Institute of Bioorganic Chemistry Polish Academy of Sciences
Laboratory of Single Cell Analyses

DOCTORAL THESIS

ANASTASIIA ZAREMBA

Bioinformatics approaches to the analysis of small non-coding RNAs in *Schmidtea mediterranea*

Supervisor: Dr. habil. Paulina Jackowiak, Assoc. Prof.

Poznań 2026

I dedicate this doctoral thesis to my parents, Svitlana Zaremba and Oleksandr Zaremba, my sister Viktoriia Zaremba, and my husband, Serhii Kozin.

TABLE OF CONTENTS

Page

ARTICLE INCLUDED IN THE DOCTORAL DISSERTATION

- Zaremba A., Marszałek-Zeńczak M., Dutta A., Samelak-Czajka A., Jackowiak P. (2025). "A modular pipeline for evidence-integrated genome annotation across species: a case study on *Schmidtea mediterranea*". Genomics 117(6): 111110.

This doctoral dissertation also contains a detailed description of the results of a publication currently in the second round of revision in Nucleic Acids Research:

- Dutta* A., Zaremba* A., Samelak-Czajka A., Marszałek-Zeńczak M., Trybus M., Osuch M., Figlerowicz M., Jackowiak P. "Knockdown of *Smed ELAC2* in *Schmidtea mediterranea* results in delayed regeneration and a reduced accumulation of the small non-coding RNA molecule, 5' tiRNA-Gly-GCC".

LIST OF ARTICLES NOT INCLUDED IN THE DOCTORAL DISSERTATION

- Dutta* A., Zaremba* A., Jackowiak P. (2025). "Ribonucleases in Mendelian disease – characterization and insight from model organisms". Genes & Diseases 12(5):101613. https://doi.org/10.1016/j.gendis.2025.101613.
- Dutta A., Samelak-Czajka A., Trybus M., Osuch M., Zaremba A., Jackowiak P. (2025). "Identification and characterization of planarian multinucleated cells in *Schmidtea mediterranea* using imaging flow cytometry". Frontiers in Cell and Developmental Biology Volume 13. https://doi.org/10.3389/fcell.2025.1611516

*these authors contributed equally

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Dr. hab. Paulina Jackowiak for giving me the opportunity to join the Laboratory of Single Cell Analyses and for her continuous support, guidance, and encouragement throughout my PhD.

I am also grateful to Annasha Dutta for our close collaboration within the same project and for her scientific insight and day-to-day support, which had a substantial impact on this dissertation.

I thank all my colleagues from the Laboratory of Single Cell Analyses, the Laboratory of Genomics, and the Department of Molecular and Systems Biology for creating a helpful and friendly environment. I am particularly thankful to Prof. Marek Figlerowicz and Dr. hab. Luiza Handschuh for valuable advice.

I also thank Dr. Irek Stolarek for mentoring me in the early years of my PhD, Dr. Małgorzata Marszałek-Zeńczak for continued support, and Jakub Kubiś for many helpful discussions. I am grateful to Marcin Osuch, Dr. Anna Samelak-Czajka, Magdalena Trybus, Dr. inż. Małgorzata Marcinkowska-Swojak, and Magdalena Rakoczy for their assistance, and to my office neighbors Dr. Agnieszka Gąbka-Buszek, Sylwia Nawrocka, and Dr. Jan Podkowiński for daily support.

Finally, I thank my husband, Serhii Kozin, for his unwavering support and for always having my back.

This work was financed by the National Science Centre, Poland, through a grant no. 2019/35/B/NZ2/02658 granted to Dr. hab. Paulina Jackowiak.

I acknowledge the use of the infrastructure developed under the project NEBI – National Research Center for Imaging in the Biological and Biomedical Sciences, POIR.04.02.00- 00-C004/19, co-financed through the European Regional Development Fund (ERDF) in the frame of Smart Growth Operational Programme 2014-2020 (Measure 4.2 Development of modern research infrastructure of the science sector).

ABSTRACT

Small non-coding RNAs (sncRNAs) are important regulators of gene expression, yet their functional interpretation in non-classical regeneration models is often constrained by incomplete reference resources. This dissertation aimed to develop and optimize bioinformatic approaches for sncRNA analysis in the planarian *Schmidtea mediterranea* and to use these approaches to clarify how perturbation of RNA metabolism influence regeneration.

First, an evidence-integrated, containerized genome annotation workflow (SmedAnno) was established to generate a reference annotation and to refine key loci of interest, including a corrected *Smed ELAC2* gene and corresponding transcript required for accurate mRNA quantification and design of sequence-dependent experiments. Next, RNA interference-mediated knockdown of *Smed ELAC2* (encoding a tRNA-processing ribonuclease) was used as a strategy to investigate the effects of alterations in the sncRNA pool on planarian regeneration.

