OXFORD BROOKES UNIVERSITY

ACME and SPLiT-seq: Quantitative and evolutionary single-cell transcriptomics in planarians

Helena García Castro

Oxford Brookes University

Faculty of Health and Life Sciences

Department of Biological and Medical Sciences

April 2023

A thesis submitted in partial fulfilment of the requirement of

Oxford Brookes University for the degree of

Doctor of Philosophy

Para Paco y Lola

Que me acompañaron en los inicios de mi carrera, allá por 1993.

Y estuvieron presentes en cada paso, literalmente.

Afirmando una y otra vez que no llegarían a ver el siguiente.

Y aquí siguen, de alguna forma extraña y misteriosa.

Y seguirán.

//

Atte. El Mochuelo Doctor

TABLE OF CONTENTS

PUBLICATIONS

I. **García-Castro H,** Kenny NJ, Iglesias M, Álvarez-Campos P, Mason V, Elek A, Schonauer A, Sleight VA, Neiro J, Aboobaker A, Permanyer J, Irimia M, Sebé-Pedrós A, Solana J. ACME dissociation: a versatile cell fixation-dissociation method for single-cell transcriptomics. *Genome Biol* 22, 89 (2021). https://doi.org/10.1186/s13059-021-02302-5

II. **García-Castro H**, Solana J. Single-cell transcriptomics in planaria: new tools allow new insights into cellular and evolutionary features. *Biochem Soc Trans* (2022) BST20210825. https://doi.org/10.1042/BST20210825

III. Leite D, Schonauer A, Blakeley G*, Harper A*, **Garcia-Castro H***, Baudouin-Gonzalez L, Wang R, Sarkis N, Nikola A, Koka VSP, Kenny N, Turetzek N, Pechmann M, Solana J, McGregor A. An atlas of spider development at single-cell resolution provides new insights into arthropod embryogenesis. *bioRxiv* 2022.06.09.495456. https://doi.org/10.1101/2022.06.09.495456

IV. **García-Castro H**, Emili E, Solana J. ACME dissociation-fixation, Flow cytometry and cell sorting of freshwater planarian cells. In L. Gentile. *The Schmidtea mediterranea Toolbox*. Pending of publication.

V. Álvarez-Campos P*, **García-Castro H***, Emili E, Pérez-Posada A, Salamanca-Díaz D, Mason V, Metzger B, Kenny N, Özpolat D, Solana J. Annelid adult cell type diversity and their pluripotent cellular origins.

** Equal contribution*

CONTRIBUTIONS

The results presented in this thesis have been generated thanks to the contribution of multiple scientists. Given the collaborative nature of the projects, as well as for stylistic reasons, some sections have been written in plural. Nonetheless, the author takes credit for most of the work presented here. All contributions are specified below:

General:

- Conception of projects and experimental design: **Jordi Solana**
- FACS facility operation: **Helen Ferry**, **Liam Hardy**, **Michal Maj** and **Robert Hedley**
- Genome annotation of *Schmidtea mediterranea* and *Dugesia japonica:* **Jakke Neiro**

Chapter II:

- Contribution to figures: **Jordi Solana** and **Nathan Kenny**
- Contribution to the evaluation of reducing agents: **Vincent Mason**
- ACME-dissociation in non-planarian model organisms: **Anna Schönauer**, **Jordi Solana, Marta Iglesias**, **Patricia Álvarez-Campos** and **Victoria Sleight**.
- ScRNA-seq analysis of cnidarian ACME-cells: **Anamaria Elek**, **Arnau Sebé-Pedrós** and **Marta Iglesias**.
- Bioinformatic analysis of planarian species: **Nathan Kenny**

Chapter III:

- GFP DNA miniprep: **Lorena Martínez Quiles**
- Contribution to dsRNA synthesis: **Vincent Mason**
- Microinjection, ACME-dissociation and monitoring (one replicate): **Elena Emili**
- Pictures from Figure 3.6: **Elena Emili**
- Data pre-processing (Linux): **Nathan Kenny**
- Contribution to data analysis: **Jordi Solana**
- DEseq2 script: **María Roselló**

Chapter IV:

- Transcriptome assembly and annotation of *Schmidtea polychroa, Girardia tigrina* and *Polycelis nigra:* **Nathan Kenny**
- Data pre-processing (Linux) of *Dugesia japonica, Girardia tigrina* and *Polycelis nigra*: **Nathan Kenny**

ABBREVIATIONS

ACME	Acetic-Methanol
aqp	Aquaporin
BrdU	Bromodeoxyuridine
BSA	Bovine serum albumin
CAGE	Cap analysis of gene expression
cDNA	Complementary DNA
cNeoblasts	Clonogenic neoblasts
ConA	Concanavalin A
CPM	Count per million
DEGs	Differentially expressed genes
DMSO	Dimethyl sulfoxide
DPBS	Dubbeco's PBS
dsRNA	Double-stranded RNA
DTT	Dithiothreitol
DVb	Dorsoventral boundary
ESTs	Expressed sequence tags
FA	Formaldehyde
FACS	Fluorescence-activated cell sorting
FISH	Fluorescence *in situ* hybridization
FPKM	Fragments per kilobase of mapped reads
FSC	Forward scatter
HNF4	Hepatocyte nuclear factor 4
HVGs	Highly variable genes
IF-1	Intermediate filament 1
KNN	K-nearest neighbour
ldlrr-1	Low density lipoprotein receptor-related 1
mRNA	Messenger RNA
Mya	Million years ago
NAC	N-acetyl-L-cysteine
NGS	Next-generation sequencing
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PBS	Phosphate-buffered saline

PCA	Principal component analysis
pgrn	Progranulin
psap	Prosaposin
psd	Pleckstrin and sec7 domain-containing
RIN	RNA integrity number
RISC	RNA-induced silencing complex
RNAi	RNA interference
RNA-seq	RNA sequencing
rRNA	Ribosomal RNA
RT	Room temperature
RT-PCR	Reverse transcription polymerase chain reaction
SAGE	Serial analysis of gene expression
scRNA-seq	Single-cell RNA-sequencing
SLC	Solute carrier
SMRT	Single-molecule real-time
snRNA-seq	Single-nuclei RNA-seq
SPLiT-seq	Split-pool ligation-based RNA-seq
TCEP	Tris (2-carboxyethyl) phosphine
TPM	Transcripts per million
TSS	Transcriptional start sites
UMI	Unique molecular identifier
X1	X-ray sensitive proliferating neoblasts
X2	X-ray sensitive cell progenitors
XIS	X-ray insensitive differentiated cells
ZMW	Zero-mode waveguides

ABSTRACT

The last decade has seen an exponential development of scRNA-seq technologies. This thesis begins by recapitulating the history of transcriptomics, from its early years to the development of modern single-cell protocols, in order to understand the evolution and current state of the field. During the experimental part, I focus on the description and validation of a versatile pipeline for scRNA-seq. This combines a novel sample preparation strategy (ACME) with a powerful *in situ* barcoding platform (SPLiT-seq) to overcome some of the restrains of current single-cell protocols. ACME provides simultaneous tissue dissociation, fixation and permeabilization, resulting in cells that can be cryopreserved, sorted by FACS and used in multiple platforms. As proof of concept, ACME-cells are combined with SPLiT-seq (Rosenberg et al., 2018), obtaining over 32,000 cell transcriptomes and profiling two comprehensive single-cell atlases for the planarian species *Schmidtea mediterranea* and *Dugesia japonica*.

Later, I use the same pipeline in a single-cell RNA interference study, in *S. mediterranea*, to describe and quantify the effects of the *hnf4* knockdown at tissue-resolution. Our results reveal cellular and genetic changes in the gut and parenchymal populations of knockdown animals, proving that our transcriptomic approach can finely dissect tissue specific effects in knockdown conditions. In the last part of the thesis, I present the single-cell atlases of multiple planarians (*Schmidtea mediterranea*, *Schmidtea polychroa, Dugesia japonica, Girardia tigrina* and *Polycelis nigra*), and compare the differences between sexual and asexual strains and life-history stages of the same species. This data includes over 116,000 cell transcriptomes, and sets the bases for a future cross-species comparison that will explore the diversity of cell types and gene expression in planarians at single-cell resolution. Overall, these examples show how single-cell transcriptomics is moving toward the end of the era of descriptive cell type atlases into more complex quantitative and evolutionary studies.

CHAPTER I: INTRODUCTION

1. BRIEF HISTORY OF TRANSCRIPTOMICS

Transcriptomics is the study of the **transcriptome**. The term 'transcriptome' was coined by Charles Auffray in 1996 to describe the sum of all the RNA molecules of an organism -known as **transcripts**- (Piétu et al., 1999). In recent decades, **transcriptomic technologies** have revolutionized biology by making it possible to measure gene expression under different conditions. This has greatly expanded our understanding of disease mechanisms, development, transcription regulation and cell physiology, among other fields.

1.1. MAKING RNA INTO cDNA

RNA is the raw material of transcriptomics, but it is highly prone to degradation by RNases. Fortunately, RNA can be reverse transcribed into more stable **complementary DNA (cDNA)** using reverse transcriptase enzymes. Reverse transcriptases, also known as viral RNA-dependent DNA polymerases, were discovered independently by David Baltimore and Howard Temin in 1970 (Baltimore, 1970; Coffin and Fan, 2016). The **reverse transcription** mechanism was first used in 1971 to synthesize cDNA from unspecific RNA templates (Spiegelman et al., 1971). Later, the use of synthetic oligos(dT)s was implemented to prime the reverse transcription of **messenger RNAs (mRNAs)** by binding to their poly-A 3'-ends (Diggelmann et al., 1973; Ross et al., 1972). Nowadays, reverse transcription protocols are routinely used in the lab, but their invention was key for the development of transcriptomics.

1.2. EARLY YEARS OF TRANSCRIPTOMICS

Decades before the development of modern transcriptomics, RNA was studied by traditional hybridization and Sanger sequencing. **Hybridization** protocols hybridize target RNAs -or cDNAs- with complementary DNA probes. The design of these probes requires prior knowledge of the sequence, limiting the use of these techniques. On the contrary, the approaches based on **Sanger sequencing** can be used in any organism, as they directly determined the sequence of the transcript (Wang et al., 2009).

1.2.1. HYBRIDIZATION-BASED APPROACHES

One of the oldest techniques for the study of RNA is the **Northern blot** protocol (Alwine et al., 1977), which is based on the analogue Southern blot protocol for DNA separation (Southern, 1975). Northern blot was designed to detect specific RNA molecules in a complex mixture (**Figure 1.1**). After years of being the standard technique for RNA studies, Northern blot was

replaced by newer protocol due its disadvantages. Despite its high specificity and simplicity, Northern blot is time-consuming, low-sensitive (not suitable for low expression transcripts) and prone to RNases degradation. Moreover, it can only process one or few genes at a time (Alwine et al., 1977; Moustafa and Cross, 2016).

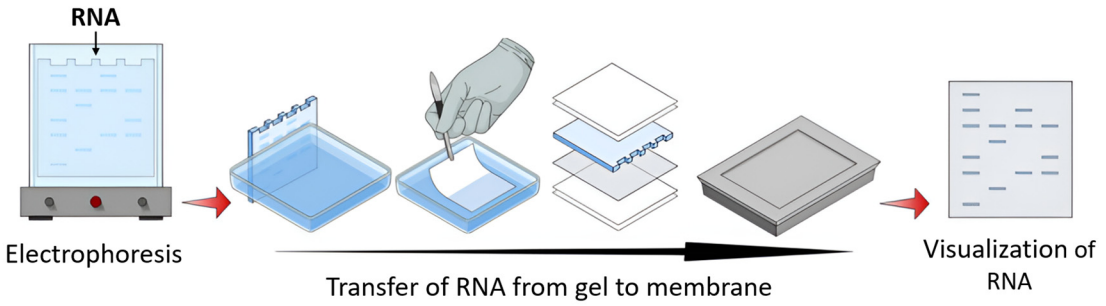


Figure 1.1 Northern blot scheme. The RNA is separated by electrophoresis and then transferred to a membrane, where it hybridizes with complementary DNA probes, labelled radioactively or chemically. The membrane is revealed by X-ray. *Adapted from: https://www.thesciencenotes.com/northern-blotting-principle-procedure-and-applications/*

In the 1980s, the **reverse transcription-polymerase chain reaction (RT-PCR)** gained popularity to quantify gene expression (Becker-André and Hahlbrock, 1989). RT-PCR combines RNA reverse transcription with the polymerase chain reaction technology, developed by Kary Mullis in 1985 (Saiki et al., 1988). RT-PCR presents multiple advantages over the Northern blot. It is fast, highly specific, high-sensitive, and suitable for amplification of very small amounts of RNA (Moustafa and Cross, 2016). Although RT-PCR can only process a few transcripts at a time, this technique remains the standard for multiple applications. A very illustrative example is the use of real-time RT-PCR in the detection of viral RNA sequences during the Covid-19 pandemic (Chung et al., 2021).

1.2.2. SANGER SEQUENCING-BASED APPROACHES

In 1977, Frederick Sanger and Walter Gilbert independently developed the **first-generation sequencing** methods for long DNA molecules (Barba et al., 2014; Maxam and Gilbert, 1977; Sanger et al., 1977). The most successful one was the **Sanger sequencing**, named after its author. This technology was soon applied to RNA studies, with the creation of **Expressed Sequence Tags (ESTs)** (Putney et al., 1983). ESTs are collections of short fragments (200-800 bp) -or **reads**- sequenced from selected clones of a cDNA library. ESTs are therefore a representation of gene expression in the library. Among other applications, ESTs were used for transcript profiling and gene discovery (Parkinson and Blaxter, 2009).

After ESTs, a more efficient method called **Serial Analysis of Gene Expression (SAGE)** was developed (Velculescu et al., 1995). In SAGE, the cDNA is digested into short tags (13-20 bp) by restriction enzymes. These tags are concatenated into longer fragments (>500 bp) and sequenced by Sanger. Concatenation allows to sequence a larger number of cDNA molecules, increasing the throughput. After sequencing, the long reads are separated again into short tags, and each tag is aligned to the reference genome to identify the gene. When the genome is not available, tags are simply used as a catalogue of markers for a certain stage or condition (Lowe et al., 2017; Velculescu et al., 1995). A variation of SAGE, known as **Cap Analysis of Gene Expression (CAGE)**, was later described (Shiraki et al., 2003). While SAGE tags derived from the 3'-end of the cDNAs, CAGE tags capture the 5'-end. Thus, CAGE additionally allows the identification of transcriptional start sites (TSS) and the analysis of the promoter usage.

1.3. THE ERA OF MICROARRAYS

From the mid-1990s to late 2000s, microarrays became the leading technology for the study of transcriptomics. Microarrays evolved from 1970s low-throughput **filter-based arrays protocols**. On those, cloned colonies were immobilized on paper filters and hybridized to RNA -or DNA- probes to detect target sequences (Gergen et al., 1979; Grunstein and Hogness, 1975). These protocols were progressively refined and automatized (Craig et al., 1990; Lennon and Lehrach, 1991) to create the microarray technology (Schena et al., 1995). **Microarrays** are chips, made of glass, plastic or silicon, with multiple probes attached to it. These probes are sets of short oligos (25-60 bp) used to hybridize complementary DNA or cDNA molecules labelled by fluorescence (Bumgarner, 2013) (**Figure 1.2**). The intensity of the **fluorescence signal** is quantified by computerized image analyses to estimate the abundance of the hybridized transcripts (Lowe et al., 2017).

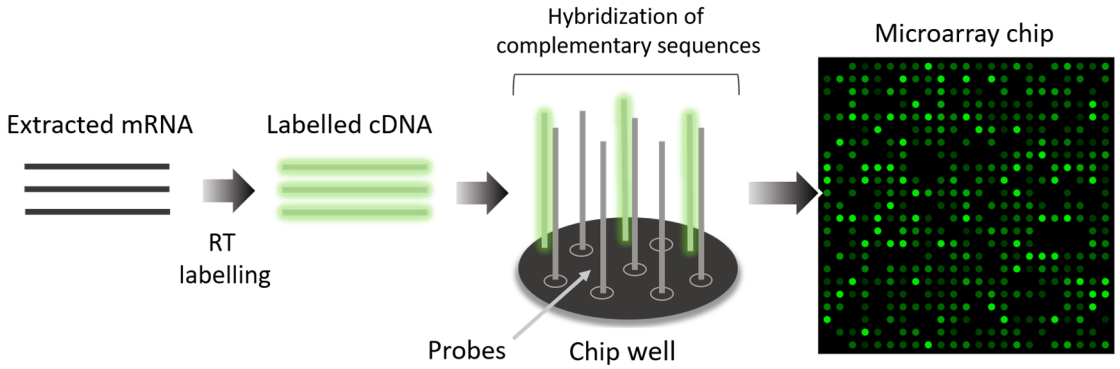


Figure 1.2 Scheme of microarrays performance. The isolated RNA is reverse transcribed into cDNA and labelled by fluorescence. Target cDNA molecules hybridize to the complementary probes attached to the microarrays chip. The fluorescence signal represents the abundance of the target transcript. *Microarray chip credit: Thomas Shafee, CC BY 4.0 <https://creativecommons.org/licenses/by/4.0>, via Wikimedia Commons.*

First microarrays were custom made. Then, private companies began to commercialize high-throughput chips to detect thousands of transcripts in parallel. The most popular choice during the golden age of microarrays was the **Affymetrix GeneChip®**. It was launched in 1994 and pioneered the use of photolithography for *in situ* probe printing (Gershon, 2004; Lipshutz et al., 1999). Commercial platforms helped to standardize and scale up microarrays to the point of offering **whole-genome chips**. In 2003, NimbleGen released the first whole-human-genome chip. Other companies soon followed, realising platforms for human, mouse, rat, and other economically relevant species (Gershon, 2004).

However, microarrays are not without drawbacks. The commercial chips, reagents and equipment needed to run the experiments are expensive. Moreover, probes design still requires prior knowledge of the sequence, which limits the use of microarrays to well-known species. From the 2010s, these limitations caused microarrays to lose ground to sequencing-based technologies, which had experienced a great development and cost reduction (Lowe et al., 2017; Moustafa and Cross, 2016).

1.4. NEXT GENERATION SEQUENCING

Next-generation sequencing (NGS) is the name given to the high-throughput sequencing technologies developed since the 2000s. NGS revolutionized genomics and transcriptomics, and quickly replaced the Sanger method by offering a much bigger throughput per experiment at a lower price. In 2005, **Roche 454**, based on pyrosequencing, became the first commercial NGS platform. One year later, **Solexa** launched the Genome Analyzer, a sequencer with a capacity of 1 Gb per run. During the following years, numerous companies launched their own NGS platforms with enhanced features and increasing throughputs. **SoLiD** (Applied Biosystems, 2008), **Ion Torrent** (LifeScience Technologies, 2010) or **HiSeq** (Illumina, 2010) are some of the most well-recognized (Liu et al., 2012; Lowe et al., 2017).

Since **Illumina** acquired Solexa in 2007 (Liu et al., 2012), this company has become one of the major actors of NGS. Nowadays, **Illumina** is leading the short-read sequencing market (**Figure 1.3.A**). Its most powerful platform to date is **NovaSeq 6000**, launched in 2017, which has capacity to generate up to 6 Tb of information (20 billion reads) per run (https://www.illumina.com/systems/sequencing-platforms/novaseq.html).

1.5. THIRD-GENERATION SEQUENCING

Since the 2010s, a third-generation of sequencing technologies has been developed by **Pacific Bioscience** and **Oxford Nanopore Technologies**. Compared to the previous generation, third-generation platforms present many advantages. First, they do not require prior PCR or bridge amplification, which reduces sample preparation time and avoids amplification bias. Second, **sequences are read in real-time**. Finally, third-generation sequencing generates much longer reads than their predecessors, earning the name of **long-read sequencing**. In contrast, they still offer lower throughputs and higher error rates (~15%) than second-generation platforms (Liu et al., 2012; van Dijk et al., 2018).

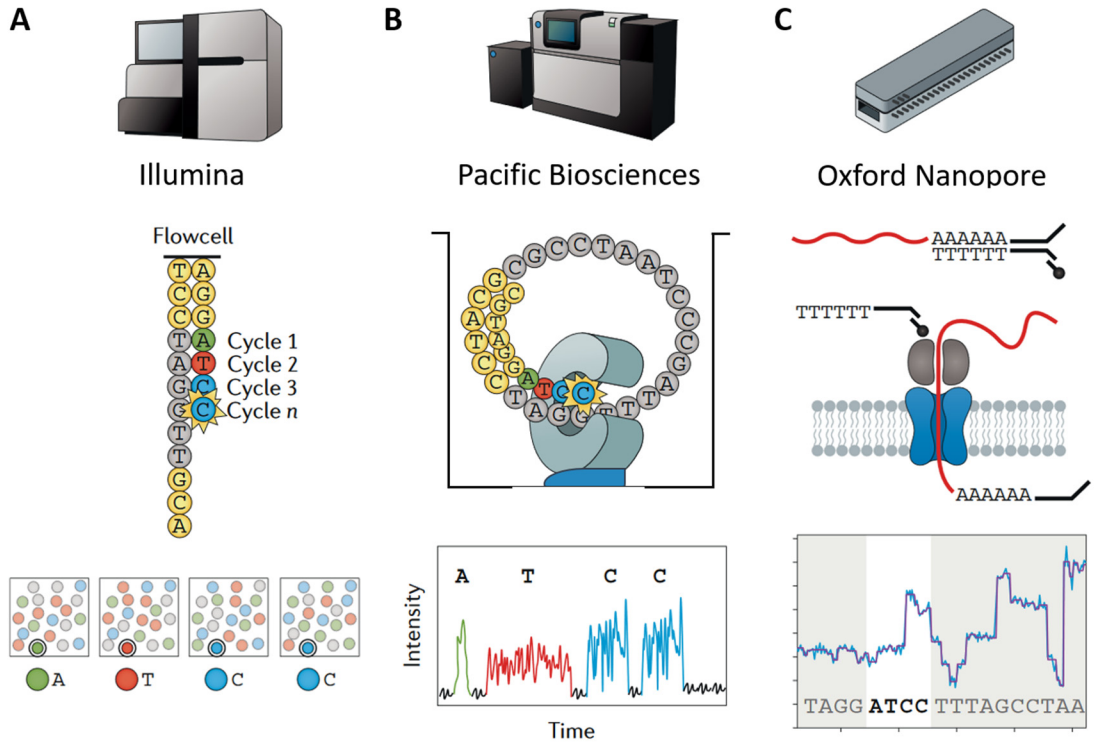


Figure 1.3 Overview of different NGS technologies. A) Illumina: Input cDNA is amplified by bridge PCR on a flow cell. The resulting clusters are sequenced through the detection of fluorescence emitted by the nucleotides incorporated during the synthesis of the complementary strand **B)** Pacific Bioscience: DNA polymerases, embedded in ZMW wells, synthesize the complementary strand of a circular cDNA template incorporating fluorescent nucleotides. Emitted florescence is read in real-time within the wells **C)** Oxford Nanopore: Molecules are polymerase translocated through the nanopores of an electro-resistant membrane. The electrical changes caused by each nucleotide are characterized in real-time. No fluorescence signal is involved. *Adapted from Stark et al., 2019.*

In 2011, **Pacific Biosciences (PacBio)** launched PacBio RS, the first commercially successful third-generation sequencer (van Dijk et al., 2018). PacBio systems are based on two different technologies: single-molecule real-time (SMRT) sequencing (Eid et al., 2009) and zero-mode waveguides (ZMW) (Levene et al., 2003). In brief, circular input templates are captured by DNA

polymerases immobilized on ZMW wells. Here, the polymerase synthesizes the complementary strand of the template using fluorescently tagged dNTPs. The platform identifies the colour of the fluorescence emitted during the incorporation of each nucleotide, reading the synthesis in real-time (Eid et al., 2009; van Dijk et al., 2018) (**Figure 1.3.B**).

On the other hand, **Nanopore Sequencing** (Manrao et al., 2012), offered by **Oxford Nanopore Technologies (ONT)**, is becoming increasingly popular. Nanopore sequencing consists of a porous lipidic membrane with a DNA polymerase embedded to each nanopore. When a DNA (or RNA) molecule passes through these nanopores, the changes in electrical current caused by the translocation of each nucleotide are detected and characterized in real-time (**Figure 1.3.C**). ONT is fluorescence free and only relies on current changes across the membrane (van Dijk et al., 2018). The read length is technically unlimited, and depends on the length and quality of the input fragment (Jain et al., 2018). Since 2013, ONT has released a series of different scale nanopore sequencing devices: **MinION** (up to 50 Gb/run), **GridION** (up to 250 Gb/run) and **PromethION** (up to 14 Tb/run) (https://nanoporetech.com/products). MinION is currently the only **pocket-size sequencer** on the market. Its portability, affordable price (from 1,000 USD) and simplicity make it one of the most interesting sequencing options for small labs, developing countries and field applications. For instance, in 2015 MinION was used in Guinea for real-time screening of Ebola during an epidemic outbreak (Quick et al., 2016).

1.6. RNA SEQUENCING

With the development of NGS platforms, RNA began to be massively sequenced. This approach was called **RNA sequencing (RNA-seq)**. First RNA-seq-like publication used a Roche 454 sequencer to profile the expression of 10.000 genes from a prostate cancer cell line (Bainbridge et al., 2006). Later, the technique was fully described and used to create the first transcriptome of *Yeast* using Illumina sequencing (Nagalakshmi et al., 2008). Since then, the use of RNA-seq grew exponentially, becoming the leading technology in transcriptomics and displacing microarrays (Lowe et al., 2017).

RNA-seq protocols start with the extraction of the RNA and its reverse transcription (RT) into cDNA. Most sequencing platforms use cDNA as input material, except ONT, which can perform direct RNA-sequencing (Smith et al., 2019; Zhao et al., 2019). Usually, mRNAs are enriched during RT using oligo(dT) primers. This reduces the capture rate of abundant ribosomal RNAs (rRNAs). After cDNA synthesis, library preparation frequently includes steps of adapters ligation, PCR amplification, cDNA fragmentation and size selection (100-500 bp), to make the input cDNA compatible with the sequencing platform. Third-generation sequencing technologies can

dispense of fragmentation and PCR amplification steps. When the library is ready, the cDNA is sequenced in one direction (single-end) or bi-directionally (paired-end), depending on the application. The resulting sequences, called **reads**, will have a variable length depending on the technology used (Lowe et al., 2017; van Dijk et al., 2018; Wang et al., 2009).

RNA-seq provides several advantages over microarrays. First, the background signal is extremely low, and the initial amount of cDNA required is minimum. The quantification of gene expression is also highly accurate, because it correlates with the total number of unique reads mapped per gene (van Dijk et al., 2018; Wang et al., 2009). In addition, RNA-seq does not require prior knowledge of the sequence and can therefore be used in non-model organisms (Carruthers et al., 2018; Chen et al., 2014). RNA-seq power and versatility makes it suitable for multiple applications, including gene identification (Chen et al., 2014), differential gene expression (Chen et al., 2016), diagnostics (Best et al., 2015), transcriptome *de novo* assembly (Carruthers et al., 2018), epitranscriptomics (Zhao et al., 2019), and the study of non-coding RNA (Westermann et al., 2016), alternative splicing (Pan et al., 2008) or single nucleotide polymorphisms (López-Maestre et al., 2016), among others.

1.7. THE EMERGENCE OF SINGLE-CELL TRANSCRIPTOMICS

Single-cell transcriptomics, also called **single-cell RNA-sequencing (scRNA-seq)**, emerged in recent years as an evolution of RNA-seq. Single-cell transcriptomics can profile gene expression in individual cells, allowing **single-cell resolution** studies. Through this new technology, we can resolve and compare cellular identities and raise a whole new repertoire of biological questions.

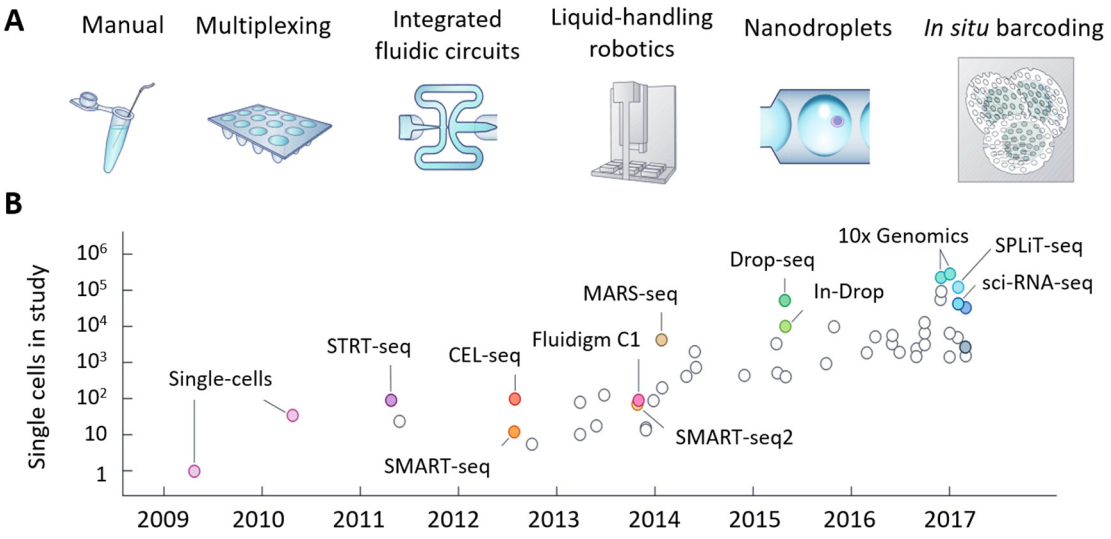


Figure 1.4 Historical evolution of single-cell RNA-seq technologies. A) Main technological developments on the field **B)** Exponential scaling of the technique, showed as number of cells per study and per year. *Adapted from Svensson et al., 2018.*

However, scRNA-seq data is also more complex, expensive to produce, and difficult to interpret. Because of this, in spite of the revolution of single-cell technologies, RNA-seq remains as a valuable tool in multiple applications. To distinguish from single-cell approaches, traditional RNA-seq is now referred to as **bulk RNA-seq** (G. Chen et al., 2019; Stark et al., 2019).

Despite being a recently developed technology, scRNA-seq has experienced a vast evolution. In 2009, the first scRNA-seq paper profiled the transcriptome of a single mouse blastomere, manually isolated in an **Eppendorf tube** (Tang et al., 2009). Since then, the number of cells profiled per experiment has grown exponentially, and the technique has gained great popularity among the scientific community (Svensson et al., 2018) (**Figure 1.4**). ScRNA-seq protocols synthesise cDNA from poly-A captured mRNA, and label this cDNA with short DNA-sequences, called **barcodes**, to identify their origin. Other than this, each single-cell approach follows different strategies to isolate the cells and process the cDNA. Depending on where the reverse transcription and barcoding take place, we can classify scRNA-seq protocols into three main categories: **plate-based**, **droplet-based**, and ***in situ* barcoding-based** (Griffiths et al., 2018; Svensson et al., 2018).

1.7.1. PLATE-BASED METHODS

In plate-based protocols, cells are isolated and processed within the wells of a plate (or chip). First plate-based publication profiled the transcriptome of 85 cells in one experiment (Islam et al., 2011). For this, each cell was **manually pipetted** into a 96-well PCR plate, and tagged with a **well-specific barcode** introduced at the 3'-end of the cDNA (**Figure 1.5**). After tagging, cDNA molecules were pooled in a single tube and amplified by PCR before sequencing.

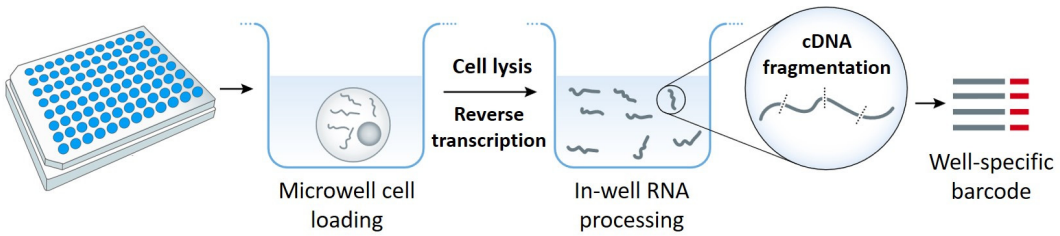


Figure 1.5 RNA processing in plate-based methods. Single cells are isolated on the microwells of a plate (or chip). Cell lysis, reverse transcription and cDNA cell-specific barcoding occurs in-well. *Adapted from Griffiths et al., 2018.*

In the following years, other plate-based protocols were developed to incorporate novel features for cDNA processing. For instance, **CEL-seq** (Hashimshony et al., 2012), introduced ***in vitro* transcription (IVT)** as an alternative to PCR amplification (Beckert and Masquida, 2011).

IVT reduces the initial amount of RNA required per experiment. Later platforms, such as MARS-seq or InDrop, also incorporated IVT in their protocols (Svensson et al., 2018).

In parallel to CEL-seq, **Smart-seq** (Ramsköld et al., 2012) was introduced as the first strategy to capture **full-length transcripts**. This opened the door to single-cell studies on single nucleotide polymorphisms and alternative splicing. Since then, improved versions of these protocols have been published: **CEL-seq2** (Hashimshony et al., 2016), **Smart-seq2** (Picelli et al., 2013) and **Smart-seq3** (Hagemann-Jensen et al., 2020).

Meanwhile, other approaches focused on **cell capture innovation** to substitute manual pipetting. In 2012, **Fluidigm C1** was the first single-cell technology based on **microfluidic cell capture**, and the first commercial platform for scRNA-seq. The C1 system distributes cell suspensions through an integrated fluidic circuit that isolates each cell into the small reaction chamber of a chip (**Figure 1.6**). At present, Fluidigm C1 devices run on Smart-seq or on a strategy known as Seq-HT, and have a maximum capacity of 800 cells per run (Chen and Ginhoux, 2018; Svensson et al., 2018).

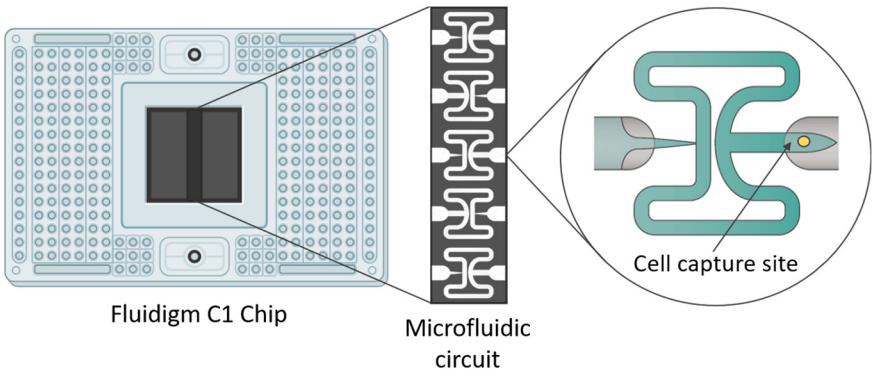


Figure 1.6 Fluidigm C1 chip. Detail of the microfluidic circuit and cell capture sites. *Adapted from Stark et al., 2019 using BioRender.com*

In 2014, the automation of scRNA-seq by **liquid handling robots** was introduced with **MARS-seq** (Jaitin et al., 2014). Robotic automation substituted some manual pipetting steps, and helped to increase throughputs up to a few thousand cells per experiment. In MARS-seq, single-cells are isolated into 384-well plates by **fluorescence-activated cell sorting (FACS)**. Additionally, the improved **MARS-seq2.0** (Keren-Shaul et al., 2019) incorporated **indexed FACS** to separate cells according to a surface marker. Indexed FACS enables depletion or enrichment of cell populations, facilitating the study of low-abundant cell-types.

After cell isolation and lysis, each cDNA molecule is tagged with plate-, cell- and molecular-specific barcodes. **Multi-tagging** innovation increased multiplexing capacity -here understood as cell processing capacity-, and introduced the molecular barcode, also known as **UMI (Unique**

Molecular Identifier), as a new concept for single-cell transcriptomics. The incorporation of UMIs allows the precise quantification of gene expression as it removes biases introduced by PCR amplification. Reads with repeated UMIs will correspond to the same original cDNA molecule, amplified multiple times (Chen and Ginhoux, 2018).

1.7.2. DROPLET-BASED METHODS

Droplet-based methods are based on **microfluidic cell capture** and were a significant step forward for single-cell transcriptomics. In 2015, two different labs independently developed the first droplet-based protocols: **InDrop** (Klein et al., 2015) and **Drop-seq** (Macosko et al., 2015). Both protocols use a microfluidic device with different flow currents for cells, synthetic barcoded beads, reagents and oil. Where these flows merge, a series of nanolitre droplets is created. The flow is adjusted in such a way that, ideally, a single cell is encapsulated into one droplet, together with a single barcoded bead and the reagents required for cell lysis and reverse transcription, which occur within the droplet (**Figure 1.7**).

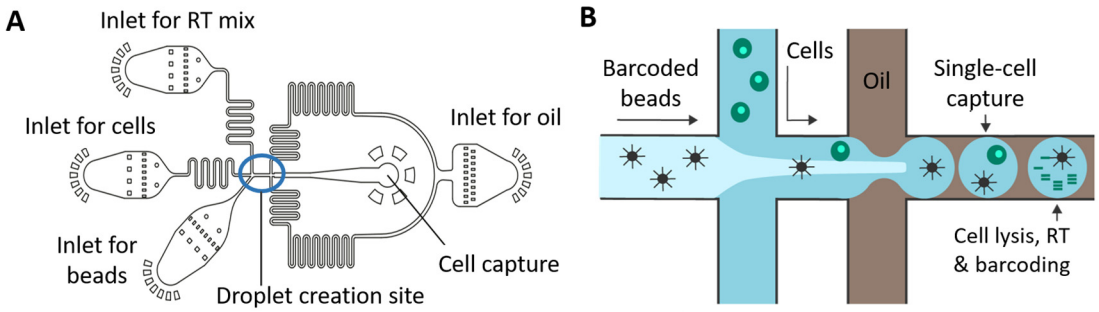


Figure 1.7 Single-cell microfluidic device. A) InDrop whole microfluidic circuit. *Adapted from Klein et al., 2015* **B)** Close-up of droplet creation and subsequent cell processing based on the Drop-seq system.

During reverse transcription, cDNAs are labelled with primers containing a common cell barcode and a unique molecular identifier (UMI). InDrop uses hydrogel microspheres to carry the primers, which are released after encapsulation. Meanwhile, Drop-seq uses beads with primers attached to their surface. After barcoding, droplets are broken, and all tagged cDNAs are pooled and amplified by PCR (Drop-seq) or IVT (InDrop) before sequencing (Klein et al., 2015; Macosko et al., 2015).

Originally, droplet methods used in-lab-operated custom microfluidic devices, with complex handling and library preparations and low rates of cell capture. However, commercial options were soon available to overcome these drawbacks. In 2015, the company **10x Genomics** presented a droplet platform built on their **GemCode** technology (Zheng et al., 2017), previously used for genome assembly and haplotyping. GemCode highly resembles InDrop and Drop-seq

protocols. The main novelty of this platform is the use of an **8-channel microfluidic chip**, which yields several thousand cells per experiment, allows the simultaneous processing of up to 8 different samples (one per channel), and improves encapsulation rate, with ~50-65% of total input cells being correctly captured.

10x Genomics Chromium, an improved version of the GemCode platform, was launched in 2016. Nowadays, 10x Chromium is **leading the market** as the most popular option for single-cell transcriptomics (Svensson et al., 2020). 10x Genomics has automatized and externalized single-cell experiments almost entirely. The only step that still takes place in the lab is sample preparation, which also follows 10x Genomics standardized protocols (https://www.10xgenomics.com/support/single-cell-gene-expression).

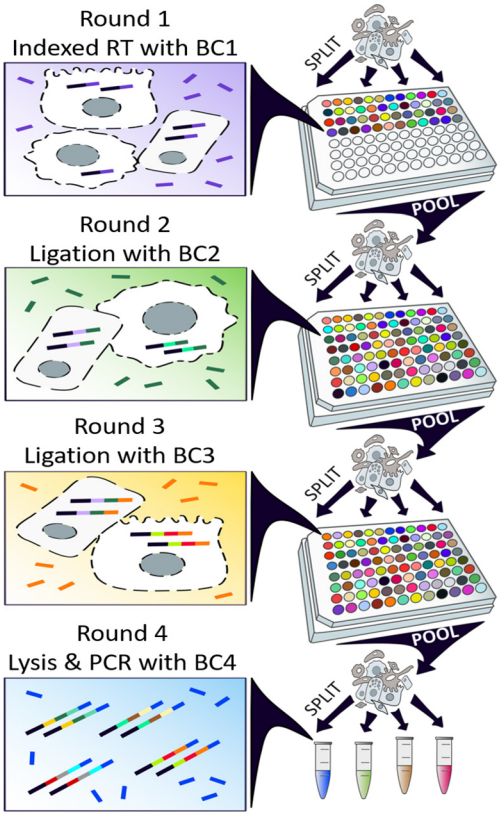
1.7.3. *IN SITU* BARCODING-BASED METHODS

***In situ* barcoding** is the most recent approach to single-cell transcriptomics. In these protocols, reverse transcription and cDNA labelling occur ***in situ* within the cell** -or nucleus-. Therefore, cells need to be fixed and permeable to allow barcodes and reagents to penetrate the membrane and reach the RNAs (Svensson et al., 2018). After barcoding, cells are lysed and library preparation continues as in bulk RNA-seq.

The barcoding process is based on **combinatorial indexing**, also known as **split-pool barcoding**. This system avoids the use of complex cell capturing devices, as cell isolation is no longer required. In combinatorial indexing, the initial pool of cells is randomly split in a plate, and all cells falling in the same well receive the same well-specific barcode. Cells are then pooled together and split again in another plate with a different set of barcodes, and the labelling process is repeated. After multiple rounds, each cell is randomly tagged with a unique combination of barcodes, as the probability of two cells falling consecutively into the same wells, and receiving the same barcodes, is minimal. Another strength of combinatorial indexing is its **high scalability**. Theoretically, cell capacity is only limited by the number of possible barcode combinations. Therefore, by increasing the number of barcodes or indexing rounds this approach could profile millions of cells per experiment (Cao et al., 2017; Rosenberg et al., 2018).

First *in situ* barcoding protocols, **sci-RNA-seq** (Cao et al., 2017) and **SPLiT-seq** (Rosenberg et al., 2018), were developed in parallel in 2017. The original **sci-RNA-seq** protocol counts with two indexing rounds (during RT and PCR amplification), and UMI sequences (Cao et al., 2017). The latest version, **sci-RNA-seq3** (Cao et al., 2020; Martin et al., 2021), incorporates an extra indexing round by hairpin ligation. In terms of cell numbers, sci-RNA-seq papers have the most impressive

datasets to date. In recent years, sci-RNA-seq3 has been used to profile the transcriptome of 2 million cells (Cao et al., 2019) and 4 million cells (Cao et al., 2020) from mice embryos and human fetal organs, respectively.



SPLiT-seq proof-of-concept paper, on the other hand, covered more than 150,000 cells from mouse brains and spinal cords, and identified over 100 distinct cell types (Rosenberg et al., 2018). SPLiT-seq uses four indexing rounds. A first indexed RT reaction, two T4 ligations and a final indexed PCR amplification. The UMI is introduced, together with the third barcode, during the second ligation (**Figure 1.8**). Some of the authors of SPLiT-seq founded **Parse Biosciences** to commercialize the technique. The company offers kits for sample fixation, barcoding and library preparation, making SPLiT-seq the first *in situ* barcoding technology commercially available (https://www.parsebiosciences.com/ technology).

Figure 1.8 SPLiT-seq workflow. Cells are randomly split in 96-well plates containing unique barcodes. The first reverse transcription captures mRNAs using an anchored-poly dT barcode (BC 1). The second (BC 2) and third barcodes (BC 3) are ligated to the cDNA in two subsequent reactions. The last barcode (BC 4) is added in a final indexed PCR reaction after cell lysis and tagmentation.

1.8. MULTIOMICS AND SPATIAL TRANSCRIPTOMICS

The future of transcriptomics points in two directions: the integration of **multiomic data** and the resolution of **spatial information**. Like scRNA-seq has become the most powerful tool in transcriptomics, other -omics have experienced equivalent revolutions driven by the rise of high-throughput technologies. For instance, the development of **ChIP-seq** (Robertson et al., 2007), **CUT&RUN** (Skene and Henikoff, 2017) or single-cell **ATAC-seq** (Buenrostro et al., 2015; Cusanovich et al., 2015) have revolutionized epigenomics, changing the way we study DNA-protein interactions and chromatin accessibility.

Multiomics aims to combine all these cutting-edge technologies in genomics, transcriptomics, proteomics, epigenomics and metabolomics to achieve more comprehensive biological insights. Multiomic studies can result from the **integration of data** of different -omics (Ranzoni et al.,

2021), or the performance of **hybrid protocols** that capture different biological features. For instance, **CITE-seq** (Stoeckius et al., 2017) and **REAP-seq** (Peterson et al., 2017) pair scRNA-seq and protein measurement. Similarly, **SHARE-seq** combines scRNA-seq with scATAC-seq for simultaneous gene expression and chromatin accessibility studies (Ma et al., 2020). In bioinformatics, novel tools and analysis pipelines are emerging to deal with the challenges of integrating multiomic data (Hao et al., 2021; Lin et al., 2022).

Spatial transcriptomics, on the other hand, study the transcriptome at spatial resolution. This is one of the major limitations of bulk and single-cell transcriptomic, which fail to preserve spatial information after tissue dissociation (Longo et al., 2021). One of the simplest approaches of spatial transcriptomics uses **laser-capture microdissection** to isolate specific tissue sections (Brosch et al., 2018). Spatial information can also be retrieved using **high-plex RNA imaging** techniques, like MERFISH (Chen et al., 2015) or seqFISH+ (Eng et al., 2019). These protocols use fluorescence *in situ* hybridization (FISH) to image panels of a few hundred transcripts on small histological sections. Finally, **spatial barcoding** approaches capture the transcripts *in situ* and label them with a location barcode. For this, permeabilized tissue sections are positioned over an array that carries the location barcodes. 'Spatial Transcriptomics' (Ståhl et al., 2016), Slide-seq (Rodriques et al., 2019; Stickels et al., 2021) and Visium, offered by 10x Genomics (https://www.10xgenomics.com/products/spatial-gene-expression), are current examples of spatial barcoding platforms.

Spatial transcriptomics are still limited by optical resolution, low coverage, low depth, and the requirement of gene markers. Thus, these techniques are usually combined with parallel bulk or single-cell RNA-seq. The **integration of data** from scRNA-seq and spatial transcriptomics is later achieved by two different computational approaches: **mapping** and **deconvolution** (Longo et al., 2021).

Transcriptomics has experienced an astonishing evolution (**Figure 1.9**), but there is still room for improvement. In the future, we will move towards the unbiased integration of the layers of information provided by multiomics, while increasing the resolution of individual techniques. These highly precise and comprehensive analyses will, foreseeably, lead to exciting breakthroughs in multiple biological fields.

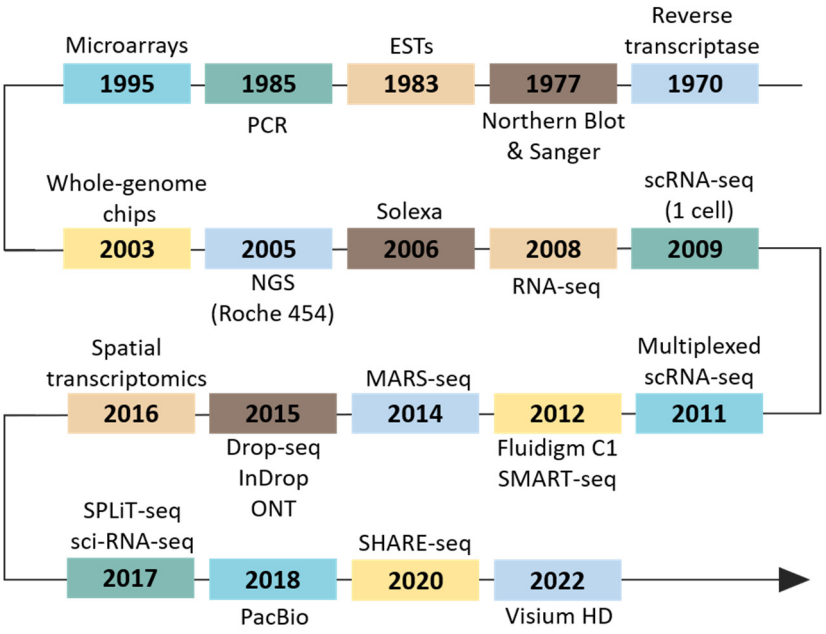


Figure 1.9 Timeline of historical innovations in transcriptomics. *Based on Hong et al., 2020.*

2. SAMPLE PREPARATION IN SINGLE-CELL TRANSCRIPTOMICS

Sample preparation is frequently overlooked, but it has a critical impact on technical variability and the final quality of the experiment. Thus, it must be carefully considered and thoroughly optimized. Sample preparation workflows are very variable but, as every protocol handles RNA, the use of **nuclease-free** consumables and reagents is always recommended (Lafzi et al., 2018). Sample preparation starts with the dissociation of solid tissues -or cell cultures- into **single-cell** or **single-nuclei** suspensions. After dissociation, sample preparation is followed by optional steps of quality control, fixation, permeabilization, cryopreservation and cell enrichment (Lafzi et al., 2018; Nguyen et al., 2018).

2.1. SINGLE-CELL DISSOCIATION METHODS

Traditional single-cell dissociation combines mechanical forces with enzymatic digestion. Dissociation of complex tissues starts with mechanical **dissection** to reduce the size and remove the hardest parts (e.g. shells). It continues with **tissue mincing**, to increase the surface area exposed to the enzyme. These steps are normally performed manually, using blades, scissors, forceps or scalpels. When tissue fragments are small enough, they are dissociated by **enzymatic digestion** using proteases such as trypsin, collagenase, dispase, liberase or papain (**Figure 1.10**). The election of the protease will depend on the specific characteristics of the tissue. The most important factors to consider during proteolysis are enzyme concentration, incubation time and temperature. These factors need to be adjusted carefully to guarantee a good dissociation without over breaking the cells (Lafzi et al., 2018; Nguyen et al., 2018; Reichard and Asosingh, 2019).

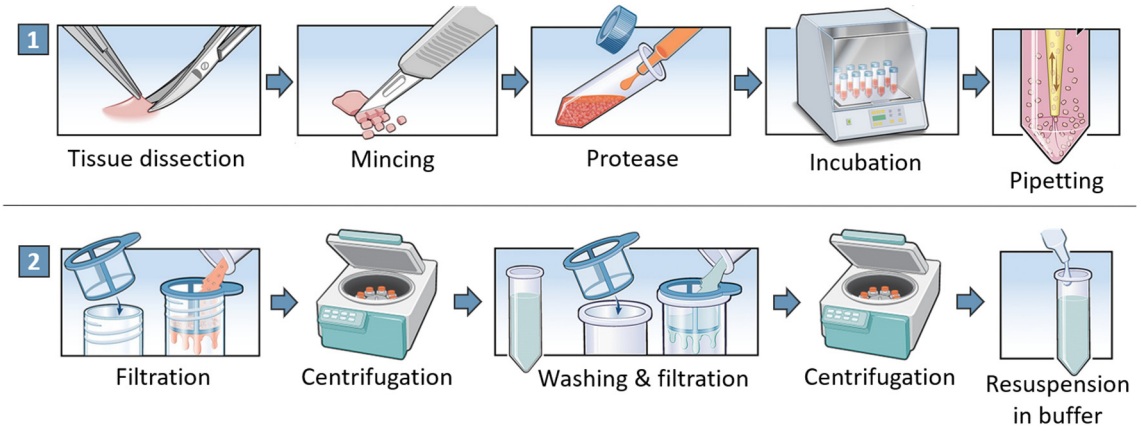


Figure 1.10 Enzymatic dissociation general workflow. *Adapted from Allan et al., 2020 and Reichard and Asosingh, 2019.*

Enzymatic digestion is usually combined with rocking and pipetting to improve tissue disruption. When this is not enough to complete dissociation, further **mechanical forces**, like vortex, blending (e.g. Polytron) or Dounce homogenization, can be applied. Mechanical dissociation can also be achieved using automatized microfluidic devices (Qiu et al., 2015) or semi-automatic homogenizer platforms, like gentleMACS (Baldan et al., 2015; Nguyen et al., 2018). After dissociation, cells are normally filtered, washed and centrifuged to remove proteases, undissociated cell clumps (**aggregates**) and broken cells (**debris**) (**Figure 1.10**). Non-adherent cells, like suspension cell cultures or blood, do not require prior dissociation and can be simply harvested and separated by centrifugation (Lafzi et al., 2018).

The main disadvantage of these techniques is that **cells are kept alive** during and after dissociation. Depending on the protocol configuration and other external factors (e.g. location of the sequencing platform), it can take several hours, or days, before the cells are further processed. If no sample preservation measures are taken, this will lead to cellular stress, transcriptional modifications and RNA degradation (Massoni-Badosa et al., 2020). Enzymatic digestion itself has also been shown to modify gene expression patterns (Denisenko et al., 2020; Huang et al., 2010; van den Brink et al., 2017). Similarly, mechanical stress induced during dissociation can alter gene expression (Mammoto et al., 2012; Martins et al., 2012).

Different strategies have been developed to minimize transcriptional bias linked to dissociation. Some examples are the optimization of **milder and less invasive protocols** for cell harvesting (Zeng et al., 2015), the use of **transcription inhibitors**, like actinomycin D (Wu et al., 2017), or the introduction of **cold-proteases** for enzymatic digestion (Adam et al., 2017; O'Flanagan et al., 2019). Cold-proteases are isolated from psychrophilic microorganisms. They remain active at cold conditions, which allows to reduce dissociation temperature from 37ºC to 6ºC. Cold digestion partially inactivates the transcriptional machinery of the cell, preventing gene alterations.

Another major inconvenience of single-cell dissociation is the requirement of small **fresh tissues** as starting material, as traditional dissociation protocols cannot be performed on large tissues, fixed or frozen samples.

2.1.1. OTHER CELL DISSOCIATION STRATEGIES

Non-enzymatic tissue dissociation has been used in microscopy long before single-cell transcriptomics existed. Some 19th and 20th centuries protocols used mechanical forces or acidic-based formulas to obtain single cell suspensions. These were combined with preservation

techniques to maintain the cell structure, like fixation with osmium tetroxide or formaldehyde (David, 1973; Schneider, 1890; Vial and Porter, 1975).

Acetic acid has been traditionally used as a dissociative agent in microscopy protocols. This kind of dissociation is known as **maceration**. Acetic acid was first used in 1890 to dissociate ***Hydra*** (Schneider, 1890). Later, the maceration technique was fully described in *Hydra attenuatta* (David, 1973). The original maceration solution contained glycerine, glacial acetic acid and water. Cells dissociated with this formula retained their characteristic morphology and their DNA content. Besides, the distribution of cell populations was stable in maceration over time, showing no signs of selective cell destruction. Therefore, maceration was suitable to classify and count cell types by microscopy in a quantitative way (David, 1973). Despite their potential, acidic formulas were never used for cell dissociation in single-cell transcriptomic protocols.

2.2. SINGLE-NUCLEI DISSOCIATION METHODS

In recent years, **single-nuclei RNA-seq (snRNA-seq)** has gained popularity as an alternative to single-cell RNA-seq protocols. In snRNA-seq, the whole cell is lysed during dissociation and only the nucleus is purified and preserved as input material for the experiment (Grindberg et al., 2013; Lacar et al., 2016). Nowadays, many single-cell platforms (e.g. Smart-seq, SPLiT-seq, sci-RNA-seq or 10x Chromium) support single-nuclei inputs and have specific protocols for nuclei isolation and processing (Cao et al., 2020; Gibbons et al., 2022; Krishnaswami et al., 2016; Rosenberg et al., 2018).

Unlike whole-cells, nuclei can be extracted from **frozen** or **fixed tissues**, as the nuclear membrane is more resistant to cryopreservation and mechanical stress. This allows for single cell studies on tissues that are difficult to dissociate by standard methods, and confers higher flexibility to collect samples from different locations or time points (Habib et al., 2017; Krishnaswami et al., 2016). Single-nuclei isolation protocols usually combine **detergents** (e.g. Triton) with **mechanical forces** (e.g. Dounce homogenizer or Polytron) to rupture the cell membrane, and then separate the nuclei by **density gradient centrifugation** (Grindberg et al., 2013; Krishnaswami et al., 2016).

Despite its advantages, snRNA-seq has an important drawback: all the information contained in the **cytoplasm** is lost during sample preparation. On a technical level, single-nuclei libraries are comparable in quality to single-cell libraries (Habib et al., 2017). The transcriptomic profiles obtained from single-nuclei protocols also resemble the ones generated by single-cell approaches and, in general, retrieve similar biological information (Grindberg et al., 2013).

However, the transcriptomes obtained from the nucleus and the cytoplasm are not completely equivalent. About 3.5% of genes show differential abundances between compartments, with some reaching up to 30-fold enrichment in the nucleus (Grindberg et al., 2013). Furthermore, cytoplasmic and nuclear RNAs are not equally matured. Cytoplasmic RNA has interesting post-transcriptional modifications, such as **alternative splicing**. The nuclear fraction, on the other hand, contains more long non-coding RNA (Zaghlool et al., 2021) and pre-mature RNA, which is richer in **intronic sequences** (~30-40% of total mapped reads) (Grindberg et al., 2013; Habib et al., 2017). Finally, single-nuclei protocols filter out **organelles** such as mitochondria, which DNA can be used to obtain valuable biological information when combined with single-cell RNA-seq (Medini et al., 2021).

2.3. SAMPLE PRESERVATION

First scRNA-seq protocols used freshly dissociated cells. However, live cells are fragile and susceptible to changes in gene expression, especially when the sample is not processed immediately after preparation. To overcome this limitation, different preservation strategies have been adopted. A simple recommendation to preserve either nuclei or cells, is to keep samples cold in as many steps of the protocol as possible. **Keeping samples at 4ºC** is a general good practice to prevent RNA degradation and reduce cellular stress (Massoni-Badosa et al., 2020).

When cell processing is delayed for more than a few hours, samples can be frozen and stored after dissociation. Long-term cryopreservation, at -80ºC or in liquid nitrogen, is suitable for both scRNA-seq (Guillaumet-Adkins et al., 2017) and snRNA-seq (Krishnaswami et al., 2016). **Cryopreservation** can induce inner crystallization and disrupt the cell membrane. For this reason, the addition of **cryoprotectants** like dimethyl sulfoxide (DMSO) is required before freezing fresh single-cell suspensions. Nevertheless, cryopreservation of live cells often results in a sharp decrease in cell **viability**, reducing the cell population by more than 50% (Guillaumet-Adkins et al., 2017).

To avoid the disadvantages of fresh samples, we can resort to **sample fixation**. There are two types of fixative reagents. **Cross-linking fixatives** (e.g. paraformaldehyde or glutaraldehyde) join the amine groups of proteins by forming intra- and intermolecular bridges. On the other hand, **precipitating fixatives** (e.g. methanol, acetone or ethanol) dehydrate cellular components, causing them to coagulate and precipitate (Howat and Wilson, 2014).

In single-cell and single-nuclei RNA-seq, the most popular options for fixation are **formaldehyde** -or paraformaldehyde- and **methanol** (Alles et al., 2017; Cao et al., 2020; J. Chen et al., 2018; Rosenberg et al., 2018). Both reagents are equally useful to stiffen the cells and stop the transcriptomic machinery but methanol, in addition, permeabilizes the cell membranes. **Permeabilization** can also be achieved after paraformaldehyde fixation by further treating samples with methanol, triton or saponins (Jamur and Oliver, 2010). Fixation and permeabilization take place right after dissociation (**Figure 1.11**).

Fixation is useful to protect samples from mechanical stress and cryopreservation temperatures. Fixed samples are accepted by most droplet-based (Alles et al., 2017; J. Chen et al., 2018) and plate-based platforms (Attar et al., 2018; Thomsen et al., 2016). ***In situ* barcoding protocols** (Cao et al., 2017; Rosenberg et al., 2018), moreover, require samples to be fixed and permeable. Finally, fixed libraries do not defer from fresh preparations in complexity or transcriptomic expression (Alles et al., 2017; Attar et al., 2018; J. Chen et al., 2018).

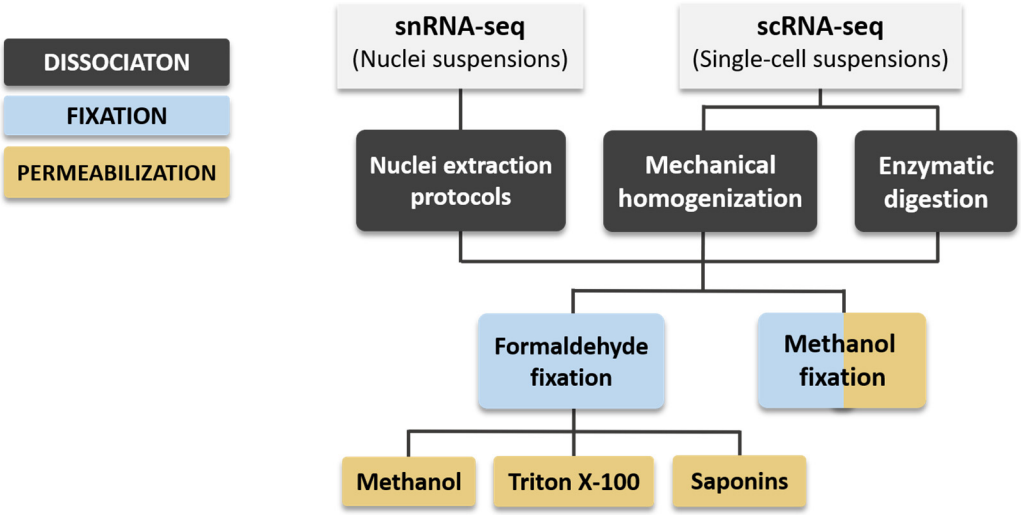


Figure 1.11 Overview of typical sample preparation workflows. Dissociation strategies aim to isolate single-nuclei (snRNA-seq) or single-cells (scRNA-seq) following different protocols. Subsequently, optional fixation and permeabilization can be performed using diverse methods.

Sample cryopreservation and fixation allow experiments to be unconstrained by time or location. This has facilitated clinical research, field sampling or time-lapse studies in single-cell transcriptomics. However, preservation protocols also have some disadvantages.

Cryopreservation of live cells can modify the transcriptomic expression and **deflect certain cell types**. Meanwhile, fixatives can be **highly toxic** and affect cellular components. Precipitating fixatives denature proteins and alter the inner structure of the cell, impeding subsequent immunohistochemistry or protein visualisation. By permeabilizing the membranes, they also

lead to a higher **RNA leakage** (Denisenko et al., 2020). Cross-linking fixatives, on the other hand, react with nucleic acids, **breaking the RNA and DNA** (Howat and Wilson, 2014; Russell et al., 2013). Reserve cross-linking protocols can help to prevent nucleic acid degradation after cross-linking fixation but, at the same time, they involve extra incubation steps at high temperatures (Thomsen et al., 2016). Importantly, fixatives do not inactivate **RNases**, but can make cells more permeable and susceptible to their action. Thus, the use of RNase inhibitors and the handling of samples under RNase-free conditions remain important when working with fixed samples.

2.4. QUALITY CONTROL

The addition of quality control steps is key to assessing whether the sample is good enough to proceed with an experiment (Nguyen et al., 2018). Quality control steps should ideally be performed **on the day of sample processing**. For instance, cryopreserved samples should be evaluated after thawing, even if the same quality controls have been already carried out before freezing. How we assess the quality will vary depending on the sample preparation protocol.

When working with **whole live cells**, viability is one of the main factors to consider. **Cell viability** measures the percentage of cells with an intact membrane. Traditionally, it has been assessed using exclusion assays to stain broken cells (e.g. Trypan blue). In those, live cells remain unstained, as dyes cannot penetrate their membranes. After staining, counting of live and dead cells can be performed by automatic counters, manual haemocytometry (e.g. Neubauer), flow cytometry or microscopy (Hanamsagar et al., 2020; Nguyen et al., 2018; Reichard and Asosingh, 2019). In good quality single-cell suspensions, viability is around 90-95%. When viability drops below 70-80%, samples must undergo a **dead cell removal** process, either using centrifugation or one of the commercial kits available for this purpose (Hanamsagar et al., 2020; Reichard and Asosingh, 2019).

Sample **homogeneity** is also important, as it indicates the proportion of cellular **aggregates** and, therefore, the efficiency of dissociation. Larger undissociated tissue fragments can be spotted by sight, while smaller aggregates can be detected by microscopy or flow cytometry. If not removed, aggregates can cause blockages on the single-cell platform and increase the ratio of capture of **doublets** (two cells or nuclei). Thus, when a cell suspension is not homogeneous enough, it requires further dissociation, purification or enrichment.

Single-nuclei viability and **homogeneity** can be assessed in a similar way, but using nucleic acid-specific dyes (e.g. Ethidium homodimer-1). In this case, only nuclei incorporate the dye, while undissociated cells and broken fragments remain unstained. A single-nuclei suspension is

considered properly lysed when the percentage of undissociated cells is below 5%. Counting of individual nuclei can be performed on an automatic counter or haemocytometer, but it is also recommended to visually assess the state of the nuclear membrane by microscopy (https://kb.10xgenomics.com/hc/en-us/articles/360050490472).

Finally, quality control should focus on **gene expression** and **RNA quality**. During the optimization of the protocol, control qPCRs for sets of marker genes can be used to screen transcriptomic changes induced by sample preparation (Nguyen et al., 2018). On the other hand, the RNA quality is assessed using a standard value, known as **RNA integrity number (RIN)**. The RIN is calculated by an algorithm that evaluates the ratio of 28S:18S ribosomal RNA bands (Schroeder et al., 2006) (**Figure 1.12**). RIN values can be measured in an Agilent 2100 Bioanalyzer or TapeStation (Padmanaban et al., 2012).

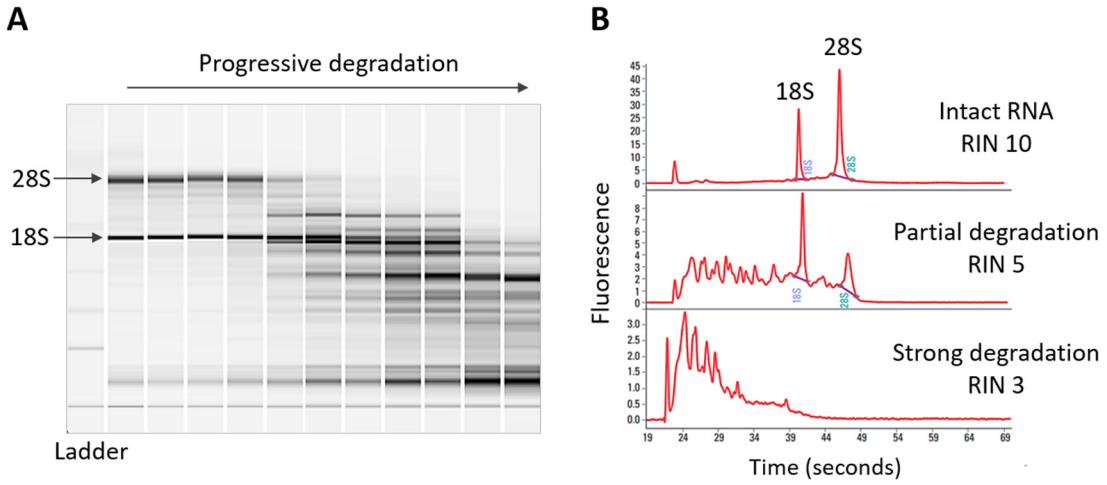


Figure 1.12 RIN values plotting. A) On-chip electrophoretic separation of RNA samples with different levels of degradation of the 18S and 28S ribosomal bands **B)** Electropherogram representation of different RIN values from intact (RIN 10), partially degraded (RIN 5) and strongly degraded (RIN 3) RNA. *Adapted from: https://www.agilent.com/cs/library/applications/5989-1165EN.pdf*

2.5. PURIFICATION AND ENRICHMENT

Current dissociation and preservation protocols are not perfect. They usually generate a large number of aggregates, debris and free-floating nucleic acids, especially when applied on complex tissues. On the other hand, our target cell populations can be difficult to detect when they are rare or diluted in heterogeneous cell suspensions. Sample purification and enrichment steps are meant to solve these limitations. Their application normally translates into much **cleaner and specific datasets**, as the likelihood of capturing free floating RNAs, doublets or unspecific cell populations gets reduced when starting from cleaner input solutions.

Washes, **centrifugation** and **size-exclusion filtration** are the simplest purification methods to remove aggregates and debris. They are usually performed as intrinsic steps of dissociation and preservation protocols. However, they can be repeated and optimized in numerous ways to improve sample quality (Lafzi et al., 2018). **Density-gradient centrifugation** can separate different sample phases and is useful for a more complex purification or enrichment. For example, density-gradients are commonly used to isolate single-nuclei after cell lysis (Grindberg et al., 2013; Krishnaswami et al., 2016).

