication, a process in which nuclei divide without cytokinesis. Adult intestine cells have a chromosome number of 32N (corresponding to 5 rounds of endoreduplication).

M cell. A cell that is produced by the embryonic MS (mesoderm) lineage. After hatching, it divides to produce post-embryonic body wall muscle cells and sex myoblasts, which in turn produce uterine and vulval muscle cells.

Neurons. There are 302 neurons in an adult hermaphrodite *C. elegans*, 80 of which are produced after hatching. The 302 neurons have been divided into 118 classes based on location, morphology, and gene/protein expression. Neuron class names are typically three capital letters (e.g. ASE, RIG, or PVQ), though some are two letters and a number (e.g. DB2 or DD5). If you see a *C. elegans* cell type that is named with three capital letters, it is probably a neuron. Ciliated neurons, which sense chemicals, odors, temperature, osmolarity, and oxygen/CO_2 concentrations, have a distinct gene expression profile from all other neurons. Not all sensory neurons are ciliated, however. Touch receptor neurons, for example, do not have cilia.

P cells. A set of twelve epithelial cells that are positioned on the ventral side of the embryo. These cells act like hypodermis in the embryo, contributing to cuticle synthesis. After hatching, they divide to produce neurons, hypodermal cells, and the vulva.

Pharynx. A muscular, tube-like organ in the head of the worm that forms the anterior-most part of the digestive system. The pharynx grinds up the worm's food (bacteria) and pumps it into the intestine. The pharynx contains two prominent sphere/oval-shaped "bulbs", one in the middle (called the "metacorpus" or "anterior bulb"), and one at the back (called the "posterior bulb"). Interestingly, the pharyngeal cavity has three-fold symmetry instead of bilateral symmetry. See https://www.wormatlas.org/hermaphrodite/pharynx/mainframe.htm for an anatomical diagram.

- **Pharyngeal epithelial cells**. A set of nine specialized epithelial cells in the anterior part of the pharynx. They contribute to producing the cuticle that forms the mouth of the worm, and are morphologically similar to the arcade cells.
- **Pharyngeal glands**. A set of five cells that secrete digestive enzymes into the pharynx. The gland cells are divided into three sub-types: g1A (two cells), g1P (one cell), and g2 (two cells).
- **Pharyngeal marginal cells**. A set of nine cells that provide structural reinforcement to the pharynx and separate different groups of pharyngeal muscles.
- **Pharyngeal muscle**. Non-striated muscles that allow the pharynx to pump material into the intestine. Pharyngeal muscles are morphologically diverse and are divided into eight classes: pm1, pm2, etc., up to pm8. The cell lineage produces 37 pharyngeal muscle cells, but many of them fuse to form syncytia.
- **Pharyngeal neurons**. A set of twenty neurons that control the pumping of the pharynx.
- **Pharyngeal-intestinal valve**. A set of six cells that form a narrow valve that connects the posterior end of the pharynx to the anterior end of the intestine.

Rectal epithelium. A set of six epithelial cells in the posterior of the worm that form the rectal passage. Each of the six cells has a specific name: B, F, K, K′ (K prime), U, and Y.

Rectal gland. A set of three cells that lie adjacent to the intestinal-rectal valve. Their function is not well-characterized, but they may secrete digestive enzymes into the posterior intestine. Confusingly, some older papers refer to them as "rectal epithelial cells", but they are not the same cell type as the B, F, K, K′, U, and Y cells, which are called "rectal epithelial cells".

Seam cells. A set of cuticle-synthesizing, hypodermis-like cells that are arranged in lines on the left and right side of the worm, spanning from the head to the tail. Unlike regular hypodermal cells, they continue to divide after hatching, producing neurons and glia as well as more seam

cells. There are 20 seam cells in the embryo and 32 in the adult worm. During post-embryonic development, each line of seam cells fuses to form a syncytium.

Sex myoblasts. A set of two progenitor cells that are born in the late L1 larval stage from divisions of the M cell. In the L3 and L4 larval stages, the sex myoblasts divide further to produce the eight uterine and eight vulval muscles.

Somatic gonad precursors (Z1/Z4). A set of two cells produced by the embryonic MS (mesoderm lineage). After hatching, they divide to produce the somatic gonad.

Somatic sheath cells. Cells that form a thin layer over the germline cells in each gonad arm. They are not present in any of the developmental stages profiled in Chapters 2 and 3.

Spermatheca. Cells that form a tube in each gonad arm that stores sperm. Eggs from the distal gonad are pushed through the spermatheca and fertilized by the sperm contained within. Spermatheca are not present in any of the developmental stages profiled in Chapters 2 and 3.

Spermatheca-uterine valve. Cells that form a valve that connects the spermatheca to the uterus. They are not present in any of the developmental stages profiled in Chapters 2 and 3.

Uterus. A tube that connects the two arms of the gonad to the vulva. Fertilized eggs are pushed through the uterus and out the vulva by the uterine and vulval muscles. The uterus is not present in the developmental stages profiled in Chapters 2 and 3.

Uterine muscles. A set of eight non-striated muscle cells that push eggs through the uterus. The uterine muscles are not present in the developmental stages profiled in Chapters 2 and 3.

Vulva. A small opening through which the worm's eggs are laid. The vulva is not present in the developmental stages profiled in Chapters 2 and 3.

Vulval precursor cells. A set of three progenitor cells that are born in the L1 larval stage from divisions of a subset of P cells. In the L3 larval stage, the vulval precursor cells divide further to produce the vulva.

Vulval muscles. A set of eight non-striated muscle cells that push eggs through the vulva. The vulval muscles are not present in the developmental stages profiled in Chapters 2 and 3.

Appendix B: *C. elegans* lineage nomenclature

Each cell in the *C. elegans* embryonic cell lineage has a unique name that specifies its ancestry, starting a "founder cell" in the early embryo. Here are a few illustrative examples:

- AB is the name of a founder cell produced by the first division of the zygote.
- The AB cell divides to produce two daughter cells. Cleavage occurs roughly perpendicular to the anterior-posterior axis of the embryo. The anterior and posterior daughters are named ABa ("a" = anterior) and ABp ("p" = posterior) respectively.
- ABa divides to produce two daughter cells. Cleavage occurs roughly parallel to the anterior-posterior axis. The left and right daughters are named ABal ("l" = left) and ABar ("r" = right) respectively.
- The name ABalapppaaa refers to a cell that was produced by nine divisions of the AB founder cell: AB -> ABa -> ABal -> ABala -> ABalap -> ABalapp -> ABalappp -> ABalapppa -> ABalapppaa -> ABalapppaaa. This cell is a terminal cell that differentiates into a neuron, called ALA. ALA is the anatomical name for the cell, while ABalapppaaa is the lineage name for the cell. Both names refer to the same cell.
- ABalapppaap is the lineage name for another neuron. You can tell that ABalapppaap is the sister of the cell ABalapppaaa because their names differ by only the last letter. They are both daughters of the cell ABalapppaa.
- The cell ABalapppapa is a "cousin" of the cell ABalapppaaa. Their most recent common ancestor is their grandparent, ABalapppa.
- MSappaap refers to a cell that produced by six divisions of the MS founder cell.
- MSappaap is bilaterally symmetric to another cell, MSpppaap. An abbreviated way to refer to this pair of cells is to use the name MSxppaap. Here, "x" means "either a or p".
- The ASJ neurons are a pair of bilaterally symmetric neurons with the same cell type. The lineage of the left ASJ neuron (ASJL) is ABalpppppppa. The lineage of the right ASJ neuron (ASJR) is ABpraaappppa. Together, the lineage of the ASJ neurons may be listed as ABalpppppppa/ABpraaappppa. Here, we cannot use the "x" notation as the most recent common ancestor of the two cells is the AB founder cell.

All of terminal cells in the *C. elegans* embryo descend from one of six specially-named founder cells in the early embryo. These founder cells are:

- AB, which founds an ectodermal lineage, producing mostly hypodermis, neurons, glia,

and pharyngeal cells.

- MS, which founds a mesodermal lineage, producing body wall muscle, pharyngeal cells, neurons, the somatic gonad, and a few other miscellaneous cell types.
- E, which founds an endodermal lineage, producing the intestine.
- C, which founds a lineage that produces both mesoderm (body wall muscle) and ectoderm (hypodermis, and two neurons).
- D, which founds a mesodermal lineage, producing body wall muscle.
- P4, which divides to produce two germline cells that don't divide further until after hatching. These cells are given special names, Z2 and Z3 (instead of being named P4a/p).

