Swapna *et al. Genome Biology* (2018) 19:124
https://doi.org/10.1186/s13059-018-1498-x

Genome Biology

RESEARCH Open Access

Comparative transcriptomic analyses and single-cell RNA sequencing of the freshwater planarian *Schmidtea mediterranea* identify major cell types and pathway conservation

Lakshmipuram Seshadri Swapna[1], Alyssa M. Molinaro[1,2], Nicole Lindsay-Mosher[1,2], Bret J. Pearson[1,2,3*] and John Parkinson[1,2,4*]

Abstract

Background: In the Lophotrochozoa/Spiralia superphylum, few organisms have as high a capacity for rapid testing of gene function and single-cell transcriptomics as the freshwater planaria. The species *Schmidtea mediterranea* in particular has become a powerful model to use in studying adult stem cell biology and mechanisms of regeneration. Despite this, systematic attempts to define gene complements and their annotations are lacking, restricting comparative analyses that detail the conservation of biochemical pathways and identify lineage-specific innovations.

Results: In this study we compare several transcriptomes and define a robust set of 35,232 transcripts. From this, we perform systematic functional annotations and undertake a genome-scale metabolic reconstruction for *S. mediterranea*. Cross-species comparisons of gene content identify conserved, lineage-specific, and expanded gene families, which may contribute to the regenerative properties of planarians. In particular, we find that the *TRAF* gene family has been greatly expanded in planarians. We further provide a single-cell RNA sequencing analysis of 2000 cells, revealing both known and novel cell types defined by unique signatures of gene expression. Among these are a novel mesenchymal cell population as well as a cell type involved in eye regeneration. Integration of our metabolic reconstruction further reveals the extent to which given cell types have adapted energy and nucleotide biosynthetic pathways to support their specialized roles.

Conclusions: In general, *S. mediterranea* displays a high level of gene and pathway conservation compared with other model systems, rendering it a viable model to study the roles of these pathways in stem cell biology and regeneration.

Keywords: *Schmidtea mediterranea*, Transcriptomics, Metabolism, Single-cell genomics, Metabolic reconstruction, Transcription factors, Tissue regeneration, Comparative genomics

* Correspondence: bret.pearson@utoronto.ca; john.parkinson@utoronto.ca
[1]Hospital for Sick Children, Toronto, ON, Canada
Full list of author information is available at the end of the article

Background

Investigations using model organisms such as *Caenorhabditis elegans*, *Drosophila melanogaster*, zebrafish, and mice continue to drive fundamental insights into the molecular mechanisms driving a variety of conserved biochemical processes [1]. However, much attention has recently turned to the use of non-traditional organisms as models to explore more specialized pathways. For example, while freshwater planarians (flatworms) have been used in a laboratory setting for more than 100 years due to their ability to regenerate following virtually any injury, the planarian *Schmidtea mediterranea* has emerged as a powerful model for dissecting the molecular basis of tissue regeneration [2, 3]. Despite significant resources put forth to develop *S. mediterranea* as a model in the lab, systematic genome-scale investigations of gene function and conservation are lacking.

Much of the interest in planarians is driven by the fact that approximately 20% of their adult cells are stem cells (called neoblasts), at least some of which are pluripotent [4–7]. In addition, planarians are one of the only models that can be used to rapidly test gene function in adult animals through RNA interference (RNAi) screening. Placing gene function in an evolutionary context is critical not only to inform on the conservation of pathways related to stem cell biology and regeneration, but also because planarians represent a key member of the otherwise neglected superphylum Lophotrochozoa/Spiralia (subsequently referred to as Lophotrochozoa), and they can further be used to model closely related parasitic flatworm species (e.g., flukes and tapeworms), which infect an estimated hundreds of millions worldwide [8].

In attempts to complement ongoing genome sequencing efforts [9, 10], several transcriptome datasets have been generated for *S. mediterranea* under various physiological conditions using a variety of experimental techniques [11–18]. In isolation, each set provides a snapshot of planarian gene expression under a specific condition; however, recent efforts have focused on integrating several transcriptomes to generate a more comprehensive overview of gene expression [9, 19]. The SmedGD repository was generated by integrating transcriptomes from whole-animal sexual and asexual worms, whereas the PlanMine database serves as a repository for the published genome as well as existing transcriptomes from the community to be deposited and queried. However, they lack systematic and comparative evolutionary and functional genomics analyses, which are required for understanding the mechanistic basis of biological processes. Together these datasets comprise more than 82,000 “transcripts” with little assessment of “completeness” from an evolutionary perspective.

Typically, transcriptome datasets are generated from entire organisms or tissues [20–22]; however, such analyses can mask the contribution of specific cell subpopulations, which can be particularly problematic when attempting to elucidate, for example, pathways expressed during key cellular events. While cell sorting offers the capability to enrich for specific cell subpopulations, the emergence of single-cell RNA sequencing (scRNAseq) offers a powerful route for interrogating gene expression profiles from individual cells [23, 24]. Applied to *S. mediterranea*, this technology is expected to yield molecular-level insights into the roles of distinct cell types, such as neoblasts, during homeostatic tissue maintenance and regeneration [7, 25–27]. Indeed, scRNAseq experiments have already been used to resolve neoblast heterogeneity and identify regulators of lineage progression [26–30].

In this study, we generate a high-confidence transcriptome pruned from an integrated transcriptome generated earlier in the lab [18], which, through combining transcriptomes from diverse physiological conditions and experimental techniques, leads to a large number of transcripts ($n = 83{,}469$) for *S. mediterranea*. Next, we apply systematic bioinformatic approaches to annotate and compare the complement with model organisms and other Platyhelminthes. This pipeline predicts putative functional annotations of the transcriptome, identifying a set of transcriptionally active transposons as well as extended families of cadherins and tumor necrosis factor (TNF) receptor associated factor (TRAF) proteins. Metabolic reconstruction further reveals an increased biochemical repertoire relative to related parasitic platyhelminths. In order to gain insights into the role of these pathways in planarian biology, high-throughput scRNA-seq was performed, capturing the transcriptional signatures from ~ 2000 cells. From the 11 distinct clusters of transcriptional profiles, we identified clusters corresponding to neoblasts, epithelial progenitors, muscle, neurons, and gut, among which neoblasts exhibit the most metabolically active profiles. We also identify a novel cluster: a *cathepsin*$^+$ cluster representing multiple unknown mesenchymal cells. Beyond giving us new insights into the evolution and dynamics of genes involved in regenerative pathways, the data and analyses presented here provide a complementary resource to ongoing genome annotation efforts for *S. mediterranea*. They are available for download from http://www.compsysbio.org/datasets/schmidtea/.

Results

A definitive transcriptome for *S. mediterranea*

A definitive transcriptome of *S. mediterranea* was generated by integrating the RNA sequencing (RNA-seq) reads generated from five separate experiments and cell purifications [18, 31–33] (National Center for Biotechnology Information [NCBI] Bioproject PRJNA215411). From an initial set of 83,469 transcripts, a tiered set of filters were applied to define a single set of 36,026 high-confidence transcripts (Fig. 1a). First, protein-coding transcripts are identified on the basis of sequence similarity to known transcripts or proteins, as well as the presence of predicted protein domains with reference to the following databases: UniProt [34], MitoCarta [35], InterPro [36], Core Eukaryotic Genes Mapping Approach (CEGMA) [37], Benchmarking Universal Single-Copy Orthologs (BUSCO) [38], and ESTs of other known platyhelminth transcriptomes deposited in the expressed sequence tag (EST) database of the NCBI: *Biomphalaria glabrata*, *Clonorchis sinensis*, *Crassostrea gigas*, *Dugesia japonica*, *Dugesia ryukyuensis*, *Echinococcus granulosus*, *Echinococcus multilocularis*, *Helobdella robusta*, *Hirudo medicinalis*, *Hymenolepis microstoma*, *Macrostomum lignano*, *Mytilus californianus*, *Opisthorchis viverrini*, *Schistosoma japonicum*, *Schistosoma mansoni*, *Taenia solium*.

Next, the protein-coding potential of the remaining transcripts was predicted using the error-tolerant ESTScan [39]. Finally, transcripts without matches to the above were parsed through a six-frame translation algorithm to identify the largest potential open reading frame (LongestORFs). ESTScan and LongestORFs predictions were further filtered such that only those predicted to have > 100 amino acid residues and also to co-localize on the genome with known *S. mediterranea* transcripts derived from complementary resources (EST database of the NCBI, SmedGD v2.0 [9] and the Oxford dataset [14]) were included in our final filtered dataset (Fig. 1a, b).

Together, this filtered set comprises 36,026 sequences, of which 28,583 map to 22,215 loci of the *S. mediterranea* genome assembly deposited in SmedGD v2.0 [9]; the remaining 7443 sequences could not be mapped. Of these unmapped transcripts, 1008 share significant sequence similarity, i.e., ≥ 80% sequence identity as assigned by the Basic Local Alignment Search Tool (BLAST) [40], with a known *S. mediterranea* protein, 106 to a protein from the closely related planarian *D. japonica*, and 65 to proteins from other Platyhelminthes. Such matches indicate that these sequences are likely bona fide transcripts that are missing from the current *S. mediterranea* genome assembly. Interestingly, among the 7443 unmapped transcripts, we also identified 794 with significant sequence identity (≥80% sequence identity as assigned by BLAST) to a non-metazoan protein in the UniProt database. Among these were 728 sequences matching sequences from *Tetrahymena thermophila* and a further 22 matching sequences from *T. pyriformis*. Such sequences likely indicate contaminants from protozoa endemic in *S. mediterranea*

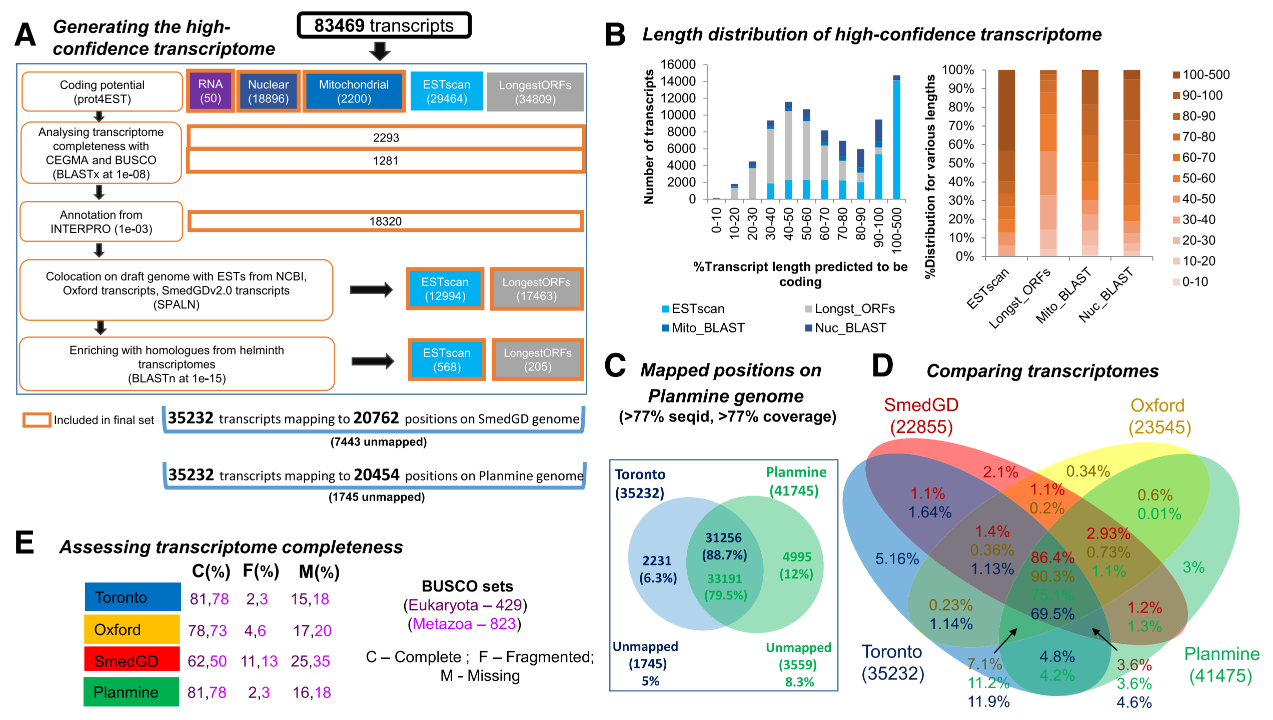


Fig. 1 Transcriptome generation and characteristics. **a** Schematic of the tiered approach used for generating the definitive transcriptome. **b** Length distribution of the transcripts generated by different methods. **c** Venn diagram showing the results for the mapping of Toronto and PlanMine transcripts onto the recent dd_Smes_g4 genome assembly. **d** Venn diagram showing the comparison of Toronto, PlanMine, SmedGD, and Oxford transcriptomes, where the transcripts are aligned using BLASTn searches customized for sensitive matches. **e** Transcriptome completeness for Toronto, PlanMine, SmedGD, and Oxford transcriptomes, estimated via CEGMA and BUSCO core eukaryotic gene sets

cultures. Further, 2 transcripts sharing ≥ 80% sequence identity to *Bos taurus* were also removed. After removal of these contaminants, we identified a final high-quality set of 35,232 transcripts, which we subsequently termed the Toronto transcriptome (Additional file 1).

Aligning the Toronto transcriptome with the recently published reference genome of *S. mediterranea* (dd_Smes_g4) [10] and applying the F1 cutoff defined by the Spaln alignment tool (corresponding to ~ 73% sequence identity and ~ 73% coverage) [41] resulted in mapping 33,487 transcripts (~ 95% of the transcriptome) to 20,483 genomic positions (Fig. 1c, Additional file 2: Figure S1A). In contrast, using similar parameters resulted in the mapping of 38,186 PlanMine transcripts (~ 91.5% of the transcriptome) to 26,510 positions. Of these, 31,286 (~ 89%) Toronto transcripts overlap with 33,191 PlanMine transcripts (79.5%), corresponding to 14,145 positions. Although both transcriptomes map a substantial proportion of their transcriptomes to the reference genome, PlanMine maps a higher number of transcripts. However, it is noteworthy that the Toronto transcriptome contributes 2231 transcripts (~ 6%) that exclusively map to the reference genome. Interestingly, while PlanMine and Toronto transcripts that map to the same loci are of similar length, PlanMine transcripts that are either unmapped or map to unique regions are significantly longer than the equivalent Toronto transcripts (Additional file 2: Figure S1B). Analyzing the distribution of sequence similarity bit scores further reveals that the unmapped transcripts from both the Toronto and PlanMine transcriptomes consist of many high-scoring matches, suggesting their likely validity (Additional file 2: Figure S1C).

