Comprehensive single-cell transcriptional profiling of the regenerative planarian *Schmidtea mediterranea*

by

Christopher T. Fincher

B.A. Honors, Biology (2012)
University of Pennsylvania, Philadelphia, PA

Submitted to the Department of Biology
In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2020

Signature of Author: _____
Department of Biology
July 22nd, 2020

Certified by: _____
Peter W. Reddien
Professor of Biology
Thesis Supervisor

Accepted by: _____
Mary Gehring
Associate Professor of Biology
Co-Director, Biology Graduate Committee

Comprehensive single-cell transcriptional profiling of the regenerative planarian *Schmidtea mediterranea*

By

Christopher T. Fincher

Submitted to the Department of Biology on July 22, 2020 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Biology

Abstract

Animals can contain hundreds of cell types, each of which has a distinct morphology and function. The transcriptome of a cell dictates this unique cell biology. Recent approaches for high throughput single-cell RNA sequencing have made it possible to generate transcriptomes easily and affordably for tens of thousands of single cells, raising the possibility that transcriptomes could be generated for all cell types and cell states in a complete animal. Planarians are freshwater flatworms renowned for their capacity for whole-body regeneration. They possess a complex body plan with multiple distinct tissues. They also possess a population of dividing cells, called neoblasts, which contain pluripotent stem cells and are the source of all new tissue, with all cell types being turned over throughout the life of the animal. Planarians also constitutively express an arrangement of regionally expressed genes in their muscle that serve as patterning information for the animal. As such, at a single time point in the adult, pluripotent stem cells, all differentiated cells, and all associated transition states from stem cell to differentiated cell can be recovered, including patterning information expressed in muscle. This makes planarians ideally suited to generating an atlas of transcriptomes for all cell types and cell states in a whole animal. We used the single-cell RNA sequencing technology Drop-seq to determine the transcriptomes for 66,783 cells from adult planarians. In doing so, we identified a number of known and novel cell populations, including a novel class of phagocytic cells. We also uncovered novel neoblast subpopulations and putative transition state populations between neoblasts and differentiated cells, as well as a number of genes with regional expression in muscle. Through the identification of known rare cell types in the data, we conclude that we have obtained near-to-complete cell type saturation for all cell types and cell states in the adult planarian. We now have full transcriptomes for each of these cell populations, which can be utilized to assay their roles in planarian biology. This approach can also be applied widely to diverse animal species, including those with limited molecular tools available.

Thesis Supervisor: Peter W. Reddien
Title: Professor of Biology

Acknowledgements

First and foremost, I would like to thank my advisor Peter Reddien for giving me the opportunity to work in your lab for my PhD work. Your scientific insight and rigor and your extensive knowledge of pretty much any topic imaginable has allowed me to grow immensely as a scientist over my time in the lab. You have also been a fantastic mentor, and your passion for science helped keep me excited about my research, even through the inevitable slumps.

I would like to thank my committee members, Omer Yilmaz, Iain Cheeseman, and Terry Orr-Weaver, for their advice and support throughout my graduate school journey. I would also like to thank Evan Macosko for taking the time to serve as the external member of my defense committee.

I would like to thank the many members of the Reddien lab that I've had the pleasure of working with over the years. You've been both amazing labmates and great friends these past years. From karaoke, to fishbowls at the Kong, to photobooths, to being robbed of best costume prizes, to woefully misguided participation in Movember, each of you have made my time in lab and in grad school a truly great experience. Also, some of my toughest times have come during grad school, and you guys were instrumental in getting me through them. I am forever grateful.

Through my time in grad school, I've had the pleasure of meeting a number of people here at MIT that have become great friends. In particular, I would like to thank Josh, Danny, Spencer, Santi, and John for helping me afford to live in Cambridge, but also my other MIT friends: Laurens, Grace, Nolan, Rohan, and everyone else. From trips to Salem, MA and other exotic locales, to ski trips and weddings, you guys have made my time in grad school absolutely amazing.

I would like to thank my boyfriend Doug. Thanks so much for all the great times we've had the past couple of years and for all of your support. You've helped me get through some truly tough times the past couple of years, and I'll never be able to thank you enough.

Finally, I wanted to thank my family, and in particular my mom and my grandparents. Your love and support and the sacrifices you've made over the years have been essential to getting me to where I am today. I truly would not have been able to accomplish this without you.

Table of Contents

Chapter 1: Introduction ……9

Foreword……..10
I. Tools for single-cell genomic analysis……..11
II. Planarians as a regenerative model system ……..42
III. Content overview……61
References……62

Chapter 2: Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*……...80

Abstract……81
Introduction……..82
Results & Figures……84
Discussion……..185
Materials and Methods……190
Table Captions ……..214
Acknowledgements……..……215
References……216

Chapter 3: Discussion ……223

I. Successful strategy for generating transcriptomes for all cell types of an animal……224
II. Identification of novel specialized neoblasts and transition states……225
III. Identification of novel differentiated cell types……229
VI. Identification of novel regionally expressed genes……232
V. Conclusion……233
References……234

Chapter 1

Introduction

Foreword

Multicellular animals can contain trillions of cells and hundreds of distinct cell types. Cell type composition changes dramatically during development from a fertilized egg to a mature adult organism. Distinct cell types in an animal are typically highly interconnected, with heterogenous cell types composing functionally distinct organs and organ systems. The actively transcribed set of genes within cells, the transcriptome, dictates the unique morphology and function of distinct cell types. Therefore, to understand the complex and interconnected biology underlying a multicellular organism, the actively transcribed genes for all cell types and cell states within an animal across its development must be determined. Recent advances in single-cell RNA sequencing technologies have made this daunting task a possibility. My thesis work focused on generating such a whole-animal cell type transcriptome atlas for the planarian *Schmidtea mediterranea*, a regenerative flatworm that is especially well suited as a case study for determining the transcriptomes of all cell types and cell states in a whole animal. The introductory chapter of this thesis will introduce the history of single-cell genomic technologies, including a summary of approaches that are currently available for the generation and analysis of single-cell genomic data, as well as an overview of how these methods are currently being used to transform diverse fields of biology. I will then introduce planarians as a model system, including the many features that make it especially well-suited for transcriptionally profiling all cells in a whole animal. My work aims to both provide insight into the cellular basis for the fascinating biology of this regenerative organism, as well as to provide a framework for generating such a whole-animal single-cell transcriptomic resource for diverse species across the animal kingdom.

I. Tools for single-cell genomic analysis

Genome sequencing projects

The complete genome sequence of the first free-living organism was released in 1995 for the bacterium *Haemophilus influenzae* (***1***). Over the next 3 years, completed genomes were released for the eukaryotic budding yeast *Saccharomyces cerevisiae* (***2***), the bacterium *Escherichia coli* K-12 (***3***), the archaeon *Methanococcus jannaschii* (***4***), and the bacterium *Mycobacterium tuberculosis* (***5***). In 1998, the complete genome sequence of a multicellular organism, the roundworm C*aenorhabditis elegans*, was completed (***6***), followed by the fruit fly *Drosophila melanogaster* in 2000 (***7***). In 2001, the first drafts of the human genome were completed (***8***, ***9***). Since that time, scores of animal genomes have been completed across the animal kingdom and individual human genomes can now be sequenced both quickly and affordably.

The complete sequencing of animal genomes has had a profound impact on biological research. The identification of genes and regulatory regions important for human health has been greatly accelerated, and we have identified many genes important for human health that are conserved across many animal species. By comparing genomes across the animal kingdom, our ability to phylogenetically group organisms, before largely limited to physiological traits, has transformed our understanding of evolutionary biology. With the ease and affordability of genome sequencing, we can now sequence the entire genomes of understudied animal, plant, microbial, and viral species, with wide ranging potential for discovering new biology and improving human health.

The advance of transcriptomic capabilities

Despite the profound impact of whole genome sequencing on biological research, animals can contain hundreds of distinct cell types, each of which contains essentially the same DNA content. It is the transcriptome, or the actively transcribed set of genes, that dictates the unique morphology and function of a cell. The capability to sequence actively transcribed genes from cells was first demonstrated in 1983, with the

sequencing of 178 clones from a rabbit muscle cDNA library using Sanger sequencing (***10***). More such cDNA sequences, later termed expressed sequence tags (ESTs), were generated by Sanger sequencing and compiled over the next decade (***11***). In 1995, two higher throughput methods for assessment of the sets of active genes from a biological sample were developed: serial analysis of gene expression (SAGE) and microarray technology (***12***, ***13***). SAGE involved the generation and concatenation of short sequence tags from a cDNA library (generated using oligo (dT) primers) into long constructs that could be cloned and sequenced, allowing for more multiplexed sequencing and the identification of thousands of transcripts from the pancreas (***12***). Microarray technology involved the printing of pre-defined complementary DNA sequences onto glass chips, allowing for the detection of corresponding gene expression using hybridization of fluorescently labeled cDNA libraries, and became the dominant method for high-throughput identification of actively transcribed genes for the next decade (***13***). In 2006, a more unbiased approach called RNA sequencing (RNA-seq) was developed, in which fragmented cDNA was amplified by polymerase chain reaction (PCR) using ligated adapter sequences, followed by annealing of sequencing primers and sequencing using Roche/454 technology, enabling detection of transcripts from 10,000 gene loci from cultured human pancreatic cells (***14***). A more high-throughput Illumina-based technology was developed in 2008 (***15***, ***16***, ***17***). RNA-seq has since become a standard laboratory tool, enabling full transcriptomes to be generated easily and affordably for tissues and populations of cells.

Single-cell RNA sequencing technologies

As with whole-genome sequencing, the ability to easily and affordably generate transcriptomes was transformational for biological research, allowing most genes actively expressed by a group of cells or tissues to be elucidated. Previously described RNA-sequencing strategies can only be performed on bulk populations of tissues or cells, however. As a result, transcriptomes generated by these methods are aggregates across multiple cell types and cell states, averaging out any heterogeneity in gene expression between transcriptionally distinct cell populations and drowning out rare

transcripts expressed by rare cell types in the sample. To overcome these limitations, a number of strategies were developed over the last decade that have allowed for the generation of transcriptomes from single cells.

Approaches for single-cell microarrays (***18***, ***19***) and single-cell RT-qPCR (***20***), which could be combined with microfluidic arrays, allowing for multiplexed measurements of hundreds of cells (***21***), were first developed in the 90's and were widely used to measure gene expression in single cells. However, these methods only enabled profiling of a limited set of pre-defined genes, not full transcriptomes. The first method for non-biased RNA sequencing of single cells, allowing for full transcriptome determination, was described in 2009 and is commonly referred to as the Tang method (***22***) (Figure 1.1A). This method involved isolating single cells into wells, where the cells were lysed and mRNA was reverse transcribed using an oligo-DT primer. A poly (A) tail was then added and the second DNA strand was generated using a poly (T) primer. The cDNA was amplified by polymerase chain reaction (PCR) using adapter sequences added during cDNA synthesis. The products were then sheared, sequencing adaptors with barcode sequences unique to each sample were ligated onto the fragments, and the libraries were pooled and sequenced. Although the method was a major breakthrough, allowing transcriptomes for single cells to be generated, the method had a fairly low sensitivity and a significant 3' bias.

In 2012, a method called SMART-seq was developed that improved on many of these issues (***23***) (Figure 1.1B). SMART-seq takes advantage of Moloney Murine Leukemia Virus (MMLV) reverse transcriptase, which adds extra C nucleotides to the 3' end of the cDNA product. In a strategy called template switching, a template switch oligo (TSO) containing an adapter sequence with added G ribonucleotides is added to the reverse transcription (RT) reaction, which also contains an oligo (dT) primer with the same adapter sequence. First strand cDNA is generated using the oligo (dT) primer, and the added G ribonucleotides of the TSO act as primers for a second round of RT, allowing both first and second strand cDNA to be generated in the same reaction (Figure 1.1B).

PCR amplification is then performed using a single primer to the adapter sequence. Barcoded sequencing libraries are generated using Nextera transposon-based tagmentation (Illumina, Inc.), in which transposons simultaneously fragment the DNA and add primers to the ends of the fragments. Illumina sequencing adapters are then added through PCR, and the libraries are pooled and sequenced. SMART-seq provided great performance increases over the Tang method and was further optimized over the next couple of years with the development of SMART-seq2, which is still widely used today (***24***, ***25***). Template switching was also adopted by many future single-cell sequencing technologies (Table 1.1).

A number of similar single-cell sequencing methods have been developed over the past decade that involve separating single cells into wells and barcoding and pooling individual samples just prior to sequencing. Quartz-seq, an approach similar to the Tang method, was released just prior to SMART-seq2 (***26***, ***27***) (Figure 1.1A). However, it was largely inferior in performance to SMART-seq2 and was not widely adopted. SUPeR-seq is another approach, also similar to the Tang method, that uses random primers for RT, allowing for single-cell sequencing of RNA species other than mRNA (***28***) (Figure 1.1A). MATQ-seq similarly allows for sequencing of RNA species other than mRNA by using both poly (dT) and internal primers for RT, followed by poly (C) tailing and second strand DNA synthesis using poly (G) primers (***29***) (Figure 1.1A). Unlike the previously described methods, MATQ-seq adds a unique barcode, called a UMI, to each RNA species in a sample during second strand cDNA synthesis, allowing for the identification of amplification artifacts. This innovation was originally demonstrated by another single-cell sequencing approach, MARS-seq, which is described below (***30***).

For each of the single-cell RNA sequencing technologies described thus far, individual samples are indistinguishable until the addition of barcoded sequencing adapters just prior to sequencing and cannot be pooled until this time. As such, these methods are generally more low throughput, are work intensive, and have a relatively high cost per cell, limiting the number of cells that can reasonably be sequenced. A number of

approaches were released that overcame this limitation by introducing unique barcodes to each sample early in the protocol. The earliest such approach was CEL-seq, in which cells are sorted into wells and RT is performed using a primer carrying a T7 promoter, an Illumina sequencing adapter, a sample-specific barcode sequence, and a poly (T) sequence (***31***) (Figure 1.1C). Second-strand DNA synthesis is then performed, followed by pooling of all samples. *in vitro* transcription (IVT) is then used for linear amplification, and the RNA is fragmented. RT is performed again, adding a second sequencing adapter, and the PCR amplified library is sequenced.

A number of similar tag-based strategies were developed in which single cells are originally separated into wells. One such approach, MARS-seq, is very similar to CEL-seq in its use of IVT amplification, but utilizes three separate tags: a molecular tag (UMI), a cellular tag, and a plate tag, providing even greater potential for multiplexing (***30***) (Figure 1.1C). As mentioned earlier, MARS-seq was the first approach to utilize a UMI strategy, which has been widely adopted and was utilized in all subsequent single-cell sequencing strategies described here, with the exception of STRT-seq. CEL-seq2 is one such strategy, which made a number of improvements to the CEL-seq protocol, including the addition of a UMI (***32***) (Figure 1.1C). Another strategy is STRT-seq, which uses a template switching mechanism for cDNA synthesis, similar to SMART-seq, but uses a barcoded TSO that is biotinylated at the 5’ end (***33***, ***34***). Streptavidin beads are then used to capture pooled cDNA, which is fragmented, end-repaired, and A-tailed, all on the streptavidin beads. Sequencing adapters are ligated, and the cDNA is PCR amplified before being sequenced. Unlike other methods, sequenced transcripts are heavily 5’ biased due to the streptavidin bead strategy. A modified STRT-seq protocol, STRT-seq/C1, has more recently simplified and adapted STRT-seq to the C1 Single-Cell Auto Prep system (Fluidigm) (***35***) (Figure 1.1D). It uses a poly (T) primer and template switching strategy for cDNA synthesis, with both the barcoded TSO and poly (T) primer biotinylated at the 5’ end, and with the poly (T) primer containing a PvuI restriction enzyme site. Library generation occurs through Nextera transposon-based tagmentation (Illumina, Inc.), and streptavidin beads are used to isolate both the 5’ and

3’ fragments, with a PvuI restriction digest removing all 3’ sequences. Unlike the earlier iteration, STRT-seq/C1 does not pool samples until after library generation. Finally, Quartz-seq2 is an optimized version of Quartz-seq that added cellular barcoding during RT, as well as UMI capability (***36***) (Figure 1.1A).

Whereas tag-based single-cell RNA sequencing methods allowed for earlier pooling of samples, making them generally more affordable and allowing for greater multiplexing capabilities, the methods described thus far are fairly low-throughput, labor intensive, and still too costly to generate transcriptomes for large numbers of cells. In 2015, two new technologies, Drop-seq and inDrop, were developed that utilized novel droplet-based approaches to allow for high throughput and affordable RNA-sequencing of thousands of single cells (***37***, ***38***). Drop-seq uses a microfluidics device to encapsulate single cells and single barcoded microbeads together within oil droplets (***37***) (Figure 1.1E). Barcoded microbeads contain a poly (T) capture sequence, a PCR priming sequence, and both a bead-specific barcode sequence that is common to each spot on the bead and a UMI barcode that differs across each spot on the bead. Only a small fraction of oil droplets contain both a cell and a bead, requiring a large starting material. Cells are lysed within the oil droplet and mRNA is captured onto the beads. The droplets are then broken, and the remaining downstream steps, from RT to sequencing, are largely identical to SMART-seq, but are performed in aggregate, greatly reducing per-cell reagent costs. InDrop also utilizes a microfluidics device to encapsulate single cells and barcoded primer sequences within oil droplets (***38***) (Figure 1.1C). However, inDrop utilizes deformable hydrogels that ensure almost all droplets contain barcoded primer sequences, thus greatly increasing cell capture efficiency and reducing the amount of starting material needed. Cell lysis, UV induced primer release from the hydrogels, mRNA capture, and RT all occur within the oil droplets. Similar to CEL-seq, the barcoded primers used by inDrop contain a T7 promoter. As such, after droplets are broken following RT, the remaining downstream steps are very similar to CEL-seq, from second strand DNA synthesis and IVT amplification to sequencing.

A few additional droplet-based methods have also been developed. 10X Genomics, Inc. has optimized and commercialized a droplet-based technology, called Chromium, that combines aspects of both inDrop and Drop-seq, including the use of deformable hydrogels and the use of cDNA fragmentation and sequencing adapter ligation, similar to inDrop, and the use of template switching/PCR amplification used in Drop-seq (***39***) (Figure 1.1E). 10X/Chromium has dramatically increased the accessibility of high-throughput single-cell RNA sequencing and has greatly improved sensitivity, precision, and noise, though with a slightly higher per-cell experimental cost prior to sequencing (***40***). DroNc-seq is another recently developed method that uses a modified Drop-seq approach that is compatible with cell nuclei, allowing for isolation of harder to dissociate cell types (***41***) (Figure 1.1E).

Although droplet-based methods have arguably been the most widely adopted single-cell RNA sequencing approaches, a number of similarly high throughput strategies have been developed that isolate cells by gravity into microwells containing barcoded beads (***42***, ***43***, ***44***). As a group, gravity-based methods are especially useful for low-input samples, requiring significantly less cells as starting material compared to droplet-based methods. The first such method, Cyto-seq, was released in 2015, around the same time as Drop-seq and inDrop (***42***). Cyto-seq utilizes a strategy by which cells are loaded by gravity into picoliter-sized wells of a Polydimethylsiloxane (PDMS) array generated from silicon wafers with an array of evenly spaced micropillars. To these wells are added barcoded beads, similar to those used in Drop-seq. A lysis buffer is then added, capturing mRNA from each cell onto the barcoded beads, and RT is performed. The beads are then pooled, and multiple rounds of PCR, initially using gene specific primers, are used to add sequencing adapters.

Additional gravity-based methods were developed over the next couple of years. One such method, Seq-well, is very similar to Cyto-seq, but adds a semipermeable polycarbonate membrane to the array to prevent cross contamination and cell loss, improving data quality (***43***) (Figure 1.1E). It then uses a largely identical protocol to

Drop-seq after pooling the beads for RT, from template switching and PCR amplification to sequencing. Another method, Microwell-seq, is very similar to Seq-well, but utilizes a simpler and more inexpensive system for generating microwell arrays (***44***) (Figure 1.1E). Specifically, it uses reusable silicon wafers with regularly spaced microwells to make PDMS micropillar arrays. These are also reusable and are used to make the agarose microwell arrays into which cells are loaded. Magnetic barcoded beads are also used to better ensure bead collection.

Finally, in addition to droplet-based and gravity-based methods, a number of high-throughput and affordable single-cell RNA sequencing technologies have been developed that rely on distributing cells across 96- or 384-well plates, and then performing one or more rounds of pooling and redistribution, barcoding the cells with each round (***45***, ***46***). Through this process, each cell is combinatorially labeled with a unique combination of barcodes. As a group, combinatorial indexing-based methods are simple, largely do not require specialized equipment, and allow for significant multiplexing, all of which lower the per-cell cost of these approaches and greatly increase their reach. The first such strategy to be released, sci-RNA-seq, relies on fluorescence activated cell sorting (FACS) sorting to distribute fixed and permeabilized cells, or isolated nuclei, into wells of a 96- or 384-well plate (***45***). Following FACS sorting, RT is performed on intact cells or nuclei, during which the first barcode and a UMI tag are introduced. Cells/nuclei are then pooled, and FACS is used to redistribute the cells/nuclei across the wells of a 96- or 384- well plate, this time at limiting numbers. Second strand synthesis, library generation, cell lysis, and PCR amplification of the libraries are all performed in these wells, with PCR primers that target the poly (T) primer on one end and the sequencing adaptor on the other introducing a second barcode specific to each well. Wells are then pooled and sequenced. A related method, SPLiT-seq, was released in 2018 and does not rely on FACS sorting of cells (***46***) (Figure 1.1F). Rather distribution of formaldehyde-fixed cells across 96-well plates occurs by manual mixing and pipetting. RT is performed on whole cells, adding a well-specific barcode. Cells are then pooled and redistributed, where an in-cell ligation

reaction adds a second barcode. A third round of pooling and splitting is performed, and a third barcode and UMI tag are added by ligation. Finally, the cells are pooled, split, and lysed. Sequencing adapters, as well as a fourth barcode, are introduced by PCR and the wells are pooled and sequenced. Because SPLiT-seq does not require FACS sorting, there is less bias in cell isolation, and the four rounds of barcoding allows for distinction between biological samples.

As described in this section, a large number of single-cell RNA sequencing technologies have been developed over the past couple of decades, each of which have distinct strengths and weaknesses (Table 1.1). Although most newer approaches allow for quick, high throughput, and affordable RNA sequencing of thousands of single cells, more low throughput technologies, such as SMART-seq, still have their advantages. Compared to more recent approaches that are strongly 3’ or 5’ biased, SMART-seq, SUPeR-seq, and MATQ-seq all have the capability of generating nearly full-length RNA sequences, and are thus better suited for certain applications, such as the analysis of splice sites. SUPeR-seq and MATQ-seq have the added benefit of allowing sequencing of non-mRNA RNA species. Furthermore, because lower-throughput methods generate fewer transcriptomes to be sequenced, a finite number of sequencing reads are spread across less cells, resulting in higher per-cell coverage and better detection of lowly expressed transcripts. More recent high-throughput methods also generally require a greater number of cells as starting material, which is not only untenable for many tissues, but the generation of thousands of single-cell transcriptomes is largely unnecessary for such samples. Newer, more high-throughput techniques do have many advantages, however. Biasing sequencing from the 3’ or 5’ end of the transcript provides strand specificity. Furthermore, the ability to sequence thousands of cells, even at lower coverage, allows for enhanced identification and characterization of subpopulations in complex tissue samples, with the sheer number of cells isolated from each subpopulation making up for missing transcripts due to low coverage from any one single cell of that subpopulation. More recent techniques have also introduced the use of molecular tags, or UMIs, that reduce noise generated from PCR amplification

artifacts. Although technologies described here vary widely in how single cells are isolated, how a unique barcode is introduced to each of those single cells, and how cDNA is amplified following RT, among other differences, virtually all approaches are capable of producing high quality single-cell RNA sequencing data. As such, the choice between technologies must largely be based on the constraints of the specific experimental question at hand.

Analysis of single-cell sequencing data

Each of the single-cell RNA sequencing methods profiled thus far generate large sequencing files containing millions of randomly ordered reads from often thousands of single cells. As such, computational methods are required to deconvolute this data and extract meaningful biological information (***47***). Most single-cell RNA sequencing methods have developed their own computational strategies for processing the raw sequencing data produced, taking into account various method-specific differences. These include Cell Ranger, which was developed to analyze Chromium 10X data (***39***), and a pipeline developed to process Drop-seq data (***37***), among many others. Despite the large number of approaches available for processing sequencing data generated by single-cell RNA sequencing, all approaches share certain computational considerations. Raw sequencing reads must be tagged with their various cell-specific barcodes, as well as the associated UMI tag, if any. The reads can then be aligned to the transcriptome or genome of choice, and those aligned reads, tagged with their cell-specific barcodes, can be used to generate a gene expression count matrix, with each unique barcode combination a distinct column and each gene or transcript a distinct row. Low-quality cells, including those with low transcript counts or low numbers of detected genes, those with very high transcript counts, indicating possible cell doublets, or those with a high percentage of mitochondrial genes, are removed from the data. The count matrix is then normalized to correct for relative gene expression differences between cells, as well as for differences in gene length for full-length RNA sequencing approaches. Regression models are also commonly used to remove technical and unwanted biological variation in the data, and a number of approaches have been developed to

correct for batch effects from different datasets, one popular example being canonical correlation analysis (CCA) (***48***). The resulting gene expression matrix is used as input for a variety of downstream applications used to uncover meaningful biological information.

Because single-cell RNA sequencing experiments generate expression data for potentially tens of thousands of genes, the data has a high dimensionality. Multiple approaches exist to make analysis of this highly complex data more manageable, better enabling real biological variation to be uncovered. A common first approach is feature selection, in which only highly informative genes, such as those with high variance across cells, are used for downstream analysis of the data (***49***). Following feature selection, algorithms for dimensionality reduction are commonly applied and are useful both for identifying inherent dimensionality in the data and for data visualization. Common algorithms for identifying inherent dimensionality include principal component analysis (PCA) (***50***), a common pre-processing step prior to clustering and visualization, and the generation of diffusion maps (***51***), which are commonly used for cell lineage reconstruction approaches reviewed in the following subsection. For visualization of the data, a number of algorithms are commonly used. The two most common are t-SNE (***52***) and UMAP (***53***), which are useful for plotting high dimensional data in low dimensional space, while largely retaining local relationships.

To determine the biological identity of each single cell in the data, cells can be clustered based on transcriptional similarities, and genes with enriched expression in each cluster can be identified. Some of the most widely used algorithms include *k*-means clustering, which iteratively defines a user-defined *k* number of centroids and assigns cells to the nearest centroid (***54***); hierarchical clustering, which either assigns each single cell as a cluster and iteratively groups clusters into ever more similar larger clusters (agglomerative approach) or iteratively splits one single cluster into ever more dissimilar smaller clusters (divisive approach); and graph-based clustering, in which cells are embedded into a multi-dimensional graph structure, edges are drawn between

transcriptionally similar cells, and the graph is divided into communities based on the degree of cell interconnectivity. A number of user-friendly packages have been released to help with the clustering workflow, from dimensionality reduction to clustering to differential expression analysis. Some popular packages include Seurat (***55***) and Cell Ranger (***39***). As a note, clustering of cells is not always the best approach for identifying biological variation. Actively differentiating cells of a single lineage commonly exhibit a continuum of gene expression. Therefore, rather than clustering, these cells are commonly placed on a one-dimensional manifold, an approach useful for cell lineage reconstruction, which is described below.

Trajectory reconstruction and lineage assessment using single-cell sequencing data

Changes in cell fate, be it during development, stem cell differentiation, or reprogramming, require dynamic and complex transcriptional changes. By sampling cells from tissues undergoing active changes in cell fate, such as in developing embryos and in tissues that are undergoing constant differentiated cell turn over in a stem cell-dependent process, a range of cells at different stages in the maturation process can be captured. Leveraging the concept that cells of similar maturation stages will share more similar transcriptional profiles, multiple methodologies have been developed that use single-cell RNA sequencing data to order cells along transcriptional trajectories, allowing for the identification of genes that vary significantly in their expression across these trajectories and that may be important for their progression.

Most trajectory reconstruction approaches first utilize dimensionality reduction, followed by construction of a minimum spanning tree (MST), definition of the path that connects the least differentiated to the most differentiated cells, and projection of the cells onto this path (***56***, ***57***, ***58***, ***59***, ***60***, ***61***, ***62***) (Figure 1.2A). Monocle, one of the first such approaches, was released in 2014, and, following user input of the cells that constitute the root state, uses independent component analysis (ICA) as its method for dimensionality reduction (***56***) (Figure 1.2A). The original iteration of Monocle could only

infer linear differentiation trajectories (Figure 1.2A). However, a later iteration, Monocle2, allows for more complex trajectories, including bifurcations, by grouping cells in higher dimensional space (***57***). The most recent iteration of Monocle was used to infer very complex trajectories of over one million cells during mouse organogenesis (***63***). Additional approaches include waterfall (***58***), TSCAN (***59***), and SLICE (***60***), each of which generate MSTs on predefined cell clusters following dimensionality reduction, reducing influence from outlier cells (Figure 1.2A). SLICE has the additional capacity to identify the trajectory start point by measuring transcriptome entropy (***60***). Yet other dimensionality reduction-based approaches include Slingshot, which fits smooth curves to a MST and projects cells onto the closest smooth curve (***61***), and SCUBA, which directly fits a smooth curve without generation of a MST (***62***) (Figure 1.2A).

An additional class of trajectory reconstruction methods are based on k-nearest neighbor graphs (k-NNGs), with each cell connected to its transcriptionally similar *k* nearest neighbors (Figure 1.2B). The first such method, Wanderlust, assigns a set of shortest walks from a manually assigned root cell, and takes the average to generate the most probable differentiation trajectories (***64***). Whereas Wanderlust is only able to predict linear trajectories, a very similar technique, Wishbone, does allow for more complex trajectories with bifurcations (***65***) (Figure 1.2B). Approximate graph abstraction (AGA) and population balance analysis (PBA) are yet other NNG-based approaches (***66***, ***67***), with AGA averaging cells into clusters before trajectory reconstruction (***66***), and PBA predicting differentiation direction by estimating the velocity of cell differentiation based on NNG local cell density (***67***). Additional methods utilizing NNG-based strategies have been used to generate complex developmental trajectories using single-cell RNA sequencing data from developing zebrafish (***68***, ***69***) and *Xenopus tropicalis* embryos (***70***), as well as differentiation trajectories of the adult *Hydra* polyp (***71***).

Other trajectory reconstruction approaches include StemID (***72***) and Mpath (***73***), which first cluster cells and connect cluster centers in high dimensional space, followed by

projection of single cells onto the edges of the connections and removal of poorly populated edges (Figure 1.2C). RNA velocity is an entirely different approach that uses unspliced mRNA as a measure of a cell's future transcriptional profile, with the gene-by-gene fraction of unspliced transcripts used to infer trajectories without a manually defined root-state (***74***) (Figure 1.2D).

The trajectory reconstruction approaches described thus far require extensive sampling of intermediate states and cannot record lineage relationships (cell division histories) of single cells. To overcome these limitations, single-cell RNA sequencing or multiplex fluorescent *in situ* hybridization (FISH) can be combined with genetic lineage-tracing strategies. A number of such methods have been developed, many of which have taken a CRISPR/Cas9-based approach. These approaches utilize the fact that Cas9, in the absence of a repair template, will generate deletions or insertions in the targeted DNA. Over time, these mutations accumulate, generating heritable marks that can be detected and used to infer lineage relationships between cells. One such method, MEMOIR, was released in 2017 and targets Cas9 to an array of genomic sequences with an associated barcode, generating progressive heritable marks that are then assayed using seq-FISH (***75***). Another such method, scGESTALT, targets inducibly expressed Cas9 to a barcode sequence contained within the 3' UTR of a transgene, and uses inDrop to identify the progressive heritable marks (***76***). Because an inducible Cas9 system is used, lineage relationships in the juvenile zebrafish brain could be examined. Two additional Cas9-based methods include LINNAEUS and ScarTrace, which both target Cas9 to multiple copies of red fluorescent protein (RFP) or green fluorescent protein (GFP) integrated into the zebrafish genome (***77***, ***78***). LINNAEUS targets 16-32 RFP sequences spread throughout the genome (***77***), whereas ScarTrace targets eight in-tandem GFP sequences (***78***). Both methods then utilize inDrop to identify progressive heritable marks. Finally, rather than CRISPR/Cas9, TracerSeq is a method that uses a Tol2 transposase system to genomically integrate GFP transcripts, each of which contain a unique barcode in their 3' UTR (***69***). Insertions occur asynchronously over many divisions, generating unique barcode combinations that are

identified by inDrop. This approach was used to identify lineage relationships in the developing zebrafish embryo.

Certain features of the previously described approaches for combining lineage tracing with single-cell RNA sequencing, such as the need to inject constructs early in development, limited their translation to a mammalian system. As such, multiple additional approaches have been released and used to profile lineage relationships in early mouse embryo development and in hematopoiesis, using 10X genomics (***79***, ***80***) or InDrop (***81***) to read out the progressive heritable changes. Two approaches utilized a CRISPR/Cas9-based approach (***79***, ***80***). One such approach uses an array of sixty genomically-integrated homing CRISPR guide RNAs (hgRNAs) that target their own genomic loci (***79***), and the other approach uses three genomically integrated guide RNAs that target a DNA sequence contained within the 3' UTR of a fluorescent transgene, multiple copies of which are spread throughout the genome (***80***), both of which were used to profile early mouse development. A third approach, termed lineage and RNA recovery (LARRY), uses a lentiviral library containing GFP constructs uniquely barcoded in their 3' UTR and under control of an EF1alpha promoter to singly infect cells and was used to profile hematopoiesis in hematopoietic stem and progenitor cells (HSPCs) cultured *in vitro* and transplanted *in vivo* (***81***). Finally, whereas previously described methods actively generate and detect heritable marks within cells, additional methods have been developed that retroactively detect endogenous tags arising from naturally occurring mutations. Single-cell RNA sequencing and single-cell ATAC-seq (Assay for Transposase-Accessible Chromatin with high-throughput sequencing), which is described in a later section, are capable of detecting endogenous mutations in mitochondrial DNA, in which somatic mutations occur at a much higher rate than genomic DNA, and can be used for lineage tracing of human cells (***82***, ***83***).

Spatial transcriptomics

All single-cell RNA sequencing approaches described thus far require dissociation of the tissue sample into a single-cell suspension prior to cell isolation. As a result, all

spatial information regarding the arrangement of cell types within the overall tissue is lost. Given the immense interconnectivity with which cell types function within an organism, this spatial information is of great importance. A number of approaches have been developed that attempt to infer this spatial information from data generated using these previously established dissociative methods. In addition, a number of novel methods have been developed that retain spatial information for each cell throughout the sequencing process. Finally, methods for highly multiplexed FISH, including *in situ* sequencing, have been developed that allow for the direct imaging and identification of thousands of transcripts within intact tissue sections.

Approaches that infer spatial information from single-cell RNA sequencing data generated using previously described dissociative methods do so by determining the spatial tissue-level expression pattern for a number of genes by *in situ* hybridization (ISH), some of which overlap spatially and some of which do not (Figure 1.3A). These expression patterns are digitized, and the unique gene expression profiles for cells in the data are used to infer the rough spatial orientation of that cell in the original tissue (Figure 1.3A). Four conceptually similar methods have been released that have demonstrated the success of this approach. The first of these methods, released in 2015, were Seurat (***55***) and an approach by the Marioni lab (***84***). Seurat was used to analyze 851 single cells isolated and sequenced by SMART-seq from the developing *Xenopus* embryo (***55***). 47 ISH patterns were used to generate a reference spatial map, which was used to assign each cell to one of 128 bins distributed along the dorsal-ventral and animal-vegetal axis. Seurat was largely able to assign cells isolated from distinct regions of the embryo to the correct bin and to fairly accurately predict ISH patterns for genes and rare cell types not provided to the algorithm. A similar approach by the Marioni lab used up to 98 ISH images to infer the spatial positions of around 139 single cells isolated and sequenced using Fluidigm C1 from the marine annelid *Platynereis dumerilii*. (***84***). A third approach, Distmap, was released in 2017 and used ISH images for 84 genes to infer spatial positions of around 1,300 single cells isolated and sequenced using Drop-seq from stage 6 *Drosophila melanogaster* embryos (***85***).

Finally, novoSparc, released in 2019, is unique in not requiring existing *in situ* patterns for spatial reconstruction, though *in situ* images can be incorporated (***86***). Using a range of 0-84 ISH images, novoSparc was able to infer spatial patterns for cells isolated from effectively 2D tissues, such as the mouse intestinal epithelium and liver lobules, as well as cells isolated from more complex tissues, such as the stage 6 *Drosophila melanogaster* and the developing *Xenopus* embryo. Although reconstruction quality was slightly improved over previous methods, and less ISH reference images were required for high-quality reconstruction, reconstructions performed using no marker genes were fairly poor.

Rather than inferring positional information from single-cell data generated using dissociative techniques, three new approaches have been developed that seek to retain this positional information throughout the sequencing process. The first such method, developed in 2016, utilizes an array of glass slide-anchored oligonucleotides containing spot-specific barcodes, UMI tags, and oligo (dT) sequences (***87***). A tissue slice is placed onto the array, where it is permeabilized, RT is performed, and cDNA is captured onto the closest oligonucleotide spot. Following RT, the tissue is fully digested, and an IVT-based method is used for amplification. Because the location of each barcode sequence on the glass slide is known, the spatial location of each read can be determined through its associated barcode. Two additional methods, developed in 2019, are conceptually very similar but utilize glass-immobilized arrays of barcoded beads, similar to those used in Drop-seq, with the spot-specific barcode sequences determined by *in situ* indexing (***88***, ***89***). One such method, Slide-seq, pools beads following RT and tissue digestion, and further downstream steps, including template switching and PCR amplification, are performed almost identically to the Drop-seq approach (***88***) (Figure 1.3B). The other such method, HDST, also pools beads following RT and tissue digestion, but downstream steps are very similar to those used in the glass slide-anchored oligonucleotide-based method described above (***87***), including IVT amplification (***89***). Although HDST reports a spatial resolution of 2 µm for mRNA capture (***89***), all three methods suffer from potential cross contamination by cells in

close proximity, and thus do not provide traditional single-cell resolution. Despite this limitation, these methods have successfully characterized roughly transcriptome-wide gene expression profiles for tissue sections of the adult mouse olfactory bulb (***87***, ***89***), the mouse hippocampus and cerebellum (***88***), and a breast cancer tumor (***89***).

Finally, a number of a multiplex FISH techniques have been developed that enable detection of individual transcripts for thousands of genes directly in intact tissue slices. One of the first such methods, sequential FISH (seqFISH), was developed in 2014 and allows for multiple rounds of probe hybridization with up to four fluorophores, followed by probe stripping (***90***). A barcoding scheme was utilized wherein individual genes were uniquely marked by a combination of the hybridization rounds in which probes to that gene were added and the fluorophores used to label the probe in each round, theoretically allowing for coverage of the entire transcriptome (Figure 1.3C). While initially only compatible with cultured cells, successive optimized versions of this approach enabled profiling of tissue sections and were used to profile up to 249 genes in 16,958 cells of the mouse hippocampus (***91***) and 10,000 genes in 2,963 cells from brain slices of the mouse subventricular zone and olfactory bulb (***92***). merFISH is a related approach, developed in 2015, that also uses multiple rounds of hybridization to barcode cells, but includes an error-correcting barcode system to lower noise in the system (***93***). While also initially only compatible with cultured cells, merFISH has also undergone successive rounds of optimization that have enabled profiling of 155 genes in one million cells of the hypothalamic preoptic region (***94***). Finally, osmFISH was developed in 2018 and does not rely on barcoding, but rather directly detects a small number of genes in each hybridization round, with multiple rounds of hybridization and stripping (***95***).

A number of FISH-based methods have also been developed to directly sequence RNA species in intact cells, all of which utilize rolling-circle amplification (RCA) to amplify signal (***96***, ***97***, ***98***). Only one such method, STARmap, is compatible with tissue sections (***98***). STARmap was developed in 2018, and utilizes two complementary DNA

probe sets, one of which contains a 5-bp barcode, to generate a template for RCA (***98***). During RCA, amine-modified nucleotides are added that enable imbedding into a tissue hydrogel, increasing optical transparency and reducing background, among other benefits. The barcode is then sequenced using fluorescent readout probe hybridization and stripping, using an error-reducing two-base sequencing scheme. STARmap was used to profile 160-1,020 genes in sections of mouse primary visual cortex and medial prefrontal cortex.

Additional single-cell genomic technologies

The single-cell sequencing techniques highlighted thus far have largely been single-cell RNA sequencing approaches. However, a number of single-cell approaches have been developed over the past decade that have enabled profiling of genomic features ranging from DNA methylation to chromatin state, as well as methods allowing for the detection of combinations of these features. Methods have been developed for single-cell detection of genomic copy-number variations (***99***) and full genome sequencing in single-cells (***100***, ***101***, ***102***). Multiple single-cell methods for analyzing the chromatin state of a cell have also been developed. These include methods for single-cell ATAC-seq to identify regions of open chromatin (***103***), single-cell whole genome bisulfite sequencing to measure DNA methylation (***104***, ***105***, ***106***, ***107***, ***108***), single-cell Hi-C to measure chromosome conformation (***109***, ***110***, ***111***, ***112***, ***113***), and single-cell Chip-seq to measure histone modifications (***114***, ***115***). Furthermore, single-cell methods have been developed that enable profiling of multiple genomic features in the same cell. These include approaches enabling both transcriptional profiling and profiling of either chromatin accessibility (***116***), DNA methylation (***117***), or protein epitopes (***118***, ***119***), as well as an approach enabling profiling of both chromatin accessibility and DNA methylation (***120***).

Applications and outlook for single-cell sequencing technologies

The explosion of single-cell genomic approaches over the past decade has transformed many aspects of biological research. The ability to quickly and affordably generate

transcriptomes for thousands of single cells has enabled cellular profiling of entire organisms and organ systems, both in the adult and during development, and has allowed evolutionary comparisons between a wide range of organisms. Furthermore, the ability to profile cellularly heterogeneous diseases, such as cancer, provides an invaluable tool in the study and treatment of such diseases, and the ease and affordability of current methods enables their use as readouts for genetic and chemical screens, with great promise for accelerated discovery in both basic science and translational research.

Animal species can contain many hundreds of distinct cell types and cell states, the composition of which varies widely over the course of development. Despite this great complexity, high throughput single-cell RNA sequencing approaches have enabled the generation of transcriptomes for most cell types in a number of animals and tissues - so called cell-type transcriptome “atlases”. These atlases have been especially transformative for emerging model organisms with limited molecular tools available. A whole-animal cell type transcriptome atlas has been generated for the asexual planarian *Schmidtea mediterranea,* as will be described in chapter 2. In addition, transcriptomes for most cell types of the adult cnidarian *Hydra* polyp (***71***), the ctenophore *Mnemiopsis leidyi* (***121***), the placozoan *Trichoplax adhaerens* (***121***), the marine annelid *Platynereis dumerilii* (***122***), and both adult and larval stages of the cnidarian *Nematostella vectensis* (***123***) and the sponge *Amphimedon queenslandica* (***121***) have been generated. Transcriptomes have also been generated from most cells of embryonic and larval C*aenorhabditis elegans* (***31***, ***45***, ***124***, ***125***) and from most cells across the life cycle stages of the ascidian *Ciona intestinalis* (***126***), as well as for a comprehensive collection of different organ systems from the mouse (***44***, ***127***). An effort to generate a human cell atlas to transcriptionally profile all human cell types and cell states is currently underway (***128***). Transcriptome atlases have also been generated for a number of organs and tissues. As an example, transcriptional profiling of cells from the lung airway epithelium revealed that the gene *CFTR*, mutations in which cause cystic fibrosis, was exclusively expressed in a previously unidentified cell type called the ionocyte (***129***, ***130***).

Animal development is a highly complex process, with the diverse array of differentiated tissue types of an adult animal all arising from a single fertilized egg. Given this daunting complexity, much remains unknown of this fundamental process. With single-cell RNA sequencing techniques enabling transcriptome generation for tens of thousands of single cells, however, the ability to transcriptionally profile this complex process has become a reality. As has been previously mentioned, single-cell RNA sequencing technologies have been applied to a wide range of animals at various stages of development, including zebrafish (***68***, ***69***), *Xenopus tropicalis* (***70***), mouse (***63***, ***131***), *C. elegans* (***31***, ***45***, ***124***, ***125***), and the ascidian *Ciona intestinalis* (***126***). Using a variety of trajectory reconstruction techniques, as reviewed above, complex developmental trajectories were generated for each organism, both confirming existing knowledge of certain developmental trajectories and refining understanding of others (***63***, ***68***, ***69***, ***70***, ***125***, ***126***, ***131***). Although all trajectory reconstruction techniques were performed on dissociated cells, precluding the direct analysis of lineage relationships, one zebrafish study used TracerSeq to combine genetic lineage tracing with single-cell RNA sequencing, confirming that most clonally related cells were in close proximity in the trajectory plot, but also identifying credible instances of divergent clones (***69***). This same zebrafish study also profiled developmental trajectories for embryos with CRISPR/Cas9 induced loss of function *chordin* mutations, revealing an expected expansion of ventral tissues and loss of dorsal tissues compared to control animals, among other findings (***69***). Together, these results demonstrate the potential for combining genetic modifications and lineage tracing with single-cell RNA sequencing to decipher the developmental logic across diverse animal embryos.

In addition to profiling development of animal embryos, other studies have used single-cell genomic approaches to profile lineage trajectories and identify stem cell populations in adult tissues. As one example, human hematopoietic cell types were recently profiled using both single-cell ATAC-seq and single-cell RNA seq, identifying chromatin-level lineage bias for various multipotent progenitors and identifying transcription factor

expression associated with these observed chromatin-level changes (***132***). As another example, single-cell RNA sequencing of mouse intestine was recently used to identify an injury-induced cell type, termed the “revival” stem cell, that is normally rare and quiescent in homeostasis, but is able to give rise to all major intestinal cell types following injury (***133***).

Systematic whole genome sequencing across the animal kingdom has revolutionized our understanding of organismal evolution. With the capacity to profile cell types at the transcriptome and chromatin level, cell type comparisons between animal species has the potential to transform our understanding of the evolution of cell types and gene regulatory networks. In fact, a number of studies have demonstrated the utility of such an approach. To give just two recent examples, comparison of transcriptional profiles from glutamatergic and GABAergic neurons isolated from reptilian and mouse brains revealed that whereas most mammalian GABAergic classes also exist in reptiles, mammals possess many more glutamatergic neuron types (***134***). As another example, comparison of cell-type-specific transcription factor expression and associated promoter sequences for sponges, ctenophores, and placozoans revealed that transcription factor motifs are highly predictive of cell type in less cellularly complex placozoans and sponges, but less so in more cellularly complex ctenophores, among other findings (***121***).

In addition to their growing importance in basic biological research, single-cell genomic technologies can also be used to profile disease states, not only enhancing our molecular understanding of diseases, but also identifying potential therapeutic targets. Tumors are highly heterogeneous in their cell type composition, and have thus been poorly characterized using bulk genomics techniques. Single-cell whole genome sequencing has been instrumental in measuring clonal evolution during tumor development (***99***, ***101***), and single-cell RNA sequencing has been used to profile a number of tumors types, as well as their microenvironment. As just one example, notoriously heterogeneous glioblastomas were recently profiled through single-cell RNA

sequencing of patient-derived organoids (***135***, ***136***). Our understanding of other human diseases can also benefit from these technologies. These include schizophrenia and other neurological disorders, for which the genetic bases are poorly understood. As one example, a recent study mapped genetic variants and gene sets associated with schizophrenia to transcriptional profiles generated by single-cell RNA (***137***). In doing so, it was recognized that gene variants were only associated with four major neuronal cell types, each of which can now be targeted for further characterization.

Finally, single-cell genomic approaches are especially effective for profiling highly heterogeneous samples and are thus ideal tools for reading out transcriptional and chromatin-level changes arising from chemical and genetic screens. Indeed, a number of such approaches have been developed. An approach for combining single-cell transcriptomics with chemical screening was recently established (***138***). Similarly, multiple methods have been released that combine CRISPR/Cas9-based screening with single-cell RNA sequencing. These include Mosaic-seq (***139***), Perturb-seq (***140***, ***141***), Crisp-seq (***142***), and CROP-seq (***143***), as well as a method that combines CRISPR/Cas9 based screening with single-cell ATAC-seq (***144***). These methods have been used to probe the genetic circuitry underlying epithelial-to-mesenchymal transitions (***145***), and have identified iPSC-derived neuron-essential genes (***146***), as well as genes involved in immune activation (***141***), as just a few examples.

Table 1.1

Methods	5' or 3' bias?	UMI capability?	Strand Specificity?	mRNA only?	Template Switching?	Amplification Method
Non Tag-Based Methods						
Tang Method	Nearly full-length	No	No	Yes	No	PCR
Smart-seq	Full-length	No	No	Yes	Yes	PCR
Smart-seq2	Full-length	No	No	Yes	Yes	PCR
Quartz-Seq	Full-length	No	No	Yes	No	PCR
SUPeR-seq	Full-length	No	No	No	No	PCR
MATQ-seq	Full-length	Yes	Yes	No	No	PCR
STRT-seq/C1*	5' biased	Yes	Yes	Yes	Yes	PCR
Tag-based Single-cell-per-well Methods						
Quartz-Seq2	3' biased	Yes	Yes	Yes	No	PCR
CEL-seq	3' biased	Yes	Yes	Yes	No	IVT
CEL-seq2	3' biased	Yes	Yes	Yes	No	IVT
MARS-seq	3' biased	Yes	Yes	Yes	No	IVT
Droplet-based Methods						
Drop-seq	3' biased	Yes	Yes	Yes	Yes	PCR
InDrop	3' biased	Yes	Yes	Yes	No	IVT
Chromium	3' biased	Yes	Yes	Yes	Yes	PCR
DroNC-seq	3' biased	Yes	Yes	Yes	Yes	PCR
Gravity-based Methods						
CytoSeq	3' biased	Yes	Yes	Yes	No	PCR
Seq-Well	3' biased	Yes	Yes	Yes	Yes	PCR
Microwell-seq	3' biased	Yes	Yes	Yes	Yes	PCR
Combinatorial Indexing Methods						
SPLiT-seq	3' biased	Yes	Yes	Yes	Yes	PCR
sci-RNA-seq	3' biased	Yes	Yes	Yes	No	PCR

Table 1.1. Summary of the capabilities of available single-cell sequencing methods.

Table adapted from (***147***). For each method, the presence of a 5’ or 3’ gene bias for sequencing reads, the capability for a UMI counting strategy, the presence of strand specificity, the capability to sequence additional types of RNA other than mRNA, the use of a template switching approach, and the amplification approach utilized are indicated. PCR, Polymerase Chain Reaction; IVT, in vitro transcription. * The earliest iteration of STRT-seq was tag-based and allowed for earlier pooling of samples. For STRT-seq/C1, samples are not pooled until following library generation, and it was thus classified as a non-tag-based method.

Figure 1.1

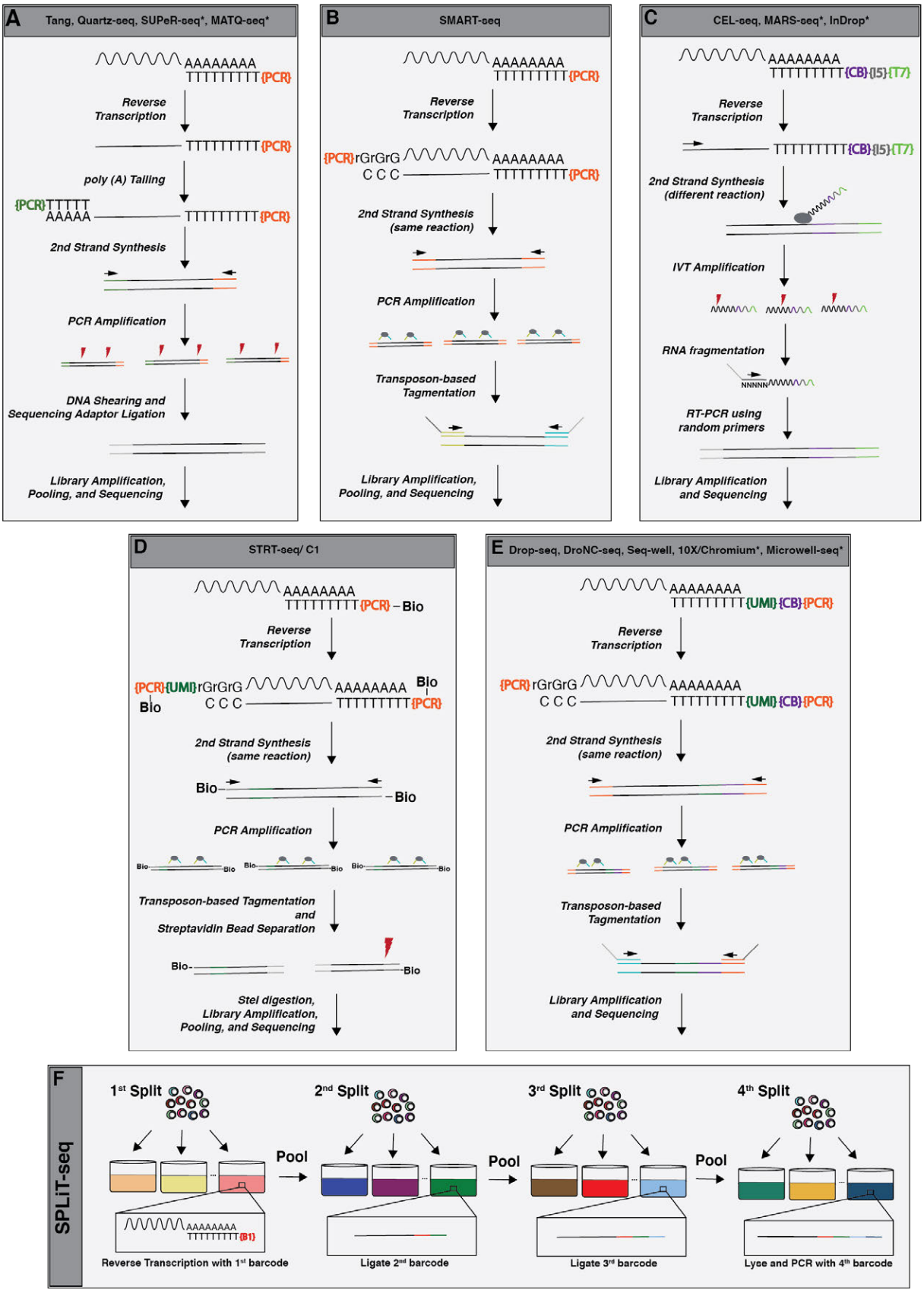

A
Tang, Quartz-seq, SUPeR-seq*, MATQ-seq*
Reverse Transcription
poly (A) Tailing
2nd Strand Synthesis
PCR Amplification
DNA Shearing and Sequencing Adaptor Ligation
Library Amplification, Pooling, and Sequencing
B
SMART-seq
Reverse Transcription
2nd Strand Synthesis (same reaction)
PCR Amplification
Transposon-based Tagmentation
Library Amplification, Pooling, and Sequencing
C
CEL-seq, MARS-seq*, InDrop*
Reverse Transcription
2nd Strand Synthesis (different reaction)
IVT Amplification
RNA fragmentation
RT-PCR using random primers
Library Amplification and Sequencing
D
STRT-seq/ C1
Reverse Transcription
2nd Strand Synthesis (same reaction)
PCR Amplification
Transposon-based Tagmentation and Streptavidin Bead Separation
Stel digestion, Library Amplification, Pooling, and Sequencing
E
Drop-seq, DroNC-seq, Seq-well, 10X/Chromium*, Microwell-seq*
Reverse Transcription
2nd Strand Synthesis (same reaction)
PCR Amplification
Transposon-based Tagmentation
Library Amplification and Sequencing
F
SPLiT-seq
1st Split
2nd Split
3rd Split
4th Split
Pool
Pool
Pool
Reverse Transcription with 1st barcode
Ligate 2nd barcode
Ligate 3rd barcode
Lyse and PCR with 4th barcode

Figure 1.1. Common paradigms for library preparation used in single-cell RNA sequencing.

Illustrations adapted from (***148***). As a note, starred methods vary slightly in the library preparation approach. **(A)** General library preparation method adopted by the Tang method, Quartz-seq (and Quartz-seq 2), SUPeR-seq, and MATQ-seq. SUPeR-seq utilizes random primers, not poly (T)-based primers, for reverse transcription. MATQ-seq utilizes random primers in addition to poly (T)-based primers for reverse transcription, and utilizes poly (C) tailing, rather than poly (A) tailing, and a poly (G) primer for 2nd strand DNA synthesis. Both MATQ-seq and Quartz-seq2 utilize a UMI strategy. Quartz-seq2 adds a cell barcoding step during RT, allowing much earlier pooling of samples. **(B)** Library preparation method adopted by SMART-seq and SMART-seq2. **(C)** General library preparation method adopted by CEL-seq (and CEL-seq2), MARS-seq, and InDrop. CEL-seq2, MARS-seq, and InDrop all utilize a UMI strategy. **(D)** Library preparation method adopted by STRT-seq/CI. **(E)** General library preparation method adopted by Drop-seq, DroNC-seq, Seq-well, 10X/Chromium, and Microwell-seq. 10X/Chromium uses cDNA fragmentation and ligation of sequencing adapters to generate libraries from amplified cDNA, and microwell-seq utilizes an expanded cell barcode scheme compared to the other methods. **(F)** Illustration adapted from (***46***). General experimental workflow for generating cell-specific barcode combinations using SPLiT-seq. As a general note, all non-tag-based methods (i.e. A, B, and D), Illumina sequencing adaptors (in two shades of grey) can possess barcoded indices, allowing for sample pooling prior to sequencing

Figure 1.2

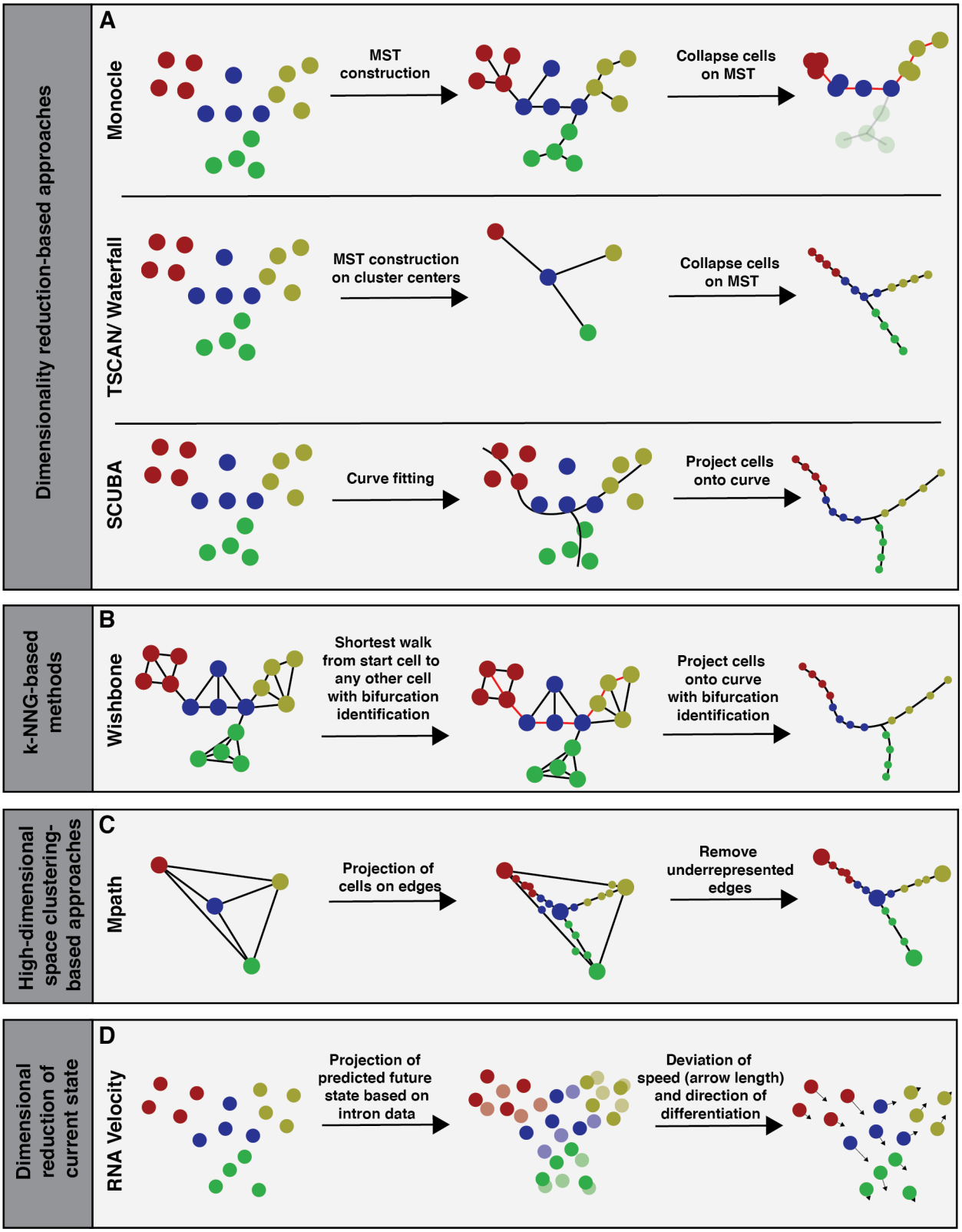

Dimensionality reduction-based approaches
A
Monocle
MST construction
Collapse cells on MST
TSCAN/ Waterfall
MST construction on cluster centers
Collapse cells on MST
SCUBA
Curve fitting
Project cells onto curve
k-NNG-based methods
B
Wishbone
Shortest walk from start cell to any other cell with bifurcation identification
Project cells onto curve with bifurcation identification
High-dimensional space clustering-based approaches
C
Mpath
Projection of cells on edges
Remove underrepresented edges
Dimensional reduction of current state
D
RNA Velocity
Projection of predicted future state based on intron data
Deviation of speed (arrow length) and direction of differentiation

Figure 1.2. Common approaches for trajectory reconstruction of single-cell data.