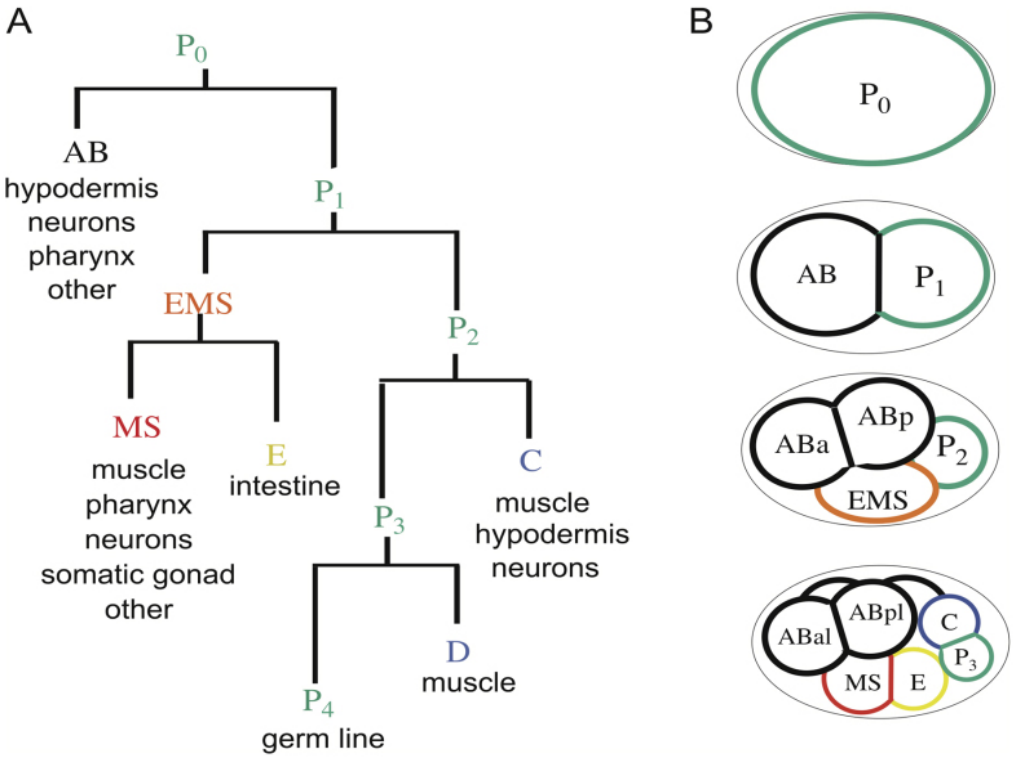


Figure reproduced from *Wormbook* (*238*) (https://www.ncbi.nlm.nih.gov/books/NBK19708/) (**A**) Lineage tree for the early *C. elegans* embryo. (**B**) Diagram of cell divisions in the early embryo.

Some cells in the *C. elegans* embryo divide further after hatching. Like the embryonic founder cells, these cells are given special names. For example, the cell with the embryonic lineage MSapaapp is called the "M" cell. It divides after hatching to produce body wall muscle, coelomocytes, and sex muscles. M.drpp refers to a body wall muscle cell that is produced by four divisions of the M cell. Here, we put a period (".") after the name of the founder cell ("M") to indicate that the divisions are postembryonic.

Appendix C: Computational analysis of sc-RNA-seq data

In this dissertation, I use data from large-scale sc-RNA-seq experiments to create an atlas of gene expression in *C. elegans* development. While the focus of my work is primarily on *C. elegans* biology and not methods development, it is important to be aware of the limitations of currently available methods and how they affect the biological interpretation of sc-RNA-seq data. Here, I will review the methods used in a "typical" sc-RNA-seq analysis workflow.

Data from single cell RNA sequencing experiments is usually represented as a large matrix in which rows correspond to genes, columns correspond to cells, and values correspond to the number of unique mRNA molecules from a given gene observed in a given cell. The sc-RNA-seq protocols used in this dissertation are able to count unique mRNA molecules because they include a random sequence barcode called a UMI ("unique molecular identifier") in each reverse transcription primer. Reads that contain the same UMI sequence can be inferred to come from the same template molecule. The sc-RNA-seq data matrix is therefore referred to as a "UMI count matrix".

The UMI count matrix of a sc-RNA-seq experiment is typically sparse. A sparse matrix is one in which most values are zero. UMI count matrices are sparse because a large fraction of cellular RNA (up to 90%) tends to be lost throughout the process of sample preparation. Genes that are expressed at low levels therefore "drop out" and are not observed in most cells, causing the UMI count matrix to have many zeroes. UMI count matrices are also "high dimensional," meaning that each cell is associated with a large number of observations (e.g. the expression values for all ~20,000 protein-coding genes in *C. elegans*).

The sparsity and high dimensionality of the UMI count matrix makes it infeasible to directly assess whether two cells have similar or dissimilar gene expression profiles. To get around this problem, the most commonly used software packages for analyzing sc-RNA-seq data, Seurat (*239*, *240*) and Monocle (*179*), convert the UMI count matrix into a smaller matrix that encodes approximately the same information, but is dense (entries are mostly non-zero) and low-dimensional (each cell is associated with ~20-100 values, rather than ~20,000 genes). This smaller matrix is generated using an algorithm called principal components analysis (PCA).

The mathematical details of PCA require a fair amount of text to describe; see Jolliffe and Cadima (*241*) for a review. Qualitatively, this process is similar to the compression of a photograph or other image using the JPEG file format: the overall structure of the image is preserved, but the finest details are intentionally lost in order to reduce the data to a more manageable size. Similarly, PCA allows one to "compress" the UMI count matrix into a small, dense matrix that preserves broad trends in gene expression, like the differences between cell types, but excludes fine-grained gene expression variation, which is assumed to be noise.

After using PCA to reduce the data to a manageable size and dimension, the next step of a typical sc-RNA-seq analysis is to make a visualization in which cells are represented as points

in a 2 or 3 dimensional space. The goal of the visualization is to place cells in positions such that cells with similar gene expression profiles are close together, and cells that have dissimilar gene expression profiles are far apart.

In this dissertation, I use two algorithms to visualize sc-RNA-seq data, t-SNE (*242*) and UMAP (*180*). In qualitative terms, t-SNE works under an assumption that if two cells have dissimilar gene expression profiles, it does not matter much whether their gene expression profiles are extremely dissimilar or only moderately so. Instead, what is important is groups of cells that have very similar gene expression profiles, as these will often correspond to specific cell types. Therefore, when t-SNE decides where to place each cell in the 2D or 3D space, cells with very similar gene expression profiles will always placed close together, but cells with dissimilar gene expression profiles can be placed flexibly, either at a moderate distance or a huge distance. The result is that cells of the same cell type tend to form clusters in t-SNE space, and these clusters are randomly distributed within the space, separated by small gaps.

t-SNE is effective when one's data consists of cells sampled from a set of distinct cell types. It is not effective however for data in which cells are sampled from a continuous biological process or spectrum, e.g. cells that are in the process of differentiation, or cells that form a spatial gradient of gene expression in a complex tissue. In these cases, there is in fact a meaningful difference between cells that have moderately dissimilar vs. extremely dissimilar gene expression profiles, violating t-SNE's main assumption.

UMAP (*180*, *181*) is an algorithm that overcomes this limitation of t-SNE, preserving both local structure in a dataset (keeping cells that are very similar to each other close together) and global structure (keeping cells that are very dissimilar far apart, more-so than cells that are only moderately dissimilar). The theoretical foundations for UMAP are based in Riemannian geometry and algebraic topology, areas of mathematics that I am not familiar with. In practice, it seems to work well.