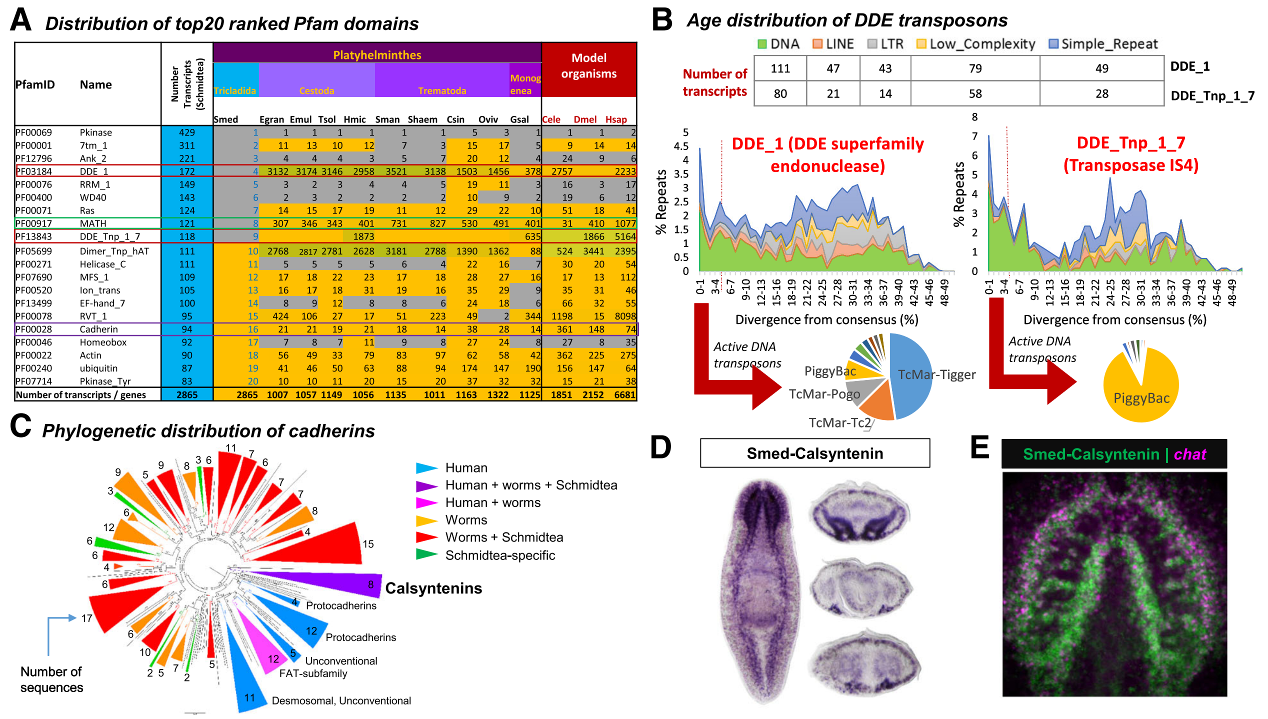
Comparisons with three previously generated transcriptomes: SmedGD v2.0 ($n = 22{,}855$, [9]), PlanMine ($n = 41{,}475$, [19]), and Oxford ($n = 23{,}545$, [14]), revealed a core set of 24,477 transcripts common to all four sets, together with 1820 transcripts unique to the Toronto set (defined as those with bit score < 40 for BLASTn [40] searches using a relaxed word size of 7 in order to maximize sensitivity); Fig. 1d). Of the unique transcripts, 371 (20.3%) share significant sequence similarity (BLAST, E-value <1e-08, % sequence identity ranging from 1.5% to 100%) to known proteins in UniProt and 1427 (78%) represent ESTScan predictions. Supporting the validity of these unique transcripts, we note that 1399 (~ 74%) map to the latest PlanMine genome dd_Smes_g4 [10]. To further assess transcriptome completeness, we performed a systematic comparison with the core eukaryotic and metazoan gene sets defined by BUSCO v1 [38], demonstrating that our high-quality transcriptome exhibits similar coverage (81% eukaryotic, 78% metazoan) as PlanMine (81% eukaryotic, 78% metazoan) and higher coverage than the Oxford (78% eukaryotic, 73% metazoan) and SmedGD (62% eukaryotic, 50% metazoan) datasets (Fig. 1e). Additionally, the Toronto transcriptome features a lower fraction of partially recovered transcript sets. However, it is noteworthy that of the 348 BUSCO genes, representing single-copy genes from 310 different eukaryotes that were *completely* recovered by the Toronto dataset, 86 appear to possess paralogs in the Toronto dataset as compared to 112 in PlanMine. Such duplicates might represent either errors during transcript assembly or alternative spliceoforms.

Functional annotation of *S. mediterranea* proteome: expanded set of transposons and TRAFs

Having compiled and validated a high-confidence set of transcripts, we next analyzed functional potential through a systematic annotation of protein domains inferred by the InterPro resource [36]. Gene Ontology (GO) assignments [42, 43] based on domain annotations of predicted proteins revealed that transport, signal transduction, biosynthetic process, cellular nitrogen compound metabolic process, and cellular protein modification process are the five most abundant biological processes, consistent with other eukaryotes (Additional file 2: Figure S2).

To identify taxon-specific gene family expansions in *S. mediterranea*, we compared the 20 most abundant Pfam [44] annotations of predicted protein sequences in our dataset to the proteomes of *Homo sapiens*, *Drosophila melanogaster*, and *Caenorhabditis elegans*, as well as several parasitic flatworms for which genome sequence data are available (cestodes: *E. granulosus*, *E. multilocularis*, *T. solium*, *H. microstoma*; trematodes: *Schistosoma mansoni*, *S. haematobium*, *C. sinensis*, *O. viverrini*; monogeneans: *Gyrodactylus salaris*) (Fig. 2a). Consistent with the other metazoans, the most abundant domains are Pkinase (PF00069), 7tm (PF00001), and Ank (PF12796). Among the remaining 17 abundant domains, three represent lineage-specific expansions: transposase-related domains, DDE_1 (PF03184) and DDE_Tnp_1_7 (PF13843) (ranked 4th and 9th most abundant, respectively) — which are significantly expanded only in *S. mediterranea* and not in other Platyhelminthes — and the meprin and TRAF homology (MATH) domain (PF00917, ranked 8th most abundant) — expanded in *S. mediterranea* in comparison to other Platyhelminthes. Another domain of interest is the cadherin domain (PF00028, ranked 16th most abundant), which is expanded throughout Platyhelminthes and also in humans, suggesting a more fundamental role for this domain.

Although *S. mediterranea* exhibits a larger ($n = 290$) repertoire of the transposase-related domains, DDE_1 and DDE_Tnp_1_7, relative to other helminths (Fig. 2a), the transcripts associated with these domains are expressed at relatively low levels: mean reads per kilobase per million mapped reads (RPKM) 1.22 +/− 0.04 and 1.10 +/− 0.42 for DDE_1 and DDE_Tnp_1_7,



A *Distribution of top20 ranked Pfam domains*

PfamID	Name	Number Transcripts (Schmidtea)	Platyhelminthes										Model organisms		
			Tricladida	Cestoda				Trematoda				Monogenea			
			Smed	Egran	Emul	Tsol	Hmic	Sman	Shaem	Csin	Oviv	Gsal	Cele	Dmel	Hsap
PF00069	Pkinase	429	1	1	1	1	1	1	1	5	3	1	1	1	2
PF00001	7tm_1	311	2	11	13	10	12	7	3	15	17	5	9	14	14
PF12796	Ank_2	221	3	4	4	4	3	5	7	20	12	4	24	9	6
PF03184	DDE_1	172	4	3132	3174	3146	2958	3521	3138	1503	1456	378	2757		2233
PF00076	RRM_1	149	5	3	2	3	4	4	5	19	11	3	16	3	17
PF00400	WD40	143	6	2	3	2	2	2	2	10	9	2	19	6	12
PF00071	Ras	124	7	14	15	17	19	11	12	29	22	10	51	18	41
PF00917	MATH	121	8	307	346	343	401	731	827	530	491	401	31	410	1077
PF13843	DDE_Tnp_1_7	118	9				1873					635		1866	5164
PF05699	Dimer_Tnp_hAT	111	10	2768	2817	2781	2628	3181	2788	1390	1362	88	524	3441	2395
PF00271	Helicase_C	111	11	5	5	5	5	6	4	22	16	7	30	20	54
PF07690	MFS_1	109	12	17	18	22	23	17	18	28	27	16	17	13	112
PF00520	Ion_trans	105	13	16	17	18	31	19	16	35	29	9	35	31	46
PF13499	EF-hand_7	100	14	8	9	12	8	8	6	24	18	6	66	32	55
PF00078	RVT_1	95	15	424	106	27	17	51	223	49	2	344	1198	15	8098
PF00028	Cadherin	94	16	21	21	19	21	18	14	38	28	14	361	148	74
PF00046	Homeobox	92	17	7	8	7	11	9	8	27	24	8	27	8	35
PF00022	Actin	90	18	56	49	33	79	83	97	62	58	42	362	225	275
PF00240	ubiquitin	87	19	41	46	50	63	88	94	174	147	190	156	147	64
PF07714	Pkinase_Tyr	83	20	10	10	11	20	15	20	37	32	32	15	21	38
Number of transcripts / genes		**2865**	**2865**	**1007**	**1057**	**1149**	**1056**	**1135**	**1011**	**1163**	**1322**	**1125**	**1851**	**2152**	**6681**

B *Age distribution of DDE transposons*

	🟩 DNA	🟧 LINE	⬜ LTR	🟨 Low_Complexity	🟦 Simple_Repeat	
Number of transcripts	111	47	43	79	49	DDE_1
	80	21	14	58	28	DDE_Tnp_1_7

C *Phylogenetic distribution of cadherins*

D

E

Fig. 2 Abundant Pfam families. **a** Comparative distribution of top 20 Pfam families in *S. mediterranea* and the ranks of these families in model organisms and closely related Platyhelminthes in terms of their abundance. The total number of transcripts for each of the species in these 20 families is indicated in the *last row*. Pfam families of particular interest are indicated in *boxes*. **b** Age distribution of DDE transposons: the classification and distribution of repeat elements in transcripts of the highly abundant DDE Pfam families is shown, along with a representation of the extent of sequence divergence of these elements from their consensus. **c** Phylogenetic distribution of cadherins from human, *C. elegans*, Platyhelminthes, and *S. mediterranea*. Clades with bootstrap support of > 600/1000 are *collapsed* and *colored* by the taxonomic representation of the species in each clade, while the number of transcripts mapping to the clade is indicated at the *edge*. **d** Whole-mount in situ hybridization of *Smed-calsyntenin*. Cross sections (*right*) are from anterior (*top*), pharyngeal (*middle*), and tail (*bottom*) regions. **e** dFISH of a single confocal slice through the brain demonstrating co-expression of *chat* in *Smed-calsyntenin* + neurons

respectively; bottom 40% of expressed transcripts (Additional file 1). Transposable elements (TEs, sequences which can change position within a genome) are classed into two types: class I (retrotransposons), which operate via a copy-and-paste mechanism and include long and short interspersed nuclear elements (LINEs and SINEs, respectively), and class II (DNA transposons), which operate via a cut-and-paste mechanism [45]. DNA transposons are the most abundant elements for transcripts with both DDE_1 and DDE_Tnp_1_7 domains. To determine whether these elements may be functionally active in the *S. mediterranea* genome, we estimated the sequence divergence of each copy relative to the consensus (Fig. 2b, [46]). Of 1641 elements, we found that 180 (13%) of DDE_1 domains and 97 (25%) of DDE_Tnp_1_7 domains exhibit relatively low sequence divergence (< 5%), indicating that they may still be functionally active. Among DDE_1 domain transcripts, almost half represent the TcMar-Tigger element, thought to be a distant relative of Mariner [47], while for DDE_Tnp_1_7 domain transcripts, the majority represent the PiggyBac element.

Beyond transposons, we found that the MATH (121 domains) domain represents *S. mediterranea*-specific expansions. MATH domains are present in mammalian tissue-specific metalloendopeptidases (meprins) and TNF receptor associated factor (TRAF) proteins. BLAST searches of MATH-domain-associated proteins in *S. mediterranea* suggest they are likely TRAF proteins (Additional file 1), important regulators of signal transduction, cell death, and cellular responses to stress [48], immune response [49], and cellular degradation [50]. Many of these domains contain transcripts that are expressed at relatively high levels (mean RPKM 18.05 +/− 5.39; top 20% of expressed transcripts; Additional file 1), suggesting an important regulatory role. Another gene family with abundant representation in Platyhelminthes is the cadherins. Cadherin-domain-containing transcripts were moderately expressed (mean RPKM 4.22 +/− 1.26; top 40% of expressed transcripts; Additional file 1). Cadherins are transmembrane proteins involved in regulating cell-cell adhesion, morphogenesis, and cell recognition [51, 52]. More than 100 cadherins have been characterized in vertebrates, belonging to four main classes [51]: classical (localized to different tissues), desmosomal, protocadherins (protocadherins and FAT subfamily of cadherins), and unconventional. A phylogenetic analysis of the 94 cadherins in *S. mediterranea* with 176 human and 211 other helminth sequences (from *C. elegans*, *E. granulosus*, *E. multilocularis*, *G. salaris*, *Hymenoloepis nana*, *S. haematobium*, *S. mansoni*, *T. solium*, *O. viverini*, and *C. sinensis*) recapitulates three of the main human clusters (desmosomal and unconventional cadherins, protocadherins (one main and one subcluster), and FAT subfamily of protocadherins (which also includes homologs in worms), as well as 8 clusters specific to other helminths, 16 clusters containing other helminths, and *S. mediterranea* sequences, 5 *Schmidtea*-specific clusters, and 1 cluster containing human, other helminths, and *S. mediterranea* sequences (Fig. 2c, Additional file 2: Figure S3). This latter cluster corresponds to calsyntenins (CLSTN), calcium-binding type I transmembrane proteins belonging to the cadherin superfamily, predominantly expressed in neurons. This cluster contains sequences from human (CLSTN1, CLSTN2), *C. elegans* (CASY-1), *C. sinensis*, *O. viverini*, and *S. mediterranea* (Smed-calsyntenin - SmedASXL_013539). Consistent with its expression in neurons in other organisms, *Smed-calsyntenin* is predominantly expressed in the brain and ventral nerve cords (with weaker expression detected in the gut), and it exhibits a high degree of co-localization with the cholinergic neuron marker *chat* (Fig. 2d, e). In the future it will be interesting to determine whether the expansion of TRAF proteins in comparison to other parasitic flatworms and the abundance of cadherins in *S. mediterranea* represent increased functional complexity in signal transduction and regeneration in planarians.