Illustrations adapted from (***149***). **(A)** Schematic of common dimensionality reduction-based approaches for trajectory reconstruction. Although the original Monocle is profiled, which did not allow for bifurcations, more recent versions do allow for bifurcations. SLICE takes a similar approach to TSCAN and Waterfall, with the added capability to infer directionality from transcriptome entropy. **(B)** Schematic of a typical k-NNG-based approach. Wanderlust is another such method that does not allow for bifurcations. **(C)** Schematic of a typical high-dimensional space clustering-based approach. StemID takes a very similar approach to Mpath, with the added capability of predicting stem cell populations. **(D)** Schematic of the RNA Velocity approach.

Figure 1.3

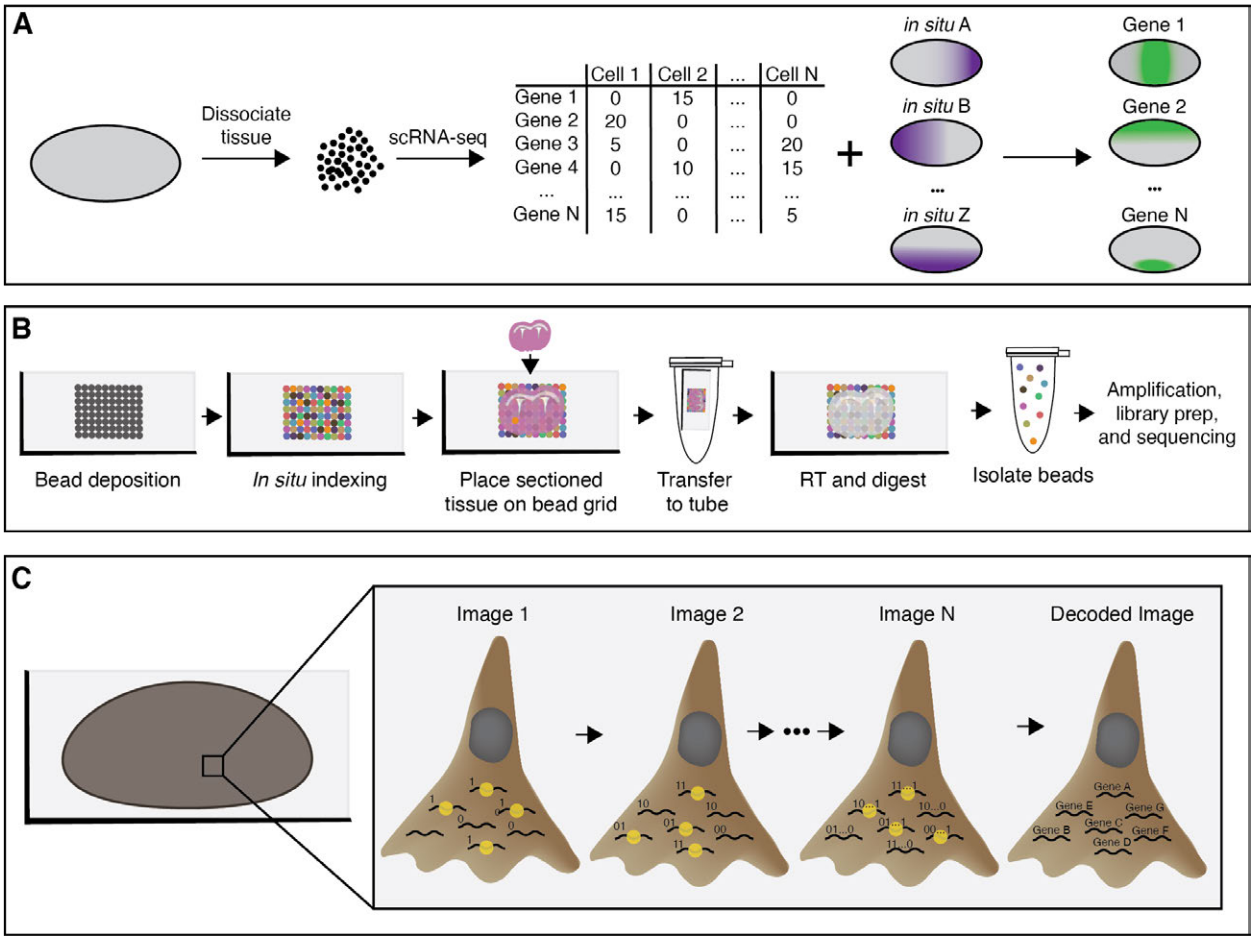

A
Dissociate tissue
scRNA-seq
Cell 1 Cell 2 ... Cell N
Gene 1 0 15 ... 0
Gene 2 20 0 ... 0
Gene 3 5 0 ... 20
Gene 4 0 10 ... 15
... ...
Gene N 15 0 ... 5
in situ A
in situ B
in situ Z
Gene 1
Gene 2
Gene N
B
Bead deposition
In situ indexing
Place sectioned tissue on bead grid
Transfer to tube
RT and digest
Isolate beads
Amplification, library prep, and sequencing
C
Image 1
Image 2
Image N
Decoded Image
Gene A
Gene E
Gene G
Gene C
Gene B
Gene F
Gene D

Figure 1.3. Common spatial transcriptomic approaches.

(A) Illustration adapted from (***86***). Schematic of the general approach used to infer spatial information from single-cell RNA sequencing data using existing ISH images. **(B)** Illustration adapted from (***88***). Schematic of the approach used by Slide-seq to retain positional information throughout the sequencing process. **(C)** Illustration adapted from (***93***). Schematic of the general approach used by seqFISH-based approaches, including merFISH, which also utilizes an error correcting barcode system. While only one fluorophore was used in the schematic, multiple fluorophores are compatible with these approaches.

II. Planarians as a regenerative model system

An introduction to planarians as a model system

Planarians are freshwater flatworms well known for their capacity for whole-body regeneration (***150***, ***151***) (Figure 1.4A). They are members of the Spiralian superphylum (***152***) and as bilaterians, possess three body axes: anterior-posterior (AP), dorsal-ventral (DV), and medial-lateral (ML) (Figure 1.4B). Planarians possess a complex anatomy, consisting of multiple organ systems, many of which are found across the animal kingdom. These include an excretory system, made up of protonephridia; both a central and peripheral nervous system; a digestive system, including an intestine and a pharynx used for feeding and defecation; a muscular system consisting of networks of muscle fibers that run in different orientations; and an epidermis (Figure 1.4C). There is also a diverse collection of mesenchymal cell types that are found in the space between the organ systems of the animal, called the parenchyma. A more detailed account of planarian anatomy will be provided in a later subsection. A number of molecular tools are available for the study of planarians, including RNA interference (RNAi) for inhibiting gene expression (***153***, ***154***) and *in situ* hybridization for visualizing gene expression (***155***). Genomic resources are also available, including a number of transcriptome assemblies (***156***), and a recently completed draft of the planarian genome (***157***). Whereas a European species, *Schmidtea mediterranea*, has been used for most recent molecular studies of planarians and will be the focus of this thesis, there are many species of planarians all over the world (***158***). Furthermore, there are two separate strains of *Schmidtea mediterranea*: a sexual cross-fertilizing hermaphroditic strain, and an asexual strain that reproduces by fissioning and subsequent regeneration. The asexual strain will be the primary focus of this thesis.

Following tissue loss, planarians generate an unpigmented outgrowth of new tissue, called the blastema, that forms at the site of injury (Figure 1.4D). Because many injuries lead to the loss of the brain or pharynx, both of which preclude eating, regeneration must occur with existing nutrients only. Only some structures are regenerated in the

blastema, depending on the site of amputation, with other structures regenerated in the pre-existing tissue (Figure 1.4D). Within seven days, all differentiated structures will have fully formed, though the proportion of these structures is generally not yet resolved by this time. Over the next several days, in a process known as morphallaxis, tissues will rescale in size and position, largely through new cell production and cell death, to generate a correctly proportioned animal (***159***). Planarians can regenerate from a wide range of injuries, including following transverse amputation along the AP axis, sagittal and parasagittal amputation along the ML axis, and a variety of irregularly shaped injuries (Figure 1.4D).

Neoblasts are the source of all new tissue in the animal

The remarkable regenerative capacity of planarians is largely derived from a population of dividing cells, called neoblasts. Neoblasts reside in the parenchyma and are very abundant (***160***) (Figure 1.5A). By electron microscopy, neoblasts exhibit a distinct morphology, with sparse cytoplasm and few mitochondria, no endoplasmic reticulum, and many free ribosomes (***161***). Neoblasts are also characterized by the presence of chromatoid bodies (***162***), similar to RNA granules in germ cells, and exhibit enriched expression of genes commonly associated with germ cells, including genes encoding PIWI (***163***), Vasa (***164***), and Bruno-like proteins (***165***). In fact, the PIWI protein-encoding gene *smedwi*-1 is commonly used as a pan-neoblast maker (***163***).

Neoblasts are the only dividing somatic cells in the planarian and are the source of all new tissue. Neoblasts not only provide the cellular material for regeneration, but they are also used to homeostatically turn over all cell types throughout the life of the animal. Because of this constant turnover, planarians actively grow and shrink depending on nutrient availability, while largely retaining proper body proportions (***166***). Neoblasts can be specifically ablated using gamma irradiation, blocking the generation of all new tissue (***163***, ***167***). Lethal doses of irradiation not only prohibit regeneration, but also prevent homeostatic cell turn over, leading to the death of the animal (***163***, ***167***). Some non-lethal doses of irradiation, on the other hand, can kill all but a few neoblasts, which

will clonally expand to generate both more neoblasts as well as differentiated cells from diverse tissue classes (***168***, ***169***). At least some individual neoblasts are pluripotent, with the injection of a single neoblast enabling the rescue of a lethally irradiated animal (***169***) (Figure 1.5B). In fact, injection of a single neoblast isolated from an asexual planarian into an irradiated sexual planarian will not only rescue the recipient animal, but will also convert that animal from the sexual genotype to that of the asexual strain through tissue turnover (***169***) (Figure 1.5B).

Neoblasts are strikingly homogenous in terms of their morphology and gene expression. However, multiple studies have identified extensive transcriptional heterogeneity within the population. As further reviewed in the following subsection, multiple studies have found that subsets of *smedwi-1*$^{+}$ neoblasts, called specialized neoblasts, express transcription factors that are required for the specification of certain tissues, including the eye (***170***, ***171***) and the protonephridia (***172***), among others. Because neoblasts are the only actively dividing somatic cells, DNA dyes, together with FACS, can be used to isolate neoblasts through their >2C DNA content (***173***). Single-cell multiplexed qPCR analysis on 4C cells revealed transcriptionally distinct populations of neoblasts, including a population called zeta-neoblasts, which function as epidermal progenitors, and gamma-neoblasts, which function as intestinal progenitors (***174***). Single-cell RNA sequencing of wounded animals further characterized these transcriptionally distinct zeta- and gamma-neoblast specialized neoblasts (***175***). It is currently unclear whether these specialized neoblasts are irreversibly committed to their cell fate, or if specialized neoblasts retain some pluripotent potential. FACS isolation of tetraspanin domain-containing protein (TSPAN-1) positive neoblasts were recently shown to be pluripotent (***176***). As discussed in chapter 2, this neoblast population is transcriptionally enriched for a number of genes associated with neuronal cell types, suggesting these cells may constitute neuronal progenitors. If so, it would imply that specialized neoblast classes possess pluripotent potential, though more work is necessary to test this hypothesis.

As neoblasts exit the cell cycle and undergo differentiation into the diverse tissues of the animal, a number of transcriptional changes must occur. A number of tools have been developed in the field to identify these transcriptional changes and to mark these post-mitotic progenitors. Following lethal irradiation, temporal loss in gene expression can be used to identify transcriptionally distinct transition state populations, with expression of genes transiently expressed early in a differentiation lineage being lost prior to expression of genes turned on later in the lineage (***177***). This approach was used to identify post mitotic transition state populations for the epidermis (***174***, ***177***, ***178***, ***179***) and for the pharynx (***180***). A number of methods are also available for labeling these post-mitotic differentiating cells. As neoblasts are the only dividing cells, Bromodeoxyuridine (BrdU) is taken up by neoblasts and is then passed on to all *smedwi-1*-negative progeny (***177***, ***181***). This approach was utilized to demonstrate that the different transient maturation states for the epidermis were in fact from the same lineage (***177***). Finally, recent post-mitotic progenitors can be identified by immunostaining for SMEDWI-1 protein. Because this protein perdures longer than mRNA after *smedwi-1* transcription ceases, recent post-mitotic progenitors that have turned off *smedwi-1* expression will still be SMEDWI-1^{+} (***165***, ***182***, ***183***).

Planarians contain diverse tissues that are maintained by distinct progenitors

Planarians possess a fairly complex anatomy, consisting of multiple distinct tissues (Figure 1.4C). For much of the 20th century, knowledge of planarian anatomy was derived from histological and electron microscopy (EM) observations. With the advent of molecular tools, however, much has been learned regarding the molecular makeup of each tissue, as well as how these tissues are both maintained and regenerated by neoblasts. This molecular characterization has been greatly accelerated by recent single-cell transcriptional profiling of the major tissues (***175***), and, as described in chapter 2, transcriptomes have now been generated for all cell types in the animal.

The planarian excretory system is made up of multiple branched structures, called protonephridia, that are spread throughout the animal and function in osmoregulation

and waste excretion (***172***, ***184***, ***185***). Protonephridia consist of ciliated flame cells that cap a tubule structure heavily enriched in the expression of genes encoding solute carrier proteins that terminates in the dorsal epithelium in a transcriptionally distinct region called the collecting duct (***172***, ***184***, ***185***). Beating of cilia in the flame cells and tubule cells generates negative pressure that induces filtration from the extracellular space into membrane fenestrations of the flame cells (***186***, ***187***). The substrate specificity of solute carriers varies spatially across the tubule and collecting duct, matching fairly well the spatial distribution of solute carriers in vertebrate nephron tubules (***185***). A set of transcription factors necessary for generating protonephridia has been identified, including *Six1/2-2*, *POU2/3*, *hunchback*, *Eya*, *Sall*, and *Osr*, many of which are evolutionarily conserved and are also required for vertebrate kidney development (***172***). *POU2/3* and *Six1/2-2* expression can be detected by FISH in *smedwi-1*$^{+}$ neoblasts, and inhibition of their expression by RNAi blocks both protonephridia progenitor formation and the formation of tubule cells. This block leads to the formation of blisters, bloating, and lysis (***172***), phenotypes that also arise following inhibition of planarian homologs to vertebrate genes essential for nephron function (***185***), suggesting a role for these neoblasts as specialized progenitors for the protonephridia.

The planarian nervous system consists of two ventrally localized cephalic ganglia and nerve cords, each of which is composed of a cortex of neuronal cell bodies surrounding a neurite-filled neuropil (***188***). Non-neuronal glial cells reside within the neuropils and display enriched expression of genes conserved in vertebrate glia that are involved in neurotransmitter reuptake and metabolism (***189***, ***190***). Cephalic ganglia are connected by an anterior commissure, and nerve cords are connected by many transverse commissures (***188***). An extensive peripheral nervous system is present, consisting of subepidermal, submuscular, gastrodermal, and pharyngeal nerve plexuses, as well as many neuronal projections throughout the parenchyma (***191***). Planarians contain a wide diversity of neuronal populations, including GABAergic (***192***), cholinergic (***193***), serotonergic (***194***), dopaminergic (***195***), and octopaminergic neurons (***196***), among

others. The brain itself displays transcriptional compartmentalization, with distinct medial/ventral, outer cortex, and lateral brain branch regions (***197***, ***198***). Planarians also possess two light-receptive eyes, containing photoreceptor neurons and pigment cup cells (***170***, ***171***), two lateral anterior regions called auricles that are thought to be responsible for chemosensation (***199***), and a number of additional putative sensory neurons, including *cintillo*$^{+}$ neurons (***200***) and *pkd1L-2*$^{+}$ neurons (***201***), among others. The planarian nervous system is highly complex, and much diversity likely remains to be uncovered. As just one example, a FISH screen of genes encoding planarian peptide hormones reveals a remarkably diverse array of unique expression patterns (***202***). Finally, a number of genes encoding transcription factors are specifically enriched in many of the known neuronal cell types. These include *ovo* for photoreceptor neurons of the eye (***171),*** *pitx* and *lhx1/5-1* for serotonergic neurons (***203***, ***204***), and *klf* for *cintillo*$^{+}$ sensory neurons (***205***), among many others. Each of these transcription factors are also expressed in *smedwi-1*$^{+}$ neoblasts and their inhibition leads to the ablation of the associated differentiated neuron during regeneration or tissue turnover, suggesting a role for these neoblasts as specialized progenitors for these neuron populations (***171, 203***, ***205***). It is currently unknown whether all neurons possess their own unique specialized neoblast, or whether hierarchies of differentiation are present.

The planarian intestine is responsible for the digestion of ingested food and consists of one anterior primary branch that splits into two posterior primary branches, with secondary, tertiary, and quaternary branches forming off of the primary branches (***206***). Based on histological data, the intestine was thought to consist of two cell types, absorptive phagocytic enterocytes and secretory goblet cells (***207***, ***208***). A transcriptionally distinct “outer” intestinal cell population also exists, as described in chapter 2. Molecular markers exist for both histologically identified cell populations. Phagocytic enterocytes were transcriptionally profiled by feeding animals iron beads and magnetically isolating cells that had phagocytosed the beads, identifying a number of enriched genes (***209***). Furthermore, lens culinaris agglutinin (LCA) stains an intestinal cell population hypothesized to be goblet cells, based on their intestinal localization and

their highly vacuolar appearance (***210***). A number of transcription factors are enriched in the intestine, including *gata4/5/6-1* (***169***), *hnf-4* (***169***), and *nkx2.2* (***209***). Single-cell multiplexed qPCR analysis on 4C cells revealed a distinct cluster of cells, termed gamma-neoblasts, that were enriched for expression of all three transcription factors, as well as an additional transcription factor *prox-1*, suggesting cells of this cluster are intestinal progenitors (***174***). Inhibition of *gata4/5/6-1* (***211***, ***212***) and *nkx2.2* (***209***), lead to breakdown of the intestine, providing further confirmation of this role. Interestingly, GATA and HNF factors are required for visceral endoderm differentiation in mice (***213***).

Planarians contain an extensive network of mononuclear muscle fibers (***214***). Planarian muscle cells express canonical actomyosin contractility genes (***214***, ***215***) but also express a number of collagens and other extracellular matrix proteins, similar to vertebrate fibroblasts (***216***). Multiple spatially and transcriptionally distinct populations of muscle are present in the body, including body-wall muscle (BWM), intestinal muscle, pharyngeal muscle, and dorsal-ventral muscle (DVM), which connects the dorsal and ventral surfaces of the animal (***214***, ***217***, ***218***). Subepidermal BWM consists of three separate layers, each with distinct orientations (***214***). These include the outermost circular layer that runs along the ML axis of the animal, a layer just below that contains a network of diagonal fibers and thin longitudinal fibers that run along the AP axis, and finally an innermost layer of thick longitudinal fibers (***214***). Distinct BWM layers are enriched for distinct transcription factors that can be used to specifically ablate a layer, including *myoD* for longitudinal fibers and *nkx1-1* for circular fibers (***217***). Interestingly, ablation of distinct muscle layers result in distinct phenotypes, with *myoD* inhibition leading to a block in regeneration and *nxk1-1* inhibition leading to defects in ML patterning (***217***). Similarly, intestinal muscle can be ablated through inhibition of the transcription factor *gata456-3*, leading to defects in intestine morphology, and DV muscle can be ablated through inhibition of the transcription factors *gata456-2* and *nk4*, leading to medial-lateral patterning defects (***218***). Intestinal muscle, DV muscle, and pharyngeal muscle can also be ablated through inhibition of the transcription factor *foxF-1*, which will be discussed further in chapters 2 and 3 regarding its additional role

in specifying a novel class of phagocytic cells (***218***). Each of the transcription factors discussed above were also expressed in *smedwi-1*$^{+}$ neoblasts, representing specialized neoblast progenitors for each of these muscle populations (***217***, ***218***).

The space between the major organs of the planarian is filled with a mesenchyme called the parenchyma (***219***). Neoblasts reside within the parenchyma, as do many of the tissue types described thus far, including muscle, protonephridia, and neurons, as well as a variety of gland cells and a poorly characterized cell type termed fixed parenchymal cells (***219***, ***220***, ***221***). Based on histochemical studies, gland cells were divided into two categories, acidophilic (eosinophilic) and cyanophilic (basophilic) (***219***, ***220***). Although gland cells have generally been poorly characterized at the molecular level, multiple markers for a population of marginal adhesive gland cells have been identified (***210***). Fixed parenchymal cells, also known as reticulocytes, were described as large cells with long processes that contained lysosomes, numerous glycogen granules, and lipid droplets, suggesting a potential phagocytic role for these cells (***221***). This capacity was further confirmed through the demonstration by EM that fixed parenchymal cells can phagocytose heat killed bacteria (***222***, ***223***). While molecular characterization of both gland cells and fixed parenchymal cells has been largely lacking, a heterogeneous cluster of cells termed “parapharyngeal” was identified from recent single-cell sequencing data, and was heavily enriched for markers with parenchymal localization patterns (***175***). Transcriptomes were also generated for a variety of putative gland cells, as well as a population of cells resembling fixed parenchymal cells, using a large-scale single-cell RNA sequencing approach, as detailed in chapter 2. Finally, planarians contain a population of pigment cells that lie between circular and longitudinal BWM and that produce their color through a mixture of ommochrome and porphyrin pigments (***224***, ***225***). Prolonged light exposure can be used to ablate planarian pigment cells through porphyrin-dependent photosensitization (***224***). Furthermore, expression of the transcription factors *albino* (***225***), *foxF-1,* and *ets-1* (***226***) is enriched in pigment cells and pigment cell ablation occurs upon their inhibition

(***225***, ***226***). Expression of each transcription factor can be found in *smedwi-1*$^{+}$ neoblasts, representing specialized progenitors for planarian pigment cells (***225***, ***226***).

The pharynx is a cylindrical structure contained within an epithelial-lined cavity, referred to as the pharyngeal pouch, that can protrude from a ventral opening in the animal, called the mouth, for feeding and defecation (***219***). The pharynx itself is made up of an epithelial layer surrounding muscle, neurons, and gland cells (***227***). A number of transcriptionally distinct domains exist within the pharynx epithelium, including the esophagus, which connects the pharynx to the intestine (***198***, ***228***), as well as a number of novel domains described in chapter 2. Because neoblasts are excluded from the pharynx (***181***), progenitors must be specified outside of the pharynx and migrate, differentiate, and incorporate into the pharynx. The *foxA* transcription factor-encoding gene is expressed in both mature and regenerating pharynges (***205***, ***229***, ***230***), as well as in *smedwi-1*$^{+}$ neoblasts surrounding the pharynx (***205***, ***229***). Inhibition of *foxA* blocks pharynx regeneration, suggesting these *smedwi-1*$^{+}$*/ foxA*$^{+}$ neoblasts are specialized neoblast progenitors for the pharynx (***205***, ***229***). The transcript dd_554 (*SmedASXL_059179*) marks an irradiation sensitive population of cells within the pharynx and just anterior to the pharynx, some of which are *smedwi-1*$^{+}$ (***180***). Following irradiation, expression of dd_554 is lost following the loss of *smedwi-1*$^{+}$ expression but prior to the loss of expression of mature pharyngeal cell markers, suggesting this transcript marks a transition state during pharynx cell differentiation (***180***).

The planarian epidermis that surrounds the animal consists of a monostratified layer of ciliated, cuboidal cells on the ventral side of the animal, and largely unciliated, more columnal cells on the dorsal side (***231***). A transcriptionally distinct differentiated cell population at the DV margin has also been identified (***179***, ***232***). From the single-cell multiplexed qPCR analysis on 4C cells that revealed gamma-neoblast intestinal progenitors, as mentioned above, an additional cluster of cells, called zeta-neoblasts, was also identified (***174***). This cluster was enriched for a number of genes, including the transcription factor *zfp-1*, inhibition of which by RNAi led to the loss of all epidermal

progenitors and differentiated epidermal cells, suggesting cells of this cluster are epidermal progenitors (***174***). In addition to *zfp-1* expressing specialized neoblasts, a number of postmitotic transition states for the epidermis have also been identified (***177***, ***178***, ***179***). Neoblasts reside in the parenchyma. Therefore, epidermal progenitors must migrate from the parenchyma, where they are generated, to the epidermis (***177***, ***178***, ***179***). In doing so, they progress through a series of roughly three distinct transcriptional states in a spatiotemporal fashion before fully differentiating into mature epidermis (***177***, ***178***, ***179***). Much remains to be learned of this spatiotemporal progression, however, as inhibition of multiple transcription factors, including *myb-1*, *soxP-3*, and *pax2/5/8*, lead to the abolition of early transcriptional states but do not affect later transcriptional stages or epidermal homeostasis in general (***233***, ***234***).

Planarians constitutively express patterning information in muscle

Following major injury, it is essential that planarians correctly replace the appropriate cell types that are lost. Planarians constitutively express regional gradients of components of known developmental signaling pathways, including Fgf, Wnt, and Bmp across the different body axes (Figure 1.6A). Interestingly, these regionally expressed genes are largely restricted in expression to muscle (***235***). Importantly, these regional gradients are rescaled following a loss of tissue to match the new size of the fragment (***175***, ***198***, ***217***, ***228***, ***235***, ***236***, ***237***, ***238***, ***239***, ***240***, ***241***, ***242***, ***243***, ***244***, ***245***). As reviewed below, inhibition of many of these regionally expressed genes by RNAi leads to defects in the regeneration and homeostatic maintenance of a proper body plan, highlighting their functional patterning role.

The Wnt signaling pathway is essential for the regeneration and maintenance of the AP axis, a role that is widely conserved across the animal kingdom (***246***). In general, Wnt pathway components, such as *wntP-2* and *wnt1*, are expressed in the posterior of the animal (***198***), whereas Wnt inhibitors, such as *notum* and *sFRP-1*, are expressed in the anterior of the animal (***198***, ***238***, ***247***) (Figure 1.6A). Nuclear localization of the Wnt effector protein Beta-catenin-1 also displays a posterior bias (***248***). Inhibition of *beta-*

catenin-1 expression by RNAi leads to the formation of two heads following a transverse amputation (***198***, ***249***, ***250***), as does inhibition of other Wnt signaling components, such as *wnt1* (***237***, ***251***) and *wntless* (***251***) (Figure 1.6B). Even in uninjured animals, RNAi of *beta-catenin-1* results in the formation of heads all along the periphery of the animal (***198***, ***249***, ***250***). Conversely, inhibition of negative regulators of the Wnt signaling pathway, including *notum* (***247***) and *APC* (***249***) by RNAi leads to the formation of two tails following a transverse amputation (Figure 1.6B). A role for FGF pathway modulators in AP patterning has also been demonstrated, with a number of kinase dead FGFR-like genes being expressed in gradients along the AP axis (***252***) (Figure 1.6A). RNAi of the FGFR-like gene *nou-darake* and the Wnt signaling components *fz5/8-4* and *wntA* leads to posterior expansion of the brain and formation of ectopic posterior eyes in both regenerating and uninjured animals (***252***, ***253***). *nou-darake* expression is confined to the head and is juxtaposed with *wntA* expression extending from the tail to just below the brain (***253***) (Figure 1.6C). Inhibition of another FGFR-like gene, *ndl-3*, and the Wnt gene *wntP-2* leads to posterior expansion of trunk structures, including the ectopic formation of mouths and pharynges (***244***, ***252***). *ndl-3* is expressed in the prepharyngeal region just above the trunk and is juxtaposed with *wntP-2* expression extending from the tail to just below the *ndl-3* domain (***244***, ***252***) (Figure 1.6C). Together, these results implicate these two gene circuits, composed of an anterior FGFRL domain and a posterior Wnt domain, in controlling head and trunk patterning along the AP axis (***252***).

As previously noted, inhibition of the anteriorly expressed Wnt inhibitor *notum* results in the formation of two tails following a transverse amputation. Expression of *notum* is largely restricted to a small cluster of muscle cells localized at the very anterior tip of the animal, referred to as the anterior pole (***247***) (Figure 1.6D), as well as a small population of neuronal cells in the brain (***254***). Multiple transcription factors are enriched in the anterior pole, including *foxD, zic-1*, *islet1*, and *pitx,* and their inhibition by RNAi leads to the ablation of the structure and consequent head patterning defects (***203***, ***204***, ***205***, ***239***, ***241***, ***255***). The posterior pole is a similar, though more diffuse, structure in the posterior of the animal that expresses a number of genes, including *wnt1* (***198***). A

number of transcription factors are also expressed in the posterior pole, including *islet1* and *pitx*, with their inhibition by RNAi leading to loss of these structures and consequent tail patterning defects (***203***, ***204***, ***239***). Through transplantation of the anterior pole region from one animal onto the posterior region of another animal, outgrowths, containing eyes and proper expression of anterior patterning genes, are generated, suggesting the anterior pole acts as an anterior organizer structure (***256***). Similarly, expansion of the *wnt1* signaling center leads to tail expansion and defects in animal proportions (***257***).

Whereas components of the Wnt and Fgf signaling pathways are important for regeneration and maintenance of the AP axis of the animal, components of the Bmp signaling pathway are important for regeneration and maintenance of the DV axis. *bmp4*, which encodes a Bmp-signaling ligand, is expressed dorsally, and its inhibition by RNAi leads to ventralization of the animal, with ventral structures, such as the nerve cords, appearing dorsally (***228***, ***236***, ***258***) (Figure 1.6A). Inhibition of genes encoding other Bmp pathway components, including *smad1*, *smad4*, and *tolloid*, also results in defects in the DV axis (***228***, ***236***). *admp* is expressed ventrally and is important for maintaining dorsal-specific *bmp* expression, with its inhibition leading to *bmp4* RNAi sensitization (***259***) (Figure 1.6A).

Planarian ML patterning is largely controlled by the genes *slit* and *wnt5*. *slit* is expressed medially in the animal, and its inhibition by RNAi leads to a collapse of tissues along the midline, including the brain lobes, eyes, and intestine (***238***, ***260***) (Figure 1.6A). Conversely, *wnt5* is expressed laterally in the animal, and its inhibition by RNAi leads to the lateral expansion of some tissues, such as the brain and nerve cords, as well as the lateral formation of ectopic eyes and pharynges (***238***) (Figure 1.6A). In addition to their roles in DV patterning, Bmp pathway components also play roles in regeneration along the ML axis of the animal. Inhibition of *bmp*, *admp*, *smad*, or *tolloid* all impair lateral regeneration following sagittal amputation and lead to midline indentations (***228***, ***236***, ***259***).

Given that inhibition of regionally expressed patterning molecules by RNAi leads to defects in the production of spatially appropriate cell types and tissues, it is likely that neoblasts can read out this positional system in some manner. Indeed, there are multiple lines of evidence that suggest that regionally expressed genes in muscle do affect neoblast fate decisions. Specialized neoblast progenitors associated with regionally localized organs are often regionally localized themselves. *ovo*$^{+}$ eye progenitors are found only in the region anterior to the pharynx (***171***) and *foxA*$^{+}$ pharynx progenitors are found only in the trunk region surrounding the pharynx (***205***, ***229***). Furthermore, inhibition of *wnt1* by RNAi leads to the presence of ectopic *ovo*$^{+}$/*smedwi-1*$^{+}$ eye progenitors during posterior regeneration (***235***). It has also been demonstrated that zeta-neoblast epidermal progenitors express different genes depending on their location along the DV axis, with dorsal zeta-neoblasts expressing *PRDM-1* and ventral zeta-neoblasts expressing *kal-1* (***179***). This development of a dorsal identity (i.e. *PRDM-1* expression) by zeta-neoblasts requires the expression of the dorsal-specifying factor *bmp-4* (***179***).

Together, these results suggest a model for how properly patterned tissues are both maintained and regenerated. During homeostasis, neoblasts read out signals from muscle, which are arranged in regional gradients along the body axes of the animal. These signals govern neoblast fate choices, ensuring appropriate cell types are made for the appropriate body regions to maintain the proper body plan. Following an injury, patterning information rescales to match the size of the new fragment, triggering neoblasts fate choices to shift and generate cell types appropriate for this new positional information. Existing cells in the wrong position are no longer maintained, leading to the gradual formation of a correctly proportioned animal. More work is needed to identify the signals read out by neoblasts to provide direct evidence for this model.

Figure 1.4

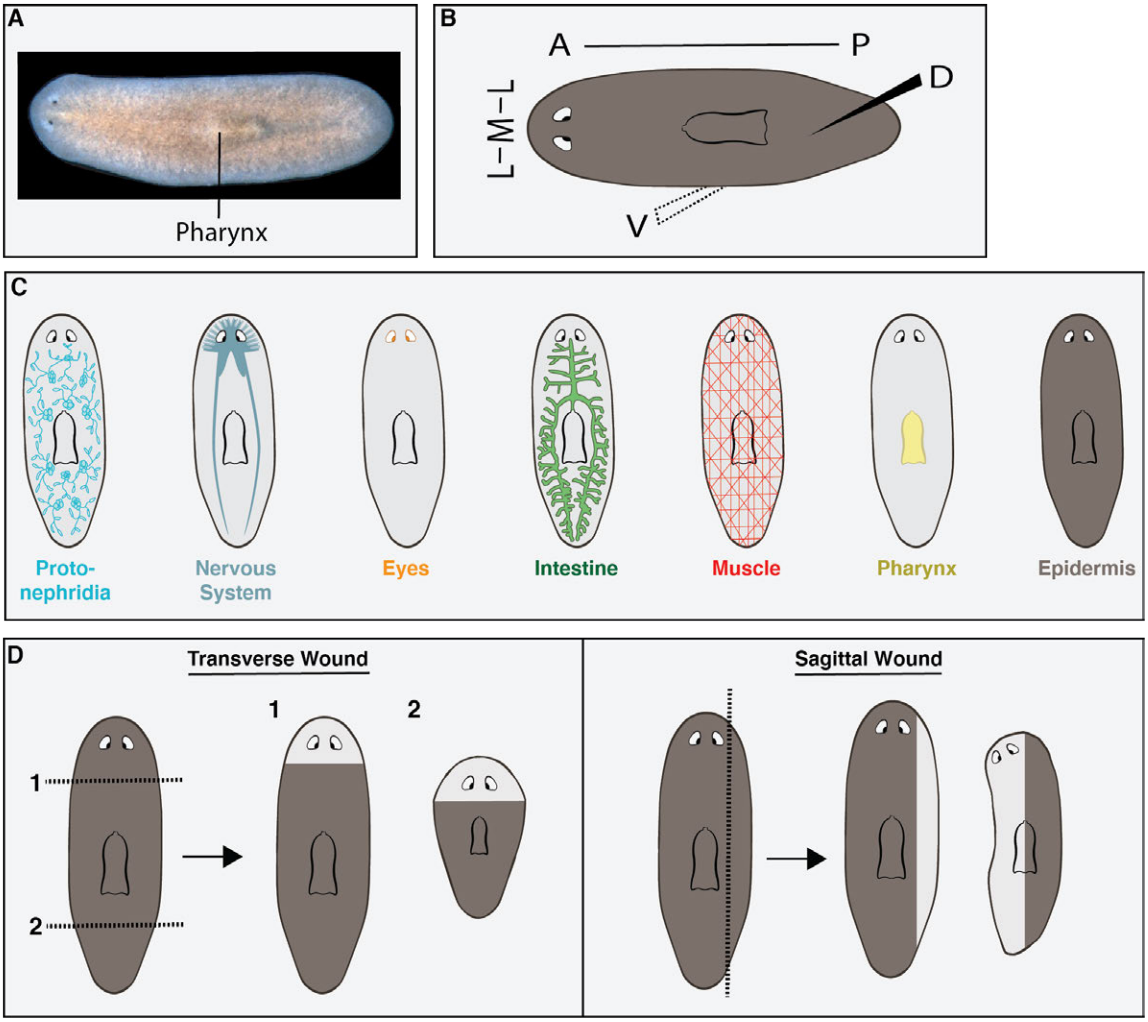

A
Pharynx
B
A
P
D
L–M–L
V
C
Proto-
nephridia
Nervous
System
Eyes
Intestine
Muscle
Pharynx
Epidermis
D
Transverse Wound
1
2
1
2
Sagittal Wound

Figure 1.4. An introduction to planarians.

(A) Live image of an adult asexual *Schmidtea mediterranea*. **(B)** Diagram illustrating the Anterior-Posterior (A-P), Dorsal-Ventral (D-V), and Medial-Lateral (M-L) axes of the animal. **(C)** Figure adapted from (***205***). Cartoons depicting a sampling of the diverse tissues that make up the planarian. **(D)** Panel adapted from (***261***). Illustration of the regenerative response to two transverse amputations (left) and a parasagittal amputation (right). Grey regions indicate the blastema.

Figure 1.5

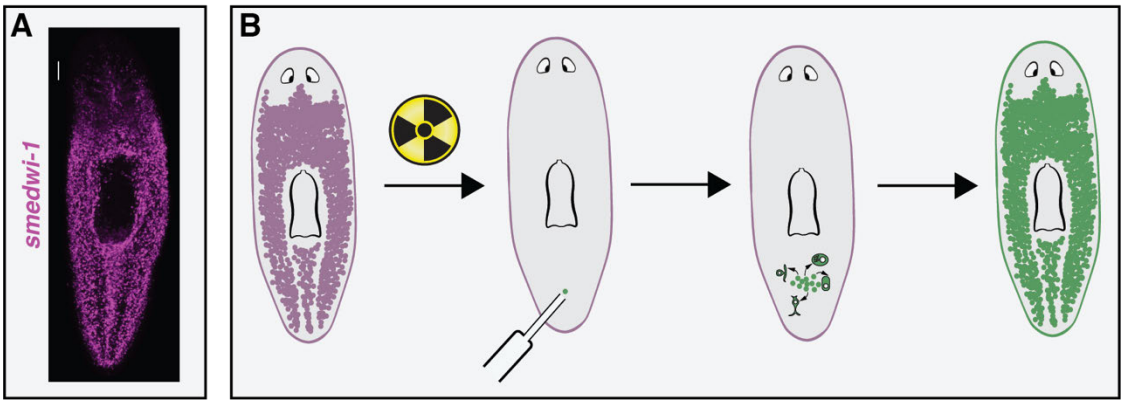


Figure 1.5. Neoblasts are the source for all new tissue.

(A) FISH image of the pan-neoblast marker *smedwi-1*. Neoblasts are located in the parenchyma and are excluded from the brain, intestine, and pharynx. **(B)** Illustration of the experiment used to determine that individual neoblasts can be pluripotent. An animal from one strain is irradiated, killing all neoblasts. A single neoblast isolated from another strain is then injected into the irradiated host. The neoblast divides and expands, generating both more neoblasts and diverse differentiated cells. Over time, the single neoblast will repopulate the neoblast population in the animal, as well as turn over all differentiated tissues, essentially converting the host animal to the donor strain.

Figure 1.6

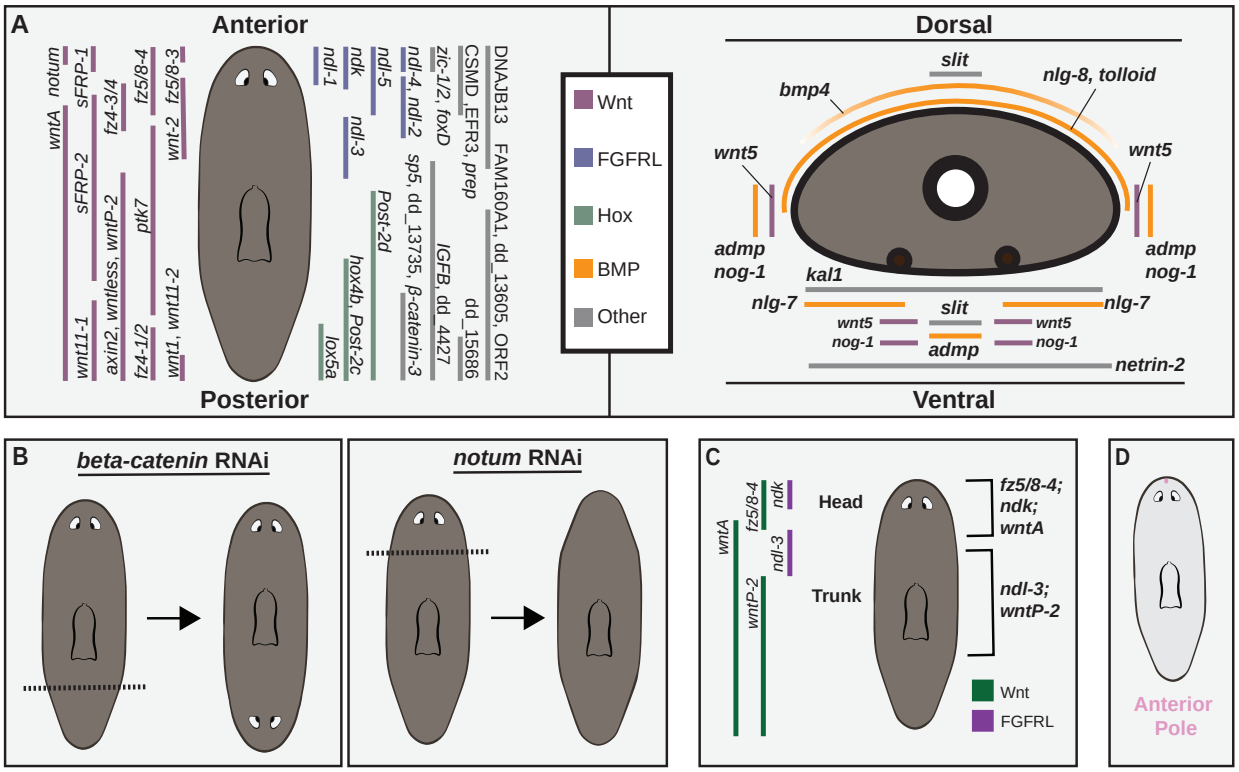

A
Anterior
Posterior
notum
sFRP-1
wntA
sFRP-2
fz4-3/4
fz5/8-4
fz5/8-3
wnt-2
ptk7
axin2, wntless, wntP-2
wnt11-1
fz4-1/2
wnt1, wnt11-2
ndl-1
ndk
ndl-3
ndl-5
Post-2d
hox4b, Post-2c
lox5a
ndl-4, ndl-2
sp5, dd_13735, β-catenin-3
zic-1/2, foxD
IGFB, dd_4427
CSMD_EFR3, prep
dd_15686
DNAJB13
FAM160A1, dd_13605, ORF2
Wnt
FGFRL
Hox
BMP
Other
Dorsal
Ventral
slit
bmp4
nlg-8, tolloid
wnt5
admp
nog-1
kal1
nlg-7
netrin-2
B
beta-catenin RNAi
notum RNAi
C
wntA
fz5/8-4
ndk
ndl-3
wntP-2
Head
Trunk
fz5/8-4; ndk; wntA
ndl-3; wntP-2
Wnt
FGFRL
D
Anterior Pole

Figure 1.6. Planarians constitutively express patterning information.

(A) Diagrams depicting the expression domains of selected regionally expressed genes in muscle (Left image modified from (***252***) and right image modified from (***262***)). Left: Dorsal view of the planarian is depicted. Right: Transverse section is depicted, with a centrally located gut branch and two ventrally biased ventral nerve cords. Genes marked in green are Wnt signaling components, genes marked in purple are FGFRL signaling components, genes marked in green are Hox genes, genes marked in orange are Bmp signaling components, and genes marked in grey represent all other classes of genes. **(B)** Illustration of the regenerative outcomes following *beta-catenin* (left) and *notum* (right) inhibition by RNAi. *beta-catenin* RNAi animals will form a head in place of a tail, and *notum* RNAi animals will form a tail in place of a head **(C)** Diagram depicting two FGFRL-Wnt circuits responsible for patterning the head and the trunk. Expression domains of the genes involved are shown to the left. Genes marked in green are Wnt signaling components and genes marked in purple are FGFRL signaling components. Note the juxtaposed domains of the trunk circuit components *ndl-3* and *wntP-2* and the head circuit components *ndk* and *wntA*. **(D)** Cartoon depicting the *notum*$^{+}$ anterior pole.

III. Content Overview

The advent of high throughput and affordable single-cell RNA sequencing approaches has enabled the transcriptional profiling of tens of thousands of single cells, raising the possibility that all cell types and cell states of a complete animal could be determined. However, this task remains daunting given the vast cellular complexity of most animals, including a fluctuating cell type composition across development. Chapter 2 of this thesis uses the single-cell RNA sequencing technique Drop-seq to generate such a whole-animal cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. Animals were divided into five body sections to enrich for rare, regionally localized cell types, with the presence of known rare cell types in the data used to guide the number of cells sequenced from each fragment. Through this iterative approach, transcriptomes could be determined for most-to-all cell types of the complete animal. Because planarians possess pluripotent neoblasts that constantly turn over all differentiated cell types, transcriptomes were determined for pluripotent stem cells, a diverse set of differentiated cells, and potentially all transition state populations from stem cell to differentiated cell for all cell types. A number of novel neoblast subclasses, transition states, and differentiated cell types were identified, including a previously undescribed class of phagocytic cell types. Transcription factors enriched in many of these cell populations were identified and allowed for ablation of those cells. Finally, because planarians constitutively express patterning information in their muscle, novel regionally expressed genes in muscle were also identified.

References

1. R. D. Fleischmann *et al.*, Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496-512 (1995).
2. A. Goffeau *et al.*, Life with 6000 genes. *Science* **274**, 546, 563-547 (1996).
3. F. R. Blattner *et al.*, The complete genome sequence of Escherichia coli K-12. *Science* **277**, 1453-1462 (1997).
4. C. J. Bult *et al.*, Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. *Science* **273**, 1058-1073 (1996).
5. S. T. Cole *et al.*, Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* **393**, 537-544 (1998).
6. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-2018 (1998).
7. M. D. Adams *et al.*, The genome sequence of Drosophila melanogaster. *Science* **287**, 2185-2195 (2000).
8. E. S. Lander *et al.*, Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
9. J. C. Venter *et al.*, The sequence of the human genome. *Science* **291**, 1304-1351 (2001).
10. S. D. Putney, W. C. Herlihy, P. Schimmel, A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* **302**, 718-721 (1983).
11. M. D. Adams *et al.*, Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651-1656 (1991).
12. V. E. Velculescu, L. Zhang, B. Vogelstein, K. W. Kinzler, Serial analysis of gene expression. *Science* **270**, 484-487 (1995).
13. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-470 (1995).
14. M. N. Bainbridge *et al.*, Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7**, 246 (2006).
15. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-628 (2008).

16. M. Sultan *et al.*, A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 956-960 (2008).
17. B. T. Wilhelm *et al.*, Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239-1243 (2008).
18. L. Luo *et al.*, Gene expression profiles of laser-captured adjacent neuronal subtypes. *Nat Med* **5**, 117-122 (1999).
19. M. K. Chiang, D. A. Melton, Single-cell transcript analysis of pancreas development. *Dev Cell* **4**, 383-393 (2003).
20. B. Lambolez, E. Audinat, P. Bochet, F. Crepel, J. Rossier, AMPA receptor subunits expressed by single Purkinje cells. *Neuron* **9**, 247-258 (1992).
21. J. Liu, C. Hansen, S. R. Quake, Solving the "world-to-chip" interface problem with a microfluidic matrix. *Anal Chem* **75**, 4718-4723 (2003).
22. F. Tang *et al.*, mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377-382 (2009).
23. D. Ramskold *et al.*, Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat Biotechnol* **30**, 777-782 (2012).
24. S. Picelli *et al.*, Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* **10**, 1096-1098 (2013).
25. S. Picelli *et al.*, Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9**, 171-181 (2014).
26. Y. Sasagawa *et al.*, Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol* **14**, R31 (2013).
27. S. Picelli, Single-cell RNA-sequencing: The future of genome biology is now. *RNA Biol* **14**, 637-650 (2017).
28. X. Fan *et al.*, Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol* **16**, 148 (2015).
29. K. Sheng, W. Cao, Y. Niu, Q. Deng, C. Zong, Effective detection of variation in single-cell transcriptomes using MATQ-seq. *Nat Methods* **14**, 267-270 (2017).
30. D. A. Jaitin *et al.*, Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776-779 (2014).
31. T. Hashimshony, F. Wagner, N. Sher, I. Yanai, CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* **2**, 666-673 (2012).

32. T. Hashimshony *et al.*, CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol* **17**, 77 (2016).
33. S. Islam *et al.*, Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res* **21**, 1160-1167 (2011).
34. S. Islam *et al.*, Highly multiplexed and strand-specific single-cell RNA 5' end sequencing. *Nat Protoc* **7**, 813-828 (2012).
35. S. Islam *et al.*, Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* **11**, 163-166 (2014).
36. Y. Sasagawa *et al.*, Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol* **19**, 29 (2018).
37. E. Z. Macosko *et al.*, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
38. A. M. Klein *et al.*, Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201 (2015).
39. G. X. Zheng *et al.*, Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017).
40. X. Zhang *et al.*, Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Mol Cell* **73**, 130-142.e135 (2019).
41. N. Habib *et al.*, Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat Methods* **14**, 955-958 (2017).
42. H. C. Fan, G. K. Fu, S. P. Fodor, Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. *Science* **347**, 1258367 (2015).
43. T. M. Gierahn *et al.*, Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat Methods* **14**, 395-398 (2017).
44. X. Han *et al.*, Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107.e1017 (2018).
45. J. Cao *et al.*, Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661-667 (2017).
46. A. B. Rosenberg *et al.*, Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176-182 (2018).
47. M. D. Luecken, F. J. Theis, Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**, e8746 (2019).

48. A. Butler, P. Hoffman, P. Smibert, E. Papalexi, R. Satija, Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* **36**, 411-420 (2018).
49. P. Brennecke *et al.*, Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* **10**, 1093-1095 (2013).
50. K. Pearson, LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 559-572 (1901).
51. R. R. Coifman *et al.*, Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the national academy of sciences* **102**, 7426-7431 (2005).
52. L. v. d. Maaten, G. Hinton, Visualizing data using t-SNE. *Journal of machine learning research* **9**, 2579-2605 (2008).
53. L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, (2018).
54. S. Lloyd, Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28**, 129-137 (1982).
55. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502 (2015).
56. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386 (2014).
57. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**, 979-982 (2017).
58. J. Shin *et al.*, Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. *Cell Stem Cell* **17**, 360-372 (2015).
59. Z. Ji, H. Ji, TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res* **44**, e117 (2016).
60. M. Guo, E. L. Bao, M. Wagner, J. A. Whitsett, Y. Xu, SLICE: determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Res* **45**, e54 (2017).
61. K. Street *et al.*, Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018).

62. E. Marco *et al.*, Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proc Natl Acad Sci U S A* **111**, E5643-5650 (2014).
63. J. Cao *et al.*, The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502 (2019).
64. S. C. Bendall *et al.*, Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157**, 714-725 (2014).
65. M. Setty *et al.*, Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol* **34**, 637-645 (2016).
66. F. A. Wolf *et al.*, PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol* **20**, 59 (2019).
67. C. Weinreb, S. Wolock, B. K. Tusi, M. Socolovsky, A. M. Klein, Fundamental limits on dynamic inference from single-cell snapshots. *Proc Natl Acad Sci U S A* **115**, E2467-e2476 (2018).
68. J. A. Farrell *et al.*, Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, (2018).
69. D. E. Wagner *et al.*, Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981-987 (2018).
70. J. A. Briggs *et al.*, The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, (2018).
71. S. Siebert *et al.*, Stem cell differentiation trajectories in Hydra resolved at single-cell resolution. *Science* **365**, (2019).
72. D. Grun *et al.*, De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell* **19**, 266-277 (2016).
73. J. Chen, A. Schlitzer, S. Chakarov, F. Ginhoux, M. Poidinger, Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. *Nat Commun* **7**, 11988 (2016).
74. G. La Manno *et al.*, RNA velocity of single cells. *Nature* **560**, 494-498 (2018).
75. K. L. Frieda *et al.*, Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107-111 (2017).
76. B. Raj *et al.*, Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nat Biotechnol* **36**, 442-450 (2018).
77. B. Spanjaard *et al.*, Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nat Biotechnol* **36**, 469-473 (2018).

78. A. Alemany, M. Florescu, C. S. Baron, J. Peterson-Maduro, A. van Oudenaarden, Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108-112 (2018).
79. R. Kalhor *et al.*, Developmental barcoding of whole mouse via homing CRISPR. *Science* **361**, (2018).
80. M. M. Chan *et al.*, Molecular recording of mammalian embryogenesis. *Nature* **570**, 77-82 (2019).
81. C. Weinreb, A. Rodriguez-Fraticelli, F. D. Camargo, A. M. Klein, Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, (2020).
82. L. S. Ludwig *et al.*, Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* **176**, 1325-1339.e1322 (2019).
83. J. Xu *et al.*, Single-cell lineage tracing by endogenous mutations enriched in transposase accessible mitochondrial DNA. *Elife* **8**, (2019).
84. K. Achim *et al.*, High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat Biotechnol* **33**, 503-509 (2015).
85. N. Karaiskos *et al.*, The Drosophila embryo at single-cell transcriptome resolution. *Science* **358**, 194-199 (2017).
86. M. Nitzan, N. Karaiskos, N. Friedman, N. Rajewsky, Gene expression cartography. *Nature* **576**, 132-137 (2019).
87. P. L. Stahl *et al.*, Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78-82 (2016).
88. S. G. Rodriques *et al.*, Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463-1467 (2019).
89. S. Vickovic *et al.*, High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* **16**, 987-990 (2019).
90. E. Lubeck, A. F. Coskun, T. Zhiyentayev, M. Ahmad, L. Cai, Single-cell in situ RNA profiling by sequential hybridization. *Nat Methods* **11**, 360-361 (2014).
91. 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).
92. C. L. Eng *et al.*, Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235-239 (2019).
93. K. H. Chen, A. N. Boettiger, J. R. Moffitt, S. Wang, X. Zhuang, RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).

94. J. R. Moffitt *et al.*, Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362**, (2018).
95. S. Codeluppi *et al.*, Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat Methods* **15**, 932-935 (2018).
96. R. Ke *et al.*, In situ sequencing for RNA analysis in preserved tissue and cells. *Nat Methods* **10**, 857-860 (2013).
97. J. H. Lee *et al.*, Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360-1363 (2014).
98. X. Wang *et al.*, Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, (2018).
99. N. Navin *et al.*, Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-94 (2011).
100. M. Gundry, W. Li, S. B. Maqbool, J. Vijg, Direct, genome-wide assessment of DNA mutations in single cells. *Nucleic Acids Res* **40**, 2032-2040 (2012).
101. Y. Wang *et al.*, Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155-160 (2014).
102. S. A. Vitak *et al.*, Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat Methods* **14**, 302-308 (2017).
103. D. A. Cusanovich *et al.*, Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910-914 (2015).
104. S. A. Smallwood *et al.*, Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* **11**, 817-820 (2014).
105. M. Farlik *et al.*, Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell Rep* **10**, 1386-1397 (2015).
106. S. J. Clark *et al.*, Genome-wide base-resolution mapping of DNA methylation in single cells using single-cell bisulfite sequencing (scBS-seq). *Nat Protoc* **12**, 534-547 (2017).
107. C. Luo *et al.*, Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600-604 (2017).
108. R. M. Mulqueen *et al.*, Highly scalable generation of DNA methylation profiles in single cells. *Nat Biotechnol* **36**, 428-431 (2018).
109. T. J. Stevens *et al.*, 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59-64 (2017).

110. L. Tan, D. Xing, C. H. Chang, H. Li, X. S. Xie, Three-dimensional genome structures of single diploid human cells. *Science* **361**, 924-928 (2018).
111. T. Nagano *et al.*, Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64 (2013).
112. I. M. Flyamer *et al.*, Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544**, 110-114 (2017).
113. V. Ramani *et al.*, Massively multiplex single-cell Hi-C. *Nat Methods* **14**, 263-266 (2017).
114. A. Rotem *et al.*, Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol* **33**, 1165-1172 (2015).
115. K. Grosselin *et al.*, High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat Genet* **51**, 1060-1066 (2019).
116. J. Cao *et al.*, Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380-1385 (2018).
117. C. Angermueller *et al.*, Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods* **13**, 229-232 (2016).
118. V. M. Peterson *et al.*, Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotechnol* **35**, 936-939 (2017).
119. M. Stoeckius *et al.*, Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**, 865-868 (2017).
120. S. Pott, Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells. *Elife* **6**, (2017).
121. A. Sebe-Pedros *et al.*, Early metazoan cell type diversity and the evolution of multicellular gene regulation. *Nat Ecol Evol* **2**, 1176-1188 (2018).
122. K. Achim *et al.*, Whole-Body Single-Cell Sequencing Reveals Transcriptional Domains in the Annelid Larval Body. *Mol Biol Evol* **35**, 1047-1062 (2018).
123. A. Sebe-Pedros *et al.*, Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq. *Cell* **173**, 1520-1534.e1520 (2018).
124. 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).
125. J. S. Packer *et al.*, A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution. *Science* **365**, (2019).
126. C. Cao *et al.*, Comprehensive single-cell transcriptome lineages of a proto-vertebrate. *Nature* **571**, 349-354 (2019).

127. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367-372 (2018).
128. O. Rozenblatt-Rosen, M. J. T. Stubbington, A. Regev, S. A. Teichmann, The Human Cell Atlas: from vision to reality. *Nature* **550**, 451-453 (2017).
129. D. T. Montoro *et al.*, A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* **560**, 319-324 (2018).
130. L. W. Plasschaert *et al.*, A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature* **560**, 377-381 (2018).
131. B. Pijuan-Sala *et al.*, A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566**, 490-495 (2019).
132. J. D. Buenrostro *et al.*, Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e1516 (2018).
133. A. Ayyaz *et al.*, Single-cell transcriptomes of the regenerating intestine reveal a revival stem cell. *Nature* **569**, 121-125 (2019).
134. M. A. Tosches *et al.*, Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles. *Science* **360**, 881-888 (2018).
135. A. Bhaduri *et al.*, Outer Radial Glia-like Cancer Stem Cells Contribute to Heterogeneity of Glioblastoma. *Cell Stem Cell* **26**, 48-63.e46 (2020).
136. F. Jacob *et al.*, A Patient-Derived Glioblastoma Organoid Model and Biobank Recapitulates Inter- and Intra-tumoral Heterogeneity. *Cell* **180**, 188-204.e122 (2020).
137. N. G. Skene *et al.*, Genetic identification of brain cell types underlying schizophrenia. *Nat Genet* **50**, 825-833 (2018).
138. S. R. Srivatsan *et al.*, Massively multiplex chemical transcriptomics at single-cell resolution. *Science* **367**, 45-51 (2020).
139. S. Xie, J. Duan, B. Li, P. Zhou, G. C. Hon, Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Mol Cell* **66**, 285-299.e285 (2017).
140. B. Adamson *et al.*, A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867-1882.e1821 (2016).
141. A. Dixit *et al.*, Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866.e1817 (2016).
142. D. A. Jaitin *et al.*, Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* **167**, 1883-1896.e1815 (2016).

143. P. Datlinger *et al.*, Pooled CRISPR screening with single-cell transcriptome readout. *Nat Methods* **14**, 297-301 (2017).
144. A. J. Rubin *et al.*, Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell* **176**, 361-376.e317 (2019).
145. J. L. McFaline-Figueroa *et al.*, A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition. *Nat Genet* **51**, 1389-1398 (2019).
146. R. Tian *et al.*, CRISPR Interference-Based Platform for Multimodal Genetic Screens in Human iPSC-Derived Neurons. *Neuron* **104**, 239-255.e212 (2019).
147. G. Chen, B. Ning, T. Shi, Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front Genet* **10**, 317 (2019).
148. C. Ziegenhain *et al.*, Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell* **65**, 631-643.e634 (2017).
149. L. Kester, A. van Oudenaarden, Single-Cell Transcriptomics Meets Lineage Tracing. *Cell Stem Cell* **23**, 166-179 (2018).
150. H. Randolph, Observations and experiments on regeneration in Planarians. *Archiv für Entwicklungsmechanik der Organismen* **5**, 352-372 (1897).
151. T. H. Morgan, Regeneration in Planarians. *Archiv für Mikroskopische Anatomie* **10**, 58-119 (1900).
152. C. E. Laumer *et al.*, Spiralian phylogeny informs the evolution of microscopic lineages. *Curr Biol* **25**, 2000-2006 (2015).
153. A. Sanchez Alvarado, P. A. Newmark, Double-stranded RNA specifically disrupts gene expression during planarian regeneration. *Proc Natl Acad Sci U S A* **96**, 5049-5054 (1999).
154. L. Rouhana *et al.*, RNA interference by feeding in vitro-synthesized double-stranded RNA to planarians: methodology and dynamics. *Dev Dyn* **242**, 718-730 (2013).
155. R. S. King, P. A. Newmark, In situ hybridization protocol for enhanced detection of gene expression in the planarian Schmidtea mediterranea. *BMC Dev Biol* **13**, 8 (2013).
156. H. Brandl *et al.*, PlanMine--a mineable resource of planarian biology and biodiversity. *Nucleic Acids Res* **44**, D764-773 (2016).
157. M. A. Grohme *et al.*, The genome of Schmidtea mediterranea and the evolution of core cellular mechanisms. *Nature* **554**, 56-61 (2018).

158. M. Vila-Farre, C. R. J, The Ecology of Freshwater Planarians. *Methods Mol Biol* **1774**, 173-205 (2018).
159. P. W. Reddien, A. Sanchez Alvarado, Fundamentals of planarian regeneration. *Annu Rev Cell Dev Biol* **20**, 725-757 (2004).
160. J. Baguna, E. Salo, C. Auladell, Regeneration and pattern formation in planarians. III. that neoblasts are totipotent stem cells and the cells. *Development* **107**, 77-86 (1989).
161. K. J. Pedersen, Cytological studies on the planarian neoblast. *Z Zellforsch Mikrosk Anat* **50**, 799-817 (1959).
162. M. Morita, J. B. Best, J. Noel, Electron microscopic studies of planarian regeneration. I. Fine structure of neoblasts in Dugesia dorotocephala. *J Ultrastruct Res* **27**, 7-23 (1969).
163. P. W. Reddien, N. J. Oviedo, J. R. Jennings, J. C. Jenkin, A. Sanchez Alvarado, SMEDWI-2 is a PIWI-like protein that regulates planarian stem cells. *Science* **310**, 1327-1330 (2005).
164. D. E. Wagner, J. J. Ho, P. W. Reddien, Genetic regulators of a pluripotent adult stem cell system in planarians identified by RNAi and clonal analysis. *Cell Stem Cell* **10**, 299-311 (2012).
165. T. Guo, A. H. Peters, P. A. Newmark, A Bruno-like gene is required for stem cell maintenance in planarians. *Dev Cell* **11**, 159-169 (2006).
166. J. Baguna, R. Romero, Quantitative analysis of cell types during growth, degrowth and regeneration in the planarians Dugesia mediterranea and Dugesia tigrina. *Hydrobiologia Hydrobiologia : The International Journal of Aquatic Sciences* **84**, 181-194 (1981).
167. C. R. Bardeen, F. H. Baetjer, The inhibitive action of the Roentgen rays on regeneration in planarians. *Journal of Experimental Zoology* **1**, 191-195 (1904).
168. A. Salvetti *et al.*, Adult stem cell plasticity: neoblast repopulation in non-lethally irradiated planarians. *Dev Biol* **328**, 305-314 (2009).
169. D. E. Wagner, I. E. Wang, P. W. Reddien, Clonogenic neoblasts are pluripotent adult stem cells that underlie planarian regeneration. *Science* **332**, 811-816 (2011).
170. S. W. Lapan, P. W. Reddien, dlx and sp6-9 Control optic cup regeneration in a prototypic eye. *PLoS Genet* **7**, e1002226 (2011).
171. S. W. Lapan, P. W. Reddien, Transcriptome analysis of the planarian eye identifies ovo as a specific regulator of eye regeneration. *Cell Rep* **2**, 294-307 (2012).
172. M. L. Scimone, M. Srivastava, G. W. Bell, P. W. Reddien, A regulatory program for excretory system regeneration in planarians. *Development* **138**, 4387-4398 (2011).

173. T. Hayashi, M. Asami, S. Higuchi, N. Shibata, K. Agata, Isolation of planarian X-ray-sensitive stem cells by fluorescence-activated cell sorting. *Dev Growth Differ* **48**, 371-380 (2006).
174. J. C. van Wolfswinkel, D. E. Wagner, P. W. Reddien, Single-cell analysis reveals functionally distinct classes within the planarian stem cell compartment. *Cell Stem Cell* **15**, 326-339 (2014).
175. O. Wurtzel *et al.*, A Generic and Cell-Type-Specific Wound Response Precedes Regeneration in Planarians. *Dev Cell* **35**, 632-645 (2015).
176. A. Zeng *et al.*, Prospectively Isolated Tetraspanin(+) Neoblasts Are Adult Pluripotent Stem Cells Underlying Planaria Regeneration. *Cell* **173**, 1593-1608.e1520 (2018).
177. G. T. Eisenhoffer, H. Kang, A. Sanchez Alvarado, Molecular analysis of stem cells and their descendants during cell turnover and regeneration in the planarian Schmidtea mediterranea. *Cell Stem Cell* **3**, 327-339 (2008).
178. K. C. Tu *et al.*, Egr-5 is a post-mitotic regulator of planarian epidermal differentiation. *Elife* **4**, e10501 (2015).
179. O. Wurtzel, I. M. Oderberg, P. W. Reddien, Planarian Epidermal Stem Cells Respond to Positional Cues to Promote Cell-Type Diversity. *Dev Cell* **40**, 491-504.e495 (2017).
180. S. J. Zhu, S. E. Hallows, K. W. Currie, C. Xu, B. J. Pearson, A mex3 homolog is required for differentiation during planarian stem cell lineage development. *Elife* **4**, (2015).
181. P. A. Newmark, A. Sanchez Alvarado, Bromodeoxyuridine specifically labels the regenerative stem cells of planarians. *Dev Biol* **220**, 142-153 (2000).
182. D. Wenemoser, P. W. Reddien, Planarian regeneration involves distinct stem cell responses to wounds and tissue absence. *Dev Biol* **344**, 979-991 (2010).
183. M. L. Scimone, J. Meisel, P. W. Reddien, The Mi-2-like Smed-CHD4 gene is required for stem cell differentiation in the planarian Schmidtea mediterranea. *Development* **137**, 1231-1241 (2010).
184. J. C. Rink, H. T. Vu, A. Sanchez Alvarado, The maintenance and regeneration of the planarian excretory system are regulated by EGFR signaling. *Development* **138**, 3769-3780 (2011).
185. H. Thi-Kim Vu *et al.*, Stem cells and fluid flow drive cyst formation in an invertebrate excretory organ. *Elife* **4**, (2015).
186. J. A. McKanna, Fine structure of the protonephridial system in Planaria. *Zeitschrift für Zellforschung und Mikroskopische Anatomie* **92**, 509-523 (1968).

187. E. E. RUPPERT, P. R. SMITH, The functional organization of filtration nephridia. *Biological Reviews* **63**, 231-258 (1988).
188. M. Morita, J. B. Best, Electron microscopic studies on Planaria. II. Fine structure of the neurosecretory system in the planarian Dugesia dorotocephala. *J Ultrastruct Res* **13**, 396-408 (1965).
189. I. E. Wang, S. W. Lapan, M. L. Scimone, T. R. Clandinin, P. W. Reddien, Hedgehog signaling regulates gene expression in planarian glia. *Elife* **5**, (2016).
190. R. H. Roberts-Galbraith, J. L. Brubacher, P. A. Newmark, A functional genomics screen in planarians reveals regulators of whole-brain regeneration. *Elife* **5**, (2016).
191. J. Baguna, R. Ballester, The nervous system in planarians: Peripheral and gastrodermal plexuses, pharynx innervation, and the relationship between central nervous system structure and the acoelomate organization. *J Morphol* **155**, 237-252 (1978).
192. K. Nishimura *et al.*, Identification of glutamic acid decarboxylase gene and distribution of GABAergic nervous system in the planarian Dugesia japonica. *Neuroscience* **153**, 1103-1114 (2008).
193. K. Nishimura, Y. Kitamura, T. Taniguchi, K. Agata, Analysis of motor function modulated by cholinergic neurons in planarian Dugesia japonica. *Neuroscience* **168**, 18-30 (2010).
194. K. Nishimura *et al.*, Identification and distribution of tryptophan hydroxylase (TPH)-positive neurons in the planarian Dugesia japonica. *Neurosci Res* **59**, 101-106 (2007).
195. K. Nishimura *et al.*, Reconstruction of dopaminergic neural network and locomotion function in planarian regenerates. *Dev Neurobiol* **67**, 1059-1078 (2007).
196. K. Nishimura *et al.*, Characterization of tyramine beta-hydroxylase in planarian Dugesia japonica: cloning and expression. *Neurochem Int* **53**, 184-192 (2008).
197. Y. Umesono, K. Watanabe, K. Agata, Distinct structural domains in the planarian brain defined by the expression of evolutionarily conserved homeobox genes. *Dev Genes Evol* **209**, 31-39 (1999).
198. C. P. Petersen, P. W. Reddien, Smed-betacatenin-1 is required for anteroposterior blastema polarity in planarian regeneration. *Science* **319**, 327-330 (2008).
199. A. Pigon, M. Morita, J. B. Best, Cephalic mechanism for social control of fissioning in planarians. II. Localization and identification of the receptors by electron micrographic and ablation studies. *J Neurobiol* **5**, 443-462 (1974).

200. N. J. Oviedo, P. A. Newmark, A. Sanchez Alvarado, Allometric scaling and proportion regulation in the freshwater planarian Schmidtea mediterranea. *Dev Dyn* **226**, 326-333 (2003).
201. K. G. Ross *et al.*, SoxB1 Activity Regulates Sensory Neuron Regeneration, Maintenance, and Function in Planarians. *Dev Cell* **47**, 331-347.e335 (2018).
202. J. J. Collins, 3rd *et al.*, Genome-wide analyses reveal a role for peptide hormones in planarian germline development. *PLoS Biol* **8**, e1000509 (2010).
203. K. W. Currie, B. J. Pearson, Transcription factors lhx1/5-1 and pitx are required for the maintenance and regeneration of serotonergic neurons in planarians. *Development* **140**, 3577-3588 (2013).
204. M. Marz, F. Seebeck, K. Bartscherer, A Pitx transcription factor controls the establishment and maintenance of the serotonergic lineage in planarians. *Development* **140**, 4499-4509 (2013).
205. M. L. Scimone, K. M. Kravarik, S. W. Lapan, P. W. Reddien, Neoblast specialization in regeneration of the planarian Schmidtea mediterranea. *Stem Cell Reports* **3**, 339-352 (2014).
206. D. J. Forsthoefel, A. E. Park, P. A. Newmark, Stem cell-based growth, regeneration, and remodeling of the planarian intestine. *Dev Biol* **356**, 445-459 (2011).
207. B. H. Willier, L. H. Hyman, S. A. Rifenburgh, A histochemical study of intracellular digestion in triclad flatworms. *J. Morphol. Journal of Morphology* **40**, 299-340 (1925).
208. S. Ishii, Electron microscopic observations on the Planarian tissues II. The intestine. *Fukushima journal of medical science* **12**, 67-87 (1965).
209. D. J. Forsthoefel *et al.*, An RNAi screen reveals intestinal regulators of branching morphogenesis, differentiation, and stem cell proliferation in planarians. *Dev Cell* **23**, 691-704 (2012).
210. R. M. Zayas, F. Cebria, T. Guo, J. Feng, P. A. Newmark, The use of lectins as markers for differentiated secretory cells in planarians. *Dev Dyn* **239**, 2888-2897 (2010).
211. N. M. Flores, N. J. Oviedo, J. Sage, Essential role for the planarian intestinal GATA transcription factor in stem cells and regeneration. *Dev Biol* **418**, 179-188 (2016).
212. A. Gonzalez-Sastre, N. De Sousa, T. Adell, E. Salo, The pioneer factor Smed-gata456-1 is required for gut cell differentiation and maintenance in planarians. *Int J Dev Biol* **61**, 53-63 (2017).

213. E. E. Morrisey *et al.*, GATA6 regulates HNF4 and is required for differentiation of visceral endoderm in the mouse embryo. *Genes Dev* **12**, 3579-3590 (1998).
214. F. Cebria, Planarian Body-Wall Muscle: Regeneration and Function beyond a Simple Skeletal Support. *Front Cell Dev Biol* **4**, 8 (2016).
215. P. W. Reddien, A. L. Bermange, K. J. Murfitt, J. R. Jennings, A. Sanchez Alvarado, Identification of genes needed for regeneration, stem cell function, and tissue homeostasis by systematic gene perturbation in planaria. *Dev Cell* **8**, 635-649 (2005).
216. L. E. Cote, E. Simental, P. W. Reddien, Muscle functions as a connective tissue and source of extracellular matrix in planarians. *Nat Commun* **10**, 1592 (2019).
217. M. L. Scimone, L. E. Cote, P. W. Reddien, Orthogonal muscle fibres have different instructive roles in planarian regeneration. *Nature* **551**, 623-628 (2017).
218. M. L. Scimone *et al.*, foxF-1 Controls Specification of Non-body Wall Muscle and Phagocytic Cells in Planarians. *Curr Biol* **28**, 3787-3801.e3786 (2018).
219. L. H. Hyman, *The invertebrates. Vol. II, Vol. II*. (McGraw-Hill book company inc., New York; London, 1951).
220. K. J. Pedersen, Some features of the fine structure and histochemistry of planarian subepidermal gland cells. *Zeitschrift fr Zellforschung Zeitschrift fr Zellforschung und Mikroskopische Anatomie* **50**, 121-142 (1959).
221. K. J. r. Pedersen, Studies on the nature of planarian connective tissue. *Zeitschrift f¸r Zellforschung Zeitschrift f¸r Zellforschung und Mikroskopische Anatomie* **53**, 569-608 (1961).
222. M. Morita, Phagocytic response of planarian reticular cells to heat-killed bacteria. *Hydrobiologia* **227**, 193-199 (1991).
223. M. Morita, Structure and function of the reticular cell in the planarian Dugesia dorotocephala. *Hydrobiologia* **305**, 189-196 (1995).
224. B. M. Stubenhaus *et al.*, Light-induced depigmentation in planarians models the pathophysiology of acute porphyrias. *Elife* **5**, (2016).
225. C. Wang *et al.*, Forkhead containing transcription factor Albino controls tetrapyrrole-based body pigmentation in planarian. *Cell Discov* **2**, 16029 (2016).
226. X. He *et al.*, FOX and ETS family transcription factors regulate the pigment cell lineage in planarians. *Development* **144**, 4540-4551 (2017).
227. S. Ishii, Electron microscopic observations on the planarian tissues I. A Survey of the pharynx. *Fukushima Journal of Medical Science* **9-10**, 51-73 (1962).

228. M. D. Molina, E. Salo, F. Cebria, The BMP pathway is essential for re-specification and maintenance of the dorsoventral axis in regenerating and intact planarians. *Dev Biol* **311**, 79-94 (2007).
229. C. E. Adler, C. W. Seidel, S. A. McKinney, A. Sanchez Alvarado, Selective amputation of the pharynx identifies a FoxA-dependent regeneration program in planaria. *Elife* **3**, e02238 (2014).
230. S. Koinuma, Y. Umesono, K. Watanabe, K. Agata, Planaria FoxA (HNF3) homologue is specifically expressed in the pharynx-forming cells. *Gene* **259**, 171-176 (2000).
231. P. Rompolas, R. S. Patel-King, S. M. King, An outer arm Dynein conformational switch is required for metachronal synchrony of motile cilia in planaria. *Mol Biol Cell* **21**, 3669-3679 (2010).
232. A. Tazaki, K. Kato, H. Orii, K. Agata, K. Watanabe, The body margin of the planarian Dugesia japonica: characterization by the expression of an intermediate filament gene. *Dev Genes Evol* **212**, 365-373 (2002).
233. S. J. Zhu, B. J. Pearson, Smed-myb-1 Specifies Early Temporal Identity during Planarian Epidermal Differentiation. *Cell Rep* **25**, 38-46.e33 (2018).
234. L. C. Cheng *et al.*, Cellular, ultrastructural and molecular analyses of epidermal cell development in the planarian Schmidtea mediterranea. *Dev Biol* **433**, 357-373 (2018).
235. J. N. Witchley, M. Mayer, D. E. Wagner, J. H. Owen, P. W. Reddien, Muscle cells provide instructions for planarian regeneration. *Cell Rep* **4**, 633-641 (2013).
236. P. W. Reddien, A. L. Bermange, A. M. Kicza, A. Sanchez Alvarado, BMP signaling regulates the dorsal planarian midline and is needed for asymmetric regeneration. *Development* **134**, 4043-4051 (2007).
237. C. P. Petersen, P. W. Reddien, A wound-induced Wnt expression program controls planarian regeneration polarity. *Proc Natl Acad Sci U S A* **106**, 17061-17066 (2009).
238. K. A. Gurley *et al.*, Expression of secreted Wnt pathway components reveals unexpected complexity of the planarian amputation response. *Dev Biol* **347**, 24-39 (2010).
239. T. Hayashi *et al.*, A LIM-homeobox gene is required for differentiation of Wnt-expressing cells at the posterior end of the planarian body. *Development* **138**, 3679-3688 (2011).
240. C. C. Chen, I. E. Wang, P. W. Reddien, pbx is required for pole and eye regeneration in planarians. *Development* **140**, 719-729 (2013).

241. C. Vasquez-Doorman, C. P. Petersen, zic-1 Expression in Planarian neoblasts after injury controls anterior pole regeneration. *PLoS Genet* **10**, e1004452 (2014).
242. J. H. Owen, D. E. Wagner, C. C. Chen, C. P. Petersen, P. W. Reddien, teashirt is required for head-versus-tail regeneration polarity in planarians. *Development* **142**, 1062-1072 (2015).
243. H. Reuter *et al.*, Beta-catenin-dependent control of positional information along the AP body axis in planarians involves a teashirt family member. *Cell Rep* **10**, 253-265 (2015).
244. R. Lander, C. P. Petersen, Wnt, Ptk7, and FGFRL expression gradients control trunk positional identity in planarian regeneration. *Elife* **5**, (2016).
245. T. Stuckemann *et al.*, Antagonistic Self-Organizing Patterning Systems Control Maintenance and Regeneration of the Anteroposterior Axis in Planarians. *Dev Cell* **40**, 248-263.e244 (2017).
246. C. P. Petersen, P. W. Reddien, Wnt signaling and the polarity of the primary body axis. *Cell* **139**, 1056-1068 (2009).
247. C. P. Petersen, P. W. Reddien, Polarized notum activation at wounds inhibits Wnt function to promote planarian head regeneration. *Science* **332**, 852-855 (2011).
248. M. Sureda-Gomez, J. M. Martin-Duran, T. Adell, Localization of planarian beta-CATENIN-1 reveals multiple roles during anterior-posterior regeneration and organogenesis. *Development* **143**, 4149-4160 (2016).
249. K. A. Gurley, J. C. Rink, A. Sanchez Alvarado, Beta-catenin defines head versus tail identity during planarian regeneration and homeostasis. *Science* **319**, 323-327 (2008).
250. M. Iglesias, J. L. Gomez-Skarmeta, E. Salo, T. Adell, Silencing of Smed-betacatenin1 generates radial-like hypercephalized planarians. *Development* **135**, 1215-1221 (2008).
251. T. Adell, E. Salo, M. Boutros, K. Bartscherer, Smed-Evi/Wntless is required for beta-catenin-dependent and -independent processes during planarian regeneration. *Development* **136**, 905-910 (2009).
252. M. L. Scimone, L. E. Cote, T. Rogers, P. W. Reddien, Two FGFRL-Wnt circuits organize the planarian anteroposterior axis. *Elife* **5**, (2016).
253. F. Cebria *et al.*, FGFR-related gene nou-darake restricts brain tissues to the head region of planarians. *Nature* **419**, 620-624 (2002).
254. E. M. Hill, C. P. Petersen, Wnt/Notum spatial feedback inhibition controls neoblast differentiation to regulate reversible growth of the planarian brain. *Development* **142**, 4217-4229 (2015).

255. M. C. Vogg *et al.*, Stem cell-dependent formation of a functional anterior regeneration pole in planarians requires Zic and Forkhead transcription factors. *Dev Biol* **390**, 136-148 (2014).
256. I. M. Oderberg, D. J. Li, M. L. Scimone, M. A. Gavino, P. W. Reddien, Landmarks in Existing Tissue at Wounds Are Utilized to Generate Pattern in Regenerating Tissue. *Curr Biol* **27**, 733-742 (2017).
257. E. G. Schad, C. P. Petersen, STRIPAK Limits Stem Cell Differentiation of a WNT Signaling Center to Control Planarian Axis Scaling. *Curr Biol* **30**, 254-263.e252 (2020).
258. H. Orii, K. Watanabe, Bone morphogenetic protein is required for dorso-ventral patterning in the planarian Dugesia japonica. *Dev Growth Differ* **49**, 345-349 (2007).
259. M. A. Gavino, P. W. Reddien, A Bmp/Admp regulatory circuit controls maintenance and regeneration of dorsal-ventral polarity in planarians. *Curr Biol* **21**, 294-299 (2011).
260. F. Cebria, T. Guo, J. Jopek, P. A. Newmark, Regeneration and maintenance of the planarian midline is regulated by a slit orthologue. *Dev Biol* **307**, 394-406 (2007).
261. P. W. Reddien, The Cellular and Molecular Basis for Planarian Regeneration. *Cell* **175**, 327-345 (2018).
262. P. W. Reddien, Constitutive gene expression and the specification of tissue identity in adult planarian biology. *Trends Genet* **27**, 277-285 (2011).

Chapter 2

Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*

Christopher T. Fincher, Omri Wurtzel, Thom de Hoog, Kellie M. Kravarik, and Peter W. Reddien

P.W.R. supervised. C.T.F. and P.W.R. designed experiments and wrote the manuscript. C.T.F., P.W.R., and O.W. analyzed data. C.T.F. and T.H. built and optimized the Drop-seq setup. C.T.F. developed the data processing pipeline. C.T.F. and K.M.K. developed the pipeline for clustering analysis. C.T.F. and P.W.R. performed planarian tissue extractions. C.T.F. performed all other planarian experiments. O.W. generated the online resource.

Published as:

Fincher, C.T., Wurtzel, O., de Hoog, T., Kravarik, K.M., and Reddien, P.W. (2018). Cell Type Transcriptome Atlas for the Planarian *Schmidtea Mediterranea*. Science *360*, eeaq1736.

Abstract

The transcriptome of a cell dictates its unique cell type biology. We used single-cell RNA sequencing to determine the transcriptomes for essentially every cell type of a complete animal: the regenerative planarian *Schmidtea mediterranea.* Planarians contain a diverse array of cell types, possess lineage progenitors for differentiated cells (including pluripotent stem cells), and constitutively express positional information, making them ideal for this undertaking. We generated data for 66,783 cells, defining transcriptomes for known and many previously unknown planarian cell types and for putative transition states between stem and differentiated cells. We also uncovered regionally expressed genes in muscle, which harbors positional information. Identifying the transcriptomes for potentially all cell types for many organisms should be readily attainable and represents a powerful approach to metazoan biology.

Introduction

The complete sequence of animal genomes, such as that of *Caenorhabditis elegans* reported in 1998 and humans in 2001, has had an immeasurable impact on research (***1***, ***2***, ***3***). Whereas the genome sequence of an organism contains the information for its development and physiology, the transcriptomes (the sets of actively transcribed genes) of the cell types in an organism define how the genome is used for the unique functions of its cells. Recent advances in RNA sequencing of individual cells have greatly enhanced the ability to determine cell type transcriptomes (***4***, ***5***), and single-cell RNA sequencing (SCS) of thousands of cells has become readily achievable (***6***). For example, the transcriptomes of most cell types of complete *C. elegans* L2 larvae and numerous mouse cells were recently reported with this approach (***7***, ***8***). We reasoned that it might be possible, given these advances, to determine the transcriptomes of essentially every cell type of a complete adult organism possessing an unknown number of cell types.

Multicellular organisms can have many millions of cells and hundreds of different cell types, and the cellular composition of organisms varies markedly over the course of development. This complexity has historically made the identification of all cell types, much less their transcriptomes, for most multicellular organisms an extreme challenge. The planarian *Schmidtea mediterranea* is an attractive case study organism for which to generate the transcriptomes for all cells in an animal. Planarians are famous for their ability to regenerate essentially any missing body part, and they possess a complex body plan containing many characterized cell types (***9***, ***10***). Despite this complexity, with an average planarian possessing ~10^5 to 10^6 cells (***11***), planarians are smaller with simpler anatomy than humans and many other model systems such as mice. Planarians are also easily dissociated into single-cell suspensions, allowing potential characterization of all cells. Because some planarian cell types, such as glia (***12***, ***13***), have only recently been defined with molecular markers, it is probable that undescribed planarian cell types exist. The combination of known and potentially unknown cell types is attractive for developing approaches that can apply to diverse organisms with varying

amounts of available cell type information. Planarians possess a population of proliferative cells called neoblasts that contain pluripotent stem cells, enabling their ability to regenerate and replace aged cells in tissue turnover (***14***). Neoblasts are the only cycling somatic cells and the source of all new tissue. Neoblasts contain multiple classes of specialized cells, with transcription factors expressed to specify cell fate (***15***, ***16***). Because of the constant turnover of planarian tissues, essentially all stages of all cell lineages, from pluripotent stem cell to differentiated cell, are anticipated to be present in the adult (***9, 17***).

Planarians also constitutively and regionally express dozens of genes that have roles in positional information (***18***). These genes, referred to as positional control genes (PCGs), are expressed in a complex spatial map spanning anterior-posterior (AP), medial-lateral (ML), and dorsal-ventral (DV) axes (***18***), and their expression is largely restricted to muscle (***19***). PCGs are hypothesized to constitute instructions for the maintenance and regeneration of the body plan. Because of these features, comprehensive SCS at a single time point (the adult) could allow transcriptome identification for all differentiated cell types, lineage precursors for these cells, and the patterning information that guides new cell production and organization. To capture this information in most organisms would require sampling the adult and many transient stages of embryogenesis.

Results & Figures

Single-cell RNA sequencing of 50,562 planarian cells

Planarians have a complex internal anatomy including a brain, ventral nerve cords, peripheral nervous system, epidermis, intestine, muscle, an excretory system (the protonephridia), and a centrally located pharynx (***10***). These major tissues are composed of multiple different cell types that, together with other gland and accessory cells, constitute the planarian anatomy.

To detect planarian cell types and states in an unbiased manner, including rare cell types, we used the SCS method Drop-seq (***6***) to determine the transcriptomes for 50,562 individual cells from adults (Figures 2.1A and 2.2A, and Table 2.1). Planarians contain 10^5 to 10^6 cells (***11***), and yet some cell types are extremely rare, such as the ~100 photoreceptor neurons of eyes (***20***). Given such rarity, sequencing random cells from entire animals might not reach cell type saturation with even 10^5 cells sequenced. Therefore, we divided animals into five sections (head, prepharyngeal region, trunk with pharynx removed, tail, and the pharynx itself) and cells from each region were dissociated, sorted by flow cytometry, and sequenced (Figures 2.1A and Figure 2.2A, and Table 2.1). Sequences were aligned to a previously assembled transcriptome (***21***). We targeted cell type saturation by assessing coverage of known, rare cell types during iterative rounds of cell isolation and sequencing in a region-by-region approach. In total, 25 separate Drop-seq runs were completed, yielding cells with an average of 3020 unique molecular identifiers (UMIs) and 1404 genes (~13% of the estimated detection limit) (Figure 2.2, A to C, Table 2.1, and Supplementary materials).

Genes with high variance and expression across cells were used to generate informative principal components using Seurat (***6, 22***). Cells were clustered using Seurat into 44 distinct major clusters using a graph-based clustering approach and were visualized by applying t-distributed stochastic neighbor embedding on transcriptomes (t-SNE) (Figures 2.1B and 2.2D). Cells from different regions were largely interspersed in the t-SNE plots, except for cells from the pharynx, which contains many unique cell

types (Figure 2.3A). Cell doublets were scarce within the data and did not affect clustering results (Figure 2.3, B to D). To determine the identity of each cluster, we identified cluster-specific genes by means of a receiver operating characteristic curve analysis and a likelihood ratio test based on zero-inflated data (Table 2.2) (***23***). Expression of established cell type markers within each cluster and fluorescence in situ hybridization (FISH) with cluster-specific markers enabled cluster assignment to one of eight previously identified planarian tissue classes: protonephridia, neural, epidermis, intestine, pharynx, muscle, neoblast, and parenchymal (Figure 2.1C). The parenchymal class was previously termed “parapharyngeal” because of localization of some enriched markers around the pharynx (***24***). However, most cell populations within this class exhibit broader localization in the planarian parenchyma. We also identified a ninth group of clusters marked by *CTSL2* (dd175) expression (Figure 2.1D). *CTSL2* (dd175) FISH revealed cells with long processes distributed broadly. We designated this group of clusters the *cathepsin*$^{+}$ class. Hierarchical clustering of a subset of 5000 cells by Euclidean distance, independently of Seurat, recapitulated assignment of cells into these nine tissue classes (Figure 2.4).

Clusters representing the major planarian tissue classes were generally heterogeneous in terms of gene expression. For example, neural clusters contained a large number of known neuronal cell types, which suggests that multiple distinct cell types could be identified within each major cluster (Figure 2.5). Therefore, we systematically subclustered each major cluster group (Figures 2.6, 2.14, 2.19, 2.31, and 2.34), identifying >150 subclusters, and determined genes with enriched expression in cells of each subcluster (Table 2.2). Subclustering proved a powerful approach to defining the collection of cell types that constituted each major cluster and identified candidate transition states between stem cells and differentiated cells.

Figure 2.1

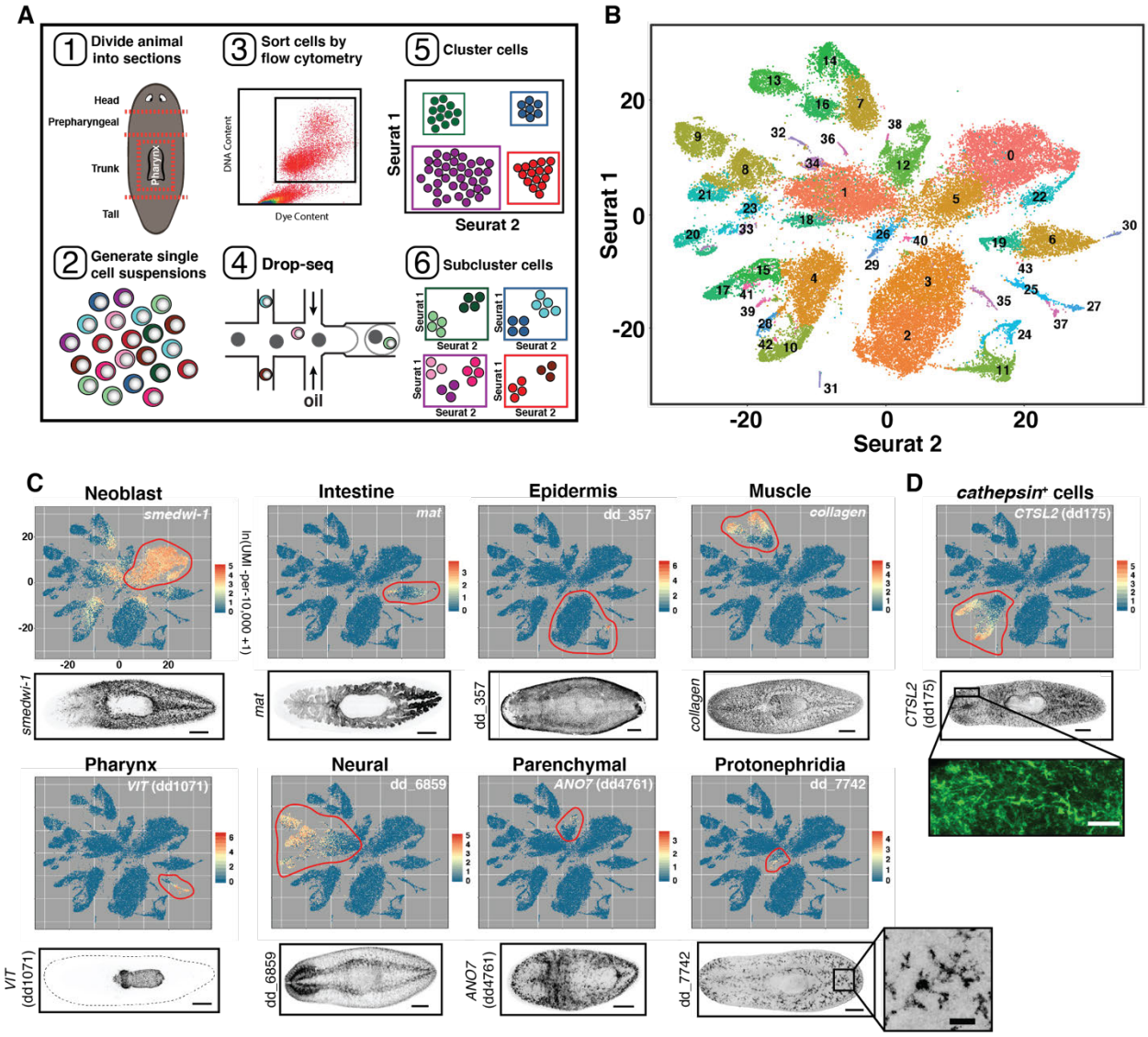

A
1 Divide animal into sections
Head
Prepharyngeal
Trunk
Tall
Pharynx
3 Sort cells by flow cytometry
DNA Content
Dye Content
5 Cluster cells
Seurat 1
Seurat 2
2 Generate single cell suspensions
4 Drop-seq
oil
6 Subcluster cells
B
Seurat 1
Seurat 2
C
Neoblast
smedwi-1
Intestine
mat
Epidermis
dd_357
Muscle
collagen
Pharynx
VIT (dd1071)
Neural
dd_6859
Parenchymal
ANO7 (dd4761)
Protonephridia
dd_7742
D
cathepsin+ cells
CTSL2 (dd175)

Figure 2.1. **Drop-seq of 50,562 planarian cells.**

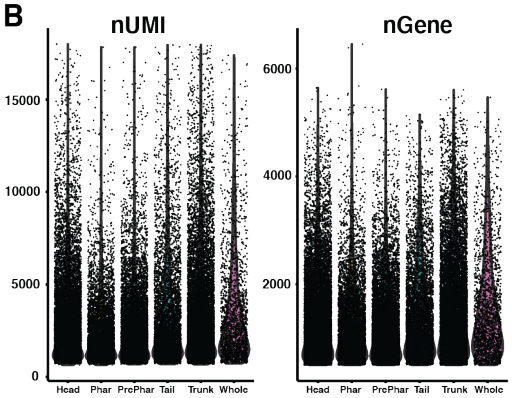
(**A**) Schematic illustrating the workflow used to isolate and cluster single cells. (**B**) t-SNE representation of 44 clusters generated from the data. (**C** and **D**) Upper panels: t-SNE plots colored according to gene expression (red, high; blue, low) for highly enriched genes from nine planarian tissue classes. Red outlines denote clusters assigned to that tissue class. Lower panels: FISH images for tissue-enriched genes. Scale bars: whole-animal images, 200 μm; insets, 50 μm.

Figure 2.2

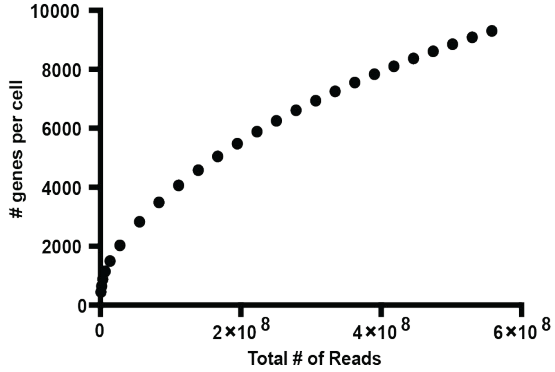
A

	Sequencing Run 1 5,876 expected cells ~449.2M reads				Sequencing Run 2 6,644 expected cells ~545.3M reads			Sequencing Run 3 7,114 expected cells ~461.4M reads			Sequencing Run 4 6,829 expected cells ~488M reads				Sequencing Run 5 8,006 expected cells ~437.9M reads				Sequencing Run 6 3,100 expected cells ~489.7M reads	Sequencing Run 7 7,784 expected cells ~413M reads			Sequencing Run 8 8,104 exptected cells ~373.8M reads		
Run #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Whole	2000 cells																								
Head		400	1531	1945							965	2113	2158	1593			1838	2940		1900					
PrePhar					2174	1350																	3904	2200	
Trunk							3120	2842											3100		2235				2000
Tail									2710	1562															
Pharynx															1607	1621						3649			

B



C



D

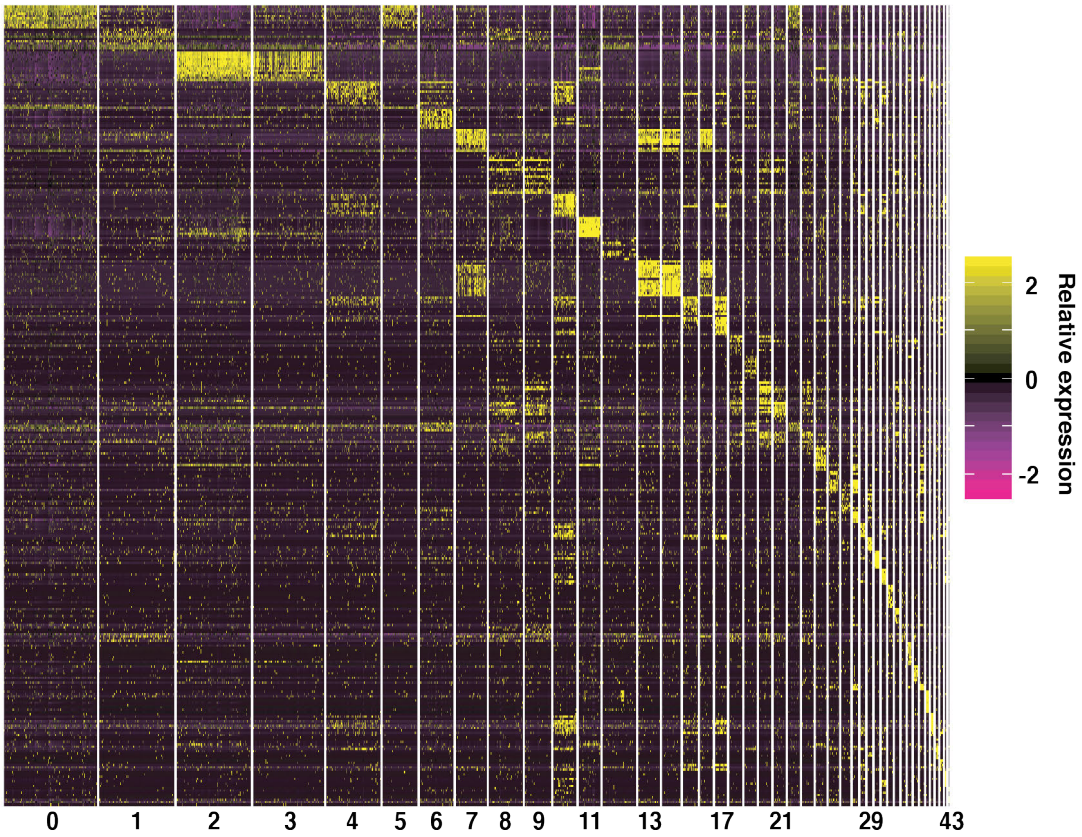
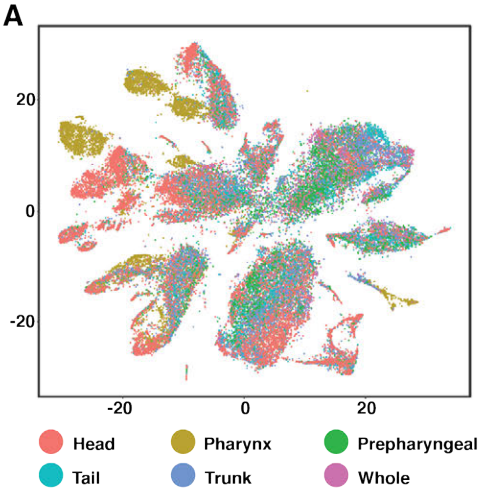


Figure 2.2. Summary of the data and primary clustering.

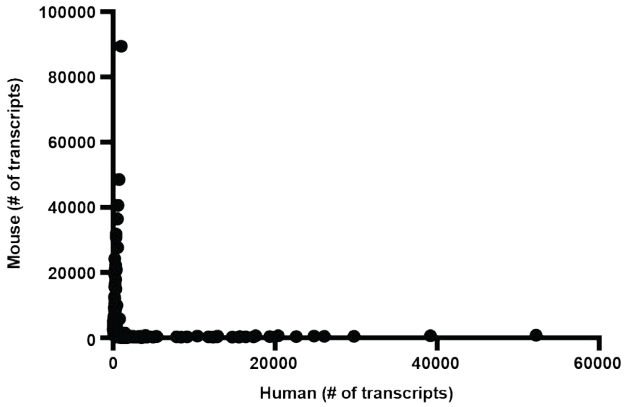
(**A**) Summary of the Drop-seq runs performed. The expected number of cells for each Drop-seq run was estimated from the bead yield (e.g., ~5% of beads are exposed to a cell). Different run colors indicate distinct biological replicates. (**B**) Violin plots indicating the average number of UMIs and average number of genes for all cells from each body region following quality filtering. (**C**) Mean number of genes per cell for 197 cells at increasing numbers of total sequencing reads. Fitting a one phase exponential association equation (y=882.4+(10647-882.4)* (1-e^(-3.29E-9*x)) to the plot identifies a plateau at 10,647 genes per cell at saturating read numbers. (**D**) Heat map of the expression of the top 10 genes from each cluster of the overall clustering, grouped by cluster number. Cells, columns; Genes, rows.

Figure 2.3

A



B



C

Cell Doublets (Total = 81/13336 cells)

	Enterocytes n=281 cells	*Mag1*+ cells n=51 cells	dd_10872+ cells n=712 cells	Ciliated epidermis n=197 cells	Serotonergic Neurons n=32 cells	Prog cells n=7773 cells	Muscle n=4211 cells
Flame cells n=160 cells	1	0	0	0	0	1	1
Enterocytes n=281 cells		2	0	6	0	12	9
***Mag1+* cells** n=51 cells			1	1	0	0	1
dd_10872+ cells n=712 cells				1	1	5	6
Ciliated Epidermis n=197 cells					0	0	3
Serotonergic Neurons n=32 cells						0	0
***Prog* cells** n=7773 cells							30

D

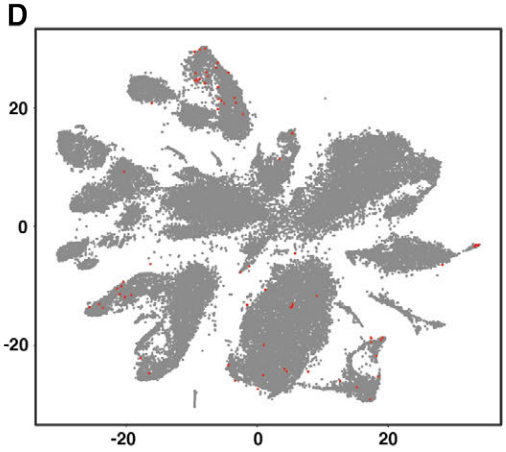


Figure 2.3. Validation of the clustering results.

(**A**) t-SNE plot colored by the body region from which each cell was isolated. Cells isolated from each body region were largely interspersed, demonstrating a general absence of batch effects. Pharynx cells did form multiple distinct clusters, consistent with the exclusive localization of many cell types to the pharynx. (**B**) Scatter plot of the number of mouse and human transcripts expressed by 268 cells of a Drop-seq run on a 1:1 mixture of HEK293T and NIH/3T3 cells (Methods). Cells almost exclusively expressed transcripts from only one species, indicating an absence of cell doublets. The mixed-species Drop-seq run was performed prior to all Drop-seq runs on planarian cells, and the same cell concentration (191 cells/µl) was used for all planarian cell runs. (**C**) Table indicating the number of cells shared between each cell type list. Cell types were assigned by the expression of at least six out of eight highly enriched genes for each cell type (Methods). (**D**) t-SNE plot colored by the 81 cell doublets identified in (C). Positive cells, red; Negative cells, grey.

Figure 2.4

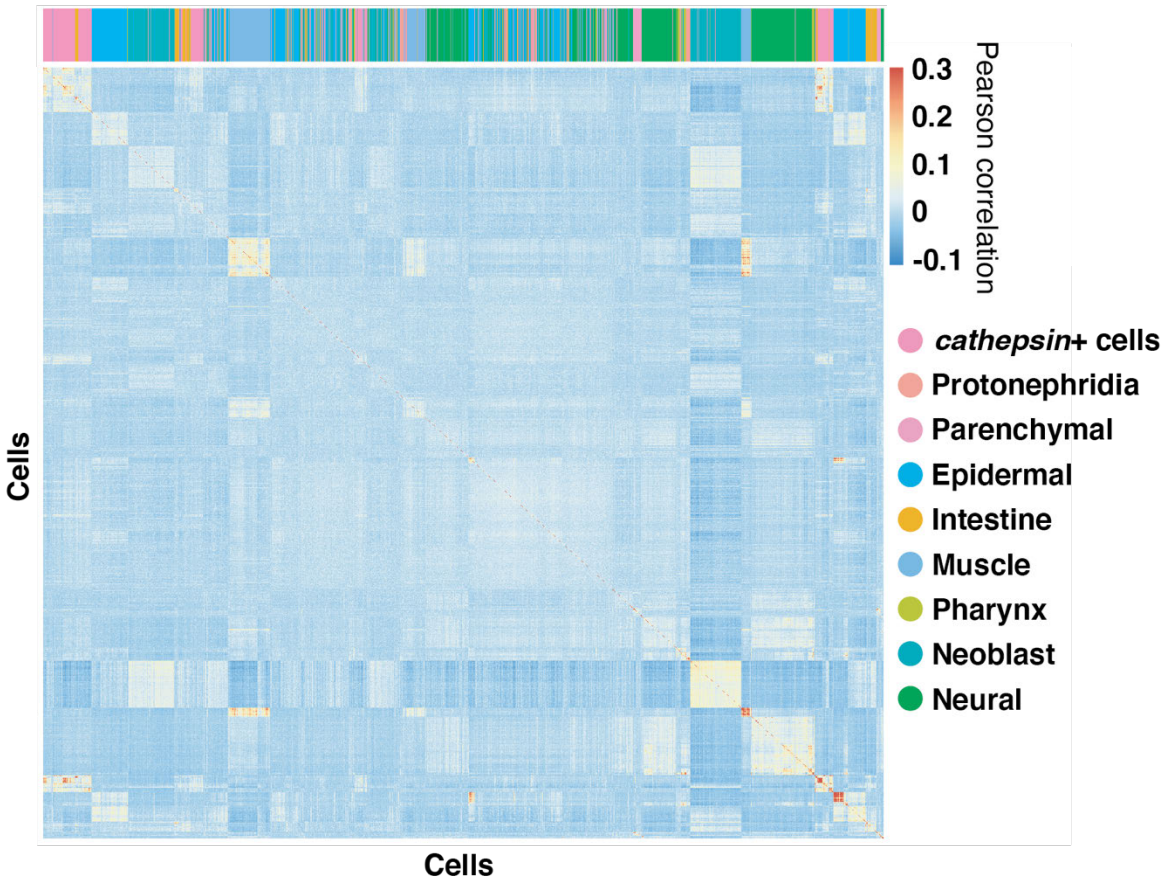

Cells
Cells
Pearson correlation
0.3
0.2
0.1
0
-0.1
cathepsin+ cells
Protonephridia
Parenchymal
Epidermal
Intestine
Muscle
Pharynx
Neoblast
Neural

Figure 2.4. Hierarchical clustering of cells independently of Seurat maintains tissue class assignment.

Heat map of the pearson correlation coefficient of 5,000 cells from the data following hierarchical clustering independently of Seurat. Top panel: cells colored by their tissue class assignment by Seurat. Pearson correlations greater than 0.3 were collapsed to 0.3.

Figure 2.5

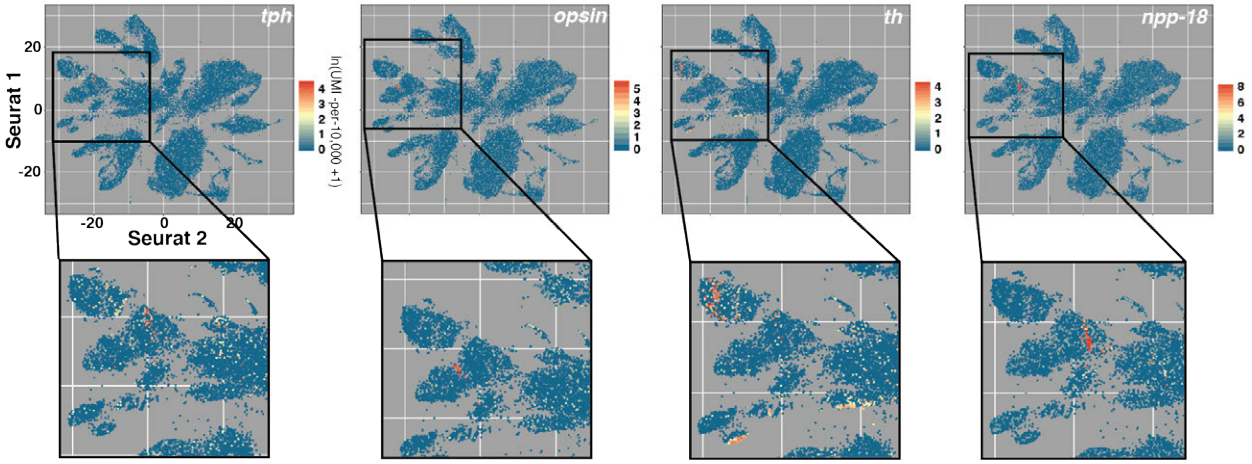

tph
opsin
th
npp-18
Seurat 1
Seurat 2
20
0
-20
-20
0
20
ln(UMI-per-10,000 +1)
4 3 2 1 0
5 4 3 2 1 0
4 3 2 1 0
8 6 4 2 0

Figure 2.5. Many clusters were a heterogeneous mix of distinct cell types.

t-SNE plots colored by expression of 4 neuronal markers: tph (***25***), opsin (***26***), th (***27***), and npp-18 (***28***), which are expressed in distinct cell populations in vivo. Insets identify distinct groups of cells positive for each marker in the neural associated clusters.

Progenitors in planarian cell lineages

Neoblasts are abundant and express canonical marker genes such as *smedwi-1* (***29***), *vasa* (***30***), and *bruli* (***31***) (Figures 2.1C and 2.7, A and B). Neoblasts are cycling cells and consequently show enrichment in expression of S/G_2/M cell cycle markers (Figure 2.7C). To identify the transcriptomes of potential neoblast subpopulations, we selected in silico and subclustered 12,212 cells with *smedwi-1* expression of ≥2.5 [ln(UMI-per-10,000 + 1)] (Figures 2.6A and 2.8A). Resulting clusters on the left of the plot were enriched in S/G_2/M cell cycle markers (Figure 2.8B). These clusters included the previously characterized major specialized neoblast classes, including γ-neoblasts (intestine progenitors) and ζ-neoblasts (epidermis progenitors) (***32***) (Figure 2.6B). A number of other subclusters were also identified, including one marked by expression of the contig dd_10988 (Figure 2.8, C and D). FISH confirmed that dd_10988 was expressed in a neoblast subset as well as in a number of *smedwi-1*$^-$ cells (Figure 2.6C).

The large number of subclustered neoblasts facilitated transcriptome determination for candidate progenitors for many planarian tissues. Clusters to the right of the plot were marked by a G_1/G_0 cell cycle status and displayed expression of various tissue markers (Figures 2.6D and 2.8B). These included a population defined by expression of *POU2/3*, a marker for protonephridia-specialized neoblasts (***33***), and a number of subclusters expressing markers also expressed in specific differentiated tissues or their postmitotic precursors, such as *ChAT* for the nervous system, *prog-1* for the epidermis, *ASCL4* (dd1854) for parenchymal cells, and *COL4A6A* (dd2337) for muscle (Figures 2.6D and 2.9A). Expression of these markers in *smedwi-1*$^+$ cells suggests that these cells could be transition states for those lineages. Several markers enriched in the dd_10988$^+$ subcluster, including dd_10988, were also expressed in cells of the two *smedwi-1*$^+$ neural subclusters, as well as in neural cells of the initial clustering (Figures 2.8C and 2.9, B and C), which suggests that the dd_10988$^+$ subcluster is enriched in neural progenitors. Likewise, many markers enriched in the *PLOD1* (dd3457)$^+$ subcluster were also expressed in the *smedwi-1*$^+$ muscle subcluster, which suggests that the *PLOD1* (dd3457)$^+$ subcluster is enriched in muscle

progenitors (Figure 2.9D). *prox-1*, *hnf-4*, *nkx2.2*, and *gata4/5/6* encode transcription factors expressed in intestinal progenitors (***32***), and in these data all four genes were expressed in γ-neoblasts (Figures 2.6, B and E, and 2.9E), with *hnf-4*, *nkx2.2*, and *gata4/5/6* also expressed in intestinal clusters (Figures 2.6E and 2.9F). *hnf-4*, but not *prox-1*, *nkx2.2*, and *gata4/5/6*, was also expressed in a *smedwi-1*$^+$ cell cluster enriched in *CTSL2* (dd175) expression (the *cathepsin*$^+$ cell marker) and in differentiated *cathepsin*$^+$ cells (Figures 2.6E and 2.9G). The additional transcription factor–encoding genes *ETS1* (dd2092) and *FOXF1* (dd6910) were expressed with *hnf-4* in these cells and also displayed expression patterns similar to that of *CTSL2* (dd175) in the animal (Figures 2.6F and 2.10, A and B) and have recently been shown to regulate the planarian pigment cell lineage (***34***). Pigment cells clustered within the *cathepsin*$^+$ cell class in our data (see below). By FISH, *hnf-4* was indeed coexpressed with *nkx2.2* and *gata4/5/6* in the intestine, but was also coexpressed with *cathepsin*$^+$ cell markers (Figure 2.10, C to E), which suggests that *hnf-4* is expressed in two distinct lineages. These data demonstrate the utility of this approach for identifying potentially novel neoblast progenitor populations and the transcription factors that define them.

Some planarian neoblasts display pluripotency in clonal assays and are hypothesized to generate all lineage-committed neoblast subpopulations, and are called clonogenic neoblasts (***14***). We selected cells expressing high levels of *smedwi-1* but that excluded ζ- and γ-neoblasts [including subclusters 2, 9, dd_10988$^+$, dd_6998$^+$, dd_17796$^+$, *SAMD15* (dd19710)$^+$, dd_11221$^+$, dd_13666$^+$, and *PLOD1* (dd3457)$^+$] and subclustered this set of neoblasts in isolation (Figure 2.11, A and B). A remnant ζ-neoblast population (clusters 4 and 6), as well as protonephridia progenitors (cluster 10) and the putative neural (clusters 2 and 5) and muscle (clusters 1 and 9) progenitor populations described above in the *smedwi-1*$^+$ cell subclustering, were identified (Figure 2.11C and Table 2.2). Clusters 0, 3, 7, and 8 were largely devoid of specifically enriched markers (Table 2.2). It is therefore possible that clonogenic neoblasts are defined by an

absence of any tissue-specific markers, as opposed to the unique expression of specific genes.

When all cells were clustered together, numerous *smedwi-1*$^+$ cells were present regionally within each of the other eight major planarian tissue clusters (Figure 2.1C). We reasoned that these *smedwi-1*$^+$ cells could represent progenitors for the cell types within each associated tissue cluster. We therefore examined these *smedwi-1*$^+$ cells after taking each tissue class in isolation and subclustering the data.

The planarian epidermis contains ciliated and nonciliated cells as well as dorsal-ventral boundary epidermis (***10, 32, 35***), and the lineage from ζ-neoblasts to epidermal cells is well characterized (***35, 36, 37***) (Figure 2.6G). SCS reveals gene expression transitions during neoblast epidermal differentiation (***35***); subclustering 11,021 epidermal lineage cells (Figure 2.1C) produced subclusters associated with each epidermal lineage stage (Figures 2.6H and 2.12, A and B). Plotting gene expression onto this t-SNE map showed a continuous progression from ζ-neoblast to differentiated cells (Figures 2.6I and 2.12C).