Both t-SNE and UMAP are "non-linear" methods, meaning that if one follows a straight line in t-SNE or UMAP space, there is not a constant change in gene expression per unit distance. The axes of a t-SNE or UMAP visualization therefore do not have any inherent meaning. Nevertheless, clusters of cells in the visualization will often correspond to a particular cell type. In this dissertation, I identify clusters of cells using the Louvain algorithm (*182*). The input to the Louvain algorithm is a "*k*-nearest-neighbor graph", which is a mathematical graph in which the vertices correspond to the sc-RNA-seq cells and edges connect cells that are nearby in t-SNE or UMAP space. Each cell is connected to the *k* cells that are closest to it (its "nearest neighbors"). The Louvain algorithm takes this graph and labels each cell as being part of a particular cluster. When making cell to cluster assignments, the Louvain algorithm seeks to maximize a statistic called "modularity", which measures how often cells within the same cluster are connected to each other, relative to a null model in which the edges of the graph are randomized. See the Methods section of Chapter 3 for more details.

Once cells have been visualized with t-SNE or UMAP and grouped into clusters, the final step of a basic analysis pipeline is to identify what cell type(s) each cluster corresponds to. This annotation process is described in Chapters 2 and 3. Annotating sc-RNA-seq cells in a biological context that one has never worked in before requires a lot of manual effort, e.g. to look up marker genes from the literature. But if one has a reference sc-RNA-seq dataset that is already annotated, one can use software such as Garnet (*243*) to build supervised classification models. One can then apply these classifiers to annotate new sc-RNA-seq datasets that have a similar cell type composition to the reference dataset.

After identifying cell types, one can move on to study-specific analyses, such as looking for genes that are differentially expressed between experimental conditions, or reconstructing developmental trajectories.

Appendix D: Quality control of sc-RNA-seq data

There are several problems that can negatively affect the quality of data produced by an sc-RNA-seq experiment. Some of these problems, such as the degradation of RNA by RNases, or incomplete tissue dissociation leading to the loss of certain cell types, need to be fixed as part of the experimental protocol. Other problems can be fixed (or at least mitigated) in downstream computational analyses. In this appendix, I will discuss two problems that were particularly relevant to my computational work: doublets and background contamination.

Doublets are instances where what looks like one cell in one's sc-RNA-seq data actually corresponds to two physical cells. Multiplets are instances where one apparent cell corresponds to more than two physical cells; these are much rarer than doublets however, so I will use the term "doublet" for this discussion. The source of doublets in sc-RNA-seq data can vary depending on the experimental protocol used. In droplet-based protocols, doublets arise when the microfluidic device produces a droplet that encompasses two cells instead of one. The doublet rate increases proportional to the number of cells loaded into the device, leading to a trade-off between throughput and data quality. The doublet rate for experiments using 10X Chromium, a commercial droplet-based sc-RNA-seq technology, is ~4-10% for typical throughputs.

In the sci-RNA-seq protocol, described in Chapter 2, doublets arise when two cells happen to be distributed to the same set of wells in the combinatorial indexing procedure. cDNA molecules from these cells thus receive the same sequence barcode. The doublet rate of sci-RNA-seq is (non-linearly) proportional to the number of cells processed in the experiment relative to the number of potential barcodes. In sci-RNA-seq with three-level combinatorial indexing, the number of possible barcodes is so high that doublets can be reduced to negligible frequencies. But doublets are still an issue for sci-RNA-seq with two-level combinatorial indexing. At typical throughputs, two-level sci-RNA-seq involves a doublet rate of 10-15%.

Independent of the protocol used, doublets can also arise when cells physically stick to other cells or cellular debris due to incomplete dissociation of a biological sample. For example, in the experiments described in Chapter 3, the theoretical doublet rate given the experimental design should be ~8%, but I estimated the actual doublet rate to be in the ballpark of 20%.

Doublets confound sc-RNA-seq data analysis. A cluster of doublets can look like a novel cell type that expresses marker genes from more than one real cell type. Doublets can also be assigned to the same cell cluster as good cells, causing cluster-level summary statistics to be inaccurate. Thus, one should always attempt to identify and remove doublets before proceeding to downstream analyses. I describe my approach to doublet filtering in the Methods section of Chapter 3. Another popular approach is a software package called Scrublet (*244*). Scrublet is more automated than my approach, but I suspected it might not be appropriate for data that consists of continuous developmental trajectories instead of discrete cell types, so I didn't use it. I did not rigorously test this hypothesis however.

Another problem that affects sc-RNA-seq data quality is background contamination. This is when cells that are damaged in the course of sample preparation leak RNA into the solution in which they are suspended. In droplet-based sc-RNA-seq protocols, this ambient RNA is uniformly distributed into each droplet and participates in the reverse transcription reactions that produce the cDNA that is later sequenced. Thus, in the final dataset, every cell will have reads that come from this contaminating RNA. In sci-RNA-seq, there are no droplets, but RNA leaked from one cell can find its way into another, because in this protocol, the cells are fixed, making them porous. The end result is the same as with droplet-based protocols: every cell gets a portion of reads that come from contaminating RNA.

Background contamination is especially annoying in sc-RNA-seq data analysis because it can make cell-type specific genes look like they are expressed in every cell type. Furthermore, since the absolute amount of contamination is approximately the same across every cell, the proportion of reads in a cell that come from contamination is inversely proportional to the amount of endogenous RNA in the cell. Small cells that have a low amounts of endogenous RNA can have their cell-type specific gene expression patterns overshadowed by the contaminating RNA.

I describe methods to remove background contamination from sc-RNA-seq data in the Methods section of Chapter 3. I developed these methods independently, but they turn out to be quite similar to those used in a software package called SoupX (*202*).

References

1. M. Kloc, B. Zagrodzinska, Chromatin elimination--an oddity or a common mechanism in differentiation and development? *Differentiation*. **68**, 84–91 (2001).

2. J. Wang, R. E. Davis, Programmed DNA elimination in multicellular organisms. *Curr. Opin. Genet. Dev.* **27**, 26–34 (2014).

3. C. H. Bassing, W. Swat, F. W. Alt, The mechanism and regulation of chromosomal V (D) J recombination. *Cell*. **109**, S45–S55 (2002).

4. O. Strassen, Embryonalentwickelung der Ascaris megalocephala. *Archiv für Entwickelungsmechanik der Organismen*. **3** (1896), pp. 133–190.

5. H. Müller, Beitrag zur Embryonalentwicklung der Ascaris megalocephala. (Aus dem Zoologischen Instsitute zu Leipzig.) (1903), , doi:10.5962/bhl.title.53422.

6. A. v. Goette, Entwicklungsgeschichte der Rhabditis nigrovenosa. *Untersuch. z. Entwick-lungsgeschichte der Würmer. Leipzig* (1882).

7. H. Spemann, Zur entwicklung des strongylus paradoxus. *Zoologische Jahrbücher* (1895) (available at http://agris.fao.org/agris-search/search.do?recordID=US201300279791).

8. E. C. Dougherty, H. G. Calhoun, Possible significance of free-living nematodes in genetic research. *Nature*. **161**, 29 (1948).

9. S. Brenner, The genetics of Caenorhabditis elegans. *Genetics*. **77**, 71–94 (1974).

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

11. U. Deppe, E. Schierenberg, T. Cole, C. Krieg, D. Schmitt, B. Yoder, G. von Ehrenstein, Cell lineages of the embryo of the nematode Caenorhabditis elegans. *Proc. Natl. Acad. Sci. U. S. A.* **75**, 376–380 (1978).

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

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

14. The C. elegans Sequencing Consortium*, Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science*. **282**, 2012–2018 (1998).

15. T. Boulin, J. F. Etchberger, O. Hobert, Reporter gene fusions. *WormBook*, 1–23 (2006).

16. R. Y. N. Lee, K. L. Howe, T. W. Harris, V. Arnaboldi, S. Cain, J. Chan, W. J. Chen, P. Davis, S. Gao, C. Grove, R. Kishore, H.-M. Muller, C. Nakamura, P. Nuin, M. Paulini, D. Raciti, F. Rodgers, M. Russell, G. Schindelman, M. A. Tuli, K. Van Auken, Q. Wang, G. Williams, A. Wright, K. Yook, M. Berriman, P. Kersey, T. Schedl, L. Stein, P. W. Sternberg, WormBase 2017: molting into a new stage. *Nucleic Acids Res.* **46**, D869–D874 (2018).

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

18. C. Mello, A. Fire, DNA transformation. *Methods Cell Biol.* **48**, 451–482 (1995).

19. J. I. Murray, T. J. Boyle, E. Preston, D. Vafeados, B. Mericle, P. Weisdepp, Z. Zhao, Z. Bao, M. Boeck, R. H. Waterston, Multidimensional regulation of gene expression in the C. elegans embryo. *Genome Res.* **22**, 1282–1294 (2012).