S. mediterranea expresses a diverse repertoire of transcription factors

We next investigated the repertoire of transcription factors in *S. mediterranea* in the context of other eukaryotes. Transcription factors were predicted for *S. mediterranea*, together with an additional 165 eukaryotes [53]. Our predictions suggest that 843 *S. mediterranea* transcripts encode transcription factors associated with 55 classes (Fig. 3a, Additional file 3); 494 (~ 59%) belong to six classes (zf-C2H2, Homeobox, zf-BED, bZIP_1, bZIP_2, and HLH), which are typically well represented across all eukaryotes. The number of predicted transcription factors in *S. mediterranea* ($n = 843$) is slightly higher than in other Lophotrochozoans ($n = 672$) or nematodes ($n = 725$), and is half the number in vertebrates ($n = 1866$) or mammals ($n = 1786$). Although several classes of transcription factors, such as Forkhead, Ets, Pax, Pou, and

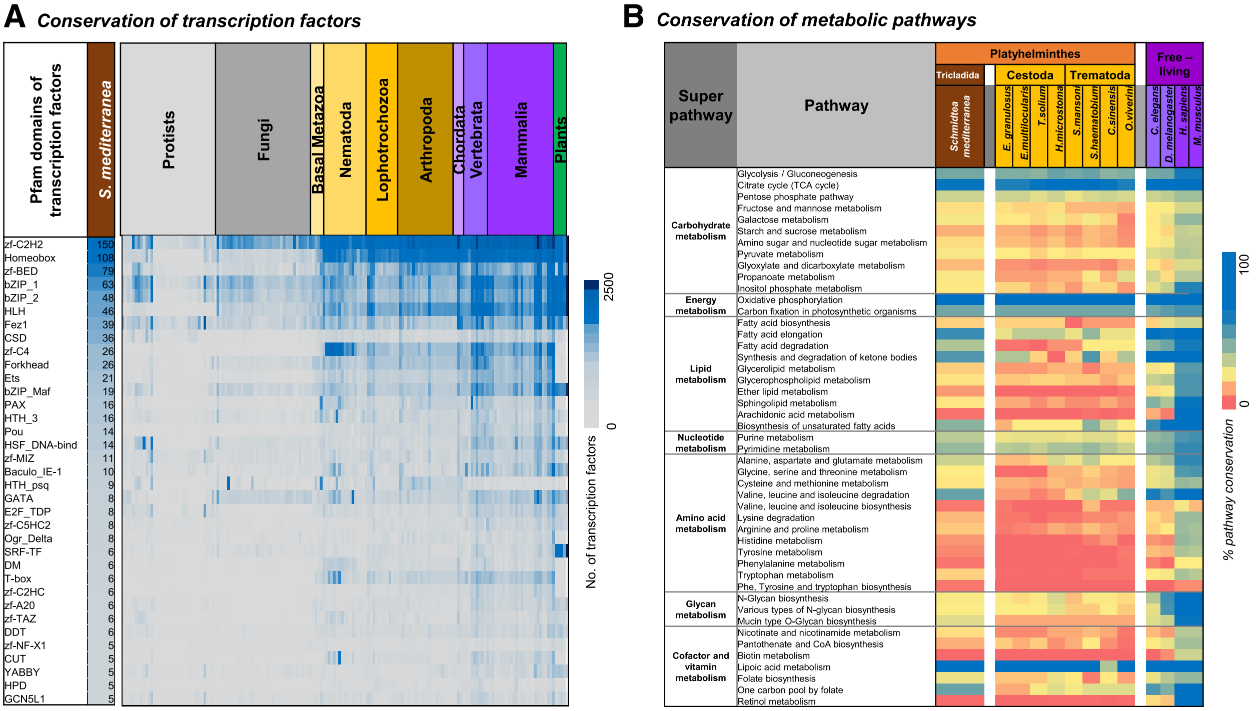


Fig. 3 Evolutionary conservation of metabolic and regulatory pathways. **a** Heatmap showing the distribution of different types of predicted transcription factors in *S. mediterranea* and eukaryotes from PhyloPro v2. The *numbers* represented are for protein predictions of transcripts in the definitive transcriptome for *S. mediterranea* and proteins for all other eukaryotes. **b** Heatmap showing the distribution of % conservation of the major classes of KEGG metabolic pathways in *S. mediterranea*, human, mouse, *C. elegans*, and closely related Platyhelminthes

GATA, have been studied in *S. mediterranea* [54, 55], several others with high abundances in *S. mediterranea* and vertebrates remain poorly characterized. These include CSD (cold-shock domain; involved in transcriptional repression and activation and in mRNA packaging, transport, localization, masking, stability, and translation) and bZIP_maf (acting as key regulators of terminal differentiation in many tissues, such as bone, brain, kidney, lens, pancreas, and retina, as well as in blood). These transcription factors have not been studied in *S. mediterranea* and are likely to be important candidates in the function of specific cell types.

Two types of transcription factors found in 75% of eukaryotic species listed in the comparative genomics resource PhyloPro v2 [53] were not predicted in *S. mediterranea*: AF-4 (a transcriptional activator that has previously been implicated in childhood lymphoblastic leukemia, mental retardation, and ataxia [56]) and Myc_N (a leucine zipper-type transcription factor implicated in cell cycle progression, cell death, and transformation). The loss of this latter transcription factor in particular suggests that planarians may have adopted an alternate mechanism of regulating Myc's canonical roles in cell proliferation and cell death.

Metabolic reconstruction reveals biochemical pathways distinct from those of parasitic helminths

Genome-scale metabolic reconstructions provide a powerful route to interrogate the metabolic capabilities of an organism [57–60]. Here we applied an integrated pipeline, developed in house (see Methods), to compare the metabolic potential of *S. mediterranea* with those derived from other helminths, human, and mouse (Fig. 3b). Among notable pathways present in *S. mediterranea* but not in other platyhelminths are several involved in fatty acid metabolism, branched chain amino acid metabolism, mucin-type O-glycan biosynthesis, and one carbon pool by folate. The loss of pathways involved in fatty acid metabolism in the parasitic flatworms may reflect their largely parasitic lifestyles. For example, schistosomes and cyclophyllidean tapeworms spend much of their life cycle in glucose-rich environments (blood and small intestine, respectively) and may therefore have adapted their metabolism to optimize glucose and glycogen as main sources of energy rather than lipids [61, 62]. Focusing on amino acid pathways, *S. mediterranea* displays similar auxotrophies as other helminths; however, a notable exception is branched chain amino acid degradation, which is largely absent from other platyhelminths yet appears to function in *S. mediterranea.* Conservation of this pathway was surprising given its role in longevity in *C. elegans*, because *S. mediterranea* exhibits no evidence of aging and is believed to be immortal [63]. Beyond core metabolic processes, *S. mediterranea* appears unique among platyhelminths in possessing enzymes required for the production of core 1 mucin-type O-glycans. Such production is likely related to the formation of the mucous secretions that coat the planarian, enabling locomotion, predation, innate immunity, and substrate adhesion [64]. Finally, our comparisons report the presence of several enzymes required for folate interconversion which are otherwise absent in parasitic flatworms. These interconversions provide additional routes for the production of various folate intermediates that are used as co-factors in a variety of metabolic processes, such as tetrahydrofolates involved in nucleotide and amino acid biosynthesis [65].

Spatial annotation of *S. mediterranea* transcripts by whole-animal scRNAseq

In order to place the annotated transcriptome data in the context of different tissues, the functional information of these transcripts was integrated with spatial information derived from single-cell RNA sequencing (scRNAseq) data of dissociated planarians obtained using Drop-seq technology [66]. The scRNAseq data consist of 51,563 transcripts expressed in 2000 cells. Pruning this dataset to only consider transcripts from our definitive set resulted in a set of 25,168 transcripts expressed in 2000 cells. The R package Seurat [67], which uses an unsupervised clustering approach by combining dimensional reduction with graph-based clustering, was used to cluster the data and discover cell types and states. Based on the set of most variable transcripts in the dataset ($n = 4586$), Seurat clusters 1195 of the 2000 cells into 11 clusters (Fig. 4a). It is noteworthy that clustering based on the larger set of 51,563 transcripts identified as expressed in the cells recapitulated a similar clustering pattern. Clusters were found to correspond to specific tissues based on the expression of previously described tissue-specific genes (Fig. 4b). In this way, clusters representing epithelial, neural, gut, muscle, parapharyngeal, and stem cells (neoblasts) were identified. Four clusters could not be identified based on previously published planarian gene expression data; however, two of these clusters displayed high expression of the cathepsin homolog *Smed-CTSL2* and were thus named cathepsin+ a and cathepsin+ b (Fig. 4c). Cluster 11 displayed enriched expression of *Smed-egr-5* and is therefore likely an epithelial subtype (discussed further below; see Fig. 5). Cluster 1 was not specifically enriched for any markers and displayed scattered expression of both neoblast and differentiated tissue markers (Fig. 4b). Its central location on the t-distributed stochastic neighbor embedding (t-SNE) plot, linking the neoblast cluster to the various tissue clusters, led us to conclude that Cluster 1 likely represents transient cell states as neoblasts differentiate along different lineages, and this idea is consistent with recently published scRNAseq studies [29, 30].

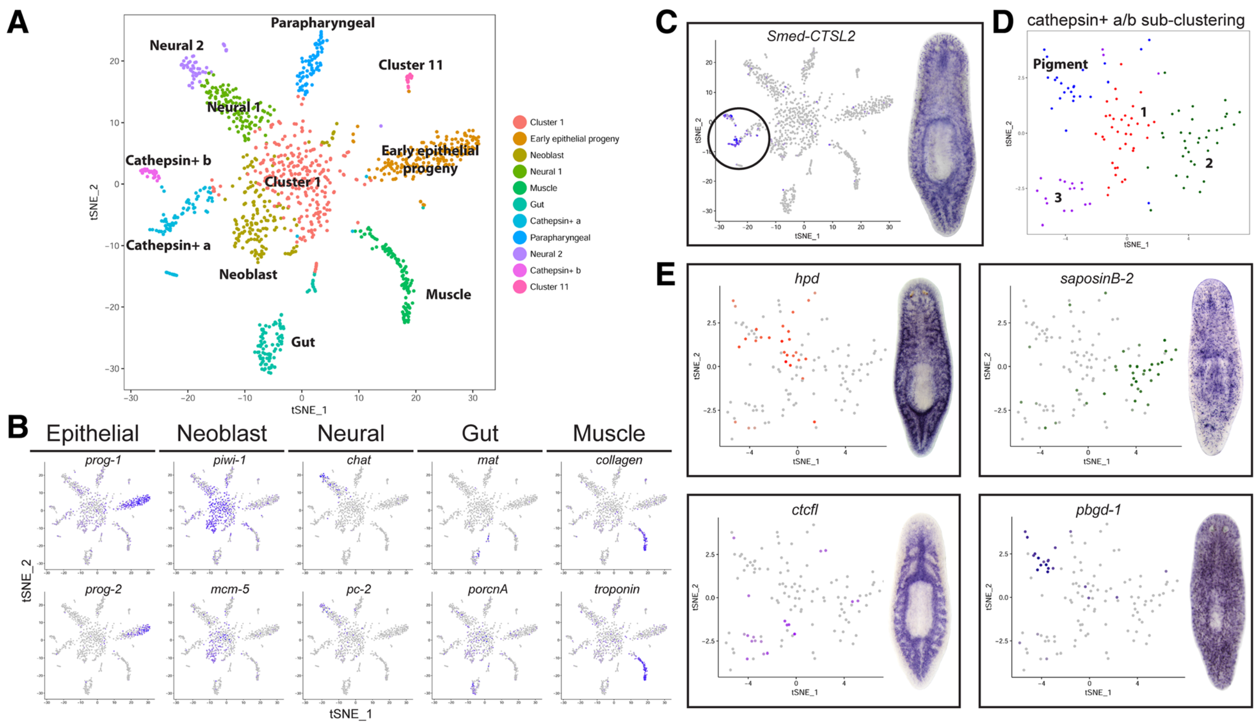


Fig. 4 Cluster separation and identification from scRNAseq data. **a** t-SNE plot of major cell clusters identified by Seurat. **b** Clusters corresponding to epithelial progenitors, neoblasts, neurons, gut, and muscles were identified based on the expression of known tissue-specific markers. **c** t-SNE plot and in situ hybridization for the cathepsin+ a/b-enriched cathepsin L homolog, *Smed-CTSL2*. **d** Re-clustering cathepsin+ a/b cells resolves 4 subclusters, with distinct expression patterns shown in t-SNE plots and by in situ hybridization in **e**

Differential expression analysis and in situ hybridization demonstrate that the cathepsin$^+$ a/b clusters represent mesenchymal populations including pigment cells

For the 11 clusters identified by Seurat, cluster markers are identified on the basis of average differential expression. This identified a larger set of cluster markers, ranging from 23 for parapharyngeal cells to 627 for neoblasts (available on figshare https://doi.org/10.6084/m9.figshare.6852896) [68]. In order to identify the most distinguishing markers, the set of highly differentially expressed genes in a cluster with respect to all other clusters was identified using pairwise assessments of differential expression using a Bayesian approach to single-cell differential expression analysis (SCDE) [69]. This approach builds probabilistic error models for individual cells, capturing both over-dispersion (greater variability than expected) as well as high magnitude outliers and dropout events, thereby providing a more robust approach for detecting differential expression signatures. The clean-up step in this approach is far more stringent than in Seurat, retaining only ~ 60% of the cells compared to the Seurat pipeline ($n = 712$). For the 11 clusters identified by Seurat and 11,538 transcripts expressed in the cells, transcripts significantly differentially expressed (q value < 0.05) in 10 out of 11 clusters are considered putative markers for the cluster (available on figshare https://doi.org/10.6084/m9.figshare.6852896) [68]. Although there is a larger set of markers detected using Seurat, SCDE also identified unique markers (available on figshare https://doi.org/10.6084/m9.figshare.6852896) [68].