Transcriptomic changes induced by *Smed ELAC2* silencing were profiled by bulk sncRNA and mRNA sequencing at 3 and 5 days post-amputation, and by time-resolved single-cell RNA sequencing. sncRNA analyses showed a selective, stage-dependent response concentrated in tRNA-derived fragments and miRNAs, with the strongest depletion observed for 5′ tiRNA-Gly-GCC. Bulk mRNA analysis revealed modest transcriptional changes. Single-cell analyses localized *Smed ELAC2* expression to specific cell populations and indicated altered differentiation trajectory from broad neoblast stem cells toward selected lineage branches under knockdown. Complementary low-input profiling of sorted fractions and scRNA-seq-guided target-activity modeling indicated cell-type-biased regulatory potential of 5′-tiRNA-Gly-GCC, particularly in *PGRN*-positive parenchymal and phagocyte populations.

Together, this doctoral dissertation provides planarian-adapted reference resources and an integrative analysis framework for determining how specific changes in sncRNA accumulation affect lineage-resolved regeneration in planarians, highlighting 5′ tiRNA-Gly-GCC as a regulatory molecule whose biogenesis is dependent on Smed ELAC2.

Keywords: *Schmidtea mediterranea*; regeneration; small non-coding RNA; tRNA-derived fragments; *ELAC2*; genome annotation; single-cell RNA-seq.

STRESZCZENIE

Małe niekodujące RNA (ang. *small non-coding RNAs*, sncRNA) są ważnymi regulatorami ekspresji genów, jednak interpretacja ich funkcji w nieklasycznych modelach regeneracji jest często ograniczona ze względu na niekompletne zasoby referencyjne. Celem niniejszej rozprawy było opracowanie i udoskonalenie metod bioinformatycznych do analizy sncRNA u wypławka *Schmidtea mediterranea* oraz zastosowanie tych narzędzi do określenia, w jaki sposób zaburzenia metabolizmu RNA wpływają na przebieg regeneracji.

Najpierw opracowano oparty na zebranych dowodach, zautomatyzowany proces adnotacji genomu (SmedAnno) w celu wygenerowania referencji i udoskonalenia kluczowych loci, w tym skorygowanego modelu genu i transkryptu *Smed ELAC2* niezbędnego do dokładnej analizy ilościowej mRNA i eksperymentów zależnych od sekwencji. Następnie zastosowano strategię wyciszania *Smed ELAC2* (kodującego rybonukleazę przetwarzającą tRNA) za pomocą interferencji RNA w celu zbadania wpływu zmian w puli sncRNA na regenerację. Zmiany transkryptomiczne wywołane wyciszeniem *Smed ELAC2* zostały zbadane poprzez sekwencjonowanie sncRNA i mRNA po 3 i 5 dniach od amputacji oraz za pomocą sekwencjonowania RNA pojedynczych komórek w różnych punktach czasowych. Analizy sncRNA wykazały selektywną, zależną od etapu odpowiedź skoncentrowaną w fragmentach pochodzących z tRNA i miRNA, przy czym najbardziej wyraźne spadki zaobserwowano w przypadku 5′ tiRNA-Gly-GCC.

Sekwencjonowanie RNA typu *bulk* wykazało niewielkie zmiany transkrypcyjne. Analizy pojedynczych komórek zlokalizowały ekspresję *Smed ELAC2* w określonych populacjach komórek i wykazały zmienione różnicowanie z ogólnych komórek macierzystych neoblastów w kierunku wybranych linii komórkowych pod wpływem wyciszenia. Uzupełniające profilowanie frakcji posortowanych przy niskim poziomie wejściowym oraz modelowanie aktywności docelowej oparte na sekwencjonowaniu scRNA wykazały potencjał regulacyjny 5′-tiRNA-Gly-GCC, szczególnie w populacjach tkanki miąższowej i fagocytów wykazujące ekspresję *PGRN*.

Wyniki te łącznie stanowią dostosowane do wypławków zasoby referencyjne oraz kompleksową strukturę analityczną, która łączy przebudowę sncRNA z programami regeneracji z rozdzielczością na poziomie linii komórkowej, podkreślając znaczenie 5′

tiRNA-Gly-GCC jako cząsteczki regulatorowej, której biogeneza jest zależna od Smed ELAC2.

Słowa kluczowe: *Schmidtea mediterranea*; regeneracja; małe niekodujące RNA; fragmenty pochodzące z tRNA; ELAC2; adnotacja genomu; sekwencjonowanie RNA pojedynczych komórek.