Other optional methods include **column-based magnetic separation** (e.g. MACS Miltenyi Biotec kits), used for dead cell removal or cell enrichment (Hanamsagar et al., 2020), and **FACS sorting**. FACS is a versatile technique. It can be used for sample purification and enrichment, but also for cell isolation, as in MARS-seq protocols (Jaitin et al., 2014; Keren-Shaul et al., 2019). Using a simple nuclear dye, like Hoechst, FACS can filter out aggregates and debris to sort **singlet-enriched** solutions. FACS can also discriminate cell populations according to their size, shape, and complexity. Furthermore, FACS indexed sorting can **isolate specific cell types** tagged with antibodies, depleting other uninformative cell populations (**Figure 1.13**) (Hu et al., 2016; Lafzi et al., 2018; Nguyen et al., 2018). Fixed-cells and nuclei can also be FACS-sorted (Marion-Poll et al., 2014).

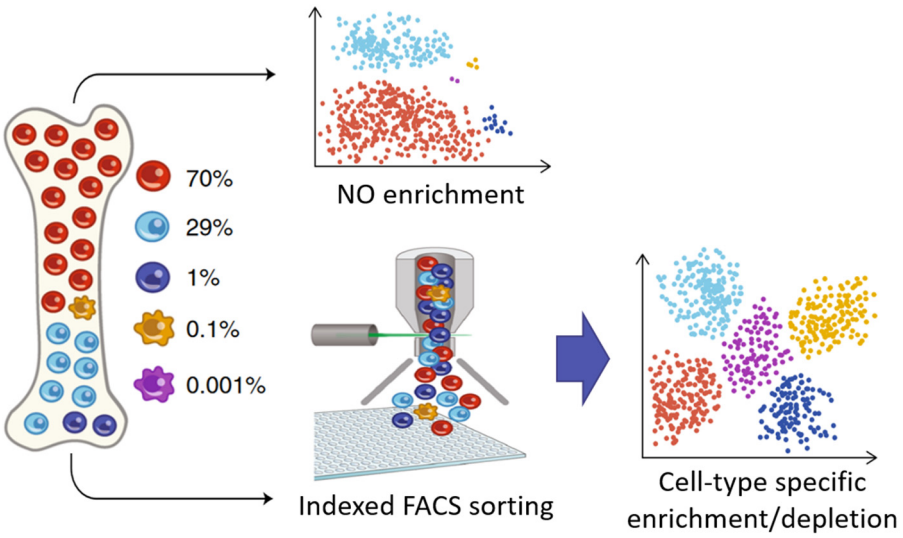


Figure 1.13 FACS index sorting scheme. Characterization of a heterogenous organ following two different strategies. The none-enriched dataset profiles a more realistic cell distribution. Meanwhile, the cell-type specific enriched/depleted dataset, sorted by indexed FACS, shows a better identification of rare cell populations. *Adapted from Keren-Shaul et al., 2019.*

Purification and enrichment methods can greatly improve the quality and accuracy of the experiment. However, they also increase sample **handling time**, which can lead to RNA degradation, cellular stress and gene expression bias. FACS sorting, in particular, exposes the sample to high pressures and osmotic changes. In live cells, this can be detrimental for viability

and contribute to transcriptomic activation (Martins et al., 2012; Nguyen et al., 2018). Another major disadvantage of FACS sorting is the requirement of large cell inputs (Hu et al., 2016).

2.6. MEDIA

Media plays an important role in sample preparation. Its composition must contribute to stabilize cell viability, RNA integrity and sample homogeneity as much as possible. Common **phosphate-buffered saline (PBS)** is the most widely used media to prevent osmotic stress in fresh and fixed samples. However, more complex commercial buffers, like Dubbeco's PBS (DPBS), can be preferred for certain live cell protocols. To prevent sample re-aggregation, small percentages of bovine serum albumin (BSA) can be added to the media (0.1-2%). Another strategy is treating the sample with DNase I, as free-floating DNA released from dead cells can induce re-aggregation (Reichard and Asosingh, 2019; Renner et al., 1993). Moreover, live cells need to be in calcium and magnesium free media, as both ions promote cell adhesion (Takeichi and Okada, 1972). To protect RNA from degradation, media must be **contaminant and nuclease free**. The use of ultrapure water and **RNases inhibitors** in the media can help to provide extra protection.

3. COMPUTATIONAL ANALYSIS IN SINGLE CELL TRANSCRIPTOMICS

After sequencing, **raw reads** have to be computer-processed and analysed to translate into meaningful biological data. Computational analysis has evolved in parallel to NGS, bulk RNA-seq and single-cell RNA-seq technologies. In recent years, this has led to the development of a complex and constantly evolving landscape of bioinformatic tools. Many of these novel tools are focusing on scRNA-seq data analysis, which is particularly challenging due its complexity, technical noise, diversity and increasing number of cells. Although it is difficult to standardize a common workflow for every application, the analysis of scRNA-seq data can be broadly divided into **pre-processing** and **downstream analysis** (Luecken and Theis, 2019).

3.1. PRE-PROCESSING

Pre-processing starts with the **processing of raw data** (**Figure 1.14**), which is common for bulk and single-cell RNA-seq. Then, scRNA-seq data is further modified with **single-cell specific pre-processing** steps, including count matrix generation, quality control, normalization, data correction, feature selection and dimensionality reduction (visualisation) (Luecken and Theis, 2019; Rostom et al., 2017).

3.1.1. RAW DATA PROCESSING

An initial **quality control** is performed to detect low-quality reads, N bases, and adapters. FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is commonly used for this quality assessment. Subsequent **reads processing** trims out low-quality sequences and other uninformative fragments (e.g. poly-A tails) using tools like Cutadapt (Martin, 2011), Trimmomatic (Bolger et al., 2014) or fastp (Chen et al., 2018). In addition, single-cell data undergoes **demultiplexing** to classify reads according to their cellular and molecular barcodes.

After that, reads are aligned -or mapped- to a reference sequence (genome or transcriptome) using different **alignment** algorithms (e.g. STAR, TopHat2, HISAT2) (Dobin et al., 2013; Kim et al., 2019, 2013). When the reference sequence is not available, the transcriptome can be **assembled *de novo*** mapping the reads to each other with software like Trinity, SOAPdenovo-Trans or StringTie2 (Grabherr et al., 2011; Kovaka et al., 2019; Xie et al., 2014). Later, gene expression is quantified as the number of reads mapped per gene (**counts**). Tools like Cufflinks, RSEM or HTseq (Anders et al., 2015; Li and Dewey, 2011; Trapnell et al., 2010) quantify counts from a reference sequence. In the absence of such reference, or when the alignment steps want to be avoided, a **pseudo-aligner** (e.g. Sailfish, Kallisto, Salmon) (Bray et al., 2016; Patro et al., 2017,

2014) can be used for simultaneous assembly and quantification of gene expression (Alser et al., 2021). Raw data processing ends with data **normalization** to correct technical bias. There are multiple normalization strategies. Some examples are the calculation of fragments per kilobase of mapped reads (FPKM) or transcripts per million (TPM) (Corchete et al., 2020).

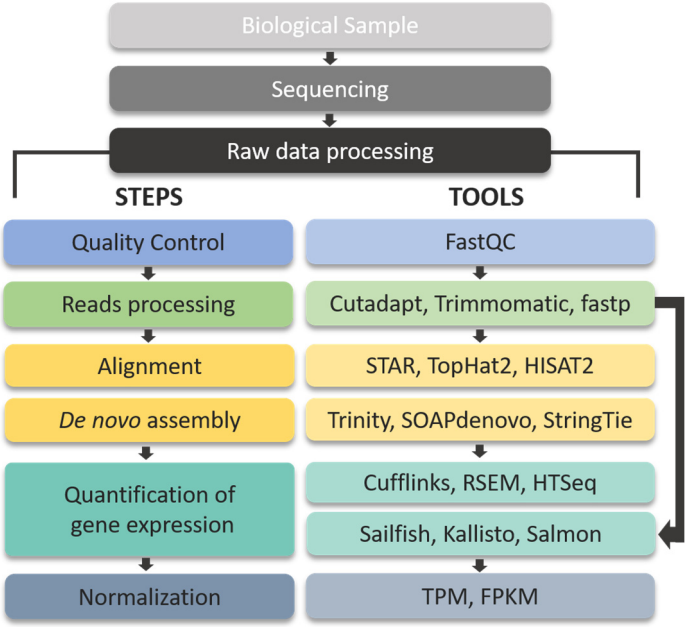


Figure 1.14 Computational analysis steps and tools for raw data processing. *Based on Hong et al., 2020.*

3.1.2. SINGLE-CELL SPECIFIC PRE-PROCESSING

After first normalization, gene expression is translated into a **count matrix** (**Figure 1.15 A**). In scRNA-seq, most genes will not be captured and, therefore, will have **zero counts** (Lafzi et al., 2018; Luecken and Theis, 2019). Further **quality control** will remove cells with an unusual number of mitochondrial counts, total counts (count depth) and total genes, as these outliers will likely correspond to broken membrane cells, with leaking RNA, or to cell-aggregates and doublets (Luecken and Theis, 2019). Doublets can also be removed using specific tools like DoubletDecon (DePasquale et al., 2019), Scrublet (Wolock et al., 2019) or Solo (Bernstein et al., 2020).

Pre-processing continues with further data **normalization** and **correction** to mend technical artefacts, batch effects (**Figure 1.15 B**), cell-specific biases or RNA content bias, among others. SCnorm (Bacher et al., 2017), SCRAN (Lun et al., 2016) and kBET (Büttner et al., 2019) are some specialized tools available for scRNA-seq normalization and batch correction. In addition, when we need to combine multiple experiments, **data integration** can be performed with tools such

as CCA (Butler et al., 2018), LIGER (Welch et al., 2019), Harmony (Korsunsky et al., 2019) or Scanorama (Hie et al., 2019).

Next, during feature selection, we usually subset **highly variable genes (HVGs)**. HVGs are those that better define cell populations or variations between experimental conditions. Therefore, they are the most informative genes for the analysis (Luecken and Theis, 2019). Finally, preprocessing ends with the visualisation of the data. ScRNA-seq datasets are multidimensional. In order to visualize them in a two-dimensional space (**Figure 1.16 A**), we need to use **dimensionality reduction** algorithms, like Principal Component Analysis (PCA) (Abdi and Williams, 2010), **t-SNE** (van der Maaten and Hinton, 2008) or **UMAP** (McInnes et al., 2020).

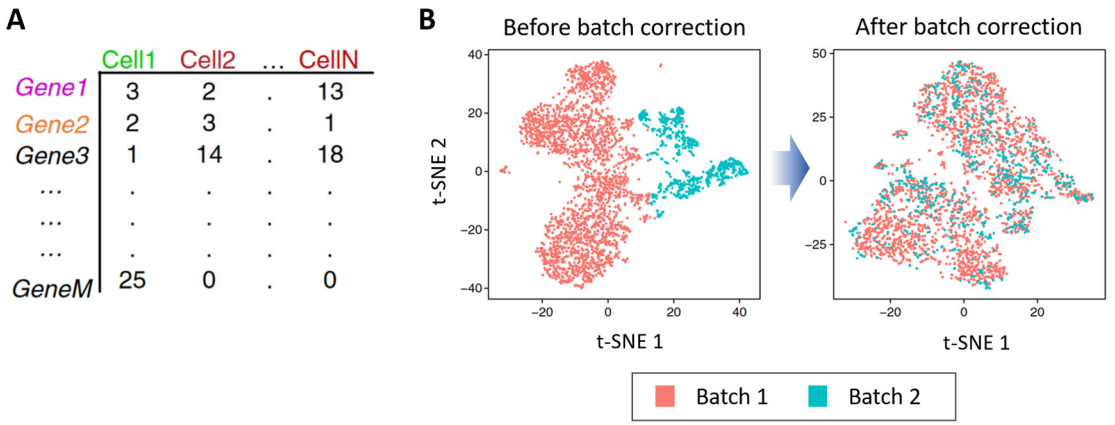


Figure 1.15 Single-cell preprocessing steps A) Count matrix: the matrix reports the number of gene counts per cell, presenting each gen in a row and each cell in a column. **B) Batch correction:** t-SNE plot visualisation of two datasets (batches) before and after batch correction. *Adapted from Butler et al. 2018, and Lafzi et al., 2018.*

3.2. DOWNSTREAM ANALYSIS

Downstream analysis starts with **cell clustering**. Cells are classified into different groups, called **clusters**, according to their transcriptomic expression. Each cluster represents an inferred cell identity (**Figure 1.16 B**). Cell clustering is performed using a graph-based approach that combines k-nearest neighbour (KNN) graph representation with community detection algorithms, like **Louvain** (Blondel et al., 2008) or **Leiden** (Traag et al., 2019). The most popular plotting options for cluster visualisation are t-SNE and UMAP. Once clusters have been defined, we can extract the list of signature genes -or **gene markers**- that characterize the expression of each cluster. These markers can be used for **cluster annotation**, to assign clusters a biological identity (**Figure 1.16 C**). Cluster annotation relies on the use of external resources, such as reference databases and literature (Luecken and Theis, 2019).

Subsequent steps will depend on our final biological question. For instance, **differential gene expression analysis** is commonly performed to identify significant gene expression changes between conditions. For this analysis, we can use bulk RNA-seq tools like DESeq2 (Love et al., 2014), edgeR (Robinson et al., 2010) and limma-voom (Ritchie et al., 2015), or specific scRNA-seq tools, such as MAST (Finak et al., 2015).

Trajectory inference is another popular scRNA-seq analysis (**Figure 1.16 D**). It was recently established, with the publication of the Monocle and Wanderlust algorithms (Bendall et al., 2014; Trapnell et al., 2014). These methods project cells over an imaginary temporal variable, known as **pseudotime**, to recreate cell differentiation pathways. More recent toolkits for the study of cell trajectories include Monocle 2 (Qiu et al., 2017), Slingshot (Street et al., 2018) and PAGA (Wolf et al., 2019). Furthermore, when pseudotime is combined with **RNA velocity** (La Manno et al., 2018), the directionality of these cell trajectories can be inferred. The RNA velocity vector is estimated by comparing the spliced and unspliced mRNA content, and it allows to predict the future state of a cell.

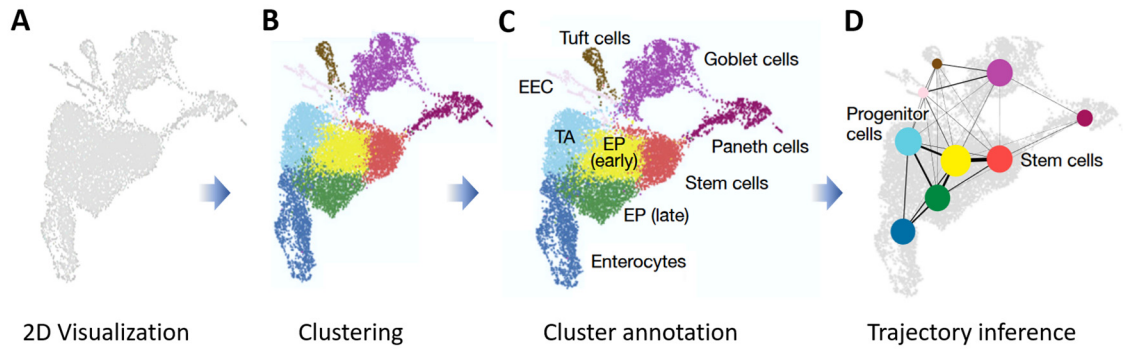


Figure 1.16 Plotting of scRNA-seq data at different points of the analysis. A) 2D visualisation after dimensionality reduction. **B)** Clusters plot. Each cluster is represented in a different colour. **C)** Cluster annotation **D)** Trajectory inference. Clusters are represented by colour dots and connected with lines of different thickness depending on the strength of their connections. *Adapted from Luecken et al., 2019.*

3.3. RESOURCES FOR SINGLE-CELL DATA ANALYSIS

Although scRNA-seq data can be analysed step by step with individual software, it is more common and convenient to perform the analysis using integrative platforms. These platforms consist of large and scalable software packages and toolkits that cover the entire analysis pipeline. Most popular ones are **Seurat** (R-based) (Satija et al., 2015), **Scanpy** (Python-based) (Wolf et al., 2018) and **Cell Ranger** (for 10x Chromium) (Zheng et al., 2017). Besides, most scRNA-seq analytical tools are open-source and can be downloaded from public software repositories like **Github** (https://github.com/) and **Bioconductor** (https://www.bioconductor.org/).

4. INTRODUCTION TO PLANARIANS

Planarians are a group of free-living flatworms belonging to the phylum *Platyhelminthes* and the order *Tricladida* (Ivankovic et al., 2019; Laumer et al., 2015). These animals have caught the attention of biologists over centuries due to their stunning regeneration capacities. Nowadays, they are consolidated **model organisms** commonly used in stem cell biology and regeneration studies. Moreover, during the last decade, planarians have become one of the most prominent models for **single-cell transcriptomics**.

4.1. GENERAL ANATOMY

Planarians are one of the simplest bilaterians. These invertebrates have unsegmented and flattened soft bodies, with no cuticle, shell or exoskeleton. Planarians have no circulatory or respiratory organs. Nutrients and oxygen diffuse through their tissues. In contrast, they present a well-organized **nervous system**, formed by the main ganglion (brain), light photoreceptors (eyespots) and longitudinal and transversal nerve cords. Planarians also have a single-opening **digestive system**, and a primitive **excretory system**, called protonephridia. In addition, sexual planarians have hermaphroditic **reproductive systems** (**Figure 1.17**) (Collins, 2017). Other planarian tissues include a single-layered epidermis, the muscle body and the connective tissue, known as parenchyma. The latest comprises some less-structured cell populations: secretory cells, parenchymal cells and neoblasts.

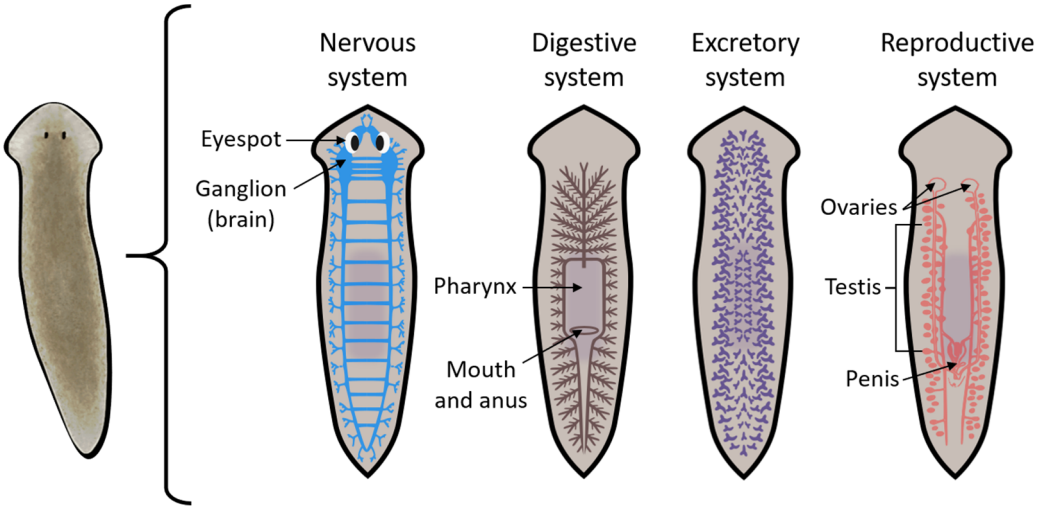


Figure 1.17 Planarians general anatomy. Planarians have some well-organized internal structures: the nervous system, digestive system, excretory system (protonephridia) and reproductive system.

4.2. REGENERATION AND CELL RENEWAL

Planarians can regenerate their whole bodies, including the brain. Moreover, they can be transversely cut into multiple pieces (**Figure 1.18 A**), or amputated in different shapes (**Figure 1.18 B**), regenerating each fragment into a full new individual in a matter of days (Baguña, 2012; Ivankovic et al., 2019). These impressive regeneration capacities are attributed to a population of self-renewing **adult pluripotent stem cells**, called **neoblasts** (Aboobaker, 2011). Neoblasts are small rounded cells, with a large nucleus, that are located throughout the planarian parenchyma (Ivankovic et al., 2019). During their whole life, planarians maintain an abundant and stable population of neoblasts, which accounts for about one fourth of their total cells.

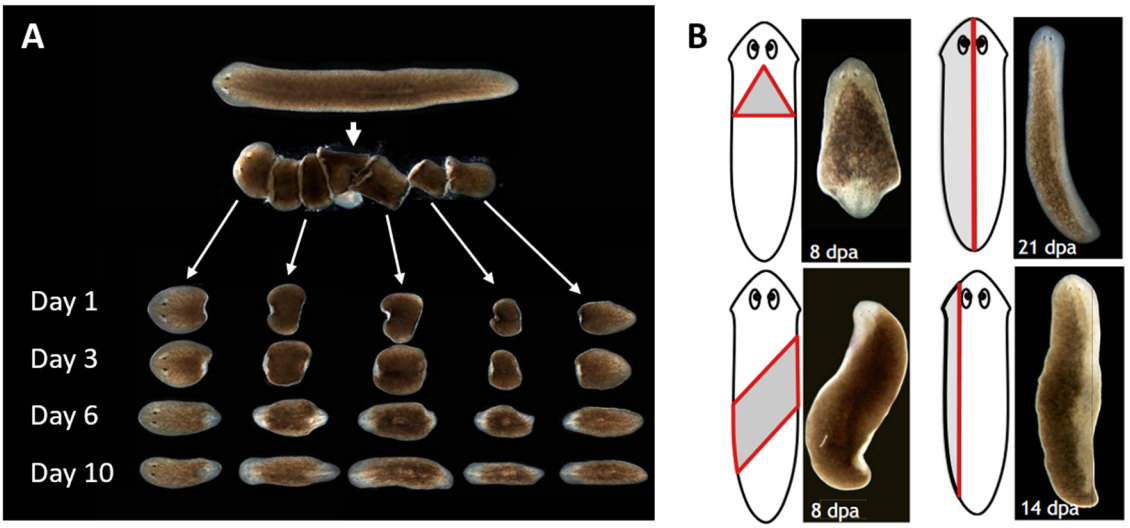


Figure 1.18 Depiction of planarian regeneration A) Regeneration of transversely cut fragments, from head to tail, at different time points. At day 10, all fragments show regenerating new individuals. *Picture credit: Jordi Solana* **B)** Regeneration of planarian fragments cut in different shapes. *Adapted from Ivancovik et al., 2019.*

Neoblasts are the only mitotic cells in asexual planarians. They intervene in **regeneration** when tissues are damaged by trauma, or after fission reproduction, by migrating to the location of the injury. There, they create a regeneration **blastema**, mostly made of undifferentiated cells, and proliferate into all somatic cell types. In sexual individuals, neoblasts can also give rise to the germline. In addition, neoblasts contribute to their own **self-maintenance** and the homeostatic **cell renewal** of the whole body (Aboobaker, 2011; Baguña, 2012; Nakagawa et al., 2012; Zhu and Pearson, 2016). Due to the neoblast continuous proliferation, the turnover of the entire animal occurs in a matter of weeks (Rink, 2013). This rapid self-renewal confers planarians an incredible plasticity. Thus, these animals can grow and degrow depending on food availability and other factors, adjusting their body proportions accordingly (Romero and Baguñà, 1991).

As proliferating cells, neoblasts can be ablated by irradiation. With the sufficient dose, irradiated animals lose the capacity to regenerate and eventually die (Abnave et al., 2017; Hayashi et al., 2006). However, it has been shown that a single healthy neoblast, transplanted on an irradiated individual, can differentiate into every cell type and colonize the whole body of the host (Wagner et al., 2011).

4.3. HETEROGENEITY OF THE NEOBLAST POPULATION

Planarian neoblasts constitute a heterogeneous cell population, with multiple subclasses described in the literature. **Clonogenic neoblasts (cNeoblasts)** were identified as a pluripotent population, able to differentiate into any cell type in asexual individuals (**Figure 1.19**) (Wagner et al., 2011).

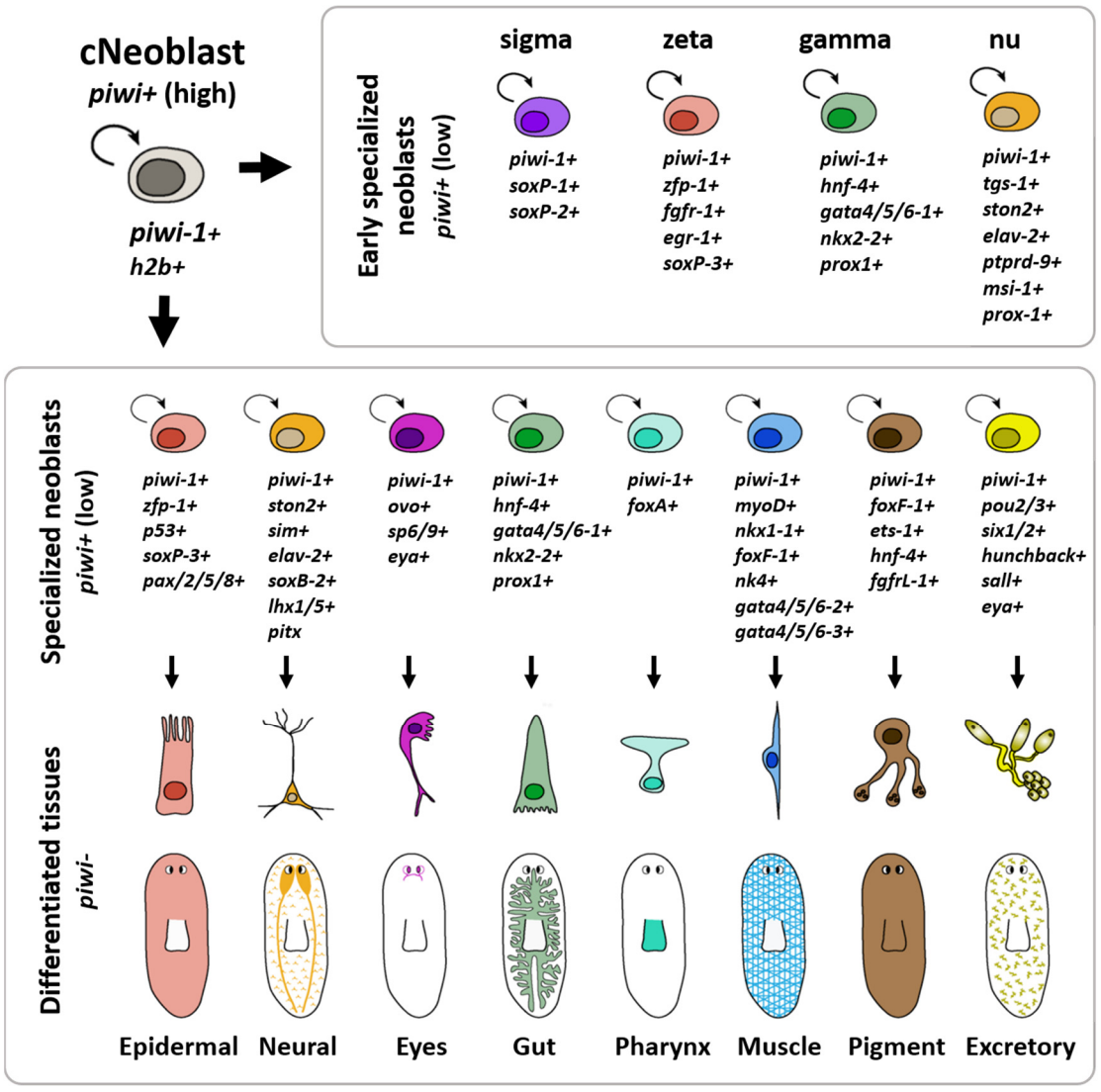


Figure 1.19 Neoblast heterogeneity. Different neoblast subpopulations express variable levels of *piwi-1*, plus specific sets of transcription factors. After differentiation, each specialized population give rise to a certain lineage or group of lineages. *Adapted from Molina and Cebrià, 2021.*

Later, neoblasts were classified in two major specialized populations: **zeta-neoblasts**, which give rise to the epidermal lineage, and **sigma-neoblasts**, which have a broader lineage potency, participate in wound response, and can regenerate the zeta-neoblasts. Within sigma-neoblasts, a third population of **gamma-neoblasts** was proposed to give rise to the gut (**Figure 1.19**). The authors also suggested that cNeoblasts were contained within the sigma population (van Wolfswinkel et al., 2014). A fourth subclass, called **nu-neoblasts**, was later proposed as neural progenitors (Molinaro and Pearson, 2016).

All neoblast subpopulations express different levels of *piwi-1*, the canonical neoblast marker (Reddien et al., 2005), plus a specific set of transcription factors. A more detailed exploration of these differences allows to classify neoblasts in even more specialized subtypes (Scimone et al., 2014; Zeng et al., 2018) (**Figure 1.19**). However, neoblast classification and differentiation pathways are still under discussion. For instance, it was recently proposed that specialization could be a reversible stage, and every neoblast may retain pluripotency to act as a cNeoblast under the right circumstances (Raz et al., 2021).

4.4. PLANARIANS AS MODEL ORGANISMS FOR TRANSCRIPTOMICS

As we have seen, planarians possess a set of biological features that make them unique, like the ability to regenerate their whole bodies and reproduce asexually, or their rapid cell turnover and body plasticity. In addition, and unlike most animals, the adult planarian contains a **snapshot of all cellular stages**, including stem cells, progenitors and differentiated cells. All of this makes planarians a very interesting model organism for single-cell resolution studies on stem cell regulation and differentiation.

Apart from their intrinsic biological interest, planarians present multiple **technical advantages** over other model organisms. First, their cultures are affordable and easy to maintain and amplify in the lab, and their use has minimal ethical implications compared to vertebrate models. Second, planarians have a long story of technical developments and multiple resources available. Since classical studies, planarians have been manipulated using dissection, X-ray irradiation, optical and electronic microscopy, and transplantation protocols (Baguña, 2012). During the molecular era, new tools like **RNA interference** (Sánchez Alvarado and Newmark, 1999) and bromodeoxyuridine **(BrdU) labelling** (Newmark and Sánchez Alvarado, 2000) were implemented to inhibit gene expression and study cell proliferation, respectively.

Then, in the early 2000s, planarian genomic and transcriptomic studies emerged together with the rise of high-throughput technologies. In 2008, the first genome of *Schmidtea mediterranea* was published (Cantarel et al., 2008) and, two years later, the first planarian transcriptome was generated for the same species using **RNA-seq** (Blythe et al., 2010). Since then, bioinformatic resources have not stopped growing to support planarian research. Nowadays, we have annotated genomes for *S. mediterranea* (Grohme et al., 2018; Guo et al., 2022) and *Dugesia japonica* (An et al., 2018; Tian et al., 2022), and transcriptomes for other multiple species. These resources are available in **PlanMine** (Brandl et al., 2016), an open-access repository that hosts shared databases and tools for the scientific community.

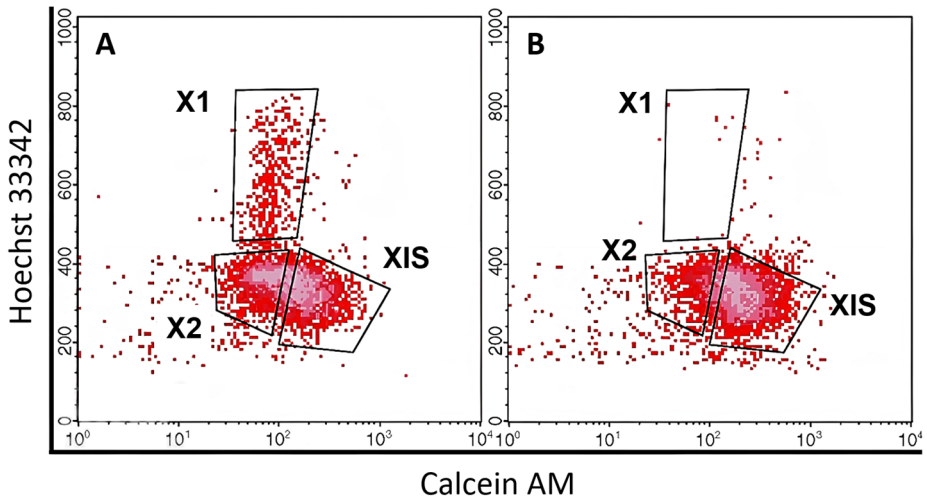


Figure 1.20 FACS sorting profiles of planarian cells stained with Hoechst 33342 and Calcein AM. A) Cells isolated from non-irradiated animals. The X1, X2 and XIS fractions are preserved. **B)** Cell populations isolated from X-ray irradiated animals. X-ray sensitive populations are absent (X1) or highly depleted (X2). *Adapted from Hayashi et al., 2006.*

Another two crucial techniques for the development of transcriptomics, **enzymatic dissociation** and **FACS sorting**, were implemented in planarians in the early 2000s. These protocols were first used to isolate head-abundant cells (Asami et al., 2002), and then further developed to compare the cell sorting profiles of non-irradiated (**Figure 1.20 A**) and X-ray irradiated animals (**Figure 1.20 B**) (Hayashi et al., 2006). Using FACS, three distinct cell populations could be isolated: X1 (X-ray sensitive proliferating neoblasts), X2 (X-ray sensitive cell progenitors) and XIS (X-ray insensitive differentiated cells). With the popularization of FACS sorting, enzymatic digestion became the standard method for dissociation in planarians (Moritz et al., 2012; Romero et al., 2012). The existence of these specialized protocols -which are rare in other model organisms- was exploited by RNA-seq studies (Labbé et al., 2012; Önal et al., 2012) and accelerated the start of **single-cell transcriptomics** in planarians.

First single-cell publications date back more than a decade, and combine FACS with reverse transcription PCR (**RT-PCR**) (Hayashi et al., 2010). Following this approach, neoblasts were characterized in multiple subtypes (van Wolfswinkel et al., 2014; Zeng et al., 2018). Later, **Smart-seq** was used in planarians to profile 1214 cells (classified into 13 identities) and study injury response at tissue resolution (Wurtzel et al., 2015). Finally, with the introduction of **droplet-based methods**, two comprehensive whole-body **single-cell atlases** of *S. mediterranea* were published (Fincher et al., 2018; Plass et al., 2018) (**Figure 1.21**). These atlases achieved a much higher resolution than previous studies, with 50,562 and 21,613 cells profiled, respectively. Their analysis allowed the identification of more than 40 cell populations and the inference of planarian **differentiation trajectories** for the first time (Plass et al., 2018).

These publications have positioned planarians as one of the most comprehensively studied animal models at single-cell resolution. Taken together, all the literature and technical development of the field make planarians a very suitable model to explore new horizons in single-cell transcriptomics.

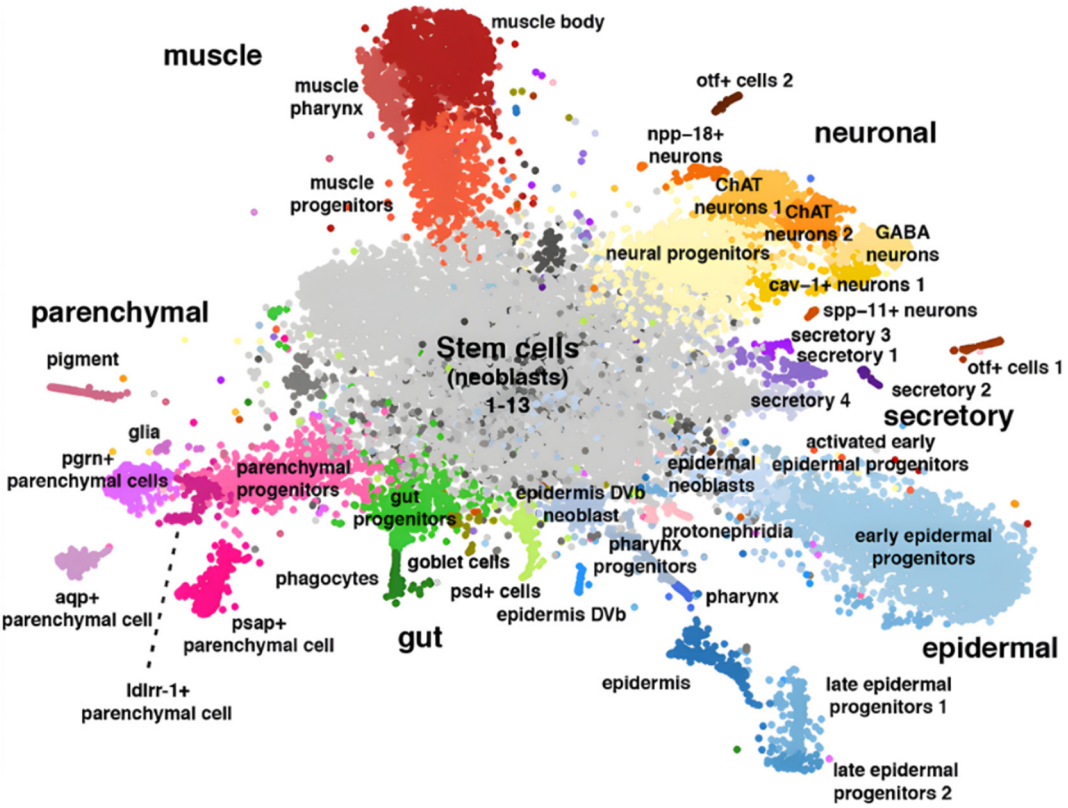


Figure 1.21 Single-cell atlas of *Schmidtea mediterranea*. The main cell groups (stem cells, muscle, neural, secretory, epidermal, protonephridia, gut and parenchymal) are represented with different colour palettes. Individual cell identities are labelled on top of -or next to- its cluster. *Adapted from Plass et al., 2018.*

5. INTRODUCTION REMARKS

In this introductory chapter, I have recapitulated the history of transcriptomics from the 1970s to the present. As we have seen, transcriptomics has experienced a tremendous evolution over the years, with an ever-increasing coverage and resolution. Then, I have delved into single-cell transcriptomics, one of the most remarkable and cutting-edge technologies for the study of the transcriptome. Since it started in 2009, single-cell transcriptomics have been applied to multiple organisms, resulting in the achievement of technological milestones (Cao et al., 2017; Islam et al., 2011; Jaitin et al., 2014; Macosko et al., 2015; Tang et al., 2009) and novel biological insights (Briggs et al., 2018; Cao et al., 2019, 2020; Fincher et al., 2018; Levy et al., 2021; Musser et al., 2021; Plass et al., 2018; Sebé-Pedrós et al., 2018a).

As an overview of the current practices and limitations of the field, I have explored sample preparation and computational analysis in single-cell transcriptomics in detail. Finally, I have presented planarians, the main model organism of this thesis, and the characteristics that make them unique and suitable for these technologies. During the following chapters, I will use scRNA-seq to study planarian biology. First, I will focus on overcoming some of the limitations of conventional protocols, presenting a pipeline that combines a novel sample preparation method (ACME) with a powerful *in situ* barcoding platform (SPLiT-seq). Then, I will use this pipeline to perform quantitative and evolutionary analyses in planarian, at tissue resolution.

AIMS OF THE THESIS

CHAPTER II:

To develop and validate a customized workflow for single-cell transcriptomics, and present its advantages over other methods. The requirements set for this workflow included non-enzymatic dissociation, low-budget, high-throughput (several thousand cells per experiment), FACS enrichment and sample cryopreservation.

CHAPTER III:

To characterize the *hnf4* knockdown in planarians and quantify its effects at tissue resolution, in order to expand our insights into the origin, differentiation and regulation of the parenchymal and gut linages.

CHAPTER IV:

To show the preliminary results on the evolutionary comparison of different planarian species at single cell resolution. The whole project aims to unravel the evolutionary conservation -or divergence- of planarian cell types and their gene expression patterns, as well as the differences between sexual and asexual planarians, at tissue-resolution.

CHAPTER II: ACME & SPLiT-seq. A VERSATILE WORKFLOW PROPOSAL FOR SINGLE-CELL TRANSCRIPTOMICS

INTRODUCTION

1. DESIGNING A CUSTOM WORKFLOW FOR SC-RNA-SEQ

In 2018, our lab aimed to establish a **custom workflow** to perform single-cell transcriptomics in **planarian**, our model organism. Planarians had been profiled at single-cell resolution in previous publications, using traditional sample preparation strategies and well-established scRNA-seq platforms (Fincher et al., 2018; Plass et al., 2018; Wurtzel et al., 2015). However, as we aimed to perform single-cell experiments on a regular basis, we decided to develop a new pipeline adapted to our research goals, budget, equipment availability, required throughputs and sample handling.

As seen in the introduction, sample preparation is one of the major technical challenges in scRNA-seq. None of the protocols available are perfect or adapt to every tissue and application (Denisenko et al., 2020). In planarians, classical **enzymatic digestion** is well-established and widely used (Hayashi et al., 2006; Moritz et al., 2012; Önal et al., 2012). However, the concerns about cellular stress and gene expression artifacts in live cell dissociation (Chaudhry, 2008; Denisenko et al., 2020; Huang et al., 2010; van den Brink et al., 2017) made us avoid this approach. **Single-nuclei isolation** protocols were also discarded, as they eliminate all mature mRNA contained in the cytoplasm, leading to a substantial loss of transcriptomic information.

We envisioned a protocol that was able to freeze the transcriptomic machinery of the cell while preserving the cytoplasmic RNA. In addition, we wanted to implement **FACS** to enrich singlets and obtain cleaner datasets with less cell debris (to reduce the likelihood of free-floating RNAs) and aggregates (to reduce the presence of doublets). However, the external location of the FACS facility made sample **cryopreservation** a must. To meet these requirements and overcome the limitations of traditional approaches we developed **ACME** (ACetic-MEthanol), a sample preparation protocol for simultaneous dissociation, fixation and permeabilization of whole-cells that, additionally, allows sample cryopreservation and FACS sorting.

To complete our custom workflow, we needed an affordable single-cell platform that was easy to implement from scratch in a new lab. This had to provide high throughputs and be flexible to multiplex different samples in one experiment. Thus, we opted for *in situ*-barcoding based platforms. Particularly, we chose **SPLiT-seq** (Split-pool ligation-based RNA-seq) because its four indexing rounds provide enough combinatorial power to process up to hundreds of thousand cells per experiment. At the same time, SPLiT-seq is easily scalable and has a lower price per cell compared to other platforms (Rosenberg et al., 2018).

2. MACERATION: AN OLD PROTOCOL FOR TISSUE DISSOCIATION

All previous single-cell studies in planarians followed a similar strategy for sample preparation. Planarian tissues were dissociated by enzymatic digestion, using trypsin or collagenase, and the resulting cell suspensions were purified by FACS (Fincher et al., 2018; Hayashi et al., 2010; Plass et al., 2018; van Wolfswinkel et al., 2014; Wurtzel et al., 2015; Zeng et al., 2018). Some studies also included sample preservation strategies, such as cryopreservation and methanol fixation, in order to process multiple samples and conditions (Fincher et al., 2018; Plass et al., 2018).

However, as seen in the introduction (section 2.1.1), non-enzymatic acidic formulas had been used since the 20th century to dissociate planarian tissues in microscopy studies. This strategy, commonly called **maceration**, preserves the morphology of the cell and allows it to be sorted under the microscope (**Figure 2.1 A**). Acetic acid maceration was developed in *Hydra* (David, 1973; Schneider, 1890) and later adapted to planarian (Baguña and Romero, 1981). To improve the preservation of planarian cell types, the authors modified the original formula adding **methanol**. Based on morphology, they characterized 13 cell identities in two planarian species, *S. mediterranea* and *Girardia tigrina* (**Figure 2.1 B**), and studied their distribution during growth, degrowth and regeneration.

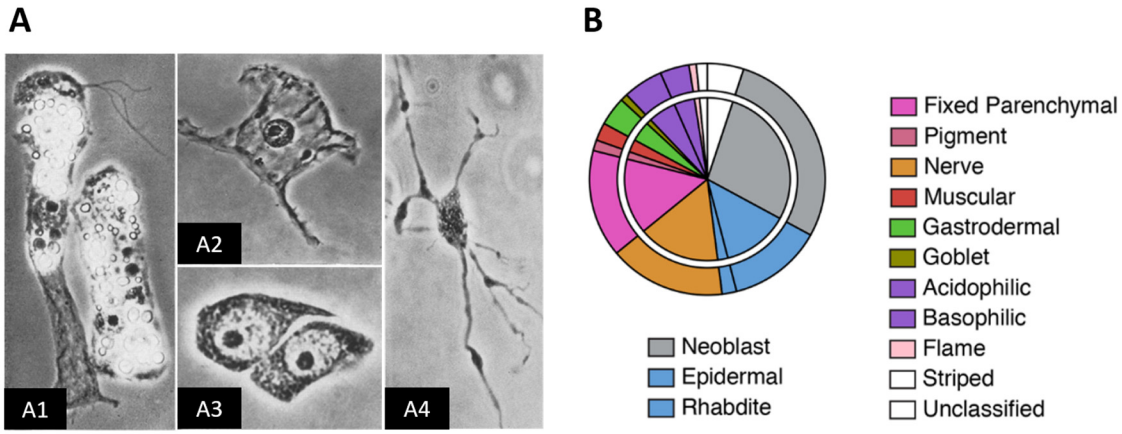


Figure 2.1 Maceration in different invertebrates. A) Macerated cells from *Hydra attenuata* exhibit distinct morphologies: epithelial cells (A1) and digestive cell (A2) at x590 magnification; I-cells (A3) and nerve cell (A4) at x1180 magnification. *Adapted from David, 1973.* **B)** Planarian cell types and their proportions, characterized by microscopy in Baguña and Romero, 1981. *Adapted from Plass et al., 2018.*

During the twentieth century, maceration continued to be used in microscopy studies (Romero and Baguña, 1991), but nowadays this method is rare and restricted to a few applications (Bradshaw et al., 2015; Thommen et al., 2019). Despite its low use, the **dissociative** and **fixative** nature of the acetic acid makes it an interesting reagent for sample preparation. For this reason, and inspired by the maceration formula, we developed **ACME**, a dissociation protocol based on

acetic acid with modifications to make it compatible with scRNA-seq. With ACME, we aimed to explore the potential of maceration in single-cell transcriptomics, where nothing similar had been tested before.

3. COMPARISON OF SINGLE-CELL METHODS

Over the years, multiple single-cell transcriptomic protocols have been developed, bringing innovation in different aspects. But some have gained more popularity than others (**Figure 2.2**). Overall, newer technologies have displaced manual **plate-based methods**. Although some automatized platforms, such as Smart-seq (Hagemann-Jensen et al., 2020), Fluidigm C1 (Achim et al., 2018) and MARS-seq (Keren-Shaul et al., 2019), continue to be updated and moderately used. The market is currently led by **droplet-based methods**. Specially by Chromium, the commercial platform offered by 10x Genomics (Svensson et al., 2020). Non-commercial approaches, like InDrop and Drop-seq, are used less frequently. Meanwhile, newer ***in situ* barcoding methods** are still settling in the scientific community. But the recent publication of massive datasets, including thousands to millions of cells (Cao et al., 2020, 2019; Martin et al., 2021; Rosenberg et al., 2018), suggests a promising future for these technologies.

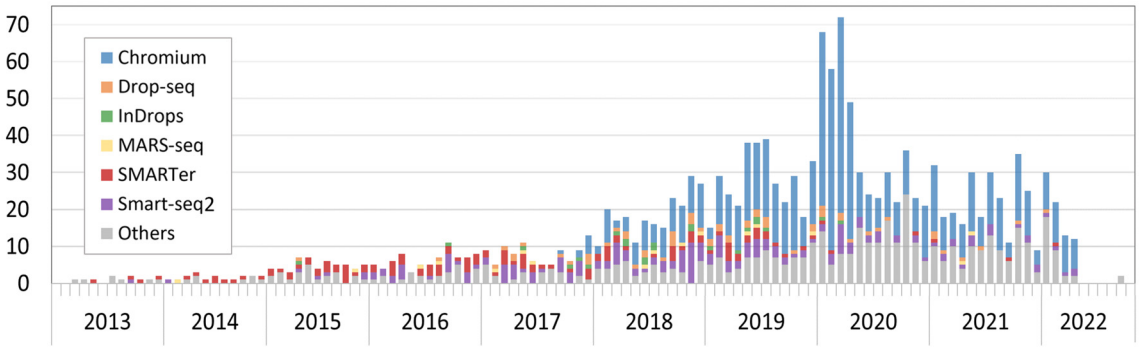


Figure 2.2 Trends in single-cell transcriptomics over the years. Single-cell studies published from 2013 to 2022. Bar graph stacked by technology. *Data extracted from www.nxn.se/single-cell-studies/gui.*

Trends aside, all single-cell protocols have advantages and disadvantages. Choosing the right one requires a careful evaluation of the research goal, the needs of the lab and the specific characteristics of each technology (**Table 2.1**).

First, depending on their **transcript coverage**, single-cell approaches can be classified as **full-length** (whole transcript capture) and **3' or 5'-end counting** (partial capture). Currently, Smart-seq and Fluidigm C1 are some of the few platforms offering full-length RNA coverage. Partial 3'/5'-end capture is usually enough to characterize gene expression, but more detailed studies (e.g. alternative splicing) may require full-length techniques. On the other hand, 3'/5'-end

protocols are tag-based, as they incorporate **UMI** tags to each individual transcript (Chen and Ginhoux, 2018; Kulkarni et al., 2019). UMIs allow to discriminate PCR duplicates, providing a more accurate quantification of gene expression. Smart-seq3 is the only full-length platform that includes UMI tagging (Hagemann-Jensen et al., 2020). In turn, full-length methods are more **sensitive**. At the same sequencing depth, they provide a better detection of lower expressed transcripts and, thus, are able to profile more **genes per cell** (Chen and Ginhoux, 2018; Ziegenhain et al., 2017).

The main limitations of full-length methods are their low **throughput** (number of cells processed per run) and high **costs**. In full-length mode, Fluidigm C1 can only profile 96 cells per run, and has a higher cost than any other platform (Chen and Ginhoux, 2018). Meanwhile, Smart-seq typically processes less than 1,000 cells, and cost about half as much as Fluidigm C1 (Hagemann-Jensen et al., 2020). The 3'/5'-end approaches are much cheaper and higher-throughput. Protocols such as MARS-seq 2.0 or Drop-seq process between 1,000-10,000 cells per run, and can be up to 10 times more affordable than full-length strategies (Keren-Shaul et al., 2019; Macosko et al., 2015). Finally, ***in situ* barcoding methods** are the most **cost-effective** (about 100 times cheaper than Fluidigm C1) and can profile up to hundreds of thousands of cells (Martin et al., 2021; Rosenberg et al., 2018).

Another strength of *in situ* barcoding methods is their **scalability**. Plate-based protocols can be scaled adding more and larger plates, but this only increases their capacity by one cell per well. Droplet-based protocols are also hard to scale up, as they rely on the design and capture rate of the microfluidic device. On the contrary, *in situ* barcoding protocols are highly scalable, because the simple use of larger plates can add thousands of cells to the experiment. For instance, the combination of three 96-well plates (96x96x96) reaches up to 100,000 cells per experiment (Cao et al., 2017; Rosenberg et al., 2018). By switching to 384-well plates (384x384x384), the same protocol can easily scale to over 1 million cells per run (Martin et al., 2021).

When the aim is to process different samples or conditions at the same time, it is important to consider the multi-sampling capacity of the platform (**multiplexing**). FACS-sorted plate-based approaches, like MARS-seq, can multiplex up to one sample per well (Jaitin et al., 2014; Keren-Shaul et al., 2019). Droplet-based protocols can only be multiplexed by increasing the number of microfluidic channels. 10x Chromium has the largest capacity, with an 8-channel chip that can run up to 8 samples simultaneously (Zheng et al., 2017). On the other hand, *in situ* barcoding technologies can multiplex as many samples as the number of wells in the plate used for the first indexing round (Cao et al., 2017).

Finally, each protocol requires a different **level of expertise**, **optimization time** and **equipment** to be established in a new lab. In this sense, *in situ* barcoding methods are very easy to perform and only require common lab equipment (Cao et al., 2017; Rosenberg et al., 2018). On the contrary, MARS-seq and Smart-seq require a FACS sorter for cell capturing (Hagemann-Jensen et al., 2020; Keren-Shaul et al., 2019), while Drop-seq and InDrop need specialized microfluidic devices and certain expertise in their use (Klein et al., 2015; Macosko et al., 2015). Commercial alternatives, such as Fluidigm C1 and 10x Chromium, require less optimization, as they use already standardized and optimized protocols. But, in turn, they demand the purchase of expensive platforms or external services.

Table 2.1 Comparison of single-cell methods (Cao et al., 2017; Hagemann-Jensen et al., 2020; Keren-Shaul et al., 2019; Klein et al., 2015; Macosko et al., 2015; Martin et al., 2021; Rosenberg et al., 2018; Zheng et al., 2017; Ziegenhain et al., 2017)**.**

	Plate-based			**Droplet-based**		***In situ* barcoding-based**	
	Fluidigm C1	Smart-seq3	MARS-seq 2.0	InDrop/Drop-seq	10x Chromium	SPLiT-seq	sci-RNA-seq3
Transcript coverage	Full-length	Full-length	3'-end	3'-end	3'/5'-end	3'-end	3'-end
UMI	No	Yes	Yes	Yes	Yes	Yes	Yes
Sensitivity	+++	+++	+	+	++	+	+
Multi-sampling	No	No	Yes	No	Yes	Yes	Yes
Cells per run	100	1k	1-10k	1-10k	1-10k	10-100k	10-400k
Special equipment	C1 platform	FACS	FACS	Microfluidic chip	Chromium system	None	None
Comercial	Yes	No	No	No	Yes	Yes	No
Cost per cell	+++++	++++	++	++	+++	+	+

In summary, full-length approaches are preferable when the aim is to deep-sequence a small number of cells (<1000) in high detail. Meanwhile, 3'/5'-end counting is a better strategy to profile thousands of cells with less coverage and detail but at a cheaper price. Among 3'/5'-end technologies, droplet-based protocols are a middle-of-the-road option in terms of throughputs and price. In particular, commercial platforms offer a more expensive but simpler alternative, which may appeal to high-budget laboratories, or to those who do not have time for optimization or want to perform single-cell transcriptomic only once or twice. On the other hand, *in situ* barcoding protocols offer multiple advantages for a more regular basis use, in terms of simplicity, price, throughput, scalability and multi-sampling capacity. For these reasons, we chose **SPLiT-seq** as a candidate for our single-cell workflow.

RESULTS

1. FROM MACERATION TO ACME

The original maceration formula described in *Hydra attenuata* contained glycerine, glacial acetic acid and water (1:1:13). *Hydra* tissues were dissociated in this formula for a few minutes at room temperature (RT), with agitation, and then fixed in 0.1 volume of 20% formaldehyde or 1% osmium tetroxide (David, 1973). Maceration was later adapted to planarian, adding methanol to improve the preservation of cell morphology (Baguña and Romero, 1981). The resulting formula contained methanol, glacial acetic acid, glycerol and water (3:1:2:14). *Planarians* were macerated at 8-10ºC for 24-48 hours, with very gentle agitation. After dissociation, cells were fixed in a similar manner as in the *Hydra* protocol. Since then, the use of the maceration solution has been limited to a few microscopy studies on hydra and planarian. In these, the authors generally used the original formula, without methanol (Bradshaw et al., 2015; Thommen et al., 2019).

We started working from **Romero and Baguñà's** formulation (3 methanol: 1 acetic acid: 2 glycerol: 14 water) to combine acetic acid dissociation with the fixative and permeabilizing properties of methanol. We coined the name **ACME (Acetic-Methanol)** in reference to these two active ingredients. Most optimization steps were performed using the planarian species *Schmidtea mediterranea.*

First, we tested different **acetic acid proportions** (3x, 2x, 1x and 0.5x). Doubling the amount of acid (3:2:2:13) resulted in faster and more homogeneous dissociations, without rupturing the cells. Then, we replaced acetic acid with different molarities of citric acid (0.25M, 0.5M, 1M, 2M and 3M), another dissociative agent, and compared the performance of both acidic solutions. **Citric acid** was discarded, as it resulted in a worse dissociation of planarians pharynxes, which were visible after incubation. We also removed the extra fixation step in formaldehyde or osmium tetroxide, and reduced the incubation time to ~1 hour, at room temperature, to keep the protocol as simple, fast, and non-toxic as possible.

These changes allowed us to obtain good tissue dissociation. However, the RNA quality was very variable. We ruled out that the maceration solution was affecting the RNA, since the same formulation and incubation conditions randomly returned good and bad results for RNA integrity. To fix this problem, we started using ultrapure water and single-use plasticware in our reactions. Moreover, we added **N-acetyl-L-cysteine (NAC)** to remove planarian mucus, since the RNases present in these secretions could be the cause of RNA degradation. NAC is a well-

established reducing agent, commonly used in planarian experiments as mucolytic (Pearson et al., 2009; Thommen et al., 2019). It breaks the disulphide bonds in proteins, avoiding RNases activity (Aldini et al., 2018; Chen et al., 2004). We used NAC for a quick cleaning prior dissociation. Planarians were soaked in 100-500 µL (enough to cover the animals) of 7.5% NAC and incubated for a couple minutes. Maceration was then poured over NAC and incubated for 1 hour. The resulting cell suspensions consistently maintained good RNA quality. We also tried to add NAC directly to the maceration solution, obtaining similar results. Moreover, we noticed NAC made tissues more permeable to the formula, facilitating dissociation.

Other **reducing agents**, like hydrochloric acid, DTT (dithiothreitol), β -mercaptoethanol and TCEP (tris (2-carboxyethyl) phosphine) (Chen et al., 2004; Han and Han, 1994), were tested to protect the RNA during and after dissociation. However, they resulted in equal or worse performance than NAC, and entailed additional problems of toxicity, cell rupture and buffer aggregation (Yang et al., 2015).

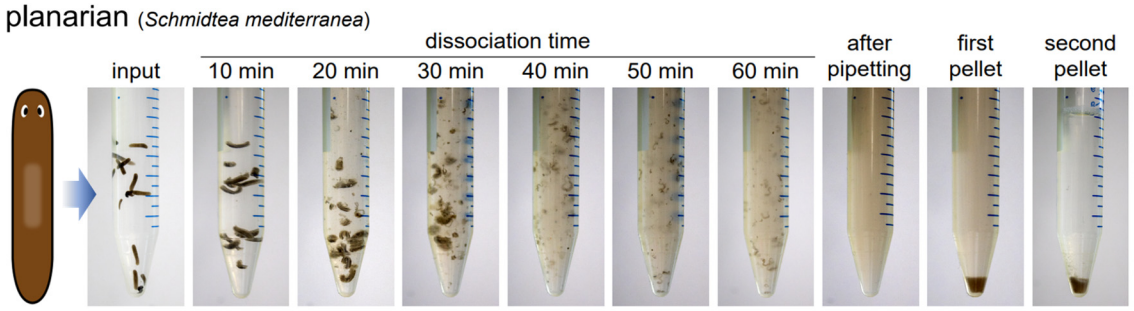


Figure 2.3 ACME dissociation in planarians. From left to right: live worms used as input, ACME dissociation after 10-60 min, cell suspension after final pipetting, pellet after first centrifugation and pellet after second centrifugation (with cell resuspended in 1x PBS 1% BSA buffer).

Thus, our final **ACME** protocol for planarians used 100-500 µL of 7.5% NAC, 1 mL glycerol, 1 mL acetic acid, 1.5 mL methanol and 6.5 mL ultrapure water (~10 mL per reaction). We normally add 100-200 µL of animal biomass (whole worms) per reaction. In ACME, samples dissociate after 45-60 minutes at RT, with see-saw agitation. After incubation, samples are pipetted up and down to complete tissue dissociation. Then, cells are centrifuged twice to remove the ACME solution, and resuspended in 1x phosphate-buffered saline (PBS) with 1% bovine serum albumin (BSA) to prevent reaggregation (**Figure 2.3**). Finally, samples are filtered to remove undissociated fragments and bigger aggregates. For more details, see Chapter V: Methods.

2. ACME-CELLS PRESERVE MORPHOLOGY AND CAN BE FACS-SORTED

For cell imaging, ACME-dissociated cells are stained with **DRAQ5**, a far-red DNA dye, and **Concanavalin A (ConA)** conjugated with Alexa Fluor 488. Concanavalin A is a lectin that binds to mannose residues and glycoproteins of the cell surface. However, in fixed and permeable cells, it can cross the membrane and stain the cytoplasm. After staining, cells can be visualized by microscopy and flow cytometry. During optimization, confocal microscopy was used to reveal the composition of ACME dissociates and to evaluate fixation. Like other dissociation protocols, ACME generates a mixture of cell **aggregates**, **singlets**, and **debris** (**Figure 2.4 A**). On the other hand, the correct sample fixation is inferred by the preservation of different cell morphologies (**Figure 2.4 B**).

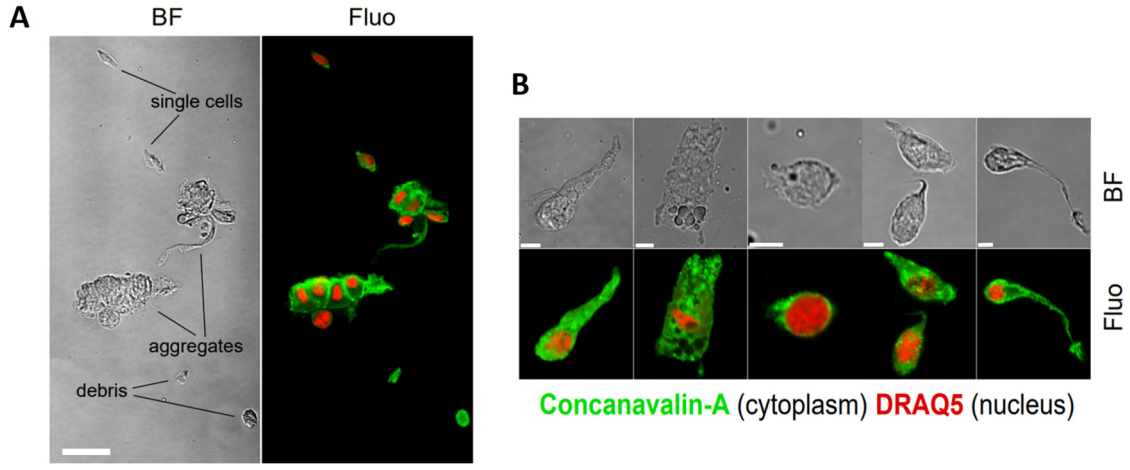


Figure 2.4 Confocal imaging of ACME dissociates. Bright field (BF) and fluorescence microscopy (Fluo) images of *S. mediterranea* ACME-cells stained with DRAQ5 (red) and ConA (green). **A)** Cell suspension mix of single cells (singlets), aggregates and debris (Scale bar = 50 µm). **B)** Detail of morphology preservation in different cell types (Scale bars = 5 µm).

Flow cytometry is used to count and enrich single cell populations. A first plot of DRAQ5 vs ConA gives us an overview of the sample (**Figure 2.5 A**). After gating out unwanted events (**Figure 2.5 B**), single-cell populations can be sorted by FACS. For gating, we apply an initial forward scatter (FSC) filter to select singlets. Then, we only select ConA (cytoplasm) and DRAQ5 (nucleus) positive events to discard cellular fragments (debris). Additionally, we apply a second singlet filter to discard more cellular aggregates using DRAQ5 correlations (**Figure 2.5 C**).

Singlets are distributed in two distinct populations, that we call G1 and G2. **G1 population** has the lowest DNA-content and corresponds to G1/G0 phases of the cell cycle. The **G2 population** has double DNA-content, matching with cells in G2/M phase. S-phase cells are difficult to resolve with our staining conditions. The proportions of both populations are stable, with ~90% of singlets in G1 and ~10% in G2 (**Figure 2.5 D**).

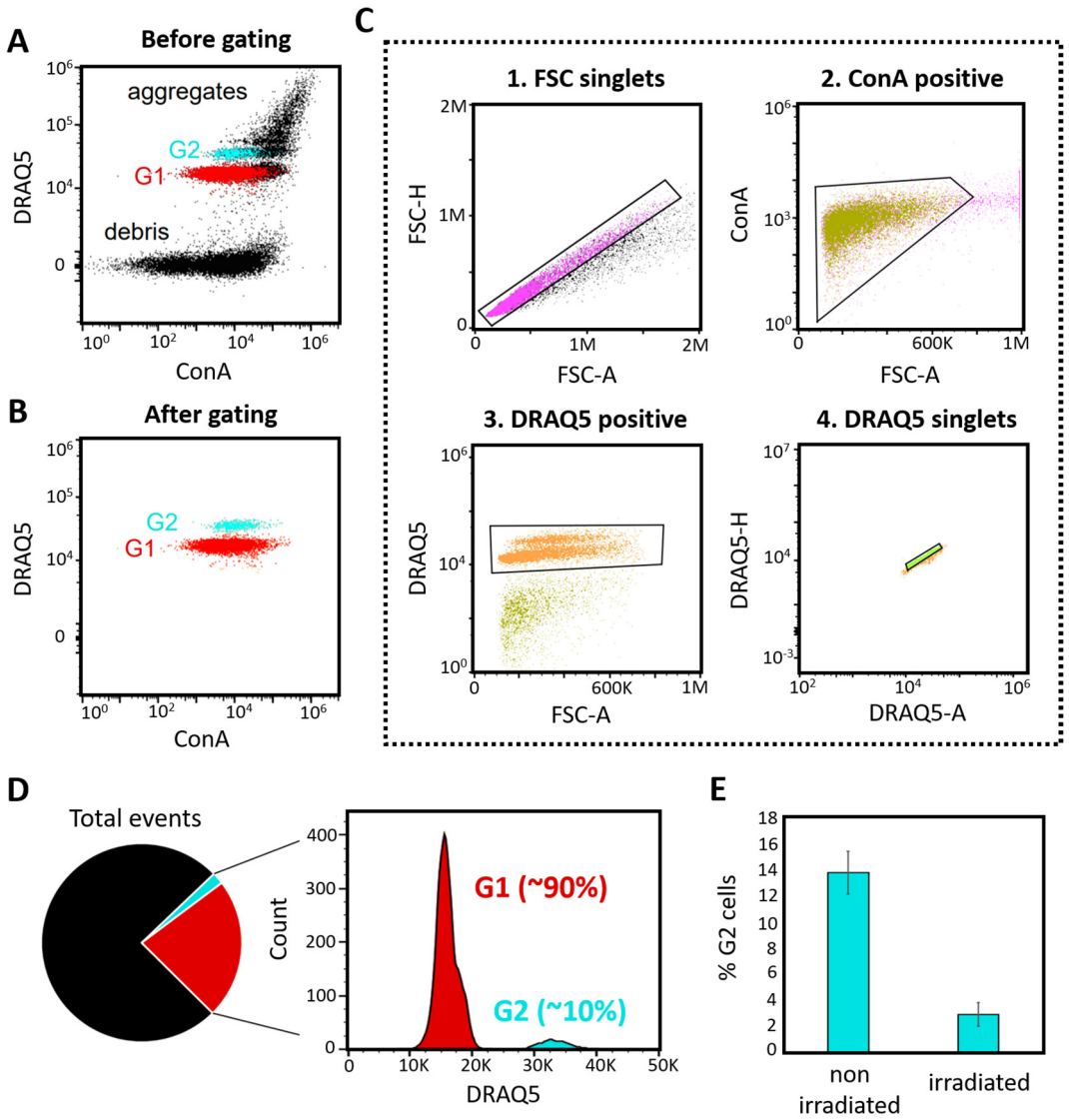


Figure 2.5 Flow cytometry profiles of *S. mediterranea* ACME-cells stained with DRAQ5 and ConA. A) Total events before gating **B)** Events after gating, corresponding to FACS-sorted populations (G1 and G2) **C)** Gating strategy: singlets filter by FSC (1), selection of ConA+ events (2), selection of DRAQ5+ events (3), singlets filter by DRAQ5 (4). Singlets are selected based on their well-correlated area vs height signal **D)** Relative proportion of singlets in a typical *S. mediterranea* sample, and histogram of their DNA content (linear scale) showing the percentages of G1 and G2 **E)** Percentage of G2 cells in non-irradiated versus irradiated animals. The values correspond to the average of 3 replicates and the error bars to the standard deviation.

The distribution of these cell populations highly correlates with the flow cytometry profiles of planarian cells obtained by enzymatic digestion (Hayashi et al., 2006). G1 would correspond to what planarian FACS protocols refer to as X2 (progenitors) and XIS (differentiated cells). Meanwhile, G2 would correspond to the X1 population (proliferating cells). Similar to X1, our G2 population is sensitive to **irradiation** (**Figure 2.5 E**).

The percentage of singlets after ACME dissociation in planarian normally oscillated around 20-40% of total events. We have observed this percentage varies between species and

experimental performance. Recent improvements in our ACME protocol (not applied to the results presented here) have enabled us to obtain up to 40-50% singlets in planarian. Despite the initial percentages, **FACS enrichment** allows our samples to reach a consistent 80-95% of singlets after sorting.

3. ACME-CELLS CAN BE CRYOPRESERVED MULTIPLE TIMES

Different factors, such as a time-spaced sample collection or logistical constraints, can make **cryopreservation** a must during sample preparation. Our workflow includes two steps of cryopreservation, before and after FACS, to completely decouple sample collection, cell sorting and sample processing (**Figure 2.6 A**).