Differential expression analysis identified a significant enrichment for a cathepsin L homolog, *Smed-CTSL2* (SmedASXL_018694), in the *cathepsin*$^+$ clusters. Cathepsin L is a lysosomal cysteine proteinase with roles in antigen processing and presentation in humans (http://www.uniprot.org/uniprot/P07711). *Smed-CTSL2* is expressed across the entire length of the animal in a pattern of branched cells surrounding the gut (Fig. 4c). Interestingly, re-clustering only the cells in the cathepsin$^+$ clusters resulted in four distinct subclusters, each with a set of putative markers identified by Seurat (Fig. 4d, Additional file 2: Figure S4A). In situ hybridization of these putative markers demonstrated their unique expression patterns: Subcluster 1 was expressed throughout the mesenchyme (although these cells did not express *piwi-1* by scRNAseq) and tightly surrounded the gut (Fig. 4e, Additional file 2: Figure S4B); Subcluster 2 had a punctate expression pattern throughout the animal with randomly localized cell aggregates (Fig. 4e, Additional file 2: Figure S4B); Subcluster 3 was expressed largely within the gut (Fig. 4e); and the final subcluster, interestingly, represented previously described planarian pigment cells based on the enriched expression of published pigment lineage

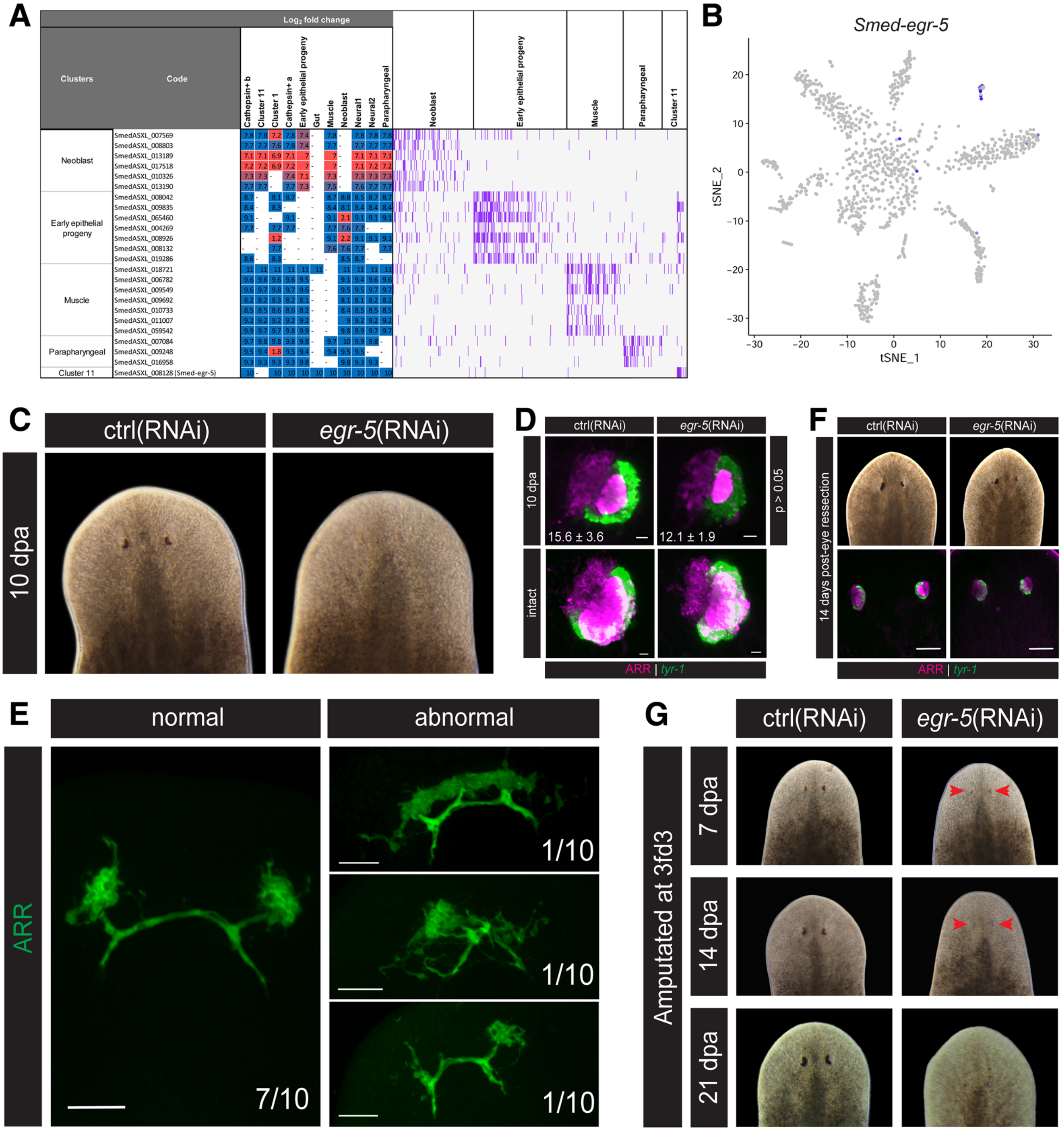


Fig. 5 *Smed-egr-5* is required for optic cup regeneration. **a** Heatmap of predicted transcription factors enriched in different clusters: the set of prediction factors significantly differentially upregulated in at least 4/10 clusters are shown, along with the $\log_2$ fold change in expression values and raw counts from corresponding cells. **b** t-SNE plot of *Smed-egr-5* expression demonstrating specificity to Cluster 11. **c** Bright field images of *Smed-egr-5*(RNAi) animals exhibiting reduced eye pigmentation. **d** In situ hybridization of *tyr-1* and immunohistochemistry for ARR in *Smed-egr-5*(RNAi) regenerating and intact animals. *Smed-egr-5* knockdown animals regenerated significantly fewer *tyr-1*$^+$ cells, while intact animals were comparable to controls. Images are 20-μm confocal z-stacks through the left eye. Scale bars = 10 μm. **e** The majority of *Smed-egr-5*(RNAi) animals displayed normal photoreceptor neuron regeneration and reinnervation by ARR staining following head amputation (70%), although some abnormalities were also observed. Images are 40–50 μm confocal z-stacks. Scale bars = 50 μm. **f** *Smed-egr-5*(RNAi) and control(RNAi) animals displayed comparable eye regeneration following right eye resection. Images are 20-μm confocal z-stacks. Scale bars = 50 μm. **g** At a lower dose of RNAi (3 feeds), *Smed-egr-5*(RNAi) animals exhibited some optic cup regeneration at 7 dpa (*red arrowheads*), which were subsequently lost at later time points. *n* = 10–20 for all experiments

markers, such as *pbgd-1* (Fig. 4e) [54]. Importantly, markers for each of these subclusters were found to be co-expressed to varying degrees in *Smed-CTSL2*$^+$ cells by double fluorescent in situ hybridization (FISH), consistent with the scRNAseq data (Additional file 2: Figure S4C–F). Interestingly, Subcluster 3 cells also expressed the neoblast marker *piwi-1* by scRNAseq (Additional file 2: Figure S4G). As an actively cycling population, the neoblast population is lost following a lethal dose of 6000 rads of irradiation. Likewise, the mesenchymal component of *ctcfl* (the Subcluster 3 marker) expression was found to be irradiation-sensitive, consistent with its partial expression in neoblasts (Additional file 2: Figure S4H).

Transcription factor analysis reveals cell type-specific expression

Mapping the 843 transcription factors to each cluster identified 30 exhibiting differential expression in specific clusters (significantly upregulated in 8/10 pairwise comparisons) (Additional file 4). Clusters that correspond to *muscle*, *epithelial*, and *parapharyngeal* cell types were associated with the most (7, 7, and 3, respectively) cluster-specific transcription factors, reflecting their generally higher number of differentially expressed transcripts (Additional file 4). Although *neoblasts* expressed a high number of transcription factors ($n = 8$), only 1 was cluster-specific. As expected, the most enriched transcription factor domains (zf-C2H2 and LIM) were also the most enriched in the cluster-specific transcripts. However, it is interesting to note that the Ets domain was associated with cluster-specific transcription factors in both *epithelial progenitors and Cluster 11, with similar patterns of expression observed in epithelial progenitors and Cluster 11.*

Aside from cluster-specific transcription factors, we identified five transcription factors that were abundant and ubiquitously expressed in all clusters (Additional file 4), comprising a Linker_histone domain involved in nucleosome assembly (SmedASXL_006919), and four CSDs, which are present in DNA- and RNA-binding proteins, and implicated in transcriptional regulation.

Analysis of differentially expressed transcription factors identifies the Cluster 11-specific *Smed-egr-5* as a regulator of optic cup regeneration

Expression of *Smed-egr-5* was specific to the unidentified Cluster 11 (Fig. 5a, b). Previous work on *Smed-egr-5* demonstrated a striking homeostatic phenotype in which worms exhibited tissue regression and ultimately lysed [70]. Consistent with previous reports, we observed *Smed-egr-5* expression subepidermally across the animal with enriched expression on the dorsal side (Additional file 2: Figure S5A) and knockdown of *Smed-egr-5* with a high dose of double-stranded RNA (dsRNA) RNAi food (2× dose) resulted in the previously described phenotype (Additional file 2: Figure S5B). dFISH revealed a very low degree of co-localization between *Smed-egr-5* and the early epithelial progenitor marker *prog-2*, but nearly 95% of *Smed-egr-5*$^+$ cells co-expressed the late epithelial progenitor marker *AGAT-1* (Additional file 2: Figure S5C). Because of the cluster specificity of *Smed-egr-5*, we sought to further characterize its function by using a lower dose of dsRNA (1× dose) to attempt to uncover further phenotypes. With our 1× RNAi food, we did not observe major defects in epithelial regeneration in *Smed-egr-5* knockdown animals (Additional file 2: Figure S5D); rather, we uncovered a new role for *Smed-egr-5* in eye regeneration. After eight feeds of 1× RNAi food, the new head tissue in *Smed-egr-5*(RNAi) regenerating animals appeared to lack eyes (Fig. 5c). To determine the extent of the missing eye tissue, *Smed-egr-5*(RNAi) animals were amputated 3 days after the eighth RNAi feed (8fd3) and were allowed to regenerate for 10 days. Regenerating animals were then stained for the optic cup marker *Smed-tyrosinase-1* (*tyr-1*) as well as anti-ARRESTIN (ARR), which marks the optic cup, photoreceptor neurons, and optic nerves. *Smed-egr-5*(RNAi) animals regenerated significantly fewer *tyr-1*$^+$ optic cup cells ($p < 0.05$), and the cells that did regenerate had noticeably weaker *tyr-1* expression (Fig. 5d). There were no apparent eye defects in homeostatic animals (Fig. 5d). ARR staining, on the other hand, revealed largely normal regeneration and reinnervation of photoreceptor neurons, although tissue organization was disrupted in a minority of animals (Fig. 5d, e). Because *tyr-1* and ARR staining in intact animals appeared largely normal, we hypothesized that *Smed-egr-5* is required specifically during optic cup regeneration.

To test this hypothesis further, an eye scratch assay was performed in which the right eye was resected without significant injury to the surrounding tissue. Previous work has demonstrated that this injury is not sufficient to illicit a regenerative response from the neoblasts; alternatively, the missing eye is restored by maintaining homeostatic levels of new cell incorporation and decreasing the rate of cell death [71]. At 14 days following eye resection, *Smed-egr-5*(RNAi) animals and *control*(RNAi) animals had comparable levels of eye restoration, supporting the hypothesis that eye homeostasis is independent of *Smed-egr-5* (Fig. 5f).

Interestingly, when *Smed-egr-5*(RNAi) animals were amputated at an earlier time point of 3fd3, optic cup regeneration was observed at 7 days post-amputation (dpa); however, these cells were subsequently lost at later time points post-amputation (Fig. 5g). The time-sensitive nature of this phenotype suggested that *Smed-egr-5* may be involved during the earliest stages of optic cup differentiation: optic cup progenitors that are still remaining after three RNAi feeds are capable of differentiating, but at later time points this progenitor population becomes exhausted and optic cup regeneration ultimately fails. From these data we hypothesize that *Smed-egr-5* plays a role in the production of optic cup progenitors. Thus, the lack of an observable homeostatic phenotype may simply be a consequence of the slow turnover of optic cup cells, and it remains possible that optic cup homeostasis may fail at later time points post-RNAi. Further studies at the neoblast level will help to elucidate the precise mechanisms by which *Smed-egr-5* promotes proper optic cup regeneration.

Systematic analysis of enriched Gene Ontology terms recapitulates cluster cell types

To provide deeper insights into functional properties associated with each cluster, we performed a GO enrichment analysis. GO mappings for 5900 transcripts expressed in the clusters were obtained through sequence similarity searches of putative homologs with GO annotations from model organisms *H. sapiens*, *Mus musculus*, *C. elegans*, *Danio rerio*, and *D. melanogaster*. Although these 5900 transcripts capture only ~ 10% of all transcripts identified in the scRNAseq data, statistically enriched terms were found to complement the previous marker gene analysis, with five of ten clusters consistent with previous cluster definitions: *muscle*, *neural1*, *neural2*, *neoblast*, and *epithelial progenitors* (Fig. 6a, Additional file 5). For example, the top ten enriched terms for *muscle* include terms such as structural constituent of muscle, muscle contraction, and muscle thin filament tropomyosin; *neoblast* is associated with many terms related to chromosomes and DNA replication, reflecting the high turnover associated with these cells; *epithelial* is enriched in terms related to endoplasmic reticulum, likely reflecting protein secretion associated with mucoid tissue [72]; and *neural1* and *2*, although displaying fewer enriched terms than the other tissues, are largely associated with neural functions. Our ability to identify similar consistent patterns of annotations in other clusters is probably limited due to the unavailability of specific GO terms for certain cell types (e.g., parapharyngeal) or due to lower numbers of cells (e.g., < 20 for gut cells) and significantly differentially expressed transcripts in these clusters.

Analyzing correlated gene expression across cell populations reveals transcriptional similarities between distinct cell clusters

In order to identify the set of known/novel subpopulations of cells sharing co-expressed sets of transcripts, we applied the Pathway and Geneset Overdispersion Analysis (PAGODA) component of the SCDE package [73]. This method identifies both the set of GO terms (assigned based on 1:1 orthologs of human) as well as de novo transcript sets consisting of well-correlated gene expression profiles. In this method, since multiple GO terms and de novo gene sets may comprise a common set of genes, clusters sharing the same set of genes are combined to arrive at a final set sharing coordinated variability in expression among the measured cells.

