Title: Cell Type Atlas and Lineage Tree Reconstruction of Whole Adult Animals by Single-Cell Transcriptomics

One Sentence Summary: Single-cell analysis reveals major planarian cell types, a single differentiation tree for the adult animal and genes linked to specific differentiation events.

Authors: Mireya Plass[1,†], Jordi Solana[1,†‡], F. Alexander Wolf[2], Salah Ayoub[1], Aristotelis Misios[1], Petar Glažar[1], Benedikt Obermayer[1,ᵀ], Fabian J. Theis[2,3], Christine Kocks[1] , and Nikolaus Rajewsky[1,*]

Affiliations:

[1] Laboratory for Systems Biology of Gene Regulatory Elements, Berlin Institute for Medical Systems Biology, Max-Delbrück Center for Molecular Medicine, Germany

[2] Helmholtz Zentrum München, German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany

[3] Department of Mathematics, Technische Universität München, Germany

[‡] Present address: Department of Biological and Medical Sciences, Oxford Brookes University, Oxford, UK

[ᵀ] Present address: Core Unit Bioinformatics, Berlin Institute of Health, Berlin, Germany

[†] These authors contributed equally to the work

* Correspondence to: rajewsky@mdc-berlin.de

Abstract:

Flatworms of the species *Schmidtea mediterranea* are immortal –adult animals contain a large pool of pluripotent stem cells that continuously differentiate to all adult cell types. Therefore, single-cell transcriptome profiling of adult animals should reveal mature and progenitor cells. Here, by combining perturbation experiments, gene expression analysis, a computational method that predicts future cell states from the transcriptional changes, and a novel lineage reconstruction method, we placed all major cell types onto a single lineage tree that connects all cells to a single stem cell compartment. We characterize gene expression changes during differentiation and discover cell types important for regeneration. Our results demonstrate the importance of single-cell transcriptome analysis for mapping and reconstructing fundamental processes of developmental and regenerative biology at unprecedented resolution.

Main Text:

Introduction

Understanding differentiation from stem cells into the different cell types that make up the human body is a central problem of basic and medical science. Although numerous basic mechanisms of cellular differentiation have been identified and many cell types have been characterized, it will require a huge coordinated undertaking to systematically map all human cell types and cellular differentiation states (*1*). Due to the advances in single-cell transcriptomics, it has already been possible to study the cell type composition of mammalian organs and tissues (*2–6*) as well as development stages (*7*, *8*). However, single-cell transcriptomics provide just a snapshot of the dynamics of the cell populations unless cells can be traced or tagged experimentally (*9–12*). Thus, reconstructing cell lineages from stem cells to differentiated cells remains a challenge. Recently, algorithms to order developmental states and compute lineage trees based on comparing single-cell transcriptomes have made considerable progress (*13–15*) and have revealed new insights into stem cell biology (*16*) and tissue differentiation (*17–20*). However, these algorithms have been developed for the study of differentiation in specific cell lineages or tissues, and are not suitable to reconstruct all the cell differentiation trajectories present in complex animals.

Given these problems, can single-cell transcriptomic approaches improve our molecular understanding of how stem cells differentiate into all the cell types of an entire complex adult animal? Freshwater planarians such as *Schmidtea mediterranea* offer a unique opportunity to answer this question. Planarians are immortal, as they contain as adults a large pool of pluripotent stem cells (neoblasts) that continuously differentiate to all mature cell types to turnover all tissues (*21*). Hence, all cell differentiation pathways are constantly active in adult

individuals. We therefore reasoned that an unbiased single-cell transcriptomics approach should yield terminally differentiated cell types as well as a large number of intermediate cellular states, making planarians an ideal model system to attempt the lineage reconstruction of a whole animal.

Here we performed highly parallel droplet-based single-cell transcriptomics, Drop-seq (*3*), to characterize planarian cell types. We molecularly characterized dozens of cell types and uncovered many new ones. By applying a newly developed algorithm, PAGA, which reconciles the principles of clustering and pseudotemporal ordering (*22*), and combining it with independent computational and experimental approaches, we derive a consolidated lineage tree that includes all identified cell types rooted to a single stem cell cluster. Along this tree, we identify 48 gene sets that are co-regulated during the differentiation of specific cell types. Finally, we used single-cell transcriptomics to characterize the cellular processes that happen during regeneration. Our results reveal a strong depletion of newly characterized cell types, suggesting that these cells are used as an energy source for regeneration.

A high resolution cell type atlas for planaria

To comprehensively characterize different cell types and progenitor stages present in adult planarians, we performed genome-wide expression profiling in individual cells using nanoliter droplets (Drop-seq) (*3*) of cells isolated from whole adult animals. These cells were obtained, after dissociation, by fluorescence-activated cell sorting (FACS), which separated intact live cells from dead cells and enucleated cellular debris (Fig. 1A). From 11 independent experiments, we captured a total of 21,612 cells. We detected on average 494 genes and ~970 transcripts (identified by using “unique molecular identifiers” (UMIs)) per cell. The individual datasets correspond to 5 wild type samples (10,866 cells), two RNAi samples (3314 cells), a high-DNA

content G2/M population corresponding to cycling planarian stem cells (typically defined as X-ray sensitive "X1 cells"; 981 cells) (*23*, *24*), and three wild type regeneration samples (6,451 cells; Table S1). Sequencing depth was comparable across samples (Fig. S1A). Biological replicates showed highly correlated gene expression profiles (Fig. S1B). Besides, all samples showed high correlation with published RNA-seq data from equivalent bulk cell populations (*24–27*) (Fig. S1C). We pooled and analyzed all single-cell datasets together using Seurat (*3*). 8 of 11 samples were fixed using methanol (*28*) or frozen with DMSO (*29*) to facilitate sample handling. To assess batch effects, we compared the overall quality across wild type samples. Cells from each batch were distributed similarly on the tSNE (Fig. S1D), which resulted in comparable proportions of cells per cluster (Fig. S1E, Table S2). Although we observed a mild bias in gene expression due to the preservation procedure of the samples, clustering was not affected (Fig. S1F). However, we observed differences in the number of UMIs per cell across clusters (Fig. S1G). Together, these analyses confirmed that sample preparation did not compromise data quality or introduce bias. Therefore, we clustered the expression profiles of the individual cells from all samples together using Seurat (*3*). In total we identified 51 cell clusters (Fig. 1B).

We elucidated the cell type identity of clusters by examining marker genes and comparing them to previous literature (Fig. S2A, B) (Supplementary Note 1). The largest cluster and 14 smaller clusters located in the center of the tSNE plot express combinations of well-known stem cell markers (Fig. S3A), such as *Smedwi-1*, *Smedtud-1*, and *bruli* (Fig. S3B). The remaining clusters corresponded to the previously described neural, epidermal, secretory, muscle, gut and protonephridia cell types (Fig. S2A, S2B). However, in each of these categories, we found several distinct clusters (Fig. 1B) that express different combinations of marker genes (Table S3,

Fig. S4). This result suggests that our approach can distinguish more cell types than previous studies.

Single-cell transcriptomics unveils previously uncharacterized cell types

In the 1980's, Baguñà and Romero used microscopy to morphologically characterize and count all major cell types in *Schmidtea mediterranea* (*30*). We used this resource as a reference to validate cell types identified by our Drop-seq data and cluster annotation. Even though the microscopy data is of a qualitative nature, we observed a strong correlation between it and our molecular,unbiased Drop-seq annotation (Fig. 1C), suggesting that FACS sorting, cryopreservation or fixation and cell capture in nanodroplets did not influence cell type proportions. We validated the identity of several clusters by designing RNA probes targeting marker genes and performing *in situ* hybridizations, both whole mount and in histological sections (Fig. S5). We could confirm major known cell types such as different types of neurons, muscle, protonephridia, epidermis and secretory cells. We identified the two main cell types of the planarian gut: phagocytes (Fig. 1D, red) and goblet cells (Fig. 1D, green), and discovered markers of planarian goblet cells for the first time, including a gene without apparent homologs in other phyla. We named this gene *bruixot*. We also distinguished body and pharynx muscle (Fig. 1E). General muscle markers colocalized with body muscle markers in all the body except in the pharynx (Fig. 1E). Pharynx muscle was characterized by the expression of *laminin* (*31*) (Fig. S5). The protonephridia cluster (0.3% of our wild type cells) contained the two main cell types of these organs, flame and tubular cells (*32*) (Fig. S4, S5). In some cases, cell clusters contain several similar subtypes that we cannot distinguish at this resolution. For instance, previously described markers of eye pigment cup cells and photoreceptor neurons (*33*) are

expressed in pigment and ChAT neurons 2 clusters respectively, indicating that the former are subtypes of the latter (Fig. S6).

We also validated a recently discovered epidermis cell type, which marks the boundary between the dorsal and ventral parts of planarians (Fig. S5) (34). Additionally, we identified an epidermal related pharynx cell type (Fig. S5) and several parenchymal cell types previously undescribed molecularly (Fig. 2B, Fig. S5). Among parenchymal clusters we found a diversity of non-overlapping cells types, including aqp+ and the psap+ parenchymal cells (Fig. 1F), which probably collectively correspond to the previously described fixed parenchymal cells (30, 35), pigment cells (cluster 44) (36, 37) and glial cells (38, 39) (cluster 47) (Fig. S4, S5). Altogether these results show that we can identify known as well as unknown cell types using single-cell transcriptomics and measure their abundances in a reproducible way.

To investigate the function of newly identified cell types we used pathway and gene set overdispersion analysis (PAGODA) (*40*) to identify variable gene sets with particular gene ontology (GO) terms annotated (Fig. S7, Supplementary Note 2). The clustering that emerges using these gene sets roughly recapitulates the one obtained with Seurat, showing the robustness of our clustering approach (Fig. S7A). This analysis revealed that neoblasts and progenitors are functionally similar, both expressing gene signatures enriched for GO terms related to RNA processing. Additionally, parenchymal clusters showed enrichment for GO terms related to *lysosome*, *extracellular region* and hydrolytic enzymes, and appear to share metabolic functions with gut cells (Fig. S7B).

Single-cell transcriptomics of purified stem cells and stem cell depleted animals reveals stem, progenitor and differentiated cell populations

The great diversity of cell types identified, which included stem cells, differentiated cells and presumably many progenitor cells, offered a unique opportunity for exploring stem cell differentiation and lineage relationships between all cell clusters. We focused on the X1 cell sample, which is enriched in G2/M neoblasts (*23*, *24*) and the *histone 2b* (*h2b*) RNAi treated whole planaria sample, in which stem and progenitor cell populations are depleted (*41*). Cells from these datasets showed a clear distribution pattern: X1 cells were located in the middle of the tSNE plot (Fig. 2A, red dots) while *h2b*(*RNAi*) resistant cells were clearly enriched in the periphery (Fig. 2A, blue dots). This distribution was specific and not the result of batch effects, as evident from the respective control samples (Fig. 2A, X1 control and *gfp*(*RNAi*) samples). Given that each dataset is enriched in particular cell populations, we reasoned that they could be used to distinguish cells in varying differentiation states. We quantified the fraction of cells per cluster from the X1 and *h2b*(*RNAi*) samples and compared them to wild type and control samples. We performed a principal component analysis (PCA) using these cellular proportions as well as the mean expression of the three top neoblast markers (*Smedwi-1*, *tub-α1* and *h2b*) in each cluster. The first two principal components resulting from this analysis separated clusters according to their gene expression profiles as neoblasts, progenitors, and differentiated cell clusters (Fig. 2B, Fig. S2A). Mapping onto the tSNE revealed that progenitor cell clusters were located between differentiated cells and neoblast clusters (Fig. 2B). To corroborate the differentiation state of the cells in the different clusters, we pooled the cells in each group and correlated their gene expression profiles to previously described FACS populations. Neoblasts clusters best correlated with X1 populations, corresponding to high content DNA G2/M

neoblasts, progenitor clusters correlated with X2 populations, a mixture of G1/S neoblasts and early progenitors, and differentiated cell clusters correlated with Xins samples, a pool of all differentiated cells (Fig. 2C). Altogether, our functional experiments reveal the stem, progenitor or differentiated status of each cell cluster.

Computational lineage reconstruction predicts a single tree for all major planarian cell differentiation trajectories

Existing methods to investigate cell differentiation using single-cell transcriptomics data were designed to study individual lineages or organs, allow few branching trajectories (*13*, *15*, *18*), and often require high sequencing depth and associated costs (*16*). To overcome these limitations, we developed the general framework of partition-based graph abstraction (PAGA), which reconciles clustering and pseudotemporal ordering algorithms and allows to infer complex cell trajectories and differentiation trees (*22*). Starting from the neighborhood-graph of single cells, in which cells are represented as nodes, the algorithm quantifies the connectivity of cell clusters and generates a much simpler abstracted graph in which nodes correspond to the clusters identified using Seurat and edges represent putative transitions between clusters. The differentiation tree is then computed as the tree-like subgraph in the abstracted graph that best explains all continuous progressions along the original single-cell graph (Supplementary Note 3).

When running this algorithm, without any assumptions about the tree structure, we obtained an abstracted graph that shows high confidence of the branching events (Fig. 3A) from which we can derive a single differentiation tree that included all the cell types and linked them to a single root, the neoblast 1 cluster. This tree defines independent differentiation branches for all the major tissues such as neurons, muscle, parenchyma and gut (Fig. 3A). Additionally, the tree reflects the relation between different groups of cells. For example, it predicts the existence of

independent progenitor cells for the epidermis dorso-ventral boundary and the pharynx cell type lineages although both lineages are related to the epidermal lineage. In contrast, it shows the presence of a shared progenitor for all parenchymal lineages despite containing cell types as different as glia and pigment cells. The connections in the tree are highly consistent with the continuity of gene expression patterns along the various lineages (Fig. S8A) except for two cases: the epidermis cluster itself is disconnected from epidermal lineage, and muscle pharynx is connected to muscle body instead of muscle progenitors (Fig. S8A). Together, from 51 clusters (with 1275 possible transitions between them) PAGA predicts 53 transitions that are mainly consistent with our marker based analysis.

Furthermore, PAGA yields a pseudotemporal ordering of individual cells within each cluster consistent with our stem cell ablation and purification experiments, and therefore confidently predicts their differentiation status, even for cell types for which separate progenitor clusters could not be identified (Fig. S8B). For instance, when we sorted the goblet cells by pseudotime, we observe a higher percentage of X1 cells in early pseudotime and *h2b*(*RNAi*) cells in the late pseudotime (Fig. S8B). To validate this observation, we performed double FISH of *bruixot*, our newly identified goblet cell marker, and *adb* (*aprenent de bruixot*), a gene expressed earlier in the goblet cluster pseudotime (Fig. S8B). Consistently, *adb* was expressed in the gut (Fig. S8C) overlapping with *bruixot*, but staining more cells located in the periphery of the gut that clearly lacked goblet cell morphology (Fig. S8D). This indicates that *adb* is a marker of immature goblet cells and that computationally estimated pseudotime correctly orders cells according to their differentiation status.

Although the tree predicts the connectivity of cell clusters, it does not give any information about the direction of the trajectories. Thus, we used the tree topology to estimate the

developmental potency of each cluster, i.e. their ability to give rise to other cells. We developed a potency score that is conceptually similar to the stemID score previously proposed to identify stem cells (*16*) but additionally estimates pluripotency vs. multi- or unipotency of cell populations. It is computed as the normalized degree of each cluster in the abstracted graph (Supplementary Note 4). This analysis showed that neoblast 1, the largest stem cell cluster, had a score of 1 (Fig. 3B), correctly assigning pluripotency to neoblasts as expected from earlier literature (*21*). We note that the potency score is independent of prior information and therefore can be used to identify stem cells from single-cell transcriptomics data alone, a feature that is particularly useful in less well-studied non-model organisms. Progenitor clusters showed lower potency than neoblasts, and higher potency scores than differentiated cells (Fig. 3B), in agreement with a gradual potency loss. To assess the stem cell and progenitor status of the clusters connected in the center of the PAGA topology, we mapped X1 and *h2b*(*RNAi*) data onto the tree. Most X1 cells were located in the neoblast 1 cluster (Fig. 3C) while *h2b*(*RNAi*) resistant cells were more enriched in the leaves of the tree (Fig. 3D). Thus, both PAGA and stem cell ablation and purification independently support the stem and progenitor status of these clusters.

The remaining neoblast clusters had lower potency scores than the neoblast cluster 1 and were connected to it. These clusters share the majority of marker genes with the neoblast 1 cluster (Table S3) and do not correspond to previously identified specialized neoblasts of the sigma, gamma and zeta class (*26*, *42*, *43*) (Fig. S9). Although some of these neoblast clusters are connected to differentiated cell types (Figure 3A), most do not give rise to differentiated cell types, raising the possibility that they represent neoblasts in different metabolic, cell cycle or activation states (Supplementary Note 5).

We detect expression of specialized neoblast markers among both neoblast and progenitor clusters (Fig. S10, S11). While present in neoblasts, sigma markers were most highly expressed in neural and muscle progenitors, gamma markers in gut and parenchymal progenitors and zeta markers in epidermal progenitors (Figure S9D). These clusters are mostly devoid of X1 cells (Figure 3C) and therefore correspond mainly to post-mitotic progenitors.

RNA velocity confirms lineage relationships predicted by PAGA

To independently validate the differentiation trajectories predicted by PAGA we used velocyto (*44*). This method computes RNA *velocity*, defined as the rate of change of mRNA levels for a gene in time, in every single cell. In differentiating cells in which changes in gene expressionare dominated by changes in transcription rates, the ratio of unspliced to spliced reads for a given gene within a cell will be proportional to the temporal change of the logarithm of spliced reads (or mature mRNAs) (*44*). Thus, one can estimate the future mRNA level of a gene by computing its *velocity* and a linear fit. By aggregating over many genes in a cell, one can estimate the cellular expression state to which the cell is apparently moving in time. We estimated mRNA velocities for each cell and projected the estimated future states of cells onto the tSNE, which describe the paths predicted by the mRNA velocity model (Fig. 3E and Fig. S12A). These paths show a highly homogenous stem cell population that moves slowly to progenitors, which will differentiate to mature cell types. The long arrows at the edges of the clusters likely are due to the averaging on the force field, as they do not appear when individual arrows are plotted (Fig. S12A). These paths largely agree with the trajectories predicted by PAGA, and also confirmed the connection between muscle progenitors and pharynx muscle predicted from gene expression changes (Fig. S12A). Additionally, velocyto can also model longer cell trajectories in order to identify their root (Fig. 3F) and terminal end points (Fig. 3G),

which corresponded to the tSNE regions containing stem cells and terminally differentiated cells, respectively. Velocyto cannot provide information from disconnected clusters. As a result, all disconnected clusters contain differentiation trajectories with independent start and end points (Fig. 3F, G).

The estimates of RNA dynamics obtained with velocyto also identified regions where genes are mainly induced or repressed compared to the steady state level. This information can be helpful to investigate relations between clusters that appear disconnected on the tSNE. We used these estimates to study the expression of marker genes from the epidermis cluster. These genes are clearly induced in epidermal progenitors and repressed in mature epidermis, where they are mainly expressed (Fig. S12B). Thus, mRNA metabolism patterns provide additional support to the differentiation trajectory connecting late epidermal progenitors to epidermis that we predicted based on gene expression changes (Fig. S8A).

A consolidated lineage tree of planarian stem cell differentiation into all major cell types

Taken together, our results show that both computational and experimental methods agree in the identification of stem cells, progenitors and differentiated cells. By combining all four independent lines of evidence (PAGA, gene expression changes, stem cell ablation and enrichment experiments, and velocyto) we provide a single consolidated tree that models stem cell differentiation trajectories into all identified cell types of adult planarians (Fig. 4A). The resulting cell lineage tree correctly recapitulates the known expression changes described during epidermal differentiation (*26*, *34*, *45*) (Fig. 4B). We observed a continuous decrease of the expression of *Smedwi-1*, a well characterized neoblast marker (Fig. S3), with pseudotime progression whereas early (*prog-1*) and late (*agat-1)* epidermal progenitor as well as mature epidermis (*vim-1*) markers increased their expression at consecutive time points (Fig. 4B).

According to the consolidated lineage tree, neoblasts (35% of our wild type cells) differentiate into at least 23 independent cell lineages. There are 6 major differentiation fates (57% of cells) (Table S2), each representing more than 1% of total cells: epidermal, parenchymal, neural, muscle, gut, and a pharynx cell type. For these major fates, we identified progenitor and differentiated states. Additionally, we identified 10 minor lineages (6% cells; each less abundant than 1% of total cells) that differentiate from the neoblasts, but for which we were unable to identify progenitors.

Self-organizing maps identify gene programs underlying cell differentiation

We used our data to identify gene sets that coordinately change their expression during differentiation. For this analysis, we discarded all cells from neoblast clusters that did not give rise to differentiated cell types in our consolidated cell lineage tree. The remaining cells were ordered following the tree for each lineage and sorted within each cluster according to their pseudotemporal ordering (Fig. 5A). Subsequently, we used self-organizing maps (SOMs) (*46*) to identify 48 sets of highly variable genes that coordinately change their expression during differentiation (*47*) (Fig. 5B, S13 and Table S4). Many of these sets contain some genes previously known to be expressed in the respective lineages and in some cases involved in their differentiation (Table S5). For instance, gene sets 10 and 11 contain genes that are highly expressed in neoblast and progenitor clusters, such as *Smedwi-1* and *tub-α-1*, whose expression drops during differentiation (Fig. 5B and S13). Similarly, we found gene sets that are regulated along muscle, neuronal, parenchymal, gut and epidermal differentiation (Fig. 5B, top row). They contained genes expressed in these lineages, such as *mhc* for the muscle and *chat* for the neuronal lineage, but also included well-known regulators of their differentiation, such as *myoD* (*48*) and *coe* (*49*). As a consequence of analyzing all detected planarian cell lineages

simultaneously, we not only identified gene sets involved in lineage specific programs but also gene sets co-regulated during the differentiation of several fates (Fig. 5B, mid and bottom row). Taken together, these results show that single-cell transcriptomics of a whole organism allows the reconstruction of specific differentiation events for many differentiation fates in parallel, enabling the identification of previously undetected combinations of co-regulated genes.

Molecular profiling of planarian regeneration by single-cell transcriptomics

Freshwater planarians are well known for their remarkable regenerative capacities. Planarians can be cut into small pieces, and each piece (except for the pharynx and the most anterior tip of the head which are devoid of neoblasts) can regenerate a complete, albeit much smaller, organism in a matter of days. This process is dynamically complex and involves the orchestration of all cellular differentiation pathways. The animal does not grow (as it cannot eat) during the process. Thus, the truncated body fragments need to reshape their body proportions to adjust to their new size by the process termed morphallaxis (*50*). It is still largely unknown how each individual cell type behaves in this process.

Given the detected cell type abundances and the cell differentiation tree of steady state adult animals, we asked if we could use Drop-Seq to profile the cellular and transcriptomic changes that occur during regeneration. We cut planarians in 5-7 pieces, discarded the head piece and prepared the remaining body pieces for single-cell transcriptomics immediately after cutting (day 0), and 2 and 4 days after cut (Fig. 6A, Fig. S14). We compared regenerating samples to day 0 using Seurat and detected hundreds of differentially expressed genes in both samples (Table S6, S7). By pooling all cells we were able to detect upregulation after 2 days of regeneration of 16 of the 128 wound induced genes described in a previous study (*42*, *51*) (Table S6, Fig.S15A). The shallowness of Drop-seq data makes difficult to assess differences in lowly expressed genes.

However, Drop-seq allows distinguishing the cell types that undergo these changes, showing that *runt-1* and *egr-2* are upregulated in the neoblast 1 cluster (Fig. S15B) and *jun-1* in the muscle body cluster (Fig. S15C; Table S7, 8).

All cells from the regenerating samples fall into clusters that are present in wild type samples, indicating an absence of regeneration-specific types or trajectories (Table S2). However our analysis revealed significant changes in cell composition during regeneration: on one hand, we observed a large increase in the number of neoblasts, consistent with an increase in mitotic activity, and of neural progenitors, reflecting active neurogenesis to replace missing brain structures after head removal (Fig. 6B, Fig. S16). On the other hand, we detected that both parenchymal progenitor cells and differentiated parenchymal cell types were depleted (Fig. 6B, Fig. S16), indicating that these cells are cleared in the process of reshaping the planarian tissue. The cell proportion changes at day 2 and 4 were clearly correlated (Fig. 6C), indicating that *aqp*+ parenchymal cells are the most depleted cell type. We experimentally confirmed this observation by *in situ* hybridization on planarian tissue sections (Figure 6D) and counting *aqp*+ parenchymal cells (Figure 6E) (Mann Whitney U-test p-value < 1e-7). Our results indicate that parenchymal cells are highly depleted upon regeneration, implying that they may be used to metabolically fuel the regeneration process (*52*).

Discussion

In this study, we used the stem cell population and the extreme regeneration capabilities of adult flatworms to generate an atlas of cell types at high resolution. We identified, quantified, and molecularly characterized 37 cell types including 23 terminally differentiated cell types, and numerous progenitor and stem cells clusters. Although our sequencing data are relatively shallow, molecular characterization of cell types using computational methods was robust,

agreed well with previously published microscopy data, and revealed progenitor and differentiated cells. This implies that the grouping of incomplete transcriptomes of thousands of cells into clusters did not suffer significantly from capture rates or other confounding factors.

The resolution of our data depends on both the number of cells sequenced and the number of genes detected per cell. Considering only wt and control samples (~11,000 cells), we can identify differentiated cell clusters containing about 10 cells. Therefore, we estimate that cells present at a frequency of <1:1000, such as such as *cintillo*+ cells (*53*) and photoreceptor neurons (*33*), will be missed by our approach. Besides, we failed to identify certain neoblast subpopulations previously described in the literature (*26*, *42*). This result could be due to the low sensitivity of Drop-seq, which captures only a fraction of mRNAs in a cell. However, we do detect the expression of the proposed marker genes of these subpopulations. They appear to be spread among neoblasts and progenitor clusters (Fig.S8), which still express neoblast markers such as *Smedwi-1* at low levels (Fig.S3, 11). This indicates that the boundary between stem cells and lineage-committed progenitors is probably not sharp. Further studies will help describing and delimiting these boundaries.