ABBREVIATIONS

A1CF	APOBEC1 complementation factor
ACA45	SCARNA15, H/ACA box small Cajal body–specific RNA 15
AGAT-1/-2/-3	Planarian epidermal markers AGAT-1, AGAT-2, AGAT-3
AGO	Argonaute
AGO2	Argonaute 2
AGO3	Argonaute 3
AlkB	Alpha-ketoglutarate-dependent dioxygenase AlkB
ALPPL2	Alkaline phosphatase, placental-like 2
AMON	Amontillado, prohormone convertase 2 ortholog
ANG	Angiogenin, RNASE5
ap2	AP-2 transcription factor
AQP	Aquaporin family water channel genes
AQP3	Aquaporin 3
AQP4	Aquaporin 4
ARM-seq	AlkB-facilitated RNA methylation sequencing
ARR2	Arrestin 2
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
ATP	Adenosine triphosphate
AWH	Arrowhead (Awh), LIM homeobox transcription factor
bambi	BMP and activin membrane-bound inhibitor (BAMBI)
BARHL1	BarH-like homeobox 1
BMP	Bone morphogenetic protein
bruli	Bruno-like RNA binding protein (bruli)
BUSCO	Benchmarking Universal Single-Copy Orthologs
cav-1	Caveolin 1
CCR4–NOT	Carbon catabolite repression 4–negative on TATA-less complex
CDC2L6	Cyclin-dependent kinase 19 alias CDK11B, historical symbol CDC2L6
cDNA	Complementary DNA
CG	Cephalic ganglion
CHAT	Choline O-acetyltransferase
ChIP-seq	Chromatin Immunoprecipitation followed by Sequencing
cNeoblast	Clonogenic neoblast
coe	Collier/olfactory-1/early B-cell factor (COE) transcription factor
COL4A1	Collagen type IV alpha 1 chain
cP	2′,3′-cyclic phosphate
CPE	Carboxypeptidase E
CRISPR	Clustered regularly interspaced short palindromic repeats
CTSL	Cathepsin L
dat/slc6a-3	Dopamine transporter, solute carrier family 6 member 3 (SLC6A3)
DCP1/2	mRNA decapping enzyme 1/2
DDX6	DEAD-box helicase 6
DGCR8	DiGeorge syndrome critical region gene 8
DICER	RNase III endoribonuclease Dicer
DLX1	Distal-less homeobox 1
DMRT2	Doublesex and mab-3 related transcription factor 2

DNAI1	Dynein axonemal intermediate chain 1
dpa	Days post amputation
DRIP	DRIP aquaporin water channel
DROSHA	RNase III endoribonuclease Drosha
Drop-seq	Droplet Sequencing
dsRNA	Double-stranded RNA
DV	Dorsal-ventral
ECM	Extracellular matrix
EGFR/ERBB4	Epidermal growth factor receptor / ERBB4-like receptor tyrosine kinases
EGR1	Early growth response 1
ELAC1	ElaC ribonuclease Z 1, short-form RNase Z
ELAC2	ElaC ribonuclease Z 2, long-form RNase Z
ENA	European Nucleotide Archive
equinox	equinox, wound-epidermis gene required for blastema formation
ERDF	European Regional Development Fund
estrella	estrella, glial marker gene
EV	Extracellular vesicle
eya	Eyes absent transcription factor
fer3l-1	Fer3-like basic helix–loop–helix transcription factor 1
fer3l-2	Fer3-like basic helix–loop–helix transcription factor 2
FGF	Fibroblast growth factor
FGFRL	Fibroblast growth factor receptor-like
foxA1	Forkhead box A1
foxC1	Forkhead box C1
foxD	Forkhead box D transcription factor
FOXJ1	Forkhead box J1
FZD4	Frizzled class receptor 4
G0	Quiescent cell-cycle phase
gad	Glutamate decarboxylase (GAD)
gata456	GATA4/5/6 family transcription factor (gata456)
GATM	Glycine amidinotransferase, mitochondrial
GEO	Gene Expression Omnibus
GLIPR1L1	GLI pathogenesis-related 1-like protein 1
GO	Gene Ontology
Gpas	Gpas, predicted heterotrimeric G protein alpha subunit gene
GTP	Guanosine triphosphate
GW182	Glycine-tryptophan 182 protein family, TNRC6A/B/C
H3K9me3	Histone H3 lysine-9 trimethylation
HB	Hunchback gap transcription factor
HENMT1	HEN1 methyltransferase 1
hmcn-1	Hemicentin 1
hnf4	Hepatocyte nuclear factor 4
HOXC12	Homeobox C12
hpa	Hours post amputation
HYDIN	Hydrocephalus-inducing protein HYDIN
IFT	Intraflagellar transport

IFT88	Intraflagellar transport protein 88
i-tRF	Internal tRNA-derived fragment
KD	Knockdown
La	La autoantigen, Sjögren syndrome antigen B, SSB
LAMB1	Laminin subunit beta 1
ldlrr-1	Low density lipoprotein receptor-related 1
lectins	Lectin family genes, epidermal secreted lectins
LHX1	LIM homeobox 1
LHX2	LIM homeobox 2
LHX5	LIM homeobox 5
LMX1A	LIM homeobox transcription factor 1 alpha
LMX1B.2	LIM homeobox transcription factor 1 beta 2 (LMX1B.2)
LNA	Locked nucleic acid
lncRNA	Long non-coding RNA
LTR	Long terminal repeat
MAGT1	Magnesium transporter 1
MEIS1, HTH	Meis homeobox 1 / homothorax
mim-tRNAseq	Modification-induced misincorporation tRNA sequencing
miRNA	MicroRNA
MMEL1/dd_5256	Membrane metallo-endopeptidase-like 1 (MMEL1), transcript dd_5256
mRNA	Messenger RNA
mt	Mitochondrial, as a prefix
musculin	Musculin (MSC), basic helix–loop–helix transcription factor
myb-1	Myb family transcription factor 1 (myb-1)
myoD	Myogenic differentiation 1
NCBI	National Center for Biotechnology Information
ncRNA	Non-coding RNA
ndl-2	Nou-darake-like 2 (ndl-2)
neuroD	Neurogenic differentiation factor (NeuroD)
NKX1-1	NK1 homeobox 1
NKX2-2	NK2 homeobox 2
NKX6-1	NK6 homeobox 1
nkx6-2	NK6 homeobox 2
notum	Notum, secreted Wnt deacylase
NPC2	Niemann–Pick disease type C2 gene
nt	Nucleotides
OAS	2′,5′-oligoadenylate synthetase
opsin	Opsin family photoreceptor genes
otp	Orthopedia homeobox transcription factor
ovo	Ovo zinc finger transcription factor
OVOL1	Ovo-like zinc finger transcription factor 1
OXPHOS	Oxidative phosphorylation
PAGE	Planarian Anatomy Gene Expression
PLANA	Planarian Anatomy Ontology
PAN2/PAN3	Poly(A)-specific ribonuclease 2/3
PARN	Poly(A)-specific ribonuclease, parent of PARN-like domain