20. 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).

21. M. E. Boeck, C. Huynh, L. Gevirtzman, O. A. Thompson, G. Wang, D. M. Kasper, V. Reinke, L. W. Hillier, R. H. Waterston, The time-resolved transcriptome of C. elegans. *Genome Res.* **26**, 1441–1450 (2016).

22. W. C. Spencer, G. Zeller, J. D. Watson, S. R. Henz, K. L. Watkins, R. D. McWhirter, S. Petersen, V. T. Sreedharan, C. Widmer, J. Jo, V. Reinke, L. Petrella, S. Strome, S. E. Von Stetina, M. Katz, S. Shaham, G. Rätsch, D. M. Miller 3rd, A spatial and temporal map of C. elegans gene expression. *Genome Res.* **21**, 325–341 (2011).

23. W. C. Spencer, R. McWhirter, T. Miller, P. Strasbourger, O. Thompson, L. W. Hillier, R. H. Waterston, D. M. Miller 3rd, Isolation of specific neurons from C. elegans larvae for gene expression profiling. *PLoS One*. **9**, e112102 (2014).

24. 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 Res.* **29**, 1036–1045 (2019).

25. T. Hashimshony, F. Wagner, N. Sher, I. Yanai, CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).

26. S. C. Tintori, E. Osborne Nishimura, P. Golden, J. D. Lieb, B. Goldstein, A Transcriptional Lineage of the Early C. elegans Embryo. *Dev. Cell*. **38**, 430–444 (2016).

27. E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, S. A. McCarroll, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. **161**, 1202–1214 (2015).

28. G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, J. H. Bielas, Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

29. J. Cao, J. S. Packer, V. Ramani, D. A. Cusanovich, C. Huynh, R. Daza, X. Qiu, C. Lee, S. N. Furlan, F. J. Steemers, A. Adey, R. H. Waterston, C. Trapnell, J. Shendure, Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*. **357**, 661–667 (2017).

30. J. Cao, M. Spielmann, X. Qiu, X. Huang, D. M. Ibrahim, A. J. Hill, F. Zhang, S. Mundlos, L. Christiansen, F. J. Steemers, C. Trapnell, J. Shendure, The single-cell transcriptional landscape of mammalian organogenesis. *Nature* (2019), doi:10.1038/s41586-019-0969-x.

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

32. D. Ramsköld, S. Luo, Y.-C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, G. P. Schroth, R. Sandberg, Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).

33. A. K. Shalek, R. Satija, X. Adiconis, R. S. Gertner, J. T. Gaublomme, R. Raychowdhury, S. Schwartz, N. Yosef, C. Malboeuf, D. Lu, J. J. Trombetta, D. Gennert, A. Gnirke, A. Goren, N. Hacohen, J. Z. Levin, H. Park, A. Regev, Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*. **498**, 236–240 (2013).

34. Q. F. Wills, K. J. Livak, A. J. Tipping, T. Enver, A. J. Goldson, D. W. Sexton, C. Holmes, Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat. Biotechnol.* **31**, 748–752 (2013).

35. A. A. Pollen, T. J. Nowakowski, J. Shuga, X. Wang, A. A. Leyrat, J. H. Lui, N. Li, L. Szpankowski, B. Fowler, P. Chen, N. Ramalingam, G. Sun, M. Thu, M. Norris, R. Lebofsky, D. Toppani, D. W. Kemp 2nd, M. Wong, B. Clerkson, B. N. Jones, S. Wu, L. Knutsson, B. Alvarado, J. Wang, L. S. Weaver, A. P. May, R. C. Jones, M. A. Unger, A. R. Kriegstein, J. A. A. West, Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32**, 1053–1058 (2014).

36. A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, S. Linnarsson, Brain structure. Cell types in the mouse cortex and

hippocampus revealed by single-cell RNA-seq. *Science*. **347**, 1138–1142 (2015).

37. B. B. Lake, R. Ai, G. E. Kaeser, N. S. Salathia, Y. C. Yung, R. Liu, A. Wildberg, D. Gao, H.-L. Fung, S. Chen, R. Vijayaraghavan, J. Wong, A. Chen, X. Sheng, F. Kaper, R. Shen, M. Ronaghi, J.-B. Fan, W. Wang, J. Chun, K. Zhang, Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*. **352**, 1586–1590 (2016).

38. I. Tirosh, B. Izar, S. M. Prakadan, M. H. Wadsworth 2nd, D. Treacy, J. J. Trombetta, A. Rotem, C. Rodman, C. Lian, G. Murphy, M. Fallahi-Sichani, K. Dutton-Regester, J.-R. Lin, O. Cohen, P. Shah, D. Lu, A. S. Genshaft, T. K. Hughes, C. G. K. Ziegler, S. W. Kazer, A. Gaillard, K. E. Kolb, A.-C. Villani, C. M. Johannessen, A. Y. Andreev, E. M. Van Allen, M. Bertagnolli, P. K. Sorger, R. J. Sullivan, K. T. Flaherty, D. T. Frederick, J. Jané-Valbuena, C. H. Yoon, O. Rozenblatt-Rosen, A. K. Shalek, A. Regev, L. A. Garraway, Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*. **352**, 189–196 (2016).

39. W. Zeng, S. Jiang, X. Kong, N. El-Ali, A. R. Ball Jr, C. I.-H. Ma, N. Hashimoto, K. Yokomori, A. Mortazavi, Single-nucleus RNA-seq of differentiating human myoblasts reveals the extent of fate heterogeneity. *Nucleic Acids Res.* (2016), doi:10.1093/nar/gkw739.

40. C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, J. L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

41. D. Ramsköld, S. Luo, Y.-C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, G. P. Schroth, R. Sandberg, Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).

42. B. B. Lake, R. Ai, G. E. Kaeser, N. S. Salathia, Y. C. Yung, R. Liu, A. Wildberg, D. Gao, H.-L. Fung, S. Chen, R. Vijayaraghavan, J. Wong, A. Chen, X. Sheng, F. Kaper, R. Shen, M. Ronaghi, J.-B. Fan, W. Wang, J. Chun, K. Zhang, Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*. **352**, 1586–1590 (2016).

43. F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, M. A. Surani, mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*. **6**, 377–382 (2009).

44. S. Picelli, Å. K. Björklund, O. R. Faridani, S. Sagasser, G. Winberg, R. Sandberg, Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods*. **10**, 1096–1098 (2013).

45. R. V. Grindberg, J. L. Yee-Greenbaum, M. J. McConnell, M. Novotny, A. L. O'Shaughnessy, G. M. Lambert, M. J. Araúzo-Bravo, J. Lee, M. Fishman, G. E. Robbins,

X. Lin, P. Venepally, J. H. Badger, D. W. Galbraith, F. H. Gage, R. S. Lasken, RNA-sequencing from single nuclei. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19802–19807 (2013).

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

47. A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, M. W. Kirschner, Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*. **161**, 1187–1201 (2015).

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

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

50. A. Adey, J. O. Kitzman, J. N. Burton, R. Daza, A. Kumar, L. Christiansen, M. Ronaghi, S. Amini, K. L. Gunderson, F. J. Steemers, J. Shendure, In vitro, long-range sequence information for de novo genome assembly via transposase contiguity. *Genome Res.* **24**, 2041–2049 (2014).

51. S. Amini, D. Pushkarev, L. Christiansen, E. Kostem, T. Royce, C. Turk, N. Pignatelli, A. Adey, J. O. Kitzman, K. Vijayan, M. Ronaghi, J. Shendure, K. L. Gunderson, F. J. Steemers, Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343–1349 (2014).

52. D. A. Cusanovich, R. Daza, A. Adey, H. A. Pliner, L. Christiansen, K. L. Gunderson, F. J. Steemers, C. Trapnell, J. Shendure, Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*. **348**, 910–914 (2015).

53. S. A. Vitak, K. A. Torkenczy, J. L. Rosenkrantz, A. J. Fields, L. Christiansen, M. H. Wong, L. Carbone, F. J. Steemers, A. Adey, Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods*. **14**, 302–308 (2017).

54. V. Ramani, X. Deng, R. Qiu, K. L. Gunderson, F. J. Steemers, C. M. Disteche, W. S. Noble, Z. Duan, J. Shendure, Massively multiplex single-cell Hi-C. *Nat. Methods*. **14**, 263–266 (2017).