Projecting high dimensional gene expression data of thousands of transcriptomes onto a two-dimensional plot (for example by the widely used tSNE method (*54*)) visually reveals clusters. However, it is impossible to infer the relationships among them, as the distances between clusters cannot be interpreted as differentiation trajectories. To solve this problem, we used computational and experimental methods to reconstruct a lineage tree. PAGA and velocyto provide two complementary approaches to study cell differentiation using single-cell transcriptomics. While velocyto allows finding the differentiation trajectories of individual cells within a cell continuum based on RNA metabolism, PAGA allows inferring the average

differentiation paths of a group of cells, even when they are disconnected. Thus, combining these two computational methods results in a robust lineage prediction. This prediction is supported by the continuity of expression of marker genes and the mapping of stem cells and differentiated cells on the tree (Fig. 4A), and validates known differentiation trajectories such as that of the epidermal lineage (Fig. 4B) (*34*).

We used PAGA (*22*) to reconstruct in an unbiased way the lineage tree of all major planarian cells. This method, although indirect, allows reconstructing the lineage information from the transcriptomic snapshot of individual cells. Besides, in contrast to high throughput lineage tracing methods (*9–12*), which rely on using transgenic or CRISPR/Cas tools, it can be applied to every species provided that single cells can be isolated and sequenced. Using this method, we identified *de novo* planarian stem cells and predict their differentiation paths to at least 23 different lineages, including several multipotent progenitor populations. Importantly, these tools can readily be applied to other organism to identify *de novo* stem cells, identify their differentiation trajectories and estimate the developmental potency of the resulting cell populations.

Pseudotemporal ordering of cells along these lineages allowed us to discover gene sets that are putatively involved in differentiation programs, highlighting the similarities and differences that exist across tissues, and identifying several genes known to be involved in cell differentiation not only in planarians but also in other species, such as *myoD*, *nkx6* and *pax6*. Further characterization of these gene sets should be the subject of future studies. As we show for the *h2b* RNAi phenotype, Drop-seq also allows profiling perturbation studies at both the transcriptomic and cellular levels. In general, and beyond planaria, we can foresee that future

studies will use single-cell transcriptomics coupled to loss-of-function experiments to unravel the specific developmental functions of genes or regulatory networks.

Furthermore, we used single-cell transcriptomics to profile cellular abundance changes upon regeneration. Our experiments revealed that several of our newly described parenchymal types are depleted in regeneration. These cell types had been largely overlooked in molecular studies but had been described, based on microscopy, in the literature decades ago. This is in part due to the unbiased nature of both microscopic and single-cell transcriptomic data. We note that these parenchymal cells are highly enriched in lysosomes and other vacuoles and might be an energy reservoir that regenerating planarians mobilize to fuel regeneration.

To make our data easily accessible, we built an interactive app that allows to query and interpret all sequencing data (https://shiny.mdc-berlin.de/psca). We also provide a detailed tutorial for the lineage reconstruction algorithm PAGA that we hope will serve as a reference for future studies (https://github.com/rajewsky-lab/planarian_lineages). Together, our results show that single-cell expression profiling can be used to systematically annotate cell types of entire animals, to reconstruct stem cell differentiation lineages of whole organisms, and to study complex processes such as regeneration (and their relation to lineages used in normal development) at single-cell resolution. Our results and methods demonstrate that single-cell approaches will become an indispensable method to study developmental and regeneration biology.

Methods Summary

Single-cell transcriptomic profiling of asexual adult planarians from the species *Schmidtea mediterranea* was performed using Drop-seq. Single cell suspensions were prepared by dissociating cells from adult planarians of 4-10 mm in length using trypsin. We used FACS to

discard broken and dead cells. Cells were either directly processed for Drop-seq or preserved in methanol or DMSO for later processing. For RNAi experiments, animals were injected with dsRNA against the coding region of *h2b* or *gfp* for three consecutive days, kept at 20°C, and their cells prepared for FACS and single cell transcriptomics 5 days after the third injection. For regeneration experiments, animals raging 4-10 mm in size were cut in 5-7 pieces, the head pieces were discarded and the remaining pieces were processed for Drop-seq immediately, 2 or 4 days after cut.

Computational analysis of the sequenced samples was done using Dropseq tools and the Seurat package (*3*). Briefly, reads were mapped to the *S. mediterranea* dd_Smed_v6 transcriptome and processed using Dropseq tools and custom perl scripts to generate Digital Gene Expression (DGE) matrices for each sample. Finally, all DGE matrices were joined. Variable genes across all clusters were used to perform a Principal Component Analysis (PCA). The first 50 PCs obtained were then tested for significance and those with a p-value < 10e-5 were used to perform clustering. The robustness of the obtained clusters was assessed and spurious clusters were merged to obtain a final set of 51 clusters.

Cell type identification was performed by calculating marker genes for each cluster. Manual inspection, comparison to previously published single-cell data, and experimental validation using in situ hybridizations of the marker genes reported allowed the identification of the different cell populations. Additional characterization of the identified cell types was performed by characterizing GO-term based gene sets using PAGODA (*40*). Experimental validation of cell types was done using whole mount *in situ* hybridization and *in situ* hybridization on histological sections as previously described and using probes complementary to marker genes (Table S9).

Lineage reconstruction was done by combining the unsupervised graph obtained with the PAGA algorithm (*22*) with velocyto (*44*), gene expression analysis, and experimental data from h2b(*RNAi*) and X1 facs sorted cells. To calculate RNA velocity with velocyto, we mapped the reads from all datasets to the planarian genome in order to extract spliced and unspliced reads. These analyses allowed us to obtain a pseudotemporal order of cells that was used to identify gene sets that change during stem cell differentiation using self-organizing maps.

To preform cell counting of regenerating planarians, positive cells were automatically counted using a custom script for ImageJ (https://imagej.net).

References and Notes

1. A. Regev *et al.*, The Human Cell Atlas. *Elife*. **6** (2017), doi:10.7554/eLife.27041.
2. D. A. Jaitin *et al.*, Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*. **343**, 776–779 (2014).
3. E. Z. Macosko *et al.*, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. **161**, 1202–1214 (2015).
4. K. Shekhar *et al.*, Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell*. **166**, 1308–1323 e30 (2016).
5. A. C. Villani *et al.*, Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*. **356,** pii: eaah4573 (2017), doi:10.1126/science.aah4573.
6. D. Grun *et al.*, Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*. **525**, 251–255 (2015).
7. N. Karaiskos *et al.*, The Drosophila embryo at single-cell transcriptome resolution. *Science*. **358**, 194–199 (2017).
8. J. Cao *et al.*, Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*. **357**, 661–667 (2017).
9. A. McKenna *et al.*, Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*. **353,** pii: aaf7907 (2016), doi: 10.1126/science.aaf7907.
10. K. L. Frieda *et al.*, Synthetic recording and in situ readout of lineage information in single cells. *Nature*. **541**, 107–111 (2016).
11. B. Spanjaard, B. Hu, N. Mitic, J. P. Junker, Massively parallel single cell lineage tracing using CRISPR/Cas9 induced genetic scars. *bioRxiv* (2017), doi:101101/205971.

12. A. Alemany, M. Florescu, C. S. Baron, J. Peterson-Maduro, A. van Oudenaarden, Whole-organism clone tracing using single-cell sequencing. *Nature*. **556**, 108–112 (2018).

13. L. Haghverdi, M. Buttner, F. A. Wolf, F. Buettner, F. J. Theis, Diffusion pseudotime robustly reconstructs lineage branching. *Nat Methods*. **13**, 845–848 (2016).

14. M. Setty *et al.*, Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol*. **34**, 637–645 (2016).

15. X. Qiu *et al.*, Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods*. **14**, 979–982 (2017).

16. D. Grün *et al.*, De Novo Prediction of Stem Cell Identity using Single-Cell Transcriptome Data. *Cell Stem Cell*. **19**, 266–277 (2016).

17. F. Notta *et al.*, Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science*. **351,** pii: aab2116 (2016), doi: 10.1126/science.aab2116.

18. J. Shin *et al.*, Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. *Cell Stem Cell*. **17**, 360–372 (2015).

19. B. Treutlein *et al.*, Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*. **509**, 371–375 (2014).

20. L. Velten *et al.*, Human haematopoietic stem cell lineage commitment is a continuous process. *Nat Cell Biol*. **19**, 271–281 (2017).

21. D. E. Wagner, I. E. Wang, P. W. Reddien, Clonogenic neoblasts are pluripotent adult stem cells that underlie planarian regeneration. *Science*. **332**, 811–816 (2011).

22. F. A. Wolf *et al.*, Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *bioRxiv* (2017), doi:101101/208819.

23. T. Hayashi, M. Asami, S. Higuchi, N. Shibata, K. Agata, Isolation of planarian X-ray-

sensitive stem cells by fluorescence-activated cell sorting. *Dev Growth Differ*. **48**, 371–380 (2006).

24. P. Onal *et al.*, Gene expression of pluripotency determinants is conserved between mammalian and planarian stem cells. *EMBO J*. **31**, 2755–2769 (2012).

25. R. M. Labbe *et al.*, A comparative transcriptomic analysis reveals conserved features of stem cell pluripotency in planarians and mammals. *Stem Cells*. **30**, 1734–1745 (2012).

26. J. C. van Wolfswinkel, D. E. Wagner, P. W. Reddien, Single-cell analysis reveals functionally distinct classes within the planarian stem cell compartment. *Cell Stem Cell*. **15**, 326–339 (2014).

27. J. Solana *et al.*, Conserved functional antagonism of CELF and MBNL proteins controls stem cell-specific alternative splicing in planarians. *Elife*. **5**, e16797 (2016).

28. J. Alles *et al.*, Cell fixation and preservation for droplet-based single-cell transcriptomics. *BMC Biol*. **15**, 44 (2017).

29. A. Guillaumet-Adkins *et al.*, Single-cell transcriptome conservation in cryopreserved cells and tissues. *Genome Biol*. **18**, 45 (2017).

30. J. Baguna, R. Romero, Quantitative-Analysis of Cell-Types during Growth, Degrowth and Regeneration in the Planarians Dugesia-Mediterranea and Dugesia-Tigrina. *Hydrobiologia*. **84**, 181–194 (1981).

31. F. Cebria, P. A. Newmark, Morphogenesis defects are associated with abnormal nervous system regeneration following roboA RNAi in planarians. *Development*. **134**, 833–837 (2007).

32. J. C. Rink, H. T. Vu, A. Sanchez Alvarado, The maintenance and regeneration of the planarian excretory system are regulated by EGFR signaling. *Development*. **138**, 3769–

3780 (2011).

33. S. W. Lapan, P. W. Reddien, Transcriptome Analysis of the Planarian Eye Identifies ovo as a Specific Regulator of Eye Regeneration. *Cell Rep*. **2**, 294–307 (2012).

34. O. Wurtzel, I. M. Oderberg, P. W. Reddien, Planarian Epidermal Stem Cells Respond to Positional Cues to Promote Cell-Type Diversity. *Dev Cell*. **40**, 491–504 e5 (2017).

35. K. J. Pedersen, Studies on the nature of planarian connective tissue. *Zeitschrift für Zellforsch und Mikroskopische Anat*. **53**, 569–608 (1961).

36. B. M. Stubenhaus *et al.*, Light-induced depigmentation in planarians models the pathophysiology of acute porphyrias. *Elife*. **5** (2016), doi:10.7554/eLife.14175.

37. C. Wang *et al.*, Forkhead containing transcription factor Albino controls tetrapyrrole-based body pigmentation in planarian. *Cell Discov*. **2**, 16029 (2016).

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

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

40. J. Fan *et al.*, Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nat Methods*. **13**, 241–244 (2016).

41. J. Solana *et al.*, Defining the molecular profile of planarian pluripotent stem cells using a combinatorial RNAseq, RNA interference and irradiation approach. *Genome Biol*. **13**, R19 (2012).

42. O. Wurtzel *et al.*, A Generic and Cell-Type-Specific Wound Response Precedes

Regeneration in Planarians. *Dev Cell*. **35**, 632–645 (2015).

43. A. M. Molinaro, B. J. Pearson, In silico lineage tracing through single cell transcriptomics identifies a neural stem cell population in planarians. *Genome Biol*. **17**, 87 (2016).

44. G. La Manno *et al.*, RNA velocity in single cells. *bioRxiv* (2017), doi: 10.1101/206052.

45. G. T. Eisenhoffer, H. Kang, A. Sanchez Alvarado, Molecular analysis of stem cells and their descendants during cell turnover and regeneration in the planarian Schmidtea mediterranea. *Cell Stem Cell*. **3**, 327–339 (2008).

46. T. Kohonen, The self-organizing map. *Proc IEEE*. **78**, 1464–1480 (1990).

47. P. Tamayo *et al.*, Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*. **96**, 2907–12 (1999).

48. M. L. Scimone, L. E. Cote, P. W. Reddien, Orthogonal muscle fibres have different instructive roles in planarian regeneration. *Nature*. **551** (2017), doi:10.1038/nature24660.

49. M. W. Cowles, K. C. Omuro, B. N. Stanley, C. G. Quintanilla, R. M. Zayas, COE Loss-of-Function Analysis Reveals a Genetic Program Underlying Maintenance and Regeneration of the Nervous System in Planarians. *PLoS Genet*. **10**, e1004746 (2014).

50. K. Agata, Y. Saito, E. Nakajima, Unifying principles of regeneration I: Epimorphosis versus morphallaxis. *Dev Growth Differ*. **49**, 73–78 (2007).

51. D. Wenemoser, P. W. Reddien, Planarian regeneration involves distinct stem cell responses to wounds and tissue absence. *Dev Biol*. **344**, 979–991 (2010).

52. C. Gonzalez-Estevez, D. A. Felix, A. A. Aboobaker, E. Salo, Gtdap-1 promotes autophagy and is required for planarian remodeling during regeneration and starvation. *Proc Natl Acad Sci U S A*. **104**, 13373–13378 (2007).

53. N. J. Oviedo, P. A. Newmark, A. Sanchez Alvarado, Allometric scaling and proportion regulation in the freshwater planarian Schmidtea mediterranea. *Dev Dyn*. **226**, 326–333 (2003).

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

55. J. Solana *et al.*, The CCR4-NOT Complex Mediates Deadenylation and Degradation of Stem Cell mRNAs and Promotes Planarian Stem Cell Differentiation. *PLoS Genet*. **9**, e1004003 (2013).

56. H. Brandl *et al.*, PlanMine - a mineable resource of planarian biology and biodiversity. *Nucleic Acids Res* (2015), doi:10.1093/nar/gkv1148.

57. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).

58. V. D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *J Stat Mech Theory Exp*. **2008**, P10008 (2008).

59. A. Cardona, J. Fernandez, J. Solana, R. Romero, An in situ hybridization protocol for planarian embryos: monitoring myosin heavy chain gene expression. *Dev Genes Evol*. **215**, 482–488 (2005).

60. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. **19**, 15 (2018).

61. S. M. Robb, E. Ross, A. Sanchez Alvarado, SmedGD: the Schmidtea mediterranea genome database. *Nucleic Acids Res*. **36**, D599-606 (2008).

62. T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, Som pak: The self-organizing map program package. *Rep A31, Helsinki Univ Technol Lab Comput Inf Sci* (1996).

63. D. van Dijk *et al.*, MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *bioRxiv* (2017), doi:10.1101/111591.

64. A. McDavid *et al.*, Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. *Bioinformatics*. **29**, 461–467 (2013).

65. J. J. Collins *et al.*, Genome-Wide Analyses Reveal a Role for Peptide Hormones in Planarian Germline Development. *PLoS Biol*. **8**, e1000509 (2010).

66. T. M. J. Fruchterman, E. M. Reingold, Graph drawing by force-directed placement. *Softw Pract Exp*. **21**, 1129–1164 (1991).

67. A. H. Rizvi *et al.*, Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nat Biotechnol*. **35**, 551–560 (2017).

68. P. Qiu *et al.*, Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat Biotechnol*. **29**, 886–891 (2011).

69. C. Trapnell *et al.*, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*. **32**, 381–386 (2014).

70. G. Giecold, E. Marco, S. P. Garcia, L. Trippa, G.-C. Yuan, Robust lineage reconstruction from high-dimensional single-cell data. *Nucleic Acids Res*. **44**, e122–e122 (2016).

71. Z. Ji, H. Ji, TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res*. **44**, e117–e117 (2016).

72. J. Chen, A. Schlitzer, S. Chakarov, F. Ginhoux, M. Poidinger, Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. *Nat Commun*. **7**, 11988 (2016).

73. X. Qiu *et al.*, Single-cell mRNA quantification and differential analysis with Census. *Nat Methods*. **14**, 309–315 (2017).

74. K. A. Janes, Single-cell states versus single-cell atlases - two classes of heterogeneity that differ in meaning and method. *Curr Opin Biotechnol*. **39**, 120–5 (2016).

Acknowledgements

We thank Jonathan Alles and Anastasiya Boltengagen for help with single-cell sequencing and members of the Rajewsky lab for many discussions. We thank Sten Linnarsson and Gioele La Manno for helping to run velocyto. We also thank Nir Friedman and Angélica García-Pérez for useful comments. The work in this project was funded by the German center for Cardiovascular Research (DZHK BER 1.2 VD) and a DFG grant RA 838/5-1. FAW acknowledges the support of the Helmholtz Postdoc Program, Initiative and Networking Fund of the Helmholtz Association. Author contributions: JS, MP, CK and NR designed the project. JS, CK, and NR designed all experiments. SA, JS, and CK performed all experiments. MP performed and coordinated computational analyses. FAW developed PAGA with the supervision of FJT. AM performed velocyto analyses. PG developed the Shiny app and performed image analysis. BO developed the visualization of gene expression along the lineage tree. All authors discussed and interpreted the data. MP, JS and NR wrote the manuscript with input from all other authors. Competing interests: authors declare no competing interests. Data and materials availability: The sequencing data generated is available in GEO under the accession GSE103633. The single cell data generated can be interactively accessed at https://shiny.mdc-berlin.de/psca. Manuals to run and reproduce velocyto and PAGA are available on Github https://github.com/rajewsky-lab/planarian_lineages.

Supplementary Materials:

Materials and Methods

Supplementary Notes 1-5

Figures S1-S22

Tables S1-S10

References (*54-74*)

Fig. 1. Cell type atlas by single-cell transcriptomics

A. Experimental workflow. **B.** tSNE representation of the single-cell transcriptomics data with clusters colored according to the expression of previously published marker genes as follows: grey, neoblasts; orange, neuronal lineage; red, muscle; purple, secretory; blue, epidermal lineage; pink, protonephridia; green, gut; magenta, parenchymal lineage. **C.** Proportions of cell types identified by Baguñà and Romero by microscopy (left) and as identified by tallying up our annotated Drop-seq clusters (right). The outer ring shows the proportion of each individual cluster, which includes neoblasts, epidermal (epidermal and rhabdite), parenchymal (fixed parenchymal), pigment, neuronal (nerve), muscular, gut (gastrodermal and goblet), secretory (acidophilic and basophilic) and protonephridia (flame) cells. We did not find "striped" cells in our dataset. Overall, we find many subtypes for each of the original cell types. **D-F** tSNE plots (upper panels) showing the expression of marker genes and their expression patterns in adult animals using double *in situ* hybridizations on tissue sections (lower panels). Nuclei in **D** and **F** were stained with *Hoechst* and are shown in blue in the overlay. Scale bars: 100 μm. The color scale for tSNE plots ranges from light grey (no expression) to blue (high expression).

Fig. 2. Neoblast ablation and enrichment experiments show stem and progenitor clusters

A. tSNE plots showing the distribution of the cells of an X1 FACS sorted sample (red) and its whole cell population control (x1 control, magenta), and a *h2b*(*RNAi*) sample with its negative control (*gfp*(*RNAi*), green). X1 cells are enriched in the center of the plot while *h2b*(*RNAi*) cells are enriched in the periphery. **B.** PCA analysis considering the expression level of neoblast marker genes and the log odds ratio of the amount of cells per cluster from *h2b*(*RNAi*) and X1 experiments compared to wt and control samples separates neoblasts (grey), progenitor clusters

(yellow) and differentiated cell clusters (blue). The location of these clusters is shown on the tSNE plot on the right. **C.** Gene expression correlation between bulk RNA-seq data from FACS sorted X1, X2, Xins populations and whole worms and the pooled clusters as defined in **B**. Neoblasts show a stronger correlation with X1, progenitors with X2, and differentiated cells with Xins and whole worms.

Fig. 3. Lineage tree reconstruction by PAGA and velocyto

A. Abstracted graph showing all the possible edges with a probability higher than 10e-6 connecting two clusters and their confidence. Each node corresponds to each of the clusters identified using Seurat. The size of nodes is proportional to the amount of cells in the cluster. The most probable path connecting the clusters is plotted on top with thicker edges. **B.** Lineage tree colored according to potency score, which ranges from blue (0) to yellow (1). **C, D.** Lineage trees colored according to the % of X1 **(C)** or *h2b*(*RNAi*) resistant **(D)** cells in each cluster. X1 cells are most abundant in the neoblast 1 cluster whereas *h2b*(*RNAi*) resistant cells are mostly located in the leaves of the tree. **E.** Velocyto force field showing the average differentiation trajectories (velocity) for cells located in different parts of the tSNE plot. **F, G.** Root **(F)** and terminal end-points **(G)** obtained after modeling the transition probabilities derived from the RNA velocity using a Markov Process. The color scale represents the density of the end points of the Markov Process and ranges from yellow (low) to blue (high).

Fig. 4. Consolidated lineage tree of planarian stem cell differentiation into all major types

A. Consolidated lineage tree including 4 independent sources of evidence. The topology of the tree is shown according to PAGA, marker-based connections are shown with red edges.

Velocyto supported connections are shown with thick edges. Progenitor and differentiated cell clusters according to neoblasts ablation and enrichment experiments are shown with yellow and blue halos, respectively. **B.** Gene expression changes of marker genes for the individual stages during epidermal differentiation (in pseudotime). Relative expression of maker genes from neoblast (*Smedwi-1*), early (*prog-1*) and late (*agat-1)* progenitors as well as from the epidermis (*vim-1*). A maximum of 1000 cells from neoblast 1, epidermal neoblasts (en), early epidermal progenitors (eep), late epidermal progenitors 1 (lep 1) and 2 (lep 2) and epidermis were sampled. Grey thin dashed lines show the expression of *Smedwi-1* after randomly permuting cells (rand 1) or after randomly sorting cells within each cluster (rand 2).

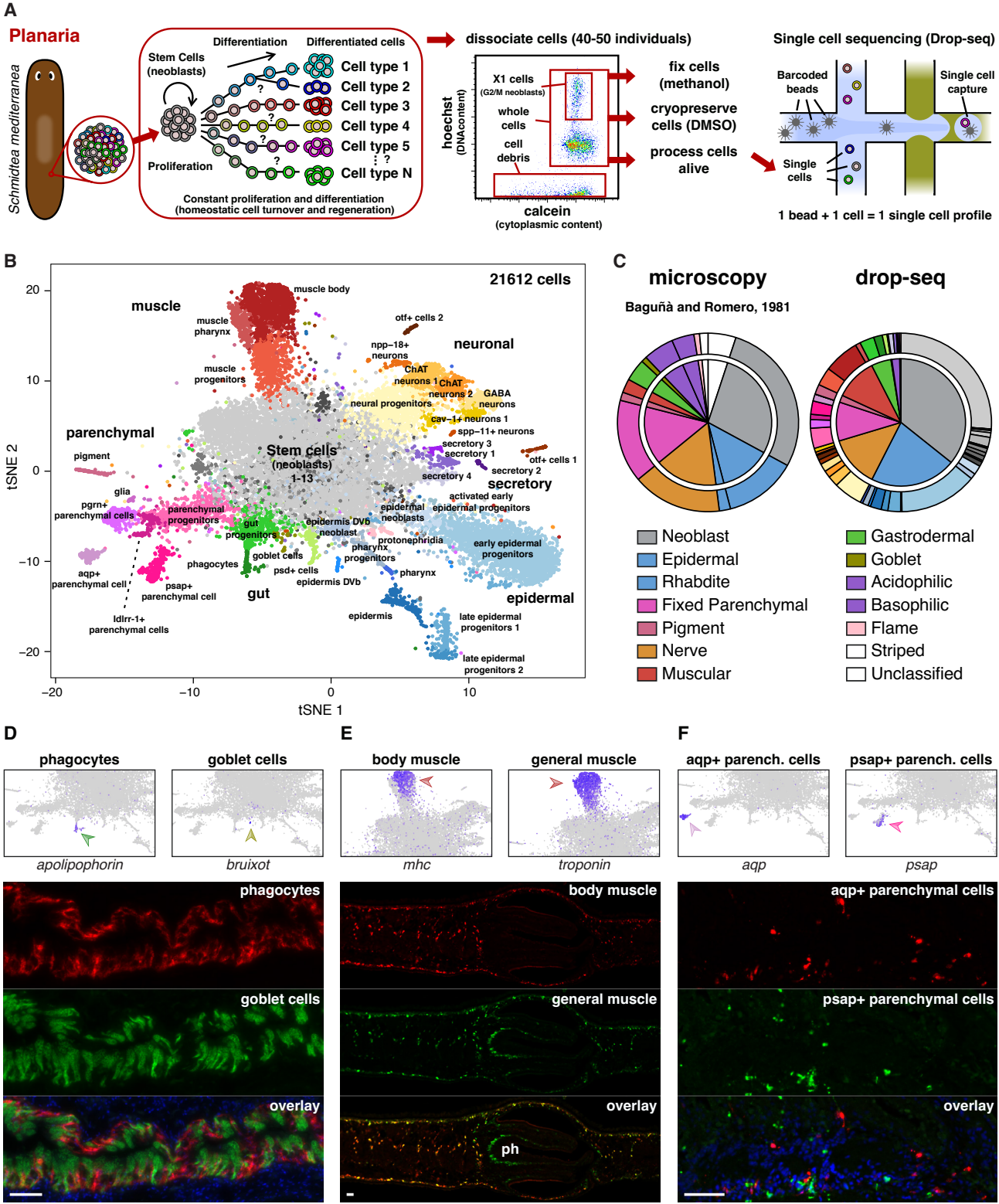
Fig. 5. Identification of gene sets regulated and corregulated in cell differentiation

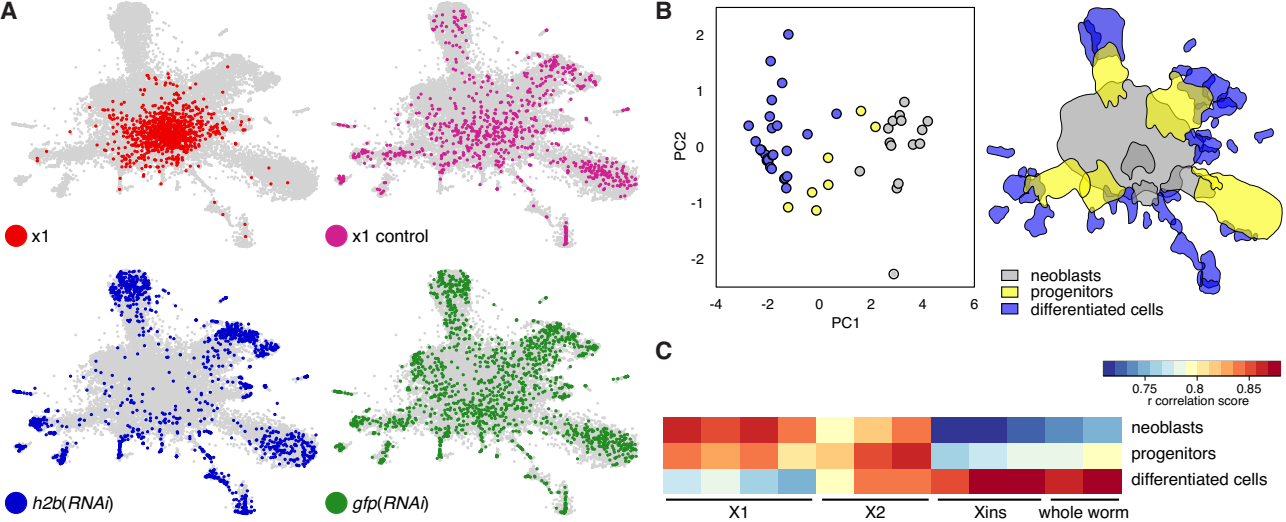
A. Schematic workflow of the analysis performed to identify gene sets involved in lineage decisions. Pseudotemporal ordering of the cells from all lineages and clustering of variable genes using SOMs allowed the identification of 48 gene sets. **B.** Graphical representation of gene expression changes during cell differentiation of 12 gene sets. For each gene set, the normalized expression of the genes is shown on the edges of the tree and ranges from blue (low expression) to red (high expression). Next to each tree, representative genes from each gene set are highlighted.