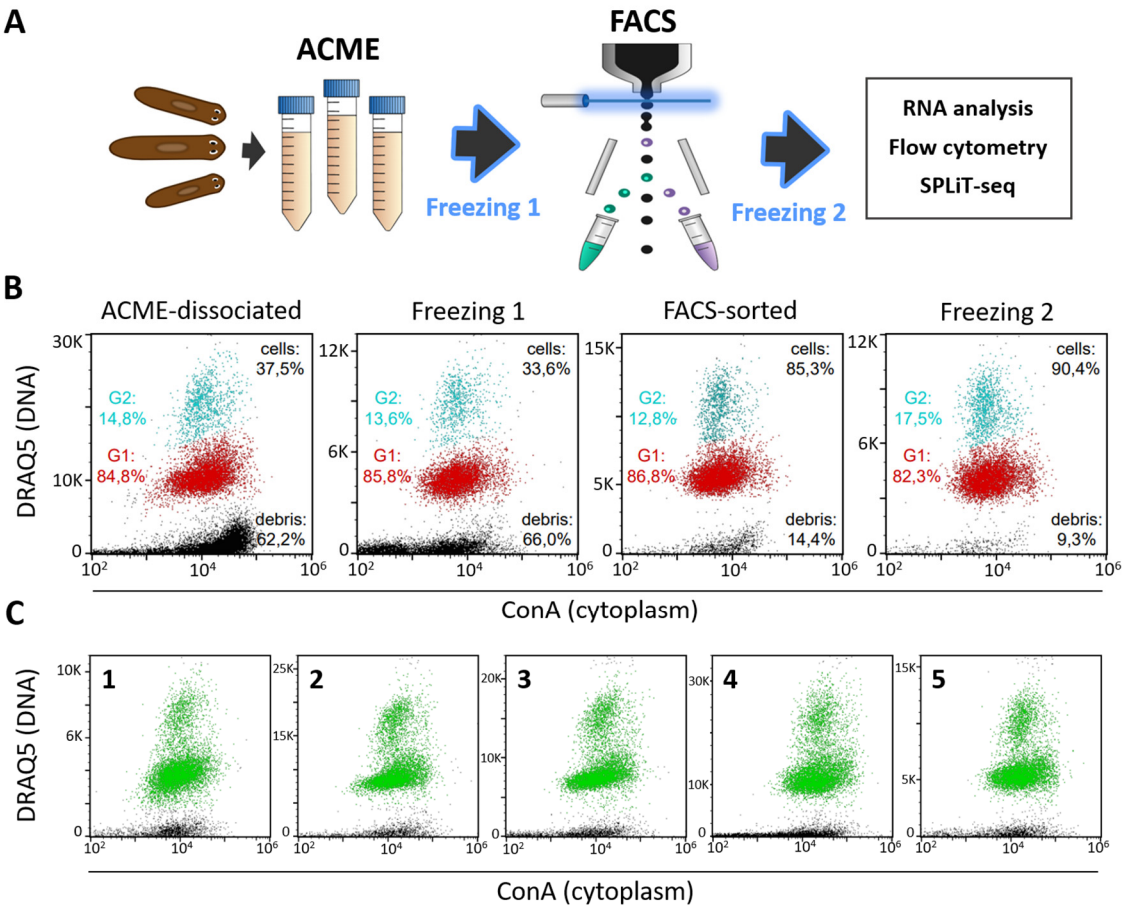


Figure 2.6 Resistance to cryopreservation and FACS sorting of *S. mediterranea* ACME-cells. A) Experimental workflow **B)** Flow cytometry profiles after different workflow steps. From left to right: ACME-dissociation, first freezing, FACS sorting and second freezing. Percentages of debris (black), total single cells (cells), G1 (red) and G2 (blue) populations are shown for each condition. Aggregates were previously removed by an FSC filter **C)** Flow cytometry profiles after 1 to 5 freezing cycles. The relative proportions of DRAQ5+ and ConA+ events (green) and debris (black) remain stable.

ACME-cells can be easily cryopreserved in PBS 1% BSA buffer by adding 10% volume of dimethyl sulfoxide (**DMSO**) (Guillaumet-Adkins et al., 2017). Cells in DMSO can be directly frozen at -20ºC or -80ºC, and stored for months. To test cryopreservation resistance, we compared the flow cytometry profiles of our cells under different conditions.

First, we tested cell integrity in the same sample after each of the following workflow stages: ACME dissociation, first cryopreservation, FACS sorting and second cryopreservation (**Figure 2.6 B**). In every case, single cell populations (G1 and G2) remained stable and maintained their relative proportions. After FACS sorting, singlet-enrichment is evidenced by a drop in the percentage of debris and an increase in the total percentage of single cells to 85-90% (**Figure 2.6 B**). To further test the resistance to cryopreservation, a dissociated sample was subjected to five subsequent freezing cycles (**Figure 2.6 C**). After each cycle, cells were centrifuged and resuspended in fresh buffer and DMSO. We found no differences in cell proportions over the cycles, proving that ACME-cells can be cryopreserved multiple times.

4. ACME-CELLS RETAIN HIGH QUALITY RNAs

Next, we checked the **RNA quality** of ACME-cells. For this assessment, we used an Agilent 2100 Bioanalyzer to obtain the **RNA integrity number (RIN)**. We found some limitations, as the algorithm used to calculate the RIN has been trained exclusively on mammalian samples (Schroeder et al., 2006). Therefore, it sometimes fails to calculate RIN values, especially in non-vertebrate samples.

In many cases, this is due to a phenomenon known as **'hidden break'**, in which the 28S ribosomal subunit is cleaved in two pieces of approximately the same size. The algorithm mistakes this hidden break for RNA degradation, underestimating the RIN value. The hidden break is commonly found in the protostome phyla, including Platyhelminthes (Natsidis et al., 2019). For this reason, RIN values in our planarian samples are never above 8, and sometimes the algorithm does not calculate them. To overcome this challenge, we registered RIN values when possible and **inferred** the rest using a linear regression. To create this regression, we correlated the percentage of RNA signal (area) contained in the ribosomal peaks with the RIN values calculated by the algorithm in other planarian samples.

By comparing the RNA quality of ACME-cells in different conditions, we found that the main factors leading to degradation were **time**, **temperature** and **FACS sorting**. As controls, we used RNA extracted from live worms directly in Trizol (**Figure 2.7 A**). To test the effects of time and temperature, we split a tube of ACME-cells in two samples, and incubated them in PBS 1% BSA

for six hours. One sample was kept at room temperature (**Figure 2.7 B**) and the other one on ice (**Figure 2.7 C**). Small aliquots were taken every hour from each sample to extract RNA. Aliquots incubated at RT started to degrade after 1 hour, while the ones kept on ice resisted degradation for up to 6 hours. Therefore, although RNA integrity progressively decreases after dissociation, keeping cells in cold conditions is sufficient to safeguard RNA for scRNA-seq experiments.

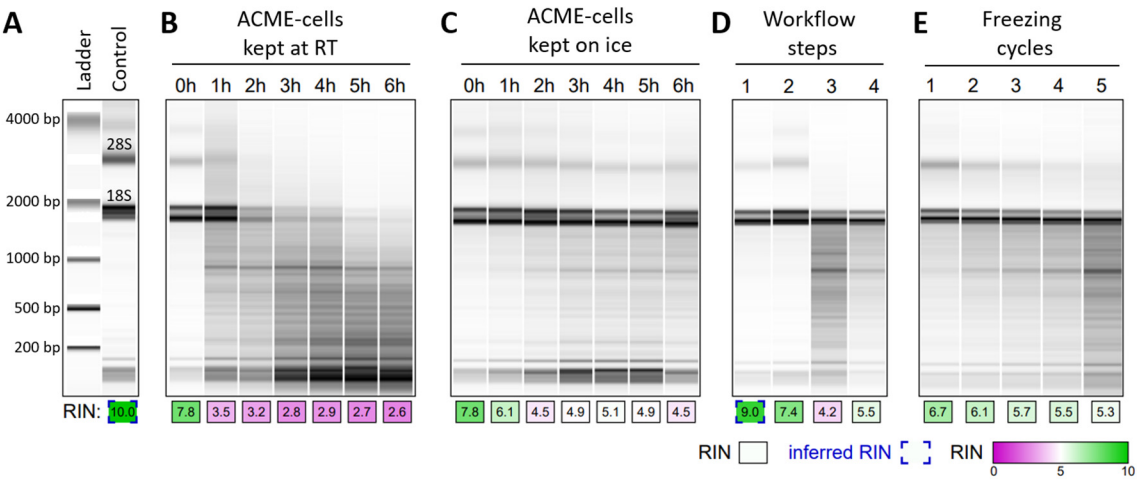


Figure 2.7 Bioanalyzer profiles and RIN values of ACME-cells under different conditions. Inferred RIN values are shown in open blue boxes. **A)** Size ladder and control RNA from fresh live worms, indicating the 18S and 28S ribosomal bands. **B-E)** RNA quality after: 0-6 hours incubation at RT **(B)**, 0-6 hours incubation on ice **(C)**, different workflow steps: dissociation (1), first cryopreservation (2), FACS sorting (3) and second cryopreservation (4) **(D)**, and 1 to 5 freeze/thaw rounds **(E)**.

We also tested RNA quality after different workflow steps: dissociation (1), first cryopreservation (2), FACS (3) and second cryopreservation (4) (**Figure 2.7 D**). Here, we observed that FACS affects RNA the most. This can be explained by the experimental time required for FACS (3-6 hours), and the difficulty to guarantee an RNases free environment during sorting. Finally, we confirmed that repeated freezing cycles (1 to 5) have a slightly negative impact on RNA, but do not lead to its complete degradation (**Figure 2.7 E**).

Overall, ACME-cells can go through cryopreservation and FACS and still retain high quality RNA, as long as kept in cold conditions. Our results provide evidence that RNA in fixed cells is highly exposed to degradation, and factors like handling time, temperature and RNases contamination can critically affect it. We have identified water and **BSA powder** as main potential sources of contamination. In our experience, contamination of these reagents can lead to complete RNA degradation.

5. ACME IS A SPECIES-VERSATILE METHOD

To test the performance of ACME in other organisms, we used it to dissociate the following species: *Nematostella vectensis* (sea anemone), *Dugesia japonica* (planarian), *Pristina leidyi* (annelid), *Lymnaea stagnalis* (snail), *Parasteatoda tepidariorum* (spider), *Drosophila melanogaster* (fruit fly), *Mus musculus* (mouse) and *Danio rerio* (zebrafish). This set of animals includes species belonging to the major **metazoan lineages**: non-bilaterian, lophotrochozoans, ecdysozoans and deuterostomes. Furthermore, it encompasses animals from terrestrial, freshwater and marine **habitats**, and from diverse **developmental stages**, such as embryos, larvae, juveniles and adults (**Figure 2.8**).

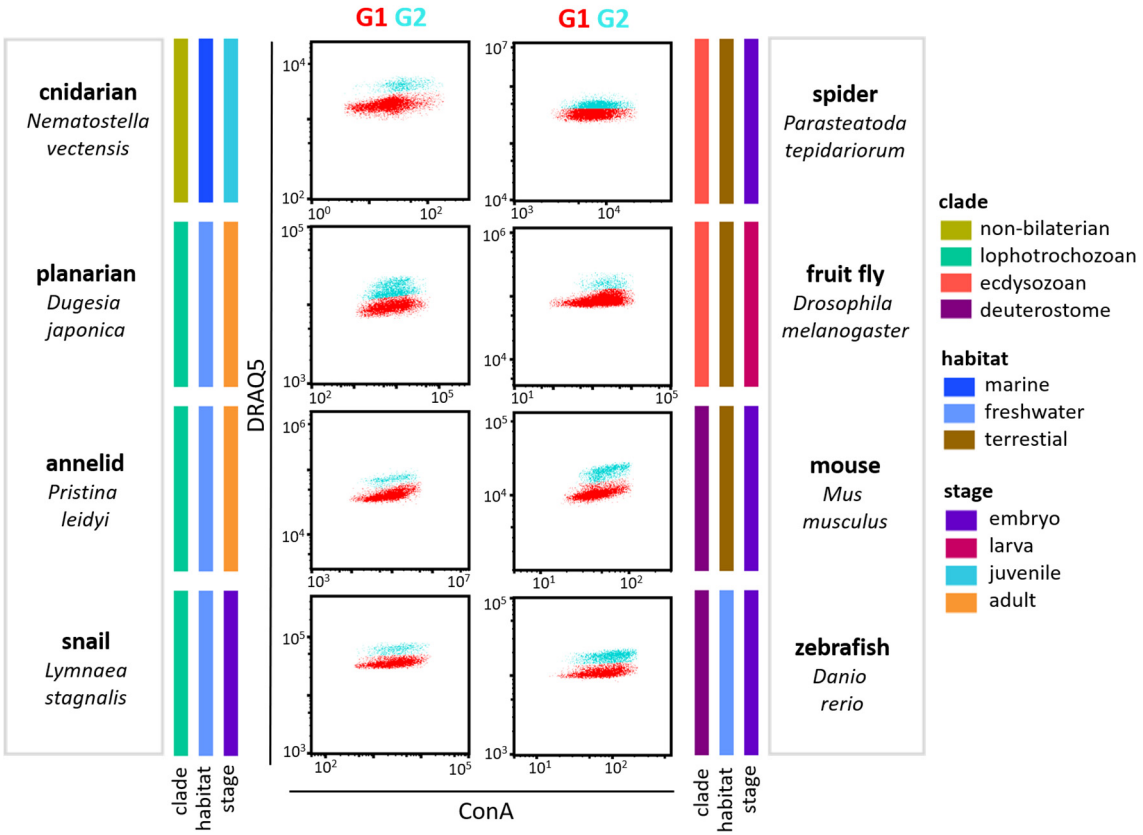


Figure 2.8 Species versatility of ACME. Flow cytometry profiles (gated) of ACME-cells from different metazoan: sea anemone juveniles (*Nematostella vectensis*), planarians (*Dugesia japonica*), annelid adults (*Pristina leidyi*), snail larvae (*Lymnaea stagnalis*), spider stage 7 embryos (*Parasteatoda tepidariorum*), fruit fly 3rd instar larvae (*Drosophila melanogaster*), mouse E11.5 embryos (*Mus musculus*), and zebrafish 1-day embryos (*Danio rerio*). All cells are stained with DRAQ5 and ConA. G1 and G2 populations are marked in red and blue, respectively. The bar colour system refers to each animal clade, habitat and developmental stage.

Sample preparation was performed by different labs and the protocol was adapted depending on the organism. Our collaborators helped with the optimization and dissociation of *Pristina*

leidyi, Nematostella vectensis, *Lymnaea stagnalis*, *Parasteatoda tepidariorum* and *Mus musculus.*

ACME cannot penetrate hard shells, chorions, cocoons, or vitelline membranes. Thus, zebrafish, spider and snail embryos required a **pre-treatment** to remove these harder layers. The **animal biomass** and **ACME volume** used per reaction were set for each species according to the animal size and sample availability. **ACME formulation** was generally the same. Only spider and snail protocols used a different formula, with half acetic acid, as they were optimized before double acetic acid conditions were tested. The **mechanical forces** applied were also heterogeneous. Soft-bodied animals, like planarians or annelids, completely dissociated with minimal mechanical forces (e.g. agitation, shaking and pipetting), but others such as zebrafish, snail and sea anemone needed of stronger homogenization (e.g. polytron pulses, gentleMACS). Finally, **incubation time** varied from 15-60 minutes between different animals. For more details, see Chapter V: Methods.

After dissociation, all samples were cryopreserved in 10% DMSO. To obtain the flow cytometry profiles, cells were thawed, stained, and gated as previously described for planarian. Despite the fact that the proportion of aggregates and debris was highly variable between species, we successfully isolated single cell populations (G1 and G2) from every animal (**Figure 2.8**).

Our results demonstrate that **ACME is a versatile method** that can be effectively used to dissociate and fix a wide range of organisms. Nonetheless, as with any other protocol, it requires optimization. The main adjustments to focus on when optimizing ACME in a new organism are the removal of harder tissues, selection of mechanical forces and incubation time. To protect RNA after dissociation, some animals can benefit from buffers with RNases inhibitors or lower percentage of BSA.

6. ACME CAN BE USED AS A FIXATIVE

ACME is a versatile protocol but is not universal, as some tissues may resist dissociation by this method. Therefore, we tested if ACME could be alternatively used to replace **formaldehyde (FA) fixation** after traditional enzymatic digestion. For this, we dissociated planarians with trypsin as previously described (Hayashi et al., 2006) and fixed the resulting cells with ACME or increasing concentrations of formaldehyde (**Figure 2.9 A**). ACME-fixation was performed with some modifications to the ACME-dissociation protocol (see Chapter V: Methods).

In trypsin-dissociated cells fixed with ACME, we detected G1 and G2 populations by flow cytometry using our normal staining and gating conditions (**Figure 2.9 B**). This indicates ACME

can be used as a fixative after trypsin, without rupturing the cells. Then, we compare the **RNA integrity** of ACME- and FA-fixed cells using an Agilent 2100 Bioanalyzer. RIN inference was performed as described in section 4. Most ACME-fixed cells had better RNA quality than cells fixed with formaldehyde, except for the 0.1% FA concentration (**Figure 2.9 C**). Arguably, these cells were unfixed, as formaldehyde is typically used at 1-4%. At these working concentrations, however, RNA was severely degraded. FA fixation is used in *in situ* barcoding-based protocols (Cao et al., 2017; Rosenberg et al., 2018) but is well-known to result in poor RNA integrity, which can make gene expression analysis difficult (Howat and Wilson, 2014; Russell et al., 2013). These experiments show that ACME can be used as an alternative to FA after enzymatic digestion to achieve higher RNA quality.

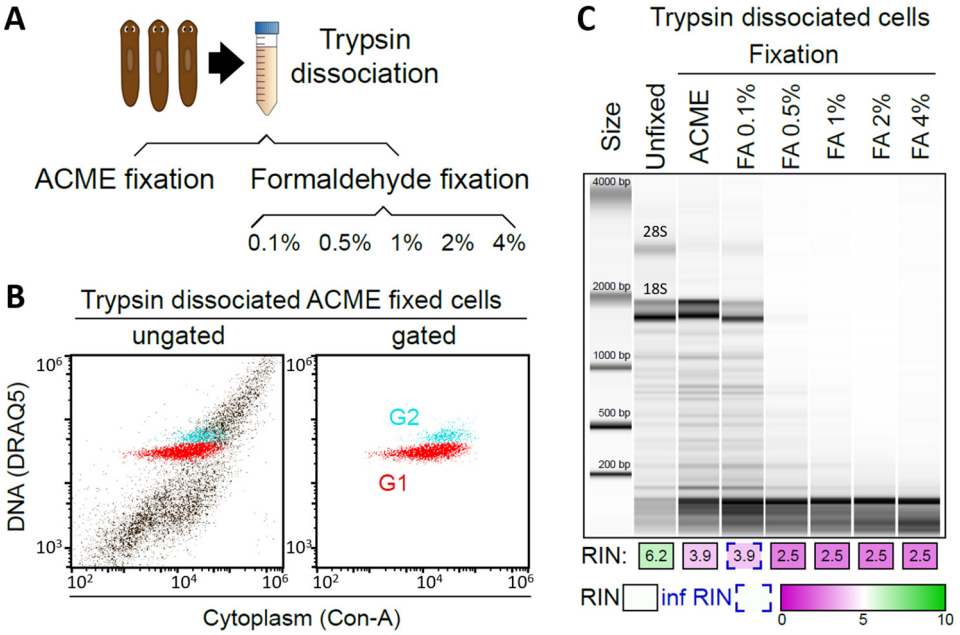


Figure 2.9 Comparison of ACME and FA fixation after trypsin dissociation. A) Experimental scheme **B)** Flow cytometry profiles (gated and ungated) of trypsin-dissociated cells fixed with ACME and stained with DRAQ5 and ConA **B)** Bioanalyzer profiles and RIN values for samples fixed with ACME or increasing concentrations of FA. Unfixed trypsin-dissociated cells were used as a control. Inferred RIN values are shown in open blue boxes.

7. ScRNA-SEQ OF CNIDARIAN ACME-CELLS USING 10x GENOMICS

Currently, 10x Chromium is the leading platform for single-cell transcriptomics (Svensson et al., 2020). For this reason, we wanted to evaluate the performance of ACME-dissociation in this technology. The results described in this section were generated by our collaborators from the Arnau Sebé-Pedrós lab, at the CRG of Barcelona. They profiled a **whole-body cell atlas** of the cnidarian ***Nematostella vectensis,*** using **ACME** in combination with **10x Chromium**. Briefly, they adapted the ACME protocol to dissociate juvenile *Nematostella* individuals and used FACS to

enrich the G1 population. Further cell processing was performed by 10x Genomics (**Figure 2.10 A**) (see Chapter V: Methods). A previous *N. vectensis* cell atlas had been published by the same lab using **enzymatic dissociation** (Liberase) and **MARS-seq** (Sebé-Pedrós et al., 2018b). This publication was used as a reference to validate our experiment.

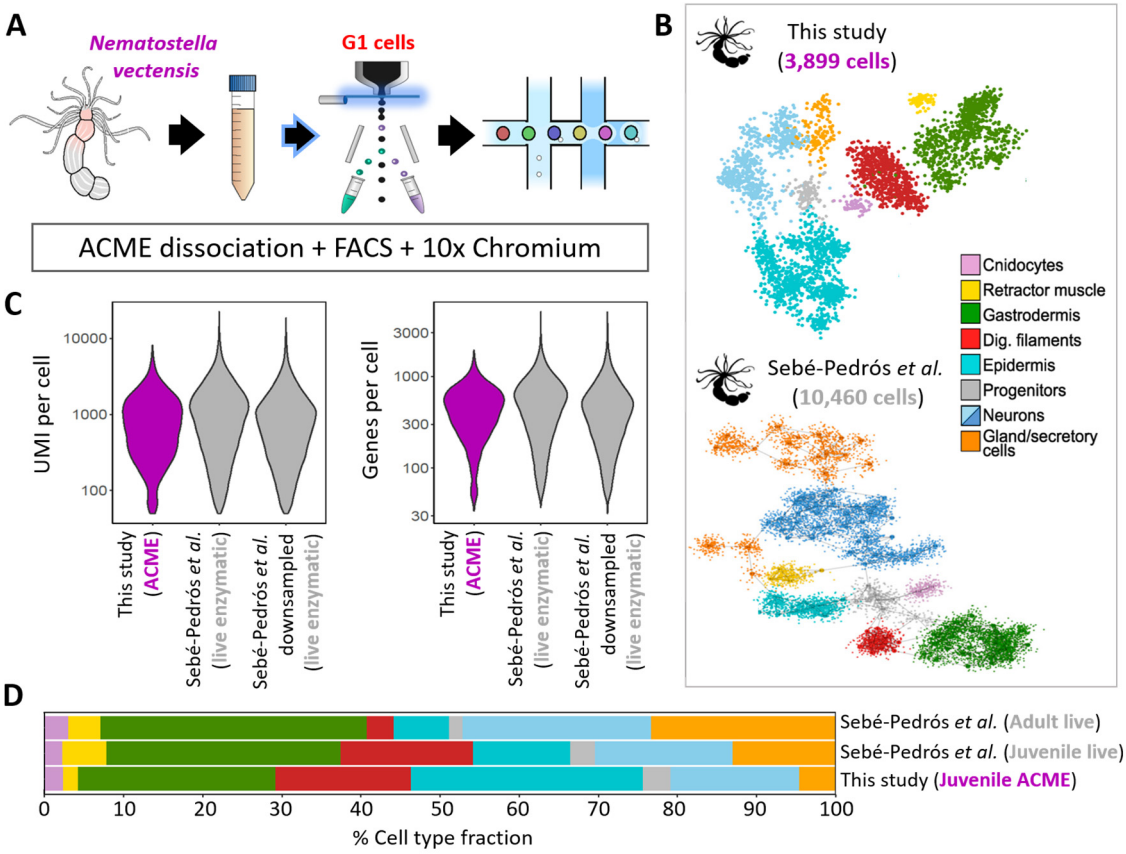


Figure 2.10 ScRNA-seq analysis of *Nematostella vectensis* ACME-cells using 10x Chromium. A) Experimental workflow **B)** 2D projections of the whole-body atlases of *N. vectensis* (juvenile) from this study and *Sebé-Pedrós et al. 2018*. Atlases are coloured by cluster identity. *Adapted from Sebé-Pedrós et al. 2018* **C-D)** Comparison of metrics between this study and previous publication (*Sebé-Pedrós et al. 2018*). Violin plots for the distribution of UMI and genes per cell **(C)**, and cell type fractions coloured by cluster identity **(D)**.

Our new *Nematostella* atlas profiled the transcriptome of **3,899 cells** and recapitulated the 8 major cell types uncovered by the previous publication: neurons, cnidocytes, epidermis, gland cells, muscle, digestive filaments, gastrodermis and progenitors (**Figure 2.10 B**); with an average of 671 UMIs per cell and 418 genes per cell. After down-sampling to a comparable number of reads per cell, these metrics highly resemble the ones from the previous study (696 UMIs and 381 genes per cell) (**Figure 2.10 C**). We also observed that **cell type abundances** were similar for the juvenile samples in both datasets (**Figure 2.10 D**), except for a more abundant epidermis, and less abundant muscle and gland cells. Despite this, our results were consistent and broadly comparable with previous observations, which demonstrates ACME can work in combination with droplet-based methods.

8. ScRNA-SEQ OF PLANARIAN ACME-CELLS USING SPLiT-SEQ

Finally, we tested our workflow proposal for single-cell transcriptomics, combining **ACME** and **SPLiT-seq** (Rosenberg et al., 2018). To assess the collision rate of our ACME-cells, we mixed two planarian species in the same experiment: ***Schmidtea mediterranea*** and ***Dugesia japonica***. Species-mixing is used in scRNA-seq to detect cells mapping to both species (collisions), which gives an estimation of cell re-aggregation and RNA leakage during the experiment. Each species was dissociated separately, and FACS-sorted to enrich singlet populations (G1 and G2). Dissociation, sorting and cell processing were performed on different days. Therefore, two cryopreservation steps were included before and after FACS (**Figure 2.11 A**). ACME-cells are fixed and permeabilized, and can be used for *in situ* barcoding without further preparation. Thus, we removed the formaldehyde fixation step included in the original SPLiT-seq protocol. Before SPLiT-seq, sorted-cells from both species were simply counted by flow cytometry, pooled together in equal amounts, and loaded on the first plate of barcodes.

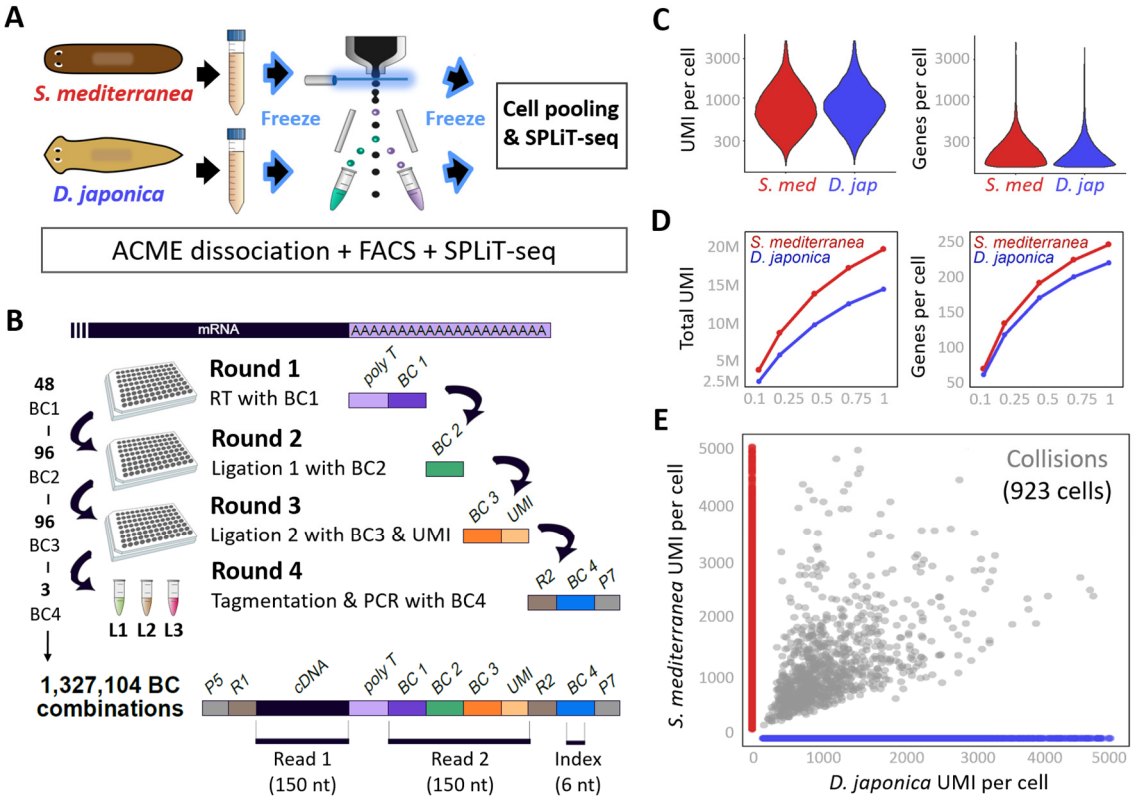


Figure 2.11 Overview and metrics of SPLiT-seq in planarian ACME-cells. A) Experimental workflow **B)** Schematic of SPLiT-seq, barcoding configuration (48x96x96x3) and sequencing plan **C)** Violin plots showing the distribution of UMIs per cell (left) and genes per cell (right) on each species **D)** Saturation plots for the total UMIs (left) and genes per cell (right) at different subsampling fractions of the complete sequencing depth **E)** Barnyard plot of *S. mediterranea* (red) and *D. japonica* (blue) UMI counts per cell. Collision events are coloured in grey.

SPLiT-seq has four rounds of barcoding: three plate reactions (indexed-RT, ligation 1 and ligation 2) and one indexed PCR amplification. In this experiment, we used a configuration of 48x96x96x3 barcodes, which gives a total of 1,327,104 potential combinations (**Figure 2.11 B**). To minimize **random collisions** (cells receiving the same combination of barcodes by chance), it is recommended to use less than 5% of the total combinations. Thus, this configuration can be used to profile up to 66,355 cells. For indexed-RT, we removed the 48 random hexamer RT barcodes described in Rosenberg *et al.* 2018*,* as they are less specific on capturing mRNAs, and only used 48 poly-dT barcodes. For PCR amplification, we used three *Round 4 barcodes,* as we generated 3 different sub-libraries (see all barcodes in Supplementary 1).

The experiment started with 480,000 total cells (10,000 per well; 5,000 from each species). After three rounds of barcoding, cells were pooled and counted again by flow cytometry. At this point, we recovered 40,000 cells (~8.5%) that were split in 3 sub-libraries. After library preparation (see Chapter V: Methods), samples were sequenced in a **NovaSeq 6000** Illumina platform following a **paired-end 150 bp** strategy. **Read 1** contained the transcript sequence, and **Read 2** the first 3 barcodes and the UMI. Additionally, we sequenced the 6 bp **index** included in the fourth barcode to identify the sub-libraries (**Figure 2.11 B**). After removing low-quality sequences, we obtained 561 M total read pairs.

The analysis of the data was performed by our former colleague and computational researcher Nathan Kenny, at Oxford Brookes University. We generated novel genome annotations for both species. With these, we mapped 96% and 85% of the total reads to *S. mediterranea* and *D. japonica*, respectively. We only selected cells mapping to a minimum of 125 genes, and also discarded cells with >5,000 UMIs to prevent the inclusion of aggregates that may have remained after FACS purification. This rendered **32,431 total cells** in one experiment: 19,025 *S. mediterranea* cells and 13,406 *D. japonica* cells. The annotation of *D. japonica* is less complete and, consequently, fewer cells crossed the minimum gene threshold. We obtained an average of 897 UMI and 240 genes per cell for *S. mediterranea*, and 949 UMI and 210 genes per cell for *D. japonica* (**Figure 2.11 C**). Compared to previous planarian publications (Plass et al., 2018), we have similar UMI counts but a lower number of genes per cell than expected.

At this sequencing depth, we observed the libraries were not yet saturated and could have been further sequenced. **Saturation** can be observed by plotting the number of total UMI and genes per cell detected in subsampled fractions of the total reads. If the curve reaches a plateau, the sample is considered saturated or over-sequenced (**Figure 2.11 D**). On the other hand, both species were effectively separated, with a **collision rate** of only 2.8% (923 cells) (**Figure 2.11 E**). Similar collision percentages have been reported in other *in* situ barcoding-based publications

(Cao et al., 2020, 2019, 2017). Overall, these results show the good **technical quality** of our scRNA-seq data.

9. CELL TYPE COMPOSITION OF TWO PLANARIAN SPECIES

After further analysis, we profiled the whole-body atlases of *S. mediterranea* and *D. japonica.* To assess the **biological quality** of the data, we compared our results with a previous single-cell atlas of *S. mediterranea* (Plass et al., 2018). This reference publication used **trypsin digestion** for tissue dissociation and **Drop-seq** for cell capturing and labelling. To improve comparability, we reanalysed Plass *et al.* data using our new annotation. The re-analysis included 21,610 cells, distributed in 38 clusters (see all clusters in Supplementary 2).

9.1. *SCHMIDTEA MEDITERRANEA*

From our SPLiT-seq data, we classified *S. mediterranea* cells in 41 clusters (**Figure 2.12** and Supplementary 2) and annotated them based on described markers (Fincher et al., 2018; Plass et al., 2018). The four central clusters (in grey) expressed neoblast markers such as *smedwi-1*, while the remaining clusters expressed markers of all previously described progenitors and cell types (Supplementary 3).

In general, this new atlas highly resembles the reference publication, with some exceptions. Our dataset showed a lower resolution of the parenchymal group, where we classified only 3 cell types. However, within them, we identified markers of all 6 parenchymal clusters retrieved in the Plass *et al.* reanalysed data. In other cases, we achieved a better resolution. We identified 6 secretory clusters, three more than in the Plass *et al.* reanalysis. We also resolved two protonephridia identities (**tubule** and **flame cells**, previously identified in Fincher et al., 2018), and extra neuronal clusters (***eye-53+*** and **serotonin neurons**) missing in the Plass *et al*. data. Furthermore, we were able to detect very low abundant cell types, such as the tubule cells (0.4%), epidermal dorsoventral boundary (DVb) (0.5%) and *eye-53+* neurons (0.2%) (Supplementary 2).

Remarkably, we found a novel *nanos*+ and *eIF3c*+ cluster that we called **germ cell progenitors**. These progenitors are well-described in the planarian literature (Wang et al., 2010, 2007). In pure asexual (*S. mediterranea*) or lab asexualized species (*D. japonica*), such as the ones used in this study, the germ cell progenitors are the remains of a truncated germline. However, none of the previous single-cell transcriptomic studies (Fincher et al., 2018; Plass et al., 2018; Zeng et al., 2018) had been able to distinguish these cells from the neoblast populations (Supplementary 4).

Germ cell progenitors are rare (1.6% of total cells), but single-cell methods can detect far rarer populations. Besides, previous publications included more cells and a higher UMI content. For this reason, we suggest the novel detection of the germ cell progenitors in our dataset relies on the early fixation provided by ACME.

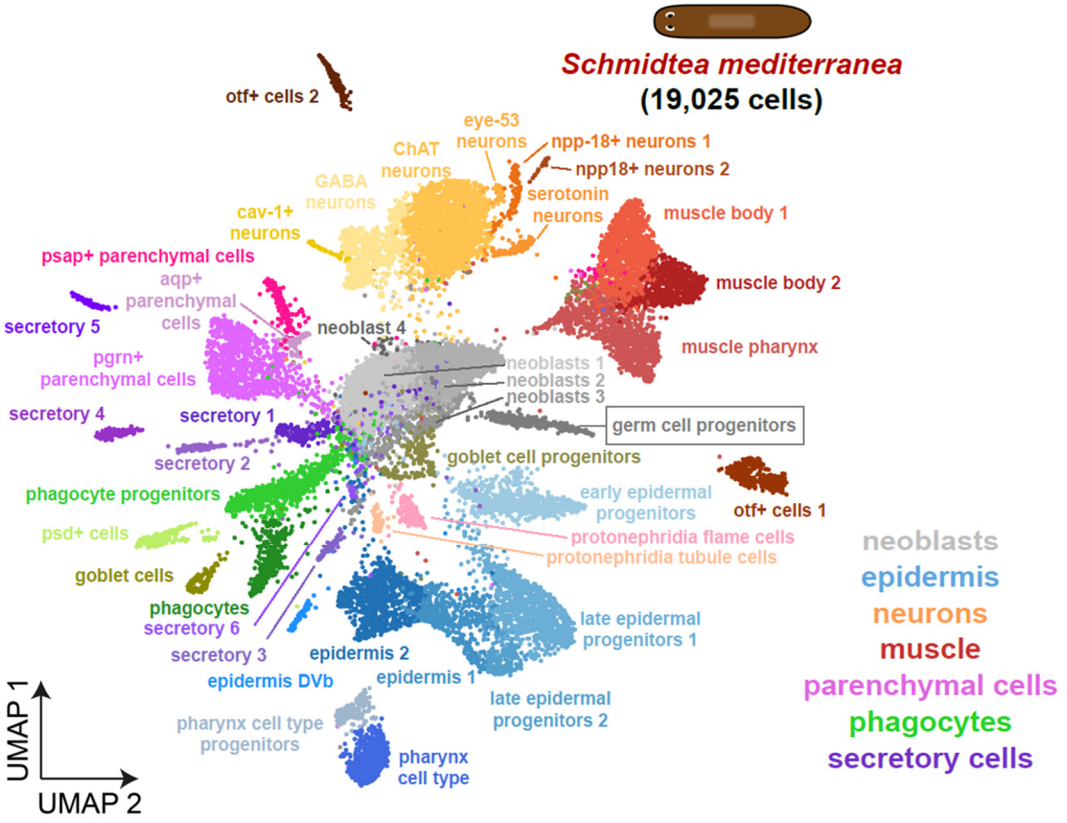


Figure 2.12 UMAP visualisation of the *Schmidtea mediterranea* single-cell atlas generated by ACME and SPLiT-seq. Cells are coloured by cluster and cell group. Major cells groups are shown on the right. Annotations are based on markers genes described in the literature.

9.2. *DUGESIA JAPONICA*

Our *Dugesia japonica* cell atlas was the first whole-body single-cell transcriptomic study performed on the species. Due to the poorer annotation and lower number of cells, we could only identify 28 clusters (**Figure 2.13** and Supplementary 2). These were annotated comparing gene markers to their *S. mediterranea* homologs. Although it is difficult to establish one-to-one homology based on top markers (Shafer, 2019), we confidently detected all major cell types (neoblasts, muscle, neurons, epidermis, parenchymal cells, phagocytes, and secretory cells) following this approach. At this resolution, we were only able to cluster two types of neurons and one type of parenchymal cells. However, we detected low abundance cell types such as epidermis DVb (1.0%) and *psd*+ cells (0.8%). Encouragingly, the **germ cell progenitors** were also identified in this species, representing 1.5% of the total cells (Supplementary 4).

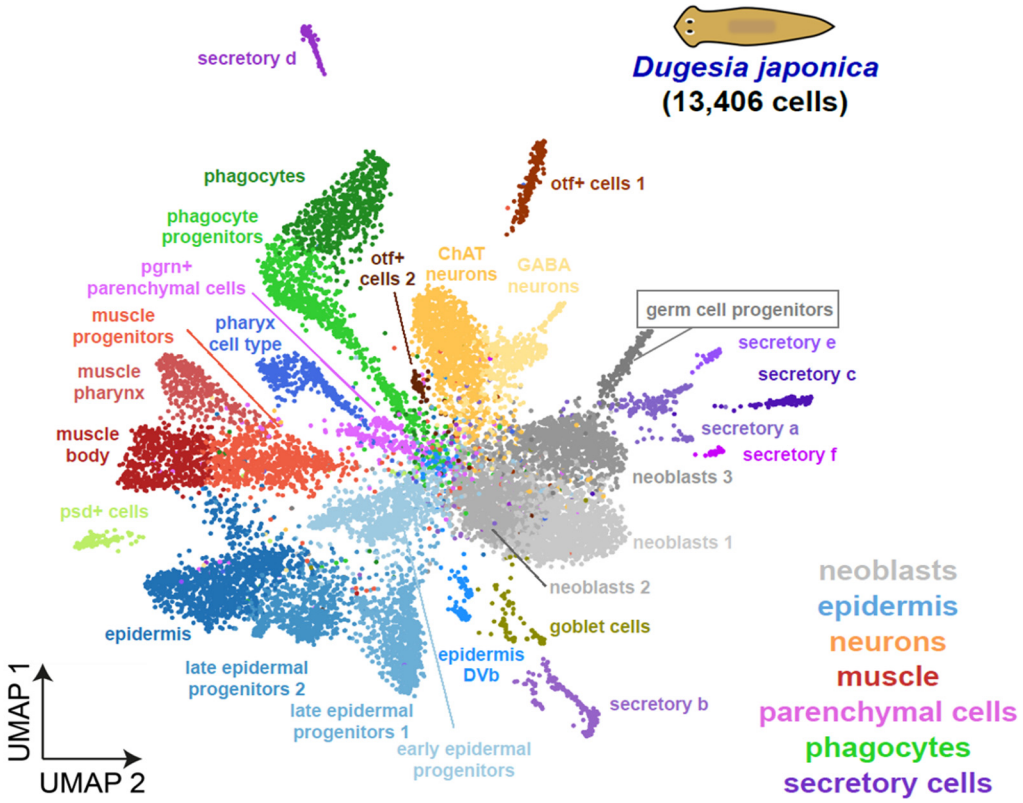


Figure 2.13 UMAP visualisation of the *Dugesia japonica* single-cell atlas generated by ACME and SPLiT-seq. Cells are coloured by cluster and cell group. Major cell groups are shown on the right. Annotations are based on the homology of cluster markers with *S. mediterranea* genes.

9.3. COMPARISON OF CELL POPULATIONS ABUNDANCE

We compared the abundance of different cluster groups (Supplementary 2) in both species with the reference data (**Figure 2.14**). We observed our *S. mediterranea* dataset contains less neoblasts (22.5%) than the Plass *et al.* reanalysis (36%). However, our neoblast percentage is more in line with microscopy estimations obtained by classical maceration (Baguña and Romero, 1981). Other abundant cell types, like epidermis, neurons, muscle, parenchymal and phagocytes, retrieved comparable proportions as those described in Plass *et al.* (Supplementary 2). When we compared both species, we observed most *D. japonica* cell groups were represented in similar proportions to *S. mediterranea*, with the exception of parenchymal cells (2.1% in *D. japonica* vs. 7.1% in *S. mediterranea*). However, it has been shown that the abundance of these cell types varies depending on the size of the animal (Baguña and Romero, 1981).

In general, our results show that ACME can be used to profile all planarian cell types. Using ACME, we were able to spot low abundance cell populations, delicate cell types (e.g. neurons) and harder planarian tissues (e.g. muscle and pharynx) at similar or higher percentages than in

trypsin-based approaches (Supplementary 2). This demonstrates ACME does not introduce biases in cell type composition. On the contrary, we propose that ACME may help to preserve certain cell types, like the secretory cells, protonephridia, neurons and germ cell progenitors.

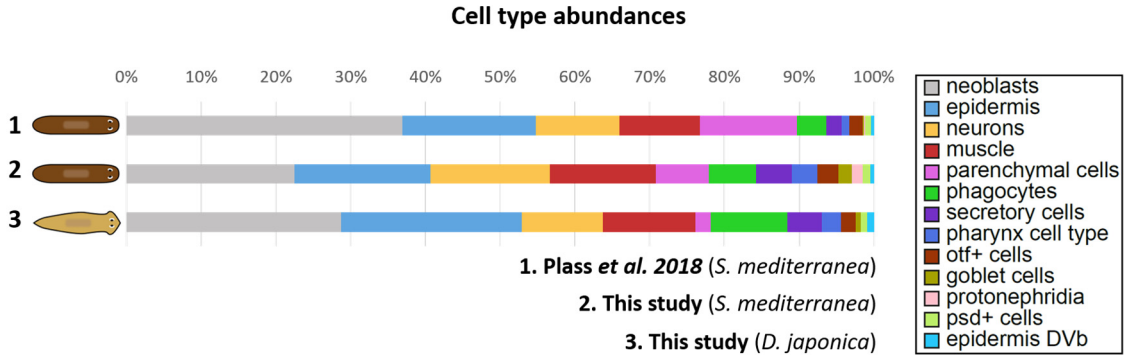


Figure 2.14 Stacked bar plot showing the abundances of different planarian cell populations. The comparison includes *Schmidtea mediterranea* data from this study and from a previous cell type atlas (Plass *et al.* 2018), and *Dugesia* japonica data from this study. Clusters were classified into groups according to the tables provided in Supplementary 2, and coloured by group.

DISCUSSION

In this chapter, I have presented **ACME** as a sample preparation protocol for single-cell transcriptomics and described its advantages over other methods. Also, I have demonstrated the performance of ACME-cells in two different single-cell platforms, **10x Chromium** and **SPLiT-seq**, proving the protocol is compatible with droplet- and *in situ* barcoding-based methods. In the same way, I have validated our initial workflow proposal for single-cell transcriptomics, which combines ACME, cryopreservation, FACS and SPLiT-seq. All the results presented have been peer-reviewed and published (García-Castro et al., 2021).

1. ASSESSMENT OF ACME IN SAMPLE PREPARATION

1.1. CELL DISSOCIATION, FIXATION AND PERMEABILIZATION

ACME relies on the principle of acetic acid dissociation. Similar approaches have been used in the past to dissociate cells for microscopy studies (Baguña and Romero, 1981; David, 1973), but acetic acid protocols had never been applied to modern single-cell transcriptomics before. One of the main advantages of ACME is the **3-in-1 action** of the acetic acid (dissociation and fixation) combined with methanol (fixation and permeabilization). This formula immediately kills the cell and freezes the transcriptomic machinery, preventing the **stress response** associated to traditional enzymatic and mechanical dissociation (Denisenko et al., 2020; Huang et al., 2010; Mammoto et al., 2012; van den Brink et al., 2017). Unlike single-nuclei approaches, ACME preserves the **whole cell**, and retains the cytoplasmic and nuclear RNA. Early fixation also allows the preservation of **cell morphology**.

In general, ACME offers a **faster route** for sample preparation. It avoids additional fixation and permeabilization steps and the use of **toxic reagents** involved in them, such as formaldehyde. Methanol is the only toxic chemical used in ACME, at a concentration of 15%, which is far below the 80-100% used in common methanol fixation protocols (Alles et al., 2017; J. Chen et al., 2018). In fact, most of the ACME formula (over 60%) is water. In most cases, ACME is easy to perform and only requires **common lab equipment** (centrifuge, freezer, and shaker) and **cheap reagents**. Moreover, and according to the literature, ACME reactions could be slowed down up to 24-48 hours in cold conditions (Baguña and Romero, 1981), making it suitable for field sampling.

1.2. CELL IMAGING AND SORTING

After dissociation, ACME-cells can be easily stained using DRAQ5 (nucleus) and ConA (cytoplasm), and visualized by flow cytometry or microscopy. This staining, combined with a simple gating strategy, allows for the distinction of two single-cell populations: **proliferating** (G2) and **non-proliferating cells** (G1). Stained ACME-cells are compatible with **FACS sorting**.

FACS is highly recommended when ACME is used for single-cell transcriptomics, since aggregates and debris account for a large fraction of the sample. The percentage of **singlets** after ACME-dissociation varies between species and protocol performance, but is usually between 20-60% of total events. To obtain cleaner datasets, it is convenient to enrich singlets, by FACS or other means, to 80-95% of total events. The generation of **aggregates** and **debris** is common in all dissociation protocols. For instance, trypsin digestion in planarian generates less aggregates but more debris than ACME. The proportion of singlets, in our experience, is similar for both techniques. We recently improved our ACME protocol by adding extra filtration steps, which has helped us to almost double the percentage of singlets in planarian samples. Surely, the future implementation of different filtration and centrifugation steps, or other strategies, will help to further improve the protocol.

Although it is recommended, FACS enrichment presents some drawbacks. First, FACS requires **large cell inputs**, which is a problem when working with scarce samples. Second, **long sorting times** can affect RNA quality. However, the flexibility of SPLiT-seq has allowed us to overcome these limitations by setting a different sorting strategy. I will present this strategy in the next chapter.

1.3. SAMPLE PRESERVATION

ACME-cells can be easily cryopreserved in 10% DMSO and stored for months. This process can be repeated several times without causing a detrimental effect on cell populations. **Cryopreservation** increases the **flexibility** of the whole pipeline, and allows time-spaced sample collection and processing. This is key to perform single-cell transcriptomic studies at different time points, developmental stages and conditions. Cryopreservation also facilitates field sampling and collaboration between institutions by the exchange of frozen samples. Multiple sample preparation protocols allow cryopreservation, but involve other disadvantages. For instance, the viability of live cells decreases significantly after freezing (Guillaumet-Adkins et al., 2017), and single-nuclei protocols fail to preserve the cytoplasmic RNA.

A general drawback of cryopreservation is **sample loss**. Although ACME-cells are very resistant to freezing temperatures, some cells break during the process. Sample loss also occurs during centrifugation and washing steps linked to thawing. This is by no means a problem when working with medium or highly concentrated samples, but it is worth considering when using low-concentrated samples with very small pellets.

On the other hand, ACME-cells generally present **high quality RNAs**. Unlike other fixatives, such as formaldehyde, the ACME formula does not break the RNA. However, cells become highly sensitive to RNases after dissociation. We have introduced different measures to prevent RNA degradation, like the use of NAC and ultrapure water. RNA in ACME-cells usually resists multiple rounds of cryopreservation and several hours in cold conditions. It also survives FACS sorting, although this process causes a partial degradation of the RNA. Therefore, when working with ACME, it is important to consider the **fragility of the RNA** and avoid contaminations, long sample handling and room temperature conditions.

1.4. PROTOCOL VERSATILITY

The original maceration formula was only used in soft-bodied planarians and cnidarians (Baguña and Romero, 1981; Bradshaw et al., 2015; David, 1973; Thommen et al., 2019). However, we have shown ACME is more **versatile** and, with proper optimization, can be used to dissociate very diverse organisms. In most cases, ACME needs to be complemented by **mechanical forces**. Depending on the animal, these forces may go from gentle pipetting and agitation to stronger homogenization and mechanical disruption. For instance, as in enzymatic dissociation, ACME cannot penetrate hard body parts (e.g. shells and cuticles). Such parts must be mechanically removed to expose the underlying tissues. Mechanical forces are applied during ACME incubation, so samples can still benefit from early fixation and stress protection.

Importantly, all dissociation protocol relies on the use of mechanical forces to a greater or lesser extent. Therefore, this is not a limitation for ACME, but an important optimization factor. Further modifications (e.g. incubation time, staining, filtering steps) could be necessary to adapt ACME to organisms other than those described in this chapter. In any case, our protocol provides a robust and broadly applicable starting point for such optimizations.

Anticipating that ACME may not be suitable for all kinds of tissues, we tested its alternative used as a **fixative** in cells previously dissociated by enzymatic digestion. Resulting ACME-fixed cells were stained and imaged in the same way as ACME-dissociated cells, and retained equal cell

populations. Moreover, ACME outperformed traditional formaldehyde fixation showing a much better preservation of the **RNA quality**.

2. ASSESSMENT OF ACME IN SINGLE-CELL TRANSCRIPTOMICS

We tested the performance of ACME-dissociated cells in **different scRNA-seq platforms** and compared the results with similar published data. First, we used 10x Chromium, a droplet-based technology, to profile 3,899 cells of the cnidarian *Nematostella vectensis.* We identified the same cell types described by a previous single-cell atlas (Sebé-Pedrós et al., 2018b), obtaining similar metrics (UMIs and genes per cell) to this publication. These results show that ACME is compatible with 10x Chromium and, foreseeably, with similar droplet-based protocols such as Drop-seq or InDrop (Klein et al., 2015; Macosko et al., 2015).

Second, we combined ACME with a modified version of SPLiT-seq (Rosenberg et al., 2018). To our knowledge, this was the **first implementation of an *in-situ* barcoding protocol in planarian**. Using this pipeline, we profiled 32,431 cell transcriptomes for *S. mediterranea* (19,025 cells) and *D. japonica* (13,406 cells). In general, our technical metrics were correct and within the expectations. However, our **gene recovery** was lower compared to other planarian single-cell atlases (Fincher et al., 2018; Plass et al., 2018). This can be attributed to several causes, including our higher percentage of intronic regions (55.7% for *S. mediterranea,* compared to 23.7% in Plass *et al.*) or different levels of **ambient RNA**.

Previous planarian atlases were generated by droplet-based methods, which I consider to be more prone to ambient RNAs. When a single-cell is captured in a droplet, free floating RNAs from broken cells can be captured and processed together, resulting in an apparently higher number of expressed genes. On the other hand, *in situ* barcoding protocols process the RNA within the cell, and have multiple rounds of centrifugation that remove free-floating RNAs and make contamination more unlikely. In planarian droplet-based publications, the ratio of UMIs to genes is about 2:1 (Fincher et al., 2018; Plass et al., 2018). In comparison, our SPLiT-seq data has a ratio of 4:1. A later publication using SPLiT-seq in planarians (Benham-Pyle et al., 2021) shows a similar 4:1 ratio, supporting the idea that the lower gene recovery is related to the platform and not to ACME-dissociation.

From a biological perspective, our new ***Schmidtea mediterranea* atlas** is highly comparable to the reference data generated by enzymatic digestion (Plass et al., 2018). Only **neoblasts** are considerably more abundant in Plass *et al.* (36% vs 22.5% in our dataset). My hypothesis is that neoblasts could be more resistant to dissociation stress. In live cell dissociation, neoblasts would

survive better than differentiated populations and, therefore, would be overrepresented in the final count. In addition, we profiled a **novel whole-body atlas of *Dugesia japonica***, opening the study of cell type evolution in this clade. We also identified a novel cluster, the **germ cell progenitors**, in both species. Although well-described in the literature (Wang et al., 2010, 2007), these progenitors had never been clustered separately from the neoblasts in single-cell studies. I suggest the early fixation provided by ACME contributes to the preservation of this cell type and, therefore, to its better resolution.

These results show that ACME is also compatible with SPLiT-seq and, most likely, with other *in situ* barcoding-based protocols, such as sci-RNA-seq (Cao et al., 2017). As ACME-cells can be FACS-sorted, they may also be suitable for plate-based methods like MARS-seq or Smart-seq (Hagemann-Jensen et al., 2020; Keren-Shaul et al., 2019). ACME does not appear to introduce biases in cell type composition, and may even help to preserve certain cell populations. Thus, ACME is a promising alternative for sample preparation in single-cell transcriptomics.

3. ASSESSMENT OF THE INTEGRATION OF ACME SINGLE-CELL DATA

A further analysis of our ACME dataset, performed by Nathan Kenny, revealed that trypsin- and ACME-dissociated experiments can be successfully integrated (García-Castro et al., 2021). This analysis shows ACME single-cell data is highly comparable to data obtained by other means of dissociation, which reinforces the biological results presented in section 9. First, integration demonstrates that most cell types can be identified in both datasets. Some low abundant cell populations (0.2-1.8%) are absent in one or the other, but these differences can be explained by cluster size and resolution parameters. Second, the **germ cell progenitors** identified using ACME are transferred scattered among the neoblast clusters of the trypsin-dissociated dataset. As stated in the previous section, this may suggest a better preservation of this cell type in ACME preparations. Finally, the integration reveals a high overlap in **differentially expressed genes** (74.3-84.3% depending on the cell type) between trypsin- and ACME-dissociated data (García-Castro et al., 2021), suggesting a high comparability not only at cellular but also at gene level.

4. ASSESSMENT OF SPLiT-SEQ FOR SINGLE-CELL TRANSCRIPTOMICS

SPLiT-seq is a relatively new protocol for scRNA-seq based on combinatorial indexing (Rosenberg et al., 2018). In the introduction, I presented the many advantages of this technology, which include high throughput (up to hundreds of thousand cells per experiment), sample multiplexing capacity, cost-effectiveness, easy implementation, and scalability. After working with SPLiT-seq,

I can reaffirm this is a really versatile and powerful method for single-cell transcriptomics. Nonetheless, it is important to comment on some of the strengths and weaknesses of the protocol.

First, the **cell retention rate** (or capture rate for microfluidic devices) of SPLiT-seq is relatively low. During cell barcoding, a high number of cells are lost due to pipetting, centrifugation and plate changes. In our experience, cell retention until library preparation depends on the species and technical performance, and can be as low as 8-12%. Yet, similar percentages can be inferred from other combinatorial barcoding publications (Cao et al., 2019). Although these values are rather hideous in the literature, none single-cell strategy captures 100% of the cells. For instance, 10x Genomics official capture rates are between 35-65%, depending on the reagent kit used (https://kb.10xgenomics.com/hc/en-us/articles/4402368510989-I-have-a-low-starting-number-of-cells-Should-I-process-my-sample-using-the-LT-or-the-standard-Single-Cell-Gene-Expression-kit-). To improve our values, we try to optimize the pipetting technique and use low binding plasticware when possible. With this, we have registered cell retention rates of up to 30-40%. Importantly, SPLiT-seq can easily reach throughputs of more than 10,000 cells even with the lowest cell recovery.

On the other hand, and like other combinatorial indexing protocols, SPLiT-seq profiles **many cells with fewer UMIs and genes per cell**. The information per cell is limited by the large number of cells profiled. Sequencing is shallower (fewer reads per cell) compared to other approaches, as the deep sequencing of many thousand cells would be prohibitively expensive. Our results in planarian show this approach is perfectly valid and highly effective to classify cell types and generate single-cell atlases, as was previously suggested (Cao et al., 2020). Even at low sequencing depth (~17K reads per cell) and with low UMI and gene counts, we identified most previously described cell types in *S. mediterranea*, and profiled the atlas of an uncharacterized species (*D. japonica*).

Finally, to exemplify **cost-effectiveness**, we can compare our SPLiT-seq data with previous droplet-based publications in *S. mediterranea*. All these datasets have a comparable number of cells and retrieve similar biological information: 50,562 cells and 44 clusters (Fincher et al., 2018), 21,613 cells and 38 clusters (Plass et al., 2018), and 32,431 cells and 41 (*S. mediterranea*) clusters (García-Castro et al., 2021). However, droplet-based approaches required numerous drop-seq experiments (25 for Fincher *et al.* and 12 for Plass *et al*.) and sequencing rounds (8 for Fincher *et al.* and 3 for Plass *et al*.) to obtain this information. We only needed a single experiment and a single sequencing run, showing that SPLiT-seq is far cheaper and less time consuming than droplet-based strategies.

After ours, another planarian dataset was published using SPLiT-seq (Benham-Pyle et al., 2021). The authors profiled ~300,000 cell transcriptomes from multiple time points and conditions to study regeneration. This new publication is another example of the great potential of SPLiT-seq for single-cell studies. In principle, SPLiT-seq is designed as a 3'-end counting protocol for the obtention of short reads. However, our preliminary results combining ACME-dissociation, un-tagmented SPLiT-seq libraries and Oxford Nanopore sequencing, suggest we will be able to use SPLiT-seq to characterize **long reads** in the future. Currently, other labs are developing similar strategies to combine SPLiT-seq with long read sequencing (Rebboah et al., 2021).

5. UMI AND GENE COUNTS IN INVERTEBRATES

The number of UMIs and genes per cell not only correlates with the sequencing depth. These metrics are also determined by the model organism, cell size and RNA content. Most single-cell protocols have been developed and optimized using mammalian cells, mainly from mice and humans (Cao et al., 2019; Jaitin et al., 2014; Macosko et al., 2015; Rosenberg et al., 2018; Zheng et al., 2017). Such models can easily yield thousands of UMIs and genes per cell. This sets **unrealistic expectations** for invertebrate studies, which generally have smaller genomes and cells. In the presentation paper of sci-RNA-seq, for instance, the UMI counts are highly variable depending on the model (for similar sequencing depth): 1123 UMIs per cell in *C. elegans*, 8611 in mouse (NIH/3T3) and 9942 in human (HEK293T) (Cao et al., 2017). This is an example of how single-cell metrics are not purely technical, but have a biological significance. In general, expected UMIs and gene counts in invertebrates will be much lower. Thus, when evaluating the quality of an experiment, it is important not to follow the standards set by mammalian datasets and to compare with more proximate model organisms.

CONCLUSION

In this chapter I have presented **ACME** as a dissociation strategy for single-cell transcriptomics. This method was developed to provide an alternative to canonical sample preparation pipelines and their limitations. ACME is a non-enzymatic and low-budget protocol for simultaneous cell fixation and dissociation. Early fixation allows the preservation of the whole-cell and its morphology, and prevents the stress response associated with dissociation. ACME is also a versatile protocol, and works in multiple animal models. After dissociation, ACME-cells can be cryopreserved multiple times, imaged by flow cytometry, and sorted by FACS. Along the process, these cells retain good quality RNAs. ACME-cells are also suitable for **droplet-** (10x Chromium) and ***in situ* barcoding-based** (SPLiT-seq) **scRNA-seq methods**. In particular, the combination of **ACME and SPLiT-seq** offers multiple advantages. ACME allows time-spaced sample collection and is compatible with cryopreservation and FACS sorting. On the other hand, SPLiT-seq can multiplex different samples in a single experiment and profile thousands of cells from a very low budget. Compared to other single-cell transcriptomic approaches, this workflow is simple, economic, and provides high experimental flexibility, opening the door to a broad range of experiments.

CHAPTER III: TISSUE-SPECIFIC EFFECTS OF THE *HNF4* KNOCKDOWN IN PLANARIAN REVEALED BY SINGLE CELL TRANSCRIPTOMICS

INTRODUCTION

1. CELL POPULATIONS DERIVED FROM GAMMA-NEOBLASTS

Gamma-neoblasts are the specialized neoblast subpopulation proposed to give rise to the gut (van Wolfswinkel et al., 2014), as their distinctive transcription factors (*prox-1, hnf4, gata4/5/6, nkx2-2*) are expressed by gut progenitors and differentiated gut cells (except *prox-1*) (Fincher et al., 2018; Forsthoefel et al., 2012; Wagner et al., 2011). The planarian gut consists of a single-opening, highly branched digestive network that extends along the whole body of the worm (**Figure 3.1 A**). It comprises the following cell identities: gut **phagocytes** and secretory **goblet cells**, which form the intestinal epithelium, and **basal cells** and **gut muscle**, which create the outer layer surrounding this epithelium (**Figure 3.1 B-C**) (Barberán et al., 2016a; Forsthoefel et al., 2020; Plass et al., 2018). Basal cells are also called **goblet cell progenitors** by some single-cell publications (García-Castro et al., 2021; Plass et al., 2018).

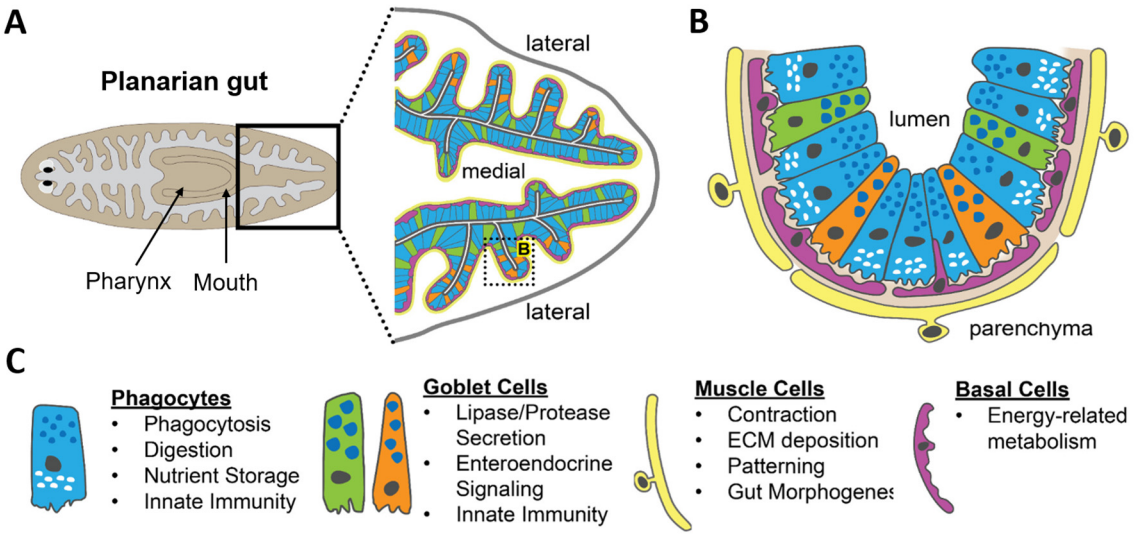


Figure 3.1 The planarian gut. A) Planarian digestive network and its main regions **B)** Close-up scheme of the gut **C)** Different gut cell types and their putative functions. *Adapted from Forsthoefel et al., 2020.*

One of the major transcription factors of gamma-neoblasts, ***hnf4***, is not only expressed in the gut but also in the **parenchymal cells** (**Figure 3.2**), suggesting that gamma-neoblasts could give rise to both populations (Fincher et al., 2018). Then, after early differentiation, parenchymal and gut fates would diverge in two separated branches with specific progenitors (**Figure 3.3 A**) (Plass et al., 2018). In addition, an **ATAC-seq** dataset recently generated by our lab (not presented), shows *hnf4* motifs are enriched in the promoters of parenchymal cells. These findings indicate *hnf4* may be involved in the regulation of the parenchymal populations.

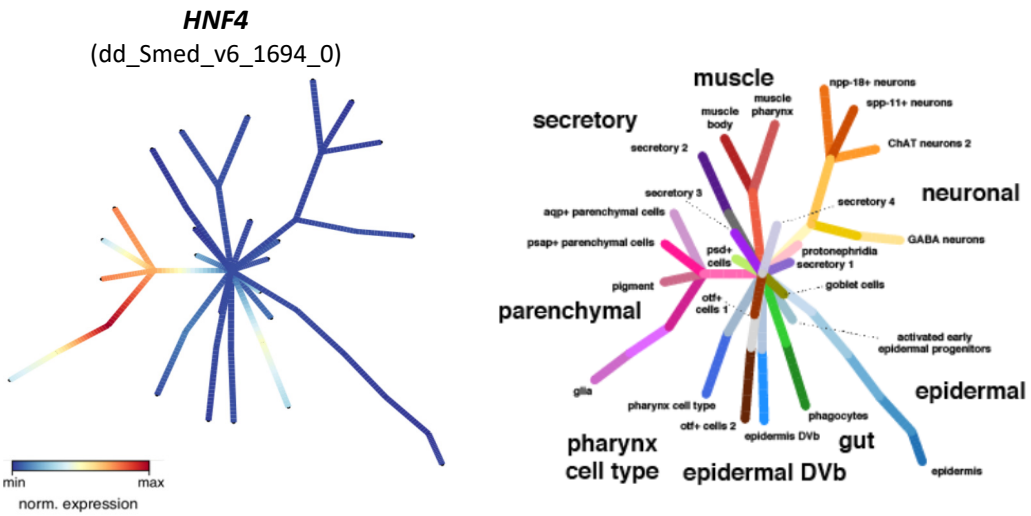


Figure 3.2 *Hnf4* expression in *Schmidtea mediterranea*. Planarian *hnf4* is strongly expressed in the gut and parenchymal branches. *From PlanMine: https://planmine.mpinat.mpg.de/planmine/report.do?id=9032914&trail=|9032914. Data extracted from Plass et al., 2018.*

2. THE PARENCHYMAL LINEAGE

Parenchymal cell populations are located in the connective tissue, which in planarians is known as the **parenchyma**. Some parenchymal populations were characterized by morphology in old microscopy literature, where they are referred as **'fixed' parenchymal cells** to distinguish them from mobile neoblasts (Baguña and Romero, 1981; Pedersen, 1961). However, parenchymal cells have been largely overlooked during the molecular era.

Single-cell transcriptomic studies have re-emerged the parenchymal as of the major cell fates in planarian and revealed new identities within this group (Fincher et al., 2018; García-Castro et al., 2021; Plass et al., 2018). Importantly, some scRNA-seq publications refer to parenchymal cells as **cathepsin+ cells**, and use the term 'parenchymal' to describe the secretory cells instead (Fincher et al., 2018). According to lineage reconstruction studies based on single-cell data, parenchymal cells can be classified into: **parenchymal progenitors**, which give rise to all other identities, ***aqp*+ cells** (aquaporin positive), ***psap*+ cells** (prosaposin positive), ***pgrn*+ cells** (progranulin positive), ***ldlrr-1*+ cells** (low density lipoprotein receptor-related 1 positive), **glia** and **pigment cells** (**Figure 3.3 B**) (Plass et al., 2018).

'**Fixed' parenchymal cells** have been classically described as highly heterogeneous. They are rich in lysosomes, hydrolytic enzymes and vacuoles, and share some metabolic functions with the gut. In addition, these cells are connected to the neoblasts through gap junctions. As planarians lack a circulatory system to transport nutrients, these populations are thought to intervene in the distribution and storage of gut metabolites, and in the transport of excretory products to

the protonephridia. Parenchymal cell populations, and specially *aqp+* cells, are depleted in starvation and regeneration conditions. On the contrary, they increase in bigger and well-fed animals, indicating their use as metabolic reservoirs in planarians (González-Estévez et al., 2007; Pedersen, 1961; Plass et al., 2018; Romero and Baguñà, 1991).

Surprisingly, lineage reconstruction analyses, using PAGA and Velocyto, have shown that pigment cells and glia are also part of the parenchymal group (**Figure 3.3 B**) (Plass et al., 2018). **Pigment cells** have a dendritic shape and are responsible for colouring planarians. In their absence, worms exhibit a white phenotype. In vertebrates, pigment cells have an ectodermal origin and arise from the neural crest (He et al., 2017). On the other hand, planarian **glial cells** are similar in shape to mammalian astrocytes, and one of their proposed functions is to metabolize the excess of neurotransmitters. Glial cells are found entwined with neurons throughout the planarian nervous system (Roberts-Galbraith et al., 2016). Like pigment cells, glia arises from the ectoderm in vertebrates (Paridaen and Huttner, 2014). ***Pgrn+*** and ***ldlrr-1+* cells** are part of the same differentiation branch than the glia (**Figure 3.3 B**). Of these, only ***ldlrr-1+* cells** have been briefly described in the literature as a whole-body expressed population predicted to be involved in wound-response regeneration (Roberts-Galbraith et al., 2016).