The gene dd_554 [SmedASXL_059179 in (***38***)] is expressed in candidate pharynx progenitors (***38***) (*smedwi-1*$^+$ cells at the pharynx base) and in *smedwi-1*$^-$ cells within the pharynx (***38***) (Figures 2.6J and 2.13). Subclustering the 1083 non-muscle, non-neuronal pharynx cluster cells (Figure 2.1C) revealed that *smedwi-1*$^+$ cells sequenced from nonpharynx midbody tissue clustered with pharynx cells, despite not being part of the pharynx itself (Figure 2.6, K and L). Because the pharynx lacks neoblasts, pharynx-specialized neoblasts must be outside of the pharynx. This clustering of neoblasts with pharynx cells clearly demonstrates the ability of SCS data clustering to associate lineage precursors with differentiated cells. Similarly, many dd_554$^+$ cells sequenced from outside of the pharynx clustered with pharynx cells (Figure 2.6, K and L). Plotting *smedwi-1*/dd_554 expression onto pharyngeal subclusters revealed a progression from *smedwi-1*$^+$ cells isolated outside the pharynx to dd_554$^+$ cells isolated outside the pharynx to dd_554$^+$ cells isolated inside the pharynx to pharyngeal cells

(Figure 2.6, K and L). These epidermis and pharynx examples demonstrate how precursor stages within cell lineages can be identified from subclustering cells within a major tissue class. Because planarians constantly generate new differentiated cells for essentially all tissue types (***17, 20***), transcriptomes for lineage precursors for essentially every cell type in the body could in principle be studied with this approach.

Cell lineages for many planarian cell types are largely uncharacterized. After tissue type subclustering, *smedwi-1*$^{+}$ cells were present with locally high expression in resultant t-SNE plots; *smedwi-1* expression level gradually declined in cells across subclusters (Figure 2.6, M and N). These *smedwi-1*$^{+}$ cells, similar to the epidermis and pharynx cases, could represent transition states between pluripotent neoblasts and differentiated cells for the various cells of the protonephridia, intestine, muscle, nervous system, parenchymal, and *cathepsin*$^{+}$ cells (Figure 2.6, M and N). The *smedwi-1*$^{+}$ cells found within subclusters of the major tissue type classes generally displayed enriched expression of at least one transcription factor. For example, *smedwi-1* expression was high within cells at the center of the parenchymal cell t-SNE plot and displayed a graded decrease projecting in all directions into seven major parenchymal subclusters (Figure 2.6N). Each projection was associated with enriched expression of one or more distinct transcription factors, identifying candidate transcription factors associated with the specification of different parenchymal cell types (Figure 2.6O).

Figure 2.6

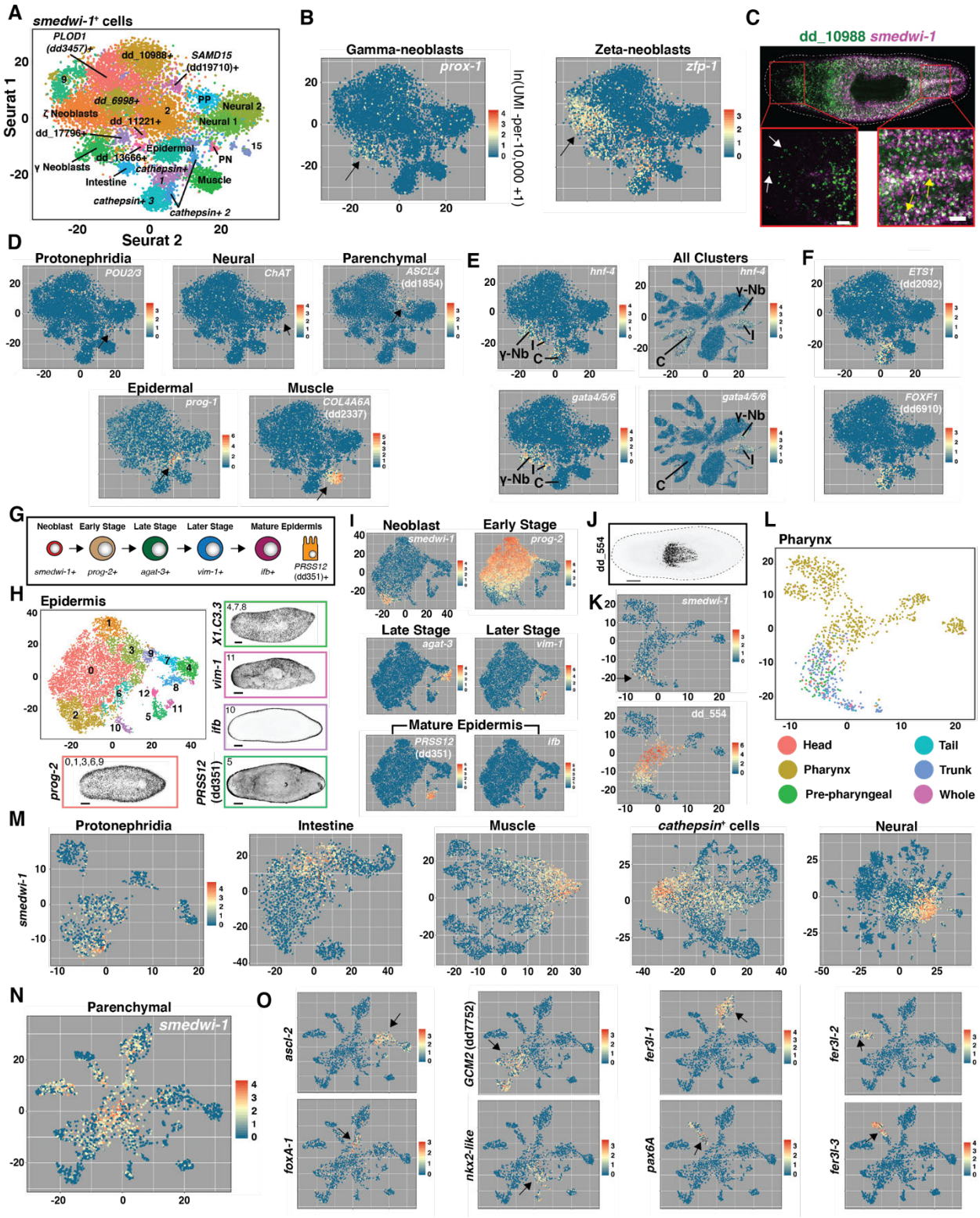

A
smedwi-1+ cells
PLOD1 (dd3457)+
dd_10988+
SAMD15 (dd19710)+
dd_6998+
PP
Neural 2
ζ Neoblasts
dd_11221+
Neural 1
dd_17796+
Epidermal
dd_13666+
PN
γ Neoblasts
cathepsin+
Intestine
Muscle
cathepsin+ 3
cathepsin+ 2
Seurat 1
Seurat 2
B
Gamma-neoblasts
prox-1
Zeta-neoblasts
zfp-1
ln(UMI-per-10,000 +1)
C
dd_10988 smedwi-1
D
Protonephridia
POU2/3
Neural
ChAT
Parenchymal
ASCL4 (dd1854)
Epidermal
prog-1
Muscle
COL4A6A (dd2337)
E
hnf-4
All Clusters
γ-Nb
I
C
gata4/5/6
F
ETS1 (dd2092)
FOXF1 (dd6910)
G
Neoblast
Early Stage
Late Stage
Later Stage
Mature Epidermis
smedwi-1+
prog-2+
agat-3+
vim-1+
ifb+
PRSS12 (dd351)+
H
Epidermis
X1.C3.3
vim-1
ifb
prog-2
PRSS12 (dd351)
I
Neoblast
smedwi-1
Early Stage
prog-2
Late Stage
agat-3
Later Stage
vim-1
Mature Epidermis
PRSS12 (dd351)
ifb
J
dd_554
K
smedwi-1
dd_554
L
Pharynx
Head
Tail
Pharynx
Trunk
Pre-pharyngeal
Whole
M
Protonephridia
Intestine
Muscle
cathepsin+ cells
Neural
smedwi-1
N
Parenchymal
smedwi-1
O
ascl-2
GCM2 (dd7752)
foxA-1
nkx2-like
fer3l-1
pax6A
fer3l-2
fer3l-3

Figure 2.6. Subclustering identifies neoblast subpopulations.

(**A**) t-SNE representation of 22 clusters generated from subclustering cells with *smedwi-1* expression ≥ 2.5 [ln(UMI-per-10,000 + 1)]. Identity of numbered clusters unknown. PP, parenchymal; PN, protonephridia. Intestine cluster is indicated by lower expression of *smedwi-1* and enriched *gata4/5/6* and *hnf-4* expression. (**B**) *smedwi-1*$^+$ t-SNE plots colored by *prox-1* and *zfp-1* expression. (**C**) Double FISH image for dd_10988 and *smedwi-1*. Yellow arrows highlight coexpression; white arrows denote absence of coexpression. (**D**) *smedwi-1*$^+$ t-SNE plots colored by expression of differentiated tissue-enriched genes. Arrows indicate gene expression sites. (**E**) Left: *smedwi-1*$^+$ t-SNE plots colored by *gata4/5/6* and *hnf-4* expression. Right: All cluster t-SNE plots colored by *gata4/5/6 and hnf-4* expression. C, *cathepsin*$^+$ cells; I, intestine; γ-Nb, γ-neoblasts. (**F**) *smedwi-1*$^+$ t-SNE plots colored by *ETS1* (dd2092) and *FOXF1* (dd6910) expression. (**G**) Epidermal cell maturation stages. (**H**) t-SNE representation of epidermal subclusters. FISH images labeled by their associated cluster(s) are shown. (**I**) Epidermal t-SNE plots colored by epidermal lineage marker expression from (G). (**J**) dd_554 FISH. (**K**) Pharynx t-SNE plots colored by *smedwi-1* and dd_554 expression. (**L**) Pharynx t-SNE plot colored by the body region from which each cell was isolated. (**M** and **N**) t-SNE plots, colored by *smedwi-1* expression, generated by subclustering cells identified as (M) protonephridia, intestine, muscle, *cathepsin*$^+$, neural, and (N) parenchymal. (**O**) Parenchymal t-SNE plots colored by expression of eight transcription factor–encoding genes enriched in (N). Arrows indicate gene expression sites. Scale bars, 50 μm (C), 200 μm [(H) and (J)].

Figure 2.7

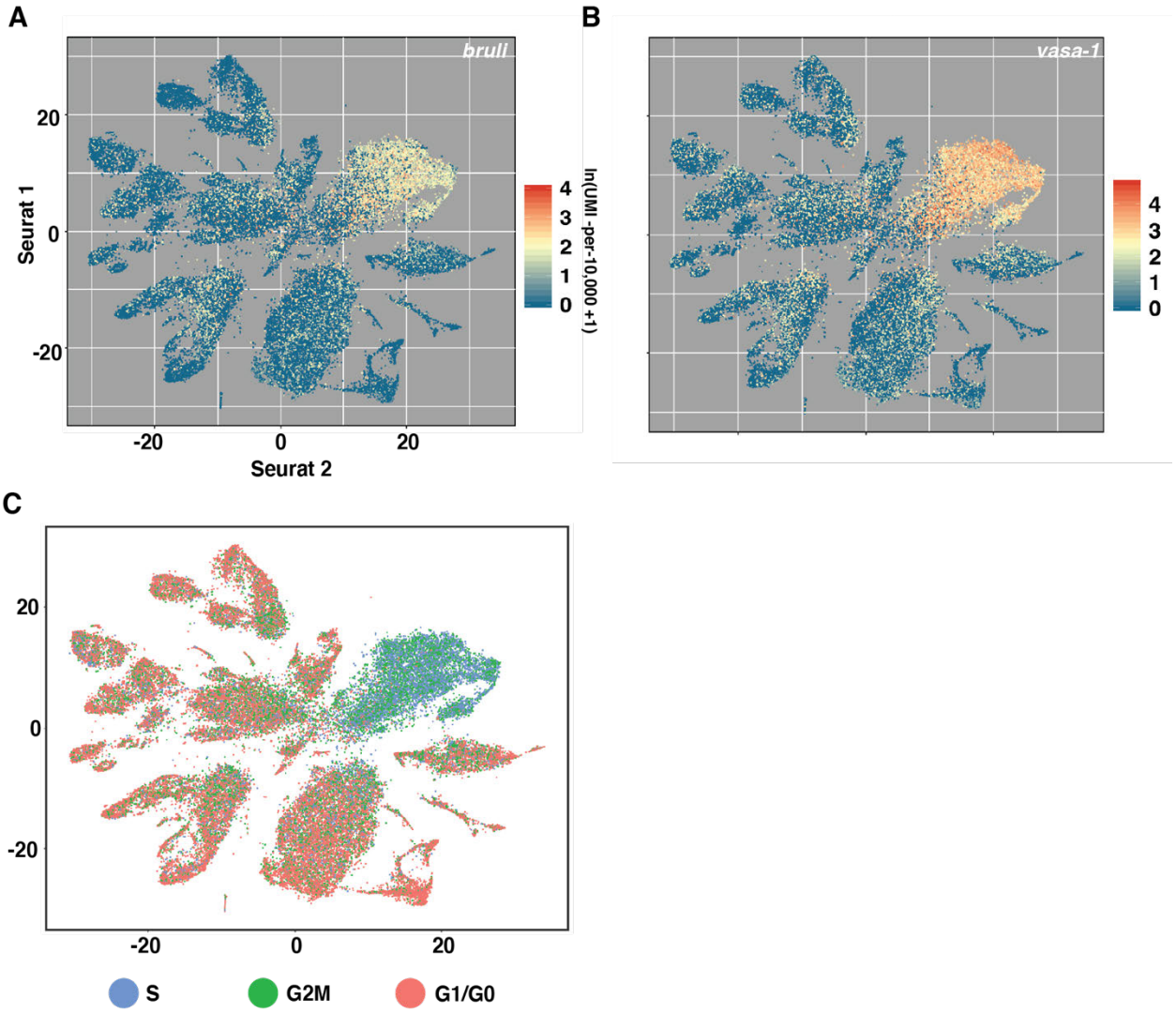

A
brul1
20
0
-20
Seurat 1
-20
0
20
Seurat 2
4
3
2
1
0
ln(UMI -per-10,000 +1)
B
vasa-1
4
3
2
1
0
C
20
0
-20
-20
0
20
S
G2M
G1/G0

Figure 2.7. Cells expressing additional neoblast markers and exhibiting a S/G2M cell cycle status display similar patterns to *smedwi-1*$^{+}$ cells in the t-SNE plot.

(**A** and **B**) t-SNE plot colored by the expression of two neoblast markers, (A) *bruli* (***31***) and (B) *vasa-1* (***30***). (**C**) t-SNE plot colored by the cell cycle status of each cell (Methods).

Figure 2.8

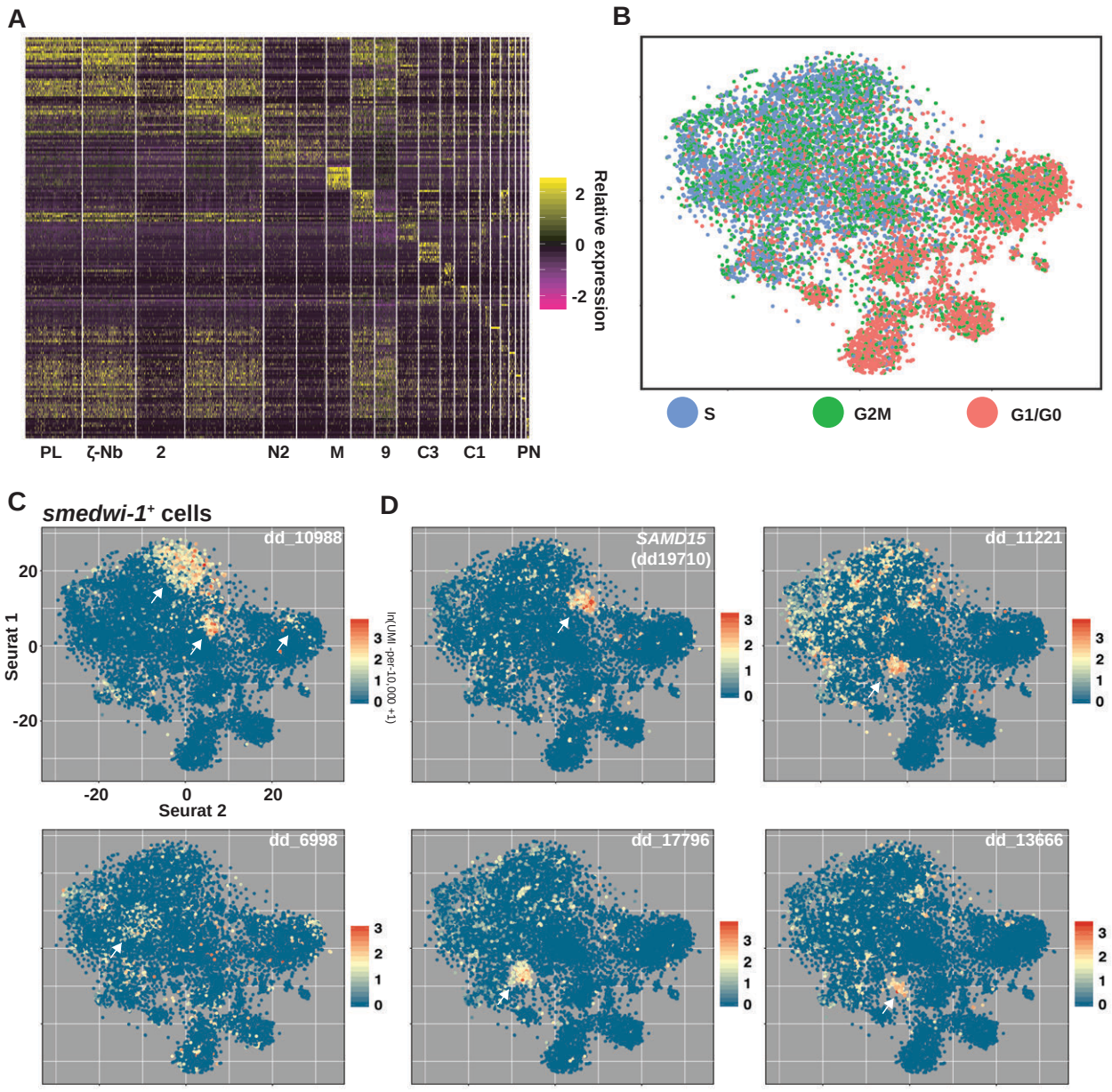

A
Relative expression
2
0
-2
PL
ζ-Nb
2
N2
M
9
C3
C1
PN
B
S
G2M
G1/G0
C
smedwi-1+ cells
dd_10988
Seurat 1
Seurat 2
20
0
-20
ln(UMI-per-10,000 +1)
D
SAMD15
(dd19710)
dd_11221
dd_6998
dd_17796
dd_13666

Figure 2.8. Additional subclusters were identified from subclustering *smedwi-1*$^+$ cells.

(**A**) Heat map of the expression of the top 10 genes from each cluster of the *smedwi-1*$^+$ cell clustering, grouped by cluster ID. Cells, columns; Genes, rows. PL, *PLOD1* (dd3457)$^+$; ζ-Nb, ζ Neoblasts; N2, Neural 2; M, Muscle; C3, *cathepsin*$^+$ 3; C1, *cathepsin*$^+$ 1; PN, Protonephridia. (**B**) *smedwi-1*$^+$ t-SNE plot colored by the cell cycle status of each cell. (**C** and **D**) *smedwi-1*$^+$ t-SNE plots colored by expression of (C) dd_10988 and (D) cluster-specific genes not included in Figure 2.6.

Figure 2.9

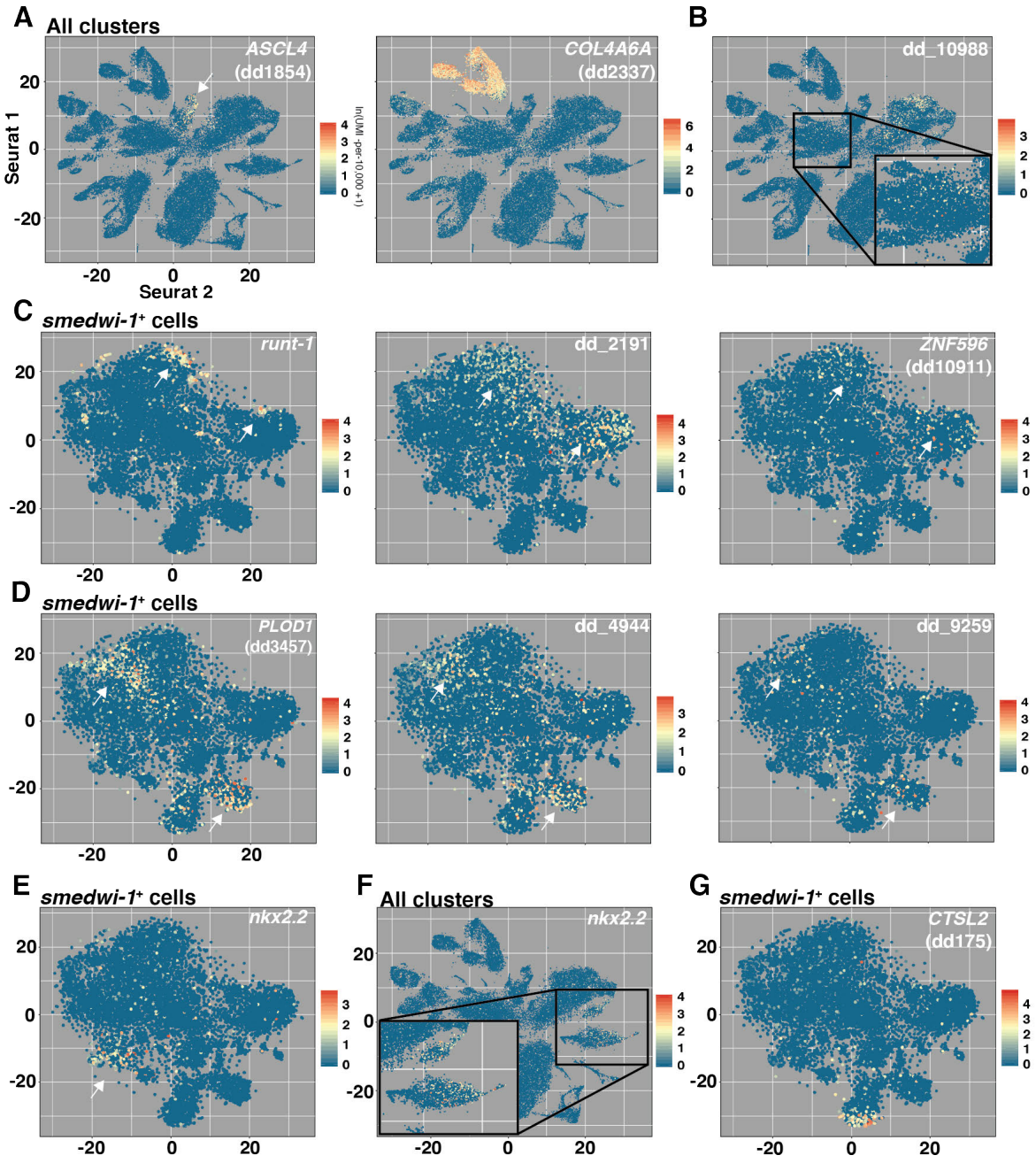

A
All clusters
ASCL4
(dd1854)
COL4A6A
(dd2337)
B
dd_10988
Seurat 1
Seurat 2
ln(UMI-per-10,000+1)
C
smedwi-1+ cells
runt-1
dd_2191
ZNF596
(dd10911)
D
smedwi-1+ cells
PLOD1
(dd3457)
dd_4944
dd_9259
E
smedwi-1+ cells
nkx2.2
F
All clusters
nkx2.2
G
smedwi-1+ cells
CTSL2
(dd175)

Figure 2.9. Additional data regarding the *smedwi-1*$^+$ cell subclustering.

(**A** and **B**) All clusters t-SNE plots colored by expression of (A) the parenchymal marker *ASCL4* (dd1854) and the muscle marker *COL4A6A* (dd2337), and (B) dd_10988. (**C**) *smedwi-1*$^+$ t-SNE plots colored by expression of genes enriched in the dd_10988$^+$ cluster and also expressed in the two neural clusters. (**D**) *smedwi-1*$^+$ t-SNE plots colored by expression of genes enriched in the *PLOD1* (dd3457)$^+$ cluster and also expressed in the muscle cluster. (**E** and **F**) (E) *smedwi-1*$^+$ and (F) all clusters t-SNE plots colored by expression of *nkx2.2*. (**G**) *smedwi-1*$^+$ t-SNE plot colored by expression of the cathepsin+ cell marker *CTSL2* (dd175). Arrows indicate sites of gene expression.

Figure 2.10

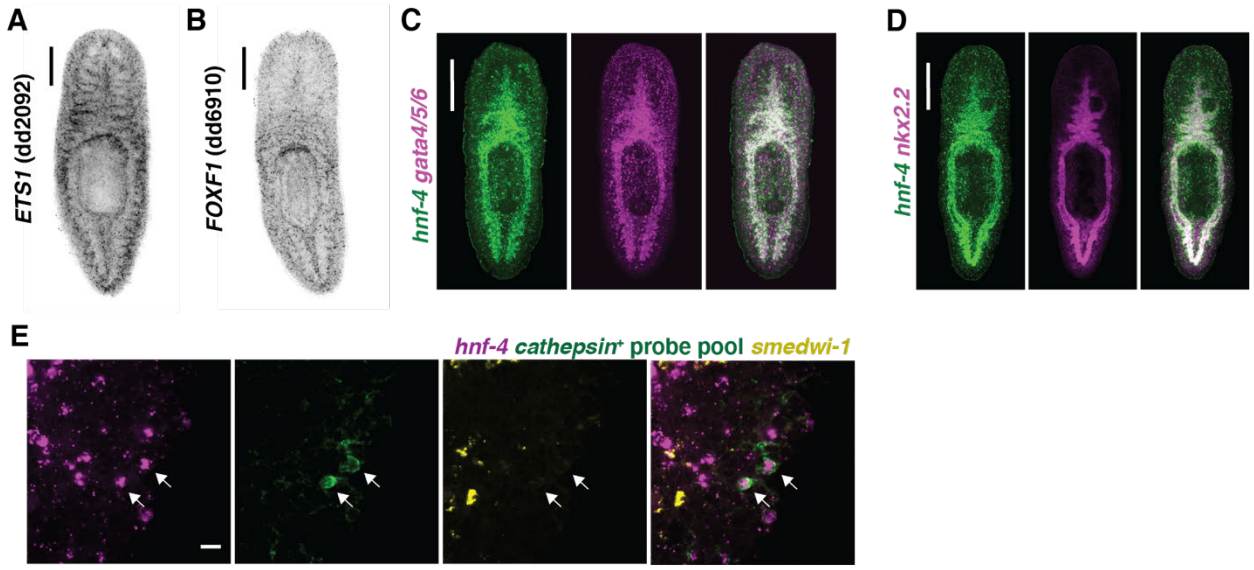

A
ETS1 (dd2092)
B
FOXF1 (dd6910)
C
hnf-4 gata4/5/6
D
hnf-4 nkx2.2
E
hnf-4 cathepsin+ probe pool smedwi-1

Figure 2.10. Additional characterization of putative intestine and *cathepsin*$^+$ cell progenitors.

(**A** and **B**) FISH images of the *cathepsin*$^+$ cell-enriched transcription factors (A) *ETS1* (dd2092) and (B) *FOXF1* (dd6910). (**C** and **D**) Double FISH images of *hnf-4* and (C) *gata4/5/6-1* and (D) *nkx2.2*. (**E**) FISH images of *hnf-4*, *smedwi-1*, and a pool of the *cathepsin*$^+$ cell markers *CTSL2* (dd582), *PTPRT* (dd10872), *TTPA* (dd6149), dd_5690, *pgbd-1*, dd_7593, and *AQP1* (dd1103). White signal in merged images indicates a positional overlap in gene expression. Arrows indicate cells positive for *hnf-4* and the pool of *cathepsin*$^+$ cell markers, but negative for *smedwi-1*. Scale bars: A-D, 200 µm; E, 10 µm.

Figure 2.11

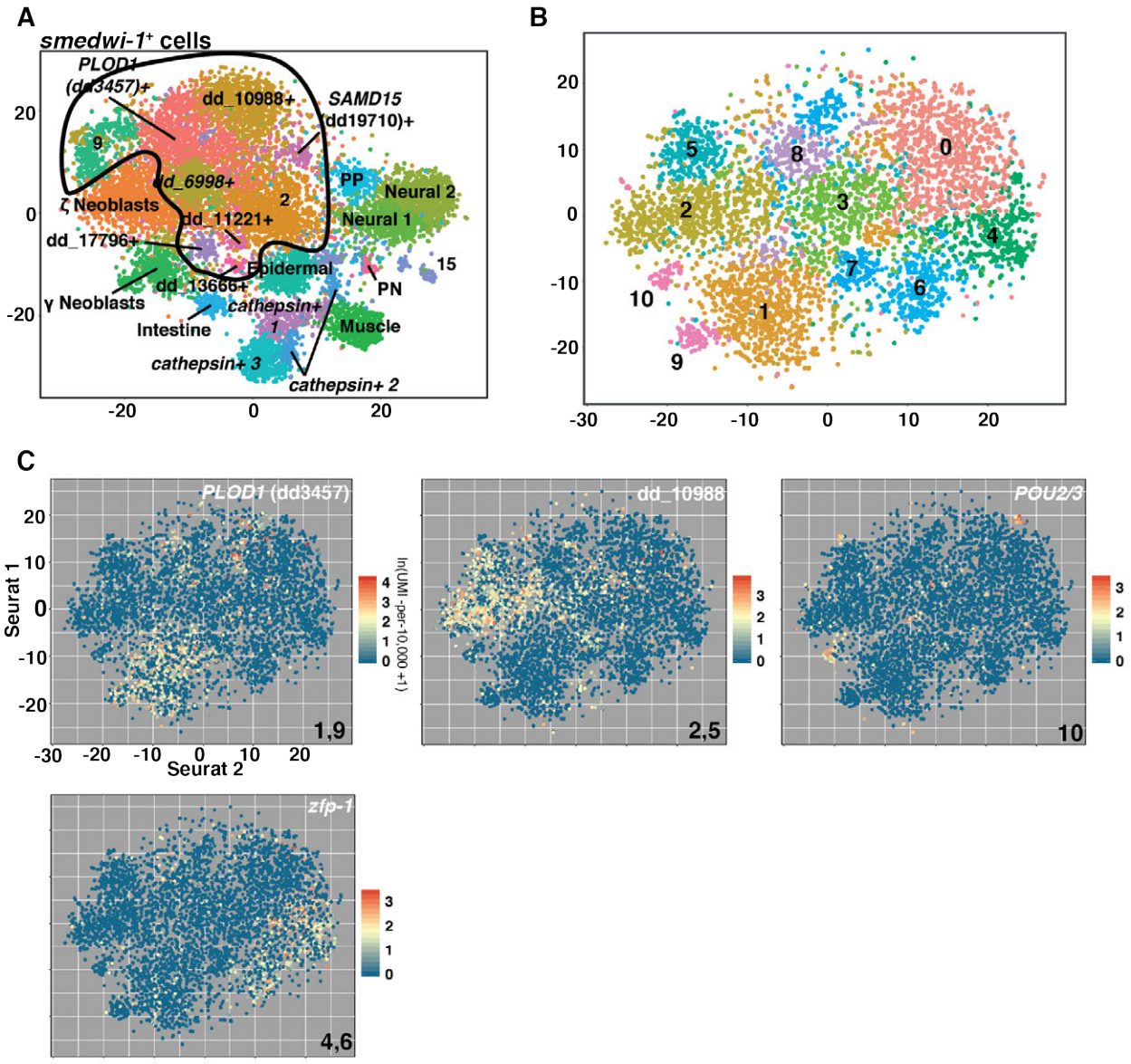

A
smedwi-1+ cells
PLOD1 (dd3457)+
dd_10988+
SAMD15 (dd19710)+
9
dd_6998+
PP
Neural 2
ζ Neoblasts
2
dd_11221+
Neural 1
dd_17796+
Epidermal
15
PN
dd_13666+
γ Neoblasts
cathepsin+ 1
Intestine
Muscle
cathepsin+ 3
cathepsin+ 2
B
C
PLOD1 (dd3457)
dd_10988
POU2/3
zfp-1
Seurat 1
Seurat 2
ln(UMI-per-10,000 +1)
1,9
2,5
10
4,6

Figure 2.11. Additional subclustering of *smedwi-1*$^+$ cell clusters with high levels of *smedwi-1*.

(**A**) t-SNE representation of *smedwi-1*$^+$ cell subclustering overlaid with a boundary indicating *smedwi-1* high clusters further subclustered in (B). (**B**) t-SNE representation of 11 clusters generated from further subclustering of (A). (**C**) *smedwi-1* high t-SNE plots colored by cluster-enriched gene expression. Numbers indicate the associated *smedwi-1* high subcluster(s).

Figure 2.12

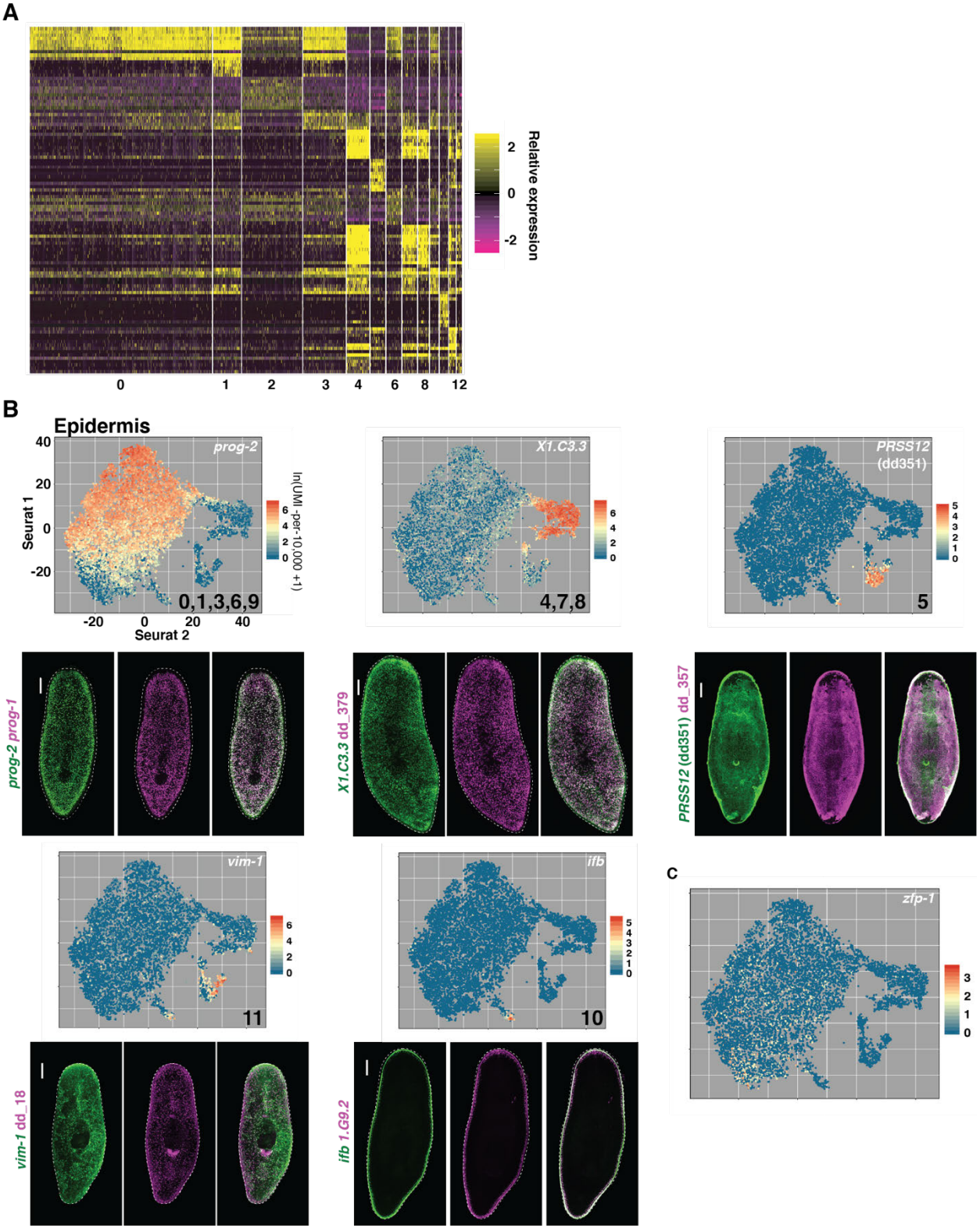

A
Relative expression
2
0
-2
0
1
2
3
4
6
8
12
B
Epidermis
prog-2
ln(UMI-per-10,000 +1)
Seurat 1
Seurat 2
0,1,3,6,9
X1.C3.3
4,7,8
PRSS12 (dd351)
5
prog-2 prog-1
X1.C3.3 dd_379
PRSS12 (dd351) dd_357
vim-1
11
ifb
10
vim-1 dd_18
ifb 1.G9.2
C
zfp-1

Figure 2.12. Additional characterization of the epidermal subclusters.

(**A**) Heat map of the expression of the top 10 genes from each cluster of the epidermal clustering, grouped by cluster number. Cells, columns; Genes, rows. (**B**) Top panels: Epidermal t-SNE plots colored by cluster-enriched gene expression. Numbers indicate the associated epidermal subcluster(s). Bottom panels: FISH images of two cluster-enriched genes. White signal in merged images indicates a positional overlap in gene expression. Scale bars: 200 µm. (**C**) Epidermal t-SNE plot colored by the expression of *zfp-1*.

Figure 2.13

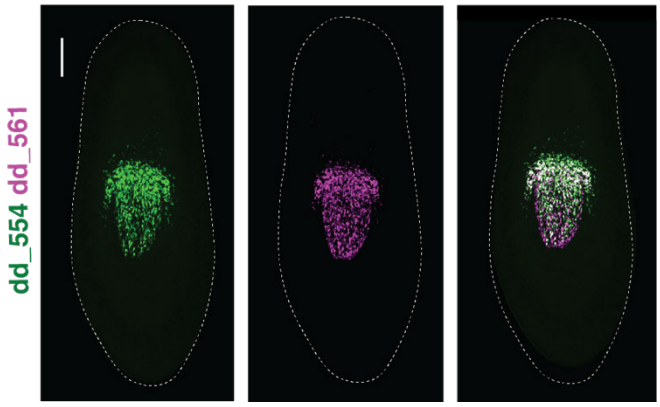


Figure 2.13. Additional data regarding cells of the pharynx lineage.

FISH images of two genes expressed in pharynx progenitor cells. White signal in merged images indicates a positional overlap in gene expression. Scale bar, 200 μm.

Subclustering cells by tissue type uncovers rare cell types

The protonephridia, the planarian excretory and osmoregulatory system, contains flame cells for filtering fluids, proximal and distal tubule cells, and a collecting duct (***33, 39, 40***). The protonephridia is a model tissue for studying organ regeneration and the evolution of kidney-like excretory systems. Subclustering of 890 protonephridia cells (Figure 2.1C) identified each known protonephridia cell type as a separate subcluster, revealing the complete transcriptomes of these cells (Figures 2.14A and 2.15, A to C). Furthermore, two protonephridia subclusters with *smedwi-1*$^+$ cells were identified (Figure 2.6M). One was enriched in flame cell gene expression (e.g., dd_2920) and the other in a proximal tubule marker (dd_10830), which suggests that they might be flame and tubule cell precursors, respectively (Figure 2.15, D and E).

Less is known regarding the full complement of cell types in other planarian tissues. Ultrastructural studies suggested that the planarian intestine contains two cell types: absorptive enterocytes and secretory goblet cells (***41, 42***). Subclustering of 3025 intestinal cells (Figure 2.1C) revealed three distinct cell populations (clusters 4, 5, and 8) (Figures 2.14B and 2.16, A and B). FISH with subcluster-enriched markers (Table 2.2) revealed distinct intestine components. Cluster 4 represented an inner intestine cell layer (Figure 2.14C) and was enriched for absorptive enterocyte markers (***43***). Cluster 8 cells were largely present within the primary intestine branches, resembling the pattern of goblet cells (***44***). A third group (cluster 5) represented an outer intestine cell layer and displayed a set of enriched genes different from that of clusters 4 and 8 (Figure 2.14C and Table 2.2). In addition to these three main intestine components, clusters representing putative transition states were also identified. Clusters 1, 3, and 6 included many *smedwi-1*$^+$ cells (Figure 2.6M). Genes with enriched expression in clusters 0 and 7 displayed expression spanning into the enterocyte cluster (cluster 4), suggesting these might be enterocyte transition states (Figure 2.17A). Genes with enriched expression in clusters 2 and 3 displayed expression spanning into the outer intestine cluster (cluster 5) and might reflect transition or variant states of these cells (Figure 2.17B). The Monocle toolkit can be used to predict cellular transitions in lineages (***45***)

and was used to build single-cell trajectories for the enterocyte and outer intestine cell lineages, closely recapitulating the candidate transition states identified by Seurat (Figures 2.14D, 2.18, and Table 2.3).

Several transcription factors required for the specification of various planarian cell types have been identified with RNA interference (RNAi) and gene expression studies. Because of constant tissue turnover, RNAi of transcription factor–encoding genes expressed in specific classes of specialized neoblasts in adult planarians can lead to steady depletion of the cell type generated by that specialized neoblast class (***33, 46, 47***). The transcriptomes identified here generate a resource of enriched gene expression for different cell types, including transcription factor–encoding genes. Accordingly, inhibition of the transcription factor–encoding *PTF1A* (dd6869) gene, which had enriched expression in candidate transition states for the outer intestine cluster, strongly reduced this cell population while not affecting absorptive enterocytes of the intestine (Figure 2.14E).

The nervous system displays by far the greatest known cell type composition complexity of the major planarian tissues. By subclustering 11,907 neuronal cells (Figure 2.1C), we identified 61 distinct subclusters representing a diversity of cell types and states (Figures 2.19A and 2.20A). Twelve subclusters had high *smedwi-1* expression, which suggested that they represent neuronal precursors (Figure 2.20B). Cluster 10 contained cells of the brain branches, as determined by expression of *gpas* (***48***) and *pds* (***49***) (Figures 2.19B and 2.20C). Three subclusters (clusters 3, 7, and 8) were defined by expression of *pc2* (encoding a neuropeptide-processing proprotein convertase) as well as an assortment of markers for rare neuron classes in the cephalic ganglia and ventral nerve cords (Figures 2.19C and 2.20, D and E). We also sequenced an additional 7766 cells from the brain region to expand the number of cells in these clusters (Figure 2.20, F to H). In addition to these large clusters, there existed a number of smaller, compact, and well-separated subclusters. These could be further divided into ciliated and nonciliated neurons according to the expression of *rootletin* (dd6573), which encodes a ciliary rootlet component (Figures 2.19D and 2.21A). Because of further heterogeneity

within these clusters (e.g., *opsin*$^+$ presumptive photoreceptors were present together, but not as a separate cluster), data from these two cell sets (ciliated, not ciliated) were each taken in isolation for further subclustering. This yielded 37 nonciliated neuron subclusters (Figures 2.19E and 2.21B, 2.22, A and B, and 2.23 to 2.25) and 25 putatively ciliated neuron subclusters (Figures 2.19F and 2.21B, 2.26, A and B, and 2.27B). We assessed the localization of cells associated with 46 of 62 of these subclusters by FISH using subcluster-specific markers. The observed cell types had a wide range of patterns including rare cell types such as photoreceptor neurons (Figures 2.19, E and F, and 2.22 to 2.27). Many genes had enriched expression in multiple clusters; the distribution of neural cell types they represented was defined by a combinatorial set of markers (Figures 2.22 to 2.27). A number of identified cell types from different subclusters displayed similar localization patterns. However, FISH demonstrated no overlap in subcluster-specific markers, consistent with the SCS data (Figure 2.19G). For several neural subtypes, we found *smedwi-1*$^+$ candidate precursor cells. Four nonciliated neuron subclusters (subclusters 1, 2, 4, and 12) and a single ciliated neuron subcluster (subcluster 1) were enriched in *smedwi-1* expression (Figure 2.28A). Nonciliated neuron subcluster 4 also expressed *gata4/5/6*, as did six *smedwi-1*$^-$ clusters (clusters 14,16/33, 24, 26, and 32) that radiated out from central *smedwi-1*$^+$ cells, raising the possibility that these *smedwi-1*$^+$ cells constitute precursors for these populations (Figure 2.28B).

The pharynx is a muscular tube used for feeding and defecation (***10***). It is contained within an epithelial cavity and connects to the intestine at its anterior end via an esophagus. Pharyngeal muscle cells and pharyngeal neurons clustered together with the other muscle cells and neurons of the body (Figures 2.29A and 2.30A). Other pharynx-associated cells, including cells from isolated pharynges and surrounding tissue, constituted the other major pharynx clusters. These non-neural, non-muscle pharynx and pharynx-associated cells (Figure 2.1C, *n* = 1083 cells) were subclustered, and FISH was performed on cluster-enriched markers (Figures 2.29, B and C, and 2.31A). Subclusters included pharyngeal cavity epithelium cells (clusters 7 and 8), the

epithelial pharynx lining (clusters 1 and 5), the mouth and esophagus (cluster 9), cells near the pharynx opening (cluster 6), and cells that constitute the connection to the planarian body (cluster 4). FISH confirmed nonoverlapping expression patterns for markers of tested separate cell populations (Figure 2.31B).

Planarian muscle expresses *collagen* in addition to canonical muscle genes such as *troponin* and *tropomyosin* (***19***). Muscle exists in a subepidermal body wall layer, in the pharynx, surrounding the intestine, and in a DV domain (***50***). Subclustering 5014 muscle cells (Figure 2.1C) revealed seven *smedwi-1*$^{+}$ candidate precursor subclusters (clusters 0, 1, 3, 4, 5, 10, and 11) (Figure 2.6M), as well as subclusters containing body wall muscle (cluster 7), pharyngeal muscle (cluster 2, 8, 9, and 12) (Figure 2.30A), a population of muscle cells enriched around the intestine (cluster 6), and an unidentified population (cluster 13) (Figures 2.30, B and C, and 2.31C). Markers for body wall muscle (cluster 7) and cluster 13 were expressed in nonoverlapping cells by FISH (Figure 2.31D).

Whereas some molecular characterization existed for the seven broad planarian tissue classes previously mentioned, very little is known regarding the cellular composition of the two remaining classes. The parenchymal class (Figure 2.1C) (***24***) was highly heterogeneous, with subclustering of 2120 cells identifying many distinct cell populations (Figures 2.31E and 2.32, A and B, and 2.33B). In addition to eight *smedwi-1*$^{+}$ putative precursor subclusters (clusters 0, 1, 2, 3, 4, 6, 8, and most of 9) (Figure 2.6N), parenchymal cell subclustering revealed 13 well-separated differentiated cell subclusters. FISH showed that each of these differentiated cell populations were present as scattered cells, presumably within a mesenchymal tissue layer called the parenchyma that surrounds major planarian organs (***10***). Previous morphological studies determined that the parenchyma is composed of multiple gland cells, neoblasts, and “fixed parenchymal cells” characterized through histological and electron microscopy studies as a likely phagocytic cell with long cellular processes filling most of the parenchymal space (***10***, ***51***, ***52***). Some identified parenchymal subclusters appeared to be gland cells, displaying processes extending to the epidermis, defining

transcriptomes for these cells. Candidate gland cell types included two that were exclusively dorsal (clusters 16 and 17), two exclusively lateral (clusters 10 and 13, including marginal adhesive gland cells and an unknown cell population), four present both dorsally and ventrally (clusters 7, 11, 14, and 15), and one present ventrally near the brain (cluster 19). Three subclusters (clusters 5, 12, and 18) contained cells with patterns similar to those of planarian neoblasts, but were not neoblasts. Finally, a single subcluster contained large cells surrounding the pharynx (small group of cluster 9 cells) and were enriched for expression of previously identified metalloprotease-encoding genes (***53***). Three pairs of parenchymal subclusters (six subclusters total) were confirmed to exist in nonoverlapping populations by FISH (Figure 2.31F).

The transcription factor–encoding gene *nkx6-like* was expressed in a parenchymal cell population marked by dd_515. Inhibition of *nkx6-like* ablated dd_515 cells, while not affecting a distinct, non-enriched parenchymal cell population marked by dd_385 (Figure 2.31G). These results further highlight the potential to use the data to ablate many specific cell types in the animal.

The final major class of cells, the *cathepsin*$^+$ group, contained 7034 cells (Figure 2.1D). This group of clusters contained recently described glia and pigment cells (***12***, ***13***, ***54***). Subclustering of *cathepsin*$^+$ cells identified four subclusters expressing *smedwi-1* that represented putative precursor cells (clusters 0, 1, 3, and 6) (Figure 2.6M), a glial subcluster (cluster 15), and two pigment cell populations (clusters 11 and 14), identifying transcriptomes for these cell types (Figures 2.34A and 2.35, A and B, and 2.36B). Eight *cathepsin*$^+$ subclusters represented previously unidentified cell populations. FISH revealed striking, elaborate morphologies for most of these cells, involving long processes and unique distributions (Figures 2.34A and figs. 2.35B and 2.36B). Cells from subclusters 5 and 10 were spread throughout the planarian body, with long processes filling substantial parenchyma space. Subcluster 8 represented cells specific to the pharynx. Subcluster 9 cells were scattered throughout the animal. Subclusters 4 and 16 identified cells with dense aggregated foci of elaborate processes at scattered locations throughout the animal that lacked definitive positions—an unusual

and unanticipated cell type distribution. FISH identified markers labeling cell bodies of these cells, revealing that the aggregates comprised many cells (Figure 2.34B). Subclusters 12 and 13 also exhibited processes with visible cell bodies. Subcluster 12 cells were largely subepidermal. The most elaborate of these newly identified cells (subclusters 5 and 10) were excluded from the intestine and brain, but had processes around the branches of the intestine and protonephridia and interspersed within the cephalic ganglia (Figures 2.34, C to E, and 2.37, A and B). FISH confirmed nonoverlapping expression patterns for two tested subclusters (Figure 2.37C).

Subcluster 7 of the *cathepsin*$^+$ group of cells was enriched in expression of genes with expression spanning into clusters 5 and 10 (Figure 2.38A). Similarly, expression of cluster 2 marker genes spanned into clusters 4 and 16 (Figure 2.38B). These cells might reflect transition or variant states of cells for clusters 5/10 and 4/16, respectively. SMEDWI-1 protein perdures in neoblast progeny after loss of *smedwi-1* mRNA, allowing detection of newly produced neoblast progeny (***31***). *MAP3K5* (dd4849)$^+$ cells, which were predicted to be expressed in cells transitioning from the *smedwi-1*$^+$ state in the *cathepsin*$^+$ cell plot, were SMEDWI-1$^+$/*smedwi-1*$^-$, supporting the interpretation that these cells are progenitors in the *cathepsin*$^+$ cell lineage (Figure 2.38, C and D). The Monocle toolkit was also used to build single-cell trajectories for these clusters, with data closely recapitulating the transition states identified by Seurat (Figures 2.34F, 2.39, and Table 2.3).

Figure 2.14

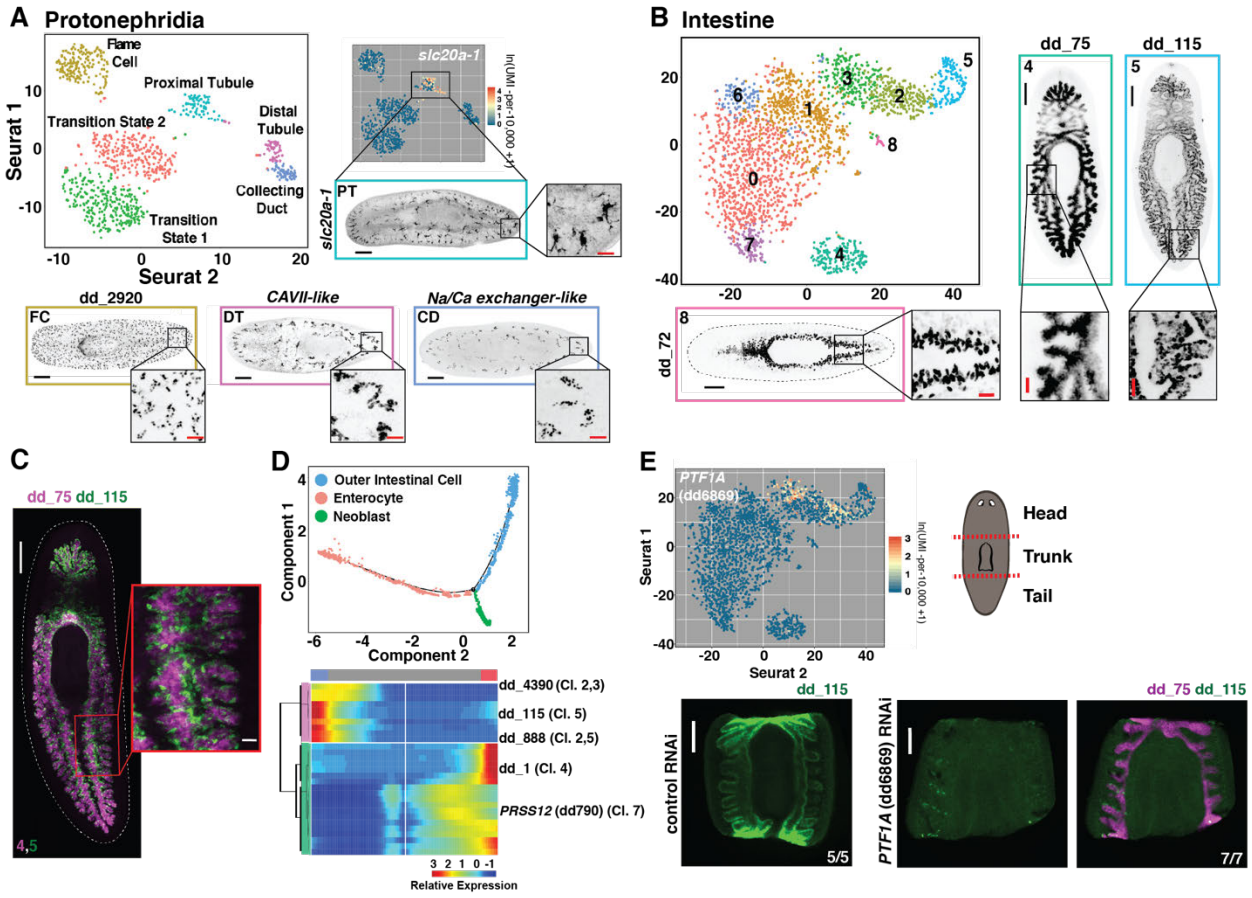

A Protonephridia
Flame Cell
Proximal Tubule
Distal Tubule
Transition State 2
Collecting Duct
Transition State 1
Seurat 1
Seurat 2
slc20a-1
PT
dd_2920
FC
CAVII-like
DT
Na/Ca exchanger-like
CD
B Intestine
dd_75
dd_115
dd_72
C
dd_75 dd_115
4,5
D
Outer Intestinal Cell
Enterocyte
Neoblast
Component 1
Component 2
dd_4390 (Cl. 2,3)
dd_115 (Cl. 5)
dd_888 (Cl. 2,5)
dd_1 (Cl. 4)
PRSS12 (dd790) (Cl. 7)
Relative Expression
E
PTF1A (dd6869)
Head
Trunk
Tail
control RNAi
PTF1A (dd6869) RNAi
5/5
7/7

Figure 2.14. Subclustering of tissues reveals transcriptomes for known and novel cell populations.

(**A**) t-SNE representation of the protonephridial subcluster. FISH images are labeled by their associated cluster. (**B**) t-SNE representation of intestinal subclusters. (**C**) Double FISH images of genes enriched in separate intestinal subclusters. Numbers indicate the associated subcluster for each marker. (**D**) Top: Cell trajectory of enterocyte and outer intestinal cell lineages produced by Monocle. Cells are colored by identity. Bottom: Heat map of branch dependent genes (*q* value < 10^{-145}) across cells plotted in pseudo-time (***45***). Cells, columns; genes, rows. Beginning of pseudo-time is at center of heat map. "Cl." annotation indicates a log-fold enrichment ≥ 1 of the gene in that intestine Seurat cluster. (**E**) Top left: Intestine t-SNE plot colored by expression of *PTF1A* (dd6869). Top right: Illustration of cutting scheme used to generate fragments. Bottom: dd_115 and dd_75 FISH of control and *PTF1A* (dd6869) RNAi animals. Animals were cut and fixed 23 days after the start of double-stranded RNA (dsRNA) feedings. Scale bars: whole-animal/fragment images, 200 µm; insets, 50 µm.

Figure 2.15

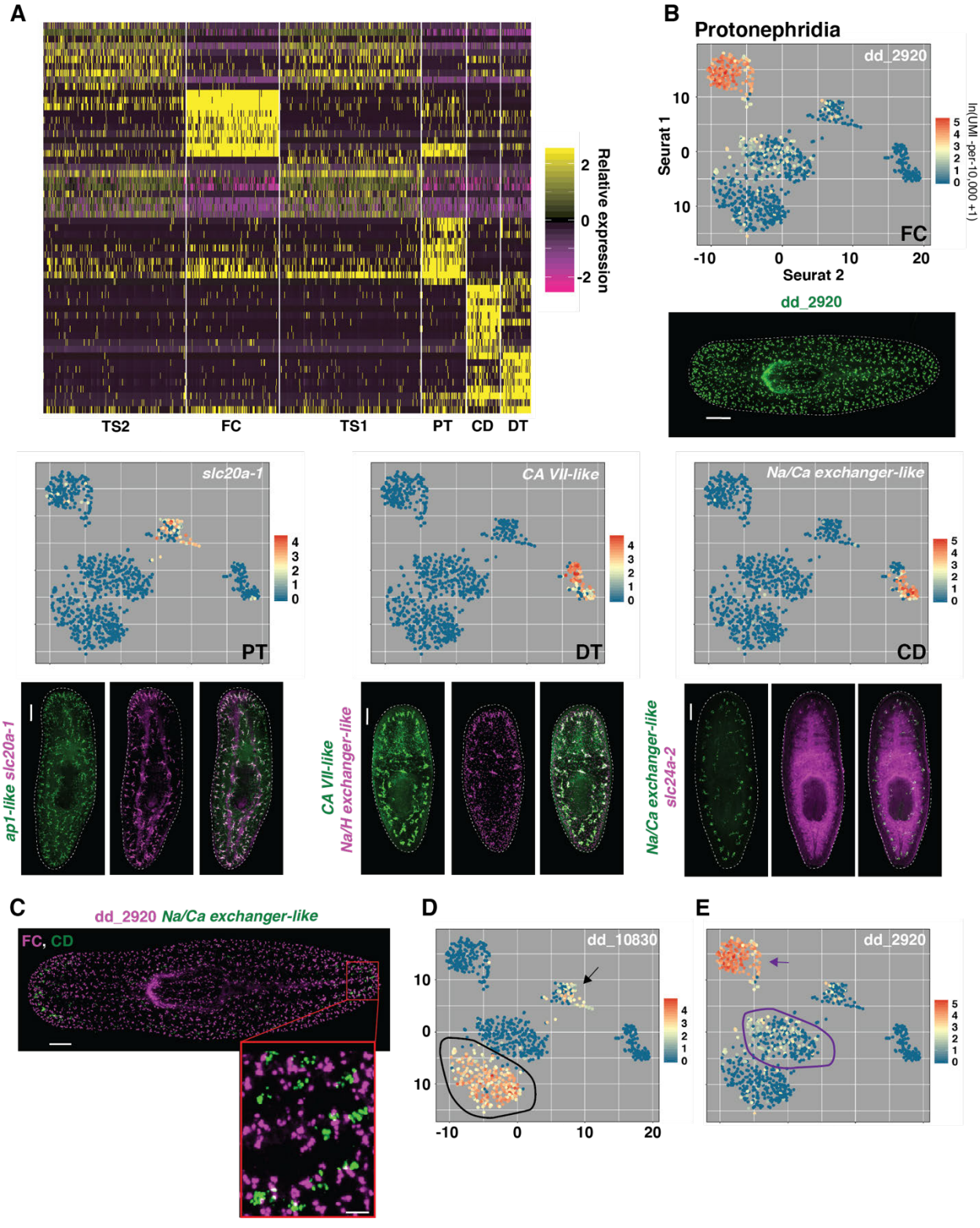

A
TS2
FC
TS1
PT
CD
DT
Relative expression
B
Protonephridia
dd_2920
Seurat 1
Seurat 2
FC
ln(UMI -per-10,000 +1)
dd_2920
slc20a-1
PT
CA VII-like
DT
Na/Ca exchanger-like
CD
ap1-like slc20a-1
CA VII-like
Na/H exchanger-like
Na/Ca exchanger-like
slc24a-2
C
dd_2920 Na/Ca exchanger-like
FC, CD
D
dd_10830
E
dd_2920

Figure 2.15. Additional characterization of the protonephridia subclusters.