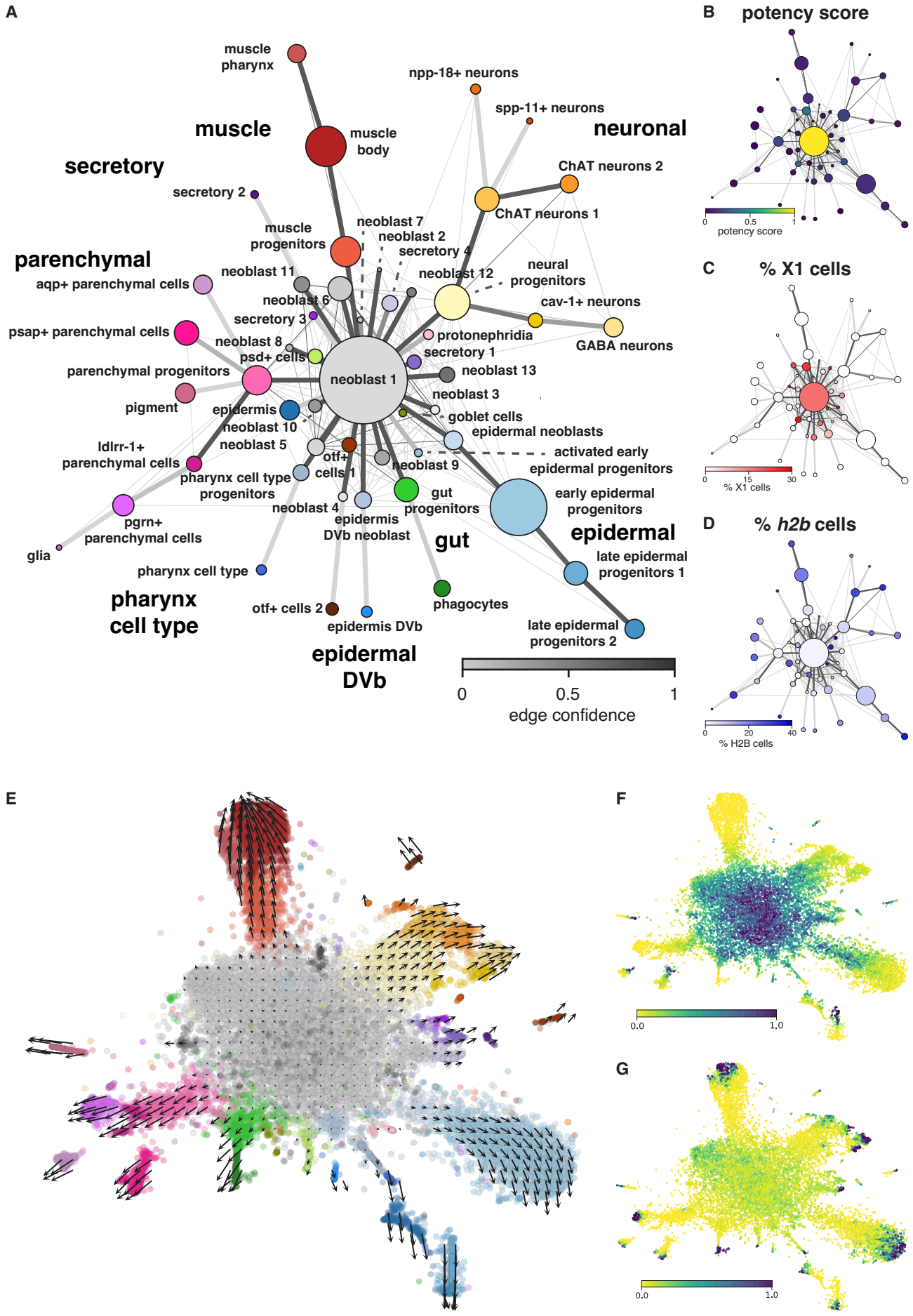
Fig. 6. Molecular profiling of regeneration by single-cell transcriptomics

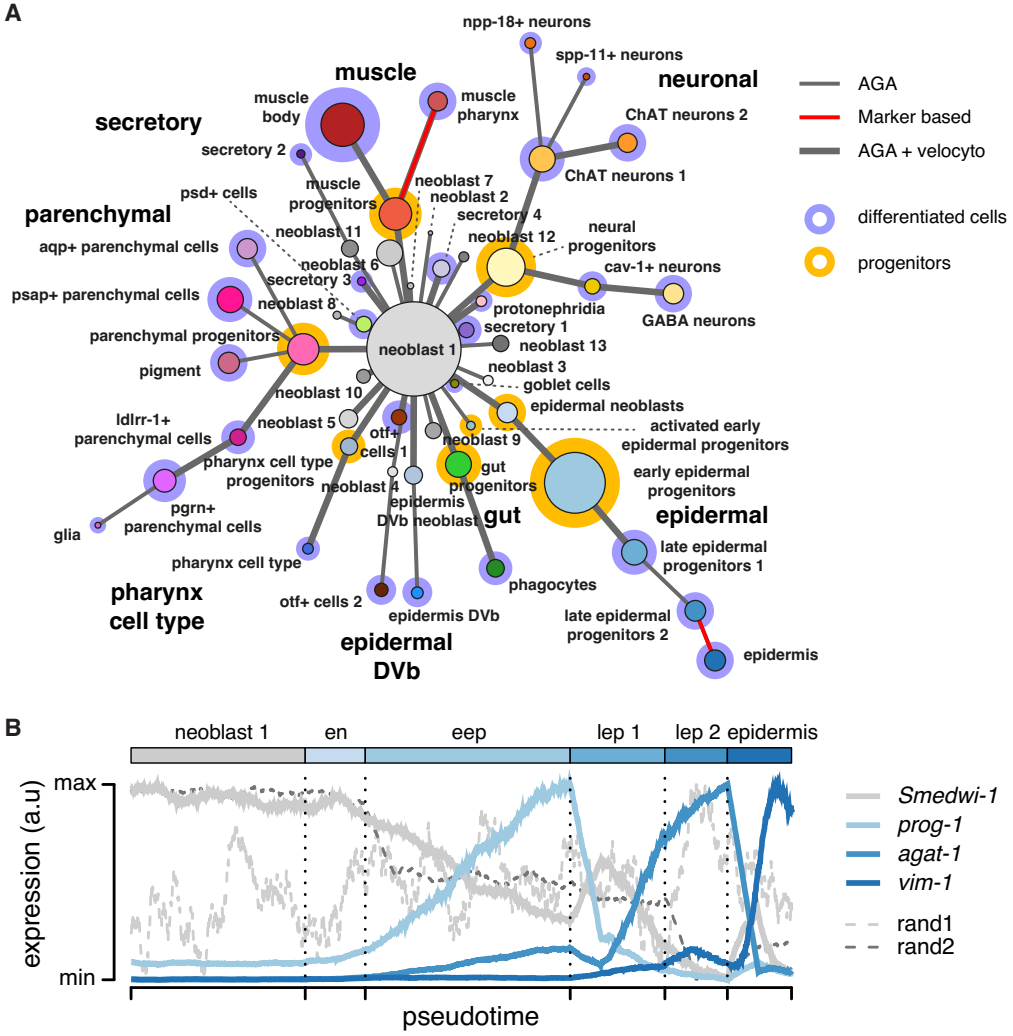
A. Experimental workflow: planarians were cut into small pieces, head pieces were discarded and the remaining pieces were processed for single-cell RNA sequencing 0, 2 and 4 days after cut. **B**. Quantification of neoblasts, neural progenitors and differentiated clusters and

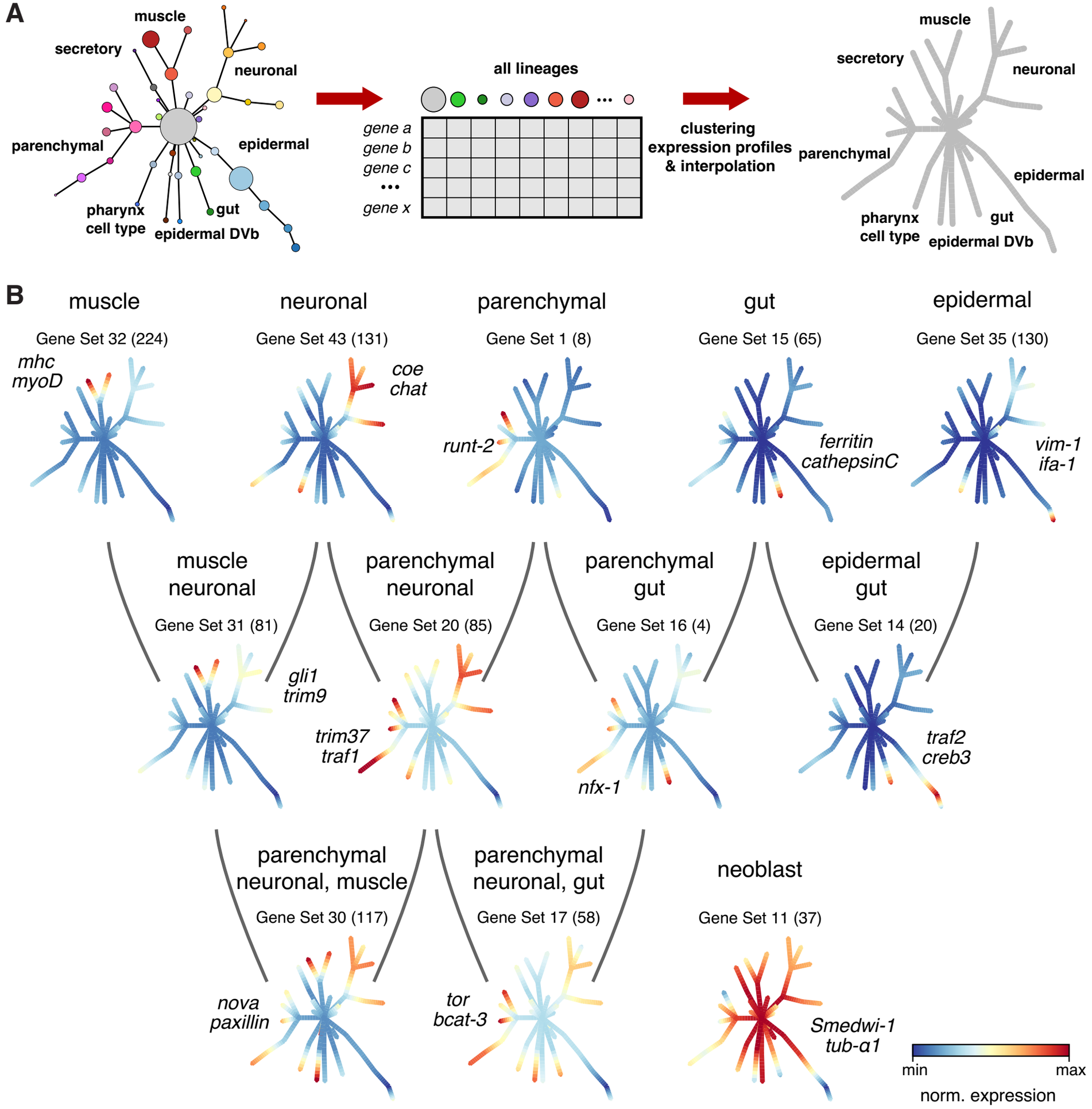
parenchymal progenitors and differentiated clusters. Significant differences calculated using a fisher test with an adjusted p-value < 0.001 are marked with **. **C.** Cluster outlines colored according to the log2 (odds ratio) of changes in regeneration at day 2 (left) and day 4 (right) vs day 0, showing enriched clusters in green colors and depleted clusters in magenta colors. Significant changes are indicated by black solid outlines. **D, E.** *In situ* hybridization on sections **(D)** and quantification (**E**) of aqp+ parenchymal cells in regenerating planarians after 0 and 4 days of regeneration. Mann Whitney U-test p-value < 10e-7. Scale bar: 100 μm.

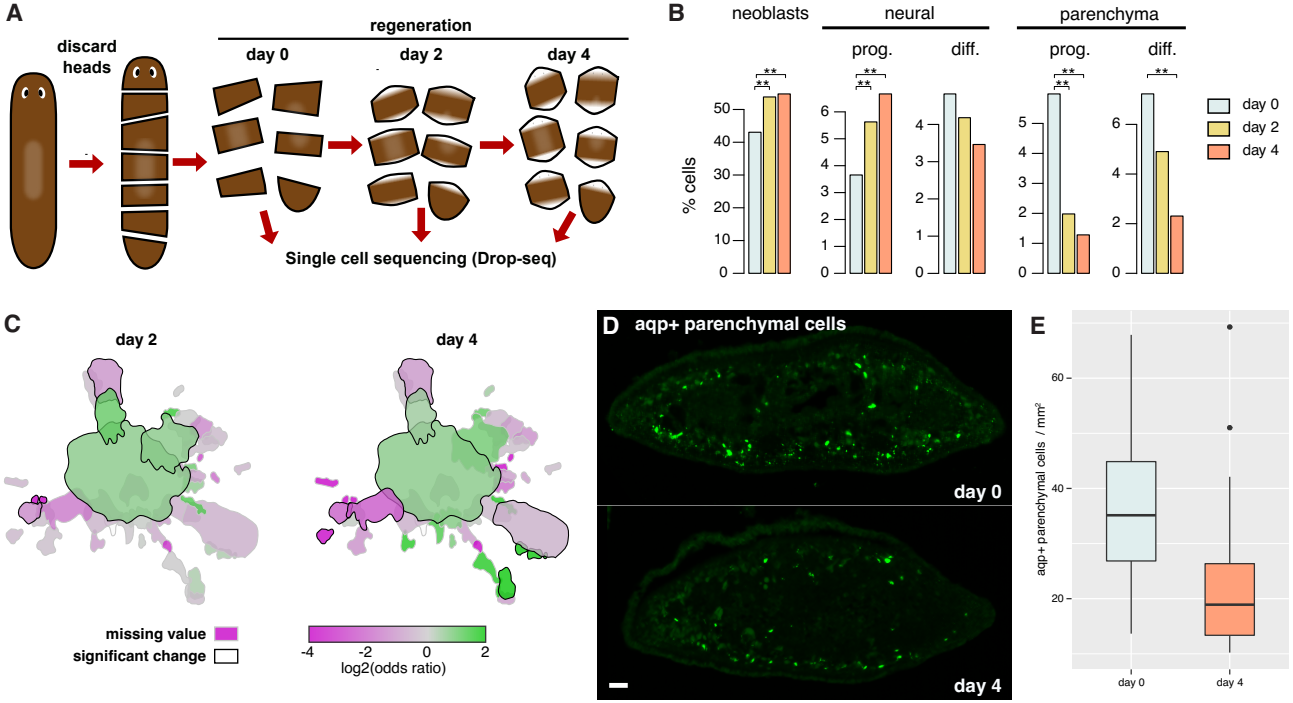

A
Planaria
Schmidtea mediterranea
Stem Cells (neoblasts)
Differentiation
Differentiated cells
Cell type 1
Cell type 2
Cell type 3
Cell type 4
Cell type 5
Cell type N
Proliferation
Constant proliferation and differentiation (homeostatic cell turnover and regeneration)
dissociate cells (40-50 individuals)
hoechst (DNAcontent)
X1 cells (G2/M neoblasts)
whole cells
cell debris
calcein (cytoplasmic content)
fix cells (methanol)
cryopreserve cells (DMSO)
process cells alive
Single cell sequencing (Drop-seq)
Barcoded beads
Single cell capture
Single cells
1 bead + 1 cell = 1 single cell profile
B
21612 cells
muscle
muscle body
muscle pharynx
muscle progenitors
otf+ cells 2
npp-18+ neurons
neuronal
ChAT neurons 1
ChAT neurons 2
GABA neurons
neural progenitors
cav-1+ neurons 1
spp-11+ neurons
secretory 3
secretory 1
otf+ cells 1
secretory 2
secretory 4
secretory
activated early epidermal progenitors
epidermal neoblasts
parenchymal
pigment
glia
pgrn+ parenchymal cells
parenchymal progenitors
Stem cells (neoblasts) 1-13
gut progenitors
epidermis DVb neoblast
protonephridia
pharynx progenitors
early epidermal progenitors
phagocytes
goblet cells
psd+ cells
epidermis DVb
pharynx
aqp+ parenchymal cell
psap+ parenchymal cell
gut
epidermal
ldlrr-1+ parenchymal cells
epidermis
late epidermal progenitors 1
late epidermal progenitors 2
tSNE 1
tSNE 2
C
microscopy
drop-seq
Baguñà and Romero, 1981
Neoblast
Epidermal
Rhabdite
Fixed Parenchymal
Pigment
Nerve
Muscular
Gastrodermal
Goblet
Acidophilic
Basophilic
Flame
Striped
Unclassified
D
phagocytes
goblet cells
apolipophorin
bruixot
E
body muscle
general muscle
mhc
troponin
F
aqp+ parench. cells
psap+ parench. cells
aqp
psap
phagocytes
goblet cells
overlay
body muscle
general muscle
overlay
ph
aqp+ parenchymal cells
psap+ parenchymal cells
overlay


A
x1
x1 control
h2b(RNAi)
gfp(RNAi)
B
PC2
PC1
2
1
0
-1
-2
-4
-2
0
2
4
6
neoblasts
progenitors
differentiated cells
C
0.75
0.8
0.85
r correlation score
neoblasts
progenitors
differentiated cells
X1
X2
Xins
whole worm


A
muscle pharynx
muscle body
muscle
secretory
secretory 2
muscle progenitors
neoblast 7
neoblast 2
secretory 4
parenchymal
neoblast 11
aqp+ parenchymal cells
neoblast 6
secretory 3
psap+ parenchymal cells
neoblast 8
psd+ cells
parenchymal progenitors
neoblast 1
pigment
epidermis
neoblast 10
neoblast 5
ldlrr-1+ parenchymal cells
otf+ cells 1
pharynx cell type progenitors
neoblast 4
epidermis DVb neoblast
pgrn+ parenchymal cells
glia
pharynx cell type
pharynx cell type
otf+ cells 2
epidermis DVb
epidermal DVb
gut progenitors
gut
phagocytes
neoblast 12
neural progenitors
neoblast 13
neoblast 3
goblet cells
epidermal neoblasts
activated early epidermal progenitors
early epidermal progenitors
epidermal
late epidermal progenitors 1
late epidermal progenitors 2
npp-18+ neurons
spp-11+ neurons
neuronal
ChAT neurons 2
ChAT neurons 1
cav-1+ neurons
GABA neurons
protonephridia
secretory 1
neoblast 9
0
0.5
1
edge confidence
B
potency score
0
0.5
1
potency score
C
% X1 cells
0
15
30
% X1 cells
D
% h2b cells
0
20
40
% H2B cells
E
F
0.0
1.0
G
0.0
1.0


A
npp-18+ neurons
spp-11+ neurons
muscle
neuronal
secretory
muscle body
muscle pharynx
ChAT neurons 2
secretory 2
ChAT neurons 1
muscle progenitors
neoblast 7
neoblast 2
psd+ cells
parenchymal
secretory 4
neoblast 11
neoblast 12
neural progenitors
aqp+ parenchymal cells
neoblast 6
cav-1+ neurons
secretory 3
psap+ parenchymal cells
neoblast 8
protonephridia
GABA neurons
secretory 1
parenchymal progenitors
neoblast 1
neoblast 13
pigment
neoblast 3
goblet cells
neoblast 10
epidermal neoblasts
ldlrr-1+ parenchymal cells
neoblast 5
activated early epidermal progenitors
otf+ cells 1
neoblast 9
pharynx cell type progenitors
gut progenitors
early epidermal progenitors
neoblast 4
pgrn+ parenchymal cells
epidermis
DVb neoblast
gut
epidermal
glia
late epidermal progenitors 1
pharynx cell type
phagocytes
pharynx cell type
otf+ cells 2
epidermis DVb
late epidermal progenitors 2
epidermis
epidermal DVb
AGA
Marker based
AGA + velocyto
differentiated cells
progenitors
B
neoblast 1
en
eep
lep 1
lep 2
epidermis
max
min
expression (a.u)
pseudotime
Smedwi-1
prog-1
agat-1
vim-1
rand1
rand2


A
muscle
secretory
neuronal
parenchymal
epidermal
pharynx cell type
gut
epidermal DVb
all lineages
gene a
gene b
gene c
gene x
clustering expression profiles & interpolation
B
muscle
Gene Set 32 (224)
mhc myoD
neuronal
Gene Set 43 (131)
coe chat
parenchymal
Gene Set 1 (8)
runt-2
gut
Gene Set 15 (65)
ferritin cathepsinC
epidermal
Gene Set 35 (130)
vim-1 ifa-1
muscle neuronal
Gene Set 31 (81)
gli1 trim9
parenchymal neuronal
Gene Set 20 (85)
trim37 traf1
parenchymal gut
Gene Set 16 (4)
nfx-1
epidermal gut
Gene Set 14 (20)
traf2 creb3
parenchymal neuronal, muscle
Gene Set 30 (117)
nova paxillin
parenchymal neuronal, gut
Gene Set 17 (58)
tor bcat-3
neoblast
Gene Set 11 (37)
Smedwi-1 tub-α1
min
max
norm. expression


A
discard heads
regeneration
day 0
day 2
day 4
Single cell sequencing (Drop-seq)
B
neoblasts
neural
parenchyma
prog.
diff.
prog.
diff.
% cells
day 0
day 2
day 4
C
day 2
day 4
missing value
significant change
-4
-2
0
2
log2(odds ratio)
D
aqp+ parenchymal cells
day 0
day 4
E
aqp+ parenchymal cells / mm²
day 0
day 4

Supplementary Materials for

Cell Type Atlas and Lineage Tree Reconstruction of Whole Adult Animals by Single-Cell Transcriptomics

Mireya Plass, Jordi Solana, F. Alexander Wolf, Salah Ayoub, Aristotelis Misios, Petar Glažar, Benedikt Obermayer, Fabian J. Theis, Christine Kocks, and Nikolaus Rajewsky

correspondence to: rajewsky@mdc-berlin.de

This file includes:

Materials and Methods

Supplementary Text

Supplemental Notes 1-5

Figs. S1 to S22

Other Supplementary Materials for this manuscript includes the following:

Table S1-S10 as xlsx files

Materials and Methods:

Animals and treatments

All animals belonged to the Berlin-1 strain of asexual type *Schmidtea mediterranea* and ranged 4-10 mm in size. For RNAi experiments, dsRNAs were synthesized as previously described (*41*, *55*). Animals were injected with dsRNA against the coding region of *h2b* or *gfp* for three consecutive days, kept at 20°C, and their cells prepared for FACS and single cell transcriptomics 5 days after the third injection. dsRNAs were delivered at a concentration of 1 μg/μl. For regeneration experiments, animals ranging 4-10 mm in size were cut in 5-7 pieces, the head pieces were discarded and the remaining pieces were processed for Drop-seq immediately, 2 or 4 days after cut.

