Branching out: how a planarian neoblast generates cellular diversity

by

Hunter O. King

B.S. Neuroscience (2015)
B.A. Biochemistry (2015)
University of Washington, Seattle

Submitted to the Department of Brain and Cognitive Sciences
In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2024

Authored by: **Hunter King**
Department of Brain and Cognitive Sciences
September 5, 2023

Certified by: **Peter W. Reddien, Ph.D.**
Professor of Biology
Thesis Supervisor

Accepted by: **Mark Harnett, Ph.D.**
Associate Professor in Brain and Cognitive Science
Graduate Officer

Branching out: how a planarian neoblast generates cellular diversity

By

Hunter O. King

Submitted to the Department of Brain and Cognitive Sciences on September 5th, 2023
in Partial Fulfillment of the Requirements for the Degree of Doctor
Of Philosophy

ABSTRACT

The extraordinary regenerative capacity of planarians is derived from their large population of pluripotent stem cells called neoblasts. Neoblasts are a heterogenous population of stem cells that begin unspecified, but are rapidly specified within a single cell cycle and give rise to every cell in the animal. In this work, we characterize the different types of specified neoblasts in the planarian *Schmidtea mediterranea* by using single-cell RNA sequencing. We also profile their post-mitotic descendants that serve as migratory progenitors. Through these analyses, we uncover many novel neoblast and post-mitotic progenitor types, and the genes that define them. Because neoblasts subtypes are often specified by their unique expression of Fate Specific Transcription Factors (FSTFs), we computationally identify planarian transcription factor genes and characterize neoblasts and post-mitotic progenitors by their expression of these putative transcription factors. We find that many cell types can be identified through their unique expression of FSTF modules, which can be used to connect cells with the same fate across neoblast, post-mitotic progenitor, and differentiated cell stages. Through gene inhibition studies, we discover that many of the FSTFs discovered for these novel precursor types have a functional role in their specification. We note that transcriptional signatures of some cell-type fates become apparent at different stages in the lifetime of the progenitor and propose a model where cellular diversity arises at different times for different tissue classes.

To better uncover the genes signatures that define different cell-types, we develop methods that use neural networks to learn patterns in gene expression in planarian post-mitotic progenitors. We find that neural network classifiers are powerful predictors of cell-type based on gene expression and the network's learned weights can be examined to uncover gene signatures that define different progenitor types. We find that autoencoders, a class of neural networks made for the efficient representation of data, can be used to learn gene signatures in cells in an unsupervised manner. We find that cells close together in UMAP space, but belonging to different cell clusters through traditional clustering, are often difficult for the neural networks to distinguish. From this, we hypothesize that cells in different clusters may be transcriptionally similar and cells existing across continuous UMAP space may exist across continuous gene-expression space, and so assigning cells to discrete clusters may not always be appropriate. We propose autoencoders, which can encode gene-expression signatures through a continuous, latent-space encoding, may be more appropriate and can be used to uncover novel gene and cell relationships.

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

Table of Contents

Chapter 1

Introduction

Introduction

Regeneration

Regeneration is an evolutionary strategy for repairing damaged and missing tissues that is widespread throughout the animal kingdom. While different species have different regenerative capabilities, most are able to regenerate to some extent (Bely & Nyberg, 2010). Studies and surveys of regeneration across species have revealed different types of regeneration employed by animals of different phyla. At the cellular level, some damaged cells can regrow processes, such as mammalian axonal regeneration following retrograde, Wallerian degeneration of peripheral neuron axons (Ramón y Cajal, 1928; Waller, 1851). Tissues like the mammalian epidermis and intestine can regenerate homeostatically (Leblond, 1981; Leblond & Stevens, 1948), resulting in their constant turnover, and can also regenerate after injury (Hageman et al., 2020; Martin, 1997). In animals with complex body plans, whole organs can regenerate like the mammalian liver after partial excision (Pack et al., 1962).

In its popular context, regeneration often refers to the replacement of complex structures. Structural regeneration is present throughout bilaterians (Bely & Nyberg, 2010), but is often specific to the environmental pressures on the animal. Zebrafish, for example, can regenerate their caudal fins (Pfefferli & Jazwinska, 2015). This is ecologically-relevant as fish can experience ‘fin rot’ (Khan et al., 1981), which results in the decay of fins from infection, and ‘fin erosion’ resulting from injuries from fighting, failed predation, abrasion, or nutrient deficiencies (Latremouille, 2003) (95+% of salmon have damaged dorsal fins (Petersson et al., 2013)). Some amphibians and lizards can regenerate their tails and limbs (Alibardi, 2018; Thornton, 1968), which is useful for

regenerating after sublethal predation and allows for the emergent survival strategy of tail dropping to escape predators (Clause & Capaldi, 2006).

Less commonly, some animals are also capable of whole-body regeneration. These animals, which include planarian flatworms, hydra, and sea stars, are able to regenerate whole organisms from small fragments missing entire organ systems (Bely & Nyberg, 2010)Cary, 2019 #5702}. T.H. Morgan reported that a planarian regenerated from a fragment he predicted was 1/279th the size of an intact animal (Morgan, 1898). This type of regeneration allows for asexual reproduction strategies like budding in hydra and fissioning in planarians and starfish.

The relevance of regeneration to the fitness of the animal is obvious: if you make an organism, why wouldn't you maintain it? However, the benefits conferred by regeneration are balanced by the complexity and energy required to perform it. August Weismann hypothesized in the late 1800s that the regenerative abilities of an animal's systems are positively influenced by how necessary the system is for the animal's survival and how likely and often it is injured (Elchaninov et al., 2021; Weismann et al., 1893). This is negatively balanced with how difficult it would be to regenerate the tissue, which he hypothesized was based on its complexity. With increasing complexity of the system, it is possible that evolution would optimize for its maintenance by preventing damage instead of regeneration (Elchaninov et al., 2021; Needham, 1952).

Regeneration in Planarians

Planarians are flatworms with a high whole-body regenerative capacity. The planarian model species *Schmidtea mediterranea* is able to regenerate whole worms from small

body fragments in about a week. Planarians have a complex anatomy with different organs distributed throughout the body (Hyman, 1951). The nervous system comprises an anterior, bilobed brain extending into a pair of ventral nerve cords that run the anterior-posterior length of the animal (Figure 1). A pair of eyes lie dorsal to the brain. The brains and ventral nerve cords consist of shell-like layers of different neuron types (Cebria, Kudome, et al., 2002), which send projections through the interior and form a neuropil core (Okamoto et al., 2005). The neuropil of the brain and ventral nerve cords contain glial cells (Roberts-Galbraith et al., 2016; Wang et al., 2016). Outside of the central nervous system, planarians also have many neurons scattered throughout the body, forming a peripheral nervous system. Planarians also have multiple classes of muscle cells with different orientations along the body (Cebrià & Vispo, 1997; Orii et al., 2002), a single layer of epidermal cells covering the surface of the animal (Glazer et al., 2010; Tazaki et al., 2002), germline-associated cells (Issigonis et al., 2022; Khan & Newmark, 2022; Newmark et al., 2008), and a feeding organ near the center of the animal on the ventral side called the pharynx, which is primarily a mix of specialized muscle, epithelial, and neural cells (Fincher et al., 2018). Small excretory organs called protonephridia are interspersed throughout the animal and consist of tubule and ciliated flame cells that pump liquid from the animal through pores in the epidermal surface (Scimone et al., 2011). Planarians also have a diverse population of gland cells (also called parenchymal cells) and phagocytic cells (Fincher et al., 2018; Scimone et al., 2018), although their functions have been less studied. These cell types are all generated by a heterogenous population of neoblasts, a pluripotent stem cell population responsible for generating all cell types in the animal, including other neoblasts. Since

planarians have a complex anatomy of many cell types that are unevenly distributed across the animal, planarians must be able to regenerate entire organ systems after injuries that completely remove them from the animal.

The regenerative capabilities of planarians give them the ability to reproduce through asexual fissioning. A chromosomal translocation in a recent ancestral *Schmidtea mediterranea* individual is thought to have caused the loss of sexual reproduction and generated the 'asexual' strain commonly used as a model today (Baguna et al., 1999). This strain reproduces purely by fissioning and generates 'clonal' individuals with each division. However, since each individual inherits a different population of neoblasts which can have unique mutations, and because neoblasts can accumulate different mutations over time, it's possible that there is some genetic diversity among clonal planarians descended from the same individual.

Between 1-6 hours following fission or an injury, planarian cells near the wound site express early wound-induced genes that help initiate regeneration (Petersen & Reddien, 2009b; Wenemoser et al., 2012). This coincides with a global increase in mitoses and a local increase in cell death around the wound site (Baguñà, 1976; Saló & Baguñà, 1984; Wenemoser & Reddien, 2010). This early response to wounding is called the generic wound response and it occurs independent of the extent of injury. At injuries that remove substantial tissue, wound-induced genes trigger a second response that occurs 24-48 hours after injury termed the 'missing tissue response' (Wenemoser & Reddien, 2010). This includes a local increase in neoblast mitosis at the wound site and a global increase in cell death (Pellettieri et al., 2010; G, 2010). The missing tissue response requires expression of the gene *follistatin*, and inhibiting it through *follistatin*

RNAi decreases the speed of regeneration , potentially by blocking an increase in proliferation of neoblasts and their descendant cells, which are post-mitotic progenitors that give rise to new tissues (Gavino et al., 2013; Roberts-Galbraith & Newmark, 2013; Tewari et al., 2018).

The proliferation of neoblasts and coinciding generation of post-mitotic progenitors near the wound site creates a front of largely undifferentiated and newly-generated cells called a ‘blastema’. This unpigmented tissue is characteristic of planarian regeneration and regeneration in some other species, such as limb regeneration in amphibians (McCusker et al., 2015). The blastema allows for the rapid replacement of missing tissues by concentrating neoblasts near the site of injury. If blastema formation is prevented by *follistatin* RNAi, regeneration is dramatically slowed (Tewari et al., 2018). Irradiation to selectively kill neoblasts also prevents regeneration, indicating new cells generated from neoblasts are required to regenerate missing tissues (Bardeen & Baetjer, 1904; Dubois, 1949; Reddien et al., 2005). The exact logic of planarian regeneration is unknown. Planarians may employ an intercalative filling of missing tissues, where all missing tissues are gradually specified and incorporated, rather than a distal outgrowth method of regeneration where progressively more distal tissues are specified and incorporated (Vogg et al., 2014). It is also unclear to what extent epimorphosis (regeneration primarily through new tissue incorporation) and morphallaxis (the reorganization of existing tissue to re-establishment proper proportions) are employed. Though, because new cells are required for regeneration, old cells die and are replaced normally by homeostatic turnover (Pellettieri & Sánchez Alvarado, 2007) which is accelerated in regeneration, and because there is a period of

time following regeneration where the animal tissues are out-of-proportion, it is likely regeneration is primarily conducted through an epimorphosis process.

Planarian Neoblasts

Neoblasts are planarian stem cells that are the only cycling cell type in the asexual animal and give rise to all planarian differentiated tissues. Neoblasts exist throughout the animal and are marked by the expression of *smedwi-1*, a planarian homolog encoding the gene for the widely conserved RNA-binding protein PIWI (Reddien et al., 2005). Previous work has identified and characterized many classes of specialized neoblasts, neoblasts that are fated to become specific tissues and can be identified by their unique expression of fate-specific transcription factors (FSTFs). The planarian eye, for example, consists of photoreceptor neural cells and pigment cells, and are specified by the shared expression of the FSTFs *ovo*, *six1/2-1* and *eya (Lapan & Reddien, 2011, 2012)*. Photoreceptor neurons specifically express the FSTF *otxA* and pigment cells express the FSTFs *sp6-9* and *dlx*. Inhibition of the shared FSTFs *ovo*, *six1/2-1,* and *eya* halt the specification of new eye cells, whereas the specific FSTFs only prevent the generation of their corresponding cell type *(Lapan & Reddien, 2011, 2012)*.

Recent work by Raz et al. has suggested that fate specification of a neoblast generally happens within a single cell division, starting with an unspecialized neoblast in the beginning of the cell cycle (G1) and showing more specification (through the expression of unique FSTFs) as a cell progresses through the cell cycle towards division (Raz et al., 2021) (Figure 2). The specialized neoblast can divide asymmetrically to produce a new, unspecialized, pluripotent neoblast that will re-enter

the cell cycle and a specialized post-mitotic progenitor that is no longer a neoblast and will migrate and differentiate into a mature cell without further divisions. Neoblasts can also divide symmetrically to generate two neoblasts, or potentially two specialized post-mitotic progenitors. Previous work analyzing neoblast colonies following sub-lethal irradiation have shown that neoblasts likely do not retain memory of previous divisions (the probability of specifying a fate is not strongly influenced by the prior fates of that neoblast lineages) (Raz et al., 2021). This ‘single-step fate model’ precludes the necessity of a complex hierarchy of progressive fate specification seen in other systems.

Previous work looking at fate specification in neoblasts has identified many classes of specialized neoblasts for many tissues. These specialized neoblasts often correspond to specific cell subtypes, instead of for broader classes of cells. For example, specialized neoblasts have been identified for almost every class of muscle, but, unlike for the planarian eye, there is no evidence for a pan-muscle class of neoblast that will further specify to a specific class of muscle (Scimone et al., 2020; Scimone et al., 2017; Scimone, Lapan, et al., 2014; Scimone et al., 2018). Because of this, there exist many classes of specialized neoblasts and there could in principle be a specialized neoblast corresponding to each of the ~150 mature cell types in the animal. However, specialized neoblast classes have been largely found though individual case studies and a comprehensive characterization of specialized neoblast classes through high-throughput methods like RNA-sequencing has not been conducted.

Planarian Position Control Genes

Planarians express a group of regionally-specific genes with a role in patterning, or predicted role in a planarian patterning pathway, called position control genes (PCGs). These genes are differentially expressed in muscle cells along at least one major body axis (Witchley et al., 2013) and several have been shown to have a role in setting the positioning of different tissue and cell types within the animal (Cebria, Kobayashi, et al., 2002; Hill & Petersen, 2015; Lander & Petersen, 2016a; Scimone et al., 2016; Sureda-Gómez et al., 2015) (Figure 3). The most characterized PCGs are members of the Wnt signaling pathway (Cebria et al., 2018; Reddien, 2018; Rink, 2018). Wnts are secreted signaling proteins that are important in a wide range of developmental processes (Wodarz & Nusse, 1998). Wnt signaling is considered 'modular' in that it is an evolutionarily-conserved signaling pathway that is used for different processes in different organisms and in different developmental contexts and times (Komiya & Habas, 2008; Srivastava, 2021). In planarians, Wnt family members are constitutively expressed, primarily in the posterior of the animal (Petersen & Reddien, 2009a). *β-catenin*, the main downstream transcription factor in Wnt signaling is also expressed in a gradient across the anterior-posterior (AP) axis, with higher levels in the posterior (Stückemann et al., 2017; Sureda-Gómez et al., 2016). In the anterior of the animal, cells express *notum* (Petersen & Reddien, 2011), an extracellular inhibitor of Wnt signaling and various soluble, non-signaling forms of the Wnt ligand Frizzled receptors (Gurley et al., 2010; Petersen & Reddien, 2008), which also act as negative regulators in Wnt signaling. Inhibition of Wnt signaling through RNAi of *β-catenin* results in cells

adopting anterior identity and leads to the formation of ectopic heads throughout the AP axis of the animal (Gurley et al., 2008; Iglesias et al., 2008; Petersen & Reddien, 2008).

Cells in the anterior region of the animal express a separate system of PCGs, a group of genes with homology to FGF receptors, but that lack the intracellular tyrosine kinase signaling domain (FGF Receptor Likes; FGFRLs) (Cebria, Kobayashi, et al., 2002; Lander & Petersen, 2016b; Scimone et al., 2016). The canonical member of this family, *nou-darake* (*ndk*), has been shown to be necessary for proper positioning of the anterior eyes and neural tissue, and inhibition of *ndk* by RNAi causes ectopic, posterior eyes to appear (Cebria, Kobayashi, et al., 2002). Since these receptors lack the intracellular signaling domain, it has been hypothesized that they are decoy receptors or help to concentrate FGFs in the anterior of the animal, but the exact mechanism by which they influence anterior patterning is not known (Cebria, Kobayashi, et al., 2002; Gerber et al., 2009).

The other major body axes of planarians express their own unique PCGs. The mediolateral (ML) axis is guided primarily by *slit* medially (Cebria et al., 2007) and a non-canonical Wnt (*wnt5*) laterally (Gurley et al., 2010). This non-canonical Wnt does not signal through β-catenin like AP-expressed Wnts. The dorsoventral (DV) axis is primarily guided by Bone Morphogenic Protein (*bmp4*), which is expressed dorsally in a medial-to-lateral gradient (Molina et al., 2007; Orii & Watanabe, 2007; Reddien et al., 2007). Anti-Dorsalizing Morphogenic Protein (*admp*) expressed laterally on the ventral surface acts in opposition to *bmp4*. Many of these PCGs have been shown to influence the patterning of the animal by the presence of ectopic, mistargeted tissue after their inhibition.

Planarian Regeneration and Turnover Involves a Coordination of These Concepts

All of these complex biological processes must coordinate during regeneration. Neoblasts concentrating at injury sites allows for rapid regeneration of missing tissues. Neoblasts must then specify to the identity of the missing cell types in order to generate the post-mitotic progenitors to replace them. While neoblasts exist throughout the entire animal, some neoblast classes are specified only in certain regions of the animal, such as eye neoblasts that are specified only in the anterior half of the animal (Lapan & Reddien, 2012). This regional specification means post-miotic migratory progenitors will have to migrate less distance and can replace tissues faster. These specification zones are also spatially course, which allows some specialized neoblasts and post-mitotic progenitors to potentially remain after amputation of certain tissues (e.g., if the head was removed, some existing eye progenitors would remain in the posterior piece) (Reddien, 2019, 2021, 2022). Since some neoblast specification zones are spatially-restricted, it is likely these neoblasts are using PCGs to infer their position in the animal and choose their fate. This has only been shown clearly with *bmp4* RNAi, which results in the ventralization of the dorsal surface. After *bmp4* RNAi, the dorsal epidermis begins to contain ciliated cells expressing the marker *kal1*, which are normally restricted to the ventral surface (Wurtzel et al., 2017). Irradiation depleting neoblasts prevents this ventralization of the dorsal surface, suggesting this ventral conversion requires new cell generation. At the same time, specialized neoblasts expressing the ventral epidermis marker *kal1* are now apparent on the dorsal side of the animal, suggesting that neoblasts, and not just their descendant post-mitotic progenitor cells, are detecting

bmp4 gradients. While the influence of PCGs on regional neoblast specification occurs during homeostasis, it is also important during regeneration. Previous studies have shown an increase in the proportion of specialized neoblasts for missing tissues after selective tissue amputation (Bohr et al., 2020; LoCascio et al., 2017). Existing evidence suggests this phenomenon is not the result of a complex feedback mechanism between differentiated tissues signaling their presence (or lack of presence) to neoblasts, but a coordination of local neoblast amplification near the missing tissue and the extracellular PCG environment likely encouraging fates corresponding to the missing tissues (the 'bystander effect') (LoCascio et al., 2017).

While PCGs are constitutively expressed, they are also dynamic. Following substantial tissue loss, entire domains of PCGs may be removed and these domains would correspond to the cell types that need to be replaced. To solve this problem, planarians rapidly rescale their PCG expression domains following injury (Petersen & Reddien, 2009b; Reddien, 2018). This rescaling of PCG domains allows the animal to quickly regenerate missing tissues and reestablish proper body proportions. Several PCGs are also wound-induced before PCG rescaling and can quickly change the extracellular PCG environment of the blastema. The posterior PCG *wnt1* is expressed rapidly after wounding in both anterior and posterior-facing wounds, favoring posterior identity (Gurley et al., 2010; Petersen & Reddien, 2009b). The extracellular Wnt inhibitor *notum*, is then specifically expressed in anterior-facing wounds, to favor anterior identities (Petersen & Reddien, 2011; Wurtzel et al., 2015). Working in tandem, these two wound-induced PCGs could rapidly regenerate anterior and posterior PCG domains.

Together these biological systems function to lend planarians the amazing ability to homeostatically turn over tissue, dynamically grow and shrink an order of magnitude in size in response to feeding state, asexually reproduce through fission, and regenerate after a diverse range of injuries.

Cell Lineage Structure for Planarians is Currently Debated

In recent years, there has been substantial development in our understanding of cell lineages in planarians. Previous work has identified neoblasts as the only dividing cell in asexual planarians, but our ability to characterize possible heterogeneity among neoblasts has been limited. Single-cell transplantation experiments have identified ‘clonogenic neoblasts’, single neoblasts that can reconstitute the neoblast population and progenitors for all cell types in the animal, and rescue animals depleted of all neoblasts (Wagner et al., 2011). Early single-cell RNA sequencing experiments have identified broad classes of neoblasts, characterized by the common enrichment of several fate-specific transcription factors (FSTFs) (van Wolfswinkel et al., 2014). These broad classes generally correspond to tissues, such as zeta-neoblasts, which are primarily epidermal-specified. Another class of neoblasts, sigma-neoblasts, were thought to generate cell types of many lineages, including zeta-neoblasts. Since these sigma neoblasts could give rise to neoblasts of other types, as well as cells making up many tissue types, this class was thought to contain pluripotent or ‘clonogenic’ neoblasts. Case studies also exist characterizing more specific neoblast types as well, such as for specific subtypes of neurons (Cowles et al., 2013; Cowles et al., 2014; Currie & Pearson, 2013; Marz et al., 2013; Molinaro & Pearson, 2016; Roberts-

Galbraith et al., 2016; Ross et al., 2017; Ross et al., 2018; Scimone, Kravarik, et al., 2014; Wenemoser et al., 2012) or muscle, which are generally defined by the expression more restricted FSTFs (Chen et al., 2013; Scimone et al., 2020; Scimone et al., 2017; Scimone, Lapan, et al., 2014; Scimone et al., 2018; Vásquez-Doorman & Petersen, 2014; Vogg et al., 2014). One example of this is the small population of sensory neurons along the anterior rim of the head which has been found to require the expression of the transcription factor *klf4* for their specification (Scimone, Kravarik, et al., 2014). Together, this work suggests that there exist neoblasts that are pluripotent and neoblasts committed to broad or specific fates.

More recently, high-yield single-cell RNA sequencing has allowed for better characterization of neoblasts and their lineage. Some studies have sought to more specifically define the population of pluripotent, clonogenic neoblasts in the animal. Zeng et al. proposed a population of neoblasts enriched in the transmembrane protein Tetraspanin as the clonogenic neoblast that could self-renew and produce specialized neoblasts of other types (Zeng et al., 2018). Other groups have proposed other classes of pluripotent neoblasts (Cui et al., 2023), as well as differing lineage structures for neoblasts based on different computation techniques applied to single-cell RNA sequencing (Plass et al., 2018). The true cell lineage structure in planarians is still debated.

In 2021 Raz et al. characterized specification in neoblasts at different stages of the cell cycle (Raz et al., 2021). Surprisingly, they found low levels of specification signatures (the expression of known FSTFs) in neoblasts at the beginning of the cell cycle (G1) and an increase in FSTF expression as the cells progress through the cell

cycle toward division. This suggested the neoblasts enter the cell cycle unspecified, meaning specification would have to occur in a single cell cycle. This ‘single-step fate model’ would not require a complex lineage structure of neoblast divisions. It doesn’t preclude the presence of a clonogenic, unspecialized neoblast, but it does suggest that specialized neoblasts could give rise to unspecialized neoblasts in G1. In this way, the clonogenic, pluripotent neoblast class could be all neoblasts in the G1 phase of the cell cycle. This is consistent with previous data identifying a population of clonogenic neoblasts, because all neoblasts could be clonogenic, but presents an alternative model supporting those results.

To come closer to a resolution between these alternative models for cell lineage in planarians, a fuller characterization of planarian neoblasts is useful. Given that G1 neoblasts can be unspecified, do we see evidence for an unspecified neoblast in later cell cycle stages? Since most neoblast characterizations have identified broad classes of neoblasts specific to different tissues, but case studies have identified neoblasts for more specific cell subtypes, are neoblasts more heterogenous than thought and consist of unique progenitors for each cell subtype in the animal, or is identity diversified later as cells progress towards differentiation (as post-mitotic progenitors)?

The Determinants of Fate Selection are Unknown

The logic of fate selection in neoblasts is also an outstanding problem in planarian biology. While it is known that certain FSTFs are important for the specification of specific neoblast types, it is likely that not all known neoblast types and their

corresponding FSTFs have been discovered. It is also not known what causes specific FSTFs to be expressed in different neoblasts.

In *Drosophila* neurogenesis, neuroblasts express a temporal series of transcription factors that generate different types of ganglion mother cells, which then give rise to different neural and/or glial subtypes. These transcriptional transitions between sequential transcription factors are cell-intrinsic, with the first transition from *hunchback* to *Krüppel* being dependent on cytokinesis and later transitions not being cytokinesis-dependent (Grosskortenhaus et al., 2005). This intrinsic, temporal method of fate specification is useful for embryogenesis, where the proportions of cell types can be tightly controlled. However, in regeneration at a given wound, not all tissue types need to be replaced, so temporal fate specification could generate unneeded precursor types and could require more time since the generation of cell types generated later in their lineage would have to be preceded by the generation of other intermediate cell types.

Cell specification in some systems also occurs temporally, but through extrinsic mechanisms (Kohwi & Doe, 2013). For example, in mammalian brain development a multipotent progenitor can give rise to deep and superficial neurons, as well as glia later in development (Guillemot, 2007; Leone et al., 2008; Molyneaux et al., 2007; Rowitch & Kriegstein, 2010). The switches between these stages are guided at least in part by extrinsic cues. For example, newly differentiated cortical neurons will express and secrete Cardiotrophin 1, which triggers a feedback mechanism in their progenitors to switch to the specification of glial cells (Barnabe-Heider et al., 2005). This mechanism

diverges from *Drosophila* neuroblasts because it is driven by extrinsic cues and because fate switches are, at least locally, synchronized.

Models of fate specification support both deterministic and stochastic mechanisms in fate determination (Zechner et al., 2020). These models can both be applied to complex lineage hierarchies. A classic example of a complex deterministic method of fate specification is *C. elegans* embryonic and neural development, where each characterized neuron is generated at predictable times and in exact proportions (Hobert, 2010; Sulston et al., 1983). Stochastic fate determination includes hematopoiesis (Weinreb et al., 2020) and retinal progenitor specification in zebrafish (Chen et al., 2012; He et al., 2012) and photoreceptor specification in *Drosophila* photoreceptors (Mikeladze-Dvali et al., 2005), where symmetric fate divisions can favor different fates along a complex lineage hierarchy.

Other methods directing fate-specification exist as well. Position and cell-cell interactions can influence fate specification extrinsically. There is some evidence for more niche methods of fate-specification. In rats, retrograde signals derived from post-synaptic neurons can cause fate switches in some neural classes. For example, retrograde Nerve growth factor (NGF) signaling is responsible for the neuropeptide Calcitonin-gene related peptide (CGRP) to be expressed in nociceptive neurons through activation and internalization of presynaptic Tropomyosin receptor kinase A (TrkA) (Harrington & Ginty, 2013; Zweifel et al., 2005). There is some debate on whether planarian neoblasts can detect the identity of missing tissues to specifically amplify the generation of their progenitors. An increase in progenitors for a missing tissue is often seen, but can be explained by the ‘bystander effect’, where the proliferation of neoblasts

near wounds leads to an increase in progenitors for neighboring tissues, because there are more neoblasts in the PCG specification zone for those cells (Bohr et al., 2020; LoCascio et al., 2017). Unlike a more specific feedback mechanism, this involves the proliferation of progenitors for neighboring tissues that were not removed. This was shown by the surgical removal of the pharynx, which results in an increase in progenitors for the pharynx, but also an increase in progenitor incorporation into the neighboring ventral nerve cords (LoCascio et al., 2017).

Because regeneration is rapid and not all tissues need to be replaced, and because neoblasts seem to be specified within a single cell division, it is unlikely that neoblast specification occurs in a strict hierarchical lineage. Neoblasts are also often specified far from the location of their differentiated tissue, like for neoblasts that give rise to planarian eye which exist in a broad region posterior to the eyes (Lapan & Reddien, 2012). This makes it unlikely that fates are determined by cell-cell interactions. Since neoblasts infer their positions by PCG gradients, and this can influence fate, it is possible fates are determined probabilistically based on a cell's position.

Limitations of Techniques and Methods

Advancements in scientific methods have allowed for better and more thorough characterization of cell classes and the fate decisions that generate them. One of the highest throughput and least biased of these are single-cell sequencing techniques, principally mRNA sequencing. Rapid microfluidic-based cell-sorting, mRNA library preparation, and next-generation sequencing methods allow for the rapid detection of mRNAs within tens of thousands of cells. These cells can be characterized in a biased

or ‘supervised’ fashion, by looking at the expression levels of genes of predetermined interest. However, the generation of large single-cell mRNA libraries allows for the use of less biased, ‘unsupervised’ characterization methods as well.

Dimensionality reduction methods can reduce cells from gene-level expression space (where each cell is represented in a high-dimensionality space equivalent to the number of genes expressed), to lower-dimensionality space representing the expression of many genes simultaneously. Classically, this was principal component space, where linear combinations of genes maximizing the variation among cells could be found. More recently, variations have been generated that casts principal components onto a manifold for better visualization in two-dimensional space (Xiang et al., 2021). Since these dimensions maximize variance across cells, they tend to visually separate cell types, since they have different gene-expression profiles.

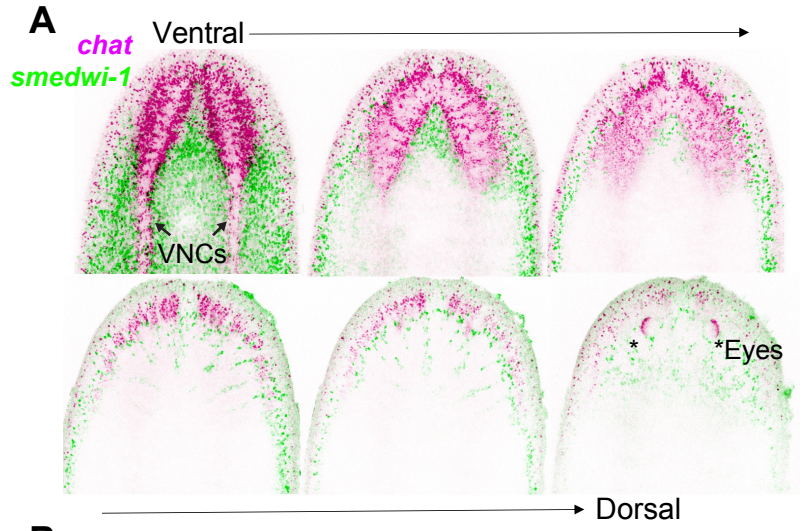
Data clustering methods can be used simultaneously to group cells with similar gene-expression profiles into different distinct clusters. This approach has some pitfalls, like deciding what constitutes a significant difference between cells to assign it a different identity. Additionally, while most clustering methods group cells into discrete clusters, some differences in cell types might not be discrete, but may be generated by changes in gene-expression along a continuous distribution (Watson E.R., 2022). Often, statistical tests are run to find genes that are expressed at different levels between cells of different clusters. This is a useful approach to find marker genes for cells of different clusters, but it has problems as well. Cells can be significantly enriched in genes also expressed in other cell types. Complex relationships in gene expression can be lost as well; for example, combinatorial or regulatory relationships that exist between multiple

expressed genes cannot be detected by the enrichment of a single gene in a given cluster.

Aims

In this work we seek to better characterize the full diversity of planarian neoblasts and their descendent post-mitotic progenitors, which are cells that migrate to the location of their mature tissue before differentiation. Using single-cell RNA sequencing, we characterize the diversity in these two precursor cell populations, in the context of regeneration, by using traditional clustering techniques and determining the modules of transcription factors that define these cell types, and at which time in the cells' lifetimes we can detect these identifying signatures. We also develop machine-learning based computational methods to infer how cells are defined by their unique combinatorial expression of different genes. From these studies we characterize unknown neoblast and post-mitotic progenitor populations and identify at what cell stages cell-type diversity arises in planarians. The results of these studies will be useful in characterizing planarian cell lineages in the context of planarian regeneration.

Figure 1


A
chat
smedwi-1
Ventral
VNCs
*
*Eyes
Dorsal

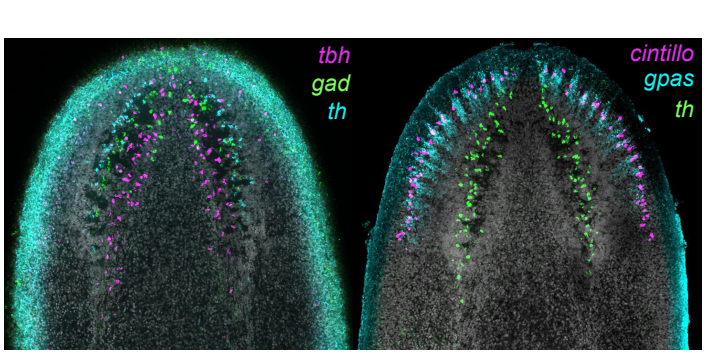

B
tbh
gad
th
cintillo
gpas
th

Figure 1. Anatomy of the planarian nervous system.
The planarian central nervous system is delineated by *chat*-expressing neurons in the anterior bilobed brain and ventral nerve cords on the ventral surface of the animal (A), surrounded by *smedwi-1*-expressing neoblasts. Brain branches extend from the brain dorsolaterally and a pair of eyes lie on the dorsal surface of the animal above the brain. The planarian central nervous system consists of a heterogenous population of neurons which exist in different spatial regions (B).

Figure 2

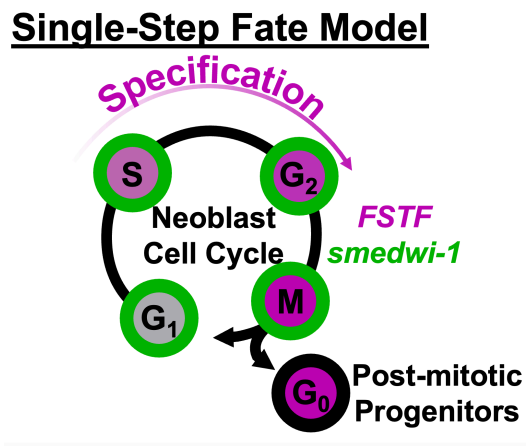

A
Single-Step Fate Model
Specification
S
G2
Neoblast
Cell Cycle
FSTF
smedwi-1
G1
M
G0
Post-mitotic
Progenitors

Figure 2. Single-step fate model
The single-step fate model introduced by Raz et al. suggests that a neoblast enters the cell cycle (G1) unspecified and express more FSTFs (a proxy for specification) as the cell progresses towards division (A). The neoblast can divide asymmetrically to give rise to another unspecialized G1 neoblast and a specialized post-mitotic progenitor.

Figure 3

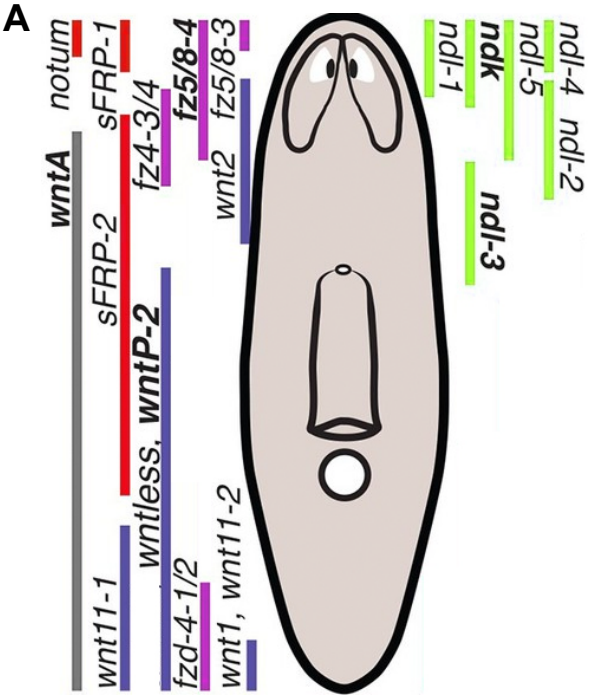

A
notum
sFRP-1
wntA
fz4-3/4
fz5/8-4
fz5/8-3
sFRP-2
wnt2
wntless, wntP-2
ndl-1
ndk
ndl-5
ndl-4
ndl-2
ndl-3
wnt11-1
fzd-4-1/2
wnt1, wnt11-2

Figure 3. Planarians express regionally-specific, position control genes
Planarians express different PCGs in different spatial domains within the animal (A). *nou-darake* (*ndk*) and other FGFRLs (*ndl*s) are expressed in the anterior (green) Wnts signaling through canonical Wnt signaling are primarily expressed in the posterior (blue), with their receptors expressed in cells in the anterior and posterior (magenta). Non-canonical *wntA* is expressed up to the posterior of the head (grey). The Wnt signaling inhibitors *notum* and *sFRP-1* and *sFRP-2* are expressed in the anterior (red).

References

Alibardi, L. (2018). Perspective: Appendage regeneration in amphibians and some reptiles derived from specific evolutionary histories. *Journal of Experimental Zoology Part B-Molecular and Developmental Evolution*, *330*(8), 396-405. https://doi.org/10.1002/jez.b.22835

Baguñà, J. (1976). Mitosis in the intact and regenerating planarian *Dugesia mediterranea* n.sp. I. Mitotic studies during growth, feeding and starvation. *J. Exp. Zool.*, *195*, 53-64.

Baguna, J., Carranza, S., Pala, M., Ribera, C., Giribet, G., Arnedo, M. A., Ribas, M., & Riutort, M. (1999). From morphology and karyology to molecules. New methods for taxonomical identification of asexual populations of freshwater planarians. A tribute to Professor Mario Benazzi. *Italian Journal of Zoology*, *66*(3), 207-214. https://doi.org/Doi 10.1080/11250009909356258

Bardeen, C. R., & Baetjer, F. H. (1904). The inhibitive action of the Roentgen rays on regeneration in planarians. *J. Exp. Zool.*, *1*, 191-195.

Barnabe-Heider, F., Wasylnka, J. A., Fernandes, K. J. L., Porsche, C., Sendtner, M., Kaplan, D. R., & Miller, F. D. (2005). Evidence that embryonic the onset of cortical neurons regulate gliogenesis via cardiotrophin-1. *Neuron*, *48*(2), 253-265. https://doi.org/10.1016/j.neuron.2005.08.037

Bely, A. E., & Nyberg, K. G. (2010). Evolution of animal regeneration: re-emergence of a field. *Trends Ecol Evol*, *25*(3), 161-170. https://doi.org/10.1016/j.tree.2009.08.005

Bohr, T. E., Shiroor, D. A., & Adler, C. E. (2020). Planarian stem cells sense the identity of missing tissues to launch targeted regeneration. *bioRxiv*, *05.05.077875*.

Cebria, F., Adell, T., & Salo, E. (2018). Rebuilding a planarian: from early signaling to final shape. *Int J Dev Biol*, *62*(6-7-8), 537-550. https://doi.org/10.1387/ijdb.180042es

Cebria, F., Guo, T., Jopek, J., & Newmark, P. A. (2007). Regeneration and maintenance of the planarian midline is regulated by a slit orthologue. *Developmental Biology*, *307*(2), 394-406. https://doi.org/S0012-1606(07)00900-1 [pii]
10.1016/j.ydbio.2007.05.006

Cebria, F., Kobayashi, C., Umesono, Y., Nakazawa, M., Mineta, K., Ikeo, K., Gojobori, T., Itoh, M., Taira, M., Sánchez Alvarado, A., & Agata, K. (2002). FGFR-related gene nou-darake restricts brain tissues to the head region of planarians. *Nature*, *419*(6907), 620-624. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12374980

Cebria, F., Kudome, T., Nakazawa, M., Mineta, K., Ikeo, K., Gojobori, T., & Agata, K. (2002). The expression of neural-specific genes reveals the structural and molecular complexity of the planarian central nervous system. *Mech Dev*, *116*(1-2), 199-204. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12128224

Cebrià, F., & Vispo, M. (1997). Myocyte differentiation and body wall muscle regeneration in the planarian *Girardia tigrina*. *Dev Genes Evol*, *207*, 306-316.

Chen, C. C., Wang, I. E., & Reddien, P. W. (2013). *pbx* is required for pole and eye regeneration in planarians [Research Support, N.I.H., Extramural]. *Development*, *140*(4), 719-729. https://doi.org/10.1242/dev.083741

Chen, Z., Li, X., & Desplan, C. (2012). Deterministic or stochastic choices in retinal neuron specification. *Neuron*, *75*(5), 739-742. https://doi.org/10.1016/j.neuron.2012.08.008

Clause, A. R., & Capaldi, E. A. (2006). Caudal autotomy and regeneration in lizards. *J Exp Zool A Comp Exp Biol*, *305*(12), 965-973. https://doi.org/10.1002/jez.a.346

Cowles, M. W., Brown, D. D., Nisperos, S. V., Stanley, B. N., Pearson, B. J., & Zayas, R. M. (2013). Genome-wide analysis of the bHLH gene family in planarians identifies factors required for adult neurogenesis and neuronal regeneration [Research Support, Non-U.S. Gov't]. *Development*, *140*(23), 4691-4702. https://doi.org/10.1242/dev.098616

Cowles, M. W., Omuro, K. C., Stanley, B. N., Quintanilla, C. G., & Zayas, R. M. (2014). COE loss-of-function analysis reveals a genetic program underlying maintenance and regeneration of the nervous system in planarians. *PLoS Genet*, *10*(10), e1004746. https://doi.org/10.1371/journal.pgen.1004746

Cui, G., Dong, K., Zhou, J. Y., Li, S., Wu, Y., Han, Q., Yao, B., Shen, Q., Zhao, Y. L., Yang, Y., Cai, J., Zhang, S., & Yang, Y. G. (2023). Spatiotemporal transcriptomic atlas reveals the dynamic characteristics and key regulators of planarian regeneration. *Nat Commun*, *14*(1), 3205. https://doi.org/10.1038/s41467-023-39016-0

Currie, K. W., & Pearson, B. J. (2013). Transcription factors *lhx1/5-1* and *pitx* are required for the maintenance and regeneration of serotonergic neurons in planarians [Research Support, Non-U.S. Gov't]. *Development*, *140*(17), 3577-3588. https://doi.org/10.1242/dev.098590

Dubois, F. (1949). Contribution á l 'ètude de la migration des cellules de règènèration chez les *Planaires dulcicoles*. *Bull. Biol. Fr. Belg.*, *83*, 213-283.