PASR	Promoter-associated small RNA
PC2	Prohormone convertase 2, PCSK2
PCG	Positional-control gene
PCSK	Prohormone convertase subtilisin/kexin family
PCSK1	Proprotein convertase subtilisin/kexin type 1
PGRN-1	Progranulin 1
piRNA	PIWI-interacting RNA
PITX3/PTX1	Paired-like homeodomain transcription factors PITX3 and PITX1
PIWI	P-element induced wimpy testis protein family
PIWIL1	Piwi-like protein 1
PKD1L-2	Polycystic kidney disease 1-like 2
PKD1L-3	Polycystic kidney disease 1-like 3
PKD2-1	Polycystic kidney disease 2-like 1
PKD2-3	Polycystic kidney disease 2-like 3
PKD2-4	Polycystic kidney disease 2-like 4
PLANA	Planarian Anatomy Ontology
PLD6	Phospholipase D family member 6, Zucchini endonuclease
PNLDC1	Poly(A)-specific ribonuclease-like domain-containing 1
pou4l-1	POU class 4 homeobox 1
prog-1	Progeny marker 1
prog-2	Progeny marker 2
PROS	Prospero homeobox transcription factor
RBM47 (APOB-1)	RNA-binding motif protein 47 (apob-1)
RCN1	Reticulocalbin 1, EF-hand calcium-binding protein
RF	RNA fragment
RNAi	RNA interference
RNase	Ribonuclease
RNA-seq	RNA sequencing
RNase Z	tRNA 3′-trailer endonuclease, ELAC1/ELAC2
RO60/TROVE2	Ro60, Y RNA-binding protein, TROVE domain family member 2
ROS	Reactive oxygen species
rRF	rRNA-derived fragment
rRNA	Ribosomal RNA
RT	Reverse transcription
S. mediterranea	*Schmidtea mediterranea*
S/G2/M	DNA synthesis, gap-2, and mitosis phases
SCARNA	Small Cajal body–specific RNA
SCARNA15	Small Cajal body–specific RNA 15, ACA45
SCG	Secretogranin protein family
scRNA-seq	Single-cell RNA sequencing
SCRO	Scarecrow homeobox transcription factor
sdRNA	snoRNA-derived RNA
sim	Single-minded bHLH-PAS transcription factor
siRNA	Small interfering RNA
SIX1	Six homeobox 1
SIX2	Six homeobox 2

slc17a-6	Solute carrier family 17 member 6
slc18a-6	Solute carrier family 18 member 6
slc20a-1	Solute carrier family 20 member 1
slc22a-6	Solute carrier family 22 member 6
slc6a-1	Solute carrier family 6 member 1
slc6a-4/sert	Solute carrier family 6 member 4, serotonin transporter (SERT)
slc6a-5	Solute carrier family 6 member 5
Smart-seq	Switching Mechanism At the 5′ RNA Template end sequencing
sncRNA	Small non-coding RNA
snDR	Small nuclear RNA-derived fragment
snoRNA	Small nucleolar RNA
snRNA	Small nuclear RNA
soxB	SoxB family transcription factor
SOXP-1	SoxP family transcription factor 1
SOXP-2	SoxP family transcription factor 2
SOXP-3	SoxP family transcription factor 3
SP1 (sp6-9)	Sp6-9 specificity protein family transcription factor
SRA	Sequence Read Archive
SSA	Sjögren syndrome antigen A, Ro autoantigen
SSB	Sjögren syndrome antigen B, La autoantigen
SSPO	SCO-spondin (SSPO)
STNB	Stonin B endocytic adaptor protein
TDC2	Tyrosine decarboxylase 2
tDR	tRNA-derived RNA, generic label
TE	Transposable element
TF	Transcription factor
TGS1	Trimethylguanosine synthase 1
th	Tyrosine hydroxylase
tiRNA	tRNA half
tlx-1	T-cell leukemia homeobox 1
TNRC6	Trinucleotide repeat–containing protein 6, GW182 family
tph	Tryptophan hydroxylase
tRF	tRNA-derived fragment
tRF-1	Pre-tRNA 3′-trailer–derived tRF
tRF-3	T-arm or acceptor-junction–derived tRF
tRF-5	D-arm or anticodon-stem–derived tRF
TRP63	Tumor protein p63 (TRP63)
TSEN	tRNA-splicing endonuclease
TSS	Transcription start site
TSSa-RNA	Transcription start site–associated RNA
UMI	Unique molecular identifier
UNC5A	Unc-5 netrin receptor A
UTR	Untranslated region
V-ATPase	Vacuolar H+-transporting ATPase
VNC	Ventral nerve cord
VVL (POU2/3)	Arrowhead/Ventral veins lacking, POU2/3 class homeobox factor