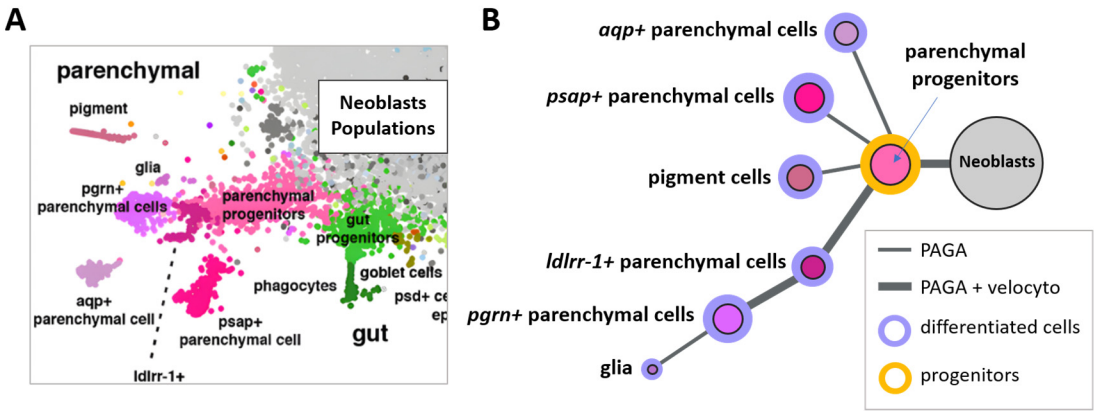


Figure 3.3 The parenchymal lineage. A) Detail of the UMAP plot of *S. mediterranea* showing all gut and parenchymal populations. **B)** Lineage reconstruction of parenchymal cells in *S. mediterranea* using PAGA and Velocyto. *Adapted from Plass et al., 2018.*

In general, the parenchymal lineage remains poorly understood. Further research is required to characterize its different cell types and understand their functions. Moreover, the apparent endodermal origin of planarian glia and pigment cells suggests a disparate evolution that has yet to be elucidated. To deepen into the differentiation and regulation of the parenchymal lineage, as well as into the connection between the parenchymal and gut cells, we decided to knockdown ***hnf4,*** a transcription factors commonly expressed by gamma-neoblasts, gut and parenchymal populations. Our goal was to describe the effects of the ***hnf4*** knockdown at tissue resolution,

and gain new insights into the cell populations that express this gene. For this, we combined RNA interference with our single-cell transcriptomic workflow, based on ACME and SPLiT-seq, previously described in Chapter II.

3. HEPATOCYTE NUCLEAR FACTOR 4 (HNF4)

The **Hepatocyte Nuclear Factor 4 (HNF4)** is an evolutionary conserved **transcription factor** mainly expressed by endodermal tissues (gut, liver and pancreas) and some mesodermal organs (kidney, urinary bladder and reproductive system). HNF4 is a nuclear receptor protein that shares family with other liver-enriched genes, like *HNF1* (*POU*), *HNF3* (*FOXA*) and *HNF6* (*ONECUT*) (Dubois et al., 2020; Lau et al., 2018; Ryffel, 2001). HNF4 was first described in the late 80s (Costa and Grayson, 1989) and, since then, has been extensively studied for its multiple roles in development, cell differentiation and gene regulation across metazoan (Barry and Thummel, 2016; Holewa et al., 1997). In mammals, HNF4 plays a key role in gastrulation, when its deletion results in embryonic lethality (Chen et al., 1994). In later development and adulthood, HNF4 is required for terminal differentiation and homeostasis of the hepatocytes, enterocytes, and renal proximal tubule cells (Cattin et al., 2009; L. Chen et al., 2019; Li et al., 2000; Marable et al., 2018). The loss or malfunction of HNF4 is also linked to the development of MODY diabetes (Urakami, 2019), as HNF4 intervenes in the regulation of glucose metabolism in the pancreatic β-cells (Wang et al., 2000).

A planarian homologous to human HNF4 (*Smed-hnf4*) has been identified in *S. mediterranea* (Wagner et al., 2011). As seen in previous sections, planarian *hnf4* is one of the main markers of **gamma-neoblasts**, and is also enriched in progenitors, proliferating, and differentiated **gut** and **parenchymal** populations (**Figure 3.4 A**) (Fincher et al., 2018; van Wolfswinkel et al., 2014; Wagner et al., 2011). Despite *hnf4* being strongly expressed by multiple planarian tissues, only a few RNAi experiments have been performed on this gene. The first *in vivo* validation of *Smed-hnf4* by RNAi resulted in a wild-type regeneration of the knockdown animals (Lobo et al., 2016). A second RNAi experiment showed *hnf4* plays a role in the regeneration of **pigment cells** (**Figure 3.4 B**). However, this study was primarily focused on pigment cells and repigmentation, not offering any further comments on the effects of the *hnf4* knockdown or its phenotype (He et al., 2017).

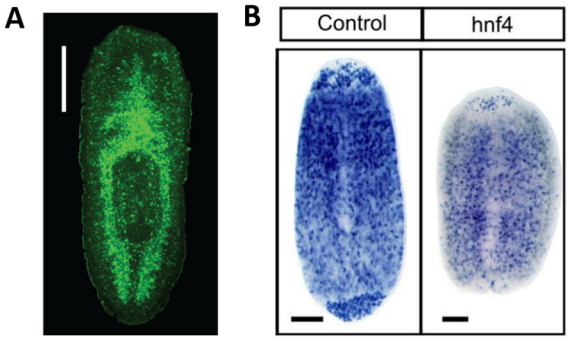


Figure 3.4 *Smed-hnf4* expression pattern and effect on repigmentation. A) FISH showing the wild-type expression of *Smed-hnf4* in the gut and the parenchymal space. Scale bar: 200 µm. *Adapted from Fincher et al., 2018* **B)** *In situ* hybridization of regenerating pigment cells (PBGD-1+), on day 7 after depigmentation, in control and *hnf4* RNAi planarians. The *hnf4* knockdown affects both regeneration and repigmentation. Scale bars: 250 µm. *Adapted from He et al., 2017.*

4. RNA INTERFERENCE

RNA interference (**RNAi**) is a commonly used technique to attenuate gene expression by introducing double-stranded RNAs (dsRNAs), homologous to a target mRNA, into the cell. This activates the RNA-induced silencing complex (RISC), which eventually mistakes target mRNAs as exogenous, cleaving them. It creates **knockdown** organisms with a significantly reduced expression of the target gene (Hannon, 2002). RNAi has been widely used in planarians to describe the role of uncharacterized genes (Scimone et al., 2017; Yamamoto and Agata, 2011). The technique can be performed by feeding (Rouhana et al., 2013) or microinjection (Sánchez Alvarado and Newmark, 1999) of the dsRNAs. Traditional ways of evaluating RNAi phenotypes consist of cutting the worms to monitor abnormal regeneration patterns, or performing *in situ* hybridizations. RNAi can also be integrated with **RNA-seq** (Solana et al., 2012) and **scRNA-seq** (Plass et al., 2018) to compare transcriptomics and cell differentiation between knock-down and wild type conditions. However, in-depth analyses of RNAi single-cell data are still scarce.

RESULTS

1. PHENOTYPIC CHARACTERIZATION OF THE *HNF4* KNOCKDOWN

To prepare our RNAi samples, we injected *Schmidtea mediterranea* worms with dsRNA from ***hnf4*** or control **GFP**. We generated two replicates that were both biological and technical, as they were injected and processed separately by different researchers. We included 25 animals per replicate and condition. Each worm was treated with 0.3 µg of dsRNA, injected in 3 consecutive days (0.1 µg/day). As we aimed to study the early effects of the *hnf4* knockdown, we decided to process the samples **9 days post-injection** (counting from the last injection). We harvested 20 worms per replica and condition and dissociated them in **ACME** as described in Chapter II, with some modifications (see Chapter V: Methods). These ACME-cells were stored at -80ºC to run a SPLiT-seq experiment. The remaining animals were kept uncut and monitored until day 15 post-injection (**Figure 3.5**).

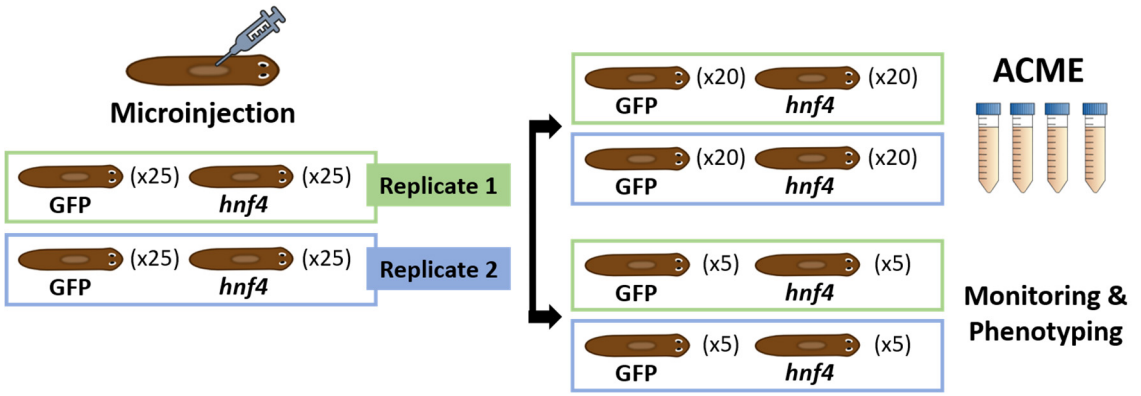


Figure 3.5 Scheme of the RNAi experiment. Worms were injected with GFP or *hnf4* dsRNA, generating two replicates. For each replicate and condition, 20 animals were used for ACME dissociation and 5 animals were kept for monitoring and phenotyping.

During monitoring, we characterized the **phenotype** of the *hnf4* knockdown and classified its development in 6 different levels of damage (**Figure 3.6**). From an initial healthy state, worms begin to show small depigmented necrotic patches. Depigmentation frequently appears in the **pre-pharyngeal region**, but can also affect other body parts. Gradually, the pre-pharynx accumulates depigmentation and tissue damage (Head damage I-III), until the head is cleaved from the body or simply disintegrates. In addition, a few worms can show a reduced mobility of the posterior body or swollenness. Surprisingly, the head can retain its mobility until a very advanced state (Head damage III). The resulting headless tails do not regenerate, and eventually die and degrade completely.

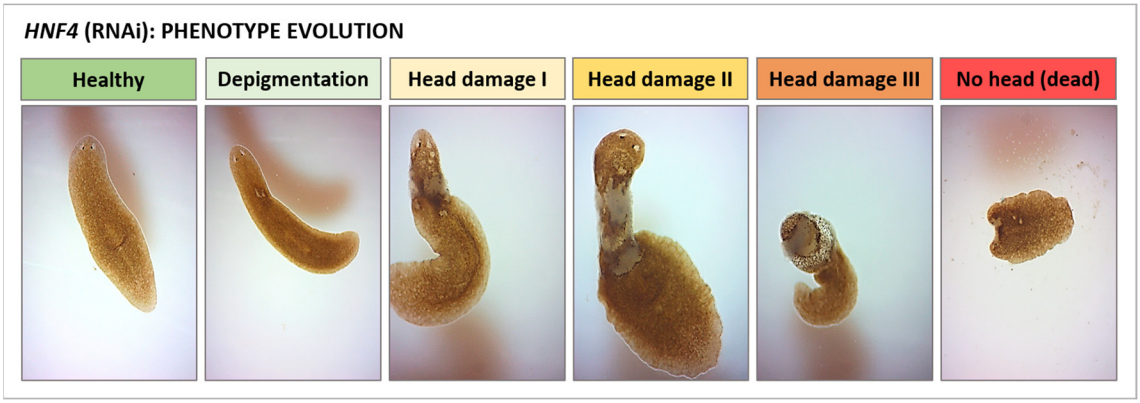


Figure 3.6 Evolution of the *hnf4* knockdown phenotype. Depigmentation and necrosis start in the pre-pharyngeal region, and extend towards the head until it is cleaved from the body. Headless animals do not regenerate and die.

Nine days post-injection, all animals treated with GFP had healthy phenotypes (**Table 3.1**). On the other hand, most worms treated with *hnf4* were healthy, but 9 out of 50 (18%) already showed one of the altered phenotypes described in **Figure 3.6**. From them, 3 out of 50 (6%) had no head. At this point, five healthy-phenotype worms were selected from each sample and replicate, and monitored from day 12 to 15. During this time, 100% of GFP worms remained healthy. On day 12, 60% of all *hnf4*-treated animals showed head damage or headless phenotypes. The remaining 40% were healthy or slightly depigmented.

Table 3.1 Phenotype monitoring of RNAi samples. All injected worms were evaluated before ACME dissociation (9 days post-injection). After sample processing, 5 whole animals from each condition were further evaluated from day 12 to 15 post-injection, and classified into 6 different groups according to their phenotype: healthy, depigmentation, head damage I, head damage II, head damage III and no head.

***HNF4* (RNAi): PHENOTYPE MONITORING**

PROCESSING TIMEPOINT (25 worms per condition)								
Sample	Days post-injection	Replicate	Healthy	Depigmentation	Head damage I	Head damage II	Head damage III	No head (dead)
GFP	9 days	1	25	0	0	0	0	0
		2	25	0	0	0	0	0
hnf4	9 days	1	19	2	2	0	0	2
		2	22	2	0	0	0	1
MONITORING (5 worms per condition)								
Sample	Days post-injection	Replicate	Healthy	Depigmentation	Head damage I	Head damage II	Head damage III	No head
GFP	12 days	1	5	0	0	0	0	0
		2	5	0	0	0	0	0
	13 days	1	5	0	0	0	0	0
		2	5	0	0	0	0	0
	14 days	1	5	0	0	0	0	0
		2	5	0	0	0	0	0
	15 days	1	5	0	0	0	0	0
		2	5	0	0	0	0	0
hnf4	12 days	1	2	0	0	0	0	3
		2	1	1	0	0	3	0
	13 days	1	1	0	1	0	0	3
		2	1	0	0	1	0	3
	14 days	1	1	0	1	0	0	3
		2	0	0	1	0	1	3
	15 days	1	0	0	1	0	1	3
		2	0	0	0	1	1	3

By day 15, every animal treated with *hnf4* had at least some degree of head damage, showing a penetrance of 100% in our *hnf4* knockdowns. These results demonstrate that *hnf4* RNAi samples have a very strong and characteristic phenotype, which mostly affects the pre-pharyngeal body and ultimately leads to head loss and death. Moreover, we were able to characterize the evolution of this phenotype in two separate replicates.

2. INTEGRATION OF RNAi AND SINGLE CELL TRANSCRIPTOMICS

We processed our RNAi samples using **SPLiT-seq** (Rosenberg et al., 2018). As shown in previous chapters, SPLiT-seq is a scRNA-seq platform with multi-sampling capacity and high experimental flexibility. Here, we took advantage of these features to introduce some improvements to the protocol described in Chapter II (see Chapter V: Methods). First, **FACS sorting** was performed in the middle of the SPLiT-seq protocol, right before the cell lysis (**Figure 3.7 A**). This greatly reduces the cell inputs required for the experiment, and the sorting time and cost. It also provides cleaner datasets, due to the later removal of aggregates and debris, and helps to prevent RNA degradation, as samples are sorted after reverse transcription. Second, we multiplexed different treatments and replicates in the same experiment, avoiding **batch effects**. We ran together two replicates of *hnf4* RNAi and two replicates of GFP RNAi. For this, we divided the first barcoding plate in 4, using 12 wells per sample (**Figure 3.7 B**). Although cells were pooled in subsequent barcoding rounds, the first barcode was used to track each sample separately.

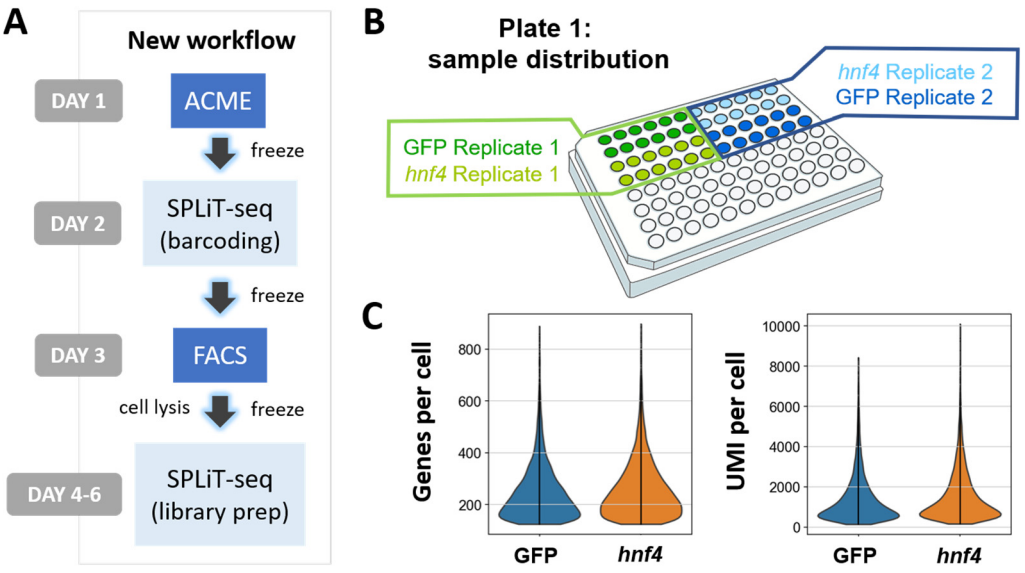


Figure 3.7 SPLiT-seq experiment configuration and technical values. A) New experimental workflow **B)** Distribution of samples and replicates in barcoding plate 1 **C)** Violin plots of genes per cell and UMI per cell in GFP and *hnf4* RNAi cells.

We introduced other **technical improvements** based on parallel optimization experiments (not shown). Briefly, we increased the concentration of Round 1 barcodes, RT enzyme and T4 ligase and reduced the number of cells loaded per well to improve the number of UMIs per cell. Besides, the tagmentation protocol was modified to work with low-concentration libraries (see Chapter V: Methods). As input for SPLiT-seq, we used **unsorted ACME-dissociated cells** after one cryopreservation step (**Figure 3.7 A**). During optimization experiments, we obtained the best performance loading 4000 cells/well. Thus, we decided to start this experiment with **5000 events/well** (240,000 total events). We counted total events instead of cells because unsorted samples also have aggregates and debris. These events occupy space, and consume part of the reagents and, therefore, need to be considered in the counting.

In this new SPLiT-seq configuration, cells were subjected to three rounds of barcoding and then pooled and frozen. Subsequently, cells were thawed and sorted by FACS using the same gating strategy described in Chapter II. Cells were sorted directly into lysis buffer. In total, we collected two libraries of 17,000 and 18,500 cells. After cell lysis, libraries were frozen again. Further library preparation was performed during the following days (**Figure 3.7 A**). For sequencing, we applied the 150 bp paired-end strategy used in Chapter II.

After data preprocessing, we profiled **24,145 total cells** of *Schmidtea mediterranea*. From them, we classified **11,271 GFP (RNAi) cells** and **12,874 *hnf4* (RNAi) cells**, with 16,489 cells from Replicate 1 and 7,656 cells from Replicate 2. On average, we obtained **1386 UMI** and **247 genes** per cell for GFP samples and **1529 UMI** and **259 genes** per cell for *hnf4* samples (**Figure 3.7 C**), with a minimum threshold of 100 genes per cell.

3. CLUSTERS ANNOTATION AND QUANTITATIVE ANALYSIS

For cluster annotation, we use an internal **reference dataset** of *Schmidtea mediterranea* already annotated, with 103,654 cells from different experiments (not shown). Labels -or cell identities- were transferred from this reference to our RNAi dataset using the Scanpy function **Ingest**. All 50 identities in the reference dataset were initially transferred, including the germ cell progenitors described in Chapter II. However, to make this annotation functional for further analyses, transferred labels had to be combined with the Leiden clustering of the RNAi dataset. Thus, each cluster was assigned with the identity of its majoritarian Ingest label. This retained 40 of the 50 identities originally transferred (**Figure 3.8** and Supplementary 5).

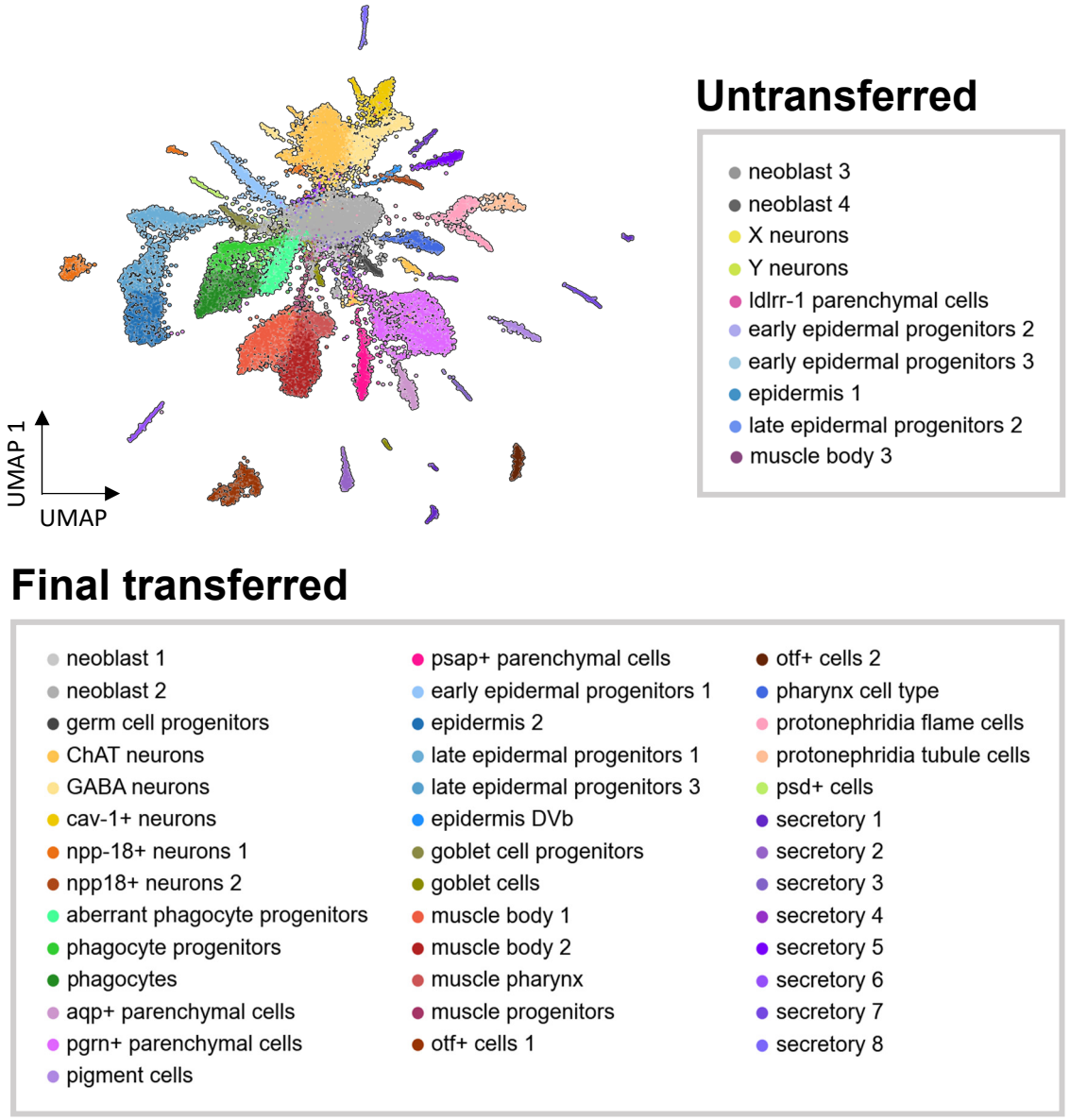


Figure 3.8 Cluster annotation. UMAP visualisation of the RNAi (GFP and *hnf4*) single-cell atlas of *Schmidtea mediterranea*. Cells are coloured by cluster identity. Untransferred identities are shown on the right legend, and final transferred identities on the bottom legend. Annotations are based on transferred labels from a reference dataset.

Parenchymal identities were successfully annotated except for **glia cells**, as they were not present in the reference dataset, and ***ldlrr-1*+ parenchymal cells**. We manually searched for these cell types looking at the expression of specific markers. The glia marker *IF-1* (Intermediate Filament 1) (Roberts-Galbraith et al., 2016) was expressed in a small set of cells that clustered together with the ***pgrn*+ parenchymal cells**, even at higher clustering resolutions. Similarly, *ldlrr-1*+ cells, marked by *ldlrr-1* (Plass et al., 2018), were widely distributed within the ***pgrn*+** and **pigment cells** populations (**Figure 3.9 A**). On the other hand, we annotated a novel phagocyte cluster with the name of **aberrant phagocyte progenitors**. This cluster was predominantly found in the ***hnf4* RNAi** knockdown samples, where normal phagocyte progenitors were mostly absent (**Figure 3.9 B**).

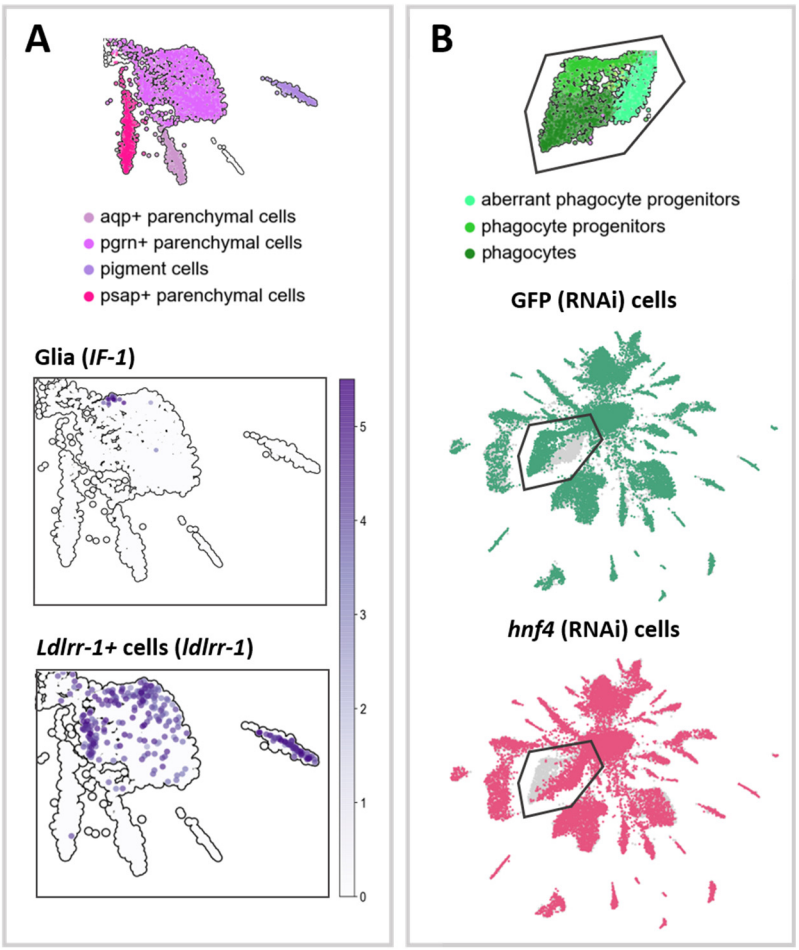


Figure 3.9 The parenchymal and phagocyte identities. A) From top to bottom: UMAP detail of all annotated parenchymal identities, and feature plots showing the expression of *IF-1* (glia marker) and *ldlrr-1* (*ldlrr-1*+ parenchymal cells marker). **B)** From top to bottom: UMAP detail of all annotated phagocyte identities, UMAP visualisation of GFP (RNAi) samples showing a predominance of normal phagocytes and phagocyte progenitors, and UMAP visualisation of *hnf4* (RNAi) samples showing the predominance of aberrant phagocyte progenitors.

After cluster annotation, we compared the **expression of *hnf4*** between control and knockdown samples. For this, we calculated the *hnf4* counts per million (CPM) and per cluster in both conditions (**Figure 3.10 A** and Supplementary 6). Surprisingly, we found a higher number of *hnf4* counts in the knockdown samples than in the controls (324 vs 206 counts, respectively). This does not necessarily mean the knockdown was ineffective, as we will see in the discussion. The expression in CPM matches our expectations, as all gut and parenchymal clusters show a high signal for *hnf4.* However, we can observe this calculation is biased by cluster size. In clusters with fewer cells, the expression of *hnf4* in only one cell can translate into an elevated number of CPM. Due to this, we see an apparent high expression in some small clusters like the epidermis DVb or the *npp-18*+ neurons 1, which do not express the gene (**Figure 3.10 A**). This could be caused by the presence of doublets or the wrong annotation of these cells.

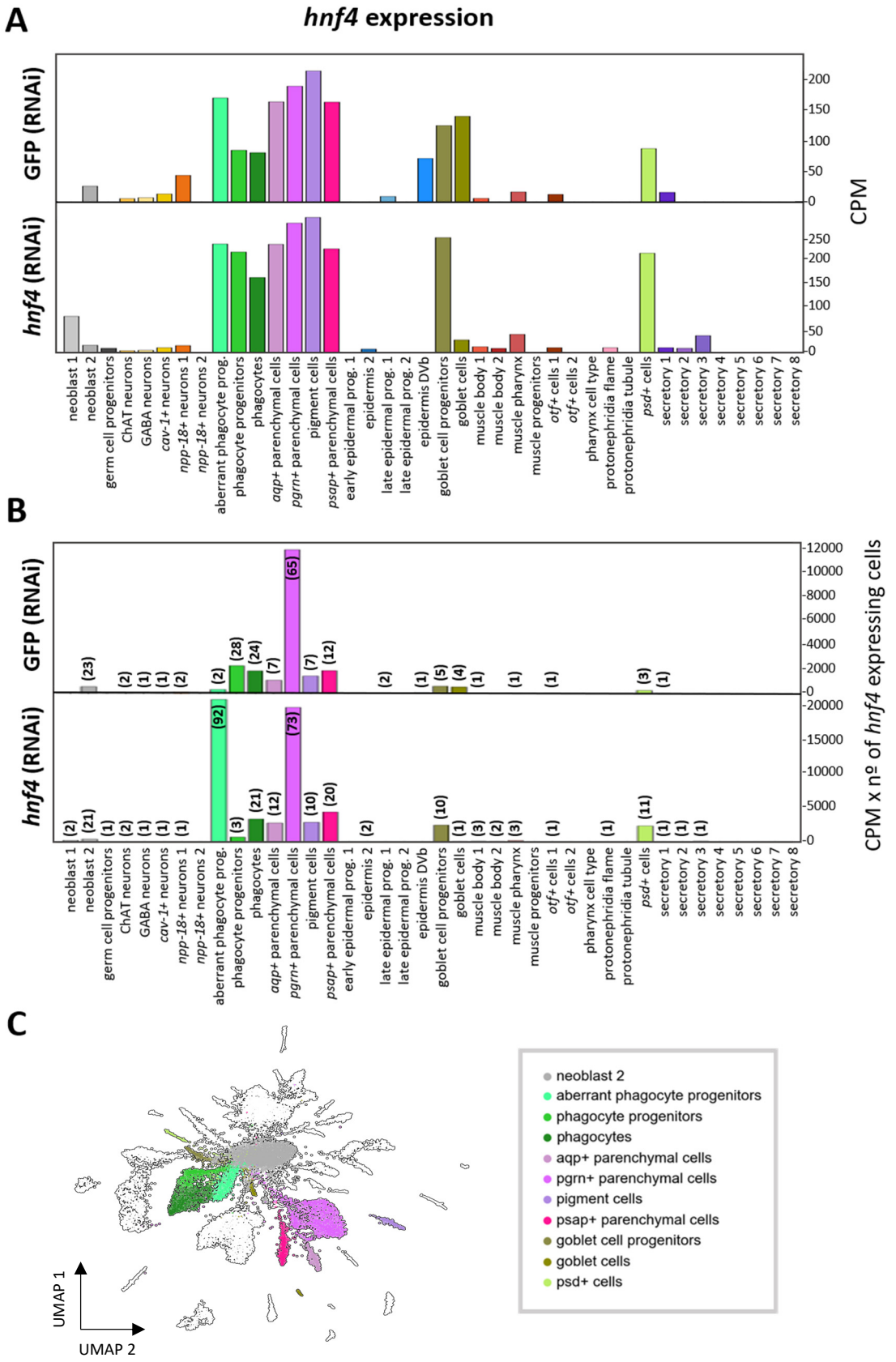


Figure 3.10 Expression of *hnf4* in different cell types. A) Bar graph of *hnf4* counts per million and per cluster in GPF (RNAi) and *hnf4* (RNAi) samples. **B)** Bar graph of *hnf4* counts per million and per cluster multiplied by the number of cells per cluster expressing *hnf4*. This number of cells is shown in parenthesis above each bar. **C)** UMAP plot of *hnf4* expression (coloured clusters) in the whole dataset (control and knockdown samples).

To get a clearer picture of *hnf4* expression, we multiplied the CPM by the number of cells per cluster expressing *hnf4* (**Figure 3.10 B** and Supplementary 6). Thus, the signal was enriched only when multiple cells of the cluster had counts for the gene. After this correction, higher expression got restricted to clusters that truly express *hnf4*, including gut (**phagocytes** and **goblet cells**), **parenchymal** cell types, **neoblast 2**, and ***psd*+ cells** (pleckstrin and sec7 domain-containing positive cells) (**Figure 3.10 B-C**). As predicted, the signal observed in other clusters was due to the presence of a few *hnf4*-expressing cells. When we compare the expression pattern between samples, the most noticeable differences are in the phagocyte lineage (**Figure 3.10 B**). GFP RNAi samples express *hnf4* in phagocyte progenitors and phagocytes. Meanwhile, *hnf4* RNAi samples barely express *hnf4* in normal phagocyte progenitors, and concentrate most of the expression in the aberrant phagocyte progenitors.

To quantify the effect of the *hnf4* knockdown in different cell types, we obtained the number of cells per cluster (cluster abundances) and ran a statistical analysis comparing these numbers in GFP and *hnf4* RNAi samples. We ran separate **Fisher tests**, with 95% confidence, for each cluster and replicate (**Table 3.2**). Significant differences in cluster abundances were classified as upregulations (cluster enriched) or downregulations (cluster depleted) in the *hnf4* knockdowns compared to the controls (**Figure 3.11**).

Table 3.2 Example of contingency tables used for Fisher tests. Phagocyte cluster contingency tables for replicate 1 (**A**) and 2 (**B**).

A

	Phagocytes	Other clusters
GFP (RNAi) Replicate 1	412	7250
hnf4 (RNAi) Replicate 1	180	8647

B

	Phagocytes	Other clusters
GFP (RNAi) Replicate 2	169	3440
hnf4 (RNAi) Replicate 2	96	3951

We found more upregulated clusters in Replicate 1 (16,489 cells) than in Replicate 2 (7,656 cells). In **Replicate 1**, *hnf4* RNAi samples were statistically more abundant in aberrant phagocyte progenitors, neoblast 2, secretory 1, epidermis 2, and multiple epidermal progenitors (**Figure 3.11 A**). In **Replicate 2**, only the aberrant phagocyte progenitors, late epidermal progenitors 3 and muscle body 1 clusters were upregulated (**Figure 3.11 B**). On the other hand, Replicate 1 and 2 had the same clusters downregulated: phagocyte progenitors, phagocytes and *pgrn*+ parenchymal cells.

Only clusters differentially expressed in both replicates were considered truly significant (**Figure 3.11 C**). Thus, the **aberrant phagocyte progenitors** and **late epidermal progenitors 3** were significantly upregulated, while normal **phagocyte progenitors**, differentiated **phagocytes** and

***pgrn+* parenchymal cells** were significantly depleted. By far, the highest level of significance was found in normal and aberrant phagocyte progenitors (**Figure 3.11 A-B,** volcano plots).

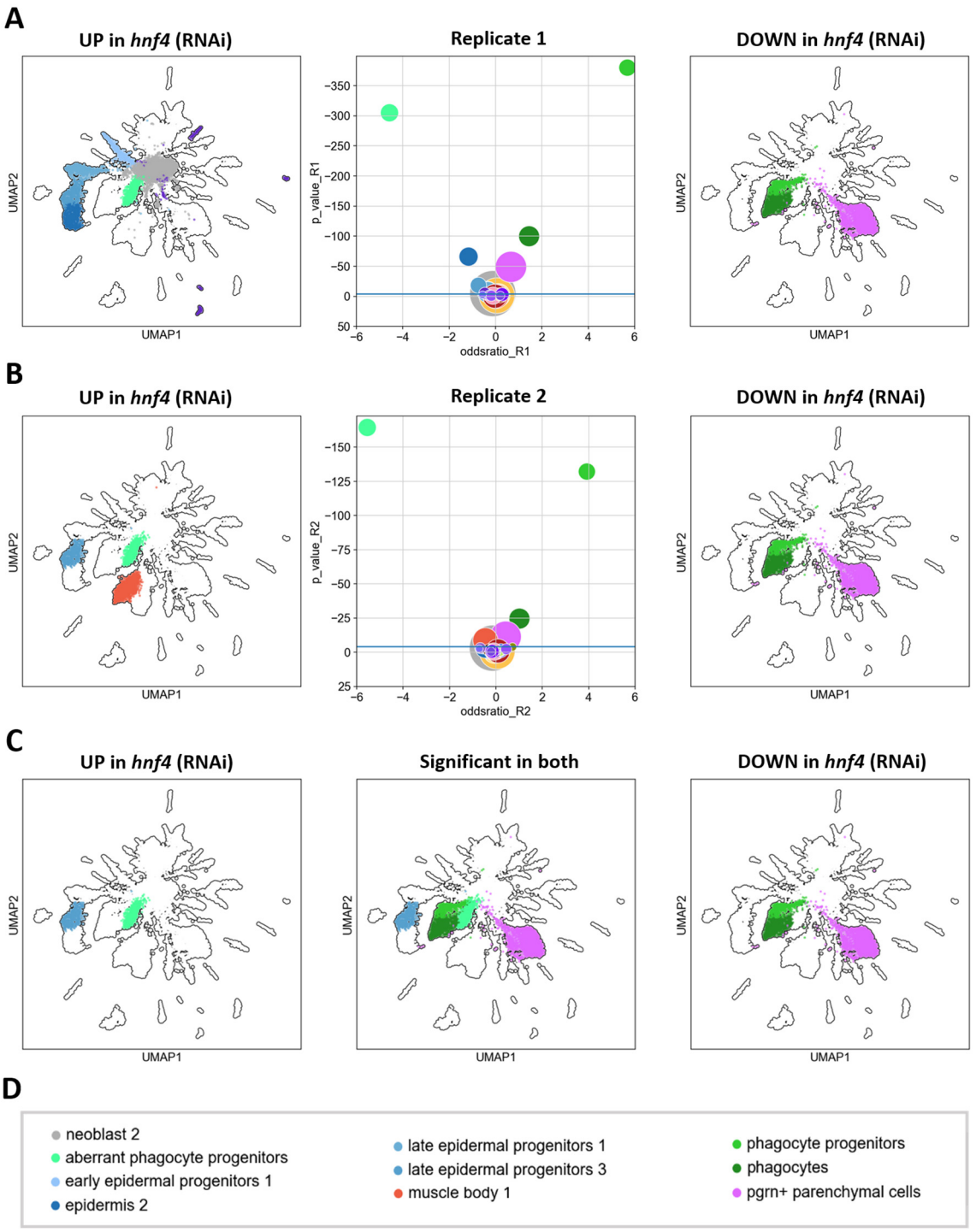


Figure 3.11 Analysis of cluster abundances. A-B) From left to right: UMAP plot of upregulated clusters in *hnf4* RNAi samples, volcano plot of p-values vs odd ratios per cluster (obtained by Fisher test), and UMAP plot of downregulated clusters in *hnf4* RNAi samples. Results for Replicate 1 (**A**) and Replicate 2 (**B**). **C)** UMAP plots showing clusters significantly upregulated (left), downregulated (right), or both (middle), in Replicates 1 and 2. **D)** General legend of cluster identities.

4. DIFFERENTIAL GENE EXPRESSION ANALYSIS

Finally, we performed a **differential gene expression** analysis between GFP and *hnf4* RNAi samples. To simplify the analysis, we classified all annotated clusters within their main cell groups: epidermis, germ cell progenitors, goblet, muscle, neoblasts, neurons, *otf*+, parenchymal, phagocytes, pharynx, protonephridia, *psd*+ and secretory (Supplementary 5). From now on, we will refer to these broad groups instead of individual clusters. We used the pseudo bulk data to obtain the counts per gene and per cell group, for each condition, and ran a **DEseq2** statistical analysis (Love et al., 2014) to spot significant differences in gene expression. For this analysis, the log2 fold change cutoff was set at ±1.0 and the unadjusted p-value at 0.05.

We identified **120 differentially expressed genes (DEGs)** (Supplementary 7), most of them in the **parenchymal** (4 DEGs: 1 upregulated and 3 downregulated) and **phagocyte** groups (111 DEGs: 28 upregulated and 83 downregulated) (**Figure 3.12**). A single DEG (SmMSTRG.20455) was found downregulated in the epidermis, muscle and neoblasts groups, while two DEGs (SmMSTRG.20455 and SmMSTRG.11706) were identified downregulated in the goblet lineage. Other groups showed no significant differences in gene expression. All DEGs found had a log2 fold change above ±2.0.

We used an internal **diamond blast** annotation of *S. mediterranea* (not shown) to identify or match these DEGs to their homologs (Supplementary 7) and unravel potential targets of *hnf4* in planarian. In the **parenchymal group**, we identified two downregulated genes (SMEST056326001 and SMEST056331001) as **Slc2a-9**. In humans, SLC2A9 -also known as GLUT-9- is a transmembrane sugar transporter that plays a role in glucose homeostasis (https://www.ncbi. nlm.nih.gov/gene/56606). In planarians, the homologous protein has been reported in pigment and parenchymal cells, gut, muscle, and the cephalic ganglia (Fincher et al., 2018; He et al., 2017).

Moving on to the **phagocyte group**, we found many interesting homologs. One gene in particular, **SmMSTRG.20455** (apolipophorin-like homology), stands out for being strongly downregulated in phagocytes and other cell groups (epidermis, muscle, neoblasts and goblet). **Apolipophorins** are invertebrates equivalent to apolipoproteins, lipid-binding proteins involved in lipid transport. In rats, HNF4 was shown to enhance the activity of the apolipoprotein A1 promoter (Chan et al., 1993), while in humans, several apolipoproteins have been identified as targets of HNF4 (Shih et al., 2000).

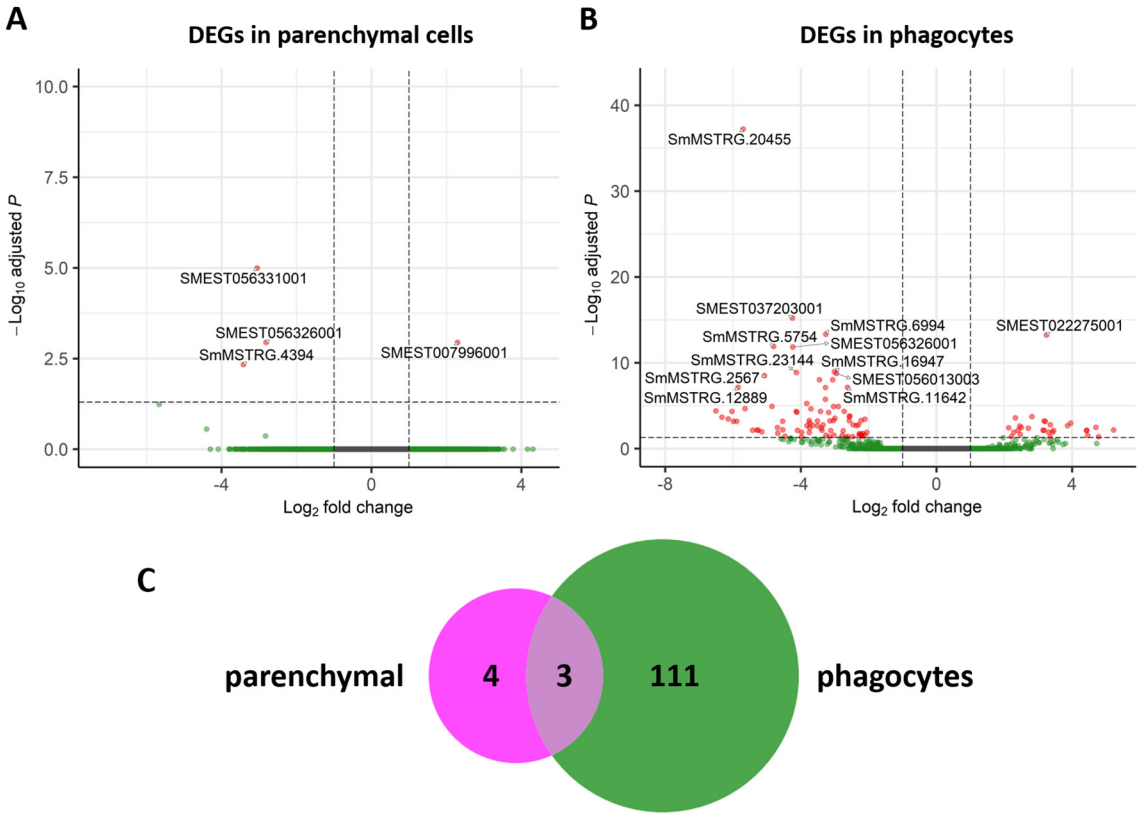


Figure 3.12 Differential gene expression analysis. A-B) Volcano plots of differential gene expression in the parenchymal (**A**) and phagocyte (**B**) groups. Genes are coloured as: non-significant (grey), non-significant but above the Log2 fold change threshold (green), or significant by Log2 fold change and p-value (red). The names of most highly significant genes are indicated. Thresholds are shown as dashed lines **C)** Venn diagram of differentially expressed genes in parenchymal cells and phagocytes.

As in the parenchymal group, three homologs of **Slc2a-9** (SMEST056331001, SMEST056326001 and SmMSTRG.16883) were found downregulated in phagocytes. Many other **solute carriers (SLCs)**, which constitute a diverse group of membrane transport proteins, were also downregulated: **Slc6a-24** (SMEST031125001), a sodium-neurotransmitter symporter; **Slc8a-2** (SMEST076607001), a sodium-calcium exchanger; **Slc16a-10** (SmMSTRG.19215), a MFS-transporter; **Slc38a-5** (SmMSTRG.9743), an amino acid transporter; and **Slc47a-4** (SmMSTRG.2567), **Slc47a-6** (SMEST045553001 and SMEST045551001) and **Slc47a-7** (SmMSTRG.9253), which are metabolic and xenobiotic cationic antiporters (https://planosphere.stowers.org/). In planarians, these SLC proteins are mainly expressed by phagocytes. Different studies carried out in mammals have reported HNF4 upregulates the expression of several SLC transporters by enhancing their promoters (Chan et al., 1993; Gallegos et al., 2012; Prestin et al., 2014; Song et al., 2020; Tümer et al., 2013).

Searching into **Harmonizome**, an online repository of datasets, we found several apolipoproteins and SLC families as predicted targets of HNF4 (based on the analysis of binding site motifs) (https://maayanlab.cloud/Harmonizome/gene_set/HNF4/MotifMap+Predicted+Transcription+Factor+Targets). In a similar way, we found many of our DEGs shared homology with HNF4 target genes predicted by Harmonizome datasets. In phagocytes, the following predicted targets were found upregulated: **acid phosphatase** (SmMSTRG.1012), **actin** (SMEST012332001), **ELKS/Rab6-interacting/CAST family** (SmMSTRG.13151) and **pleckstrin domains** (SMEST058735001). And the following ones downregulated: **alpha-2 macroglobulin** (SMEST037144001), **cytochrome P450** (SmMSTRG.18127), **dynein 1 heavy chain** (SMEST064358001), **laminin alpha** (SMEST056013003), **cAMP-specific 3',5'-cyclic phosphodiesterase** (SmMSTRG.17073), **metabotropic glutamate receptor** (SMEST064049001), **prosaponin** (SmMSTRG.6994), **semaphorin** (SMEST069640002 and SMEST016303003), **Rho-GAP domain** (SmMSTRG.11383), **TGF-β-induced** (SmMSTRG.11642) and **tyrosine-protein phosphatase** (SMEST055800001).

Genes affected by the *hnf4* knockdown cover a wide range of biological functions. As we have seen, *hnf4* has an impact on glucose and lipid transport, but is also involved in processes as diverse as general catabolism, immunity, or cell proliferation. One interesting example is **PIM-1** (SMEST000602001), a proto-oncogene downregulated in phagocytes. This gene promotes cell proliferation and has been linked to HNF4 in humans (Vuong et al., 2015).

Overall, DEseq2 has revealed that phagocytes concentrate most of the differential gene expression, with only a few specific genes affected in other cell groups. Unfortunately, many of our DEGs could not be assigned to any homolog, or were matched with undescribed hypothetical proteins (Supplementary 7). However, we have found two well documented examples of genes upregulated by human HNF4 that were downregulated in our knockdown samples: an apolipophorin and several solute carrier transporters. These results indicate the *hnf4* transcriptional programme in planarians could be very similar to that of mammalian.

DISCUSSION

Previous publications have linked the expression of *hnf4* in planarians to gamma-neoblasts (gut progenitors), differentiated gut cells (van Wolfswinkel et al., 2014; Wagner et al., 2011) and parenchymal cells (Fincher et al., 2018). Moreover, single-cell ATAC-seq studies carried out in our lab have found that *hnf4* motifs are enriched in both parenchymal and gut cells (not shown). Thus, parenchymal cells appear to be linked to **endodermal tissues** by the expression of this transcription factor. This poses a challenge for evolutionary biology, since the parenchymal lineage in planarian contains cell types as **pigment cells** or **glia** (Plass et al., 2018), which are typically associated to ectodermal tissues (He et al., 2017). To learn more about the relationship and regulation of the gut and parenchymal lineages, we studied the *hnf4* knockdown at single-cell resolution. In this chapter, I have described its effects at **phenotypic**, **cluster** and **gene level**.

1. PHENOTYPIC CHARACTERIZATION OF THE *HNF4* KNOCKDOWN

There are few studies describing *hnf4* knockdowns in planarians. The first *in vivo* validation of *Smed-hnf4* by RNAi resulted in a wild-type morphology regeneration of the trunks, amputated right after 3 days of microinjections (Lobo et al., 2016). On a different publication, the *hnf4* knockdown was achieved by feeding planarian on bacterial cultures expressing the gene construct. Feeding was performed on days 5 and 8 during a depigmentation experiment. In this study, the authors focused mainly on pigment cells. Therefore, although the *hnf4* knockdown was reported to be detrimental for pigment cells regeneration, no further comments were mentioned on the phenotype (He et al., 2017).

Taking these previous experiments as reference, it was surprising to see such strong phenotypes in our knockdown animals. According to our observations, the *hnf4* RNAi phenotype starts with the necrotic depigmentation of the **pre-pharyngeal area**, and progresses accumulating **tissue damage in the head**, that eventually splits off from the body. The headless animal does not regenerate and finally dies (**Figure 3.6**). From the first symptoms, the phenotype evolves in a matter of 2-4 days. For our RNAi experiments, we treated planarians for 3 consecutive days following a standard **microinjection** protocol, and collected the samples 9 days post-treatment. Some injected animals were kept and monitored from day 12 to 15, without amputation.

Our stronger phenotype, compared to other publications, may be explained by the longer **experimental time**, since even after 9 days post-treatment most of our worms still looked

healthy. Another reason could be that the monitoring was performed on non-amputated animals, and not under regeneration conditions as in previous studies.

Additionally, we have performed parallel experiments **feeding** the animals with *hnf4* dsRNA (not shown). Although these experiments were not properly monitored, after 5 feedings we could observe a similar progression, leading to the headless phenotype previously described. Overall, the recurrence of the same phenotype in our lab experiments, carried out by different people and using different techniques, makes me feel confident about the robustness of this novel phenotypic characterization.

2. TECHNICAL IMPROVEMENTS IMPLEMENTED IN SPLiT-SEQ

For our scRNA-seq experiment, I followed a similar pipeline to the one described in Chapter II but implementing some technical improvements. The most relevant one is the use of **FACS in the middle of the SPLiT-seq protocol**. In SPLiT-seq, the full cell is used as a reaction chamber during barcoding. Thus, the cell integrity is maintained on this first part of the protocol (Rosenberg et al., 2018). This is unique to *in situ* barcoding-based technologies and, as demonstrated in this chapter, it makes it possible to sort cells after labelling them.

In this new configuration, the SPLiT-seq plates are loaded with unsorted cells. Therefore, the barcoding occurs on unclean samples (singlets, aggregates and debris) that are later purified by FACS before the cell lysis. This pipeline has many advantages. First, it **reduces the initial cell input** (ACME samples required per experiment) from eight to 1-2 tubes. Second, as many cells are naturally lost during the barcoding process, the final number of cells to be sorted is much lower, resulting in **less time and money** spent in the FACS facility. Third, it provides **cleaner datasets**, because it sorts out all the debris generated during the barcoding. Finally, it **prevents RNA degradation**, as cells are sorted after reverse transcription.

Sorting cDNA-containing cells has been a key improvement to the protocol. For other species studied in the lab, like *Parasteatoda tepidariorum* (spider) or *Pristina leidyi* (annelid), we encountered serious difficulties to run the original pipeline, in which FACS was performed before SPLiT-seq. The little amount of RNA per cell in these species, and its fragility, made it impossible to keep working on our envisioned projects. In these cases, sorting RNA-containing cells for long hours resulted in complete RNA degradation. Running FACS after barcoding allowed us to perform SPLiT-seq with minimal processing time after tissue dissociation, protecting the RNA. This greatly broadened the **range of species and experimental designs** (e.g. mixing several conditions) within the scope of our pipeline.

3. CLUSTER ANALYSIS

After clustering, I was able to annotate most of our target populations, except for glia and *ldlrr-1+* parenchymal cells. However, both cell types were found contained within their closest related cluster, the *pgrn+* parenchymal cells (Plass et al., 2018) (**Figure 3.9 A**). At the cluster level, the strongest and most visual effect of the *hnf4* knockdown was the **depletion of the phagocyte progenitors** and the emergence of a new parallel population, that we called **aberrant phagocyte progenitors** (**Figure 3.9 B**). In control samples, phagocyte progenitors differentiate into phagocytes. Meanwhile, knockdown samples follow an alternative differentiation pathway in which aberrant phagocyte progenitors emerge from the neoblasts. These aberrant progenitors appear to get stuck or struggle to differentiate into phagocytes, as the proportion of the latest also decreases in the knockdown.

When I evaluated the **expression of *hnf4*,** most gene counts were found in the expected clusters: phagocytes, goblet cells, parenchymal cells and some neoblast subpopulations (**Figure 3.10**). In addition, I found ***psd+* cells** also expressed *hnf4* counts. The role of *psd+* cells is still unknown, but they are located around the pharynx area and, therefore, may be physiologically related with the digestive system of planarians.

In *hnf4* knockdown samples, most clusters showed an *hnf4* expression similar to the controls. Except in phagocyte progenitors, where the expression was attenuated, and in aberrant phagocyte progenitors, where it was highly increased. But overall, the number of *hnf4* counts in the knockdowns was surprisingly high (**Figure 3.10 B** and Supplementary 6). This can have multiple explanations. The expression levels could be coming back to normal after 9 days post injection, or we could be capturing dsRNAs from the injections. Another possibility is that, because *hnf4* is a key transcription factor, cells may try to overcompensate the degradation of these mRNAs with the upregulation of the gene expression. However, it is intriguing that the most divergent cluster, the aberrant phagocyte progenitors, concentrates such a high number of *hnf4* counts. A more in-depth study of these aberrant phagocytes, by *in situ* hybridization, microscopy, or other methods, would be required to understand more about their nature.

Finally, I ran a statistical analysis based on the **Fisher test** to evaluate the differences in **cluster abundances** between *hnf4* and GFP RNAi samples. Only results that were significant in our two replicates were considered as valid. The following clusters were **downregulated** in the *hnf4* knockdowns of both replicates: phagocyte progenitors, differentiated phagocytes and *pgrn+* parenchymal cells (**Figure 3.11**). **Upregulated** clusters, on the other hand, differed between replicates. The aberrant phagocyte progenitors, which were clearly enriched in the *hnf4*

knockdowns, were significantly upregulated in Replicate 1 and 2. However, the late epidermal progenitors 3 were also upregulated in both replicates, even though this cell type does not express *hnf4.* Some muscle, secretory and epidermal clusters were also found upregulated in one of the two replicates (**Figure 3.11**).

In general, it is difficult to evaluate compositional changes in single-cell transcriptomics using traditional statistical approaches. Because when some clusters are depleted, others have to increase their relative frequencies to compensate, leading to false positives (Büttner et al., 2021). I suggest the upregulation of certain muscle and epidermal clusters, like the late epidermal progenitors 3, correspond to **false positives** caused by the rearrangement of cluster abundances after the depletion of phagocytes and parenchymal groups. In principle, these artefacts would only affect upregulated clusters, while downregulated clusters would be truly significant. In the future, I will explore novel strategies like sccODA (Büttner et al., 2021), a Bayesian model for single-cell compositional data analysis, to avoid this unspecific effect.

4. DIFFERENTIAL GENE EXPRESSION ANALYSIS

In planarians, gene interactions with *hnf4* remain poorly understood. To unravel some of these interactions, I performed a differential gene expression analysis at tissue resolution to detect genes affected by the *hnf4* knockdown. The analysis revealed most differentially expressed genes were concentrated in the **parenchymal** and **phagocyte** populations, which coincides with the previous cluster analysis. Other tissues showed none or unclear results.

For instance, a single gene (SmMSTRG.20455) was downregulated in epidermis, muscle and neoblasts. This could be a real case of downregulation, but could also be attributed to the presence of **doublets**. The expression of SmMSTRG.20455 is much higher in phagocytes than in other tissues (**Figure 3.13**). Therefore, capturing one singlet of any cell type together with a phagocyte could be enough to significantly increase the gene counts in the first group. The number of doublets in our datasets is usually low, and **doublet discrimination** is not routinely performed. However, it would be interesting to use a doublet detection software during data pre-processing to resolve this ambiguity.

In the case of **goblet cells**, two different genes appear downregulated: SmMSTRG.20455, mentioned above, and SmMSTRG.11706. Again, both genes are much more expressed in phagocytes and could be seen as a consequence of doublets (**Figure 3.13**). However, goblet cells are part of the gut and express *hnf4*, suggesting these downregulations could be real.

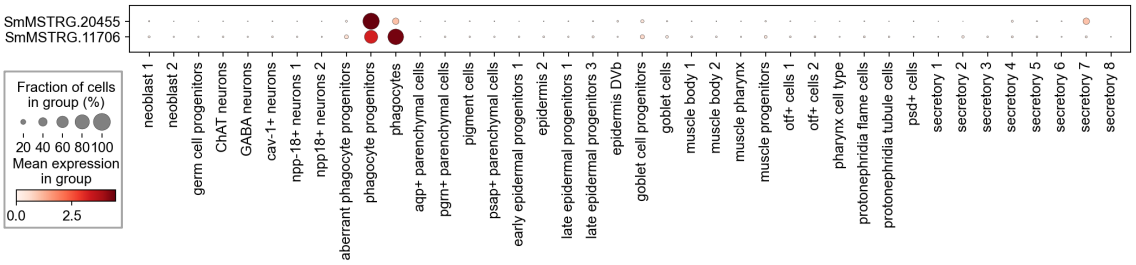


Figure 3.13 Dotplot showing the fraction of cells expressing different genes and their mean expression per cluster. The plot shows SmMSTRG.20455 and SmMSTRG.11706, genes downregulated in the goblet, epidermis, muscle and/or neoblasts populations.

In the **parenchymal group**, three genes were significantly downregulated and one upregulated. Two of the downregulated genes were homologs to a sugar transporter, supporting that parenchymal cells are involved in nutrient distribution (Baguña and Romero, 1981; González-Estévez et al., 2007; Pedersen, 1961). Since *pgrn*+ parenchymal cells are depleted at cellular level, I expected more genes to be differentially expressed in this population. Finally, in **phagocytes**, 111 DEGs were found down- or upregulated. According to their homologs, these genes cover a **wide range of biological functions**, including glucose and lipid transport, general membrane transport, metabolism, cell proliferation or immunity. The functional diversity of HNF4 target genes had been described in humans (Bolotin et al., 2010), but not in planarians. Besides, although I found various homologs of our DEGs were predicted targets of HNF4, the experimental literature linking them was scarce.

In summary, this differential gene expression analysis provides an extensive list of **putative target genes of *hnf4*** in planarians. Some of them have clear homologs, but many others do not. Therefore, the role of these genes and the regulation mechanisms by which they interact with *hnf4* are still unknown and open to further studies.

5. GENERAL DISCUSSION AND FUTURE PERSPECTIVES

In this chapter, I aimed to characterize the effects of the *hnf4* knockdown in planarians. When worms were dissociated at day 9 post-injection, only 9/50 individuals presented a non-wild-type phenotype, and 3/50 were headless. This was a good indicator that the knockdown was effective, but likely on its **early stage**. At this point, the **strong effects** of the knockdown include the disruption of phagocyte differentiation, the almost complete depletion of phagocyte progenitors, and the emergence of an alternative pathway, characterized by aberrant phagocyte progenitors. In addition, the phagocyte populations experience a high number of significant

transcriptional changes. As **moderate effects**, we find the depletion of the *pgrn*+ parenchymal cells**,** as well as some gene expression changes in the parenchymal group. Given the results presented, I can neither confirm nor rule out a moderate effect on goblet cells. Either way, it is clear that different **gut cells** are unequally affected by *hnf4*. While we observe dramatic changes in phagocytes, the effect on goblet cells, if real, is very mild.

As *hnf4* is a transcription factor of gamma-neoblasts (van Wolfswinkel et al., 2014; Wagner et al., 2011), I expected to see significant changes in some **neoblast** subpopulations. But the gamma-neoblasts are the only ones expressing *hnf4*, so the effect may be diluted amidst the whole neoblast population. This could be resolved by sub-clustering the neoblasts, running the same statistical analysis on the sub-clustered populations, and observing whether any differences emerge.

At the beginning of the study, I also expected to find specific changes in **pigment cells**, as a previous publication had shown they were somehow affected by the lack of *hnf4* (He et al., 2017). However, our results are difficult to compare, because the authors introduced additional experimental conditions such as light treatments and regeneration. Furthermore, the effect described in this paper is partial, not quantified and based on the observation of a single marker, so it is difficult to evaluate what level of depletion is to be expected.

My phenotypic observations pointed out depigmentation as one of the early symptoms of the knockdown. Yet, this depigmentation could be caused by tissue necrosis rather than by a specific depletion of pigment cells, as no changes were detected at cluster or gene level. If there is any defect on pigment cells, it does not appear at this experimental time point, or it is very mild to be detected in such small population by our quantitative methods.

It is important to remark that our experiments were carried out in **non-amputated animals**, where changes occur in the context of normal **cell turnover**. In regeneration conditions all tissues are missing, and neoblasts have to fast proliferate and differentiate into every cell type. On the contrary, during normal homeostasis, the neoblasts replace damage cells at a slower pace. Although the general self-renewal ratio in planarians is quite high, some cell populations could certainly be replaced faster than others. In this sense, pigment cells could be more stable and long-lasting than other cell types (e.g. cell progenitors), and thus not being so rapidly affected by the *hnf4* knockdown.

On the other hand, my findings at gene level suggest the **evolutionary conservation of *hnf4* target genes** between planarians and mammals. The similarities with the mammalian HNF4 transcriptional program provide evidence of planarians suitability as model organisms for gut

differentiation, metabolism and metabolic diseases studies. However, not all HNF4 mechanisms are conserved. For instance, mouse HNF4 is expressed in nephrons and required for renal differentiation and homeostasis (Marable et al., 2018), while planarian *hnf4* so does not appear to be involved in the regulation of the **protonephridia**.

Taking all these results into account, I propose the following model for the *hnf4* knockdown: In planarians, *hnf4* would be necessary for the specialization of gamma-neoblasts, but also for the differentiation of the phagocyte and parenchymal branches. The lack of *hnf4* would affect all progenitor populations expressing the gene but, in homeostasis conditions (non-amputated animals), the effect would be more pronounced in tissues with a **higher turnover**. My proposal is that, due to their location and role in digestion, phagocytes have a higher renewal ratio compared to goblet or parenchymal cells and, therefore, are the most affected population. The **aberrant phagocyte progenitors** would be the result of normal phagocyte progenitors getting stuck during differentiation. In consequence, the population of terminal phagocytes would decrease (as we see it happens). Since the biggest parenchymal population (*pgrn+* cells) is also depleted in the knockdown, we can expect something similar is happening with the **parenchymal progenitors** (here clustered together with the *pgrn+* cells), but at a slower pace. Terminal cell identities of the parenchymal lineage, like pigment cells, would be the last ones to be affected, or could only be visibly affected under **regeneration conditions**. If animals survived longer than 15 days, I would expect to see more changes in the parenchymal linage, and possibly on the goblet and *psd+* cells. But it is also likely that the absence of *hnf4* causes the death of the animal before these changes can be perceived in non-regeneration conditions.

To study if different cell types are affected at earlier or later stages, or under regeneration, I propose a future single-cell experiment using samples at **different time points**. This configuration should include animals from day 7 to 14 post-injection, to observe the full evolution of the knockdown, and ideally have 3 replicates per condition. To complement this single-cell data, the evolution of the phenotype should be assessed using ***in-situ* hybridization** to compare controls and *hnf4* knockdowns at different times points. The *in-situ* hybridizations should target *hnf4* and other markers of parenchymal and gut populations, to evaluate changes in these tissues. *In situ*s should be performed on both, non-amputated and regenerating animals.

Finally, one of the goals of this chapter was to improve our understanding of **parenchymal cells**. But the impact of the *hnf4* knockdown on the parenchymal lineage was milder than expected, and there are still many questions around the origin, function and regulation of these populations. Further studies in planarians, and other model organisms, are needed to shed light

on the evolution of controversial cell types, like glia. In the same way, further experiments are required to unravel the role of all parenchymal subpopulations. The results presented here have shown how **single-cell RNAi** can be used to characterize the tissue specific effects of knockdowns at much higher resolution than classical protocols. Thus, I believe the pipeline proposed in this chapter will be a powerful tool for future RNAi studies related with the parenchymal or other cell populations.

CONCLUSION

In this chapter, I have unravelled the effects of the ***hnf4* knockdown** in planarians, at tissue resolution, using single-cell transcriptomics. For this, **RNA interference** by microinjection was combined with our scRNA-seq pipeline (**ACME & SPLiT-seq**) to multiplex different samples and replicates in the same experiment, avoiding batch effects. Following this strategy, the *hnf4* knockdown was described and quantified at the **phenotypic**, **cellular** and **gene level**. First, we found the knockdown leads to a headless phenotype, causing the death of the animal. Second, we discovered *hnf4* does not equally affect all **gut cell types**. *Hnf4* plays a crucial role on **phagocyte regulation and differentiation**, as progenitors and differentiated phagocytes are strongly depleted in knockdown samples. On the contrary, the proportion of goblet cells is not significantly affected. Third, the knockdown of *hnf4* affects the parenchymal lineage by reducing its larger population, the ***pgrn*+ cells**. Finally, through a differential gene expression analysis, we found multiple **putative target genes of *hnf4*** in phagocytes and parenchymal cells. Overall, the effects described here were very tissue-specific and not broadly distributed among all cell types, validating the sensitivity and specificity of the technique. These results show how **single-cell RNAi** can enormously improve the resolution of knockdown studies, offering more comprehensive insights than the combination of RNAi with other techniques, such as bulk RNA-seq or *in situ* hybridization.