FACS experiments

FACS experiments were performed as previously described (*23*, *24*). Essentially, planarians were cut into little pieces on ice and in the presence of trypsin to help cell dissociation. Cells were then sequentially filtered through 40 μm and 20 μm filters and stained with the cytoplasmic dye Calcein-AM (BD Biosciences, at a final concentration of 0.5 μg/ml) and the nuclear dye Hoechst 33,342 (Fluka Biochemika, at a final concentration of 20 μg/ml). Propidium iodide was used to discard dead cells. Cells were then sorted with a BD FACSAria III. Two wild type and RNAi samples were sorted directly into ice-cold methanol 80% in PBS (*28*). The final percentage of methanol was adjusted to 70%, after being diluted with the incoming FACS sorted cells. An additional wild type sample was sorted in a 10% DMSO in PBS solution and frozen directly after FACS collection. Furthermore, an additional wild type sample, as well as the X1 population together with its whole cell population wild type control, were sorted into PBS and

directly processed for Drop-seq. Whole cell population gating was achieved by sorting the Hoechst positive, Calcein positive, PI negative population of cells. X1 gating is achieved by sorting the population of cells with double content of DNA.

Drop-seq procedure, single cell library generation and sequencing

Monodisperse droplets of about 1 nl in size were generated using microfluidic PDMS devices (Drop-SEQ chips, FlowJEM, Toronto, Canada; pre-coated with Aquapel). Barcoded microparticles (Barcoded Beads SeqB; ChemGenes Corp., Wilmington, MA, USA) were prepared and flowed in using a self-built Drop-seq set up (*3*) (Online-Dropseq-Protocol-v.3.1 http://mccarrolllab.com/dropseq/). Cell preparations were kept on ice and handled in the cold throughout. Live and cryopreserved (*29*) thawed cells were centrifuged at 300 x g for 3 minutes. Cells fixed in methanol (*28*) were centrifuged at 3000 x g for 5 minutes and resuspended in the presence of 1 unit/µl RiboLock RNAse inhibitor (Thermo Scientific). Cell pellets were resuspended in 1 ml PBS + 0.01% BSA, centrifuged again, resuspended in 0.5 ml in PBS + 0.01% BSA, passed through a 35 µm cell strainer, counted with a Nanoentek Eve Automatic cell counter (with default settings; min. size 5 µm), and diluted to 1.5 ml. Table S1 lists the final cell concentrations. Droplets were collected in 50 ml Falcon tubes for ~13 minutes, corresponding to ~1 ml of combined aqueous flow volume (1 ml cells and 1 ml of beads). Droplets were broken promptly after collection and barcoded beads with captured transcriptomes were reverse transcribed and exonuclease-treated. First strand cDNA was amplified by equally distributing beads from one run to 24 or 48 PCR reactions (50 µl volume; 4 + 9 cycles). 20 or 10 µl fractions of each PCR reaction were pooled (total = 480 µl), then double-purified with 0.6x volumes of Agencourt AMPure XP beads (Beckman Coulter, A63881) and eluted in 12 µl. 1 µl of the amplified cDNA libraries were quantified on a BioAnalyzer High Sensitivity Chip (Agilent). If

necessary, more cDNA was purified from the PCR reactions. 600 pg cDNA library were fragmented and amplified (12 cycles) for sequencing with the Nextera XT v2 DNA sample preparation kit (Illumina) using custom primers enabling 3'-targeted amplification as described. The libraries were double-purified with AMPure XP Beads (0.6x, 1x), quantified and sequenced (paired end) on Illumina Nextseq500 sequencers (library concentration 1.8 pM; Nextseq 500/550 High Output v2 kit (75 cycles) in paired-end mode; read1 = 20 bp using the custom primer Read1CustSeqB (*3*), read 2 = 64 bp).

scRNA processing and quantification of gene expression using Digital Expression Matrices

Raw scRNA-seq data was preprocessed to create a Digital Gene Expression Matrix (DGE) using Dropseq tools v 1.12. Briefly, paired-end reads, one containing the cellular and the molecular barcodes and the other containing an RNA fragment, were joined in a bam file using picard-tools v 1.95 (http://broadinstitute.github.io/picard) and sorted using samtools 1.3. Reads were then tagged with the cell and the molecular barcodes (UMIs), trimmed at the 5'end to remove adapter sequences, and at the 3' end to remove polyA tails using Dropseq tools with default parameters. Next, reads were mapped to the *S. mediterranea* dd_Smed_v6 transcriptome assembly (http://planmine.mpi-cbg.de) (*56*) using STAR v 2.5.1b (*57*). Using a custom perl script, we added the corresponding gene tag to each mapped read in the sam file equivalently to what is done using Dropseq tools when mapping to the genome. Finally, Dropseq tools were used to correct for bead synthesis errors affecting the last position of the cell barcode.

To distinguish real cell transcriptomes captured in the drops from ambient RNA, we plotted the cumulative fraction of uniquely mapped reads assigned to each cellular barcoded for each of the sequenced barcodes ordered decreasingly, and identified the inflection point of the curve using the R package dropbead (https://github.com/rajewsky-lab/dropbead). The final DGE for

each of the samples was obtained using Dropseq tools by counting the number of UMIs assigned to each gene in each of the selected cells using all mapped reads. The summary of all the stats for each sample can be found in Table S1. The raw fastq files as well as the individual DGE matrices for each dataset can be found on GEO with the accession GSE103633.

Data clustering

DGE expression matrices were joined and analyzed using the R package Seurat v1.4 (https://github.com/satijalab/seurat). Only genes expressed in at least 3 cells and cells with a minimum of 200 genes were kept. For each cell UMI counts per gene were normalized to the total UMI count of the cell and log transformed as norm = log(UMI+1). To prevent the inclusion in the analysis of doublets, cells with more than 2500 genes were also discarded (Table S1). A total of 21612 cells from the 11 experiments were included in the final analysis. Next, we used Seurat to regress on the number of UMI counts per cell to remove possible bias coming from the differences in the number of UMIs per cell observed. The clustering of the cells was done using only variable genes, which were those with a mean expression between 0.01 and 3, and a dispersion bigger than 0.4 using the function MeanVarPlot. The variable genes were then used to perform a Principal Component Analysis (PCA). The top 50 PCs obtained were then tested for significance using a JackStraw test with 100 replicates and those with a p-value < 1e-5 were used to perform the clustering. This is a Louvain based clustering on the cell neighborhood knn graph (*58*). Clusters were identified using the function FindClusters from the same package using a resolution of 6. To prevent obtaining spurious clusters result of overclustering, the robustness of the clusters was calculated using the function AssessNodes from Seurat. For each cluster, the average expression of all variable genes (4910) is computed and a phylogenetic tree based on the distance matrix in gene expression space is computed. Next, it computes an Out of Bag error for

a random forest classifier trained on each internal node split of the tree. We recursively build a tree and assessed all its nodes, merging all clusters with an out of bag error bigger than 0.1 until no such nodes were found. This procedure allowed us to identify a final set of 51 clusters with large differences in their cell number, from neoblast 2 (13 cells) or the glia (24 cells), to the neoblast 1 cluster that contains more than 6000 cells.

Identification of cluster specific marker genes

To identify differentially expressed genes in each of the clusters relative to the rest of the cells in the analysis, we used the function FindAllMarkers from Seurat package, using the likelihood-ratio test. Only genes enriched and expressed at least in 25% of the cells in one of the two populations and with a log fold difference bigger that 0.25 were considered. Manual inspection and experimental validation using *in situ* hybridizations of the marker genes reported allowed the identification of the different cell populations as described.

***In situ* hybridization, immunohistochemistry and imaging**

Riboprobes were generated by *in vitro* transcription of PCR products generated as described above, with only one T7 promoter linked to the 3' end of the amplicon, and in the presence of digoxigenin-labeled UTP (Roche). The products of *in vitro* transcription reactions were then treated with DNAse. Riboprobes were then precipitated in ethanol in the presence of LiCl and glycogen and resuspended in 50% formamide in TE buffer, 0.01% Tween. The primers used are listed in Table S9.

Whole mount *in situ* hybridization and *in situ* hybridization on histological sections were performed as previously described (*59*, *55*). Briefly, for whole mount *in situ* hybridization animals were killed in a 2% HCl solution, fixed in Carnoy's solution, bleached in an 8%

H_2O_2/methanol solution and rehydrated. For *in situ* hybridization on histological sections animals were killed in 37% formaldehyde (Roth) for 3-5 min, fixed in 4% paraformaldehyde in PBS for 4 hours, washed in PBS overnight, dehydrated through a series of ethanol solutions and xylene substitute (Sigma) and embedded in paraffin. Specimens were sectioned in 10 µm slices and laid in poly-L-lysine pre-coated slides. Slides were then rehydrated through xylene substitute (Sigma) and a series of ethanol solutions. Animals or slides were then permeabilized with Proteinase K (Sigma), treated with 0.25% and 0.5% acetic anhydride in 0.1M triethanolamine pH 7.6, prehybridized and hybridized with digoxigenin labelled riboprobes (0.2 ng/µl, O/N at 56°C), washed, immunolabeled with anti-digoxigenin-alkaline-phosphatase antibody (Roche) and developed in the presence of NBT and BCIP (Roche). For double fluorescent *in situ* hybridization on sections, probes were labeled with digoxigenin and fluorescein, detected with an anti-digoxigenin-peroxidase antibody and an anti-fluorescein peroxidase antibody respectively (Roche), and the signal was developed with the Tyramide Signal Amplification kit (Perkin Elmer). Whole mount animals were imaged with a Keyence microscope using the full focus function. Sections were also imaged with a Keyence microscope. For aqp+ cell counting, *in situ* hybridization of *aqp* was performed on 6 slides containing sections of regenerating animals of day 0 and 4. Positive cells were automatically counted using a custom script for ImageJ.

Sample correlation

DGE expression matrices were processed independently and analyzed using the R package Seurat v1.4 (https://github.com/satijalab/seurat). Only genes expressed in at least 3 cells and cells expressing a minimum of 200 genes were kept. To prevent the inclusion in the analysis of doublets, cells with more than 2500 genes were also discarded.

We assigned to each gene the log of the sum of UMIs for all the cells included + 1. For each pair of samples, we compared their expression profiles and calculated the Pearson correlation coefficient across samples (Fig. S1B). We used the same approach to correlate the expression profiles of the cells in each sample per cluster (Fig. S1F).

Sample correlation to RNA-seq datasets

We downloaded X1, X2, and Xins FACS sorted samples and whole worm RNA-seq samples from GEO database (Table S10). The data was mapped to planarian transcriptome as described for the Drop-seq samples. All the reads mapped to different transcripts belonging to the same gene were counted once and the total read count was then assigned to the gene. The final expression of each gene was normalized to the number of reads in the dataset and expressed as tags per million mapped reads (TPMs). The Pearson correlation coefficient between the log of TPMs +1 and the log of the normalized UMIs +1 was calculated for each pair of samples (Fig. S1C).

PAGODA analysis

We used PAGODA (*40*) to functionally characterize the cell clusters identified with Seurat. For each gene, we obtained their GO term annotations from Planmine (*56*). All the genes containing the same GO term were joined in a gene set, and those sets containing at least 5 genes were kept to perform the clustering with PAGODA. We used 5% of the cells to fit the knn error model and only valid cells were kept for further analyses. We selected as top aspects those with a z-score < 0.01. Finally, we reduced the number of aspects based on their PCA loadings and on their redundancy (distance threshold = 0.6).

To prevent that the amount of cells in the different clusters would bias the identification of variable pathways in the analysis of all the cells, we sampled the 100 cells with the highest UMI count per cluster (if possible). This amounts to a total of 4395 cells used for the analysis. In the parenchyma and gut analysis, all the cells belonging to these clusters were used (3306 cells).

Lineage Tree Reconstruction

We use the partition-based graph abstraction (PAGA) (*22*) for inferring the lineage tree. The graph abstraction algorithm reconciles clustering and trajectory inference by explaining data variability both in terms of discrete and continuous latent variables. For this, it considers a partitioned graph of neighborhood relations among data points. By quantifying both the distances between nodes using a random-walk based measure and the connectivity of given partitions, it generates a much simpler abstracted graph in which nodes correspond to partitions of the original graph. The differentiation tree arises as the tree-like subgraph in the abstracted graph that best explains the global topology of the original graph. In simple cases, this can be found by fitting a minimum spanning tree into the abstracted graph, weighted with inverse connectivity. For the complex topology of the planaria data though, it is necessary to fit a tree-like subgraph that optimizes continuity when transitioning between edges.

We use a clustering of the data that we inferred using the Louvain algorithm (*58*) in the implementation of Seurat (*3*). The clusters that arise from the Louvain algorithm define, by construction, partitions of the graph of neighborhood relations among data points and we can readily use graph abstraction for inferring an abstracted graph and, within that, a lineage tree (Fig. 3A).

The code is available within the single-cell analysis framework Scanpy (*60*) on Github https://github.com/theislab/scanpy. Results for the present study as well as more detailed

information on the analysis can be found at https://github.com/rajewsky-lab/planarian_lineages. Additional information on the computational method can be found on the Supplementary Note 3.

RNA velocity analysis

We applied velocyto (*44*) to calculate RNA velocity on the data. RNA velocity distinguishes between spliced and unspliced reads in order to estimate the rate of transcriptional change of each cell. It then extrapolates the transcriptional states of the cells in the near future, plotting the RNA dynamics on a given embedding.

In order to extract spliced and unspliced reads, we mapped the reads from all the single cell runs to the planarian genome (SmedAsxl Genome Assembly v1.1) downloaded from SmedGD (http://smedgd.stowers.org/) (*61*). To obtain an annotation, we mapped the transcriptome to this genome version using STARlong (*57*) with the following parameters (--outSAMattributes NH HI NM MD --outFilterMultimapScoreRange 1 --outFilterMismatchNmax 2000 --scoreGapNoncan -20 --scoreGapGCAG -4 --scoreGapATAC -8 --scoreDelOpen -1 --scoreDelBase -1 --scoreInsOpen -1 --scoreInsBase -1 --alignEndsType Local --seedSearchStartLmax 50 --seedPerReadNmax 100000 --seedPerWindowNmax 1000 --alignTranscriptsPerReadNmax 100000 --alignTranscriptsPerWindowNmax 10000). After mapping, we discarded all transcripts that were multiple mapped and kept only genes for which all its transcripts were mapped to the same contig for which we generated a gff annotation file using custom perl scripts.

We ran the Command Line Interface (CLI) of velocyto with permissive logic settings. In this analysis, we used the mapping to the transcriptome only for those cells present in the final Seurat analysis, as described above. We further discarded cells that either did not have enough UMIs after the new mapping, or they did not have unspliced reads. Using the remaining 19140 cells, we selected a subset of genes based on their coefficient of variation (CV), average

expression, and a minimum expression of spliced and unspliced reads in total. With these criteria, we selected 2795 genes that were used to perform a PCA. Using the 75 first principal components, the data was imputed/smoothened (using k=200 nearest neighbors) and normalized. We used the standard python implementation of velocyto with default parameters for fitting gene models, predicting velocity, extrapolating and plotting (a jupyter notebook, with all the analysis in detail, is available in the project's github repository: https://github.com/rajewsky-lab/planarian_lineages) (Fig. 3E; Fig. S12). For these plots we used the tSNE embedding as produced by the Seurat analysis. Finally, we estimated the differentiation start and end points (Fig. 3F,G), using the functions prepare_markov and run_markov included in velocyto. These use backward and forward markov processes on the transition probability matrix of the cells to determine high density regions for the start and end points of trajectories, respectively.

Gene expression clustering

We used the R package som (*62*) to cluster gene expression profiles of cells sorted according to pseudo time. First, we used MAGIC (*63*) to impute the single cell UMI counts per gene in the whole dataset in order to better recover the gene-gene correlations underlying in the data. For the imputation, we used 100 PC and t = 10. Next, the expression values for each gene were normalized.

For the lineage specific SOMs (Fig. S8A), the clustering analysis was restricted to a subset of variable genes expressed in at least 15 cells from the clusters of interest. We selected variable genes using the function MeanVarPlot from the Seurat package (parameters x.low.cutoff = 0.3, y.cutoff = 0.1). In the global SOMs (Fig. 6B and Fig. S13), we selected variable genes (expressed in at least 15 cells) belonging to the neoblast 1 cluster or any of the clusters included

in lineage trajectories finishing in differentiated cell type clusters (Fig. 4A). To select variable genes, we also used the function MeanVarPlot (parameters x.low.cutoff = 0.3, y.cutoff = 0.01).

To create the SOMs, we generated a matrix in which the cells from each of the clusters were added consecutively and within each cluster sorted according to pseudotime. We used this matrix as input for the function *som* (xdim = 1, ydim = 20 or 50 for lineage specific and global SOMs respectively). In the case of the global clustering analysis, the clustering provided 48 gene sets. The genes belonging to each set are provided in Table S4.

For each cluster, we calculated the average gene expression of each gene set. Changes in expression of gene sets on connections between clusters were inferred by linear interpolation. We used the python packages networkx and matplotlib to draw the lineage trees and associated information.

Differential gene expression analysis

To identify differentially expressed genes in the regeneration samples to the control we used the function DiffExpTest from the Seurat Package, which implements a likelihood-ratio test (LTR) method for the identification of differentially expressed genes in single cell data (*64*). We calculated differentially expressed genes for each cluster independently, comparing day 4 and day 2 samples to regeneration day 0 sample. Additionally, we pooled all the cells from each sample and calculated differentially expressed genes for the whole sample. We finally kept as differentially expressed genes those having an adjusted p-value < 0.05 and a $|\log2FC| > 0.58$ (Table S6, S7 and S8).

Supplementary Note 1: Cell type characterization and nomenclature

To characterize the cell types included in each of the clusters we first correlated their gene expression patterns with those of previously published single cell sequencing experiments. First, we scored each of the predicted clusters using the list of marker genes from Wurtzel *et al*. (*42*). For each cluster, we defined the score of a gene set as the log of the average number of normalized UMIs of the set in a given cluster. The distribution of all the scores for all clusters using the marker gene sets is shown in Fig. S2A. We used these scores to classify the clusters in major groups, assigning each cluster the identity of the cell population with the highest score. This comparison readily showed that our tSNE representation is roughly distributed in 7 territories that showed markers of neoblasts (clusters 1-15), neural (clusters 16-24), muscle (clusters 25-27), parapharyngeal (clusters 28-31), epidermal (clusters 32-39), protonephridia (cluster 40) and gut (clusters 41-51) described previously (*42*) (Fig. 1B, S2B). Interestingly, neoblast clusters were at the center of the tSNE plot, and expressed well-known neoblast marker such as *Smedwi-1* (dd_Smed_v6_659_0) (Fig. S3).

Epidermal clusters expressed markers of different epidermal populations, including progenitors (clusters 32-36) indicating that other clusters connected to the central neoblast cluster (Fig. S2B) and that expressed neoblast markers to a lower extent (Fig. S2A, top row) might correspond to progenitors as well. Thus, our single cell sequencing experiment contains not only neoblasts and differentiated cells but also their progenitors.

To further characterize differentiated cell types, we compared the groups highlighted in Fig. S2B to the cellular proportions found by Baguñà and Romero in their 1981 description of major planarian cell types (*30*). We used as a reference their quantification of 7 mm planarians, roughly comparable to the ones used in our experiments. We took several assumptions:

a) Baguñà and Romero did not distinguish mature from progenitor states and only left a 5% of cells “unclassified”, in which they counted “cells that are difficult to assign to a specific cell type”, “mainly cells in the process of differentiation”. However, only our early epidermal progenitors accounted for roughly 12-15% of our total cells, exceeding the 5% of unclassified cells. Thus, it is possible that these early epidermal progenitors already look like epidermal cells morphologically and that Baguñà and Romero counted them in their “epidermal” class. Similarly, other progenitors might resemble their differentiated state morphologically and might have been counted in their respective classes.

b) “Rhabdite” cells were “presumed to be the precursors of epidermal cells”, so we assumed they represent some of our epidermal clusters.

c) “Acidophilic” and “Basophilic” secretory cells might correspond to our “secretory clusters”.

d) We did not find an equivalent to their “Striped” class.

e) We initially did not detect an equivalent for the “fixed parenchymal cells” at first. These are an abundant cell type (15% of planarian cells) located in the parenchyma of planarians -the tissue between the body wall and the gut of the animal (Fig. S2A). However, we observed a group of clusters (Fig. 1B, magenta, clusters 41, 43-47 and 49) that expressed markers from cells located in the parenchyma discovered in another study (*39*) (Fig. S4). Their abundance in our dataset (11.7%) was consistent with the literature, indicating that they could represent the “fixed parenchymal cells” described by microscopy (Fig. 1C).

To validate and elucidate the identity of our cell clusters we performed *in situ* hybridization experiments (Fig. 1D; Fig. S5). These revealed that markers of clusters 43 and 46 are also expressed in parenchymal cells (Fig. S5), indicating that these new cell types are the molecular

equivalent to the morphologically described fixed parenchymal cells (*35*). However, we find several classes of parenchymal cells that are non-overlapping (Fig. 1F). To name parenchymal cell types we looked at their marker genes. Cluster 41 expressed the known marker *ldlrr-1* (dd_Smed_v6_1581_0) (Fig. S4) (*39*), cluster 43 expressed the marker *aqp* (dd_Smed_v6_1103_0) (Fig. 1F, S5), cluster 45 expressed the marker *pgrn* (dd_Smed_v6_67_0) (Fig. S5) and cluster 46 expressed the marker *psap* (dd_Smed_v6_663_0) (Fig. S5). We also found glial cell markers such as *estrella* (dd_Smed_v6_1792_0) , *if-1* (dd_Smed_v6_12254_0) and *cali* (dd_Smed_v6_9961_0) (Fig. S4) (cluster 47) (*38*, *39*) and pigment cell markers such as *pbgd-1* (dd_Smed_v6_626_0) and *kmo-1* (dd_Smed_v6_6831_0) (cluster 44) (*36*, *37*).

Our *in situ* experiments showed that markers of both clusters 42 and 48 are expressed in the gut (Fig. S5) and are non-overlapping, showing that they are different cell types (Fig. 1D). Cluster 42 expressed the marker *apolipophorin* (dd_Smed_v6_194_0) (Fig. 1D, S5) and cluster 48 expressed a novel gene that we called *bruixot* (dd_Smed_v6_131_0), a novel marker of goblet cells. Clusters 42 and 48, together with cluster 51 had a combined abundance of 5.2% of wild type cells, much more comparable to the 4% and 1% that Baguñà and Romero had found for phagocytes and goblet cells.

We could not confirm the identity of cluster 50, which is also related to the gut or the parenchymal cells by marker expression, and we named it after the marker *psd* (dd_Smed_v6_9018_0) (Fig. S4).

We also confirmed that cluster 37 contained cells from a previously unknown pharynx cell type (Fig. S5) related to the epidermis (Fig. S2). Cluster 39 contains a special type of epidermal cells located at the dorso-ventral boundary (epidermis DVb) (Fig. S5), already described by

Wurtzel and coworkers (*34*). Cluster 34 contains a subtype of epidermal progenitors that are more frequent in regeneration samples and that we called activated early epidermal progenitors.

We also confirmed that clusters 25-27 contained different types of muscle cells. The general marker *troponin* (dd_Smed_v6_323_0) and the body muscle marker *mhc* (dd_Smed_v6_432_0) (cluster 25) colocalized in all muscle except for the pharynx (Fig. 1E). Cluster 26 had the known pharynx marker *laminin* (dd_Smed_v6_8356_0) and therefore represents the pharynx muscle.

We also confirmed that markers of our neuronal clusters were expressed in the brain and CNS (Fig. S5). We examined the markers of these clusters to name these cell types and compared them to the literature (*49*, *65*). Clusters 19 and 24 expressed the markers *chat* (dd_Smed_v6_6208_0) and *chat2* (dd_Smed_v6_11968_0), cluster 23 expressed the GABA receptor *gbrb-1* (dd_Smed_v6_19336_0), cluster 22 expressed the marker *cav-1* (dd_Smed_v6_8555_0), cluster 16 expressed the marker *spp-11* (dd_Smed_v6_1724_0) and cluster 17 expressed the marker *npp-18* (dd_Smed_v6_1117_0). The neuronal related clusters 18 and 21 (Fig. S2) expressed the marker *otf* (dd_Smed_v6_5915_0) (Fig S4).

Supplementary Note 2: Functional characterization of cell clusters using PAGODA

We used PAGODA (*40*) to characterize the function of the different cell clusters identified. This method is designed to overcome the noise in single cell data and identify gene sets whose expression is coordinated in specific groups of cells. In this analysis, we used as gene sets groups of genes sharing the same GO term. When analyzing all clusters, we observe that the clustering of the sampled cells using GO term based gene sets recapitulates the graph-based clustering of cells done with Seurat. However, PAGODA is not able to separate neoblasts and progenitor clusters (Fig. S7A). These clusters are mostly characterized by a higher expression of genes involved in *RNA processing*. Additionally, this analysis also confirms the functional identity of some of the clusters, such as muscle cells, which are enriched in genes related to *contractile fiber* and *collagen trimer*. In this joint analysis we also find similarities between gut and parenchymal clusters, which are grouped together. The clusters from these cell types are closely related and share several GO terms such as *lysosome*, *extracellular region*, *lipid catabolic process* and *metallopeptidase activity* (Fig. S7A). Consistent with these results, Baguñà and Romero already noted a high incidence of vacuoles and lipid droplets in gut cells as well as in the fixed parenchyma cells (*30*).