wi-1	Wound-induced gene 1 (wi-1)
wi-2	Wound-induced gene 2 (wi-2)
Wnt	Wingless/Int signaling family
wnt11-1	Wnt family member 11-1
wnt11-2	Wnt family member 11-2
wntP-1	Posterior Wnt family member P-1
wntP-2	Posterior Wnt family member P-2
WT	Wild type
XRN1	5′→3′ exoribonuclease 1
YBX1	Y-box binding protein 1
yRF	Y RNA-derived fragment
ZFPL1	Zinc finger protein-like 1
zic-2	Zic family zinc finger protein 2
ZicA	ZicA zinc finger transcription factor
zpc family genes	Zona pellucida C domain–containing (zpc) family genes

1. INTRODUCTION

Small non-coding ribonucleic acids (sncRNAs) are key regulators involved in diverse biological processes (L.-L. Chen and Kim 2024). Over the last two decades, their repertoire has expanded from the three canonical classes, microRNA (miRNA), PIWI-interacting RNA (piRNA), and small interfering RNA (siRNA), to a broader landscape that includes molecules produced from other non-coding RNAs, called RNA fragments (RFs). This expanded catalog has revealed regulatory logic that extends beyond Argonaute (AGO)-centered RNA interference, including mechanisms linked to RNA modification, stress responses, and RNA–protein assembly (Ivanov et al. 2011; Wajahat et al. 2021). Despite progress, open questions remain about the origin, specificity, and physiological significance of many classes of RFs and their interaction with the canonical pathways. These uncertainties motivate systematic study of sncRNAs in different organisms and conditions (Gou et al. 2023) .

Regeneration is a biologically and medically important phenomenon in which sncRNAs are involved but their functional impact remains incompletely understood. The freshwater planarian *Schmidtea mediterranea* is a leading *in vivo* invertebrate model for regeneration: in this animal adult pluripotent stem cells repopulate all tissues after injury, enabling replacement of missing body parts and restoration of function (Reddien 2018). Cellular and molecular studies have mapped stem-cell states, patterning signals, and lineage outputs with increasing resolution, yet most work has emphasized protein-coding genes (Reddien 2021a). The extent to which sncRNAs coordinate wound responses, lineage specification, and tissue remodeling in planarians is not fully defined. Bridging this gap promises insights into conserved regeneration mechanisms and expands the comparative framework needed to translate principles across species.

Addressing these questions requires bioinformatic strategies that are tailored to the non-classical model. This thesis develops and applies approaches to profile sncRNAs during regeneration of *S. mediterranea* and address their functional implications by integrating sncRNA dynamics with mRNA expression and lineage-state transitions.

1.1. Landscape of sncRNAs

sncRNAs (typically 18–40 nucleotides (nt)) are often defined by size, however, in practice they are best distinguished by how they are produced and by their mechanisms of action. In animals, three canonical classes are recognized: miRNAs, siRNAs and

piRNAs (P. Zhang et al. 2019; George et al. 2024). In addition to these principal classes, numerous sncRNAs are generated from ncRNAs, in particular from tRNAs, rRNAs, snoRNAs, snRNAs, and Y RNAs. Such molecules are respectively called tRNA fragments, snoRNA-derived fragments, snRNA-derived fragments, rRNA-derived fragments, and Y RNA-derived fragments. Then, there are also promoter-associated small RNAs and transcription start site–associated RNAs. Across classes, functions of these molecules include post-transcriptional repression of messenger RNA (mRNA), modulation of translation, regulation of transposon and influence on chromatin (Y. Wang et al. 2016; Q. Chen and Zhou 2023; Onishi et al. 2021). **Figure 1** summarizes the taxonomy of ncRNA used in this thesis.