CHAPTER IV: EVOLUTIONARY COMPARISON OF PLANARIAN SPECIES BY SINGLE CELL TRANSCRIPTOMICS (PRELIMINARY RESULTS)

INTRODUCTION

1. BIOLOGICAL DIVERSITY OF PLATYHELMINTHES

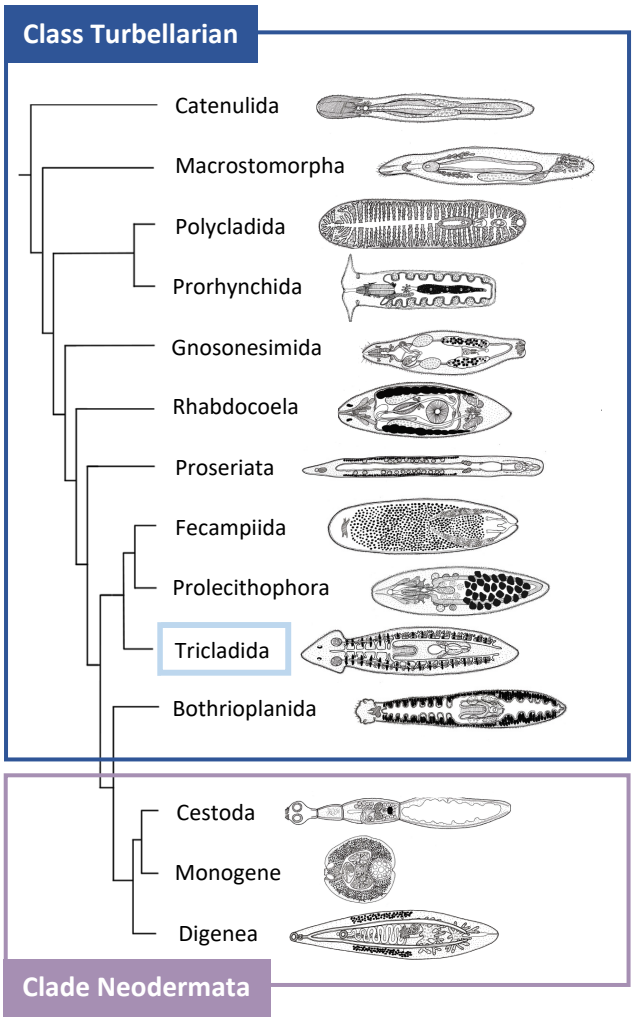
Platyhelminthes (from the Greek *platy*=flat and *helminth*=worm) are one of the simplest bilaterian and the fourth-largest phylum in the animal kingdom, with more than 20,000 described species (Riutort et al., 2012). Platyhelminthes are worldwide spread and inhabit a variety of ecosystems. They divide in multiple subgroups with highly diverse biological features. According to their lifestyle, they are classified in **free-living** (Turbellaria) or **parasitic** (Neodermata) groups (**Figure 4.1**). **Turbellarian** are mostly carnivorous, while **Neodermata** are obligate parasites. Neodermata classes are responsible for multiple human, fish, and livestock diseases, as tapeworms or flukes. Platyhelminthes also exhibit uneven **regeneration capacities**. Neoblast-like cells have been found throughout the whole phylum. However, multiple taxa have limited or no ability to regenerate. For instance, *Macrostomum lignano* (order Macrostomorpha) can regenerate any missing body parts, except the brain. Meanwhile, species like *Phaenocora unipunctata* (order Rhabdocoela) show no regeneration whatsoever (Egger et al., 2007). In the same way, Platyhelminthes exhibit diverse **reproductive biotypes**, including different models of sexual and asexual reproduction. Generally, asexuality correlates with better regeneration capacities (Collins, 2017).

2. CELL DIVERSITY IN PLANARIANS

Platyhelminthes have different lifestyles, habitats, reproductive strategies and regeneration capacities. Therefore, from an evolutionary perspective, arises the question of whether all Platyhelminthes have the same cell types. Given the extension of the phylum, this enquiry is difficult to cover. Thus, we focused on a smaller phylogenetic group within the Platyhelminthes, **the planarians**, to approach this question.

Planarian is the common name given to free-living flatworms of the order **Tricladida**. There is a great variety of planarian species. Some of them are widely studied, but others are barely described. Planarians have different morphologies, reproductive strategies, behaviours, and environmental distributions (Sluys et al., 2009; Vila-Farré et al., 2011). Among the Platyhelminthes, planarians have the most outstanding **regeneration capacities**, with stem cells (neoblasts) capable of regenerating even most challenging tissues, such as the pharynx and the brain. Planarians have long been model organisms in regeneration studies, for which they have an **extensive literature** and multiple **bioinformatic resources** available (Brandl et al., 2016, https://planosphere.stowers.org/). Besides, the **single-cell atlases** of *Schmidtea mediterranea*

Platyhelminthes



and *Dugesia japonica* have been published in recent years (Fincher et al., 2018; García-Castro et al., 2021; Plass et al., 2018), as seen in previous chapters.

Finally, the adult planarian contains a snapshot of all possible cell types, from pluripotent neoblasts to differentiated cells, which allows cell **differentiation trajectories** to be clearly traced (Wolf et al., 2019). These unique characteristics make planarians an ideal starting group to study evolutionary cell diversity in Platyhelminthes. Focusing on phylogenetically closer species also has the advantage of allowing us to answer more subtle questions about gene expression and cell differentiations.

Figure 4.1 Phylogeny of the Platyhelminthes. Terminal taxa show the names and drawings of the orders of class Turbellarian and the different classes of Neodermata. (*Adapted from Laumer et al., 2015*).

Hence, the biological questions posed for this study are as follows:

- Do all planarians species have the same **cell types** in the same proportions?
- How does **gene expression** vary across species?
- Are **differentiation trajectories** common between species?
- What are the differences between **sexual** and **asexual** strains?

3. PLANARIAN PHYLOGENY

Planarians are spread all over the world and occupy multiple ecosystems. In old literature, planarians were classified in three major groups based solely on their habitat: Maricola (marine), Paludicola (freshwater) and Terricola (land). Nonetheless, this nomenclature has been discontinued based on new molecular evidence (Sluys et al., 2009). Nowadays, the order

Tricladida divides in three suborders: **Maricola** (marine species), **Cavernicola** (freshwater species that inhabit caves), and **Continenticola** (freshwater and land species). The suborder Continenticola comprises two superfamilies, **Planarioidea** and **Geoplanoidea**. Within Planarioidea, all families (Planariidae, Kenkiidae and Dendrocoelidae) inhabit in fresh water. In turn, Geoplanoidea includes freshwater (Dugesiidae) and land species (Geoplanidae) (Ronald Sluys and Marta Riutort, 2018). In this chapter, we will focus on the following freshwater species: *Schmidtea mediterranea*, *Schmidtea polychroa, Dugesia japonica* and *Girardia tigrina,* of the family **Dugesiidae**, and *Polycelis nigra,* of the family **Planariidae** (**Figure 4.2**) (Ronald Sluys and Marta Riutort, 2018; Sluys et al., 2009).

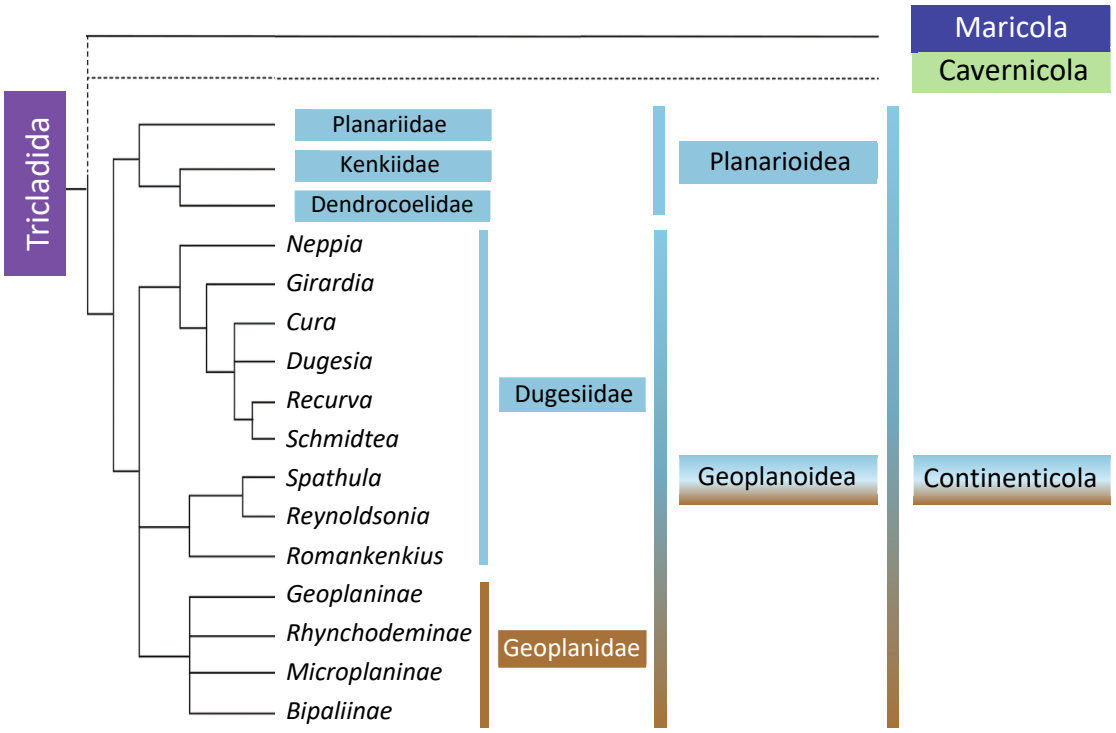


Figure 4.2 Phylogeny of the Tricladida. Planarian species are divided in three suborders: Maricola, Cavernicola and Continenticola. The latest comprises the superfamilies of Planarioidea and Geoplanoidea. Each superfamily includes multiple families (in colour squares) and genera. *Adapted from Ronald Sluys and Marta Riutort, 2018.*

4. INTRODUCTION TO THE PLANARIAN SPECIES

Schmidtea mediterranea and *Schmidtea polychroa* are the phylogenetically closest species included in this chapter (**Figure 4.3 A**), diverging ~43 million years ago (Mya) (Lázaro et al., 2011). ***Schmidtea mediterranea*** (Benazzi et al., 1975) is the most well-known planarian species. In recent years, it has been the favourite option for molecular biology studies (Plass et al., 2018; Wurtzel et al., 2015; Zeng et al., 2018). Moreover, it counts with different assembled genomes and annotations (Grohme et al., 2018; Guo et al., 2022). Geographically, the species is

distributed in a few coastal regions of the **Western Mediterranean**: Catalonia (Barcelona and Girona), the Balearic Islands (Mallorca and Menorca), Tunisia (Lebna), Corsica, Sardinia, and Sicily. These animals are mainly **diploids**, although there are some triploid populations (Lázaro et al., 2011). There are sexual and asexual *Schmidtea mediterranea* strains (**Figure 4.3 B-B'**). **Sexual** individuals are cross-fertilizing hermaphrodites and lay cocoons, while **asexual** planarians are fissiparous -they reproduce by fission, tearing themselves in two or more pieces- (Chong et al., 2011). In *S. mediterranea,* asexuality is caused by a translocation between the 1st and 3rd chromosome and cannot be reverted. The asexual strain has only been identified in Catalonia and the Balearic Islands. Typically, laboratory strains used around the world are asexual and originates from the same pond system in Montjuïc (Barcelona) (Lázaro et al., 2011).

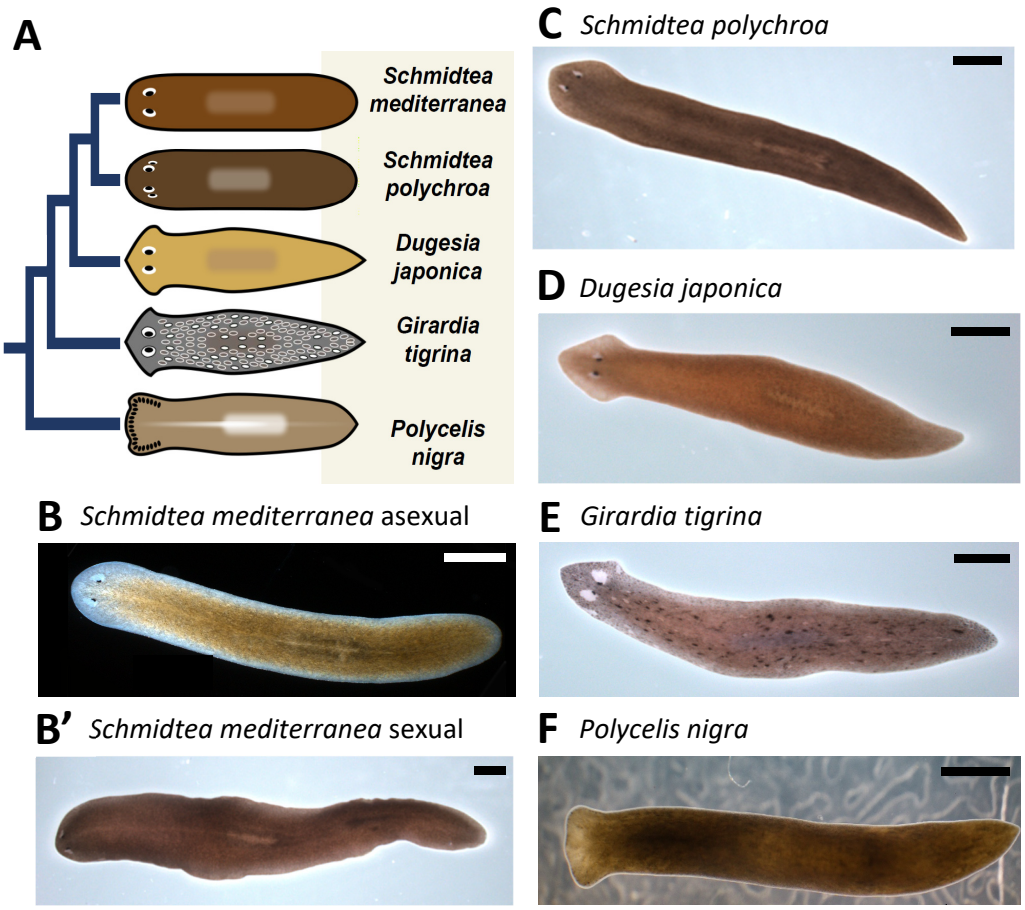


Figure 4.3 Planarian species used in this study. A) Cartoon of phylogenetic relations between species **B-F)** Microscopy images of *Schmidtea mediterranea* asexual (B) and sexual (B') strains, *Schmidtea polychroa* (C), *Dugesia japonica* (D), *Girardia tigrina* (E) and *Polycelis nigra* (F). Scale bars are equivalent to 1 mm. *Images B and F were taken from Planmine:*

<https://planmine.mpibpc.mpg.de/planmine/report.do?id=2000001#ad-image-0>
<https://planmine.mpibpc.mpg.de/planmine/report.do?id=2000002#ad-image-0>

Schmidtea polychroa (Schmidt, 1861) (**Figure 4.3 C**) is widely distributed in Europe (as far north as Southern Sweden), North América, the Azores and North Africa (Vila-Farré et al., 2011). In this species, **diploid** strains reproduce **sexually** while **polyploid** ones (triploids, tetraploids or pentaploids) reproduce by **parthenogenesis**, laying unfertilized eggs. These animals are cross-fertilizing hermaphrodites; therefore, parthenogenetic cocoons need **allosperm** to trigger embryonic development. Sexual and parthenogenic individuals can coexist and even interbreed, producing fertile offspring. Parthenogens, in addition, can occasionally reproduced sexually. *S. polychroa* does not reproduce by fission (D'Souza et al., 2006).

Dugesia japonica (Ichikawa and Kawakatsu, 1964) (**Figure 4.3 D**) belongs to the same monophyletic clade than the genus *Schmidtea*. This species has been extensively used in planarian studies from classical to modern literature, and thus has genomes (An et al., 2018; Tian et al., 2022) and transcriptomes available. *D. japonica* inhabits **East Asia**, including the Japanese Islands -where it is the most common planarian-, Taiwan, Korea, Eastern China and Russia (Primorsky) (Kawakatsu et al., 1995). These individuals are self-fertilizing hermaphrodites, mostly diploids, triploids or mixoploids. Diploids can reproduce **sexually** or **asexually** (fissiparous), while other karyotypes reproduce by fission, parthenogenesis or self-fertilization (Tamura et al., 1995).

Girardia tigrina (Girard, 1850) (**Figure 4.3 E**) belongs to a different clade than *Dugesia* and *Schmidtea* (Álvarez-Presas et al., 2008). The species is native to America, where it is widespread, but has also become invasive in Asia (Australia and Japan) and Europe (Vila-Farré et al., 2011). In America, *G. tigrina* can be **asexual** (fissiparous), **sexual**, or alternate between seasons. On the contrary, most invasive populations described in other continents were asexual. Sexual individuals are diploids, while asexual strains can show diploidy, triploidy or mixoploidy (Stocchino et al., 2019).

Finally, ***Polycelis nigra*** (Müller, 1774) (**Figure 4.3 F**) is the most distant species, belonging to a different family (Planariidae) and superfamily (Planarioidea) (Sluys et al., 2009). These planarians are widespread in Western Europe and the British Isles. *P. nigra* has a distinctive morphological feature compared to the other species mentioned: a line of abundant small eyes that extends along the margins of the head. Besides, under desiccation and starvation conditions, *P. nigra* can wrap itself in a mucous capsule to survive (Vila-Farré et al., 2011), suggesting they could possess specialized secretory glands. This species has two reproductive modes: **diploid** individuals are sexual cross-fertilizing hermaphrodites, while **polyploid** individuals (mainly triploids) reproduce by pseudogamous parthenogenesis (Beukeboom et al., 1998).

5. THE REPRODUCTIVE SYSTEM OF PLANARIANS

The planarian reproductive system develops post-embryonically from the neoblasts, until the animal reaches reproductive maturity (**Figure 4.4 A-B**), and can be regenerated after injury (Newmark et al., 2008). Sexual and parthenogenic planarians are **simultaneous cross-fertilizing hermaphrodites**, although some species can also self-fertilize (D'Souza and Michiels, 2009; Tamura et al., 1995). Asexual planarians do not develop their gonads and complementary organs, but still possess a truncated germ line (Wang et al., 2010, 2007). Reproduction modes are always affected by the environment. For instance, under **culturing conditions**, some species can totally or partially lose their sexual organs, limiting them to asexual reproduction.

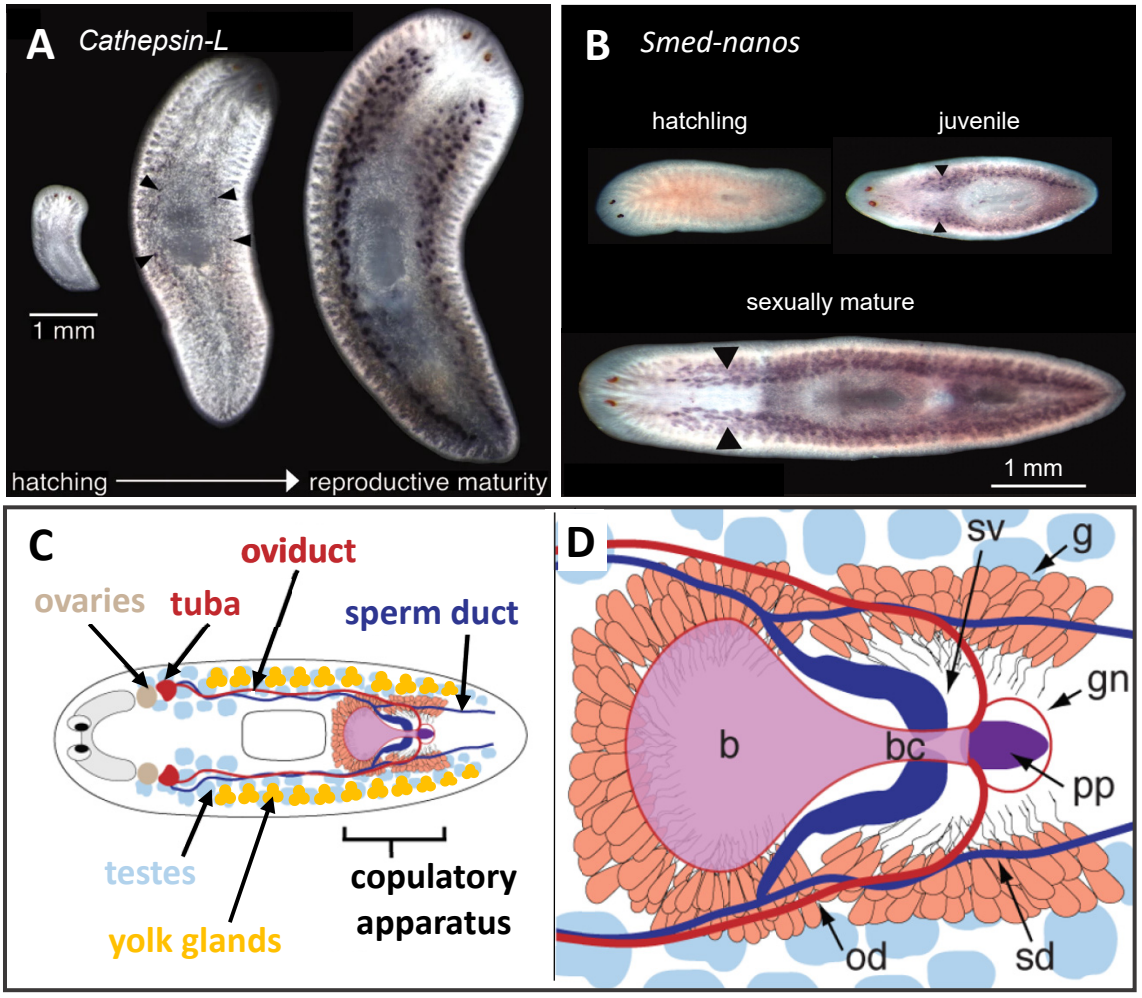


Figure 4.4 The reproductive system of planarians. A-B) Whole-mount *in situ* hybridizations of *Cathepsin-L* and *Smed-nanos* (testes markers) in *Schmidtea mediterranea*. Transcripts are undetectable in hatchlings, but begin to be expressed in juvenile animals. *Adapted from Wang et al., 2007 and Zayas et al., 2005.* **C)** Cartoon of female and male gonads, and complementary organs **D)** Detail of the copulatory apparatus showing the bursa (b), bursa canal (bc), gonopore (gn), penis papilla (pp), seminal vesicles (sv), sperm ducts (sd), oviducts (od) and other glands (g). *Adapted from Chong et al., 2011.*

Hermaphrodite planarians have a complex reproductive system with simultaneous **female and male gonads**, which arose from germline stem cells derived from the neoblasts, and their complementary organs (e.g. the yolk glands). The only Tricladida known to have pure females and males is the marine species *Sabussowia dioica* (Charbagi-Barbirou and Tekaya, 2009). Sexual planarians have two ventral **ovaries** located after the cephalic ganglia (brain). The **oviducts** and **yolk glands** (vitelline) run along to the nerve cords and connect the ovaries with the **copulatory apparatus**, located in the post pharyngeal region. Parallel to them, planarians have two dorsolateral lines with multiple **testes**, connected to the copulatory apparatus through the **sperm ducts**. The copulatory apparatus comprises the copulatory **bursa** and bursal canal, the **gonopore** -through which the sperm is exchanged during the copula-, the **penis papilla**, the seminal vesicles -that store the mature sperm-, and other glands (**Figure 4.4 C**).

During mating, sperm is reciprocally exchanged between partners and deposited in the bursa, from where it is transferred through the oviducts. The sperm arrives to the **tuba**, a neck-shaped region at the end of the oviducts, next to the ovaries. Here, sperm can be stored for months. When the mature oocyte leaves the ovary, it is fertilized in the tuba. Fertilized eggs go down the oviducts, adding yolk cells produced by the yolk glands. In the bursa, the glands around the copulatory apparatus synthesize the **egg capsule**, where multiple embryos and yolk cells are packaged. This creates the **cocoon**, which is finally released through the gonopore (**Figure 4.4 D**) (Chong et al., 2011; D'Souza and Michiels, 2009; Saló, 2006).

The reproductive system of planarians has been studied using different techniques, such as microscopy, whole-mount *in situ* hybridization, microarrays or RNA-seq. However, the complexity of this system has not been profiled by single-cell transcriptomics, as all publications to date have used only asexual strains (Benham-Pyle et al., 2021; Fincher et al., 2018; García-Castro et al., 2021; Plass et al., 2018). To our knowledge, this chapter presents the **first study at single-cell resolution** that includes **sexual** planarian species (*S. polychroa* and *P. nigra*) and strains (*S. mediterranea* and *G. tigrina*), and compares them to asexual planarians (*D. japonica* and *S. mediterranea*). In addition, to deepen into the development of the reproductive system, we included different **life-history stages** of *S. polychroa*: non-sexualized hatchlings and juveniles, and sexualized adults.

6. CROSS-SPECIES SINGLE-CELL TRANSCRIPTOMICS

Using single-cell transcriptomics scientists have succeeded in profiling the **single-cell atlases**, transcriptomic signatures, and cell differentiation trajectories of multiple tissues and whole organisms across Metazoan. However, from invertebrates (Davie et al., 2018; Packer et al., 2019; Sebé-Pedrós et al., 2018a; Siebert et al., 2019) to vertebrates (Briggs et al., 2018; Cao et al., 2019; Farrell et al., 2018; Vento-Tormo et al., 2018), most single-cell publications are still focused on individual animals. More complex **cross-species comparisons** are still scarce.

Cross-species single-cell transcriptomics will help to elucidate the diversity and evolution of cell types and their gene regulatory programmes. Yet, these studies remain challenging, specially across phyla. As evolutionary distances increase, transcriptomic signatures, gene sequences, and cell types differ more and more, and grows the complexity to assign gene homologies. This poses a problem, as most cross-species analyses only include **one-to-one gene orthologs**, losing much of the species-specific information (Shafer, 2019; Tarashansky et al., 2021). In addition, technical variability across experiments requires new normalization and batch correction strategies that maintain real biological diversity between samples.

Despite the challenges, the scientific community is moving towards this direction, as evidenced by the publication of multiple cross-species single-cell studies in the last years (Baron et al., 2016; Geirsdottir et al., 2019; Lust et al., n.d.; Ton et al., 2022; Tosches et al., 2018). This, in turn, has been accompanied by the development of novel **bioinformatic tools** for data integration, such as Harmony (Korsunsky et al., 2019), LIGER (Welch et al., 2019), Scanorama (Hie et al., 2019) or SAMap (Tarashansky et al., 2021). In the future, we will implement some of these novel approaches to perform a single-cell comparison at **short phylogenetic distance**, as all the species included in this project belong to the same order. Our aim will be comparing cell type abundances, differentiation trajectories and gene expression among different flatworm species, as well as assess the variations between sexual and asexual strains.

RESULTS

1. ScRNA-SEQ OF PLANARIAN SPECIES USING ACME & SPLiT-SEQ

We profiled the single-cell transcriptomes of different planarian species (***Schmidtea mediterranea, Schmidtea polychroa, Dugesia japonica, Girardia tigrina*** and ***Polycelis nigra***), strains (***S. mediterranea*** **sexual** and **asexual**) and life-history stages (***S. polychroa*** **hatchlings**, **juveniles** and **adults**) using ACME and SPLiT-seq. All samples were dissociated by separate in **ACME** as described in Chapter II. ACME-cells were frozen once before being used for SPLiT-seq. Each sample (species, strain or life stage) ran in a different batch, except for *S. polychroa* hatchling and juvenile cells, which were loaded in the same plate (but separated in different wells). In total, **7 SPLiT-seq experiments** were performed using the same workflow and technical improvements implemented in Chapter III. The initial SPLiT-seq plates were loaded with **5000 events/well** (240,000 total events), and cells were **FACS-sorted** after the third round of barcoding. Each batch was sorted into two separate sub-libraries. Library preparation and sequencing proceeded as in previous chapters (see more in Chapter V: Methods).

We mapped *S. mediterranea* and *D. japonica* reads to their respective genomes (An et al., 2018; Guo et al., 2022) using the annotations generated for Chapter II. *S. polychroa, G*. tigrina, and *P. nigra* were mapped to *de novo* Iso-Seq transcriptomes, which were annotated using BLAST against the Swiss-Prot protein sequence database (https://www.uniprot.org/). The percentages of mapped reads were as follows: **94.9%** for *S. mediterranea* asexual, **93.7%** for *S. mediterranea* sexual, **93.6%** for *S. polychroa* adults, **94.7%** for *S. polychroa* hatchlings and juveniles, **96%** for *D. japonica,* **89.3%** for *G. tigrina* and **84.8%** for *P. nigra.* The pre-processed reads from both *S. mediterranea* strains were merged directly into the same count matrix, without batch correction. *S. polychroa* life-history stages were merged in the same way.

Table 4.1 Average number of genes and UMIs per cell and per sample.

	Metrics	
	Genes/cell	UMI/cell
Schmidtea mediterranea (asexual)	344	525
Schmidtea mediterranea (sexual)	257	345
Schmidtea polychroa (adult)	305	398
Schmidtea polychroa (juvenile)	232	296
Schmidtea polychroa (hatchling)	204	255
Dugesia japonica	239	376
Girardia tigrina	271	368
Polycelis nigra	217	283

After matrices pre-processing, we profiled **24,809 total cells** of *S. mediterranea*, **7,912** from asexual and **16,897** from sexual planarians. For *S. polychroa*, we profiled **48,990 total cells**. Of these, **15,581** cells were from adults, **15,382** from juveniles and **18,027** from hatchlings. The number of cell transcriptomes profiled for others species were: **10,500 cells** (*D. japonica*), **20,190 cells** (*G.* tigrina) and **12,134 cells** *(P. nigra*). All cells included in the analysis had a minimum of 100 genes. Metrics for the average number of genes and UMIs per cell are shown in **Table 4.1**.

2. CLUSTER ANNOTATION OF *SCHMIDTEA MEDITERRANEA*

Most *Schmidtea mediterranea* clusters were annotated using an internal **reference dataset** of the same species (not shown). Cell identities were first transferred from this reference dataset using the function **Ingest**. Then, transferred identities were combined with the clustering of the new dataset at resolution 3.0, and clusters were assigned to their majoritarian Ingest identity.

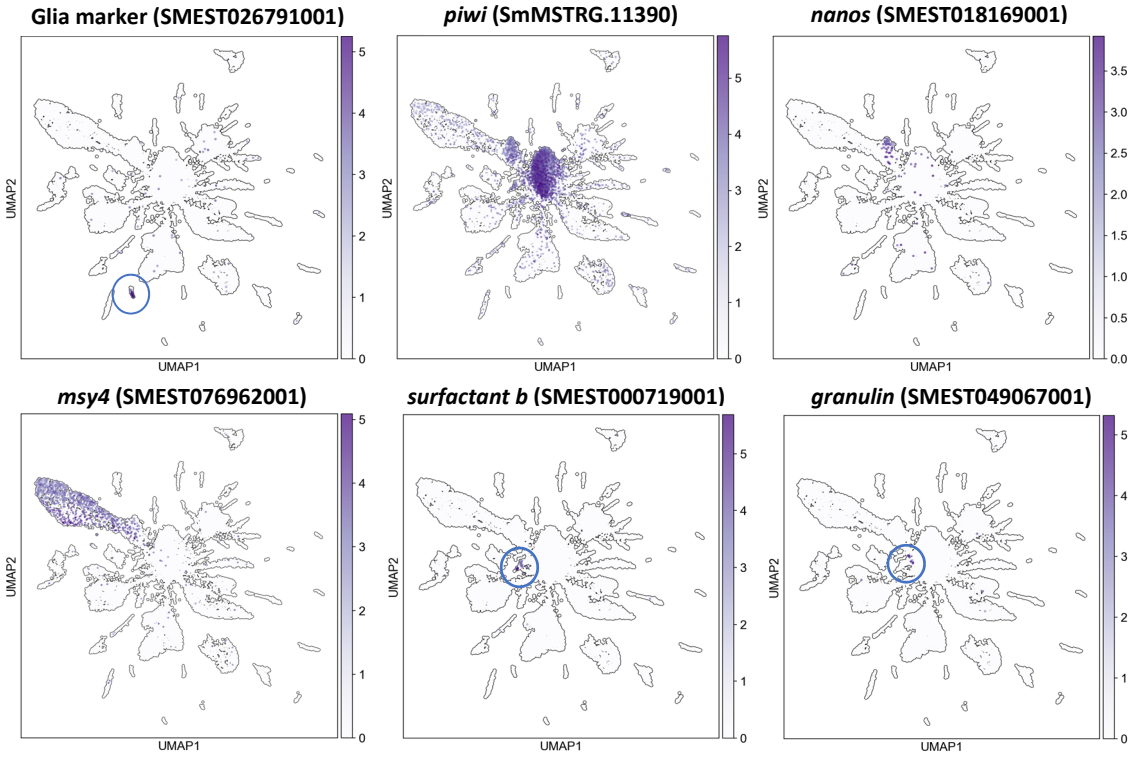


Figure 4.5 Feature plots of different *S. mediterranea* markers used for manual annotation. Marker sequences were obtained from the following sources: glia marker (*Plass et al., 2018*), *piwi* (DQ186985, *Reddien et al., 2005*), *nanos* (EF035555, *Wang et al., 2007*), *msy4* (BK007101, *Wang et al., 2010*), *surfactant b* (KY847536.1, *Rouhana et al., 2017*) and *granulin* (DN304193.1, *Zayas et al., 2005*).

Our reference dataset lacks the annotation of the glia, and consists solely of *S. mediterranea* asexual cells. Thus, the **glia** and all **neoblast**, **reproductive system** (germline, gonads and complementary organs), and **unknown** clusters were manually annotated (at Leiden resolution

3.0) according to the expression -or lack of expression- of different markers described in the literature: SMEST026791001 (glia marker), SmMSTRG.11390 (*piwi*, neoblast marker), SMEST018169001 (*nanos*, germ line progenitors and testes marker), SMEST076962001 (*msy4*, male germ cells marker), SMEST000719001 (*surfactant b*, yolk glands marker) and SMEST049067001 (*granulin*, sperm ducts and seminal vesicles marker) (**Figure 4.5**) (Chong et al., 2011; Plass et al., 2018; Reddien et al., 2005; Rouhana et al., 2017; Steiner et al., 2016; Wang et al., 2010, 2007; Zayas et al., 2005). Some of these gene were between the top 20 markers of the clustering. SmMSTRG.11390 was a top marker for clusters 0, 2, 41, 49 and 64, annotated as neoblasts, and cluster 70, annotated as reproductive system. SMEST000719001 was top marker of cluster 57 and SMEST026791001 of cluster 63, annotated as reproductive system and glia, respectively. In total, 56 different cell types were annotated (**Figure 4.6** and Supplementary 8).

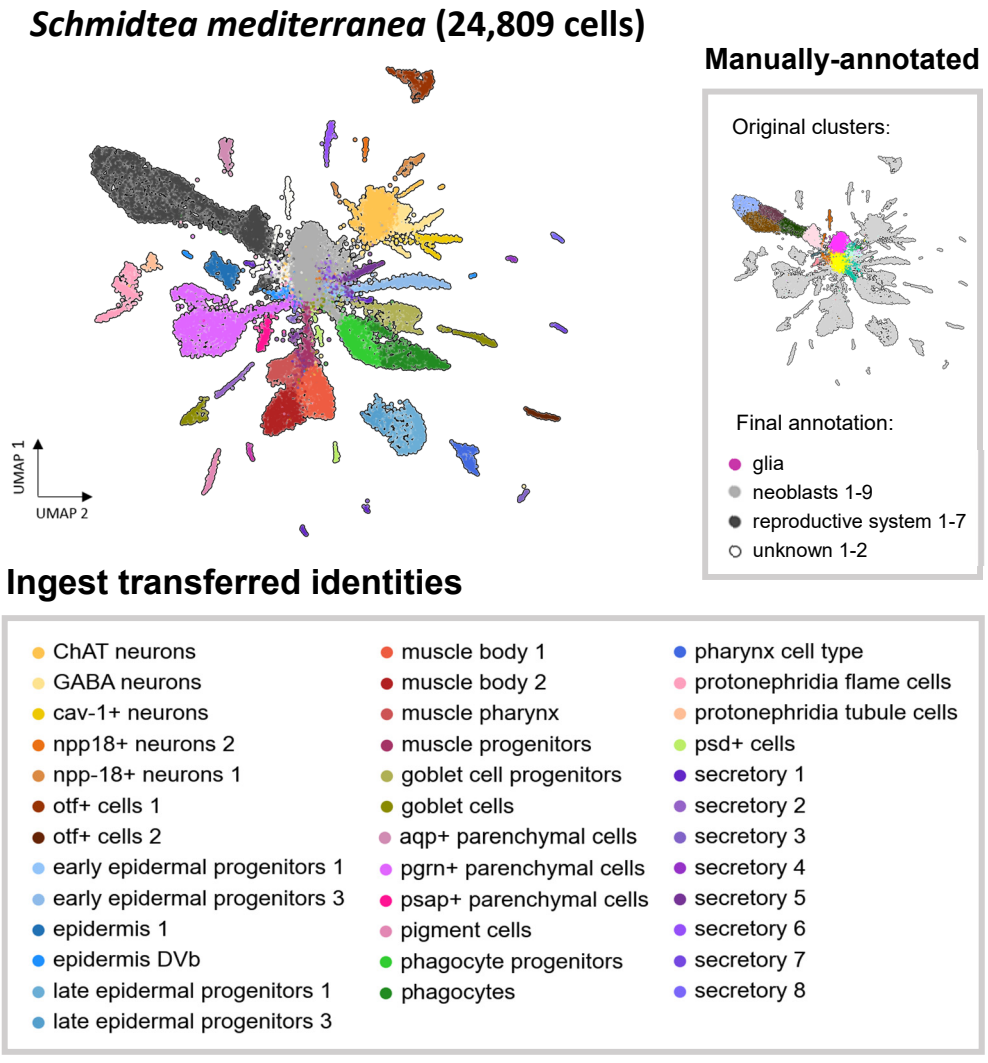


Figure 4.6 Annotated single-cell atlas of *Schmidtea mediterranea*. 2D UMAP visualisation of merged *S. mediterranea* asexual and sexual datasets. Cells are coloured according to their cluster identities (Supplementary 8). Cluster annotation was performed using Ingest (transferring labels from a reference dataset) or by evaluating the expression of markers shown in Figure 4.5 (manually-annotated).

3. PRELIMINARY ANNOTATION OF BROAD CELL TYPES IN OTHER PLANARIAN SPECIES

Schmidtea polychroa, Dugesia japonica, Girardia tigrina and *Polycelis nigra* clusters were preliminary annotated within the following **broad cell types**: epidermis, epidermis DVb, eyes, goblet, muscle, neoblasts, neurons, *otf*+, parenchymal, phagocytes, pharynx, protonephridia, *psd*+, reproductive system, secretory and unknown (Supplementary 8). The annotation was performed manually, by assessing the expression of multiple markers per tissue (not shown). These markers were extracted from our *S. mediterranea* **reference dataset** clustered at very low resolution (0.2). Markers sequences were BLAST to the different species to find **homologous genes**. The expression of the homologs with lower E-values was used to manually assigned clusters to their broad cell type identities (**Figure 4.7**).

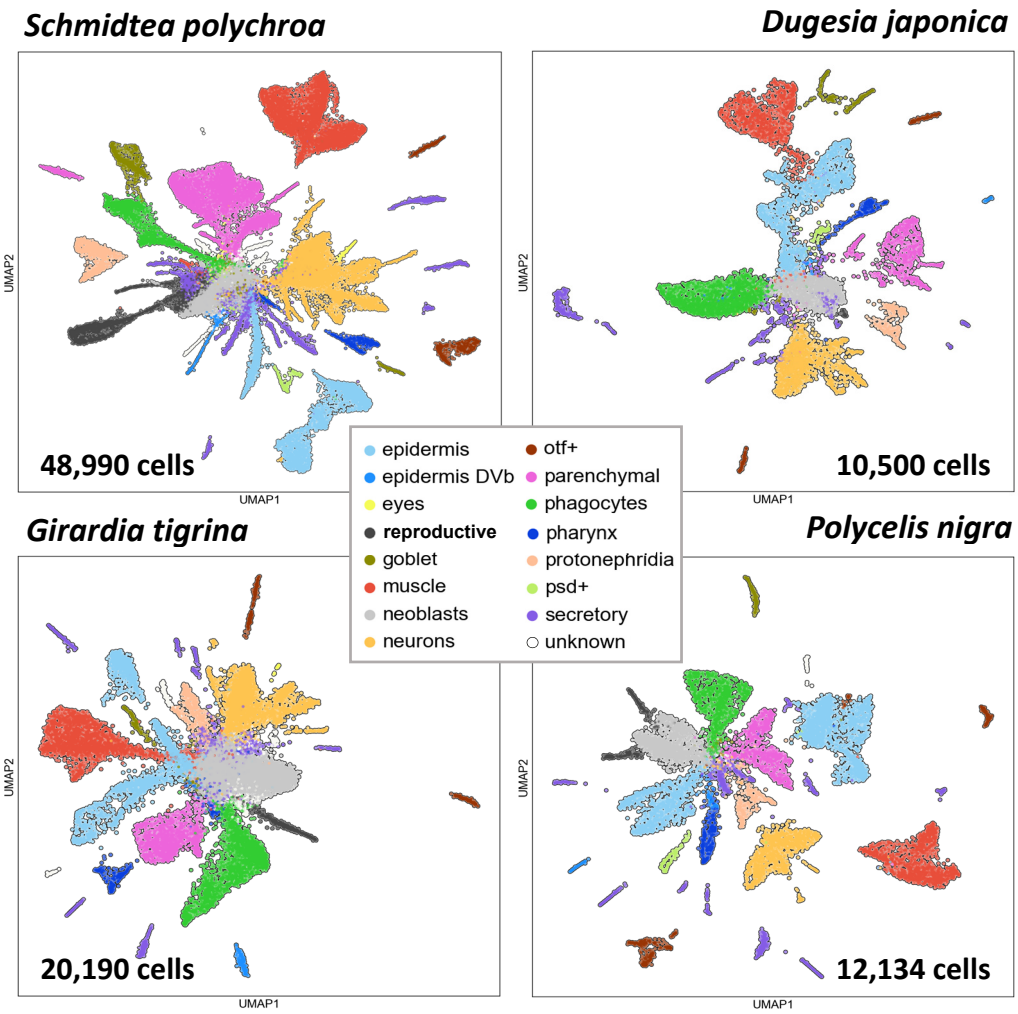


Figure 4.7 Preliminary annotation of the single-cell atlases of different planarian species. 2D UMAP visualisations of *S. polychroa* (adult, juveniles and hatchlings), *D. japonica*, *G. tigrina* and *P. nigra* datasets. Cells are coloured according to their broad cell types (Supplementary 8). Annotation was performed manually by assessing the expression of homologous markers to our references *S. mediterranea* dataset.

4. ANALYSIS OF CLUSTER ABUNDANCES IN *S. MEDITERRANEA*

The *Schmidtea mediterranea* dataset is a merge of cells from **sexual** and **asexual** individuals, cultured separately. Although they belong to the same species, both strains present different UMAP distributions (**Figure 4.8 A**). To explore these differences at tissue resolution, we performed a preliminary analysis on **cluster abundances**. For this, we extracted the percentage of cells per cluster and per condition (Supplementary 8). Then, for each cluster, a ratio was calculated dividing the percentage in sexual samples by the percentage in asexual samples. Finally, these ratios were transformed into **log ratios** (Chapter V: Methods).

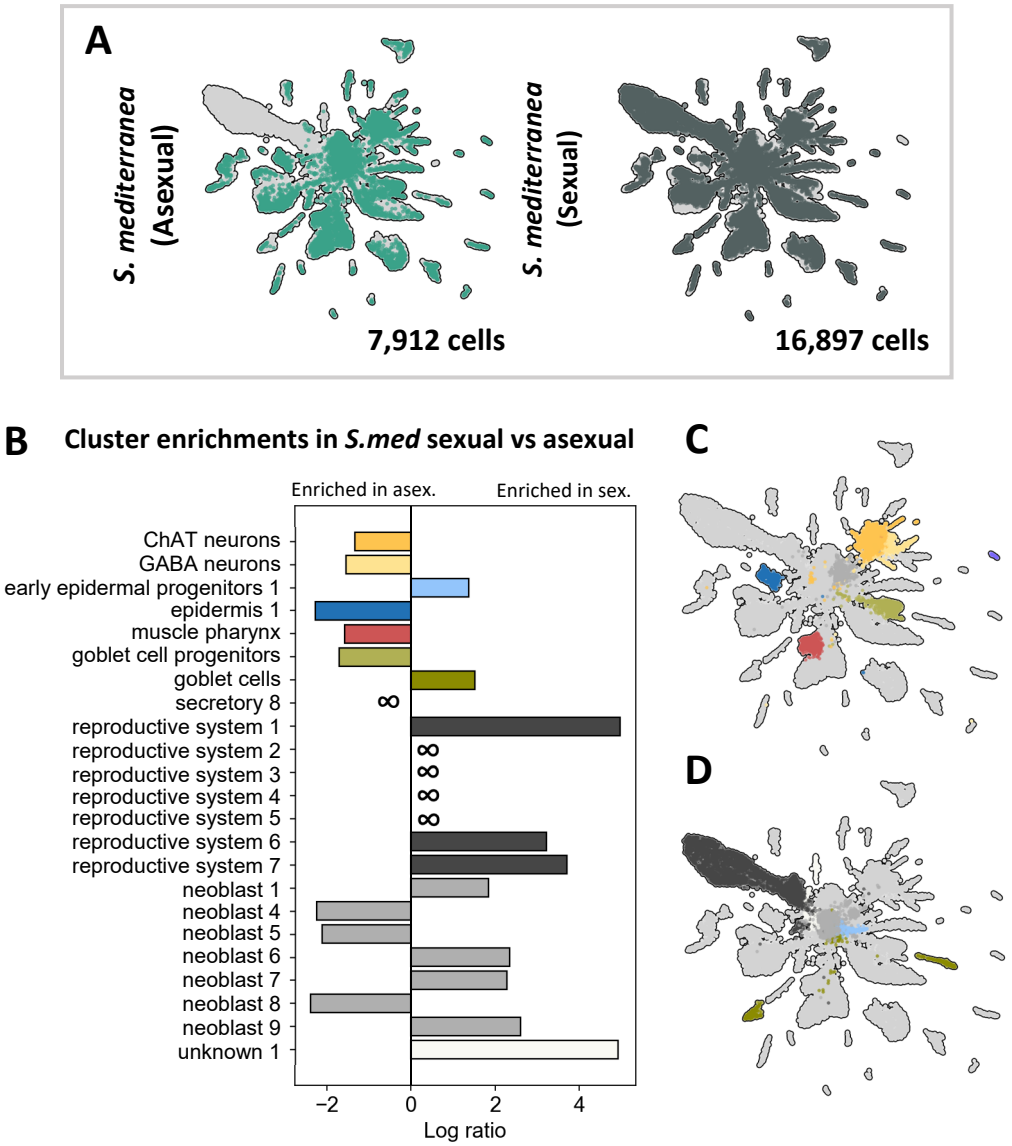


Figure 4.8 Analysis of cluster abundances in *Schmidtea mediterranea*. A) UMAP plots and total number of cells from sexual and asexual samples. **B)** Bar plot of cluster enrichments in *S. mediterranea*. Clusters with a negative log ratio are enriched in the asexual strain, and clusters with positive log ratios are enriched in sexual animals. Infinite symbols indicate that the opposite condition has no counts for this cluster. **C)** UMAP plot showing clusters enriched in the asexual strain. **D)** UMAP plot showing clusters enriched in the sexual strain.

Clusters with log ratios below -1 were considered as enriched in the asexual strain, while those with log ratios above +1 were considered enriched in sexual individuals. In total, **23 clusters** were enriched in one condition or another (**Figure 4.8 B**). Asexual planarians had a higher number of ChAT and GABA neurons, epidermis 1, muscle pharynx and goblet progenitors. Besides, **secretory 8** cells were only found in this sample. In turn, sexual planarians had higher number of early epidermal progenitors 1 and goblet cells. As expected, all seven **reproductive system** clusters were either more abundant or only present in the sexual strain. One of clusters classified as unknown (**unknown 1**) was also much more abundant in sexual animals, suggesting it could be part of the reproductive system. Finally, most **neoblast** clusters were more abundant in one of the two strains (**Figure 4.8 B-D**).

5. ANALYSIS OF CLUSTER ABUNDANCES IN *S. POLYCHROA*

The *Schmidtea polychroa* dataset contains cell transcriptomes from different life-history stages of the species: **hatchling**, **juvenile** and **adult** planarians. **Adult samples** were from **fully sexual individuals**, selected from a general culture by checking the size (taking the largest animals), the presence of gonopore, and the ability to lay cocoons. Candidates were separated in different boxes to assess these features. **Juvenile samples** were generated selecting mid-to-large size individuals, with no gonopore and no ability to lay cocoons. Finally, **hatchling samples** were obtained by separating cocoons in a different box, and collecting animals only a few days after hatching.

For each life-history stage, we can observe variations in the UMAP distribution (**Figure 4.9 A**). As for *S. mediterranea,* we made a preliminary analysis on cluster abundances to assess these variations. The analysis was performed as described in the previous section (Chapter V: Methods and Supplementary 8). In this case, there were three different conditions to compare, so the analysis was run by pairs: adult vs hatchling (**Figure 4.9 B-B'**), adult vs juvenile (**Figure 4.9 C-C'**) and juvenile vs hatchling (**Figure 4.9 D-D'**).

The largest differences occurred between the most distant conditions, adults and hatchlings. In this pair, **25 clusters** were enriched with log ratios above ±1. Planarian hatchlings presented higher proportions of multiple **neuronal**, **secretory** and **neoblasts** clusters, and of ***psd+*** cells. **Cluster 66 (unknown)** and **cluster 12 (epidermal)** were also more abundant in hatchlings compared to adults. In fact, these two clusters are visually depleted in the adult UMAP (**Figure 4.9 A**). As expected, the **reproductive system** populations were enriched in adult planarians. **Cluster 36 (unknown)** and **cluster 54 (parenchymal)** were also more frequent in adults, suggesting they could be somehow related with sexual reproduction (**Figure 4.9 B-B'**).

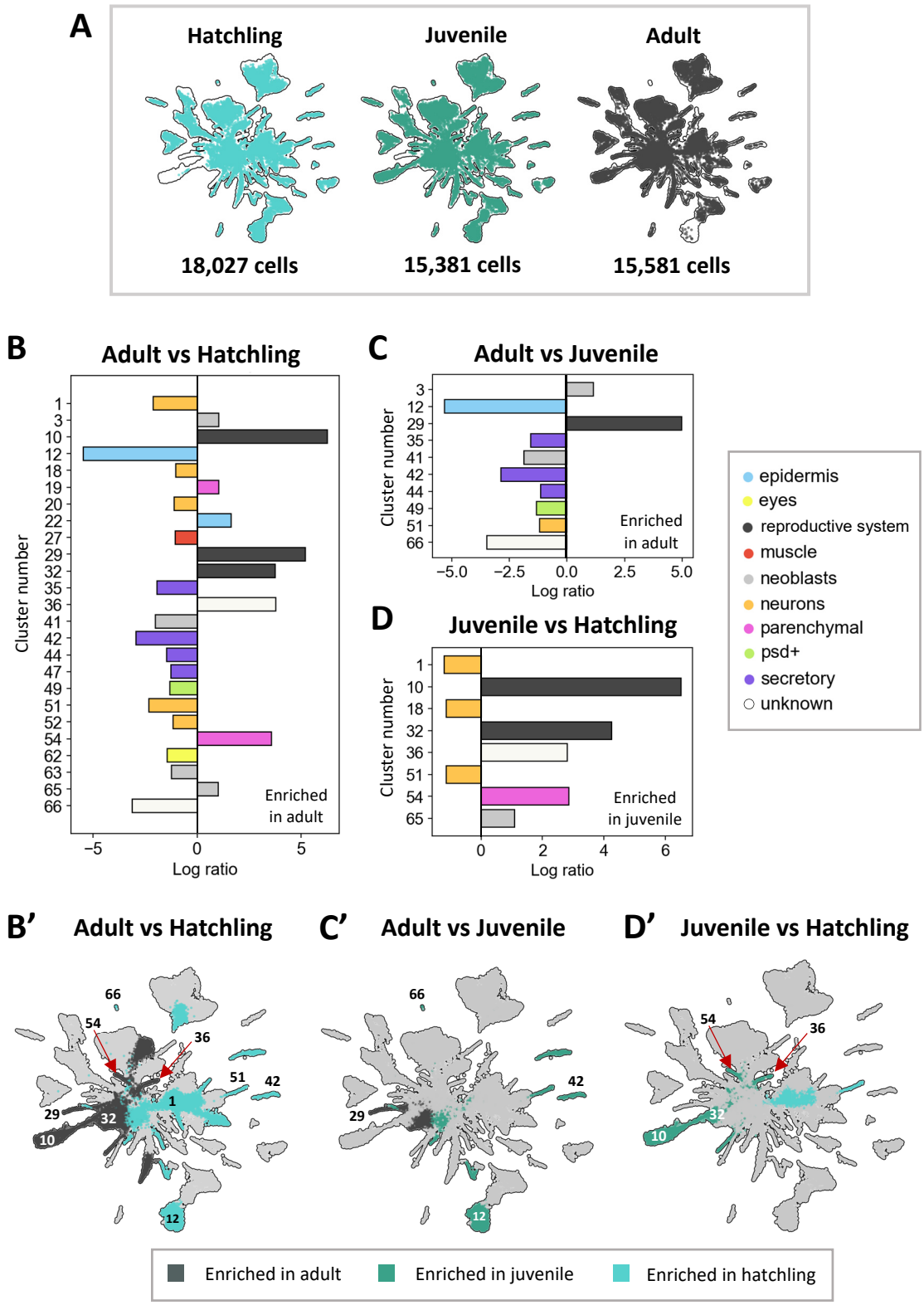


Figure 4.9 Analysis of cluster abundances in *Schmidtea polychroa.* A) UMAP plots and number of total cells of hatchling, juvenile and adult samples. **B-D)** Bar plots of cluster enrichments between pairs of conditions. Clusters with negative log ratios are enriched in hatchlings **(B and D)** and juveniles **(C)**. Clusters with positive log ratios are enriched in adults **(B and C)** and juveniles **(D)**. **B'-D')** UMAP plots showing clusters enriched in each pair of conditions. Clusters with log ratios above ±2 are indicated in the plots.

In the adult vs juvenile pair, we obtained 10 clusters with log ratios above ±1. Juvenile samples are midway between hatchlings and adults. They keep **similar enrichments than hatchlings** in neurons, secretory, neoblats and *psd*+ cells. Clusters 66 and 12 are also highly enriched in juveniles compared to adults. But surprisingly, adults only have two populations enriched, cluster 3 (neoblast) and cluster 29 (reproductive system) (**Figure 4.9 C-C'**). These results suggest the **reproductive system is almost as developed in juveniles as it is in the adult planarian**. Finally, when comparing juvenile vs hatchling, we find 8 clusters with log ratios above ±1. Planarian hatchlings are enriched for **neuronal clusters**. In juveniles, clusters 10 and 32 (main branch of the reproductive system), cluster 36 (unknown) and cluster 54 (parenchymal) are the ones more enriched (**Figure 4.9 D-D'**). Overall, these results profile the progressive development of the reproductive system at cluster resolution in *S. polychroa,* and provides an overview on the changes occurred in other cell types along different life stages of the animal.

DISCUSSION

The results presented in this chapter are exploratory and still in the **early stages of bioinformatic analysis**. Nonetheless, they serve to visually show the completeness of the experimental part, offer some preliminary insights and identify potential challenges.

1. ASSESSMENT OF DATA PRE-PROCESSING

Starting at the pre-processing level, the **percentage of mapped reads** per species oscillates between 85-95%, which confirms the quality of our reference sequences. In this chapter, we used novel transcriptome assemblies for *S. polychroa, G. tigrina* and *P. nigra*, and already published genomes and annotations for *S. mediterranea* and *D. japonica* (An et al., 2018; García-Castro et al., 2021; Guo et al., 2022). These references are giving good results at the individual level, but using a combination of **transcriptomes and genomes** could become more challenging at the time of integrating all data.

Different species also have **varying numbers of total cells**. The *S. mediterranea* (24,809 cells) and *S. polychroa* (48,990 cells) datasets are bigger because they combine different conditions and experimental batches. *G. tigrina* has the largest number of cells for a single experiment (20,190 cells), while *D. japonica* (10,500 cells) and *P. nigra* (12,134 cells) have the lowest numbers. One of the aims of this project is to compare the proportion of different cell types between species, but less abundant populations, such as the glia or the eyes, cannot be resolved below a certain resolution. For instance, eyes have only been identified in *S. polychroa* and G. *tigrina* (**Figure 4.7**). Thus, in future analyses, these rare cell identities could not be included and comparisons could be limited to broader cell groups. However, less abundant cell types might still be present in other datasets and not have been clustered out. If this is the case, I expect the **integration of all species** into a single dataset, with a much larger number of total cells, will make these low abundant populations cluster out more easily. Yet, I predict one of the main challenges will be to distinguish between real **biological and technical cell diversity**.

2. ASSESSMENT OF CLUSTER ANNOTATION

Most ***Schmidtea mediterranea*** cell identities have been easily annotated by **transferring labels** from a reference dataset. However, certain populations are still missing. This is the case of the *ldlrr-1*+ parenchymal cells and some epidermal and neuronal cell types included in previous publications (García-Castro et al., 2021). In the future, these missing clusters should be looked

for more carefully using specific gene markers. On the other hand, the **reproductive system** has not been properly annotated, as all previous *S. mediterranea* single-cell atlases were of the asexual strain. In future annotations, the reproductive system will be evaluated, cluster by cluster, using published markers for each of the planarian sexual organs and complementary tissues. These markers will need to be very specific (e.g. ovary-specific, testes-specific, etc) and will not necessary be part of the top markers for these clusters.

All **other planarian species** included in the study have been preliminary classified in **broad cell types**. Annotation was performed manually by assessing the expression of genes homologous to *Schmidtea mediterranea* markers. This system would be laborious and imprecise to identify every individual cell type. Hence, for future analyses, I will create a common reference gene annotation based on *S. mediterranea*. For this, I will use software like **OrthoFinder** (Emms and Kelly, 2015) to identify homologs between species. The initial strategy will consist on retaining only **one-to-one homologs**, and use them to integrate all datasets and re-annotate clusters identities.

Despite the shallow annotation, we can draw some preliminary conclusions from the results presented in this chapter. For instance, the ***psd*+ cells** were the only broad cell type (excluding the eyes) that was not found in every sample. In *Girardia tigrina*, *psd*+ cells could not be identified, although this species has the largest individual dataset (**Figure 4.6**). On the contrary to other smaller cell types, *psd*+ cells easily cluster out and are detectable even at very low resolutions. This suggests *G. tigrina* may not possess this particular cell type or, if it did, the transcriptomic signature may be very different from that of other species.

On the other hand, we have observed **variable complexities in the reproductive system** of sexual planarians at single-cell level. The sexual strain of *Schmidtea mediterranea* has the most abundant and complex reproductive system. It comprises 19% of the total cells and has seven clusters annotated at clustering resolution 3.0 (Supplementary 8). Considering the comparisons with the asexual strain, the cluster *unknown 1* (1.5% of cells) could also be related to sexual reproduction (**Figure 4.8 B**). The adult sample of *Schmidtea polychroa* has the second most complex reproductive system, comprising 9% of the total cells (Supplementary 8) and 3 clusters annotated at clustering resolution 3.0. Moreover, clusters 36 and 54 could also be linked to reproduction in this species (**Figure 4.9 B-D**). Although the reproductive system of *S. mediterranea* is apparently more complex, *S. polychroa* has proven to lay more viable cocoons in the lab. These results suggest both species present sufficient complexity at the single-cell level to be considered sexed.

The other sexual species included in the study, ***Girardia tigrina*** and ***Polycelis nigra***, only have one and two reproductive system clusters, respectively (~ 2% of the total cells, Supplementary 8). These results are in line with our expectations, as in the lab these planarian cultures never - or very rarely- have laid cocoons. Both species appear to be **asexualized by laboratory conditions** and have no sexual organs. Therefore, the clusters annotated as reproductive system may simply correspond to truncated branches or **germ cell progenitors**. To profile the single-cell atlas of truly sexual individuals, the reproductive capacity of the animals should be carefully evaluated, as we did in the case of *Schmidtea polychroa*. However, many long-term lab cultures remain permanently asexualized. Collecting samples directly from nature is not always viable (given the location of the species) and involves many challenges, like collecting the right species, genotyping and establishing the lab culture. Finding a lab, or company, to provide a recently established culture would be the best option, although is not always available.

3. ASSESSMENT OF CLUSTER ABUNDANCES IN *S. MEDITERRANEA* AND *S. POLYCHROA*

The analysis of cluster abundances was performed by calculating the log ratio, which does not assess significance. The results obtained are **descriptive** but not statistically validated. Besides, although different experiments were merged together, **no batch correction** was applied. Nonetheless, these preliminary results suggest samples from the same species integrate very well with no need of batch correction. The single-cell atlases of *S. mediterranea* and *S. polychroa* are very cohesive, with samples homogenously distributed among most of the clusters (**Figure 4.8-9 A**). Major differences focus on the reproductive system, neoblasts and unknown populations likely related to the reproductive system, but this is to be expected when comparing sexual and asexual planarians. Within these clusters, we can highlight the presence of an almost fully-developed reproductive system in the *S. polychroa* juvenile sample.

Out of these groups, we find other interesting enrichments. In both species, **neuronal populations** were more abundant in non-sexual planarians (*S. mediterranea* asexual and *S. polychroa* hatchlings). These samples are smaller than sexual individuals, so the differences could be size related, as it has been shown that neuronal cells are enriched in smaller planarians (Baguña and Romero, 1981). In this publication, the authors also describe an increase of fixed parenchymal cells in larger animals. This is not observed in *S. mediterranea* but, interestingly, there is a small parenchymal population (cluster 54) enriched in *S. polychroa* larger samples (adult and juvenile).

Other intriguing cases in *S. polychroa* include cluster 12 (epidermal) and cluster 66 (unknown), which are practically absent in the adult sample, and the enrichment of **secretory populations** in hatchling and juvenile planarians. In *S.* mediterranea, the most notorious case is that of the secretory 8, which is exclusive to the asexual strain. In general, and without taking the reproductive system into account, the life-history stages of *S. polychroa* appear to be more diverse at the cellular level than the two strains of *S. mediterranea.*

The differences presented here may encompass novel biological insights that are worth exploring further. However, proper batch correction and statistical analysis will be required to confirm these assertions and discard technical artefacts.

4. FUTURE DIRECTIONS

For the continuation of this project, I will start optimizing the preprocessing of the raw matrices by testing different values of the following parameters: top highly variable genes, PCs, n-neighbours, and maximum number of genes and total counts. The goal will be obtaining the best possible resolution for each species. Later, **TransDecoder** (https://github.com/TransDecoder/) will be run to identify coding regions in our reference genomes and transcriptomes, and generate their predicted proteomes. These proteomes will be used to find gene orthologs between species with software like **OrthoFinder** (Emms and Kelly, 2015) or **Possvm** (Grau-Bové and Sebé-Pedrós, 2021). One-to-one orthologs will be used to homogenize the annotation of all species and integrate them into a single matrix.

Integration and batch-correction will be performed in Scanpy using **Scanorama** (Hie et al., 2019). The annotation of cell identities in the integrated dataset will be achieved by combining label transferring tools with the evaluation of published gene markers of *Schmidtea mediterranea*. In parallel, datasets will be integrated using the **SAMap** algorithm (Tarashansky et al., 2021), which combines orthologs and paralogs to find relationships between phylogenetically distant species. The results of both integrations will be compared. To re-evaluate cluster abundances on the integrated dataset, I will use **sccoDA** (Büttner et al., 2021), a Bayesian approach for differential cell-type composition analysis. Differential gene expression will be assessed at tissue resolution using **DEseq2** (Love et al., 2014), as in Chapter III, or other similar method. Finally, cell differentiation trajectories will be evaluated by separate for each species dataset using **PAGA** (Wolf et al., 2019).

CONCLUSION

In this chapter, I have presented the whole-body **single-cell atlases of five planarian species**: *Schmidtea mediterranea* (asexual strain) and *Dugesia japonica,* which had been previously profiled at single-cell resolution (Fincher et al., 2018; García-Castro et al., 2021; Plass et al., 2018), and *Schmidtea mediterranea* (sexual strain), *Schmidtea polychroa, Girardia tigrina* and *Polycelis nigra*, whose atlases are presented here for the first time. In addition, I have performed comparative **analyses on cluster abundances** between two *S. mediterranea* strains (sexual and asexual) and between three different life-history stages of *S. polychroa* (hatchlings, juveniles and adults). These analyses offer some preliminary insights into the development of the reproductive system in *S. polychroa*. They also reflect the uneven enrichment of certain populations, including neoblast, neuronal, secretory and epidermal cell types, between the different samples. However, these observations have to be properly validated by means of statistical analysis. In the future, this project aims to integrate all datasets presented here, and become a **cross-species single-cell evolutionary comparison** that explore the conservation -and divergence- of cell types and gene expression across the different planarian species.

FINAL DISCUSSION

The development, optimization and repurposing of experimental methods play a fundamental role in scientific progress. A large part of this thesis has been focused on these topics, with the study of new **methodologies for single-cell transcriptomics**. Indeed, the first goal of the thesis was to design and validate a customized pipeline for scRNA-seq.

This pipeline had to meet certain requirements for **sample preparation**, including the use of non-enzymatic dissociation, to avoid cellular stress, and the possibility of cell cryopreservation and enrichment. In Chapter II, I presented **ACME** as a novel protocol to prepare single-cell suspensions that provided simultaneous tissue dissociation, fixation and permeabilization. This protocol was based on a 19th century maceration technique used in microscopy. ACME generates samples as flexible as nuclei preparations, but preserving the whole cell information. ACME-cells can be easily stored, frozen and sorted by FACS, facilitating time-spaced sample collection. On the other hand, we needed a **scRNA-seq platform** that was affordable to run on a regular basis, profile several thousand cells per experiment, was easily scalable, had a flexible configuration to allow sample multiplexing, and required basic lab equipment. For this reason, throughout the thesis, I have been using and implementing optimizations on **SPLiT-seq** (Rosenberg et al., 2018), a previously developed *in situ* barcoding-based technology that met all these requirements. One of the key optimizations was the use of FACS after cell barcoding, which was presented in Chapter III.

The combination of ACME and SPLiT-seq allows to overcome many of the traditional limitations of scRNA-seq. Nowadays, this pipeline is been successfully **implemented in numerous single-cell transcriptomic projects**, carried out by our lab, collaborators and external research groups. In particular, ACME has been very helpful to dissociate diverse invertebrate organisms for which there was no dissociation protocols available, including annelid, mayfly, snail or ascidian species. Since ACME preserves the morphology of the cell, this protocol could have other potential applications beyond single-cell transcriptomics. On the other hand, thanks to SPLiT-seq, we have been able to generate datasets of thousands of cells, such as those presented in this thesis. Nowadays, *in situ* barcoding methods are still making their way through the scientific community, which has a general preference for more stablished microfluidic platforms, like 10x Genomics Chromium. However, microfluidic-based technologies are currently more expensive and less flexible. After working with SPLiT-seq, I believe *in situ* barcoding platforms will be invaluable tools for the future development of single-cell transcriptomics.

In relation to the biological insights presented in this thesis, **ACME-cells** have been shown suitable for scRNA-seq on both microfluidic and *in situ* barcoding platforms. ACME has also provided **better preservation** of certain cell types than other dissociation strategies. This claim is supported by the finding of a novel cluster, the **germ cell progenitors** (Wang et al., 2010, 2007), in different datasets along Chapters II, III and IV. In addition, I have presented the novel **whole-body cell atlases** of multiple planarian species, including *Schmidtea mediterranea* (sexual strain), *Dugesia japonica, Schmidtea polychroa, Girardia tigrina* and *Polycelis nigra.* This will open the door to studies on cell type evolution in different Platyhelminthes clades other than the canonical species *Schmidtea mediterranea*.

Moreover, in Chapter III, I characterized the **knockdown of *hnf4*** at tissue resolution. After treating planarians by microinjection of dsRNAs, a strong headless phenotype is unravelled after day 9 to 15. At cellular level, this knockdown has shown to strongly **disrupt phagocytes and phagocytic differentiation**, and deplete *pgrn*+ parenchymal cells. The *hnf4* knockdown also affects different cell types at gene level, but the large majority of transcriptional changes occur in phagocyte populations. This experiment has revealed a list of **putative targets of *hnf4*** in planarians. Many of them are homologs to well-described target of mammalian HNF4, suggesting a strong conservation of the *hnf4* transcriptional program between planarians and mammals.

As already discussed in each individual chapter, these results and the analyses performed to obtain them are not without limitations. In Chapter III, for instance, I was able to include two **replicates** per sample thanks to the flexibility of our pipeline. This could be considered insufficient, as more traditional techniques easily include three or more replicates. However, most single-cell experiments are performed without replicates because of the technical challenges they involve. In the same chapter, I comment on the limitations of traditional **statistical methods** to assess differential cluster abundances. Later, in Chapter IV, I highlight some of the challenges of **integrating technically diverse multi-species** datasets from different experimental batches. These examples are a reminder that our experiments are running at the cutting edge of current technical development, thus it is normal to experience some technical constraints.

Single-cell transcriptomics officially started in 2009, representing a revolution in biology. However, after more than a decade, this **technology is still in full development**. Modern microfluidic-based single-cell platforms were created from 2015 (Klein et al., 2015; Macosko et al., 2015; Zheng et al., 2017), while *in* situ barcoding technologies did not arrive until 2017 (Cao et al., 2017; Rosenberg et al., 2018). Regarding the analysis of single cell data, most of the tools

used or cited in this thesis, such as Scanpy (Wolf et al., 2018), PAGA (Wolf et al., 2019) or DoubletDecon (DePasquale et al., 2019), have also been developed very recently. At present, we count with multiple guidance for sample preparation, experimental configuration and bioinformatic analysis (Lafzi et al., 2018; Luecken and Theis, 2019; Nguyen et al., 2018). But there is no clear consensus, as the options available are enormous and all technologies have their own strengths and weaknesses.

Beyond current technical challenges, however, a horizon of new possibilities extends. Nowadays, single-cell transcriptomics is still living **the cell type atlas era**. In planarians, different resolution atlases have been published in the last years (Fincher et al., 2018; Plass et al., 2018; Wurtzel et al., 2015). These atlases have been useful to identify novel cell types, define cell differentiation trajectories and study gene expression at tissue resolution. But we are approaching the end of this era, which will foreseeably be surpassed by **more complex comparative and quantitative studies**. In those, experiments will have a larger number of cells, and single-cell datasets will be integrated and compared across multiple conditions, replicates and species. This will pose great technical challenges, but will also lead to major functional genomics and evolutionary insights.