(**A**) Heat map of the expression of the top 10 genes from each cluster of the protonephridia clustering, grouped by cluster ID. Cells, columns; Genes, rows. CD, Collecting Duct; DT, Distal Tubule; FC, Flame Cell; PT, Proximal Tubule; TS1, Transition State 1; TS2, Transition State 2. (**B**) Top panels: Protonephridia subcluster t-SNE plots colored by cluster-enriched gene expression. Label indicates the associated protonephridia subcluster. Bottom panels: FISH images of one or two cluster-enriched genes. (**C**) Double FISH image of two protonephridia markers enriched in separate clusters. Colored labels indicate the associated protonephridia subcluster for each marker, demonstrating a lack of co-expression. (**D** and **E**) Protonephridia t-SNE plots colored by expression of (D) dd_10830, which marks the proximal tubule cluster (black arrow) and the transition state 1 cluster (black circle), and (E) dd_2920, which marks the flame cell cluster (purple arrow) and the transition state 2 cluster (purple circle). White signal in merged images indicates a positional overlap in expression. Scale bars: whole-animal images, 200 μm; inset, 50 μm.

Figure 2.16

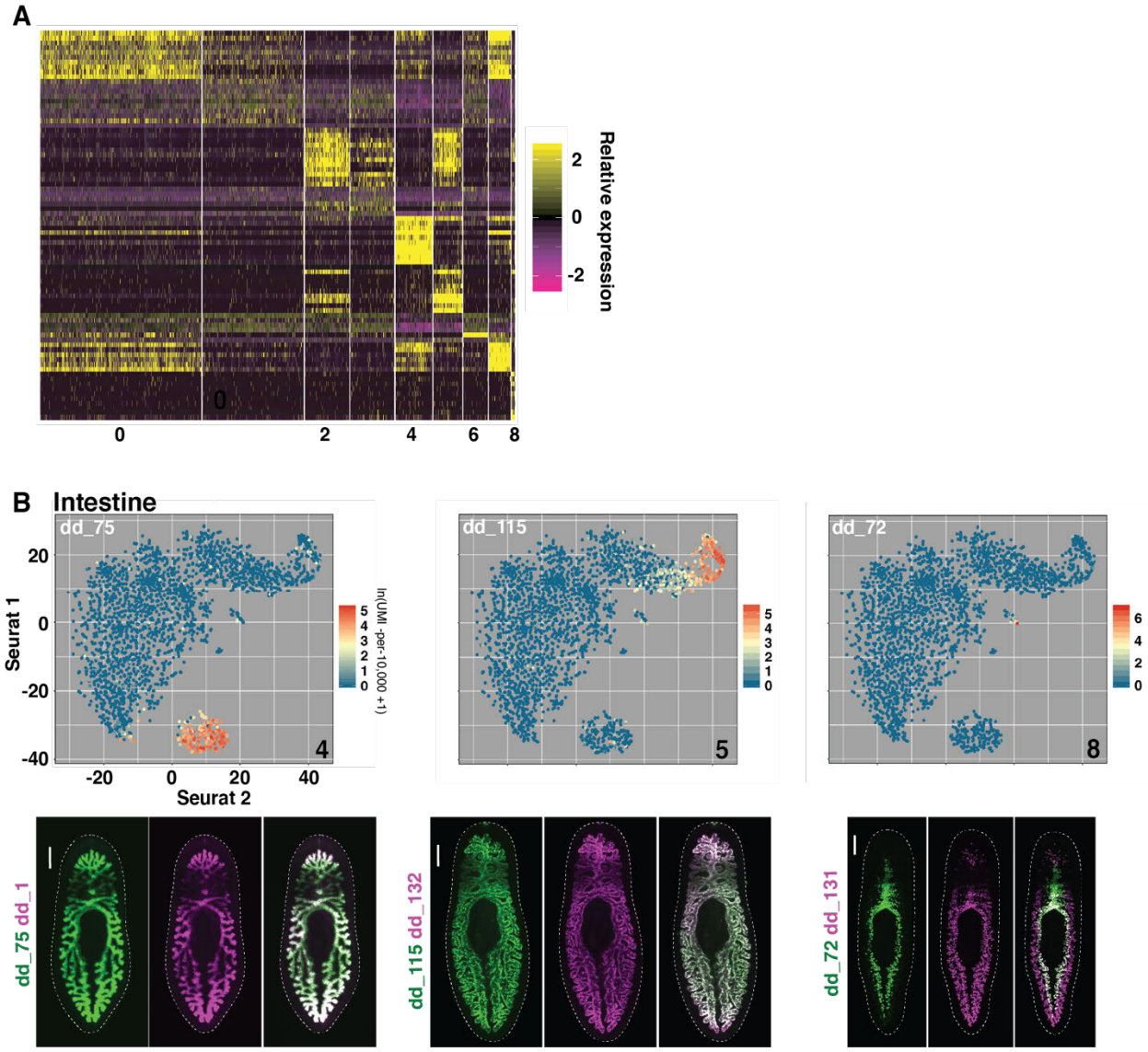

A
Relative expression
2
0
-2
0
0
2
4
6
8
B Intestine
dd_75
dd_115
dd_72
Seurat 1
Seurat 2
ln(UMI per 10,000 +1)
4
5
8
dd_75 dd_1
dd_115 dd_132
dd_72 dd_131

Figure 2.16. Additional characterization of the major differentiated intestine subclusters.

(**A**) Heat map of the expression of the top 10 genes from each cluster of the intestine clustering, grouped by cluster number. Cells, columns; Genes, rows. (**B**) Top panel: Intestine t-SNE plots colored by cluster-enriched gene expression for the major differentiated cell clusters. Numbers indicate the associated intestine subcluster. Bottom panel: FISH images of two cluster-enriched genes. White signal in merged images indicates a positional overlap in expression between the two genes. Scale bars, 200 μm.

Figure 2.17

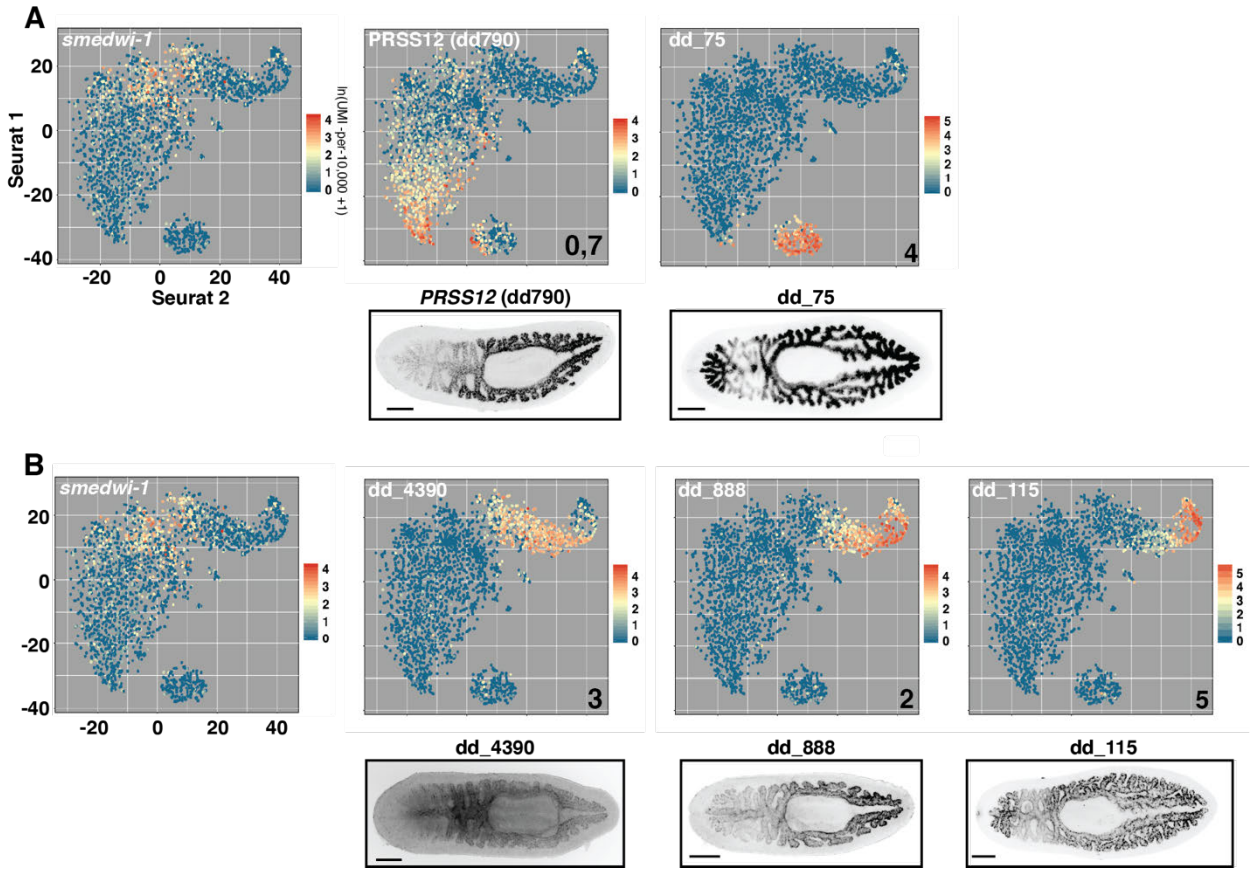

A
smedwi-1
PRSS12 (dd790)
dd_75
0,7
4
Seurat 1
Seurat 2
ln(UMI-per-10,000 +1)
PRSS12 (dd790)
dd_75
B
smedwi-1
dd_4390
dd_888
dd_115
3
2
5
dd_4390
dd_888
dd_115

Figure 2.17. Additional characterization of the putative transition state intestine subclusters.

(**A** and **B**) Top panel: Intestine t-SNE plots colored by cluster-enriched gene expression for the putative transition state clusters. Numbers indicate the associated intestine subcluster. Bottom panel: FISH images of cluster-enriched genes. Scale bars, 200 μm.

Figure 2.18

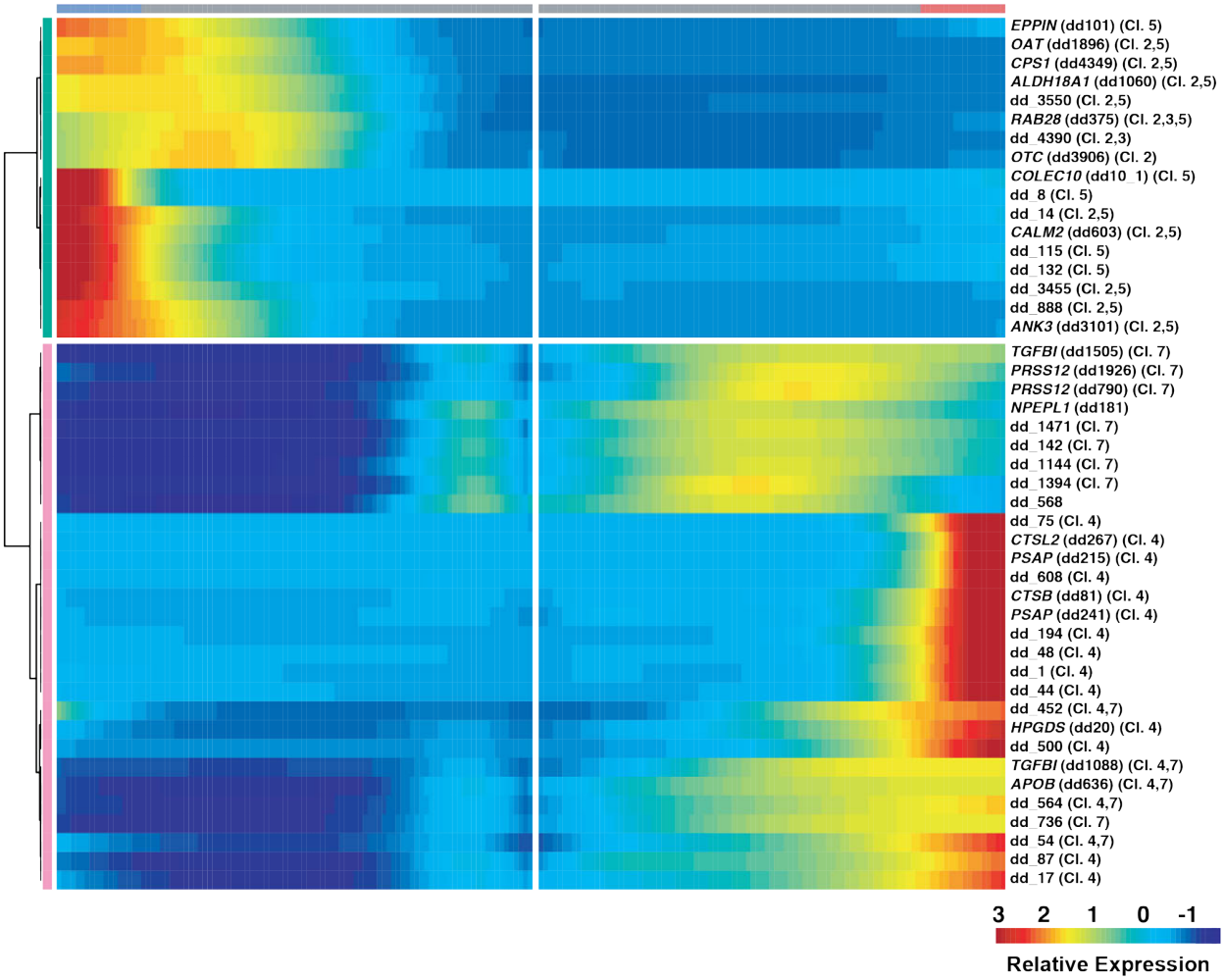

EPPIN (dd101) (Cl. 5)
OAT (dd1896) (Cl. 2,5)
CPS1 (dd4349) (Cl. 2,5)
ALDH18A1 (dd1060) (Cl. 2,5)
dd_3550 (Cl. 2,5)
RAB28 (dd375) (Cl. 2,3,5)
dd_4390 (Cl. 2,3)
OTC (dd3906) (Cl. 2)
COLEC10 (dd10_1) (Cl. 5)
dd_8 (Cl. 5)
dd_14 (Cl. 2,5)
CALM2 (dd603) (Cl. 2,5)
dd_115 (Cl. 5)
dd_132 (Cl. 5)
dd_3455 (Cl. 2,5)
dd_888 (Cl. 2,5)
ANK3 (dd3101) (Cl. 2,5)
TGFBI (dd1505) (Cl. 7)
PRSS12 (dd1926) (Cl. 7)
PRSS12 (dd790) (Cl. 7)
NPEPL1 (dd181)
dd_1471 (Cl. 7)
dd_142 (Cl. 7)
dd_1144 (Cl. 7)
dd_1394 (Cl. 7)
dd_568
dd_75 (Cl. 4)
CTSL2 (dd267) (Cl. 4)
PSAP (dd215) (Cl. 4)
dd_608 (Cl. 4)
CTSB (dd81) (Cl. 4)
PSAP (dd241) (Cl. 4)
dd_194 (Cl. 4)
dd_48 (Cl. 4)
dd_1 (Cl. 4)
dd_44 (Cl. 4)
dd_452 (Cl. 4,7)
HPGDS (dd20) (Cl. 4)
dd_500 (Cl. 4)
TGFBI (dd1088) (Cl. 4,7)
APOB (dd636) (Cl. 4,7)
dd_564 (Cl. 4,7)
dd_736 (Cl. 7)
dd_54 (Cl. 4,7)
dd_87 (Cl. 4)
dd_17 (Cl. 4)
3 2 1 0 -1
Relative Expression

Figure 2.18. Lineage reconstruction of the enterocyte and outer intestine cell lineages.

Heat map of branch dependent genes (q-value < 1E-100) across cells plotted in pseudotime. Cells, columns; Genes, rows. Beginning of pseudotime at center of heatmap. “Cl.” annotation indicates a log-fold enrichment ≥ 1 of that gene in that intestine Seurat cluster.

Figure 2.19

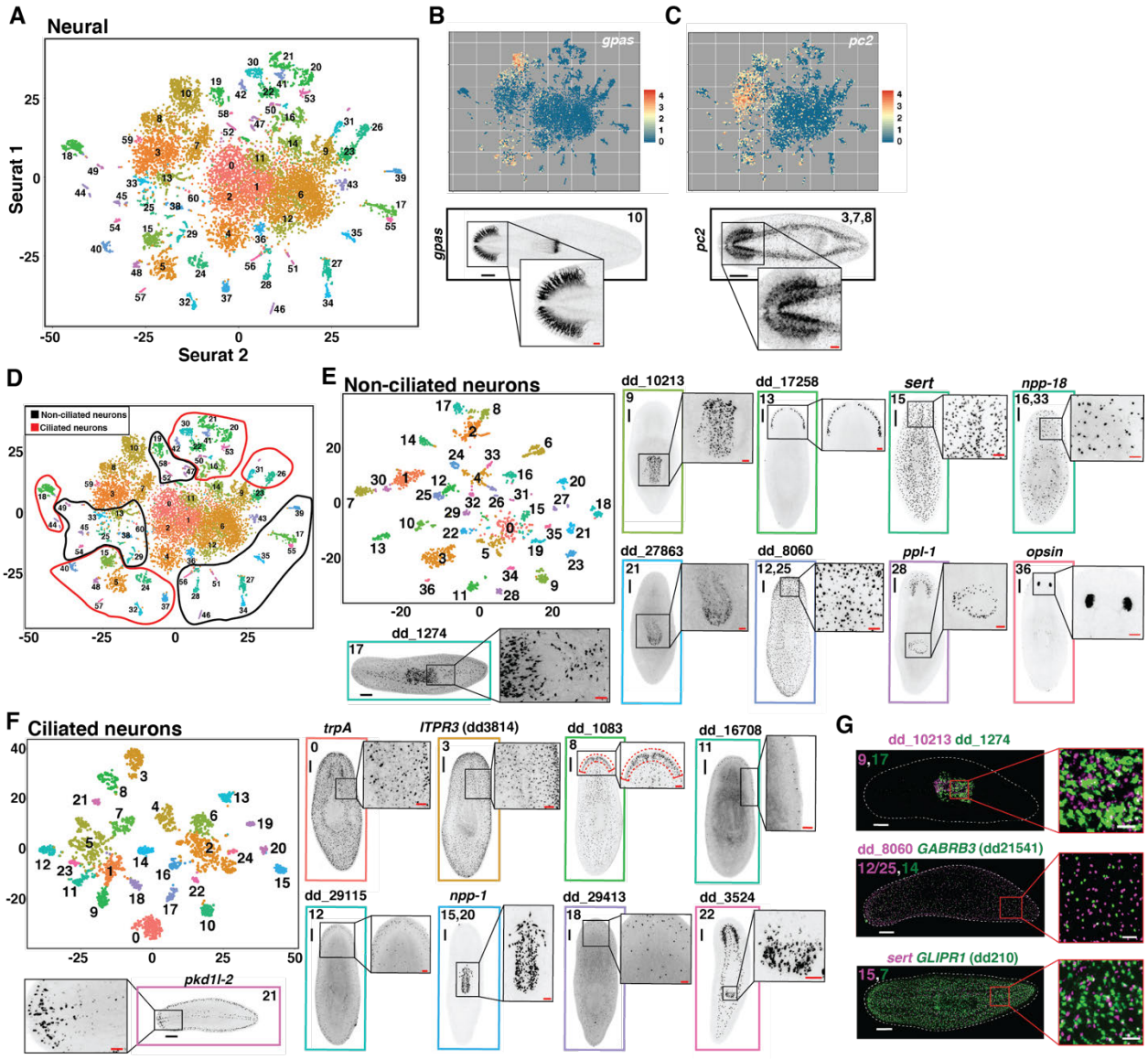

A
Neural
Seurat 1
Seurat 2
B
gpas
C
pc2
D
Non-ciliated neurons
Ciliated neurons
E
Non-ciliated neurons
dd_10213
dd_17258
sert
npp-18
dd_27863
dd_8060
ppl-1
opsin
dd_1274
F
Ciliated neurons
trpA
ITPR3 (dd3814)
dd_1083
dd_16708
dd_29115
npp-1
dd_29413
dd_3524
pkd1l-2
G
dd_10213 dd_1274
dd_8060 GABRB3 (dd21541)
sert GLIPR1 (dd210)

Figure 2.19. Subclustering of neural cells reveals known and novel cell populations.

(**A**) t-SNE representation of the neural subcluster. (**B** and **C**) Top: t-SNE plots colored by expression of *gpas* (B) and *pc-2* (C). Bottom: FISH for *gpas* (B) and *pc-2* (C) labeled with the associated neural subcluster. (**D**) t-SNE plot in (A) overlaid with outlines indicating the ascribed identity of each subcluster as ciliated or nonciliated. (**E** and **F**) t-SNE representation of subclustered cells identified in (D) as nonciliated (E) or ciliated (F). (**G**) Double FISH images of three sets of nonciliated neuron genes enriched in separate subclusters. Numbers indicate the associated nonciliated neuron subcluster(s) for each marker. Scale bars: whole-animal images, 200 µm; insets, 50 µm.

Figure 2.20

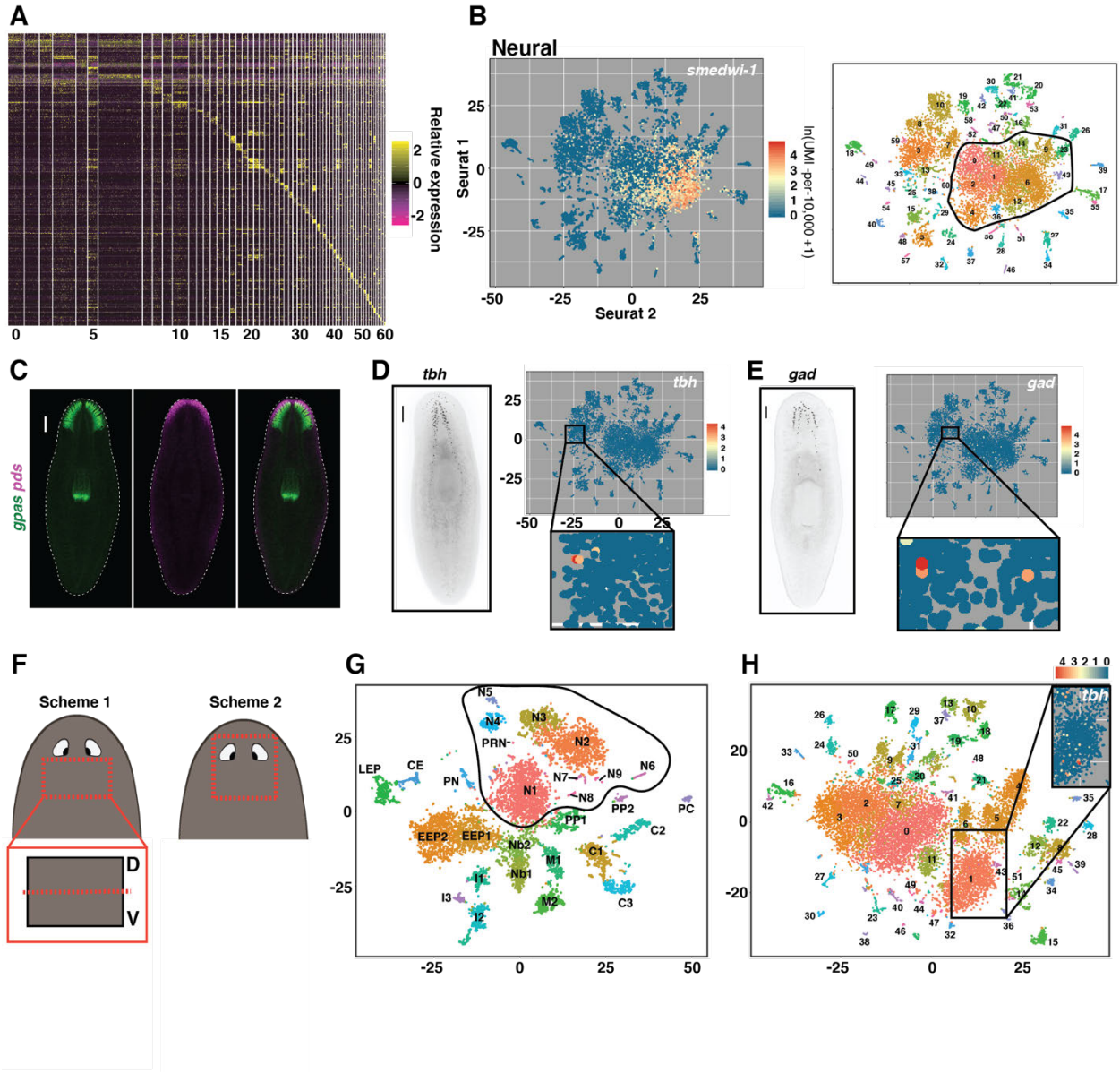

A
Relative expression
2
0
-2
0
5
10
15
20
30
40
50
60
B
Neural
smedwi-1
Seurat 1
Seurat 2
25
0
-25
-50
-25
0
25
ln(UMI-per-10,000 +1)
4
3
2
1
0
C
gpas pds
D
tbh
E
gad
F
Scheme 1
Scheme 2
D
V
G
N5
N4
N3
PRN
N2
LEP
CE
PN
N7
N9
N6
N1
N8
PP2
PC
PP1
C2
EEP2
EEP1
Nb2
C1
M1
I1
Nb1
I3
M2
C3
I2
H
4 3 2 1 0
tbh

Figure 2.20. Additional characterization of the neural subclusters.

(**A**) Heat map of the expression of the top 10 genes from each cluster of the neural clustering, grouped by cluster number. Cells, columns; Genes, rows. (**B**) Left: Neural t-SNE plot colored by *smedwi-1* expression. Right: t-SNE representation of neural subclustering overlaid with a circle, indicating clusters marked as putative transition states. (**C**) FISH images of *gpas* (***48***) from Figure 2.19B and *pds* (***49***). White signal in the merged image indicates positional overlap in gene expression. (**D** and **E**) Left: FISH images for (D) *tbh* (***55***) and (E) *gad* (***56***). Right: Neural t-SNE plots colored by (D) *tbh* and (E) *gad* expression. Insets identify ~7 *tbh*$^{+}$ and ~4 *gad*$^{+}$ neurons in neural subclusters 3 and 7. (**F**) Amputation schemes for isolating planarian brain cells. Left, the ventral half of the fragment was used. (**G**) t-SNE representation of 28 clusters generated from cells isolated in (F). C = *cathepsin*$^{+}$ cells, CE = ciliated epidermis, EEP = early epidermal progenitors, I = intestine, LEP = late epidermal progenitors, M = muscle, N = neural, Nb = neoblast, PC = pigment cells, PN = protonephridia, PP = parenchymal, PRN = photoreceptor neurons. (**H**) t-SNE representation of 52 clusters generated from combining and re-clustering all cells identified as neural from the brain data in (G) and from all cells identified as neural in the original data. Inset identifies ~13 *tbh*$^{+}$ neurons, representing an almost doubling of these cell types in the data. Scale bars, 200 μm.

Figure 2.21

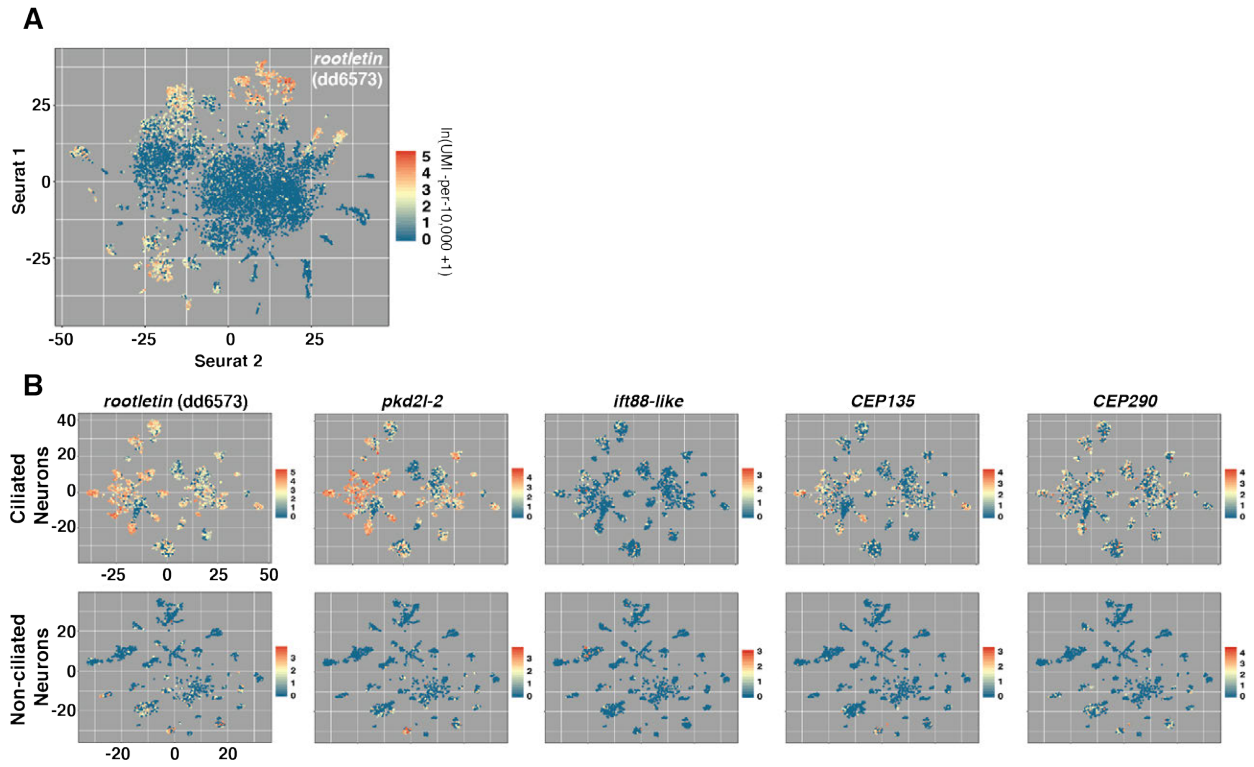

A
rootletin
(dd6573)
Seurat 1
Seurat 2
ln(UMI -per-10,000 +1)
B
rootletin (dd6573)
pkd2l-2
ift88-like
CEP135
CEP290
Ciliated Neurons
Non-ciliated Neurons

Figure 2.21. Additional data regarding the grouping of neuronal clusters into ciliated and non-ciliated neurons.

(**A**) Neural t-SNE plot colored by *rootletin* (dd6573) expression. Expression used to distinguish ciliated and non-ciliated neuron subclusters in Figure 2.19D. (**B**) Non-ciliated and ciliated neuron t-SNE plots colored by expression of five genes encoding cilia or centrosome components (***57***). Cells of the ciliated neuron subcluster are heavily enriched in genes encoding cilia or centrosome components, but at least one non-ciliated neuron subcluster is also enriched in such genes. Further work would be needed to assess cilia presence/absence in these subclusters.

Figure 2.22

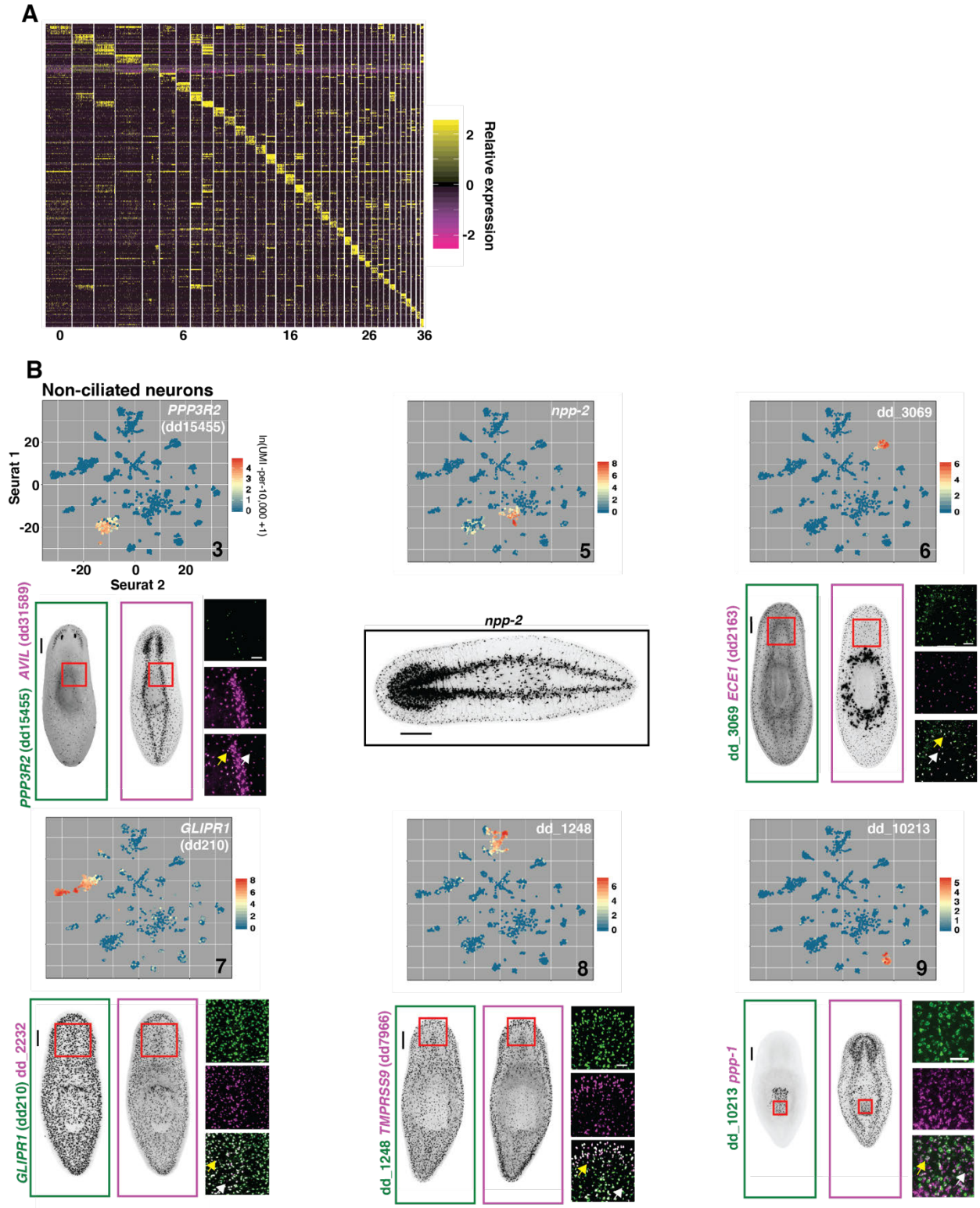

A
Relative expression
2
0
-2
0
6
16
26
36
B
Non-ciliated neurons
PPP3R2 (dd15455)
ln(UMI-per-10,000 +1)
Seurat 1
Seurat 2
3
npp-2
5
dd_3069
6
PPP3R2 (dd15455) AVIL (dd31589)
npp-2
dd_3069 ECE1 (dd2163)
GLIPR1 (dd210)
7
dd_1248
8
dd_10213
9
GLIPR1 (dd210) dd_2232
dd_1248 TMPRSS9 (dd7966)
dd_10213 ppp-1

Figure 2.22 Additional characterization of the non-ciliated neuron subclusters.

(**A**) Heat map of the expression of the top 10 genes from each cluster of the non-ciliated neuron clustering, grouped by cluster number. Cells, columns; Genes, rows. (**B**) Top panel: Non-ciliated neuron t-SNE plots colored by cluster-enriched gene expression. Number indicates the associated non-ciliated neuron subcluster. Bottom panel: FISH images of one or two cluster-enriched genes. The region in the red box is shown at higher magnification to the right. White signal in merged images indicates co-expression. Yellow arrows: co-expression. White arrows: no co-expression. Scale bars: whole-animal images, 200 µm; insets, 50 µm.

Figure 2.23

B continued

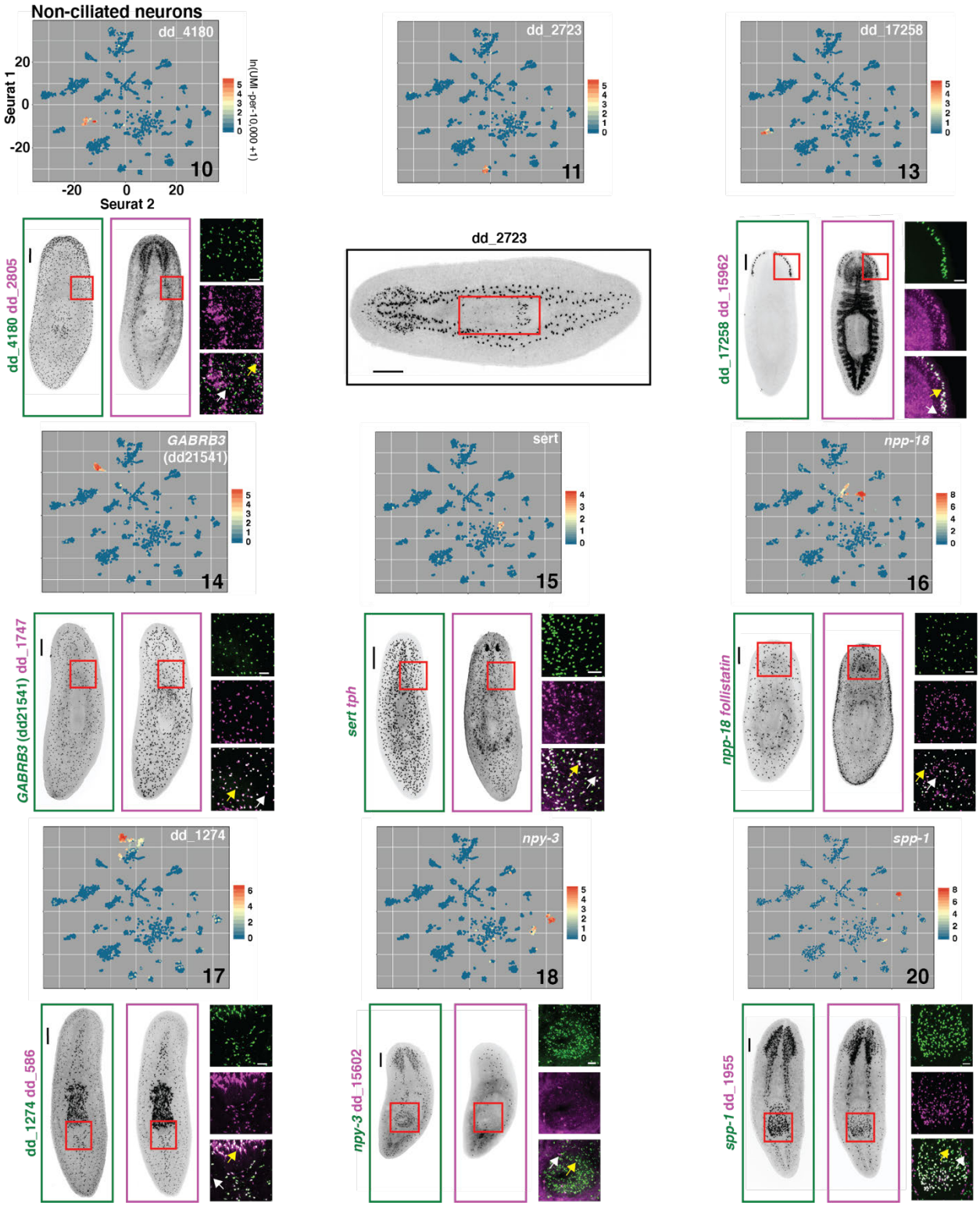


Figure 2.23. Continuation of Figure 2.22.

(**B**) Top panel: Non-ciliated neuron t-SNE plots colored by cluster-enriched gene expression. Number indicates the associated non-ciliated neuron subcluster. Bottom panel: FISH images of one or two cluster-enriched genes. The region in the red box is shown at higher magnification to the right. White signal in merged images indicates co-expression. Yellow arrows: co-expression. White arrows: no co-expression. Scale bars: whole-animal images, 200 µm; insets, 50 µm.

Figure 2.24

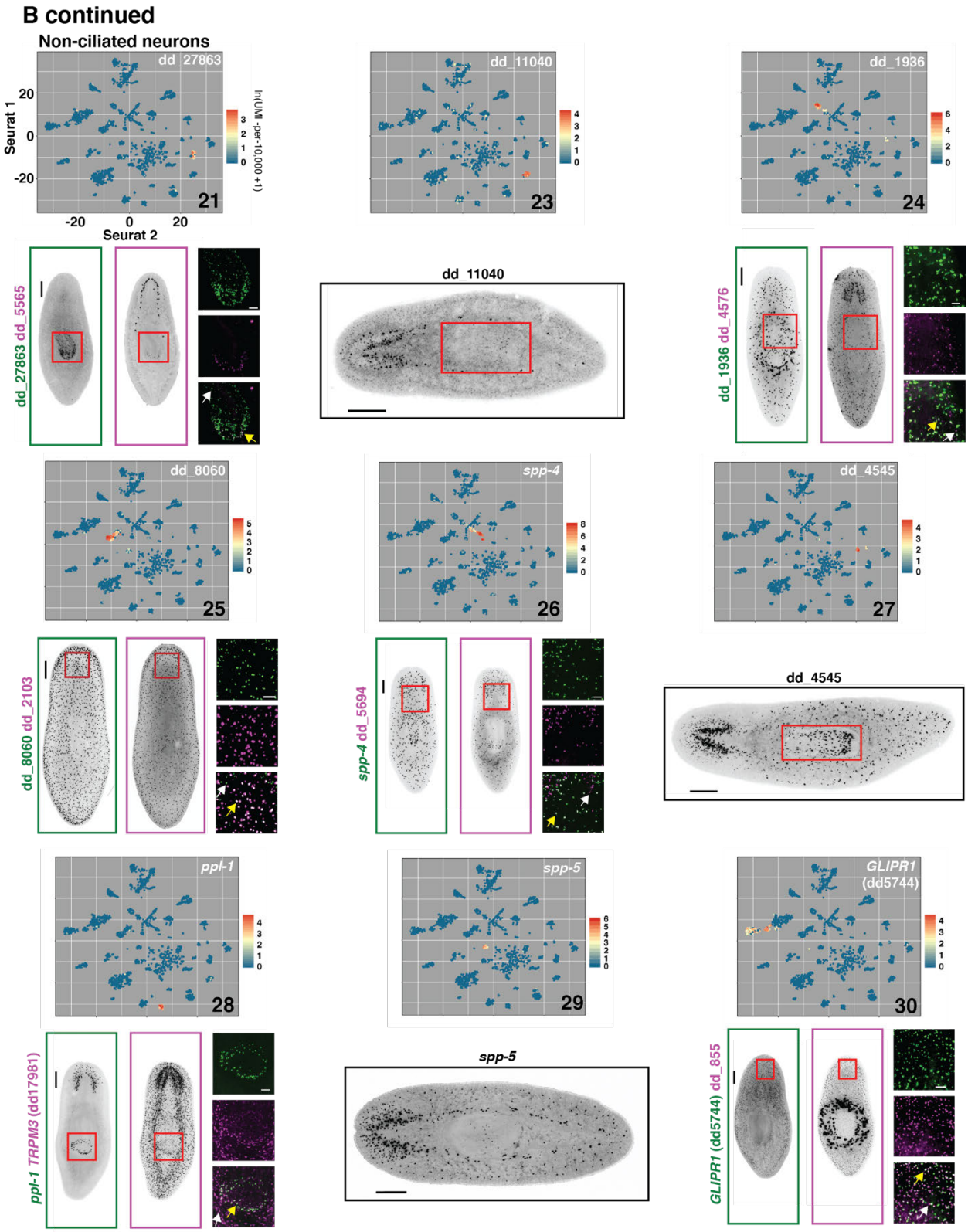

B continued
Non-ciliated neurons
dd_27863
21
Seurat 1
Seurat 2
ln(UMI-per-10,000+1)
dd_11040
23
dd_1936
24
dd_27863 dd_5565
dd_11040
dd_1936 dd_4576
dd_8060
25
spp-4
26
dd_4545
27
dd_8060 dd_2103
spp-4 dd_5694
dd_4545
ppl-1
28
spp-5
29
GLIPR1 (dd5744)
30
ppl-1 TRPM3 (dd17981)
spp-5
GLIPR1 (dd5744) dd_855

Figure 2.24. Continuation of Figure 2.22.

(**B**) Top panel: Non-ciliated neuron t-SNE plots colored by cluster-enriched gene expression. Number indicates the associated non-ciliated neuron subcluster. Bottom panel: FISH images of one or two cluster-enriched genes. The region in the red box is shown at higher magnification to the right. White signal in merged images indicates co-expression. Yellow arrows: co-expression. White arrows: no co-expression. Scale bars: whole-animal images, 200 µm; insets, 50 µm.

Figure 2.25

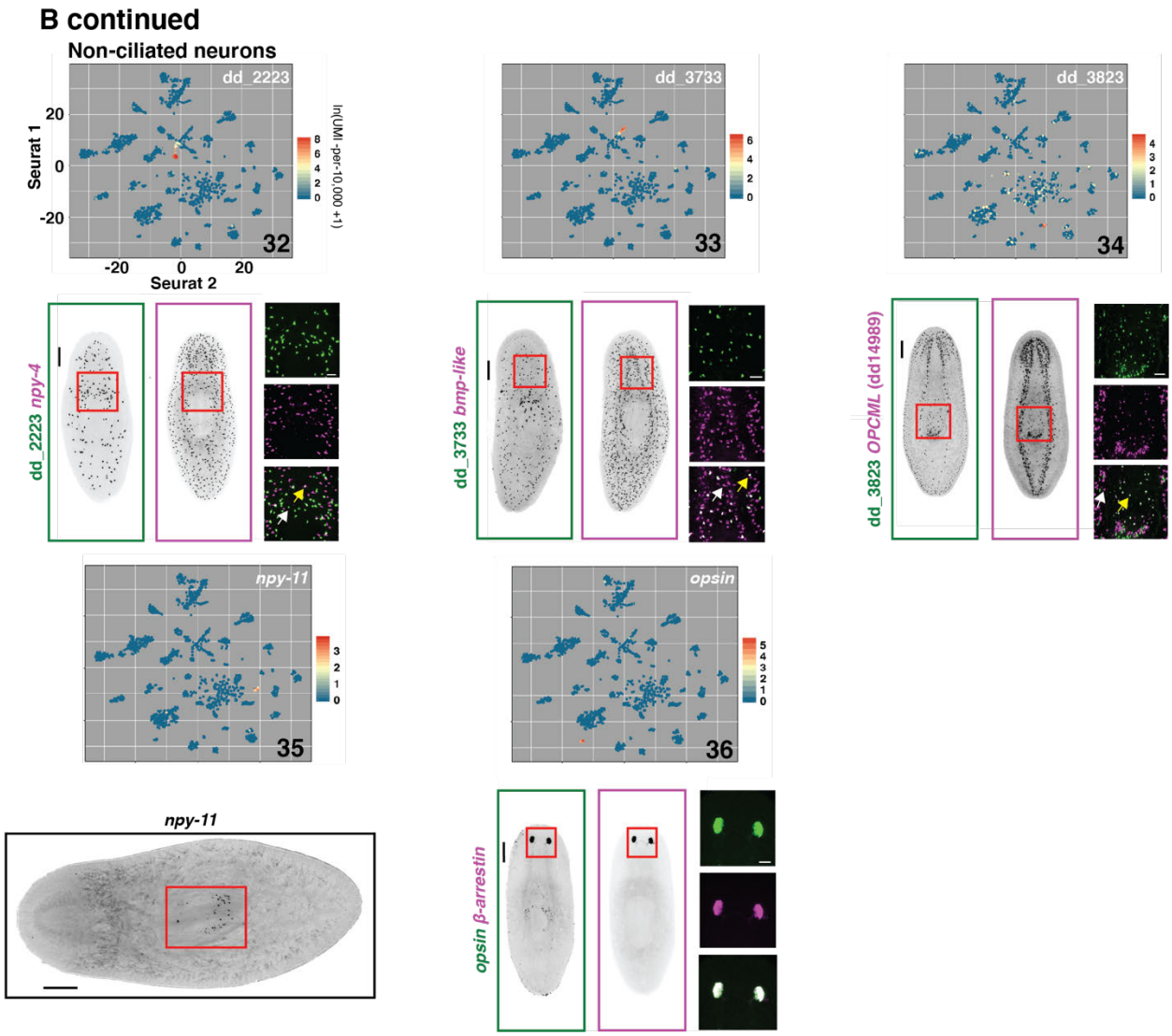

B continued
Non-ciliated neurons
dd_2223
dd_3733
dd_3823
Seurat 1
Seurat 2
ln(UMI-per-10,000 +1)
32
33
34
dd_2223 npy-4
dd_3733 bmp-like
dd_3823 OPCML (dd14989)
npy-11
opsin
35
36
npy-11
opsin β-arrestin

Figure 2.25. Continuation of Figure 2.22.

(**B**) Top panel: Non-ciliated neuron t-SNE plots colored by cluster-enriched gene expression. Number indicates the associated non-ciliated neuron subcluster. Bottom panel: FISH images of one or two cluster-enriched genes. The region in the red box is shown at higher magnification to the right. White signal in merged images indicates co-expression. Yellow arrows: co-expression. White arrows: no co-expression. Scale bars: whole-animal images, 200 µm; insets, 50 µm.

Figure 2.26

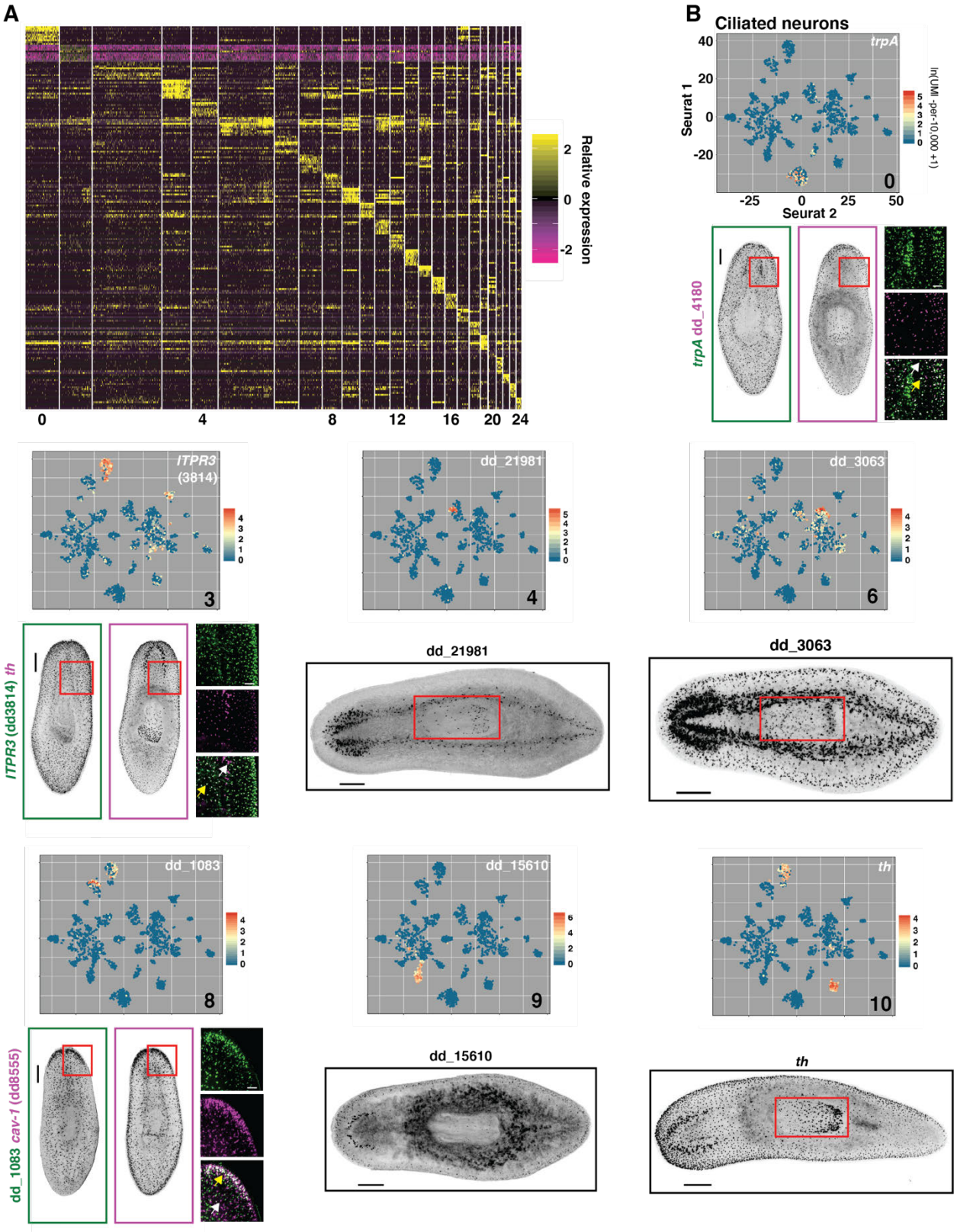

A
Relative expression
2
0
-2
0
4
8
12
16
20
24
B
Ciliated neurons
trpA
Seurat 1
Seurat 2
ln(UMI -per-10,000 +1)
0
trpA dd_4180
ITPR3 (3814)
3
dd_21981
4
dd_3063
6
ITPR3 (dd3814) th
dd_21981
dd_3063
dd_1083
8
dd_15610
9
th
10
dd_1083 cav-1 (dd8555)
dd_15610
th

Figure 2.26. Additional characterization of the ciliated neuron subclusters.

(**A**) Heat map of the expression of the top 10 genes from each cluster of the ciliated neuron clustering, grouped by cluster number. Cells, columns; Genes, rows. (**B**) Top panel: Ciliated neuron t-SNE plots colored by cluster-enriched gene expression. Number indicates the associated ciliated neuron subcluster. Bottom panel: FISH images of one or two cluster-enriched genes. The region in the red box is shown at higher magnification to the right. White signal in merged images indicates co-expression. Yellow arrows: co-expression. White arrows: no co-expression. Scale bars: whole-animal images, 200 µm; insets, 50 µm.

Figure 2.27

B continued

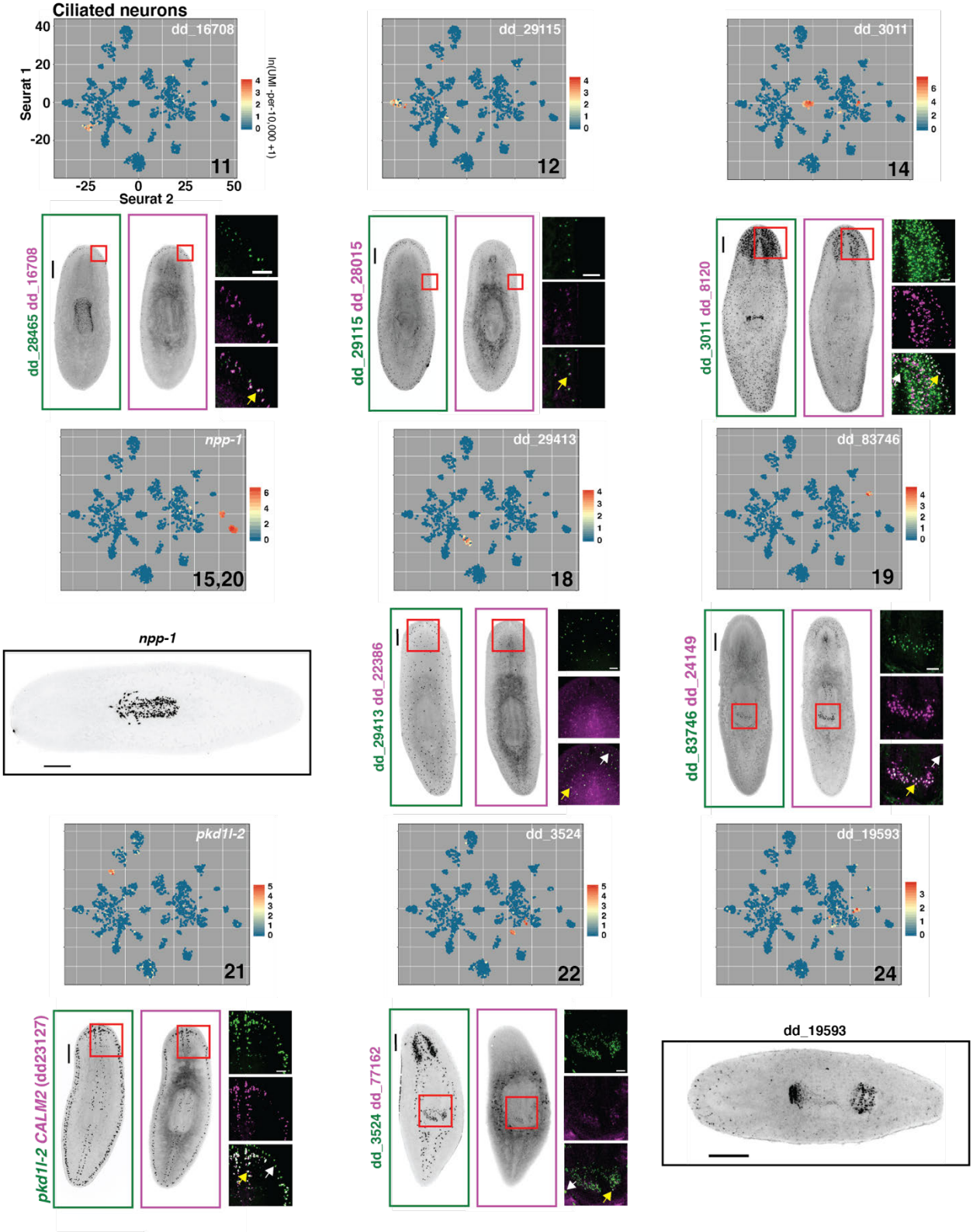


Figure 2.27. Continuation of Figure 2.26.

(**B**) Top panel: Ciliated neuron t-SNE plots colored by cluster-enriched gene expression. Number indicates the associated ciliated neuron subcluster. Bottom panel: FISH images of one or two cluster-enriched genes. The region in the red box is shown at higher magnification to the right. White signal in merged images indicates co-expression. Yellow arrows: co-expression. White arrows: no co-expression. Scale bars: whole-animal images, 200 µm; insets, 50 µm.