Elchaninov, A., Sukhikh, G., & Fatkhudinov, T. (2021). Evolution of Regeneration in Animals: A Tangled Story. *Frontiers in Ecology and Evolution*, *9*. https://doi.org/ARTN 621686
10.3389/fevo.2021.621686

Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M., & Reddien, P. W. (2018). Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science*, *360*(6391), 874. https://doi.org/10.1126/science.aaq1736

Gavino, M. A., Wenemoser, D., Wang, I. E., & Reddien, P. W. (2013). Tissue absence initiates regeneration through Follistatin-mediated inhibition of Activin signaling. *eLife*, *2*, e00247. https://doi.org/10.7554/eLife.00247

Gerber, S. D., Steinberg, F., Beyeler, M., Villiger, P. M., & Trueb, B. (2009). The murine Fgfrl1 receptor is essential for the development of the metanephric kidney. *Developmental Biology*, *335*(1), 106-119. https://doi.org/10.1016/j.ydbio.2009.08.019

Glazer, A. M., Wilkinson, A. W., Backer, C. B., Lapan, S. W., Gutzman, J. H., Cheeseman, I. M., & Reddien, P. W. (2010). The Zn finger protein Iguana impacts Hedgehog signaling by promoting ciliogenesis. *Developmental Biology*, *337*(1), 148-156. https://doi.org/S0012-1606(09)01284-6 [pii]
10.1016/j.ydbio.2009.10.025

Grosskortenhaus, R., Pearson, B. J., Marusich, A., & Doe, C. Q. (2005). Regulation of temporal identity transitions in Drosophila neuroblasts. *Dev Cell*, *8*(2), 193-202. https://doi.org/10.1016/j.devcel.2004.11.019

Guillemot, F. (2007). Cell fate specification in the mammalian telencephalon. *Prog Neurobiol*, *83*(1), 37-52. https://doi.org/10.1016/j.pneurobio.2007.02.009

Gurley, K. A., Elliott, S. A., Simakov, O., Schmidt, H. A., Holstein, T. W., & Sánchez Alvarado, A. (2010). Expression of secreted Wnt pathway components reveals unexpected complexity of the planarian amputation response. *Developmental Biology*, *347*(1), 24-39. https://doi.org/S0012-1606(10)00991-7 [pii]
10.1016/j.ydbio.2010.08.007
Gurley, K. A., Rink, J. C., & Sánchez Alvarado, A. (2008). Beta-catenin defines head versus tail identity during planarian regeneration and homeostasis. *Science*, *319*(5861), 323-327. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18063757
Hageman, J. H., Heinz, M. C., Kretzschmar, K., van der Vaart, J., Clevers, H., & Snippert, H. J. G. (2020). Intestinal Regeneration: Regulation by the Microenvironment. *Dev Cell*, *54*(4), 435-446. https://doi.org/10.1016/j.devcel.2020.07.009
Harrington, A. W., & Ginty, D. D. (2013). Long-distance retrograde neurotrophic factor signalling in neurons. *Nat Rev Neurosci*, *14*(3), 177-187. https://doi.org/10.1038/nrn3253
He, J., Zhang, G., Almeida, A. D., Cayouette, M., Simons, B. D., & Harris, W. A. (2012). How variable clones build an invariant retina. *Neuron*, *75*(5), 786-798. https://doi.org/10.1016/j.neuron.2012.06.033
Hill, E. M., & Petersen, C. P. (2015). Wnt/Notum spatial feedback inhibition controls neoblast differentiation to regulate reversible growth of the planarian brain. *Development*, *142*(24), 4217-4229. https://doi.org/10.1242/dev.123612
Hobert, O. (2010). Neurogenesis in the nematode Caenorhabditis elegans. *WormBook*, 1-24. https://doi.org/10.1895/wormbook.1.12.2
Hyman, L. H. (1951). *The Invertebrates: Platyhelminthes and Rhynchocoela The acoelomate bilateria* (Vol. II). McGraw-Hill Book Company Inc.
Iglesias, M., Gomez-Skarmeta, J. L., Salo, E., & Adell, T. (2008). Silencing of *Smed-betacatenin1* generates radial-like hypercephalized planarians. *Development*, *135*(7), 1215-1221. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18287199
Issigonis, M., Redkar, A. B., Rozario, T., Khan, U. W., Mejia-Sanchez, R., Lapan, S. W., Reddien, P. W., & Newmark, P. A. (2022). A Kruppel-like factor is required for development and regeneration of germline and yolk cells from somatic stem cells in planarians. *PLoS Biol*, *20*(7), e3001472. https://doi.org/10.1371/journal.pbio.3001472
Khan, R. A., Campbell, J., & Lear, H. (1981). Mortality in captive Atlantic cod, Gadus morhua, associated with fin rot disease. *J Wildl Dis*, *17*(4), 521-527. https://doi.org/10.7589/0090-3558-17.4.521
Khan, U. W., & Newmark, P. A. (2022). Somatic regulation of female germ cell regeneration and development in planarians. *Cell Rep*, *38*(11), 110525. https://doi.org/10.1016/j.celrep.2022.110525
Kohwi, M., & Doe, C. Q. (2013). Temporal fate specification and neural progenitor competence during development. *Nat Rev Neurosci*, *14*(12), 823-838. https://doi.org/10.1038/nrn3618
Komiya, Y., & Habas, R. (2008). Wnt signal transduction pathways. *Organogenesis*, *4*(2), 68-75. https://doi.org/10.4161/org.4.2.5851

Lander, R., & Petersen, C. P. (2016a). Wnt, Ptk7, and FGFRL expression gradients control trunk positional identity in planarian regeneration. *Elife*, *5*(e12850). https://doi.org/10.7554/eLife.12850

Lander, R., & Petersen, C. P. (2016b). Wnt, Ptk7, and FGFRL expression gradients control trunk positional identity in planarian regeneration. *Elife*, *5*. https://doi.org/10.7554/eLife.12850

Lapan, S. W., & Reddien, P. W. (2011). *dlx* and *sp6-9* control optic cup regeneration in a prototypic eye. *PLoS Genet*, *7*(8), e1002226. https://doi.org/10.1371/journal.pgen.1002226

PGENETICS-D-11-00610 [pii]

Lapan, S. W., & Reddien, P. W. (2012). Transcriptome Analysis of the Planarian Eye Identifies *ovo* as a Specific Regulator of Eye Regeneration. *Cell Reports*, *2*(2), 294-307. https://doi.org/10.1016/j.celrep.2012.06.018

Latremouille, D. N. (2003). Fin erosion in aquaculture and natural environments. *Reviews in Fisheries Science*, *11*(4), 315-335. https://doi.org/10.1080/10641260390255745

Leblond, C. P. (1981). The life history of cells in renewing systems. *Am J Anat*, *160*(2), 114-158. https://doi.org/10.1002/aja.1001600202

Leblond, C. P., & Stevens, C. E. (1948). The constant renewal of the intestinal epithelium in the albino rat. *Anat Rec*, *100*(3), 357-377. https://doi.org/10.1002/ar.1091000306

Leone, D. P., Srinivasan, K., Chen, B., Alcamo, E., & McConnell, S. K. (2008). The determination of projection neuron identity in the developing cerebral cortex. *Curr Opin Neurobiol*, *18*(1), 28-35. https://doi.org/10.1016/j.conb.2008.05.006

LoCascio, S. A., Lapan, S. W., & Reddien, P. W. (2017). Eye Absence Does Not Regulate Planarian Stem Cells during Eye Regeneration. *Dev Cell*, *40*(4), 381-391 e383. https://doi.org/10.1016/j.devcel.2017.02.002

Martin, P. (1997). Wound healing--aiming for perfect skin regeneration. *Science*, *276*(5309), 75-81. https://doi.org/10.1126/science.276.5309.75

Marz, M., Seebeck, F., & Bartscherer, K. (2013). A Pitx transcription factor controls the establishment and maintenance of the serotonergic lineage in planarians. *Development*, *140*(22), 4499-4509. https://doi.org/10.1242/dev.100081

McCusker, C., Bryant, S. V., & Gardiner, D. M. (2015). The axolotl limb blastema: cellular and molecular mechanisms driving blastema formation and limb regeneration in tetrapods. *Regeneration (Oxf)*, *2*(2), 54-71. https://doi.org/10.1002/reg2.32

Mikeladze-Dvali, T., Wernet, M. F., Pistillo, D., Mazzoni, E. O., Teleman, A. A., Chen, Y. W., Cohen, S., & Desplan, C. (2005). The growth regulators warts/lats and melted interact in a bistable loop to specify opposite fates in Drosophila R8 photoreceptors. *Cell*, *122*(5), 775-787. https://doi.org/10.1016/j.cell.2005.07.026

Molina, M. D., Saló, E., & Cebria, F. (2007). The BMP pathway is essential for re-specification and maintenance of the dorsoventral axis in regenerating and intact planarians. *Developmental Biology*, *311*(1), 79-94. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17905225

Molinaro, A. M., & Pearson, B. J. (2016). In silico lineage tracing through single cell transcriptomics identifies a neural stem cell population in planarians. *Genome Biol*, *17*, 87. https://doi.org/10.1186/s13059-016-0937-9

Molyneaux, B. J., Arlotta, P., Menezes, J. R., & Macklis, J. D. (2007). Neuronal subtype specification in the cerebral cortex. *Nat Rev Neurosci*, *8*(6), 427-437. https://doi.org/10.1038/nrn2151

Morgan, T. H. (1898). Experimental studies of the regeneration of Planaria maculata. *Archiv für Entwicklungsmechanik der Organismen*, *7*, 364-397.

Needham, A. E. (1952). *Regeneration of wound-healing*. London, Methuen & Co.

Newmark, P. A., Wang, Y., & Chong, T. (2008). Germ cell specification and regeneration in planarians. *Cold Spring Harb Symp Quant Biol*, *73*, 573-581. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=19022767

Okamoto, K., Takeuchi, K., & Agata, K. (2005). Neural projections in planarian brain revealed by fluorescent dye tracing. *Zoolog Sci*, *22*(5), 535-546. https://doi.org/10.2108/zsj.22.535

Orii, H., Ito, H., & Watanabe, K. (2002). Anatomy of the planarian Dugesia japonica I. The muscular system revealed by antisera against myosin heavy chains. *Zoolog Sci*, *19*(10), 1123-1131. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12426474

Orii, H., & Watanabe, K. (2007). Bone morphogenetic protein is required for dorso-ventral patterning in the planarian *Dugesia japonica*. *Dev Growth Differ*, *49*(4), 345-349. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17501910

Pack, G. T., Islami, A. H., Hubbard, J. C., & Brasfield, R. D. (1962). Regeneration of human liver after major hepatectomy. *Surgery*, *52*, 617-623. https://www.ncbi.nlm.nih.gov/pubmed/14483060

Pellettieri, J., Fitzgerald, P., Watanabe, S., Mancuso, J., Green, D. R., & Sánchez Alvarado, A. (2010). Cell death and tissue remodeling in planarian regeneration. *Developmental Biology*, *338*(1), 76-85. https://doi.org/S0012-1606(09)01193-2 [pii]
10.1016/j.ydbio.2009.09.015

Pellettieri, J., & Sánchez Alvarado, A. (2007). Cell turnover and adult tissue homeostasis: from humans to planarians. *Annu Rev Genet*, *41*, 83-105. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18076325

Petersen, C. P., & Reddien, P. W. (2008). *Smed-betacatenin-1* is required for anteroposterior blastema polarity in planarian regeneration. *Science*, *319*(5861), 327-330. https://doi.org/1149943 [pii]
10.1126/science.1149943

Petersen, C. P., & Reddien, P. W. (2009a). Wnt signaling and the polarity of the primary body axis. *Cell*, *139*(6), 1056-1068. https://doi.org/S0092-8674(09)01493-7 [pii]
10.1016/j.cell.2009.11.035

Petersen, C. P., & Reddien, P. W. (2009b). A wound-induced Wnt expression program controls planarian regeneration polarity. *Proc Natl Acad Sci U S A*, *106*(40), 17061-17066. https://doi.org/0906823106 [pii]
10.1073/pnas.0906823106
Petersen, C. P., & Reddien, P. W. (2011). Polarized *notum* activation at wounds inhibits Wnt function to promote planarian head regeneration. *Science*, *332*(6031), 852-855. https://doi.org/332/6031/852 [pii]
10.1126/science.1202143
Petersson, E., Karlsson, L., Ragnarsson, B., Bryntesson, M., Berglund, A., Stridsman, S., & Jonsson, S. (2013). Fin erosion and injuries in relation to adult recapture rates in cultured smolts of Atlantic salmon and brown trout. *Canadian Journal of Fisheries and Aquatic Sciences*, *70*(6), 915-921. https://doi.org/10.1139/cjfas-2012-0247
Pfefferli, C., & Jazwinska, A. (2015). The art of fin regeneration in zebrafish. *Regeneration (Oxf)*, *2*(2), 72-83. https://doi.org/10.1002/reg2.33
Plass, M., Solana, J., Wolf, F. A., Ayoub, S., Misios, A., Glazar, P., Obermayer, B., Theis, F. J., Kocks, C., & Rajewsky, N. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, *360*(6391), 875. https://doi.org/10.1126/science.aaq1723
Ramón y Cajal, S. (1928). *Degeneration and regeneration of the nervous system*. Clarendon Press.
Raz, A. A., Wurtzel, O., & Reddien, P. W. (2021). Planarian stem cells specify fate yet retain potency during the cell cycle. *Cell Stem Cell*. https://doi.org/10.1016/j.stem.2021.03.021
Reddien, P. W. (2018). The Cellular and Molecular Basis for Planarian Regeneration. *Cell*, *175*(2), 327-345. https://doi.org/10.1016/j.cell.2018.09.021
Reddien, P. W. (2019). The cells of regeneration. *Science*, *365*(6451), 314-316. https://doi.org/10.1126/science.aay3660
Reddien, P. W. (2021). Principles of regeneration revealed by the planarian eye. *Curr Opin Cell Biol*, *73*, 19-25. https://doi.org/10.1016/j.ceb.2021.05.001
Reddien, P. W. (2022). Positional Information and Stem Cells Combine to Result in Planarian Regeneration. *Cold Spring Harb Perspect Biol*, *14*(4). https://doi.org/10.1101/cshperspect.a040717
Reddien, P. W., Bermange, A. L., Kicza, A. M., & Sánchez Alvarado, A. (2007). BMP signaling regulates the dorsal planarian midline and is needed for asymmetric regeneration. *Development*, *134*(22), 4043-4051. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17942485
Reddien, P. W., Oviedo, N. J., Jennings, J. R., Jenkin, J. C., & Sánchez Alvarado, A. (2005). SMEDWI-2 is a PIWI-like protein that regulates planarian stem cells. *Science*, *310*, 1327-1330.
Rink, J. C. (2018). Stem Cells, Patterning and Regeneration in Planarians: Self-Organization at the Organismal Scale. *Methods Mol Biol*, *1774*, 57-172. https://doi.org/10.1007/978-1-4939-7802-1_2

Roberts-Galbraith, R. H., Brubacher, J. L., & Newmark, P. A. (2016). A functional genomics screen in planarians reveals regulators of whole-brain regeneration. *Elife*, *5*. https://doi.org/10.7554/eLife.17002

Roberts-Galbraith, R. H., & Newmark, P. A. (2013). Follistatin antagonizes Activin signaling and acts with Notum to direct planarian head regeneration. *Proc Natl Acad Sci U S A*, *110*(4), 1363-1368. https://doi.org/10.1073/pnas.1214053110

Ross, K. G., Currie, K. W., Pearson, B. J., & Zayas, R. M. (2017). Nervous system development and regeneration in freshwater planarians. *Wiley Interdiscip Rev Dev Biol*, *6*(3). https://doi.org/10.1002/wdev.266

Ross, K. G., Molinaro, A. M., Romero, C., Dockter, B., Cable, K. L., Gonzalez, K., Zhang, S., Collins, E. S., Pearson, B. J., & Zayas, R. M. (2018). SoxB1 activity regulates sensory neuron regeneration, maintenance, and function in planarians. *Dev Cell*, *47*(3), 331-347 e335. https://doi.org/10.1016/j.devcel.2018.10.014

Rowitch, D. H., & Kriegstein, A. R. (2010). Developmental genetics of vertebrate glial-cell specification. *Nature*, *468*(7321), 214-222. https://doi.org/10.1038/nature09611

Saló, E., & Baguñà, J. (1984). Regeneration and pattern formation in planarians. I. The pattern of mitosis in anterior and posterior regeneration in *Dugesia (G) tigrina*, and a new proposal for blastema formation. *J. Embryol. Exp. Morphol.*, *83*, 63-80.

Scimone, M. L., Atabay, K. D., Fincher, C. T., Bonneau, A. R., Li, D. J., & Reddien, P. W. (2020). Muscle and neuronal guidepost-like cells facilitate planarian visual system regeneration. *Science*, *368*(6498). https://doi.org/10.1126/science.aba3203

Scimone, M. L., Cote, L. E., & Reddien, P. W. (2017). Orthogonal muscle fibres have different instructive roles in planarian regeneration. *Nature*, *551*(7682), 623-628. https://doi.org/10.1038/nature24660

Scimone, M. L., Cote, L. E., Rogers, T., & Reddien, P. W. (2016). Two FGFRL-Wnt circuits organize the planarian anteroposterior axis. *eLife*, *5*. https://doi.org/10.7554/eLife.12845

Scimone, M. L., Kravarik, K. M., Lapan, S. W., & Reddien, P. W. (2014). Neoblast Specialization in Regeneration of the Planarian *Schmidtea mediterranea*. *Stem cell reports*, *3*(2), 339-352. https://doi.org/10.1016/j.stemcr.2014.06.001

Scimone, M. L., Lapan, S. W., & Reddien, P. W. (2014). A *forkhead* transcription factor is wound-induced at the planarian midline and required for anterior pole regeneration [Research Support, N.I.H., Extramural

Research Support, Non-U.S. Gov't]. *PLoS genetics*, *10*(1), e1003999. https://doi.org/10.1371/journal.pgen.1003999

Scimone, M. L., Srivastava, M., Bell, G. W., & Reddien, P. W. (2011). A regulatory program for excretory system regeneration in planarians. *Development*, *138*(20), 4387-4398. https://doi.org/138/20/4387 [pii]

10.1242/dev.068098

Scimone, M. L., Wurtzel, O., Malecek, K., Fincher, C. T., Oderberg, I. M., Kravarik, K. M., & Reddien, P. W. (2018). foxF-1 Controls Specification of Non-body Wall Muscle and Phagocytic Cells in Planarians. *Curr Biol*, *28*(23), 3787-3801 e3786. https://doi.org/10.1016/j.cub.2018.10.030

Srivastava, M. (2021). Beyond Casual Resemblance: Rigorous Frameworks for Comparing Regeneration Across Species. *Annu Rev Cell Dev Biol*, *37*, 415-440. https://doi.org/10.1146/annurev-cellbio-120319-114716

Stückemann, T., Cleland, J. P., Werner, S., Thi-Kim Vu, H., Bayersdorf, R., Liu, S. Y., Friedrich, B., Julicher, F., & Rink, J. C. (2017). Antagonistic Self-Organizing Patterning Systems Control Maintenance and Regeneration of the Anteroposterior Axis in Planarians. *Dev Cell*, *40*(3), 248-263 e244. https://doi.org/10.1016/j.devcel.2016.12.024

Sulston, J. E., Schierenberg, E., White, J. G., & Thomson, J. N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. *Developmental Biology*, *100*(1), 64-119. https://doi.org/10.1016/0012-1606(83)90201-4

Sureda-Gómez, M., Martín-Durán, J. M., & Adell, T. (2016). Localization of planarian beta-CATENIN-1 reveals multiple roles during anterior-posterior regeneration and organogenesis. *Development*, *143*(22), 4149-4160. https://doi.org/10.1242/dev.135152

Sureda-Gómez, M., Pascual-Carreras, E., & Adell, T. (2015). Posterior Wnts Have Distinct Roles in Specification and Patterning of the Planarian Posterior Region. *Int J Mol Sci*, *16*(11), 26543-26554. https://doi.org/10.3390/ijms161125970

Tazaki, A., Kato, K., Orii, H., Agata, K., & Watanabe, K. (2002). The body margin of the planarian Dugesia japonica: characterization by the expression of an intermediate filament gene. *Dev Genes Evol*, *212*(8), 365-373. https://doi.org/10.1007/s00427-002-0253-0

Tewari, A. G., Stern, S. R., Oderberg, I. M., & Reddien, P. W. (2018). Cellular and Molecular Responses Unique to Major Injury Are Dispensable for Planarian Regeneration. *Cell Reports*, *25*(9), 2577-2590 e2573. https://doi.org/10.1016/j.celrep.2018.11.004

Thornton, C. S. (1968). Amphibian limb regeneration. *Adv Morphog*, *7*, 205-249. https://doi.org/10.1016/b978-1-4831-9954-2.50010-0

van Wolfswinkel, J. C., Wagner, D. E., & Reddien, P. W. (2014). Single-Cell Analysis Reveals Functionally Distinct Classes within the Planarian Stem Cell Compartment. *Cell Stem Cell*, *15*(3), 326-339. https://doi.org/10.1016/j.stem.2014.06.007

Vásquez-Doorman, C., & Petersen, C. P. (2014). *zic-1* Expression in planarian neoblasts after injury controls anterior pole regeneration [Research Support, N.I.H., Extramural
Research Support, Non-U.S. Gov't]. *PLoS genetics*, *10*(7), e1004452. https://doi.org/10.1371/journal.pgen.1004452

Vogg, M. C., Owlarn, S., Perez Rico, Y. A., Xie, J., Suzuki, Y., Gentile, L., Wu, W., & Bartscherer, K. (2014). Stem cell-dependent formation of a functional anterior regeneration pole in planarians requires Zic and Forkhead transcription factors [Research Support, Non-U.S. Gov't]. *Developmental Biology*, *390*(2), 136-148. https://doi.org/10.1016/j.ydbio.2014.03.016

Wagner, D. E., Wang, I. E., & Reddien, P. W. (2011). Clonogenic neoblasts are pluripotent adult stem cells that underlie planarian regeneration. *Science*, *332*(6031), 811-816. https://doi.org/332/6031/811 [pii]
10.1126/science.1203983

Waller, A. (1851). Experiments on the Section of the Glosso-Pharyngeal and Hypoglossal Nerves of the Frog, and Observations of the Alterations Produced Thereby in the Structure of Their Primitive Fibres. *Edinb Med Surg J*, *76*(189), 369-376. https://www.ncbi.nlm.nih.gov/pubmed/30332247

Wang, I. E., Lapan, S. W., Scimone, M. L., Clandinin, T. R., & Reddien, P. W. (2016). Hedgehog signaling regulates gene expression in planarian glia. *Elife*, *5*. https://doi.org/10.7554/eLife.16996

Watson E.R., M. A., Fard A.T., Mar J.C. (2022). How does the structure of data impact cell–cell similarity? Evaluating how structural properties influence the performance of proximity metrics in single cell RNA-seq data *Briefings in Bioinformatics*, *23*(6).

Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D., & Klein, A. M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science*, *367*(6479). https://doi.org/10.1126/science.aaw3381

Weismann, A., Parker, W. N., & Rönnfeldt, H. (1893). *The germ-plasm : a theory of heredity*. Scribner's.

Wenemoser, D., Lapan, S. W., Wilkinson, A. W., Bell, G. W., & Reddien, P. W. (2012). A molecular wound response program associated with regeneration initiation in planarians. *Genes & Development*, *26*(9), 988-1002. https://doi.org/10.1101/gad.187377.112

Wenemoser, D., & Reddien, P. W. (2010). Planarian regeneration involves distinct stem cell responses to wounds and tissue absence. *Developmental Biology*, *344*(2), 979-991. https://doi.org/S0012-1606(10)00837-7 [pii]
10.1016/j.ydbio.2010.06.017

Witchley, J. N., Mayer, M., Wagner, D. E., Owen, J. H., & Reddien, P. W. (2013). Muscle cells provide instructions for planarian regeneration [Research Support, N.I.H., Extramural
Research Support, Non-U.S. Gov't]. *Cell Reports*, *4*(4), 633-641. https://doi.org/10.1016/j.celrep.2013.07.022

Wodarz, A., & Nusse, R. (1998). Mechanisms of Wnt signaling in development. *Annu Rev Cell Dev Biol*, *14*, 59-88. https://doi.org/10.1146/annurev.cellbio.14.1.59

Wurtzel, O., Cote, L. E., Poirier, A., Satija, R., Regev, A., & Reddien, P. W. (2015). A Generic and Cell-Type-Specific Wound Response Precedes Regeneration in Planarians. *Dev Cell*, *35*(5), 632-645. https://doi.org/10.1016/j.devcel.2015.11.004

Wurtzel, O., Oderberg, I. M., & Reddien, P. W. (2017). Planarian Epidermal Stem Cells Respond to Positional Cues to Promote Cell-Type Diversity. *Dev Cell*, *40*(5), 491-504 e495. https://doi.org/10.1016/j.devcel.2017.02.008

Xiang, R., Wang, W., Yang, L., Wang, S., Xu, C., & Chen, X. (2021). A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data. *Front Genet*, *12*, 646936. https://doi.org/10.3389/fgene.2021.646936

Zechner, C., Nerli, E., & Norden, C. (2020). Stochasticity and determinism in cell fate decisions. *Development*, *147*(14). https://doi.org/10.1242/dev.181495

Zeng, A., Li, H., Guo, L., Gao, X., McKinney, S., Wang, Y., Yu, Z., Park, J., Semerad, C., Ross, E., Cheng, L. C., Davies, E., Lei, K., Wang, W., Perera, A., Hall, K., Peak, A., Box, A., & Sanchez Alvarado, A. (2018). Prospectively Isolated Tetraspanin(+) Neoblasts Are Adult Pluripotent Stem Cells Underlying Planaria Regeneration. *Cell*, *173*(7), 1593-1608 e1520. https://doi.org/10.1016/j.cell.2018.05.006

Zweifel, L. S., Kuruvilla, R., & Ginty, D. D. (2005). Functions and mechanisms of retrograde neurotrophin signalling. *Nat Rev Neurosci*, *6*(8), 615-625. https://doi.org/10.1038/nrn1727

Chapter 2

A transcription factor atlas of stem cell fate specification in planarians

A transcription factor atlas of stem cell fate specification in planarians

Hunter O. King[2,4†], Kwadwo E. Owusu-Boaitey[2,3,5,†], Christopher T. Fincher[2,3], and Peter W. Reddien[1,2,3,*]

[1]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

[2]Whitehead Institute for Biomedical Research, Cambridge, MA 02142, USA.

[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

[4]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

[5]Harvard/MIT MD-PhD, Harvard Medical School, Boston, MA 02115, USA.

*Corresponding author. Email: reddien@wi.mit.edu

[†]These authors contributed equally to this work

Whole-body regeneration requires the ability to produce the full repertoire of adult cell types. The planarian *Schmidtea mediterranea* contains roughly 150 distinct cell types, which can be regenerated from a stem cell population called neoblasts. Neoblast fate choice can be regulated by the expression of fate-specific transcription factors (FSTFs). How fate choices are made for many cell types and distributed across neoblasts versus their post-mitotic progeny remains unclear. We used single-cell RNA-sequencing (scRNA-seq) to systematically map fate choices made in S/G_2/M neoblasts and, separately, in their post-mitotic progeny that serve as progenitors. We defined transcription factor expression signatures associated with all detected fate choices, identifying numerous new progenitor classes and FSTFs that regulate them. Our work generates an atlas of stem cell fate-specification paths with associated transcription factor signatures for most cell types in a complete adult organism.

Introduction

Adult stem cells in many regenerative organisms must be capable of making a large array of possible fate choices. Planarians take this challenge to the extreme, where stem cells and/or their post-mitotic descendant cells must choose between over a hundred possible fates (Fincher et al., 2018). Planarians are a powerful model for studying how adult stem cells generate cell-type diversity (Reddien, 2018). Planarians can regenerate any missing body part, including the entire animal from small body fragments and also constitutively replace all tissues throughout the body during homeostatic tissue turnover. Adult stem cells called neoblasts are responsible for the production of all adult somatic cell types in both regeneration and homeostatic tissue turnover (Baguñà et al., 1989; Wagner et al., 2011).

In many developing animals, the enormous challenge of generating dozens to hundreds of different cell types is accomplished gradually and can involve progressive fate restriction, sometimes with a long lineage hierarchy (Sulston et al., 1983). In regeneration, by contrast, new tissues can return quickly and concurrently (King and Newmark, 2012; Reddien, 2018). For instance, in planarians, new cell types can differentiate to replace missing tissues within three days of amputation despite cell division of neoblasts taking 12-24 hours (Newmark and Sánchez Alvarado, 2000; Wenemoser and Reddien, 2010; Reddien, 2018). Furthermore, the identity of missing cell types at wounds is unpredictable, potentially making a strict hierarchical lineage from a multipotent progenitor impractical for tailoring production to the identity of missing cell types. Therefore, addressing the mechanisms of fate choice specification in stem and progenitor cells is central for understanding regeneration.

Substantial data indicate that neoblasts are a major site of fate specification in planarians (Reddien, 2018). Neoblasts are molecularly and functionally heterogenous, being comprised of distinct subsets called specialized neoblasts (Adler et al., 2014; Scimone et al., 2014a; Zhu and Pearson, 2016; Ivankovic et al., 2019). Specialized neoblasts express transcription factors associated with particular fates, termed fate-specific transcription factors

(FSTFs) (Raz et al., 2021). Many FSTFs are required for the formation of different mature cell types (Lapan and Reddien, 2011; Scimone et al., 2011; Lapan and Reddien, 2012; Wenemoser et al., 2012; Cowles et al., 2013; Currie and Pearson, 2013; März et al., 2013; Adler et al., 2014; Cowles et al., 2014; Scimone et al., 2014a; Scimone et al., 2014b; van Wolfswinkel et al., 2014; Vásquez-Doorman and Petersen, 2014; Vogg et al., 2014; Molinaro and Pearson, 2016; Roberts-Galbraith et al., 2016; Wang et al., 2016; He et al., 2017; Ross et al., 2017; Scimone et al., 2017; Fincher et al., 2018; Ross et al., 2018; Scimone et al., 2018). Specialized neoblasts, therefore, serve as the precursors for the mature cell types of the animal. FSTF expression is a frequent feature of $S/G_2/M$ neoblasts and neoblast specification can commonly occur as cells transit through the cell cycle (Raz et al., 2021). Previous work supports a non-hierarchical model for the neoblast lineage called the single-step fate model, in which neoblasts can specify one of a diverse set of possible fates during the cell cycle and can divide with an asymmetric outcome with a neoblast descendant able to specify a new fate, thus retaining potency (Raz et al., 2021).

Comprehensive adult planarian single-cell RNA sequencing (scRNA-seq) data estimated that ~150 distinct mature cell types/cell states exist in adult planarians (Fincher et al., 2018). Single-cell RNA sequencing (scRNA-seq) has also recently made it possible to identify neoblast subtypes as clusters corresponding to the major tissue classes of the animal (Fincher et al., 2018; Plass et al., 2018; Zeng et al., 2018; Niu et al., 2021). However, little additional structure has been described for such stem cell clusters, leaving unclear the scope of specification that happens in neoblasts. For instance, are all ~150 fate decisions that must occur in the adult made in the neoblasts? Alternatively, are some or even many fate choices made in post-mitotic neoblast descendant cells that serve as migratory progenitors? What transcription factor (TF) signatures and dynamics characterize and regulate the generation of cell-type diversity as progenitors mature and potentially diversify over time? To address these questions, we performed scRNA-seq on dividing neoblasts ($S/G_2/M$) and G_0 post-mitotic neoblast progeny during regeneration and identified novel cell types and TF expression signatures in these

stages. We find differences across several major tissue types in how cell type diversity is generated. We assembled a curated catalog of all putative planarian transcription factors, and utilized it in conjunction with scRNA-seq data to uncover new FSTFs and specialized neoblast populations. Our results generate a stem cell fate-specification atlas for the majority of cell types of a complete adult animal.

Results

Single-cell RNA-sequencing of S/G_2/M neoblasts identifies diverse specialized neoblast classes

Recent work showed that FSTF expression is high in S/G_2/M neoblasts, suggesting that the complete set of neoblast fate choices could be investigated in the 4C neoblast stage (Raz et al., 2021). S/G_2/M neoblasts can be isolated with Hoechst labeling and fluorescence-activated cell sorting (FACS) based on their 4C DNA content (the X1 gate) (Figure 1A) (Hayashi et al., 2006). We purified 4C neoblasts 72 hours after amputation that removed either the anterior or posterior, yielding neoblasts involved in anterior and posterior regeneration, respectively (Figure 1A). This time point involves substantial neoblast amplification and reestablishment of zones of positional information that influence regeneration outcomes (Reddien, 2022). 23,873 S/G_2/M neoblasts that formed 12 distinct clusters were analyzed (Figure 1B).

Planarian tissues can be categorized into nine major classes: neoblasts, epidermis, muscle, neurons, intestine, protonephridia (a filtration organ system), phagocytic/*cathepsin*+ cells, glandular/parenchymal cells, and pharynx (feeding tube) (Fincher et al., 2018; Plass et al., 2018). Within each major tissue class there exists numerous cell types with distinct features and functional roles. We clustered single-cell neoblast transcriptomes and found neoblasts for all eight major differentiated planarian tissue types were represented by at least one large cluster (Figure 1B, 1C; Supplemental Figure 1A). Specialized neoblasts for rare cell types, such as

photoreceptor neurons, were present in the $S/G_2/M$ neoblast transcriptomes, indicating that even the rarest specialized neoblast classes should be present in this dataset.

Because multiple FSTFs often characterize a cell type, we hypothesized that ensembles of correlated transcription factors could be used to identify new cell types or states as well as to identify FSTFs associated with previously known or new states. We computed pairwise correlation values between previously identified FSTFs across all sequenced 4C neoblasts and performed hierarchical clustering to group correlated/coexpressed FSTFs into modules (Figure 1D). FSTFs known to be associated with fate specification of the same tissue type from prior work displayed correlated expression and clustered together into modules representing the different tissues they mutually specify. We then identified cells coexpressing components of the transcription factor (TF) ensembles to annotate regions of cells in the scRNA-seq UMAP visualizations as associated with distinct fates (Figure 1E). This approach also identified spatial regions containing specialized neoblasts for the intestine, epidermis, protonephridia, parenchyma, muscle, eyes, and neurons. By identifying cells through the shared expression of a module of transcription factors instead of individual canonical transcription factors, cells could more precisely be classified to a specific cell type, because many individual transcription factors were expressed in multiple cell types. For example, epidermal-specialized neoblasts express the FSTFs *zfp-1*, *p53*, and *soxP-3*, but cells expressing each individual gene were not all epidermis fated (70%, 67%, 86% were epidermal-fated, respectively; Supplemental Figure 1B-D). However, 97% of cells expressing all three FSTFs were epidermis fated. This approach is applied broadly below in ascribing identity to cells and for identifying FSTFs associated with particular neoblast states.

Defining the transcriptomes of post-mitotic progenitor cells with scRNA-seq

4C neoblasts produce post-mitotic (post-neoblast) G_0 cells that are progenitors for differentiated cell types. Whether all cell-fate decisions are made in the neoblasts or whether decisions

continue to be made in post-neoblast G_0 stages is unknown. Investigating these G_0 cells is therefore critical for understanding how the complete complement of ~150 adult cell fate choices for planarian anatomy emerge. We used FACS to isolate G_1/G_0 cells (the X2 gate) (Hayashi et al., 2006) during head and tail regeneration (Figure 1A). G_0 cells were computationally isolated from G_1 cells as previously described (Raz et al, 2021). 31,084 G_0 cells resulted in 26 cell clusters (Figure 1F), which broadly represented progenitors for the known tissue types of the animal (Figure 1G). FSTFs and differentiated cell-type-specific markers were often co-expressed in distinct subsets of cells within G_0 clusters (Supplemental Figure 1E). Furthermore, post-mitotic progenitors for rare cell types such as photoreceptor neurons were present, suggesting that this dataset, like the neoblast dataset, contains even the rarest progenitor classes.

In silico identification of planarian transcription factors

Expression of transcription factors is one of the earliest features that distinguishes classes of progenitors. We therefore generated an *in silico* catalog of all putative planarian transcription factors (Supplementary Figure 1F) and utilized them to enable the identification of neoblast types and the transcription factors that define them. Additional cataloging of planarian transcription factors has recently been undertaken (Neiro et al., 2022). We first searched the translated planarian transcriptome for genes encoding known eukaryotic DNA-binding domain (DBD) motifs with 67 Pfam DNA-binding domain models using HMMER (Eddy, 2011; Wheeler and Eddy, 2013). This identified 1,317 transcripts representing 970 unique genes with at least one potential DBD. Many genes in this set were previously characterized, with sequences deposited in GenBank. To classify uncharacterized genes, we performed BLAST analysis against the human proteome. Genes with significant similarity to non-TFs, such as other classes of DNA-binding proteins (e.g., histones, transposons) were removed, resulting in 716 unique

genes with domains resembling characterized DBDs or that have been previously identified in planarians.

Within our DBD-containing catalog, 77 genes (11%) were previously classified as FSTFs in planarians and another 102 (14%) have been studied in other contexts in planarians (Supplementary Figure 1F). The remaining candidate TF genes consist of those with BLAST hits to human TFs (242 genes; 34%) and genes with no BLAST hit to any human protein (295 genes; 41%). Among known planarian FSTFs, homeodomains, zinc fingers, and forkhead domains collectively made up the majority of domains within these gene products (63% total; 33%, 19%, and 11% respectively) (Figure 1H). Among known planarian TFs with no characterized FSTF function, the homeodomain, zinc finger, and HLH domains were the most abundant (15%, 18%, and 28%, respectively). Similarly, non-characterized genes with clear blast hits to human TFs consisted of many homeodomain-containing proteins and zinc finger-containing proteins (28% and 30%, respectively).

We calculated the expression specificity of each putative transcription factor for each major tissue type in the neoblast data (Supplementary Figure 1G). In our TF catalog, homeodomain transcription factor-encoding genes were the most represented family among known FSTFs and many were expressed in neoblasts and enriched in different progenitor types (Figure 1H). Previous work has shown that the complete set of all 118 *C. elegans* neuron classes can be delineated by unique combinations of homeodomain protein expression (Reilly et al., 2020). Most tissue type progenitor classes had enriched expression for at least one homeodomain transcription factor (Supplemental Figure 1H), except for pharynx-specialized neoblasts, which displayed highly enriched expression for only the forkhead transcription factors *foxD* and *foxA* (Supplementary Figure 1G), overall indicating that neoblast classes could also be delineated by their expression of transcription factors.

Muscle cell-type diversity emerges in the S/G_2/M neoblast stage

To compare the degree of cell-type specification in the S/G_2/M (4C) state and in the post-mitotic progenitor stage, we performed subclustering of cells by tissue type. We first examined the three clusters expressing *PLOD1*, which is known to be expressed in muscle and phagocyte/*cathepsin*+ precursors. Muscle progenitors are some of the best understood planarian progenitor classes (Scimone et al., 2017; Fincher et al., 2018; Scimone et al., 2018). Planarians possess a diverse array of muscle subtypes with different roles in their physiology and in regulating regeneration (Figure 2A) (Witchley et al., 2013; Cebrià, 2016; Scimone et al., 2016; Scimone et al., 2017; Scimone et al., 2018; Cote et al., 2019; Scimone et al., 2020). Body wall muscle (BWM) is comprised of longitudinal fibers oriented on the anterior-posterior (AP) axis, circular fibers oriented perpendicular to the AP axis, and diagonal fibers (Cebrià, 2016). Two dorsoventral muscle (DVM) fiber types (medial and lateral) connect dorsal and ventral animal regions (Scimone et al., 2018). A supportive set of muscle surrounds the intestine and a separate network of muscle exists within the pharynx. Finally, unique muscle-like cells at the anterior tip of the animal (the anterior pole) and at the posterior tip (the posterior pole) function as patterning organizers during regeneration (Reddien, 2018). Each muscle subtype has been shown to be specified by distinct transcription factors (Scimone et al., 2017; Scimone et al., 2018).

Subclustering of *PLOD1*+ S/G_2/M cells identified 13 total subclusters (Figure 2B) from 5,633 cells. Most subclusters had unique expression of TFs specifying the different muscle subtypes, and almost all muscle cell types were represented by their own subcluster (Figure 2C). *gata4/5/6-2* (specifying medial DVM (Scimone et al., 2018)) and *nk4* (specifying lateral DVM (Scimone et al., 2018)) expression was enriched in cluster 8 and cluster 10, respectively (Figure 2C). *nkx1-1* (specifying circular muscle (Scimone et al., 2017)) and *gata4/5/6-3* (specifying intestinal muscle (Scimone et al., 2018)) were expressed in different regions of subcluster 5. *myoD* (specifying longitudinal muscle (Scimone et al., 2017)) was expressed in two clusters (cluster 12 and cluster 2; Figure 2C). Subcluster 2 had unknown fate and displayed

enriched expression of multiple distinct TFs, including *hesl-3*, *hesl-2*, *musculin*, *myoD*, *nk4*, and *hnf4* (Supplementary Figure 2A). Clusters 6, 7, 9, and 11 collectively expressed markers associated with pharynx muscle (Figure 2C) (Fincher et al., 2018). Each pharynx muscle cluster also expressed separate TF sets, including *musculin*, *dmd-3*, *ap2*, *coe*, *sim*, *foxC1*, and *zicA*]. Anterior pole cells (defined by coexpression of *foxD* and *prep*) were also present, although not as their own unique cluster (Supplementary Figure 2A). *foxF-1* is expressed broadly across non-BW muscle (Scimone et al., 2018) and, as expected, was expressed across multiple regions, including the clusters specifying pharynx muscle and other non-BW muscle (Scimone et al., 2018) (Figure 2C). Expression of positional control genes (PCGs), which constitute positional information and are primarily expressed in muscle cells (Witchley et al., 2013), was detectable in different muscle specialized neoblasts in the data (Supplemental Figure 2A). Finally, we used the *in silico*-generated transcription factor catalog to define transcription factor modules for different muscle subtypes (Figure 2D and 2E). Together, these data define transcriptomes and TF modules for unique clusters of $S/G_2/M$ cells with fates specified to essentially each of the different muscle subtypes of the animal.

We subclustered 2,823 post-mitotic G_0 muscle progenitor cells, revealing 17 distinct clusters (Figure 2F). Most subclusters displayed enriched expression in both cell-type specific FSTFs and post-mitotic markers (Figure 2G and Supplemental Figure 2E). Cluster 4 and 12 represented DVM and were enriched in *nk4* and *gata4/5/6-2*. Cluster 13 represented intestinal muscle and was enriched in expression for *gata4/5/6-3* and the known intestinal muscle marker *PTPRD* (Scimone et al., 2018). Clusters 0, 1, 6, and 15 expressed the known body-wall muscle marker *bwm-1*, with cluster 0 being enriched for circular fiber marker expression and clusters 1 and 15 being enriched for longitudinal fiber marker expression (Figure 2G). Cluster 15 also expressed the midline and dorsal PCGs *slit* and *bmp4*. Clusters 3, 8, and 10 collectively represented pharynx muscle. Finally, a distinct cluster of anterior pole cells could be identified in cluster 11 based on the enriched expression of *foxD*, *prep*, and *zicA*, along with PCGs known to

be expressed in the anterior pole (e.g., *sFRP-1*) (Supplemental Figure 2E). These results are consistent with known muscle cell-type specification events that indicate distinct fates emerge at the outset of production from the earliest possible cellular stage (the neoblast, stem cell stage), without intermediate branches of fate refinement.

Lack of neoblast diversity of *cathepsin*$^+$ cell populations

Subclustering of the *PLOD1*+ S/G_2/M cells also identified a single population of cells (cluster 4) putatively specifying the phagocytic/*cathepsin*+ cells of the animal (Figure 2B). The phagocytic group of planarian cells is comprised of a diverse population of cell types (at least 9) with broad phagocytic capacity and complex morphologies, including pigment and glial cells (Fincher et al., 2018) (Figure 2H; adapted from Fincher et al. 2018). *foxF-1* is broadly expressed across phagocytic cells and is required for specifying many of these cell types (Scimone et al., 2018). *ets-1* promotes pigment cell specification (He et al., 2017), and its expression is broad throughout the phagocytic population as well (Fincher et al., 2018), suggesting that it could regulate multiple phagocytic subtypes. *hnf4* is also expressed in neoblasts expressing phagocytic cell markers (Fincher et al., 2018). S/G_2/M cells associated with the phagocytic state were enriched in expression for these three TF-encoding genes (*foxF-1*, *ets-1*, and *hnf4*), along with *zfp-1* and *hunchback* (Figure 2C and Supplemental Figure 2A). Expression of these genes was broad throughout cluster 4, without overt substructure. Subclustering generated seven not well-separated clusters and some of these cells expressed markers for phagocytic cell subtypes (Supplemental Figure 2B-D). However, these markers largely had expression scattered broadly across subclusters.

Subclustering of the post-mitotic G_0 phagocytic progenitor cells identified clusters more specifically enriched in cell-type-specific marker expression, consistent with these clusters representing progenitors for several of the known differentiated phagocytic cell subtypes (Figure 2I and 2J, Supplemental Figure 2F). However, not all phagocytic cell-type diversity was

apparent in the G_0 post-neoblast progenitors either. These findings raise the possibility that diversification of states for *cathepsin*+ cells might not occur until after the $S/G_2/M$ stage in post-mitotic progenitors, or is not clearly represented by distinct transcriptomes. These findings are in stark contrast with muscle progenitors, for which transcriptome heterogeneity matching the different classes of differentiated muscle was readily identified in the $S/G_2/M$ neoblast stage.

Parenchymal progenitor fate diversity emerges in neoblasts

The parenchymal cell population is a highly heterogenous collection of distinct cell types scattered within a mesenchymal tissue surrounding major organs, and includes gland cell types (Hyman, 1951; Fincher et al., 2018; Plass et al., 2018). Fate specification of these cells is poorly understood. Recent scRNA-seq data identified 13 distinct differentiated parenchymal cell clusters, associated with cell types with diverse spatial distributions (Fincher et al., 2018). In Fincher et al. 2018, an additional eight clusters of putative transition states for different parenchymal subtypes were also identified, with enriched TF expression (Figure 3A; adapted from Fincher et al. 2018). It was unclear from this prior work whether these transition states represented neoblasts or post-mitotic neoblast progeny.