55. R. M. Mulqueen, D. Pokholok, S. Norberg, A. J. Fields, D. Sun, K. A. Torkenczy, J. Shendure, C. Trapnell, B. J. O'Roak, Z. Xia, F. J. Steemers, A. C. Adey, Scalable and efficient single-cell DNA methylation sequencing by combinatorial indexing (2017), , doi:10.1101/157230.

56. Supplemental online materials.

57. R. V. Grindberg, J. L. Yee-Greenbaum, M. J. McConnell, M. Novotny, A. L.

O'Shaughnessy, G. M. Lambert, M. J. Araúzo-Bravo, J. Lee, M. Fishman, G. E. Robbins, X. Lin, P. Venepally, J. H. Badger, D. W. Galbraith, F. H. Gage, R. S. Lasken, RNA-sequencing from single nuclei. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 19802–19807 (2013).

58. J. Gertz, K. E. Varley, N. S. Davis, B. J. Baas, I. Y. Goryshin, R. Vaidyanathan, S. Kuersten, R. M. Myers, Transposase mediated construction of RNA-seq libraries. *Genome Res.* **22**, 134–141 (2012).

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

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

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

62. R. Ruksana, K. Kuroda, H. Terami, T. Bando, S. Kitaoka, T. Takaya, Y. Sakube, H. Kagawa, Tissue expression of four troponin I genes and their molecular interactions with two troponin C isoforms in Caenorhabditis elegans. *Genes Cells*. **10**, 261–276 (2005).

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

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

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

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

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

68. C. L. Araya, T. Kawli, A. Kundaje, L. Jiang, B. Wu, D. Vafeados, R. Terrell, P. Weissdepp, L. Gevirtzman, D. Mace, W. Niu, A. P. Boyle, D. Xie, L. Ma, J. I. Murray, V. Reinke, R. H. Waterston, M. Snyder, Corrigendum: Regulatory analysis of the C. elegans genome with spatiotemporal resolution. *Nature*. **528**, 152 (2015).

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

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

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

72. B. D. Harfe, A. Vaz Gomes, C. Kenyon, J. Liu, M. Krause, A. Fire, Analysis of a Caenorhabditis elegans Twist homolog identifies conserved and divergent aspects of mesodermal patterning. *Genes Dev.* **12**, 2623–2635 (1998).

73. M. Horn, C. Geisen, L. Cermak, B. Becker, S. Nakamura, C. Klein, M. Pagano, A. Antebi, DRE-1/FBXO11-dependent degradation of BLMP-1/BLIMP-1 governs C. elegans developmental timing and maturation. *Dev. Cell*. **28**, 697–710 (2014).

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

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

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

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

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

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

80. M. Kudron, W. Niu, Z. Lu, G. Wang, M. Gerstein, M. Snyder, V. Reinke, Tissue-specific direct targets of Caenorhabditis elegans Rb/E2F dictate distinct somatic and germline programs. *Genome Biol.* **14**, R5 (2013).

81. M. M. Kudron, A. Victorsen, L. Gevirtzman, L. W. Hillier, W. W. Fisher, D. Vafeados, M. Kirkey, A. S. Hammonds, J. Gersch, H. Ammouri, M. L. Wall, J. Moran, D. Steffen, M. Szynkarek, S. Seabrook-Sturgis, N. Jameel, M. Kadaba, J. Patton, R. Terrell, M. Corson, T. J. Durham, S. Park, S. Samanta, M. Han, J. Xu, K.-K. Yan, S. E. Celniker, K. P. White, L. Ma, M. Gerstein, V. Reinke, R. H. Waterston, The ModERN Resource: Genome-Wide Binding Profiles for Hundreds of Drosophila and Caenorhabditis elegans Transcription Factors. *Genetics*. **208**, 937–949 (2018).

82. P.-Y. Tung, J. D. Blischak, C. J. Hsiao, D. A. Knowles, J. E. Burnett, J. K. Pritchard, Y. Gilad, Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.* **7**, 39921 (2017).

83. A. McKenna, G. M. Findlay, J. A. Gagnon, M. S. Horwitz, A. F. Schier, J. Shendure, Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*. **353**, aaf7907 (2016).

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

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

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

87. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).

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

89. A. Adey, J. N. Burton, J. O. Kitzman, J. B. Hiatt, A. P. Lewis, B. K. Martin, R. Qiu, C. Lee, J. Shendure, The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature*. **500**, 207–211 (2013).

90. X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. Pliner, C. Trapnell, Reversed graph embedding resolves complex single-cell developmental trajectories. *bioRxiv* (2017), p. 110668.

91. N. Habib, Y. Li, M. Heidenreich, L. Swiech, I. Avraham-Davidi, J. J. Trombetta, C. Hession, F. Zhang, A. Regev, Div-Seq: Single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science*. **353**, 925–928 (2016).

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

93. M. B. Gerstein, Z. J. Lu, E. L. Van Nostrand, C. Cheng, B. I. Arshinoff, T. Liu, K. Y. Yip, R. Robilotto, A. Rechtsteiner, K. Ikegami, P. Alves, A. Chateigner, M. Perry, M. Morris, R. K. Auerbach, X. Feng, J. Leng, A. Vielle, W. Niu, K. Rhrissorrakrai, A. Agarwal, R. P. Alexander, G. Barber, C. M. Brdlik, J. Brennan, J. J. Brouillet, A. Carr, M.-S. Cheung, H. Clawson, S. Contrino, L. O. Dannenberg, A. F. Dernburg, A. Desai, L. Dick, A. C. Dosé, J. Du, T. Egelhofer, S. Ercan, G. Euskirchen, B. Ewing, E. A. Feingold, R. Gassmann, P. J. Good, P. Green, F. Gullier, M. Gutwein, M. S. Guyer, L. Habegger, T. Han, J. G. Henikoff,

S. R. Henz, A. Hinrichs, H. Holster, T. Hyman, A. L. Iniguez, J. Janette, M. Jensen, M. Kato, W. J. Kent, E. Kephart, V. Khivansara, E. Khurana, J. K. Kim, P. Kolasinska-Zwierz, E. C. Lai, I. Latorre, A. Leahey, S. Lewis, P. Lloyd, L. Lochovsky, R. F. Lowdon, Y. Lubling, R. Lyne, M. MacCoss, S. D. Mackowiak, M. Mangone, S. McKay, D. Mecenas, G. Merrihew, D. M. Miller 3rd, A. Muroyama, J. I. Murray, S.-L. Ooi, H. Pham, T. Phippen, E. A. Preston, N. Rajewsky, G. Rätsch, H. Rosenbaum, J. Rozowsky, K. Rutherford, P. Ruzanov, M. Sarov, R. Sasidharan, A. Sboner, P. Scheid, E. Segal, H. Shin, C. Shou, F. J. Slack, C. Slightam, R. Smith, W. C. Spencer, E. O. Stinson, S. Taing, T. Takasaki, D. Vafeados, K. Voronina, G. Wang, N. L. Washington, C. M. Whittle, B. Wu, K.-K. Yan, G. Zeller, Z. Zha, M. Zhong, X. Zhou, modENCODE Consortium, J. Ahringer, S. Strome, K. C. Gunsalus, G. Micklem, X. S. Liu, V. Reinke, S. K. Kim, L. W. Hillier, S. Henikoff, F. Piano, M. Snyder, L. Stein, J. D. Lieb, R. H. Waterston, Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*. **330**, 1775–1787 (2010).

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

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

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

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

98. M. Harterink, D. H. Kim, T. C. Middelkoop, T. D. Doan, A. van Oudenaarden, H. C. Korswagen, 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).

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

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

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

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

103. T. M. Tabuchi, B. Deplancke, N. Osato, L. J. Zhu, M. I. Barrasa, M. M. Harrison, H. R.

Horvitz, A. J. M. Walhout, K. A. Hagstrom, Chromosome-biased binding and gene regulation by the Caenorhabditis elegans DRM complex. *PLoS Genet.* **7**, e1002074 (2011).

104. I. Latorre, M. A. Chesney, J. M. Garrigues, P. Stempor, A. Appert, M. Francesconi, S. Strome, J. Ahringer, The DREAM complex promotes gene body H2A.Z for target repression. *Genes Dev.* **29**, 495–500 (2015).