Figure 2.28

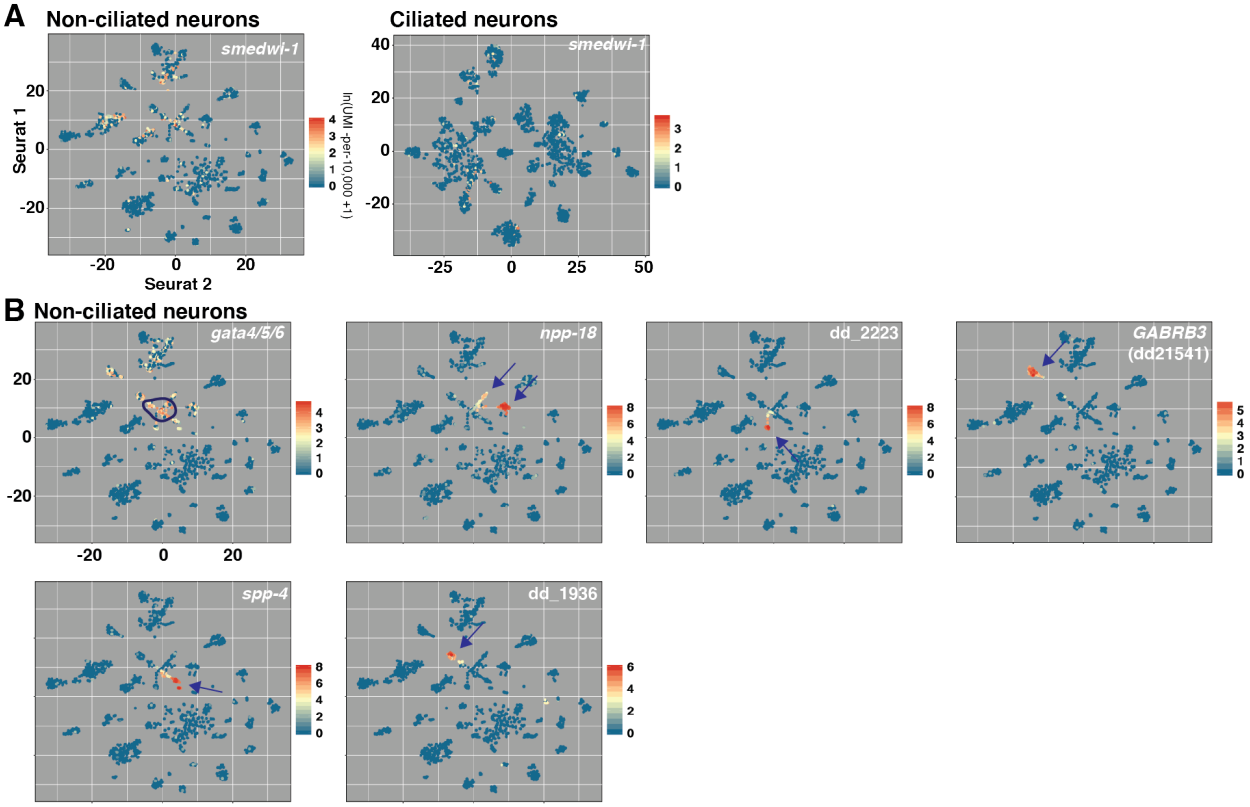

A
Non-ciliated neurons
Ciliated neurons
smedwi-1
smedwi-1
Seurat 1
Seurat 2
ln(UMI-per-10,000 +1)
B
Non-ciliated neurons
gata4/5/6
npp-18
dd_2223
GABRB3
(dd21541)
spp-4
dd_1936

Figure 2.28. Identification of putative transition states in the ciliated and non-ciliated neuron subclusters.

(**A**) Ciliated and non-ciliated neuron t-SNE plots colored by *smedwi-1* expression. (**B**) Non-ciliated neuron t-SNE plots colored by *gata4/5/6-1* expression. Circled subcluster indicates one domain of co-expression with *smedwi-1*. Arrows indicate sites of cluster-specific gene expression in 5 additional subclusters that radiate out from the circled *smedwi-1*$^{+}$ subcluster.

Figure 2.29

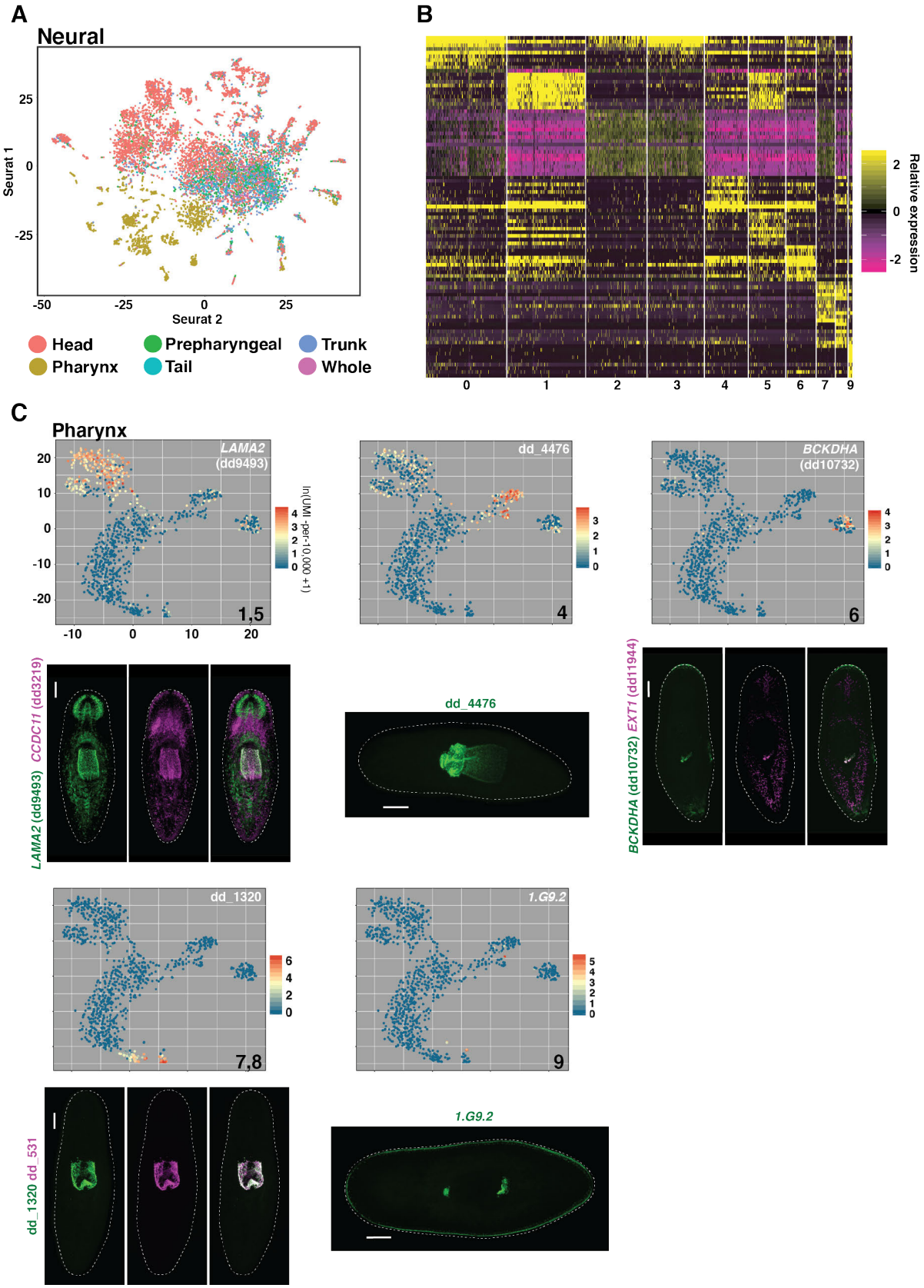

A
Neural
Seurat 1
Seurat 2
Head
Pharynx
Prepharyngeal
Tail
Trunk
Whole
B
Relative expression
C
Pharynx
LAMA2 (dd9493)
dd_4476
BCKDHA (dd10732)
ln(UMI-per-10,000 +1)
1,5
4
6
LAMA2 (dd9493) CCDC11 (dd3219)
dd_4476
BCKDHA (dd10732) EXT1 (dd11944)
dd_1320
1.G9.2
7,8
9
dd_1320 dd_531
1.G9.2

Figure 2.29. Additional characterization of the pharynx subcluster.

(**A**) Neural t-SNE plot colored by the body region from which each cell was isolated. Many pharynx-derived subclusters were present. (**B**) Heat map of the expression of the top 10 genes from each cluster of the pharynx clustering, grouped by cluster number. Cells, columns; Genes, rows. (**C**) Top panel: Pharynx t-SNE plots colored by cluster-enriched gene expression. Number indicates the associated pharynx subcluster. Bottom panel: FISH images of one or two cluster-enriched genes. White signal in merged images indicates co-expression. Scale bar, 200 μm.

Figure 2.30

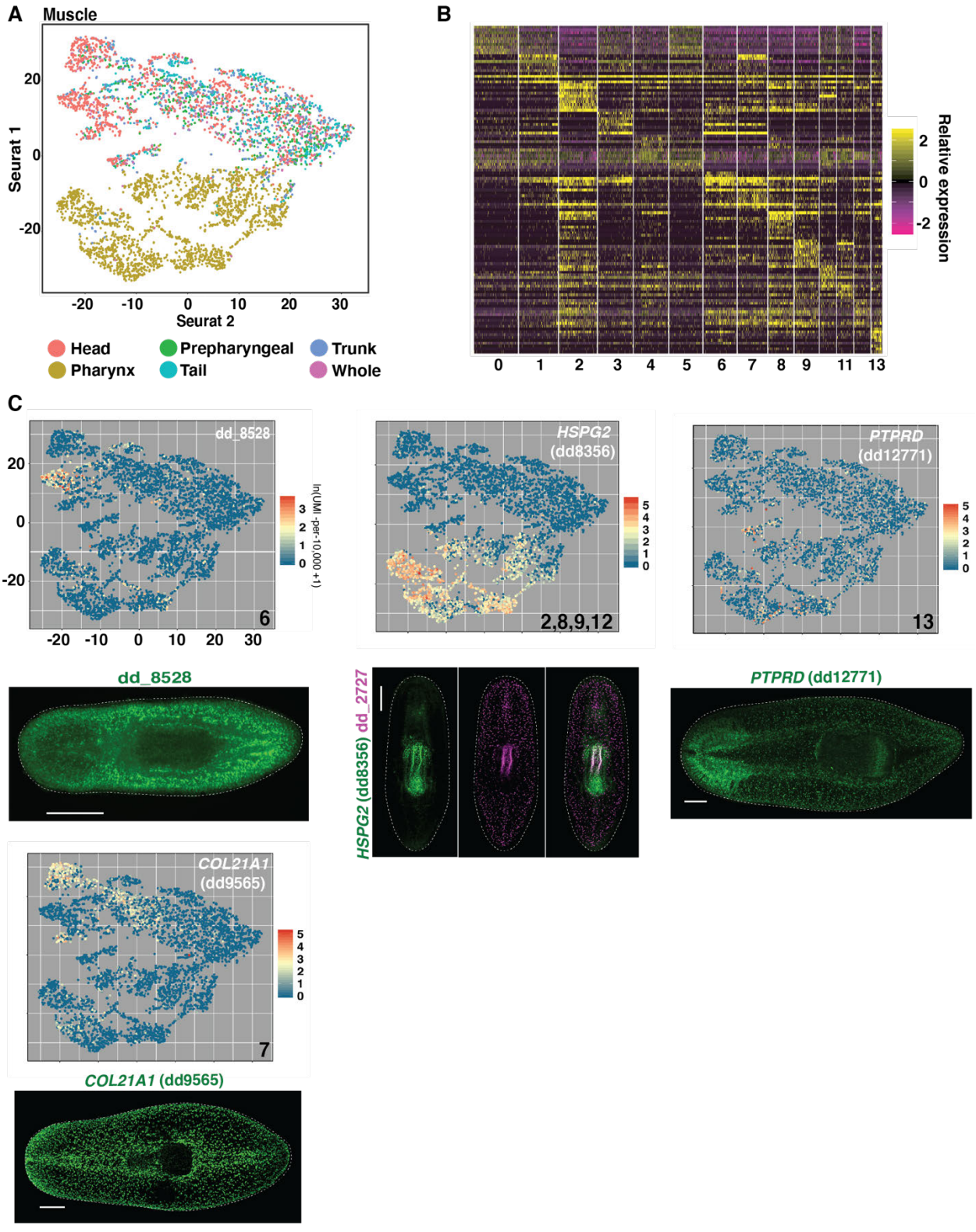

A
Muscle
Seurat 1
Seurat 2
Head
Pharynx
Prepharyngeal
Tail
Trunk
Whole
B
Relative expression
C
dd_8528
ln(UMI-per-10,000+1)
HSPG2 (dd8356)
2,8,9,12
PTPRD (dd12771)
13
dd_8528
HSPG2 (dd8356) dd_2727
PTPRD (dd12771)
COL21A1 (dd9565)
7
COL21A1 (dd9565)

Figure 2.30. Additional characterization of the muscle subcluster.

(**A**) Muscle t-SNE plot colored by the body region from which each cell was isolated. Many pharynx-derived subclusters were present. (**B**) Heat map of the expression of the top 10 genes from each cluster of the muscle clustering, grouped by cluster number. Cells, columns; Genes, rows. (**C**) Top panel: Muscle t-SNE plots colored by cluster-enriched gene expression. Number in bottom right corner of plot indicates the associated muscle subcluster number. Bottom panel: FISH images of one or two cluster-enriched genes. White signal in merged images indicates co-expression. Scale bar, 200 μm.

Figure 2.31

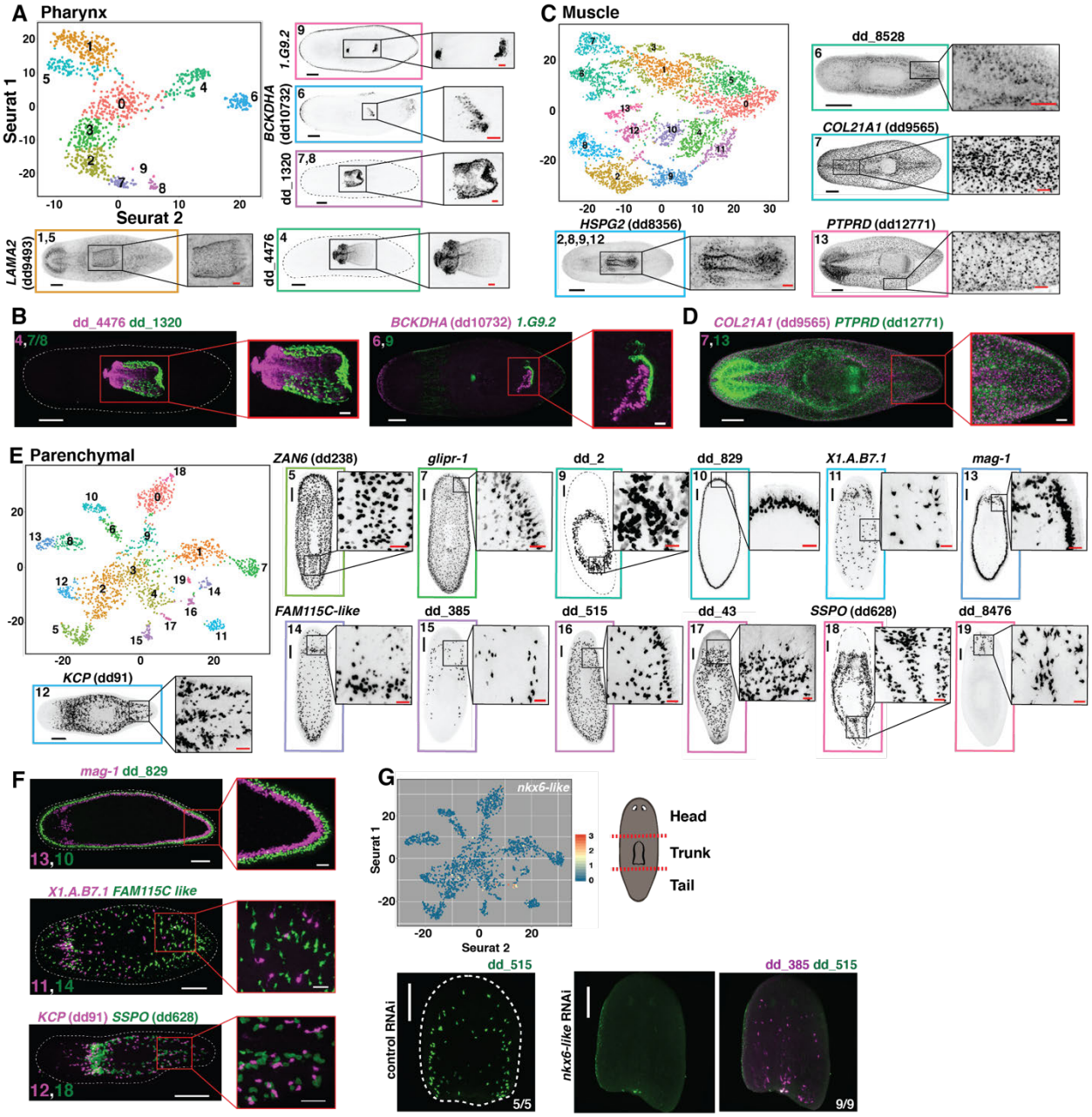

A Pharynx
Seurat 1
Seurat 2
1.G9.2
BCKDHA (dd10732)
dd_1320
LAMA2 (dd9493)
dd_4476
C Muscle
dd_8528
COL21A1 (dd9565)
HSPG2 (dd8356)
PTPRD (dd12771)
B
dd_4476 dd_1320
4,7/8
BCKDHA (dd10732) 1.G9.2
6,9
D
COL21A1 (dd9565) PTPRD (dd12771)
7,13
E Parenchymal
KCP (dd91)
ZAN6 (dd238)
glipr-1
dd_2
dd_829
X1.A.B7.1
mag-1
FAM115C-like
dd_385
dd_515
dd_43
SSPO (dd628)
dd_8476
F
mag-1 dd_829
13,10
X1.A.B7.1 FAM115C like
11,14
KCP (dd91) SSPO (dd628)
12,18
G
nkx6-like
Seurat 1
Seurat 2
Head
Trunk
Tail
dd_515
control RNAi
5/5
dd_385 dd_515
nkx6-like RNAi
9/9

Figure 2.31 Tissue subclustering identifies cell populations of poorly characterized tissues.

(**A**) t-SNE representation of the pharynx subcluster. FISH images are labeled by their associated cluster(s). (**B**) Double FISH images of pharynx markers enriched in separate subclusters. Numbers indicate the associated pharynx subcluster(s) for each marker. (**C**) t-SNE representation of the muscle subcluster. (**D**) Double FISH images of two muscle markers enriched in separate subclusters. Numbers indicate the associated muscle subcluster for each marker. (**E**) t-SNE representation of the parenchymal subcluster. (**F**) Double FISH images of three sets of parenchymal markers enriched in separate subclusters. Numbers indicate the associated parenchymal subcluster for each marker. (**G**) Top left: Parenchymal t-SNE plot colored by expression of *nkx6-like.* Top right: Illustration of cutting scheme used to generate fragments. Bottom: dd_515 and dd_385 FISH of control and *nkx6-like* RNAi animals. Animal sections were cut and fixed 23 days after the start of dsRNA feedings. Scale bars: whole-animal/fragment images, 200 µm; insets, 50 µm.

Figure 2.32

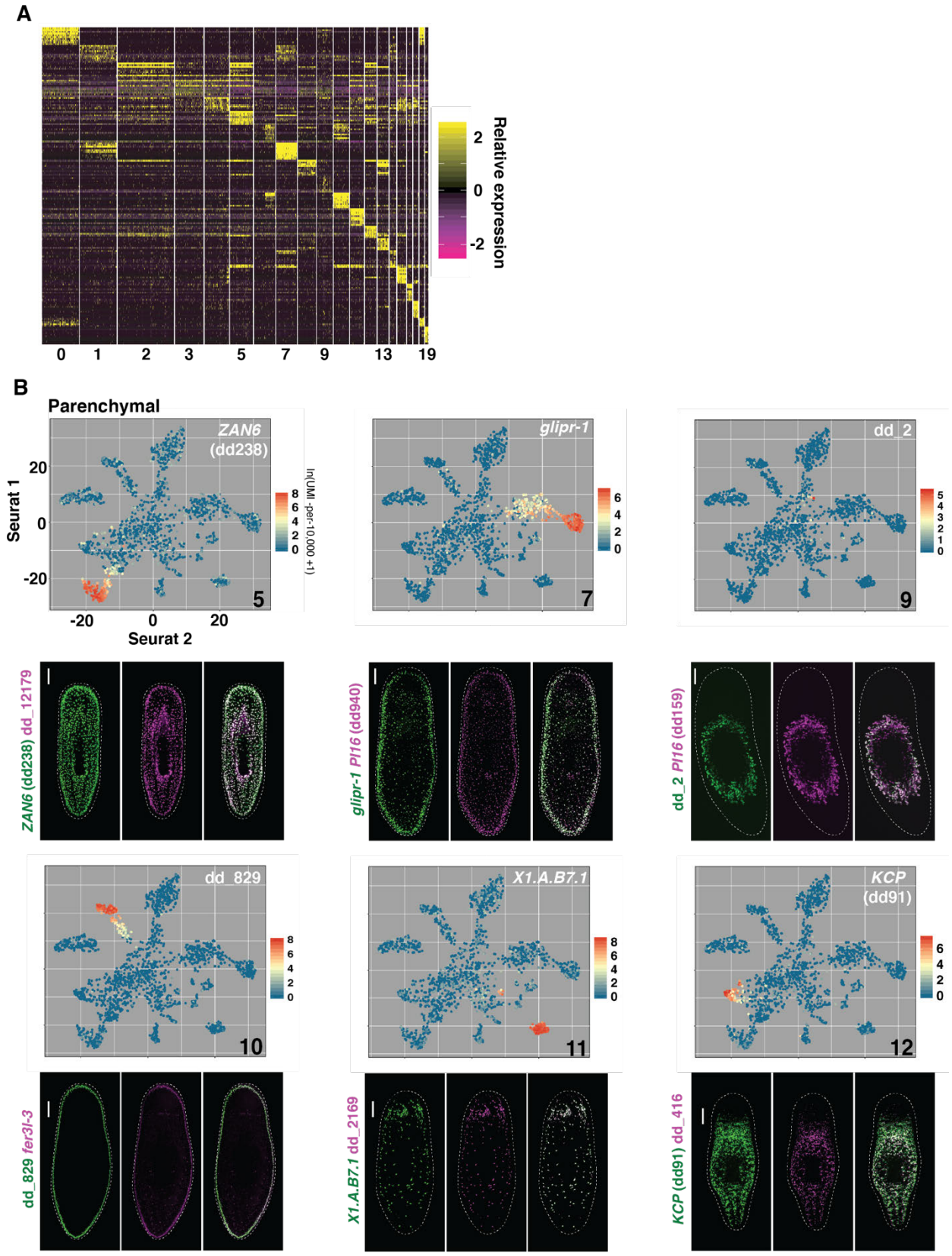

A
Relative expression
2
0
-2
0
1
2
3
5
7
9
13
19
B
Parenchymal
ZAN6
(dd238)
glipr-1
dd_2
Seurat 1
Seurat 2
ln(UMI-per-10,000 +1)
5
7
9
ZAN6 (dd238) dd_12179
glipr-1 Pl16 (dd940)
dd_2 Pl16 (dd159)
dd_829
X1.A.B7.1
KCP
(dd91)
10
11
12
dd_829 fer3l-3
X1.A.B7.1 dd_2169
KCP (dd91) dd_416

Figure 2.32. Additional characterization of the parenchymal subcluster.

(**A**) Heat map of the expression of the top 10 genes from each cluster of the parenchymal clustering, grouped by cluster number. Cells, columns; Genes, rows. (**B**) Top panel: Parenchymal t-SNE plots colored by cluster-enriched gene expression. Numbers indicate the associated parenchymal subcluster. Bottom panel: FISH images of one or two cluster-enriched genes. White signal in merged images indicates co-expression. Scale bar, 200 μm.

Figure 2.33

B continued

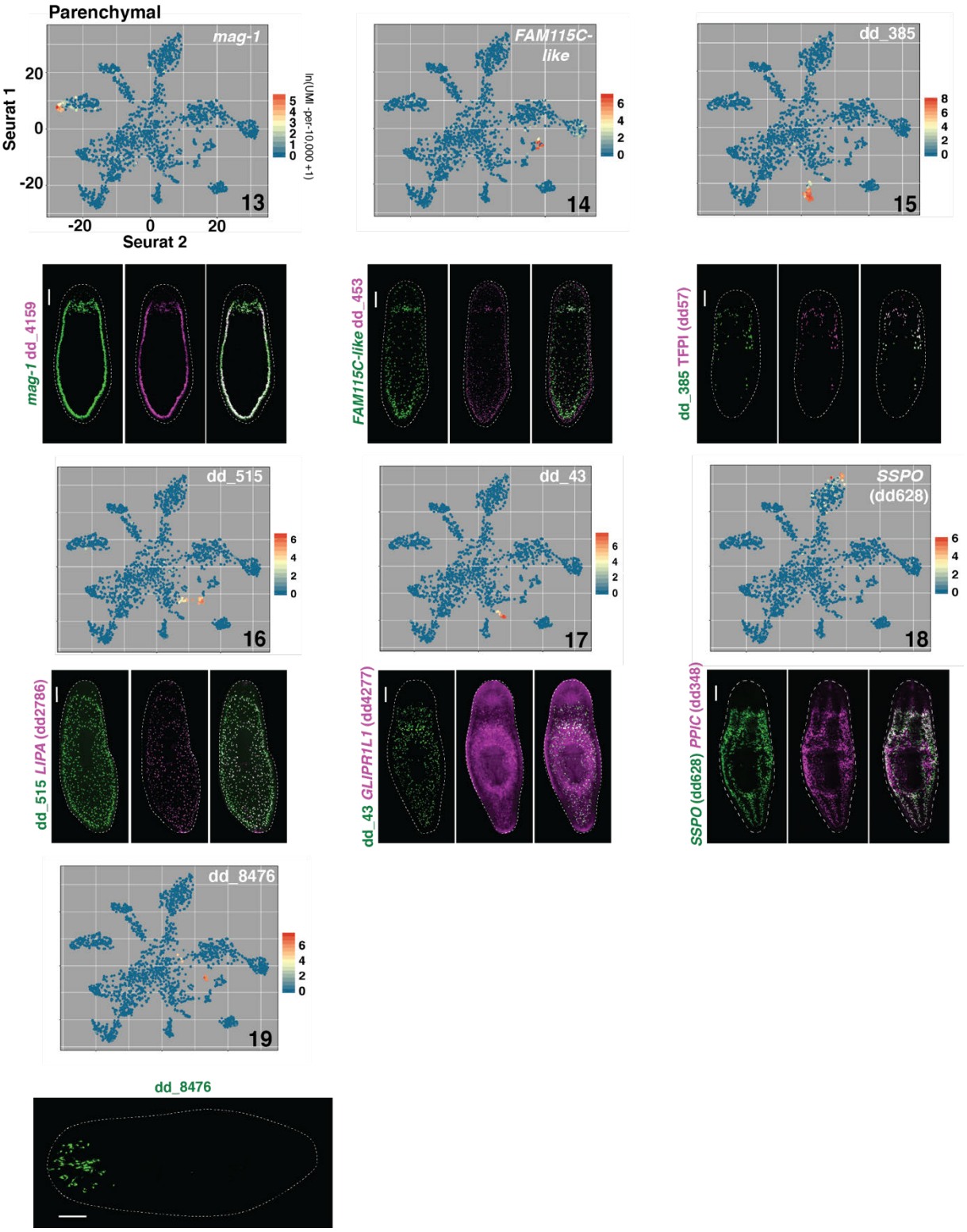


Figure 2.33. Continuation of Figure 2.32.

(**B**) Top panel: Parenchymal t-SNE plots colored by cluster-enriched gene expression. Numbers indicate the associated parenchymal subcluster. Bottom panel: FISH images of one or two cluster-enriched genes. White signal in merged images indicates co-expression. Scale bar, 200 μm.

Figure 2.34

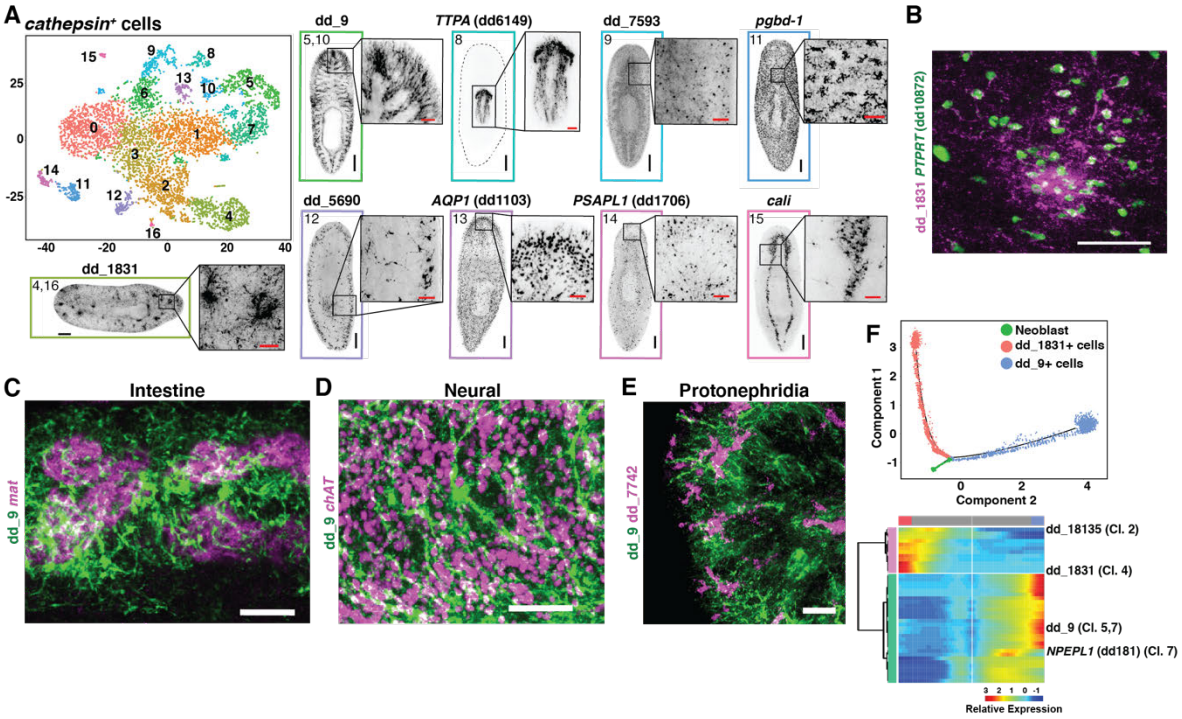

A
cathepsin+ cells
dd_9
TTPA (dd6149)
dd_7593
pgbd-1
dd_5690
AQP1 (dd1103)
PSAPL1 (dd1706)
cali
dd_1831
B
dd_1831 PTPRT (dd10872)
C
Intestine
dd_9 mat
D
Neural
dd_9 chAT
E
Protonephridia
dd_9 dd_7742
F
Neoblast
dd_1831+ cells
dd_9+ cells
Component 1
Component 2
dd_18135 (Cl. 2)
dd_1831 (Cl. 4)
dd_9 (Cl. 5,7)
NPEPL1 (dd181) (Cl. 7)
Relative Expression

Figure 2.34 Tissue subclustering reveals a previously unidentified class of cells.

(**A**) t-SNE representation of the *cathepsin*$^{+}$ cell subcluster. FISH images are labeled by their associated cluster(s). Images associated with subclusters 5/10 and 8 are single slices in the animal. All other images are maximum intensity projections. (**B**) Double FISH for two *cathepsin*$^{+}$ cell markers enriched in the same subclusters, 4 and 16. (**C** to **E**) FISH for dd_9 and *mat* (C), *ChAT* (D), and dd_7742 (E). (**F**) Top: Cell trajectory of dd_1831^{+} and dd_9^{+} *cathepsin*$^{+}$ cell lineages produced by Monocle. Cells are colored by identity. Bottom: Heat map of branch dependent genes (*q* value < 10^{-175}) across cells plotted in pseudo-time (***45***). Cells, columns; genes, rows. Beginning of pseudo-time is at center of heat map. “Cl.” annotation indicates a log-fold enrichment ≥ 1 of the gene in that *cathepsin*$^{+}$ cell Seurat cluster. Scale bars: whole-animal images, 200 μm; insets and (B) to (E), 50 μm.

Figure 2.35

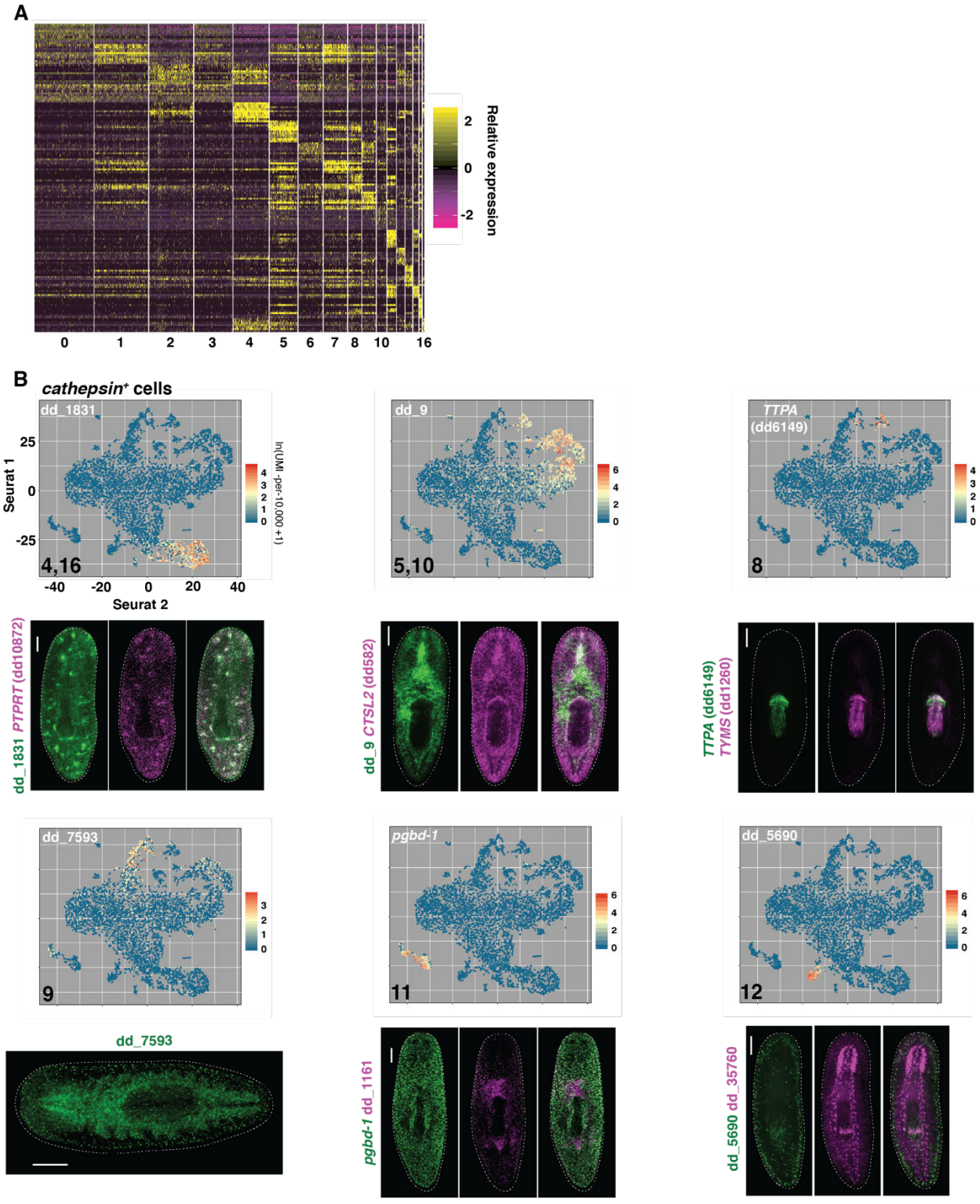

A
Relative expression
2
0
-2
0 1 2 3 4 5 6 7 8 10 16
B
cathepsin+ cells
dd_1831
ln(UMI-per-10,000 +1)
Seurat 1
Seurat 2
4,16
dd_1831 PTPRT (dd10872)
dd_9
5,10
dd_9 CTSL2 (dd582)
TTPA (dd6149)
8
TTPA (dd6149) TYMS (dd1260)
dd_7593
9
dd_7593
pgbd-1
11
pgbd-1 dd_1161
dd_5690
12
dd_5690 dd_35760

Figure 2.35. Additional characterization of the *cathepsin*$^+$ cell subcluster.

(**A**) Heat map of the expression of the top 10 genes from each cluster of the *cathepsin*$^+$ cell clustering, grouped by cluster number. Cells, columns; Genes, rows. (**B**) Top panel: *cathepsin*$^+$ cell t-SNE plots colored by cluster-enriched gene expression. Number indicates the associated *cathepsin*$^+$ cell subcluster. Bottom panel: FISH images of one or two cluster-enriched genes. White signal in merged images indicates co-expression. Scale bar, 200 μm.

Figure 2.36

B continued

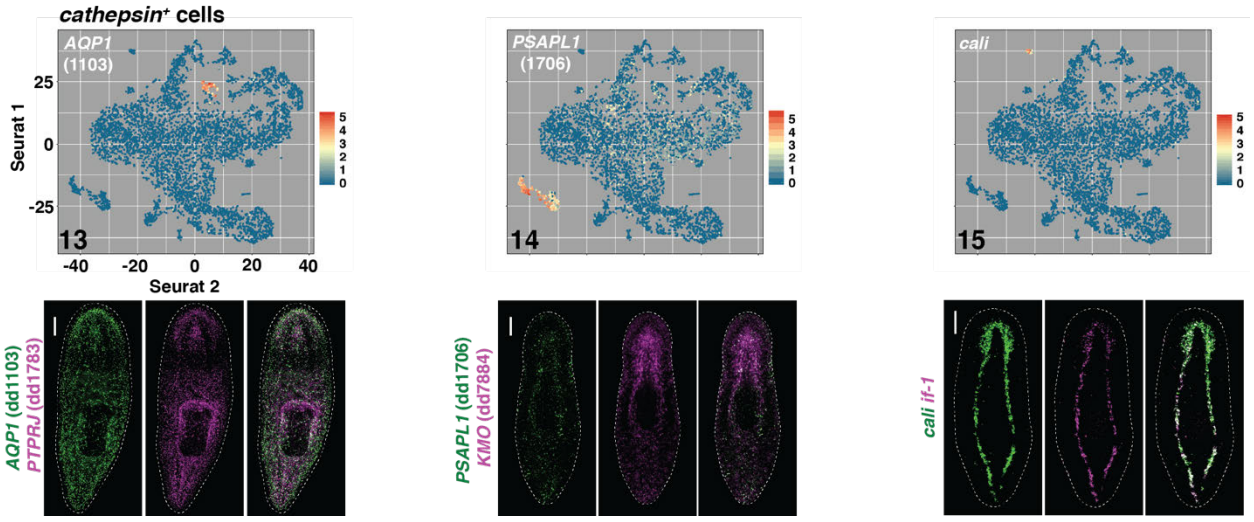


Figure 2.36. Continuation of Figure 2.35.

(**B**) Top panel: *cathepsin*$^+$ cell t-SNE plots colored by cluster-enriched gene expression. Numbers indicate the associated *cathepsin*$^+$ cell subcluster. Bottom panel: FISH images of one or two cluster-enriched genes. White signal in merged images indicates co-expression. Scale bar, 200 µm.

Figure 2.37

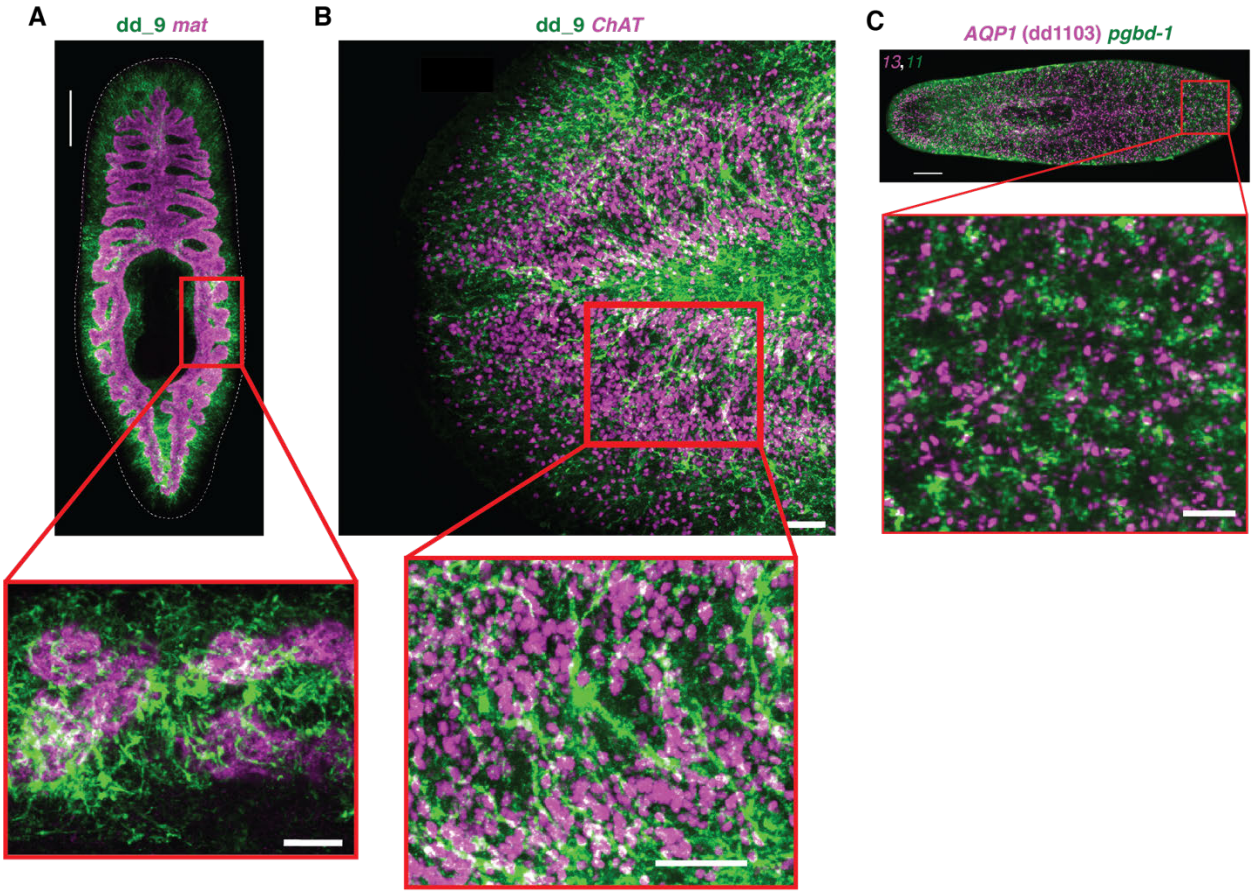


Figure 2.37. Additional images characterizing *cathepsin*$^{+}$ cells.

(**A** and **B**) Double FISH images of dd_9 and (A) *mat* (***58***) or (B) *ChAT* (***14***, ***59***). Red box indicates the region magnified in the inset below. Insets also in Figure 2.34C and 2.34D, respectively. (**C**) FISH images of two *cathepsin*$^{+}$ cell markers enriched in separate clusters, demonstrating an absence of co-expression. Scale bars: whole-animal images, 200 µm; insets, 50 µm.

Figure 2.38

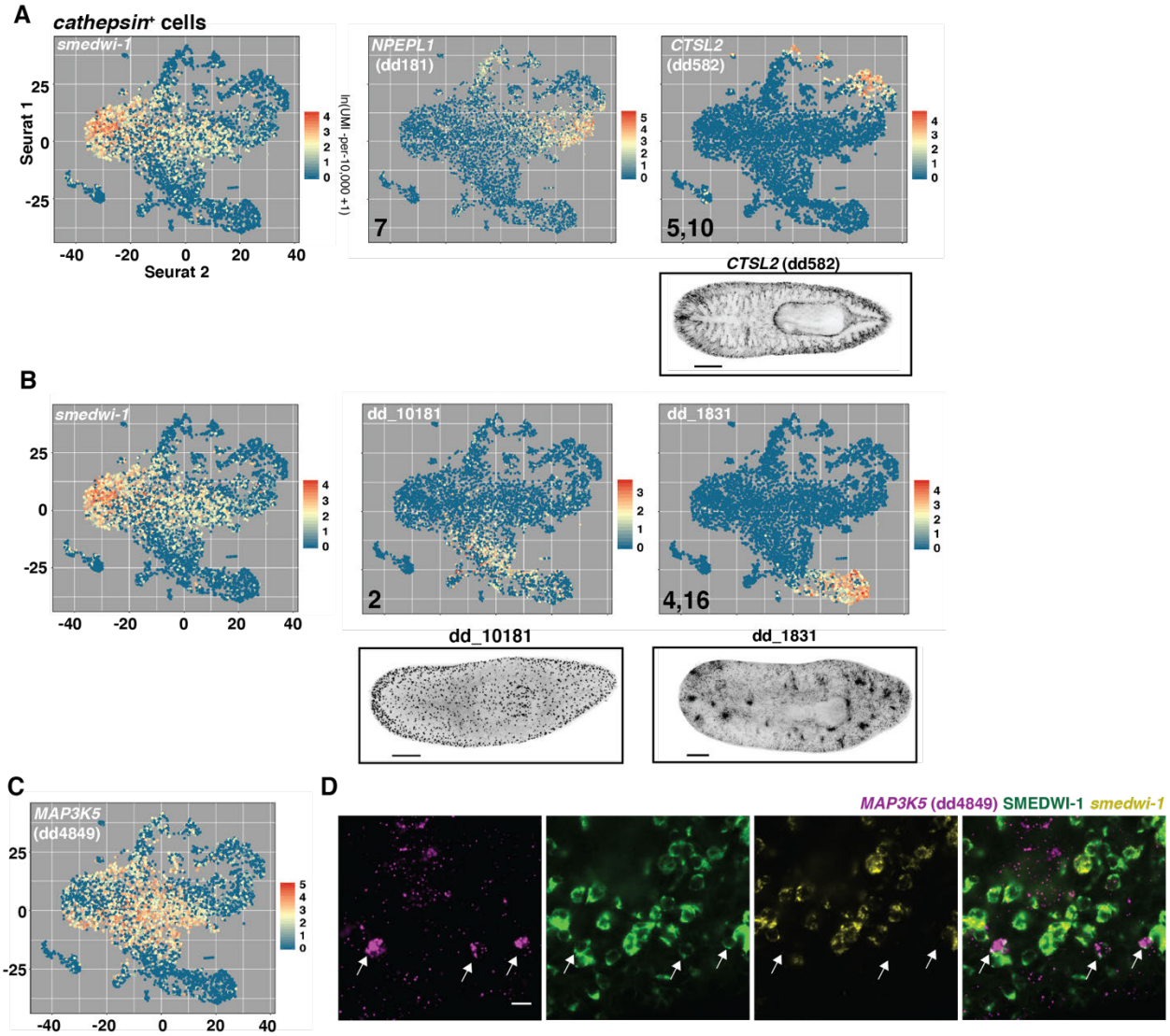

A
cathepsin+ cells
smedwi-1
NPEPL1 (dd181)
CTSL2 (dd582)
7
5,10
Seurat 1
Seurat 2
ln(UMI-per-10,000 +1)
CTSL2 (dd582)
B
smedwi-1
dd_10181
dd_1831
2
4,16
dd_10181
dd_1831
C
MAP3K5 (dd4849)
D
MAP3K5 (dd4849) SMEDWI-1 smedwi-1

Figure 2.38. Identification of putative *cathepsin*$^+$ cell transition states.

(**A** and **B**) Top panel: *cathepsin*$^+$ cell t-SNE plots colored by expression of genes enriched in putative transition state clusters associated with (A) subclusters 5/10 and (B) subclusters 4/16. Number indicates the associated *cathepsin*$^+$ cell subcluster. Bottom panel: FISH images of cluster-enriched genes. (**C**) *cathepsin*$^+$ cell t-SNE plot colored by expression of the gene *MAP3K5* (dd4849). (**D**) FISH/antibody stain for *MAP3K5* (dd4849), *smedwi-1* RNA, and SMEDWI-1 protein. Arrows indicate *MAP3K5* (dd4849)$^+$ cells that co-express SMEDWI-1 protein, but do not co-express *smedwi-1* RNA. Scale bars: whole-animal images, 200 μm; D, 10 μm.

Figure 2.39

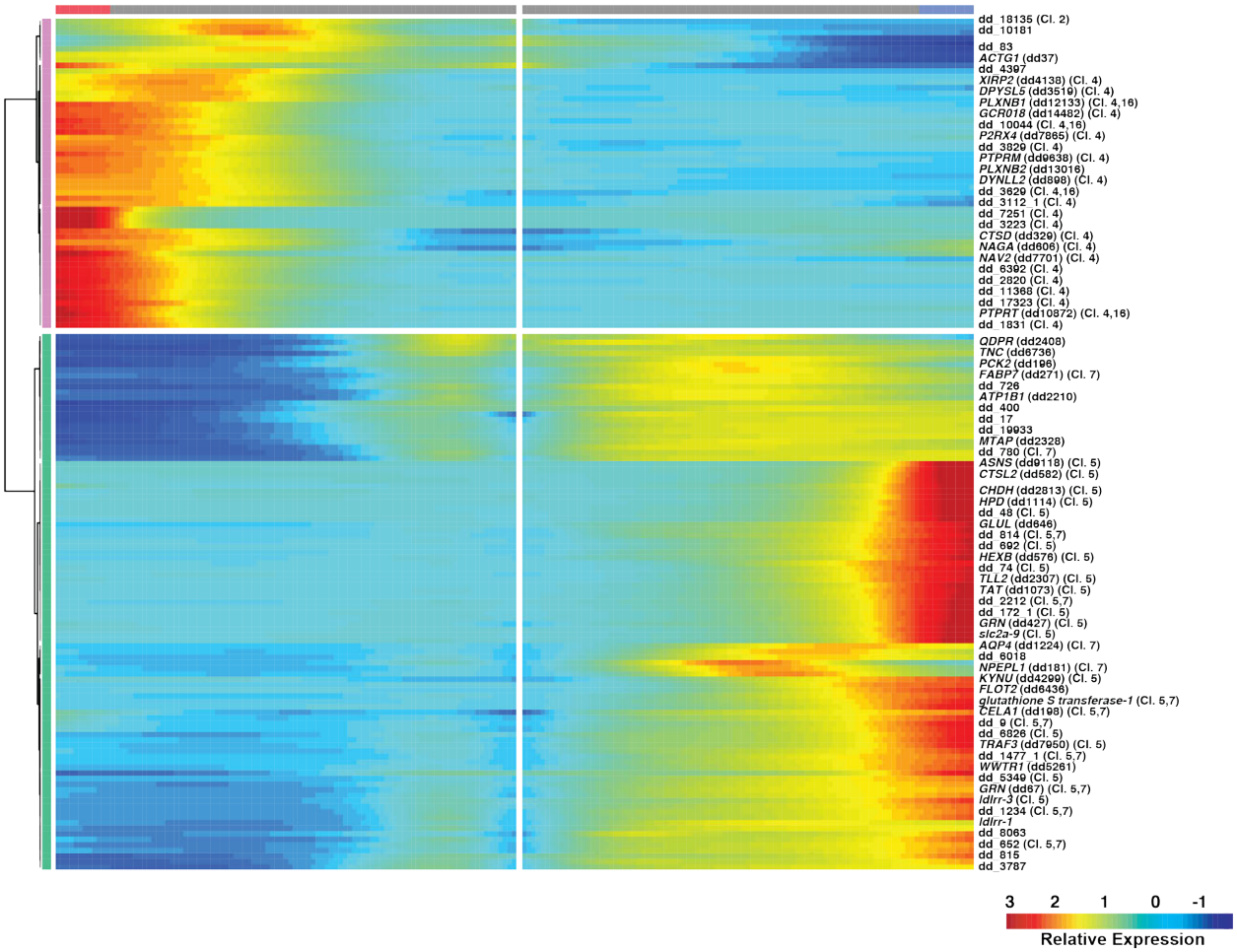

dd_18135 (Cl. 2)
dd_10181
dd_83
ACTG1 (dd37)
dd_4397
XIRP2 (dd4138) (Cl. 4)
DPYSL5 (dd3516) (Cl. 4)
PLXNB1 (dd12133) (Cl. 4,16)
GCR018 (dd14482) (Cl. 4)
dd_10044 (Cl. 4,16)
P2RX4 (dd7865) (Cl. 4)
dd_3820 (Cl. 4)
PTPRM (dd9838) (Cl. 4)
PLXNB2 (dd13018)
DYNLL2 (dd898) (Cl. 4)
dd_3629 (Cl. 4,16)
dd_3112_1 (Cl. 4)
dd_7251 (Cl. 4)
dd_3223 (Cl. 4)
CTSD (dd329) (Cl. 4)
NAGA (dd606) (Cl. 4)
NAV2 (dd7701) (Cl. 4)
dd_6392 (Cl. 4)
dd_2820 (Cl. 4)
dd_11368 (Cl. 4)
dd_17323 (Cl. 4)
PTPRT (dd10872) (Cl. 4,16)
dd_1831 (Cl. 4)
QDPR (dd2408)
TNC (dd6736)
PCK2 (dd196)
FABP7 (dd271) (Cl. 7)
dd_726
ATP1B1 (dd2210)
dd_400
dd_17
dd_19033
MTAP (dd2328)
dd_780 (Cl. 7)
ASNS (dd9119) (Cl. 5)
CTSL2 (dd582) (Cl. 5)
CHDH (dd2813) (Cl. 5)
HPD (dd1114) (Cl. 5)
dd_48 (Cl. 5)
GLUL (dd646)
dd_814 (Cl. 5,7)
dd_892 (Cl. 5)
HEXB (dd576) (Cl. 5)
dd_74 (Cl. 5)
TLL2 (dd2307) (Cl. 5)
TAT (dd1073) (Cl. 5)
dd_2212 (Cl. 5,7)
dd_172_1 (Cl. 5)
GRN (dd427) (Cl. 5)
slc2a-9 (Cl. 5)
AQP4 (dd1224) (Cl. 7)
dd_6018
NPEPL1 (dd181) (Cl. 7)
KYNU (dd4299) (Cl. 5)
FLOT2 (dd6436)
glutathione S transferase-1 (Cl. 5,7)
CELA1 (dd198) (Cl. 5,7)
dd_9 (Cl. 5,7)
dd_6826 (Cl. 5)
TRAF3 (dd7950) (Cl. 5)
dd_1477_1 (Cl. 5,7)
WWTR1 (dd5261)
dd_5349 (Cl. 5)
GRN (dd67) (Cl. 5,7)
ldlrr-3 (Cl. 5)
dd_1234 (Cl. 5,7)
ldlrr-1
dd_8063
dd_652 (Cl. 5,7)
dd_815
dd_3787
3
2
1
0
-1
Relative Expression

Figure 2.39. Lineage reconstruction of the dd_1831^{+} and dd_9^{+} *cathepsin*$^{+}$ cell lineages.

Heat map of branch dependent genes (q-value < 1E-75) across cells plotted in pseudotime. Cells, columns; Genes, rows. Beginning of pseudotime at center of heatmap. “Cl.” annotation indicates a log-fold enrichment ≥ 1 of that gene in that *cathepsin*$^{+}$ cell Seurat cluster.

A near complete discovery of planarian cell type transcriptomes

The number of cell types identified in this study vastly exceeded prior planarian SCS data (***24***). Within the neuronal subclusters, a 17-cell subcluster represented photoreceptor neurons (Figure 2.19E), which are present at ~100 total cells in a medium-sized (~2 to 3 mm) animal (***20***). Therefore, our data should have readily included unknown cell types as rare as photoreceptor neurons. Similarly, an average-sized planarian has ~60 *cintillo*$^+$ neurons, and our data included 10 *cintillo*$^+$ neurons (Figure 2.40A). These cells were grouped within a larger subcluster (cluster 3) of nonciliated neurons (Figure 2.40B), suggesting that even further subclustering of this "subcluster 3" could reveal additional distinct cell types. Indeed, *cintillo*$^+$ cells emerged as a unique cluster from such additional (fourth tier) subclustering of original data (Figure 2.40C and Table 2.2). Esophagus cells, connecting pharynx to intestine, clustered with mouth cells in the pharynx subclustering data (Figure 2.40D). About 50 of these cells exist in an average-sized animal, and three such cells were present in the data (Figure 2.40, A and D). Several known rare cell types did not separate into individual clusters, although most could still be identified in the data, which suggests that the data are largely saturated for rare cell types. These include anterior pole cells, which function as an anterior organizer (***60***, ***61***, ***62***); *notum*$^+$ neurons in the brain (***63***); and posterior pole cells (***49***), each of which are among the rarest known cell types in the animal, with only ~10 each present in an average animal (Figure 2.40A). Five anterior pole cells, 10 *notum*$^+$ neurons, and one posterior pole cell were identified in the data (Figure 2.40, E to G). Similarly, ~25 *ovo*$^+$ eye progenitors (***46***) and ~90 *nanos*$^+$ germ cells (***64***) are present in an average animal (Figure 2.40A). Two eye progenitors and 19 germ cells were identified in the data (Figure 2.40, H and I). In addition to the asexual strain of *S. mediterranea* used in this study, a sexual strain of cross-fertilizing hermaphrodites exists. We sequenced 8455 cells from this strain, adding sexual strain cells to this resource as well, including seven yolk cells and seven testes cells in addition to the 19 germ cells described above (Figure 2.41, A to D). Further sequencing of sexual cell types could be a target for future studies. Together, our data indicate that

we have essentially reached saturation for determining the cell type transcriptomes of asexual planarians.