We found two TF-encoding genes, *RREB1* and *RREB2*, that had abundant and enriched expression across mature parenchymal tissues and that were expressed in two $S/G_2/M$ neoblast clusters (Figure 1C). These clusters displayed mutually exclusive expression of the transcription factors *NKX2-4* and *foxA*. *foxA* is required for pharynx regeneration (Adler et al., 2014) and *foxA*+ neoblasts include pharynx progenitors in a broad domain surrounding the pharynx (Adler et al., 2014; Scimone et al., 2014a). Two $S/G_2/M$ populations expressed *foxA* (Figure 1C), one parenchymal cell-associated and the other being pharynx neoblasts.

We subclustered the 1,018 *foxA*+ 4C parenchymal neoblasts and the 647 *NKX2-4*+ 4C parenchymal neoblasts separately (Figure 3B & Figure 3C). Multiple subclusters displayed enriched expression for TFs specific to distinct differentiated parenchymal subtypes (Figure 3D,

3E). *foxA*+ parenchymal neoblasts produced five subclusters. Cluster 3 had enriched expression of *fer3l-1*, which is expressed in putative precursors for differentiated *SSPO* (dd_9342) cells (Fincher et al., 2018). Cluster 1 had enriched expression of *fer3l-2*, a TF-encoding gene expressed in the putative precursor population for *mag-1*+ cells (Fincher et al., 2018). Cluster 1 also had enriched expression of *ptf-1*, but in a separate domain. *ptf-1* expression was enriched in putative precursors for dd_8476 cells, a unique parenchymal population localized to the planarian head (Fincher et al., 2018). Clusters 2 and 4 expressed the non-TF-encoding dd_2 gene, which marks a parapharyngeal parenchymal population (parenchymal cluster 9 in Fincher et al. 2018). Correlation analysis among all transcription factors in the TF catalog identified several transcription factor modules corresponding to identified parenchymal neoblast types (Figure 3F and 3G). Notably, this revealed two transcription factor-encoding genes, dd_14712 and dd_17476, that were co-expressed in clusters 2 and 4 of the *foxA*+ cluster. In total, of the 13 known differentiated parenchymal cell clusters (Fincher et al., 2018), four were represented within the *foxA*+ $S/G_2/M$ parenchymal-specialized neoblast cluster.

Subclustering of the *NKX2-4*+ $S/G_2/M$ parenchymal neoblasts identified several subclusters corresponding to the other differentiated parenchymal cells (Figure 3C). Cluster 5 was enriched in the expression of *IRX1* (Figure 3E). In prior data (Fincher et al., 2018), *IRX1* expression was enriched in putative precursors for differentiated *ZAN6*+ (dd_238) cells (Figure 3A). Cluster 5 had enriched expression of the TF-encoding *GCM2* gene (Figure 3E). In Fincher et al 2018, *GCM2* expression was enriched in putative parenchymal precursors; a subset of these expressed *KCP*+ (dd_91) and another subset coexpressed *IRX1*. Finally, cluster 1 displayed enriched expression of the TF-encoding genes *nkx2-like* and *nkx6-like*. *nkx2-like* was expressed broadly throughout cluster 1 and *nkx6-like* was expressed in a subset of cluster 1 cells. *nkx2-like* was expressed in several putative parenchymal precursors, as well as three differentiated parenchymal clusters, including candidate precursors for the dd_515 population

(Figure 3A). Within the *NKX2-4*+ S/G_2/M parenchymal cell specialized neoblasts, we also found a small fraction of cells (primarily in cluster 0) that expressed the TF-encoding gene *ascl-2* (Figure 3E).

Taken together, these results suggest that during the S/G_2/M neoblast state, parenchymal cell diversity is generated through ten populations of specialized neoblasts that can be grouped into two major categories: *foxA*+ cells that can be subdivided into subsets of *fer3l-1*+, *fer3l-2*+*, ptf-1*+, and dd_2+ neoblasts and *NKX2-4*+ cells that can be subdivided into *IRX1*+, *GCM-2*+, *nkx2-like*+, *ascl-2*+, and *nkx6-like*+ subsets.

To assess how this subtype diversity was reflected in the early G_0 stage, we combined G_0 post-neoblast parenchymal progenitors (1,996 cells) and performed subclustering (Figure 3H). We found seven total clusters with six representing likely descendants of parenchymal neoblast subtypes from the 4C state (Figure 3H and Supplemental Figure 3A-D). One cluster (cluster 6) corresponded to the population of parenchymal cells marked by *GATA3* expression from Fincher et al. 2018 parenchymal clusters 6 and 10 (Figure 3A and 3I). These cells were enriched in the TF-encoding genes *GATA3*, *pax6A*, *Post-2b*, *EVX1*, and *fer3l-3* (Supplemental Figure 3A) – similar to a precursor state from Fincher et al. 2018. By contrast, no 4C parenchymal population clearly expressed the combined transcription factor signature corresponding to this population. However, low levels of *GATA3* and *Post-2b* were present sparsely in one subcluster (Supplemental Figure 3E). This raises the possibility that this parenchymal cell type might be substantially matured through TF-encoding gene expression post-mitotically.

We tested the functional requirement for parenchymal lineage-associated TFs in specifying parenchymal subtypes using RNA interference (RNAi). Because of constant cell turnover in planarians, RNAi of a TF-encoding gene required for fate specification of neoblasts can lead to steady depletion of the cell type generated by the associated neoblast class

(Reddien, 2018). Furthermore, amputation challenges the animal to produce missing cell types in a blastema. RNAi animals were subjected to head and tail amputation and were fixed after regeneration (>10 days post-amputation, dpa). Inhibition of many TFs led to reduced numbers of parenchymal cells expressing cell-type-specific markers. This could reflect depletion of those cells, or in some cases loss of marker expression without loss of the cell type. In many cases, RNAi resulted in fewer cells robustly expressing the marker gene used, rather than lower levels of marker gene expression throughout the cell population, consistent with cell loss (Figure 3J, Supplementary Figure F-L). *IRX1* RNAi and *GCM2* RNAi led to depletion of *ZAN6*+ (dd_238) and *KCP*+ (dd_91) cells, respectively. *ZAN6* (dd_238) expression marks differentiated parenchymal cluster 5 in Fincher et al. 2018 data and *KCP* (dd_91) expression marks differentiated parenchymal cluster 12. RNAi of *ASCL4* led to the depletion of dd_801+ cells, which comprise two differentiated parenchymal subtypes: cluster 12 and cluster 5 of Fincher et al. 2018 data (Figure 3J and Supplemental Figure 3J). *ascl-2* RNAi eliminated cells expressing *glipr-1* (dd_924), as did RNAi of dd_10911, which encodes a zinc finger protein enriched in *glipr-1* cells (Figure 3J and Supplemental Figure 3F, 3H). We also found that whereas *RREB1* and *RREB2* expression globally marked 4C parenchymal neoblasts, only *RREB2* was functionally required for forming any of the parenchymal subtypes tested (Figure 3J and Supplemental 3I). Finally, RNAi of *Post-2b*, which was expressed in G_0 cell-specific cluster 6 resulted in loss of dd_829-expressing parenchymal cells (Figure 3J and Supplemental Figure 3G). Overall, these results identify numerous FSTFs associated with new specialized neoblast classes and show that specification of most-to-all of the cell-type diversity of planarian parenchymal cells is independently generated in the 4C stem cell state.

Specification of fates in progenitors of the nervous system

The planarian nervous system contains more cell-type diversity than any other planarian tissue (Fincher et al., 2018). Clustering of S/G_2/M neoblasts revealed two distinct populations

containing neural progenitors; one was labeled "Neural" and a separate cluster was unified by expression of the transcription factor *six1/2-1* (Figure 1B). We subclustered the 4C "Neural" and *six1/2-1*$^{+}$ populations independently (Figure 4A, 4B). Two subclusters of the *six1/2-1* population were eye progenitors, expressing FSTFs for the eye such as *ovo* and *eya* (Figure 4B). The *six1/2-1* cluster also contained cells expressing the transcription factor-encoding *MECOM* and *dmd-1* genes, which are expressed in somatic support cells for the germline. Subclustering of the 4C "Neural" population (6,526 cells) revealed six main clusters, with known neural FSTFs expressed within them (Figure 4A). Cluster 2 had enriched expression of neural FSTFs such as *nkx6-like* (Scimone et al., 2014a) and *prox-1* (Figure 4A). Cluster 5 had enriched expression of the neural FSTFs *pou4l-1*, *neuroD-1*, and *nkx2-like* (Cowles et al., 2013; Cowles et al., 2014; Brown et al., 2018). Most other known neural FSTFs were expressed broadly across multiple of the subclusters, including *pax6A*, *ski-3*, *fli1*, *sp6-9*, and *scratch* (Scimone et al., 2014a; Molinaro and Pearson, 2016). These results indicate that at least some neural signatures emerge in the 4C neoblast state and that unique transcriptomes for the cells at these stages can be defined.

Correlation analysis between transcription factors among neural neoblast-containing clusters (“Neural” and “*six1/2-1*+”) identified several transcription factor modules (Supplemental Figure 4A) associated with cells that occupied domains in the UMAP subclusters (Supplemental Figure 4B). For example, a transcription factor-encoding gene with similarity to human *GFI1B* (dd_14824) had correlated expression with known neural FSTFs *pou4l-1* and *neuroD-1* (Supplemental Figure 4A). *GFI1B* was expressed in a small number of ciliated neuron populations from Fincher et al. 2018, such as clusters 11 and 18 (marked by dd_28465 and dd_29413, respectively) (Figure 4C). *GFI1B* RNAi led to loss of both of these differentiated neuron populations (Figure 4E). *ascl-2* (dd_14753) and dd_14712, which encodes an ETS-domain containing transcription factor, had correlated expression with each other and the neural FSTF *prox-1* (Supplementary 4A). These TF-encoding genes were expressed in a subset of non-ciliated peripheral neurons marked by dd_3069, and RNAi of *ascl-2* resulted in the loss of

peripheral neurons cells expressing dd_3069 (Figure 4E). Using scRNA-seq data from Fincher et al. 2018, we found that *IRX6* was expressed in a unique subset of non-ciliated neurons marked by expression of the neural peptide-encoding *npp-18* gene. RNAi of *IRX6* resulted in a loss of *npp-18*$^{+}$ cells (Figure 4E). *INSM2* (dd_28888) had modestly correlated expression with neural FSTFs, including *tcf/lef* and *otxB* (R=0.08 and R =0.12). *INSM2* was expressed in a subset of non-ciliated neurons marked by the expression of the secreted neural peptide-encoding gene *spp-4* and *INSM2* RNAi resulted in reduction of *spp-4*+ cells (Figure 4E). These results identify novel neoblasts classes for mature neural subtypes and FSTFs that delineate them.

To study neural fate in neoblast progeny cells, we subclustered the 9,993 post-mitotic G_0 neural progenitors (Figure 4D). A total of 77 G_0 neural progenitor clusters were identified – many more than the eight neural X1 clusters identified above. In Fincher et al. 2018, at least 70 differentiated neuron clusters were identified (Figure 4C). There is therefore a similar scale of identified heterogeneity in G_0 neural progenitor subclusters and mature neurons, but further heterogeneity likely remains hidden in the clusters of both of these datasets. We were able to assign 37 subclusters as candidate G_0 neural progenitors for distinct neural subtypes based on the shared expression of unique markers between differentiated neuron subtypes and G_0 neural progenitor clusters (Figure 4D). An additional 30 subclusters could not readily be connected to a differentiated cluster from Fincher et al. 2018, but displayed expression of unique markers (mostly FSTFs). Therefore, in total at least 67 classes of post-mitotic neural progenitor states were identified, with their transcriptomes and associated FSTFs defined.

Because the number of neural G_0 progenitor classes in this data was far greater than the number of clusters of neural S/G_2/M neoblasts (77 vs. 8), it is possible that fate heterogeneity increases for neural progenitors in post-mitotic stages. To explore this possibility further, we first assessed the expression of neural FSTFs that had enriched expression in G_0 subclusters within S/G_2/M neoblasts. Many of these genes were expressed in a similar number of both neoblasts

and their post-mitotic descendants. However, whereas they were expressed in specific subclusters within G_0 cells, they were not expressed in a specific subcluster (or region of UMAP space) in S/G_2/M neoblasts (Supplemental Figure 4C). This differed from neoblasts for other cell types, such as muscle, where cells expressing FSTFs clustered together. This could indicate that neural neoblasts are more homogenous and less defined than G_0 neural progenitors, even when they express neural FSTFs. We next asked if FSTFs that are co-expressed in the G_0 neural progenitor state tended to also be co-expressed in neoblasts even if the cells were not clustered. It was previously shown that *pitx* and *lhx1/5-1* specify serotonergic neuron progenitors (Currie and Pearson, 2013; März et al., 2013), and these two transcription factors were coexpressed in G_0 cells (Supplemental Figure 4D). However, despite both being individually expressed in some S/G_2/M neoblasts, they were not coexpressed and their gene expression correlation was lower than in G_0 progenitors (R=0 vs. R=0.29). We also found a transcription factor-encoding gene, *UNCX*, that had highly correlated expression to *pitx* in post-mitotic progenitors. In S/G_2/M neoblasts, *UNCX* was expressed in a relatively similar number of cells, but *UNCX* expression was less correlated to *pitx* in neoblasts than in G_0 progenitors (R=0.07 vs. R=0.41) (Supplemental Figure 4E). Another example population of G_0 progenitors (cluster 7; Figure 4D) expressed both *otxA* and *otxB*, yet these genes were uncorrelated in S/G_2/M neoblasts (R=0.01 vs R=0.33) (Supplemental Figure 4F). Among pairs of correlated TFs identified in G_0 neural progenitors, many showed a similar increase in correlation from neoblast state to post-mitotic progenitor (Figure 4F). These examples indicate that at least some neural progenitors did not express the full complement of transcription factors that define them in the neoblast stage, but activate them as post-mitotic progenitors.

We next utilized the atlas of TFs expressed in different neural G_0 progenitors to identify novel FSTFs required for the presence of neural populations. Expression of *PHOX2A*, which encodes a paired-like homeodomain TF, defined a unique G_0 neural progenitor population with a signature corresponding to dd_8060+ peripheral neurons (Supplementary Figure 4S). RNAi of

PHOX2A eliminated dd_8060-expressing neurons from the animal (Figure 4E). Another TF, *POU4F3*, was specifically expressed in a G_0 neural progenitor population corresponding to *CALM2*+ (dd_23127) neurons (ciliated cluster 21 from Fincher et al. 2018) and *POU4F3* RNAi led to loss of *CALM2*+ neurons (Figure 4E, 4G; Supplementary Figure 4T). *PHOX2A* was expressed in 4C neural neoblasts, but in a scattered pattern. Similarly, *POU4F3* was also expressed in 4C neural neoblasts by sequencing data, and this was confirmed in uninjured animals by FISH (Figure 4G). *POU4F3* was coexpressed with the known neural FSTF *scratch* in G_0 cells, but it was not in neoblasts. The neural population represented by *CALM2* has been shown to be regulated by the FSTF *soxB1-2* (Ross et al., 2018). In the Fincher et al. 2018 atlas, *soxB1-2* is expressed in the cluster marked by *CALM2*, but also in several other clusters. *Tbx2/3b* was expressed robustly in a G_0 neural progenitor population that is associated with *GLIPR1*+ (dd_210) peripheral neurons (Figure 4H). RNAi of *Tbx2/3b* resulted in the loss of cells expressing *GLIPR1*+, establishing *Tbx2/3b* as an FSTF for this neural population (Figure 4E). *Tbx2/3b*+ neoblasts were identified in uninjured animals by FISH (Figure 4H), but coexpression of *Tbx2/3b* and *lhx1/5-1* (another TF-encoding gene enriched in this population) was relatively rare in 4C neoblasts compared to post-mitotic progenitors. An additional eight examples of FSTF-cell type ablation by RNAi were found. This included (i) *IRX2* RNAi, which led to loss of tyrosine hydroxylase (*th*)-positive dopaminergic neurons, (ii) *UNCX* RNAi, which led to loss of *sert*-positive serotonergic neurons, (iii) RNAi of dd_10911 (encoding a zinc finger protein family member), which led to loss of dd_69653-positive peripheral neurons, and (iv) RNAi of three TFs that each independently led to loss of dd_29413-positive ciliated neurons (including *Tbx2/3*c and a gene with similarity to human *SOX2*) (Figure 4E and Supplemental Figures 4G-4U).

Together these analyses demonstrate that numerous unique neural progenitor states were identified in the G_0 post-mitotic progenitor data and that they could be associated with mature specific neuron types through transcription factor and differentiated marker expression

similarity. For several of these, identified FSTFs were shown to be required for the presence of the mature cell type. The progenitors for many of these cell types were not overtly present in the 4C neural-specialized neoblast scRNA-seq data as unique subclusters or through expression of their characteristic transcription factor modules, consistent with fate diversification or maturation occurring post-mitotically for many neural subtypes. The neural population of 4C neoblasts will be an intriguing target for continued detailed investigation to unravel the underpinnings of neural fate choice in neoblasts.

Expansion of cell-type diversity post-mitotically

As described above, a large diversity of neural and *cathepsin*+ cells are known to exist, but the neoblasts for these populations contained far fewer identifiable cell classes than their differentiated counterparts. To further explore the possibility of fate diversification occurring in post-mitotic stages for these tissues, we compared cells in the neoblast and post-neoblast progenitor stages for both classes. Both neural and phagocytic/*cathepsin*+ tissue had more post-mitotic clusters compared to neoblasts, irrespective of the number of principal components used for data analysis. We used k-means clustering of cells in these classes using the top principal components in the dataset as another way to assess the number of cell types in the data. To measure the number of clusters/states in the data we used the Ratkowsky index, which is a measure of the quality of clustering of a dataset. By using the Ratkowsky index for a variable number of cell-state identities, the number of groups in each tissue progenitor class (for 4C and post-mitotic cells) that best explains the differences between cells can be found. For neural progenitors, more groups explained the variance in the data for the post-mitotic progenitors compared to neoblasts, consistent with a greater number of cell states emerging post-mitotically after cells left the neoblast stage (Supplemental Figure 4V). *cathepsin*+-fated progenitors also showed an increase in group number in post-mitotic progenitors compared to 4C neoblasts, with a more modest difference compared to neural states. By contrast, muscle-

fated cells had similar numbers of groups in the neoblast and post-mitotic progenitor datasets. This analysis lends further support to the idea that some cell types continue to diversify after initial specification as neoblasts.

A *six1/2-1*+ cluster associated with germline niche cells

Post-mitotic progenitor cluster 17 (Figure 1F) had enriched expression of *six1/2-1, dmd-1*, and *MECOM,* and likely represents progeny of cells from within the *six1/2-1*$^{+}$ S/G_2/M neoblast cluster that also expressed these genes (Figure 5A, 5B). These cells did not express neural markers indicating that they are not neural progenitors (Supplemental Figure 5A). *dmd-1* is expressed in testis gonad support cells in sexually-reproducing *S. mediterranea* (Chong et al., 2013). These cluster 17 cells also expressed *ophis*, *aadc*, and *laminA* (Figure 5B), other genes that are expressed in the male and female sexual accessory cell types in sexual *S. mediterranea* (Saberi et al., 2016; Issigonis et al., 2022; Khan and Newmark, 2022). Some of these genes are also known to be expressed in asexuals, which possess germ cells but not mature gametes (Wang et al., 2007). Two *notch*-family genes were found to be expressed in the ovaries in sexual planarians (Khan and Newmark, 2022) and were also expressed in subclusters of these G_0 cells. The gene referred to as *notch2* (dd_7067) in Khan et al. 2022 was expressed in a subcluster devoid of *dmd-1* expression, as expected given this gene is expressed in ovaries and *dmd-1* is expressed in the testis. These results suggest this cluster of cells likely contains progenitors for the somatic support cells for the germline in asexual planarians.

Distinct S/G_2/M intestine states correspond to mature intestinal subtypes

The planarian intestine is comprised of three main cell types: phagocytic enterocytes, basal/outer intestinal cells, and secretory goblet cells dispersed within the enterocyte layer (Fincher et al., 2018; Forsthoefel et al., 2020) (Figure 5C). Several FSTFs regulate specification of intestine neoblasts (also called gamma neoblasts), including *gata4/5/6-1*, *nkx2.2*, *prox-1*, and

hnf4 (Wagner et al., 2011; van Wolfswinkel et al., 2014; Flores et al., 2016; González-Sastre et al., 2017). Intestinal neoblasts have largely been considered a homogenous specialized neoblast population, despite mature intestine cell-type heterogeneity. Clustering of $S/G_2/M$ neoblasts identified two distinct intestinal neoblast populations (Figure 1B). We combined both clusters and performed subclustering with 2,065 intestinal neoblasts, generating five total subclusters (Figure 5D). All subclusters shared the expression of known intestine FSTFs. Four clusters (subclusters 0, 1, 3, and 4) were identified as putative enterocyte precursors based on the expression of enterocyte-specific markers (Figure 5E). These clusters were also enriched in the expression of the transcription factors *zfp-1* and *osr* – FSTFs not previously known to be associated with gamma neoblasts. Cells of subcluster 2 did not express *zfp-1* and *osr,* but were characterized by expression of the TF-encoding *hunchback* and *RREB2* genes. *RREB2* has been previously shown to be expressed in goblet and basal intestine cells (Forsthoefel et al., 2020). Correlation analysis between transcription factors identified transcription factor modules that associated broadly into two main groups – the enterocyte clusters and subcluster 2 (Figure 5F, 5G). *zfp-1* and *p53*, which are associated with the epidermal lineage, were coexpressed in enterocyte neoblasts, but another epidermal TF (*soxP-3*) was not strongly expressed (Supplemental Figure 5B). Subcluster 2 neoblasts coexpressed the known FSTFs *gata4/5/6-1* and *prox-1*, as well as *Tbx2/3*c.

In post-mitotic G_0 cells, two clusters of apparent intestinal progenitor cells existed (586 total cells) (Figure 1F). Each displayed gene expression corresponding to either enterocytes or basal/outer intestinal cells (Figure 1G). We pooled these cells and subjected them to subclustering, revealing five subclusters (Figure 5H). The presumptive G_0 enterocyte progenitors shared a similar transcriptional program to the presumptive 4C enterocyte-specialized neoblasts: expression of enterocyte-specific markers and *zfp-1* and *osr* (Figure 5I; Supplemental 5C). The presumptive G_0 basal/outer intestinal cell progenitors shared a transcriptional program similar to subcluster 2 of the 4C gamma neoblasts (*zfp-1*-, *osr*-,

RREB2+), suggesting that subcluster 2 of the 4C gamma neoblasts represents outer intestinal cell-specialized neoblasts (Figure 5I; Supplemental Figure 5C). These clusters also expressed some goblet cell markers, indicating they may also give rise to basal and goblet cells (Figure 5I). Correlation analysis in the gamma neoblasts showed that *zfp-1* is negatively correlated with both *RREB2* and *hunchback*, supporting the hypothesis that *zfp-1+/osr*+ neoblasts and the *RREB2+/hunchback*+ neoblasts define distinct states within the gamma neoblast compartment (Supplement Figure 5D). Despite the fact that all intestinal neoblasts share expression of many canonical intestine FSTFs (e.g., *gata4/5/6-1*, *nkx2.2*, *prox-1,* and *hnf4*), these data indicate intestinal neoblasts are heterogenous and generate cell-type diversity through at least two broad gamma neoblast subtypes.

Neoblast fates for the epidermis

The epidermis originates from neoblasts called zeta neoblasts (van Wolfswinkel et al., 2014). Dorsal and ventral epidermis are functionally distinct, and a unique population also exists at the dorsal-ventral (DV) animal boundary (Wagner et al., 2012; Wurtzel et al., 2017) (Figure 6A). Dorsal and ventral identities originate in zeta neoblasts, marked by the expression of *PRDM1-1* and *kal1*, respectively (Wurtzel et al., 2017). Subclustering of the 6,504 S/G_2/M epidermal neoblasts identified unique clusters corresponding to dorsal and ventral precursors (Figure 6B). A third cluster displayed lower expression of canonical epidermal FSTFs (*soxP-3*, *p53*, and *zfp-1*) (Figure 6C). TF-correlation analysis identified modules that correspond to the ventral and dorsal epidermal clusters (Figure 6D and 6E). An additional module had enriched expression of *Post-2a* and *Post-2b,* transcription factors previously found to be expressed in the mature DV boundary epidermis (Wurtzel et al., 2017). This indicates that the epidermal DV-boundary likely arises in neoblasts, not later in G_0 progenitors where expression of the previously identified DV boundary FSTF *tlx-1* first arises (Wurtzel et al., 2017). Subclustering of the 2,742 G_0 epidermal progenitor cells identified dorsal and ventral epidermal progenitors, although the cells clustered

primarily by canonical epidermal lineage progression markers, such as *prog-2*, rather than dorsal/ventral markers (Figure 6F and 6G). An offshoot G_0 progenitor subcluster (cluster 4) consisted of cells expressing transcription factors in the DV boundary TF module.

Neoblast fates for the protonephridia

The planarian protonephridia functions as a waste excretion and osmoregulatory system that includes flame cells, proximal tubule cells, distal tubules cells, and collecting duct cells (Rink et al., 2011; Scimone et al., 2011; Vu et al., 2015) (Figure 6H). Several TFs are important for protonephridia formation (Scimone et al., 2011). Subclustering of $S/G_2/M$ protonephridia neoblasts (550 total) revealed a seemingly homogenous population of cells (Figure 6I). Protonephridia FSTFs such as *POU2/3*, *six1/2-2*, and *sall* were expressed throughout these $S/G_2/M$ neoblasts (Figure 6J). *hunchback*, an FSTF required for forming all subtypes of protonephridia, displayed enriched expression primarily in cluster 1. Correlation analysis between transcription factors expressed in protonephridia neoblasts identified modules of correlated transcription factors that did not mark distinct regions of cells within clustered UMAP space, potentially indicating these correlated transcription factors are expressed broadly in protonephridia neoblasts (Figure 6K).

Subclustering of post-mitotic G_0 protonephridia progenitor cells (1,179) revealed a small number of clusters and regions corresponding to nephridia cell types (Figure 6L). Cluster 5, along with a small population of cells in regions of clusters 0 and 2, was positive for the flame cell markers dd_2920 and *ECE2* (dd_5256) (Figure 6M). A small region of cluster 0 expressed differentiated proximal tubule cells markers such dd_9435 and *ALPPL2* (dd_8942). These results support a previously proposed model that distinct protonephridia subtypes emerge from a common pool of 4C progenitors defined by expression of several FSTFs (Scimone et al., 2011), and suggest that cell-type diversity for these cells emerges later in the post-mitotic states.

Discussion

Generating cell-type diversity is a foundational task not only for animal development, but also in adult homeostasis and regeneration. Regeneration poses a number of unique challenges compared to development and generating cell-type diversity might therefore involve different mechanisms in these contexts. First, cell-fate specification must be tailored to the identity of unpredictable missing tissue types in regeneration rather than initiating from a fixed starting point. Second, fate specification in regeneration can occur in adult progenitors existing within organized, mature tissues presenting different contextual environments for influencing fate. Finally, whereas amplification of cells prior to differentiation is critical in development, wound sites in principle could have many progenitors from early time points. Planarians can generate every adult cell type (~150 types) in adulthood during tissue turnover and in regeneration from pluripotent stem cells called neoblasts (Fincher et al., 2018; Plass et al., 2018; Reddien, 2018). Planarians therefore present an attractive venue for uncovering solutions to the challenge of generating large-scale cell-type diversity in an adult context. Many fundamental questions about the nature of fate choice in planarian regeneration remain unaddressed. For example, are there neoblast classes corresponding to fate choices for each adult cell type of the entire animal? Or, alternatively, is some decision-making made in the post-mitotic descendant cells of neoblasts that serve as progenitors? Systematic identification of fate choices and the transcription factor expression signatures associated with those fate choices in neoblasts and post-mitotic progenitors will help understand the logic of fate choice in a regenerative context.

Here, we utilized scRNA-seq on 4C neoblasts and 2C post-mitotic progenitors to generate an atlas of transcription factors associated with all stem cell differentiation paths that could be annotated. In total, 133 separate clusters of post-mitotic progenitors were identified – similar to the estimated total number of planarian cell types. We were able to connect the majority of these progenitors to particular known differentiated states and to annotate the

transcription factor expression signatures of each of these progenitor types (TF Atlas: Supplemental Table 5). For many of these cell states and FSTFs that had not been previously investigated we utilized RNAi to establish a requirement of the FSTF for regeneration and maintenance of the predicted target differentiated cell type.

This work combined with prior studies on neoblast specialization identifies distinct principles for generating cell-type diversity across tissues (Figure 7A, 7B): within muscle and parenchymal cells, distinct transcriptional regulatory programs involving unique FSTF expression signatures establish diversity in the 4C neoblast state, and these states are maintained into post-mitotic progenitors and differentiated cells. For intestine cell-type formation all intestinal 4C cells are united by the expression of some major FSTFs, and exclusive expression of other FSTFs distinguish specialized neoblasts for distinct intestinal cell types. The epidermis is similar to these other tissues, in that dorsal and ventral epidermal states were apparent in 4C cells, as previously suggested using the markers *prdm-1* and *kal-1* (Wurtzel et al., 2017), along with newly identified neoblasts for the DV boundary. These four tissues all represent cases of fate diversification at the earliest possible step: in the neoblasts.

For the protonephridia and phagocytic/*cathepsin*+ cells, 4C neoblasts specialized for these tissue classes were apparent but not overtly heterogeneous, in contrast to muscle, intestine, parenchymal cells, and epidermis. These 4C neoblasts lacked clear signatures of cell type-specific markers and were not comprised of groups of cells with multiple unique FSTF signatures. This raises the possibility that generation of cell-type diversity for these tissues could arise from a common pool of 4C precursors that diversifies in the G_0 stage, as was previously proposed for the protonephridia (Scimone et al., 2011) (Figure 7B). Other molecular and functional studies will be important for continued assessment of this possibility.

Finally, neural differentiation involves heterogeneity in the 4C neoblast state, but this heterogeneity was very limited when compared to the large degree of heterogeneity of mature neuron types. Several known neural FSTFs displayed enriched expression in a few common

subclusters, whereas others were broadly expressed throughout the entire population and others were rare and scattered in their expression. By contrast, substantial transcriptome heterogeneity was readily apparent for neural progenitors in the form of many distinct cell-type-specific clusters in the post-mitotic progenitor state. The fate choice for many mature neural cell types could be mapped onto these post-mitotic progenitor clusters. These post-mitotic neural progenitors were distinguished by their expression of unique transcription factors (Figure 7A), many of which were not robustly co-expressed in neural neoblasts. Migratory targeting in post-mitotic progenitors could in principle provide a means for extrinsic cues to further influence fate choice by local cell position.

Overall, we also identified 20 different TFs with either previously unknown FSTF function or FSTF function for a cell type different than the one previously known. Many of these new FSTFs span the parenchymal and neural lineages—two of the more diverse tissue types of the animal. At least 40 different choices associated with distinct transcriptome states could be documented within the neoblasts. This number will likely continue to increase somewhat through continued molecular dissection of neoblast states. How neoblasts can choose between such a plethora of options is unknown. 133 states were apparent in the G_0 state, with transcriptomes displaying further maturation and distinction from one another. The timing of many choices in the trajectory of cells towards differentiation still remain unaccounted for, and can be the subject of future investigation. Overall, this work helps define the scope, timing, and organization principles that define the generation of cell-type diversity in an adult context, in which tissue turnover and regeneration can produce all cells of the organism. We also present an atlas of transcription factor signatures associated with stem cell differentiation into most cell types of a complete adult animal. This resource can enable probing the mechanisms of cell-fate specification in regeneration and the study of the function and evolution of broadly conserved transcription factors in animal cell-type formation.

Materials and Methods

Dissociation and fluorescence-activated cell sorting (FACS)

Cell preparation for FACS was performed as described (Raz et al, 2021). Briefly: starved animals were diced in polystyrene dishes with Calcium-Magnesium Free solution (400mg/L NaH2PO4, 800mg/L NaCl, 1200 mg/L KCl, 800 mg/L NaHCO3, 240 mg/L glucose, 15mM HEPES, pH7.3) with 1%BSA and 0.1mg/mL collagenase (termed CMFB-C solution). Diced fragments were transferred to a 50mL conical tube and vigorously agitated in 50mL of CMFB-C using transfer pipettes for 10 minutes to generate cell suspensions. The cell suspension was passed through a 40µm filter and centrifuged at 300g for 5 minutes. Supernatant was decanted and cells were resuspended in 3.5mL CMFB (no collagenase) before passing through a 35µm filter. 10mg/mL Hoechst solution (Invitrogen) was added to cell suspensions at a 1:50 dilution and the suspension was incubated for 45 minutes in the dark at room temperature. Propidium iodide was added to cells (5mg/mL) directly before sorting. Cells were sorted on a FACS Aria II (BD Biosciences) for live (propidium iodide negative), nucleated (Hoechst positive) cells.

Cell sorting, library preparation, and 10X Genomics scRNA-seq

Cells were sorted into X1 and X2 gates as described (Hayashi et all, 2006). After sorting, cells were resuspended in CMF with 1% BSA in a concentration range between 700 cells/uL to 1200 cells/uL. Cells were processed using the 10X Genomics Chromium Controller and Chromium Single Cell Library and Gel Bread Kit following the standard manufacturer protocol. Amplified cDNA libraries were quantified using a bioanalyzer and size selected using AMPure beads. Sequencing was performed on an Illumina HighSeq 2500 using the ‘Drop-Seq core computation protocol’ as described in (Fincher et al, 2018). Reads were mapped to the v6 Dresden transcriptome assembly (Rozanski et al, 2019) and a gene cell matrix was generated for analysis using the Seurat 3.0.2 pipeline.

scRNA-seq data analysis, clustering, and subclustering

The Seurat 3.0.2 package was used for pre-processing, quality control, clustering, and subclustering analyses. Each individual library was processed separately. First, contig isoforms for sequences were merged by summing the mapped reads to each isoform. Transcripts representing ribosomal and mitochondrial reads were removed as described (Fincher et al, 2018), and cells with fewer than 500 UMIs or greater than 18,000 UMIs were also removed. Seurat objects were generated for each individual library. Cells were then normalized using the NormalizeData function ("LogNormalize" method; scale factor = 10,000) and variable genes were identified using the FindVariableGenes function ("vst" selection method; 2,000 features). Variations in global gene expression were regressed out by scaling and centering to the number of unique molecular identifiers (UMIs) using the ScaleData function. The same variable genes were used as input for principal component analysis using the RunPCA function. Cells were clustered using the FindNeighbors and FindClusters functions, and subsequently plotted in two dimensions using Uniform Manifold Approximation and Projection (UMAP; RunUMAP function). The number of principal components (PCs) used as input for FindClusters and RunUMAP was determined empirically by testing a range of different PCs as input until optimal clustering occurred, as described (Fincher et al, 2018). Cells from individual libraries/objects and of the same FACS gating were merged using the merge function, whenever the same 10X Genomics Chromium Kit chemistry was used in library production. Upon generation, two clusters from merged X1-gated object contained no exclusively enriched genes, but rather enriched genes highly expressed in cells from regions of most other clusters, suggesting that these clusters represented artifacts (Fincher et al, 2018). These clusters were removed from the data. Additionally, two other clusters represented more differentiated populations of muscle and epidermis (clustering away from cycling muscle and epidermal neoblasts)—likely representing non-cycling cells—and were removed. Cells with expression of *smedwi-1* > 0.5ln (UMI-per-

10,000+1) were retained for the final merged X1 dataset. X1 clusters were identified by the presence of known markers or novel markers associated with known differentiated tissues. For the merged early G_0 dataset, initial clustering yielded 20 clusters (PC=50, resolution=0.5). Finally, cells with expression of *smedwi-1* < 3.5ln (UMI-per-10,000+1) and *p4hb* > 1.5ln (UMI-per-10,000+1) were retained. Early G_0 cell cluster labels were identified by the presence of known markers for differentiated tissues (Fincher et al, 2018). For both datasets (merged X1 and G_0 progenitor cells), genes enriched in clusters and subclusters were identified using the FindMarkers functions (log-scaled fold difference threshold of 0.5).

When library production across multiple runs involved different 10X Genomics Chromium Kit chemistries (e.g., Chromium v3 vs. v3.1), the Seurat Integration pipeline was used rather than the merge function alone. Briefly, objects were first merged (merge function) then split into lists of two objects (SplitObject function). Cells were normalized, variable genes identified, and features repeatedly variable across datasets were selected (SelectIntegrationFeatures function). Cross-data pairs of cells in a matched biological state (called anchors) were identified (FindIntegrationAnchors function) and an integrated data assay was formed (IntegrateData function). The default assay was then set to “integrated”, and standard scaling, PCA, UMAP, and clustering were performed. The default assay was switched to “RNA” when visualizing gene expression and switched back to “integrated” when subclustering. Subclustering analysis was performed by subsetting data into new Seurat objects (subset function) and performing processing similar to initial cluster formation (variable features, scaling data, PCA, finding neighbors and clusters, UMAP).

RNAi

RNAi was performed as described in Scimone et al. 2017. Briefly, dsRNA was synthesized by *in vitro* transcription (Promega) using PCR-generated templates with flanking T7 promoters,

followed by ethanol precipitation, and annealing after suspension in water. dsRNA was mixed with planarian food (liver) and ~2 uL of this mixture was given per animal for feedings. All animals received at least 8 feedings. For regeneration experiments, animals were amputated intro three pieces (head, trunk, and tail) several days after the last RNAi feeding. Fragments were fixed for labeling and further analysis.

Cell *fluorescence in situ* hybridization

This protocol was performed as described (Raz et al, 2021). Briefly, sorted cells were resuspended at 4,000 cells/μL in CMF and plated on poly-lysine-D coverslips in 24 well plates. Cells were allowed to settle for 30 minutes before fixation in 4% paraformaldehyde in CMF. Cells were washed in 1% phosphate-buffered saline with 0.1% Triton-x (PBSTx) and a 1:1 PBSTx:PreHyb solution (King and Newmark, 2013). Cells were incubated in PreHyb for two hours at 56°C and hybridized with RNA probes at 1:800 in Hyb for >16 hours at 56°C (digoxigenin-, fluorescein isothiocynate-, and dintriophenol-labeled riborpobes were synthesized as described in Pearson et al. 2009). Cells were then blocked 30 minutes prior to incubating overnight with anti-DIG-POD (1:1500, Roche), anti-FITC-POD (1:2000, Roche), or anti-DNP-HRP (1:150 Perkin-Elmer) in blocking solutions of PBSTx with 10% western block reagent (Roche) for ant-DIG-POD, combined 5% western block reagent and 5% casein solution (Sigma) for anti-FITC-POD, or combined 5% western block reagent and 5% heat-inactivated horse serum anti-DNP-HRP. Fluorescent tyramide development and amplification was performed by placing cells in borate buffer for 5 min (0.1M boric acid, 2M NaCl, pH8.5), followed by 10 min in borate buffer with rhodamine (1:1000) or fluorescein (1:1500) tyramide, and 0.0003% hydrogen peroxide. After development, peroxidase inactivation was performed in a 1% sodium azide solution for >1 hour, followed by antibody labeling for the second probe. Samples were mounted in Vectashield with DAPI (Vector Labs).

Whole-mount fixation and *in situ* hybridization

Animals were incubated in 5% N-acetylcysteine for 5 minutes prior to fixation with 4% formaldehyde in PBSTx for 20 minutes (with a brief PBSTx wash in between). Animals were then then washed in PBSTx once, placed in PBSTx:Methanol for 10 minutes, and methanol alone for 10 minutes before storage at -20C overnight. Animals were then bleached for 2 hours in 5% formamide, 0.5x SSC, and 1.2% hydrogen peroxide solution on a light source. Animals were then incubated in 5 µg/mL proteinase K, before post-fixation in 4% formaldehyde in PBSTx. Hybridization and tyramide development was performed as described above.

Imaging acquisition

Fluorescent micrographs were acquired using the Leica Stellaris Sp8 laser scanning confocal microscope. ImageJ software (Fiji) was used for processing images from FISH data.

Transcription factor catalog construction

A FASTA database of planarian protein sequences was generated from six-frame translations of all dd_Smed_v6 transcripts. Pfam Hidden Markov models (HMMs) were downloaded from pfam.xfam.org for known DNA-binding domains (DBD) and collated to a single file. Using HMMER v3, the translated transcriptome was searched (hmmsearch) using each DBD HHM (hmmfetch) to identify transcripts encoding possible transcription factors (Supplementary Table 2; raw output). Of transcripts representing multiple isoforms of the same gene, only the longest transcript containing a DBD was recorded. Planarian transcripts with at least one match to a DBD HHM were compiled into an intermediate catalog and were BLAST searched (blastx; evalue < 1e-5) against the human proteome. Genes that have been previously identified in planarians were identified by constructing a blast database from *Schmidtea mediterranea* proteins and transcripts previously deposited into GenBank, and BLAST searching against that (blastx and blastn; evalue < 1e-5, percent identity > 90%). Probable transcription factors were

identified by homology to human transcription factors and the subset of validated FSTFs were identified by literature searching for planarian transcription factors previously deposited in GenBank. Transcripts that blast to non-transcription factors were categorized as non-transcription factor DNA-related proteins or miscellaneous proteins and were excluded. Transcript products without similarity to a human protein were classified as having no blast hit and were retained for future analysis.

Transcription factor correlation analysis

Gene-cell matrices of sequenced X1 and G_0 progenitor cells were exported from Seurat after preprocessing and clustering as csv files of the gene-cell matrix, cell identities, and transcript contig identifiers and imported to MATLAB. Putative transcription factors from the catalog were isolated from both gene cell matrices, collapsing all isoforms for each gene. Pearson correlation coefficients (corrcoef) were calculated between gene pairs across all X1 or G_0 cells. The technique was validated by making a correlation matrix of known FSTFs from the transcription factor catalog and verifying high correlations between transcription factors with overlapping expression. Putative transcription factors from the catalog with high correlations ($r > 0.14$) between either each other or a known FSTF were inspected. Heatmaps were generated comparing correlations between correlated putative transcription factors and between putative transcription factors and known FSTFs. Genes in heatmaps were ordered by hierarchical clustering across both axes. Correlation matrices were exported to Python as csv files for figure generation using the Seaborn library. Final correlation matrices were reclustered by Euclidian distance using the UPGMA algorithm, and gene pairs with the highest correlations were included in the final figure.

Transcription factor modules

Transcription factor module figures were generated by extracting UMAP positions of cells expressing the entire module of transcription factors (> 0.5ln (UMI-per-10,000 +1)). The cell positions were binned to a 2D heatmap and contours were generated by convolving a gaussian kernel over it (with a standard deviation of 2*the minimal dimension of the UMAP plot) and thresholding (0.5-0.7 of the maximal value).

Ratkowsky index

Cell principal component embeddings were extracted for specific clusters in Seurat. The number of principal components chosen for each tissue type was determined by the elbow method, corresponding to the number of principal components where additional components no longer explain more variance than what is expected by splitting unstructured data. The same number of principal components were used for $S/G_2/M$ neoblast clusters and for G_0 progenitor clusters (Neural: 40, Muscle: 20, Cathepsin: 20). Ratkowsky indicies were calculated using the NbCluster package in R for these principal component embeddings using kmeans clustering and euclidean distance. NbClust function source code was altered to remove hard-coded seed state and replaced with a random number generator and this was run 1,000 times per tissue type to generate an average Ratkowsky index and a 95% confidence interval. The maximum Ratkowsky value is indicated on Ratkowsky index plots.

Transcription factor enrichment figures

Transcription factor tissue-type expression plots for $S/G_2/M$ neoblasts were made by averaging expression of each transcription factor across tissue-associated clusters. An expression specificity score for the most enriched cluster was measured by calculating the average expression of each transcription factor across the tissue-associated cells with the highest average expression of the transcription factor, divided by the average expression of the next highest tissue class. These were subdivided by domain, and plotted on a log-scaled line plot

and arranged radially by tissue-type. The homeodomain-containing transcription factor heatmap was generated by averaging expression for homeodomain-containing genes across each tissue type as before. A ratio of average expression for each tissue divided by max average expression was generated for each transcription factor. The heatmap was plotted using the seaborn package in python and hierarchically clustered.