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

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

107. L. Tilleman, S. De Henau, M. Pauwels, N. Nagy, I. Pintelon, B. P. Braeckman, K. De Wael, S. Van Doorslaer, D. Adriaensen, J.-P. Timmermans, L. Moens, S. Dewilde, An N-myristoylated globin with a redox-sensing function that regulates the defecation cycle in Caenorhabditis elegans. *PLoS One*. **7**, e48768 (2012).

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

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

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

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

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

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

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

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

116. L. Hong, T. Elbl, J. Ward, C. Franzini-Armstrong, K. K. Rybicka, B. K. Gatewood, D. L. Baillie, E. A. Bucher, MUP-4 is a novel transmembrane protein with functions in epithelial

cell adhesion in Caenorhabditis elegans. *J. Cell Biol.* **154**, 403–414 (2001).

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

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

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

120. D. Sieburth, Q. Ch'ng, M. Dybbs, M. Tavazoie, S. Kennedy, D. Wang, D. Dupuy, J.-F. Rual, D. E. Hill, M. Vidal, G. Ruvkun, J. M. Kaplan, Systematic analysis of genes required for synapse structure and function. *Nature*. **436**, 510–517 (2005).

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

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

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

124. A. Yoshida, S. Nakano, T. Suzuki, K. Ihara, T. Higashiyama, I. Mori, A glial K+/Cl- cotransporter modifies temperature-evoked dynamics in Caenorhabditis elegans sensory neurons. *Genes Brain Behav.* (2015) (available at http://onlinelibrary.wiley.com/doi/10.1111/gbb.12260/pdf).

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

126. M. E. Gruidl, P. A. Smith, K. A. Kuznicki, J. S. McCrone, J. Kirchner, D. L. Roussell, S. Strome, K. L. Bennett, Multiple potential germ-line helicases are components of the germ-line-specific P granules of Caenorhabditis elegans. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 13837–13842 (1996).

127. I. Kawasaki, A. Amiri, Y. Fan, N. Meyer, S. Dunkelbarger, T. Motohashi, T. Karashima, O. Bossinger, S. Strome, The PGL family proteins associate with germ granules and function redundantly in Caenorhabditis elegans germline development. *Genetics*. **167**,

645–661 (2004).

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

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

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

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

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

133. B. D. Ackley, J. R. Crew, H. Elamaa, T. Pihlajaniemi, C. J. Kuo, J. M. Kramer, The NC1/endostatin domain of Caenorhabditis elegans type XVIII collagen affects cell migration and axon guidance. *J. Cell Biol.* **152**, 1219–1232 (2001).

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

135. H. Komatsu, M. Y. Chao, J. Larkins-Ford, M. E. Corkins, G. A. Somers, T. Tucey, H. M. Dionne, J. Q. White, K. Wani, M. Boxem, A. C. Hart, OSM-11 facilitates LIN-12 Notch signaling during Caenorhabditis elegans vulval development. *PLoS Biol.* **6**, e196 (2008).

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

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

138. J. D. McGhee, M. C. Sleumer, M. Bilenky, K. Wong, S. J. McKay, B. Goszczynski, H. Tian, N. D. Krich, J. Khattra, R. A. Holt, D. L. Baillie, Y. Kohara, M. A. Marra, S. J. M. Jones, D. G. Moerman, A. G. Robertson, The ELT-2 GATA-factor and the global regulation of transcription in the C. elegans intestine. *Dev. Biol.* **302**, 627–645 (2007).

139. A. Patton, S. Knuth, B. Schaheen, H. Dang, I. Greenwald, H. Fares, Endocytosis function of a ligand-gated ion channel homolog in Caenorhabditis elegans. *Curr. Biol.* **15**, 1045–1050 (2005).

140. Y. Zhang, C. Ma, T. Delohery, B. Nasipak, B. C. Foat, A. Bounoutas, H. J. Bussemaker, S. K. Kim, M. Chalfie, Identification of genes expressed in C. elegans touch receptor neurons. *Nature*. **418**, 331–335 (2002).

141. T. Kawano, H. Zheng, D. C. Merz, Y. Kohara, K. K. Tamai, K. Nishiwaki, J. G. Culotti, C. elegans mig-6 encodes papilin isoforms that affect distinct aspects of DTC migration, and interacts genetically with mig-17 and collagen IV. *Development*. **136**, 1433–1442 (2009).

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

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

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

145. T. Janssen, E. Meelkop, M. Lindemans, K. Verstraelen, S. J. Husson, L. Temmerman, R. J. Nachman, L. Schoofs, Discovery of a cholecystokinin-gastrin-like signaling system in nematodes. *Endocrinology*. **149**, 2826–2839 (2008).

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

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

148. R. Nass, M. K. Hahn, T. Jessen, P. W. McDonald, L. Carvelli, R. D. Blakely, A genetic screen in Caenorhabditis elegans for dopamine neuron insensitivity to 6-hydroxydopamine identifies dopamine transporter mutants impacting transporter biosynthesis and trafficking. *J. Neurochem.* **94**, 774–785 (2005).

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

150. A. Oishi, K. Gengyo-Ando, S. Mitani, A. Mohri-Shiomi, K. D. Kimura, T. Ishihara, I. Katsura, FLR-2, the glycoprotein hormone alpha subunit, is involved in the neural control of intestinal functions in Caenorhabditis elegans. *Genes Cells*. **14**, 1141–1154 (2009).

151. R. J. Hobson, V. M. Hapiak, H. Xiao, K. L. Buehrer, P. R. Komuniecki, R. W. Komuniecki, SER-7, a Caenorhabditis elegans 5-HT7-like receptor, is essential for the 5-HT stimulation of pharyngeal pumping and egg laying. *Genetics*. **172**, 159–169 (2006).

152. M. Furuya, H. Qadota, A. D. Chisholm, A. Sugimoto, The C. elegans eyes absent ortholog EYA-1 is required for tissue differentiation and plays partially redundant roles with

PAX-6. *Dev. Biol.* **286**, 452–463 (2005).

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

154. C. O. Ortiz, J. F. Etchberger, S. L. Posy, C. Frøkjaer-Jensen, S. Lockery, B. Honig, O. Hobert, Searching for neuronal left/right asymmetry: genomewide analysis of nematode receptor-type guanylyl cyclases. *Genetics*. **173**, 131–149 (2006).

155. M. Lindemans, T. Janssen, S. J. Husson, E. Meelkop, L. Temmerman, E. Clynen, I. Mertens, L. Schoofs, A neuromedin-pyrokinin-like neuropeptide signaling system in Caenorhabditis elegans. *Biochem. Biophys. Res. Commun.* **379**, 760–764 (2009).

156. H. Inada, H. Ito, J. Satterlee, P. Sengupta, K. Matsumoto, I. Mori, Identification of guanylyl cyclases that function in thermosensory neurons of Caenorhabditis elegans. *Genetics*. **172**, 2239–2252 (2006).

157. K. Yamada, T. Hirotsu, M. Matsuki, R. A. Butcher, M. Tomioka, T. Ishihara, J. Clardy, H. Kunitomo, Y. Iino, Olfactory plasticity is regulated by pheromonal signaling in Caenorhabditis elegans. *Science*. **329**, 1647–1650 (2010).

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

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

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

161. D. A. Birnby, E. M. Link, J. J. Vowels, H. Tian, P. L. Colacurcio, J. H. Thomas, A transmembrane guanylyl cyclase (DAF-11) and Hsp90 (DAF-21) regulate a common set of chemosensory behaviors in caenorhabditis elegans. *Genetics*. **155**, 85–104 (2000).

162. J. F. Etchberger, A. Lorch, M. C. Sleumer, R. Zapf, S. J. Jones, M. A. Marra, R. A. Holt, D. G. Moerman, O. Hobert, The molecular signature and cis-regulatory architecture of a C. elegans gustatory neuron. *Genes Dev.* **21**, 1653–1674 (2007).

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

164. J. M. Gray, D. S. Karow, H. Lu, A. J. Chang, J. S. Chang, R. E. Ellis, M. A. Marletta, C. I. Bargmann, Oxygen sensation and social feeding mediated by a C. elegans guanylate

cyclase homologue. *Nature*. **430**, 317–322 (2004).