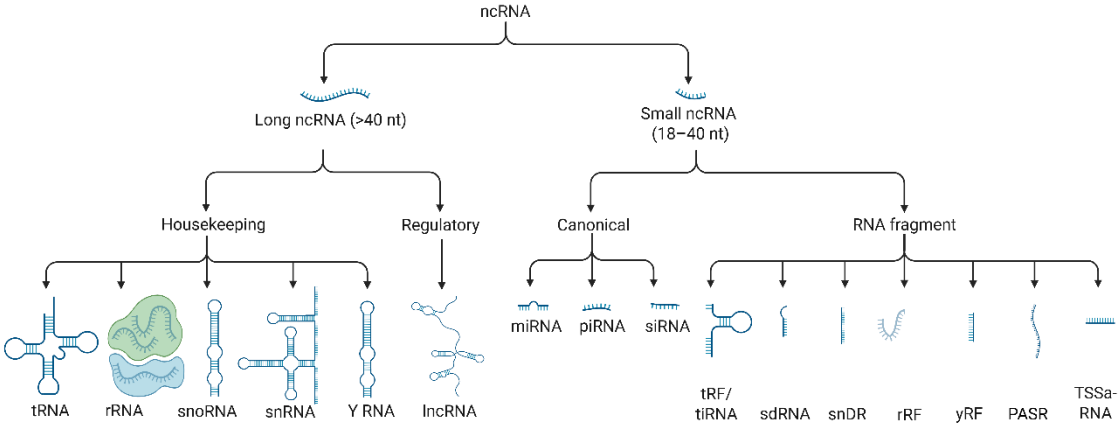

Figure 1. Taxonomy of ncRNAs. The diagram divides ncRNA first by size into two groups: long ncRNA (>40 nt) and sncRNA (18–40 nt). The label "long ncRNA (>40 nt)" is a size-based category only and must not be associated with long non-coding RNA (lncRNA), which is a distinct regulatory class shown under the separate branch. Within each size group, ncRNAs are then grouped by role (housekeeping and regulatory) or biogenesis (canonical and RFs).

1.1.1. Canonical sncRNAs

As introduced in **Section 1.1**, there are three canonical sncRNA classes. miRNAs are Drosha- and Dicer-processed products that are loaded onto AGO proteins and repress target mRNAs predominantly through seed-matched sites (Carthew and Sontheimer 2009a). piRNAs associate with PIWI proteins to silence transposable elements (Halic and Moazed 2009). siRNAs derive from long double-stranded RNA (dsRNA) and mediate RNA interference through AGO2-catalyzed cleavage of perfectly or near-perfectly complementary RNAs (Fire et al. 1998).

1.1.1.1. miRNAs

miRNAs are approximately 22 nucleotides in length regulatory RNAs that guide AGO proteins to complementary sites in mRNAs to repress gene expression post-transcriptionally. They are produced from endogenous hairpin precursors and act through conserved effector mechanisms that are now well characterized (Bartel 2018; Ha and Kim 2014; Jonas and Izaurralde 2015). Targeting is driven primarily by "seed" pairing: Watson–Crick base-pairing between miRNA nucleotides 2–8 (the seed) and complementary sequences typically located in the 3′ untranslated region (3′ UTR) of mRNAs (H. Guo et al. 2010; Bartel 2009, 2018). The most effective canonical mRNA target site types are 8-mer (perfect 2–8 pairing plus an A opposite miRNA position 1), 7-mer-m8 (perfect 2–8 pairing), and 7-mer-A1 (perfect 2–7 pairing with an A opposite position 1). Weaker 6-mer sites can act cooperatively when present in multiples. Pairing outside the seed contributes in two ways: (i) 3′-supplementary/compensatory pairing (typically to miRNA nt 13–16/17–19) can enhance site efficacy or partially rescue a seed mismatch, and (ii) rarer centered sites with contiguous pairing across miRNA nt 4–15 can mediate repression in specific contexts. Target site efficacy is further influenced by local sequence context (AU-rich flanks), site positioning within the 3′ UTR, and multiplicity/spacing of sites within a transcript (H. Guo et al. 2010; Bartel 2009, 2018).

Biological roles. miRNAs regulate gene expression mainly by triggering mRNA deadenylation and decay, with translational inhibition as an additional, often earlier but smaller, effect (H. Guo et al. 2010; Eichhorn et al. 2014; Jonas and Izaurralde 2015; Bartel 2018). Through these molecular outcomes, miRNAs modulate multiple essential processes, including cell-fate transitions, developmental timing, stress responses, and tissue homeostasis, acting within gene-regulatory networks together with transcription factors and RNA-binding proteins (Gurtan and Sharp 2013; Michlewski and Cáceres 2019).

Biogenesis. miRNA maturation proceeds through a two-step Ribonuclease (RNase) III pathway including: (i) DROSHA (nuclear processing) and (ii) DICER (cytoplasmic processing). The Microprocessor complex, with DROSHA as the catalytic RNase III enzyme and DGCR8 as the double-stranded RNA–binding partner, cleaves primary miRNA transcripts (pri-miRNAs) to release ~65-nt hairpin precursors (pre-miRNAs) (Ha and Kim 2014; Bartel 2018). Exportin-5/Ran- guanosine triphosphate (GTP) transports pre-miRNAs to the cytoplasm, where DICER cleaves near the loop to

produce a ~22-nt duplex with characteristic 2-nt 3′ overhangs. The resulting miRNA duplex is then loaded onto an AGO protein. The thermodynamic asymmetry of the duplex and the identity of the 5′ nucleotide determine which strand is retained as the guide, while the passenger strand is displaced (Ha and Kim 2014; Bartel 2018). The guide strand bound by AGO forms the core of the RNA-induced silencing complex (RISC). In animals, AGO–miRNA complex recruits Trinucleotide repeat-containing gene 6A protein (TNRC6), which then brings in enzymes that shorten mRNA tail (poly(A)-nuclease complex (PAN2–PAN3) and the Carbon catabolite repression 4 (CCR4)–negative on TATA-less (NOT) deadenylase complex) and the DEAD (Asp-Glu-Ala-Asp) box polypeptide 6 (DDX6) helicase. This leads to removal of mRNA cap by Decapping Protein 2 (DCP2)–DCP1 and subsequent 5′ to 3′ decay by 5'-3' exoribonuclease 1 (XRN1) (Jonas and Izaurralde 2015; Bartel 2018; Eichhorn et al. 2014).