Transcription factor atlas

The transcription factor atlas table for G_0 progenitor cells was generated by creating a Seurat object of all G_0 cells where the cluster identity of each cell was determined by the subclustering of the clusters from the tissue-specific analysis. Enriched genes for each subcluster were then found in Seurat and filtered for transcription factors with at least a one-fold log enrichment, an adjusted p-value of 0.05, and expressed in at least 10% of cells within the subcluster they are enriched in. Transcription factors were sorted by increasing adjusted p-value.

Neural G_0 characterization

Neural G_0 clusters were individually subclustered, with neighboring clusters 12 and 15 subclustered together. Enriched genes were identified for each subcluster in the same manner as other tissue types. To relate G_0 subclusters to their corresponding subcluster in the Fincher et al. 2018 atlas of differentiated cells, we created a merged Seurat object of all G_0 neural cells, with each cell maintaining their identity from subclustering. We generated a similar merged Seurat object for non-ciliated, ciliated, and “All Neural” clusters 3, 4, 7, 8, 10, 36, 43, and 59 (clusters that were not annotated as ciliated or non-ciliated, and were not enriched in expression of the neoblast genes *smedwi-1* or *smedwi-2*) from the Fincher dataset.

From these merged Seurat datasets, we found genes commonly enriched in at least one subcluster in both datasets and calculated the Euclidian distance between all G_0 and differentiated/Fincher subclusters based on the average z-scored expression of just these

mutually-enriched genes. For G_0 and differentiated pairs with low distances, we looked for the expression of unique marker genes shared between these cells to indicate a possible shared lineage. We annotated these predicted corresponding differentiated subclusters on UMAP plots of each neural G_0 subclustering. We also annotated neural G_0 UMAP plots with genes enriched in each subcluster.

Gene annotation and nomenclature

Genes were labeled as previously described (Fincher et al., 2018). Briefly, previously published planarian genes that had been submitted to the NCBI Nucleotide database (https://www.ncbi.nlm.nih.gov/nucleotide/) appear in *italics* within text and figures. Sequences not found in the planarian database, but that have a human best-blast (blastx) are labeled in uppercase with their human gene name, followed by the contig ID of the dd_v6 transcriptome assembly in parentheses.

Acknowledgements

We thank M.L. Scimone for help with dissociation and FACS for scRNA-seq preparation. We also thank Christopher Rodriguez for helpful discussions on algorithms and figure design and Giselle Valdes for discussions on computational analysis. We also thank Amelie Raz for fruitful conversations and all other Reddien lab members for discussion and comments. We acknowledge NIH R35 GM145345 for the support of this work. The project described was supported by award Number T32GM007753 and T32GM144273 from the National Institute of General Medical Sciences. P.W.R. is an Investigator of the Howard Hughes Medical Institute and an associate member of the Broad Institute of Harvard and MIT.

Author Contributions

H.O.K., K.E.O., and P.W.R. conceived and designed experiments. H.O.K., K.E.O., and C.T.F. performed the experiments. H.O.K. and K.E.O analyzed the data. H.O.K, K.E.O, and P.W.R. wrote the manuscript.

Figure 1

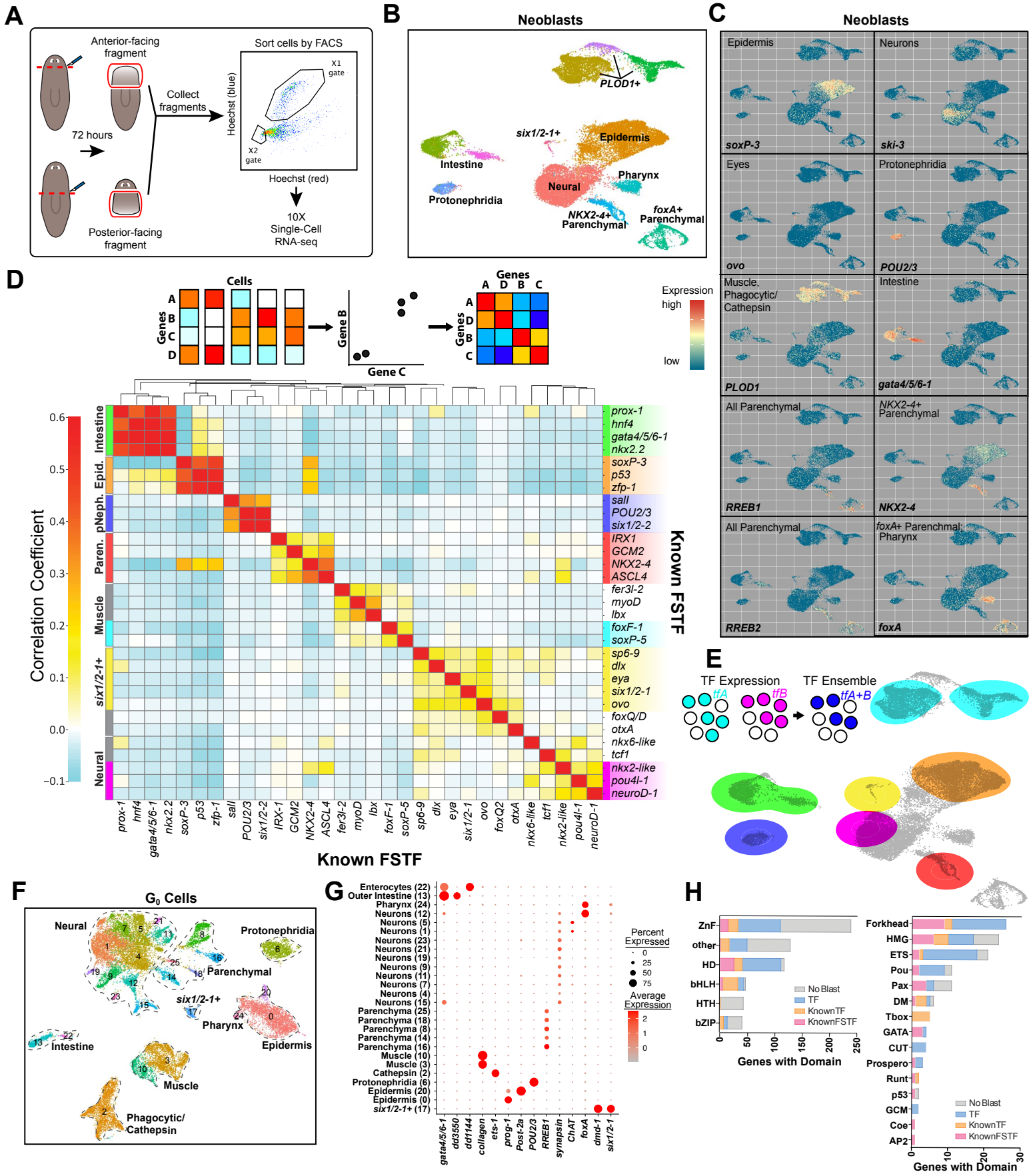

A
Anterior-facing fragment
Collect fragments
72 hours
Posterior-facing fragment
Sort cells by FACS
Hoechst (blue)
Hoechst (red)
X1 gate
X2 gate
10X Single-Cell RNA-seq
B
Neoblasts
PLOD1+
six1/2-1+
Epidermis
Intestine
Pharynx
Neural
Protonephridia
NKX2-4+ Parenchymal
foxA+ Parenchymal
C
Neoblasts
Epidermis
soxP-3
Neurons
ski-3
Eyes
ovo
Protonephridia
POU2/3
Muscle, Phagocytic/ Cathepsin
PLOD1
Intestine
gata4/5/6-1
All Parenchymal
RREB1
NKX2-4+ Parenchymal
NKX2-4
All Parenchymal
RREB2
foxA+ Parenchmal; Pharynx
foxA
D
Cells
Genes
Gene B
Gene C
Expression
high
low
Correlation Coefficient
Known FSTF
Intestine
Epid.
pNeph.
Paren.
Muscle
six1/2-1+
Neural
prox-1
hnf4
gata4/5/6-1
nkx2.2
soxP-3
p53
zfp-1
sall
POU2/3
six1/2-2
IRX-1
GCM2
NKX2-4
ASCL4
fer3l-2
myoD
lbx
foxF-1
soxP-5
sp6-9
dlx
eya
six1/2-1
ovo
foxQ2
otxA
nkx6-like
tcf1
nkx2-like
pou4l-1
neuroD-1
E
TF Expression
TF Ensemble
tfA
tfB
tfA+B
F
G0 Cells
Neural
Protonephridia
Parenchymal
six1/2-1+
Pharynx
Intestine
Epidermis
Muscle
Phagocytic/ Cathepsin
G
Enterocytes (22)
Outer Intestine (13)
Pharynx (24)
Neurons (12)
Neurons (5)
Neurons (1)
Neurons (23)
Neurons (21)
Neurons (19)
Neurons (9)
Neurons (11)
Neurons (7)
Neurons (4)
Neurons (15)
Parenchyma (25)
Parenchyma (18)
Parenchyma (8)
Parenchyma (14)
Parenchyma (16)
Muscle (10)
Muscle (3)
Cathepsin (2)
Protonephridia (6)
Epidermis (20)
Epidermis (0)
six1/2-1+ (17)
gata4/5/6-1
dd3550
dd1144
collagen
ets-1
prog-1
Post-2a
POU2/3
RREB1
synapsin
ChAT
foxA
dmd-1
six1/2-1
Percent Expressed
Average Expression
H
ZnF
other
HD
bHLH
HTH
bZIP
No Blast
TF
KnownTF
KnownFSTF
Genes with Domain
Forkhead
HMG
ETS
Pou
Pax
DM
Tbox
GATA
CUT
Prospero
Runt
p53
GCM
Coe
AP2

Figure 1. Single-cell RNA sequencing of S/G_2/M neoblasts and G_0 progeny

(A) Schematic illustration of workflow used to isolate and sequence cells.

(B) UMAP plot visualizing 12 clusters of 23,873 total neoblasts from X1 FACS gate.

(C) UMAP plots visualizing gene expression (red/high, blue/low) of enriched marker genes.

(D) Top: Schematic illustration displaying gene correlation analysis procedure. Pairwise Pearson correlations between genes expressed for each cell are displayed in a heatmap followed by hierarchical clustering. Bottom: Correlation analysis heatmap of subset of known FSTFs across all S/G_2/M cells.

(E) Top: Illustration defining TF ensemble as co-expression of multiple transcription factors. Bottom: TF ensemble expression domain is overlaid onto UMAP plot.

(F) UMAP plot visualizing 26 clusters of 31,084 total G_0 cells from the X2 FACS gate.

(G) Dot plot visualizing the expression of known tissue markers genes across clusters.

(H) Categorization of genes within *in silico* TF catalog. Bar graphs display number of unique genes from TF catalog encoding a DNA binding domain (DBD) motif of a given family. Genes are further labeled as: “Known FSTF” (previously published as planarian FSTF); “Known TF” (previously published as planarian TF, but without characterized FSTF attribution); “TF” (Not published as planarian TF, but with blast hit to Human TF”); or “No Blast” (no planarian TF characterization and no human blast hit).

Figure 2

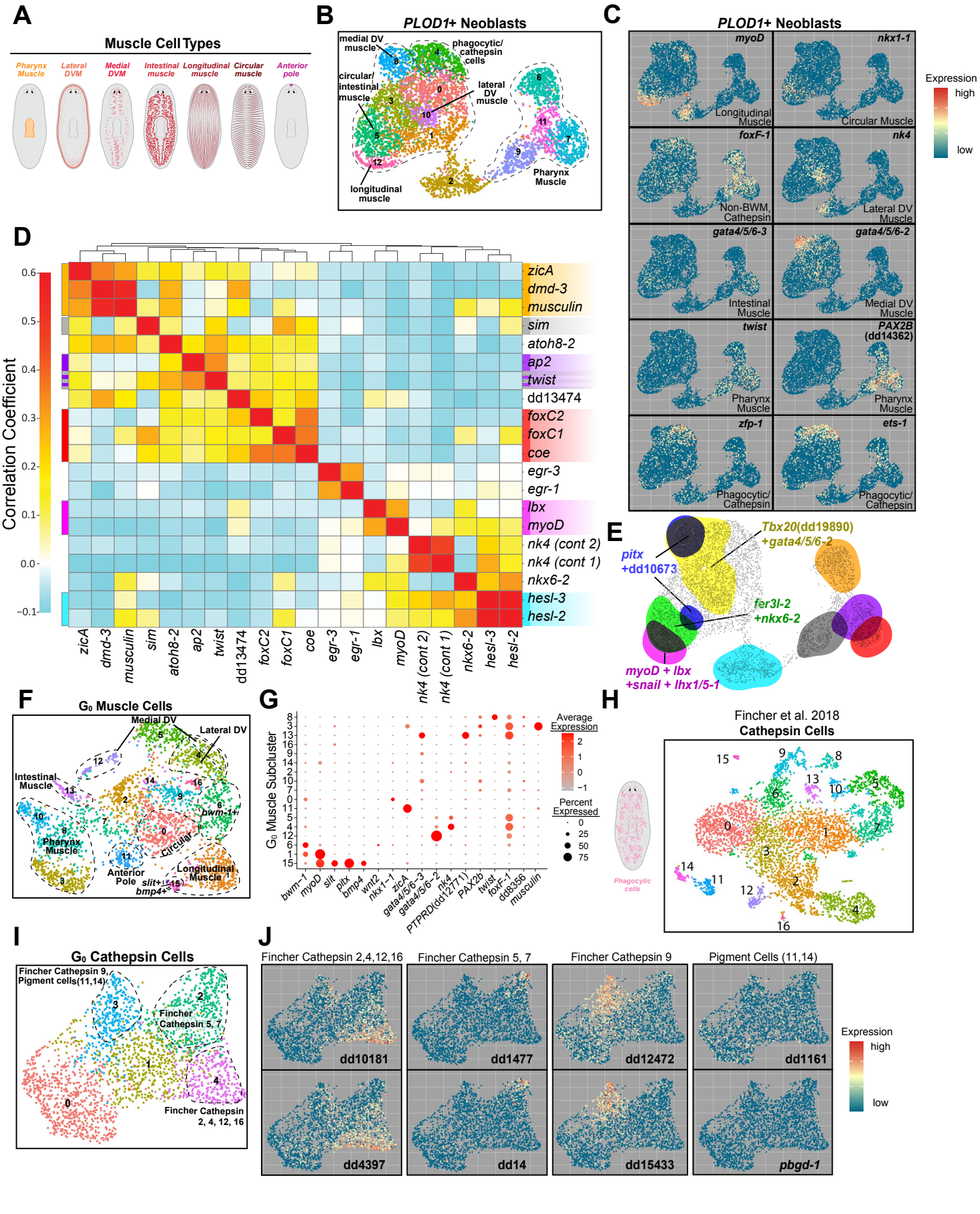

A
Muscle Cell Types
Pharynx Muscle
Lateral DVM
Medial DVM
Intestinal muscle
Longitudinal muscle
Circular muscle
Anterior pole
B
PLOD1+ Neoblasts
medial DV muscle
phagocytic/ cathepsin cells
circular/ intestinal muscle
lateral DV muscle
longitudinal muscle
Pharynx Muscle
C
PLOD1+ Neoblasts
myoD
Longitudinal Muscle
nkx1-1
Circular Muscle
foxF-1
Non-BWM, Cathepsin
nk4
Lateral DV Muscle
gata4/5/6-3
Intestinal Muscle
gata4/5/6-2
Medial DV Muscle
twist
Pharynx Muscle
PAX2B (dd14362)
Pharynx Muscle
zfp-1
Phagocytic/ Cathepsin
ets-1
Phagocytic/ Cathepsin
Expression
high
low
D
Correlation Coefficient
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
zicA
dmd-3
musculin
sim
atoh8-2
ap2
twist
dd13474
foxC2
foxC1
coe
egr-3
egr-1
lbx
myoD
nk4 (cont 2)
nk4 (cont 1)
nkx6-2
hesl-3
hesl-2
E
pitx +dd10673
Tbx20(dd19890) +gata4/5/6-2
fer3l-2 +nkx6-2
myoD + lbx +snail + lhx1/5-1
F
G0 Muscle Cells
Medial DV
Lateral DV
Intestinal Muscle
bwm-1+
Pharynx Muscle
Circular
Anterior Pole
slit+ bmp4+
Longitudinal Muscle
G
G0 Muscle Subcluster
Average Expression
Percent Expressed
bwm−1
myoD
slit
pitx
bmp4
wnt2
nkx1−1
zicA
gata4/5/6−3
gata4/5/6−2
nk4
PTPRD(dd12771)
PAX2b
twist
foxF−1
dd8356
musculin
H
Fincher et al. 2018
Cathepsin Cells
Phagocytic cells
I
G0 Cathepsin Cells
Fincher Cathepsin 9, Pigment cells(11,14)
Fincher Cathepsin 5, 7
Fincher Cathepsin 2, 4, 12, 16
J
Fincher Cathepsin 2,4,12,16
Fincher Cathepsin 5, 7
Fincher Cathepsin 9
Pigment Cells (11,14)
dd10181
dd1477
dd12472
dd1161
dd4397
dd14
dd15433
pbgd-1
Expression
high
low

Figure 2. Cell-type transcriptome diversity emerges readily in the S/G_2/M neoblast stage for muscle, but not for phagocytic/*cathepsin*+ cells.

(A) Illustration of different planarian muscle cell types.

(B) Labeled UMAP plot showing subclustered *PLOD1*+ S/G_2/M cells (encompassing muscle and *cathepsin*+ cells)

(C) UMAP plots visualizing expression of subtype-specific transcription factors on *PLOD1*+ S/G_2/M cells.

(D) Correlation analysis between transcription factors showing TF modules for different muscle subtypes. Modules were determined by computing pairwise correlations between all transcription factors (from the TF catalog) of *PLOD1*+ S/G_2/M cells.

(E) Visualization of TF module expression domain on UMAP plot for *PLOD1*+ S/G_2/M cells.

(F) Labeled UMAP plot showing subclustered G_0 muscle cells.

(G) Dot plot showing gene expression of muscle subtype markers across G_0 muscle clusters.

(H) Clusters of identified *cathepsin*+ cell populations from Fincher et al 2018 scRNA-seq data.

(I) Labeled UMAP plot showing subclustered G_0 *cathepsin*+ cells.

(J) UMAP plots showing expression of subtype-specific phagocytic markers in G_0 *cathepsin*+ clusters.

Figure 3

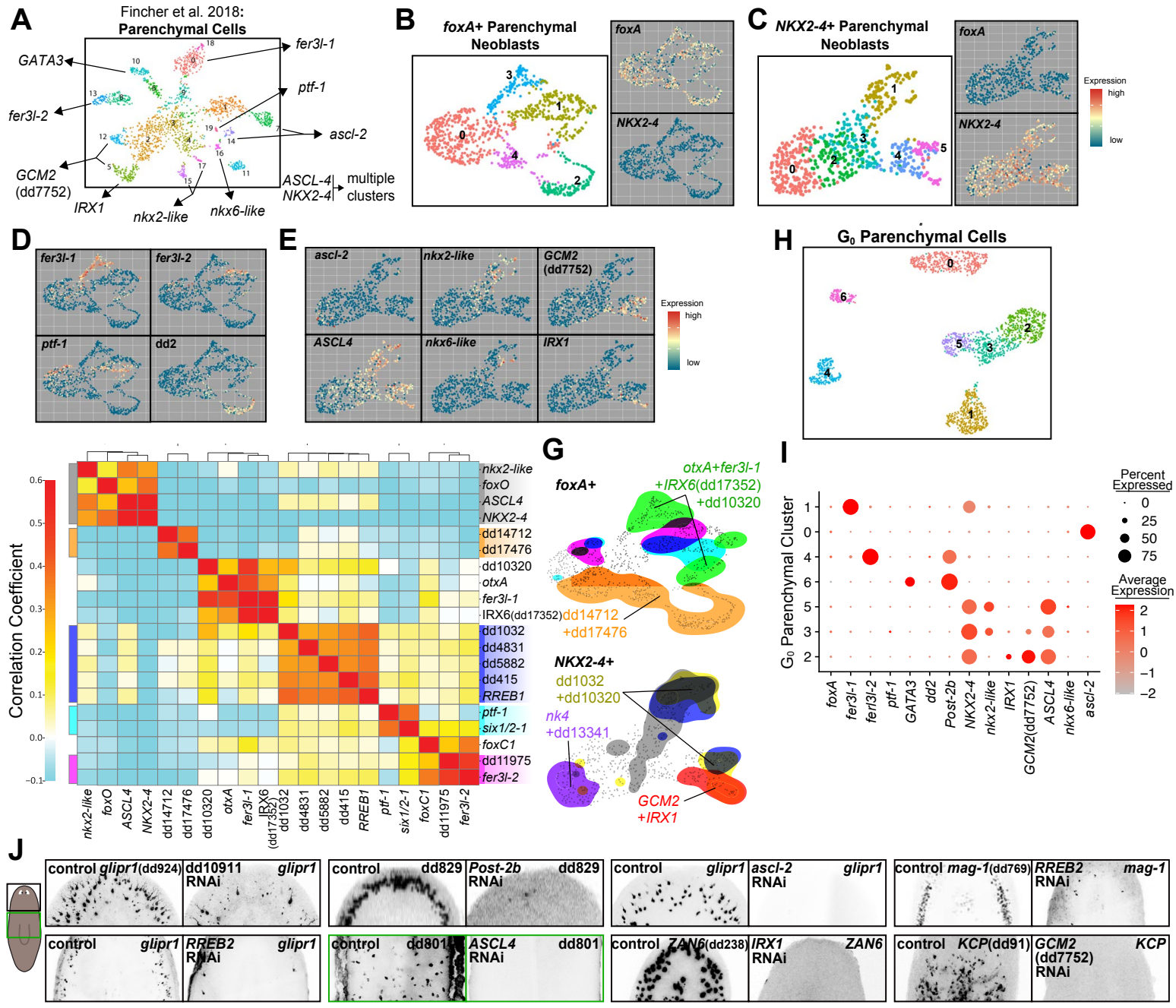

A
Fincher et al. 2018:
Parenchymal Cells
GATA3
fer3l-1
ptf-1
fer3l-2
ascl-2
GCM2
(dd7752)
ASCL-4
NKX2-4
multiple
clusters
IRX1
nkx2-like
nkx6-like
B
foxA+ Parenchymal
Neoblasts
foxA
NKX2-4
C
NKX2-4+ Parenchymal
Neoblasts
foxA
NKX2-4
Expression
high
low
D
fer3l-1
fer3l-2
ptf-1
dd2
E
ascl-2
nkx2-like
GCM2
(dd7752)
ASCL4
nkx6-like
IRX1
Expression
high
low
H
G0 Parenchymal Cells
F
Correlation Coefficient
nkx2-like
foxO
ASCL4
NKX2-4
dd14712
dd17476
dd10320
otxA
fer3l-1
IRX6(dd17352)
dd1032
dd4831
dd5882
dd415
RREB1
ptf-1
six1/2-1
foxC1
dd11975
fer3l-2
G
foxA+
otxA+fer3l-1
+IRX6(dd17352)
+dd10320
dd14712
+dd17476
NKX2-4+
dd1032
+dd10320
nk4
+dd13341
GCM2
+IRX1
I
G0 Parenchymal Cluster
Percent
Expressed
0
25
50
75
Average
Expression
2
1
0
−1
−2
foxA
fer3l-1
fer3l-2
ptf-1
GATA3
dd2
Post-2b
NKX2-4
nkx2-like
IRX1
GCM2(dd7752)
ASCL4
nkx6-like
ascl-2
J
control glipr1(dd924)
dd10911
RNAi
glipr1
control
dd829
Post-2b
RNAi
dd829
control
glipr1
ascl-2
RNAi
glipr1
control mag-1(dd769)
RREB2
RNAi
mag-1
control
glipr1
RREB2
RNAi
glipr1
control
dd801
ASCL4
RNAi
dd801
control ZAN6(dd238)
IRX1
RNAi
ZAN6
control KCP(dd91)
GCM2
(dd7752)
RNAi
KCP

Figure 3. Parenchymal cell-type diversity emerges in neoblasts.

(A) Clusters of identified parenchymal cell populations from Fincher et al 2018 scRNA-seq data, with enriched TF-encoding genes labeled.

(B and C) Subclustering of *foxA*+ and *NKX2-4*+ $S/G_2/M$ parenchymal cells with UMAP plots visualizing *foxA* and *NKX2-4* expression.

(D and E) UMAP plots visualizing transcription factor and marker gene expression across *foxA*+ and *NKX2-4* 4C parenchymal cells.

(F) Correlation analysis between transcription factors showing TF modules for different parenchymal subtypes. Modules were determined by computing pairwise correlations between all transcription factors (from TF catalog) of all $S/G_2/M$ parenchymal cells (*NKX2-4*+ and *foxA*+).

(G) Visualization of TF module expression domain on UMAP plots for all $S/G_2/M$ parenchymal cells.

(H) Labeled UMAP plot showing subclustered G_0 parenchymal cells.

(I) Dot plot showing gene expression of parenchymal subtype markers across G_0 parenchymal clusters.

(J) FISH images showing loss of parenchymal cell types following RNAi of different subtype-specific TF-encoding genes. RNAi condition in top left of each box. Marker gene in top right of each box. Head (black box) or trunk (green box) regions shown.

Figure 4

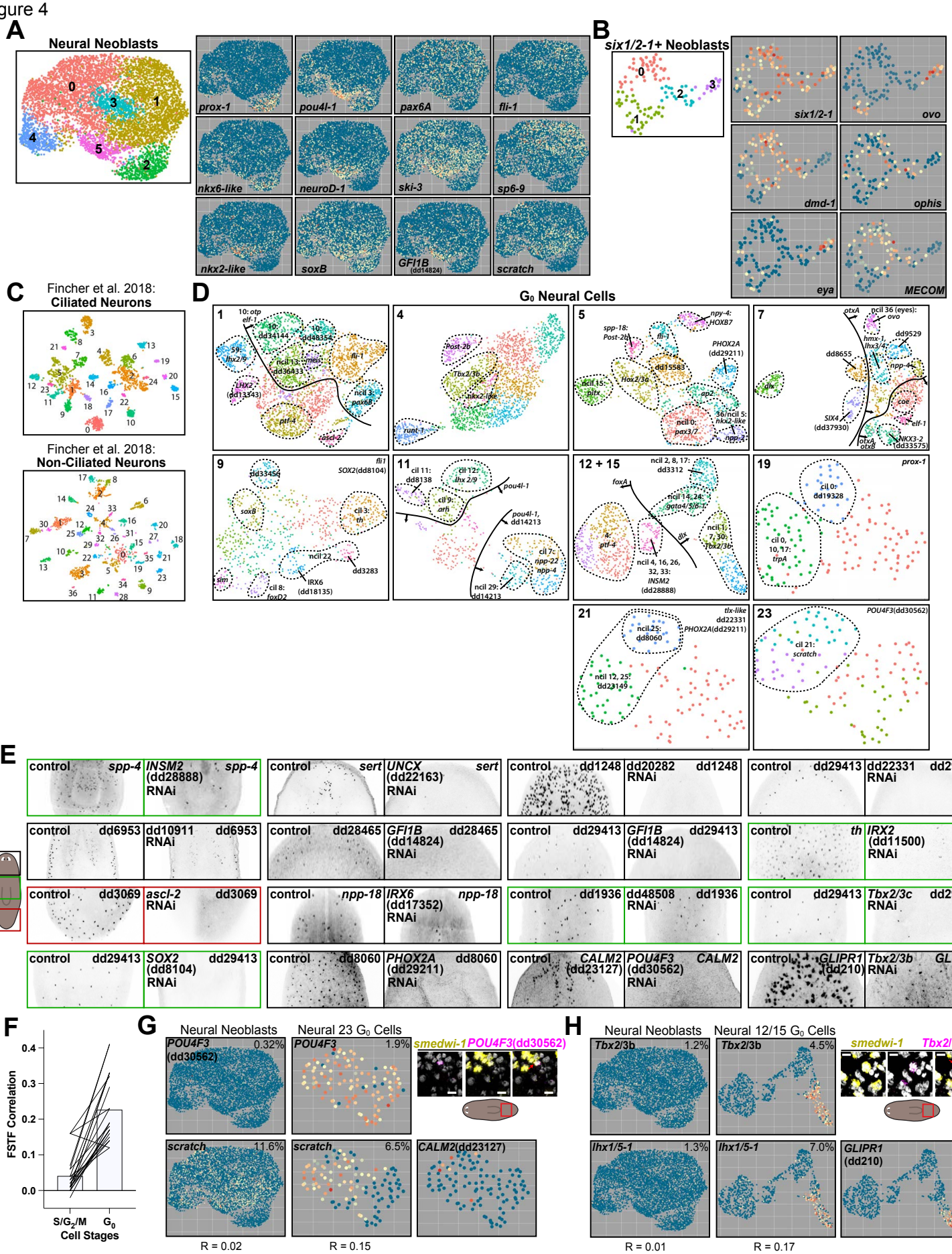

A
Neural Neoblasts
prox-1
pou4l-1
pax6A
fli-1
nkx6-like
neuroD-1
ski-3
sp6-9
nkx2-like
soxB
GFI1B (dd14824)
scratch
B
six1/2-1+ Neoblasts
six1/2-1
ovo
dmd-1
ophis
eya
MECOM
C
Fincher et al. 2018:
Ciliated Neurons
Fincher et al. 2018:
Non-Ciliated Neurons
D
G0 Neural Cells
E
control spp-4 INSM2 (dd28888) RNAi spp-4
control sert UNCX (dd22163) RNAi sert
control dd1248 dd20282 RNAi dd1248
control dd29413 dd22331 RNAi dd29413
control dd6953 dd10911 RNAi dd6953
control dd28465 GFI1B (dd14824) RNAi dd28465
control dd29413 GFI1B (dd14824) RNAi dd29413
control th IRX2 (dd11500) RNAi th
control dd3069 ascl-2 RNAi dd3069
control npp-18 IRX6 (dd17352) RNAi npp-18
control dd1936 dd48508 RNAi dd1936
control dd29413 Tbx2/3c RNAi dd29413
control dd29413 SOX2 (dd8104) RNAi dd29413
control dd8060 PHOX2A (dd29211) RNAi dd8060
control CALM2 (dd23127) POU4F3 (dd30562) RNAi CALM2
control GLIPR1 (dd210) Tbx2/3b RNAi GLIPR1
F
FSTF Correlation
S/G2/M
G0
Cell Stages
G
Neural Neoblasts
Neural 23 G0 Cells
smedwi-1 POU4F3(dd30562)
POU4F3 (dd30562) 0.32%
POU4F3 1.9%
scratch 11.6%
scratch 6.5%
CALM2(dd23127)
R = 0.02
R = 0.15
H
Neural Neoblasts
Neural 12/15 G0 Cells
smedwi-1 Tbx2/3b
Tbx2/3b 1.2%
Tbx2/3b 4.5%
lhx1/5-1 1.3%
lhx1/5-1 7.0%
GLIPR1 (dd210)
R = 0.01
R = 0.17

Figure 4. More cell states are present in neural post-mitotic progenitors than in S/G_2/M neoblasts.

(A) UMAP plot showing subclustered neural S/G_2/M cells (from “Neural” of Figure 1F) and expression of neural transcription factors on neural S/G_2/M cells.

(B) UMAP plot showing subclustered *six1/2-1*+ S/G_2/M cells and expression of genes encoding neural (*six1/2-1*) and eye *(ovo, eya*) transcription factors and germline markers (*ophis, laminA, nhr-1, MECOM*) on *six1/2-1* S/G_2/M cells.

(C) Clusters of identified neurons (ciliated and non-ciliated) from Fincher et al 2018 scRNA-seq data (Fincher et al. 2018).

(D) Labeled UMAP plot showing subclustered G_0 neurons. Numbers in upper left refer to cluster number from Figure 1F that was used for subclustering. Enriched transcription factors labeled for several subcluster regions. ncil: non-ciliated; cil: ciliated; ncil-#s and cil-#’s refer to non-ciliated and ciliated subclusters from Fincher et al 2018 scRNA-seq data. Numbers not preceded with a label refer to subclusters from “All Neural” clusters of Fincher et al 2018 scRNA-seq data.

(E) FISH images showing loss of neural cell types following RNAi of different subtype-specific TF-encoding genes. RNAi condition in top left of each box. Marker gene in top right of each box. Animals were fed dsRNA at least 8 times over 4 weeks before head and tail amputation. Regenerating trunks were allowed to regenerate at least 10 days before fixation. Head (black box), trunk (green box), or tail (red box) regions shown.

(F) Changes in expression correlation values among neural transcription factor pairs between S/G_2/M cells and G_0 cells. Each line connects the same transcription factor correlation pair between cell states.

(G) UMAP plots showing expression of transcription factors *POU4F3* and *scratch* in S/G_2/M neural cells and G_0 neural cells. *CALM2* expression in G_0 neural cells. FISH image shows co-expression of *smedwi-1* and *POU4F3*.

(H) UMAP plots showing expression of transcription factors *Tbx2/3b* and *Ihx1/5-1* in $S/G_2/M$ neural cells and G_0 neural cells. *GLIPR1* expression in G_0 neural cells. FISH image shows co-expression of *smedwi-1* and *Tbx2/3b*.

Figure 5

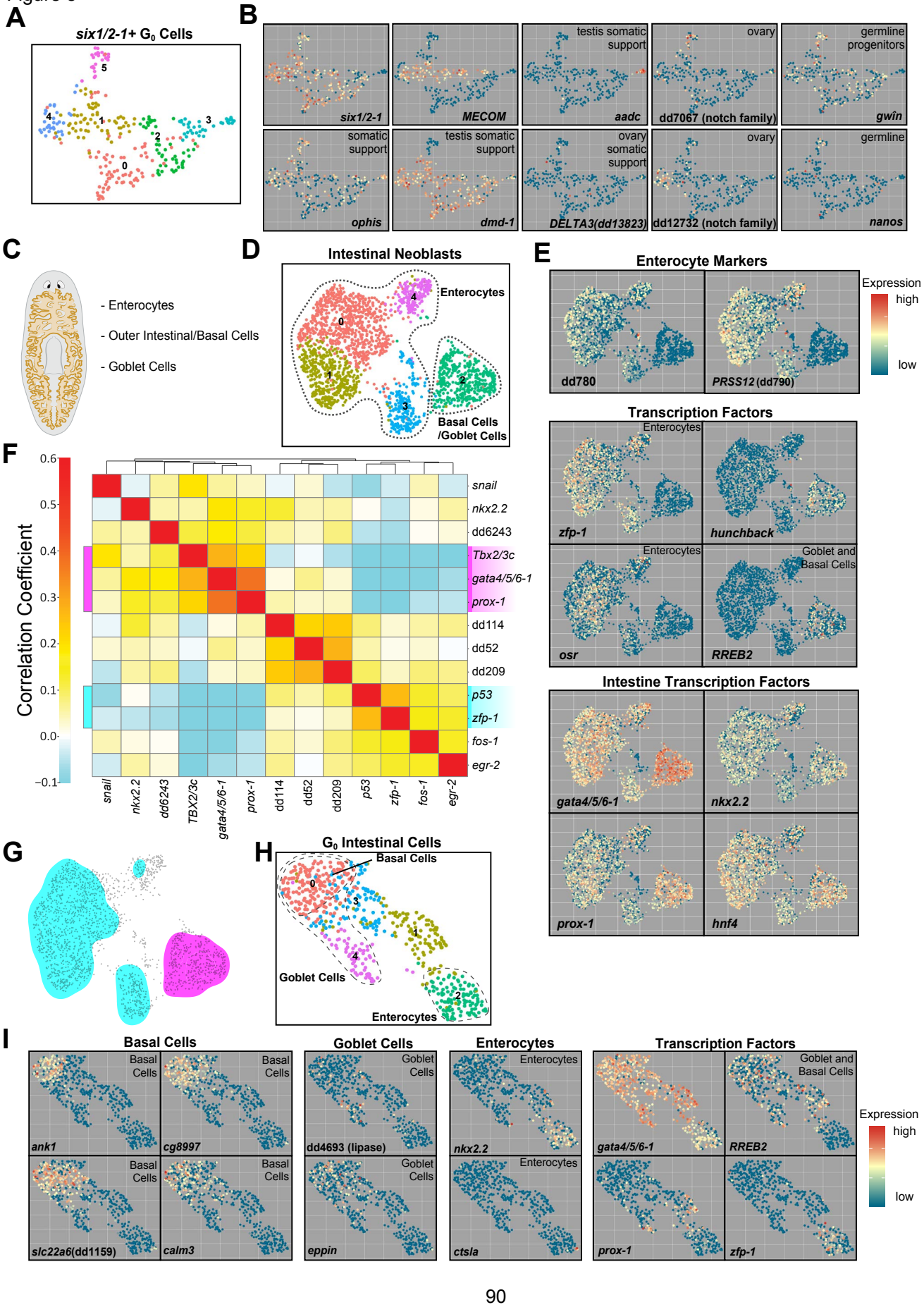

A
six1/2-1+ G0 Cells
B
six1/2-1
MECOM
testis somatic support
aadc
ovary
dd7067 (notch family)
germline progenitors
gwin
somatic support
ophis
testis somatic support
dmd-1
ovary somatic support
DELTA3(dd13823)
ovary
dd12732 (notch family)
germline
nanos
C
- Enterocytes
- Outer Intestinal/Basal Cells
- Goblet Cells
D
Intestinal Neoblasts
Enterocytes
Basal Cells /Goblet Cells
E
Enterocyte Markers
dd780
PRSS12 (dd790)
Expression
high
low
Transcription Factors
Enterocytes
zfp-1
hunchback
Enterocytes
osr
Goblet and Basal Cells
RREB2
Intestine Transcription Factors
gata4/5/6-1
nkx2.2
prox-1
hnf4
F
Correlation Coefficient
snail
nkx2.2
dd6243
Tbx2/3c
gata4/5/6-1
prox-1
dd114
dd52
dd209
p53
zfp-1
fos-1
egr-2
G
H
G0 Intestinal Cells
Basal Cells
Goblet Cells
Enterocytes
I
Basal Cells
Goblet Cells
Enterocytes
Transcription Factors
ank1
cg8997
dd4693 (lipase)
nkx2.2
gata4/5/6-1
RREB2
slc22a6(dd1159)
calm3
eppin
ctsla
prox-1
zfp-1
Expression
high
low

Figure 5. Cell transcriptomes identify two distinct S/G_2/M intestinal neoblast states and germline support cell progenitors.

(A) Labeled UMAP plot showing subclustered *six1/2-1+* germline-associated G_0 cells.

(B) UMAP plots visualizing expression of germline-associated markers in *six1/2-1+* G_0 cells.

(C) Illustration of different planarian intestinal cell types.

(D) Labeled UMAP plot showing subclustering of all S/G_2/M intestinal cells.

(E) UMAP plots visualizing expression of intestinal markers and transcription factors on intestinal S/G_2/M cells.

(F) Correlation analysis between transcription factors showing TF modules for different intestinal subtypes. Modules were determined by computing pairwise correlations between all transcription factors (from TF catalog) of intestinal S/G_2/M cells.

(G) Visualization of TF module expression domain on UMAP plot for all intestinal S/G_2/M cells.

(H) Labeled UMAP plot showing subclustering of all G_0 intestinal cells.

(I) UMAP plots visualizing expression of intestinal subtype markers and intestinal transcription factors on intestinal G_0 cells.

Figure 6

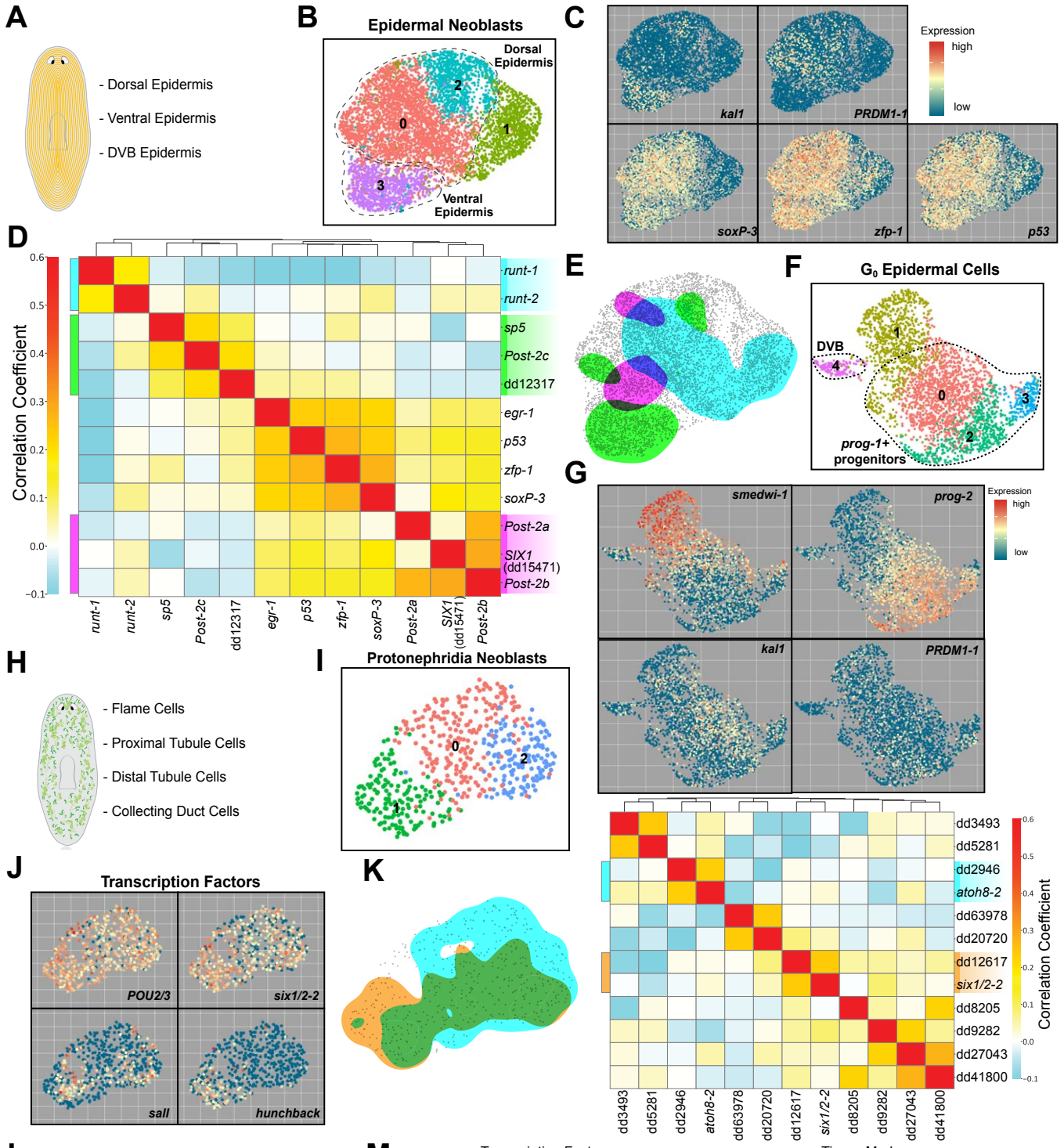

A
- Dorsal Epidermis
- Ventral Epidermis
- DVB Epidermis
B
Epidermal Neoblasts
Dorsal Epidermis
Ventral Epidermis
C
Expression
high
low
kal1
PRDM1-1
soxP-3
zfp-1
p53
D
Correlation Coefficient
runt-1
runt-2
sp5
Post-2c
dd12317
egr-1
p53
zfp-1
soxP-3
Post-2a
SIX1 (dd15471)
Post-2b
E
F
G0 Epidermal Cells
DVB
prog-1+ progenitors
G
smedwi-1
prog-2
kal1
PRDM1-1
H
- Flame Cells
- Proximal Tubule Cells
- Distal Tubule Cells
- Collecting Duct Cells
I
Protonephridia Neoblasts
J
Transcription Factors
POU2/3
six1/2-2
sall
hunchback
K
dd3493
dd5281
dd2946
atoh8-2
dd63978
dd20720
dd12617
six1/2-2
dd8205
dd9282
dd27043
dd41800
L
G0 Protonephridia Cells
Flame cells
Proximal Tubule
M
Transcription Factors
Tissue Markers
dd2920
ECE2(dd5256)
Flame Cells
dd9435
ALPPL2(dd8942)
Proximal Tubule

Figure 6. Diversification of epidermal cells arises from distinct neoblast subtypes, whereas protonephridia diversity arises from a common neoblast pool.

(A) Illustration of different planarian epidermis cell types.

(B) Labeled UMAP plot showing subclustering of all S/G_2/M epidermal cells.