Our analysis reveals a set of four non-redundant clusters, two of which are shown in Fig. 6b. Note, while cell labels were not used during PAGODA, hierarchical clustering of the significantly correlated modules largely recapitulated the patterns of cell clustering generated by the Seurat analysis, especially for muscle, epithelial progenitor, and neural cells. Indeed, *epithelial progenitor* cells display the most distinct pattern of coordination, which PAGODA associates with Cluster 11 cells. The hierarchical clustering also places the gut and cathepsin + cells together, suggesting that they share transcriptionally co-regulated transcripts. One of the clusters corresponds to a set of cytoskeletal-related proteins in epithelial progenitor cells, as it is enriched in actins, dyneins, and FERM-domain-containing protein (found in several cytoskeletal-associated proteins [74]). The cluster also consists of several unannotated proteins, suggesting their likely involvement in cytoskeleton-related aspects. Although cytoskeletal-related proteins are found in all eukaryotic cells, they are likely to be enriched in epithelial cell types given the role of the cytoskeleton in epithelial cell polarity and intracellular trafficking [75, 76]. Although the second “cathepsin+ specific” cluster consists of proteins annotated to be involved in the lipid metabolic process in the lysosome [77, 78], phosphorylation/dephosphorylation [79], and cytoskeletal processes, it is unclear as to why these transcripts are co-expressed, opening up novel avenues for experimental interrogation. Reassuringly, *Smed-CTSL2* and *SmedASXL_009754* (encoding the cathepsin domain) are also identified in this cluster, emphasizing its abundant and unique expression in these cells.

scRNAseq data reveal tissue-specific patterns of metabolic pathway expression

The availability of cell-specific expression profiles generated through scRNAseq raises the intriguing possibility of identifying tissue-specific expression patterns for metabolic enzymes. Applying the hypergeometric test to mean enzyme expression (calculated using SCDE) for each cluster allowed the identification of significantly upregulated or downregulated metabolic pathways, as defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG) [80] (Additional file 6). Consistent with expectations, *neoblasts* were identified as the most metabolically active cell type followed by *muscle* and *epithelial progenitors* (Fig. 7a). The most significantly upregulated pathways are glycolysis/gluconeogenesis in

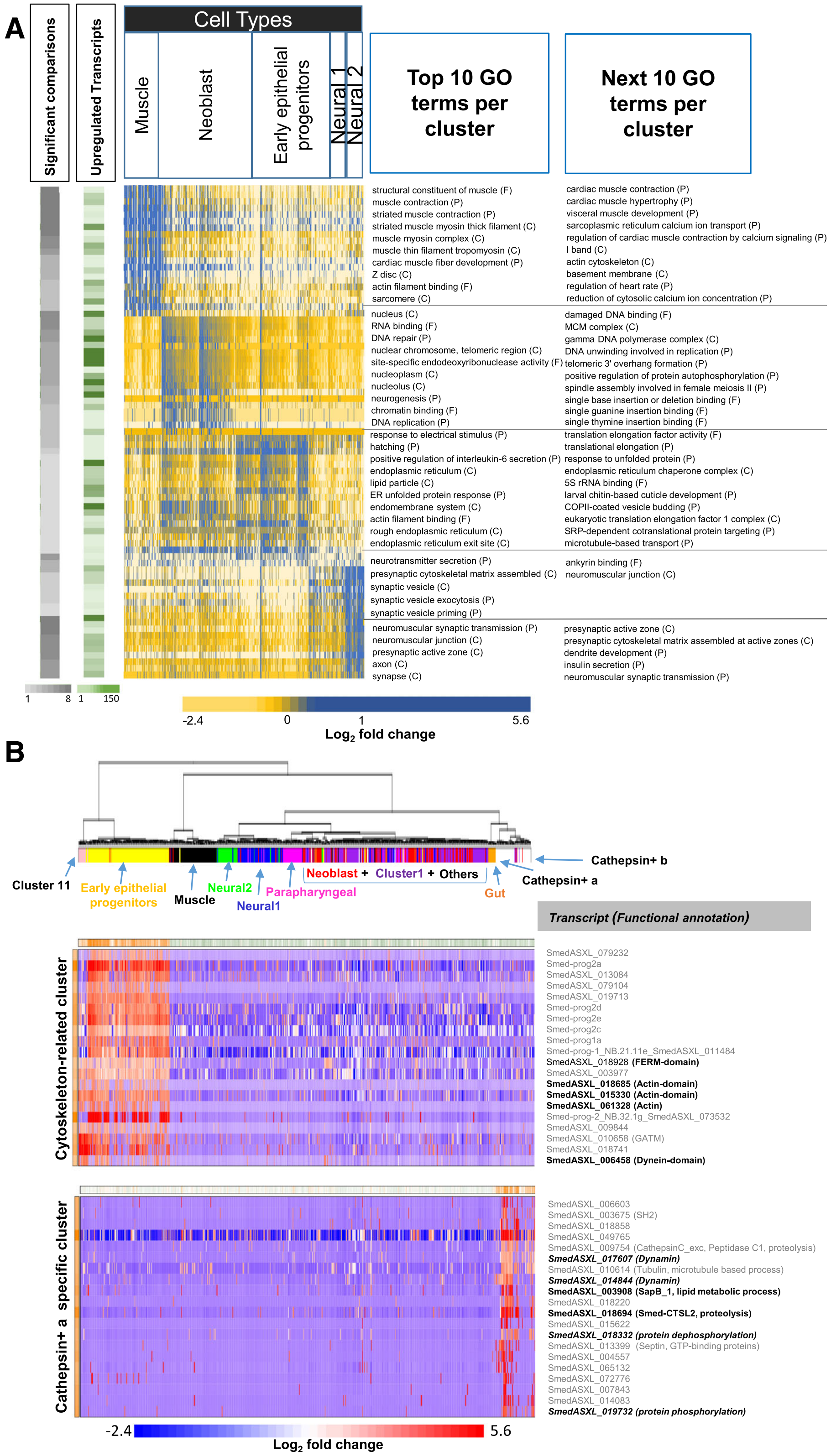


Fig. 6 (See legend on next page.)

(See figure on previous page.)

Fig. 6 Co-expressed sets. **a** Heatmap depicting the top 20 GO terms significantly enriched in each cluster along with the average expression of transcripts per GO term. The total numbers of statistically significant comparisons and upregulated transcripts for each GO term are also indicated alongside. **b** Unlabeled hierarchical clustering of cells based on GO gene sets and de novo gene sets consisting of significantly co-expressed offsets of transcripts with very similar gene expression profiles, generated using PAGODA. Two of the most significantly co-expressed modules are indicated, along with the changes in their expression

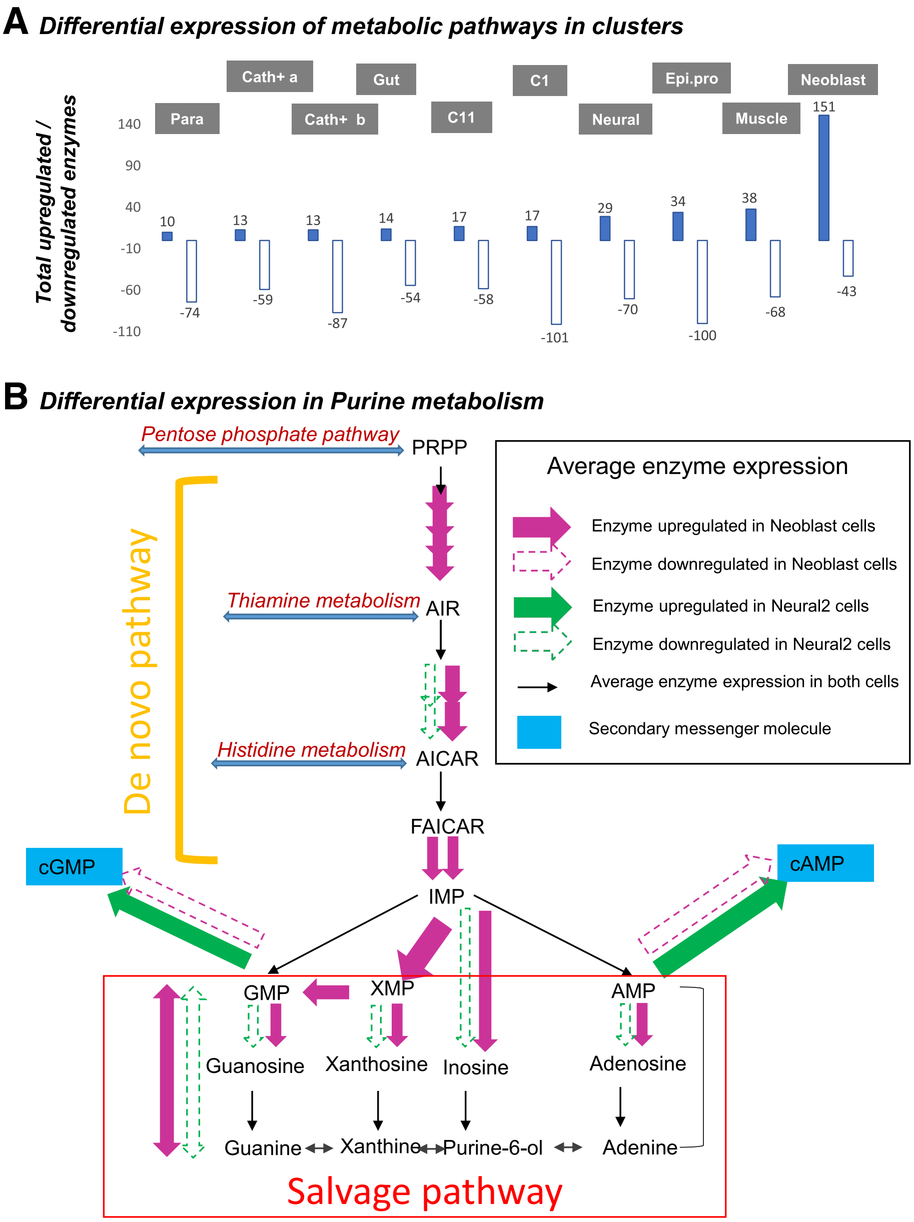


Fig. 7 Differential expression of metabolic pathways in clusters. **a** Distribution of significantly upregulated and downregulated enzymes in each cluster based on pairwise comparisons of $\log_2$ fold change in expression between clusters. **b** Schematic of differential expression in purine metabolism in neoblast and neural cell types

muscle (13/21 enzymes upregulated), supporting an increased need for energy production, and purine metabolism in *neoblast* (25/35 enzymes upregulated) and *neural2* (9/35 enzymes upregulated) cell types (Additional file 6). The purine metabolites adenine and guanine can be synthesized in two distinct pathways: the de novo pathway from CO_2, glycine, glutamine, aspartate, N^{10}-formyltetrahydrofolate and ribose-5-phosphate, starting with phosphoribosyl pyrophosphate (PRPP) and ending in inosine monophosphate (IMP) synthesis; and the salvage pathway, which recycles purine bases by degradation of nucleic acids and nucleotides (Fig. 7b). The purine nucleotides adenosine monophosphate (AMP), guanosine monophosphate (GMP), and xanthosine monophosphate (XMP) are synthesized from IMP. The corresponding trinucleotides lead to generation of intracellular secondary messengers, such as cyclic AMP (cAMP) and cyclic GMP (cGMP). Conversely, the purine nucleotide monophosphates can also be generated by the salvage pathway, by attaching free purine bases to PRPP: via the hypoxanthine-guanine phosphoribosyltransferase (HGPRT) enzyme for IMP, XMP, and GMP synthesis and adenine phosphoribosyltransferase (APRT) for AMP synthesis. As expected, several enzymes of the de novo pathway are upregulated in *neoblasts*, along with HGPRT of the salvage pathway; however, synthesis of secondary messengers is downregulated. In contrast, there is a significant upregulation of enzymes producing cAMP and cGMP in cells of the neural2 cluster. It is worth noting that *neoblasts*, in addition to upregulated purine metabolism, are also enriched for pyrimidine metabolism (21/24 enzymes) and one carbon pool by folate (10/11 enzymes upregulated). The enriched synthesis of folate derivatives likely provides the carbon units powering the de novo synthesis of purines and pyrimidines.

Discussion

In this study, starting with an initial set of 83,469 transcripts, we used a hierarchical tiered approach based on protein prediction algorithms of varying stringency and genome assembly mapping to define a high-confidence set of 35,232 transcripts, with 33,487 transcripts (~ 95% of transcriptome) mapping to 20,483 loci associated with the recently published dd_Smes_g4 *S. mediterranea* genome [10]. The number of mapped loci is consistent with the number of gene models supported by RNA sequencing (RNAseq) data ($n = 19{,}794$) for the closely related regeneration-competent flatworm *Macrostomum lignano* [81], supporting the quality of the filtered transcriptome. The usage of a tiered approach, which differs from that used to generate other integrated transcriptomes, i.e., PlanMine [19], Oxford [14], and SmedGD [9], reveals that there are 5% unique transcripts in the Toronto transcriptome — of which 20% are supported by homology mapping and 74% by genome assembly mapping, adding to the existing *S. mediterranea* repertoire. Further, assessment of transcriptome completeness in terms of core eukaryotic and metazoan gene sets as defined by BUSCO v1 [38] reveals that, although the Toronto and PlanMine transcriptomes have the greatest coverage (81% of "core" eukaryotic genes, 78% of "core" metazoan genes), the Toronto dataset also comprises the fewest duplicates in comparison. However, we note that this could also be an artifact of transcript length, potential fusion products from mis-assembly, or spliceoforms, which we did not assess and may be superior in other datasets.

A systematic and comparative bioinformatics analysis of the Toronto transcriptome with the genomes of *human*, *mouse*, *C. elegans*, and close platyhelminth relatives reveals an abundance of transposase-related domains (270 transcripts; DNA transposons of type DDE_1 and DDE_Tnp_1_7), MATH domains (99 transcripts; matrix metalloproteases and TNF-receptor associated factors) and cadherins (100 transcripts) in the planarian. Although the presence of transposable elements is corroborated by previous studies in *S. mediterranea* [82–84] and the basal flatworm *M. lignano* [81], it is important to note that they are expressed at low RPKM and only a small percentage appear active. Of the 99 transcripts with MATH domains, most are likely to be homologs of TRAF proteins, involved in signal transduction, on the basis of their top homologs. In light of studies supporting the role of homologs of human TRAF-3 and TRAF-6 proteins in immune response in the closely related planarian *Dugesia japonica* [85], the repertoire of putative TRAF proteins identified in this study provides candidate transcripts that can be tested for their role in planarian immunity. Cadherins are involved in regulating cell-cell adhesion, morphogenesis, and cell recognition [51, 52], with additional roles in cellular positioning and maintenance during and after development [86]. Phylogenetic analysis of putative cadherins obtained from humans, *S. mediterranea*, and other helminths predicts SmedASXL_013539 to be a calsyntenin-like protein, an ortholog of CASY-1 in *C. elegans*, which has been shown to be essential for learning [87], and CLSTN-1 and CLSTN-2 in humans, implicated in axonal anterograde transport and modulation of post-synaptic signals [88]. Functional characterization of these genes by RNAi may provide novel insights regarding immunity and learning, respectively, in planarians.

Our current understanding of *S. mediterranea* metabolism is limited [89]. Here we used an established enzyme prediction pipeline [90] to perform a metabolic reconstruction for *S. mediterranea*. Comparative analyses with other flatworms reveal that *S. mediterranea* encodes pathways for alternate sources of energy production, such as fatty acid metabolism and branched chain amino acid degradation. Our analyses also identified enzymes responsible for core 1 mucin-type O-glycosylation (notably absent in parasitic flatworms), which may be involved in the formation of the mucous coating, which is involved in locomotion, predation, innate immunity, and substrate adhesion [64].