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

166. X. Wang, W. Zhang, T. Cheever, V. Schwarz, K. Opperman, H. Hutter, D. Koepp, L. Chen, The C. elegans L1CAM homologue LAD-2 functions as a coreceptor in MAB-20/Sema2–mediated axon guidance. *J. Cell Biol.* **180**, 233–246 (2008).

167. X. Qiu, A. Hill, J. Packer, D. Lin, Y.-A. Ma, C. Trapnell, Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods*. **14**, 309–315 (2017).

168. C. T. Fincher, O. Wurtzel, T. de Hoog, K. M. Kravarik, P. W. Reddien, Cell type transcriptome atlas for the planarian Schmidtea mediterranea. *Science* (2018), doi:10.1126/science.aaq1736.

169. M. Plass, J. Solana, F. A. Wolf, S. Ayoub, A. Misios, P. Glažar, B. Obermayer, F. J. Theis, C. Kocks, N. Rajewsky, Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*. **360** (2018), doi:10.1126/science.aaq1723.

170. Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, Principal investigators, Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*. **562**, 367–372 (2018).

171. A. Zeisel, H. Hochgerner, P. Lönnerberg, A. Johnsson, F. Memic, J. van der Zwan, M. Häring, E. Braun, L. E. Borm, G. La Manno, S. Codeluppi, A. Furlan, K. Lee, N. Skene, K. D. Harris, J. Hjerling-Leffler, E. Arenas, P. Ernfors, U. Marklund, S. Linnarsson, Molecular Architecture of the Mouse Nervous System. *Cell*. **174**, 999–1014.e22 (2018).

172. A. Sebé-Pedrós, B. Saudemont, E. Chomsky, F. Plessier, M.-P. Mailhé, J. Renno, Y. Loe-Mie, A. Lifshitz, Z. Mukamel, S. Schmutz, S. Novault, P. R. H. Steinmetz, F. Spitz, A. Tanay, H. Marlow, Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq. *Cell*. **173**, 1520–1534.e20 (2018).

173. N. Karaiskos, P. Wahle, J. Alles, A. Boltengagen, S. Ayoub, C. Kipar, C. Kocks, N. Rajewsky, R. P. Zinzen, The Drosophila embryo at single-cell transcriptome resolution. *Science*, eaan3235 (2017).

174. J. A. Briggs, C. Weinreb, D. E. Wagner, S. Megason, L. Peshkin, M. W. Kirschner, A. M. Klein, The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* (2018), doi:10.1126/science.aar5780.

175. D. E. Wagner, C. Weinreb, Z. M. Collins, J. A. Briggs, S. G. Megason, A. M. Klein, Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo.

Science (2018), doi:10.1126/science.aar4362.

176. J. A. Farrell, Y. Wang, S. J. Riesenfeld, K. Shekhar, A. Regev, A. F. Schier, Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* (2018), doi:10.1126/science.aar3131.

177. B. Pijuan-Sala, J. A. Griffiths, C. Guibentif, T. W. Hiscock, W. Jawaid, F. J. Calero-Nieto, C. Mulas, X. Ibarra-Soria, R. C. V. Tyser, D. L. L. Ho, W. Reik, S. Srinivas, B. D. Simons, J. Nichols, J. C. Marioni, B. Göttgens, A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* (2019), doi:10.1038/s41586-019-0933-9.

178. M. Sarov, J. I. Murray, K. Schanze, A. Pozniakovski, W. Niu, K. Angermann, S. Hasse, M. Rupprecht, E. Vinis, M. Tinney, E. Preston, A. Zinke, S. Enst, T. Teichgraber, J. Janette, K. Reis, S. Janosch, S. Schloissnig, R. K. Ejsmont, C. Slightam, X. Xu, S. K. Kim, V. Reinke, A. F. Stewart, M. Snyder, R. H. Waterston, A. A. Hyman, A genome-scale resource for in vivo tag-based protein function exploration in C. elegans. *Cell*. **150**, 855–866 (2012).

179. X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, C. Trapnell, Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods*. **14**, 979–982 (2017).

180. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018), (available at http://arxiv.org/abs/1802.03426).

181. E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, E. W. Newell, Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* (2018), doi:10.1038/nbt.4314.

182. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *arXiv [physics.soc-ph]* (2008), (available at http://arxiv.org/abs/0803.0476).

183. C. L. Araya, T. Kawli, A. Kundaje, L. Jiang, B. Wu, D. Vafeados, R. Terrell, P. Weissdepp, L. Gevirtzman, D. Mace, W. Niu, A. P. Boyle, D. Xie, L. Ma, J. I. Murray, V. Reinke, R. H. Waterston, M. Snyder, Regulatory analysis of the C. elegans genome with spatiotemporal resolution. *Nature*. **512**, 400–405 (2014).

184. G. Broitman-Maduro, M. Owraghi, W. W. K. Hung, S. Kuntz, P. W. Sternberg, M. F. Maduro, 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).

185. 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. *Dev. Biol.* **374**, 12–23 (2013).

186. M. Hu, D. Krause, M. Greaves, S. Sharkis, M. Dexter, C. Heyworth, T. Enver,

Multilineage gene expression precedes commitment in the hemopoietic system. *Genes Dev.* **11**, 774–785 (1997).

187. P. Laslo, C. J. Spooner, A. Warmflash, D. W. Lancki, H.-J. Lee, R. Sciammas, B. N. Gantner, A. R. Dinner, H. Singh, Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell*. **126**, 755–766 (2006).

188. M. Thomson, S. J. Liu, L.-N. Zou, Z. Smith, A. Meissner, S. Ramanathan, Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. *Cell*. **145**, 875–889 (2011).

189. E. W. Brunskill, J.-S. Park, E. Chung, F. Chen, B. Magella, S. S. Potter, Single cell dissection of early kidney development: multilineage priming. *Development*. **141**, 3093–3101 (2014).

190. W. Wang, X. Niu, T. Stuart, E. Jullian, W. M. Mauck, R. G. Kelly, R. Satija, L. Christiaen, A single-cell transcriptional roadmap for cardiopharyngeal fate diversification. *Nature Cell Biology*. **21** (2019), pp. 674–686.

191. S. J. Husson, T. Janssen, G. Baggerman, B. Bogert, A. H. Kahn-Kirby, K. Ashrafi, L. Schoofs, Impaired processing of FLP and NLP peptides in carboxypeptidase E (EGL-21)-deficient Caenorhabditis elegans as analyzed by mass spectrometry. *J. Neurochem.* **102**, 246–260 (2007).

192. T. R. Sarafi-Reinach, P. Sengupta, The forkhead domain gene unc-130 generates chemosensory neuron diversity in C. elegans. *Genes Dev.* **14**, 2472–2485 (2000).

193. O. Hobert, A map of terminal regulators of neuronal identity in Caenorhabditis elegans. *Wiley Interdiscip. Rev. Dev. Biol.* **5**, 474–498 (2016).

194. 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).

195. 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).

196. 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).

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

198. K. Brunschwig, C. Wittmann, R. Schnabel, T. R. Bürglin, H. Tobler, F. Müller, Anterior organization of the Caenorhabditis elegans embryo by the labial-like Hox gene ceh-13.

Development. **126**, 1537–1546 (1999).

199. 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. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 15479–15484 (2010).

200. L. Kester, A. van Oudenaarden, Single-Cell Transcriptomics Meets Lineage Tracing. *Cell Stem Cell*. **23**, 166–179 (2018).

201. J. Packer, C. Trapnell, Single-Cell Multi-omics: An Engine for New Quantitative Models of Gene Regulation. *Trends Genet.* **34**, 653–665 (2018).

202. M. D. Young, S. Behjati, SoupX removes ambient RNA contamination from droplet based single cell RNA sequencing data. *bioRxiv* (2018), p. 303727.

203. 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).

204. J. Bruin, FAQ. What are pseudo-R-squareds (2006), (available at https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/).

205. G. Deltas, The Small-Sample Bias of the Gini Coefficient: Results and Implications for Empirical Research. *Rev. Econ. Stat.* **85**, 226–234 (2003).

206. R. Kolde, Pheatmap: pretty heatmaps. *R package version*. **61**, 915 (2012).

207. S. Aibar, C. B. González-Blas, T. Moerman, V. A. Huynh-Thu, H. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, J. Aerts, J. van den Oord, Z. K. Atak, J. Wouters, S. Aerts, SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods*. **14**, 1083–1086 (2017).