The methods presented here are likely to contribute to this leap, as has been addressed throughout this thesis. Altogether, I have validated a new pipeline for scRNA-seq, developing ACME as a novel dissociation strategy and profiling the whole-body cell atlases of two planarian species at high resolution. In addition, I have performed a single-cell RNAi study on the planarian *hnf4*, revealing specific cellular and genetic effects in knockdown conditions. Finally, I have showed the preliminary results of a cross-species evolutionary comparison in planarian at single-cell level. With this, most of the **aims** proposed at the beginning of the thesis has been achieved, offering a good example of the evolution of single-cell transcriptomics and its still enormous potential.

CHAPTER V: METHODS

ANIMAL CULTURE

The asexual planarian species use in this thesis originate from the Berlin-1 clonal strain of *Schmidtea mediterranea* (Solana et al., 2016) and the SSP-9T-5 clonal strain of *Dugesia japonica* (Nishimura et al., 2015). Sexual strains of *Schmidtea mediterranea, Schmidtea polychroa* and *Girardia tigrina* were obtained from the laboratory cultures of Emili Saló and Teresa Adell at the University of Barcelona, while *Polycelis nigra* animals were purchased from Blades Biological Ltd (https://blades-bio.co.uk/). All species were maintained at Oxford Brookes University using 'Montjuïc' culture water (1.6 mM NaCl, 1.0 mM $CaCl_2$, 1.0 mM $MgSO_4$, 0.1 mM $MgCl_2$, 0.1 mM KCl and 1.2 mM NaHCO3) adjusted to pH 7.0. Animals were kept on plastic boxes at 18 or 20ºC, in the dark, and fed with raw live at least once every two weeks. Planarians selected for the experiments underwent starvation for 7-14 days.

ACME DISSOCIATION IN PLANARIAN

The ACME solution was prepared fresh using 6.5 mL of commercial nuclease-free ultrapure water, 1.5 mL of methanol, 1 mL of acetic acid and 1 mL of glycerol per reaction (13:3:2:2 ratio). For each reaction, 10-30 mixed size adult planarians were added to a 15 mL Falcon tube to a final biomass of ~100-200 µL. Culture water was completely removed and animals were soaked in 100-500 µL of freshly prepared 7.5% N-acetyl L-cysteine (NAC, Sigma A7250) diluted in 1x PBS. Samples were briefly incubated in NAC at RT (1 min). Then, 10 mL of ACME solution were added per reaction. Samples were incubated at RT on a see-saw shaker at 35-45 rpm.

Chapter II and **IV** samples were incubated in ACME for 1 hour. Afterwards, they were pipetted up and down several times to complete dissociation.

Chapter III samples were incubated for 35 min, and pipetted up and down to help dissociation. With the cells still in ACME, we filtered twice through 50 µm and 30 µm strainers (CellTrics) into new Falcon tubes, and centrifuged at 1000 g for 5 min (4ºC). We discarded 8-9 mL of supernatant, and resuspended the pellet in the remaining volume. This volume was filter through a 40 µm strainer for 1000 µL pipette tips (Flowmi) into a new Falcon tube.

From this point, cells were kept in cold conditions to prevent RNA degradation. We centrifuged samples at 1000 g for 5 min (4ºC) to remove ACME. Cells were resuspended in 7 mL of freshly prepared 1x PBS 1% BSA (Thermo Fisher, cat. BP9700100) buffer and centrifuged again. Supernatants were discarded and final pellets were resuspended in 900 µL of buffer and transferred to 1.5 mL Eppendorf tubes. For cryopreservation, we added 100 µL of DMSO per tube (10%) and stored cells directly at -80ºC. After thawing, samples were centrifuged twice at

1000 g for 5 min (4ºC) to remove DMSO and resuspended in 200 µL to 1 mL of fresh buffer, depending on the application.

ACME DISSOCIATION IN OTHER MODEL ORGANISMS

ACME dissociation in other animals was performed with modifications to the above protocol. Snail and spider embryos were dissociated using a 14:3:1:2 ratio of nuclease-free water, methanol, acetic acid and glycerol. Except for *Nematostella vectensis*, the 7.5% NAC dilution was mixed directly with ACME, adding 50-500 µL per sample. Zebrafish embryos were dechorionated before dissociation, and mechanically disrupted in ACME using a Polytron homogenizer. Snail embryos were decapsulated by passing them through a syringe, and then mechanically disrupted in ACME using Polytron. To remove broken eggshells, dissociated mixes from zebrafish and snail were passed through a 100 µm CellTrics filter (Sysmex). Spider egg capsules were mechanically disrupted in ACME using short pulses (30 sec) of Polytron, and then filtered through a 40 µm cell strainer (Corning). Spider embryos were then incubated for 1h at RT with see-saw agitation. *Pristina leidyi* adults (100-120 animals/sample) were incubated for 30 min at RT, and manually shaken every 10 minutes to help dissociation. *Nematostella vectensis* juveniles (10 animals/sample) were incubated in 4 mL of ACME, for 1 hour at RT, on a GentlMACS dissociator (Miltenyi Biotech, 130-093-237). *Nematostella* cells were centrifuged at 1500 g and resuspended in 1x PBS 0.5% BSA buffer with RNases inhibitor (40U/mL).

TRYPSIN DISSOCIATION

Between 50-70 mixed size adult planarians (*S. mediterranea*) were chopped into small pieces in a petri dish using a sterile razor blade. Chopped worms were transferred to a 15 mL Falcon tube containing 10 mL of 1% trypsin diluted in 1x PBS. Tissues were incubated for 30 min at room temperature in a see-saw shaker (35-45 rpm). The reaction was pipetted up and down every 10 minutes to help dissociation. After incubation, we diluted trypsin adding 4 mL of fresh buffer (1x PBS 1% BSA) and centrifuged the cells at 1000 g for 5 min (4ºC). Cells were resuspended in 10 mL of buffer, passed through a 50 µm CellTrics filter (Sysmex) and through a 20 µm nylon net filter (Millipore) into a new 15 mL Falcon tube. Cells were centrifuged at 1000 g for 5 min (4ºC). The supernatant was discarded and the pellet was resuspended in 1-2 mL of buffer and transferred to a 2.0 mL Eppendorf tube. Trypsinized cells were fixed, using ACME or formaldehyde, or directly stained for flow cytometry visualisation.

ACME FIXATION

We transferred 300-600 µL of trypsinized cells to a 15 mL Falcon tube, and added 8.5 mL of modified ACME formula, without methanol: 6.5 mL of buffer (1x PBS 1% BSA), 1 mL of glycerol, 1 mL of acetic acid and 100 µL of 7.5% NAC diluted in 1x PBS. We used PBS buffer instead of water to avoid an osmotic shock to dissociated cells. We incubated cells for 15-20 min at RT in a see-saw shaker (35-45 rpm). Subsequently, we added 1.5 mL of methanol and incubated for another 15-20 min. Methanol was added later to allow partial fixation in acetic acid prior to permeabilization. After incubation, cells were washed and resuspended as described for ACME dissociation.

FORMALDEHYDE FIXATION

We started from 100-200 µL of trypsinized cells per sample, and resuspended them in 8 mL of buffer (1x PBS 1% BSA) in a 15 mL Falcon tube. We added 2 mL of formaldehyde (FA) diluted in 1x PBS to a final concentration of 0.1%, 0.5%, 1%, 2% or 4%. Immediately, tubes were gently shaken to homogenize the concentration of FA. Samples were incubated for 10 min at 4ºC in a see-saw shaker (35-45 rpm). Then, we centrifuged twice at 1000 g for 5 min (4ºC) to remove FA. Fixed cells were resuspended in 5-7 mL of buffer after the first centrifugation, and in 1 mL of Trizol after the second (for RNA extraction).

ASSESSMENT OF RNA QUALITY

All RNA extractions were performed using Trizol, or Trizol LS, following the manufacturer's protocol. RNA quality was assessed using an Agilent 2100 Bioanalyzer, according to the Agilent RNA 6000 Nano Kit Guide. RIN values were obtained directly from the bioanalyzer or inferred using a linear regression. This linear regression was created by correlating the percentage of RNA in the ribosomal peaks (values provided by the bioanalyzer) with the RIN in planarian samples resolved by the algorithm. As control samples, we used RNA extractions from live worms directly on Trizol.

CELL STAINING

Prior to flow cytometry and FACS, ACME-cells were resuspended in 1 mL of fresh buffer (1x PBS 1% BSA) and filtered through a 50 µm CellTrics strainer (Sysmex). Most cells were stained with 0.5-1µL/mL of DRAQ5 (5 mM stock, eBioscience) and 2 µL/mL of Concanavalin-A conjugated with

AlexaFluor 488 (1 mg/mL stock, Invitrogen). Unfixed trypsin-dissociated cells were stained using 10 µL/mL of DRAQ5 (5 mM stock) and 1 µL/mL of Calcein (0.5 mg/mL stock, eBioscience). *Nematostella vectensis* cells were stained with 0.33 µL/mL of DRAQ5 (5 mM stock). All samples were incubated in the dark for 20-40 min (RT or ice).

FLOW CYTOMETRY AND FACS

For cell visualisation and counting, we used a CytoFlex S Flow Cytometer (Beckman Coulter). For cell sorting, we used a BD FACS Aria III Cell Sorter (BD Biosciences) and the BD FACSDiva Software, setup in 4-Way Purity mode, with an 85 µm nozzle and moderate-pressure separation (45 Psi). The gating strategy to select DRAQ5+ and ConA+ singlets was the following: (1) FSC-Height vs FSC-Area, to select events with well-correlated sizes (singlets), (2) ConA-Area vs FSC-Area, to select ConA+ events, (3) DRAQ5-Area vs FSC-Area, to select DRAQ+ events, and (4) DRAQ5-Height vs DRAQ5-Area, to select singlets. We used a red laser (780/60 nm filter) for DRAQ5 and a yellow-green laser (525/40 nm filter) for ConA. Gated events were sorted and collected in 1.5 mL Eppendorf tubes.

For **Chapter II** planarian samples, the FACS was decontaminated with bleach and precooled to avoid RNases contamination. The injection and collection chambers were kept at 4ºC. Cells were sorted in 100 µL of collection buffer (1x PBS 1% BSA), up to 500,000 cells per tube. To complete a sorting run took 3-6 hours. After sorting, samples were centrifuged at 1000 g for 5 min (4ºC), and resuspended in 900 µL of fresh buffer. At this point, cells were cryopreserved adding 100 µL of DMSO, and stored at -80ºC.

For **Chapter II** *N. vectensis* samples, a total of 6,250 G1 singlets were sorted into Master Mix without RT enzyme C (from the 10x Chromium single cell 3' reagents kit v3.1), using 1-Drop purity mode, a 100 µm nozzle and 20 Psi.

Chapter III and **IV** samples were sorted at room temperature, directly in 50 µL of *Lysis buffer* (SPLiT-seq protocol). We collected 10,000-25,000 cells per tube. Sorting took 1,5 hours.

IRRADIATION

We irradiated three petri dishes with 20 planarians each (*S. mediterranea*) at 60 Gy in a Gamma (Cs-137) Cabinet Irradiator. As negative controls, we used three equivalent non-irradiated plates. Irradiated samples and controls were dissociated 72 hours post-treatment as described (ACME dissociation). ACME-cells were resuspended in 1 mL of buffer (1x PBS 1% BSA), passed through a 50 µm CellTrics filter (Sysmex), and stained with 2 µL/mL of DRAQ5 (5 mM stock) and

2 µL/mL of ConA (1 mg/mL stock) for 40 min at RT. Then, samples were profiled by flow cytometry to obtains the percentage of G2 cells.

10x CHROMIUM SINGLE-CELL TRANSCRIPTOMICS

The *Nematostella vectensis* library was prepared using a Chromium Next GEM Single Cell 3´Reagent Kit (v3.1), and loaded onto a Chromium Next GEM Chip according to the manufacturer's protocol. The cDNA was amplified for 12 cycles and indexed by PCR for another 16 cycles. Library was sequenced on a HiSeq 2500 (50 bp paired end). The bioinformatic analysis was performed using MetaCell (Baran et al., 2019) as described in **García-Castro *et al.*, 2021**.

SPLiT-SEQ

SPLiT-seq was performed as previously described (Rosenberg et al., 2018), with modifications. *All oligo sequences used are provided in Supplementary 1.

1. Plates preparation

Barcodes were provided lyophilized by Integrated DNA technologies on 96-well Stock-Plates: Stock-1 (well-specific anchored poly(dT) **Round 1 barcodes*), Stock-2 (well-specific **Round 2 barcodes*) and Stock-3 (well-specific **Round 3 barcodes*). Lyophilized barcodes were resuspended in nuclease-free water to a final concentration of 100 µM/well. From these stocks, we prepared three working dilution 96-well plates (WD-1, WD-2 and WD-3). WD-1 was prepared with 12 µL of Stock-1 and 88 µL of nuclease-free water per well. WD-2 with 12 µL of Stock-2, 11 µL of **Linker_1* (100 µM) and 77 µL of nuclease-free water per well. WD-3 with 14 µL of Stock-3, 13 µL of **Linker_2* (100 µM) and 73 µL of nuclease-free water per well. WD-2 and WD-3 were heated to 95ºC for 2 min and ramped down to 20ºC at a rate of -0.1ºC/s, to anneal the 5' end of each barcode to the universal linker oligos.

2. Flow cytometry and sample dilution

After thawing, ACME-cells were resuspended in 200-400 µL of fresh buffer (1x PBS 1% BSA) and filtered through a 50 µm CellTrics strainer (Sysmex). We diluted 50-100 µL per sample in buffer at 1:3 or 1:10. Dilutions were stained with 0.15 µL of DRAQ5 and 0.6 µL of ConA, for 20 min at RT, and counted 3 times by flow cytometry to quantify the concentration of total events and cells (singlets). According to these calculations, the original (undiluted) samples were diluted to their final working concentrations.

In **Chapter II**, the *D. japonica* and *S. mediterranea* sorted samples were mixed 1:1 and diluted in 0.5x PBS to a final concentration of 1,250 cells/µL. The first barcoding plate was loaded with 10,000 cells/well of this mix (5,000 per species). In **Chapter III** and **IV**, the unsorted samples were diluted in 0.5x PBS to a concentration of 625 events/µL, and first barcoding plates were loaded with 5,000 events/well.

3. Round 1 of barcoding: Reverse transcription

We prepared a new 96-well plate with 4 µL/well (**Chapter II**) or 8 µL/well (**Chapter III** and **IV**) of **Round 1 barcodes* from WD-1. Then, we added 8 µL/well of RT mix: 4 µL of 5x Maxima H Minus RT Buffer (Thermo Scientific), 0.375 µL of SUPERase-In RNase Inhibitor (20 U/µL, Invitrogen), 1 µL of 10 mM/each dNTPs (NEB), 1.65 µL (**Chapter II**) or 0.65 µL (**Chapter III** and **IV**) of nuclease-free water, and 1 µL (**Chapter II**) or 2 µL (**Chapter III** and **IV**) of Maxima H Minus RT (200 U/µL, Thermo Scientific). Finally, we added 8 µL/well of previously diluted cells. The plate was incubated in a thermocycler for 35 min at 50ºC. Individual reactions were then pooled in a 15 mL Falcon tube, on ice. We added 10% Triton X-100 to the pooled cells (0.1% final concentration) and centrifuged them at 1200 g for 5 min (4ºC). We discarded the supernatant and resuspended the pellet in 2 mL of 1x NEB buffer 3.1 (NEB) with 20 µL of SUPERase-In RNase inhibitor.

4. Round 2 of barcoding: Ligation 1

A new 96-well plate was prepared with 10 µL/well of **Round 2 barcodes* from WD-2. Then, 2.04 mL of ligation mix (500 µL of T4 Ligase Buffer 10x (NEB), 100 µL of T4 DNA Ligase (400 U/µL, NEB), 100 µL of 1x PBS 1% BSA buffer and 1,340 µL of nuclease-free water) were added to the cells in 1x NEB buffer 3.1, and mixed thoroughly into a disposable basin. We added 40 µL/well of cells and ligation mix to the plate, and covered it with adhesive PCR plate seal. The plate was incubated in a thermocycler for 30 min at 37ºC. To block **Linker_1* after incubation, 10 µL/well of blocking solution (264 µL of **Blocker_1* (26.4 µM final concentration), 250 µL of T4 Ligase Buffer 10x and 486 µL of nuclease-free water) were added to the plate and incubated for another 30 min at 37ºC.

5. Round 3 of barcoding: Ligation 2

After blocking, cells were pooled into a disposable basin. We added 100 µL (**Chapter II**) or 150 µL (**Chapter III** and **IV**) of T4 DNA Ligase (400 U/µL, NEB) and mixed thoroughly with the cells. A new 96-well plate was prepared with 10 µL/well of **Round 3 Barcodes* from WD-3. Plate was loaded with 50-55 µL/well of cells, sealed, and incubated for 30 min at 37ºC. After ligation, 20 µL/well of termination solution (288 µL of **Blocker_2* (11.5 µM final concentration), 625 µL of

0.5M EDTA and 1,587 µL of nuclease-free water) were added without further incubation. Cells were pooled into a 15 mL Falcon tube, on ice. We added 10% Triton-X 100 (0.1% final concentration) and centrifuged at 1200 g for 5 min (4ºC). The supernatant was discarded, and cells were resuspended in 4.04 mL of washing buffer (4000 µL of 1x PBS and 40 µL of 10% Triton X-100). Cells were centrifuged again at 1200 g for 5 min (4ºC), and further processed according to the following sections.

6. Cell lysis (Chapter II)

After washing, cells were resuspended in 50 µL of 1x PBS buffer. Of these, 5 µL were diluted in 195 µL of 1x PBS (1:40), and counted by flow cytometry to decide the number of sub-libraries. We divided the remaining 45 µL in three sub-libraries of ~13,000 cells/each. The volume of each sub-library was adjusted to 50 µL with 1x PBS. We added 50 µL of lysis buffer (20 mM Tris pH 8.0, 400 mM NaCl, 100 mM EDTA, 4.4% SDS) and 10 µL of Proteinase K (20 mg/mL) per sub-library and incubated the lysates at 55ºC for 2 hours, in agitation. After incubations, lysates were frozen at -80ºC.

7. FACS (Chapters III and IV)

After washing, cells were resuspended in 800 µL of buffer (1x PBS 1% BSA) and split in two 1.5 mL Eppendorf tubes (400 µL/each). We added 44 µL of DMSO per tube and stored cells at -80ºC. For sorting, samples were thawed and centrifuged twice at 1200 g for 5 min (4º), adding 10% Triton X-100 (0.1% final concentration) to help precipitation. Cells were resuspended in 400-500 µL of buffer and stained with 0.5 µL of DRAQ5 and 1.0 µL of ConA. FACS sorting was performed as described in previous sections. Samples were sorted in 50 µL of lysis buffer (20 mM Tris pH 8.0, 400 mM NaCl, 100 mM EDTA, 4.4% SDS) up to 10,000-25,000 cells/tube. After sorting, volumes were adjusted to 100 µL when necessary. We added 10 µL of Proteinase K per library and incubated for 2 hours at 55 ºC, in agitation. After incubations, lysates were frozen at -80ºC.

8. cDNA purification

We used magnetic Dynabeads MyOne Streptavidin C1 (Invitrogen) to purify the cDNA in the lysates. Streptavidin in beads binds to the biotin molecule at the 3' end of the third barcode. We followed the manufacturer's protocol for Nucleic Acid Purification, with the modifications included in **Rosenberg *et al*., 2018**.

9. Template Switch

The template switch mix was prepared with 44 µL of 5x RT Buffer, 44 µL of 20% Ficoll PM 400 (Sigma Aldrich), 22 µL of 10 mM/each dNTPs, 5.5 µL of **TSO* (100 µM), 5.5 µL of SUPERase-In RNase inhibitor, 11 µL of Maxima H Minus RT (200 U/µL) and 88 µL of nuclease-free water per sample. Using a magnetic rack, the Dynabeads binding cDNAs were washed with 250 µL of nuclease-free water (no resuspension) and resuspended in 200 µL of template switch mix. Samples were incubated for 30 min at room temperature and then for 90 min at 42ºC, with agitation. After incubation, the mix was removed using a magnetic rack, and beads were resuspended in 250 µL of Tris-T buffer (10 mM Tris pH 8.0, 0.1% Tween-20 and 0.2% SUPERase-In RNase inhibitor) and kept at 4ºC.

10. PCR amplification

Dynabeads in Tris-T buffer were placed in a magnetic rack, washed (250 µL of water), and resuspended in 220 µL of PCR mix: 110 µL of 2x KAPA HiFi HotStart ReadyMix (Roche), 8.8 µL of **PCR_PF* (10 µM), 8.8 µL of **PCR_PR* (10 µM) and 92.4 µL of nuclease-free water. Each sample was split in 4 PCR reactions (55 µL/each) and amplified using the following program: 95ºC (3 min), and five cycles at 98ºC (20 s), 65ºC (45 s) and 72ºC (3 min). Afterwards, the 4 PCR reactions were combined, and Dynabeads were held using a magnetic rack. We collected the supernatant, containing amplified cDNA in suspension, and split it into 4 wells of a qPCR plate (50 µL/well). We added 2.5 µL of 20x EvaGreen (Biotium) per well and ran the following program in a qPCR thermocycler: 95ºC (3 min), and 9-11 cycles at 98ºC (20 s), 65ºC (20 s) and 72ºC (3 min).

11. Size selection

Amplified qPCR reactions were pooled and purified by SPRI size selection to remove fragments smaller than 300 bp. For this, we used KAPA Pure Beads (Roche) according to the manufacturer's protocol, with two modifications from **Rosenberg *et al*., 2018**: washing steps were performed with 750 µL of 85% ethanol, and cDNA was eluted in 20 µL of nuclease-free water at 37ºC for 10 min. Samples were purified using a 0.8x (**Chapter II**) or 0.7x ratio (**Chapter III** and **IV**) of KAPA Pure Beads. After size selection, libraries were quantified running a dsDNA High sensitivity Qubit assay (Thermo Fisher). Fragments distribution was checked in an Agilent 2100 Bioanalyzer following the Agilent High Sensitivity DNA Kit Guide.

12. Tagmentation

Chapter II libraries were tagmented using the Nextera DNA Library Preparation Kit (Illumina, discontinued). The tagmentation reaction was prepared mixing 50 ng of cDNA (diluted in 20 µL of nuclease-free water), 25 µL of Tagmentation Buffer and 5 µL of Enzyme 1. Samples were incubated in a pre-heated thermocycler for 5 min at 55ºC. We neutralized the reaction by immediately cleaning with the Monarch PCR & DNA Cleanup Kit (NEB). Samples were eluted in a final volume of 20 µL of nuclease-free water.

Chapter III and **IV** libraries were tagmented using the Nextera XT DNA library Preparation Kit (Illumina). We diluted 1 ng of cDNA in 5 µL of nuclease-free water, and mixed it with 10 µL of Tagmentation Buffer (TD) and 5 µL of Tagmentation Enzyme (ATM). Samples were incubated in a pre-heated thermocycler for 5 min at 55ºC. Immediately after, we neutralized the reaction with 5 µL of Neutralizing buffer (NT), incubating for 5 min at RT.

13. Round 4 of barcoding: PCR amplification

For **Chapter II**, we prepared the following PCR mix per sample: 20 µL of tagmented cDNA, 25 µL of 2x KAPA HiFi HotStart ReadyMix, 1.5 µL of **P5_oligo* (10 µM), 1.5 µL of **Round 4 Barcode_X* (10 µM) and 2.5 µL of 20x EvaGreen. Then, a qPCR amplification was run as follows: 95ºC (30 s), and 8-10 cycles (until plateau) at 95ºC (10 s), 55ºC (30 s) and 72ºC (30 s). In **Chapter III** and **IV**, PCR reactions were prepared using 20 µL of tagmented cDNA, 15 µL of Nextera XT PCR mix, 1 µL of **P5_oligo* (10 µM) and 1 µL of **Round 4 Barcode_X* (10 µM). Samples were amplified in a PCR thermocycler: 72 ºC (3 min); 95 ºC (30 s); 12 cycles at 95 ºC (10 s), 55 ºC (30 s) and 72 ºC (30 s); and 72 ºC (5 min). In all cases, different Round 4 Barcodes were used for each sub-library.

14. Final size selection and quality assessment

After amplification, samples were size selected using a 0.7x (**Chapter II**) or 0.6x ratio (**Chapter III** and **IV**) of KAPA Pure Beads. Final libraries were quantified by Qubit, and fragment distribution was checked in an Agilent 2100 Bioanalyzer.

RNA INTERFERENCE

1. Input material

For ***hnf4***, we extracted RNA from fresh wild type *S. mediterranea* worms using Trizol. For reverse transcription, we mixed 1 µg of RNA, 1 µL of 50 µM oligo(dT)s, 1 µL of 10 mM dNTPs and up to 15.5 µL of nuclease-free water, and incubated for 5 min at 65ºC. Then, we added 4 µL of 5x RT buffer and 0.5 µL of Maxima H Minus RT (200 U/µL). We incubated for 30 min at 50ºC, and for 10 min at 85ºC. Final cDNA was diluted at 1:5 in nuclease-free water. For **GFP**, we used a DNA miniprep of EGFP amplified in a pAGW vector. The vector was provided by the Drosophila Genomics Resource Center (https://dgrc.bio.indiana.edu/Repository/1071.txt ?file=140).

2. Primary PCR

The primary PCR mix was prepared with 2 µL of cDNA, 2 µL of 10x Standard *Taq* Reaction Buffer (NEB), 0.4 µL of dNTPs (2.5 µM), 0.2 µL of Hot Start *Taq* DNA Polymerase (NEB), 4 µL of Primer Forward (2.5 µM), 4 µL of Primer Reverse (2.5 µM) and 7.4 µL of water. The primers being: ggccgcggCGCTGAAATAGCCAGTCACA (*hnf4*-F), gccccggccGCCGCTTCAGGTGATATGTT (*hnf4*-R), ggccgcggGTCTATATCATGGCCGACAAG (GFP-F) and gccccggccACTGGGTGCTCAGGTAGTGGT (GFP-R). *Hnf4* primers were designed for the GenBank sequence JF802199.1. Both primer pairs included linkers for Universal T7 primers: ggccgcgg (linker-F) and gccccggcc (linker-R). The reactions were amplified in a thermocycler as follows: 94ºC (30 s); 35 cycles at 94ºC (20 s), 55 ºC (20 s) and 68ºC (30 s); and 68ºC (5 min). We ran the PCR products in a 1% agarose electrophoresis gel and cut the bands under UV light. Gel bands were placed in Eppendorf tubes with 50 µL of nuclease-free water, and frozen at -20ºC.

3. Secondary PCR

Gel bands were thawed and centrifuged 1 minute at maximum speed. The supernatants were collected and used as input cDNA for the secondary PCR. We prepared 100 µL reactions with 3 µL of cDNA, 2 µL of dNTPs (2.5 µM), 10 µL of 10x Standard *Taq* Reaction Buffer, 1 µL of Hot Start *Taq* DNA Polymerase, 82 µL of water, 1 µL of Universal T7-F5' primer (25 µM, gagaattctaatacgactcactatagggccgcgg) and 1 µL of Universal T7-R3' primer (25 µM, agggatcctaatacgactcactataggccccggc). We used the same pair of primers for *hnf4* and GFP. The reactions ran in a thermocycler as follows: 94 ºC (30 s); 5 cycles at 94 ºC (20 s), 50 ºC (20 s) and 68ºC (30 s); then 35 cycles at 94ºC (20 s), 65ºC (20 s) and 68ºC (30 s); and 68ºC (5 min). The size of the bands was checked running 5-10 µL/sample in 1% agarose gel. The remaining volume was

purified by SPRI size selection using a 0.75x (*hnf4*) or 1.6x ratio (GFP) of KAPA Pure Beads according to the manufacturer's protocol. Samples were eluted in 20 µL of nuclease-free water.

4. dsRNA synthesis

For each reaction, we mixed 1 µg of purified cDNA, 12.5 µL of 2x Express Buffer (T7 RiboMAX, Promega), 2.5 µL of Express Mix (T7 RiboMAX, Promega), and up to 25 µL of nuclease-free water. We incubated for 4 hours at 37ºC. Then, we added 2.5 µL of DNase (1 U/µL, T7 RiboMAX, Promega) and incubated for 30 min at 37ºC. Finally, we added 375 µL of Stop Solution (1M NH_4OAc, 10 mM EDTA, 0.2% SDS). The resulting dsRNA was purified using Phenol:Chloroform. We added 1 µL of GlycoBlue (to improve pellet visualisation) and 400 µL of Acid-Phenol:Chloroform (pH 4.5, Thermo Fisher) per reaction and vortexed thoroughly. We centrifuged for 5 min and transferred the aqueous top phases to new tubes. We added 400 µL of chloroform, centrifuged for 5 min, and collected the top phases again. To precipitate pellets, we added 1 mL of cold ethanol, vortexed and centrifuged for 15 min. Pellets were washed in 1 mL of 70% ethanol and centrifuged for 10 min. We discarded the supernatants and let the pellets dry for 5 min at 37ºC. Afterwards, pellets were resuspended in 10-20 µL of nuclease-free water. All centrifugations were performed at 4ºC and maximum speed. As quality check, we ran 0.5 µL of purified dsRNA in a 1% agarose gel. Finally, we measured the concentration in a Nanodrop.

5. Injections and phenotyping

All dsRNA preparations were diluted to a working concentration of 1 µg/µL. For injections, we used *Schmidtea mediterranea* worms of approximately 6-8 mm in length. We injected 50 animals (25 per replicate) with *hnf4* dsRNA and 50 animals (25 per replicate) with GFP dsRNA, using a Nanoject II Auto-Nanoliter Injector (Drummond Scientific Company). Each animal was treated with 0.1 µg of dsRNA for 3 consecutive days (0.3 µg in total). Phenotypes were monitored by observing the uncut animals from days 9 to 15 after the last round of injections.

RNA EXTRACTION FOR ISO-SEQ RNA-SEQUENCING

For each species, one or two live planarians were collected in a 2 mL Eppendorf tube. Culture water was removed and RNA extractions were performed using Trizol according to the standard manufacturer's protocol. To help dissociation, planarians were mechanically disrupted on Trizol using tissue grinding pestles. Samples were preserved on ice during the extraction and frozen at -80ºC immediately after. RNA quality was assessed from an aliquot of each sample, using an Agilent 2100 Bioanalyzer according to the Agilent RNA 6000 Nano Kit Guide.

BIOINFORMATIC ANALYSIS

This section does not include the analysis of *N. vectensis* data. An extended version of all the analyses mentioned in Chapter II can be found in **García-Castro *et al*., 2021**.

1. PREPROCESSING

1.1 Quality control (Linux)

All the sub-libraries from the same experiment -or SPLiT-seq batch- were pooled together and sequenced on a NovaSeq 6000 platform by Novogene (https://www.novogene.com), obtaining 150 bp length paired-end reads. The quality of the raw reads was assessed with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Then, CutAdapt v2.8 (Martin, 2011) was used to trim universal adaptors, low-quality sequences, terminal Ns and short reads. We discarded Read 1 sequences (transcripts) shorter than 60 bp and Read 2 sequences (UMI and barcodes) shorter than 94 bp, the minimum to span all barcodes. Read 2 sequences were also checked for 'phase' using *grep*, and discarded when barcodes were not in the correct position. Makerpairs (https://github.com/sestaton/Pairfq/wiki/makepairs) was used to select only paired reads, and FastQC was ran again to confirm the removal of low-quality sequences.

1.2 Reference files (Linux)

For *Schmidtea mediterranea,* the S2F2 genome (Grohme et al., 2018) used in **Chapter II**, and the S2 genome (Guo et al., 2022) used in **Chapters III** and **IV**, were downloaded from PlanMine (https://planmine.mpinat.mpg.de/planmine/begin.do). *Dugesia japonica* v1.0 genome (An et al., 2018) was obtained from http://www.planarian.jp. *De novo* annotations were created for both species using public RNA-seq datasets from the NCBI Sequence Read Archive and the DNA Data Bank of Japan. These datasets were aligned to the reference genomes with HiSat 2.1.0 (Kim et al., 2019) and merged to existing gene models (SMESG- for *S. mediterranea,* and v1 Augustus-derived for *D. japonica*) using StringTie (Pertea et al., 2016). In **Chapter II**, the *fasta* and *gtf* files of these species were concatenated and analysed as a combined dataset.

For *Schmidtea polychroa, Girardia tigrina* and *Polycelis nigra,* we generated new transcriptome sequences using IsoSeq on a PacBio Sequel II platform. Transcriptomes were assembled *de novo* using Trinity (Grabherr et al., 2011), and BUSCO (Simão et al., 2015) was ran to assessed the completeness of the assembly. Transcriptomes were annotated using BLAST against the Swissprot protein set (https://www.uniprot.org/).

For all planarian species, Drop-seq_tools 2.3.0 (https://github.com/broadinstitute/Drop-seq) was used to create the sequence dictionaries, refFlat, reduced GTF and interval files; and STAR 2.7.3a (Dobin et al., 2013) to generate the indexes.

1.3 Demultiplexing, mapping and matrix generation (Linux)

All sub-libraries were processed by separate using the SPLiT-seq toolbox (https://github.com/RebekkaWegmann/splitseq_toolbox) to extract, tag and correct UMIs and barcodes (with hamming distance ≤ 1), remove poor quality reads, and trim poly-A tails from Read 1. After this, mapping was performed with STAR-2.7.3a (Dobin et al., 2013), and Picard v2.21.1 (http://broadinstitute.github.io/picard/) was used to re-order and merged the aligned reads. Drop-seq_tools 2.3.0 was run afterwards to tag reads with interval files and create the digital expression matrices (including intronic regions).

In **Chapter II**, reads mapping to *D. japonica* and *S. mediterranea* were used to create a Barnyard plot for species collision. Collisions were defined as cells sharing over 10% of their UMIs with the minority species. To assay library saturation, final reads were down-sampled randomly using Seqtk (https://github.com/lh3/seqtk) to 10%, 25%, 50% and 75% of total read depth. From these, we generated subsampled matrices and identified their number of UMI and gene per cell, which were used to obtain the saturation plots.

1.4 Matrix preprocessing (Seurat & Scanpy)

In **Chapter II**, the digital expression matrices were loaded into Seurat v3.1.0 (Stuart et al., 2019), in R, and a Seurat object was created for each sub-library with a threshold of 125 minimum genes per cell. Cells with more than 5000 UMI counts were excluded. Data was normalized and scaled in Seurat using default parameters, and the top 10,000 variable features (genes) were selected. Sub-libraries were merged and data was normalized again. Doublets were removed using DoubletDecon v1.15 (DePasquale et al., 2019). For dimensionality reduction, a Principal Component Analysis (PCA) was run with 50 PCs.

In **Chapters III** and **IV**, the digital expression matrices were converted into Seurat objects using Seurat v3.1.0, with no thresholds applied. Sub-libraries were normalizing separately, and then merged by experiment (**Chapter III**) or species (**Chapter IV**). Write10xCounts was used to convert the matrices to 10x format and transfer them to Scanpy (Wolf et al., 2018) for further processing. In Scanpy, biological samples from the same SPLiT-seq batch were separated according to their Round 1 barcodes. Then, cell with less than 100 genes (low informative), and genes with more than 1M counts (very highly expressed) were filtered out.

In **Chapter III**, the matrix was further sliced to removed cells with an outlier number of genes (>900) or total counts (>1800). Data was normalized to 10,000 reads per cell and transformed to logarithmic scale. The top 12,000 highly variable genes were selected, and data was regressed out and scaled. A PCA was performed using 150 PCs, and a K-nearest neighbour (kNN) graph was calculated with 95 PCs and 75 n-neighbours. Finally, UMAP was used for 2D data projection (min_dist = 0.5, spread = 1, alpha = 1, gamma = 1.0).

In **Chapter IV,** outliers were removed from the matrices applying different filters depending on the species. Cells were filtered out over 1200 genes or 1700 total counts for *S. mediterranea* and *S. polychroa,* 800 genes or 1500 total counts for *D. japonica*, 800 genes or 1300 total counts for *P. nigra,* and 1000 genes or 1500 total counts for *G. tigrina.* Afterwards, all species were normalized to 10,000 reads per cell, subset for the top 10,000 highly variable genes and scaled. PCAs and K-nearest neighbours were performed using different parameters: 90 PCs and 40 k (*S. mediterranea*), 60 PCs and 40 k (*S. polychroa*), 40 PCs and 20 k (*D. japonica*), 50 PCs and 35 k (*G. tigrina*), or 50 PCs and 30 k (*P. nigra*). UMAP was used for final data projections (min_dist = 0.4, spread = 1).

1.5 Reanalysis of the Plass *et al.* dataset (Seurat)

The original dataset from **Plass *et al.*, 2018** was downloaded from the NCBI GEO (GSE103633) and re-analysed using our novel *S. mediterranea* annotation. Data was pre-processed with Drop-seq tools 2.3.0 using the Drop-seq Core Computational Protocol v2.0.0. To construct the matrices, we used the barcode list from the original study. Cells with more than 2500 genes were excluded. Seurat v 3.1.0 was run for further processing, as described in section 1.4 (Chapter II).

2. DATA ANALYSIS

2.1 Clustering, gene markers and cluster annotation

Chapter II (Seurat):

Clustering was performed using the Louvain algorithm at 1.6 resolution -or 'granularity'-. Clusters were colour as indicated in Supplementary 2, and visualized by UMAP using the following settings: dims = 1:50, reduction = 'pca', spread = 1, metric = 'euclidean', seed.use = 1, n.neighbors = 45, min.dist = 0. Gene markers were extracted using FindAllMarkers. Cluster frequencies (Supplementary 2) were calculated using the Idents function, and transferred to a stacked bar plot to compare cell abundances (**Figure 2.14**). The expression of individual gene

markers was presented in FeaturePlots, and used for cluster identification (Supplementary 3). Violin plots were generated with the VlnPlot function.

To annotate *S. mediterranea*, we cross-referenced our new cluster markers with those from the previous publication (Plass et al., 2018). Novel clusters *eye-53*+ neurons, serotonin neurons, protonephridia tubule cells and protonephridia flame cells, were named after the expression of markers *eye-53-1* (Collins et al., 2010), *sert* (Currie and Pearson, 2013; März et al., 2013), *CAVII-like* (Scimone et al., 2011) and *egfr-5* (Barberán et al., 2016b), respectively. To annotate *D. japonica*, we found the *S. mediterranea* homologues for the top markers, and examined their expression in both species using feature plots. Finally, to stablish homologies between our novel gene models and published gene sequences, we used blastn megablast, Standalone BLAST, and tblastn. Secretory clusters were arbitrarily named 1-7 (*S. mediterranea*) and a-h (*D. japonica*) according to their abundances, as the homology between species, or with other published data, requires further research.

Chapter III (Scanpy):

Clustering was performed using the Leiden algorithm at resolution 2.5. The top 10 markers per cluster were extracted using the Wilcoxon method. Cells were first annotated using the Ingest function to transfer labels from a reference *S. mediterranea* dataset with 103,654 cells (not shown). Ingest annotations were combined to the Leiden clustering, and each cluster was assigned to its majoritarian Ingest identity. Annotated clusters were classified in groups, and coloured as indicated in Supplementary 5. Violin plots, feature plots and dotplots were generated using sc.pl.violin, sc.pl.umap and sc.pl.dotplot, respectively.

Chapter IV (Scanpy):

Clustering was performed using the Leiden algorithm at resolution 3.0 (*Schmidtea mediterranea* and *Schmidtea polychroa*), 2.0 (*Dugesia japonica*) or 1.5 (*Girardia tigrina* and *Polycelis nigra*). The top 10 markers per cluster were extracted using the Wilcoxon method. Most *Schmidtea mediterranea* identities were annotated using the Ingest tool as indicated for Chapter III. The glia, neoblasts, reproductive system and unknown clusters were manually identified according to the expression of SMEST026791001 (glia) (Plass et al., 2018), SmMSTRG.11390 (*piwi*) (Reddien et al., 2005), SMEST018169001 (*nanos*) (Wang et al., 2007), SMEST076962001 (*msy4*) (Wang et al., 2010), SMEST000719001 (*surfactant b*) (Rouhana et al., 2017) and SMEST049067001 (*granulin*) (Zayas et al., 2005). *Schmidtea polychroa*, *Dugesia japonica*, *Girardia tigrina* and *Polycelis nigra* were annotated in broad cell types (Supplementary 8) by assessing the expression of homologous genes to *S. mediterranea* markers. These markers were

extracted from the reference dataset clustered at resolution 0.2, to retain only broader cell identities. Homologies were assigned using BLAST and selecting the hits with the lowest E-values.

2.2 Visualisation of *hnf4* CPM and per cluster (Scanpy)

In **Chapter III**, the *hnf4* counts and the number of cells expressing *hnf4* counts were extracted for each cluster using Pandas (Supplementary 6). These values were used to calculate *hnf4* CPM and multiplied them by the number of cells expressing *hnf4* (**Figure 3.10 A-B**). The results were represented in bar graphs generated with matplotlib.

2.3 Quantification of cluster abundances (Scanpy)

Chapter III:

To quantify cluster abundances, we extracted the number of cells per cluster, for each replicate and RNAi condition (*hnf4* and GFP), using Pandas (Supplementary 6). These numbers were statistically compared with a Fisher test at 95% confidence. We ran separated tests for each cluster and replicate, using contingency tables as the one showed below (**Table 5.1**). Clusters were considered significantly up- or downregulated when the Fisher test was significant for both replicates. Final results were visualized in UMAP and Volcano plots (for the *p* and *odd ratio* values of the statistical test). Volcano plots were generated using Seaborn (sns.scatterplot).

Table 5.1 Generic contingency table used for Fisher test.

REPLICATE X	**GFP**	***hnf4***
Cluster N	counts	counts
All other clusters	counts	counts

Chapter IV:

We extracted the percentage of cells per cluster and per condition (Supplementary 8) using Pandas. For each cluster, these percentages were divided by pair of samples to obtain the ratio, which was later transformed into logarithmic scale and used to estimate cluster enrichments. Clusters with log ratios above +1 were considered enriched in the sample used as numerator to calculate the ratio. Conversely, clusters with ratios below -1 were considered enriched in the denominator sample. Enrichments were presented in bar plots generated with matplotlib.

2.4 Differential gene expression analysis (Scanpy and R)

In **Chapter III**, we used the pseudo-bulk unprocessed data (before normalization) to obtain the raw counts per gene in each condition (*hnf4* and GFP). Clustering information was copied from the processed matrix to assigned clusters and cluster group identities (Supplementary 5). For each cluster group, we ran a pseudo-bulk DEseq2 (Love et al., 2014) analysis in R. The log2 fold change threshold was set at ±1.0, and the minimum p-value at 0.05. Volcano plots were generated in R using EnhancedVolcano. The differentially expressed genes resulted from the analysis were matched to their top homolog identities using an internal diamond blast annotation for *S. mediterranea* (not shown).

REFERENCES

- Abdi, H. and Williams, L.J. (2010) 'Principal component analysis: Principal component analysis', *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), pp. 433–459. Available at: https://doi.org/10.1002/wics.101.

- Abnave, P. *et al.* (2017) 'An X-ray shielded irradiation assay reveals EMT transcription factors control pluripotent adult stem cell migration *in vivo* in planarians', *Development*, p. dev.154971. Available at: https://doi.org/10.1242/dev.154971.

- Aboobaker, A.A. (2011) 'Planarian stem cells: a simple paradigm for regeneration', *Trends in Cell Biology*, 21(5). Available at: https://doi.org/doi:10.1016/j.tcb.2011.01.005.

- Achim, K. *et al.* (2018) 'Whole-Body Single-Cell Sequencing Reveals Transcriptional Domains in the Annelid Larval Body', *Molecular Biology and Evolution*. Edited by G. Wagner, 35(5), pp. 1047–1062. Available at: https://doi.org/10.1093/molbev/msx336.

- Adam, M., Potter, A.S. and Potter, S.S. (2017) 'Psychrophilic proteases dramatically reduce single cell RNA-seq artifacts: A molecular atlas of kidney development', *Development*, p. dev.151142. Available at: https://doi.org/10.1242/dev.151142.

- Aldini, G. *et al.* (2018) 'N-Acetylcysteine as an antioxidant and disulphide breaking agent: the reasons why', *Free Radical Research*, 52(7), pp. 751–762. Available at: https://doi.org/10.1080/10715762.2018.1468564.

- Allan, K. *et al.* (2020) 'Preparing a Single Cell Suspension from Zebrafish Retinal Tissue for Flow Cytometric Cell Sorting of Müller Glia', *Cytometry Part A*, 97(6), pp. 638–646. Available at: https://doi.org/10.1002/cyto.a.23936.

- Alles, J. *et al.* (2017) 'Cell fixation and preservation for droplet-based single-cell transcriptomics', *BMC Biology*, 15(1), p. 44. Available at: https://doi.org/10.1186/s12915-017-0383-5.

- Álvarez-Presas, M., Baguñà, J. and Riutort, M. (2008) 'Molecular phylogeny of land and freshwater planarians (Tricladida, Platyhelminthes): From freshwater to land and back', *Molecular Phylogenetics and Evolution*, 47(2), pp. 555–568. Available at: https://doi.org/10.1016/j.ympev.2008.01.032.

- Alwine, J.C., Kemp, D.J. and Stark, G.R. (1977) 'Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes.', *Proceedings of the National Academy of Sciences*, 74(12), pp. 5350–5354. Available at: https://doi.org/10.1073/pnas.74.12.5350.

- An, Y. *et al.* (2018) 'Draft genome of Dugesia japonica provides insights into conserved regulatory elements of the brain restriction gene nou-darake in planarians', *Zoological Letters*, 4(1), p. 24. Available at: https://doi.org/10.1186/s40851-018-0102-2.

- Anders, S., Pyl, P.T. and Huber, W. (2015) 'HTSeq--a Python framework to work with high-throughput sequencing data', *Bioinformatics*, 31(2), pp. 166–169. Available at: https://doi.org/10.1093/bioinformatics/btu638.

- Asami, M. *et al.* (2002) 'Cultivation and Characterization of Planarian Neuronal Cells Isolated by Fluorescence Activated Cell Sorting (FACS)', *Zoological Science*, 19(11), pp. 1257–1265. Available at: https://doi.org/10.2108/zsj.19.1257.

- Attar, M. *et al.* (2018) 'A practical solution for preserving single cells for RNA sequencing', *Scientific Reports*, 8(1), p. 2151. Available at: https://doi.org/10.1038/s41598-018-20372-7.

- Bacher, R. *et al.* (2017) 'SCnorm: robust normalization of single-cell RNA-seq data', *Nature Methods*, 14(6), pp. 584–586. Available at: https://doi.org/10.1038/nmeth.4263.

- Baguña, J. (2012) 'The planarian neoblast: the rambling history of its origin and some current black boxes', *The International Journal of Developmental Biology*, 56(1-2–3), pp. 19–37. Available at: https://doi.org/10.1387/ijdb.113463jb.

- Baguña, J. and Romero, R. (1981) 'Quantitative analysis of cell types during growth, degrowth and regeneration in the planarians Dugesia mediterranea and Dugesia tigrina', *Hydrobiologia*, 84, pp. 181–194.

- Bainbridge, M.N. *et al.* (2006) 'Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach', *BMC Genomics*, 7(1), p. 246. Available at: https://doi.org/10.1186/1471-2164-7-246.

- Baldan, V. *et al.* (2015) 'Efficient and reproducible generation of tumour-infiltrating lymphocytes for renal cell carcinoma', *British Journal of Cancer*, 112(9), pp. 1510–1518. Available at: https://doi.org/10.1038/bjc.2015.96.

- Baltimore, D. (1970) 'Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of RNA Tumour Viruses', *Nature*, 226(5252), pp. 1209–1211. Available at: https://doi.org/10.1038/2261209a0.

- Baran, Y. *et al.* (2019) 'MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions', *Genome Biology*, 20(1), p. 206. Available at: https://doi.org/10.1186/s13059-019-1812-2.

- Barba, M., Czosnek, H. and Hadidi, A. (2014) 'Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology', *Viruses*, 6(1), pp. 106–136. Available at: https://doi.org/10.3390/v6010106.

- Barberán, S., Fraguas, S. and Cebrià, F. (2016) 'The EGFR signaling pathway controls gut progenitor differentiation during planarian regeneration and homeostasis', *Development*, p. dev.131995. Available at: https://doi.org/10.1242/dev.131995.

- Barberán, S., Martín-Durán, J.M. and Cebrià, F. (2016) 'Evolution of the EGFR pathway in Metazoa and its diversification in the planarian Schmidtea mediterranea', *Scientific Reports*, 6(1), p. 28071. Available at: https://doi.org/10.1038/srep28071.

- Baron, M. *et al.* (2016) 'A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure', *Cell Systems*, 3(4), pp. 346-360.e4. Available at: https://doi.org/10.1016/j.cels.2016.08.011.

- Barry, W.E. and Thummel, C.S. (2016) 'The Drosophila HNF4 nuclear receptor promotes glucose-stimulated insulin secretion and mitochondrial function in adults', *eLife*, 5, p. e11183. Available at: https://doi.org/10.7554/eLife.11183.

- Becker-André, M. and Hahlbrock, K. (1989) 'Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcript titration assay (PATITY)', *Nucleic Acids Research*, 17(22), pp. 9437–9446.

- Beckert, B. and Masquida, B. (2011) 'Synthesis of RNA by In Vitro Transcription', in H. Nielsen (ed.) *RNA*. Totowa, NJ: Humana Press (Methods in Molecular Biology), pp. 29–41. Available at: https://doi.org/10.1007/978-1-59745-248-9_3.

- Benazzi, M. *et al.* (1975) 'Further Contribution to the Taxonomy of the « *Dugesia Lugubris-Polychroa Group* » with Description of *Dugesia Mediterranea* N.SP. (Tricladida, Paludicola)', *Bolletino di zoologia*, 42(1), pp. 81–89. Available at: https://doi.org/10.1080/11250007509430132.

- Bendall, S.C. *et al.* (2014) 'Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B Cell Development', *Cell*, 157(3), pp. 714–725. Available at: https://doi.org/10.1016/j.cell.2014.04.005.

- Benham-Pyle, B.W. *et al.* (2021) 'Identification of rare, transient post-mitotic cell states that are induced by injury and required for whole-body regeneration in Schmidtea mediterranea', *Nature Cell Biology*, 23(9), pp. 939–952. Available at: https://doi.org/10.1038/s41556-021-00734-6.

- Bernstein, N.J. *et al.* (2020) 'Solo: Doublet Identification in Single-Cell RNA-Seq via Semi-Supervised Deep Learning', *Cell Systems*, 11(1), pp. 95-101.e5. Available at: https://doi.org/10.1016/j.cels.2020.05.010.

- Best, M.G. *et al.* (2015) 'RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics', *Cancer Cell*, 28(5), pp. 666–676. Available at: https://doi.org/10.1016/j.ccell.2015.09.018.

- Beukeboom, L.W., Sharbel, T.F. and Michiels, N.K. (1998) 'Reproductive modes, ploidy distribution, and supernumerary chromosome frequencies of the flatworm Polycelis nigra (Platyhelminthes: Tricladida)', *Hydrobiologia*, 383(1), pp. 277–285. Available at: https://doi.org/10.1023/A:1003460132521.

- Blythe, M.J. *et al.* (2010) 'A Dual Platform Approach to Transcript Discovery for the Planarian Schmidtea Mediterranea to Establish RNAseq for Stem Cell and Regeneration Biology', *PLoS ONE*. Edited by J. Jaeger, 5(12), p. e15617. Available at: https://doi.org/10.1371/journal.pone.0015617.

- Bolger, A.M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. Available at: https://doi.org/10.1093/bioinformatics/btu170.

- Bolotin, E. *et al.* (2010) 'Integrated approach for the identification of human hepatocyte nuclear factor 4α target genes using protein binding microarrays', *Hepatology*, 51(2), pp. 642–653. Available at: https://doi.org/10.1002/hep.23357.

- Bradshaw, B., Thompson, K. and Frank, U. (2015) 'Distinct mechanisms underlie oral vs aboral regeneration in the cnidarian Hydractinia echinata', *eLife*, 4, p. e05506. Available at: https://doi.org/10.7554/eLife.05506.

- Brandl, H. *et al.* (2016) 'PlanMine – a mineable resource of planarian biology and biodiversity', *Nucleic Acids Research*, 44(D1), pp. D764–D773. Available at: https://doi.org/10.1093/nar/gkv1148.

- Bray, N.L. *et al.* (2016) 'Near-optimal probabilistic RNA-seq quantification', *Nature Biotechnology*, 34(5), pp. 525–527. Available at: https://doi.org/10.1038/nbt.3519.

- Briggs, J.A. *et al.* (2018) 'The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution', *Science*, 360(6392), p. eaar5780. Available at: https://doi.org/10.1126/science.aar5780.

- van den Brink, S.C. *et al.* (2017) 'Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations', *Nature Methods*, 14(10), pp. 935–936. Available at: https://doi.org/10.1038/nmeth.4437.

- Brosch, M. *et al.* (2018) 'Epigenomic map of human liver reveals principles of zonated morphogenic and metabolic control', *Nature Communications*, 9(1), p. 4150. Available at: https://doi.org/10.1038/s41467-018-06611-5.

- Buenrostro, J.D. *et al.* (2015) 'Single-cell chromatin accessibility reveals principles of regulatory variation', *Nature*, 523(7561), pp. 486–490. Available at: https://doi.org/10.1038/nature14590.

- Bumgarner, R. (2013) 'Overview of DNA Microarrays: Types, Applications, and Their Future', *Current Protocols in Molecular Biology*, 101(1). Available at: https://doi.org/10.1002/0471142727.mb2201s101.

- Butler, A. *et al.* (2018) 'Integrating single-cell transcriptomic data across different conditions, technologies, and species', *Nature Biotechnology*, 36(5), pp. 411–420. Available at: https://doi.org/10.1038/nbt.4096.

- Büttner, M. *et al.* (2019) 'A test metric for assessing single-cell RNA-seq batch correction', *Nature Methods*, 16(1), pp. 43–49. Available at: https://doi.org/10.1038/s41592-018-0254-1.

- Büttner, M. *et al.* (2021) 'scCODA is a Bayesian model for compositional single-cell data analysis', *Nature Communications*, 12(1), p. 6876. Available at: https://doi.org/10.1038/s41467-021-27150-6.

- Cantarel, B.L. *et al.* (2008) 'MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes', *Genome Research*, 18(1), pp. 188–196. Available at: https://doi.org/10.1101/gr.6743907.

- Cao, J. *et al.* (2017) 'Comprehensive single-cell transcriptional profiling of a multicellular organism', *Science*, 357(6352), pp. 661–667. Available at: https://doi.org/10.1126/science.aam8940.

- Cao, J. *et al.* (2019) 'The single-cell transcriptional landscape of mammalian organogenesis', *Nature*, 566(7745), pp. 496–502. Available at: https://doi.org/10.1038/s41586-019-0969-x.

- Cao, J. *et al.* (2020) 'A human cell atlas of fetal gene expression', *Science*, 370(6518), p. eaba7721. Available at: https://doi.org/10.1126/science.aba7721.

- Carruthers, M. *et al.* (2018) 'De novo transcriptome assembly, annotation and comparison of four ecological and evolutionary model salmonid fish species', *BMC Genomics*, 19(1), p. 32. Available at: https://doi.org/10.1186/s12864-017-4379-x.

- Cattin, A.-L. *et al.* (2009) 'Hepatocyte Nuclear Factor 4α, a Key Factor for Homeostasis, Cell Architecture, and Barrier Function of the Adult Intestinal Epithelium', *Molecular and Cellular Biology*, 29(23), pp. 6294–6308. Available at: https://doi.org/10.1128/MCB.00939-09.

- Chan, J., Nakabayashi, H. and Wong, N.C.W. (1993) 'HNF-4 increases activity of the rat Apo Al gene', *Nucleic Acids Research*, 21(5), pp. 1205–1211. Available at: https://doi.org/10.1093/nar/21.5.1205.

- Charbagi-Barbirou, K. and Tekaya, S. (2009) 'Sexual differentiation and karyological study in the gonochoristic planarian Sabussowia dioica (Platyhelminthes, Tricladida)', *Cahiers De Biologie Marine*, 50, pp. 303–309.

- Chaudhry, M.A. (2008) 'Induction of Gene Expression Alterations by Culture Medium from Trypsinized Cells', *Jornal of Biological Sciences*, 8(1), pp. 81–87. Available at: https://doi.org/10.3923/jbs.2008.81.87.

- Chen, G., Ning, B. and Shi, T. (2019) 'Single-Cell RNA-Seq Technologies and Related Computational Data Analysis', *Frontiers in Genetics*, 10, p. 317. Available at: https://doi.org/10.3389/fgene.2019.00317.

- Chen, J. *et al.* (2014) 'RNA-Seq for gene identification and transcript profiling of three Stevia rebaudiana genotypes', *BMC Genomics*, 15(1), p. 571. Available at: https://doi.org/10.1186/1471-2164-15-571.

- Chen, J. *et al.* (2018) 'PBMC fixation and processing for Chromium single-cell RNA sequencing', *Journal of Translational Medicine*, 16(1), p. 198. Available at: https://doi.org/10.1186/s12967-018-1578-4.

- Chen, J. and Ginhoux, F. (2018) 'A Single-Cell Sequencing Guide for Immunologists', *Frontiers in Immunology*, 9, p. 13. Available at: https://doi.org/10.3389/fimmu.2018.02425.

- Chen, K.H. *et al.* (2015) 'Spatially resolved, highly multiplexed RNA profiling in single cells', *Science*, 348(6233), p. aaa6090. Available at: https://doi.org/10.1126/science.aaa6090.

- Chen, L. *et al.* (2019) 'A reinforcing HNF4–SMAD4 feed-forward module stabilizes enterocyte identity', *Nature Genetics*, 51(5), pp. 777–785. Available at: https://doi.org/10.1038/s41588-019-0384-0.

- Chen, S. *et al.* (2018) 'fastp: an ultra-fast all-in-one FASTQ preprocessor', *Bioinformatics*, 34(17), pp. i884–i890. Available at: https://doi.org/10.1093/bioinformatics/bty560.

- Chen, W. *et al.* (2016) 'Identification and Comparative Analysis of Differential Gene Expression in Soybean Leaf Tissue under Drought and Flooding Stress Revealed by RNA-Seq', *Frontiers in Plant Science*, 7. Available at: https://doi.org/10.3389/fpls.2016.01044.

- Chen, W.S. *et al.* (1994) 'Disruption of the HNF-4 gene, expressed in visceral endoderm, leads to cell death in embryonic ectoderm and impaired gastrulation of mouse embryos', *Genes & Development*, 8, pp. 2466–2477.

- Chen, Z., Ling, J. and Gallie, D. (2004) 'RNase activity requires formation of disulfide bonds and is regulated by the redox state', *Plant Molecular Biology*, 55(1), pp. 83–96. Available at: https://doi.org/10.1007/s11103-004-0438-1.

- Chong, T. *et al.* (2011) 'Molecular markers to characterize the hermaphroditic reproductive system of the planarian Schmidtea mediterranea', *BMC Developmental Biology*, 11(1), p. 69. Available at: https://doi.org/10.1186/1471-213X-11-69.

- Chung, Y.-S. *et al.* (2021) 'Validation of real-time RT-PCR for detection of SARS-CoV-2 in the early stages of the COVID-19 outbreak in the Republic of Korea', *Scientific Reports*, 11(1), p. 14817. Available at: https://doi.org/10.1038/s41598-021-94196-3.

- Coffin, J.M. and Fan, H. (2016) 'The Discovery of Reverse Transcriptase', *Annual Review of Virology*, 3(1), pp. 29–51. Available at: https://doi.org/10.1146/annurev-virology-110615-035556.

- Collins, J.J. *et al.* (2010) 'Genome-Wide Analyses Reveal a Role for Peptide Hormones in Planarian Germline Development', *PLoS Biology*. Edited by V. Hartenstein, 8(10), p. e1000509. Available at: https://doi.org/10.1371/journal.pbio.1000509.

- Collins, J.J. (2017) 'Platyhelminthes', *Current Biology*, 27(7), pp. R252–R256. Available at: https://doi.org/10.1016/j.cub.2017.02.016.

- Corchete, L.A. *et al.* (2020) 'Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis', *Scientific Reports*, 10(1), p. 19737. Available at: https://doi.org/10.1038/s41598-020-76881-x.

- Costa, R.H. and Grayson, D.R. (1989) 'Multiple Hepatocyte-Enriched Nuclear Factors Function in the Regulation of Transthyretin and otl-Antitrypsin Genes', *MOL. CELL. BIOL.*, 9, p. 11.

- Craig, A.G. *et al.* (1990) 'Ordering of cosmid clones covering the Herpes simplex virus type I (HSV-l) genome: a test case for fingerprinting by hybridisation', *Nucleic Acids Research*, 18(9), pp. 2953–60.

- Currie, K.W. and Pearson, B.J. (2013) 'Transcription factors *lhx1/5-1* and *pitx* are required for the maintenance and regeneration of serotonergic neurons in planarians', *Development*, 140(17), pp. 3577–3588. Available at: https://doi.org/10.1242/dev.098590.

- Cusanovich, D.A. *et al.* (2015) 'Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing', *Science*, 348(6237), pp. 910–914. Available at: https://doi.org/10.1126/science.aab1601.

- David, C.N. (1973) 'A quantitative method for maceration of hydra tissue', *Wilhelm Roux' Archiv fur Entwicklungsmechanik der Organismen*, 171(4), pp. 259–268. Available at: https://doi.org/10.1007/BF00577724.

- Davie, K. *et al.* (2018) 'A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain', *Cell*, 174(4), pp. 982-998.e20. Available at: https://doi.org/10.1016/j.cell.2018.05.057.

- Denisenko, E. *et al.* (2020) 'Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows', *Genome Biology*, 21(1), p. 130. Available at: https://doi.org/10.1186/s13059-020-02048-6.

- DePasquale, E.A.K. *et al.* (2019) 'DoubletDecon: Deconvoluting Doublets from Single-Cell RNA-Sequencing Data', *Cell Reports*, 29(6), pp. 1718-1727.e8. Available at: https://doi.org/10.1016/j.celrep.2019.09.082.

- Diggelmann, H., Faust, C.H. and Mach, B. (1973) 'Enzymatic Synthesis of DNA Complementary to Purified 14S Messenger RNA of Immunoglobulin Light Chain', *Proceedings of the National Academy of Sciences*, 70(3), pp. 693–696. Available at: https://doi.org/10.1073/pnas.70.3.693.

- van Dijk, E.L. *et al.* (2018) 'The Third Revolution in Sequencing Technology', *Trends in Genetics*, 34(9), pp. 666–681. Available at: https://doi.org/10.1016/j.tig.2018.05.008.

- Dobin, A. *et al.* (2013) 'STAR: ultrafast universal RNA-seq aligner', *Bioinformatics*, 29(1), pp. 15–21. Available at: https://doi.org/10.1093/bioinformatics/bts635.

- D'Souza, T.G. *et al.* (2006) 'Paternal inheritance in parthenogenetic forms of the planarian Schmidtea polychroa', *Heredity*, 97(2), pp. 97–101. Available at: https://doi.org/10.1038/sj.hdy.6800841.

- D'Souza, T.G. and Michiels, N.K. (2009) 'Sex in Parthenogenetic Planarians: Phylogenetic Relic or Evolutionary Resurrection?', in I. Schön, K. Martens, and P. Dijk (eds) *Lost Sex*. Dordrecht: Springer Netherlands, pp. 377–397. Available at: https://doi.org/10.1007/978-90-481-2770-2_18.

- Dubois, V. *et al.* (2020) 'Control of Cell Identity by the Nuclear Receptor HNF4 in Organ Pathophysiology', *Cells*, 9(10), p. 2185. Available at: https://doi.org/10.3390/cells9102185.

- Egger, B., Gschwentner, R. and Rieger, R. (2007) 'Free-living flatworms under the knife: past and present', *Development Genes and Evolution*, 217(2), p. 89. Available at: https://doi.org/10.1007/s00427-006-0120-5.

- Eid, J. *et al.* (2009) 'Real-Time DNA Sequencing from Single Polymerase Molecules', *Science*, 323(5910), pp. 133–138. Available at: https://doi.org/10.1126/science.1162986.

- Emms, D.M. and Kelly, S. (2015) 'OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy', *Genome Biology*, 16(1), p. 157. Available at: https://doi.org/10.1186/s13059-015-0721-2.

- Eng, C.-H.L. *et al.* (2019) 'Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+', *Nature*, 568(7751), pp. 235–239. Available at: https://doi.org/10.1038/s41586-019-1049-y.

- Farrell, J.A. *et al.* (2018) 'Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis', *Science*, 360(6392), p. eaar3131. Available at: https://doi.org/10.1126/science.aar3131.

- Finak, G. *et al.* (2015) 'MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data', *Genome Biology*, 16(1), p. 278. Available at: https://doi.org/10.1186/s13059-015-0844-5.

- Fincher, C.T. *et al.* (2018) 'Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*', *Science*, 360(6391), p. eaaq1736. Available at: https://doi.org/10.1126/science.aaq1736.

- Forsthoefel, D.J. *et al.* (2012) 'An RNAi Screen Reveals Intestinal Regulators of Branching Morphogenesis, Differentiation, and Stem Cell Proliferation in Planarians', *Developmental Cell*, 23(4), pp. 691–704. Available at: https://doi.org/10.1016/j.devcel.2012.09.008.

- Forsthoefel, D.J. *et al.* (2020) 'Cell-type diversity and regionalized gene expression in the planarian intestine', *eLife*, 9, p. e52613. Available at: https://doi.org/10.7554/eLife.52613.

- Gallegos, T.F. *et al.* (2012) 'Organic Anion and Cation SLC22 "Drug" Transporter (Oat1, Oat3, and Oct1) Regulation during Development and Maturation of the Kidney Proximal Tubule', *PLoS ONE*. Edited by S.R. Singh, 7(7), p. e40796. Available at: https://doi.org/10.1371/journal.pone.0040796.

- García-Castro, H. *et al.* (2021) 'ACME dissociation: a versatile cell fixation-dissociation method for single-cell transcriptomics', *Genome Biology*, 22(1), p. 89. Available at: https://doi.org/10.1186/s13059-021-02302-5.

- Geirsdottir, L. *et al.* (2019) 'Cross-Species Single-Cell Analysis Reveals Divergence of the Primate Microglia Program', *Cell*, 179(7), pp. 1609-1622.e16. Available at: https://doi.org/10.1016/j.cell.2019.11.010.

- Gergen, J.P., Stern, R.H. and Wensink, P.C. (1979) 'Filter replicas and permanent collections of recombinant DNA plasmids', *Nucleic Acids Research*, 7(8), pp. 2115–2136. Available at: https://doi.org/10.1093/nar/7.8.2115.

- Gershon, D. (2004) 'Microarrays go mainstream', *Nature Methods*, 1(3), pp. 263–270. Available at: https://doi.org/10.1038/nmeth1204-263.

- Gibbons, M. *et al.* (2022) 'Rapid, scalable isolation of human tumor nuclei for single cell genomics', *Cancer Research*, 82(12_Supplement), p. 3392. Available at: https://doi.org/10.1158/1538-7445.AM2022-3392.

- Girard, C. (1850) 'Description of North American Planariae', *Proceedings of the Boston Society of Natural History 1848-1851*, 3, pp. 264–265.

- González-Estévez, C. *et al.* (2007) '*Gtdap-1* promotes autophagy and is required for planarian remodeling during regeneration and starvation', *Proceedings of the National Academy of Sciences*, 104(33), pp. 13373–13378. Available at: https://doi.org/10.1073/pnas.0703588104.

- Grabherr, M.G. *et al.* (2011) 'Full-length transcriptome assembly from RNA-Seq data without a reference genome', *Nature Biotechnology*, 29(7), pp. 644–652. Available at: https://doi.org/10.1038/nbt.1883.

- Grau-Bové, X. and Sebé-Pedrós, A. (2021) 'Orthology Clusters from Gene Trees with Possvm', *Mol Biol Evol*, 38(5). Available at: https://doi.org/10.1093/molbev/msab234.

- Griffiths, J.A., Scialdone, A. and Marioni, J.C. (2018) 'Using single-cell genomics to understand developmental processes and cell fate decisions', *Molecular Systems Biology*, 14(4). Available at: https://doi.org/10.15252/msb.20178046.

- Grindberg, R.V. *et al.* (2013) 'RNA-sequencing from single nuclei', *Proceedings of the National Academy of Sciences*, 110(49), pp. 19802–19807. Available at: https://doi.org/10.1073/pnas.1319700110.

- Grohme, M.A. *et al.* (2018) 'The genome of Schmidtea mediterranea and the evolution of core cellular mechanisms', *Nature*, 554(7690), pp. 56–61. Available at: https://doi.org/10.1038/nature25473.

- Grunstein, M. and Hogness, D.S. (1975) 'Colony hybridization: A method for the isolation of clofied DNAs that contain a specific gene', *Proc. Nat. Acad. Sci. USA*, 72(10), pp. 3961–3965. Available at: https://doi.org/10.1073/pnas.72.10.3961.

- Guillaumet-Adkins, A. *et al.* (2017) 'Single-cell transcriptome conservation in cryopreserved cells and tissues', *Genome Biology*, 18(1), p. 45. Available at: https://doi.org/10.1186/s13059-017-1171-9.

- Guo, L. *et al.* (2022) 'Island-specific evolution of a sex-primed autosome in a sexual planarian', *Nature*, 606(7913), pp. 329–334. Available at: https://doi.org/10.1038/s41586-022-04757-3.