Figure 2.40

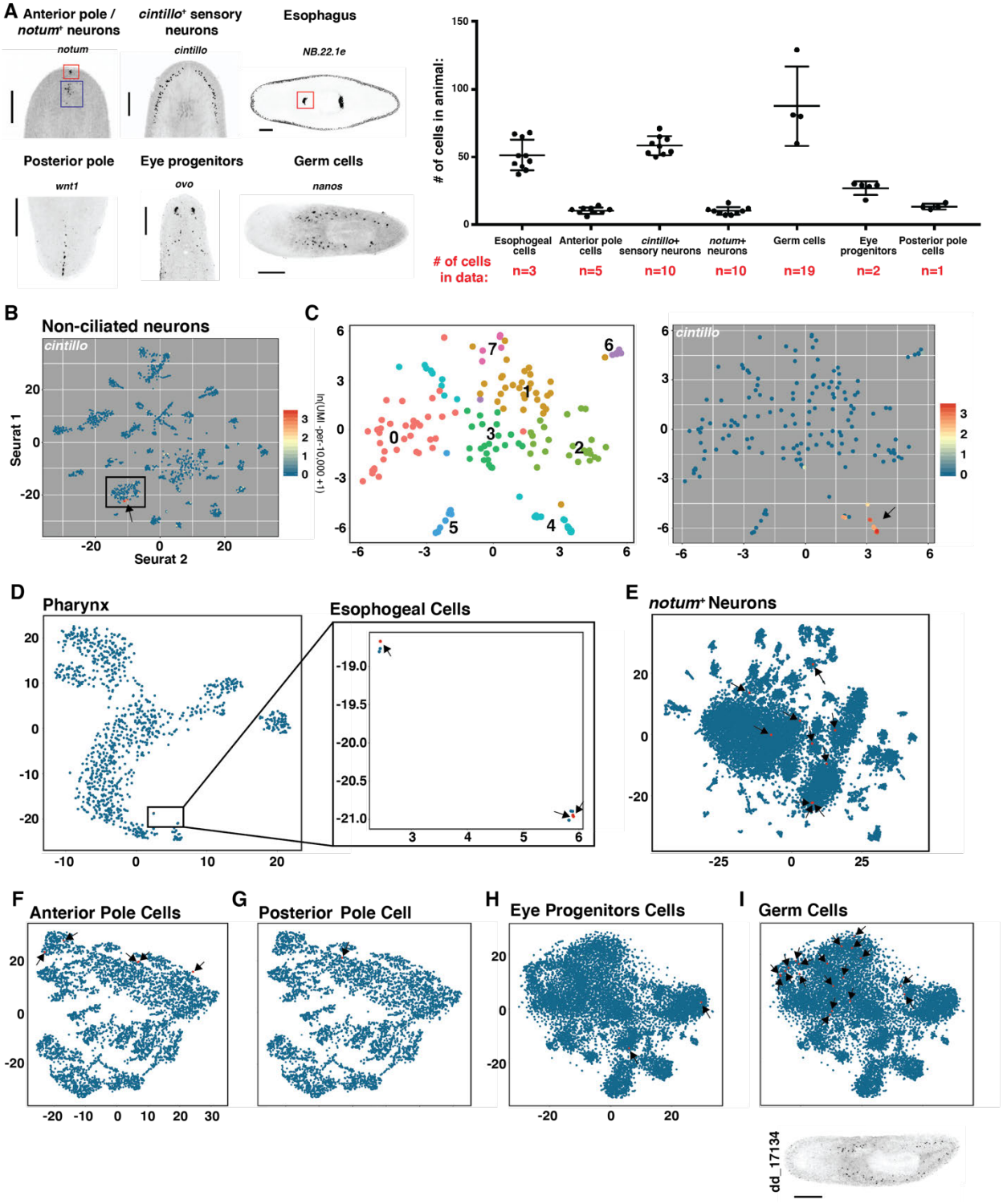

A
Anterior pole / notum+ neurons
cintillo+ sensory neurons
Esophagus
notum
cintillo
NB.22.1e
Posterior pole
Eye progenitors
Germ cells
wnt1
ovo
nanos
# of cells in animal:
Esophageal cells
Anterior pole cells
cintillo+ sensory neurons
notum+ neurons
Germ cells
Eye progenitors
Posterior pole cells
# of cells in data:
n=3
n=5
n=10
n=10
n=19
n=2
n=1
B
Non-ciliated neurons
cintillo
Seurat 1
Seurat 2
C
ln(UMI-per-10,000 +1)
cintillo
D
Pharynx
Esophogeal Cells
E
notum+ Neurons
F
Anterior Pole Cells
G
Posterior Pole Cell
H
Eye Progenitors Cells
I
Germ Cells
dd_17134

Figure 2.40. Identification of rare cell types in the data.

(**A**) Left: FISH images for the anterior pole (red box) and notum+ neurons (blue box), esophagus (red box), *cintillo*$^{+}$ sensory neurons, eye progenitors, posterior pole, and germ cells as marked by *notum* (***62***), *NB.22.1e* (***32***), *cintillo* (***65***), *ovo* (***46***), *wnt1* (***49***), and *nanos* (***64***), respectively. Right: Number of anterior pole cells, *notum*$^{+}$ neurons, esophageal cells, *cintillo*$^{+}$ neurons, eye progenitors, posterior pole cells, and germ cells in adult animals. Red text indicates the number of each cell in the data. (**B**) Non-ciliated neuron t-SNE plot colored by *cintillo* expression. Arrow indicates *cintillo*$^{+}$ cells. Box surrounds subcluster 3. (**C**) Left: t-SNE representation of cells boxed in (B) following additional subclustering. Right: Same t-SNE plot colored by *cintillo* expression. Arrow indicates *cintillo*$^{+}$ cells. (**D**) Pharynx t-SNE plot colored by esophageal cells positive for expression (>0.5, ln(UMI-per-10,000+1)) of *NB.22.1e* (***66***), *wntP-3* (***49***), and *bmp4* (***67***). t-SNE plot of cells in boxed region (pharynx subcluster 9) in isolation without further subclustering is shown in inset. Arrows indicate the 3 positive cells. (**E**) Combined neural t-SNE plot from Figure 2.20H colored by *notum*$^{+}$ neurons positive for expression of *chat* (>2.5, ln(UMI-per-10,000+1)) and *notum* ((>2, ln(UMI-per-10,000+1)) (***63***). Arrows indicate the ten positive cells. (**F**) Muscle t-SNE plot colored by anterior pole cells positive for expression (>0.5, ln(UMI-per-10,000+1)) of *notum* (***62***), *zic-1* (***61***, ***68***), and *foxD* (***60***, ***61***). Arrows indicate the five positive cells. (**G**) Muscle t-SNE plot colored by posterior pole cells positive for expression (>0.5, ln(UMI-per-10,000+1)) of *wnt1* and *wnt11-1* (***49***), *pitx* (***69***), and *fz4-1* (***70***). Arrow indicates one positive cell. (**H**) *smedwi-1*$^{+}$ t-SNE plot colored by eye progenitor cells positive for expression (>0.5, ln(UMI-per-10,000+1)) of *ovo*, *eya*, *six-1/2*, and *smedwi-1* (***46***, ***47***). Arrows indicate the two positive cells. (**I**) Top: *smedwi-1*$^{+}$ t-SNE plot colored by germ cells positive for expression (>0.5, ln(UMI-per-10,000+1)) of *nanos* and *gH4* (***64***), dd_17134, and *smedwi-1*. Arrows indicate the 19 positive cells. Bottom: FISH image for dd_17134, demonstrating expression in germ cells. Positive cells, red; negative cells, blue. Scale bars, 200 μm.

Figure 2.41

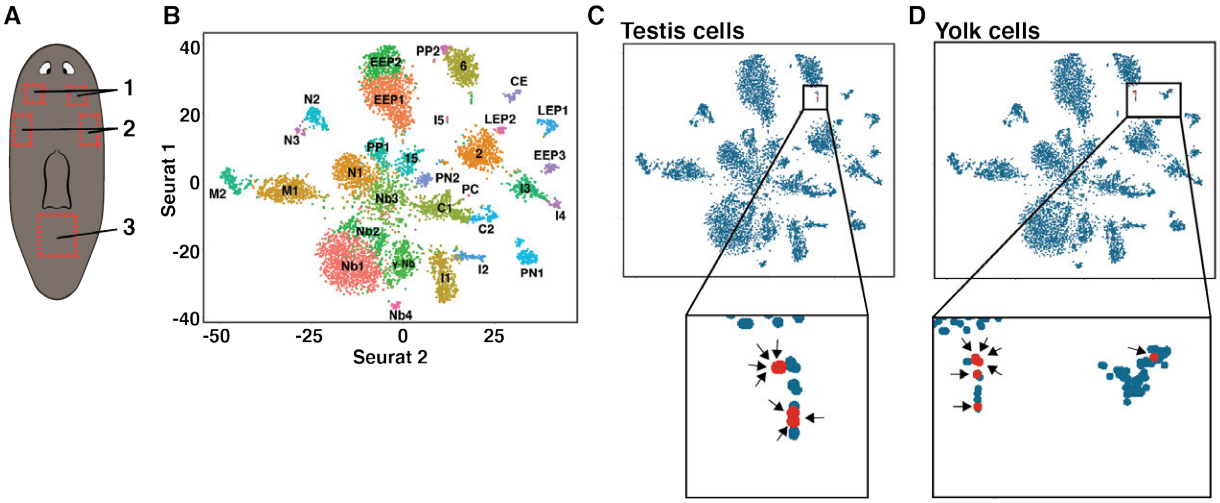

A
1
2
3
B
Seurat 1
Seurat 2
40
20
0
-20
-40
-50
-25
0
25
EEP2
PP2
6
CE
N2
EEP1
I5
LEP2
LEP1
N3
PP1
I6
2
EEP3
N1
PN2
PC
M2
M1
Nb3
C1
I3
I4
Nb2
C2
Nb1
Y-Nb
I2
PN1
I1
Nb4
C
Testis cells
D
Yolk cells

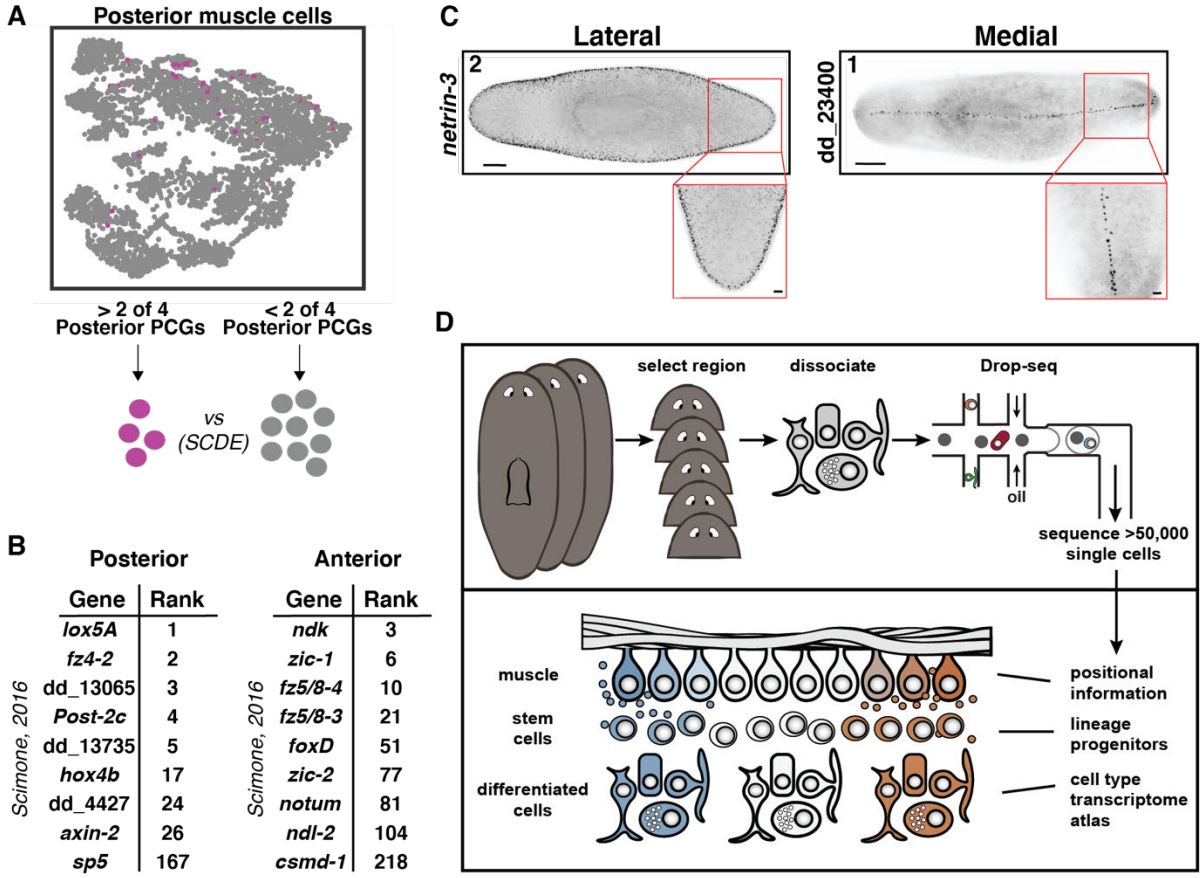
Figure 2.41. Targeted sequencing of sexual planarian anatomy.

(**A**) Diagram depicting 3 regions isolated from mature hermaphrodites. (**B**) t-SNE representation of 31 clusters generated from the sexual planarian sequencing data. C = *cathepsin*$^{+}$ cells, CE = ciliated epidermis, EEP = early epidermal progenitors, I = intestine, LEP = late epidermal progenitors, M = muscle, N = neural, Nb = neoblast, PC = pigment cell, PN = protonephridia, PP = parenchymal. Identity of numbered clusters unknown. (**C**) Sexual t-SNE plot colored by testis cells positive for expression (>0.5, ln(UMI-per-10,000+1)) of *Y box protein 4-like protein*, Contig50287, and *PLS3* (Contig40669). Arrows indicate the seven positive cells. (**D**) Sexual t-SNE plot colored by yolk cells positive for expression (>0.5, ln(UMI-per-10,000+1)) of Contig10743, Contig5529, *C-type lectin-like protein*, Contig45120, *FTMT* (Contig47570), Contig50285, Contig27235, and *putative surfactant B-associated protein*. Arrows indicate the seven positive cells.

Discovery of novel patterning genes

Planarians constitutively express dozens of genes associated with patterning (PCGs) in complex spatial patterns across body axes (***18***). PCGs are almost exclusively expressed in muscle (***19***). AP-axis PCGs are well established, including with muscle SCS (***71***). Muscle cells did not subcluster according to their anatomical positions (Figures 2.42A and 2.43, A to C). However, we reasoned that expression of known PCGs could ascribe locations to muscle cells in the data. Because of variability in the expression of any one PCG, muscle cell regional identity was determined on the basis of expression of at least two PCGs. For example, posterior muscle cells were identified by coexpression of at least two of the four posterior PCGs *wnt11-1*, *wnt11-2*, *fz4-1*, and *wntP-2*, yielding 163 cells (Figures 2.42A and 2.43D). Differential expression analysis using the algorithm SCDE (***72***) was performed on these 163 cells against the 4851 other muscle cells (Figure 2.42A and Table 2.4). Strikingly, nine of the differentially expressed genes were identified by Scimone *et al*. (***71***) as posterior-enriched; eight of these were within the top 26 genes identified by differential expression analysis (Figure 2.42B, hypergeometric $P = 2.75 \times 10^{-9}$). A similar analysis on 837 anterior muscle cells was also performed (Figure 2.43, A and D, and Table 2.4). Nine of the differentially expressed genes were identified by Scimone *et al*. (***71***); four of the genes were within the top 25 genes identified by SCDE (Figure 2.42B, hypergeometric $P = 5.80 \times 10^{-5}$). We also applied this approach to the less well studied ML axis. We identified 62 lateral muscle cells, and FISH with 15 of the top genes identified seven with lateral muscle expression (Figures 2.42C and 2.43, B and D to G) (***73, 74***). We isolated 90 medial muscle cells, and the top-ranked gene displayed a striking thin stripe of expression down the dorsal midline (Figures 2.42C and 2.43, C, D, and H). Together, these results demonstrate the power of deep SCS for identifying regional gene expression, such as that involved in patterning, in adult animal tissues.

Figure 2.42



B

Scimone, 2016

Posterior		Anterior	
Gene	Rank	Gene	Rank
lox5A	1	*ndk*	3
fz4-2	2	*zic-1*	6
dd_13065	3	*fz5/8-4*	10
Post-2c	4	*fz5/8-3*	21
dd_13735	5	*foxD*	51
hox4b	17	*zic-2*	77
dd_4427	24	*notum*	81
axin-2	26	*ndl-2*	104
sp5	167	*csmd-1*	218

Figure 2.42. Identification of new regionally expressed genes in muscle.

(**A**) Top: t-SNE plot colored by muscle cells positive for expression ≥ 0.5 [ln(UMI-per-10,000 + 1)] of two of the four posterior PCGs *wnt11-1*, *wnt11-2*, *fz-4*, and *wntP-2*. Pink, positive cells; gray, negative cells. Bottom: Transcriptomes for posterior muscle cells were compared to all other muscle cells by SCDE. (**B**) List of differentially expressed genes in posterior and anterior muscle cells that were identified in Scimone *et al*. (***71***). Rank indicates the rank of the gene in our analysis. (**C**) FISH images of one lateral and one medial expressed gene ranked highly in this analysis (***73***). Number indicates gene rank in the list generated by SCDE. Scale bars: whole-mount images, 200 µm; insets, 50 µm. (**D**) Illustration highlighting the capacity of the data set to identify almost all cell types in the planarian, as well as specialized neoblast progenitors and novel patterning information from the adult animal.

Figure 2.43

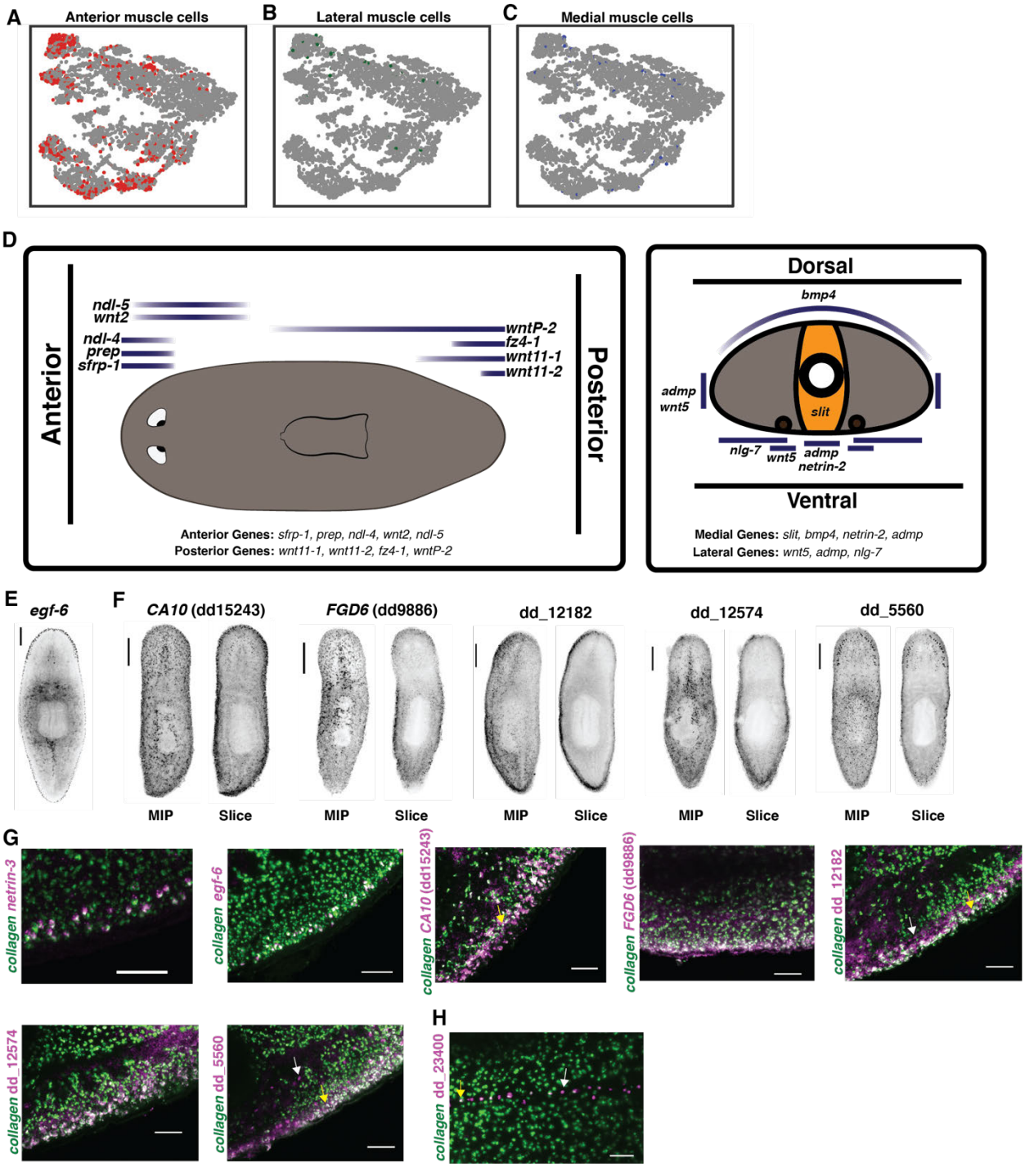

A
Anterior muscle cells
B
Lateral muscle cells
C
Medial muscle cells
D
Anterior
Posterior
ndl-5
wnt2
ndl-4
prep
sfrp-1
wntP-2
fz4-1
wnt11-1
wnt11-2
Anterior Genes: sfrp-1, prep, ndl-4, wnt2, ndl-5
Posterior Genes: wnt11-1, wnt11-2, fz4-1, wntP-2
Dorsal
bmp4
admp
wnt5
slit
nlg-7
wnt5
admp
netrin-2
Ventral
Medial Genes: slit, bmp4, netrin-2, admp
Lateral Genes: wnt5, admp, nlg-7
E
egf-6
F
CA10 (dd15243)
FGD6 (dd9886)
dd_12182
dd_12574
dd_5560
MIP
Slice
G
collagen netrin-3
collagen egf-6
collagen CA10 (dd15243)
collagen FGD6 (dd9886)
collagen dd_12182
collagen dd_12574
collagen dd_5560
H
collagen dd_23400

Figure 2.43. Additional data regarding the identification of regionally expressed muscle genes.

(**A**) t-SNE plot colored by muscle cells positive for expression (>0.5, ln(UMI-per-10,000+1)) of 2 of the 5 anterior PCGs (***19***) *sfrp-1* (***49***), *ndl-4* (***75***), *prep* (***76***), *wnt2* (***49***), and *ndl-5* (***71***). Positive cells, red; negative cells, grey. (**B**) Same as (A) for muscle cells positive for expression of 2/3 lateral PCGs (***19***) *admp* (***77***, ***78***), *wnt5* (***79***), and *nlg-7* (***80***). Positive cells, green; negative cells, grey. (**C**) Same as (A) for muscle cells positive for expression of 2/4 medial PCGs (***19***) *slit* (***81***), *bmp4* (***67***, ***82***, ***83***), *netrin-2* (***84***), and *admp* (***77***, ***78***). Positive cells, blue; negative cells, grey. (**D**) Diagrams depicting the expression domains of the PCGs used in this study (modified from (***18***)). Left: Dorsal view of the planarian is depicted. Right: Transverse section is depicted, with a centrally located gut branch and two ventrally biased ventral nerve cords. (**E**) FISH image of the laterally expressed gene *egf-6* (***73***). (**F**) FISH images of five genes that display lateral expression in muscle cells, but that also express other domains of expression. MIP, maximum intensity projection; Slice, single optical slice through the animal at the DV median plane. (**G**) Double FISH images of *collagen* (***19***) and the seven laterally expressed genes from (E) and (F). White signal indicates co-expression. Yellow arrows: co-expression. White arrows: no co-expression. (**H**) Same as (G) for the medially expressed gene dd_23400. Scale bars: whole-animal images, 200 µm; G and H, 50 µm.

Discussion

RNA sequencing of >50,000 cells (in total, 66,783 cells were sequenced) of the planarian *S. mediterranea* allowed the identification of transcriptomes for most to all cell types of an adult animal. This includes transcriptomes for cell types present as rarely as 10 cells in an animal with 10^5 to 10^6 cells, which strongly suggests that we have reached near-saturation. Sequencing of different body regions and assessment of rare cell type coverage in an iterative process enabled us to reach this saturation level. Some cell types might escape detection by this technique if they are exceptionally rare or hard to dissociate from the animal. Our data did indicate that some cell types were preferentially recovered according to the abundance of that cell type by FISH, whereas others were less represented (Figure 2.44, A and B). In particular, *prog-1*$^+$ epidermal progenitor cells were highly overrepresented in the data relative to their prevalence in the animal, perhaps because their small size made their isolation easier (Figure 2.44A). Absent *prog-1*$^+$ cells, most other cell types analyzed were represented similarly to their relative abundance in the animal (Figure 2.44B). Regardless of differences in ease of dissociation between cell types, we recovered data from all known cell types assessed. Not every known rare cell type emerged as a separable cluster; that is, these cells were sometimes embedded within a larger cluster. In some instances, further rounds of subclustering based on such knowledge resulted in splitting of subclusters into additional subclusters. Therefore, further subclustering analyses and even deeper sequencing will likely continue to enhance the capacity to computationally isolate rare cell types from other clusters. Nonetheless, the transcriptomes for such rare cell types are present in our data and can be studied by searching for the desired cells. Another challenge inherent in assessing saturation of cell type sequencing is ambiguity with the term cell type. Gene expression heterogeneity exists within well-defined clusters and could reflect differences attributable to technical sampling error, cell type state differences, or robust differences in biological function. Further in vivo morphological and functional studies with identified cell clusters, further computational analyses, and

even more sequencing data can continue to refine the knowledge of biologically important cell type differences.

Cell types have been previously identified largely through morphological descriptions and perhaps a few marker genes. Determining cell type transcriptomes with large-scale SCS is a powerful new approach to defining the cell type constitution of a tissue, an organ, or even a complete animal. In our study, we identified a large number of previously uncharacterized planarian cell populations across multiple tissues. This included multiple cell populations (in the *cathepsin*$^+$ group) previously undescribed at the molecular level. One cell population, defined by dd_9 expression, had long processes filling parenchymal space and surrounding, but excluded from, other planarian tissues. This pattern is reminiscent of “fixed parenchymal cells,” a largely uncharacterized cell population described by histology and electron microscopy (EM) (***52***). Previous EM work suggested that fixed parenchymal cells are likely phagocytic, with clearly observed lysosomes; dd_9$^+$ cells highly expressed genes encoding a variety of digestive enzymes and endocytosis proteins, providing further support for this hypothesis (Table 2.2). The biology of these *cathepsin*$^+$ cells and all the other diverse cell types identified in this work can now be studied in depth using identified transcriptomes and the tools of planarian biology research. For instance, we show for two case studies above that RNAi of a gene encoding a transcription factor with enriched expression in a candidate cell lineage leads to ablation of the predicted differentiated cell.

Generating transcriptomes for most to all cell types in an animal will be invaluable for studying gene function and the biology and evolution of a large range of important cell types. Because of their phylogenetic position within the Spiralian superphylum (***85***), major cell types found across diverse bilaterians (e.g., shared between humans and *Drosophila*, *C. elegans*, molluscs, annelids, and/or other bilaterians) should have been present in the last common ancestor of planarians and humans. As such, studying the transcriptomes and associated genes with cell type–enriched expression in this data set can allow characterization of the gene function underlying the biology of these cells.

Planarian biology presents many features that made this organism attractive for comprehensive SCS. Planarians are a model for studying numerous important problems in regeneration, stem cell biology, patterning, and evolution. At a single time point—the adult—there exist progenitors for essentially all cell types and the patterning information for guiding new cell type production. We identified the transcriptomes of numerous candidate transition states in lineages from pluripotent stem cell to diverse differentiated cell types. Furthermore, we used the data to identify novel regionally expressed genes in planarian muscle (the site of patterning gene expression). Together, these results illustrate the capacity of our data set to define cell type transcriptomes, identify lineage transition states, and ascertain novel patterning information, all from a single time point (Figure 2.42D). We propose that this atlas-like data set of cell type transcriptomes, much like the genome sequence of an animal, can serve as a resource fueling an immense amount of research, not only in planarians but in other bilaterians with similar cell types. To facilitate such study, we developed an online resource that generates cluster expression data, for any gene, across all clusters and subclusters (digiworm.wi.mit.edu). Case study model organisms have proved to be valuable testing grounds for developing approaches to complete genome sequencing; these planarian SCS data demonstrate an approach to near-to-complete cell type transcriptome identification that could be applied broadly to diverse organisms with varying degrees of information about cell type composition. The remarkable ability of single-cell RNA sequencing to reach nearly complete saturation of transcriptome identification for all the cell types of an animal represents a powerful approach for describing the anatomy of complete organisms at the molecular level.

Figure 2.44

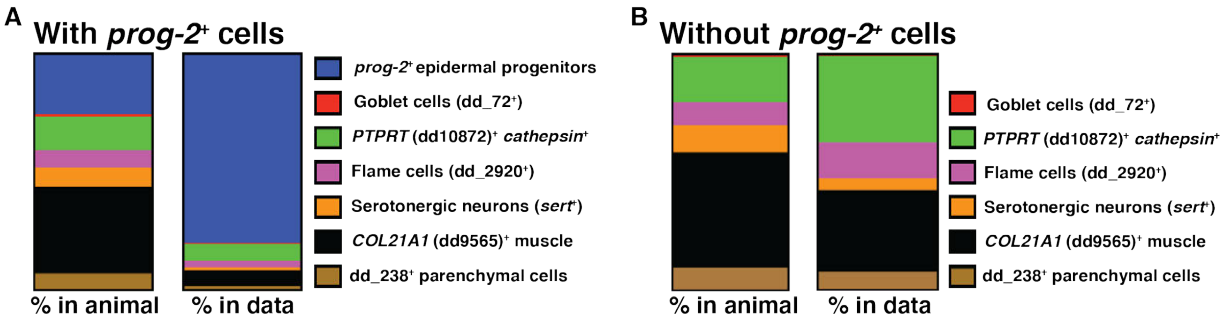

A
With prog-2+ cells
prog-2+ epidermal progenitors
Goblet cells (dd_72+)
PTPRT (dd10872)+ cathepsin+
Flame cells (dd_2920+)
Serotonergic neurons (sert+)
COL21A1 (dd9565)+ muscle
dd_238+ parenchymal cells
% in animal
% in data
B
Without prog-2+ cells
Goblet cells (dd_72+)
PTPRT (dd10872)+ cathepsin+
Flame cells (dd_2920+)
Serotonergic neurons (sert+)
COL21A1 (dd9565)+ muscle
dd_238+ parenchymal cells
% in animal
% in data

Figure 2.44. Comparison of relative cell type proportions in the animal and in the data.

(**A**) Left: Plot of the relative proportions of 7 distinct cell types in the planarian head, as determined by FISH across 4-5 animals. Right: Plot of the relative proportion of these cell types in the data. *prog-2*$^{+}$ epidermal progenitors are significantly over-enriched in the animal compared to in the data. (**B**) Plot of the relative proportions of the cell types in (A), excluding *prog-2*$^{+}$ epidermal progenitors.

Materials and Methods

Animal care

Asexual *Schmidtea mediterranea* strain (CIW4) and sexual strain (S2F1L3F2L3) animals were used. Animals were starved for at least 7 days prior to experiments.

Cell culture HEK293T and NIH/3T3 cell lines (ATCC) were cultured in DMEM (Invitrogen) supplemented with 10% FBS (Life Technologies) and 1% penicillin-streptomycin (ThermoFisher, Inc.).

Prior to Drop-seq run, cells were grown to confluence. Cells were washed once with 1X PBS. NIH/3T3 cells were then treated with TrypLE (Invitrogen) for five minutes and quenched with an equal volume of growth medium. HEK293T cells were resuspended in 1X PBS by manual dissociation. Cells were spun down at 300xg for five minutes and resuspended in 1X PBS + 0.01% BSA. Cells were spun down again at 300xg for three minutes and resuspended in 1 mL 1X PBS. Cells were passed through a 40µm filter and counted. 1X PBS + 0.01% BSA was used to generate a 1:1 mixture of NIH/3T3:HEK293T cells at a final concentration of 191 cells/µl.

Gene annotation and nomenclature

Genes were labeled as previously described (***35***) with one modification. Briefly, planarian genes that were previously reported and submitted to the Nucleotide database in the NCBI website (https://www.ncbi.nlm.nih.gov/nucleotide/) appear in italics. Sequences that are not found in the planarian nucleotide database, but have a human best-blast hit (blastx; E-value < 10-5; (***86***)), are labeled in uppercase with their human gene name, followed by the contig ID of the appropriate transcriptome assembly (***21***, ***87***) in parentheses. dd_Smed_v4 (***21***) contigs are prefixed by dd. Contigs that have no planarian identifier or human Blast hit are labeled according to the contig ID. dd_Smed_v4 (***21***) contigs are prefixed by dd_.Table 2.5 includes mapping of gene labels to contig IDs.

Delineation of planarian body sections

Animals were divided into five body sections for the principal sequencing (Figure 2.1A): (1) head – extending from head tip to below the auricles; (2) prepharyngeal – below the head region and above the pharynx; (3) trunk – starting above the pharynx and reaching below the pharynx. Pharynx was removed from trunk; (4) tail – starting below pharynx and extending to tail tip; (5) pharynx – the pharynx was isolated by making two shallow incisions perpendicular to one another at the base of the pharynx and along the length of the pharynx (***20***). For the targeted sequencing of brain cells, animals were divided into two body sections (Figure 2.20F): (1) a square directly below the eyes, enriched for the ventral half of the section and (2) a square surrounding the eyes and extending three eye-lengths below the eyes. For the sequencing from sexual animals, two prepharyngeal and one postpharyngeal region were used (Figure 2.41A).

Generation of planarian single cell suspensions

Cell suspensions were generated as recently described (***24***), by using 25 to 90 whole animals or body sections in each experiment. Briefly, samples were cut into small fragments (<1mm) and transferred to a 50 mL tube with CMFB (400mg/L NaH2PO4, 800mg/L NaCl, 1200 mg/L KCl, 800 mg/L NaHCO3, 240 mg/L glucose, 1% BSA, 15mM HEPES, pH 7.3). Volume was brought to 45 mL with CMFB, to which 5 mL collagenase (1 mg/ml; Sigma-aldrich, C0130) was added. Suspension was generated through agitation of the mixture for 5-10 minutes using pipettes. Cell suspension was passed through a 40µm filter and centrifuged at 1250 rpm for 5 minutes. The supernatant was removed and cells were resuspended in 1 mL of CMFB. Cells were incubated in Hoechst 33342 (40 µl/ml, Invitrogen) for 45 minutes. Propidium iodide (3 µl/ml, Sigma-aldrich) was added immediately prior to fluorescence-activated cell sorting (FACS). Planarian cell fractions (***88***) were defined, and sorted into CMFB (0.01% BSA). In seven samples, constituting Drop-seq runs 11-14, 17-21, 23-25, and brain schemes 1 and 2 (Figures 2.2A and 2.20F), cells with 4C DNA content as determined by high Hoechst signal were gated out.

Methods for Drop-seq, library preparation, and sequencing

Planarian Drop-seq runs were performed as described previously (***6***), using the optimized 'Drop-seq Laboratory Protocol', v 3.1 (http://mccarrolllab.com/download/905/) with minor modifications. Briefly, cells were diluted to a concentration of 191 cells/µl in 0.01% CMFB, and beads were diluted to a concentration of 148 beads/µl in lysis buffer. Droplet generation oil (Bio Rad, #1864006), cells, and "Barcoded Bead SeqB" beads (ChemGenes Corporation, Wilmington MA) were concurrently pumped through a PDMS co-flow microfluidic droplet generation device (Nanoshift LLC, Emeryville CA), encompassing cells and beads within oil droplets. Cells were lysed within the droplets and cellular mRNA was captured by the barcoded bead. Depending on the cell yield for each biological sample, approximately 1 mL of cells and 1 mL of beads were collected in droplets, which were broken by 1H,1H,2H,2H-Perfluoro-1-octanol (Sigma-aldrich). Beads were isolated and reverse transcription was performed, adding a unique cellular barcode, as well as a molecular barcode, to each cDNA. Following an exonuclease treatment, PCR amplification was performed on aliquots of 2,000 beads using 4 + 12 cycles. Pairs of PCR reactions were purified using 0.6X AMPure XP beads (Agencourt). Concentration of amplified DNA was measured by Qubit (ThermoFisher, Inc.), and equal amounts of DNA from each pair of PCR reactions were combined to 600 pg total in 5 µl H2O for each Drop-seq run. Pooled samples were tagmented using the Nextera XT v2 DNA sample preparation kit (Illumina, Inc.) by incubating the sample for 5 minutes at 55°C. Libraries were purified by two rounds of AMPure XP purification (0.6X, 1X) and sequenced on Illumina NextSeq 500 (Illumina, Inc.). Read 1 was 20 bp; Read 1 index was 8 bp; Read 2 (paired end) was 63 bp.

The Drop-seq run on cultured cells was performed as above, with a few modifications. Cell mixture was diluted to a concentration of 191 cells/µl in 1X PBS + 0.01% BSA. PCR amplification was performed on aliquots of 2,000 beads using 4 + 9 cycles. Single PCR reactions were purified using 0.6X AMPure XP beads. Separate libraries were

generated for each PCR reaction and were purified by one round of AMPure XP purification (0.6X). Pooled libraries were sequenced on Illumina Miseq (Illumina, Inc.). Read1 was 20 bp; Read 1 index was 8 bp; Read 2 (paired end) was 80 bp.

Description of Drop-seq runs

As summarized in Figure 2.2A, 25 Drop-seq runs were performed on 15 distinct biological samples for the principal sequencing. Indexed Drop-seq libraries were combined to ~6,000-8,000 cells per sequencing run. Combined libraries were sequenced across 8 different sequencing runs, yielding ~375-550 million reads per run. Drop-seq run 19 (sequencing run 6) was downsampled to 300 million reads prior to data processing and alignment.

The choice of body section for each Drop-seq run was determined by the presence of rare cell types from each body section in the data. Initially, Drop-seq runs 1-10 were performed on approximately equal numbers of cells from each body section, excluding the pharynx, to assess the number of rare cell types present in the data from each section. For example, it was found upon analyzing these preliminary results that rare neuronal cell types, such as the photoreceptors of the eye (***20***), were present at very low numbers in the data. As such, 4 additional Drop-seq runs were performed on cells from the head. This logic was used to guide the decision regarding choice of body section for all other principal Drop-seq runs.

For the targeted sequencing of brain cells, two Drop-seq runs were performed on two distinct biological samples, as described above. Indexed Drop-seq libraries were combined to ~8,000 cells and sequenced on one sequencing run, yielding ~550 million reads. For the sequencing from sexual animals, three Drop-seq runs were performed on one biological sample, as described above. Indexed libraries were combined to ~8,500 cells and sequenced on one sequencing run, yielding ~450 million reads. For the sequencing of cultured cells, one Drop-seq run was performed. Indexed Drop-seq

libraries were combined to ~300 cells and sequenced on one sequencing run, yielding ~18.5 million reads.

Data processing and alignment of Drop-seq data

The ‘Drop-seq core computational protocol’, v1.2, as developed by Jim Nemesh in the McCaroll lab (http://mccarrolllab.com/wp-content/uploads/2016/03/DropseqAlignmentCookbookv1.2Jan2016.pdf), was used to process and align the sequencing data with one modification. Briefly, sequencing reads were tagged with their associated cell and molecular barcodes, followed by the trimming of 5’ primer and 3’ polyA sequences. Sequences generated from asexual animals were mapped using bowtie to the dd_Smed_v4 assembly (***21***) with the following parameters [--best -S]. The following four sequences were added to the transcriptome for mapping.

> SMED_11901_V2 -
ATGAATGAAATTTTGGAAAAGGATATGAAAGCGATTGAATCCATTAAAGTAAAAGAA
AAAAAGGCTGTTGATGGTTTTATGGGTACCTCATCGTTTCATGGAGTGATTCAAGC
ATATCATAAACGAAATAAAATTGATAAAGGGAGCTGGTTCATCAGTTTAGTTATTTG
TATGTTTGGCTTAATTGGGCATCTCTACCTAATAATCAGTAGATATATAAGTTTGCC
CACAACTATTGACATGGTCTCTTCAGTGAATTTTGATCCTTTTCCTGCTGTCGCAAT
ATGTCCGGTTACCTTTATTAGCAGGGATAAATTCACCAAGTATTACAATACAACTCA
AGTTTCCCTTAATAAAAAGCTAGTTGGGGATATTTTCTACGTCGATGTAAGTGCCTT
GAATTTCTGGAGGTCCCTAAGTAAACAACAAGGCAAAGACATAAACAGTAGTTCAG
TTCTTGGAAAGTATTGGGATGAAGCTGAAACCACTTTCTATAGATTCCAGAAAATGA
TGAATGTTTCAATAGGTCATCGAAATTATGAAATGATTTTCTTTTGTGAAATTAACAA
TAAACCTTGCTCATGGGAACATTTCCTTGAATTCGATCATCCGATTTATAAGCGATG
TTTTAAATTCTCCTATCCGGTAACTGATGAAGATGAAATTCCAGATAAATTGATATTG
GGGCTTTATGTTGATGATGACTATCAAAGAGACACTGATGATATTAAAACGATAATA
ACCTCTCATGGAGGAAAGGTTACTATAAATGAAGCAAGTATTTACCCTGGAACTGA
AAGTTCATTTGAACATTTTCCGTCAGGATTCCAAACGATGTTTCGATTGAAACAAGA
AGGTAGCAGTCAAATCAATAAACCAAGGTCTCCATGCCAAGTTAATACTGATTCAG

TGATCAACGTTTTCAACGATTATGAATATGATGGCTCAACAAATATCACAATACCAT
ATAAATACAATGTGATACTTTGCAGACAATACCATCAACAAATAGAATGCGTTAAAA
GATGCAAGTGTTTAAATCCGAACATTCCAGTATTTGTTGATGCTATTAAGAATTCTG
AAAATAAATCATTCTTTTGCGATGAGATTCAGCTTAATTCTTCCTTTTCAAGCATTAT
TAATCAGCTTGATTGTCTTTATAATTTAGATTATGATCAGTATTTTAATGAGAATGTT
ATATCATTATGTTCGGGATTGTGTAATCAGGTAGAATATTCAATGTATTCTTATACTA
TGCCTTGGTTCGGTAAAACAATGATCAAAGAAATGGAGTTTGTCNAATGAGAAATT
CATGGCCCATTACAACAGTCTAATTAATTCCATCCAATTGTGAAGGATTATGGAACC
ATTGAATAGAGCACGCAATTGCGTAATCAAATCCATGAAAGATAATGATCAAGCCA
GCTTGTGTTTCGCAATGATTAATATTCAATTTGAATCTCCCAGAAAAGAAATTATTC
GAGAATATGAGGCATATTTATTGGGGAATTTACTCAGTGATTTCGGCGGGATTTTA
GGACTGTGGATTGGAATGTCTCTGATAACAATTATTGAAATCATATACTTAGCATGC
TCGTTGAGTAAACACAAAACTGAACGCGCTGCTTCAGTTTTCAAAAAGTCAATCCA
CAAGAGAAGTCTGAAAAGGAATTCCGATAAAAACAAAATTATCAGAATCGGAATAG
AAAATGAGGCGTATGAAAATTAG

>dd_Smed_v4_0_0_1 -
TTTTTTTTTTTTTTTTCTAAGCAGTGGTATCAACGCAGAGTACGGGGGGTATTGCAC
TGTTTAGTTGTGATATTTTCCTTTTTGTATTACGGTTTGTAGGTATTTTTATGTTTTTA
TTCTCTCGAATGGAATATGATATATCTTTTGTGTGTTTTGTTTTATTTTCTTTGGTTAA
CATTAGAATTACATTAATTGATAGTAGTGATATGTTTTCGGATTTCTTTATCTAGTTTT
TAATTGTCTTTAGGATAAGCTTTTGTTTTACTTGTTTTTTTTTCTTATTTGTTTAGTTAT
TCTTAATAGTATTTCTTTAGATTGTTAAAGATTTGTTAGTCGTTTGTTTGCTTATATG
GTTTTAGTAATTCTTTTTAGCAATTACGTAATTAATATGAATTATGCTATTTATAACTC
AATGCGTCTACAACTGTTTTTTAAAAACATTTCATTTTTTGTAAAATGTAGTCCCTGC
TCACTGATAAGTTAAATAGCTGCAGTACTTTGACTGTACGAAGGTAGCATAATTACT
TGTCTACTAATTCTAGAATTGTTTGAATGGGTTTATTGATAGATAGCAAGTTTTTATT
TAGCCTTGTTTTTTAATTTATACTTTCTGTAAAGATACAGTTTGTTATTTCAAGGACG
AAAAGACCCTAGAGAGTTTTTAACTTAGTGGTGTTTCTTACTTTAGTTATTTGTTGG
GGTAACGGTATTTATTTTGAATACTTATTTATTACATTATGAACTTCCTTAGGGATAA

CAGGGATATAGAATCTTGGAGGACATATCGAAGATTTTGTTTTCTACCTCGATGTTG
AATTGTAGTTAAAGATAGGTGTAGAGGCCTTTACTTTTAGTCTGTTCGACTAATAAT
ATTATTCGTGATTTGAGTTTAGACCGATGTGAATCAGGTTGGTTTTTATCCTGAATTT
CTGTCCATTGTACGAAAGGAATTGGATTGGTATTAAAAAAAAAAAAAAAAAAAAAA

>mtRNA_2 -

TCTAAGCAGTGGTATCAACGCAGAGTACGGGGGGTATTGCACTGTTTAGTTGTGAT
ATTTTCCTTTTTGTATTACGGTTTGTAGGTATTTTTATGTTTTTATTCTCTCGAATGGA
ATATGATATATCTTTTGTGTGTTTTGTTTTATTTTCTTTGGTTAACATTAGAATTACAT
TAATTGATAGTAGTGATATGTTTTCGGATTTCTTTATCTAGTTTTTAATTGTCTTTAG
GATAAGCTTTTGTTTTACTTGTTTTTTTTTCTTATTTGTTTAGTTATTCTTAATAGTATT
TCTTTAGATTGTTAAAGATTTGTTAGTCGTTTGTTTGCTTATATGGTTTTAGTAATTCT
TTTTAGCAATTACGTAATTAATATGAATTATGCTATTTATAACTCAATGCGTCTACAA
CTGTTTTTTTAAAAACATTTCATTTTTTGTAAAATGTAGTCCCTGCTCACTGATAAGTT
AAATAGCTGCAGTACTTTGACTGTACGAAGGTAGCATAATTACTTGTCTACTAATTC
TAGAATTGTTTGAATGGGTTTATTGATAGATAGCAAGTTTTTATTTAGCCTTGTTTTT
TAATTTATACTTTCTGTAAAGATACAGTTTGTTATTTCAAGGACGAAAAGACCCTAG
AGAGTTTTTAACTTAGTGGTGTTTCTTACTTTAGTTATTTGTTGGGGTAACGGTATTT
ATTTTGAATACTTATTTATTACATTATGAACTTCCTTAGGGATAACAGGGATATAGAA
TCTTGGAGGACATATCGAAGATTTTGTTTTCTACCTCGATGTTGAATTGTAGTTAAA
GATAGGTGTAGAGGCCTTTACTTTTAGTCTGTTCGACTAATAATATTATTCGTGATTT
GAGTTTAGACCGATGTGAATCAGGTTGGTTTTTATCCTGAATTTCTGTCCATTGTAC
GAAAGGAATTGGATTGGTATTAAA

>mtRNA_1 -

AGTTGGTGTTGTTGTTTTGTGCAGGTAAGTTAATTAAAACTAGCAGATTCATGTTCT
GTCTATGAGTCCTTTCTCTGTATATGTGGTTAAGATAGTTTATTCAGAATGTTAATTT
GTGGAGTTAATGGTAAAAGACTTGTTTTTCTTAATATTTGTTTTAATAGCTTAATTTA
AAGTGGAGTTTTGAAGAAGCTCAGTTGTTGGTTTCCTTCTTGAAACAATATAGTTTT
GTCTTGGCTTTTTAAACTGGCATGCTGTTAAACGCTATAGTTTATTGGTGGTTGTTT

ACTATTAATTTTGTTTAATGGTTTTATTAGGCGTGTATATTTTAAAATTTAATGTAAAT
TGATTGCTTGAGTCGGTATATGCTATTAGGAGATCAAATGAGTGCCAGCTTCTGCG
GTTACACTTTGTATTACTATGTTAGTTTATTATTTGGTTTAAATTGGTTAAGTTTCAAT
AAGAGACTTTATGTATGACTAGTGGTAGATTTTAATACTTTTATTAGTTTTACTTCCT
TTTTAGACATGAATCTGGCTTTATTTATAAGGGTTGTTTATTTTATTTCCTCATCAAA
ATGAAAAGACTTGGCAGTTGTTCTAATTATTTGGGGAGTGTGGGTTTAGAAAAGAG
TATCCGCTCAATATCTCGCTAAGATTATGGTTAGTGTACGGTTGTACATATGTGAAT
GGCCTTATAGTTATGCTTTCTTTAATGCAAATCATTGTGCTGCTTATCTTAGATTATG
CTTTCACTACATTGGTTAGATACCTTTTGAAATAATTGGTGTTGATCAGGACTAAATA
GTAAATTTAGATGAATTGGCTTTTTTGAATCTTTTCTAGGACTTAGTACACACCGCC
CGTCAATCTCCGTTCTTTAAGAGGAGTTAAGTCGTAACATGGCG

Sequences generated from sexual animals were mapped using bowtie to a separate transcriptome assembly generated from *S. mediterranea* hermaphrodites (***87***) with the following parameters [--best -S]. The above four sequences were not added to the transcriptome for mapping. Sequences generated from cultured cells were mapped using STAR to a mixed human + mouse genome (http://mccarrolllab.com/dropseq/).

Using the DetectBeadSynthesisErrors module [NUM_BARCODES=2X expected cell number, PRIMER_SEQUENCE=AAGCAGTGGTATCAACGCAGAGTAC], errors in barcode sequence associated with bead synthesis were detected and were either corrected, if possible, or the associated reads were removed. The number of cells in a run was estimated by plotting the cumulative distribution of number of reads per cell. The number of cells at the inflection point was used as the estimated cell number in the data. A gene expression matrix was generated for the expected number of cells using the module DigitalExpression [NUM_CORE_BARCODES= # cells at inflection point]. Contig isoforms for sequences generated from asexual animals were then merged by summing the mapped reads to each isoform. Finally, cell IDs were tagged with their body section of origin, and the resulting expression matrices were combined for all

sequencing runs. Generation of the mixed species plot for cultured cell data was performed as described in the 'Drop-seq core computational protocol.'

Quality filtering of single cell data

For the principal sequencing, cells with more than 18,000 unique molecular identifiers (UMIs) and cells expressing less than 500 genes were removed from the data to ensure that low quality cells and potential cell doublets were not present. Five transcripts identified as ribosomal or mitochondrial by BLAST (dd_Smed_v4_0_0_1, dd_Smed_v4_7_0_1, dd_Smed_v4_4_1_1, mtRNA_1, and mtRNA_2) were also removed as previously described (***24***). The average numbers of genes and UMIs for cells from each body section following quality filtering are included in the table below (Figure 2.2B).

Body Section	Mean nUMIs / Cell	Mean nGenes / Cell
Whole	3680	1604
Head	2921	1369
Prepharyngeal	2601	1258
Trunk	3535	1571
Tail	3334	1449
Pharynx	2446	1295
All Cells	3020	1404

Using the Seurat package, v2.0 (***22***), cells were normalized using the function Setup [is.expr=0.1, names.field = 2, names.delim = "_", total.expr=1e4, do.logNormalize = T], which divides cell UMIs in the expression matrix by the total number of UMIs per cell, then multiplies by 10,000 before transforming to log-scale (ln(UMIs – per – 10,000 +1)) (***6***). The RegressOut function from Seurat was then used to eliminate variation resulting from the number of UMIs in each cell with parameters [latent.vars = "nUMI"].

For the targeted sequencing of brain cells, cells with more than 15,400 unique molecular identifiers (UMIs) and cells expressing less than 500 genes were removed from the data. Five transcripts identified as ribosomal or mitochondrial by BLAST (dd_Smed_v4_0_0_1, dd_Smed_v4_7_0_1, dd_Smed_v4_4_1_1, mtRNA_1, and mtRNA_2) were also removed as previously described (***24***). The remaining cells had an average of 1,372 genes and 2,925 UMIs following quality filtering. For the sequencing from sexual animals, cells with more than 12,000 unique molecular identifiers (UMIs) and cells expressing less than 500 genes were removed from the data. The remaining cells had an average of 1,586 genes and 3,108 UMIs following quality filtering.

Gene saturation analysis

To determine the number of total sequencing reads required to comprehensively detect the landscape of gene expression in a single cell by Drop-seq, a sequencing library was generated from 197 cells isolated from a whole animal, as listed below. These cells were not included in the clustering analysis.

AGGAGGAATTAT, GCTATGCTTGAC, ATGAGCGATTCN, CTGTGTTCACGA, GAAATCAATGGC, CTTAAGGTATGT, ATCCAATAAGGT, CCTAACAGAATG, CTGAGCTTTTAT, TGTGACACCCCA, GGTAAACGATTA, TATGATCACCAN, AAGCTCCACCAA, ATGTGGAAGTCC, CATGTCAAAATN, AATGGATCCGAA, CTCTATCAGTGN, CCTGATTTCAGA, ATCATACCAGAA, CCCGCTTTCTAC, ACGTCTATAATT, TTCAATTGCTCG, CTAGTGCTGGGC, TGGTTTGCCCCC, AGACCTGGTGTA, CAATGAGAGGCA, TAATGTCGACGC, TAACAGAGAATT, CTATAGTGTTCG, GATAAGTAATAC, AACAGATATGCG, GAAGTAGCCATT, GTTCTGAGTAAG, TCTTTGAGGAAC, GGCAGCTGCGCG, CAAGGTCGACGA, GTTTTAGCTGCT, TGCCACCCATGC, GCGCCAAAGCGC, GAGATGCAATCG, CCCTCGCTCAGG, GGACTAGTGGAC, GTACCGGCGCGC, ACCAGCCCGGCG, AGCGAAAACAGT, CTAACGGACACG, TACAGTGTATGC, CAAGAATCGTGG, TCGACGCTGACC, CAATAAAAGTGC, GCCTAGCGATCT, TCCCACGTTCAG,

CGGGGGCGCCTN, CTCTAGTGGGAG, GCCGTCTTTGAC, TTCAACTCAATT, ATCGTCATTTCA, CCAATATGAAGA, TATTCCCTGGTC, GGAGAAATTGTT, ATGTAGATCCTA, GCGTTCTTGCCN, ACATACCGGTTG, GCTGCGGTGGGC, TCGTTTGCATCT, GTCGTCACACTT, TAGCAGAGACTT, GTTCATGACAAA, GTTTCTACGGAN, CGGTTGCCAAGG, AGGTGTTCGGGC, TCCCTATGGGAT, TACACATCGCCG, CCGTGCCTCCGA, ACAAGTAGCAAG, CCTCCATCGGAG, TATACAGGAGAT, ACATGTTCGGAT, AGAGCTATGCCC, TTGCGGCAGTGT, CAAGCCGGCGAC, CACGACGGTTTG, CGATGAAAACGC, TTAATTCCCCAG, AGCCCACCTGAC, ACAAGAATAGAC, TTTGATATGGCN, CTACTTCCTTCG, TCGGATTAGGCG, CCAGATGTGGCA, CTCAACACAGGA, TTAACATTCACT, CGATGTTTCTTT, TAAGGACCCAGA, CCTCGTATGCTG, CTACATTCGTGN, CGGGTTACTCAA, CCTCCTTACGTC, ACCCCTCACCTA, TTGAAATATCCA, AAATTACCTTCA, GAGTTGCTTGTA, TTGATACAATTG, CTCGTGTCAATN, CCTGTTATCCCT, TGGCTTCGAACT, TCTTAGTGGTTC, AACTGGTCCAAA, TCTTTTACACTT, GAGGCGGACTAG, GGCTTGGGCCCC, AAAGCCATCACT, AAGATGCCAAAC, GTACAGCTCTGA, GTTTTGCATGTG, ACCAAAACCTCG, GGGGTAATTAGA, TTTAGTAGAGCG, CCCCCCGGCGCT, AGAACCCAGACG, CCGGTTTGTGTT, GCGTGGGTGTCA, TACCGTGTTGCA, CCGTAATTTTAC, CGATGCCTTAGA, GTACAACAGAAT, TAGGGCTACACA, TACTCCAGAGGG, AGGCGAGTTTTA, TCCAATCCCTAC, AGCACGATTCAG, TCTTAATTTCCT, GAGTGGCTTTGC, TTATGTACGGGG, TCTTAACTCCCG, CTCCTTACGTCN, TAACTTCGCGCA, TGCAACCGGGCC, CGTCAGCGTTGC, CGAACTACCAAC, AATACGTTCCGC, CGCAAGCAATTT, TGATCGTTGACT, AAGACGTTTCGG, CTTTCGACTACT, CCAGGGCACCGT, GCCCCTATTCAC, TCCGCTTCAGTN, TTTTGGACGGGT, TAAGTCTTAATG, GAGCACTCTGAG, GGCTTGCCCCCT, CATTTTTAAGAG, GAAGGGTTGGTN, TCTCTGGAGAGC, GTCCTCTCCTTG, AATGTTGACAGC, AGTACGGGGTGC, ATAACCCCACGC, ACTAATTTTCCA, TTATCGGGGCAC, CCCGACATAACA, CACTGAGCCTCC, AGTCCATCACGG, CCACAAATCCTG, TCCCACAAGCGT, AAACGGTACCAC, TGTAAACGGGCA, CATTGTCAAAAT, AACCCCAACGTC, GCCTGTATCCCG, GTATTCGCGGCC,

CCTGGTACGTGC, AGCACCCATCAC, GATGCCTTAGAN, TCTCTGCTTTTA, TTTAGGTCTGTG, GTAGTATAGCAA, AAATTATATAGT, GGGGTTCGTTTA, CACCCCTCTGTA, TACCCATTCGTA, TGGGCAATTTAG, AAAATGTCCACT, CGCGGCTATTCC, CCAGCGGTAGTC, TTTGAAACATTT, TGGGCAGCTCGG, AGTACAGTGTAG, TAATTGCTATGA, CTCCCAACACCG, TAGGCCCGATAC, CGACAAGGAAAN, CCCATTTAATTC, ATCATCAGCTAC, CACCAAAATCCG, CTCATTCCGCTG, TTCCCGGCCACT

Cells were sequenced on the Illumina NextSeq 500, generating over 550 million reads, and the data was aligned and processed as described above. Following the detection of bead synthesis errors, the data was downsampled to 2.5% of the total read count (14M total reads). Cells with > 9,000 UMIs and cells expressing < 500 genes were removed from the data, yielding 197 cells expressing an average of 1,500 genes per cell, similar to the average 1,404 genes per cell from the main data. The full data was then progressively downsampled 23 additional times to a range of 530 million to 870,000 reads prior to the generation of expression matrices. Expression matrices for each of these downsampled datasets were then generated for the 197 cells identified above and the average number of genes expressed per cell was calculated and plotted as a function of total read count (Figure 2.2C). Fitting a one-phase exponential association function to the data ($y = 882.4 + (10647 - 882.4) * (1 - e^{-3.29E-9*x})$) reveals a theoretical plateau of 10,647 genes.

Initial clustering of all cells

The Seurat package, v2.0 (***22***), was used for all steps of clustering, following the Seurat package documentation (http://satijalab.org/seurat/). Briefly, the Seurat function MeanVarPlot was used to identify genes with high variance and high expression using the parameters [y.cutoff = -.5, x.low.cutoff=.2, x.high.cutoff=15, fxn.x = expMean, fxn.y=logVarDivMean, set.var.genes = TRUE]. These genes were then used as input for principal component analysis using the function PCA [pcs.store = 150]. Cells were clustered using the function FindClusters [pc.use = c(1:150), resolution = 2], which

utilizes a graph-based clustering approach, and plotted in 2 dimensions by t-distributed stochastic neighbor embedding (t-SNE). The number of principal components used as input for FindClusters was determined empirically for each set of clusters. Namely, 150 principal components (PCs) were initially used as input for all clustering performed in this work, before decreasing the number of PCs used as input until optimal clustering occurred. This approach was superior to the identification of significant PCs using automated methods, such as bootstrapping (***24***). 63 clusters were generated from the initial clustering of all cells using 150 PCs and a resolution of 2. Cluster 10 contained no exclusively enriched genes. Rather, enriched genes were also highly expressed in cells from regions of most other clusters, suggesting cluster 10 in fact represented an artifact. Cells from cluster 10 were removed from the data, and the remaining cells were reclustered as above, again generating 63 clusters with similar identities. Eight of these clusters (8, 28, 29, 38, 42, 53, 56, and 57) possessed a number of genes with exclusively enriched expression. These cluster-enriched genes were not associated with any known planarian cell types or tissues, however, and were largely undetectable by fluorescence in-situ hybridization (FISH) despite being very highly expressed in the data. As such, these cells were removed from further analysis. Finally, three sets of clusters: 10, 55, 59, and 60; 9, 45, 50, 52, and 58; and 13, 31, 44, 46, and 54 were largely interspersed with one another in the t-SNE plot. These three sets of clusters were merged into three clusters. Following these changes, all clusters were re-numbered to reflect the new total cluster number.