Several studies have analyzed the role of transcription factors in *S. mediterranea* — involving pigmentation [91], gametogenesis [92], epidermal lineage differentiation [93], regeneration [94], and glial cells [95]. Interestingly, Scimone et al. combined RNA sequencing of neoblasts from wounded planarians with expression screening to identify 33 transcription factors and proposed that cell fate for almost all cell types is decided by expression of distinct transcription factors in the neoblast cells [55]. In this study, we used a combination of profile-based approaches to predict 841 putative transcription factors in *S. mediterranea*. A comparative analysis of putative transcription factors with other eukaryotic species reveals that transcription factor classes belonging to zf-C2H2, Homeobox, zf-BED, bZIP, and HLH are well represented in most species. Several others, such as CSD, Ets, and bZIP-map, well represented in *S. mediterranea* and vertebrates, have not been studied in the planarian. Studying these transcription factors in *S. mediterranea* might provide insights into the understanding of the regeneration process.

Several whole-organism as well as tissue-specific bulk RNAseq analyses investigating gene expression differences between two or more treatment conditions have been undertaken in *S. mediterranea*. To date, 32 RNA-seq/transcriptome datasets are currently available through the NCBI Gene Expression Omnibus (GEO). These experiments provide insights into factors required for restricting injury responses in planarians [96], signaling in planarian glia [95], tissue embryogenesis, homeostasis, and regeneration [97], and transcriptional changes in neoblasts [98]. However, recent developments in scRNAseq technology [99] have provided a novel approach to more directly assess functional differences between different cell populations [100, 101]. Recently, scRNAseq has been adopted by studies in *S. mediterranea*. A comprehensive study by Wurtzel et al. [26] https://doi.org/10.1016/j.devcel.2015.11.004 using smart-seq2 scRNAseq technology on 619 cells predicted 13 distinct cell clusters and defined 1214 unique tissue markers. This landmark study showed that a generic wound response transcriptional program is activated in almost all cells irrespective of the injury, with most wound-induced genes expressed in muscle, epidermis, and stem cells [26]. A comparison of the cluster markers in our study with those from Wurtzel et al. [102] shows that, although the majority of the cluster markers are shared for muscle (109/122), neural (67/74), and neoblast (87/94) cells, several unique cluster markers are found from this study. Further, Cluster 11 shares 105/133 cluster markers with epithelial cell types, consistent with the presence of *AGAT-1*$^+$ *Smed-egr-5*$^+$ cells in this cluster (Additional file 7).

In this study, to better understand the dynamics of the transcriptome in a spatial context, we applied scRNAseq to ~ 2000 cells, from which 25,168 transcripts were identified as expressed in at least one cell. Cluster analysis revealed 11 major clusters, with marker mapping identifying them to be associated with muscle, neural, neoblast, epithelial, and gut tissues, as well as a large cluster of cells likely representing transient transition states during neoblast differentiation (Cluster 1). Further, three novel clusters were identified: two cathepsin+ clusters consisting of four distinct mesenchymal cell types and a *Smed-egr-5*$^+$ cluster involved in optic cup regeneration. Reassuringly, the cell types of four clusters — muscle, neural, neoblast, and epithelial cells — were recapitulated on the basis of GO term assignments from 1:1 orthologs of model organisms for the most differentially enriched transcripts in these clusters, demonstrating the ability to identify cell types solely on the basis of enrichment of GO terms if GO term assignments are available for differentially enriched transcripts. Differential expression analysis of transcription factors in these clusters identified several cluster-specific factors likely associated with driving the morphogenesis and maintenance of tissue-specific biochemical processes. Analyzing the differential expression of metabolic pathways in these clusters identified neoblast cells as the most metabolically active cell type in *S. mediterranea*, with highly upregulated purine and pyrimidine metabolism and folate interconversions for providing the key metabolic precursors for nucleotide production. Analysis of purine metabolism with respect to different cell types revealed additional cell-specific patterns of expression, including the upregulation of both de novo and salvage biosynthetic pathways in neoblast cells, as well as the upregulation of intracellular secondary messengers involved in neuronal signaling. Furthermore, our study revealed four cadherin and two MATH domain proteins to be significantly upregulated in neoblast cells, whereas one cadherin and four MATH domain proteins are significantly upregulated in neural cells, providing testable hypotheses for learning more about immunity and learning in planarians.

It should be noted that during the revision of this manuscript, two new studies describing single-cell sequencing in *S. mediterranea* were published [29, 30]. Reassuringly, despite these new studies generating sequence data from ~ 22,000 and ~ 67,000 cells respectively, the results presented in both papers are consistent with our own findings. For example, our finding that pigment cells form a subcluster within the larger cathepsin+ cluster is consistent with the subclustering analysis performed in the Fincher study, in which *pbgd-1* was found to mark a specific cathepsin+ subcluster [29]. Further, saposinB-2, which we found to be a specific marker for the cathepsin+ subcluster 2, is expressed in a cathepsin+ subcluster from the same study. This suggests that smaller scale datasets, such as the one presented here, are sufficient to recapitulate many of the conclusions of larger-scale studies and consequently represent a valuable experimental template to assay specific RNAi phenotypes with single-cell sequencing in the future.

Conclusions

Here we present a definitive set of transcripts for the freshwater planarian *Schmidtea mediterranea*. We further annotate all genes with identifiable homology and identify gene family expansions and losses. Interestingly, TRAF proteins have been disproportionately increased, while Myc and AF-4 transcription factors are absent. A genome-scale metabolic reconstruction was then performed to identify metabolic pathways conserved in platyhelminths, those that have been lost in parasitic flatworms and those that represent lineage-specific innovations in *S. mediterranea*. Sequencing transcripts associated with 2000 individual cells identified cell types by differential gene expression and further revealed additional genes and pathways specific to each cell type. These analyses also uncovered a novel cell type associated with a novel mesenchymal cell population. In summary, these analyses build a foundation of cell types and gene conservation profiles that will inform future gene function studies.

Methods

Culturing of *S. mediterranea*, in situ hybridization, and RNA interference

Asexual individuals of *S. mediterranea* CIW4 strain were reared as previously described [103]. In situ hybridization was performed as previously described [18, 104]. RNAi was performed as previously described [54], with either three or eight feeds as indicated in the text.

Generating a high-confidence *S. mediterranea* transcriptome

The initial transcriptome of 83,469 transcripts was an assembly collated from five separate experiments and more than 1 billion RNA-seq reads from whole animals, purified tissues, RNAi conditions, and irradiated whole animals [18, 31–33] (NCBI Bioproject PRJNA215411). The resulting transcriptome was filtered using various criteria in order to arrive at a high-confidence set of putative protein-coding transcripts (Fig. 1a). As a first step, likely contaminants were identified by a BLASTn (from BLAST+ 2.2.28) [40] search against the protein nucleotide (nt) database (2016) [105] to remove sequences matching other species at a sequence identity and query coverage cutoff of 95% ($n = 237$) as well as those matching vector sequences ($n = 8$). Next, likely mis-assembled transcripts were removed by identifying all transcripts with ≥ 25 unmapped bases to the transcriptome ($n = 2387$). Clustering approaches did not reduce the initial transcriptome to the expected range observed in regeneration-competent species such as *M. lignano* and *D. japonica*, suggesting the presence of contaminants, misassembled transcripts, split transcripts, alternative splice variants, and/or leaky transcripts. Therefore, the initial transcriptome was scrutinized via a multi-layered approach to identify potential protein-coding transcripts. The transcriptome was parsed through the prot4EST v3.1b [106] pipeline, an integrated approach which overcomes deficits in training data in order to convert transcripts into proteins. This multi-tiered program identifies coding transcripts in various stages. The first step identifies homologs of known RNA and protein sequences using the BLAST suite [40] — BLASTn (from BLAST 2.2.28) against the SILVA database (release 115) [107] at an E-value of 1e-65 for identifying RNA transcripts, BLASTx against the MitoMiner database (v3.1) [35] at an E-value of 1e-08 and against the UniProt database [34] at an e value of 1e-05 for identifying mitochondrial and nuclear transcripts, respectively. From the remaining transcripts, the second step identifies likely protein-coding transcripts using ESTscan [v3.0.3] [39], a hidden Markov model (HMM)-based model trained to be error-tolerant, using a simulated *S. mediterranea* training set. Finally, the remaining transcripts are processed to identify the longest string of amino acids uninterrupted by stop codons from a six-frame translation of the sequence (LongestORFs). From the set of categorized transcripts, all transcripts with query coverage spanning two thirds of the reference sequence in RNA/mitochondrial/nuclear databases are retained. The rest of the transcripts are retained only if there is any support in terms of the following: (1) homology with respect to conserved eukaryotic gene sets (CEGMA v2.5 [37] and BUSCO v1.1 [38] using BLASTx at an E-value of 1e-08) and other helminth transcriptome EST datasets obtained from the NCBI (*B. glabrata*, *C. sinensis*, *C. gigas*, *D. japonica*, *D. ryukyuensis*, *E. granulosus*, *E. multilocularis*, *H. robusta*, *H. medicinalis*, *H. microstoma*, *M. lignano*, *M. californianus*, *O. viverrini*, *S. japonicum*, *S. mansoni*, *T. solium*) using BLASTn at an E-value of 1e-15; (2) annotation by InterPro [36] at an E-value of 1e-03; and (3) co-location of the draft *S. mediterranea* genome with ESTs from NCBI, transcripts from the Oxford dataset (v0.1) [14], or transcripts from SmedGD v2.0 using Spaln v2 [41] at a stringency filtering of F2 (corresponding to alignment length > 200 bp, sequence identity ≥ 93%, query coverage ≥ 93%).

Comparison with PlanMine genome and transcriptome

The Toronto transcriptome was mapped onto the PlanMine genome [10] using Spaln v2 [41] at stringency filtering cutoffs corresponding to F2 (sequence identity ≥ 93%, query coverage ≥ 93%) and F1 (sequence identity ≥ 75%, query coverage ≥ 75%) in order to identify the extent of overlap. Subsequently, the transcriptomes were compared using BLASTn [40] searches against each other using a relaxed word size ($n = 7$) in order to improve the stringency of the searches. BLASTn matches of the Toronto transcriptome to the PlanMine transcriptome were pruned based on the nearest bit score cutoff corresponding to the number of overlapping matches to the genome identified at F1 cutoff (corresponding to a bit score value ≥ 40). Based on this cutoff, matches were identified between the Toronto, PlanMine, Oxford, and SmedGD transcriptomes.

Functional annotation of the transcriptome

The predicted protein sequences generated from the high-confidence transcriptome were functionally annotated by (1) HMM searches against the curated Pfam-A database v31 using the PfamScan tool with hmmer-3.1b1 [44] at default cutoffs. Only those matches with an E-value cutoff of < 0.001 were considered for further analysis; (2) InterProScan v5.15.54.0 [108] searches against profiles from High-quality Automated and Manual Annotation of Poteins (HAMAP), ProDom, Protein Information Resource SuperFamily (PIRSF), Simple Modular Architecture Research Tool (SMART), Pfam, Gene3D, Coils, Prosite, TIGRFAM, PRINTS, and Superfamily databases; and (3) GO annotation based on Interpro2GO (2016) mappings [109].

RPKM calculation

The expression levels of the transcripts were calculated by mapping the reads from 58 RNA-seq results (listed as the column headers under the RPKM section in Additional file 1) onto the initial transcriptome using Burrows-Wheeler Aligner (BWA) [110] and obtaining the number of reads mapped for each transcript. The normalized expression levels were quantified in RPKM units for each transcript for each RNA-seq experiment using the formula:

RPKM = Number of Reads/(Transcript Length/1000 * Total Num Reads/1,000,000) where Total Num Reads consisted only of those transcripts with ≥ 10 reads mapped to them in a sample. Next, the mean, standard deviation, and median RPKM values for each transcript were calculated based on the number of RNA-seq experiments where the transcript was expressed. The mean values of all transcripts in the definitive transcriptome were used to derive a percentile distribution of RPKM values, which is used as a guide to derive the average level of expression of a transcript (low < 50th percentile, high > 20th percentile, medium ≤ 20th percentile and ≥ 50th percentile).

Phylogenetic analysis of cadherins

A set of 94 *S. mediterranea* transcripts with predicted cadherin domains from Pfam-A [44] at an E-value < 0.0001 were collected. 1:1 orthologs of these transcripts were identified using Inparanoid v2.0 [111] for *C. elegans* ($n = 3$), *E. granulosus* ($n = 24$), *E. multilocularis* ($n = 23$), *G. salaris* ($n = 16$), *H. nana* ($n = 24$), *S. haematobium* ($n = 21$), *S. mansoni* ($n = 20$), *T. solium* ($n = 37$), *O. viverini* ($n = 21$), and *C. sinensis* ($n = 22$). A set of 176 Ensembl [112] isoforms annotated as cadherins were also retrieved. A non-redundant set from the set of 481 sequences was generated using the online version of CD-HIT (weizhongli-lab.org) [113] at 50% sequence identity cutoff, yielding 249 clusters. From each cluster, only the longest sequence was retained, unless they were helminth sequences, leading to 331 sequences. These sequences were aligned using the Multiple Alignnment using Fast Fourier Transform (MAFFT) web tool (https://mafft.cbrc.jp/alignment/software/) [114] and trimmed using trimAl 1.4 [115] (with the -gappyout setting) and a maximum likelihood phylogenetic tree constructed using PhyML package v20140412 [116] with 1024 bootstrap replicates.

Enzyme annotation of the predicted proteome

For each of the predicted protein sequences, an initial set of enzyme commission (EC) predictions was obtained from several methods: (1) density estimation tool for enzyme classification (DETECT) v1.0 run using default parameters (here we retained hits with Integrated Likelihood Score (ILS) cutoff ≥ 0.9 from the top predictions file which also had ≥ 5 positive hits) [57]; (2) BLASTP (from BLAST+ 2.2.28) run against the Swiss-Prot database (release 2014-08) at an E-value cutoff of 1e-10; the enzyme annotations of top hits in the Swiss-Prot database were mapped to the query sequence [40]; and (3) PRIAM enzyme rel. Feb-2014 run using relaxed cutoffs specified for genome-wide annotations of organisms (minimum probability > 0.5, profile coverage > 70%, check catalytic - TRUE) [58]. From these assignments, a set of consolidated high-confidence predictions was derived using in-house scripts by retaining only those predictions identified by both PRIAM and BLASTP and combining them with the predictions from DETECT. Percent pathway conservation was calculated for the set of metabolic pathways as defined by KEGG v70 [80] using the following formula: (Number of predicted ECs in a KEGG pathway × 100)/Total number of ECs in the KEGG pathway.

Transcription factor prediction

The InterProScan v5.15.54.0 [108] outputs for all 35,235 high-confidence predicted protein sequences were scanned as follows in order to identify a set of putative transcription factors: (1) InterProScan hits with the description "transcription factor", (2) InterProScan hits to the Pfam families listed in the curated transcription factor database DNA-binding domain (DBD) v2.0 [117], (3) InterProScan hits to the Superfamily families listed in DBD v2.0. The hits from all of the above criteria were consolidated to arrive at the final predicted set of transcription factors for the organism.