1.1.1.2 piRNAs

piRNAs are approximately 24–35 nt molecules that form complexes with PIWI-family proteins. They are produced from long single-stranded precursors transcribed from genomic piRNA clusters, carry a 5′ monophosphate, and acquire a 2′-O-methyl group at the 3′ end that provides stability. These sncRNAs are amplified when target cleavage events generate new piRNA 5′ ends, which increases both the abundance and the sequence diversity of guides directed against active transposable elements (Czech and Hannon 2016; Ozata et al. 2019; J. Yu et al. 2024).

Biological roles. PIWI–piRNA complexes protect the genome in animal germlines by silencing TE through two cooperating routes (Czech and Hannon 2016). In the cytoplasm, PIWI proteins guided by piRNAs slice complementary transposon RNAs, which depletes these transcripts and seeds the production of new piRNAs. In the nucleus, PIWI–piRNA complexes direct co-transcriptional repression at transposon loci, for example by recruiting factors that establish H3K9me3-marked heterochromatin and, in mammals, by guiding *de novo* DNA methylation in the fetal male germline (Ozata et al. 2019). Beyond transposons, piRNAs also regulate subsets of germline mRNAs, including the turnover of spermatogenic transcripts, and they are essential for fertility, with the specific targets and outcomes varying across species and developmental stages of the germline (Du et al. 2024; Goh et al. 2015).

Biogenesis. Two coupled routes generate piRNAs (Riquelme et al. 2021). First, slicer-dependent “ping-pong” amplification occurs when complementary PIWI–piRNA complexes reciprocally cleave targets. Here, “slicer” refers to the endonuclease activity resident within the PIWI protein itself, not an external nuclease. Each cleavage defines the partner piRNA 5′ end, producing guide pairs with a diagnostic 10-nt 5′→5′ overlap and characteristic base preferences, typically1U bias in antisense partner and 10A bias in sense partner. Second, in primary, also called phased, processing, a mitochondrion-associated endonuclease, Zucchini, also called Phospholipase D family member 6 (PLD6), cleaves a single-stranded precursor to define an initial 5′ end on the outer mitochondrial membrane. Downstream piRNAs are then formed processively in a 3′→5′ direction along the transcript (Sun et al. 2022). Trimming by the exonuclease Poly(A)-specific ribonuclease (PARN)-like domain containing 1 (PNLDC1) and 2′-O-methylation by HEN methyltransferase 1 (HENMT1) finalize maturation. The relative contribution of ping-pong amplification and phased processing differs by species and by developmental stage of the germline (Sun et al. 2022; Weick and Miska 2014).

1.1.1.3 siRNAs

siRNAs are 21–23-nt duplex RNAs with 2-nt 3′ overhangs and 5′ phosphates that guide AGO proteins to cleave highly complementary RNAs (Fire et al. 1998; Bernstein et al. 2001; Elbashir et al. 2001). They are generated from long double-stranded RNA (dsRNA) by Dicer and subsequently loaded onto AGO2 to form the RISC complex.

Biological roles. siRNAs primarily mediate post-transcriptional silencing by AGO2-catalyzed slicing in the cytoplasm. They also are involved in antiviral defense and silencing of repetitive elements (Carthew and Sontheimer 2009a). Endogenous siRNAs (endo-siRNAs) can originate from cellular double-stranded RNA generated by convergent transcription and other sense-antisense duplexes, including pseudogene-derived antisense transcripts to their cognate mRNAs, as well as from structured hairpins. In contrast, exogenous siRNA (exo-siRNAs) can originate from viral infection or through experimental delivery (Carthew and Sontheimer 2009b). In plants and many fungi, RNA-dependent RNA polymerases copy target-related RNAs into new dsRNA, which Dicer converts into secondary siRNAs. This amplifies the silencing signal at the RNA level and supports long-range or systemic effects. siRNAs can also guide chromatin repression via RNA-directed DNA methylation in plants and heterochromatin assembly in fission yeast

(Baulcombe 2004; Asanuma et al. 2022). In animals, additional programs include endogenous siRNA regulation of coding transcripts in mouse oocytes and gene-specific repression from dedicated hairpin RNA loci in *Drosophila*, whose siRNA products are loaded into AGO2 (Watanabe et al. 2008; Okamura et al. 2008).