For the targeted sequencing of brain cells, cells were clustered as above using 100 PCs and a resolution of 2. Cluster-enriched genes for cluster 7 of the clustering results were not associated with any known planarian cell types or tissues and were largely undetectable by fluorescence in-situ hybridization (FISH) despite being very highly expressed in the data. As such, these cells were removed from further analysis. For the sequencing from sexual animals, cells were clustered as above using 75 PCs and a resolution of 2. Clusters 6, 7, 18, 20, 21, 27, 30, and 32 were not associated with any known planarian cell types or tissues and were largely undetectable by fluorescence in-

situ hybridization (FISH) despite being very highly expressed in the data. As such, these cells were removed from further analysis. Cells were re-clustered using the same number of PCs, and a resolution of 3.

Subclustering subsets of cells

Clusters were assigned a tissue identity based on expression of tissue-specific markers (***24***) or expression of highly enriched transcripts (Table 2.2). Cells assigned to clusters that express genes from the same tissue class (***24***) were isolated to new Seurat objects using the SubsetData function. These cells were then re-clustered as described above. Resulting subclusters with similar cluster-enriched genes were combined for each of the tissue subclusters (see table below). All subclusters were then re-numbered to reflect the new total subcluster number. For the parenchymal subcluster, cluster 15 was manually split into two clearly distinct cell populations (14 and 19 in the final numbering). For the pharynx subcluster, clusters 8 and 9 were manually split into three clearly distinct cell populations (7,8, and 9 in the final numbering). These changes, along with the PCs and resolution parameters used as input for subclustering each tissue, are summarized in the table below.

Subcluster	**Combined Clusters**	**Split Clusters**	**PCs used**	**Resolution**
Parenchymal	2,5	15	25	2
Protonephridia	0,4; 2,3	None	10	1.5
Epidermal	0,1,2,3,5,7; 6,9,15; 8,11	None	15	2
All neural	3,4; 7,8,12,13,14; 15,16	None	50	5
Non-ciliated neurons	3,14; 0,38	None	75	5
Ciliated neurons	2,4,16; 6,10,15	None	35	4
***smedwi-1*$^{+}$**	None	None	35	2

Muscle	None	None	10	1.5
Intestine	0,1; 2,3	None	35	1.5
***cathepsin*⁺ cells**	None	None	25	1
Pharynx	None	8,9	20	1.5

Cells from the all neural subclustering above and cells assigned a neural identity in the targeted brain sequencing (Figure 2.20G) were combined and re-clustered using 75 PCs and a resolution of 2.

Identification of cluster-enriched genes

Cluster-enriched genes were identified using the FindAllMarkers function from Seurat using the following parameters [thresh.use = 0.25, test.use = "bimod" or “roc”, only.pos = T]. Specifically, enriched genes were identified using both a receiver operating characteristic curve (ROCC) analysis and a likelihood ratio test (LRT) test based on zero-inflated data (***23***), thresholding for genes that show at least a 0.25 fold average difference (log-scale) between all other clusters. Resulting p-values were adjusted for multiple hypothesis correction using the p.adjust R function with default parameters and transcripts with corrected p-values greater than 1E-4 were discarded (Table 2.2). Transcripts with corrected p-values greater than 1E-2 were discarded for subclustering of non-ciliated neuron cluster 3. A threshold for genes that show at least a 0.5 fold average difference (log-scale) between all other clusters was used for clustering of cells from the targeted brain sequencing and for the sequencing from sexual animals.

Cell type assignment for cell doublet analysis

Cell types used for the cell doublet analysis were identified as expressing six of eight cluster-enriched contigs (Table 2.2) using the Seurat function WhichCells [subset.name=, accept.low=0.5]. The cluster-enriched contigs used for the analysis are listed below.

Flame cells

dd_Smed_v4_5256_0_1

dd_Smed_v4_4268_0_1

dd_Smed_v4_2920_0_1

dd_Smed_v4_7255_0_1

dd_Smed_v4_6287_0_1

dd_Smed_v4_11608_0_1

dd_Smed_v4_16519_0_1

dd_Smed_v4_5409_0_1

Enterocytes

dd_Smed_v4_1_0_1

dd_Smed_v4_48_0_1

dd_Smed_v4_44_0_1

dd_Smed_v4_75_0_1

dd_Smed_v4_194_0_1

dd_Smed_v4_215_0_1

dd_Smed_v4_20_0_1

dd_Smed_v4_267_0_1

mag-1^{+} cells

dd_Smed_v4_451_0_1

dd_Smed_v4_769_0_1

dd_Smed_v4_14_0_1

dd_Smed_v4_557_0_1

dd_Smed_v4_1041_0_1

dd_Smed_v4_2759_0_1

dd_Smed_v4_8929_0_1

dd_Smed_v4_6728_0_1

dd_10872^{+} cells

dd_Smed_v4_10872_0_1

dd_Smed_v4_1831_0_1

dd_Smed_v4_551_0_1

dd_Smed_v4_10044_0_1

dd_Smed_v4_266_0_1

dd_Smed_v4_9638_0_1

dd_Smed_v4_8942_0_1

dd_Smed_v4_663_0_1

Ciliated Epidermis

dd_Smed_v4_357_0_1

dd_Smed_v4_298_0_1

dd_Smed_v4_181_0_1

dd_Smed_v4_877_0_1

dd_Smed_v4_817_1_1

dd_Smed_v4_351_0_1

dd_Smed_v4_155_2_1

dd_Smed_v4_709_0_1

Serotonergic neurons

dd_Smed_v4_585_0_1

dd_Smed_v4_8392_0_1

dd_Smed_v4_12700_0_1

dd_Smed_v4_5999_0_1

dd_Smed_v4_15253_0_1

dd_Smed_v4_11320_0_1

dd_Smed_v4_12323_0_1

dd_Smed_v4_20712_0_1

prog-2^{+} epidermal progenitors

dd_Smed_v4_478_0_1

dd_Smed_v4_332_0_1

dd_Smed_v4_213_0_1

dd_Smed_v4_6912_0_1

dd_Smed_v4_363_0_1

dd_Smed_v4_61_0_1

dd_Smed_v4_3549_0_1

dd_Smed_v4_69_0_1

Generic muscle cell

dd_Smed_v4_323_0_1

dd_Smed_v4_2337_0_1

dd_Smed_v4_2197_0_1

dd_Smed_v4_223_0_1

dd_Smed_v4_1579_0_1

dd_Smed_v4_579_0_1

dd_Smed_v4_436_0_1

dd_Smed_v4_402_0_1

Cell lineage reconstruction

The Monocle package, v2.6 (***45***), was used for all steps of cell lineage reconstruction for the intestine and *cathepsin*$^{+}$ cell lineages, following the Monocle package documentation (http://cole-trapnell-lab.github.io/monocle-release). Briefly, expression matrices for cells from all clusters of the intestine lineage or clusters 0,1,2,3,4,5,7, and 16 of the *cathepsin*$^{+}$ cell lineage were used to create CellDataSet objects using the function newCellDataSet [expressionFamily=negbinomial.size()]. After estimating size factors and dispersions using the estimateSizeFactors and estimateDispersions functions, respectively, a dispersion table was generated using the dispersionTable function. A subset of genes to be used for cell clustering was then chosen using the

following parameters [mean_expression >= 0.5 & dispersion_empirical >= 2 * dispersion_fit]. Dimensionality of the data was reduced using the function reduceDimension [max_components=2,reduction_method ="DDRTree"] and cells were ordered using the function orderCells [reverse=FALSE]. Expression of the neoblast marker *smedwi-1* was used to set the root state, using the function orderCells [root_state=x]. Finally, branch dependent genes were identified using the BEAM function [branch_point=1, cores=10,branch_labels=c (1,2)] and filtered for q-values < 1E-2 (Table 2.3).

Cell cycle state assignment

The CellCycleScoring function in Seurat was used to assign a S or G2M cell cycle score to cells of the data using established cell cycle state markers from (***89***). A subset of S and G2M markers with clear planarian homologs were used in the analysis, as follows.

S phase contigs	Contig ID	G2M phase contigs	Contig ID
dd_Smed_v4_5764_0_1	MCM5	dd_Smed_v4_14261_0_1	BIRC5
dd_Smed_v4_5688_0_1	PCNA	dd_Smed_v4_6668_0_1	CKS2
dd_Smed_v4_1260_0_1	TYMS	dd_Smed_v4_970_0_1	MKI67
dd_Smed_v4_8206_0_1	FEN1	dd_Smed_v4_5865_0_1	SMC4
dd_Smed_v4_5956_0_1	MCM4	dd_Smed_v4_15869_0_1	AURKB
dd_Smed_v4_1651_0_1	RRM1	dd_Smed_v4_13972_0_1	KIF20B
dd_Smed_v4_1568_0_1	UNG	dd_Smed_v4_17887_0_1	TTK
dd_Smed_v4_15389_0_1	GINS2	dd_Smed_v4_88923_0_1	CDC25C
dd_Smed_v4_15465_0_1	DTL	dd_Smed_v4_7553_0_1	RANGAP1
dd_Smed_v4_11558_0_1	PRIM1	dd_Smed_v4_14243_0_1	ECT2
dd_Smed_v4_4341_0_1	RPA2	dd_Smed_v4_9585_0_1	KIF23
dd_Smed_v4_5663_0_1	NASP	dd_Smed_v4_2016_0_1	LBR
dd_Smed_v4_9628_0_1	SLBP	dd_Smed_v4_12391_0_1	NEK2
dd_Smed_v4_4379_0_1	UBR7		

dd_Smed_v4_11434_0_1	MSH2		
dd_Smed_v4_8626_0_1	RAD51		
dd_Smed_v4_20567_0_1	RRM2		
dd_Smed_v4_16942_0_1	CDC45		
dd_Smed_v4_18778_0_1	CDC6		
dd_Smed_v4_14547_0_1	EXO1		
dd_Smed_v4_12580_0_1	TIPIN		
dd_Smed_v4_17862_0_1	DSCC1		
dd_Smed_v4_10119_0_1	BLM		
dd_Smed_v4_7575_0_1	CLSPN		
dd_Smed_v4_9168_0_1	POLA1		
dd_Smed_v4_5543_0_1	CHAF1B		

Gene cloning

All genes presented in this study were cloned prior to their use for FISH or RNAi, as previously described (***35***). Briefly, gene-specific primers were used to amplify specific gene sequences from planarian cDNA. Sequences were then inserted into a pGEM vector according to the manufacturer's protocol (Promega). Plasmids were transformed into DH10B competent bacteria, which were plated onto LB + 2% agar plates containing 0.1 mg/ml carbenicillin, 0.5 mM IPTG, and 0.08 mg/ml X-gal. Clones were screened for the presence of a DNA fragment of the predicted insert size by colony PCR. Positive colonies were cultured overnight and plasmid DNA was extracted using a Qiagen QIAprep Miniprep kit (Quiagen, 27106). Plasmid sequence was validated by Sanger sequencing (Genewiz, Inc.).

Fluorescence *in situ* hybridizations and immunofluorescence

Fluoresence *in situ* hybridizations were performed as previously described (***90***). Briefly, 5% N-acetylcysteine was used to remove mucus prior to fixation. Animals were also incubated in proteinase k (2 µg/ml) following bleaching. Following an overnight

hybridization at 56°C, animals were incubated in pre-hybridization buffer, 1:1 prehybridization buffer:2X-SSC, 2X-SSC, 0.2X-SSC, and PBST (twice each). Prior to overnight incubation in primary antibody, animals were incubated for 1.5 hours in block consisting of 0.5% Roche western blocking reagent (RWBR) (Roche, 11921673001) and 5% inactivated horse serum in PBST for anti-DIG antibody, 0.5% RWBR and 5% casein in PBST for anti-DNP antibody, and 1% RWBR in PBST for anti-FITC antibody. Peroxidase inactivation with 1% sodium azide was performed at room temperature for 1.5 hours. SMEDWI-1 immunofluorescence was performed following all FISH steps. Samples were incubated in block consisting of 0.5% RWBR and 5% HIHS in PBST for 1.5 hours and incubated overnight with 1:1000 rabbit anti-SMEDWI-1 primary antibody. Antibody was washed out and samples were incubated overnight with 1:300 goat antirabbit IgG-HRP secondary antibody (Invitrogen, T20924). Antibody was washed out and samples were incubated in tyramide solution (1:200 Alexa fluor 488 tyramide [Invitrogen, T20912], 1:20,000 30% H2O2, amplification buffer) for 1 hour. Fluorescent images were taken with a Zeiss LSM 700 confocal microscope and Fiji/ImageJ was used to process images. The color of inverted images was set to grey in Fiji before inversion. Unless otherwise indicated, all images shown were maximum intensity projections of representative results observed in 3 - 6 animals.

We cloned and made probes for FISH for genes with enriched expression in identified clusters. In a number of cases, markers for a cluster were also expressed in other clusters. In 71/72 probe pairs tested where FISH signal was detected for both probes, expression within the same cells was detected. In 72/72 of these cases signal was also consistent with the predicted tissue of origin of that cluster. In 24/24 cases in which only one probe yielded detectable FISH signal, signal was also consistent with the predicted tissue of origin of that cluster.

RNAi

Template PCR reactions with flanking T7 promoters were generated from plasmids and used to generate dsRNA by in vitro transcription (Promega). Transcription was

performed overnight, followed by the addition of DNAse (Promega) for 45 minutes and precipitation with 3M sodium acetate (pH 5.2) and 100% ethanol (-20°C). Following a wash with 80% ethanol (-20°C), dsRNA was resuspended in water and annealed at 95°C for five minutes, followed by a 30 minute incubation at room temperature. 6 µl of dsRNA was mixed with 13 µl of liver and used for feedings (***91***). Animals were kept on gentamycin reagent solution (1:1000, Gibco) and fed seven times over 21 days. Two days following the seventh feeding, animals were cut into 2-3 pieces. Head and trunk pieces were fixed, as above. Control animals were fed dsRNA generated from the *C. elegans* gene *unc-22*.

Single cell differential expression analysis of muscle cells

Regionally expressed muscle cells did not cluster based on their positional identity. As such, they were identified in the data by their expression of genes known to be expressed in those regions (***71***). Because of inherent variability in expression of any single gene, muscle cells were required to express at least two genes specific to a region to be identified as from that region (***71***). Posterior muscle cells were defined by their expression of at least two of the four genes *wnt11-1*, *wnt11-2*, *wntP-2*, and *fz4-1* (***49***, ***70***, ***92***, ***93***,). To identify these cells within the data, the Seurat function WhichCells [subset.name=, accept.low=0.5] was used on the muscle subcluster for each of the regionally expressed genes. Cells determined to express at least two of the genes were then selected for further analysis. An expression matrix for all muscle cells was generated after tagging the name of each muscle cell as positive or negative for the two posterior markers. This expression matrix was then used as input for the R package SCDE (***72***). SCDE analysis revealed a list of genes enriched in the posterior muscle cells. Genes were sorted by the SCDE output parameter “conservative estimate” of fold enrichment, resulting in *wnt11-1*, *fz4-1*, *wnt11-2*, and *wntP-2* (the genes used to identify cells as posterior) at ranks 1, 2, 3, and 6 of the data, respectively. These genes were highlighted in grey and were excluded from the ranking determination of additional genes (Table 2.3). Genes identified as posterior in Scimone et al. (***71***) were highlighted

in green (Table 2.3). All genes with a conservative estimate < 1 were removed from the data.

A similar analysis was performed for anterior, lateral, and medial muscle cells. In short, anterior muscle cells were defined by their expression of at least two of the five genes (***19***) *sfrp-1* (***49***), *ndl-4* (***75***), *prep* (***76***), *wnt2* (***49***), and *ndl-5* (***71***). Lateral muscle cells were defined by their expression of at least two of the three genes (***19***) *admp* (***77***, ***78***), *wnt5* (***79***), and *nlg-7* (***80***). Medial muscle cells were defined by their expression of at least two of the four genes (***19***) *slit* (***81***), *bmp4* (***82***, ***83***, ***67***), *netrin-2* (***84***), and *admp* (***77***, ***78***). Genes identified as anterior in Scimone et al. (***71***) were similarly highlighted in green in the list of enriched genes generated by SCDE.

Cell counting and volume measurement

Counting of rare cells (Figure 2.40) was performed on whole-animal FISH images manually in FIJI/ImageJ. Counting of cells for the comparison of cell type proportions in the animal and in the data (Figure 2.44) was performed using the spots functionality of the 3D image processing software Imaris. 20X images of the head were cropped below the posterior end of the cephalic ganglia before cell counting. The following parameters were used for each cell population, as determined by visual inspection.

Marker Gene	Quality	Diameter (in µm)
prog-2	5.07	12.5
dd_72	5.55	6.25
***PTPRT* (dd10872)**	4.76	9
dd_2920	4.32	6.25
sert	3.94	6.25
***COL21A1* (dd9585)**	5.22	4.5
dd_238	2.26	7

Total head volume was measured using the surface functionality of Imaris, using a diameter of 5 µm for all images. Quality scores were determined for each individual image by visual inspection. Cell counts from each animal were normalized to the total volume of the head fragment. The average normalized count for each cell type across 4- 5 animals was then used to determine the proportion of each cell type, compared to the other cell types analyzed, in the animal.

Table Captions

Table 2.1. Description of single cells sequenced in this report.

Table 2.2. Cluster- and subcluster-enriched genes.

Table 2.3. Branch-dependent genes from Monocle analysis.

Table 2.4. Genes enriched in regionally localized muscle cells.

Table 2.5. Contig annotation of all genes mentioned in this study.

*All tables can be accessed at https://hdl.handle.net/1721.1/126304.

Acknowledgements

We thank M. L. Scimone, C. McQuestion, and K. D. Atabay for their help in the targeted dissociation of tissue from the planarian brain; all Reddien Lab members for valuable comments and discussion; and E. Z. Macosko, M. Goldman, and S. A. McCarroll for making protocols available. We acknowledge NIH (R01GM080639) support. P.W.R. is an Investigator of the Howard Hughes Medical Institute and an associate member of the Broad Institute of Harvard and MIT. We thank the Eleanor Schwartz Charitable Foundation for support.

References

1. Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-2018 (1998).
2. E. S. Lander *et al.*, Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
3. J. C. Venter *et al.*, The sequence of the human genome. *Science* **291**, 1304-1351 (2001).
4. D. A. Jaitin *et al.*, Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776-779 (2014).
5. A. K. Shalek *et al.*, Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363-369 (2014).
6. E. Z. Macosko *et al.*, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
7. J. Cao *et al.*, Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661-667 (2017).
8. X. Han *et al.*, Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107.e1017 (2018).
9. P. W. Reddien, A. Sanchez Alvarado, Fundamentals of planarian regeneration. *Annu Rev Cell Dev Biol* **20**, 725-757 (2004).
10. L. H. Hyman, *The invertebrates. Vol. II, Vol. II*. (McGraw-Hill book company inc., New York; London, 1951).
11. J. Baguna, R. Romero, Quantitative analysis of cell types during growth, degrowth and regeneration in the planarians Dugesia mediterranea and Dugesia tigrina. *Hydrobiologia Hydrobiologia : The International Journal of Aquatic Sciences* **84**, 181-194 (1981).
12. I. E. Wang, S. W. Lapan, M. L. Scimone, T. R. Clandinin, P. W. Reddien, Hedgehog signaling regulates gene expression in planarian glia. *Elife* **5**, (2016).
13. R. H. Roberts-Galbraith, J. L. Brubacher, P. A. Newmark, A functional genomics screen in planarians reveals regulators of whole-brain regeneration. *Elife* **5**, (2016).
14. D. E. Wagner, I. E. Wang, P. W. Reddien, Clonogenic neoblasts are pluripotent adult stem cells that underlie planarian regeneration. *Science* **332**, 811-816 (2011).

15. M. L. Scimone, K. M. Kravarik, S. W. Lapan, P. W. Reddien, Neoblast specialization in regeneration of the planarian Schmidtea mediterranea. *Stem Cell Reports* **3**, 339-352 (2014).
16. P. W. Reddien, Specialized progenitors and regeneration. *Development* **140**, 951-957 (2013).
17. P. A. Newmark, A. Sanchez Alvarado, Bromodeoxyuridine specifically labels the regenerative stem cells of planarians. *Dev Biol* **220**, 142-153 (2000).
18. P. W. Reddien, Constitutive gene expression and the specification of tissue identity in adult planarian biology. *Trends Genet* **27**, 277-285 (2011).
19. J. N. Witchley, M. Mayer, D. E. Wagner, J. H. Owen, P. W. Reddien, Muscle cells provide instructions for planarian regeneration. *Cell Rep* **4**, 633-641 (2013).
20. S. A. LoCascio, S. W. Lapan, P. W. Reddien, Eye Absence Does Not Regulate Planarian Stem Cells during Eye Regeneration. *Dev Cell* **40**, 381-391.e383 (2017).
21. S. Y. Liu *et al.*, Reactivating head regrowth in a regeneration-deficient planarian species. *Nature* **500**, 81-84 (2013).
22. R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, A. Regev, Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495-502 (2015).
23. A. McDavid *et al.*, Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics* **29**, 461-467 (2013).
24. O. Wurtzel *et al.*, A Generic and Cell-Type-Specific Wound Response Precedes Regeneration in Planarians. *Dev Cell* **35**, 632-645 (2015).
25. K. Nishimura *et al.*, Identification and distribution of tryptophan hydroxylase (TPH)-positive neurons in the planarian Dugesia japonica. *Neurosci Res* **59**, 101-106 (2007).
26. A. Sanchez Alvarado, P. A. Newmark, Double-stranded RNA specifically disrupts gene expression during planarian regeneration. *Proc Natl Acad Sci U S A* **96**, 5049-5054 (1999).
27. K. Nishimura *et al.*, Reconstruction of dopaminergic neural network and locomotion function in planarian regenerates. *Dev Neurobiol* **67**, 1059-1078 (2007).
28. J. J. Collins, 3rd *et al.*, Genome-wide analyses reveal a role for peptide hormones in planarian germline development. *PLoS Biol* **8**, e1000509 (2010).
29. P. W. Reddien, N. J. Oviedo, J. R. Jennings, J. C. Jenkin, A. Sanchez Alvarado, SMEDWI-2 is a PIWI-like protein that regulates planarian stem cells. *Science* **310**, 1327-1330 (2005).

30. D. E. Wagner, J. J. Ho, P. W. Reddien, Genetic regulators of a pluripotent adult stem cell system in planarians identified by RNAi and clonal analysis. *Cell Stem Cell* **10**, 299-311 (2012).
31. T. Guo, A. H. Peters, P. A. Newmark, A Bruno-like gene is required for stem cell maintenance in planarians. *Dev Cell* **11**, 159-169 (2006).
32. J. C. van Wolfswinkel, D. E. Wagner, P. W. Reddien, Single-cell analysis reveals functionally distinct classes within the planarian stem cell compartment. *Cell Stem Cell* **15**, 326-339 (2014).
33. M. L. Scimone, M. Srivastava, G. W. Bell, P. W. Reddien, A regulatory program for excretory system regeneration in planarians. *Development* **138**, 4387-4398 (2011).
34. X. He *et al.*, FOX and ETS family transcription factors regulate the pigment cell lineage in planarians. *Development* **144**, 4540-4551 (2017).
35. O. Wurtzel, I. M. Oderberg, P. W. Reddien, Planarian Epidermal Stem Cells Respond to Positional Cues to Promote Cell-Type Diversity. *Dev Cell* **40**, 491-504.e495 (2017).
36. G. T. Eisenhoffer, H. Kang, A. Sanchez Alvarado, Molecular analysis of stem cells and their descendants during cell turnover and regeneration in the planarian Schmidtea mediterranea. *Cell Stem Cell* **3**, 327-339 (2008).
37. K. C. Tu *et al.*, Egr-5 is a post-mitotic regulator of planarian epidermal differentiation. *Elife* **4**, e10501 (2015).
38. S. J. Zhu, S. E. Hallows, K. W. Currie, C. Xu, B. J. Pearson, A mex3 homolog is required for differentiation during planarian stem cell lineage development. *Elife* **4**, (2015).
39. H. Thi-Kim Vu *et al.*, Stem cells and fluid flow drive cyst formation in an invertebrate excretory organ. *Elife* **4**, (2015).
40. J. C. Rink, H. T. Vu, A. Sanchez Alvarado, The maintenance and regeneration of the planarian excretory system are regulated by EGFR signaling. *Development* **138**, 3769-3780 (2011).
41. B. H. Willier, L. H. Hyman, S. A. Rifenburgh, A histochemical study of intracellular digestion in triclad flatworms. *J. Morphol. Journal of Morphology* **40**, 299-340 (1925).
42. S. Ishii, Electron microscopic observations on the Planarian tissues II. The intestine. *Fukushima journal of medical science* **12**, 67-87 (1965).
43. D. J. Forsthoefel *et al.*, An RNAi screen reveals intestinal regulators of branching morphogenesis, differentiation, and stem cell proliferation in planarians. *Dev Cell* **23**, 691-704 (2012).

44. R. M. Zayas, F. Cebria, T. Guo, J. Feng, P. A. Newmark, The use of lectins as markers for differentiated secretory cells in planarians. *Dev Dyn* **239**, 2888-2897 (2010).
45. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381-386 (2014).
46. S. W. Lapan, P. W. Reddien, Transcriptome analysis of the planarian eye identifies ovo as a specific regulator of eye regeneration. *Cell Rep* **2**, 294-307 (2012).
47. S. W. Lapan, P. W. Reddien, dlx and sp6-9 Control optic cup regeneration in a prototypic eye. *PLoS Genet* **7**, e1002226 (2011).
48. F. Cebria *et al.*, Dissecting planarian central nervous system regeneration by the expression of neural-specific genes. *Dev Growth Differ* **44**, 135-146 (2002).
49. C. P. Petersen, P. W. Reddien, Smed-betacatenin-1 is required for anteroposterior blastema polarity in planarian regeneration. *Science* **319**, 327-330 (2008).
50. F. Cebria, Planarian Body-Wall Muscle: Regeneration and Function beyond a Simple Skeletal Support. *Front Cell Dev Biol* **4**, 8 (2016).
51. K. J. Pedersen, Some features of the fine structure and histochemistry of planarian subepidermal gland cells. *Zeitschrift fr Zellforschung Zeitschrift fr Zellforschung und Mikroskopische Anatomie* **50**, 121-142 (1959).
52. K. J. r. Pedersen, Studies on the nature of planarian connective tissue. *Zeitschrift f¸r Zellforschung Zeitschrift f¸r Zellforschung und Mikroskopische Anatomie* **53**, 569-608 (1961).
53. P. A. Newmark, P. W. Reddien, F. Cebria, A. Sanchez Alvarado, Ingestion of bacterially expressed double-stranded RNA inhibits gene expression in planarians. *Proc Natl Acad Sci U S A* **100 Suppl 1**, 11861-11865 (2003).
54. B. M. Stubenhaus *et al.*, Light-induced depigmentation in planarians models the pathophysiology of acute porphyrias. *Elife* **5**, (2016).
55. K. Nishimura *et al.*, Characterization of tyramine beta-hydroxylase in planarian Dugesia japonica: cloning and expression. *Neurochem Int* **53**, 184-192 (2008).
56. K. Nishimura *et al.*, Identification of glutamic acid decarboxylase gene and distribution of GABAergic nervous system in the planarian Dugesia japonica. *Neuroscience* **153**, 1103-1114 (2008).
57. J. Azimzadeh, M. L. Wong, D. M. Downhour, A. Sanchez Alvarado, W. F. Marshall, Centrosome loss in the evolution of planarians. *Science* **335**, 461-463 (2012).

58. D. Wenemoser, P. W. Reddien, Planarian regeneration involves distinct stem cell responses to wounds and tissue absence. *Dev Biol* **344**, 979-991 (2010).
59. K. Nishimura, Y. Kitamura, T. Taniguchi, K. Agata, Analysis of motor function modulated by cholinergic neurons in planarian Dugesia japonica. *Neuroscience* **168**, 18-30 (2010).
60. M. L. Scimone, S. W. Lapan, P. W. Reddien, A forkhead transcription factor is wound-induced at the planarian midline and required for anterior pole regeneration. *PLoS Genet* **10**, e1003999 (2014).
61. M. C. Vogg *et al.*, Stem cell-dependent formation of a functional anterior regeneration pole in planarians requires Zic and Forkhead transcription factors. *Dev Biol* **390**, 136-148 (2014).
62. C. P. Petersen, P. W. Reddien, Polarized notum activation at wounds inhibits Wnt function to promote planarian head regeneration. *Science* **332**, 852-855 (2011).
63. E. M. Hill, C. P. Petersen, Wnt/Notum spatial feedback inhibition controls neoblast differentiation to regulate reversible growth of the planarian brain. *Development* **142**, 4217-4229 (2015).
64. Y. Wang, R. M. Zayas, T. Guo, P. A. Newmark, nanos function is essential for development and regeneration of planarian germ cells. *Proc Natl Acad Sci U S A* **104**, 5901-5906 (2007).
65. N. J. Oviedo, P. A. Newmark, A. Sanchez Alvarado, Allometric scaling and proportion regulation in the freshwater planarian Schmidtea mediterranea. *Dev Dyn* **226**, 326-333 (2003).
66. C. E. Adler, C. W. Seidel, S. A. McKinney, A. Sanchez Alvarado, Selective amputation of the pharynx identifies a FoxA-dependent regeneration program in planaria. *Elife* **3**, e02238 (2014).
67. M. D. Molina, E. Salo, F. Cebria, The BMP pathway is essential for re-specification and maintenance of the dorsoventral axis in regenerating and intact planarians. *Dev Biol* **311**, 79-94 (2007).
68. C. Vasquez-Doorman, C. P. Petersen, zic-1 Expression in Planarian neoblasts after injury controls anterior pole regeneration. *PLoS Genet* **10**, e1004452 (2014).
69. M. Marz, F. Seebeck, K. Bartscherer, A Pitx transcription factor controls the establishment and maintenance of the serotonergic lineage in planarians. *Development* **140**, 4499-4509 (2013).

70. K. A. Gurley, J. C. Rink, A. Sanchez Alvarado, Beta-catenin defines head versus tail identity during planarian regeneration and homeostasis. *Science* **319**, 323-327 (2008).
71. M. L. Scimone, L. E. Cote, T. Rogers, P. W. Reddien, Two FGFRL-Wnt circuits organize the planarian anteroposterior axis. *Elife* **5**, (2016).
72. P. V. Kharchenko, L. Silberstein, D. T. Scadden, Bayesian approach to single-cell differential expression analysis. *Nat Methods* **11**, 740-742 (2014).
73. S. Barberan, J. M. Martin-Duran, F. Cebria, Evolution of the EGFR pathway in Metazoa and its diversification in the planarian Schmidtea mediterranea. *Sci Rep* **6**, 28071 (2016).
74. M. L. Scimone, L. E. Cote, P. W. Reddien, Orthogonal muscle fibres have different instructive roles in planarian regeneration. *Nature* **551**, 623-628 (2017).
75. J. C. Rink, K. A. Gurley, S. A. Elliott, A. Sanchez Alvarado, Planarian Hh signaling regulates regeneration polarity and links Hh pathway evolution to cilia. *Science* **326**, 1406-1410 (2009).
76. D. A. Felix, A. A. Aboobaker, The TALE class homeobox gene Smed-prep defines the anterior compartment for head regeneration. *PLoS Genet* **6**, e1000915 (2010).
77. M. D. Molina *et al.*, Noggin and noggin-like genes control dorsoventral axis regeneration in planarians. *Curr Biol* **21**, 300-305 (2011).
78. M. A. Gavino, P. W. Reddien, A Bmp/Admp regulatory circuit controls maintenance and regeneration of dorsal-ventral polarity in planarians. *Curr Biol* **21**, 294-299 (2011).
79. T. Adell, E. Salo, M. Boutros, K. Bartscherer, Smed-Evi/Wntless is required for beta-catenin-dependent and -independent processes during planarian regeneration. *Development* **136**, 905-910 (2009).
80. M. D. Molina, E. Salo, F. Cebria, Expression pattern of the expanded noggin gene family in the planarian Schmidtea mediterranea. *Gene Expr Patterns* **9**, 246-253 (2009).
81. F. Cebria, P. A. Newmark, Morphogenesis defects are associated with abnormal nervous system regeneration following roboA RNAi in planarians. *Development* **134**, 833-837 (2007).
82. H. Orii, K. Kato, K. Agata, K. Watanabe, Molecular Cloning of Bone Morphogenetic Protein (BMP) Gene from the Planarian *Dugesia japonica*. *Zoological Science* **15**, 871-877, 877 (1998).
83. P. W. Reddien, A. L. Bermange, A. M. Kicza, A. Sanchez Alvarado, BMP signaling regulates the dorsal planarian midline and is needed for asymmetric regeneration. *Development* **134**, 4043-4051 (2007).

84. F. Cebria, P. A. Newmark, Planarian homologs of netrin and netrin receptor are required for proper regeneration of the central nervous system and the maintenance of nervous system architecture. *Development* **132**, 3691-3703 (2005).
85. C. E. Laumer *et al.*, Spiralian phylogeny informs the evolution of microscopic lineages. *Curr Biol* **25**, 2000-2006 (2015).
86. C. Camacho *et al.*, BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
87. L. Rouhana, A. P. Vieira, R. H. Roberts-Galbraith, P. A. Newmark, PRMT5 and the role of symmetrical dimethylarginine in chromatoid bodies of planarian stem cells. *Development* **139**, 1083-1094 (2012).
88. T. Hayashi, M. Asami, S. Higuchi, N. Shibata, K. Agata, Isolation of planarian X-ray-sensitive stem cells by fluorescence-activated cell sorting. *Dev Growth Differ* **48**, 371-380 (2006).
89. I. Tirosh *et al.*, Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196 (2016).
90. R. S. King, P. A. Newmark, In situ hybridization protocol for enhanced detection of gene expression in the planarian Schmidtea mediterranea. *BMC Dev Biol* **13**, 8 (2013).
91. L. Rouhana *et al.*, RNA interference by feeding in vitro-synthesized double-stranded RNA to planarians: methodology and dynamics. *Dev Dyn* **242**, 718-730 (2013).
92. K. A. Gurley *et al.*, Expression of secreted Wnt pathway components reveals unexpected complexity of the planarian amputation response. *Dev Biol* **347**, 24-39 (2010).
93. C. P. Petersen, P. W. Reddien, A wound-induced Wnt expression program controls planarian regeneration polarity. *Proc Natl Acad Sci U S A* **106**, 17061-17066 (2009).

Chapter 5

Discussion

I. Successful strategy for characterizing the transcriptomes of all cell types of an animal

Even with the advent of affordable, high throughput single-cell RNA sequencing approaches, determining the transcriptomes for all cell types in a whole animal has remained a daunting prospect. Many animals contain trillions of cells and hundreds of cell types, many of which are quite rare, and cell type composition can change dramatically over the course of development. Our success in generating such a whole-animal cell-type transcriptome atlas for the planarian flatworm was grounded in two factors. First, planarians contain a population of pluripotent stem cells that they use to constantly turnover all cell types throughout the life of the animal, and they constitutively express patterning information to guide the placement of these cells. As such, at a single time point in the adult, we were able to isolate all stem cell populations, including pluripotent stem cells; all differentiated cells; the transition state populations associated with each cell lineage; and patterning information. For most organisms, obtaining all of this information would require sampling across many stages of development. Second, our strategy for capturing rare, regionally localized cells types was crucial to obtaining complete cell type saturation. Rather than dissociate whole animals, iterative Drop-seq runs were performed on body fragments to differing depths, allowing saturation of cell-type coverage in regions with rare cell types. Were we to dissociate whole animals, prevalent cell types present throughout the body would largely crowd out rare, regionally localized cell types, requiring many more cells to be sequenced for saturation. Following each Drop-seq run, the data was processed and used to assess the presence of known rare cell types, allowing active determination of which fragments needed additional sequencing. Sequencing runs were performed for each fragment until full cell type saturation had been reached.

Single-cell sequencing approaches can be applied to most organisms, raising the prospect of generating whole-animal atlases for diverse animal species. Our strategy for achieving cell type saturation can be employed towards this goal, even in animals with little previous molecular characterization. By identifying genes that were enriched for

expression in each population of cells in the planarian data, we were able to design FISH probes to determine the anatomical localization pattern of most cell populations, including for previously uncharacterized cell types. As such, for animals with little molecular characterization, FISH can be performed following each round of single-cell sequencing using markers for each population to assay localization and rarity *in vivo*. Some cells that are rare in the sequencing data might be rare in the animal or difficult to capture at their endogenous frequency following dissociation. FISH experiments allow assessment of cell-type-frequency in vivo and aid in assessment of saturation for rare cell types. Furthermore, combinatorial indexing-based single-cell RNA sequencing methods have enabled very high throughput single-cell RNA sequencing in the absence of any specialized equipment, lowering the costs and increasing the reach of such technologies to enable single cell transcriptional profiling of diverse, understudied animals (***1***, ***2***). Indeed, as described in chapter 1, whole animal cell type atlases with varying degrees of cell type saturation have now been generated for a wide range of other adult animals using a variety of single-cell RNA sequencing approaches (***3***, ***4***, ***5***, ***6***). The ability to generate transcriptomes for all cell types in a complete animal has meaningful implications for evolutionary biology. With many cell types, tissues, and the transcriptional machinery used to generate those cells and tissues present across diverse animal species, cross-species comparisons of single-cell RNA sequencing data could provide insights into cell-type evolution and the evolution of gene regulatory networks that control tissue development.

II. Identification of novel specialized neoblasts and transition states

The regenerative capacity of planarians is largely derived from neoblasts, at least some of which are pluripotent (***7***). Neoblasts are heterogeneous as a population (***8***). Multiple subpopulations, called specialized neoblasts, have been identified that express transcription factors essential for the specification of various differentiated cell lineages (***8***). Thousands of neoblasts were identified in our single-cell sequencing data by their expression of *smedwi-1*. To facilitate the identification of specialized neoblasts, cells

with clear *smedwi-1* expression were subclustered in isolation. Many of the resultant subclusters were enriched for known tissue-specific markers, generally exhibited lower expression of *smedwi-1*, and were assigned a predicted G1/G0 cell cycle state, indicating these are likely post-mitotic progenitors that haven't yet fully turned off *smedwi-1* expression. However, a number of subclusters exhibited high *smedwi-1* expression and were assigned a predicted S or G2/M cell cycle state. Two of these subclusters included previously identified specialized neoblast classes: zeta-neoblast epidermal progenitors and gamma-neoblast intestinal progenitors. A group of specialized neoblasts for the protonephridia were also identified. Interestingly, a subcluster marked by the previously undescribed transcript dd_10988 was enriched for a number of genes also expressed in neurons, suggesting this subcluster could represent a transcriptionally unified subpopulation of neuronal specialized neoblasts. Another subcluster marked by the gene *PLOD1* (dd3457) was similarly enriched for a number of genes also expressed in muscle, suggesting this subcluster could represent a transcriptionally unified subpopulation of muscle specialized neoblasts. More work will be necessary to assess the biological role of these novel neoblast subpopulations, however.

While it has been clearly demonstrated that individual neoblasts exhibit pluripotent potential, it is unclear whether all neoblasts have this potential or only a subset. Is there a single transcriptionally distinct pluripotent neoblast class? Do specialized neoblasts retain pluripotent potential? Although a number of neoblast subclusters with high *smedwi-1* expression were identified that were not associated with any differentiated lineages, these subclusters generally lacked specifically enriched genes. Furthermore, when these subclusters with high *smedwi-1* expression, excluding zeta- and gamma-neoblast clusters, were subclustered again in isolation, clusters of remnant zeta-neoblasts, protonephridia progenitors, dd_10988^{+} putative neuronal progenitors, and *PLOD1* (3457)$^{+}$ putative muscle progenitors were identified, but the remaining neoblasts subclusters largely lacked specifically enriched markers. Interestingly, these transcriptionally non-distinct clusters were also the highest in *smedwi-1* expression,

suggesting more naïve neoblasts may be defined by the absence of tissue specific markers, rather than the unique expression of specific genes. If there is no transcriptionally distinct pluripotent neoblast class, do all neoblasts, specialized neoblasts included, retain pluripotent potential? Some evidence does exist to support this idea. An antibody was recently developed to the protein product of the gene *tgs-1*, which is specifically enriched in the dd_10988^{+} subcluster of neoblasts (***9***). Cells from this neoblast subclass were isolated by FACS and were shown to individually possess pluripotent potential (***9***). Because the dd_10988^{+} cluster likely represents neuronal specialized neoblasts, this result would suggest that specialized neoblasts do retain their pluripotent potential, though more work will need to be done to test this hypothesis.

Two planarian tissues, the epidermis and the pharynx, have been shown to exhibit transcriptionally distinct post-mitotic states as they differentiate (***10***, ***11***, ***12***, ***13***). The differentiation process is much less well characterized for other planarian tissues, however. *smedwi-1*$^{+}$ cells were present locally within each of the broad tissue class clusters, suggesting transition state populations for these tissues may be present. Indeed, following subclustering of intestine and *cathepsin*$^{+}$ cells in isolation, transcriptionally distinct subclusters were found to separate local areas with *smedwi-1* expression from areas enriched for differentiated marker expression. Trajectory reconstruction of the intestine and *cathepsin*$^{+}$ cell lineages using Monocle2 identified a number of genes with variable expression across the predicted trajectories. There is even some molecular evidence validating these predicted trajectories *in vivo*. *MAP3K5* (dd4849), which is predicted to be expressed early in the differentiation trajectory for one of the *cathepsin*$^{+}$ cell lineages, was commonly expressed in cells positive for SMEDWI-1 protein and negative for *smedwi-1* mRNA, a combination that marks recent post-mitotic progenitors (***14***, ***15***, ***16***). *smedwi-1*/SMEDWI-1 assays are only capable of marking very recent post-mitotic progenitors, however, and are thus incapable of validating much of the predicted trajectories. Other approaches include temporally tracking gene expression loss following lethal irradiation, which has been utilized to identify genes expressed temporally along differentiation trajectories for the epidermis

and the pharynx (***10***, ***13***). Irradiated animals begin dying within two weeks following irradiation, however, so this approach is only appropriate for lineages with rapid turnover. A related approach would be to mark planarian neoblasts and their progeny with BrdU (***10***, ***17***) and temporally track the incidence of BrdU co-expression with putative transition state markers.

Trajectory reconstruction of the intestine and *cathepsin*$^+$ cell lineages predicted a number of genes that vary temporally across these differentiation trajectories. With *smedwi-1*$^+$ cells clustering locally in each of the differentiated tissue clusters, trajectory reconstruction tools could be applied to characterize differentiation in other tissue lineages as well. Furthermore, multiple trajectory reconstruction algorithms have been developed, including the most recent iteration of Monocle (Monocle3), that enable complex, multi-branching trajectory reconstructions that have been used to analyze early vertebrate embryogenesis (***18***, ***19***, ***20***, ***21***). Application of such approaches could provide insights into a number of questions surrounding planarian cell type differentiation. Whereas it is known that neoblasts are the source of all new tissue in the animal, and specialized neoblasts have been identified for many known differentiated cell types, it is currently unclear whether hierarchies exist for cell-fate decisions within lineages. For example, does a neoblast first commit to making a neuron before committing to make a serotonergic neuron specifically? Furthermore, the molecular relatedness of distinct planarian tissues is currently unknown. Clustering of approximately 600 cells isolated by SMART-seq clustered intestine together with *cathepsin*$^+$ cells, with a number of genes, including the transcription factor *hnf4*, enriched in both intestinal and *cathepsin*$^+$ cells (***22***). What developmental relation do these distinct tissues share, if any? Although trajectory reconstruction could provide initial hints to questions such as these, any results would require further validation.

Finally, transcription factors were identified with enriched expression in putative transition state populations for most differentiated cell types in the data. It has been previously shown that inhibition by RNAi of transcription factors necessary for the

specification of a cell type can be used to ablate that cell through tissue turnover (***23***, ***24***, ***25***, ***26***, ***27***, ***28***, ***29***, ***30***). In my work, RNAi of two such transcription factor-encoding genes identified for the newly described outer intestinal cell lineage and a cell population from the parenchymal cell lineage led to the loss of differentiated cell markers for these cells in the animal. While suggestive of cell ablation, loss of a few cell markers could also indicate disruption of a limited transcriptional network. Further EM evidence or FISH using a panel of enriched markers for each cell is needed to confirm the cells were truly ablated. Because transcription factors were associated with most differentiated cell populations in the data, many of which were previously undescribed, a wide range of cell types in the animal could be ablated using this approach, allowing their biological function in the animal to be determined. This is especially important for planarians as a model system, in that transgenic approaches for targeted cell-type ablation are currently unavailable.

III. Identification of novel differentiated cell types

Planarians possess a complex anatomy made up of a number of functionally distinct tissues containing a diverse array of differentiated cell types. A number of novel differentiated cell populations were identified in the Drop-seq data, spanning across almost all planarian tissues. Even for some molecularly well-characterized tissues, new constituent cell populations were identified. For example, a previously undescribed “outer” layer of cells was identified for the intestine. Many tissue classes in the planarian have had limited characterization at the molecular level, and a significant number of novel cell populations were identified for these tissues. A novel major tissue class termed *cathepsin*$^{+}$ cells was also discovered, encompassing eight transcriptionally distinct cell populations.

The *cathepsin*$^{+}$ cell class consists of transcriptionally distinct cell populations, including previously described pigment and glial populations. In general, *cathepsin*$^{+}$ clusters exhibited enriched expression of genes encoding digestive enzymes and endocytic

machinery components, suggesting these cells may collectively function in phagocytosis. One *cathepsin*$^+$ cell population, marked by the transcript dd_9, exhibited an interesting morphology by FISH, with long processes spread throughout the parenchyma of the animal. This morphology and anatomical localization was highly reminiscent of fixed parenchymal cells, as described in chapter 1, which had been demonstrated by EM to phagocytose bacteria (***31***, ***32***, ***33***). Indeed, each *cathepsin*$^+$ cell population was recently shown to be not only capable of phagocytosing fluorescently labeled heat-killed bacteria, but were the primary cells that did so (***34***).The transcription factors *ets-1* and *foxF-1* are enriched in all *cathepsin*$^+$ cell populations and are necessary for their specification, with RNAi of *foxF-1* leading to a loss of all *cathepsin*$^+$ cells and RNAi of *ets-1* leading to a loss of pigment cells (the effect on other *cathepsin*$^+$ cells was not assessed) (***34***, ***35***). Both *ets-1* and *foxF-1* animals lyse not long after depigmentation, suggesting an essential biological role for *cathepsin*$^+$ cells in the animal (***34***, ***35***). The cause of this lethality is currently unclear, however, as is what biological role these cells plays in the animal. Are they responsible for phagocytosing cellular debris and/or apoptotic cells? Given the lack of a circulatory system in planarians, could they serve such a role through their long, parenchyma-filling processes? Given their ability to phagocytose bacteria, could they serve as an innate immune system? More work is needed to begin to understand the biological role of this fascinating new class of cells.

As described in chapter 1, planarians contain a number of gland cell populations identified largely through EM and histological studies (***31***, ***36***, ***37***), with a number of gene markers identified for a marginal adhesive gland population (***38***). Single-cell RNA sequencing of approximately 600 wounded planarians by SMART-seq identified a heterogenous cluster of cells, termed “parapharyngeal”, that were enriched for a number of parenchymal-localized genes, as well as the markers for the marginal adhesive gland population (***22***), suggesting this cluster may represent multiple distinct gland cell populations. From the Drop-seq data, the parenchymal subcluster contained a number of distinct cell populations, many markers for which were also enriched in the

previously described “parapharyngeal” cluster. FISH for these marker genes revealed multiple cell populations with long processes that terminated at the epidermis or the pharynx, suggesting a potential glandular role. Though more work will need to be done to conclusively confirm the identity and function of these putative gland cells, at least one transcription factor was found to be enriched in each parenchymal cell population, potentially allowing the targeted ablation of each cell population in the animal.

The nervous system is by far the most cellularly complex of the planarian tissues. Subclustering of cells marked as neuronal identified a large number of distinct clusters, many of which had not been previously described. Some clusters lacked specifically enriched marker genes and were instead defined by a combination of markers shared between multiple cell populations, further highlighting neuronal complexity. Future work will be needed to determine the function of these novel cell populations. The clusters that were identified were largely peripheral neurons, as determined by FISH localization patterns. Known cell types in the brain, such as octopaminergic neurons (***39***), were largely present in low numbers and were found within just four very heterogenous clusters. This suggests that even more single-cell sequencing could be beneficial for identification of neurons that constitute the brain. Although targeted sequencing of the region surrounding the brain did help in enriching for known brain-localized neurons, the enrichment was limited, suggesting even more targeted isolation of cells from the brain is necessary. Irradiation-based depletion of neoblasts and early epidermal progenitors (***10***, ***11***), which were shown to be massively overrepresented in the data compared to their relative abundance in the animal, would increase the capture rate of neurons and other differentiated cells. Furthermore, two recently developed methods allow for specific isolation of transcriptionally distinct cells. One approach uses single-cell RNA sequencing to identify FACS gates that specifically isolate transcriptionally distinct cell populations (***40***). Another approach, termed Probe-seq, uses fluorescently labeled *in situ* probes to isolate certain cell populations by FACS (***41***).

Finally, there are two distinct strains of *Schmidtea mediterranea*: an asexual strain, for which we generated this cell type transcriptome atlas, and a sexual strain of cross-fertilizing hermaphrodites. An initial sequencing run of sexual planarians in the work of this thesis yielded only a fraction of the sexual anatomy, with transcriptomes for only a couple of testis and yolk cells recovered. More targeted isolation of the sexual anatomy could be performed to further enrich for these cells in the data. Additionally, only homeostatic animals were profiled for the cell type transcriptome atlas. It would be interesting to profile animals at various stages of regeneration to determine cell type-specific changes in gene expression and changes in the relative abundance of differentiated cell types throughout the regenerative process.

IV. Identification of novel regionally expressed genes

Planarians constitutively express patterning molecules regionally across the different body axes of the animal that are essential for proper regeneration and maintenance of the body plan (***42***). These patterning molecules are largely expressed in muscle (***43***). Regionally expressed genes along the anterior-posterior (AP) axis have been profiled through single-cell RNA sequencing of muscle cells isolated from specific regions along the AP axis (***44***). Although a number of genes have been identified with regional expression in muscle along the dorsal-ventral (DV) and medial-lateral (ML) axes, the full complement of genes regionally expressed across these axes has not been as well profiled. Although muscle cells in the Drop-seq data did not cluster by position, but rather by broad muscle class, we reasoned that expression of known regionally expressed genes could be used to assign a positional identity to muscle cells in the data, enabling other genes with similar regional expression to be identified. This approach proved quite effective. When transcriptomes from muscle cells assigned an anterior or posterior identity were compared to transcriptomes from all other muscle cells, most previously identified genes with regional expression along the AP axis were recovered. This approach was further used to identify regionally expressed genes along the ML axis, identifying a number of novel genes with lateral domains of expression, as

well as one gene, dd_23400, that was expressed medially in a line of cells straight down the center of the animal. Functional patterning roles for these genes can now be assessed using RNAi. Interestingly, it has recently been shown that inhibition of genes encoding components of the STRIPAK complex results in expansion of *wnt1* expression up the midline in dd_23400^{+} muscle cells, leading to a dramatic expansion in tail length, and suggesting this dd_23400^{+} medial muscle cell population may have a distinct patterning role in the animal (***45***). Finally, the approach undertaken here is conceptually similar to the more sophisticated spatial transcriptomic approaches reviewed in Chapter 1. Given the numerous FISH images available for regionally expressed muscle genes in planarians, approaches such as novoSparc (***46***) could be used with the muscle cells from the Drop-seq data to infer gene expression localization patterns for all planarian genes, potentially identifying others with regional expression across the various body axes, including the DV axis, which was not profiled here.

V. Conclusion

As described in this thesis, high throughput single-cell RNA sequencing was used to transcriptionally profile the regenerative planarian *Schmidtea mediterranea*. Through iterative sequencing of body fragments, transcriptomes for most-to-all cell types of the complete animal were determined. Using this approach, a number of novel neoblast subclasses, transition state populations, and differentiated cell types were identified, as were a number of novel genes with regional expression in muscle, which provides patterning information for the animal. We now have full transcriptomes for each of these cell populations, enabling their biological function in the animal to be assessed. Furthermore, our approach for reaching full cell type saturation can be applied broadly to diverse animal species, including emerging model organisms with little previous molecular characterization.

References

1. J. Cao *et al.*, Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661-667 (2017).
2. A. B. Rosenberg *et al.*, Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176-182 (2018).
3. A. Sebe-Pedros *et al.*, Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq. *Cell* **173**, 1520-1534.e1520 (2018).
4. A. Sebe-Pedros *et al.*, Early metazoan cell type diversity and the evolution of multicellular gene regulation. *Nat Ecol Evol* **2**, 1176-1188 (2018).
5. K. Achim *et al.*, Whole-Body Single-Cell Sequencing Reveals Transcriptional Domains in the Annelid Larval Body. *Mol Biol Evol* **35**, 1047-1062 (2018).
6. S. Siebert *et al.*, Stem cell differentiation trajectories in Hydra resolved at single-cell resolution. *Science* **365**, (2019).
7. D. E. Wagner, I. E. Wang, P. W. Reddien, Clonogenic neoblasts are pluripotent adult stem cells that underlie planarian regeneration. *Science* **332**, 811-816 (2011).
8. P. W. Reddien, Specialized progenitors and regeneration. *Development* **140**, 951-957 (2013).
9. A. Zeng *et al.*, Prospectively Isolated Tetraspanin(+) Neoblasts Are Adult Pluripotent Stem Cells Underlying Planaria Regeneration. *Cell* **173**, 1593-1608.e1520 (2018).
10. G. T. Eisenhoffer, H. Kang, A. Sanchez Alvarado, Molecular analysis of stem cells and their descendants during cell turnover and regeneration in the planarian Schmidtea mediterranea. *Cell Stem Cell* **3**, 327-339 (2008).
11. K. C. Tu *et al.*, Egr-5 is a post-mitotic regulator of planarian epidermal differentiation. *Elife* **4**, e10501 (2015).
12. O. Wurtzel, I. M. Oderberg, P. W. Reddien, Planarian Epidermal Stem Cells Respond to Positional Cues to Promote Cell-Type Diversity. *Dev Cell* **40**, 491-504.e495 (2017).
13. S. J. Zhu, S. E. Hallows, K. W. Currie, C. Xu, B. J. Pearson, A mex3 homolog is required for differentiation during planarian stem cell lineage development. *Elife* **4**, (2015).
14. T. Guo, A. H. Peters, P. A. Newmark, A Bruno-like gene is required for stem cell maintenance in planarians. *Dev Cell* **11**, 159-169 (2006).
15. D. Wenemoser, P. W. Reddien, Planarian regeneration involves distinct stem cell responses to wounds and tissue absence. *Dev Biol* **344**, 979-991 (2010).

16. M. L. Scimone, J. Meisel, P. W. Reddien, The Mi-2-like Smed-CHD4 gene is required for stem cell differentiation in the planarian Schmidtea mediterranea. *Development* **137**, 1231-1241 (2010).
17. P. A. Newmark, A. Sanchez Alvarado, Bromodeoxyuridine specifically labels the regenerative stem cells of planarians. *Dev Biol* **220**, 142-153 (2000).
18. J. A. Farrell *et al.*, Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, (2018).
19. D. E. Wagner *et al.*, Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981-987 (2018).
20. J. A. Briggs *et al.*, The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, (2018).
21. J. Cao *et al.*, The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496-502 (2019).
22. O. Wurtzel *et al.*, A Generic and Cell-Type-Specific Wound Response Precedes Regeneration in Planarians. *Dev Cell* **35**, 632-645 (2015).
23. S. W. Lapan, P. W. Reddien, dlx and sp6-9 Control optic cup regeneration in a prototypic eye. *PLoS Genet* **7**, e1002226 (2011).
24. M. L. Scimone, M. Srivastava, G. W. Bell, P. W. Reddien, A regulatory program for excretory system regeneration in planarians. *Development* **138**, 4387-4398 (2011).
25. S. W. Lapan, P. W. Reddien, Transcriptome analysis of the planarian eye identifies ovo as a specific regulator of eye regeneration. *Cell Rep* **2**, 294-307 (2012).
26. D. Wenemoser, S. W. Lapan, A. W. Wilkinson, G. W. Bell, P. W. Reddien, A molecular wound response program associated with regeneration initiation in planarians. *Genes Dev* **26**, 988-1002 (2012).
27. M. W. Cowles *et al.*, Genome-wide analysis of the bHLH gene family in planarians identifies factors required for adult neurogenesis and neuronal regeneration. *Development* **140**, 4691-4702 (2013).
28. K. W. Currie, B. J. Pearson, Transcription factors lhx1/5-1 and pitx are required for the maintenance and regeneration of serotonergic neurons in planarians. *Development* **140**, 3577-3588 (2013).
29. M. L. Scimone, K. M. Kravarik, S. W. Lapan, P. W. Reddien, Neoblast specialization in regeneration of the planarian Schmidtea mediterranea. *Stem Cell Reports* **3**, 339-352 (2014).

30. M. L. Scimone, S. W. Lapan, P. W. Reddien, A forkhead transcription factor is wound-induced at the planarian midline and required for anterior pole regeneration. *PLoS Genet* **10**, e1003999 (2014).
31. K. J. r. Pedersen, Studies on the nature of planarian connective tissue. *Zeitschrift f¸r Zellforschung Zeitschrift f¸r Zellforschung und Mikroskopische Anatomie* **53**, 569-608 (1961).
32. M. Morita, Phagocytic response of planarian reticular cells to heat-killed bacteria. *Hydrobiologia* **227**, 193-199 (1991).
33. M. Morita, Structure and function of the reticular cell in the planarian Dugesia dorotocephala. *Hydrobiologia* **305**, 189-196 (1995).
34. M. L. Scimone *et al.*, foxF-1 Controls Specification of Non-body Wall Muscle and Phagocytic Cells in Planarians. *Curr Biol* **28**, 3787-3801.e3786 (2018).
35. X. He *et al.*, FOX and ETS family transcription factors regulate the pigment cell lineage in planarians. *Development* **144**, 4540-4551 (2017).
36. L. H. Hyman, *The invertebrates. Vol. II, Vol. II*. (McGraw-Hill book company inc., New York; London, 1951).
37. K. J. Pedersen, Some features of the fine structure and histochemistry of planarian subepidermal gland cells. *Zeitschrift fr Zellforschung Zeitschrift fr Zellforschung und Mikroskopische Anatomie* **50**, 121-142 (1959).
38. R. M. Zayas, F. Cebria, T. Guo, J. Feng, P. A. Newmark, The use of lectins as markers for differentiated secretory cells in planarians. *Dev Dyn* **239**, 2888-2897 (2010).
39. K. Nishimura *et al.*, Characterization of tyramine beta-hydroxylase in planarian Dugesia japonica: cloning and expression. *Neurochem Int* **53**, 184-192 (2008).
40. C. S. Baron *et al.*, Cell Type Purification by Single-Cell Transcriptome-Trained Sorting. *Cell* **179**, 527-542.e519 (2019).
41. R. Amamoto *et al.*, Probe-Seq enables transcriptional profiling of specific cell types from heterogeneous tissue by RNA-based isolation. *Elife* **8**, (2019).
42. P. W. Reddien, Constitutive gene expression and the specification of tissue identity in adult planarian biology. *Trends Genet* **27**, 277-285 (2011).
43. J. N. Witchley, M. Mayer, D. E. Wagner, J. H. Owen, P. W. Reddien, Muscle cells provide instructions for planarian regeneration. *Cell Rep* **4**, 633-641 (2013).
44. M. L. Scimone, L. E. Cote, T. Rogers, P. W. Reddien, Two FGFRL-Wnt circuits organize the planarian anteroposterior axis. *Elife* **5**, (2016).

45. E. G. Schad, C. P. Petersen, STRIPAK Limits Stem Cell Differentiation of a WNT Signaling Center to Control Planarian Axis Scaling. *Curr Biol* **30**, 254-263.e252 (2020).
46. M. Nitzan, N. Karaiskos, N. Friedman, N. Rajewsky, Gene expression cartography. *Nature* **576**, 132-137 (2019).