Transposon analysis

RepeatMasker (2013) was used to predict repeats for the SmedAsxl genome v1.1. All transcripts assigned DDE transposase domains were mapped onto the masked SmedAsxl genome with the F2 cutoff of Spaln v2 [41] and searched for the presence of repetitive elements. For repetitive elements found within the mapped region, sequence regions flanking 1000 bp on either side of the repetitive element were extracted and its sequence divergence with the consensus of the repeat element calculated using the Needleman-Wunsch algorithm from the European Molecular Biology Open Software Suite (EMBOSS) package. A histogram of the extent of sequence divergence was analyzed in order to identify likely active elements, characterized by sequence divergence ≤5% from consensus element [118].

Generation of single-cell RNA-seq data

For single-cell RNA sequencing, a whole-animal cell suspension (in calcium-magnesium-free (CMF) + 10% glucose solution) was stained with the cell viability dye calcein (0.2 μg/ml), and calcein-positive cells were collected by fluorescence-activated cell sorting (FACS). Cells were then processed through a Drop-seq instrument and complementary DNA (cDNA) libraries were prepared as described in [66]. Libraries were sequenced on an Illumina NextSeq500 to a total depth of ~ 480 million reads. The data are available at the NCBI GEO database under accession number GSE115280 (https://www.ncbi.nlm.nih.gov/gds/?term=GSE115280) [119]. Reads were aligned to the *S. mediterranea* SmedASXL transcriptome assembly under NCBI BioProject PRJNA215411 using Bowtie2 with 15-bp 3′ trimming.

Identification of clusters and cluster markers using Seurat

To identify cell clusters enriched for transcriptionally co-expressed profiles, single-cell RNA-seq data were processed against the definitive Toronto transcriptome using the Seurat [67] pipeline while considering the standard default quality cutoffs optimized for a dataset of size ~ 3000 cells, i.e., min.genes = 200, min.cells = 3, tot.expr = 1e4. The resolution parameter in the FindClusters function was varied from 0.4 to 4, and a resolution of 1 was chosen as it yielded the most visually distinct clustering pattern. In Seurat [67], cluster markers were identified using the FindAllMarkers function of the Seurat pipeline by considering transcripts that are expressed in at least 25% of the cells in the cluster, with an average expression ≥ 25% in comparison to their expression in all other clusters. The significance of the differential expression is calculated using the "bimod" likelihood-ratio test for single-cell gene expression [120] for all cells in one cluster vs all other cells and expressed as *p* values.

Differential expression of transcripts and identification of cluster markers in SCDE

Differential expression of transcripts between clusters was calculated using the SCDE R package, which employs a Bayesian approach to single-cell differential expression analysis [69], considering only those cells with a minimum library size of 500, and only those transcripts mapping to ≥ 10 reads and detected in ≥ 5 cells, since this yielded at least ten cells per cluster. Differential expression was calculated for all-vs-all pairwise combinations of clusters classified using Seurat, and the $\log_2$ fold change and *p* values were noted. All transcripts that are significantly upregulated in 9/10 pairwise comparisons are considered as cluster markers.

Hypergeometric test for KEGG metabolic pathways

The enrichment of differentially expressed transcripts (both upregulated, corresponding to a $\log_2$ fold change > 1; and downregulated, corresponding to a $\log_2$ fold change < − 1, according to SCDE) was assessed using a hypergeometric test (using the *phyper* function in R) for all pairwise combinations of clusters classified using Seurat. All KEGG pathways with a *p* value < 0.05 were considered to be enriched.

Hypergeometric test for analyzing enrichment of Gene Ontology terms

Gene Ontology (GO) refers to a database providing a structured vocabulary for annotating genes [43]. The genes are annotated using specific biologically relevant terms corresponding to three main categories: Biological Process (BP), Molecular Function (MF), and Cellular Compartment (CC). *Schmidtea* transcripts were annotated with the GO terms from 1:1 orthologs from five model organisms: *H. sapiens*, *M. musculus*, *D. rerio*, *C. elegans*, and *D. melanogaster*, as identified by Inparanoid (annotations downloaded from GO website http://geneontology.org/page/download-annotations). The annotations were transferred for GO terms designated by all methods other than Inference by Electronic Annotation (non-IEA) on the basis of Inparanoid mapping, using in-house scripts. The enrichment of significantly upregulated transcripts associated with the GO term ($\log_2$ fold change > 1 calculated using SCDE) was assessed using a hypergeometric test (using the *phyper* function in R) for all pairwise combinations of clusters classified using

Seurat. All statistically significant GO terms associated with more upregulated transcripts than downregulated transcripts and containing at least two significantly upregulated transcripts were considered to be enriched.

Identifying co-expressed modules in cell types

Using the Pathway and Geneset Overdispersion Analysis (PAGODA) component of the SCDE package [73], the set of co-expressed gene sets characterized by statistically significant coordinated variability in sets of cells was identified. For the pre-defined gene sets, GO term annotations assigned based on 1:1 Inparanoid orthologs of *H. sapiens* were considered. The initial dataset was cleaned using parameters similar to those used for SCDE, i.e., min.genes = 500, resulting in a set of 11,542 transcripts and 720 cells. The *k* nearest neighbors (KNN)-based error modeling step was carried out by considering 11 subpopulations (for the 11 Seurat clusters). The results were viewed in the PAGODA application.

Additional files

Additional file 1: Characteristics of the Toronto definitive transcriptome. Details of *S. mediterranea* transcripts in the definitive transcriptome. (XLSB 18937 kb)

Additional file 2: Figure S1. Comparative analysis of the mapped and unmapped transcripts of Toronto and PlanMine transcriptomes onto dd_Smes_g4 genome assembly. **Figure S2.** Phylogenetic profiling of *S. mediterranea* based on GO Slim terms. Piecharts showing phylogenetic breakdown of various GO Slim groupings. **Figure S3.** Phylogenetic distribution of cadherins from human, *C. elegans*, Platyhelminthes, and *S. mediterranea*. Phylogenetic tree of cadherins. **Figure S4.** Gene expression profiles of Cluster 7 subclusters. **Figure S5.** *Smed-egr5* gene expression and phenotypes. (PDF 1428 kb)

Additional file 3: List of predicted transcription factors identified in Toronto transcriptome dataset. (XLSX 52 kb)

Additional file 4: Distribution of predicted transcription factors in clusters. Three tables indicating distribution of transcription factors identified in the Toronto transcriptome with clusters identified by Seurat. (XLSX 4752 kb)

Additional file 5: Heatmap of all Gene Ontology (GO) terms. Table showing distribution of various GO categories across cell-type clusters. (XLSX 2009 kb)

Additional file 6: Distribution of significantly upregulated/downregulated enzymes in the clusters. Table indicating pathways displaying patterns of significantly upregulated or downregulated enzymes. (XLSX 25 kb)

Additional file 7: Comparison of cluster markers identified in this study with results of Wurtzel et al. [26]. Table comparing cell-type markers identified in this study with those of a previously published study. (XLSX 9 kb)

Abbreviations

dpa: days post-amputation; EST: Expressed sequence tag; FISH: fluorescent in situ hybridization; GO: Gene ontology; MATH: Meprin and TRAF homology; RNAi: RNA interference; RPKM: Reads per kilobase per million mapped reads; scRNAseq: single-cell RNA sequencing; TNF: Tumor necrosis factor; TRAF: TNF receptor associated factor; tSNE: t-distributed stochastic neighbor embedding

Funding

This research was supported by grants from the Natural Sciences and Engineering Research Council (NSERC; RGPIN-2014-06664 to JP and RGPIN-2016-06354 to BJP) and the National Institutes for Health (R21AI126466 to JP). AMM was funded by Canadian Institutes of Health Research Banting and Best Ph.D. fellowship GSD-152379. BJP was supported by Ontario Institute for Cancer Research (OICR) grant number IA-026. NLM was supported by a student Restracomp fellowship given by the Hospital for Sick Children. Computing resources were provided by the SciNet High Performance Computing (HPC) Consortium; SciNet is funded by the Canada Foundation for Innovation under the auspices of Compute Canada, the Government of Ontario, Ontario Research Fund - Research Excellence, and the University of Toronto.

Availability of data and materials

The Toronto transcriptome dataset is available at http://compsysbio.org/datasets/schmidtea/Toronto_transcriptome.fa [121] and an augmented version also containing a set of non-overlapping PlanMine transcripts that map onto the dd_Smes_g4 genome is available at http://compsysbio.org/datasets/schmidtea/Toronto_transcriptome_plus.fa [122] as well as the PlanMine genomic resource site (http://planmine.mpi-cbg.de/planmine/begin.do). These sequences are annotated using BLASTx against non-redundant protein (NR, Dec 2017) at E-value ≤ 1e-10 and BLASTn against non-redundant nucleotide (NT, Dec 2017) at E-value ≤ 1e-50. Single-cell RNA sequence data are available at the NCBI Gene Expression Omnibus (GEO) database with accession number GSE115280 (https://www.ncbi.nlm.nih.gov/gds/?term=GSE115280) [119]. Data corresponding to cluster markers for the 11 clusters identified in this study are available from figshare (https://doi.org/10.6084/m9.figshare.6852896) [68]. All other data generated or analyzed during this study are included in this published article and its additional files.

Authors' contributions

JP and BJP together designed and supervised the project. LSS assembled transcriptome datasets, performed comparative and metabolic analyses, and drafted the first version of the manuscript. AMM performed single-cell genomics analyses and follow-up characterizations in *Schmidtea*. NLM performed in vivo characterization of cathepsin+ clusters. LSS, AMM, BJP, and JP all contributed to writing of the manuscript, and NLM helped with edits. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Hospital for Sick Children, Toronto, ON, Canada. [2]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada. [3]Ontario Institute for Cancer Research, Toronto, ON, Canada. [4]Department of Biochemistry, University of Toronto, Toronto, ON, Canada.

Received: 15 January 2018 Accepted: 1 August 2018
Published online: 24 August 2018

References

1. Duronio RJ, O'Farrell PH, Sluder G, Su TT. Sophisticated lessons from simple organisms: appreciating the value of curiosity-driven research. Dis Model Mech England. 2017;10:1381–9.
2. Newmark PA, Reddien PW, Cebria F, Alvarado AS. Ingestion of bacterially expressed double-stranded RNA inhibits gene expression in planarians. Proc Natl Acad Sci. 2003;100:11861–5.
3. Karami A, Tebyanian H, Goodarzi V, Shiri S. Planarians: an in vivo model for regenerative medicine. Int J Stem Cells. 2015;8:128–33.
4. Sánchez Alvarado A. Planarian regeneration: its end is its beginning. Cell. 2006;124:241–5.
5. Newmark PA, Wang Y, Chong T. Germ cell specification and regeneration in planarians. Cold Spring Harb Symp Quant Biol. 2008;73:573–81.
6. Reddien PW, Alvarado AS. Fundamentals of planarian regeneration. Annu Rev Cell Dev Biol. 2004;20:725–57.