208. 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. *Dev. Biol.* **347**, 154–166 (2010).

209. M. T. Weirauch, A. Yang, M. Albu, A. G. Cote, A. Montenegro-Montero, P. Drewe, H. S. Najafabadi, S. A. Lambert, I. Mann, K. Cook, H. Zheng, A. Goity, H. van Bakel, J.-C. Lozano, M. Galli, M. G. Lewsey, E. Huang, T. Mukherjee, X. Chen, J. S. Reece-Hoyes, S. Govindarajan, G. Shaulsky, A. J. M. Walhout, F.-Y. Bouget, G. Ratsch, L. F. Larrondo, J. R. Ecker, T. R. Hughes, Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*. **158**, 1431–1443 (2014).

210. Q. Zhu, J. I. Murray, K. Tan, J. Kim, *qinzhu/VisCello.celegans: VisCello.celegans v1.1.0 release* (2019; https://zenodo.org/record/3262315).

211. Q. Zhu, J. I. Murray, K. Tan, J. Kim, *qinzhu/VisCello: VisCello v1.0.0* (2019;

https://zenodo.org/record/3262313).

212. S. R. Wicks, R. T. Yeh, W. R. Gish, R. H. Waterston, R. H. Plasterk, Rapid gene mapping in Caenorhabditis elegans using a high density polymorphism map. *Nat. Genet.* **28**, 160–164 (2001).

213. J. S. Reece-Hoyes, B. Deplancke, J. Shingles, C. A. Grove, I. A. Hope, A. J. M. Walhout, A compendium of Caenorhabditis elegans regulatory transcription factors: a resource for mapping transcription regulatory networks. *Genome Biol.* **6**, R110 (2005).

214. P. J. Skene, S. Henikoff, An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife Sciences*. **6**, e21856 (2017).

215. O. Uchida, H. Nakano, M. Koga, Y. Ohshima, The C. elegans che-1 gene encodes a zinc finger transcription factor required for specification of the ASE chemosensory neurons. *Development*. **130**, 1215–1224 (2003).

216. A. S. Wenick, O. Hobert, Genomic cis-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in C. elegans. *Dev. Cell*. **6**, 757–770 (2004).

217. K. Good, R. Ciosk, J. Nance, A. Neves, R. J. Hill, J. R. Priess, The T-box transcription factors TBX-37 and TBX-38 link GLP-1/Notch signaling to mesoderm induction in C. elegans embryos. *Development*. **131**, 1967–1978 (2004).

218. M. F. Maduro, R. J. Hill, P. J. Heid, E. D. Newman-Smith, J. Zhu, J. R. Priess, J. H. Rothman, Genetic redundancy in endoderm specification within the genus Caenorhabditis. *Dev. Biol.* **284**, 509–522 (2005).

219. T. Fukushige, M. Krause, The myogenic potency of HLH-1 reveals wide-spread developmental plasticity in early C. elegans embryos. *Development*. **132**, 1795–1805 (2005).

220. B. Tursun, T. Patel, P. Kratsios, O. Hobert, Direct conversion of C. elegans germ cells into specific neuron types. *Science*. **331**, 304–308 (2011).

221. O. Hobert, Terminal Selectors of Neuronal Identity. *Curr. Top. Dev. Biol.* **116**, 455–475 (2016).

222. V. Agarwal, J. Shendure, Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *bioRxiv* (2018), p. 416685.

223. A. Coghlan, K. H. Wolfe, Fourfold faster rate of genome rearrangement in nematodes than in Drosophila. *Genome Res.* **12**, 857–867 (2002).

224. Z. Zhao, T. J. Boyle, Z. Bao, J. I. Murray, B. Mericle, R. H. Waterston, Comparative analysis of embryonic cell lineage between Caenorhabditis briggsae and Caenorhabditis

elegans. *Dev. Biol.* **314**, 93–99 (2008).

225. D. Yin, E. M. Schwarz, C. G. Thomas, R. L. Felde, I. F. Korf, A. D. Cutter, C. M. Schartner, E. J. Ralston, B. J. Meyer, E. S. Haag, Rapid genome shrinkage in a self-fertile nematode reveals sperm competition proteins. *Science*. **359**, 55–61 (2018).

226. S. Zhong, S. Zhang, X. Fan, Q. Wu, L. Yan, J. Dong, H. Zhang, L. Li, L. Sun, N. Pan, X. Xu, F. Tang, J. Zhang, J. Qiao, X. Wang, A single-cell RNA-seq survey of the developmental landscape of the human prefrontal cortex. *Nature*. **555**, 524–528 (2018).

227. T. J. Nowakowski, A. Bhaduri, A. A. Pollen, B. Alvarado, M. A. Mostajo-Radji, E. Di Lullo, M. Haeussler, C. Sandoval-Espinosa, S. J. Liu, D. Velmeshev, J. R. Ounadjela, J. Shuga, X. Wang, D. A. Lim, J. A. West, A. A. Leyrat, W. J. Kent, A. R. Kriegstein, Spatiotemporal gene expression trajectories reveal developmental hierarchies of the human cortex. *Science*. **358**, 1318–1323 (2017).

228. J. D. Buenrostro, M. R. Corces, C. A. Lareau, B. Wu, A. N. Schep, M. J. Aryee, R. Majeti, H. Y. Chang, W. J. Greenleaf, Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* (2018), doi:10.1016/j.cell.2018.03.074.

229. D. R. Farnsworth, L. Saunders, A. C. Miller, A Single-Cell Transcriptome Atlas for Zebrafish Development, , doi:10.1101/738344.

230. B. Raj, D. E. Wagner, A. McKenna, S. Pandey, A. M. Klein, J. Shendure, J. A. Gagnon, A. F. Schier, Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat. Biotechnol.* **36**, 442–450 (2018).

231. B. Spanjaard, B. Hu, N. Mitic, P. Olivares-Chauvet, S. Janjuha, N. Ninov, J. P. Junker, Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat. Biotechnol.* **36**, 469–473 (2018).

232. K. L. Frieda, J. M. Linton, S. Hormoz, J. Choi, K.-H. K. Chow, Z. S. Singer, M. W. Budde, M. B. Elowitz, L. Cai, Synthetic recording and in situ readout of lineage information in single cells. *Nature*. **541**, 107–111 (2017).

233. 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).

234. S. Shah, E. Lubeck, W. Zhou, L. Cai, seqFISH Accurately Detects Transcripts in Single Cells and Reveals Robust Spatial Organization in the Hippocampus. *Neuron*. **94**, 752–758.e1 (2017).

235. Z. F. Altun, D. H. Hall, Handbook of C. elegans Anatomy. *WormAtlas. http://www.wormatlas.org/hermaphrodite/hermaphroditehomepage.htm* (2019).

236. S. Ward, J. S. Carrel, Fertilization and sperm competition in the nematodeCaenorhabditis elegans. *Dev. Biol.* **73**, 304–321 (1979).

237. J. Hodgkin, T. Doniach, Natural variation and copulatory plug formation in Caenorhabditis elegans. *Genetics*. **146**, 149–164 (1997).

238. L. Rose, P. Gönczy, Polarity establishment, asymmetric division and segregation of fate determinants in early C. elegans embryos. *WormBook*, 1–43 (2014).

239. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411 (2018).

240. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck 3rd, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive Integration of Single-Cell Data. *Cell*. **177**, 1888–1902.e21 (2019).

241. I. T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments. *Philos. Trans. A Math. Phys. Eng. Sci.* **374**, 20150202 (2016).

242. L. van der Maaten, G. Hinton, Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

243. H. A. Pliner, J. Shendure, C. Trapnell, Supervised classification enables rapid annotation of cell atlases. *bioRxiv* (2019), p. 538652.

244. S. L. Wolock, R. Lopez, A. M. Klein, Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell systems* (2019) (available at https://www.sciencedirect.com/science/article/pii/S2405471218304745).

Vita

Jonathan Packer was born in 1994 and has lived a mostly uneventful life. He grew up in Westchester, New York and attended Hackley School. He received a B.S. in Operations Research from Columbia University in 2015. By the time anyone reads this paragraph, he will hopefully have received his PhD in Genome Sciences from the University of Washington. As of 2019, he lives in Cambridge, Massachusetts and works for Foresite Labs, a subsidiary of Foresite Capital that does translational research for drug target discovery.