(C) UMAP plots visualizing expression of epidermal subtype markers and transcription factors on epidermal S/G_2/M cells (*kal-1* ventral epidermis, *PRDM1-1* dorsal epidermis).

(D) Correlation analysis between transcription factors showing TF modules for different epidermal subtypes. Modules were determined by computing pairwise correlations between all transcription factors (from TF catalog) of intestinal S/G_2/M cells.

(E) Visualization of TF module expression domain on UMAP plot for all epidermal S/G_2/M cells.

(F) Labeled UMAP plot showing subclustering of all G_0 epidermal cells.

(G) UMAP plots visualizing expression of epidermal progression (*prog*-2) and subtype markers (*kal-1, PRDM1-1*) on epidermal G_0 cells.

(H) Illustration of different planarian protonephridia cell types.

(I) UMAP plot showing subclustering of all SG_2/M protonephridia cells.

(J) UMAP plots visualizing expression of protonephridia FSTFs on protonephridia S/ G_2/M cells.

(K) Left: Visualization of TF module expression domain on UMAP plot for all protonephridia S/G_2/M cells. Right: Correlation analysis between transcription factors showing TF modules for across S/ G_2/M protonephridia cells. Modules were determined by computing pairwise correlations between all transcription factors (from TF catalog) of protonephridia S/G_2/M cells.

(L) Labeled UMAP plot showing subclustering of all G_0 protonephridia cells.

(M) UMAP plots visualizing expression of protonephridia FSTFs and subtype specific markers on protonephridia G_0 cells (*kal-1* ventral epidermis, *PRDM1-1* dorsal epidermis).

Figure 7

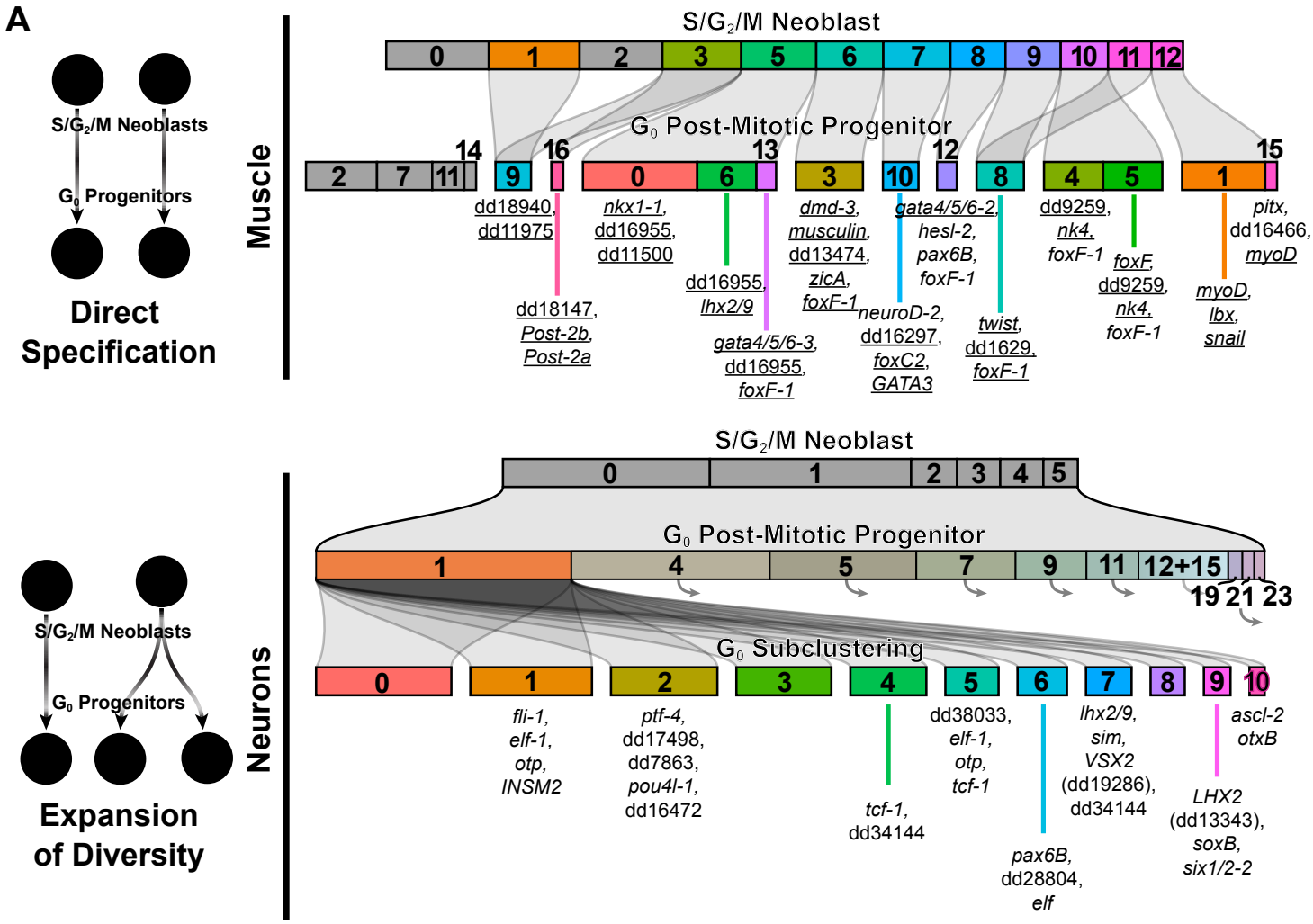

A
S/G2/M Neoblasts
G0 Progenitors
Direct Specification
Muscle
S/G2/M Neoblast
0 1 2 3 5 6 7 8 9 10 11 12
G0 Post-Mitotic Progenitor
14 16 13 12 15
2 7 11 9 0 6 3 10 8 4 5 1
dd18940, dd11975
nkx1-1, dd16955, dd11500
dd16955, lhx2/9
dmd-3, musculin, dd13474, zicA, foxF-1
gata4/5/6-2, hesl-2, pax6B, foxF-1
dd9259, nk4, foxF-1
foxF, dd9259, nk4, foxF-1
pitx, dd16466, myoD
myoD, lbx, snail
dd18147, Post-2b, Post-2a
gata4/5/6-3, dd16955, foxF-1
neuroD-2, dd16297, foxC2, GATA3
twist, dd1629, foxF-1
S/G2/M Neoblasts
G0 Progenitors
Expansion of Diversity
Neurons
S/G2/M Neoblast
0 1 2 3 4 5
G0 Post-Mitotic Progenitor
1 4 5 7 9 11 12+15 19 21 23
G0 Subclustering
0 1 2 3 4 5 6 7 8 9 10
fli-1, elf-1, otp, INSM2
ptf-4, dd17498, dd7863, pou4l-1, dd16472
tcf-1, dd34144
dd38033, elf-1, otp, tcf-1
pax6B, dd28804, elf
lhx2/9, sim, VSX2 (dd19286), dd34144
LHX2 (dd13343), soxB, six1/2-2
ascl-2 otxB
B
Protonephridia
Phagocytes
Diversity at Maturation
Mature
G0
S/G2/M
G1
Neural
Muscle
Parenchyma
Intestine
Epidermis
Direct Specification
G0 Progenitor Diversification

Figure 7. A transcription factor atlas of stem cell fate specification identifies organization principles used to generate cell-type diversity in planarians.

(A) An atlas of neoblast and G_0 post-mitotic progenitor cell types, and the transcription factor signatures that are associated with them, was used to connect G_0 cell-type identity to neoblast identity. Cluster numbers correspond to UMAP subcluster identities. Connections between neoblast and G_0 clusters correspond to matching fate identities. Transcription factor signatures for fate paths are denoted below the G_0 clusters. Underline denotes TF is expressed in corresponding neoblast cluster and G_0 cells.

(B) Cell-type diversity emerges at different stages across planarian tissues. For some tissues (muscle, parenchyma, intestine, epidermis), fate diversification occurs at the earliest possible step, in the neoblasts. For other tissues, such as protonephridia and phagocytes, fate diversification at the level of distinct transcriptomes has not emerged by the neoblast or fully by the G_0 stage and likely occurs later. For tissues such as the nervous system, substantial fate diversification at the level of distinct transcriptomes has not occurred by the neoblast stage, but emerges by the post-mitotic G_0 stage.

Figure S1

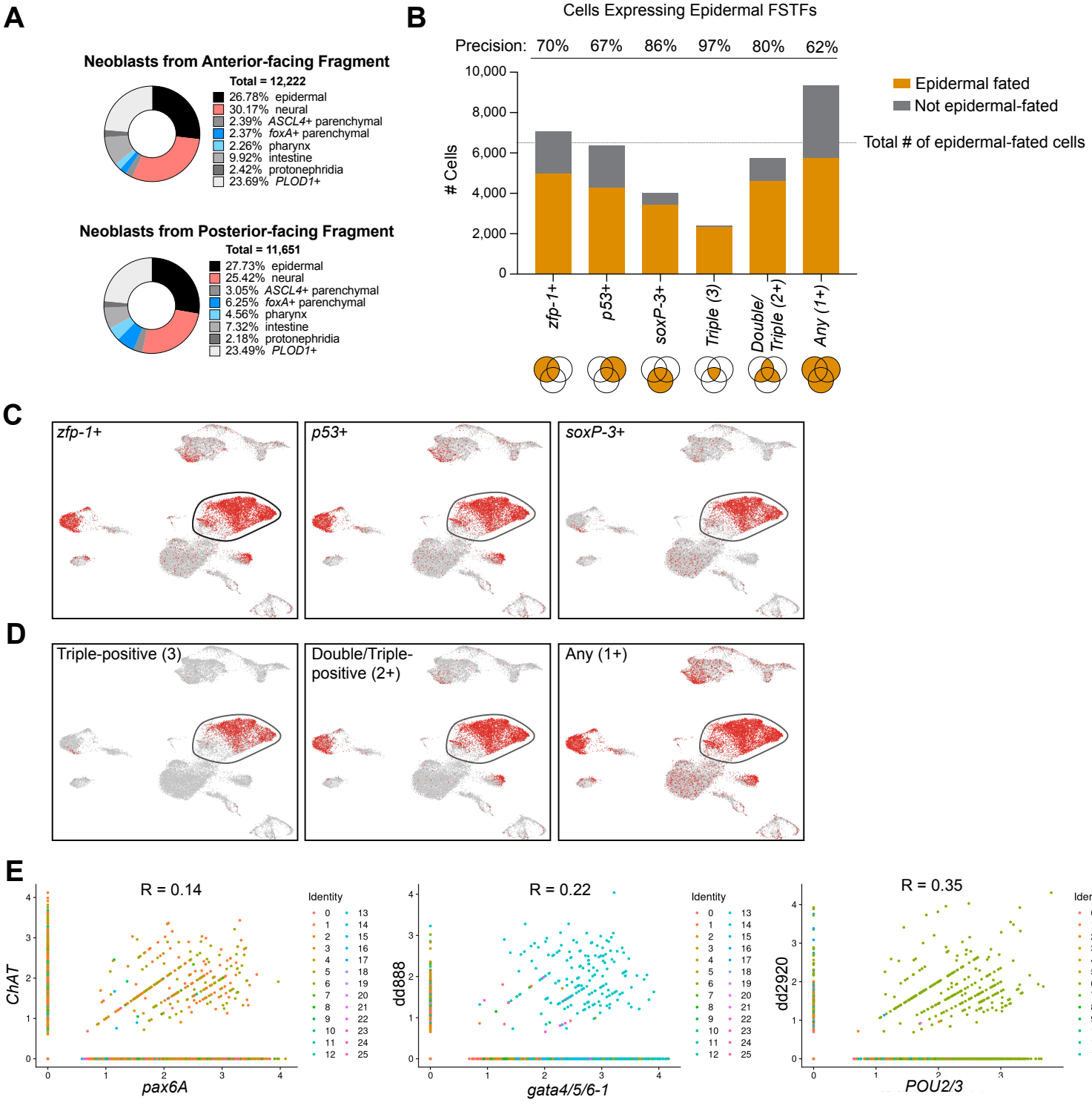
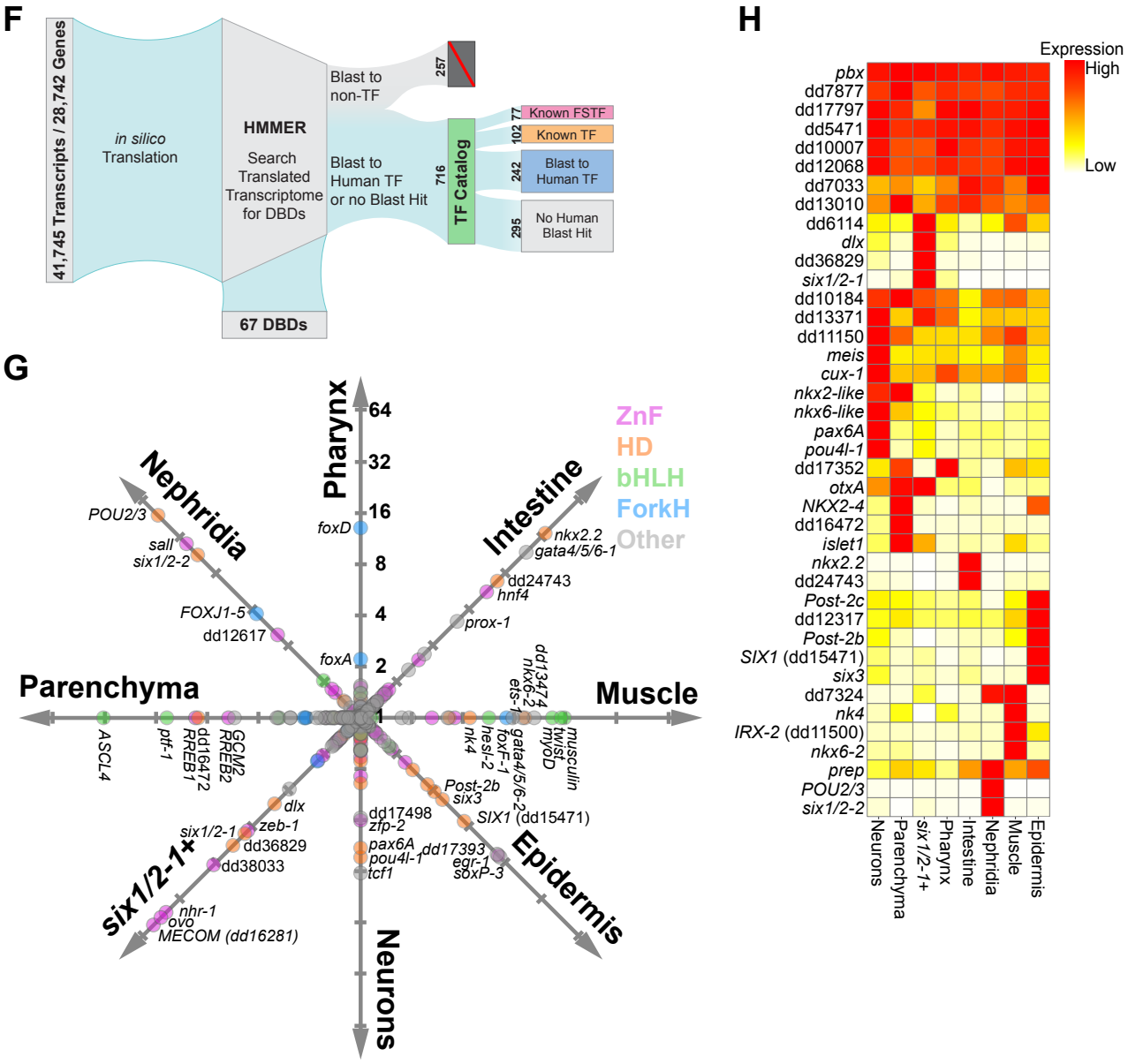

A
Neoblasts from Anterior-facing Fragment
Total = 12,222
26.78% epidermal
30.17% neural
2.39% ASCL4+ parenchymal
2.37% foxA+ parenchymal
2.26% pharynx
9.92% intestine
2.42% protonephridia
23.69% PLOD1+
Neoblasts from Posterior-facing Fragment
Total = 11,651
27.73% epidermal
25.42% neural
3.05% ASCL4+ parenchymal
6.25% foxA+ parenchymal
4.56% pharynx
7.32% intestine
2.18% protonephridia
23.49% PLOD1+
B
Cells Expressing Epidermal FSTFs
Precision: 70% 67% 86% 97% 80% 62%
Cells
10,000
8,000
6,000
4,000
2,000
0
zfp-1+
p53+
soxP-3+
Triple (3)
Double/ Triple (2+)
Any (1+)
Epidermal fated
Not epidermal-fated
Total # of epidermal-fated cells
C
zfp-1+
p53+
soxP-3+
D
Triple-positive (3)
Double/Triple-positive (2+)
Any (1+)
E
R = 0.14
ChAT
pax6A
R = 0.22
dd888
gata4/5/6-1
R = 0.35
dd2920
POU2/3
Identity

Figure S1


F
41,745 Transcripts / 28,742 Genes
in silico Translation
HMMER
Search Translated Transcriptome for DBDs
67 DBDs
Blast to non-TF
257
Blast to Human TF or no Blast Hit
716
TF Catalog
77
Known FSTF
102
Known TF
242
Blast to Human TF
295
No Human Blast Hit
G
Pharynx
Intestine
Muscle
Epidermis
Neurons
six1/2-1+
Parenchyma
Nephridia
64
32
16
8
4
2
1
ZnF
HD
bHLH
ForkH
Other
foxD
nkx2.2
gata4/5/6-1
dd24743
hnf4
prox-1
foxA
dd13474
nkx6-2
ets-1
twist
myoD
musculin
nk4
hesl-2
foxF-1
gata4/5/6-2
Post-2b
six3
SIX1 (dd15471)
dd17498
zfp-2
pax6A
pou4l-1
tcf1
dd17393
egr-1
soxP-3
dlx
six1/2-1
zeb-1
dd36829
dd38033
nhr-1
ovo
MECOM (dd16281)
POU2/3
sall
six1/2-2
FOXJ1-5
dd12617
ASCL4
ptf-1
dd16472
RREB1
GCM2
RREB2
H
Expression
High
Low
pbx
dd7877
dd17797
dd5471
dd10007
dd12068
dd7033
dd13010
dd6114
dlx
dd36829
six1/2-1
dd10184
dd13371
dd11150
meis
cux-1
nkx2-like
nkx6-like
pax6A
pou4l-1
dd17352
otxA
NKX2-4
dd16472
islet1
nkx2.2
dd24743
Post-2c
dd12317
Post-2b
SIX1 (dd15471)
six3
dd7324
nk4
IRX-2 (dd11500)
nkx6-2
prep
POU2/3
six1/2-2
Neurons
Parenchyma
six1/2-1+
Pharynx
Intestine
Nephridia
Muscle
Epidermis

Supplemental Figure 1, related to Figure 1. Analysis of sequenced S/G_2/M cells, sequenced G_0 cells, and *in silico* TF catalog construction.

(A) Fraction of neoblast classes organized by fragment from which cells originated (anterior-facing fragment or posterior-facing fragment).

(B) Percentage of epidermal fated and non-epidermal-fated neoblasts (S/G_2/M cells) expressing individual or multiple genes. Epidermal fated defined as cells in “Epidermis” cluster of S/G_2/M cells.

(C and D) UMAP plot visualizing S/G_2/M cells expressing a single epidermal FSTF or coexpressing multiple epidermal FSTFs.

(E) Correlation values for tissue-specific FSTFs and tissue-specific differentiated markers in neural, intestinal, and protonephridia G_0 cells (*pax6A* and *ChAT* coexpression for neurons; *gata4/5/6-1* and dd_888 coexpression for intestine; *POU2/3* and dd_2920 coexpression for protonephridia).

(F) Construction of *in silico* TF catalog: All planarian transcripts from the transcriptome were translated *in silico* and searched via HHMER for Pfam DNA-binding domains (DBDs) to categorize transcripts into a TF catalog. Genes were further labeled as: “Known FSTF” (previously published as planarian FSTF); “Known TF” (previously published as planarian TF, but without characterized FSTF attribution); “TF” (Not published as planarian TF, but with blast hit to Human TF”); or “No Blast” (no planarian TF characterization and no human blast hit).

(G) Expression specificity score for each TF among different neoblast classes.

(H) Heatmap showing expression specificity score for select homeodomain TFs among different neoblast classes.

Figure S2

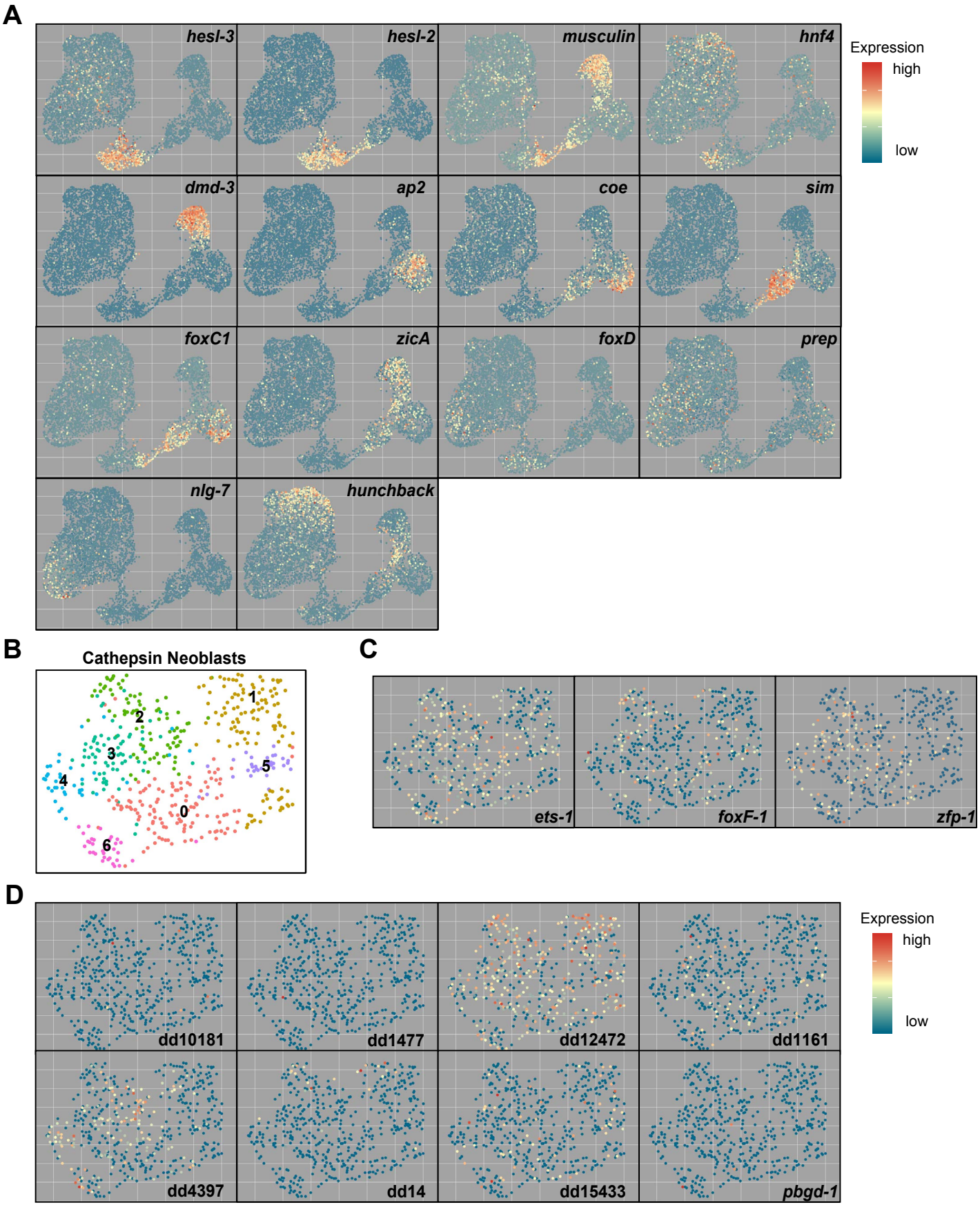

A
hesl-3
hesl-2
musculin
hnf4
dmd-3
ap2
coe
sim
foxC1
zicA
foxD
prep
nlg-7
hunchback
Expression
high
low
B
Cathepsin Neoblasts
1
2
3
4
5
0
6
C
ets-1
foxF-1
zfp-1
D
dd10181
dd1477
dd12472
dd1161
dd4397
dd14
dd15433
pbgd-1
Expression
high
low

Figure S2

E

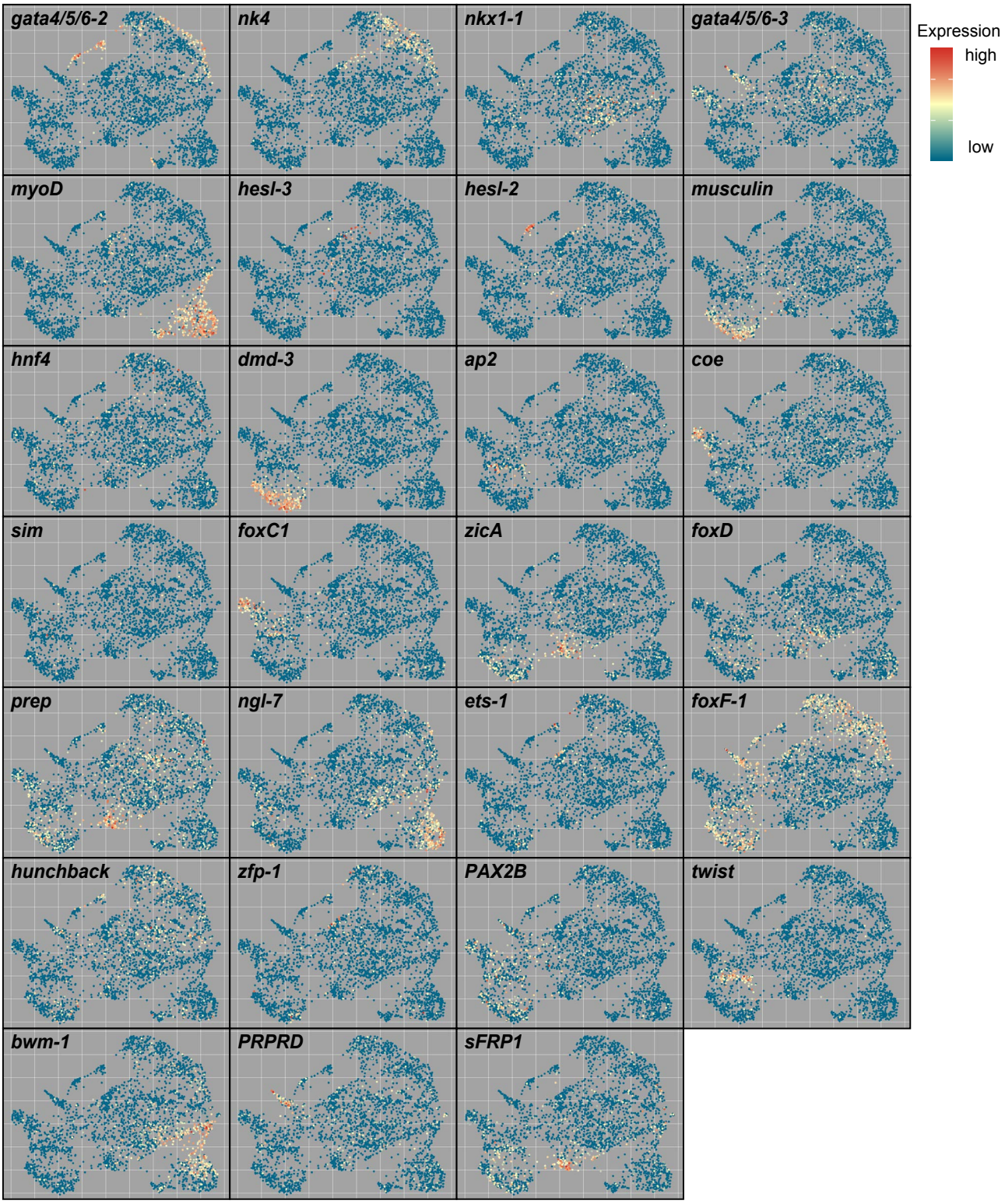
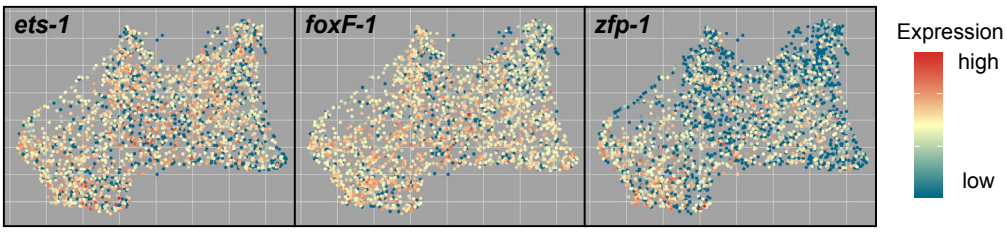

gata4/5/6-2
nk4
nkx1-1
gata4/5/6-3
Expression
high
low
myoD
hesl-3
hesl-2
musculin
hnf4
dmd-3
ap2
coe
sim
foxC1
zicA
foxD
prep
ngl-7
ets-1
foxF-1
hunchback
zfp-1
PAX2B
twist
bwm-1
PRPRD
sFRP1

Figure S2

F



Supplemental Figure 2, related to Figure 2. Expression of muscle and phagocytic transcription factors and genes in *PLOD1*+ S/G_2/M cells, G_0 muscle cells, and G_0 *cathepsin*+ cells.

(A) UMAP plots visualizing expression of muscle-associated transcription factors (e.g., *hesl-3, musculin, foxD*), positional control genes (e.g., *nlg-7*), and phagocytic-associated transcription factors (e.g., *hunchback, hnf4*) in *PLOD1*+ S/G_2/M cells.

(B) "Cathepsin" cells from *PLOD1*+ S/G_2/M cells were subjected to an additional round of subclustering. Unlabeled UMAP plot of subclustered cathepsin-fated S/G_2/M cells.

(C) UMAP plots depict expression of phagocytic-associated transcription factors in subclustered cathepsin-fated S/G_2/M cells.

(D) UMAP plots visualizing expression of subtype-specific phagocytic cell markers in *PLOD1+/cathepsin*+ S/G_2/M cells. Little neoblast diversity is apparent in cathepsin-fated S/G_2/M cells.

(E) UMAP plots visualizing expression of muscle-associated transcription factors, positional control genes, and phagocytic-associated transcription factors in G_0 muscle cells.

(F) UMAP plots visualizing expression of phagocytic-associated transcription factors in cathepsin-fated G_0 cells.

Figure S3

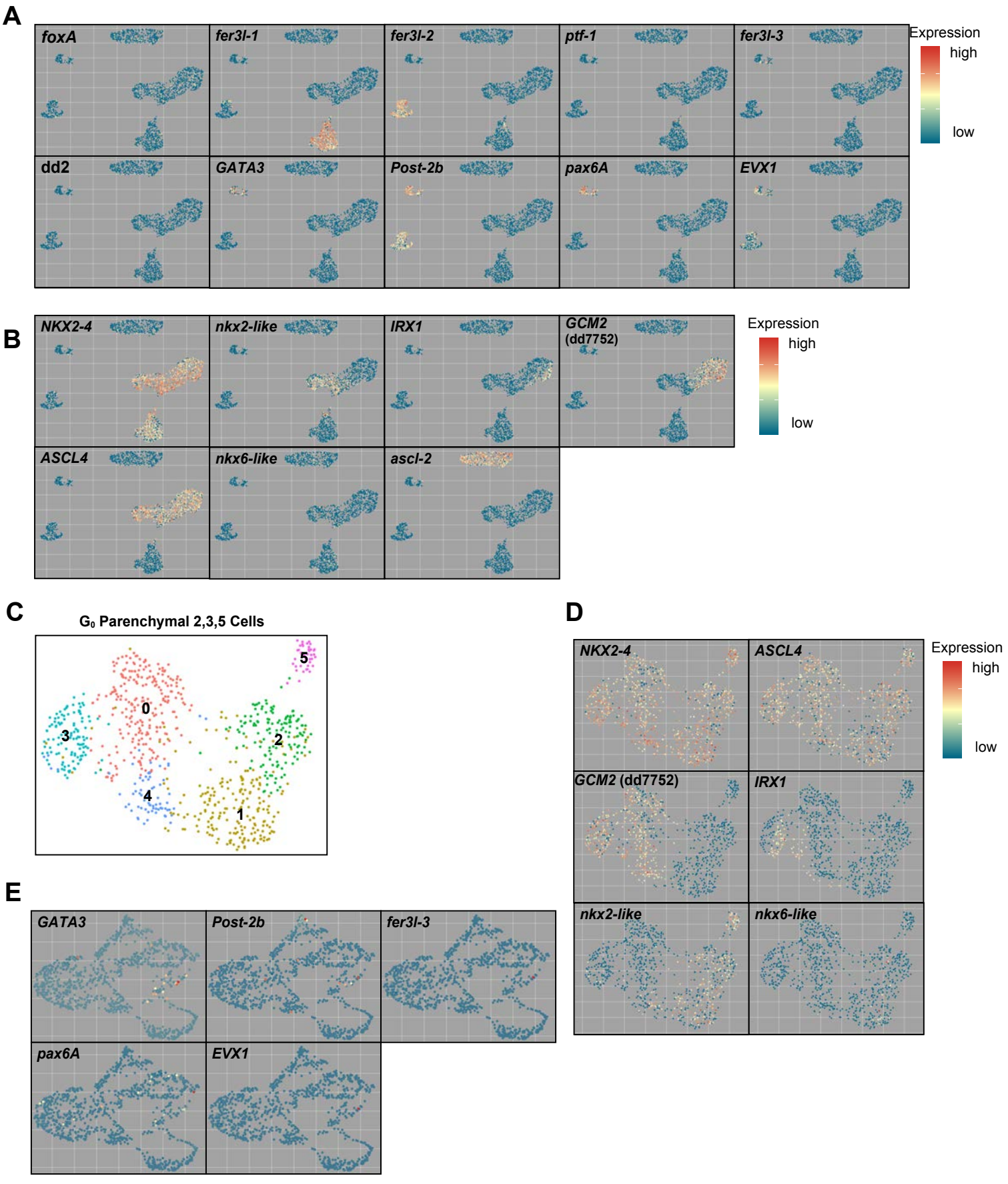

A
foxA
fer3l-1
fer3l-2
ptf-1
fer3l-3
dd2
GATA3
Post-2b
pax6A
EVX1
Expression
high
low
B
NKX2-4
nkx2-like
IRX1
GCM2
(dd7752)
ASCL4
nkx6-like
ascl-2
Expression
high
low
C
G0 Parenchymal 2,3,5 Cells
0
1
2
3
4
5
D
NKX2-4
ASCL4
GCM2 (dd7752)
IRX1
nkx2-like
nkx6-like
Expression
high
low
E
GATA3
Post-2b
fer3l-3
pax6A
EVX1

Figure S3

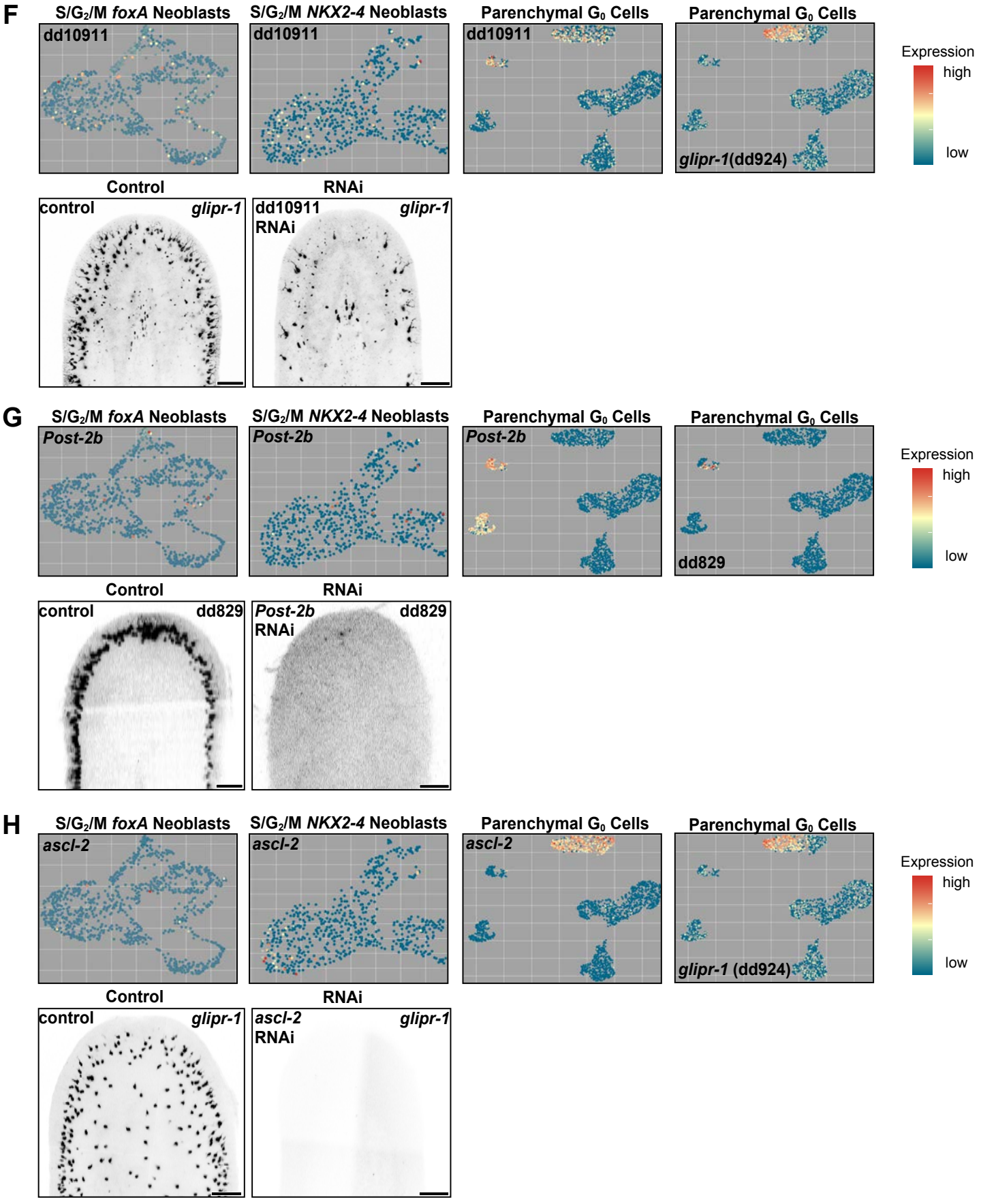

F
S/G2/M foxA Neoblasts
S/G2/M NKX2-4 Neoblasts
Parenchymal G0 Cells
Parenchymal G0 Cells
dd10911
dd10911
dd10911
glipr-1(dd924)
Expression
high
low
Control
RNAi
control
glipr-1
dd10911
RNAi
glipr-1
G
S/G2/M foxA Neoblasts
S/G2/M NKX2-4 Neoblasts
Parenchymal G0 Cells
Parenchymal G0 Cells
Post-2b
Post-2b
Post-2b
dd829
Expression
high
low
Control
RNAi
control
dd829
Post-2b
RNAi
dd829
H
S/G2/M foxA Neoblasts
S/G2/M NKX2-4 Neoblasts
Parenchymal G0 Cells
Parenchymal G0 Cells
ascl-2
ascl-2
ascl-2
glipr-1 (dd924)
Expression
high
low
Control
RNAi
control
glipr-1
ascl-2
RNAi
glipr-1

Figure S3

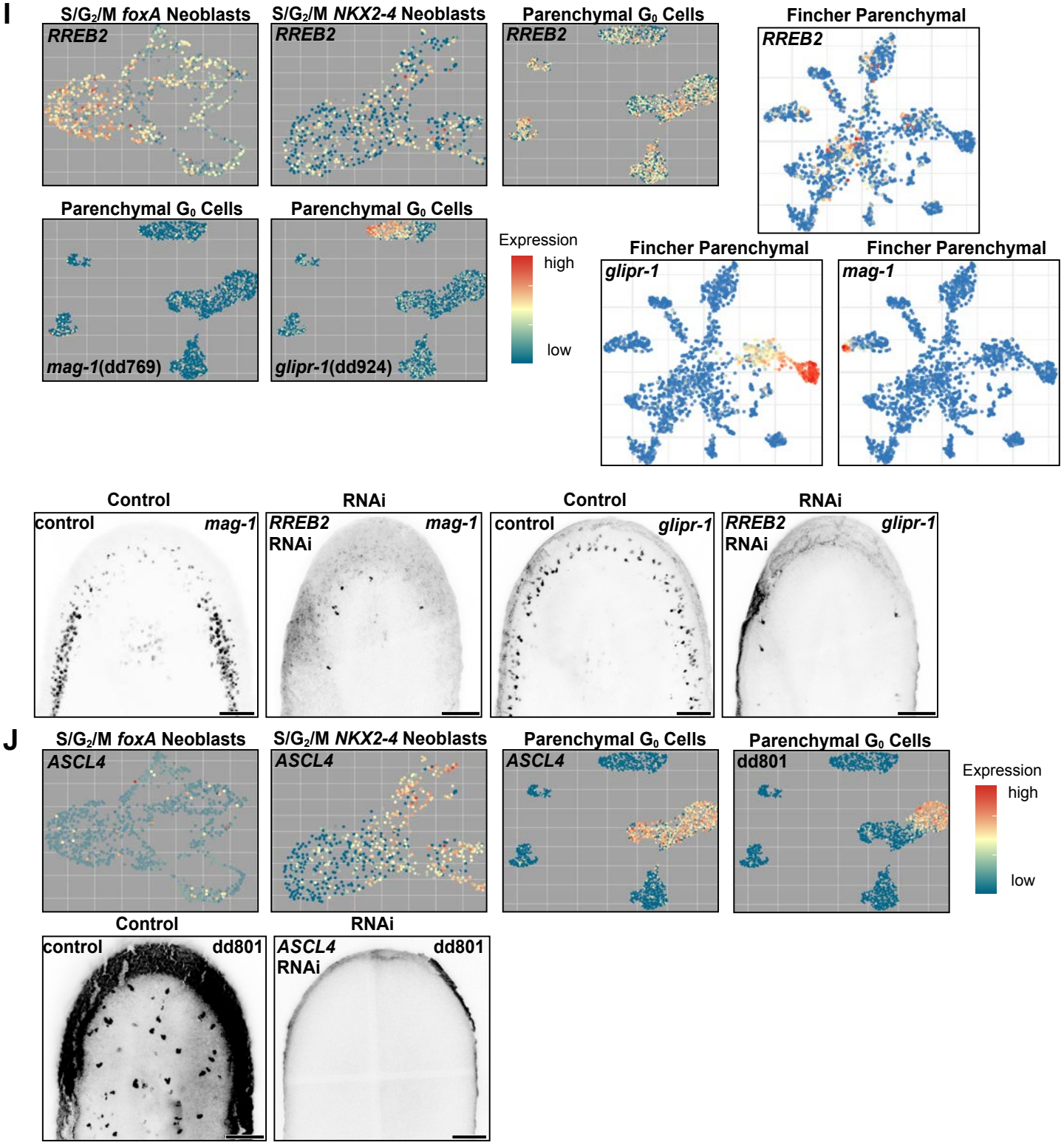
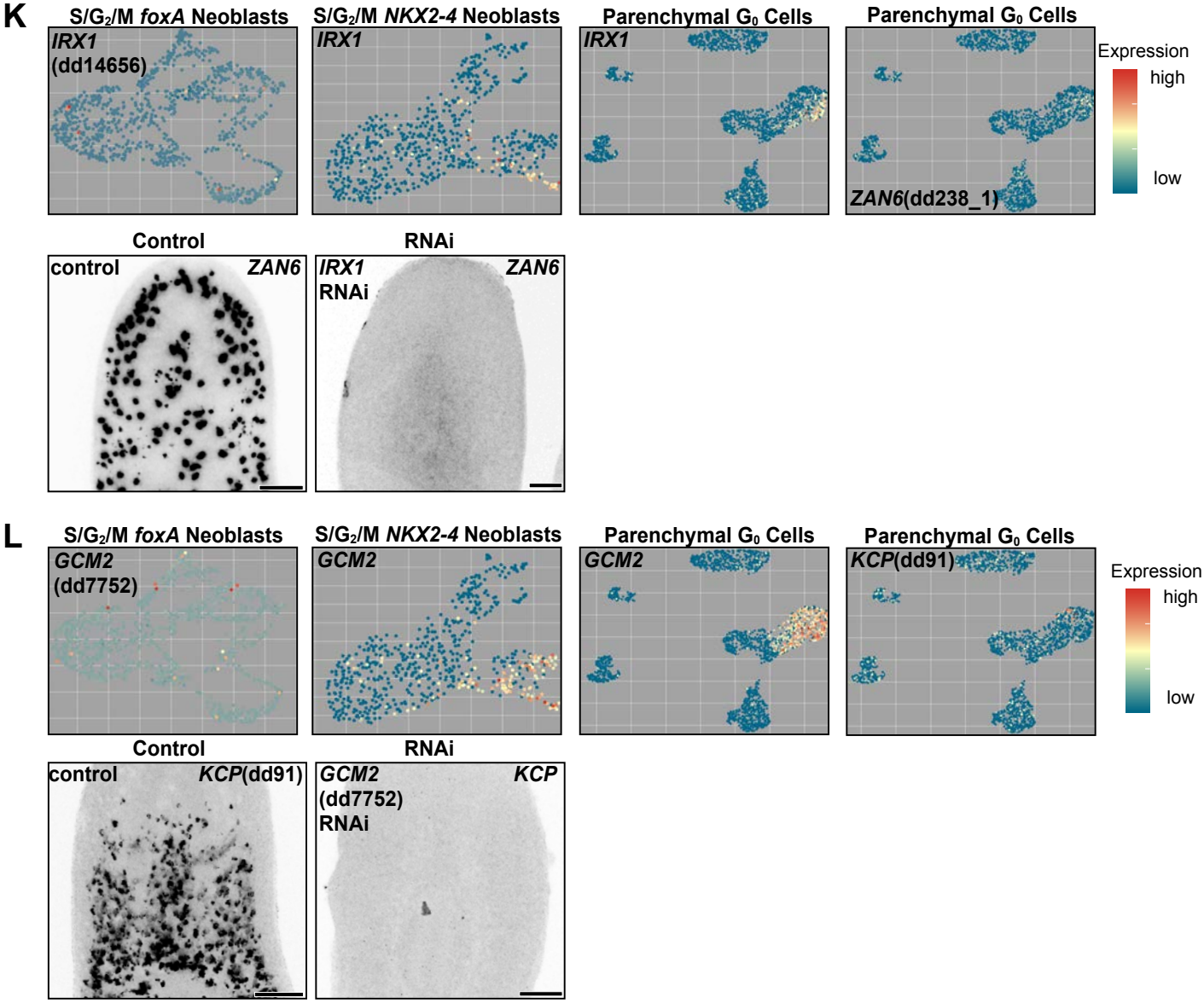

I
S/G2/M foxA Neoblasts
S/G2/M NKX2-4 Neoblasts
Parenchymal G0 Cells
Fincher Parenchymal
RREB2
Parenchymal G0 Cells
Parenchymal G0 Cells
mag-1(dd769)
glipr-1(dd924)
Expression
high
low
Fincher Parenchymal
Fincher Parenchymal
glipr-1
mag-1
Control
RNAi
Control
RNAi
control
mag-1
RREB2
RNAi
mag-1
control
glipr-1
RREB2
RNAi
glipr-1
J
S/G2/M foxA Neoblasts
S/G2/M NKX2-4 Neoblasts
Parenchymal G0 Cells
Parenchymal G0 Cells
ASCL4
ASCL4
ASCL4
dd801
Expression
high
low
Control
RNAi
control
dd801
ASCL4
RNAi
dd801

Figure S3


K
S/G2/M foxA Neoblasts
S/G2/M NKX2-4 Neoblasts
Parenchymal G0 Cells
Parenchymal G0 Cells
IRX1 (dd14656)
IRX1
IRX1
ZAN6(dd238_1)
Expression
high
low
Control
RNAi
control
ZAN6
IRX1 RNAi
ZAN6
L
S/G2/M foxA Neoblasts
S/G2/M NKX2-4 Neoblasts
Parenchymal G0 Cells
Parenchymal G0 Cells
GCM2 (dd7752)
GCM2
GCM2
KCP(dd91)
Expression
high
low
Control
RNAi
control
KCP(dd91)
GCM2 (dd7752) RNAi
KCP

Supplemental Figure 3, related to Figure 3. Analysis of parenchymal S/G_2/M cells, G_0 **cells, and functional characterization of parenchymal enriched TFs in cell fate specification.**

(A and B) UMAP plot visualizing expression of TFs or other markers (e.g., dd_2) enriched in different subtypes of parenchymal G_0 cells.