To investigate further the differences between these related lineages, we performed an additional PAGODA analysis including only gut and parenchymal lineages (Fig. S7B). In this analysis, we found that cells from progenitor clusters are quite similar and appear together in the clustering. These cells show high expression of genes annotated with *mRNA metabolic function*, *mitotic cell cycle*, and *transcription*, suggesting that they may still share undifferentiated traits with neoblasts. The clustering of the rest of the cells using the GO term defined gene sets recapitulates the groups obtained using Seurat. As seen in the previous analyses, several of the

terms are shared by gut and parenchymal clusters, such as *vacuole*, *cytoplasmic vesicle* as well as several other term related with enzymatic activity. We also see a clear enrichment of terms involving transport and enzymatic activity in phagocytes, such as *symporter activity*, *secondary active transmembrane transporter activity*, *oligosaccharide metabolic process* and *serine type endopeptidase activity*. In contrast, we find clear enrichment of other GO terms like to *protoporphyrinogen IX biosynthetic process* and *metallopeptidase activity* in parenchymal cells (Fig. S7B). Top markers of the parenchymal clusters include lysosomal enzymes (dd_Smed_v6_642_0) and amino acid synthetases (dd_Smed_v6_646_0) among others. These results suggest that the parenchymal lineage shares metabolic functions with the gut and that it may be responsible for some biosynthetic processes as well as containing specialized functions such as those from glial and pigment cells.

Supplementary Note 3: Fundamentals and robustness of PAGA

As Planarians regenerative capabilities suggest that a comprehensive snapshot of the cellular spectrum of an adult planaria is likely to contain many transitional states between different cell types, planarians offer a unique system for studying the statistical inference of its cellular lineage relations. In this supplementary note, we show why previous methods are not able to accomplish this task while our approach yields a robust inference of an unprecedented rich structure of topological relations among cell types. In the main text, we validate that the large majority of these topological relations corresponds to actual biological lineage relations. Finally, we inspect more closely inspect the epidermal lineage. To reproduce the figures of these supplemental notes, go to https://github.com/rajewsky-lab/planarian_lineages.

Manifold-learning-based data embeddings cannot represent complex topological properties of high-dimensional data

Low-dimensional non-linear embeddings can only provide a topologically faithful visualization of a dataset if its topology is sufficiently simple. In many data science problems, one is merely interested in distinguishing discrete, clearly separated categories into which observations may be classified. This occurs for single-cell datasets that only contain clearly separated cell types. Topologically, this corresponds to a simple collection of point-like objects, for which the tSNE algorithm provides faithful visualizations, good abstractions, in which distance locally has a qualitative interpretation in terms of biological similarity. However, the rich continuous structure of the planarian single-cell data, with its mostly connected components, cannot be captured by low-dimensional embeddings, which are insufficient, as seen by the comparison of the tSNE (*54*) and the force-directed drawing of the single-cell graph using the

Fruchterman-Reingold algorithm (*66*). (Fig. S17). This observation has been previously noted upon studying examples with much simpler topology (*67*).

Conceptual differences between graph abstraction and previous methods

To infer topological relations, we study the connectivity patterns of a partitioned neighborhood graph of single cells; an approach that some of us have termed partition-based graph abstraction (PAGA). All previous methods try to infer lineage relations based on estimates of average distances between clusters in a usually dimensionally-reduced high dimensional space. However, the assumption that justifies this strategy is fundamentally wrong: any unsupervised distance measure can only reflect biological similarity in the limit of very small distances, that is, on a local scale. This can be understood as follows: let us quantify biological similarity $s_{\mathrm{bio}} \in [0,1]$ of cells using a function f of a fixed, unsupervised distance measure $\mathrm{d}_{\mathcal{X}}$ in a high-dimensional space $\mathcal{X}$ of the measured variables, for instance, euclidian or correlation distance,

$$s_{\mathrm{bio}}(i,j) = f(d_{\mathcal{X}}(x_i - x_j)), \quad (1)$$

where i and j label measured cells and $\mathrm{x_i}$ and $\mathrm{x_j}$ their molecular profile; for instance, the vector of gene expression measurements. Clearly, the function f is unknown. However, we know that in the limit of $\mathrm{d}_{\mathcal{X}} \to 0$, it holds

$$\mathrm{d}_{\mathcal{X}}(\mathrm{x_i} - \mathrm{x_j}) \to 0 \quad \Rightarrow \quad \mathrm{s_{bio}}(\mathrm{i},\mathrm{j}) \to 1, \quad (2)$$

if not with certainty, at least with high probability. Irrespective of the precise definition of the distance measure $\mathrm{d}_{\mathcal{X}}$ as a function of the difference of two cells $(\mathrm{x_i} - \mathrm{x_j})$, we can use the Taylor

expansion of $\mathrm{f} \circ \mathrm{d}_{\mathcal{X}}$ to measure biological similarity of cells. Because of Eq (2), the 0th order term vanishes,

$$s_{\mathrm{bio}}(i,j) = \nabla(f \circ d_{\mathcal{X}})|_{(x_i - x_j)=0} \cdot (x_i - x_j) + \mathcal{O}((x_i - x_j)^2) \quad (3)$$

PAGA infers the topological structure of data by solely relying on the limit $\mathrm{d}_{\mathcal{X}} \to 0$. In this limit, cells that are close can reasonably be assumed to be biological similar and are hence more likely to be continuously related –for instance, by a developmental process –than distant cells. This assumption can be represented by constructing the graph $\mathcal{G}$ of neighborhoods of single cells. Instead of invoking distances between cluster centroids to make inferences about lineage relations, like other methods, PAGA proceeds by studying the "connectivity" of partitions of the graph $\mathcal{G}$. By that, it establishes topological relations between partitions, clusters or "cell types", instead of distance measurements in a geometric space, like other methods. Through this highly different design, PAGA achieves a much higher level of robustness.

For single data points, this can only be observed for dense sampling. On the scale of clusters, this can almost never be observed. With its statistical tests for connectivity of partitions of the single-cell graph, PAGA leverages the smallest observed scale in the data –in contrast to all previous methods. For more background, see detailed comparisons with all previous methods in (22).

Monocle 2 prediction of a lineage tree for planaria shows qualitative inconsistencies

In order to infer a tree of lineage relations, Monocle 2 (*15*) uses "reversed graph embedding", which aims to fit a geometrical graph-model to projections of the data in a low-dimensional latent space. Even though, in principle, any model could be used for that, in practice, only tree-like models are computationally tractable. Hence, Monocle 2 tries to force

data into a tree-like topology without providing a statistical measure for how reliable the resulting fit is. The result on the data show, for instance, the epidermal and gut lineages are distributed across the full tree (Fig. S18).

Robustness of the PAGA cell map for Planaria

Since the advent of computational methods for reconstructing lineage trees from molecular data (*68*), these methods (*16*, *69–73*) have been plagued by robustness issues and their reliability has been rightfully questioned. Here, we address this criticism by showing robustness of our results with respect to subsampling, the choice of samples, and parameters of the graph representation. The pseudotime coordinate in the PAGA cell maps is an extension of Diffusion Pseudotime (*13*) for partially disconnected graphs. Its robustness has been demonstrated before and will not be discussed here. Further robustness studies on different datasets for the PAGA abstracted graph can be found in (22).

Robustness across subsampling

To assess the robustness of PAGA lineage reconstruction to number of cells, we run PAGA after subsampling 100%, 80% and 10% of the data. Even when subsampling 80% of the data, the abstracted graph obtained with 100% of the data is perfectly reproduced, and only when rerunning PAGA using 10% of the data we observe some major deviations in the topology of the original graph (Fig. S19).

Robustness using only wild type samples

The abstracted graph obtained when running PAGA using only wt samples (Fig. S20) shows that the inference of the lineage relations for the wt samples yields nearly the same result

as the inference using all samples together (Fig. S19A). This reflects that the use of the regenerating planarian samples does not affect the inference of differentiation trajectories.

Robustness when varying the number of neighbors in the graph representation

The number of nearest neighbors in the single-cell graph is the only parameter of PAGA if we consider data preprocessing as an independent step. We used 4 different numbers of neighbors (5, 20, 50 and 100) and compared the resulting abstracted graphs (Fig. S21) with that of the original plot, which uses 30 neighbors (Fig. S19A). All the graphs show high similarity thus confirming the robustness of PAGA regardless of the number of nearest neighbors used from the single-cell graph. Large sampling studies of parameters also showed highly robust reconstruction of topologies using other datasets (*22*).

The epidermal lineage

Analysis of marker gene expression shows that between the late epidermal progenitors cluster and the epidermis cluster, the columns marked dark blue in the "clusters" row, there is a strong discontinuity of the epidermal lineage (Fig. S22A). This explains why PAGA (Fig. 3, Fig. S22) fails to place the epidermis in its biologically correct location, even when only the clusters from the epidermal lineage are explored (Fig. S22B).

Supplementary Note 4: The potency score

In the literature, there are two main views on how potency of cells can be measured. One view assumes that the transcriptome of cells with higher potency is less biased towards certain specific gene sets than those of mature cells. To quantify the degree of "bias of the transcriptome", several entropy-based measures have been developed. Another view alludes to the definition of potency: cells that can develop towards a diversity of fates have higher potency than cells that can only develop to a small number of cell fates. Grün et al. realized that in their inferred differentiation trees they only needed to count the number of outgoing edges for each cell type. They used this number and multiplied it with an entropy-measure to define the "stemID" score, which they propose to measure potency (*16*).

We follow this suggestion but propose to account for confidence in the presence of edges in the differentiation tree, or more generally, in the abstracted graph. In the framework of graph abstraction, the confidence weight per edge is given by the attachedness or connectivity of partitions on the neighborhood graph. Even though there are several ways to define such a measure, they are all very similar (*22*). We then define a potency score for each node as the degree in the abstracted graph, in which weights quantify confidence in the presence of edges (Fig. 3B). This degree-based potency score agrees better with markers that quantify potency than Grün et al.'s edge-number score and an expression-entropy based score for measuring potency. In particular, in contrast to Grün et al. (*16*), the framework of graph abstraction makes an expression-entropy score unnecessary as the degree-based potency score fully distinguishes stem from more differentiated cells.

Supplementary Note 5: Characterization of specialized neoblast markers

In planarians, it has been shown the presence of specialized neoblast markers that give raise to specific lineages (*26*, *42*, *43*). In our analysis, we did not find specific clusters containing these subpopulations, which may be related to the sensitivity of our single-cell sequencing method. However, these specialized neoblast markers are expressed both among neoblast and progenitor clusters (Fig. S9-S11). In the lineage analysis done with PAGA, neoblast 1 has the highest potency as all other clusters are connected to it. All the other neoblast clusters have lower potency scores and were connected to neoblast 1 (Fig. 3B). These neoblast clusters share the majority of marker genes with the neoblast 1 cluster (Table S3) and do not correspond to the previously identified specialized neoblasts (*26*, *42*, *43*).

Some of these neoblast clusters indeed are connected to differentiated types (neoblast 4 to otf+ cells 2, neoblast 11 to secretory 2, epidermis DVb neoblasts to epidermis DVb and epidermal neoblasts to early epidermal progenitors). All other neoblast clusters are connected only to the central neoblast 1 cluster.

When ordering the cells in neoblast 1 and progenitor clusters according to pseudotime, we observe that classical neoblast markers quickly drop with pseudotime whereas specialized neoblast marker expression increases (Fig. S11). Therefore, the progenitor clusters identified contain both neoblasts and post-mitotic progenitors that are *Smedwi-1* negative. These results suggest that specialized neoblasts are subtypes of an heterogeneous cell type rather than different cell types that cannot be properly profiled using Drop-seq (*74*).

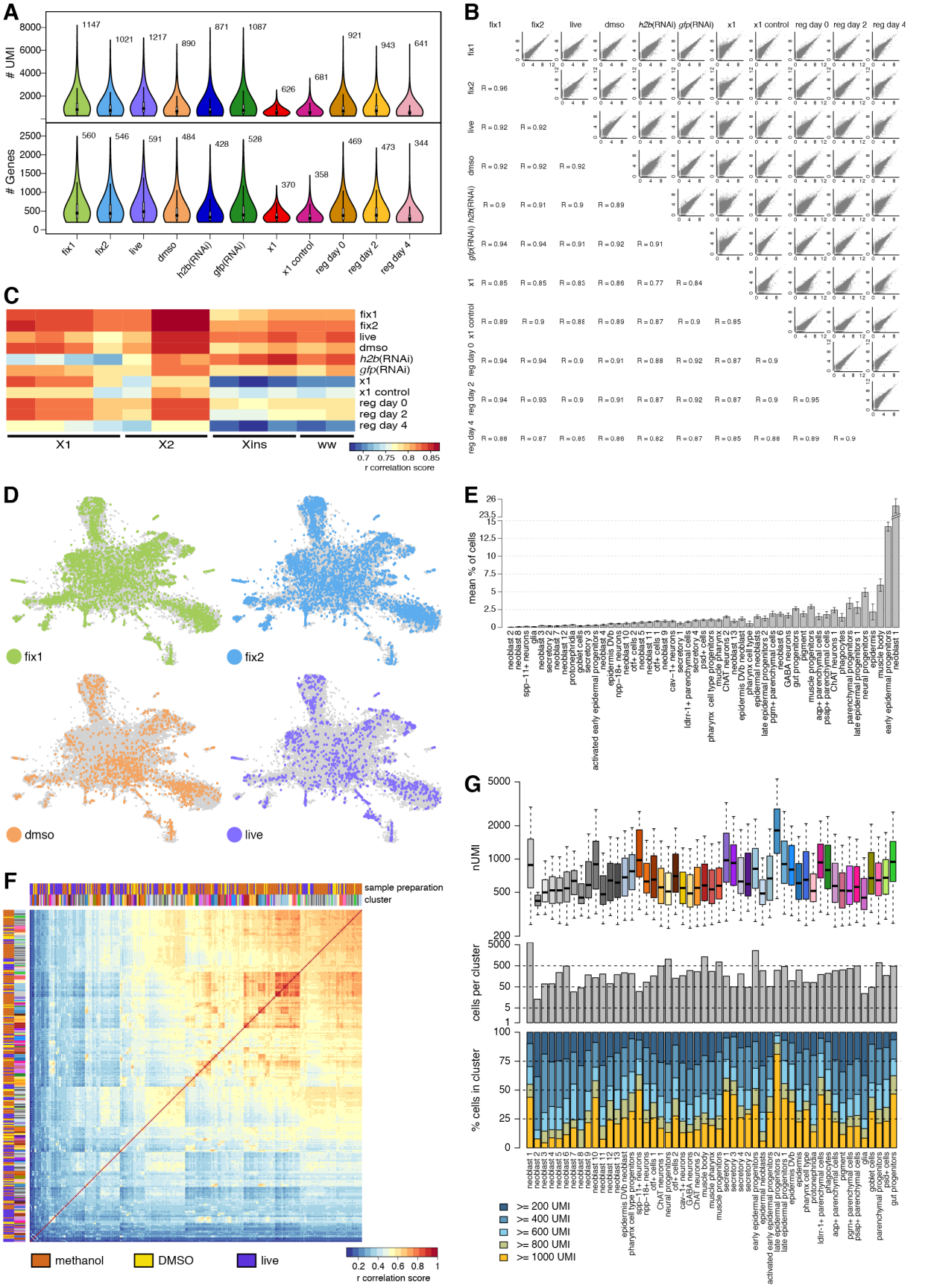

A
UMI
Genes
fix1
fix2
live
dmso
h2b(RNAi)
gfp(RNAi)
x1
x1 control
reg day 0
reg day 2
reg day 4
B
R = 0.96
R = 0.92
R = 0.9
R = 0.91
R = 0.89
R = 0.94
R = 0.85
R = 0.83
R = 0.86
R = 0.77
R = 0.84
R = 0.88
R = 0.87
R = 0.92
R = 0.95
C
X1
X2
Xins
ww
r correlation score
D
E
mean % of cells
F
sample preparation
cluster
methanol
DMSO
live
r correlation score
G
nUMI
cells per cluster
% cells in cluster
>= 200 UMI
>= 400 UMI
>= 600 UMI
>= 800 UMI
>= 1000 UMI

Fig. S1. Drop-seq run statistics and comparison with previous data

A. Violin plots showing the distribution of UMIs (top) and genes (bottom) per sample. The numbers on top of each violin plot indicate the mean value in each sample. **B**. Scatter plots showing the correlation in gene expression across all samples. The axes represent log (UMI+1) counts. The lower half of the matrix shows the Pearson correlation coefficient R for the comparisons in the upper half. Although all correlations are quite high (R >= 0.7), the lowest is observed between the *h2b*(*RNAi*) and the X1 samples, which contain different subsets of the total cell population. **C.** Heatmap showing the correlation between gene expression for each Drop-seq sample and previously published RNA-seq datasets from FACS sorted populations and whole worms. **D.** tSNE plots showing the distribution of the cells of the 4 wild type replicates: live (purple), dmso fixed (orange), methanol fix 1 (green) and methanol fix 2 (blue), showing that the processing of the samples does not affect the distribution of cells on the tSNE. **E.** Mean number of cells per cluster calculated across wild type datasets. The error bars show the standard error of the mean indicating a high similarity in cell proportions across datasets. **F.** Gene expression correlation across clusters for the wild type samples analyzed. The sample preparation protocol as well as the cluster identity of each cluster in each sample is shown. Clustering across samples is mainly driven by the identity of the cluster rather than the sample preparation protocol. **G.** Distribution of UMIs (top) and number of cells (middle) in each of the clusters. The average number of UMI changes significantly across clusters, with smaller clusters having generally less UMIs. If higher quality cells, i.e. cells with more UMIs, would be used in the analyses, the proportion of cells per cluster would change unevenly.

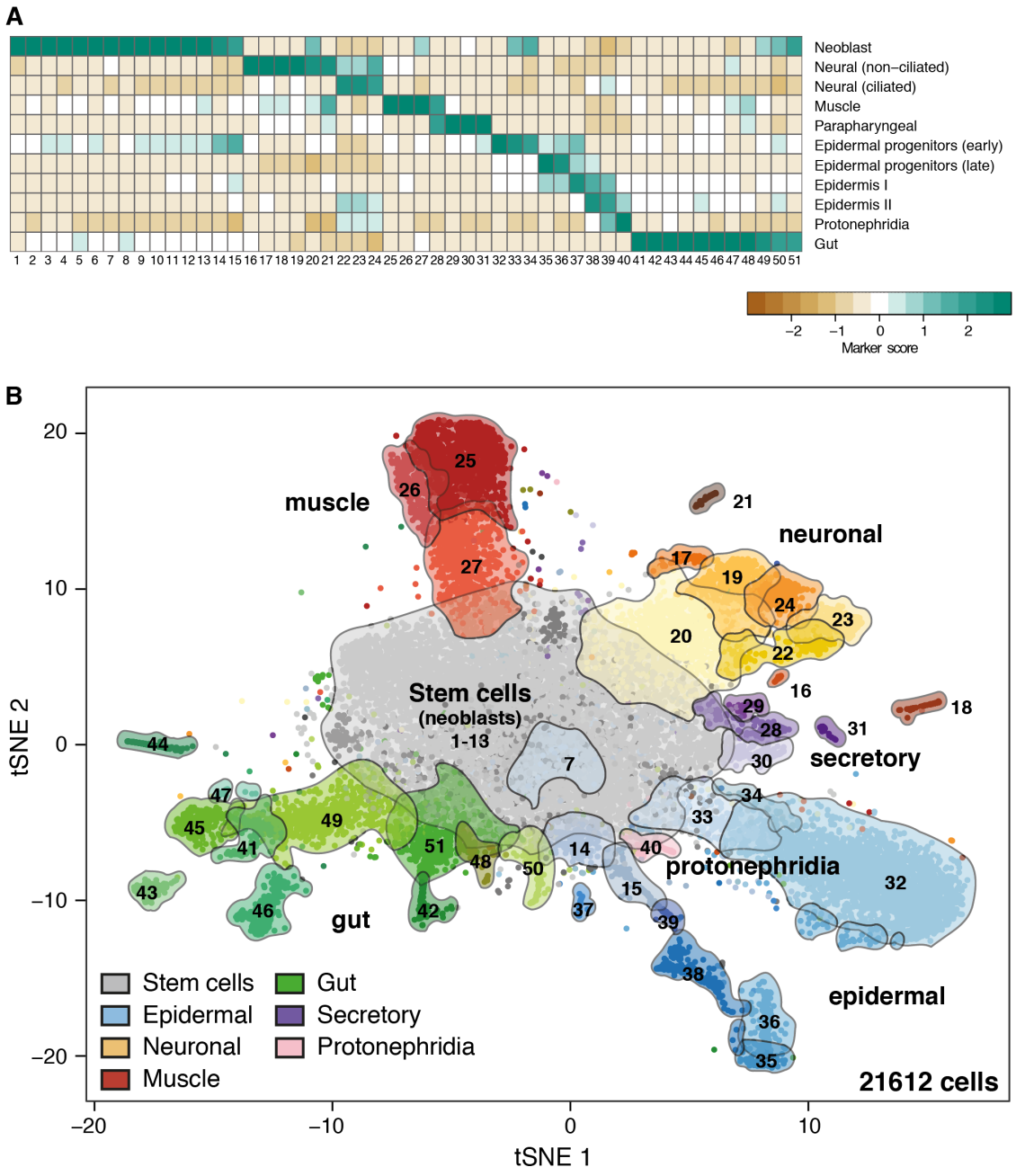


Fig. S2. Characterization of clusters with published markers

A. Scoring of the Drop-seq cell clusters using previously defined marker genes (*42*). Each of the identified clusters can be associated with the previously defined cell types but we find higher cell type diversity. **B.** tSNE plot showing clusters labeled according to their expression of markers shown in (**A**).

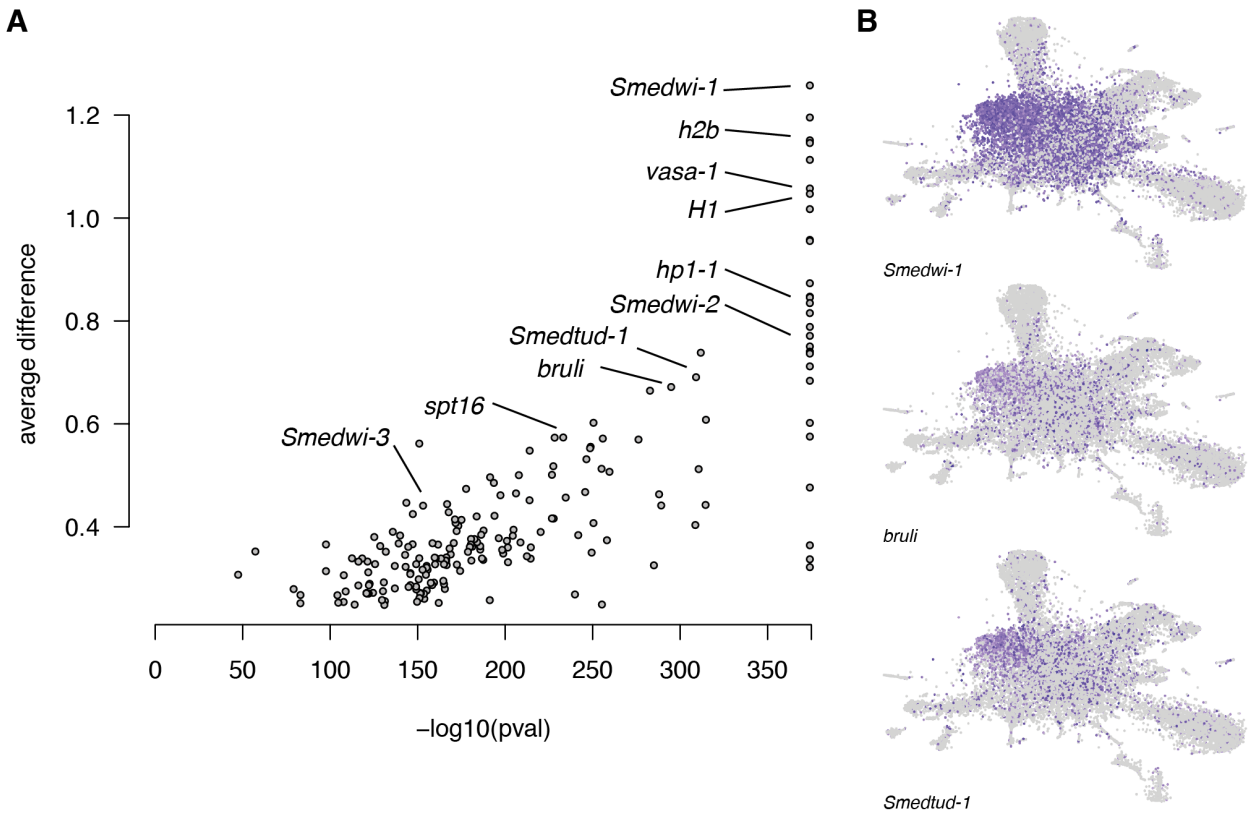


Fig. S3. Marker expression in neoblast clusters

A. Scatter plot of the identified markers of the cluster neoblast 1, including well-known neoblast marker genes. **B.** tSNE plots showing expression of the classic neoblast marker genes *Smedwi-1* (dd_Smed_v6_659_0), *bruli* (dd_Smed_v5_2592) and *Smedtud-1* (dd_Smed_v5_1582) in the central clusters of the tSNE. Color scale for tSNE plots ranges from light grey (no expression) to blue (high expression).