- Habib, N. *et al.* (2017) 'Massively parallel single-nucleus RNA-seq with DroNc-seq', *Nature Methods*, 14(10), pp. 955–958. Available at: https://doi.org/10.1038/nmeth.4407.

- Hagemann-Jensen, M. *et al.* (2020) 'Single-cell RNA counting at allele and isoform resolution using Smart-seq3', *Nature Biotechnology*, 38(6), pp. 708–714. Available at: https://doi.org/10.1038/s41587-020-0497-0.

- Han, J.C. and Han, G.Y. (1994) 'A procedure for quantitative determination of tris(2-carboxyethyl)phosphine, an odorless reducing agent more stable and effective than dithiothreitol', *Analytical Biochemistry*, 220, pp. 5–10. Available at: https://doi.org/10.1006/ABIO.1994.1290.

- Hanamsagar, R. *et al.* (2020) 'An optimized workflow for single-cell transcriptomics and repertoire profiling of purified lymphocytes from clinical samples', *Scientific Reports*, 10(1), p. 2219. Available at: https://doi.org/10.1038/s41598-020-58939-y.

- Hannon, G.J. (2002) 'RNA interference', *Nature*, 418, pp. 244–251. Available at: https://doi.org/10.1038/418244a.

- Hao, Y. *et al.* (2021) 'Integrated analysis of multimodal single-cell data', *Cell*, 184(13), pp. 3573-3587.e29. Available at: https://doi.org/10.1016/j.cell.2021.04.048.

- Hashimshony, T. *et al.* (2012) 'CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification', *Cell Reports*, 2(3), pp. 666–673. Available at: https://doi.org/10.1016/j.celrep.2012.08.003.

- Hashimshony, T. *et al.* (2016) 'CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq', *Genome Biology*, 17(1), p. 77. Available at: https://doi.org/10.1186/s13059-016-0938-8.

- Hayashi, T. *et al.* (2006) 'Isolation of planarian X-ray-sensitive stem cells by fluorescence-activated cell sorting', *Development, Growth and Differentiation*, 48(6), pp. 371–380. Available at: https://doi.org/10.1111/j.1440-169X.2006.00876.x.

- Hayashi, T. *et al.* (2010) 'Single-cell gene profiling of planarian stem cells using fluorescent activated cell sorting and its "index sorting" function for stem cell research: Single-cell gene profiling of stem cell', *Development, Growth & Differentiation*, 52(1), pp. 131–144. Available at: https://doi.org/10.1111/j.1440-169X.2009.01157.x.

- He, X. *et al.* (2017) 'FOX and ETS family transcription factors regulate the pigment cell lineage in planarians', *Development*, p. dev.156349. Available at: https://doi.org/10.1242/dev.156349.

- Hie, B., Bryson, B. and Berger, B. (2019) 'Efficient integration of heterogeneous single-cell transcriptomes using Scanorama', *Nature Biotechnology*, 37(6), pp. 685–691. Available at: https://doi.org/10.1038/s41587-019-0113-3.

- Holewa, B. *et al.* (1997) 'HNF4β, a New Gene of the HNF4 Family with Distinct Activation and Expression Profiles in Oogenesis and Embryogenesis of Xenopus laevis', *MOL. CELL. BIOL.*, 17, p. 8.

- Hong, M. *et al.* (2020) 'RNA sequencing: new technologies and applications in cancer research', *Journal of Hematology & Oncology*, 13(1), p. 166. Available at: https://doi.org/10.1186/s13045-020-01005-x.

- Howat, W.J. and Wilson, B.A. (2014) 'Tissue fixation and the effect of molecular fixatives on downstream staining procedures', *Methods*, 70(1), pp. 12–19. Available at: https://doi.org/10.1016/j.ymeth.2014.01.022.

- Hu, P. *et al.* (2016) 'Single Cell Isolation and Analysis', *Frontiers in Cell and Developmental Biology*, 4. Available at: https://doi.org/10.3389/fcell.2016.00116.

- Huang, H.-L. *et al.* (2010) 'Trypsin-induced proteome alteration during cell subculture in mammalian cells', *Journal of Biomedical Science*, 17(36). Available at: https://doi.org/10.1186/1423-0127-17-36.

- Ichikawa and Kawakatsu (1964) 'Dugesia japonica', in *Continenticola*. Miller SE, Rycroft S.

- Islam, S. *et al.* (2011) 'Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq', *Genome Research*, 21(7), pp. 1160–1167. Available at: https://doi.org/10.1101/gr.110882.110.

- Ivankovic, M. *et al.* (2019) 'Model systems for regeneration: planarians', *Development*, 146(17), p. dev167684. Available at: https://doi.org/10.1242/dev.167684.

- Jain, M. *et al.* (2018) 'Nanopore sequencing and assembly of a human genome with ultra-long reads', *Nature Biotechnology*, 36(4), pp. 338–345. Available at: https://doi.org/10.1038/nbt.4060.

- Jaitin, D.A. *et al.* (2014) 'Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types', *Science*, 343(6172), pp. 776–779. Available at: https://doi.org/10.1126/science.1247651.

- Jamur, M.C. and Oliver, C. (2010) 'Permeabilization of Cell Membranes', in C. Oliver and M.C. Jamur (eds) *Immunocytochemical Methods and Protocols*. Totowa, NJ: Humana Press, pp. 63–66. Available at: https://doi.org/10.1007/978-1-59745-324-0_9.

- Kawakatsu, M., Oki, I. and Tamura, S. (1995) 'Taxonomy and geographical distribution of Dugesia japonica and D. ryukyuensis in the Far East', *Hydrobiologia*, 305, pp. 55–61. Available at: https://doi.org/10.1007/BF00036363.

- Keren-Shaul, H. *et al.* (2019) 'MARS-seq2.0: an experimental and analytical pipeline for indexed sorting combined with single-cell RNA sequencing', *Nature Protocols*, 14(6), pp. 1841–1862. Available at: https://doi.org/10.1038/s41596-019-0164-4.

- Kim, D. *et al.* (2013) 'TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions', *Genome Biology*, 14(4), p. R36. Available at: https://doi.org/10.1186/gb-2013-14-4-r36.

- Kim, D. *et al.* (2019) 'Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype', *Nature Biotechnology*, 37(8), pp. 907–915. Available at: https://doi.org/10.1038/s41587-019-0201-4.

- Klein, A.M. *et al.* (2015) 'Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells', *Cell*, 161(5), pp. 1187–1201. Available at: https://doi.org/10.1016/j.cell.2015.04.044.

- Korsunsky, I. *et al.* (2019) 'Fast, sensitive and accurate integration of single-cell data with Harmony', *Nature Methods*, 16(12), pp. 1289–1296. Available at: https://doi.org/10.1038/s41592-019-0619-0.

- Kovaka, S. *et al.* (2019) 'Transcriptome assembly from long-read RNA-seq alignments with StringTie2', *Genome Biology*, 20(1), p. 278. Available at: https://doi.org/10.1186/s13059-019-1910-1.

- Krishnaswami, S.R. *et al.* (2016) 'Using single nuclei for RNA-seq to capture the transcriptome of postmortem neurons', *Nature Protocols*, 11(3), pp. 499–524. Available at: https://doi.org/10.1038/nprot.2016.015.

- Kulkarni, A. *et al.* (2019) 'Beyond bulk: a review of single cell transcriptomics methodologies and applications', *Current Opinion in Biotechnology*, 58, pp. 129–136. Available at: https://doi.org/10.1016/j.copbio.2019.03.001.

- La Manno, G. *et al.* (2018) 'RNA velocity of single cells', *Nature*, 560(7719), pp. 494–498. Available at: https://doi.org/10.1038/s41586-018-0414-6.

- Labbé, R.M. *et al.* (2012) 'A Comparative Transcriptomic Analysis Reveals Conserved Features of Stem Cell Pluripotency in Planarians and Mammals', *STEM CELLS*, 30(8), pp. 1734–1745. Available at: https://doi.org/10.1002/stem.1144.

- Lacar, B. *et al.* (2016) 'Nuclear RNA-seq of single neurons reveals molecular signatures of activation', *Nature Communications*, 7(1), p. 11022. Available at: https://doi.org/10.1038/ncomms11022.

- Lafzi, A. *et al.* (2018) 'Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies', *Nature Protocols*, 13(12), pp. 2742–2757. Available at: https://doi.org/10.1038/s41596-018-0073-y.

- Lau, H.H. *et al.* (2018) 'The molecular functions of hepatocyte nuclear factors – In and beyond the liver', *Journal of Hepatology*, 68(5), pp. 1033–1048. Available at: https://doi.org/10.1016/j.jhep.2017.11.026.

- Laumer, C.E., Hejnol, A. and Giribet, G. (2015) 'Nuclear genomic signals of the "microturbellarian" roots of platyhelminth evolutionary innovation', *eLife*, 4, p. e05503. Available at: https://doi.org/10.7554/eLife.05503.

- Lázaro, E.M. *et al.* (2011) 'Schmidtea mediterranea phylogeography: an old species surviving on a few Mediterranean islands?', *BMC Evolutionary Biology*, 11(1), p. 274. Available at: https://doi.org/10.1186/1471-2148-11-274.

- Lennon, G.G. and Lehrach, H. (1991) 'Hybridization analyses of arrayed cDNA libraries', *Perspectives*, 7(10), pp. 314–317.

- Levene, M.J. *et al.* (2003) 'Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations', *Science*, 299(5607), pp. 682–686. Available at: https://doi.org/10.1126/science.1079700.

- Levy, S. *et al.* (2021) 'A stony coral cell atlas illuminates the molecular and cellular basis of coral symbiosis, calcification, and immunity', *Cell*, 184(11), pp. 2973-2987.e18. Available at: https://doi.org/10.1016/j.cell.2021.04.005.

- Li, B. and Dewey, C.N. (2011) 'RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome', *BMC Bioinformatics*, 12(323). Available at: https://doi.org/10.1186/1471-2105-12-323.

- Li, J., Ning, G. and Duncan, S.A. (2000) 'Mammalian hepatocyte differentiation requires the transcription factor HNF-4α', *Genes & Development*, 14, pp. 464–474.

- Lin, Y. *et al.* (2022) 'scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning', *Nature Biotechnology*, 40(5), pp. 703–710. Available at: https://doi.org/10.1038/s41587-021-01161-6.

- Lipshutz, R.J. *et al.* (1999) 'High density synthetic oligonucleotide arrays', *Nature Genetics*, 21(S1), pp. 20–24. Available at: https://doi.org/10.1038/4447.

- Liu, L. *et al.* (2012) 'Comparison of Next-Generation Sequencing Systems', *Journal of Biomedicine and Biotechnology*, 2012, pp. 1–11. Available at: https://doi.org/10.1155/2012/251364.

- Lobo, D., Morokuma, J. and Levin, M. (2016) 'Computational discovery and *in vivo* validation of *hnf4* as a regulatory gene in planarian regeneration', *Bioinformatics*, 32(17), pp. 2681–2685. Available at: https://doi.org/10.1093/bioinformatics/btw299.

- Longo, S.K. *et al.* (2021) 'Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics', *Nature Reviews Genetics*, 22(10), pp. 627–644. Available at: https://doi.org/10.1038/s41576-021-00370-8.

- Lopez-Maestre, H. *et al.* (2016) 'SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence', *Nucleic Acids Research*, p. gkw655. Available at: https://doi.org/10.1093/nar/gkw655.

- Love, M.I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), p. 550. Available at: https://doi.org/10.1186/s13059-014-0550-8.

- Lowe, R. *et al.* (2017) 'Transcriptomics technologies', *PLOS Computational Biology*, 13(5), p. e1005457. Available at: https://doi.org/10.1371/journal.pcbi.1005457.

- Luecken, M.D. and Theis, F.J. (2019) 'Current best practices in single-cell RNA-seq analysis: a tutorial', *Molecular Systems Biology*, 15(6). Available at: https://doi.org/10.15252/msb.20188746.

- Lun, A.T.L., McCarthy, D.J. and Marioni, J.C. (2016) 'A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor', *F1000Research*, 5(2122). Available at: https://doi.org/doi: 10.12688/f1000research.9501.2.

- Lust, K. *et al.* (no date) 'Single-cell analyses of axolotl telencephalon organization, neurogenesis, and regeneration', *Science*, 377(6610), p. eabp9262. Available at: https://doi.org/10.1126/science.abp9262.

- Ma, S. *et al.* (2020) 'Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin', *Cell*, 183(4), pp. 1103-1116.e20. Available at: https://doi.org/10.1016/j.cell.2020.09.056.

- van der Maaten, L. and Hinton, G. (2008) 'Visualizing Data using t-SNE', *Journal of Machine Learning Research*, 9, pp. 2579–2605.

- Macosko, E.Z. *et al.* (2015) 'Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets', *Cell*, 161(5), pp. 1202–1214. Available at: https://doi.org/10.1016/j.cell.2015.05.002.

- Mammoto, A., Mammoto, T. and Ingber, D.E. (2012) 'Mechanosensitive mechanisms in transcriptional regulation', *Journal of Cell Science*, p. jcs.093005. Available at: https://doi.org/10.1242/jcs.093005.

- Manrao, E.A. *et al.* (2012) 'Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase', *Nature Biotechnology*, 30(4), pp. 349–353. Available at: https://doi.org/10.1038/nbt.2171.

- Marable, S.S. *et al.* (2018) 'Hnf4a deletion in the mouse kidney phenocopies Fanconi renotubular syndrome', *JCI Insight*, 3(14), p. e97497. Available at: https://doi.org/10.1172/jci.insight.97497.

- Marion-Poll, L. *et al.* (2014) 'Fluorescence-activated sorting of fixed nuclei: a general method for studying nuclei from specific cell populations that preserves post-translational modifications', *European Journal of Neuroscience*, 39(7), pp. 1234–1244. Available at: https://doi.org/10.1111/ejn.12506.

- Martin, B.K. *et al.* (2021) 'An optimized protocol for single cell transcriptional profiling by combinatorial indexing', *arXiv* [Preprint]. Available at: https://doi.org/10.48550/arXiv.2110.15400.

- Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing reads', *EMBnet.journal*, 17(1), pp. 10–12. Available at: https://doi.org/DOI: https://doi.org/10.14806/ej.17.1.200.

- Martins, R.P. *et al.* (2012) 'Mechanical Regulation of Nuclear Structure and Function', *Annual Review of Biomedical Engineering*, 14(1), pp. 431–455. Available at: https://doi.org/10.1146/annurev-bioeng-071910-124638.

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

- Massoni-Badosa, R. *et al.* (2020) 'Sampling time-dependent artifacts in single-cell genomics studies', *Genome Biology*, 21(1), p. 112. Available at: https://doi.org/10.1186/s13059-020-02032-0.

- Maxam, A.M. and Gilbert, W. (1977) 'A new method for sequencing DNA.', *Proceedings of the National Academy of Sciences*, 74(2), pp. 560–564. Available at: https://doi.org/10.1073/pnas.74.2.560.

- McInnes, L., Healy, J. and Melville, J. (2020) 'UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction'. arXiv: arXiv. Available at: http://arxiv.org/abs/1802.03426.

- Medini, H., Cohen, T. and Mishmar, D. (2021) 'Mitochondrial gene expression in single cells shape pancreatic beta cells' sub-populations and explain variation in insulin pathway', *Scientific Reports*, 11(1), p. 466. Available at: https://doi.org/10.1038/s41598-020-80334-w.

- Molinaro, A.M. and Pearson, B.J. (2016) 'In silico lineage tracing through single cell transcriptomics identifies a neural stem cell population in planarians', *Genome Biology*, 17(1), p. 87. Available at: https://doi.org/10.1186/s13059-016-0937-9.

- Moritz, S. *et al.* (2012) 'Heterogeneity of planarian stem cells in the S/G2/M phase', *The International Journal of Developmental Biology*, 56(1-2–3), pp. 117–125. Available at: https://doi.org/10.1387/ijdb.113440sm.

- Moustafa, K. and Cross, J. (2016) 'Genetic Approaches to Study Plant Responses to Environmental Stresses: An Overview', *Biology*, 5(2), p. 20. Available at: https://doi.org/10.3390/biology5020020.

- Müller, O. (1774) 'Polycelis nigra', in *GBIF Backbone Taxonomy*. GBIF Secretariat (2022).

- Musser, J.M. *et al.* (2021) 'Profiling cellular diversity in sponges informs animal cell type and nervous system evolution', *Science*, 374(6568), pp. 717–723. Available at: https://doi.org/10.1126/science.abj2949.

- Nagalakshmi, U. *et al.* (2008) 'The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing', *Science*, 320(5881), pp. 1344–1349. Available at: https://doi.org/10.1126/science.1158441.

- Nakagawa, H. *et al.* (2012) 'Drpiwi-1 is essential for germline cell formation during sexualization of the planarian Dugesia ryukyuensis', *Developmental Biology*, 361(1), pp. 167–176. Available at: https://doi.org/10.1016/j.ydbio.2011.10.014.

- Natsidis, P. *et al.* (2019) 'Computational discovery of hidden breaks in 28S ribosomal RNAs across eukaryotes and consequences for RNA Integrity Numbers', *Scientific Reports*, 9(1), p. 19477. Available at: https://doi.org/10.1038/s41598-019-55573-1.

- Newmark, P.A. and Sánchez Alvarado, A. (2000) 'Bromodeoxyuridine Specifically Labels the Regenerative Stem Cells of Planarians', *Developmental Biology*, 220(2), pp. 142–153. Available at: https://doi.org/10.1006/dbio.2000.9645.

- Newmark, P.A., Wang, Y. and Chong, T. (2008) 'Germ Cell Specification and Regeneration in Planarians', *Cold Spring Harbor Symposia on Quantitative Biology*, 73(0), pp. 573–581. Available at: https://doi.org/10.1101/sqb.2008.73.022.

- Nguyen, Q.H. *et al.* (2018) 'Experimental Considerations for Single-Cell RNA Sequencing Approaches', *Frontiers in Cell and Developmental Biology*, 6, p. 108. Available at: https://doi.org/10.3389/fcell.2018.00108.

- Nishimura, O. *et al.* (2015) 'Unusually Large Number of Mutations in Asexually Reproducing Clonal Planarian Dugesia japonica', *PLOS ONE*. Edited by B. Fu, 10(11), p. e0143525. Available at: https://doi.org/10.1371/journal.pone.0143525.

- O'Flanagan, C.H. *et al.* (2019) 'Dissociation of solid tumor tissues with cold active protease for single-cell RNA-seq minimizes conserved collagenase-associated stress responses', *Genome Biology*, 20(1), p. 210. Available at: https://doi.org/10.1186/s13059-019-1830-0.

- Önal, P. *et al.* (2012) 'Gene expression of pluripotency determinants is conserved between mammalian and planarian stem cells: Conserved determinants of planarian pluripotency', *The EMBO Journal*, 31(12), pp. 2755–2769. Available at: https://doi.org/10.1038/emboj.2012.110.

- Packer, J.S. *et al.* (2019) 'A lineage-resolved molecular atlas of *C. elegans* embryogenesis at single-cell resolution', *Science*, 365(6459), p. eaax1971. Available at: https://doi.org/10.1126/science.aax1971.

- Padmanaban, A., Salowsky, R. and Cher, C. (2012) 'RNA quality control using the agilent 2200 TapeStation system–assessment of the RIN e quality metric', *Agilent technologies application notes* [Preprint].

- Pan, Q. *et al.* (2008) 'Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing', *Nature Genetics*, 40(12), pp. 1413–1415. Available at: https://doi.org/10.1038/ng.259.

- Paridaen, J.T. and Huttner, W.B. (2014) 'Neurogenesis during development of the vertebrate central nervous system', *EMBO reports*, 15(4), pp. 351–364. Available at: https://doi.org/10.1002/embr.201438447.

- Parkinson, J. and Blaxter, M. (2009) 'Expressed Sequence Tags: An Overview', in J. Parkinson (ed.) *Expressed Sequence Tags (ESTs)*. Totowa, NJ: Humana Press (Methods in Molecular Biology), pp. 1–12. Available at: https://doi.org/10.1007/978-1-60327-136-3_1.

- Patro, R. *et al.* (2017) 'Salmon provides fast and bias-aware quantification of transcript expression', *Nature Methods*, 14(4), pp. 417–419. Available at: https://doi.org/10.1038/nmeth.4197.

- Patro, R., Mount, S.M. and Kingsford, C. (2014) 'Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms', *Nature Biotechnology*, 32(5), pp. 462–464. Available at: https://doi.org/10.1038/nbt.2862.

- Pearson, B.J. *et al.* (2009) 'Formaldehyde-based whole-mount in situ hybridization method for planarians', *Developmental Dynamics*, 238(2), pp. 443–450. Available at: https://doi.org/10.1002/dvdy.21849.

- Pedersen, K.J. (1961) 'Studies on the nature of planarian connective tissue', *Zeitschrift für Zellforschung und Mikroskopische Anatomie*, 53(5), pp. 569–608. Available at: https://doi.org/10.1007/BF00339508.

- Pertea, M. *et al.* (2016) 'Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown', *Nature Protocols*, 11(9), pp. 1650–1667. Available at: https://doi.org/10.1038/nprot.2016.095.

- Peterson, V.M. *et al.* (2017) 'Multiplexed quantification of proteins and transcripts in single cells', *Nature Biotechnology*, 35(10), pp. 936–939. Available at: https://doi.org/10.1038/nbt.3973.

- Picelli, S. *et al.* (2013) 'Smart-seq2 for sensitive full-length transcriptome profiling in single cells', *Nature Methods*, 10(11), pp. 1096–1098. Available at: https://doi.org/10.1038/nmeth.2639.

- Piétu, G. *et al.* (1999) 'The Genexpress IMAGE Knowledge Base of the Human Brain Transcriptome: A Prototype Integrated Resource for Functional and Computational Genomics', *Genome Research*, 9(2), pp. 195–209. Available at: https://doi.org/10.1101/gr.9.2.195.

- Plass, M. *et al.* (2018) 'Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics', *Science*, 360(6391), p. eaaq1723. Available at: https://doi.org/10.1126/science.aaq1723.

- Prestin, K. *et al.* (2014) 'Transcriptional regulation of urate transportosome member SLC2A9 by nuclear receptor HNF4α', *American Journal of Physiology-Renal Physiology*, 307(9), pp. F1041–F1051. Available at: https://doi.org/10.1152/ajprenal.00640.2013.

- Putney, S.D., Herlihy, W.C. and Schimmel, P. (1983) 'A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing', *Nature*, 302, pp. 718–721.

- Qiu, X. *et al.* (2015) 'Microfluidic device for mechanical dissociation of cancer cell aggregates into single cells', *Lab on a Chip*, 15(1), pp. 339–350. Available at: https://doi.org/10.1039/C4LC01126K.

- Qiu, X. *et al.* (2017) 'Single-cell mRNA quantification and differential analysis with Census', *Nature Methods*, 14(3), pp. 309–315. Available at: https://doi.org/10.1038/nmeth.4150.

- Quick, J. *et al.* (2016) 'Real-time, portable genome sequencing for Ebola surveillance', *Nature*, 530(7589), pp. 228–232. Available at: https://doi.org/10.1038/nature16996.

- Ramsköld, D. *et al.* (2012) 'Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells', *Nature Biotechnology*, 30(8), pp. 777–782. Available at: https://doi.org/10.1038/nbt.2282.

- Ranzoni, A.M. *et al.* (2021) 'Integrative Single-cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Haematopoiesis', *Cell Stem Cell*, 28(3), pp. 472-487.e7. Available at: https://doi.org/10.1016/j.stem.2020.11.015.

- Raz, A.A., Wurtzel, O. and Reddien, P.W. (2021) 'Planarian stem cells specify fate yet retain potency during the cell cycle', *Cell Stem Cell*, 28(7), pp. 1307-1322.e5. Available at: https://doi.org/10.1016/j.stem.2021.03.021.

- Rebboah, E. *et al.* (2021) 'Mapping and modeling the genomic basis of differential RNA isoform expression at single-cell resolution with LR-Split-seq', *Genome Biology*, 22(1), p. 286. Available at: https://doi.org/10.1186/s13059-021-02505-w.

- Reddien, P.W. *et al.* (2005) 'SMEDWI-2 Is a PIWI-Like Protein That Regulates Planarian Stem Cells', *Science*, 310(5752), pp. 1327–1330. Available at: https://doi.org/10.1126/science.1116110.

- Reichard, A. and Asosingh, K. (2019) 'Best Practices for Preparing a Single Cell Suspension from Solid Tissues for Flow Cytometry', *Cytometry Part A*, 95(2), pp. 219–226. Available at: https://doi.org/10.1002/cyto.a.23690.

- Renner, W.A. *et al.* (1993) 'Cell-cell adhesion and aggregation: Influence on the growth behavior of CHO cells', *Biotechnology and Bioengineering*, 41(2), pp. 188–193. Available at: https://doi.org/10.1002/bit.260410204.

- Rink, J.C. (2013) 'Stem cell systems and regeneration in planaria', *Development Genes and Evolution*, 223(1–2), pp. 67–84. Available at: https://doi.org/10.1007/s00427-012-0426-4.

- Ritchie, M.E. *et al.* (2015) 'limma powers differential expression analyses for RNA-sequencing and microarray studies', *Nucleic Acids Research*, 43(7), pp. e47–e47. Available at: https://doi.org/10.1093/nar/gkv007.

- Riutort, M. *et al.* (2012) 'Evolutionary history of the Tricladida and the Platyhelminthes: an up-to-date phylogenetic and systematic account', *The International Journal of Developmental Biology*, 56(1-2–3), pp. 5–17. Available at: https://doi.org/10.1387/ijdb.113441mr.

- Roberts-Galbraith, R.H., Brubacher, J.L. and Newmark, P.A. (2016) 'A functional genomics screen in planarians reveals regulators of whole-brain regeneration', *eLife*, 5, p. e17002. Available at: https://doi.org/10.7554/eLife.17002.

- Robertson, G. *et al.* (2007) 'Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing', *Nature Methods*, 4(8), pp. 651–657. Available at: https://doi.org/10.1038/nmeth1068.

- Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2010) 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*, 26(1), pp. 139–140. Available at: https://doi.org/10.1093/bioinformatics/btp616.

- Rodriques, S.G. *et al.* (2019) 'Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution', *Science*, 363, pp. 1436–1467. Available at: https://doi.org/10.1126/science.aaw1219.

- Romero, B.T., Evans, D.J. and Aboobaker, A.A. (2012) 'FACS Analysis of the Planarian Stem Cell Compartment as a Tool to Understand Regenerative Mechanisms', in K.A. Mace and K.M. Braun (eds) *Progenitor Cells*. Totowa, NJ: Humana Press (Methods in Molecular Biology), pp. 167–179. Available at: https://doi.org/10.1007/978-1-61779-980-8_13.

- Romero, R. and Baguñà, J. (1991) 'Quantitative cellular analysis of growth and reproduction in freshwater planarians (Turbellaria; Tricladida). I. A cellular description of the intact organism', *Invertebrate Reproduction & Development*, 19(2), pp. 157–165. Available at: https://doi.org/10.1080/07924259.1991.9672170.

- Ronald Sluys and Marta Riutort (2018) 'Planarian Diversity and Phylogeny', *Methods Mol Biol*, 1774, pp. 1–56. Available at: https://doi.org/10.1007/978-1-4939-7802-1_1.

- Rosenberg, A.B. *et al.* (2018) 'Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding', *Science*, 360(6385), pp. 176–182. Available at: https://doi.org/10.1126/science.aam8999.

- Ross, J. *et al.* (1972) 'In Vitro Synthesis of DNA Complementary to Purified Rabbit Globin mRNA', *Proceedings of the National Academy of Sciences*, 69(1), pp. 264–268. Available at: https://doi.org/10.1073/pnas.69.1.264.

- Rostom, R. *et al.* (2017) 'Computational approaches for interpreting scRNA-seq data', *FEBS Letters*, 591(15), pp. 2213–2225. Available at: https://doi.org/10.1002/1873-3468.12684.

- Rouhana, L. *et al.* (2013) 'RNA interference by feeding in vitro–synthesized double-stranded RNA to planarians: Methodology and dynamics', *Developmental Dynamics*, 242(6), pp. 718–730. Available at: https://doi.org/10.1002/dvdy.23950.

- Rouhana, L. *et al.* (2017) 'Genetic dissection of the planarian reproductive system through characterization of Schmidtea mediterranea CPEB homologs', *Developmental Biology*, 426(1), pp. 43–55. Available at: https://doi.org/10.1016/j.ydbio.2017.04.008.

- Russell, J.N., Clements, J.E. and Gama, L. (2013) 'Quantitation of Gene Expression in Formaldehyde-Fixed and Fluorescence-Activated Sorted Cells', *PLoS ONE*. Edited by Y. Wu, 8(9), p. e73849. Available at: https://doi.org/10.1371/journal.pone.0073849.

- Ryffel, G. (2001) 'Mutations in the human genes encoding the transcription factors of the hepatocyte nuclear factor (HNF)1 and HNF4 families: functional and pathological consequences', *Journal of Molecular Endocrinology*, 27(1), pp. 11–29. Available at: https://doi.org/10.1677/jme.0.0270011.

- Saiki, R.K. *et al.* (1988) 'Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase', *Science*, 239, pp. 487–491. Available at: https://doi.org/10.1126/science.2448875.

- Saló, E. (2006) 'The power of regeneration and the stem-cell kingdom: freshwater planarians (Platyhelminthes)', *BioEssays*, 28(5), pp. 546–559. Available at: https://doi.org/10.1002/bies.20416.

- Sánchez Alvarado, A. and Newmark, P.A. (1999) 'Double-stranded RNA specifically disrupts gene expression during planarian regeneration', *Proceedings of the National Academy of Sciences*, 96(9), pp. 5049–5054. Available at: https://doi.org/10.1073/pnas.96.9.5049.

- Sanger, F., Nicklen, S. and Coulson, A.R. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the National Academy of Sciences*, 74(12), pp. 5463–5467. Available at: https://doi.org/10.1073/pnas.74.12.5463.

- Satija, R. *et al.* (2015) 'Spatial reconstruction of single-cell gene expression data', *Nature Biotechnology*, 33(5), pp. 495–502. Available at: https://doi.org/10.1038/nbt.3192.

- Schena, M. *et al.* (1995) 'Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray', *Science*, 270(5235), pp. 467–470. Available at: https://doi.org/10.1126/science.270.5235.467.

- Schmidt, O. (1861) 'Uber Planaria torva Autorum', *Z wiss Zool*, 11(1), pp. 89–94.

- Schneider, K.C. (1890) 'Histologie von Hydra fusca mit besonderer Berücksichtigung des Nervensystems der Hydropolypen', *Archiv für mikroskopische Anatomie*, 35(1), pp. 321–379.

- Schroeder, A. *et al.* (2006) 'The RIN: an RNA integrity number for assigning integrity values to RNA measurements', *BMC Molecular Biology*, 7(1), p. 3. Available at: https://doi.org/10.1186/1471-2199-7-3.

- Scimone, M.L. *et al.* (2011) 'A regulatory program for excretory system regeneration in planarians', *Development*, 138(20), pp. 4387–4398. Available at: https://doi.org/10.1242/dev.068098.

- Scimone, M.L. *et al.* (2014) 'Neoblast Specialization in Regeneration of the Planarian Schmidtea mediterranea', *Stem Cell Reports*, 3(2), pp. 339–352. Available at: https://doi.org/10.1016/j.stemcr.2014.06.001.

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

- Sebé-Pedrós, A., Saudemont, B., *et al.* (2018) 'Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq', *Cell*, 173(6), pp. 1520-1534.e20. Available at: https://doi.org/10.1016/j.cell.2018.05.019.

- Sebé-Pedrós, A., Chomsky, E., *et al.* (2018) 'Early metazoan cell type diversity and the evolution of multicellular gene regulation', *Nature Ecology & Evolution*, 2(7), pp. 1176–1188. Available at: https://doi.org/10.1038/s41559-018-0575-6.

- Shafer, M.E.R. (2019) 'Cross-Species Analysis of Single-Cell Transcriptomic Data', *Frontiers in Cell and Developmental Biology*, 7, p. 175. Available at: https://doi.org/10.3389/fcell.2019.00175.

- Shih, D.Q. *et al.* (2000) 'Genotype/phenotype relationships in HNF-4alpha/MODY1: haploinsufficiency is associated with reduced apolipoprotein (AII), apolipoprotein (CIII), lipoprotein(a), and triglyceride levels.', *Diabetes*, 49(5), pp. 832–837. Available at: https://doi.org/10.2337/diabetes.49.5.832.

- Shiraki, T. *et al.* (2003) 'Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage', *Proceedings of the National Academy of Sciences*, 100(26), pp. 15776–15781. Available at: https://doi.org/10.1073/pnas.2136655100.

- Siebert, S. *et al.* (2019) 'Stem cell differentiation trajectories in *Hydra* resolved at single-cell resolution', *Science*, 365(6451), p. eaav9314. Available at: https://doi.org/10.1126/science.aav9314.

- Simão, F.A. *et al.* (2015) 'BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs', *Bioinformatics*, 31(19), pp. 3210–3212. Available at: https://doi.org/10.1093/bioinformatics/btv351.

- Skene, P.J. and Henikoff, S. (2017) 'An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites', *eLife*, 6, p. e21856. Available at: https://doi.org/10.7554/eLife.21856.

- Sluys, R. *et al.* (2009) 'A new higher classification of planarian flatworms (Platyhelminthes, Tricladida)', *Journal of Natural History*, 43(29–30), pp. 1763–1777. Available at: https://doi.org/10.1080/00222930902741669.

- Smith, A.M. *et al.* (2019) 'Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing', *PLOS ONE*. Edited by H.-J. Wieden, 14(5), p. e0216709. Available at: https://doi.org/10.1371/journal.pone.0216709.

- Solana, J. *et al.* (2012) 'Defining the molecular profile of planarian pluripotent stem cells using a combinatorial RNA-seq, RNA interference and irradiation approach', *Genome Biology*, 13(3), p. R19. Available at: https://doi.org/10.1186/gb-2012-13-3-r19.

- Solana, J. *et al.* (2016) 'Conserved functional antagonism of CELF and MBNL proteins controls stem cell-specific alternative splicing in planarians', *eLife*, 5, p. e16797. Available at: https://doi.org/10.7554/eLife.16797.

- Song, H. *et al.* (2020) 'HNF4A-AS1/hnRNPU/CTCF axis as a therapeutic target for aerobic glycolysis and neuroblastoma progression', *Journal of Hematology & Oncology*, 13(1), p. 24. Available at: https://doi.org/10.1186/s13045-020-00857-7.

- Southern, E.M. (1975) 'Detection of Specific Sequences Among DNA Fragments Separated by Gel Electrophoresis', *J. Mol. Biol.*, 98, pp. 503–517.

- Spiegelman, S., Watson, K.F. and Kacian, D.L. (1971) 'Synthesis of DNA Complements of Natural RNAs: A General Approach', *Proceedings of the National Academy of Sciences*, 68(11), pp. 2843–2845. Available at: https://doi.org/10.1073/pnas.68.11.2843.

- Ståhl, P.L. *et al.* (2016) 'Visualization and analysis of gene expression in tissue sections by spatial transcriptomics', *Science*, 353(6294), pp. 78–82. Available at: https://doi.org/10.1126/science.aaf2403.

- Stark, R., Grzelak, M. and Hadfield, J. (2019) 'RNA sequencing: the teenage years', *Nature Reviews Genetics*, 20(11), pp. 631–656. Available at: https://doi.org/10.1038/s41576-019-0150-2.

- Steiner, J.K., Tasaki, J. and Rouhana, L. (2016) 'Germline Defects Caused by Smed-boule RNA-Interference Reveal That Egg Capsule Deposition Occurs Independently of Fertilization, Ovulation, Mating, or the Presence of Gametes in Planarian Flatworms', *PLOS Genetics*. Edited by R.S. Hawley, 12(5), p. e1006030. Available at: https://doi.org/10.1371/journal.pgen.1006030.

- Stickels, R.R. *et al.* (2021) 'Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2', *Nature Biotechnology*, 39(3), pp. 313–319. Available at: https://doi.org/10.1038/s41587-020-0739-1.

- Stocchino, G.A. *et al.* (2019) 'The invasive alien freshwater flatworm Girardia tigrina (Girard, 1850) (Platyhelminthes, Tricladida) in Western Europe: new insights into its morphology, karyology and reproductive biology', *Contributions to Zoology*, pp. 1–21. Available at: https://doi.org/10.1163/18759866-20191406.

- Stoeckius, M. *et al.* (2017) 'Simultaneous epitope and transcriptome measurement in single cells', *Nature Methods*, 14(9), pp. 865–868. Available at: https://doi.org/10.1038/nmeth.4380.

- Stuart, T. *et al.* (2019) 'Comprehensive Integration of Single-Cell Data', *Cell*, 177(7), pp. 1888-1902.e21. Available at: https://doi.org/10.1016/j.cell.2019.05.031.

- Svensson, V., da Veiga Beltrame, E. and Pachter, L. (2020) 'A curated database reveals trends in single-cell transcriptomics', *Database*, 2020, baaa073. Available at: https://doi.org/10.1093/database/baaa073.

- Svensson, V., Vento-Tormo, R. and Teichmann, S.A. (2018) 'Exponential scaling of single-cell RNA-seq in the past decade', *Nature Protocols*, 13(4), pp. 599–604. Available at: https://doi.org/10.1038/nprot.2017.149.

- Takeichi, M. and Okada, T.S. (1972) 'Roles of magnesium and calcium ions in cell-to-substrate adhesion', *Experimental Cell Research*, 74(1), pp. 51–60. Available at: https://doi.org/10.1016/0014-4827(72)90480-6.

- Tamura, S., Oki, I. and Kawakatsu, M. (1995) 'A review of chromosomal variation in Dugesia japonica and D. ryukyuensis in the Far East', *Hydrobiologia*, 305, pp. 79–84. Available at: https://doi.org/10.1007/BF00036366.

- Tang, F. *et al.* (2009) 'mRNA-Seq whole-transcriptome analysis of a single cell', *Nature Methods*, 6(5), pp. 377–382. Available at: https://doi.org/10.1038/nmeth.1315.

- Tarashansky, A.J. *et al.* (2021) 'Mapping single-cell atlases throughout Metazoa unravels cell type evolution', *eLife*, 10, p. e66747. Available at: https://doi.org/10.7554/eLife.66747.

- Thommen, A. *et al.* (2019) 'Body size-dependent energy storage causes Kleiber's law scaling of the metabolic rate in planarians', *eLife*, 8, p. e38187. Available at: https://doi.org/10.7554/eLife.38187.

- Thomsen, E.R. *et al.* (2016) 'Fixed single-cell transcriptomic characterization of human radial glial diversity', *Nature Methods*, 13(1), pp. 87–93. Available at: https://doi.org/10.1038/nmeth.3629.

- Tian, Q. *et al.* (2022) 'Whole-genome sequence of the planarian Dugesia japonica combining Illumina and PacBio data', *Genomics*, 114(2), p. 110293. Available at: https://doi.org/10.1016/j.ygeno.2022.110293.

- Ton, M.-L.N. *et al.* (2022) *Rabbit Development as a Model for Single Cell Comparative Genomics*. preprint. Developmental Biology. Available at: https://doi.org/10.1101/2022.10.06.510971.

- Tosches, M.A. *et al.* (2018) 'Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles', *Science*, 360(6391), pp. 881–888. Available at: https://doi.org/10.1126/science.aar4237.

- Traag, V.A., Waltman, L. and van Eck, N.J. (2019) 'From Louvain to Leiden: guaranteeing well-connected communities', *Scientific Reports*, 9(1), p. 5233. Available at: https://doi.org/10.1038/s41598-019-41695-z.

- Trapnell, C. *et al.* (2010) 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation', *Nature Biotechnology*, 28(5), pp. 511–515. Available at: https://doi.org/10.1038/nbt.1621.

- Trapnell, C. *et al.* (2014) 'The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells', *Nature Biotechnology*, 32(4), pp. 381–386. Available at: https://doi.org/10.1038/nbt.2859.

- Tümer, E. *et al.* (2013) 'Enterocyte-specific Regulation of the Apical Nutrient Transporter SLC6A19 (B0AT1) by Transcriptional and Epigenetic Networks', *Journal of Biological Chemistry*, 288(47), pp. 33813–33823. Available at: https://doi.org/10.1074/jbc.M113.482760.

- Urakami, T. (2019) 'Maturity-onset diabetes of the young (MODY): current perspectives on diagnosis and treatment', *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, Volume 12, pp. 1047–1056. Available at: https://doi.org/10.2147/DMSO.S179793.

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

- Velculescu, V.E. *et al.* (1995) 'Serial Analysis of Gene Expression', *Science*, 270, pp. 484–487.

- Vento-Tormo, R. *et al.* (2018) 'Single-cell reconstruction of the early maternal–fetal interface in humans', *Nature*, 563(7731), pp. 347–353. Available at: https://doi.org/10.1038/s41586-018-0698-6.

- Vial, J. and Porter, K.R. (1975) 'Scanning microscopy of dissociated tissue cells', *Journal of Cell Biology*, 67(2), pp. 345–360. Available at: https://doi.org/10.1083/jcb.67.2.345.

- Vila-Farré, M. *et al.* (2011) 'Freshwater planarians (Platyhelminthes, Tricladida) from the Iberian Peninsula and Greece: diversity and notes on ecology', *Zootaxa*, 2779(1), p. 1. Available at: https://doi.org/10.11646/zootaxa.2779.1.1.

- Vuong, L.M. *et al.* (2015) 'Differential Effects of Hepatocyte Nuclear Factor 4α Isoforms on Tumor Growth and T-Cell Factor 4/AP-1 Interactions in Human Colorectal Cancer Cells', *Molecular and Cellular Biology*, 35(20), pp. 3471–3490. Available at: https://doi.org/10.1128/MCB.00030-15.

- Wagner, D.E., Wang, I.E. and Reddien, P.W. (2011) 'Clonogenic Neoblasts Are Pluripotent Adult Stem Cells That Underlie Planarian Regeneration', *Science*, 332(6031), pp. 811–816. Available at: https://doi.org/10.1126/science.1203983.

- Wang, H. *et al.* (2000) 'Hepatocyte Nuclear Factor 4α Regulates the Expression of Pancreatic β-Cell Genes Implicated in Glucose Metabolism and Nutrient-induced Insulin Secretion', *Journal of Biological Chemistry*, 275(46), pp. 35953–35959. Available at: https://doi.org/10.1074/jbc.M006612200.

- Wang, Y. *et al.* (2007) '*nanos* function is essential for development and regeneration of planarian germ cells', *Proceedings of the National Academy of Sciences*, 104(14), pp. 5901–5906. Available at: https://doi.org/10.1073/pnas.0609708104.

- Wang, Y. *et al.* (2010) 'A functional genomic screen in planarians identifies novel regulators of germ cell development', *Genes & Development*, 24(18), pp. 2081–2092. Available at: https://doi.org/10.1101/gad.1951010.

- Wang, Z., Gerstein, M. and Snyder, M. (2009) 'RNA-Seq: a revolutionary tool for transcriptomics', *Nature Reviews Genetics*, 10(1), pp. 57–63. Available at: https://doi.org/10.1038/nrg2484.

- Welch, J.D. *et al.* (2019) 'Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity', *Cell*, 177(7), pp. 1873-1887.e17. Available at: https://doi.org/10.1016/j.cell.2019.05.006.

- Westermann, A.J. *et al.* (2016) 'Dual RNA-seq unveils noncoding RNA functions in host–pathogen interactions', *Nature*, 529(7587), pp. 496–501. Available at: https://doi.org/10.1038/nature16547.

- Wolf, F.A. *et al.* (2019) 'PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells', *Genome Biology*, 20(1), p. 59. Available at: https://doi.org/10.1186/s13059-019-1663-x.

- Wolf, F.A., Angerer, P. and Theis, F.J. (2018) 'SCANPY: large-scale single-cell gene expression data analysis', *Genome Biology*, 19(1), p. 15. Available at: https://doi.org/10.1186/s13059-017-1382-0.

- Wolock, S.L., Lopez, R. and Klein, A.M. (2019) 'Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data', *Cell Systems*, 8(4), pp. 281-291.e9. Available at: https://doi.org/10.1016/j.cels.2018.11.005.

- Wurtzel, O. *et al.* (2015) 'A Generic and Cell-Type-Specific Wound Response Precedes Regeneration in Planarians', *Developmental Cell*, 35(5), pp. 632–645. Available at: https://doi.org/10.1016/j.devcel.2015.11.004.

- Xie, Y. *et al.* (2014) 'SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads', *Bioinformatics*, 30(12), pp. 1660–1666. Available at: https://doi.org/10.1093/bioinformatics/btu077.

- Yamamoto, H. and Agata, K. (2011) 'Optic chiasm formation in planarian I: Cooperative netrin- and robo-mediated signals are required for the early stage of optic chiasm formation: Optic chiasm formation in planarian', *Development, Growth & Differentiation*, 53(3), pp. 300–311. Available at: https://doi.org/10.1111/j.1440-169X.2010.01234.x.

- Yang, M., Dutta, C. and Tiwari, A. (2015) 'Disulfide-Bond Scrambling Promotes Amorphous Aggregates in Lysozyme and Bovine Serum Albumin', *The Journal of Physical Chemistry B*, 119(10), pp. 3969–3981. Available at: https://doi.org/10.1021/acs.jpcb.5b00144.

- Zaghlool, A. *et al.* (2021) 'Characterization of the nuclear and cytosolic transcriptomes in human brain tissue reveals new insights into the subcellular distribution of RNA transcripts', *Scientific Reports*, 11(1), p. 4076. Available at: https://doi.org/10.1038/s41598-021-83541-1.

- Zayas, R.M. *et al.* (2005) 'The planarian Schmidtea mediterranea as a model for epigenetic germ cell specification: Analysis of ESTs from the hermaphroditic strain.', *Proc. Nat. Acad. Sci. USA*, 102(51), pp. 18491–6. Available at: https://doi.org/10.1073/pnas.0509507102.

- Zeng, A. *et al.* (2018) 'Prospectively Isolated Tetraspanin+ Neoblasts Are Adult Pluripotent Stem Cells Underlying Planaria Regeneration', *Cell*, 173(7), pp. 1593-1608.e20. Available at: https://doi.org/10.1016/j.cell.2018.05.006.

- Zeng, J. *et al.* (2015) 'A Minimally Invasive Method for Retrieving Single Adherent Cells of Different Types from Cultures', *Scientific Reports*, 4(1), p. 5424. Available at: https://doi.org/10.1038/srep05424.

- Zhao, L. *et al.* (2019) 'Analysis of Transcriptome and Epitranscriptome in Plants Using PacBio Iso-Seq and Nanopore-Based Direct RNA Sequencing', *Frontiers in Genetics*, 10, p. 253. Available at: https://doi.org/10.3389/fgene.2019.00253.

- Zheng, G.X.Y. *et al.* (2017) 'Massively parallel digital transcriptional profiling of single cells', *Nature Communications*, 8(1), p. 14049. Available at: https://doi.org/10.1038/ncomms14049.

- Zhu, S.J. and Pearson, B.J. (2016) '(Neo)blast from the past: new insights into planarian stem cell lineages', *Current Opinion in Genetics & Development*, 40, pp. 74–80. Available at: https://doi.org/10.1016/j.gde.2016.06.007.

- Ziegenhain, C. *et al.* (2017) 'Comparative Analysis of Single-Cell RNA Sequencing Methods', *Molecular Cell*, 65(4), pp. 631-643.e4. Available at: https://doi.org/10.1016/j.molcel.2017.01.023.

SUPPLEMENTARY MATERIALS

SUPPLEMENTARY 1

GENERAL OLIGONUCLEOTIDES

Table S1.1 General oligonucleotides used in SPLiT-seq. Names, working concentrations (WC), oligo descriptions and sequences are indicated.

Oligo Name	WC	Description	Sequence
Linker_1	100 µM	Link Round 1 & Round 2 Barcodes	CGAATGCTCTGGCCTCTCAAGCACGTGGAT
Linker_2	100 µM	Link Round 2 & Round 3 Barcodes	AGTCGTACGCCGATGCGAAACATCGGCCAC
Blocker_1	100 µM	Block Linker_1	ATCCACGTGCTTGAGAGGCCAGAGCATTCG
Blocker_2	100 µM	Block Linker_2	GTGGCCGATGTTTCGCATCGGCGTACGACT
TSO	100 µM	Secure the 5´end (HPLC purified)	AAGCAGTGGTATCAACGCAGAGTGAATrGrG+G
PCR_PR	10 µM	Bind to Round 3 Barcodes	CAGACGTGTGCTCTTCCGATCT
PCR_PF	10 µM	Bind to TSO	AAGCAGTGGTATCAACGCAGAGT
P5_oligo	10 µM	P5 Illumina adapter (N501)	AATGATACGGCGACCACCGAGATCTACACTAGATCGCTCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
Round 4 Barcode 1	10 µM	**Barcode 4** + P7 Illumina adapter (TSBC07)	CAAGCAGAAGACGGCATACGAGATGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Round 4 Barcode 2	10 µM	**Barcode 4** + P7 Illumina adapter (TSBC08)	CAAGCAGAAGACGGCATACGAGATTCAAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Round 4 Barcode 3	10 µM	**Barcode 4** + P7 Illumina adapter (TSBC09)	CAAGCAGAAGACGGCATACGAGATCTGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Round 4 Barcode 4	10 µM	**Barcode 4** + P7 Illumina adapter (TSBC10)	CAAGCAGAAGACGGCATACGAGATAAGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Round 4 Barcode 5	10 µM	**Barcode 4** + P7 Illumina adapter (TSBC11)	CAAGCAGAAGACGGCATACGAGATGTAGCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Round 4 Barcode 6	10 µM	**Barcode 4** + P7 Illumina adapter (TSBC12)	CAAGCAGAAGACGGCATACGAGATTACAAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Round 4 Barcode 7	10 µM	**Barcode 4** + P7 Illumina adapter (TSBC13)	CAAGCAGAAGACGGCATACGAGATTTGACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Round 4 Barcode 8	10 µM	**Barcode 4** + P7 Illumina adapter (TSBC14)	CAAGCAGAAGACGGCATACGAGATGGAACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

ROUND 1 BARCODES

Table S1.2 Round 1 barcodes used in SPLiT-seq.

Plate Well	Name	Sequence
A1	Round1_1	/5Phos/AGGCCAGAGCATTCG**AACGTGAT**TTTTTTTTTTTTTTTVN
A2	Round1_2	/5Phos/AGGCCAGAGCATTCG**AAACATCG**TTTTTTTTTTTTTTTVN
A3	Round1_3	/5Phos/AGGCCAGAGCATTCG**ATGCCTAA**TTTTTTTTTTTTTTTVN
A4	Round1_4	/5Phos/AGGCCAGAGCATTCG**AGTGGTCA**TTTTTTTTTTTTTTTVN
A5	Round1_5	/5Phos/AGGCCAGAGCATTCG**ACCACTGT**TTTTTTTTTTTTTTTVN
A6	Round1_6	/5Phos/AGGCCAGAGCATTCG**ACATTGGC**TTTTTTTTTTTTTTTVN
A7	Round1_7	/5Phos/AGGCCAGAGCATTCG**CAGATCTG**TTTTTTTTTTTTTTTVN
A8	Round1_8	/5Phos/AGGCCAGAGCATTCG**CATCAAGT**TTTTTTTTTTTTTTTVN
A9	Round1_9	/5Phos/AGGCCAGAGCATTCG**CGCTGATC**TTTTTTTTTTTTTTTVN
A10	Round1_10	/5Phos/AGGCCAGAGCATTCG**ACAAGCTA**TTTTTTTTTTTTTTTVN
A11	Round1_11	/5Phos/AGGCCAGAGCATTCG**CTGTAGCC**TTTTTTTTTTTTTTTVN
A12	Round1_12	/5Phos/AGGCCAGAGCATTCG**AGTACAAG**TTTTTTTTTTTTTTTVN
B1	Round1_13	/5Phos/AGGCCAGAGCATTCG**AACAACCA**TTTTTTTTTTTTTTTVN
B2	Round1_14	/5Phos/AGGCCAGAGCATTCG**AACCGAGA**TTTTTTTTTTTTTTTVN
B3	Round1_15	/5Phos/AGGCCAGAGCATTCG**AACGCTTA**TTTTTTTTTTTTTTTVN
B4	Round1_16	/5Phos/AGGCCAGAGCATTCG**AAGACGGA**TTTTTTTTTTTTTTTVN
B5	Round1_17	/5Phos/AGGCCAGAGCATTCG**AAGGTACA**TTTTTTTTTTTTTTTVN
B6	Round1_18	/5Phos/AGGCCAGAGCATTCG**ACACAGAA**TTTTTTTTTTTTTTTVN
B7	Round1_19	/5Phos/AGGCCAGAGCATTCG**ACAGCAGA**TTTTTTTTTTTTTTTVN
B8	Round1_20	/5Phos/AGGCCAGAGCATTCG**ACCTCCAA**TTTTTTTTTTTTTTTVN
B9	Round1_21	/5Phos/AGGCCAGAGCATTCG**ACGCTCGA**TTTTTTTTTTTTTTTVN
B10	Round1_22	/5Phos/AGGCCAGAGCATTCG**ACGTATCA**TTTTTTTTTTTTTTTVN
B11	Round1_23	/5Phos/AGGCCAGAGCATTCG**ACTATGCA**TTTTTTTTTTTTTTTVN
B12	Round1_24	/5Phos/AGGCCAGAGCATTCG**AGAGTCAA**TTTTTTTTTTTTTTTVN
C1	Round1_25	/5Phos/AGGCCAGAGCATTCG**AGATCGCA**TTTTTTTTTTTTTTTVN
C2	Round1_26	/5Phos/AGGCCAGAGCATTCG**AGCAGGAA**TTTTTTTTTTTTTTTVN
C3	Round1_27	/5Phos/AGGCCAGAGCATTCG**AGTCACTA**TTTTTTTTTTTTTTTVN
C4	Round1_28	/5Phos/AGGCCAGAGCATTCG**ATCCTGTA**TTTTTTTTTTTTTTTVN
C5	Round1_29	/5Phos/AGGCCAGAGCATTCG**ATTGAGGA**TTTTTTTTTTTTTTTVN
C6	Round1_30	/5Phos/AGGCCAGAGCATTCG**CAACCACA**TTTTTTTTTTTTTTTVN
C7	Round1_31	/5Phos/AGGCCAGAGCATTCG**GACTAGTA**TTTTTTTTTTTTTTTVN
C8	Round1_32	/5Phos/AGGCCAGAGCATTCG**CAATGGAA**TTTTTTTTTTTTTTTVN
C9	Round1_33	/5Phos/AGGCCAGAGCATTCG**CACTTCGA**TTTTTTTTTTTTTTTVN
C10	Round1_34	/5Phos/AGGCCAGAGCATTCG**CAGCGTTA**TTTTTTTTTTTTTTTVN
C11	Round1_35	/5Phos/AGGCCAGAGCATTCG**CATACCAA**TTTTTTTTTTTTTTTVN
C12	Round1_36	/5Phos/AGGCCAGAGCATTCG**CCAGTTCA**TTTTTTTTTTTTTTTVN
D1	Round1_37	/5Phos/AGGCCAGAGCATTCG**CCGAAGTA**TTTTTTTTTTTTTTTVN
D2	Round1_38	/5Phos/AGGCCAGAGCATTCG**CCGTGAGA**TTTTTTTTTTTTTTTVN
D3	Round1_39	/5Phos/AGGCCAGAGCATTCG**CCTCCTGA**TTTTTTTTTTTTTTTVN
D4	Round1_40	/5Phos/AGGCCAGAGCATTCG**CGAACTTA**TTTTTTTTTTTTTTTVN
D5	Round1_41	/5Phos/AGGCCAGAGCATTCG**CGACTGGA**TTTTTTTTTTTTTTTVN
D6	Round1_42	/5Phos/AGGCCAGAGCATTCG**CGCATACA**TTTTTTTTTTTTTTTVN
D7	Round1_43	/5Phos/AGGCCAGAGCATTCG**CTCAATGA**TTTTTTTTTTTTTTTVN

D8	Round1_44	/5Phos/AGGCCAGAGCATTCGCTGAGCCATTTTTTTTTTTTTTTVN
D9	Round1_45	/5Phos/AGGCCAGAGCATTCGCTGGCATATTTTTTTTTTTTTTTVN
D10	Round1_46	/5Phos/AGGCCAGAGCATTCGGAATCTGATTTTTTTTTTTTTTTVN
D11	Round1_47	/5Phos/AGGCCAGAGCATTCGCAAGACTATTTTTTTTTTTTTTTVN
D12	Round1_48	/5Phos/AGGCCAGAGCATTCGGAGCTGAATTTTTTTTTTTTTTTVN

ROUND 2 BARCODES

Table S1.3 Round 2 barcodes used in SPLiT-seq.

Plate Well	Name	Sequence
A1	Round2_1	/5Phos/CATCGGCGTACGACTAACGTGATATCCACGTGCTTGAG
A2	Round2_2	/5Phos/CATCGGCGTACGACTAAACATCGATCCACGTGCTTGAG
A3	Round2_3	/5Phos/CATCGGCGTACGACTATGCCTAAATCCACGTGCTTGAG
A4	Round2_4	/5Phos/CATCGGCGTACGACTAGTGGTCAATCCACGTGCTTGAG
A5	Round2_5	/5Phos/CATCGGCGTACGACTACCACTGTATCCACGTGCTTGAG
A6	Round2_6	/5Phos/CATCGGCGTACGACTACATTGGCATCCACGTGCTTGAG
A7	Round2_7	/5Phos/CATCGGCGTACGACTCAGATCTGATCCACGTGCTTGAG
A8	Round2_8	/5Phos/CATCGGCGTACGACTCATCAAGTATCCACGTGCTTGAG
A9	Round2_9	/5Phos/CATCGGCGTACGACTCGCTGATCATCCACGTGCTTGAG
A10	Round2_10	/5Phos/CATCGGCGTACGACTACAAGCTAATCCACGTGCTTGAG
A11	Round2_11	/5Phos/CATCGGCGTACGACTCTGTAGCCATCCACGTGCTTGAG
A12	Round2_12	/5Phos/CATCGGCGTACGACTAGTACAAGATCCACGTGCTTGAG
B1	Round2_13	/5Phos/CATCGGCGTACGACTAACAACCAATCCACGTGCTTGAG
B2	Round2_14	/5Phos/CATCGGCGTACGACTAACCGAGAATCCACGTGCTTGAG
B3	Round2_15	/5Phos/CATCGGCGTACGACTAACGCTTAATCCACGTGCTTGAG
B4	Round2_16	/5Phos/CATCGGCGTACGACTAAGACGGAATCCACGTGCTTGAG
B5	Round2_17	/5Phos/CATCGGCGTACGACTAAGGTACAATCCACGTGCTTGAG
B6	Round2_18	/5Phos/CATCGGCGTACGACTACACAGAAATCCACGTGCTTGAG
B7	Round2_19	/5Phos/CATCGGCGTACGACTACAGCAGAATCCACGTGCTTGAG
B8	Round2_20	/5Phos/CATCGGCGTACGACTACCTCCAAATCCACGTGCTTGAG
B9	Round2_21	/5Phos/CATCGGCGTACGACTACGCTCGAATCCACGTGCTTGAG
B10	Round2_22	/5Phos/CATCGGCGTACGACTACGTATCAATCCACGTGCTTGAG
B11	Round2_23	/5Phos/CATCGGCGTACGACTACTATGCAATCCACGTGCTTGAG
B12	Round2_24	/5Phos/CATCGGCGTACGACTAGAGTCAAATCCACGTGCTTGAG
C1	Round2_25	/5Phos/CATCGGCGTACGACTAGATCGCAATCCACGTGCTTGAG
C2	Round2_26	/5Phos/CATCGGCGTACGACTAGCAGGAAATCCACGTGCTTGAG
C3	Round2_27	/5Phos/CATCGGCGTACGACTAGTCACTAATCCACGTGCTTGAG
C4	Round2_28	/5Phos/CATCGGCGTACGACTATCCTGTAATCCACGTGCTTGAG
C5	Round2_29	/5Phos/CATCGGCGTACGACTATTGAGGAATCCACGTGCTTGAG
C6	Round2_30	/5Phos/CATCGGCGTACGACTCAACCACAATCCACGTGCTTGAG
C7	Round2_31	/5Phos/CATCGGCGTACGACTGACTAGTAATCCACGTGCTTGAG
C8	Round2_32	/5Phos/CATCGGCGTACGACTCAATGGAAATCCACGTGCTTGAG
C9	Round2_33	/5Phos/CATCGGCGTACGACTCACTTCGAATCCACGTGCTTGAG
C10	Round2_34	/5Phos/CATCGGCGTACGACTCAGCGTTAATCCACGTGCTTGAG
C11	Round2_35	/5Phos/CATCGGCGTACGACTCATACCAAATCCACGTGCTTGAG
C12	Round2_36	/5Phos/CATCGGCGTACGACTCCAGTTCAATCCACGTGCTTGAG

D1	Round2_37	/5Phos/CATCGGCGTACGACT**CCGAAGTA**ATCCACGTGCTTGAG
D2	Round2_38	/5Phos/CATCGGCGTACGACT**CCGTGAGA**ATCCACGTGCTTGAG
D3	Round2_39	/5Phos/CATCGGCGTACGACT**CCTCCTGA**ATCCACGTGCTTGAG
D4	Round2_40	/5Phos/CATCGGCGTACGACT**CGAACTTA**ATCCACGTGCTTGAG
D5	Round2_41	/5Phos/CATCGGCGTACGACT**CGACTGGA**ATCCACGTGCTTGAG
D6	Round2_42	/5Phos/CATCGGCGTACGACT**CGCATACA**ATCCACGTGCTTGAG
D7	Round2_43	/5Phos/CATCGGCGTACGACT**CTCAATGA**ATCCACGTGCTTGAG
D8	Round2_44	/5Phos/CATCGGCGTACGACT**CTGAGCCA**ATCCACGTGCTTGAG
D9	Round2_45	/5Phos/CATCGGCGTACGACT**CTGGCATA**ATCCACGTGCTTGAG
D10	Round2_46	/5Phos/CATCGGCGTACGACT**GAATCTGA**ATCCACGTGCTTGAG
D11	Round2_47	/5Phos/CATCGGCGTACGACT**CAAGACTA**ATCCACGTGCTTGAG
D12	Round2_48	/5Phos/CATCGGCGTACGACT**GAGCTGAA**ATCCACGTGCTTGAG
E1	Round2_49	/5Phos/CATCGGCGTACGACT**GATAGACA**ATCCACGTGCTTGAG
E2	Round2_50	/5Phos/CATCGGCGTACGACT**GCCACATA**ATCCACGTGCTTGAG
E3	Round2_51	/5Phos/CATCGGCGTACGACT**GCGAGTAA**ATCCACGTGCTTGAG
E4	Round2_52	/5Phos/CATCGGCGTACGACT**GCTAACGA**ATCCACGTGCTTGAG
E5	Round2_53	/5Phos/CATCGGCGTACGACT**GCTCGGTA**ATCCACGTGCTTGAG
E6	Round2_54	/5Phos/CATCGGCGTACGACT**GGAGAACA**ATCCACGTGCTTGAG
E7	Round2_55	/5Phos/CATCGGCGTACGACT**GGTGCGAA**ATCCACGTGCTTGAG
E8	Round2_56	/5Phos/CATCGGCGTACGACT**GTACGCAA**ATCCACGTGCTTGAG
E9	Round2_57	/5Phos/CATCGGCGTACGACT**GTCGTAGA**ATCCACGTGCTTGAG
E10	Round2_58	/5Phos/CATCGGCGTACGACT**GTCTGTCA**ATCCACGTGCTTGAG
E11	Round2_59	/5Phos/CATCGGCGTACGACT**GTGTTCTA**ATCCACGTGCTTGAG
E12	Round2_60	/5Phos/CATCGGCGTACGACT**TAGGATGA**ATCCACGTGCTTGAG
F1	Round2_61	/5Phos/CATCGGCGTACGACT**TATCAGCA**ATCCACGTGCTTGAG
F2	Round2_62	/5Phos/CATCGGCGTACGACT**TCCGTCTA**ATCCACGTGCTTGAG
F3	Round2_63	/5Phos/CATCGGCGTACGACT**TCTTCACA**ATCCACGTGCTTGAG
F4	Round2_64	/5Phos/CATCGGCGTACGACT**TGAAGAGA**ATCCACGTGCTTGAG
F5	Round2_65	/5Phos/CATCGGCGTACGACT**TGGAACAA**ATCCACGTGCTTGAG
F6	Round2_66	/5Phos/CATCGGCGTACGACT**TGGCTTCA**ATCCACGTGCTTGAG
F7	Round2_67	/5Phos/CATCGGCGTACGACT**TGGTGGTA**ATCCACGTGCTTGAG
F8	Round2_68	/5Phos/CATCGGCGTACGACT**TTCACGCA**ATCCACGTGCTTGAG
F9	Round2_69	/5Phos/CATCGGCGTACGACT**AACTCACC**ATCCACGTGCTTGAG
F10	Round2_70	/5Phos/CATCGGCGTACGACT**AAGAGATC**ATCCACGTGCTTGAG
F11	Round2_71	/5Phos/CATCGGCGTACGACT**AAGGACAC**ATCCACGTGCTTGAG
F12	Round2_72	/5Phos/CATCGGCGTACGACT**AATCCGTC**ATCCACGTGCTTGAG
G1	Round2_73	/5Phos/CATCGGCGTACGACT**AATGTTGC**ATCCACGTGCTTGAG
G2	Round2_74	/5Phos/CATCGGCGTACGACT**ACACGACC**ATCCACGTGCTTGAG
G3	Round2_75	/5Phos/CATCGGCGTACGACT**ACAGATTC**ATCCACGTGCTTGAG
G4	Round2_76	/5Phos/CATCGGCGTACGACT**AGATGTAC**ATCCACGTGCTTGAG
G5	Round2_77	/5Phos/CATCGGCGTACGACT**AGCACCTC**ATCCACGTGCTTGAG
G6	Round2_78	/5Phos/CATCGGCGTACGACT**AGCCATGC**ATCCACGTGCTTGAG
G7	Round2_79	/5Phos/CATCGGCGTACGACT**AGGCTAAC**ATCCACGTGCTTGAG
G8	Round2_80	/5Phos/CATCGGCGTACGACT**ATAGCGAC**ATCCACGTGCTTGAG
G9	Round2_81	/5Phos/CATCGGCGTACGACT**ATCATTCC**ATCCACGTGCTTGAG
G10	Round2_82	/5Phos/CATCGGCGTACGACT**ATTGGCTC**ATCCACGTGCTTGAG
G11	Round2_83	/5Phos/CATCGGCGTACGACT**CAAGGAGC**ATCCACGTGCTTGAG
G12	Round2_84	/5Phos/CATCGGCGTACGACT**CACCTTAC**ATCCACGTGCTTGAG

H1	Round2_85	/5Phos/CATCGGCGTACGACT**CCATCCTC**ATCCACGTGCTTGAG
H2	Round2_86	/5Phos/CATCGGCGTACGACT**CCGACAAC**ATCCACGTGCTTGAG
H3	Round2_87	/5Phos/CATCGGCGTACGACT**CCTAATCC**ATCCACGTGCTTGAG
H4	Round2_88	/5Phos/CATCGGCGTACGACT**CCTCTATC**ATCCACGTGCTTGAG
H5	Round2_89	/5Phos/CATCGGCGTACGACT**CGACACAC**ATCCACGTGCTTGAG
H6	Round2_90	/5Phos/CATCGGCGTACGACT**CGGATTGC**ATCCACGTGCTTGAG
H7	Round2_91	/5Phos/CATCGGCGTACGACT**CTAAGGTC**ATCCACGTGCTTGAG
H8	Round2_92	/5Phos/CATCGGCGTACGACT**GAACAGGC**ATCCACGTGCTTGAG
H9	Round2_93	/5Phos/CATCGGCGTACGACT**GACAGTGC**ATCCACGTGCTTGAG
H10	Round2_94	/5Phos/CATCGGCGTACGACT**GAGTTAGC**ATCCACGTGCTTGAG
H11	Round2_95	/5Phos/CATCGGCGTACGACT**GATGAATC**ATCCACGTGCTTGAG
H12	Round2_96	/5Phos/CATCGGCGTACGACT**GCCAAGAC**ATCCACGTGCTTGAG

ROUND 3 BARCODES

Table S1.4 Round 3 barcodes used in SPLiT-seq.