(C) Subclusters 2, 3, and 5 of parenchymal G_0 cells were subjected to an additional round of subclustering together.

(D) UMAP plots depict expression of parenchymal-associated transcription factors in subclustered parenchymal G_0 cells from clusters 2, 3, and 5 specifically (combined).

(E) S/G_2/M parenchymal cell UMAP plot showing expression of transcription factors known to define one subtype of differentiated parenchymal cells (Fincher et al 2018 scRNA-seq data clusters 6 and 10).

(F) Top: expression of dd_10911 (TF-encoding gene) in all parenchymal S/G_2/M cells (*foxA*+ population and *NKX2-4* population) and all parenchymal G_0 cells. Expression of *glipr1* (dd_924) in parenchymal G_0 cells. Bottom: RNAi showing loss of *glipr1*+ cells following dd_10911 RNAi.

(G) Top: expression of *Post-2b* in all parenchymal S/G_2/M cells (*foxA*+ population and *NKX2-4* population) and all parenchymal G_0 cells. Expression of (dd_829) in parenchymal G_0 cells. Bottom: RNAi showing loss of dd_829 cells following *Post-2b* RNAi.

(H) Top: expression of *ascl-2* in all parenchymal S/G_2/M cells (*foxA*+ population and *NKX2-4* population) and all parenchymal G_0 cells. Expression of *glipr1* in parenchymal G_0 cells. Bottom: RNAi showing loss of *glipr1*+ cells following *ascl-2* RNAi.

(I) Top panels: expression of genes encoding parenchymal-associated TFs (*RREB2*) and differentiated parenchymal subtype markers (*mag-1*, *glipr-1*) in all parenchymal S/G_2/M cells (*foxA*+ population and *NKX2-4* population), all parenchymal G_0 cells, and all parenchymal cells from Fincher et al 2018 data. Bottom panels: RNAi showing loss of *glipr1*+ cells and *mag-1*+ cells following *RREB2* RNAi.

(J) Top: expression of *ASCL4* in all parenchymal S/G_2/M cells (*foxA*+ population and *NKX2-4* population) and all parenchymal G_0 cells. Expression of dd_801 in parenchymal G_0 cells. Bottom: RNAi showing loss of dd_801 cells following *ASCL4* RNAi.

(K) Top: expression of *IRX1* (dd_14656) in all parenchymal S/G_2/M cells (*foxA*+ population and *NKX2-4* population) and all parenchymal G_0 cells. Expression of *ZAN6* (dd_238_1) in parenchymal G_0 cells. Bottom: RNAi showing loss of *ZAN6*+ cells following *IRX1* RNAi.

(L) Top: expression of *GCM2* (dd_7752) in all parenchymal S/G_2/M cells (*foxA*+ population and *NKX2-4* population) and all parenchymal G_0 cells. Expression of *KCP* (dd_91) in parenchymal G_0 cells. Bottom: RNAi showing loss of *KCP*+ cells following *GCM2* RNAi.

(F-L) FISH images are the same as in Figure 3J, but zoomed out.

Figure S4

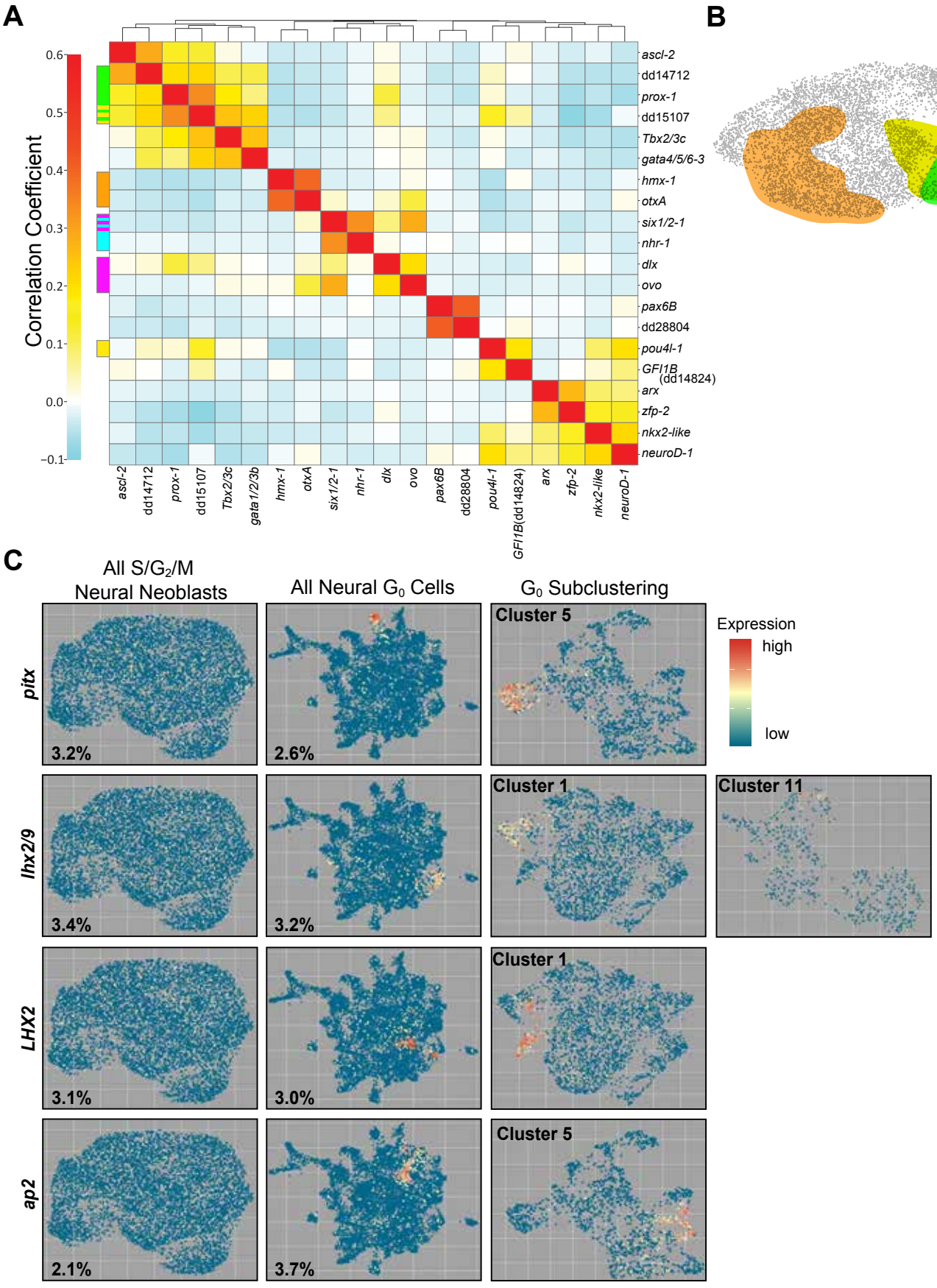

A
Correlation Coefficient
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
ascl-2
dd14712
prox-1
dd15107
Tbx2/3c
gata1/2/3b
hmx-1
otxA
six1/2-1
nhr-1
dlx
ovo
pax6B
dd28804
pou4l-1
GFI1B(dd14824)
arx
zfp-2
nkx2-like
neuroD-1
B
C
All S/G2/M Neural Neoblasts
All Neural G0 Cells
G0 Subclustering
Expression
high
low
pitx
lhx2/9
LHX2
ap2
Cluster 5
Cluster 1
Cluster 11
3.2%
2.6%
3.4%
3.2%
3.1%
3.0%
2.1%
3.7%

Figure S4

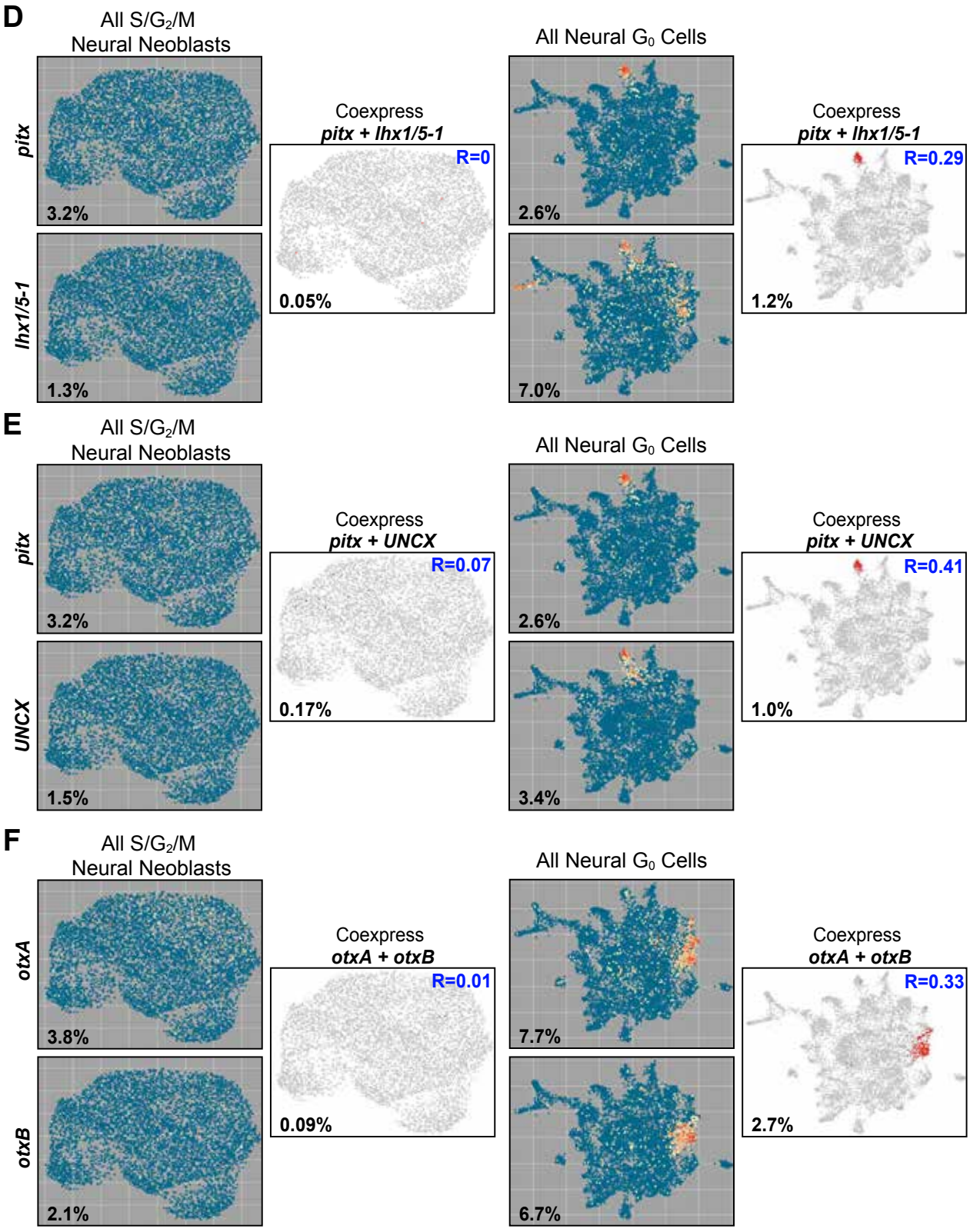

D
All S/G2/M Neural Neoblasts
All Neural G0 Cells
pitx
lhx1/5-1
Coexpress pitx + lhx1/5-1
R=0
R=0.29
3.2%
1.3%
0.05%
2.6%
7.0%
1.2%
E
All S/G2/M Neural Neoblasts
All Neural G0 Cells
pitx
UNCX
Coexpress pitx + UNCX
R=0.07
R=0.41
3.2%
1.5%
0.17%
2.6%
3.4%
1.0%
F
All S/G2/M Neural Neoblasts
All Neural G0 Cells
otxA
otxB
Coexpress otxA + otxB
R=0.01
R=0.33
3.8%
2.1%
0.09%
7.7%
6.7%
2.7%

Figure S4

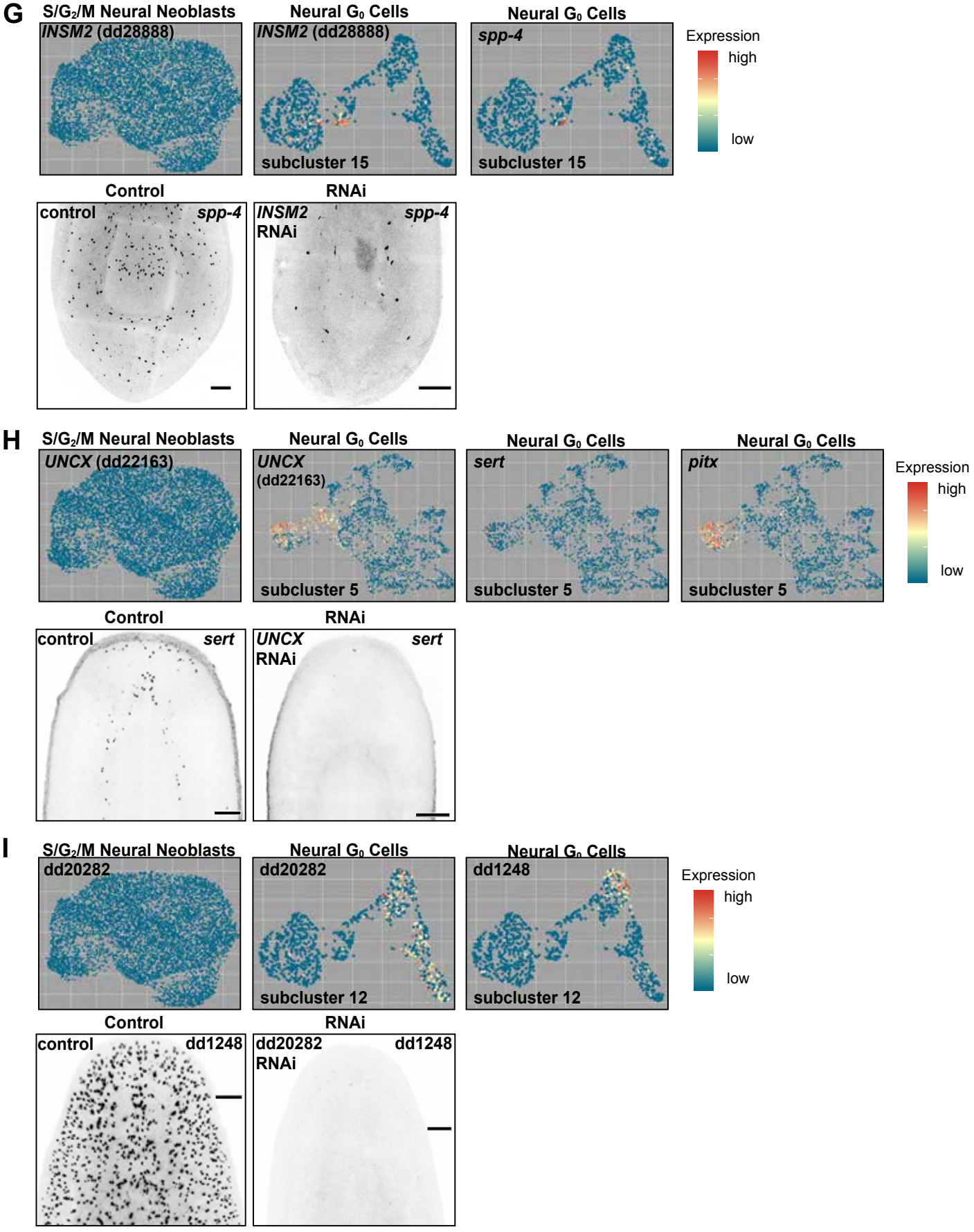

G
S/G2/M Neural Neoblasts
INSM2 (dd28888)
Neural G0 Cells
INSM2 (dd28888)
subcluster 15
Neural G0 Cells
spp-4
subcluster 15
Expression
high
low
Control
RNAi
control
spp-4
INSM2
RNAi
spp-4
H
S/G2/M Neural Neoblasts
UNCX (dd22163)
Neural G0 Cells
UNCX
(dd22163)
subcluster 5
Neural G0 Cells
sert
subcluster 5
Neural G0 Cells
pitx
subcluster 5
Expression
high
low
Control
RNAi
control
sert
UNCX
RNAi
sert
I
S/G2/M Neural Neoblasts
dd20282
Neural G0 Cells
dd20282
subcluster 12
Neural G0 Cells
dd1248
subcluster 12
Expression
high
low
Control
RNAi
control
dd1248
dd20282
RNAi
dd1248

Figure S4

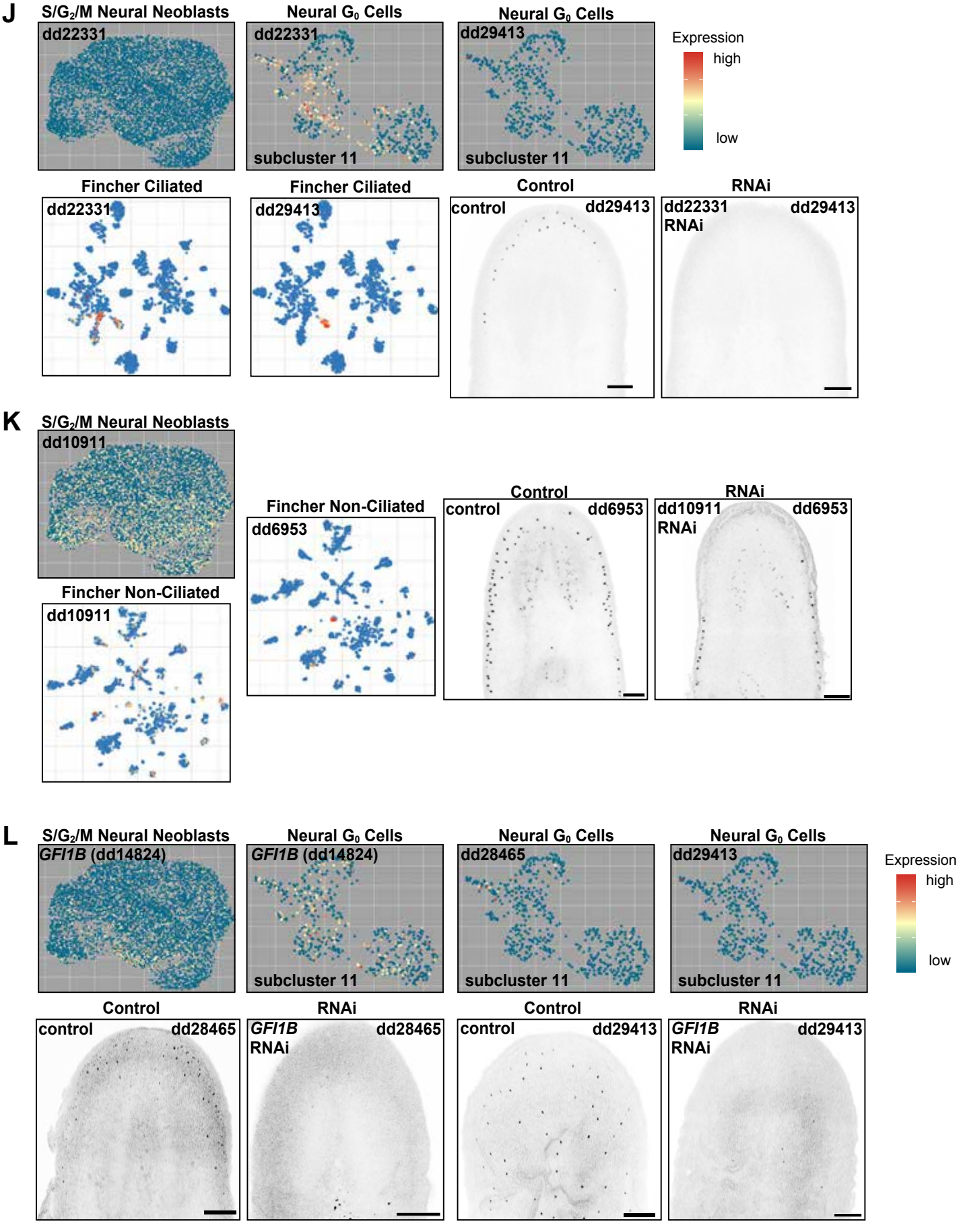

J
S/G2/M Neural Neoblasts
dd22331
Neural G0 Cells
dd22331
subcluster 11
Neural G0 Cells
dd29413
subcluster 11
Expression
high
low
Fincher Ciliated
dd22331
Fincher Ciliated
dd29413
Control
control
dd29413
RNAi
dd22331
RNAi
dd29413
K
S/G2/M Neural Neoblasts
dd10911
Fincher Non-Ciliated
dd10911
Fincher Non-Ciliated
dd6953
Control
control
dd6953
RNAi
dd10911
RNAi
dd6953
L
S/G2/M Neural Neoblasts
GFI1B (dd14824)
Neural G0 Cells
GFI1B (dd14824)
subcluster 11
Neural G0 Cells
dd28465
subcluster 11
Neural G0 Cells
dd29413
subcluster 11
Expression
high
low
Control
control
dd28465
RNAi
GFI1B
RNAi
dd28465
Control
control
dd29413
RNAi
GFI1B
RNAi
dd29413

Figure S4

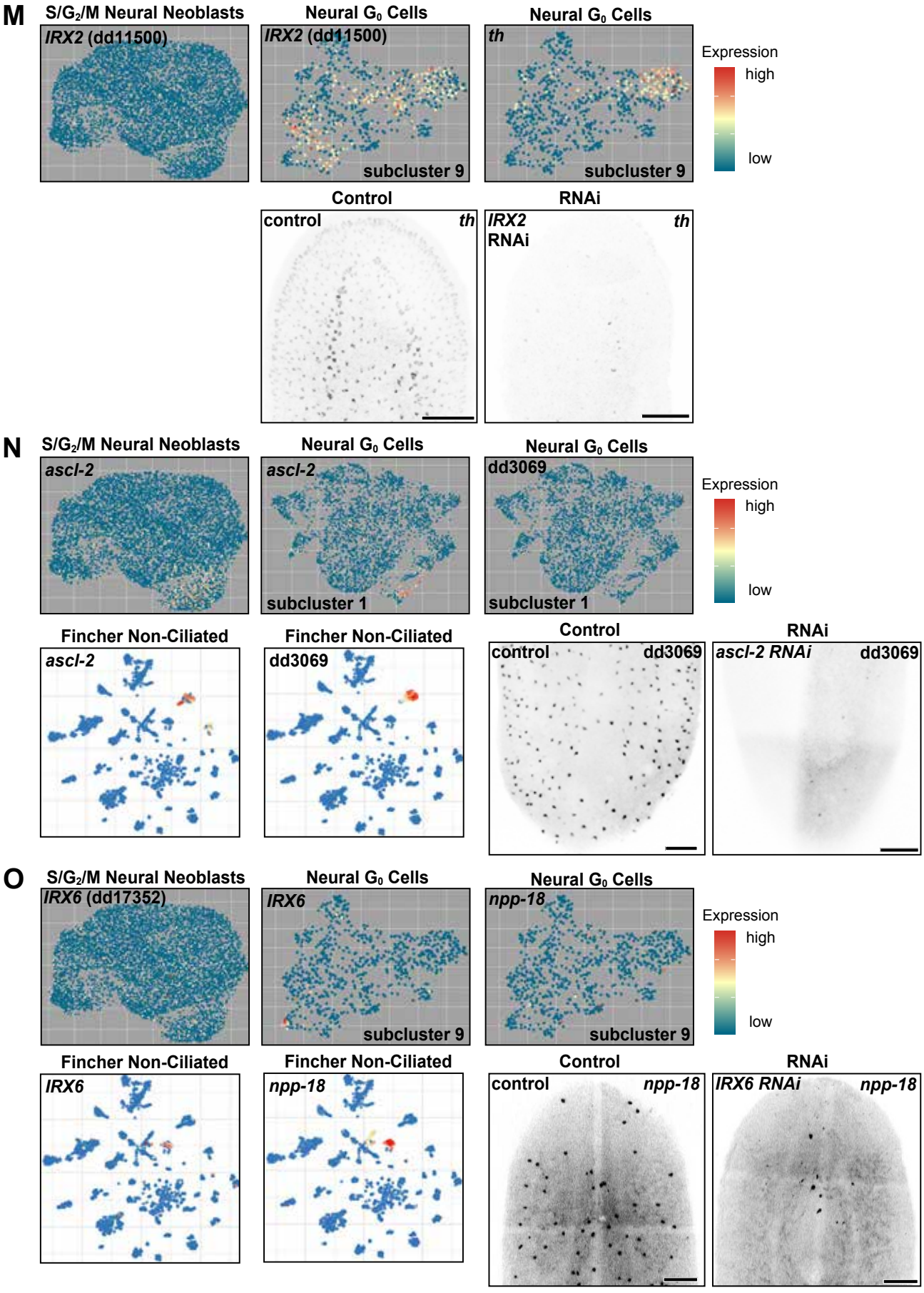

M
S/G2/M Neural Neoblasts
IRX2 (dd11500)
Neural G0 Cells
IRX2 (dd11500)
subcluster 9
Neural G0 Cells
th
subcluster 9
Expression
high
low
Control
control
th
RNAi
IRX2
RNAi
th
N
S/G2/M Neural Neoblasts
ascl-2
Neural G0 Cells
ascl-2
subcluster 1
Neural G0 Cells
dd3069
subcluster 1
Expression
high
low
Fincher Non-Ciliated
ascl-2
Fincher Non-Ciliated
dd3069
Control
control
dd3069
RNAi
ascl-2 RNAi
dd3069
O
S/G2/M Neural Neoblasts
IRX6 (dd17352)
Neural G0 Cells
IRX6
subcluster 9
Neural G0 Cells
npp-18
subcluster 9
Expression
high
low
Fincher Non-Ciliated
IRX6
Fincher Non-Ciliated
npp-18
Control
control
npp-18
RNAi
IRX6 RNAi
npp-18

Figure S4

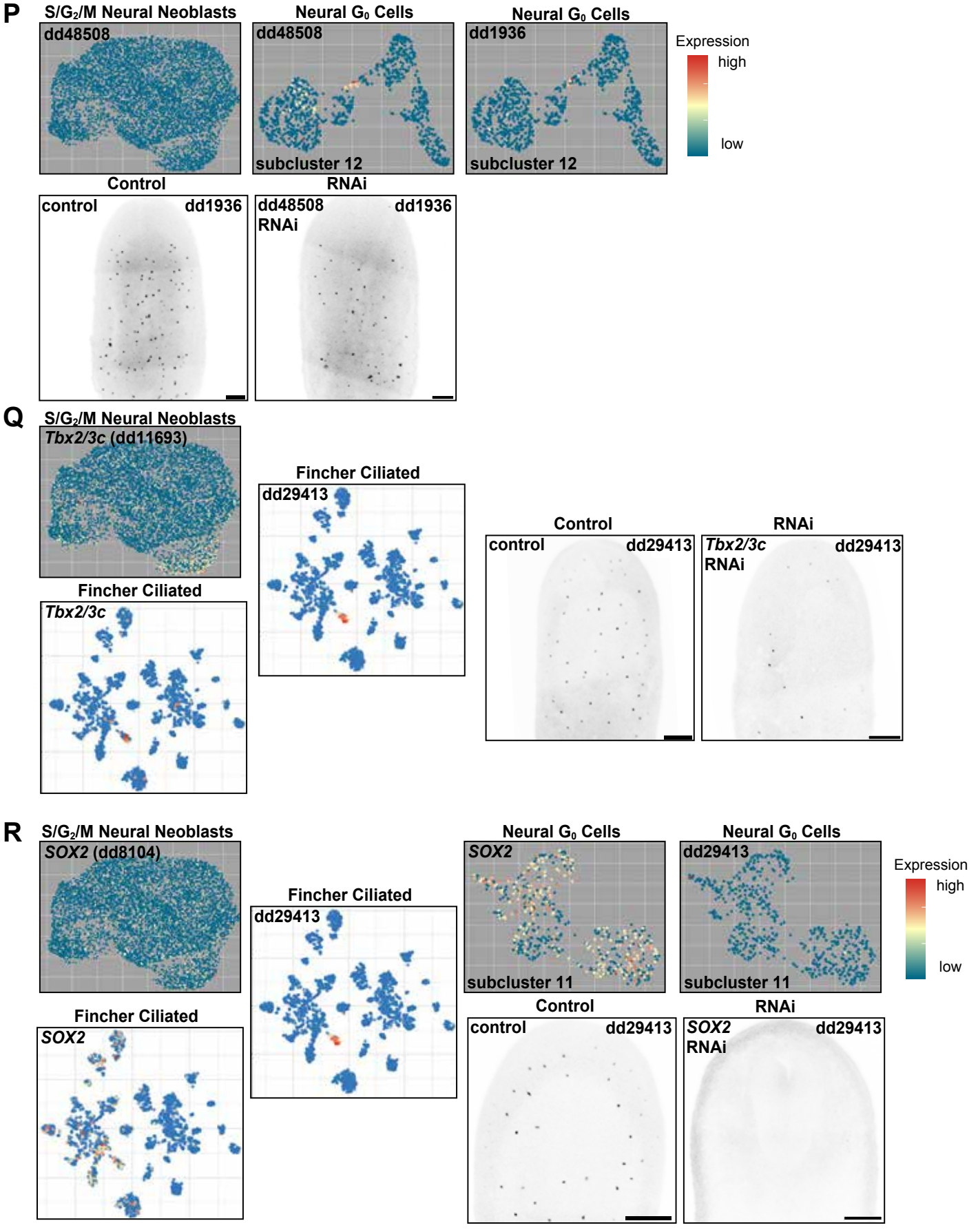

P
S/G2/M Neural Neoblasts
dd48508
Neural G0 Cells
dd48508
subcluster 12
Neural G0 Cells
dd1936
subcluster 12
Expression
high
low
Control
control
dd1936
RNAi
dd48508
RNAi
dd1936
Q
S/G2/M Neural Neoblasts
Tbx2/3c (dd11693)
Fincher Ciliated
dd29413
Fincher Ciliated
Tbx2/3c
Control
control
dd29413
RNAi
Tbx2/3c
RNAi
dd29413
R
S/G2/M Neural Neoblasts
SOX2 (dd8104)
Fincher Ciliated
dd29413
Fincher Ciliated
SOX2
Neural G0 Cells
SOX2
subcluster 11
Neural G0 Cells
dd29413
subcluster 11
Expression
high
low
Control
control
dd29413
RNAi
SOX2
RNAi
dd29413

Figure S4

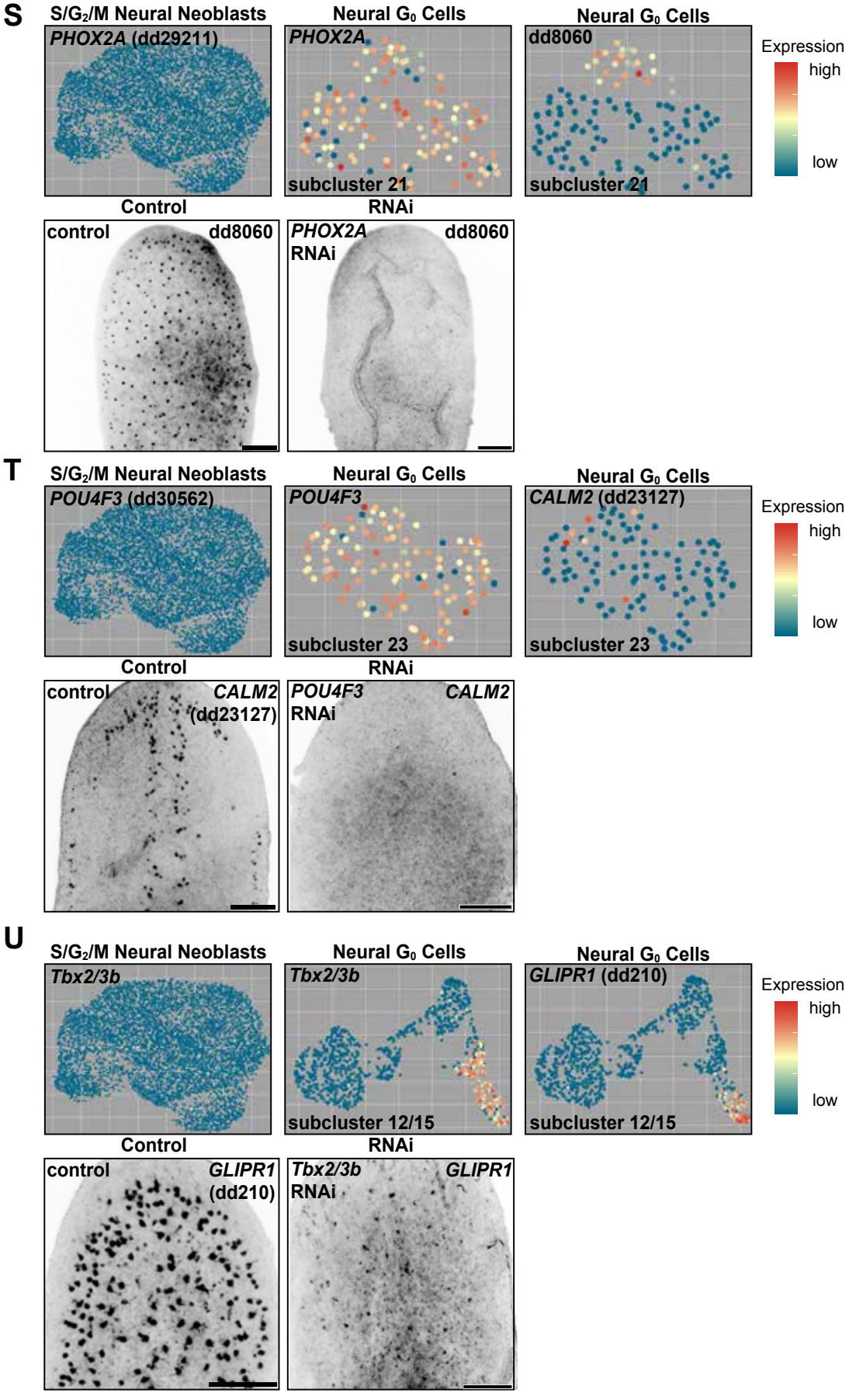
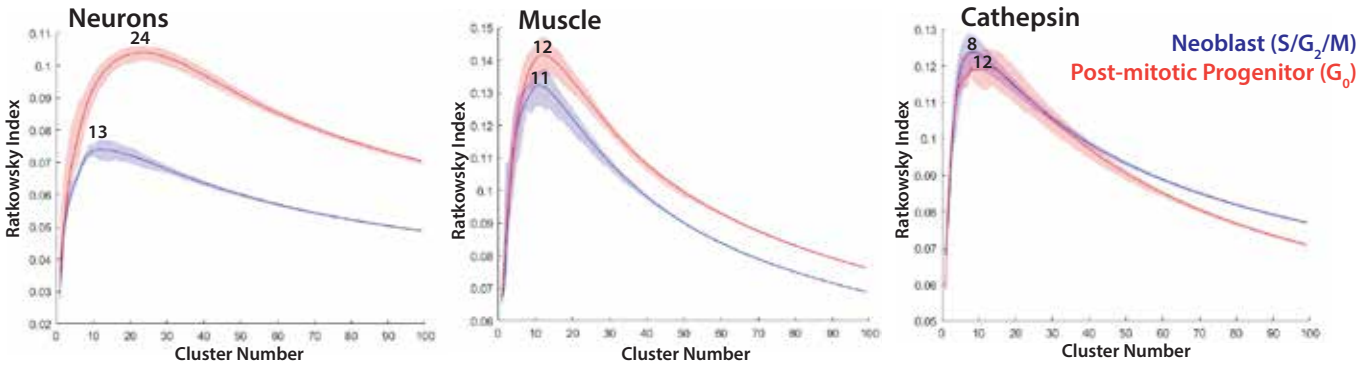

S
S/G2/M Neural Neoblasts
PHOX2A (dd29211)
Neural G0 Cells
PHOX2A
subcluster 21
Neural G0 Cells
dd8060
subcluster 21
Expression
high
low
Control
RNAi
control
dd8060
PHOX2A
RNAi
dd8060
T
S/G2/M Neural Neoblasts
POU4F3 (dd30562)
Neural G0 Cells
POU4F3
subcluster 23
Neural G0 Cells
CALM2 (dd23127)
subcluster 23
Expression
high
low
Control
RNAi
control
CALM2
(dd23127)
POU4F3
RNAi
CALM2
U
S/G2/M Neural Neoblasts
Tbx2/3b
Neural G0 Cells
Tbx2/3b
subcluster 12/15
Neural G0 Cells
GLIPR1 (dd210)
subcluster 12/15
Expression
high
low
Control
RNAi
control
GLIPR1
(dd210)
Tbx2/3b
RNAi
GLIPR1

Figure S4

V


Neurons
24
13
Ratkowsky Index
Cluster Number
Muscle
12
11
Ratkowsky Index
Cluster Number
Cathepsin
8
12
Ratkowsky Index
Cluster Number
Neoblast (S/G2/M)
Post-mitotic Progenitor (G0)

Supplemental Figure 4, related to Figures 4. Neural FSTF expression in neural progenitors (S/G_2/M and G_0) and the role of novel neural FSTFs in fate specification of distinct neural subtypes.

(A) Correlation analysis between transcription factors showing TF modules for all neuron subtypes ("Neural" and "*six1/2-1+"* cells combined). Modules were determined by computing pairwise correlations between all transcription factors (from TF catalog) of all S/G_2/M neural cells ("Neural" and "*six1/2-1+"* cells).

(B) Visualization of TF module expression domain on UMAP plot for all S/G_2/M cells from "Neural" and "*six1/2-1+*" clusters.

(C) Expression of neural FSTFs in neural S/G_2/M cells, G_0 cells, and respective G_0 subclusters. Measurements in lower left of boxes refer to percentages of cells expressing a given FSTF among noted neural progenitors.

(D, E, and F) Expression of individual neural TFs and coexpression of multiple neural TFs in neural S/G_2/M cells and neural G_0 cells. FSTFs defining unique subtypes of neurons are expressed but lowly-to-not correlated in S/G_2/M cells, and become substantially correlated in the G_0 state.

(G) Top: expression of *INSM2* in neural S/G_2/M cells and a subset of neural G_0 cells expressing *spp-4*. Bottom: RNAi showing loss of *spp-4+* neurons following *INSM2* RNAi.

(H) Top: expression of *UNCX* in neural S/G_2/M cells and in serotonergic G_0 cells (defined by *sert* and *pitx* coexpression). Bottom: RNAi showing loss of serotonergic neurons (*sert+*) following *UNCX* RNAi.

(I) Top: expression of dd_20282 (TF-encoding gene) in neural S/G_2/M cells and in a subset of non-ciliated peripheral neural G_0 cells defined by dd_1248 expression. Bottom: RNAi showing loss of dd_1248+ non-ciliated peripheral neurons following dd_20282 RNAi.

(J) Top: expression of dd_22331 (TF-encoding gene) in neural S/G_2/M cells and subset of ciliated peripheral neural G_0 cells defined by dd_29413 expression. Bottom: Expression of

dd_22331 in a subset of ciliated neurons from Fincher et al 2018 data. RNAi showing loss of dd_29413+ ciliated peripheral neurons following dd_22331 RNAi.

(K) Top Left: expression of dd_10911 (TF-encoding gene) in neural S/G_2/M cells. Bottom Left/Middle: expression of dd_10911 and dd_6953 (differentiated neural subtype marker) in non-ciliated cells of Fincher et al 2018 scRNA-seq data. Right: RNAi showing loss of dd_6953+ cells following dd_10911 RNAi.

(L) Top: expression of *GFI1B* in neural S/G_2/M cells and a subset of ciliated peripheral neural G_0 cells defined by dd_29413 expression. Bottom: RNAi showing loss of dd_29413+ ciliated peripheral neurons following *GFI1B* RNAi.

(M) Top: expression of *IRX2* in neural S/G_2/M cells and in subset of dopaminergic G_0 cells (defined by *tyrosine hydroxylase (th)* expression). Bottom: RNAi showing loss of dopaminergic neurons (*th*+) following *IRX2* RNAi.