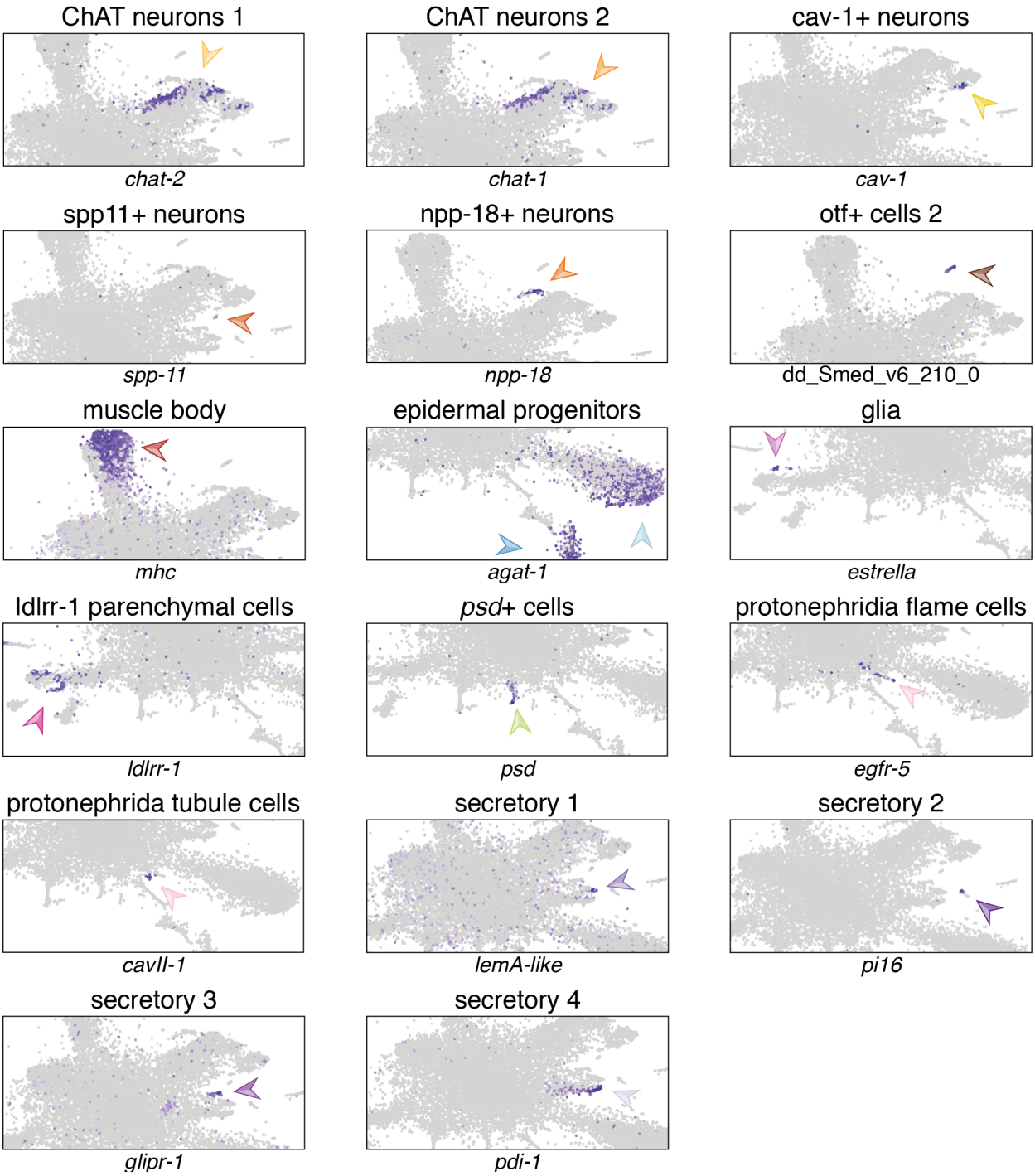


Fig. S4. Marker expression of differentiated cell clusters

tSNE plots showing the expression of several cluster markers, showing differentially expressed genes of each: ChAT neurons 1 (*chat-2*, dd_Smed_v6_11968_0), ChAT neurons 2 (*chat-1*,

dd_Smed_v6_6208_0), cav-1+ neurons (*cav-1*, dd_Smed_v6_8555_0), spp-11+ neurons (*spp-11*, dd_Smed_v6_1724_0), npp-18+ neurons (*npp-18*, dd_Smed_v6_1117_0), otf+ cells 2 (dd_Smed_v6_210_0), muscle body (*mhc*, dd_Smed_v6_432_0), late epidermal progenitors 2 (*agat-1*, dd_Smed_v6_920_0), ldlrr-1+ parenchymal cells (*ldlrr-1*, dd_Smed_v6_1581_0_1), psd+ cells (*psd*, dd_Smed_v6_9018_0_1), glia (*estrella*, dd_Smed_v6_1792_0), protonephridia flame cells (*egfr-5*, dd_Smed_v6_11310_0), protonephridia tubule cells (*cavII-1*, dd_Smed_v6_4841_0), secretory 1 (*lemA-like*, dd_Smed_v6_750_0), secretory 2 (*pi16*, dd_Smed_v6_7651_0), secretory 3 (*glipr-1*, dd_Smed_v6_924_0), secretory 4 (*pdi-1*, dd_Smed_v6_821_0). Color scale for tSNE plots ranges from light grey (no expression) to blue (high expression).

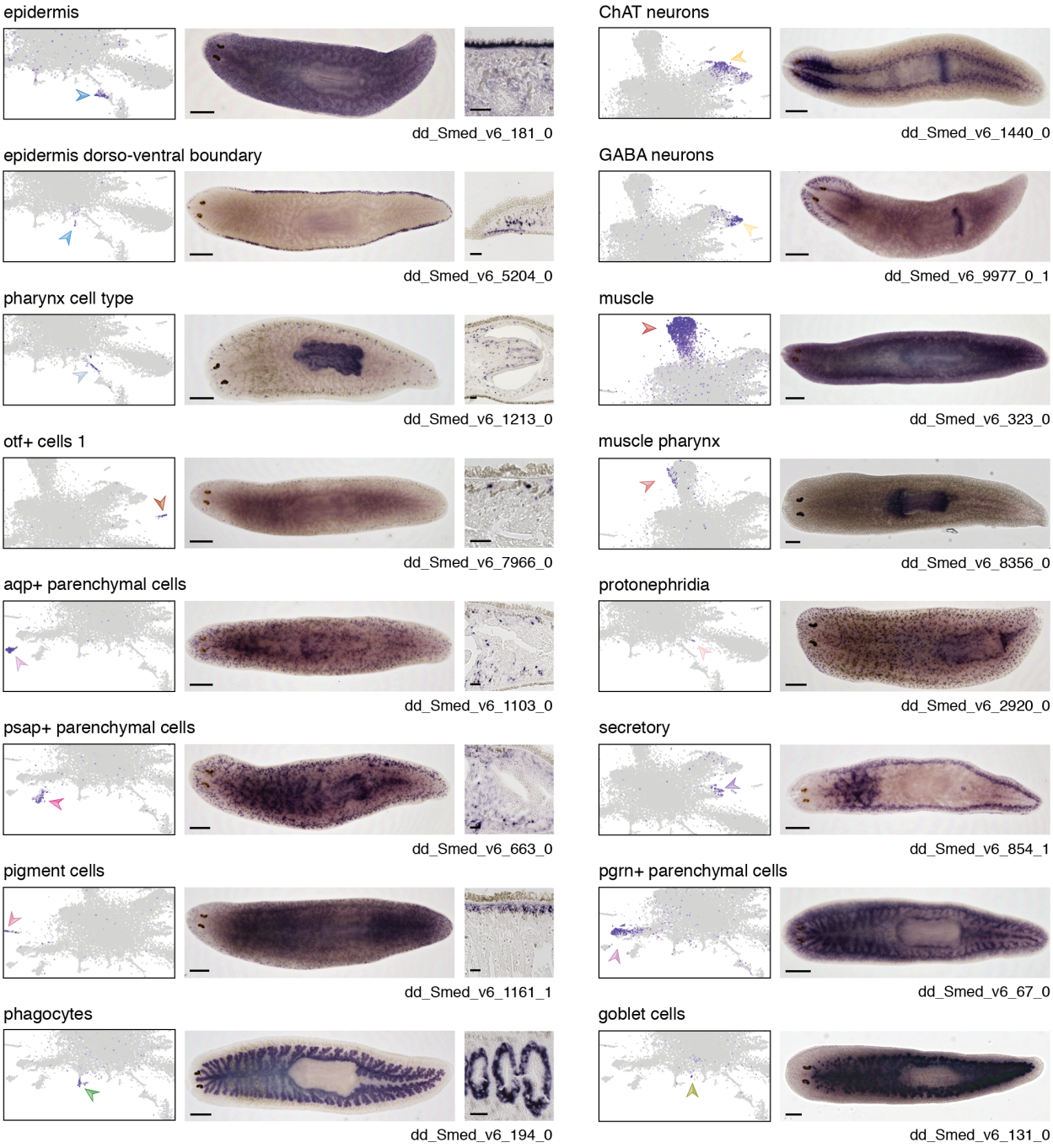


Fig. S5. Validation of marker gene expression by whole mount *in situ* hybridization

tSNE plots (left panels) and expression of the identified marker genes of known and novel cell types and their expression patterns in adult animals using *in situ* hybridizations in whole mount and tissue sections (right panels, left column). Color scale of the tSNE plots ranges from light

grey (no expression) to blue (high expression). Scale bars: 500 μm for whole mount *in situ* hybridizations, 100 μm for *in situ* on sections.

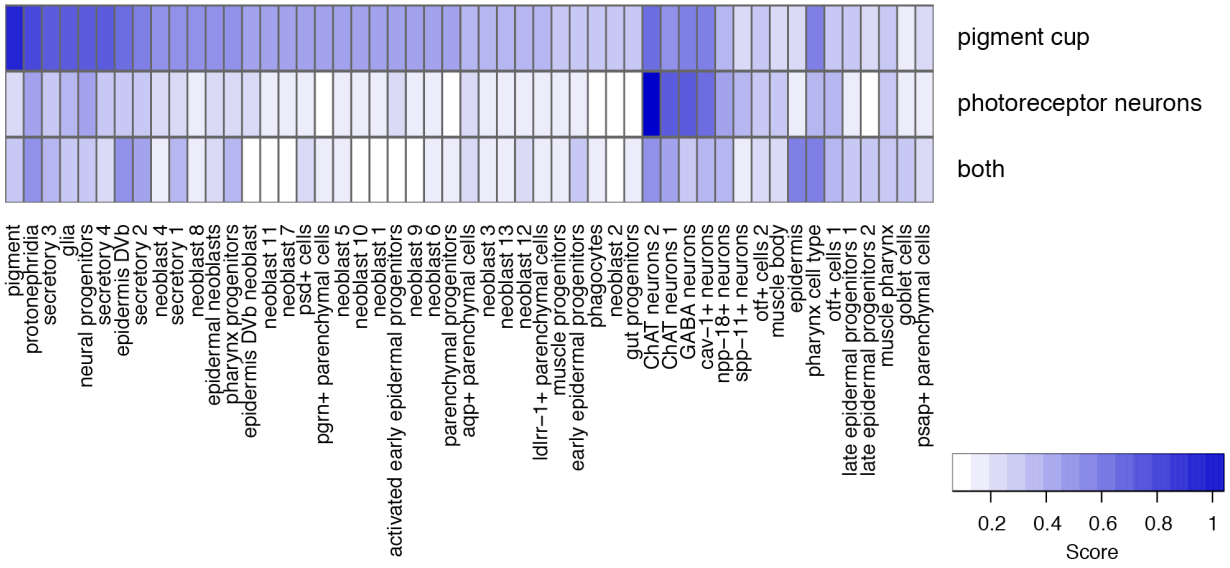


Fig. S6. Expression of eye rare cell subtypes.

Heatmap showing the scoring of the Drop-seq cell clusters using previously described marker genes of the eye, including markers of photoreceptor neurons and pigment cup cells. The color scale ranges from white (low score) to blue (high score).

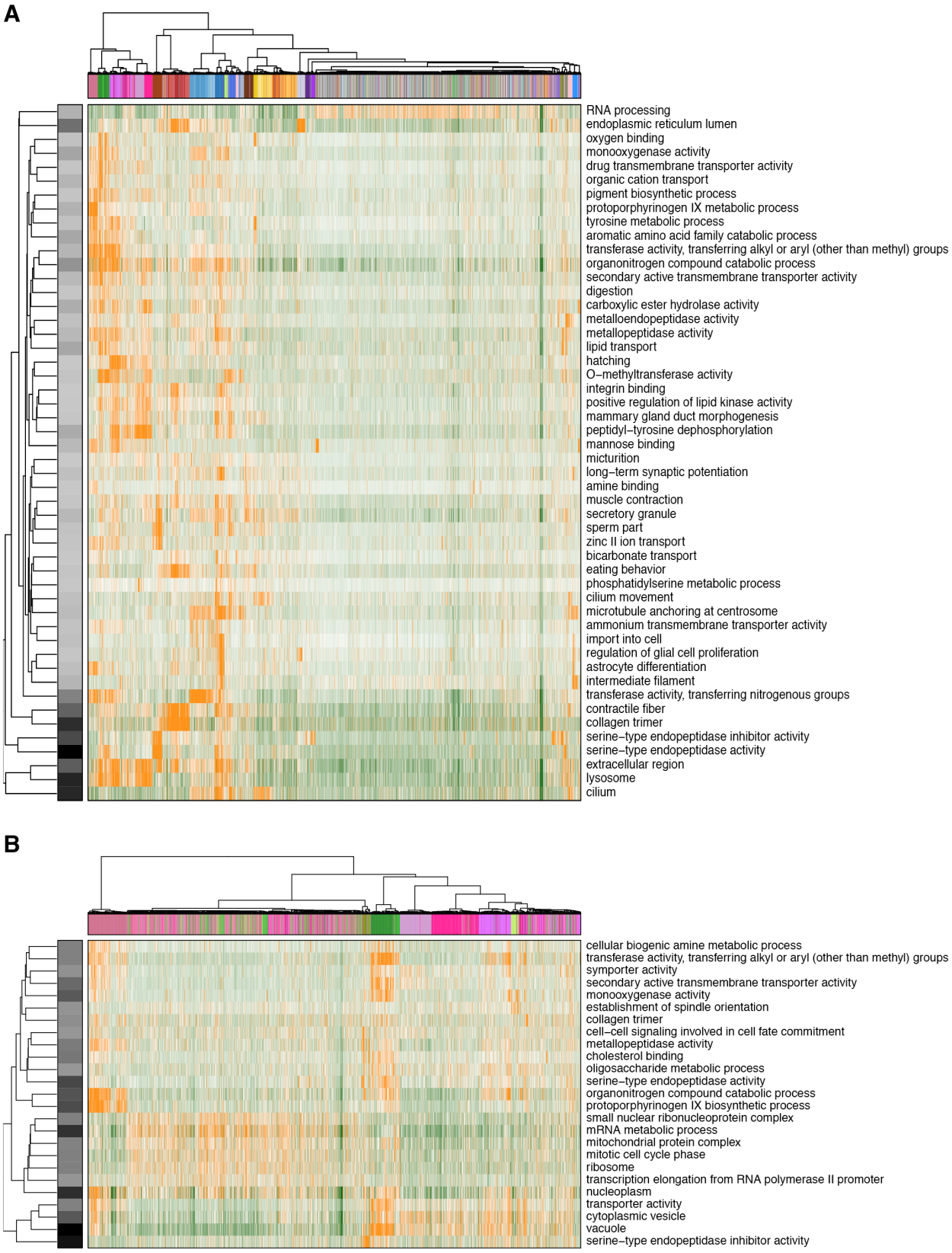

A
RNA processing
endoplasmic reticulum lumen
oxygen binding
monooxygenase activity
drug transmembrane transporter activity
organic cation transport
pigment biosynthetic process
protoporphyrinogen IX metabolic process
tyrosine metabolic process
aromatic amino acid family catabolic process
transferase activity, transferring alkyl or aryl (other than methyl) groups
organonitrogen compound catabolic process
secondary active transmembrane transporter activity
digestion
carboxylic ester hydrolase activity
metalloendopeptidase activity
metallopeptidase activity
lipid transport
hatching
O−methyltransferase activity
integrin binding
positive regulation of lipid kinase activity
mammary gland duct morphogenesis
peptidyl−tyrosine dephosphorylation
mannose binding
micturition
long−term synaptic potentiation
amine binding
muscle contraction
secretory granule
sperm part
zinc II ion transport
bicarbonate transport
eating behavior
phosphatidylserine metabolic process
cilium movement
microtubule anchoring at centrosome
ammonium transmembrane transporter activity
import into cell
regulation of glial cell proliferation
astrocyte differentiation
intermediate filament
transferase activity, transferring nitrogenous groups
contractile fiber
collagen trimer
serine−type endopeptidase inhibitor activity
serine−type endopeptidase activity
extracellular region
lysosome
cilium
B
cellular biogenic amine metabolic process
transferase activity, transferring alkyl or aryl (other than methyl) groups
symporter activity
secondary active transmembrane transporter activity
monooxygenase activity
establishment of spindle orientation
collagen trimer
cell−cell signaling involved in cell fate commitment
metallopeptidase activity
cholesterol binding
oligosaccharide metabolic process
serine−type endopeptidase activity
organonitrogen compound catabolic process
protoporphyrinogen IX biosynthetic process
small nuclear ribonucleoprotein complex
mRNA metabolic process
mitochondrial protein complex
mitotic cell cycle phase
ribosome
transcription elongation from RNA polymerase II promoter
nucleoplasm
transporter activity
cytoplasmic vesicle
vacuole
serine−type endopeptidase inhibitor activity

Fig. S7. Pagoda analysis of GO terms

A, B. Heatmap showing the top aspects ($p < 0.01$) identified with PAGODA using GO term based gene sets that vary across all cells **(A)** or that distinguish parenchyma and gut cells **(B)**. The first row shows the colors assigned to the cells according to the clusters obtained with Seurat. Each row shows the Cell PC scores for each of the cells (columns). The colors range from green (low score) to orange (high score). The GO terms associated to each aspect are shown on the right. The clustering obtained with PAGODA defines groups of cells that strongly agree with Seurat clustering except for the clusters closer to the neoblast cloud which are heavily intermixed, which suggests that they may correspond to less differentiated cell types.

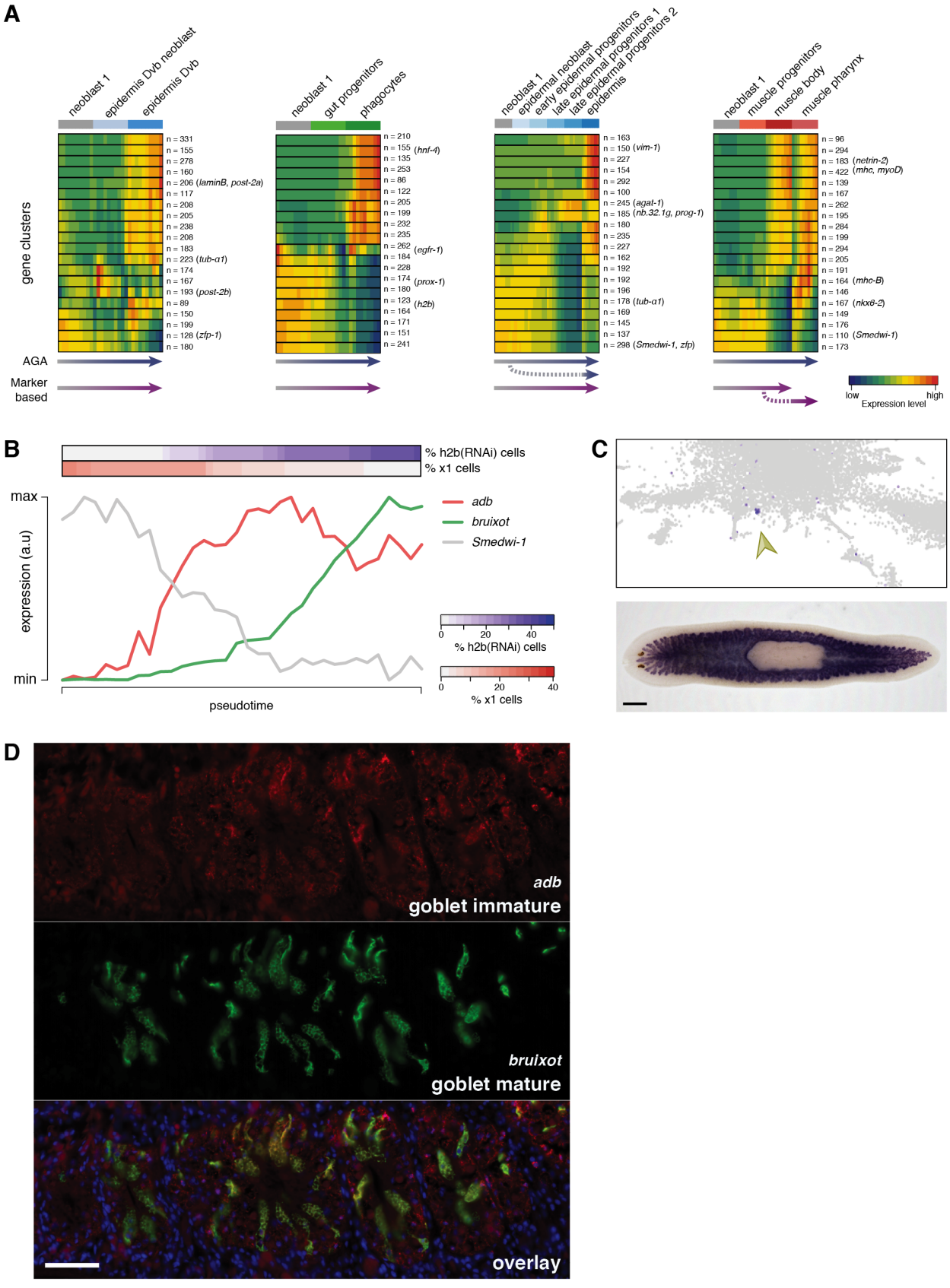

A
neoblast 1
epidermis Dvb neoblast
epidermis Dvb
neoblast 1
gut progenitors
phagocytes
neoblast 1
epidermal neoblast
early epidermal progenitors 1
late epidermal progenitors 1
late epidermal progenitors 2
epidermis
neoblast 1
muscle progenitors
muscle body
muscle pharynx
gene clusters
n = 331
n = 155
n = 278
n = 160
n = 206 (laminB, post-2a)
n = 117
n = 208
n = 205
n = 238
n = 208
n = 183
n = 223 (tub-α1)
n = 174
n = 167
n = 193 (post-2b)
n = 89
n = 150
n = 199
n = 128 (zfp-1)
n = 180
n = 210
n = 155 (hnf-4)
n = 135
n = 253
n = 86
n = 122
n = 205
n = 199
n = 232
n = 235
n = 262 (egfr-1)
n = 184
n = 228
n = 174 (prox-1)
n = 180
n = 123 (h2b)
n = 164
n = 171
n = 151
n = 241
n = 163
n = 150 (vim-1)
n = 227
n = 154
n = 292
n = 100
n = 245 (agat-1)
n = 185 (nb.32.1g, prog-1)
n = 180
n = 235
n = 227
n = 162
n = 192
n = 192
n = 196
n = 178 (tub-α1)
n = 169
n = 145
n = 137
n = 298 (Smedwi-1, zfp)
n = 96
n = 294
n = 183 (netrin-2)
n = 422 (mhc, myoD)
n = 139
n = 167
n = 262
n = 195
n = 284
n = 199
n = 294
n = 205
n = 191
n = 164 (mhc-B)
n = 146
n = 167 (nkx6-2)
n = 149
n = 176
n = 110 (Smedwi-1)
n = 173
AGA
Marker based
low
high
Expression level
B
% h2b(RNAi) cells
% x1 cells
max
min
expression (a.u)
pseudotime
adb
bruixot
Smedwi-1
0
20
40
% h2b(RNAi) cells
0
20
40
% x1 cells
C
D
adb
goblet immature
bruixot
goblet mature
overlay

Fig. S8. Gene expression changes in pseudotemporally ordered cells evaluate lineage relationships

A. Average expression of variable gene sets identified using SOMs during the differentiation of neoblasts to epidermis DVb, phagocytes, epidermis and muscle. The changes in expression of these gene sets in the cells ordered according to pseudotime is gradual and supports the differentiation trajectories predicted by PAGA in the epidermis DVb and phagocyte lineages. However, gene expression contradicts the differentiation trajectories proposed by the PAGA in the case of the epidermal and the muscle lineages, which show continuity (epidermis) or discontinuity (muscle) in the gene expression along the trajectory. **B.** Relative expression of *adb, bruixot*, and *Smedwi-1* in pseudotemporally ordered cells from the goblet cluster. On top of the plot, the cumulative fraction of *h2b*(*RNAi*) resistant as well as X1 cells is shown. The expression of the genes along the cluster suggests that these genes are markers of immature (*adb*) and mature (*bruixot*) goblet cells **C.** Expression of *adb* on a tSNE plot (top) and in a whole mount *in situ* hybridization (bottom). **D.** Double fluorescent *in situ* hybridization of *adb* (red) and *bruixot* (green). The two genes overlap but *adb* has a broader expression in cells that lack the goblet cell morphology. Nuclei were stained with *Hoechst* and shown in blue in the overlay.