Well	Name	Sequence
A1	R3_01	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****AACGTGAT**GTGGCCGATGTTTCG
A2	R3_02	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****AAACATCG**GTGGCCGATGTTTCG
A3	R3_03	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****ATGCCTAA**GTGGCCGATGTTTCG
A4	R3_04	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****AGTGGTCA**GTGGCCGATGTTTCG
A5	R3_05	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****ACCACTGT**GTGGCCGATGTTTCG
A6	R3_06	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****ACATTGGC**GTGGCCGATGTTTCG
A7	R3_07	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****CAGATCTG**GTGGCCGATGTTTCG
A8	R3_08	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****CATCAAGT**GTGGCCGATGTTTCG
A9	R3_09	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****CGCTGATC**GTGGCCGATGTTTCG
A10	R3_10	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****ACAAGCTA**GTGGCCGATGTTTCG
A11	R3_11	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****CTGTAGCC**GTGGCCGATGTTTCG
A12	R3_12	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****AGTACAAG**GTGGCCGATGTTTCG
B1	R3_13	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****AACAACCA**GTGGCCGATGTTTCG
B2	R3_14	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****AACCGAGA**GTGGCCGATGTTTCG
B3	R3_15	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****AACGCTTA**GTGGCCGATGTTTCG
B4	R3_16	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****AAGACGGA**GTGGCCGATGTTTCG
B5	R3_17	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****AAGGTACA**GTGGCCGATGTTTCG
B6	R3_18	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****ACACAGAA**GTGGCCGATGTTTCG
B7	R3_19	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****ACAGCAGA**GTGGCCGATGTTTCG
B8	R3_20	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****ACCTCCAA**GTGGCCGATGTTTCG
B9	R3_21	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****ACGCTCGA**GTGGCCGATGTTTCG
B10	R3_22	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****ACGTATCA**GTGGCCGATGTTTCG
B11	R3_23	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****ACTATGCA**GTGGCCGATGTTTCG
B12	R3_24	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****AGAGTCAA**GTGGCCGATGTTTCG
C1	R3_25	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****AGATCGCA**GTGGCCGATGTTTCG
C2	R3_26	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****AGCAGGAA**GTGGCCGATGTTTCG
C3	R3_27	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****AGTCACTA**GTGGCCGATGTTTCG
C4	R3_28	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****ATCCTGTA**GTGGCCGATGTTTCG
C5	R3_29	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****ATTGAGGA**GTGGCCGATGTTTCG
C6	R3_30	/5Biosg/CAGACGTGTGCTCTTCCGATCT**NNNNNNNNNN****CAACCACA**GTGGCCGATGTTTCG

C7	R3_31	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGACTAGTAGTGGCCGATGTTTCG
C8	R3_32	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCAATGGAAGTGGCCGATGTTTCG
C9	R3_33	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCACTTCGAGTGGCCGATGTTTCG
C10	R3_34	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCAGCGTTAGTGGCCGATGTTTCG
C11	R3_35	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCATACCAAGTGGCCGATGTTTCG
C12	R3_36	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCAGTTCAGTGGCCGATGTTTCG
D1	R3_37	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCGAAGTAGTGGCCGATGTTTCG
D2	R3_38	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCGTGAGAGTGGCCGATGTTTCG
D3	R3_39	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCTCCTGAGTGGCCGATGTTTCG
D4	R3_40	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCGAACTTAGTGGCCGATGTTTCG
D5	R3_41	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCGACTGGAGTGGCCGATGTTTCG
D6	R3_42	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCGCATACAGTGGCCGATGTTTCG
D7	R3_43	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCTCAATGAGTGGCCGATGTTTCG
D8	R3_44	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCTGAGCCAGTGGCCGATGTTTCG
D9	R3_45	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCTGGCATAGTGGCCGATGTTTCG
D10	R3_46	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGAATCTGAGTGGCCGATGTTTCG
D11	R3_47	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCAAGACTAGTGGCCGATGTTTCG
D12	R3_48	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGAGCTGAAGTGGCCGATGTTTCG
E1	R3_49	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGATAGACAGTGGCCGATGTTTCG
E2	R3_50	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGCCACATAGTGGCCGATGTTTCG
E3	R3_51	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGCGAGTAAGTGGCCGATGTTTCG
E4	R3_52	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGCTAACGAGTGGCCGATGTTTCG
E5	R3_53	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGCTCGGTAGTGGCCGATGTTTCG
E6	R3_54	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGGAGAACAGTGGCCGATGTTTCG
E7	R3_55	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGGTGCGAAGTGGCCGATGTTTCG
E8	R3_56	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGTACGCAAGTGGCCGATGTTTCG
E9	R3_57	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGTCGTAGAGTGGCCGATGTTTCG
E10	R3_58	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGTCTGTCAGTGGCCGATGTTTCG
E11	R3_59	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGTGTTCTAGTGGCCGATGTTTCG
E12	R3_60	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTAGGATGAGTGGCCGATGTTTCG
F1	R3_61	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTATCAGCAGTGGCCGATGTTTCG
F2	R3_62	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTCCGTCTAGTGGCCGATGTTTCG
F3	R3_63	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTCTTCACAGTGGCCGATGTTTCG
F4	R3_64	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTGAAGAGAGTGGCCGATGTTTCG
F5	R3_65	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTGGAACAAGTGGCCGATGTTTCG
F6	R3_66	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTGGCTTCAGTGGCCGATGTTTCG
F7	R3_67	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTGGTGGTAGTGGCCGATGTTTCG
F8	R3_68	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNTTCACGCAGTGGCCGATGTTTCG
F9	R3_69	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAACTCACCGTGGCCGATGTTTCG
F10	R3_70	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAAGAGATCGTGGCCGATGTTTCG
F11	R3_71	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAAGGACACGTGGCCGATGTTTCG
F12	R3_72	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAATCCGTCGTGGCCGATGTTTCG
G1	R3_73	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAATGTTGCGTGGCCGATGTTTCG
G2	R3_74	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACACGACCGTGGCCGATGTTTCG
G3	R3_75	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNACAGATTCGTGGCCGATGTTTCG
G4	R3_76	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGATGTACGTGGCCGATGTTTCG
G5	R3_77	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGCACCTCGTGGCCGATGTTTCG
G6	R3_78	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGCCATGCGTGGCCGATGTTTCG
G7	R3_79	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNAGGCTAACGTGGCCGATGTTTCG

G8	R3_80	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNATAGCGACGTGGCCGATGTTTCG
G9	R3_81	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNATCATTCCGTGGCCGATGTTTCG
G10	R3_82	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNATTGGCTCGTGGCCGATGTTTCG
G11	R3_83	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCAAGGAGCGTGGCCGATGTTTCG
G12	R3_84	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCACCTTACGTGGCCGATGTTTCG
H1	R3_85	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCATCCTCGTGGCCGATGTTTCG
H2	R3_86	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCGACAACGTGGCCGATGTTTCG
H3	R3_87	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCTAATCCGTGGCCGATGTTTCG
H4	R3_88	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCCTCTATCGTGGCCGATGTTTCG
H5	R3_89	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCGACACACGTGGCCGATGTTTCG
H6	R3_90	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCGGATTGCGTGGCCGATGTTTCG
H7	R3_91	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNCTAAGGTCGTGGCCGATGTTTCG
H8	R3_92	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGAACAGGCGTGGCCGATGTTTCG
H9	R3_93	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGACAGTGCGTGGCCGATGTTTCG
H10	R3_94	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGAGTTAGCGTGGCCGATGTTTCG
H11	R3_95	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGATGAATCGTGGCCGATGTTTCG
H12	R3_96	/5Biosg/CAGACGTGTGCTCTTCCGATCTNNNNNNNNNNGCCAAGACGTGGCCGATGTTTCG

SUPPLEMENTARY 2

Tables S2.1-S2.3 These tables present all identified cell types (cluster names) with their corresponding cluster group, as well as the number of cells per cluster (freq.), percentage of cells per cluster (%), number of cells per group (group freq.) and percentage of cells per group (group %). Total cells and percentages are specified at the bottom. The colour codes per cluster use in the UMAPs are also indicated.

SCHMIDTEA MEDITERRANEA

Table S2.1 Cluster information from the *Schmidtea mediterranea* dataset (Chapter II).

***Schmidtea mediterranea* (ACME & SPLiT-seq)**						
Cluster name	**Cluster group**	**Freq.**	**%**	**Group freq.**	**Group %**	**Colour**
neoblasts 1	neoblasts	2002	10.5	4274	22.5	#C8C8C8
neoblasts 2		1598	8.4			#AFAFAF
neoblasts 3		338	1.8			#969696
neoblast 4		29	0.2			#646464
germ cell progenitors		307	1.6			#7D7D7D
early epidermal progenitors	epidermis	908	4.8	3467	18.2	#9ECAE1
late epidermal progenitors 1		984	5.2			#6BAED6
late epidermal progenitors 2		187	1.0			#56A0CE
epidermis 1		481	2.5			#4292C6
epidermis 2		907	4.8			#2171B5
ChAT neurons	neurons	2003	10.5	3029	15.9	#FEC44F
GABA neurons		659	3.5			#FEE391
serotonin neurons		124	0.7			#FA9632
npp-18+ neurons 1		114	0.6			#EC7014
npp-18+ neurons 2		40	0.2			#AD4818
eye-53+ neurons		37	0.2			#FCAE3F
cav-1+ neurons		52	0.3			#EEC900
muscle body 1	muscle	1080	5.7	2703	14.2	#EE5C42
muscle body 2		639	3.4			#B22222
muscle pharynx		984	5.2			#CD5555
pgrn+ parenchymal cells	parenchymal	1039	5.5	1352	7.1	#E066FF
psap+ parenchymal cells		243	1.3			#FF1493
aqp+ parenchymal cells		70	0.4			#CD96CD
phagocyte progenitors	phagocytes	821	4.3	1192	6.3	#32CD32
phagocytes		371	2.0			#228B22
secretory 1	secretory	243	1.3	916	4.8	#6428C8
secretory 2		169	0.9			#9664C8
secretory 3		167	0.9			#8264C8
secretory 4		131	0.7			#9632C8
secretory 5		128	0.7			#7800FA
secretory 6		78	0.4			#9650FA

pharynx progenitors	pharynx	191	1.0	658	3.5	#9FB6CD
pharynx cell type		467	2.5			#4169E1
otf+ cells 1	otf	324	1.7	523	2.7	#993404
otf+ cells 2		199	1.0			#662506
goblet cell progenitors	goblet	240	1.3	353	1.9	#8B8B46
goblet cells		113	0.6			#8B8B00
protonephridia flame cells	protonephridia	182	1.0	264	1.4	#FFA0BE
protonephridia tubule cells		82	0.4			#FFBE96
psd+ cells	psd	196	1.0	196	1.0	#BCEE68
epidermis DVb	epidermis DVb	98	0.5	98	0.5	#1E90FF
	TOTAL	19025	100	19025	100	

DUGESIA JAPONICA

Table S2.2 Cluster information from the *Dugesia japonica* dataset (Chapter II).

***Dugesia japonica* (ACME & SPLiT-seq)**						
Cluster name	**Cluster group**	**Freq.**	**%**	**Group freq.**	**Group %**	**Colour**
neoblasts 1	neoblasts	1457	10.9	3848	28.7	#C8C8C8
neoblasts 2		1184	8.8			#AFAFAF
neoblasts 3		1006	7.5			#969696
germ cell progenitors		201	1.5			#7D7D7D
early epidermal progenitors	epidermis	856	6.4	3243	24.2	#9ECAE1
late epidermal progenitors 1		715	5.3			#6BAED6
late epidermal progenitors 2		564	4.2			#4292C6
epidermis		1108	8.3			#2171B5
ChAT neurons	neurons	1042	7.8	1452	10.8	#FEC44F
GABA neurons		410	3.1			#FEE391
muscle progenitors	muscle	758	5.7	1658	12.4	#EE5C42
muscle body		564	4.2			#B22222
muscle pharynx		336	2.5			#CD5555
pgrn+ parenchymal cells	parenchymal	278	2.1	278	2.1	#E066FF
phagocyte progenitors	phagocytes	747	5.6	1378	10.3	#32CD32
phagocytes		631	4.7			#228B22
secretory a	secretory	180	1.3	615	4.6	#8264C8
secretory b		171	1.3			#9664C8
secretory c		103	0.8			#5014B4
secretory d		93	0.7			#9632C8
secretory e		43	0.3			#9650FA
secretory f		25	0.2			#C800FA
pharynx cell type	pharynx	338	2.5	338	2.5	#4169E1
otf+ cells 1	otf	222	1.7	270	2.0	#993404
otf+ cells 2		48	0.4			#662506

goblet cells	goblet	92	0.7	92	0.7	#8B8B00
psd+ cells	psd	108	0.8	108	0.8	#BCEE68
epidermis DVb	epidermis DVb	126	0.9	126	0.9	#1E90FF
	TOTAL	13406	100	13406	100	

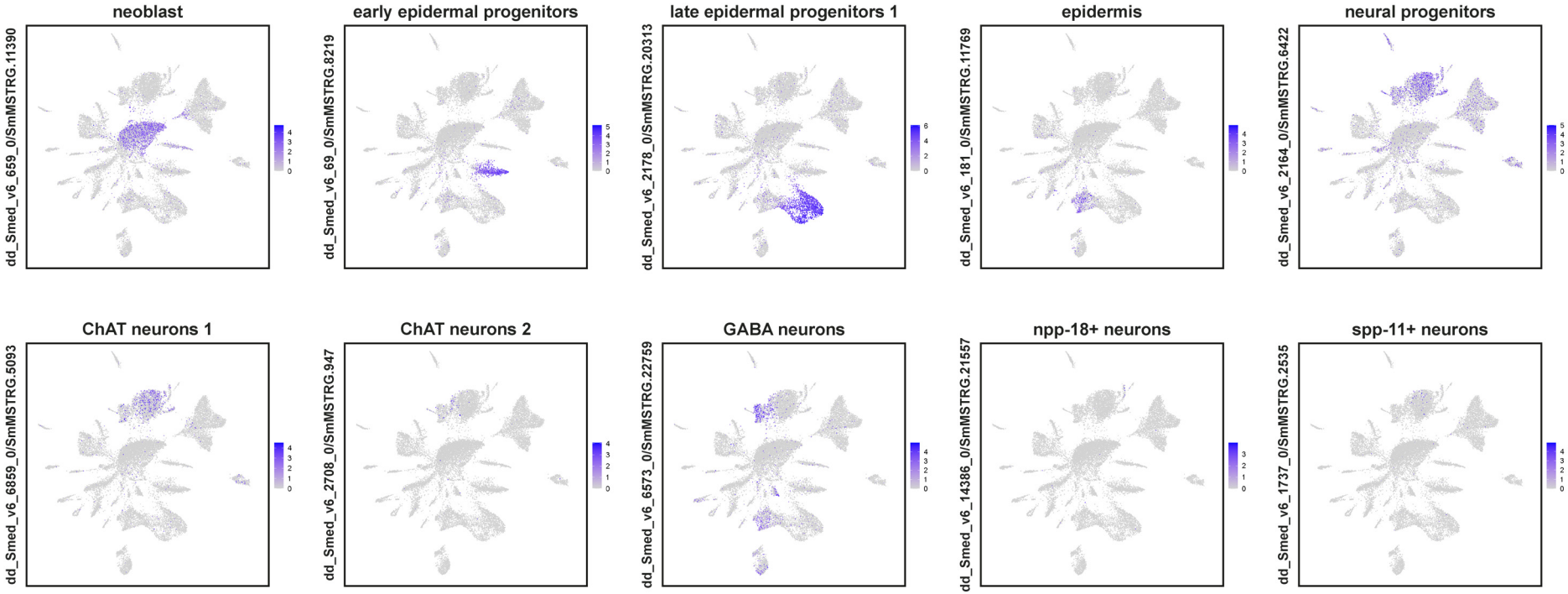
REANALYSIS OF PLASS *ET AL.*, 2018

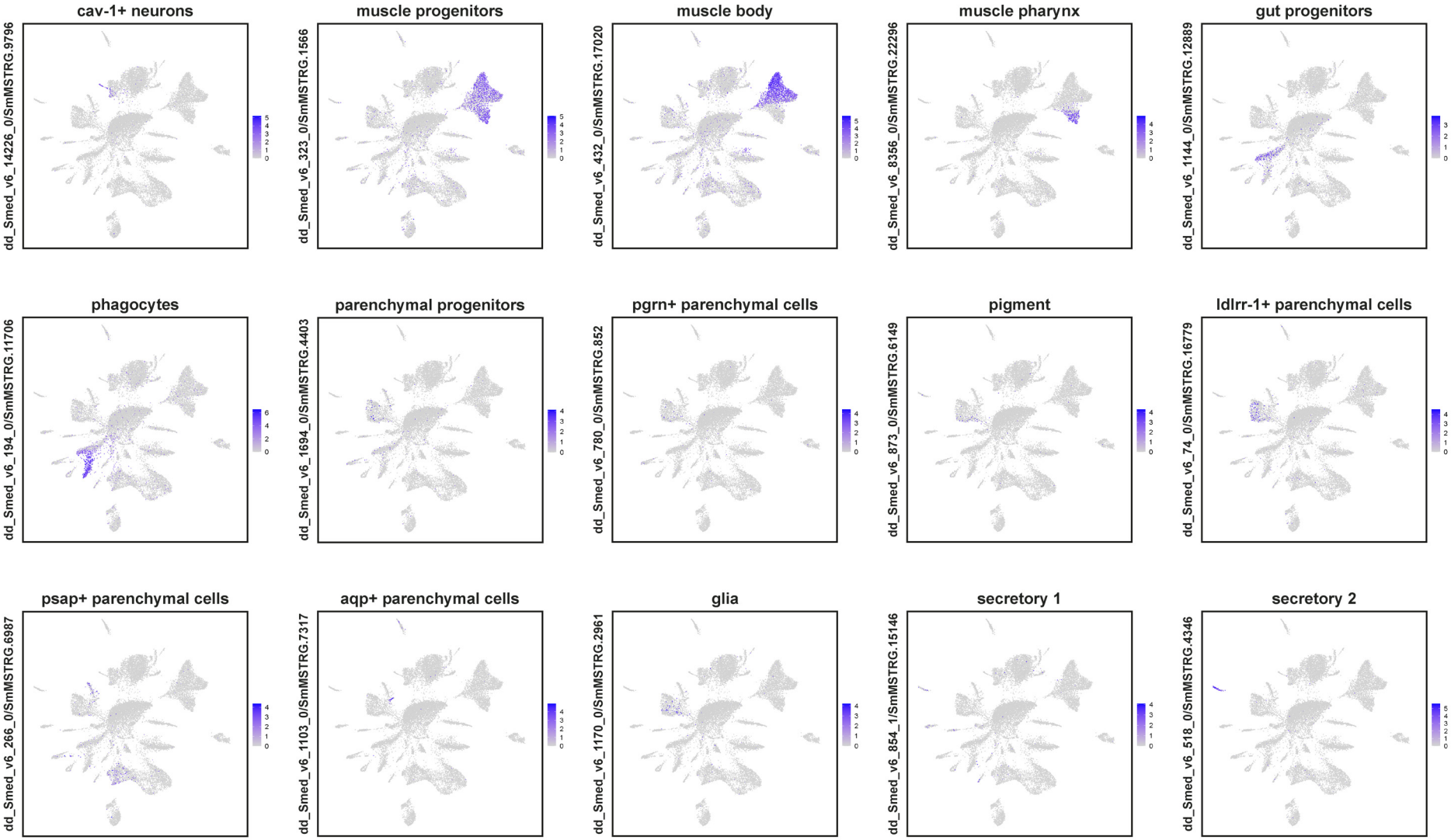
Table S2.3 Cluster information from the reanalysis of Plass *et al.*, 2018 (Chapter II).

***Schmidtea mediterranea* (Reanalysis of Plass *et al*. 2018)**						
Cluster name	**Cluster group**	**Freq.**	**%**	**Group freq.**	**Group %**	**Colour**
neoblasts 1	neoblasts	2252	10.4	7774	36.0	#DCDCDC
neoblasts 2		1642	7.6			#C8C8C8
neoblasts 3		1451	6.7			#B4B4B4
neoblasts 4		1090	5.0			#A0A0A0
neoblasts 5		594	2.7			#8C8C8C
neoblasts 6		379	1.8			#787878
neoblasts 7		280	1.3			#646464
neoblasts 8		86	0.4			#505050
early epidermal progenitors 1	epidermis	1507	7.0	4181	19.3	#9ECAE1
early epidermal progenitors 2		1601	7.4			#6BAED6
late epidermal progenitors		758	3.5			#56A0CE
epidermis		315	1.5			#2171B5
neural progenitors	neurons	1288	6.0	2612	12.1	#FFF7BC
ChAT neurons 1		609	2.8			#FEC44F
ChAT neurons 2		313	1.4			#FE9929
GABA neurons		402	1.9			#FEE391
muscle progenitors	muscle	911	4.2	2311	10.7	#EE5C42
muscle body		1044	4.8			#B22222
muscle body 2		356	1.6			#CD5555
parenchymal progenitors	parenchymal	855	4.0	2408	11.1	#FF69B4
pgrn+ parenchymal cells		373	1.7			#E066FF
psap+ parenchymal cells		507	2.3			#FF1493
aqp+ parenchymal cells		295	1.4			#CD96CD
ldlrr+ parenchymal cells		61	0.3			#F75394
pigment cells		317	1.5			#CD6889
phagocyte progenitors	phagocytes	446	2.1	614	2.8	#32CD32
phagocytes		168	0.8			#228B22
secretory 1	secretory	409	1.9	559	2.6	#6428C8
secretory 2		107	0.5			#9664C8
secretory 3		43	0.2			#8264C8
pharynx cell type	pharynx	208	1.0	208	1.0	#4169E1

otf+ cells 1	otf	165	0.8	290	1.3	#993404
otf+ cells 2		125	0.6			#662506
goblet cells progenitors	goblet	340	1.6	359	1.7	#8B8B46
goblet cells		19	0.1			#8B8B00
protonephridia	protonephridia	109	0.5	109	0.5	#FFA0BE
psd+ cells	psd	75	0.3	75	0.3	#BCEE68
epidermis DVb	epidermis DVb	110	0.5	110	0.5	#1E90FF
	TOTAL	21610	100	21610	100	

SUPPLEMENTARY 3




cav-1+ neurons
dd_Smed_v6_14226_0/SmMSTRG.9796
muscle progenitors
dd_Smed_v6_323_0/SmMSTRG.1566
muscle body
dd_Smed_v6_432_0/SmMSTRG.17020
muscle pharynx
dd_Smed_v6_8356_0/SmMSTRG.22296
gut progenitors
dd_Smed_v6_1144_0/SmMSTRG.12889
phagocytes
dd_Smed_v6_194_0/SmMSTRG.11706
parenchymal progenitors
dd_Smed_v6_1694_0/SmMSTRG.4403
pgrn+ parenchymal cells
dd_Smed_v6_780_0/SmMSTRG.852
pigment
dd_Smed_v6_873_0/SmMSTRG.6149
ldlrr-1+ parenchymal cells
dd_Smed_v6_74_0/SmMSTRG.16779
psap+ parenchymal cells
dd_Smed_v6_266_0/SmMSTRG.6987
aqp+ parenchymal cells
dd_Smed_v6_1103_0/SmMSTRG.7317
glia
dd_Smed_v6_1170_0/SmMSTRG.2961
secretory 1
dd_Smed_v6_854_1/SmMSTRG.15146
secretory 2
dd_Smed_v6_518_0/SmMSTRG.4346

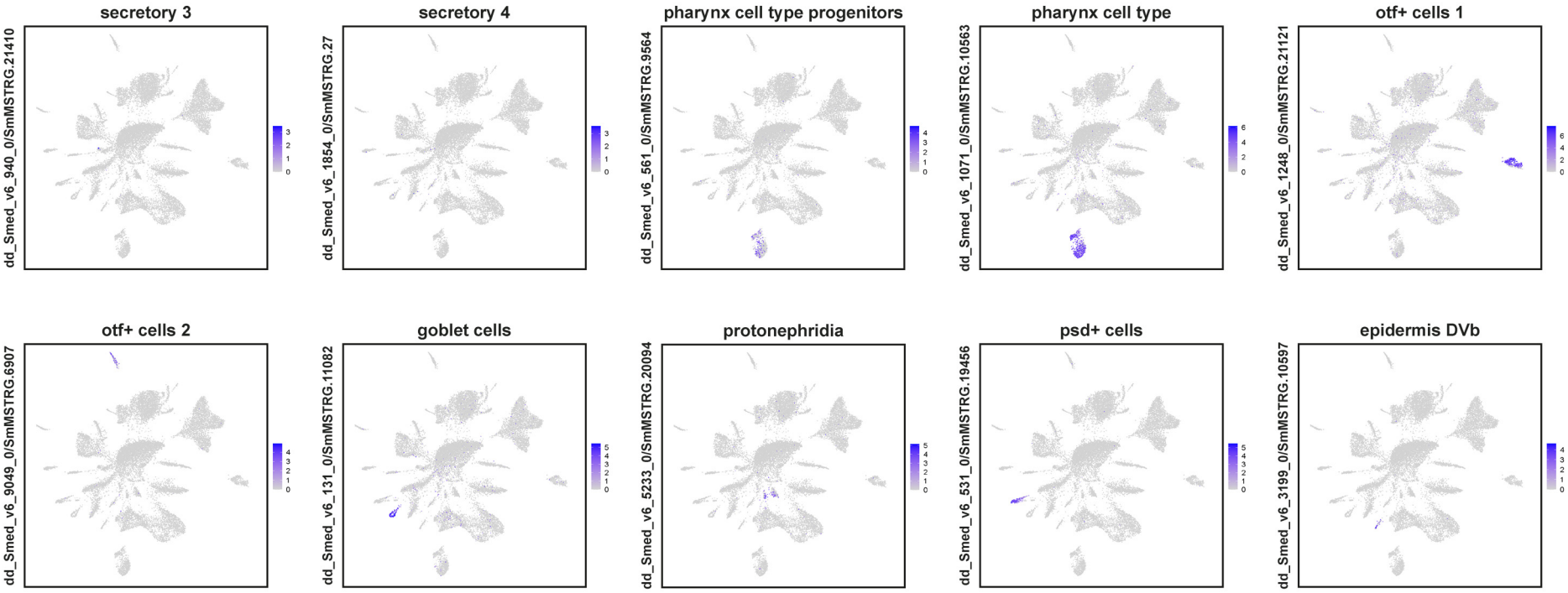


Figure S3 Expression of previously published markers in the new *Schmidtea mediterranea* dataset. Feature plots of markers (from Plass *et al.,* 2018) use to annotate the *S. mediterranea* dataset generated by SPLiT-seq in Chapter II.

SUPPLEMENTARY 4

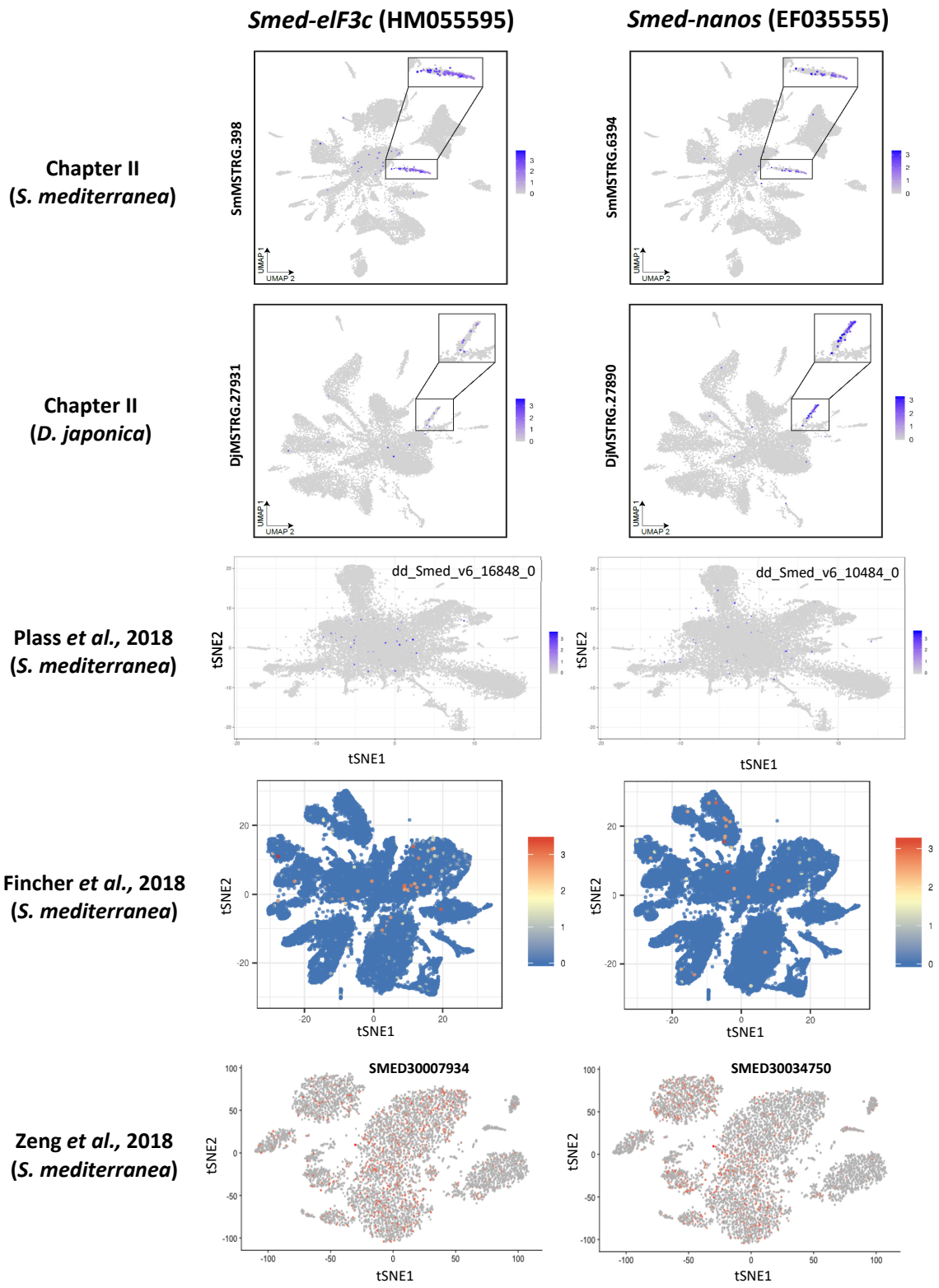


Figure S4 Expression of germ line progenitors' markers. Feature plots of *Smed-elF3c* and *Smed-nanos* in different datasets. From top to bottom: Chapter II (*Schmidtea mediterranea* and *Dugesia japonica*), Plass *et al.* (downloaded from https://shiny.mdc-berlin.de/psca/), Fincher *et al.* (downloaded from https://digiworm.wi.mit.edu/) and Zeng *et al.* (downloaded from https://planosphere.stowers.org).

SUPPLEMENTARY 5

Table S5 Cluster information from the *Schmidtea* mediterranea RNAi dataset (Chapter III).

***Schmidtea mediterranea* (RNAi dataset)**		
Cluster name	**Cluster group**	**Colour**
neoblast 1	neoblasts	'#c8c8c8',
neoblast 2		'#afafaf',
germ cell progenitors	germ cell progenitors	#464646'
ChAT neurons	neurons	'#fec44f',
GABA neurons		'#fee391',
cav-1+ neurons		'#eec900',
npp-18+ neurons 1		'#ec7014',
npp-18+ neurons 2		'#ad4818',
aberrant phagocyte progenitors	phagocytes	#44FF98'
phagocyte progenitors		'#32cd32',
phagocytes		'#228b22',
aqp+ parenchymal cells	parenchymal	'#cd96cd',
pgrn+ parenchymal cells		'#e066ff',
pigment cells		'#af88e3',
psap+ parenchymal cells		'#ff1493',
early epidermal progenitors 1	epidermis	'#93c5fb',
epidermis 2		'#2171b5',
late epidermal progenitors 1		'#6baed6',
late epidermal progenitors 3		'#56a0ce',
epidermis DVb		'#1e90ff',
goblet cell progenitors	goblet	'#8b8b46',
goblet cells		'#8b8b00',
muscle body 1	muscle	'#ee5c42',
muscle body 2		'#b22222',
muscle pharynx		'#cd5555',
muscle progenitors		'#a53467',
otf+ cells 1	*otf+*	'#993404',
otf+ cells 2		'#662506',
pharynx cell type	pharynx	'#4169e1',
protonephridia flame cells	protonephridia	'#ffa0be',
protonephridia tubule cells		'#ffbe96',
psd+ cells	*psd+*	'#bcee68',
secretory 1	secretory	'#6428c8',
secretory 2		'#9664c8',
secretory 3		'#8264c8',
secretory 4		'#9632c8',
secretory 5		'#7800fa',
secretory 6		'#9650fa',
secretory 7		'#7548e1',
secretory 8		'#7b68fa',

SUPPLEMENTARY 6

Table S6.1 Total counts per cluster for all genes (Chapter III).

Total counts (all genes)			
Cluster name	**Whole data**	**GFP RNAi**	***hnf4* RNAi**
neoblast 1	42028	15917	26111
neoblast 2	2321326	1007775	1313551
germ cell progenitors	172094	67940	104154
ChAT neurons	832234	379627	452607
GABA neurons	329960	149525	180435
cav-1+ neurons	171477	79018	92459
npp-18+ neurons 1	113022	47895	65127
npp-18+ neurons 2	42563	20765	21798
aberrant phagocyte progenitors	484507	12192	472315
phagocyte progenitors	381642	367460	14182
phagocytes	466934	321564	145370
aqp+ parenchymal cells	107666	50622	57044
pgrn+ parenchymal cells	669868	383245	286623
pigment cells	77423	38716	38707
psap+ parenchymal cells	177085	76241	100844
early epidermal progenitors 1	429794	170723	259071
epidermis 2	358599	107624	250975
late epidermal progenitors 1	558672	232445	326227
late epidermal progenitors 3	264940	93858	171082
epidermis DVb	32800	14597	18203
goblet cell progenitors	82853	41515	41338
goblet cells	74048	37015	37033
muscle body 1	411316	179127	232189
muscle body 2	394432	179818	214614
muscle pharynx	140465	63194	77271
muscle progenitors	22544	10319	12225
otf+ cells 1	177385	84901	92484
otf+ cells 2	87067	34868	52199
pharynx cell type	157651	66239	91412
protonephridia flame cells	160825	68765	92060
protonephridia tubule cells	73357	36531	36826
psd+ cells	88218	35652	52566
secretory 1	158977	66401	92576
secretory 2	170782	68597	102185
secretory 3	53029	25363	27666
secretory 4	29108	14646	14462
secretory 5	270386	123655	146731
secretory 6	142573	66215	76358
secretory 7	51918	24187	27731
secretory 8	51798	17313	34485

Table S6.2 *Hnf4* counts per million and per cluster (Chapter III).

***hnf4* counts per million**			
Cluster name	**Whole dataset**	**GFP RNAi**	***hnf4* RNAi**
neoblast 1	47.59	0.00	76.60
neoblast 2	19.82	24.81	15.99
germ cell progenitors	5.81	0.00	9.60
ChAT neurons	4.81	5.27	4.42
GABA neurons	6.06	6.69	5.54
cav-1+ neurons	11.66	12.66	10.82
npp-18+ neurons 1	26.54	41.76	15.35
npp-18+ neurons 2	0.00	0.00	0.00
aberrant phagocyte progenitors	227.03	164.04	228.66
phagocyte progenitors	86.47	81.64	211.54
phagocytes	102.80	77.75	158.22
aqp+ parenchymal cells	195.05	158.03	227.89
pgrn+ parenchymal cells	220.94	182.65	272.13
pigment cells	245.41	206.63	284.19
psap+ parenchymal cells	192.00	157.40	218.16
early epidermal progenitors 1	0.00	0.00	0.00
epidermis 2	5.58	0.00	7.97
late epidermal progenitors 1	3.58	8.60	0.00
late epidermal progenitors 3	0.00	0.00	0.00
epidermis DVb	30.49	68.51	0.00
goblet cell progenitors	181.04	120.44	241.91
goblet cells	81.03	135.08	27.00
muscle body 1	9.72	5.58	12.92
muscle body 2	5.07	0.00	9.32
muscle pharynx	28.48	15.82	38.82
muscle progenitors	0.00	0.00	0.00
otf+ cells 1	11.27	11.78	10.81
otf+ cells 2	0.00	0.00	0.00
pharynx cell type	0.00	0.00	0.00
protonephridia flame cells	6.22	0.00	10.86
protonephridia tubule cells	0.00	0.00	0.00
psd+ cells	158.70	84.15	209.26
secretory 1	12.58	15.06	10.80
secretory 2	5.86	0.00	9.79
secretory 3	18.86	0.00	36.15
secretory 4	0.00	0.00	0.00
secretory 5	0.00	0.00	0.00
secretory 6	0.00	0.00	0.00
secretory 7	0.00	0.00	0.00
secretory 8	0.00	0.00	0.00

Table S6.3 *Hnf4* counts and number of cells expressing *hnf4* counts per tissue (Chapter III).

	***hnf4* counts**			**Cells expressing *hnf4***		
Cluster name	**Whole data**	**GFP RNAi**	***hnf4* RNAi**	**Whole data**	**GFP RNAi**	***hnf4* RNAi**
neoblast 1	2	0	2	2	0	2
neoblast 2	46	25	21	44	23	21
germ cell progenitors	1	0	1	1	0	1
ChAT neurons	4	2	2	4	2	2
GABA neurons	2	1	1	2	1	1
cav-1+ neurons	2	1	1	2	1	1
npp-18+ neurons 1	3	2	1	3	2	1
npp-18+ neurons 2	0	0	0	0	0	0
aberrant phagocyte progenitors	110	2	108	94	2	92
phagocyte progenitors	33	30	3	31	28	3
phagocytes	48	25	23	45	24	21
aqp+ parenchymal cells	21	8	13	19	7	12
pgrn+ parenchymal cells	148	70	78	138	65	73
pigment cells	19	8	11	17	7	10
psap+ parenchymal cells	34	12	22	32	12	20
early epidermal progenitors 1	0	0	0	0	0	0
epidermis 2	2	0	2	2	0	2
late epidermal progenitors 1	2	2	0	2	2	0
late epidermal progenitors 3	0	0	0	0	0	0
epidermis DVb	1	1	0	1	1	0
goblet cell progenitors	15	5	10	15	5	10
goblet cells	6	5	1	5	4	1
muscle body 1	4	1	3	4	1	3
muscle body 2	2	0	2	2	0	2
muscle pharynx	4	1	3	4	1	3
muscle progenitors	0	0	0	0	0	0
otf+ cells 1	2	1	1	2	1	1
otf+ cells 2	0	0	0	0	0	0
pharynx cell type	0	0	0	0	0	0
protonephridia flame cells	1	0	1	1	0	1
protonephridia tubule cells	0	0	0	0	0	0
psd+ cells	14	3	11	14	3	11
secretory 1	2	1	1	2	1	1
secretory 2	1	0	1	1	0	1
secretory 3	1	0	1	1	0	1
secretory 4	0	0	0	0	0	0
secretory 5	0	0	0	0	0	0
secretory 6	0	0	0	0	0	0
secretory 7	0	0	0	0	0	0
secretory 8	0	0	0	0	0	0
TOTAL	530	206	324			

SUPPLEMENTARY 7

Table S7 Differentially expressed genes per cluster group. This table includes all DEGs, their cluster groups, gene expression tendencies (up or down-regulation), reference or accession number of the closest homolog (obtained by diamond blast), E-value of the diamond blast, and closest homologs identities.

Cluster group	Gene	DGE	Reference	E-Value	Homolog identity
epidermis	SmMSTRG.20455	down	XP_022332232.1	8.61E-101	apolipophorins-like [Crassostrea virginica]
muscle	SmMSTRG.20455	down	XP_022332232.1	8.61E-101	apolipophorins-like [Crassostrea virginica]
neoblasts	SmMSTRG.20455	down	XP_022332232.1	8.61E-101	apolipophorins-like [Crassostrea virginica]
goblet	SmMSTRG.20455	down	XP_022332232.1	8.61E-101	apolipophorins-like [Crassostrea virginica]
goblet	SmMSTRG.11706	down	PAA80758.1	1.79E-155	hypothetical protein BOX15_Mlig027085g4 [Macrostomum lignano]
parenchymal	SMEST007996001	up			
parenchymal	SMEST056331001	down	AKN21403.1	5.5E-219	slc2a-9, partial [Schmidtea mediterranea]
parenchymal	SMEST056326001	down	AKN21403.1	6.08E-104	slc2a-9, partial [Schmidtea mediterranea]
parenchymal	SmMSTRG.4394	down			
phagocytes	SmMSTRG.1012	up	QGW52011.1	8.28E-134	acid phosphatase type 6 [Dugesia japonica]
phagocytes	SMEST012332001	up	XP_031571867.1	8.86E-267	actin, cytoplasmic-like [Actinia tenebrosa]
phagocytes	SmMSTRG.20129	up	THD22761.1	4.74E-35	Calsyntenin-3 [Fasciola hepatica]
phagocytes	SmMSTRG.10417	up	QAU32669.1	0	Dach-2 [Schmidtea mediterranea]
phagocytes	SMEST074789007	up	WP_065380720.1	3.27E-18	DDE-type integrase/transposase/recombinase [Candidatus Glomeribacter gigasporarum]
phagocytes	SMEST022275001	up			
phagocytes	SmMSTRG.13151	up	KAA0200365.1	1.74E-88	ELKS/Rab6-interacting/CAST family member 1 [Fasciolopsis buski]
phagocytes	SmMSTRG.12022	up	XP_026536998.1	8.74E-120	FAD-dependent oxidoreductase domain-containing protein 1 [Notechis scutatus]
phagocytes	SMEST063471002	up	CAB3372235.1	5.5E-95	Hypothetical predicted protein [Cloeon dipterum]
phagocytes	SmMSTRG.21594	up	GBN94221.1	3.64E-26	hypothetical protein AVEN_200951-1 [Araneus ventricosus]
phagocytes	SmMSTRG.21736	up	PAA93291.1	8.45E-70	hypothetical protein BOX15_Mlig000210g3 [Macrostomum lignano]
phagocytes	SmMSTRG.12339	up	PAA55441.1	3.12E-190	hypothetical protein BOX15_Mlig020537g1 [Macrostomum lignano]
phagocytes	SMEST050123007	up	ELT99958.1	1.93E-307	hypothetical protein CAPTEDRAFT_223727 [Capitella teleta]

phagocytes	SMEST063805001	up	ELT96483.1	1.98E-104	hypothetical protein CAPTEDRAFT_228599 [Capitella teleta]
phagocytes	SMEST008191001	up	VVC44406.1	2.22E-193	Hypothetical protein CINCED_3A019335 [Cinara cedri]
phagocytes	SMEST022272002	up	TNN13942.1	3.26E-25	hypothetical protein EWB00_002503, partial [Schistosoma japonicum]
phagocytes	SmMSTRG.16518	up	KAF5269493.1	4.33E-28	hypothetical protein FQA39_LY08682 [Lamprigera yunnana]
phagocytes	SmMSTRG.2217	up	XP_009065072.1	1.56E-19	hypothetical protein LOTGIDRAFT_183988 [Lottia gigantea]
phagocytes	SMEST067002001	up	AKN21724.1	0	polycystic kidney disease protein 1 (PKD1L-1) [Schmidtea mediterranea]
phagocytes	SMEST005882002	up	XP_018653986.1	0	putative tolloid [Schistosoma mansoni]
phagocytes	SMEST058735001	up	TPP56899.1	4.27E-56	Ras-associated and pleckstrin domains-containing protein 1 [Fasciola gigantica]
phagocytes	SmMSTRG.19702	up	XP_024348697.1	2.19E-40	Thyrotropin-releasing hormone receptor [Echinococcus granulosus]
phagocytes	SMEST047737002	up	XP_022166445.1	3.03E-49	uncharacterized protein LOC111030999 isoform X1 [Myzus persicae]
phagocytes	SmMSTRG.6035	up	VZI29060.1	9.43E-126	unnamed protein product [Spirometra erinaceieuropaei]
phagocytes	SmMSTRG.5729	up	AEJ87267.1	0	voltage operated calcium channel Cav1A [Dugesia japonica]
phagocytes	SmMSTRG.10483	up			
phagocytes	SMEST070747001	up			
phagocytes	SmMSTRG.4181	up			
phagocytes	SMEST037144001	down	QDF60598.1	0	alpha-2-macroglobulin [Dugesia japonica]
phagocytes	SMEST037203001	down	QDF60598.1	0	alpha-2-macroglobulin [Dugesia japonica]
phagocytes	SmMSTRG.20455	down	XP_022332232.1	8.61E-101	apolipophorins-like [Crassostrea virginica]
phagocytes	SmMSTRG.15864	down	VDH90145.1	9.54E-307	Cu+-exporting ATPase [Mytilus galloprovincialis]
phagocytes	SmMSTRG.18127	down	XP_013408119.1	3.15E-100	cytochrome P450 3A29 [Lingula anatina]
phagocytes	SMEST064358001	down	TNN11253.1	0	Cytoplasmic dynein 1 heavy chain 1 isoform 1 [Schistosoma japonicum]
phagocytes	SMEST061955001	down	CAD5113772.1	3.26E-172	DgyrCDS2941 [Dimorphilus gyrociliatus]
phagocytes	SmMSTRG.22294	down	XP_013400269.1	3.94E-284	diacylglycerol kinase beta isoform X3 [Lingula anatina]
phagocytes	SmMSTRG.18862	down	ANO39025.1	3.49E-286	GCR050 [Schmidtea mediterranea]
phagocytes	SMEST000653001	down	XP_026479896.1	1.4E-71	GPI ethanolamine phosphate transferase 2-like [Ctenocephalides felis]
phagocytes	SMEST032522001	down	VDI31805.1	8.85E-13	Hypothetical predicted protein, partial [Mytilus galloprovincialis]
phagocytes	SmMSTRG.1359	down	PAA82774.1	1.78E-96	hypothetical protein BOX15_Mlig002181g3 [Macrostomum lignano]
phagocytes	SmMSTRG.112	down	PAA89199.1	1.6E-270	hypothetical protein BOX15_Mlig003482g1 [Macrostomum lignano]
phagocytes	SMEST053202001	down	PAA84507.1	3.27E-219	hypothetical protein BOX15_Mlig008771g3 [Macrostomum lignano]

phagocytes	SmMSTRG.8128	down	PAA78691.1	5.85E-102	hypothetical protein BOX15_Mlig012691g1 [Macrostomum lignano]
phagocytes	SmMSTRG.17703	down	PAA88602.1	0	hypothetical protein BOX15_Mlig016190g1 [Macrostomum lignano]
phagocytes	SmMSTRG.12495	down	PAA88602.1	0	hypothetical protein BOX15_Mlig016190g1 [Macrostomum lignano]
phagocytes	SmMSTRG.5753	down	PAA71212.1	2.75E-280	hypothetical protein BOX15_Mlig019787g1 [Macrostomum lignano]
phagocytes	SmMSTRG.5754	down	PAA71212.1	8.34E-279	hypothetical protein BOX15_Mlig019787g1 [Macrostomum lignano]
phagocytes	SmMSTRG.11706	down	PAA80758.1	1.79E-155	hypothetical protein BOX15_Mlig027085g4 [Macrostomum lignano]
phagocytes	SmMSTRG.3195	down	PAA59456.1	4.66E-67	hypothetical protein BOX15_Mlig034563g2 [Macrostomum lignano]
phagocytes	SMEST055398002	down	ELT99789.1	3.62E-42	hypothetical protein CAPTEDRAFT_183609 [Capitella teleta]
phagocytes	SMEST001708001	down	TGZ74088.1	8.18E-229	hypothetical protein CRM22_001133 [Opisthorchis felineus]
phagocytes	SmMSTRG.15711	down	KAF2070342.1	7.15E-133	hypothetical protein CYY_008345 [Polysphondylium violaceum]
phagocytes	SMEST048502001	down	KAF5278802.1	5.17E-34	hypothetical protein FQR65_LT03489 [Abscondita terminalis]
phagocytes	SmMSTRG.9436	down	XP_009046384.1	2.72E-80	hypothetical protein LOTGIDRAFT_111188, partial [Lottia gigantea]
phagocytes	SmMSTRG.7846	down	KAF8567825.1	2.48E-07	hypothetical protein P879_02004 [Paragonimus westermani]
phagocytes	SmMSTRG.14602	down	XP_018652399.1	1.03E-07	hypothetical protein Smp_136860 [Schistosoma mansoni]
phagocytes	SmMSTRG.6039	down	KAE8576283.1	1.02E-52	hypothetical protein XENTR_v10004131 [Xenopus tropicalis]
phagocytes	SMEST056013003	down	GAA52816.1	4.47E-141	laminin alpha 1/2, partial [Clonorchis sinensis]
phagocytes	SmMSTRG.16947	down	QDP17043.1	0	leucine-rich repeat serine/threonine-protein kinase 2-like protein, partial [Dugesia japonica]
phagocytes	SmMSTRG.17073	down	XP_029716999.1	9.7E-170	LOW QUALITY PROTEIN: cAMP-specific 3',5'-cyclic phosphodiesterase-like [Aedes albopictus]
phagocytes	SMEST005054001	down	KAF5399472.1	4.46E-256	Lysosomal alpha-glucosidase [Paragonimus heterotremus]
phagocytes	SmMSTRG.9782	down	XP_032897887.1	4.55E-152	lysosomal protective protein isoform X3 [Amblyraja radiata]
phagocytes	SMEST064049001	down	XP_013385784.1	4.33E-210	metabotropic glutamate receptor 1 [Lingula anatina]
phagocytes	SmMSTRG.23615	down	XP_013390140.1	2.4E-198	P protein-like isoform X2 [Lingula anatina]
phagocytes	SmMSTRG.14441	down	XP_029054704.1	9.95E-22	piggyBac transposable element-derived protein 4-like [Osmia bicornis bicornis]
phagocytes	SMEST000602001	down	AVM18990.1	1.88E-36	PIM-1-like protein [Schmidtea mediterranea]
phagocytes	SmMSTRG.8273	down	VDI62030.1	2.57E-227	polypeptide N-acetylgalactosaminyltransferase [Mytilus galloprovincialis]
phagocytes	SMEST045007001	down	XP_018114576.1	2.47E-15	PREDICTED: cadherin EGF LAG seven-pass G-type receptor 3-like, partial [Xenopus laevis]
phagocytes	SMEST079513001	down	XP_013070480.1	9.63E-115	PREDICTED: probable flavin-containing monoamine oxidase A [Biomphalaria glabrata]
phagocytes	SmMSTRG.6994	down	XP_013861451.1	0.00029	PREDICTED: prosaposin isoform X2 [Austrofundulus limnaeus]
phagocytes	SMEST031117001	down	XP_019616611.1	7.35E-27	PREDICTED: uncharacterized protein LOC109464110 isoform X2 [Branchiostoma belcheri]

phagocytes	SMEST061340001	down	XP_023300757.1	3.1E-114	pseudouridine-metabolizing bifunctional protein C1861.05 [Lucilia cuprina]
phagocytes	SMEST025889001	down	THD27271.1	0.0000621	putative beta thymosin [Fasciola hepatica]
phagocytes	SmMSTRG.18305	down	XP_039252277.1	1.1E-62	putative L-aspartate dehydrogenase isoform X1 [Styela clava]
phagocytes	SmMSTRG.11383	down	OON14849.1	4.32E-70	RhoGAP domain protein, partial [Opisthorchis viverrini]
phagocytes	SmMSTRG.6009	down	KAF6037976.1	1.3E-78	SDR16C5 [Bugula neritina]
phagocytes	SMEST069640002	down	AJA72716.1	5.23E-288	semaphorin 1, partial [Schmidtea mediterranea]
phagocytes	SMEST016303003	down	AJA72717.1	8.9E-75	semaphorin-like protein, partial [Schmidtea mediterranea]
phagocytes	SmMSTRG.21830	down	AAW24762.1	1.12E-41	SJCHGC04453 protein [Schistosoma japonicum]
phagocytes	SmMSTRG.19215	down	AKN21505.1	9.86E-56	slc16a-10 [Schmidtea mediterranea]
phagocytes	SMEST056331001	down	AKN21403.1	5.5E-219	slc2a-9, partial [Schmidtea mediterranea]
phagocytes	SMEST056326001	down	AKN21403.1	6.08E-104	slc2a-9, partial [Schmidtea mediterranea]
phagocytes	SmMSTRG.16883	down	AKN21403.1	3.59E-144	slc2a-9, partial [Schmidtea mediterranea]
phagocytes	SmMSTRG.9743	down	AKN21664.1	2.03E-304	slc38a-5 [Schmidtea mediterranea]
phagocytes	SmMSTRG.2567	down	AKN21701.1	2.37E-217	slc47a-4, partial [Schmidtea mediterranea]
phagocytes	SMEST045553001	down	AKN21703.1	5.93E-150	slc47a-6 [Schmidtea mediterranea]
phagocytes	SMEST045551001	down	AKN21703.1	2.37E-277	slc47a-6 [Schmidtea mediterranea]
phagocytes	SmMSTRG.9253	down	AKN21704.1	0	slc47a-7 [Schmidtea mediterranea]
phagocytes	SMEST031125001	down	AKN21444.1	8.51E-274	slc6a-24 [Schmidtea mediterranea]
phagocytes	SMEST076607001	down	AKN21456.1	0	slc8a-2 [Schmidtea mediterranea]
phagocytes	SmMSTRG.8818	down	XP_021943595.1	5.08E-134	sorbitol dehydrogenase isoform X1 [Folsomia candida]
phagocytes	SMEST039062001	down	KXJ06092.1	6.67E-46	TNF receptor-associated factor 6 [Exaiptasia diaphana]
phagocytes	SmMSTRG.11642	down	TPP57686.1	7.27E-127	Transforming growth factor-beta-induced protein ig-h3, partial [Fasciola gigantica]
phagocytes	SMEST070472001	down	XP_020894935.1	4.74E-62	TRPM8 channel-associated factor homolog [Exaiptasia diaphana]
phagocytes	SmMSTRG.23144	down	XP_015795966.1	5.31E-273	tubulin alpha chain [Tetranychus urticae]
phagocytes	SMEST055800001	down	WP_162473876.1	8.93E-16	tyrosine-protein phosphatase [endosymbiont 'TC1' of Trimyema compressum]
phagocytes	SmMSTRG.12536	down	XP_022851230.1	4.77E-15	ubiquitin carboxyl-terminal hydrolase 27 isoform X3 [Olea europaea var. sylvestris]
phagocytes	SmMSTRG.15439	down	KAA3678091.1	1.03E-22	uncharacterized protein DEA37_0010981 [Paragonimus westermani]
phagocytes	SMEST031101001	down	XP_034307373.1	1.67E-08	uncharacterized protein LOC105333658 isoform X2 [Crassostrea gigas]
phagocytes	SmMSTRG.3916	down	XP_013407971.1	2.69E-30	uncharacterized protein LOC106171979 [Lingula anatina]

phagocytes	SMEST047314001	down	XP_021350418.1	0.000142	uncharacterized protein LOC110448482 [Mizuhopecten yessoensis]
phagocytes	SMEST031104001	down	XP_034318688.1	1.19E-20	uncharacterized protein LOC117686918 [Crassostrea gigas]
phagocytes	SmMSTRG.23048	down	XP_028406786.1	5.36E-32	zinc finger MYM-type protein 1-like [Dendronephthya gigantea]
phagocytes	SmMSTRG.15564	down			
phagocytes	SmMSTRG.9113	down			
phagocytes	SmMSTRG.17772	down			
phagocytes	SmMSTRG.4394	down			
phagocytes	SmMSTRG.4400	down			
phagocytes	SMEST010944001	down			
phagocytes	SmMSTRG.6993	down			
phagocytes	SmMSTRG.12889	down			

SUPPLEMENTARY 8

Table S8.1 Cluster information from the *Schmidtea mediterranea* dataset (Chapter IV). Table shows the annotated cell types (cluster name) and their corresponding broad lineages (cluster group). The number (frequency, freq.) and percentage (%) of cells per cluster are indicated for both strains included in the study (asexual and sexual). The sums of these values are specified at the bottom. Colour codes used in UMAP visualisations are indicated for each cluster (colour).

***Schmidtea mediterranea* (species comparison)**						
Cluster name	**Cluster group**	**Asexual**		**Sexual**		**Colour**
		Freq.	**%**	**Freq.**	**%**	
ChAT neurons	neurons	979	12.4	829	4.9	#FEC44F
GABA neurons		357	4.5	263	1.6	#FEE391
cav-1+ neurons		149	1.9	179	1.1	#EEC900
npp-18+ neurons 1		63	0.8	85	0.5	#EC7014
npp-18+ neurons 2		27	0.3	61	0.4	#DB8B44
otf+ cells 1	*oft+*	136	1.7	212	1.3	#993404
otf+ cells 2		76	1.0	129	0.8	#662506
early epidermal progenitors 1	epidermis	23	0.3	128	0.8	#93C5FB
early epidermal progenitors 3		330	4.2	403	2.4	#8FBBEB
epidermis 1		314	4.0	139	0.8	#2171B5
late epidermal progenitors 1		262	3.3	401	2.4	#6BAED6
late epidermal progenitors 3		73	0.9	185	1.1	#56A0CE
epidermis DVb	epidermis DVb	26	0.3	96	0.6	#1E90FF
muscle body 1	muscle	462	5.8	613	3.6	#EE5C42
muscle body 2		427	5.4	519	3.1	#B22222
muscle pharynx		307	3.9	219	1.3	#CD5555
muscle progenitors		116	1.5	256	1.5	#A53467
goblet cell progenitors	goblet	328	4.1	215	1.3	#B0B054
goblet cells		60	0.8	369	2.2	#8B8B00
aqp+ parenchymal cells	parenchymal	59	0.7	172	1.0	#D18CB3
glia		17	0.2	28	0.2	#C936AA
pgrn+ parenchymal cells		601	7.6	1656	9.8	#E066FF
psap+ parenchymal cells		93	1.2	178	1.1	#FF1493
pigment cells		92	1.2	179	1.1	#E388B4
phagocyte progenitors	phagocytes	301	3.8	659	3.9	#32CD32
phagocytes		180	2.3	687	4.1	#228B22
pharynx cell type	pharynx	113	1.4	196	1.2	#4169E1
protonephridia flame cells	protonephridia	233	2.9	342	2.0	#FFA0BE
protonephridia tubule cells		39	0.5	73	0.4	#FFBE96
psd+ cells	*psd+*	44	0.6	76	0.4	#BCEE68
secretory 1	secretory	73	0.9	134	0.8	#6428C8
secretory 2		56	0.7	157	0.9	#9664C8
secretory 3		46	0.6	55	0.3	#8264C8
secretory 4		23	0.3	32	0.2	#9632C8
secretory 5		115	1.5	250	1.5	#763596

secretory 6	secretory	90	1.1	146	0.9	#9650FA
secretory 7		29	0.4	46	0.3	#7548E1
secretory 8		53	0.7	0	0.0	#7B68FA
reproductive system 1	reproductive system	13	0.2	869	5.1	#464646
reproductive system 2		0	0.0	760	4.5	#464646
reproductive system 3		0	0.0	636	3.8	#464646
reproductive system 4		0	0.0	432	2.6	#464646
reproductive system 5		0	0.0	424	2.5	#464646
reproductive system 6		3	0.0	60	0.4	#464646
reproductive system 7		1	0.0	28	0.2	#464646
neoblast 1	neoblasts	205	2.6	1580	9.4	#AFAFAF
neoblast 2		567	7.2	959	5.7	#AFAFAF
neoblast 3		91	1.2	229	1.4	#AFAFAF
neoblast 4		124	1.6	56	0.3	#AFAFAF
neoblast 5		71	0.9	35	0.2	#AFAFAF
neoblast 6		8	0.1	87	0.5	#AFAFAF
neoblast 7		3	0.0	31	0.2	#AFAFAF
neoblast 8		22	0.3	9	0.1	#AFAFAF
neoblast 9		1	0.0	13	0.1	#AFAFAF
unknown 1	unknown	4	0.1	259	1.5	#F9F9F3
unknown 2		27	0.3	63	0.4	#F9F9F3
	TOTAL	7912	100.0	16897	100.0	

Table S8.2 Cluster information from the *Schmidtea polychroa* dataset (Chapter IV). Table shows all cluster numbers and their assigned broad cell types. The number (frequency, freq.) and percentage (%) of cells per cluster are indicated for each life-history stage: hatchling, juvenile and adult. The sums of these values are specified at the bottom. Colour codes used in UMAP visualisations are indicated for each cluster (colour).

***Schmidtea polychroa* (species comparison)**								
Cluster number	**Broad cell type (Cluster group)**	**Hatchling**		**Juvenile**		**Adult**		**Colour**
		Freq.	**%**	**Freq.**	**%**	**Freq.**	**%**	
9	epidermis	706	3.9	503	3.3	365	2.3	#8ACFF4
12		720	4.0	552	3.6	14	0.1	
15		390	2.2	301	2.0	269	1.7	
22		148	0.8	203	1.3	399	2.6	
28		260	1.4	230	1.5	118	0.8	
48	epidermis DVb	99	0.5	75	0.5	70	0.4	#1E90FF
62	eyes	66	0.4	36	0.2	21	0.1	#F6FC57
10	reproductive system	10	0.1	772	5.0	657	4.2	#464646
29		17	0.1	17	0.1	541	3.5	
32		17	0.1	276	1.8	200	1.3	
14	goblet	342	1.9	277	1.8	451	2.9	#8B8B00
40		129	0.7	132	0.9	121	0.8	

4	muscle	892	4.9	691	4.5	580	3.7	#E84E3A
7		560	3.1	572	3.7	588	3.8	
8		501	2.8	578	3.8	613	3.9	
27		324	1.8	166	1.1	134	0.9	
59		41	0.2	47	0.3	61	0.4	
0	neoblasts	1003	5.6	798	5.2	846	5.4	#C8C8C8
3		621	3.4	490	3.2	1100	7.1	
41		189	1.0	148	1.0	41	0.3	
63		60	0.3	27	0.2	22	0.1	
65		21	0.1	38	0.2	37	0.2	
1	neurons	1520	8.4	563	3.7	305	2.0	#FEC44F
2		933	5.2	573	3.7	741	4.8	
17		449	2.5	226	1.5	250	1.6	
18		484	2.7	189	1.2	205	1.3	
20		435	2.4	242	1.6	176	1.1	
38		168	0.9	135	0.9	111	0.7	
39		178	1.0	104	0.7	118	0.8	
43		173	1.0	83	0.5	96	0.6	
46		98	0.5	77	0.5	90	0.6	
51		143	0.8	56	0.4	25	0.2	
52		111	0.6	62	0.4	43	0.3	
57		53	0.3	57	0.4	51	0.3	
64		38	0.2	26	0.2	35	0.2	
26	*otf*+	266	1.5	213	1.4	150	1.0	#993404
33		153	0.8	169	1.1	171	1.1	
5	parenchymal	638	3.5	734	4.8	764	4.9	#ED64DB
6		562	3.1	691	4.5	782	5.0	
11		426	2.4	461	3.0	437	2.8	
19		223	1.2	250	1.6	397	2.5	
30		165	0.9	152	1.0	250	1.6	
50		55	0.3	77	0.5	93	0.6	
54		11	0.1	68	0.4	114	0.7	
56		60	0.3	46	0.3	65	0.4	
13	phagocytes	457	2.5	366	2.4	403	2.6	#32CD32
21		312	1.7	236	1.5	233	1.5	
31		184	1.0	144	0.9	172	1.1	
25	pharynx	190	1.1	232	1.5	220	1.4	#0B40DE
16	protonephridia	298	1.7	329	2.1	325	2.1	#FFBE96
34		153	0.8	147	1.0	188	1.2	
49	*psd*+	108	0.6	94	0.6	38	0.2	#BCEE68
60		59	0.3	45	0.3	39	0.3	
23	secretory	260	1.4	262	1.7	188	1.2	#855CE6
24		289	1.6	245	1.6	170	1.1	
35		243	1.3	161	1.0	55	0.4	
37		171	0.9	125	0.8	144	0.9	
42		186	1.0	152	1.0	21	0.1	
44		172	1.0	118	0.8	54	0.3	

45	secretory	71	0.4	80	0.5	117	0.8	#855CE6
47		121	0.7	82	0.5	44	0.3	
53		80	0.4	65	0.4	56	0.4	
55		69	0.4	65	0.4	42	0.3	
36	unknown	24	0.1	143	0.9	286	1.8	#F9F9F3
58		66	0.4	56	0.4	30	0.2	
61		46	0.3	41	0.3	38	0.2	
66		10	0.1	11	0.1	1	0.0	
	TOTAL	18027	100	15382	100	15581	100.0	

Table S8.3 Cluster information from the *Dugesia japonica*, *Girardia tigrina* and *Polycelis nigra* datasets (Chapter IV). The number (frequency, freq.) and percentage (%) of cells per broad cell type are specified for each species. The sums of these values are indicated at the bottom of the table. Colour codes used for UMAP visualisations are specified in the last column.

Other planarian species (species comparison)							
Broad cell types	***D. japonica***		***G. tigrina***		***P. nigra***		**Colour**
	Freq.	**%**	**Freq.**	**%**	**Freq.**	**%**	
epidermis	1756	16.7	2783	13.8	3245	26.7	#8ACFF4
epidermis DVb	65	0.6	160	0.8	145	1.2	#1E90FF
eyes	0	0	24	0.1	0	0.0	#F6FC57
reproductive system	20	0.2	346	1.7	245	2.0	#464646
goblet	215	2.0	151	0.7	143	1.2	#8B8B00
muscle	1195	11.4	2432	12.0	1392	11.5	#E84E3A
neoblasts	1337	12.7	5071	25.1	1557	12.8	#C8C8C8
neurons	1768	16.8	2951	14.6	1363	11.2	#FEC44F
otf+	170	1.6	360	1.8	353	2.9	#993404
parenchymal	869	8.3	1492	7.4	985	8.1	#ED64DB
phagocytes	1861	17.7	2346	11.6	1075	8.9	#32CD32
pharynx	184	1.8	279	1.4	285	2.3	#0B40DE
protonephridia	297	2.8	505	2.5	328	2.7	#FFBE96
psd+	84	0.8	0	0.0	110	0.9	#BCEE68
secretory	679	6.5	970	4.8	861	7.1	#855CE6
unknown	0	0.0	320	1.6	47	0.4	#F9F9F3
TOTAL	10500	100	20190	100.0	12134	100.0	

ACKNOWLEDGEMENTS

DIRECTED BY
JORDI SOLANA

with NATHAN KENNY as
DIRECTOR OF BIOINFORMATICS

and PATRICIA ÁLVAREZ CAMPOS as
SCIENCE COACH

DISSERTATION COMMITTEE

DIRECTOR OF STUDIES	DIANNE NEWBURY
EXTERNAL REVIEWER	TERESA ADELL
INTERNAL REVIEWER	SAAD ARIF

CAST

SOLANA LAB

PLANARIAN CUISINE CHEF	VINCENT MASON
ACETYLGLUCOSAMINE	ELENA EMILI
SCIENTIFIC ILLUSTRATOR	CIRENIA ARIAS BALDRICH
EL JOVEN QUE SE ENCARGA DE LA BIOINFORMATICA EN NUESTRO LABORATORIO	DAVID SALAMANCA
S/G2/M/GO/G1	ALBERTO PEREZ POSADA
THALASSOPHILE	SOPHIE PERON
FASHION ICON	OLENA MARIA STOICA
TED TALKER	DIANALI RODRIGUEZ
COFFEE ENTHUSIAST	VIRGINIA VANNI

ÖZPOLAT LAB

BIOLOGIST & CERAMIC ARTIST	DUYGU ÖZPOLAT
POSTDOCTORAL SALSA DANCER	RANNY PASOS RIBEIRO
BIOLOGIST & WRITTER	BRIA METZGER

FRIENDLY VISITORS

MANIFESTOR	MARÍA ROSELLÓ
GOBBO	PHILIP BERTEMES
PRISTINA LADY	IRENE DEL OLMO
LA PYTHONISA	PATRICIA MEDINA BURGOS

ACME PAPER COLLABORATORS

CNIDARIAN PI	ARNAU SEBÉ-PEDRÓS
CNIDARIAN RESEARCHER	MARTA IGLESIAS GARCÍA
CNIDARIAN BIOINFORMATICIAN	ANAMARIA ELEK
SPIDER RESEARCHER	ANNA SCHÖNAUER
SNAIL RESEARCHER	VICTORIA SLEIGHT
PLANARIAN BIOINFORMATICIAN	JAKKE NEIRO
PLANARIAN PI	AZIZ ABOOBAKER
ZEBRAFISH RESEARCHER	JON PERMANYER
ZEBRAFISH PI	MANUEL IRIMIA

FLOW CYTOMETRY FACILITY

FACILITY MANAGER	ROBERT ´BOBBY´ HEADLEY

ITAÑOLOS

ALICANTIÑOLA	DR LORENA MARTÍNEZ
ALBACETEÑOLA	DR TERESA MINGUEZ
BASQUE ITAÑOLO	DR VICENTE PAGALDAY
NAPOLI ITAÑOLA	DR EMMANUELA CAROLLO

73b HOUSEHOLD

SPIRITUAL GUIDE	ROXANA BARSAN
ASSOCIATE HOUSEMATE	ANDY YAO
SOFTWARE DEVELOPER & PROUD MUM	ELENA DEL VALLE MATUTE
POLISH CUISINE CHEF	TOM KATZMAREK
BOHEMIAN NURSE	ANA CELIA
PROFESSIONAL JOB INTERVIEWEE	RISHI
FOREST GUARD	ALE
CINAMMON MANTECADO TASTER	CHARLOTTE
LANDLORD	NIGEL COWELL

EXTRAS

DOG	CONGA
FATHER	JULIAN
MOTHER	MARIAN
FRIEND I	GABY
FRIEND II	CAROLA
THE CANDIDATE WHO REJECTED THIS PHD	ANONYMOUS
MISTERIOUS EXTRA	AE

MODEL ORGANISMS

LEADING MODEL ORGANISM	*SCHMIDTEA MEDITERRANEA*
SUPPORTING MODEL ORGANISM I	*DUGESIA JAPONICA*
SUPPORTING MODEL ORGANISM II	*SCHMIDTEA POLYCHROA*
SUPPORTING MODEL ORGANISM III	*POLYCELIS NIGRA*
SUPPORTING MODEL ORGANISM IV	*GIRARDIA TIGRINA*

FUNDING BODIES

PHD SCHOLARSHIP	NIGEL GROOME-OXFORD BROOKES
FUNDER II	MRC
FUNDER III	BBSRC

WRITTEN AND PRODUCED BY

HELENA GARCÍA CASTRO

AT

OXFORD BROOKES UNIVERSITY

DEPARTMENT OF BIOLOGICAL AND MEDICAL SCIENCES

PLANARIAN SINGLE CELL TRANSCRIPTOMICS ASSOCIATION