(N) Top: expression of *aslc-2* in neural S/G_2/M cells and in a subset of non-ciliated peripheral G_0 neurons defined by dd_3069 expression. Bottom: Expression of *ascl-2* in a subset of non-niliated neurons from Fincher et al 2018 data. RNAi showing loss of dd_3069+ non-ciliated peripheral neurons following *ascl-2* RNAi.

(O) Top: expression of *IRX6* in neural S/G_2/M cells and in *npp-18*+ G_0 neurons. Bottom: Expression of *IRX6* in non-ciliated neurons from Fincher et al 2018 data. RNAi showing loss of *npp-18*+ neurons following *IRX6* RNAi.

(P) Top: expression of dd_48508 (TF-encoding gene) in a subset of non-ciliated neurons defined by dd_1936 expression. Bottom: RNAi showing loss of dd_1936+ non-ciliated neurons following dd_48508 RNAi.

(Q) Top Left: expression of *Tbx2/3c* in neural S/G_2/M cells. Bottom Left/Middle: expression of *Tbx2/3c* and dd_29413 in a subcluster of ciliated neurons from Fincher et al 2018 scRNA-seq data. Right: RNAi showing loss of dd_29413+ ciliated neurons following *Tbx2/3c* RNAi.

(R) Top: expression of *SOX2* in neural S/G_2/M cells and a subset of dd_29413+ neural G_0 cells. Bottom: Expression of SOX2 in a subset of ciliated neurons from Fincher et al 2018 data. RNAi showing loss of dd_29413+ ciliated neurons following *SOX2* RNAi.

(S) Top: expression of *PHOX2A* in neural S/G_2/M cells and a subset of dd_8060+ neural G_0 cells. Bottom: RNAi showing loss of dd_8060+ neurons following *PHOX2A* RNAi.

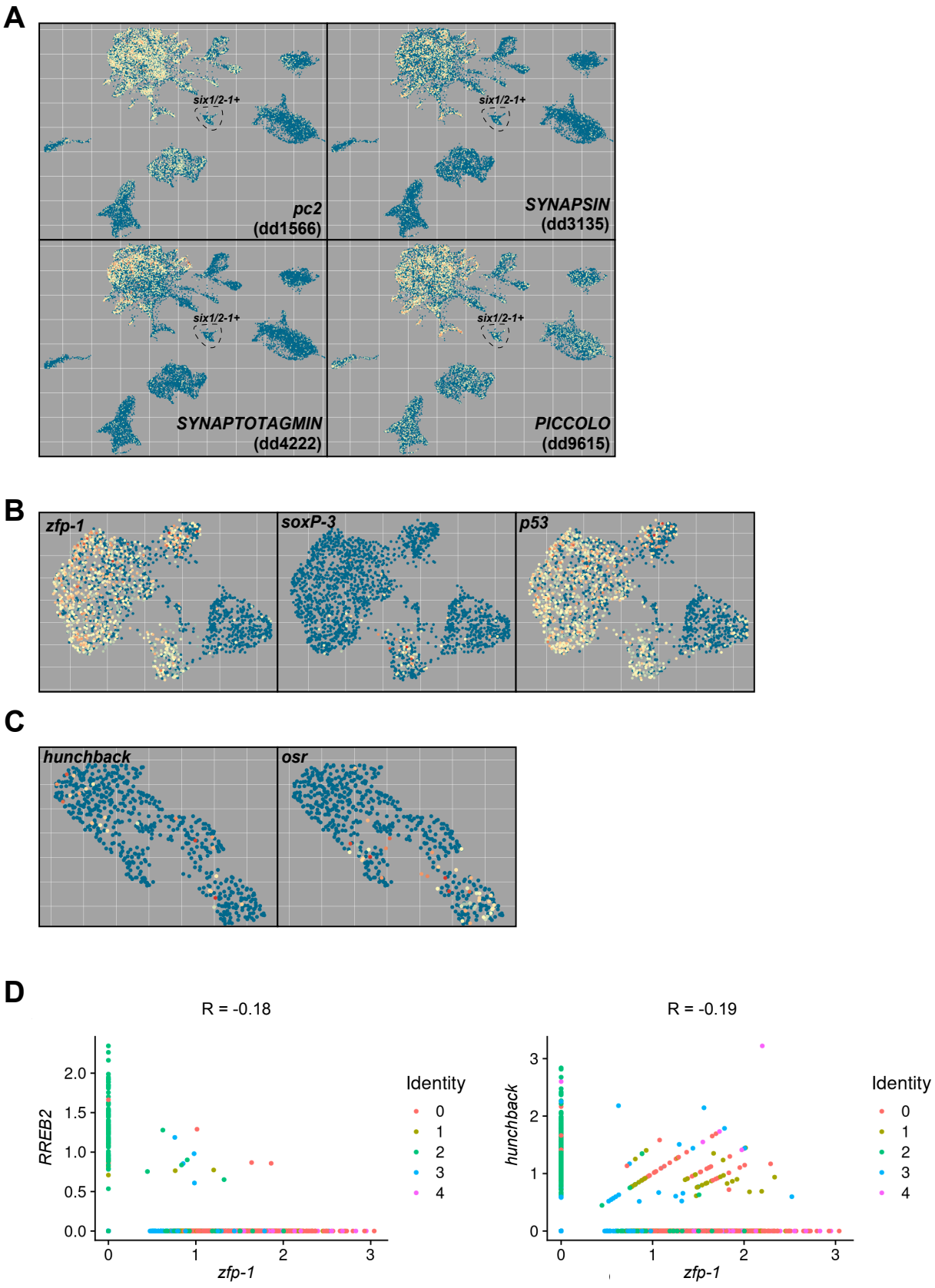
(T) Top: expression of *POU4F3* in neural S/G_2/M cells and a subset of *CALM2*+ neural G_0 cells. Bottom: RNAi showing loss of *CALM2*+ neurons following *POU4F3* RNAi.

(U) Top: expression of *Tbx2/3b* in neural S/G_2/M cells and a subset of *GLIPR1*+ neural G_0 cells. Bottom: RNAi showing loss of *GLIPR1*+ neurons following *Tbx2/3b* RNAi.

(G-U) FISH images are the same as in Figure 4E, but zoomed out.

(V) Ratkowsky index (the number of groups that best explain differences between cells) for different tissue types during the neoblast (S/G_2/M) and post-mitotic (G_0) stages. The number of groups that best explain differences between cells is roughly matched in neoblast and post-mitotic states for muscle and *cathepsin*+ tissues, but greatly expanded for neurons. Numbers above each curve note the number of groups with the highest Ratkowsky index.

Figure S5


A
six1/2-1+
pc2
(dd1566)
SYNAPSIN
(dd3135)
SYNAPTOTAGMIN
(dd4222)
PICCOLO
(dd9615)
B
zfp-1
soxP-3
p53
C
hunchback
osr
D
R = -0.18
RREB2
zfp-1
Identity
0
1
2
3
4
R = -0.19
hunchback
zfp-1
Identity
0
1
2
3
4

Supplemental Figure 5, related to Figures 5 and 6. Gene expression patterns and correlations in G_0 cells and subsets of $S/G_2/M$ cells.

(A) UMAP plot depicting expression of neural marker genes across all G_0 cells

(B) UMAP plot depicting expression of *zfp-1*, *soxP-3,* and *p53* (canonical epidermal FSTFs) in intestinal neoblasts ($S/G_2/M$ cells).

(C) UMAP plot depicting expression of *hunchback* and *osr* in G_0 intestinal cells

(D) Correlation values amongst $S/G_2/M$ cells for TFs associated with outer intestinal/basal cell fates (*RREB2, hunchback)* versus enterocyte states (*zfp-1*). TFs associated with cells of these two states are negatively correlated.

Click here to access/download
Supplemental Videos and Spreadsheets
Supplemental Table 1.xlsx

Click here to access/download
Supplemental Videos and Spreadsheets
Supplemental Table 2.xlsx

Click here to access/download
Supplemental Videos and Spreadsheets
Supplemental Table 3.xlsx

Click here to access/download
Supplemental Videos and Spreadsheets
Supplemental Table 4.xlsx

Click here to access/download
Supplemental Videos and Spreadsheets
Supplemental Table 5.xlsx

References

Adler, C.E., Seidel, C.W., McKinney, S.A., and Sánchez Alvarado, A. (2014). Selective amputation of the pharynx identifies a *FoxA*-dependent regeneration program in planaria. Elife *3*, e02238.

Baguñà, J., Saló, E., and Auladell, C. (1989). Regeneration and pattern formation in planarians. III. Evidence that neoblasts are totipotent stem cells and the source of blastema cells. Development *107*, 77- 86.

Brown, D.D.R., Molinaro, A.M., and Pearson, B.J. (2018). The planarian TCF/LEF factor Smed-tcf1 is required for the regeneration of dorsal-lateral neuronal subtypes. Dev Biol *433*, 374-383.

Cebrià, F. (2016). Planarian Body-Wall Muscle: Regeneration and Function beyond a Simple Skeletal Support. Front Cell Dev Biol *4*, 8.

Chong, T., Collins, J.J., 3rd, Brubacher, J.L., Zarkower, D., and Newmark, P.A. (2013). A sex-specific transcription factor controls male identity in a simultaneous hermaphrodite. Nat Commun *4*, 1814.

Cote, L.E., Simental, E., and Reddien, P.W. (2019). Muscle functions as a connective tissue and source of extracellular matrix in planarians. Nat Commun *10*, 1592.

Cowles, M.W., Brown, D.D., Nisperos, S.V., Stanley, B.N., Pearson, B.J., and Zayas, R.M. (2013). Genome-wide analysis of the bHLH gene family in planarians identifies factors required for adult neurogenesis and neuronal regeneration. Development *140*, 4691-4702.

Cowles, M.W., Omuro, K.C., Stanley, B.N., Quintanilla, C.G., and Zayas, R.M. (2014). COE loss-of-function analysis reveals a genetic program underlying maintenance and regeneration of the nervous system in planarians. PLoS Genet *10*, e1004746.

Currie, K.W., and Pearson, B.J. (2013). Transcription factors *lhx1/5-1* and *pitx* are required for the maintenance and regeneration of serotonergic neurons in planarians. Development *140*, 3577-3588.

Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput Biol *7*, e1002195.

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*, 874.

Flores, N.M., Oviedo, N.J., and Sage, J. (2016). Essential role for the planarian intestinal GATA transcription factor in stem cells and regeneration. Dev Biol *418*, 179-188.

Forsthoefel, D.J., Cejda, N.I., Khan, U.W., and Newmark, P.A. (2020). Cell-type diversity and regionalized gene expression in the planarian intestine. Elife *9*.

González-Sastre, A., De Sousa, N., Adell, T., and Saló, E. (2017). The pioneer factor *Smed-gata456-1* is required for gut cell differentiation and maintenance in planarians. Int J Dev Biol *61*, 53-63.

Hayashi, T., Asami, M., Higuchi, S., Shibata, N., and Agata, K. (2006). Isolation of planarian X-ray-sensitive stem cells by fluorescence-activated cell sorting. Dev Growth Differ *48*, 371-380.

He, X., Lindsay-Mosher, N., Li, Y., Molinaro, A.M., Pellettieri, J., and Pearson, B.J. (2017). FOX and ETS family transcription factors regulate the pigment cell lineage in planarians. Development *144*, 4540-4551.

Hyman, L.H. (1951). The Invertebrates: Platyhelminthes and Rhynchocoela The acoelomate bilateria, Vol II (New York: McGraw-Hill Book Company Inc.).

Issigonis, M., Redkar, A.B., Rozario, T., Khan, U.W., Mejia-Sanchez, R., Lapan, S.W., Reddien, P.W., and Newmark, P.A. (2022). A Kruppel-like factor is required for development and regeneration of germline and yolk cells from somatic stem cells in planarians. PLoS Biol *20*, e3001472.

Ivankovic, M., Haneckova, R., Thommen, A., Grohme, M.A., Vila-Farre, M., Werner, S., and Rink, J.C. (2019). Model systems for regeneration: planarians. Development *146*.

Khan, U.W., and Newmark, P.A. (2022). Somatic regulation of female germ cell regeneration and development in planarians. Cell Rep *38*, 110525.
King, R.S., and Newmark, P.A. (2012). The cell biology of regeneration. The Journal of cell biology *196*, 553-562.
Lapan, S.W., and Reddien, P.W. (2011). *dlx* and *sp6-9* control optic cup regeneration in a prototypic eye. PLoS Genet *7*, e1002226.
Lapan, S.W., and Reddien, P.W. (2012). Transcriptome Analysis of the Planarian Eye Identifies *ovo* as a Specific Regulator of Eye Regeneration. Cell Reports *2*, 294-307.
März, M., Seebeck, F., and Bartscherer, K. (2013). A Pitx transcription factor controls the establishment and maintenance of the serotonergic lineage in planarians. Development *140*, 4499-4509.
Molinaro, A.M., and Pearson, B.J. (2016). In silico lineage tracing through single cell transcriptomics identifies a neural stem cell population in planarians. Genome Biol *17*, 87.
Neiro, J., Sridhar, D., Dattani, A., and Aboobaker, A. (2022). Identification of putative enhancer-like elements predicts regulatory networks active in planarian adult stem cells. Elife *11*.
Newmark, P., and Sánchez Alvarado, A. (2000). Bromodeoxyuridine specifically labels the regenerative stem cells of planarians. Dev Biol *220*, 142-153.
Niu, K., Xu, H., Xiong, Y.Z., Zhao, Y., Gao, C., Seidel, C.W., Pan, X., Ying, Y., and Lei, K. (2021). Canonical and early lineage-specific stem cell types identified in planarian SirNeoblasts. Cell Regen *10*, 15.
Plass, M., Solana, J., Wolf, F.A., Ayoub, S., Misios, A., Glazar, P., Obermayer, B., Theis, F.J., Kocks, C., and Rajewsky, N. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. Science *360*, 875.
Raz, A.A., Wurtzel, O., and Reddien, P.W. (2021). Planarian stem cells specify fate yet retain potency during the cell cycle. Cell Stem Cell.
Reddien, P.W. (2018). The Cellular and Molecular Basis for Planarian Regeneration. Cell *175*, 327-345.
Reddien, P.W. (2022). Positional Information and Stem Cells Combine to Result in Planarian Regeneration. Cold Spring Harb Perspect Biol *14*.
Reilly, M.B., Cros, C., Varol, E., Yemini, E., and Hobert, O. (2020). Unique homeobox codes delineate all the neuron classes of C. elegans. Nature *584*, 595-601.
Rink, J.C., Vu, H.T., and Sánchez Alvarado, A. (2011). The maintenance and regeneration of the planarian excretory system are regulated by EGFR signaling. Development *138*, 3769-3780.
Roberts-Galbraith, R.H., Brubacher, J.L., and Newmark, P.A. (2016). A functional genomics screen in planarians reveals regulators of whole-brain regeneration. Elife *5*.
Ross, K.G., Currie, K.W., Pearson, B.J., and Zayas, R.M. (2017). Nervous system development and regeneration in freshwater planarians. Wiley Interdiscip Rev Dev Biol *6*.
Ross, K.G., Molinaro, A.M., Romero, C., Dockter, B., Cable, K.L., Gonzalez, K., Zhang, S., Collins, E.S., Pearson, B.J., and Zayas, R.M. (2018). SoxB1 activity regulates sensory neuron regeneration, maintenance, and function in planarians. Dev Cell *47*, 331-347 e335.
Saberi, A., Jamal, A., Beets, I., Schoofs, L., and Newmark, P.A. (2016). GPCRs Direct Germline Development and Somatic Gonad Function in Planarians. PLoS Biol *14*, e1002457.
Scimone, M.L., Atabay, K.D., Fincher, C.T., Bonneau, A.R., Li, D.J., and Reddien, P.W. (2020). Muscle and neuronal guidepost-like cells facilitate planarian visual system regeneration. Science *368*.
Scimone, M.L., Cote, L.E., and Reddien, P.W. (2017). Orthogonal muscle fibres have different instructive roles in planarian regeneration. Nature *551*, 623-628.
Scimone, M.L., Cote, L.E., Rogers, T., and Reddien, P.W. (2016). Two FGFRL-Wnt circuits organize the planarian anteroposterior axis. eLife *5*.

Scimone, M.L., Kravarik, K.M., Lapan, S.W., and Reddien, P.W. (2014a). Neoblast Specialization in Regeneration of the Planarian *Schmidtea mediterranea*. Stem cell reports *3*, 339-352.
Scimone, M.L., Lapan, S.W., and Reddien, P.W. (2014b). A *forkhead* transcription factor is wound-induced at the planarian midline and required for anterior pole regeneration. PLoS genetics *10*, e1003999.
Scimone, M.L., Srivastava, M., Bell, G.W., and Reddien, P.W. (2011). A regulatory program for excretory system regeneration in planarians. Development *138*, 4387-4398.
Scimone, M.L., Wurtzel, O., Malecek, K., Fincher, C.T., Oderberg, I.M., Kravarik, K.M., and Reddien, P.W. (2018). foxF-1 Controls Specification of Non-body Wall Muscle and Phagocytic Cells in Planarians. Curr Biol *28*, 3787-3801 e3786.
Sulston, J.E., Schierenberg, E., White, J.G., and Thomson, J.N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev Biol *100*, 64-119.
van Wolfswinkel, J.C., Wagner, D.E., and Reddien, P.W. (2014). Single-Cell Analysis Reveals Functionally Distinct Classes within the Planarian Stem Cell Compartment. Cell Stem Cell *15*, 326-339.
Vásquez-Doorman, C., and Petersen, C.P. (2014). *zic-1* Expression in planarian neoblasts after injury controls anterior pole regeneration. PLoS genetics *10*, e1004452.
Vogg, M.C., Owlarn, S., Perez Rico, Y.A., Xie, J., Suzuki, Y., Gentile, L., Wu, W., and Bartscherer, K. (2014). Stem cell-dependent formation of a functional anterior regeneration pole in planarians requires Zic and Forkhead transcription factors. Developmental Biology *390*, 136-148.
Vu, H.T., Rink, J.C., McKinney, S.A., McClain, M., Lakshmanaperumal, N., Alexander, R., and Sánchez Alvarado, A. (2015). Stem cells and fluid flow drive cyst formation in an invertebrate excretory organ. Elife *4*.
Wagner, D.E., Ho, J.J., and Reddien, P.W. (2012). Genetic regulators of a pluripotent adult stem cell system in planarians identified by RNAi and clonal analysis. Cell Stem Cell *10*, 299-311.
Wagner, D.E., Wang, I.E., and Reddien, P.W. (2011). Clonogenic neoblasts are pluripotent adult stem cells that underlie planarian regeneration. Science *332*, 811-816.
Wang, C., Han, X.S., Li, F.F., Huang, S., Qin, Y.W., Zhao, X.X., and Jing, Q. (2016). Forkhead containing transcription factor Albino controls tetrapyrrole-based body pigmentation in planarian. Cell Discov *2*, 16029.
Wang, Y., Zayas, R.M., Guo, T., and Newmark, P.A. (2007). *nanos* function is essential for development and regeneration of planarian germ cells. Proc Natl Acad Sci U S A *104*, 5901-5906.
Wenemoser, D., Lapan, S.W., Wilkinson, A.W., Bell, G.W., and Reddien, P.W. (2012). A molecular wound response program associated with regeneration initiation in planarians. Genes & Development *26*, 988-1002.
Wenemoser, D., and Reddien, P.W. (2010). Planarian regeneration involves distinct stem cell responses to wounds and tissue absence. Dev Biol *344*, 979-991.
Wheeler, T.J., and Eddy, S.R. (2013). nhmmer: DNA homology search with profile HMMs. Bioinformatics *29*, 2487-2489.
Witchley, J.N., Mayer, M., Wagner, D.E., Owen, J.H., and Reddien, P.W. (2013). Muscle cells provide instructions for planarian regeneration. Cell Reports *4*, 633-641.
Wurtzel, O., Oderberg, I.M., and Reddien, P.W. (2017). Planarian Epidermal Stem Cells Respond to Positional Cues to Promote Cell-Type Diversity. Dev Cell *40*, 491-504 e495.
Zeng, A., Li, H., Guo, L., Gao, X., McKinney, S., Wang, Y., Yu, Z., Park, J., Semerad, C., Ross, E.*, et al.* (2018). Prospectively Isolated Tetraspanin(+) Neoblasts Are Adult Pluripotent Stem Cells Underlying Planaria Regeneration. Cell *173*, 1593-1608 e1520.

Zhu, S.J., and Pearson, B.J. (2016). (Neo)blast from the past: new insights into planarian stem cell lineages. Curr Opin Genet Dev *40*, 74-80.

Chapter 3

Interpretable Neural Networks for Single-Cell Gene Expression Signature Analysis

Interpretable Neural Networks for Single-Cell Gene Expression Signature Analysis

Hunter King and Peter Reddien

Abstract

Single-cell RNA sequencing (scRNA-seq) is a powerful approach for identifying and characterizing the transcriptomes of cell types. Current methods in scRNA-seq data analysis primarily use unsupervised data clustering methods paired with statistical measurements and techniques to identify cell types, find enriched gene expression in those cell types, and to infer relationships between cell types, such as identifying transition states or lineage relationships with methods such as pseudotime-based gene expression trajectory analysis. We previously studied known and novel planarian progenitor cell types using these methods. Here, we develop flexible machine-learning based methods to analyze patterns in gene expression within different populations of post-mitotic progenitors, uncovering gene signatures that best define these cell types. We use neural network encodings of gene expression to represent the cells themselves, and use these to relate and categorize cells.

Background

Planarians are flatworms with the capacity to regenerate missing tissues, or even whole organisms, after a diverse range of injuries. This ability comes from the large population of adult pluripotent stem cells, called neoblasts. These neoblasts are responsible for homeostatically replacing cells throughout the lifetime of the animal (Pellettieri & Sánchez Alvarado, 2007) and also regenerating missing tissues during regeneration (Dubois, 1948). Neoblasts begin the cell cycle unspecified, but become specified and express fate-specific transcription factors (FSTFs) as they progress towards division (Raz, 2000). These neoblasts can then divide to generate post-mitotic progenitors that will migrate towards their target location in the animal before differentiating (Atabay et al., 2018; Eisenhoffer et al., 2008; Oderberg et al., 2017; Tu et al., 2015; Wurtzel et al., 2017). Previous work has identified many classes of specialized neoblasts and post-mitotic progenitor types (Chapter 2).

We previously obtained scRNA-seq data from planarian $S/G_2/M$ neoblasts and from the G_0 post-mitotic progenitors they give rise to (Chapter 2). These cells were collected from both anterior- and posterior-facing wound sites at 72 hours post-amputation, a time of substantial cell fate specification and differentiation in planarian regeneration (Wenemoser & Reddien, 2010). The greatest diversity of cell types was found in the post-mitotic progenitors, which were more representative of the diversity observed in the differentiated cell types of the animal, which these progenitors will eventually become (Chapter 2). Through Louvian-based clustering of all G_0 progenitors based on expression of their top variable genes, we identified 26 major clusters. Further subclustering of these 26 primary G_0 progenitors clusters uncovered additional cell

types, many of which clearly corresponded to previously identified mature planarian cell types by comparing signatures of enriched gene expression between these states. Using known markers of differentiated tissues, we found that most planarian tissue classes were represented by multiple progenitor clusters, some of which were adjacent and continuous with each other in UMAP space, suggesting that some of these cell types might not be fully distinct from each other and could instead exist along a continuous gene-expression spectrum. These findings led us to question the appropriateness of forcing discrete clusters on single-cell RNA sequencing data. We therefore considered whether methods that do not sort the cells into mutually exclusive bins could yield additional insights into what transcriptional signatures define cell-types and what constitutes differences among transcriptionally-similar cell-types.

Here, we report on simple and interpretable neural network architectures that efficiently learn characteristics in gene expression among single cells utilizing this dataset. We then examine the final structure of the networks to infer these learned characteristics. This approach identifies groups of cells expressing shared transcriptional signatures that span multiple Louvain-based clusters and were not found during traditional clustering analyses, and highlights similarities and differences between cells-types through their network encoding.

Principles and Parameters

We first trained a multi-class perceptron to use gene expression data to predict cell-type identity based on information from previous clustering results. A perceptron is a simple neural network with no hidden layers. Using the z-scored expression of the top 2,000

most variable genes expressed among sequenced G_0 post-mitotic progenitors, the perceptron predicted which of the 26 primary cluster classes a cell belonged to with 100% accuracy on the training dataset and 90% accuracy on a testing dataset (Figure 1A). The discrepancy between training and testing accuracy indicates overfitting, whereby the network memorizes patterns unique to training data without general applicability to non-trained data. To prevent overfitting, we can inject noise into our network to prevent it from simply memorizing the training data. Injecting gaussian noise with a standard deviation of 1.25, brought training and testing accuracy into parity at 92%, indicating less reliance on non-generalizable patterns specific to the training data that would result in overfitting (Figure 1B).

Because a perceptron is a linear classifier, it is straightforward to examine the learned weights to uncover the genes used by the network for cell classification (Figure 1C). The weights for the genes utilized in categorizing cells as belonging to the primary epidermis progenitor cluster (cluster 0), had a mean of 2.5e-04 and standard deviation of 0.091. The top three weights were for the genes *soxP-3*, *EGR-1*, and *p53* (weights of 0.85, 0.76, and 0.58, respectively), which are three canonical epidermal fate-specific transcription factors (FSTFs) (Cheng et al., 2018; van Wolfswinkel et al., 2014) (Figure 1C, bottom). The fact that the perceptron prominently utilized FSTFs genes uncovered from previous detailed molecular studies of this fate trajectory indicates the utility of this approach for identifying candidate regulatory factors for less studied fate trajectories. The gene with the most negative weight for this classification was a gene encoding a gamma-actin (dd37, weight:-0.37). This gene was therefore a negative predictor of epidermis-fated progenitors and was de-enriched in both epidermal and intestinal G_0

progenitors identified by clustering in our dataset. In single-cell sequencing datasets generated in previous studies, mature epidermal cells remain de-enriched in dd37, along with the intestine and pharynx (Fincher et al., 2018).

In addition to weights, we also examined the biases the network learned. Here, a single bias is learned for each class, adjusting the probability of classification based on the prior probability of the class and how well it is predicted by its features. In the case of the epidermis, the bias is higher than the average (0.22 vs -0.6779), representing a higher prior probability because the epidermal cluster contained the most cells. In contrast the small epidermal dorsal-ventral boundary (DVB) cluster, a small population of specialized epidermal cells surrounding the animal at the dorsal-ventral median plane, had a large negative bias (-1.5), representing the unlikelihood any given cell is a DVB cell (essentially, this is a probability adjustment based on the prior, given this is a rare cell-type). The highest weighted gene for classification of a cell to the DVB epidermal cluster was *Post-2a* (0.67), which was previously found to be an FSTF associated with DVB epidermal cells (Wurtzel et al., 2017).

Although the perceptron achieved high accuracies, it assigned high weights to many genes (239 genes had weights higher than 0.3), making analysis of the genes truly important for classification difficult. By introducing L1 regularization to the loss function, which penalizes the network in proportion to the sum of all weights, the network is forced to optimize the genes it chooses to use in the classification to only the most important ones, which also reduces overfitting. Training with regularization (λ= 0.01) dropped testing accuracy modestly from 92% to 89%. After training with regularization, *EGR-1*, *soxP-3*, and *p53* remained the genes with highest weights for

classification into the main epidermal G0 cluster (with weights 0.37, 0.37, and 0.31 respectively). However, now only 10 genes had a weight greater than 0.1 (compared to 187 previously) and the standard deviation of the weights dropped from 0.09 to 0.02 (Figure 1D). The novel zinc finger-encoding gene dd2946, which had the fourth highest weight, was confirmed to be expressed in mature epidermal cells in single-cell sequencing datasets from previous studies (Fincher et al., 2018). 27 genes had weights greater than a standard deviation above the mean, compared to 218 previously. The skewness, a measure of deviation from a normal distribution, of the distribution of weights was ten times higher (15.4 vs 1.6) and the kurtosis, a measure of the proportion of outliers in the data, was 24 times higher (294.5 vs 12.4) after introducing regularization, indicating the weights were less normally-distributed and more outlier-prone. This is consistent with fewer genes being assigned higher weights in the network. Whereas total network accuracy only dropped modestly, accuracy on some classes decreased significantly. Statistical recall of pharynx-specified progenitors decreased from 85% to 21%, reflecting the majority of pharynx-specified cells now being misclassified as *foxA*+ neurons (cluster 12; pharynx progenitors are also *foxA*+), epidermal cells (cluster 0), and cluster 4 neurons. Because there are fewer pharynx cells than other cell types in the dataset, the network can perform better if it invests regularized weight in more common cell types. This is reflected in the negative bias of the pharynx output class (-0.83) compared to epidermal, *foxA*+ neurons, and neural cluster 4 (0.62, -0.26, and 1.78, respectively).

To counter imbalances in class membership, we incorporated class weights to increase the importance of rare classes/cell-types in the loss function. This modestly

decreased network accuracy from 92% to 87%, while improving classification class parity. The recall rate of pharynx-specialized progenitors increased from 21% to 97%. This was reflected in an increase in the bias (0.32 vs. -0.83), an increase in weight for the pharynx marker gene *foxA* (0.38 vs 0.09), and an increase in the number of genes with weights over a standard deviation above the mean (33 vs 22).

Together, these studies show that neural networks can learn to classify cells into predetermined classes based on their unique expression of genes. The trained neural networks can then be assessed by their learned weights to determine features important for the classification and the training parameters can be used to better facilitate learning based on analysis goals, such as only relying on the most important features.

G0 Neural Progenitors

Within the primary major G_0 progenitor clusters, 11 constitute G_0 neural progenitors. There exist up to 77 neuron G_0 progenitor subtypes that we previously found through subclustering. Many of these subtypes were enriched in the expression of unique transcription factors and differentiated marker genes related to their specified function (Chapter 2). By looking at genes with enriched expression in these progenitor clusters and finding which differentiated cells also express them, we identified the likely fate of roughly half of these neural progenitor types. To characterize the cell types by their gene expression, we trained a perceptron on the 1,000 most variable genes expressed across all neural progenitors (Figure 2A). This network achieved a classification accuracy of 70%, and placed a high weight on many genes (131 genes had weights

greater than 0.17, one standard deviation above the mean of weights; Figure 2B). Many of the highly weighted genes were known neural FSTFs that we had previously identified, or that were previously known in the literature. Individual subclusters were associated with high, positive weights for genes with expression specific to those subclusters (Figure 2C). Neural cluster 23, which likely gives rise to differentiated ciliated neuron cluster 21 from Fincher et. al, had enriched expression of the TFs *POU4F3* (dd30562) and *scratch*, and both of these genes had high weights in our network. Other genes with high weights had specific expression for different subclusters derived from neural major cluster 23, including the protocadherin-encoding dd15033 gene. These genes represent the gene expression signature that best defined these individual subtypes in comparison to cells making up all other neural G0 progenitors. Through regularization, these also represented the minimal signature for accurately calling these cell types.

The output layer activations of softmax-transformed log probabilities (logits) represent the probability that a given cell belongs to each particular cell class (cluster). We can measure the confidence of cell classification by measuring the probability the network assigned to the predicted class for that cell. The confidence was significantly lower for misclassified cells compared to correctly classified cells (median 0.12 vs. 0.36; $p<0.0001$, Mann-Whitney U test; Figure 2D). When only considering cells for which the network had high confidence in class membership prediction, the prediction accuracy converged to 100%, making classification ‘confidence’ a good measure for the probability of correct classification (Figure 2E). By mapping classification confidence onto UMAP space, we observed that cells with highly confident classifications lie in the

heart of clusters, away from cluster boundaries denoted by dotted lines (Figure 2F). This could indicate that cells near cluster boundaries are difficult to distinguish from neighboring cells of other clusters. Furthermore, the low classification confidence and lower classification accuracy observed for cells in the border regions between clusters could indicate that some of these neural progenitor types are less discrete and more continuous with each other in gene expression space. Cluster 4 cells are mostly low confidence, which is supported by our previous work using enriched marker genes that failed to identify any specific cluster 4 neural fate and relatively few distinguishing genes.

Because transcription factors are often the most useful markers for defining planarian progenitor types and because many of the high weights among variable genes were transcription factors, we asked if classification of neural G_0 progenitors would be possible using the information regarding transcription factor expression in a cell alone. We trained a network to classify cells using only the expression of 714 previously identified putative planarian transcription factors, from a complete catalog of predicted TF genes in the planarian genome (see chapter 2). Our network obtained a training accuracy of 59% and a testing accuracy of 52% using TF expression data alone (Figure 3A). 87 transcription factors had weights one standard deviation above the mean, indicating relatively few transcription factors were used by the network to classify G_0 neural cells (Figure 3B). Individual subclusters were classified using far fewer transcription factors, as indicated by the corresponding weights applied to all TFs for cells classified to each cluster. Neural cluster 11 for example, had high weights for FSTFs broadly expressed in the cluster, such as *pou4l-1* which is expressed in all

neural 11 subclusters except 3 and 2, and therefore had high weights for several neural 11 subclusters (Figure 3C). *sim*, dd18207, and dd26877 only had high weights for a single subcluster each, and correspondingly had expression enriched in only those same subclusters. Interestingly, *soxP-4* had a negative weight for subcluster 6 and 1, and plotting expression of this gene showed it was de-enriched in those subclusters alone. Because we are using a regularizer that punishes the sum of absolute values of weights for the network, genes that are broadly expressed in but a few cell types are more likely to have a negative weight for those few de-enriched cell types compared to many high weights in all the others. *pou4l-1*, which has many high weights in cluster 11, has low weights in other neural clusters so it therefore has a few high weights instead of many negative. Classification confidence exhibited a similar distribution to the network trained on all variable genes (Figure 3D), also showed increasing accuracy with an increase in confidence stringency (Figure 3E), and similarly had higher confidence in classifications for cells further from other clusters in UMAP space (Figure 3F).

We found that cells misclassified by the network tended to have larger pairwise distances in UMAP space from other cells in their true identity cluster, indicating they may be misclassified because they are more different in gene expression to other cells in the cluster (Figure 3G). The position of misclassified cells in UMAP space was significantly closer to random cells within their wrongly-called cluster identity, compared to random cells from any cluster, suggesting that they are being misclassified to cell clusters more similar in gene expression than would be expected by random chance. Misclassified cells that were closer in UMAP space to other cells of the same true identity than random also had more confident classifications compared to cells further

away from cells of their same cluster (Figure 3H). This indicates that, for misclassified cells with high confidence classifications, the network is misclassifying them to neighboring clusters. In some instances, this could reflect overclustering, with biological identities not cleanly separated by called cluster boundaries. This approach can help reveal such candidate cluster boundaries of low confidence, cells that confidently represent a given cluster, and even multiple subregions within a cluster with high confident cells (Figure 3F), where cells might have been underclustered.

These studies show that simple neural networks can learn to classify and distinguish a diverse and large population (77 subtypes) of similar cells (neural-fated post-mitotic progenitors). The networks can also classify these cells using a functional subset of genes, in this case transcription factor-encoding genes, and can report trained weights to uncover genes important for these classifications. By looking at classification probabilities, we infer the confidence the network had in the classification and see that higher confidences are more likely to be classified correctly. Lower confidence classifications are more likely to occur near cluster boundaries and misclassifications are more likely to be made to neighboring clusters, which are transcriptionally more similar. One scenario in which this could occur is if some neural progenitor types exist along a continuous gene-expression space as a consequence of transcriptionally maturing from common pools of neural neoblasts (see chapter 2).

Alternative representations of cells identify unique characterizing features

Classifying cells on the basis of their previously assigned, discrete cluster identity as performed above has some limitations. First, it requires previous identification of called

"true" cell type identity, which can be attempted through cell clustering (as done above) or through manual curation based on gene expression thresholds of known cell type markers identified through prior studies. But because we found the neural network had low confidence near the interface between continuous clusters and misclassifications were usually made between neighboring clusters, it is possible that some cell types (especially diversifying progenitors) do not fit easily into the discrete "true" identities ascribed. We can potentially utilize alternative representations of cells to circumvent these pitfalls.

Our previous work (chapter 2) identified modules of transcription factors with correlated expression in different cell types. We hypothesize that the expression of a full module of transcription factors is more specific for defining a cell type than the expression of individual transcription factors, which are often non-specific. For instance, even the canonical epidermal marker *zfp-1* is also expressed in intestinal clusters. To represent cells by their expression of transcription factor modules, we sought to use an autoencoder, which is a neural network with at least one internal layer that is smaller in width than the input and that is trained to simply reconstruct the input data. This network architecture has been traditionally used for efficient data compression and removing noise from data, and has been previously used in single-cell RNA sequencing analysis for imputation (predicting the expression of dropout genes) (Talwar et al., 2018), denoising (Eraslan et al., 2019), and dimensionality reduction for visualization (Lin et al., 2020). Here, we predicted that the hidden layers (the latent space and other intermediate hidden layers), would represent high-level features, such as modules of transcription factors. In this case, latent space is a small set of nodes that carry different

continuous activation values for individual cells based on their expression levels of different genes. Because the latent space is smaller than the input dimensionality (the total number of transcription factors), we hypothesized that efficient encoding of gene expression in latent space would be representative of cell types and states. In other words, the expression of all 714 transcription factors is represented by a few variables in latent space, which form an encoding of the cell's gene expression signature and indirectly the cell itself. The latent space encoding of a cell could be used similarly to their cluster identity, but unlikely discrete clustering, cells can exist continuously throughout latent space.

We trained a symmetric autoencoder with one non-linear 50-node hidden layer between the latent-space layer and the input and output layers (Figure 4A). We were able to train the autoencoder to generate a reconstruction of the expression of the 714 putative transcription factors expressed across G0 neural progenitors. However, examining the intermediate hidden layer weights showed high, uniform weights distributed widely across all nodes (not shown). Similarly, looking at the hidden layer activations, the pattern of activity of the hidden nodes (the numerical value each node has) when given the gene expression for each cell, showed a uniform distribution of activations across nodes without clear pattern. This outcome is difficult to interpret and biologically irrelevant, even if accurate, and such combinatorial encoding runs the risk of memorizing noisy features through unique coding in the latency space, risking overfitting. To prevent this possibility we incorporated dropout, injected noise, and added L1 regularization of weights and layer activations, to prevent overfitting and to encourage a sparse representation of transcription factor modules. We also weighted

the loss of reconstructed features by the original variance of the z-scored gene expression level to encourage learning variable features.

Because larger latent spaces can better represent the full expression state of a cell, as latent-space dimension size increased, reconstruction error (the mean squared error of reconstructed versus actual gene expression levels) decreased (Figure 4B). We found that error decreased sharply with the first 15-nodes in latent-space. Error continued to decrease at a slower rate until reaching a latent-space dimension of 50-nodes, where error plateaued because it was restricted by the other 50-node layers. Interestingly, between a latent space size of 25 and 40, an alternative encoding solution arose with a near linear relationship between the latent-space dimension size and error. This likely represents some combinatorial encoding of transcription factors. We proceeded with a small latent dimension space of 7-nodes for subsequent analysis, forcing an efficient coding or transcription factors that might favor the representation of transcription factor modules characteristic of cell types.

We performed hierarchical clustering of the 7-dimension latent-space activations of all 9,913 neural post-mitotic progenitors to cluster cells with similar latent-space representations (Figure 4C, top). We plotted the position of cells in these latent-space clusters in UMAP space and found that many of these clustered cells were present in similar, continuous regions of UMAP space. However, the hierarchical clusters of cells with latent-space information sometimes differed from previous Louvain-based clustering of cells based on using all variable genes, as is typically done in processing scRNA-seq data for representation by UMAP (Figure 4C, bottom). When cells clustered with similar 7-dimension latent space activations were mapped onto previously

generated UMAP plots that involved Louvain clustering, some cells were present within a cluster and other cells distributed to a nearby cluster or clusters (Figure 4C). By looking at the output reconstruction for the latent activation of each cell, we can infer which transcription factors were encoded by the autoencoder. Transcription factors with high output reconstruction activations (the activation of the output layer, which is the networks attempt to reconstruct/decode expression based on the latent space) were expressed in cells of the corresponding latent activation clusters (Figure 4D, top). Transcription factor markers for latent-space clusters that do not correlate with previous Louvian clusters can also be found through their output activations, such as dd6778 which is expressed in latent activation cluster 26, but spans multiple Louvain clusters. Average activations of cells belonging to the same subclusters from previous Louvain-based clustering showed similar latent activations among subclusters belonging to the same major clusters (number preceding underscore is major cluster, number after is subcluster; Figure 4E). Distances in latent space and in intermediate hidden layer space were also continuous in UMAP space. For example, distances from the mean activation of cells belonging to major Louvain cluster 5, subcluster 6 in latent space and hidden layer 1 space show that neighboring cells in UMAP space have more similar activations than further cells in UMAP space (Figure 4F).

Because the latent-space encoding of cells can represent cell types, we hypothesized that we could use the latent activations to classify cells to their previous Louvain clusters. We trained a neural network with a 50-node hidden layer (similar to the autoencoder), to classify latent-space representations to their Louvain clusters. This attained a testing accuracy of 39%. A similar network architecture attained a testing

accuracy of 52% when trained on the expression of all 714 transcription factors. The small decrease in accuracy by training on the latent-space activations alone is impressive given the latent space had only 7 dimensions, a 102x decrease in the input dimensionality. This suggests that differences between cell types were preserved in latent space.

These analyses show that autoencoders can learn an efficient representation of cells in a lower dimensional space. These latent-space representations are both efficient encodings of gene expression and also serve as a non-discrete cellular identifier, similar to clustering. The autoencoder representation of a cell type can be used to uncover populations of cells and the genes expressed among them that differs from Louvain clustering. Louvain subclusters belonging to the same major cluster produce similar latent-space activations, indicating similarity, and distance in latent-space correlates to distance in UMAP space, but not in a way that directly follows the boundaries of cell clusters from Louvain clustering. For instance, cells could share expression of modules of genes associated with some particular cellular function (e.g., ciliogenesis or serotonin synthesis) but otherwise differ for other genes or gene modules.

Discussion

Recent years have seen a large increase in the accessibility of technologies for collecting massive amounts of data. Sequencing of mRNA transcripts from individual cells can generate large datasets of tens of thousands of transcript reads per cell. Tools for analysis of these large datasets include dimensionality reduction methods for visualization, unsupervised clustering algorithms, gene expression enrichment analysis,

and pseudotime trajectory inference. Although these tools are incredibly powerful, they have some drawbacks. Clustering of cells forces cells into discrete memberships which can generate arbitrary distinctions between transcriptionally similar cells. Common gene enrichment analysis methods depend on these discrete clusters and often use statistical tests to find the most specific single markers for any of these defined cell clusters. However, cells can be defined by more complex, combinatorial gene expression patterns that could be missed by simple enrichment analysis.

This paper utilizes small neural network architectures to learn gene expression patterns in cells. These methods can be supervised, in which labels for cells are previously generated through clustering or manual curation, or they can be unsupervised, in which efficient encodings of gene expression profiles are learned using autoencoders. By looking at network weights, outputs, and the types of mistakes made by the network, defining characteristics of the cells can be inferred.

Machine learning techniques can uncover relationships in scRNA-seq data that are difficult or limited with more traditional clustering-based methods. Neural networks with hidden layers can learn more complex relationships between features, such as combinatorial expression of genes that define cell types. These networks can also be flexibly constructed around different problems. For instance, cells can be labelled by multiple properties, (e.g., position in UMAP space, position physically in the animal, cluster membership, biological state from the experiment, such as in this case deriving from anterior or posterior regeneration datasets), and the network can learn what distinguishes cell types by these arbitrary properties. Different data modalities can even be merged, so that a cell type is predicted by various combinations of different features,

such as gene expression, protein expression, cell morphology, or even metadata like developmental stage. Genes can be subset to identify whether and how cells are defined by different classes of genes (such as transcription factors, subsets of transcription factors, or cell adhesion molecules). Networks could learn to uncover gene regulatory networks by training a network to predict gene expression based on the expression of transcription factors. Using autoencoders to represent gene expression and cell types can allow for better comparisons between cells on the basis of latent-space distance. Genes important for the identities of cells or modules of genes associated with particular biological properties of cells can also be uncovered by looking at autoencoder weights and activations. Many of the principles explored here on how network architecture and training properties (such as regularization methods) affect network encoding can also apply more broadly to generate interpretable neural networks for other uses.