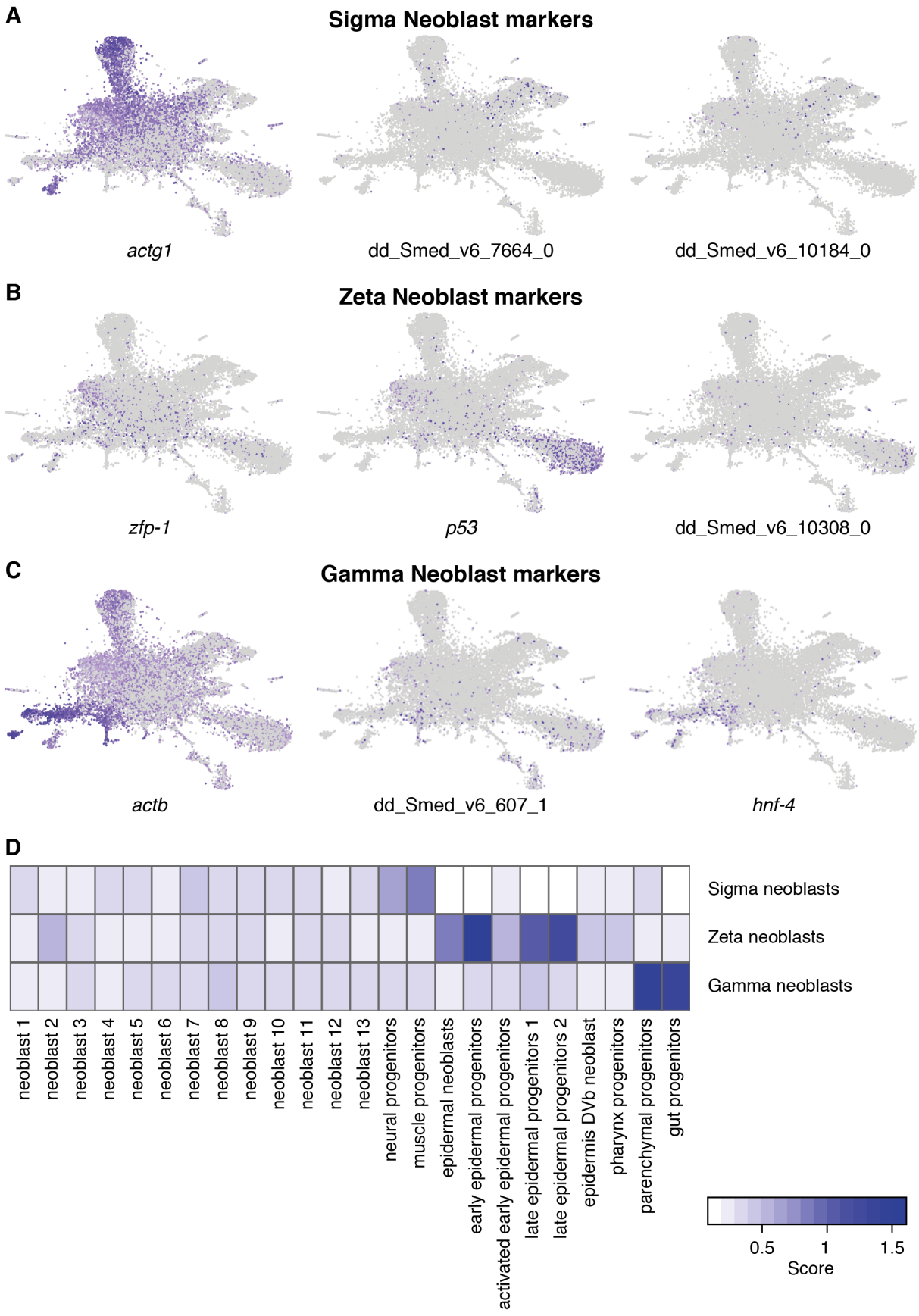

A
Sigma Neoblast markers
actg1
dd_Smed_v6_7664_0
dd_Smed_v6_10184_0
B
Zeta Neoblast markers
zfp-1
p53
dd_Smed_v6_10308_0
C
Gamma Neoblast markers
actb
dd_Smed_v6_607_1
hnf-4
D
Sigma neoblasts
Zeta neoblasts
Gamma neoblasts
neoblast 1
neoblast 2
neoblast 3
neoblast 4
neoblast 5
neoblast 6
neoblast 7
neoblast 8
neoblast 9
neoblast 10
neoblast 11
neoblast 12
neoblast 13
neural progenitors
muscle progenitors
epidermal neoblasts
early epidermal progenitors
activated early epidermal progenitors
late epidermal progenitors 1
late epidermal progenitors 2
epidermis DVb neoblast
pharynx progenitors
parenchymal progenitors
gut progenitors
0.5
1
1.5
Score

Fig. S9. Specialized neoblasts markers

tSNE plots showing the expression of specialized neoblast marker genes of sigma **(A)**, zeta **(B)**, and gamma **(C)** neoblasts. Color scale from tSNE plots as described previously. **D.** Heatmap showing the scoring of the Drop-seq cell clusters using previously described marker genes of sigma, gamma and zeta neoblasts. The color scale ranges from white (low score) to blue (high score).

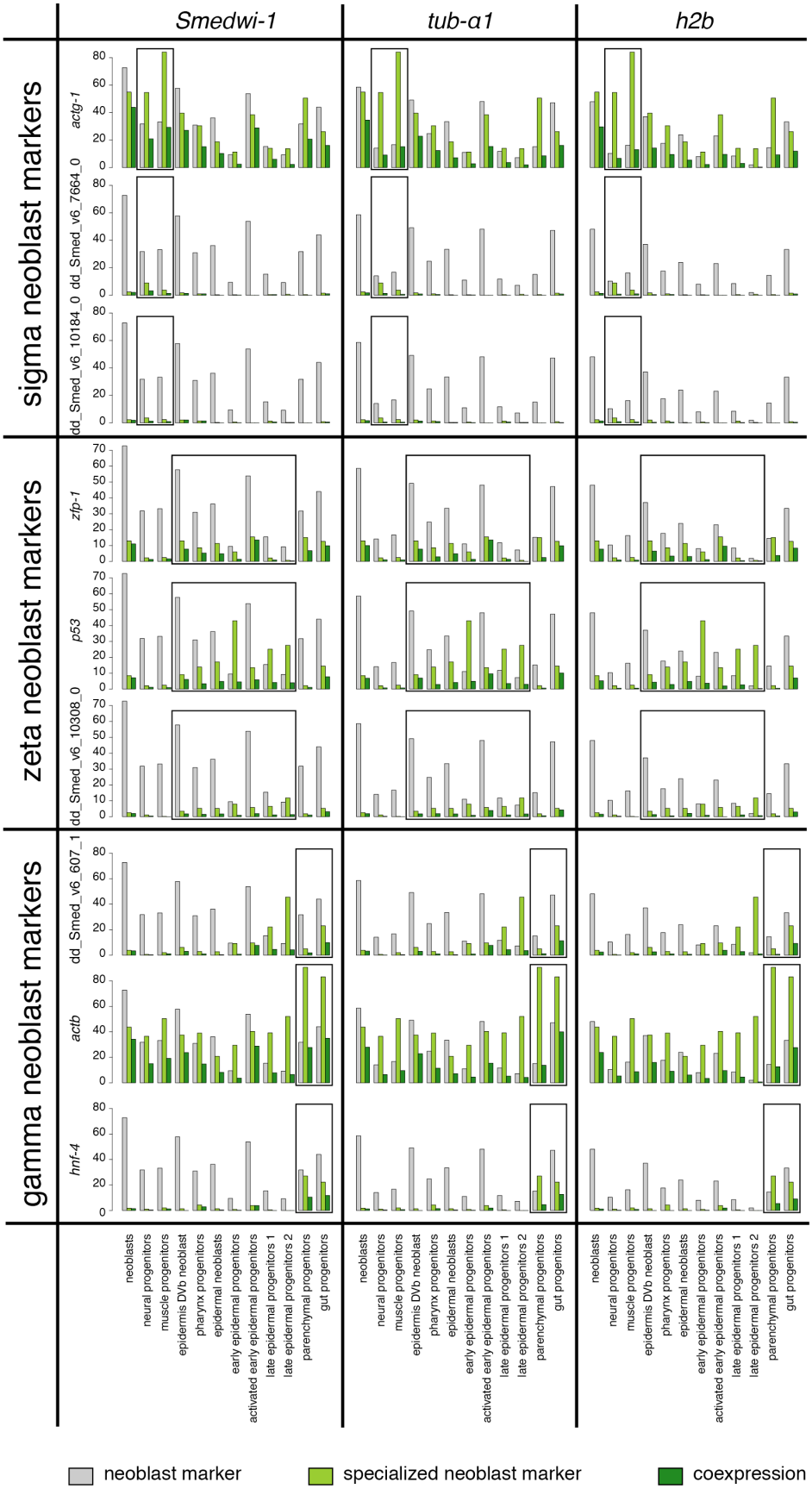

gamma neoblast markers
zeta neoblast markers
sigma neoblast markers
Smedwi-1
tub-α1
h2b
hnf-4
actb
dd_Smed_v6_607_1
dd_Smed_v6_10308_0
p53
zfp-1
dd_Smed_v6_10184_0
dd_Smed_v6_7664_0
actg-1
neoblasts
neural progenitors
muscle progenitors
epidermis DVb neoblast
pharynx progenitors
epidermal neoblasts
early epidermal progenitors
activated early epidermal progenitors
late epidermal progenitors 1
late epidermal progenitors 2
parenchymal progenitors
gut progenitors
neoblast marker
specialized neoblast marker
coexpression

Fig. S10. Co-expression of neoblast markers and specialized neoblasts markers

Barplots showing the percentage of cells expressing general neoblasts markers (grey bars), specialized neoblast markers (light green) and their coexpression (dark green) in cells from neoblasts and progenitor clusters.

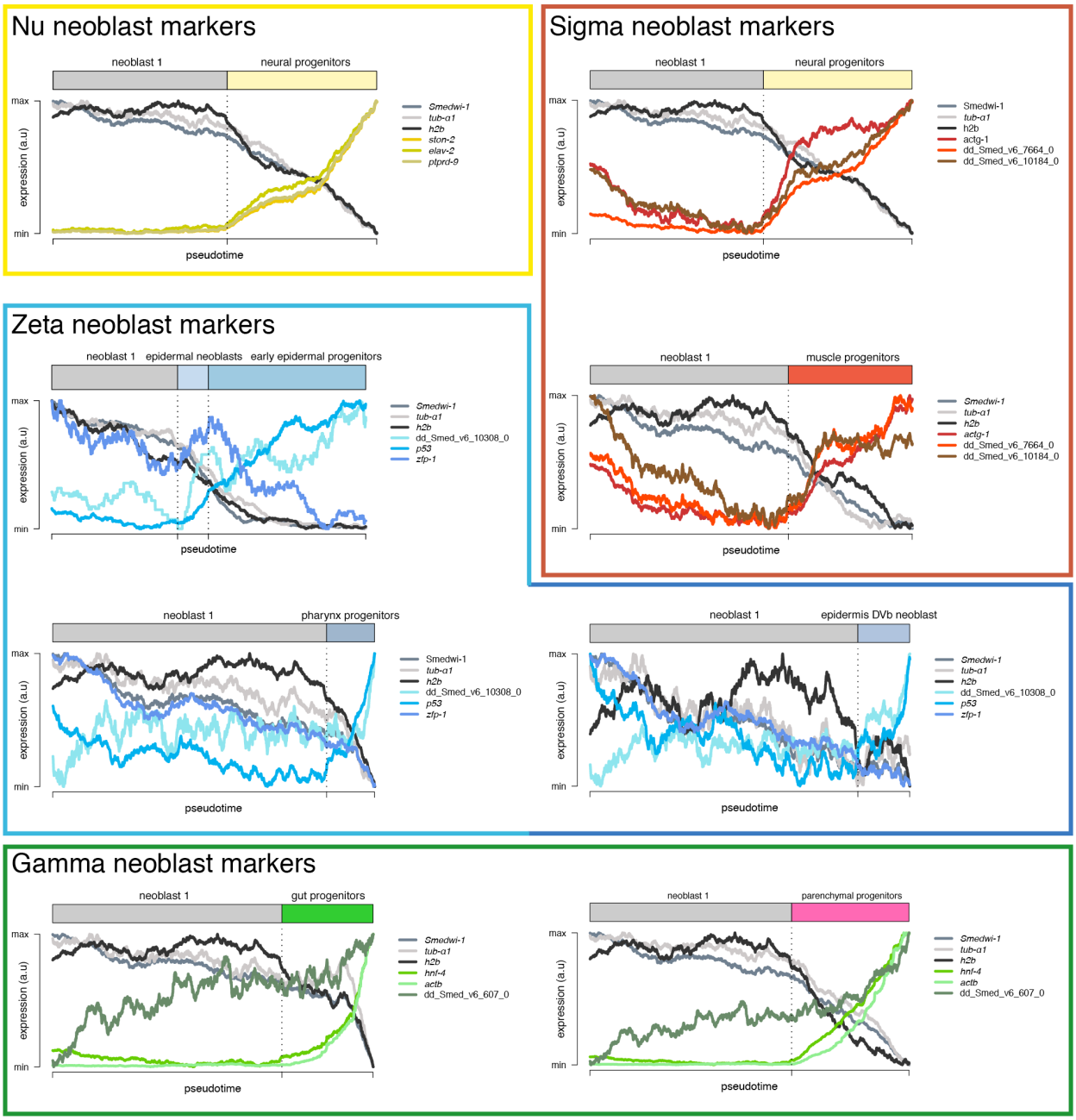


Fig. S11. Expression of neoblast markers in pseudotemporally ordered neoblast and progenitor cells

Relative gene expression of general neoblasts markers (*Smedwi-1*, *tub-α1* and *h2b*; grey lines), and nu (yellow), sigma (red), zeta (blue) and gamma (green) neoblast markers in cells from neoblast 1 and progenitor clusters of different lineages. In all the included lineages, we observe a

progressive decrease in the expression of general neoblast markers as well as an increase of specialized neoblast markers with pseudotime. A maximum of 1000 cells per cluster were randomly sampled for each cluster.

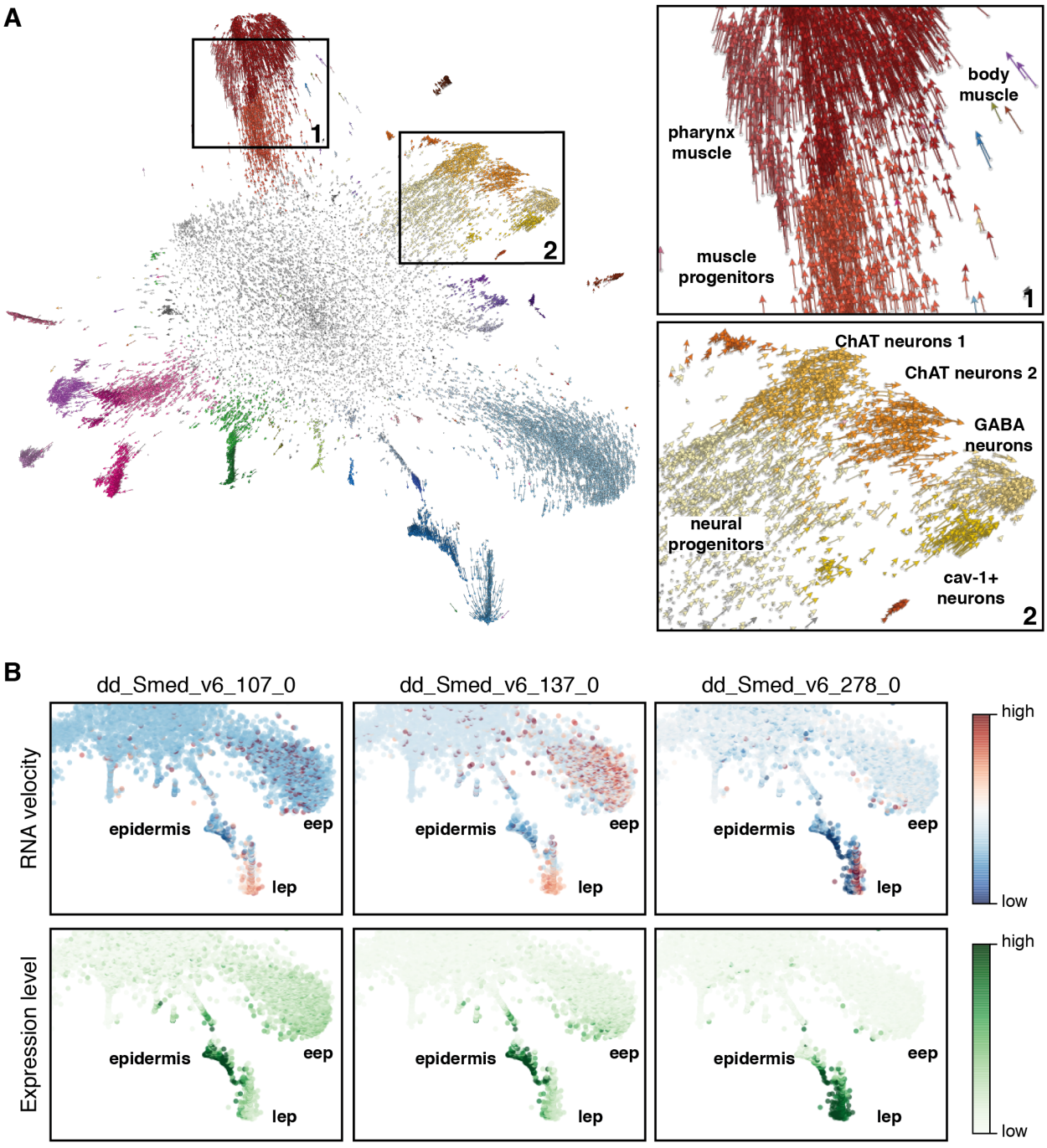


Fig. S12. RNA Velocity identifies differentiation trajectories for individual cells

A. tSNE plot showing the predicted RNA velocity trajectories for the individual cells (left) and a detailed view of the muscle (1) and neural (2) differentiation trajectories. (1) Cells from the muscle progenitor cluster have trajectories going towards pharynx and body muscle. (2) Neural

progenitor cell trajectories point on the one hand towards ChAT neurons 1 and these to ChAT neurons 2 and on the other hand towards cav-1+ neurons and these to GABA neurons. These paths corroborate the previous trajectories identified using PAGA (neural lineages) or based on marker gene expression changes along pseudotemporally ordered cells (muscle lineages). **B.** Comparison of RNA velocity (top panels) and expression levels (bottom panels) of the epidermis marker genes dd_Smed_v6_107_0, dd_Smed_v6_137_0 and dd_Smed_v6_278_0. In the tree cases, we observe an induction of the expression of marker genes in epidermal progenitors clusters and a repression of the expression in the epidermis cluster compared to steady state. These plots therefore suggest a connection between late epidermal progenitors and epidermis.

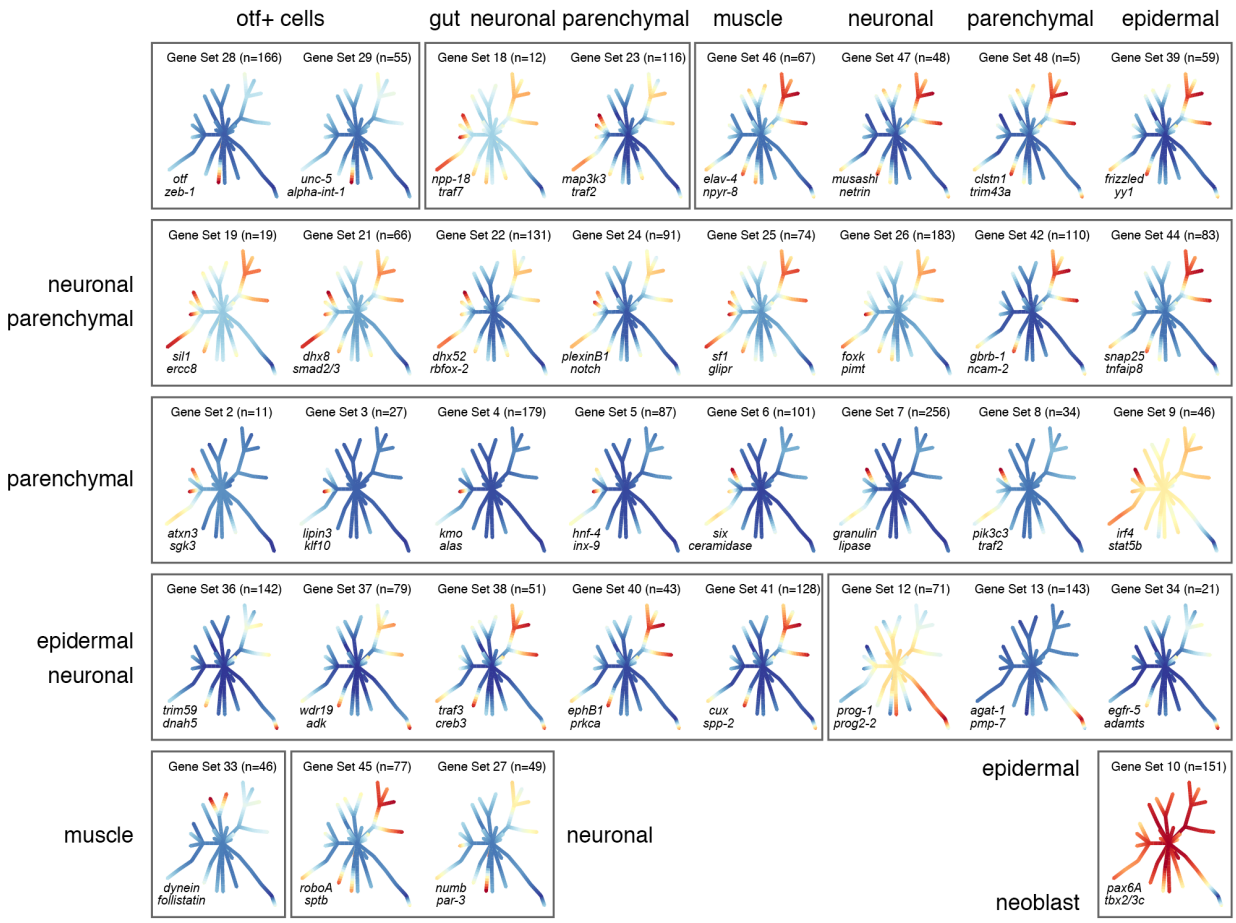


Fig. S13. **Identification of gene sets regulated and coregulated in cell differentiation**

Graphical representation of gene expression changes during cell differentiation of 36 gene sets. For each gene set, the normalized expression of the genes is shown on the edges of the tree and ranges from blue (low expression) to red (high expression). Next to each tree, two representative genes from the cluster are highlighted.

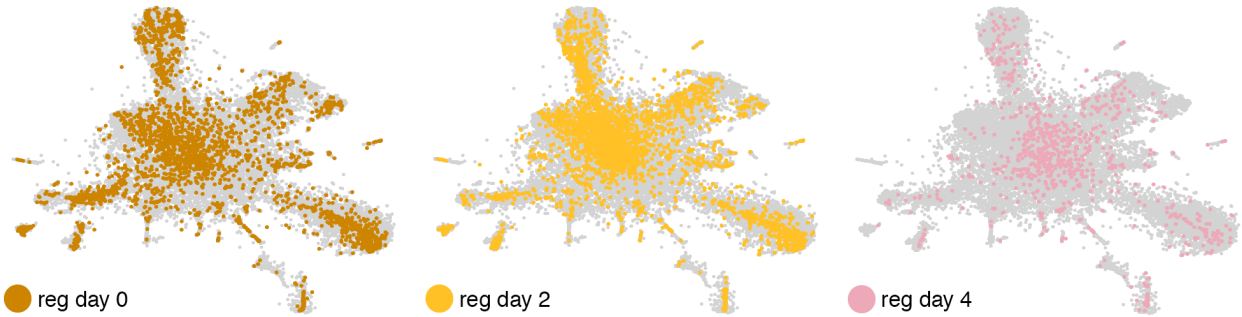


Fig. S14. tSNE plots of regeneration samples

A. tSNE plots showing the distribution of the cells of regeneration samples after 0 (reg day 0), 2 (reg day 2) and 4 (reg day 4) days of regeneration.

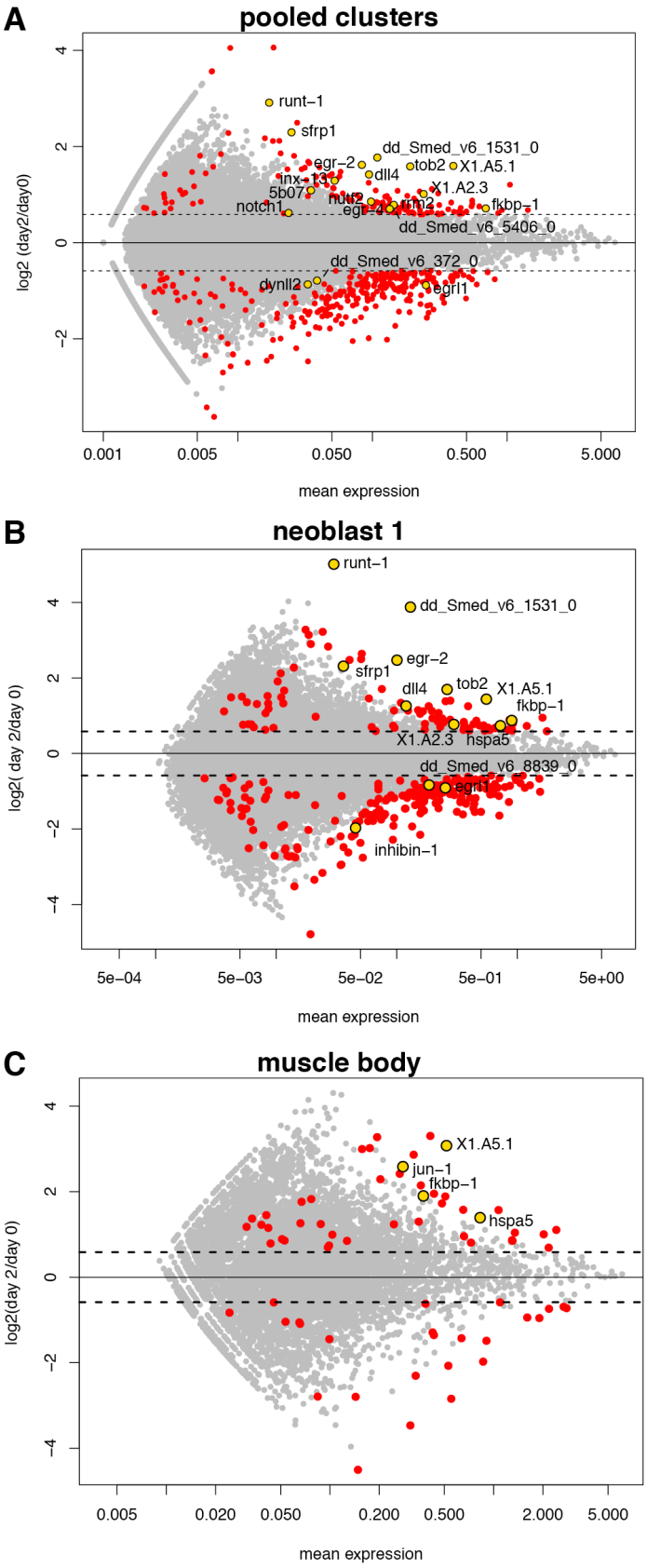

A
pooled clusters
log2 (day2/day0)
mean expression
runt−1
sfrp1
egr−2
dd_Smed_v6_1531_0
tob2
X1.A5.1
dll4
inx−13
5b07
X1.A2.3
notch1
fkbp−1
dd_Smed_v6_5406_0
dd_Smed_v6_372_0
dynll2
egrl1
B
neoblast 1
log2(day 2/day 0)
mean expression
runt−1
dd_Smed_v6_1531_0
egr−2
sfrp1
dll4
tob2
X1.A5.1
fkbp−1
X1.A2.3
hspa5
dd_Smed_v6_8839_0
egrl1
inhibin−1
C
muscle body
log2(day 2/day 0)
mean expression
X1.A5.1
jun−1
fkbp−1
hspa5

Fig. S15. Quantification of gene expression changes during regeneration

A-C. MA plots showing differential gene expression 2 days after regeneration compared to day 0 in pooled clusters (**A**), neoblast 1 (**B**) and muscle body (**C**), highlighting significantly differentially expressed genes (red points, log2FC > |0.58|, adjusted p-value < 0.05) and its overlap with wound-induced significant upregulated genes previously described (yellow dots).