7. Rink JC. Stem cell systems and regeneration in planaria. Dev Genes Evol. 2013;223:67–84.
8. el Kouni MH. Pyrimidine metabolism in schistosomes: a comparison with other parasites and the search for potential chemotherapeutic targets. Comp Biochem Physiol Part - B Biochem Mol Biol. 2017;213:55–80.
9. Robb SMC, Gotting K, Ross E, Sánchez Alvarado A. SmedGD 2.0: the Schmidtea mediterranea genome database. Genesis. 2015;53:535–46.
10. Grohme MA, Schloissnig S, Rozanski A, Pippel M, Young GR, Winkler S, et al. The genome of Schmidtea mediterranea and the evolution of core cellular mechanisms. Nature. England. 2018;554:56–61.
11. Abril JF, Cebrià F, Rodríguez-Esteban G, Horn T, Fraguas S, Calvo B, et al. Smed454 dataset: unravelling the transcriptome of Schmidtea mediterranea. BMC Genomics. 2010;11:731.
12. Adamidi C, Wang Y, Gruen D, Mastrobuoni G, You X, Tolle D, et al. De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. Genome Res. 2011;21: 1193–200.
13. Blythe MJ, Kao D, Malla S, Rowsell J, Wilson R, Evans D, et al. A dual platform approach to transcript discovery for the planarian Schmidtea mediterranea to establish RNAseq for stem cell and regeneration biology. PLoS One. 2010;5:e15617.
14. Kao D, Felix D, Aboobaker A. The planarian regeneration transcriptome reveals a shared but temporally shifted regulatory program between opposing head and tail scenarios. BMC Genomics. 2013;14:797.
15. Lapan SW, Reddien PW. Transcriptome analysis of the planarian eye identifies ovo as a specific regulator of eye regeneration. Cell Rep. 2012;2:294–307.
16. Resch AM, Palakodeti D, Lu YC, Horowitz M, Graveley BR. Transcriptome analysis reveals strain-specific and conserved stemness genes in Schmidtea mediterranea. PLoS One. 2012;7:e34447.
17. Sandmann T, Vogg MC, Owlarn S, Boutros M, Bartscherer K. The head-regeneration transcriptome of the planarian Schmidtea mediterranea. Genome Biol. 2011;12:R76.
18. Zhu SJ, Hallows SE, Currie KW, Xu C, Pearson BJ. A mex3 homolog is required for differentiation during planarian stem cell lineage development. elife. 2015;4:1–23.
19. Brandl H, Moon HK, Vila-Farré M, Liu SY, Henry I, Rink JC. PlanMine—a mineable resource of planarian biology and biodiversity. Nucleic Acids Res. 2016;44:D764–73.
20. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. The human transcriptome across tissues and individuals. Science. 2015;348:660–5.
21. Hostelley TL, Lodh S, Zaghloul NA. Whole organism transcriptome analysis of zebrafish models of Bardet-Biedl syndrome and Alström syndrome provides mechanistic insight into shared and divergent phenotypes. BMC Genomics. 2016;17:318.
22. Parkinson J, Wasmuth JD, Salinas G, Bizarro CV, Sanford C, Berriman M, et al. A transcriptomic nalysis of Echinococcus granulosus larval stages: implications for parasite biology and host adaptation. PLoS Negl Trop Dis. 2012;6:e1897.
23. Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, et al. Quantitative assessment of single-cell RNA-sequencing methods. Nat Methods. 2014;11:41–6.
24. Dalerba P, Kalisky T, Sahoo D, Rajendran PS, Rothenberg ME, Leyrat AA, et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nat Biotechnol. 2011;29:1120–7.
25. Gehrke AR, Srivastava M. Neoblasts and the evolution of whole-body regeneration. Curr Opin Genet Dev. 2016;40:131–7.
26. Wurtzel O, Cote LE, Poirier A, Satija R, Regev A, Reddien PW. A generic and cell-type-specific wound response precedes regeneration in planarians. Dev Cell. 2015;35:632–45.
27. Zhu SJ, Pearson BJ. (Neo)blast from the past: new insights into planarian stem cell lineages. Curr Opin Genet Dev. 2016;40:74–80.
28. Molinaro AM, Pearson BJ. In silico lineage tracing through single cell transcriptomics identifies a neural stem cell population in planarians. Genome Biol. 2016;17:87.
29. Fincher CT, Wurtzel O, de Hoog T, Kravarik KM, Reddien PW. Cell type transcriptome atlas for the planarian Schmidtea mediterranea. Science. 2018;360 https://doi.org/10.1126/science.aaq1736.
30. Plass M, Solana J, Wolf FA, Ayoub S, Misios A, Glazar P, et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. Science. 2018;360 https://doi.org/10.1126/science.aaq1723.
31. Lin AYT, Pearson BJ. Planarian yorkie/YAP functions to integrate adult stem cell proliferation, organ homeostasis and maintenance of axial patterning. Development. 2014;141:1197–208.
32. Currie KW, Pearson BJ. Transcription factors lhx1/5-1 and pitx are required for the maintenance and regeneration of serotonergic neurons in planarians. Development. 2013;140:3577–88.
33. Labbé RM, Irimia M, Currie KW, Lin A, Zhu SJ, Brown DDR, et al. A comparative transcriptomic analysis reveals conserved features of stem cell pluripotency in planarians and mammals. Stem Cells. 2012;30:1734–45.
34. Pundir S, Martin MJ, O'Donovan C. UniProt protein knowledgebase. In: Wu C, Arighi C, Ross K, editors. Protein bioinformatics. Methods in molecular biology, vol. 1558. New York: Humana Press; 2017. p. 41–55.
35. Calvo SE, Clauser KR, Mootha VK. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. Nucleic Acids Res. 2016;44: D1251–7.
36. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, et al. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. 2015;43:D213–21.
37. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23:1061–7.
38. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
39. Iseli C, Jongeneel CV, Bucher P. ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol. 1999:138–48.
40. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
41. Iwata H, Gotoh O. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. Nucleic Acids Res. 2012;40:e161.
42. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
43. Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2015;43:D1049–56.
44. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44:D279–85.
45. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.
46. Huang CRL, Burns KH, Boeke JD. Active transposition in genomes. Annu Rev Genet. 2012;46:651–75.
47. Smit AF, Riggs AD. Tiggers and DNA transposon fossils in the human genome. Proc Natl Acad Sci U S A. 1996;93:1443–8.
48. Bradley JR, Pober JS. Tumor necrosis factor receptor-associated factors (TRAFs). Oncogene. 2001;20:6482–91.
49. Walsh MC, Lee J, Choi Y. Tumor necrosis factor receptor-associated factor 6 (TRAF6) regulation of development, function, and homeostasis of the immune system. Immunol Rev. 2015;266:72–92.
50. Yang XD, Sun SC. Targeting signaling factors for degradation, an emerging mechanism for TRAF functions. Immunol Rev. 2015;266:56–71.
51. Tepass U, Truong K, Godt D, Ikura M, Peifer M. Cadherins in embryonic and neural morphogenesis. Nat Rev Mol Cell Biol. 2000;1:91–100.
52. Maître JL, Heisenberg CP. Three functions of cadherins in cell adhesion. Curr Biol. 2013;23:PR626–33.
53. Cromar GL, Zhao A, Xiong X, Swapna LS, Loughran N, Song H, et al. PhyloPro2.0: a database for the dynamic exploration of phylogenetically conserved proteins and their domain architectures across the Eukarya. Database. 2016;2016 pii:baw013
54. He X, Lindsay-Mosher N, Li Y, Molinaro AM, Pellettieri J, Pearson BJ. FOX and ETS family transcription factors regulate the pigment cell lineage in planarians. Development. 2017;144:4540–51.
55. Scimone ML, Kravarik KM, Lapan SW, Reddien PW. Neoblast specialization in regeneration of the planarian schmidtea mediterranea. Stem Cell Rep. 2014; 3:339–52.
56. Esposito G, Cevenini A, Cuomo A, de Falco F, Sabbatino D, Pane F, et al. Protein network study of human AF4 reveals its central role in RNA Pol II-mediated transcription and in phosphorylation-dependent regulatory mechanisms. Biochem J. 2011;438:121–31.

57. Hung SS, Wasmuth J, Sanford C, Parkinson J. DETECT—a density estimation tool for enzyme classification and its application to Plasmodium falciparum. Bioinformatics. 2010;26:1690–8.
58. Claudel-Renard C, Chevalet C, Faraut T, Kahn D. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res. 2003;31:6633–9.
59. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007;35:W182–5.
60. Arakaki AK, Tian W, Skolnick J. High precision multi-genome scale reannotation of enzyme function by EFICAz. BMC Genomics. 2006;7:315.
61. Skelly PJ, Da'dara AA, Li XH, Castro-Borges W, Wilson RA. Schistosome feeding and regurgitation. PLoS Pathog. 2014;10:e1004246.
62. Rodriguez-Contreras D, Skelly PJ, Landa A, Shoemaker CB, Laclette JP. Molecular and functional characterization and tissue localization of 2 glucose transporter homologues (TGTP1 and TGTP2) from the tapeworm Taenia solium. Parasitology. 1998;117(Pt 6):579–88.
63. Gallo M, Riddle DL. Regulation of metabolism in Caenorhabditis elegans longevity. J Biol. 2010;9:7.
64. Bocchinfuso DG, Taylor P, Ross E, Ignatchenko A, Ignatchenko V, Kislinger T, et al. Proteomic profiling of the planarian *Schmidtea mediterranea* and its mucous reveals similarities with human secretions and those predicted for parasitic flatworms. Mol Cell Proteomics. 2012;11:681–91.
65. Bailey LB, Gregory JF 3rd, Jesse F. Folate metabolism and requirements. J Nutr. 1999;129:779–82.
66. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161:1202–14.
67. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33:495–502.
68. Swapna LS, Molinaro AM, Lindsay-Mosher N, Pearson BJ, Parkinson J. Comparative transcriptomic analyses and single-cell RNA sequencing of the freshwater planarian Schmidtea mediterranea identifies major cell types and pathway conservation. Data Sets. figshare. https://doi.org/10.6084/m9.figshare.6852896.
69. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11:740–2.
70. Tu KC, Cheng LC, Vu HTK, Lange JJ, McKinney SA, Seidel CW, et al. Egr-5 is a post-mitotic regulator of planarian epidermal differentiation. elife. 2015;4:e10501.
71. LoCascio SA, Lapan SW, Reddien PW. Eye absence does not regulate planarian stem cells during eye regeneration. Dev Cell. 2017;40:381–391.e3.
72. Cheng LC, Tu KC, Seidel CW, Robb SMC, Guo F, Sánchez Alvarado A. Cellular, ultrastructural and molecular analyses of epidermal cell development in the planarian Schmidtea mediterranea. Dev Biol. 2018;433:357–73.
73. Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat Methods. 2016;13:241–4.
74. Chishti AH, Kim AC, Marfatia SM, Lutchman M, Hanspal M, Jindal H, et al. The FERM domain: a unique module involved in the linkage of cytoplasmic proteins to the membrane. Trends Biochem Sci. 1998;23:281–2.
75. Pasti G, Labouesse M. Epithelial junctions, cytoskeleton, and polarity. WormBook. 2014:1–35. https://doi.org/10.1895/wormbook.1.56.2.
76. van Niel G, Heyman M. The epithelial cell cytoskeleton and intracellular trafficking. II. Intestinal epithelial cell exosomes: perspectives on their structure and function. Am J Physiol Gastrointest Liver Physiol. 2002;283:G251–5.
77. Getz GS, Reardon CA. The mutual interplay of lipid metabolism and the cells of the immune system in relation to atherosclerosis. Clin Lipidol. 2014;9:657–71.
78. Hubler MJ, Kennedy AJ. Role of lipids in the metabolism and activation of immune cells. J Nutr Biochem. 2016;34:1–7.
79. Clark AR, Dean JLE. The control of inflammation via the phosphorylation and dephosphorylation of tristetraprolin: a tale of two phosphatases. Biochem Soc Trans. 2016;44:1321–37.
80. Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, Bork P, et al. KEGG atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res. 2008;36:W423–6.
81. Wasik K, Gurtowski J, Zhou X, Ramos OM, Delás MJ, Battistoni G, et al. Genome and transcriptome of the regeneration-competent flatworm, Macrostomum lignano. Proc Natl Acad Sci. 2015;112:201516718.
82. Novick P, Smith J, Ray D, Boissinot S. Independent and parallel lateral transfer of DNA transposons in tetrapod genomes. Gene. 2010;449:85–94.
83. Tang Z, Zhang H-H, Huang K, Zhang X-G, Han M-J, Zhang Z. Repeated horizontal transfers of four DNA transposons in invertebrates and bats. Mob DNA. 2015;6:3.
84. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 2007;41:331–68.
85. Pang Q, Gao L, Hu W, An Y, Deng H, Zhang Y, et al. De novo transcriptome analysis provides insights into immune related genes and the RIG-I-like receptor signaling pathway in the freshwater planarian (Dugesia japonica). PLoS One. 2016;11:e0151597.
86. Nelson WJ, Dickinson DJ, Weis WI. Roles of cadherins and catenins in cell-cell adhesion and epithelial cell polarity. Prog Mol Biol Transl Sci. 2013;116:3–23.
87. Ikeda DD, Duan Y, Matsuki M, Kunitomo H, Hutter H, Hedgecock EM, et al. CASY-1, an ortholog of calsyntenins/alcadeins, is essential for learning in Caenorhabditis elegans. Proc Natl Acad Sci U S A. 2008;105:5260–5.
88. Ponomareva OY, Holmen IC, Sperry AJ, Eliceiri KW, Halloran MC. Calsyntenin-1 regulates axon branching and endosomal trafficking during sensory neuron development in vivo. J Neurosci. 2014;34:9235–48.
89. Mouton S, Willems M, Houthoofd W, Bert W, Braeckman BP. Lack of metabolic ageing in the long-lived flatworm Schmidtea polychroa. Exp Gerontol. 2011;46:755–61.
90. Hung SS, Parkinson J. Post-genomics resources and tools for studying apicomplexan metabolism. Trends Parasitol. 2011;27:131–40.
91. Wang C, Han X-S, Li F-F, Huang S, Qin Y-W, Zhao X-X, et al. Forkhead containing transcription factor Albino controls tetrapyrrole-based body pigmentation in planarian. Cell Discov. 2016;2:16029.
92. Iyer H, Issigonis M, Sharma PP, Extavour CG, Newmark PA. A premeiotic function for boule in the planarian Schmidtea mediterranea. Proc Natl Acad Sci U S A. 2016;113:E3509–18.
93. Abnave P, Aboukhatwa E, Kosaka N, Thompson J, Hill MA, Aboobaker AA. Epithelial-mesenchymal transition transcription factors control pluripotent adult stem cell migration in vivo in planarians. Development. 2017;144:3440-53.
94. Flores NM, Oviedo NJ, Sage J. Essential role for the planarian intestinal GATA transcription factor in stem cells and regeneration. Dev Biol. 2016;418:179–88.
95. Wang IE, Lapan SW, Scimone ML, Clandinin TR, Reddien PW. Hedgehog signaling regulates gene expression in planarian glia. elife. 2016;5 pii:e16996
96. Lin AYT, Pearson BJ. Yorkie is required to restrict the injury responses in planarians. PLoS Genet. 2017;13 https://doi.org/10.1371/journal.pgen1006874.
97. Davies EL, Lei K, Seidel CW, Kroesen AE, McKinney SA, Guo L, et al. Embryonic origin of adult stem cells required for tissue homeostasis and regeneration. elife. 2017;6 pii:e21052
98. Rodríguez-Esteban G, González-Sastre A, Rojo-Laguna JI, Saló E, Abril JF. Digital gene expression approach over multiple RNA-Seq data sets to detect neoblast transcriptional changes in Schmidtea mediterranea. BMC Genomics. 2015;16:361.
99. Perkel JM. Single-cell sequencing made simple. Nature. 2017;547:125–6.
100. Villani A-C, Satija R, Reynolds G, Sarkizova S, Shekhar K, Fletcher J, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science. 2017;356 pii:eaah4573
101. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015;525:251–5.
102. Wurtzel O, Oderberg IM, Reddien PW. Planarian epidermal stem cells respond to positional cues to promote cell-type diversity. Dev Cell. 2017;40:491–504.e5.
103. Alvarado AS. The Schmidtea mediterranea database as a molecular resource for studying platyhelminthes, stem cells and regeneration. Development. 2002;129:5659–65.
104. Pearson BJ, Eisenhoffer GT, Gurley KA, Rink JC, Miller DE, Alvarado AS. Formaldehyde-based whole-mount in situ hybridization method for planarians. Dev Dyn. 2009;238:443–50.
105. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2017;45:D12–7.
106. Wasmuth JD, Blaxter ML. prot4EST: translating expressed sequence tags from neglected genomes. BMC Bioinformatics. 2004;5:187.
107. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41:D590–6.
108. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40.

109. Burge S, Kelly E, Lonsdale D, Mutowo-Muellenet P, McAnulla C, Mitchell A, et al. Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. Database (Oxford). 2012;2012:bar068.
110. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics England. 2009;25:1754–60.
111. Östlund G, Schmitt T, Forslund K, Köstler T, Messina DN, Roopra S, et al. Inparanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 2010;38:D196–203.
112. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res. 2016;44:D710–6.
113. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
114. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 2017; Available from: https://mafft.cbrc.jp/alignment/server/large.html
115. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.
116. Guindon S, Delsuc F, Dufayard JF, Gascuel O. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol. 2009;537:113–37.
117. Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA. DBD—taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res. 2008;36:D88–92.
118. Blazejewski T, Nursimulu N, Pszenny V, Dangoudoubiyam S, Namasivayam S, Chiasson MA, et al. Systems-based analysis of the Sarcocystis neurona genome identifies pathways that contribute to a heteroxenous life cycle. MBio. 2015;6:e02445–14.
119. NCBI Gene Expression Omnibus Accession. Available from: https://www.ncbi.nlm.nih.gov/gds/?term=GSE115280.
120. McDavid A, Finak G, Chattopadyay PK, Dominguez M, Lamoreaux L, Ma SS, et al. Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinformatics. 2013;29:461–7.
121. High confidence Toronto transcriptome. Available from: http://compsysbio.org/datasets/schmidtea/Toronto_transcriptome.fa.
122. High confidence Toronto transcriptome augmented with Planmine transcripts. Available from: http://compsysbio.org/datasets/schmidtea/Toronto_transcriptome_plus.fa.