Methods

Perceptron classification of G_0 progenitors

Single-cell RNA sequencing dataset was prepared previously using the 10X platform and processed in Seurat (Chapter 2). Gene-cell matrices of normalized and z-scored expression levels were exported from Seurat into Python. Testing dataset was constructed with a 3:7 testing:training dataset split. Perceptron neural networks were constructed in Python using the TensorFlow library as dense layers of sizes matching the gene number inputs and cell label outputs. Gaussian noise and L1 regularizers were used during training, through their built-in layers. Training was conducted with a 60-sample batch size, and conducted with 50 epochs for training without a L1 regularizer and 200 epochs with a L1 regularizer. Weights were visualized in Prism. Gene expression was visualized on UMAP plots generated in Seurat as previously described (Chapter 2).

Perceptron classification of G_0 neural-fated progenitors

Gene-cell matrices of normalized expression levels were exported from Seurat into Python and z-scored. Testing dataset was constructed with a 2:8 testing:training dataset split. Cluster/class weights were taken by the inverse abundance of each class label and used for training. Training was conducted with a 60-sample batch size, and conducted with 300 epochs for training. Training was stopped when loss on the testing dataset stopped improving, and the network weights producing the highest testing accuracy was used. Classification confidence was taken as the maximum, softmax-

transformed logit in the classification prediction and differences in confidence between correctly and incorrectly classified cells were measured by the Mann-Whitney U Test. Weights were visualized in Prism and gene expression and confidence was visualized on UMAP plots generated in Seurat. Simulations on distances between cells of different clusters were performed in MATLAB. Briefly, random cells that were classified correctly or incorrectly were measured to a random cell of the same 'true' or called identity in UMAP space. This simulation was ran 100,000 times for each classification category. The distances between two random cells, regardless of identity, was calculated in a similar manner for comparison.

Autoencoder training on G_0 neural-fated progenitors

Autoencoders were created using a single network with an input and output layer of size 714, a hidden layer after the input and before the output layer of size 50, and a middle latent-space layer of variable size (7-node in later experiments). A testing dataset was generated by a 1:9 testing:training dataset split. Network loss was taken at the mean squared error of the network reconstructed output compared to gene-expression input. Genes were weighted in the loss function by their original variance. Layer activations were made by generating functions with learned weights and biases in MATLAB that could be applied to each cell's gene expression vectors. Activations were hierarchically clustered and visualized in Python with the Seaborn library. Cells in these clusters were visualized in Seurat. Distances between cells in latent-space were calculated by Euclidean distance. Activation distance heatmaps were generated based on the normalized distances from cell latent activations and the average activations for cells of

different clusters, the closest 50% of cells were plotted using the heatmap legend and remained plotted in grey (shown is cluster 5_6). Latent space classifiers were made with L1 regularizers on both layers and the L1 coefficient set to bring training and testing accuracy close to parity (L = 0.004 for the control, 0.001 for the latent-space classifier).

Fig1

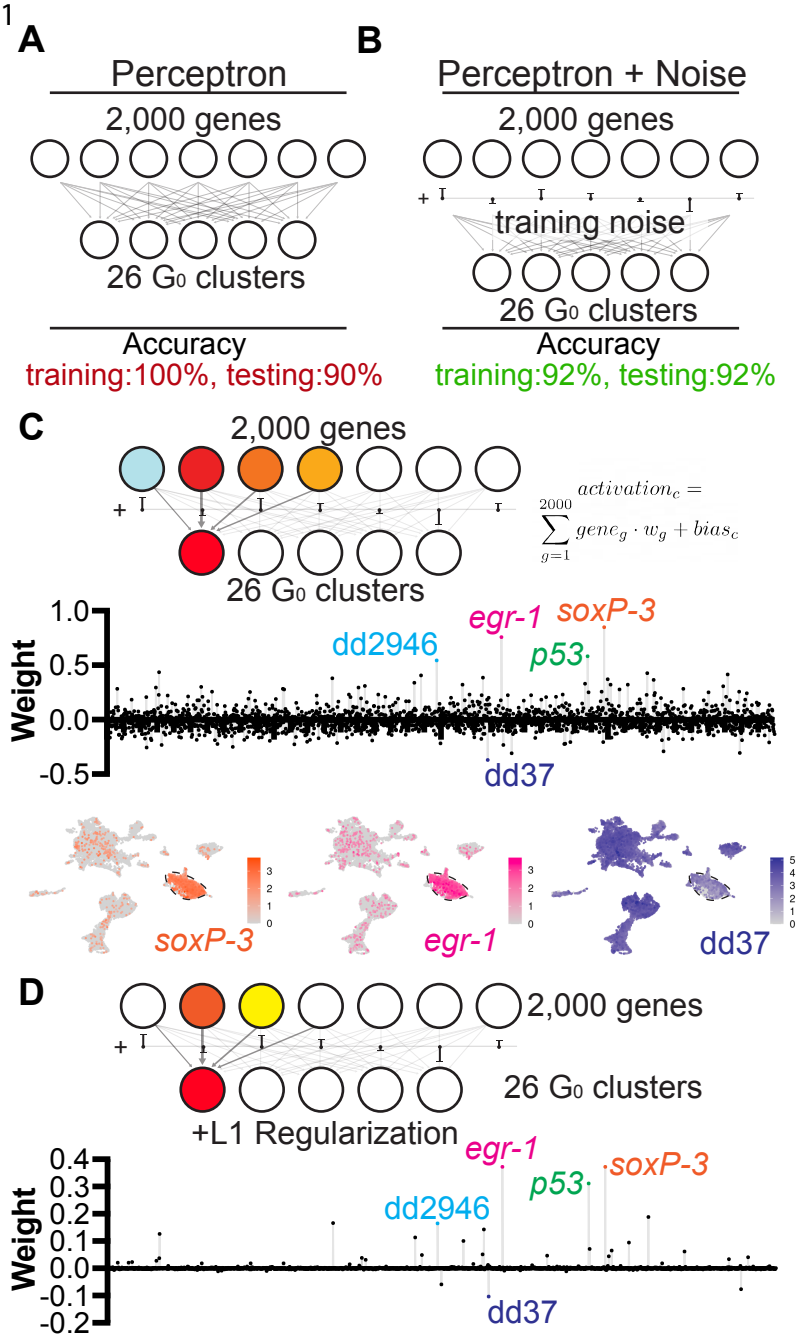

A
Perceptron
2,000 genes
26 G0 clusters
Accuracy
training:100%, testing:90%
B
Perceptron + Noise
2,000 genes
training noise
26 G0 clusters
Accuracy
training:92%, testing:92%
C
2,000 genes
26 G0 clusters
$activation_c = \sum_{g=1}^{2000} gene_g \cdot w_g + bias_c$
Weight
1.0
0.5
0.0
-0.5
dd2946
egr-1
soxP-3
p53
dd37
soxP-3
egr-1
dd37
D
2,000 genes
26 G0 clusters
+L1 Regularization
Weight
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
egr-1
p53
soxP-3
dd2946
dd37

Figure 1. Neural network architecture and training parameters influence the classification of post-mitotic progenitors to their major cell type identity.
A basic perceptron neural network (A) can accurately classify G0 post-mitotic progenitors into their major Louvain-derived clusters based on the expression of their top 2,000 most variable genes. Injecting noise during training brings the training and testing accuracy into parity, improving overfitting (B). Cartoon representation of how weights on different genes can contribute to different classification outputs (C, top-left). Output activations can be represented as a linear sum of weighted inputs with the addition of a learned bias factor (C, top-right). Weights of each gene for the classification of the epidermal G0 post-mitotic progenitor cluster (C, center) with highest weighted genes enriched or de-enriched in the epidermal cluster (circled; C, bottom). L1 regularization decreases the number of genes with high weights (D).

Fig2

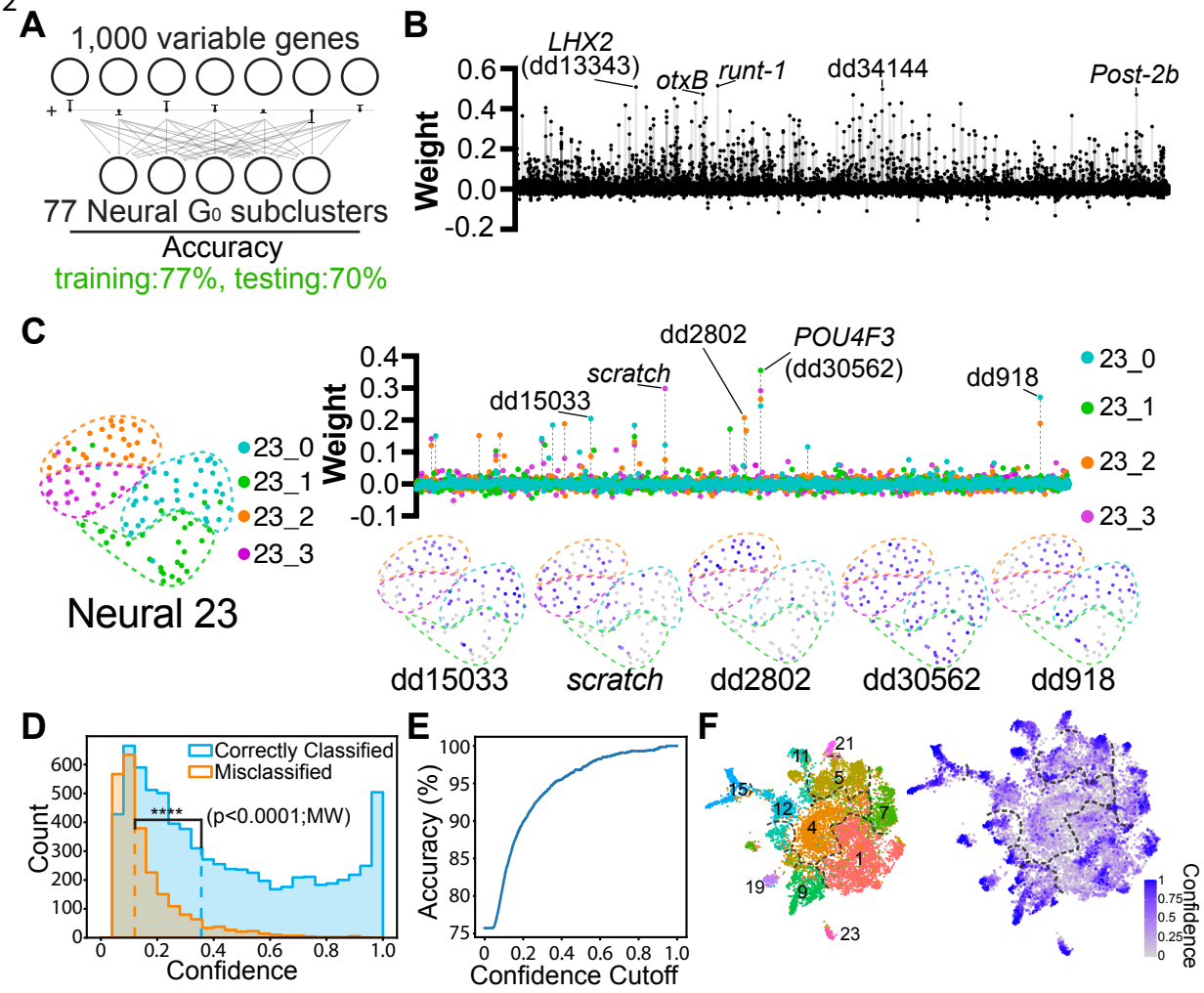

A
1,000 variable genes
77 Neural G0 subclusters
Accuracy
training:77%, testing:70%
B
LHX2
(dd13343)
otxB
runt-1
dd34144
Post-2b
Weight
C
dd2802
POU4F3
(dd30562)
scratch
dd15033
dd918
23_0
23_1
23_2
23_3
Neural 23
dd15033
scratch
dd2802
dd30562
dd918
D
Correctly Classified
Misclassified
****
(p<0.0001;MW)
Count
Confidence
E
Accuracy (%)
Confidence Cutoff
F
Confidence

Figure 2. Neural networks can accurately classify neural-specified post-mitotic progenitors to their neural subtype identity by variable gene expression.
A perceptron neural network (A) with training noise can accurately classify G0 neural post-mitotic progenitors into their Louvain-derived subclusters based on the expression of their top 1,000 most variable genes. Weights of each gene for the classification of all 77 neural subtypes (B), with high weighted genes highlighted. Louvain subclustering of neural major cluster 23 (C, left). Gene weights for cluster 23 subclusters (C, right) from neural network trained to classify 77 neural subtypes. Highest weighted genes are enriched in specific neural 23 subclusters (C, bottom). The neural network confidence is higher for cells that were correctly classified (D). If thresholding classifications by neural network confidence, accuracy increases with increasing confidence threshold (E). In UMAP space, cells have higher confidences further from Louvain cluster (F), and lower confidence near the interface between clusters.

Fig3

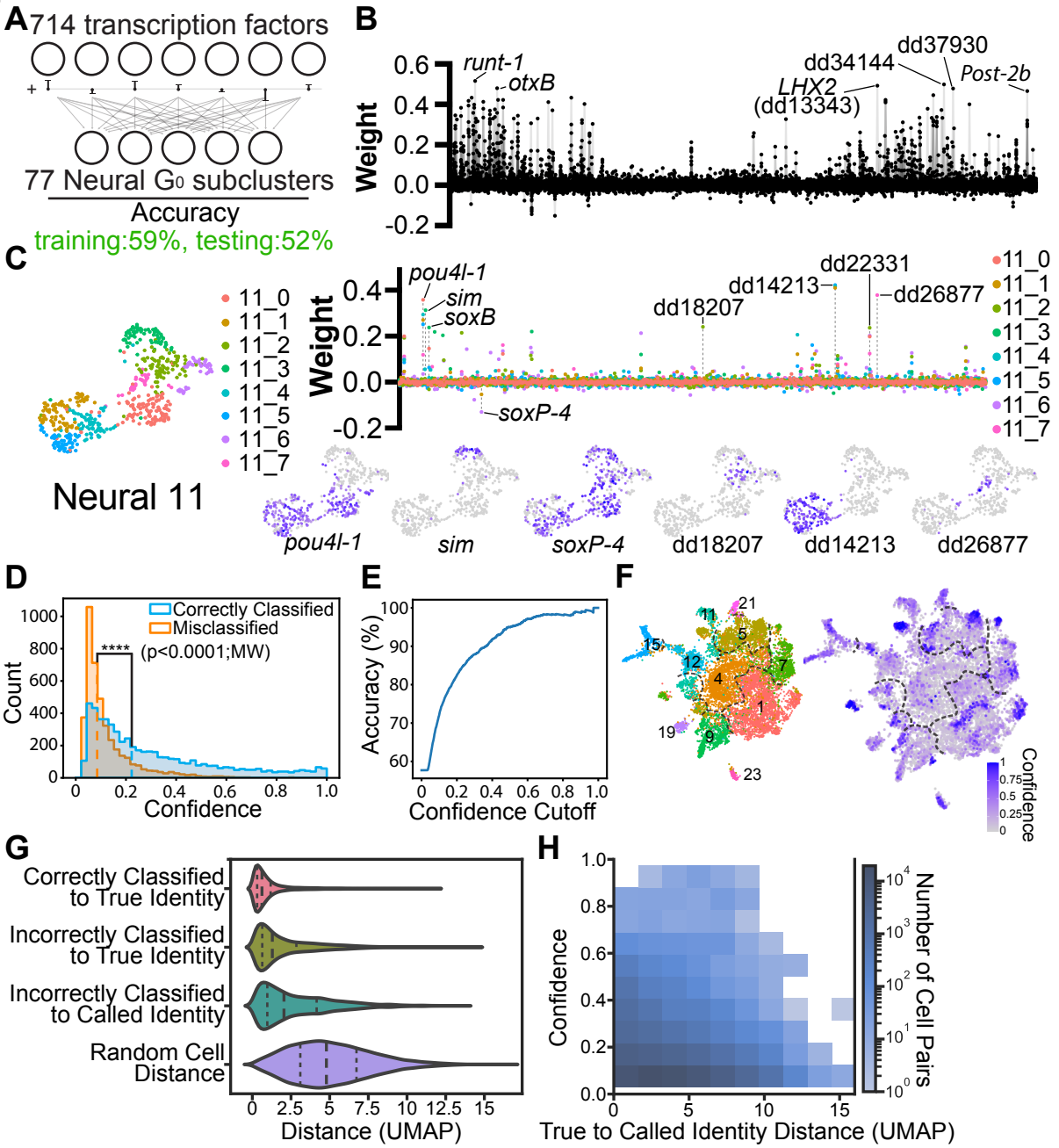

A
714 transcription factors
77 Neural G0 subclusters
Accuracy
training:59%, testing:52%
B
Weight
runt-1
otxB
LHX2
(dd13343)
dd34144
dd37930
Post-2b
C
Neural 11
11_0
11_1
11_2
11_3
11_4
11_5
11_6
11_7
pou4l-1
sim
soxB
soxP-4
dd18207
dd14213
dd22331
dd26877
D
Correctly Classified
Misclassified
****
(p<0.0001;MW)
Count
Confidence
E
Accuracy (%)
Confidence Cutoff
F
Confidence
G
Correctly Classified to True Identity
Incorrectly Classified to True Identity
Incorrectly Classified to Called Identity
Random Cell Distance
Distance (UMAP)
H
Confidence
Number of Cell Pairs
True to Called Identity Distance (UMAP)

Figure 3. Neural networks can accurately classify neural-specified post-mitotic progenitors to their neural subtype identity by transcription factor expression. A perceptron neural network (A) with training noise can accurately classify G0 neural post-mitotic progenitors into their Louvain-derived subclusters based on the expression of the 714 predicted transcription factors in planarians. Weights of each transcription factor for the classification of all 77 neural subtypes (B), with high weighted genes highlighted. Louvain subclustering of neural major cluster 11 (C, left). Gene weights for cluster 11 subclusters (C, right) from neural network trained to classify 77 neural subtypes. Highest weighted genes are enriched in specific neural 11 subclusters (C, bottom), and a negatively weighted gene is enriched broadly in neural 11 subclusters. The neural network confidence is higher for cells that were correctly classified (D). If thresholding classifications by neural network confidence, accuracy increases with increasing confidence threshold (E). In UMAP space, cells have higher confidences further from Louvain cluster (F), and lower confidence near the interface between clusters. Cells that are correctly classified are closer to other cells of their labelled, true identity than cells that are misclassified (G). Cells that are misclassified are closer to cells of their incorrect, called identity than is expected by chance (G). For misclassified cells, classification confidence is higher when the cell is closer to other cells of their incorrect, called identity (H).

Fig4

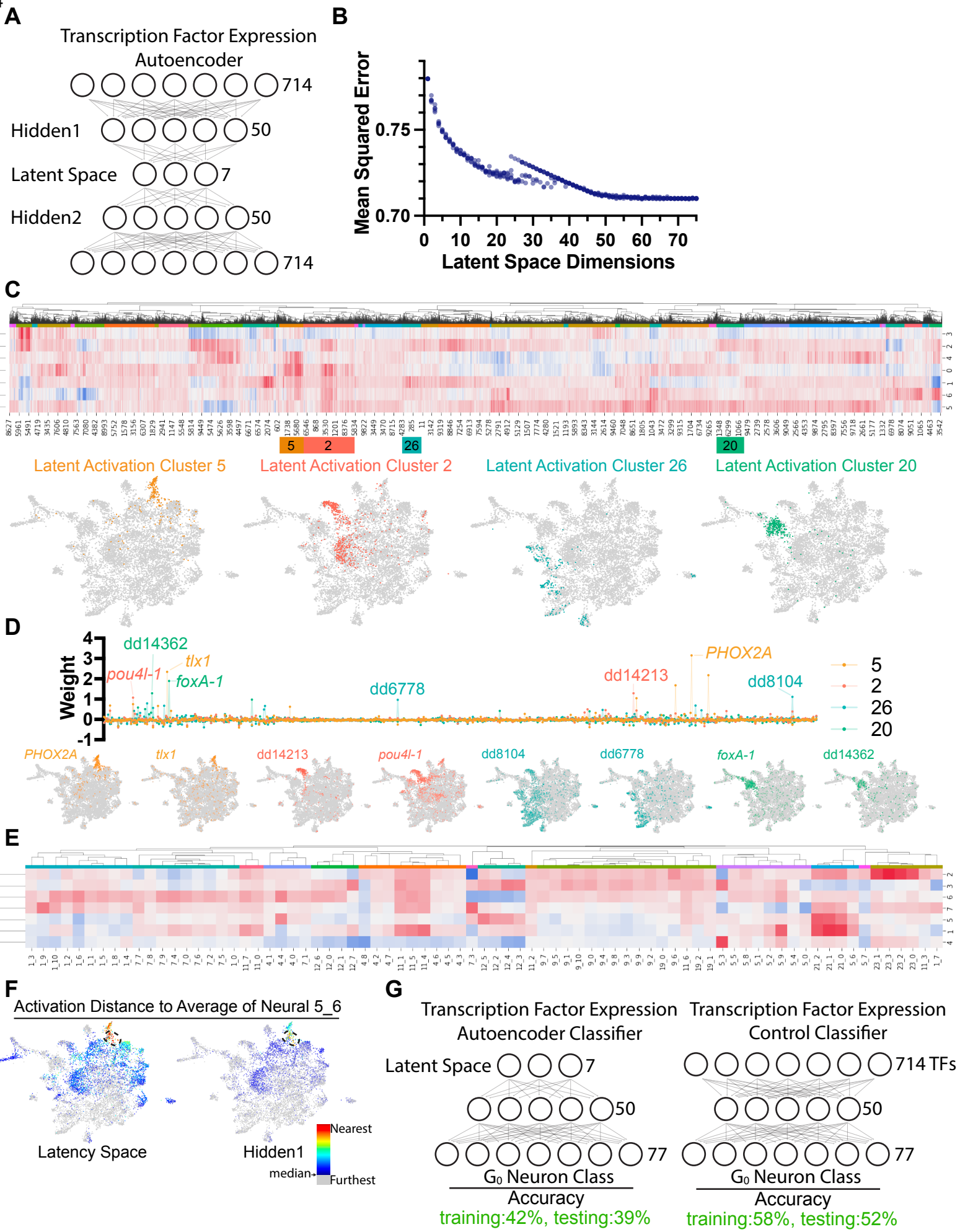

A
Transcription Factor Expression
Autoencoder
714
Hidden1
50
Latent Space
7
Hidden2
50
714
B
Mean Squared Error
0.75
0.70
0
10
20
30
40
50
60
70
Latent Space Dimensions
C
5
2
26
20
Latent Activation Cluster 5
Latent Activation Cluster 2
Latent Activation Cluster 26
Latent Activation Cluster 20
D
Weight
4
3
2
1
0
-1
dd14362
tlx1
pou4l-1
foxA-1
dd6778
PHOX2A
dd14213
dd8104
5
2
26
20
PHOX2A
tlx1
dd14213
pou4l-1
dd8104
dd6778
foxA-1
dd14362
E
F
Activation Distance to Average of Neural 5_6
Latency Space
Hidden1
Nearest
median
Furthest
G
Transcription Factor Expression
Autoencoder Classifier
Latent Space
7
50
77
G0 Neuron Class
Accuracy
training:42%, testing:39%
Transcription Factor Expression
Control Classifier
714 TFs
50
77
G0 Neuron Class
Accuracy
training:58%, testing:52%

Figure 4. Autoencoders can learn efficient representations of single-cell transcriptomes

A diagram of an autoencoder architecture (A) used to encode the expression of 714 putative transcription factors for neural post-mitotic progenitors. Error of transcription factor reconstruction decreases with increasing latent space size (B). Hierarchical clustering of cell's latent space activations (C, top), showing groups of cells with similar activations lying in continuous UMAP space (C, bottom). Reconstructed outputs for example cells with similar activation clusters (D, top) showing transcription factor expression in same cells (D, bottom). Average latent space activation for cells in traditional Louvain clustering (E) with major clusters denoted by first number and subclusters denoted by number trailing underscore. Distances between cells activation in latent space and hidden layer 1 to cells of neural cluster 5, subcluster 6 (circled) showing distances correlate with UMAP distance (F). 7-node latent space activations of cells can be used to classify post-mitotic neurons to their Louvain subcluster identity (G, left) almost as well as a classification based on all 714 transcription factors (G, right).

References

Atabay, K. D., LoCascio, S. A., de Hoog, T., & Reddien, P. W. (2018). Self-organization and progenitor targeting generate stable patterns in planarian regeneration. *Science*, *360*(6387), 404-409. https://doi.org/10.1126/science.aap8179

Cheng, L. C., Tu, K. C., Seidel, C. W., Robb, S. M. C., Guo, F., & Sánchez Alvarado, A. (2018). Cellular, ultrastructural and molecular analyses of epidermal cell development in the planarian *Schmidtea mediterranea*. *Developmental Biology*, *433*(2), 357-373. https://doi.org/10.1016/j.ydbio.2017.08.030

Dubois, F. (1948). Sur les conditions de la migration des cellules de régénération chez les planaires d'eau douce. *Soc Biol Strasbourg*, 533-535.

Eisenhoffer, G. T., Kang, H., & Sánchez Alvarado, A. (2008). Molecular analysis of stem cells and their descendants during cell turnover and regeneration in the planarian *Schmidtea mediterranea*. *Cell Stem Cell*, *3*(3), 327-339. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18786419

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., & Theis, F. J. (2019). Single-cell RNA-seq denoising using a deep count autoencoder. *Nat Commun*, *10*(1), 390. https://doi.org/10.1038/s41467-018-07931-2

Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M., & Reddien, P. W. (2018). Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science*, *360*(6391), 874. https://doi.org/10.1126/science.aaq1736

Lin, E., Mukherjee, S., & Kannan, S. (2020). A deep adversarial variational autoencoder model for dimensionality reduction in single-cell RNA sequencing analysis. *BMC Bioinformatics*, *21*(1), 64. https://doi.org/10.1186/s12859-020-3401-5

Oderberg, I. M., Li, D. J., Scimone, M. L., Gavino, M. A., & Reddien, P. W. (2017). Landmarks in Existing Tissue at Wounds Are Utilized to Generate Pattern in Regenerating Tissue. *Curr Biol*, *27*(5), 733-742. https://doi.org/10.1016/j.cub.2017.01.024

Pellettieri, J., & Sánchez Alvarado, A. (2007). Cell turnover and adult tissue homeostasis: from humans to planarians. *Annu Rev Genet*, *41*, 83-105. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18076325

Raz, E. (2000). The function and regulation of vasa-like genes in germ-cell development. *Genome Biol*, *1*(3), REVIEWS1017. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11178242

Talwar, D., Mongia, A., Sengupta, D., & Majumdar, A. (2018). AutoImpute: Autoencoder based imputation of single-cell RNA-seq data. *Sci Rep*, *8*(1), 16329. https://doi.org/10.1038/s41598-018-34688-x

Tu, K. C., Cheng, L. C., H, T. K. V., Lange, J. J., McKinney, S. A., Seidel, C. W., & Sánchez Alvarado, A. (2015). *Egr-5* is a post-mitotic regulator of planarian epidermal differentiation. *Elife*, *4*, e10501. https://doi.org/10.7554/eLife.10501

van Wolfswinkel, J. C., Wagner, D. E., & Reddien, P. W. (2014). Single-Cell Analysis Reveals Functionally Distinct Classes within the Planarian Stem Cell Compartment. *Cell Stem Cell*, *15*(3), 326-339. https://doi.org/10.1016/j.stem.2014.06.007

Wenemoser, D., & Reddien, P. W. (2010). Planarian regeneration involves distinct stem cell responses to wounds and tissue absence. *Developmental Biology*, *344*(2), 979-991. https://doi.org/S0012-1606(10)00837-7 [pii]
10.1016/j.ydbio.2010.06.017

Wurtzel, O., Oderberg, I. M., & Reddien, P. W. (2017). Planarian Epidermal Stem Cells Respond to Positional Cues to Promote Cell-Type Diversity. *Dev Cell*, *40*(5), 491-504 e495. https://doi.org/10.1016/j.devcel.2017.02.008

Chapter 4

Discussion

Discussion

The remarkable regenerative capabilities of planarians have captivated scientists for centuries. Studies on regeneration date back to the 18th century. Pivotal experiments by Abraham Trembley in 1744 described whole-body regeneration in hydra (Trembley, 1744), which could generate two regenerated individuals from one cut animal. Lazzaro Spallanzani described regeneration in several species in the late 1700s, including limb and tail regeneration in salamanders (Tsonis & Fox, 2009). Nearly a century later, Thomas Hunt Morgan became fascinated with regeneration and extensively documented the regenerative abilities of planarians. Through methodical amputation experiments, he found that planarians could be divided into tiny fragments that would each regenerate into a new worm (Morgan, 1898; Morgan, 1901). Though, pieces from the pharynx or anterior to the eyes, he noted, would not go on to regenerate (Morgan, 1898). We now know that those two regions lack the essential cells for regeneration (Newmark & Sánchez Alvarado, 2000), neoblasts.

Recent advancements in genetics and molecular biology methods have led to a resurgence in planarian research. Planarians have become a powerful model system for not only regeneration, but also for the basic principles of body plan patterning, fate specification and stem cell biology, tissue formation, and wound signaling that underly the phenomenon of regeneration. In this work, I primarily expand on knowledge of planarian neoblast biology.

Initial studies of planarian anatomy utilized traditional microscopy, histology, and electron microscopy to characterize different tissues in the animal (Carpenter et al., 1974; Hyman, 1951; MacRae, 1967). Later studies used in-situ hybridization and

antibody staining to identify different cell types in the animal, but this was largely done on the basis of genetic homology to known cell types in more studies organisms, or though screening gene cDNA libraries (Sánchez Alvarado et al., 2002) or antibody libraries made against planarian homogenate (Ross et al., 2015). In 2014, a single-cell gene expression analysis of planarian neoblasts using quantitative PCR on known neoblast genes and fate-specific transcription factors (FSTFs) revealed three major classes of neoblasts (van Wolfswinkel et al., 2014). In 2018, a large single-cell RNA sequencing dataset in planarians was generated and used to characterize most mature cell types in the animal (Fincher et al., 2018). The discovery of over a hundred, transcriptionally and spatially-distinct cell types in the animal greatly expanded the known cellular diversity in planarians. However, because neoblasts were not enriched in this dataset and older sequencing platforms didn't reliably detect rare transcripts, this study did not uncover the known diversity of planarian neoblasts.

To better characterize the full diversity of planarian neoblasts, we isolated dividing neoblasts using fluorescent-activated cell sorting (FACS) on the basis of DNA content. Previous work has identified actively dividing (4C) neoblasts as the neoblast stage of specification (Raz et al., 2021), so this enriches for specified neoblasts. We then constructed single-cell libraries from these cells for RNA sequencing. We also isolated post-mitotic progenitors, the specialized migratory cells that are produced by neoblasts and incorporate into existing tissues or nucleate new ones before differentiation. The population of progenitor cells at this stage is relatively understudied, but have been characterized for the epidermal fate, where it was found that epidermal neoblasts are specified far from the epidermal surface and their post-mitotic descendent

cells migrate towards the epidermis with changes in progressive changes in gene expression as they migrate (Eisenhoffer et al., 2008; Tu et al., 2015; Wurtzel et al., 2017).

The migratory, post-mitotic progenitor stage has been the focus of recent studies on self-organization and tissue structure in planarians (Atabay et al., 2018; Oderberg et al., 2017). This intermediate cell type, which is a transition state from stem cell to differentiated cell, could also be a site of further fate specification, after initial neoblast specification.

From these studies, we uncovered neoblast and post-mitotic progenitor types for many known mature cell types in the animal. By using the shared expression of defining FSTFs, we were able to connect our sequenced neoblast and post-mitotic progenitors to differentiated cell types characterized in previous studies. From this data, we were also able to characterize new and known precursor cell types by their full gene expression. This led to the discovery of novel FSTFs for most of our identified neoblast and post-mitotic progenitor cell types, many of which we found to be functionally required for the specification or survival of their associated differentiated cell type through RNAi-based gene inhibition experiments.

We were able to identify neoblast classes for most cell types, comprising most major tissue classes. However, we failed to identify many clear neural-specified neoblast classes, despite differentiated neuronal subtypes being the most diverse in the animal (Fincher et al., 2018) and in our post-mitotic progenitor population. We characterized neural-specified neoblasts by the combinatorial expression of transcription factors that define them in their later, post-mitotic progenitor stage and

found that whereas many of these FSTFs are expressed in neoblasts, they were not coexpressed in the transcription factor modules that define them later. This increase in FSTF signature specificity with cell stage and increase in the number of identifiable clusters led us to hypothesize that neural-specified cells expand in diversity in the post-mitotic progenitor stage. This could for some states represent further maturation of state post-mitotically of initial choices made in neoblasts. Post-mitotic diversification of fate has been shown previously for protonephridial cells, which seem to exist as a common pool as neoblasts with FSTFs distinguishing different mature protonephridial cell types being coexpressed in the neoblast (Scimone et al., 2011). In both neoblast and post-mitotic progenitors, we also observed that these FSTFs, which have unique expression in different mature protonephridial cells, were coexpressed. The correlation in expression among these FSTFs decreased in post-mitotic progenitors compared to neoblasts. This refinement of fate or expansion of cell-type diversity seemed to happen in neurons, protonephridia, and possibly phagocytic cathepsin cells, while not occurring for cell precursors for other tissues like muscle, where evidence for fate choice for almost all cell types exist as early as the specified neoblast. These results suggest diversity in different tissues can arise at different stages in the lifetime of the cell. It is unknown whether this apparent expansion of diversity in post-mitotic progenitors is truly further fate specification, with fate decisions being made in the actual post-mitotic progenitors. An alternative possibility is that the final fate selection of some neoblasts cannot be identified with transcriptome analysis alone. For example, other determinants of fate could precede FSTF signature expression, such as chromatin accessibility, and not be assayed here. Further experiments looking at chromatin accessibility and

modifications in neoblasts and post-mitotic progenitors could shed light on this possibility.

The development of transgenesis in planarians would allow for the genetic labelling of cells for lineage tracing. In the future, FSTF-expressing cells could be genetically labeled for tracking descendent cells. This still would be logistically challenging. Under the single-step fate model, specified neoblasts can divide asymmetrically to give rise to unspecified neoblasts that can adopt other fates. So stable labelling of a class of specified neoblasts would label other fates over time. But, precise temporal labelling of neoblast classes found here may allow the verification of lineages through the post-mitotic progenitor and mature cell stages. This would be useful for lineage tracing between neural-specified neoblasts and post-mitotic progenitor populations, which are difficult to connect computationally because of a lack of transcriptional similarity. Because the labelling of neoblast classes is only as specific as the FSTF driver(s) selected, this work can provide a basis for that selection.

While this work does not specifically look at some debated subjects in neoblast biology, it can still provide some insight into them. The presence of an unspecified clonogenic neoblast is not clearly present in our data. This supports the idea that there may be no dedicated clonogenic neoblast and that specialized neoblasts dividing to generate unspecified neoblasts can be clonogenic. Further work on this could include a dedicated search for dividing neoblast lacking FSTF signatures associated with specific fates, which would be expected for a dedicated clonogenic neoblast. The mechanism of asymmetric neoblast division generating an unspecified G_1 neoblast from a specified dividing neoblast has also not been studied. Dividing neoblasts in cytokinesis seem to

show an asymmetric localization of the neoblast marker *smedwi-1* and transcripts for FSTFs (Raz et al., 2021), indicating asymmetric divisions may involve the localization of transcripts by selective segregation or degradation. Though in cases where neoblasts numbers must expand, such as after sublethal irradiation or during animal growth, neoblasts must divide symmetrically to generate two neoblasts. It is not known if all neoblasts are specified before division in the context of neoblast expansion, if some specified neoblasts can lose fates in both neoblast daughter cells, or if recently divided G_1 neoblasts can retain fates in this specific context.

Modern single-cell RNA sequencing methods allow for the detection of over 30% of transcripts in each cell, with tens to hundreds of thousands of cells being sequenced (Zheng et al., 2017). Advancements are also being made in methods for the analysis of single-cell sequencing datasets, but most of these are minor improvements in existing methods like data clustering, dimensionality reduction-based visualization, and computational methods to connect cells by gene expression similarity to relate them in pseudotime. These have limitations, such as grouping cells into discrete clusters, which is convenient for analysis but not always representative of biology. Generally, gene enrichment analysis is then performed through statistical tests of gene expression levels between these discrete clusters, but genes can exist in regulatory networks and have complex roles that simple enrichment wouldn't necessarily uncover. We developed a machine learning method to learn gene expression features defining single cells. We used these machine learning networks in the classification of cells according to discrete clustering, which uncovers gene signatures that best classify the cell types. While genes important in the classification are often enriched in cells of those clusters, neural

networks can learn more complex relationships among genes that define cell types. Since the training of the network only adds numerical weights and biases to the network, the final classification network can be expressed as a formula for the classification probability of each class in terms of the expression of every gene. Analysis of the learned weights of the network shows genes that are most useful for an accurate classification of the cell, which is the gene signature that best defines it. Since neural networks learn gene expression patterns from individual cells instead of statistical measures across entire averaged discrete clusters, it can learn differences even among cells of the same identity and use different logic for classifying those cells. For example, if two cell types existed within one cluster, such as two transcriptionally similar mature cell types, the network can learn two classifications rulesets for the cells or it can learn a classification ruleset that works on all cells of the cluster, based on the average transcriptome of the cell. Which approach the network uses in learning the classification depends on the power of the network (size and depth of the network), its training rules (regularization forces more efficient encodings, which will favor learning fewer classification rules in some cases), and which produces a more accurate classification.

The architectures of neural networks are flexible. Instead of training a network to classify a cell to a discrete cluster, which we noted previously can be flawed, we can instead train a neural network to learn gene expression patterns based on each cell individually. An autoencoder is a special type of neural network architecture that efficiently learns and encodes patterns in data. It does this by compressing the data with each subsequent layer, and trying to recreate the original data using this highly

compressed version. Through this, it learns the most efficient way to encode the data. For example, in our autoencoder the deepest encoding layer (termed the ‘latent space’), is only 7-nodes which must represent and be used to reconstruct the input 714 transcription factor expression levels for each cell. By doing this, the autoencoder learns patterns in gene expression in different cell types by treating each cell independently, but learning patterns that occur across many cells.

Neural networks are powerful because they are customizable. One can change the architecture of the network for your specific task and can modify training parameters to change how the network learns. We implement various forms of regularization in our training to favor the network to learn based on as few genes as possible, which forces the network to learn which genes are the most important defining a cell type. Neural networks can also learn to classify cells based on parameters other than gene expression, such as characteristics of the cell’s morphology, and can be used to classify based on labels other than cluster identity, such as position in the animal.

Future work using neural networks to learn gene expression patterns in single cells could focus on different implementations and improve the interpretability of the data. In this thesis, I examine weights and biases, network activation patterns, classification confidence, and utilize accuracy metrics and error analysis to interpret what the network has learned about the data and use it to infer information about the sequenced cells. Uncovering complex patterns learned by neural networks is difficult however. Examining weights or randomly permutating input genes and looking for if and how cells are misclassified can give some insight into how each gene is used to classify cells. The functional importance of these genes could be even tested using gene

inhibition, to see if these misclassifications are biologically meaningful. Potentially, more specific neural network architectures can be used for examining specific regulatory relationships between genes. For example, if two genes are required for the specification of a cell type, a traditional neural network might assign high weights to both genes, which would cause a high activation of a common output node that sums both gene's expression values (e.g., 1rpm gene1 + 1rpm gene2 = 2 output activation). However, high expression of either gene alone might also be enough to activate the output node (e.g., 2rpm gene1 + 0rpm gene2 = 2 output activation). A neural network could instead be constructed that multiplies inputs to allow for a coexpression rule, where both genes would need to be expressed to activate the output (e.g., 2rpm gene1 * 0rpm gene2 = 0 output activation).

Overall, this thesis characterizes planarian neoblast and post-mitotic progenitor classes by gene expression and, more specifically, through their expression of FSTFs, many of which act to specify or maintain cell identity. We used this data to propose a model in which cell type diversity arises at different stages for different tissue types. We also develop a neural network-based method to analyze and learn gene expression patterns from single-cell sequencing datasets. Hopefully this work will serve as a foundation for future studies in the biology of planarian fate specification and expand the methods used to analyze it.

References

Atabay, K. D., LoCascio, S. A., de Hoog, T., & Reddien, P. W. (2018). Self-organization and progenitor targeting generate stable patterns in planarian regeneration. *Science*, *360*(6387), 404-409. https://doi.org/10.1126/science.aap8179

Carpenter, K., Morita, M., & Best, J. (1974). Ultrastructure of the photoreceptor of the planarian *Dugesia dorotocephala*. I. Normal eye. *Cell Tissue Res.*, *148*(2), 143-158.

Eisenhoffer, G. T., Kang, H., & Sánchez Alvarado, A. (2008). Molecular analysis of stem cells and their descendants during cell turnover and regeneration in the planarian *Schmidtea mediterranea*. *Cell Stem Cell*, *3*(3), 327-339. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18786419

Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M., & Reddien, P. W. (2018). Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science*, *360*(6391), 874. https://doi.org/10.1126/science.aaq1736

Hyman, L. H. (1951). *The Invertebrates: Platyhelminthes and Rhynchocoela The acoelomate bilateria* (Vol. II). McGraw-Hill Book Company Inc.

MacRae, E. (1967). The fine structure of sensory receptor processes in the auricular epithelium of the planarian, Dugesia tigrina. *Z. Zellf.*, *82*, 479-494.

Morgan, T. H. (1898). Experimental studies of the regeneration of Planaria maculata. *Archiv für Entwicklungsmechanik der Organismen*, *7*, 364-397.

Morgan, T. H. (1901). *Regeneration*. The Macmillan Company.

Newmark, P., & Sánchez Alvarado, A. (2000). Bromodeoxyuridine specifically labels the regenerative stem cells of planarians. *Dev. Biol.*, *220*(2), 142-153.

Oderberg, I. M., Li, D. J., Scimone, M. L., Gavino, M. A., & Reddien, P. W. (2017). Landmarks in Existing Tissue at Wounds Are Utilized to Generate Pattern in Regenerating Tissue. *Curr Biol*, *27*(5), 733-742. https://doi.org/10.1016/j.cub.2017.01.024

Raz, A. A., Wurtzel, O., & Reddien, P. W. (2021). Planarian stem cells specify fate yet retain potency during the cell cycle. *Cell Stem Cell*. https://doi.org/10.1016/j.stem.2021.03.021

Ross, K. G., Omuro, K. C., Taylor, M. R., Munday, R. K., Hubert, A., King, R. S., & Zayas, R. M. (2015). Novel monoclonal antibodies to study tissue regeneration in planarians. *BMC Dev Biol*, *15*, 2. https://doi.org/10.1186/s12861-014-0050-9

Sánchez Alvarado, A., Newmark, P. A., Robb, S. M., & Juste, R. (2002). The *Schmidtea mediterranea* database as a molecular resource for studying platyhelminthes, stem cells and regeneration. *Development*, *129*(24), 5659-5665. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=12421706

Scimone, M. L., Srivastava, M., Bell, G. W., & Reddien, P. W. (2011). A regulatory program for excretory system regeneration in planarians. *Development*, *138*(20), 4387-4398. https://doi.org/138/20/4387 [pii]
10.1242/dev.068098

Trembley, A. (1744). *Mémoires pour servir à l'histoire d'un genre de polypes d'eau douce, à bras en forme de cornes, par A. Trembley*. chez Jean & Hermann Verbeek.

Tsonis, P. A., & Fox, T. P. (2009). Regeneration according to Spallanzani. *Dev Dyn*, *238*(9), 2357-2363. https://doi.org/10.1002/dvdy.22057

Tu, K. C., Cheng, L. C., H, T. K. V., Lange, J. J., McKinney, S. A., Seidel, C. W., & Sánchez Alvarado, A. (2015). *Egr-5* is a post-mitotic regulator of planarian epidermal differentiation. *Elife*, *4*, e10501. https://doi.org/10.7554/eLife.10501

van Wolfswinkel, J. C., Wagner, D. E., & Reddien, P. W. (2014). Single-Cell Analysis Reveals Functionally Distinct Classes within the Planarian Stem Cell Compartment. *Cell Stem Cell*, *15*(3), 326-339. https://doi.org/10.1016/j.stem.2014.06.007

Wurtzel, O., Oderberg, I. M., & Reddien, P. W. (2017). Planarian Epidermal Stem Cells Respond to Positional Cues to Promote Cell-Type Diversity. *Dev Cell*, *40*(5), 491-504 e495. https://doi.org/10.1016/j.devcel.2017.02.008

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., . . . Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nat Commun*, *8*, 14049. https://doi.org/10.1038/ncomms14049