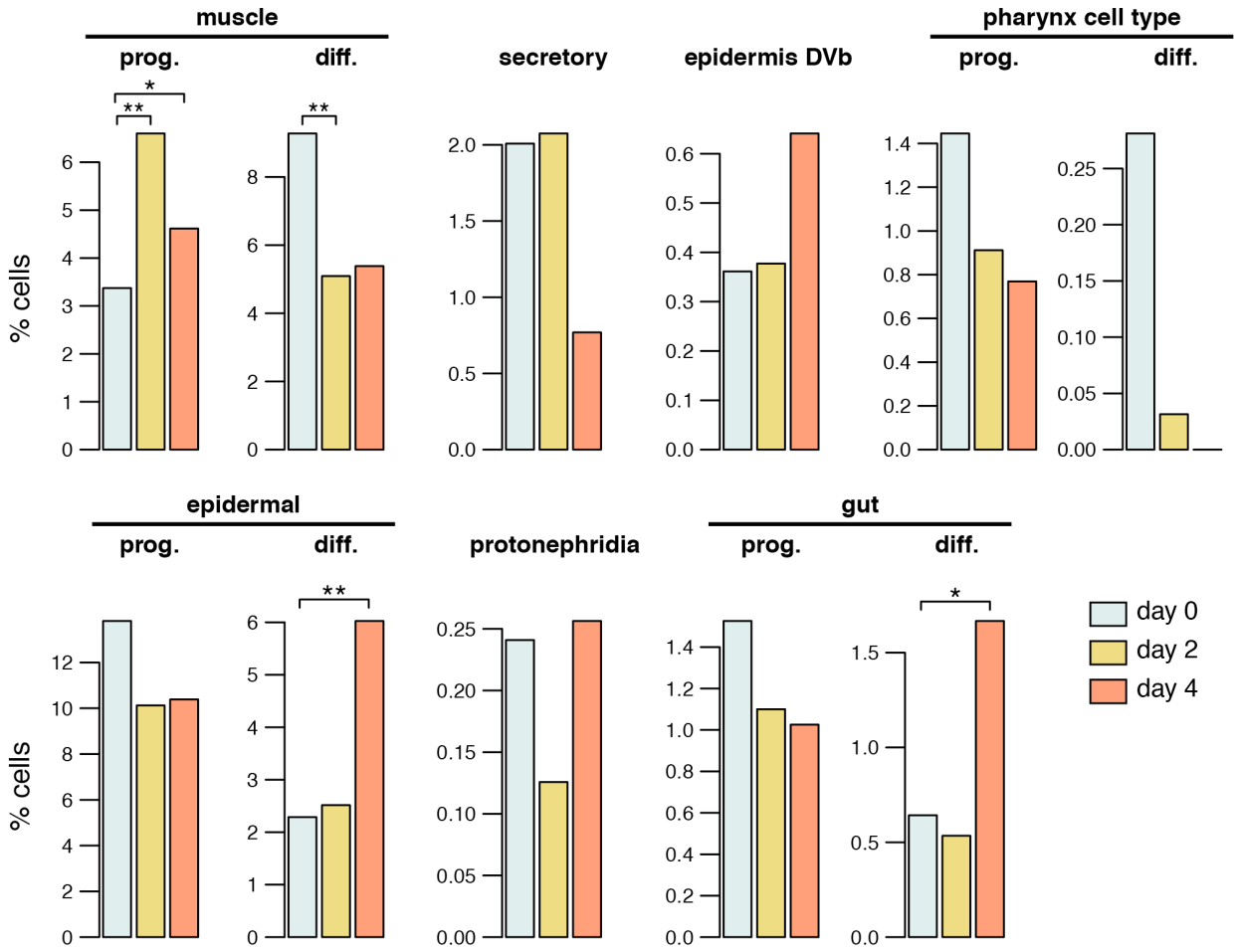


Fig. S16. Quantification of cellular abundances during regeneration

Quantification of progenitors, if defined, and differentiated cell populations of muscle, secretory, epidermis DVb, pharynx cell, epidermal, protonephridia and gut. Significant differences calculated using a fisher test with adjusted p-value < 0.05 or p-value < 0.001 are marked with * and ** respectively.

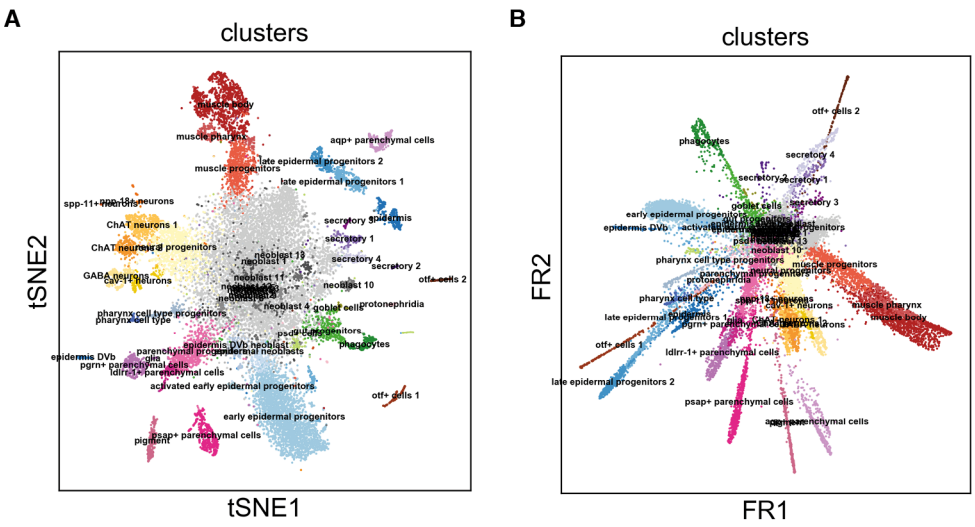


Fig. S17. Non-linear embeddings violate topological properties of high-dimensional data

tSNE **(A)** and force-directed drawing **(B)** of the single-cell graph using the Fruchterman-Reingold algorithm. The latter has only recently been suggested for visualizing single-cell data.

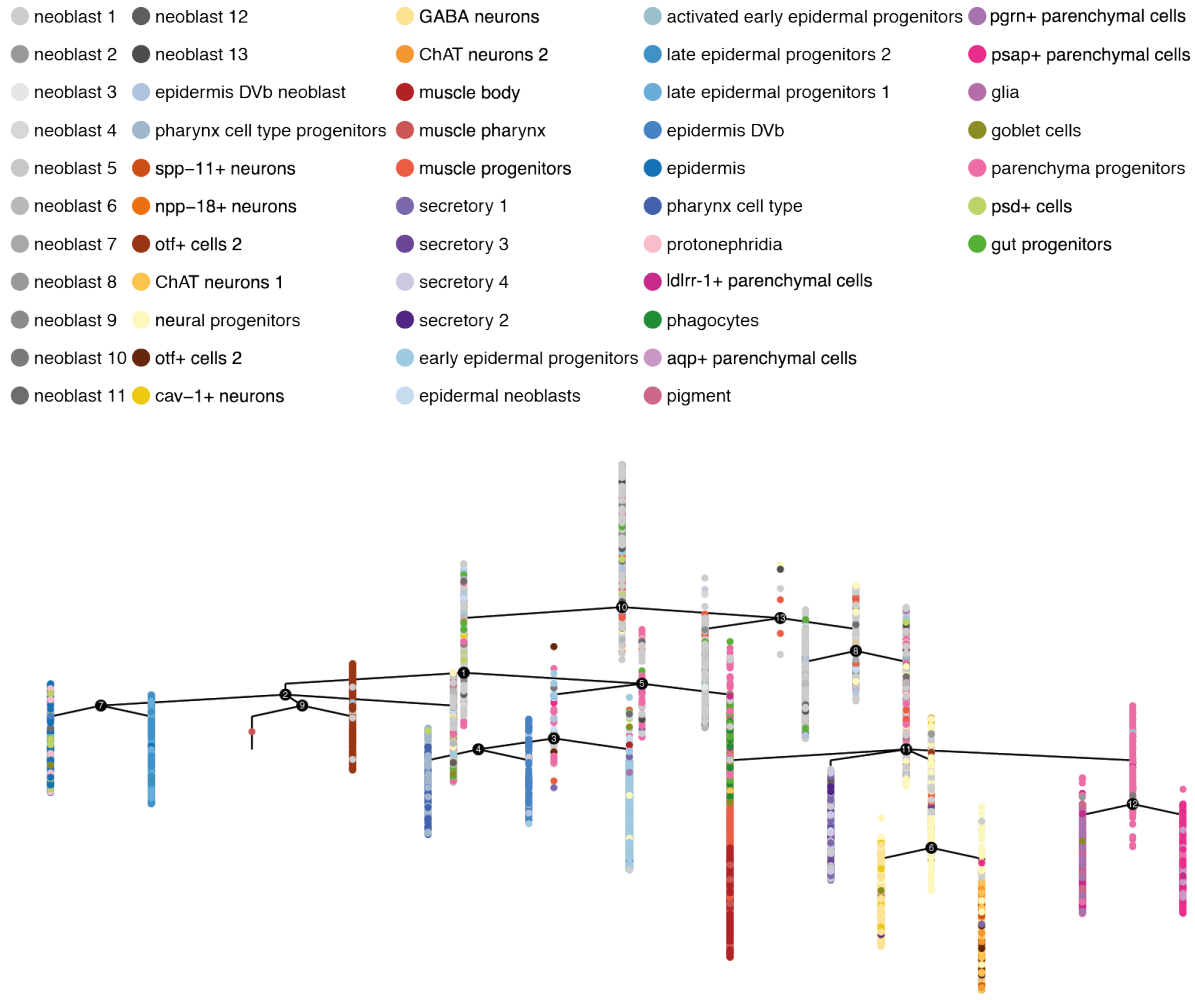


Fig. S18. Planaria lineage tree predicted by Monocle 2

The prediction of the planarian lineage tree shows qualitative inconsistencies compared to PAGA and does not confirms the previously described epidermal lineage.

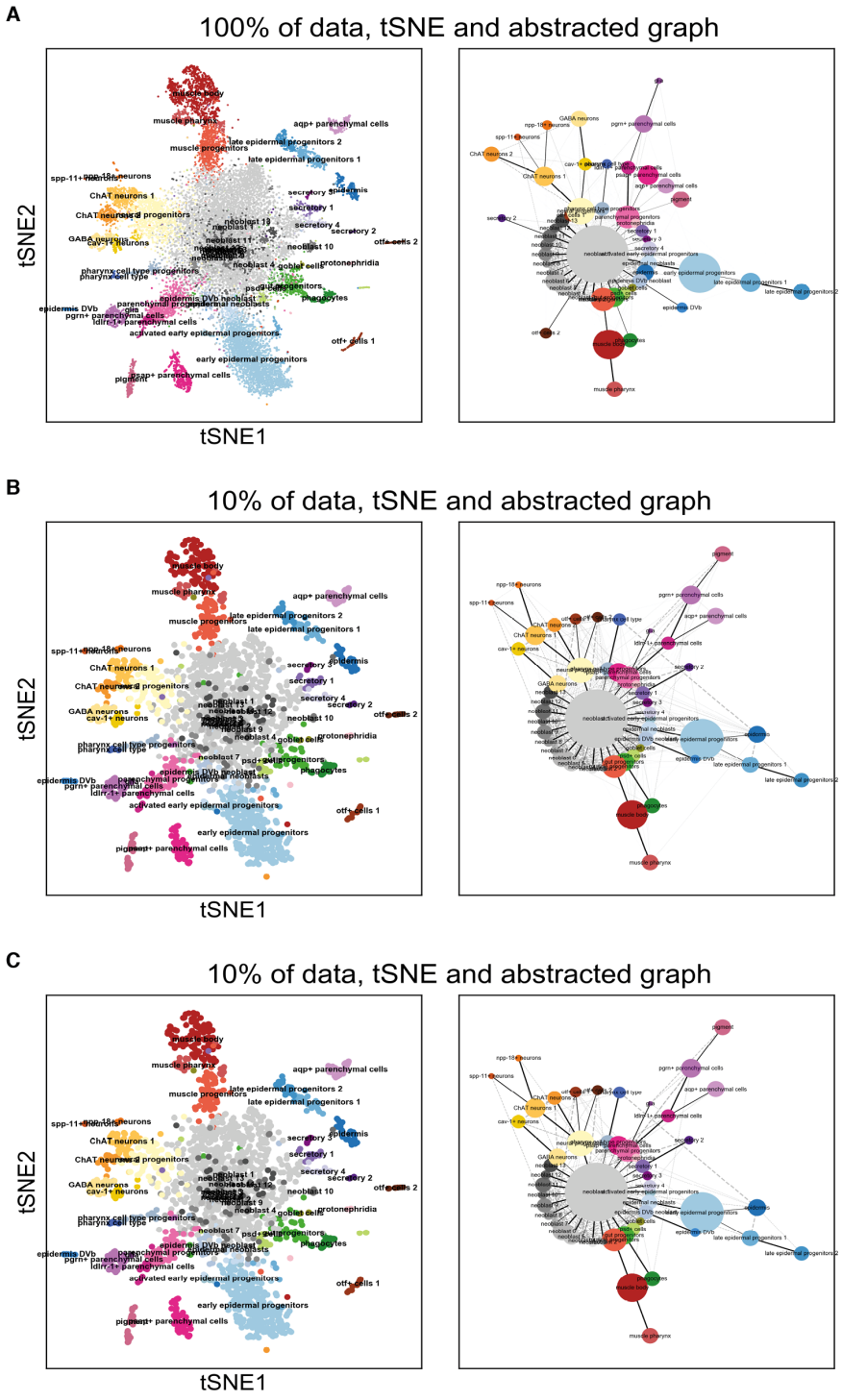


Fig. S19. PAGA robustness to subsampling

tSNE plots (left) and abstracted graphs (right) resulting from subsampling 100% **(A)**, 80% **(B)** and 10% **(C)** of the original cells.

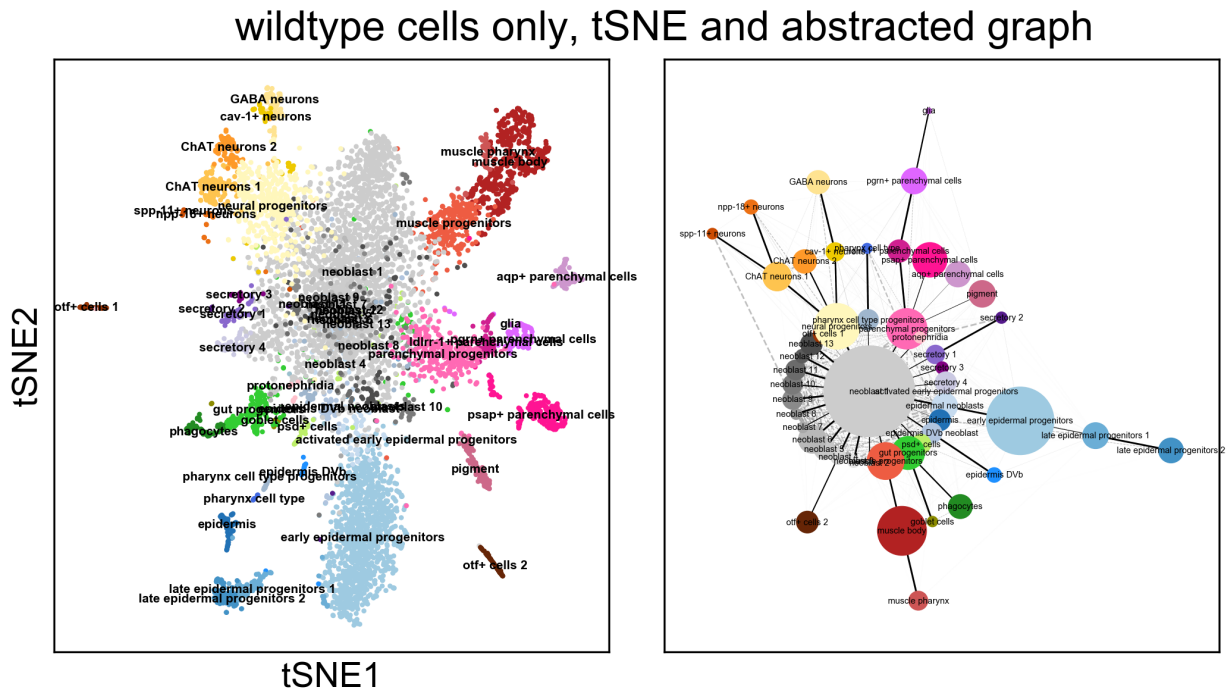


Fig. S20. Lineage reconstruction using only wild type samples

tSNE (left) and abstracted graph (right) obtained using only wild type samples.

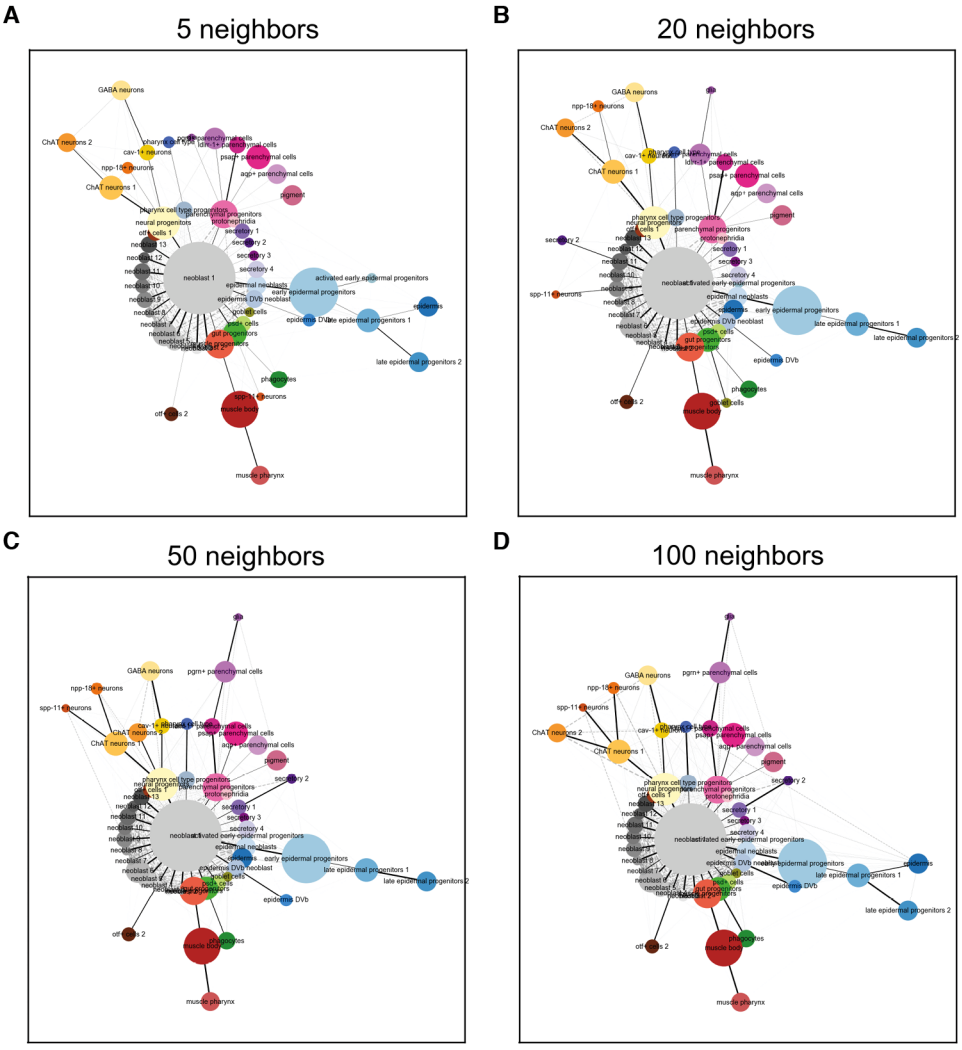


Fig. S21. Robust reconstruction for different numbers of nearest neighbors

Comparison of the effects of varying the number of nearest neighbors on the lineage reconstruction using PAGA. Abstracted graphs predicted using 5 **(A)**, 20 **(B)**, 50 **(C)** and 100 **(D)** neighbors are shown. All the abstracted graphs show a very similar topology compared to the one used in the analyses, which is done with 30 neighbors (Fig. S19A).

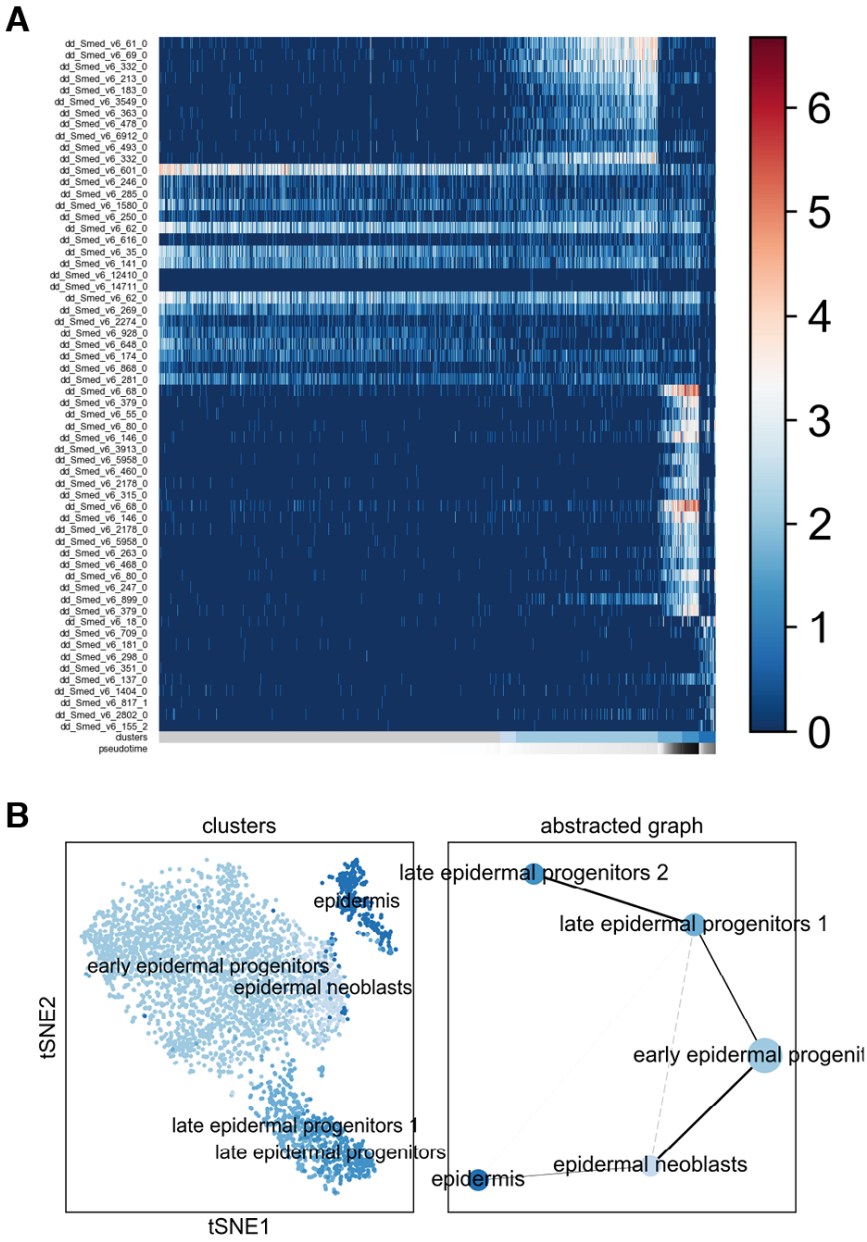


Fig. S22. Details of the epidermal lineage.

A. Log transformed raw expression of the 10 top-ranked markers of the 6 clusters in the epidermal lineage versus pseudotemporal order. **B.** tSNE (left) and abstracted graph (right) obtained with PAGA using only the clusters belonging to the epidermal lineage.

Additional Table S1

Statistics of Drop-seq samples

Additional Table S2

Number of cells per cluster

Additional Table S3

Marker genes identified for each cluster.

Additional Table S4

Gene sets identified using SOMs.

Additional Table S5

Examples of known genes included in the SOMs.

Additional Table S6

Comparison of log2FC in wound-induced genes from Wurtzel et al. 2015 and the regeneration samples.

Additional Table S7

Differentially expressed genes in regeneration day 2 vs day 0.

Additional Table S8

Differentially expressed genes in regeneration day 4 vs day 0.

Additional Table S9

List of primers used for *in situ* probes.

Additional Table S10

Published RNA-seq datasets used in the paper