Biogenesis. Dicer cleaves long dsRNA into 21–23-nt siRNA duplexes bearing 2-nt 3′ overhangs (Bernstein et al. 2001). The duplex then is loaded into AGO2 that typically cleaves the passenger strand, leaving the guide strand bound (Matranga et al. 2005; Rand et al. 2005). The AGO2–siRNA complex recognizes near-perfectly complementary target RNAs and cleaves the phosphodiester bond opposite guide positions 10–11, followed by exonucleolytic decay of the resultant cleavage products (Hammond et al. 2000; J. Liu et al. 2004).

1.1.2. RFs

In addition to canonical miRNAs, siRNAs, and piRNAs, there exists a diverse class of sncRNAs, called RFs, that originate from a variety of other non-coding RNAs. Their nomenclature consistently reflects their source molecule type and, in certain cases, specifies the particular region within the parental RNA from which they are derived.

1.1.2.1. tRNA-derived fragments

In general, tRNA-derived fragments can be categorized into two principal groups: (i) tiRNAs (also termed tRNA halves) and (ii) shorter tRFs, also referred to as tDRs. tiRNAs (~30–35 nt) are generated by a cleavage in the anticodon loop that yields a 5′ half (from 5′ end to the anticodon loop) and a 3′ half (from the anticodon loop to the 3′ CCA). Although mostly stress induced, tiRNAs are also abundant under physiological conditions, for example in mammalian sperm, brain – including the hippocampus – in hematopoietic tissues and circulating blood, in human breast milk extracellular vesicles, in seminal plasma exosomes, and across other steady-state biofluids (Ivanov et al. 2011; Sharma et al. 2016; Godoy et al. 2018; Joseph Mohsen Dhahbi 2015; Gaylord et al. 2024; Jehn et al. 2020; Joseph M. Dhahbi et al. 2013). Shorter tRFs (~14–30 nt) include tRF-5, tRF-3, internal tRFs (i-tRFs), and tRF-1. tRF-5 species initiate at the mature 5′ end and terminate within the D-arm/anticodon stem, whereas tRF-3 species initiate within the T-arm and terminate at or just upstream of the 3′ CCA; i-tRFs derive from internal segments; tRF-1 is derived from the 3' trailers of precursor tRNA (pre-tRNA). In this thesis, tRFs is

hereafter used as a collective term for all tRNA-derived fragments, and species with a length of approximately 14–30 nt are specifically referred as short tRFs (tRF-5, tRF-3, i-tRF, and tRF-1) to avoid confusion. **Figure 2** summarizes the tRNA cleavage sites and the resulting classes of tRFs.

Biological roles. tRFs act through several well-characterized mechanisms. 5′ tiRNAs bearing terminal oligoguanine motifs repress cap-dependent initiation by displacing initiation factors and by promoting stress-granule formation (Ivanov et al. 2011). Short 18–22 nt tRF-3 species can be loaded onto AGO and repress mRNAs with complementary sites within 3′-UTR, in a miRNA-like manner (Kuscu et al. 2018). In turn, 3′ CCA-containing tRF-3 molecules recognize primer-binding sites on long terminal repeat (LTR) retrotransposons and restrict their activity (Schorn et al. 2017). Sperm-borne tiRNAs and short tRFs convey regulatory information to the zygote and shape early embryonic gene expression, and a tRF-1 species, tRF-1001, is required for rapid proliferation in human cell lines (Sharma et al. 2016; Y. S. Lee et al. 2009). Additional reports describe binding to ribosomal or ribosome-associated proteins and consequent modulation of translational efficiency (Mleczko et al. 2018). Furthermore, some tRFs modulate gene expression independently of AGO by associating with other RNA-binding proteins, such as Y box binding protein 1 (YBX1), and competitively displacing it from CA-rich sites on pro-metastatic mRNAs, which reduces the stability of those transcripts; this was shown for hypoxia-induced 5′ tRFs from tRNA-Glu, tRNA-Asp, tRNA-Gly and tRNA-Tyr in breast cancer models (Goodarzi et al. 2015).

Biogenesis. Currently, the tRF-1 series represents the only well-characterized class of tRFs derived from tRNA precursors. These molecules are generated from the 3′ trailers of pre-tRNAs, following their cleavage by the tRNA 3′ endonuclease RNase Z (ELAC2 in metazoans). Most classes of tRFs, however, arise from mature or nearly mature tRNAs (Jiang et al. 2022). Their biogenesis is mediated by several endonucleases, whose usage depends on the organism and context, for example physiological state or stress. In vertebrates, pancreatic-type RNase A-family enzymes, notably angiogenin (RNase 5) and RNase 1, cleave the anticodon loop to produce tiRNAs (Elder et al. 2024; B.-F. Fu and Xu 2022). RNase T2-family nucleases play analogous roles in many non-vertebrates and in plants, where T2 enzymes generate both tiRNAs and shorter tRFs (Su et al. 2020). During interferon responses, the 2′,5′-oligoadenylate synthetase (OAS)–RNase L antiviral endoribonuclease also produces tiRNAs from selected tRNAs (Elder

et al. 2024; B.-F. Fu and Xu 2022; Takenaka et al. 2025). In addition, DICER is involved in the biogenesis of specific ~18–22-nt tRF-3 and tRF-5 species, for example in human lymphoid cells and in *Arabidopsis*.(Yang et al. 2024; S.-I. Kim et al. 2025).

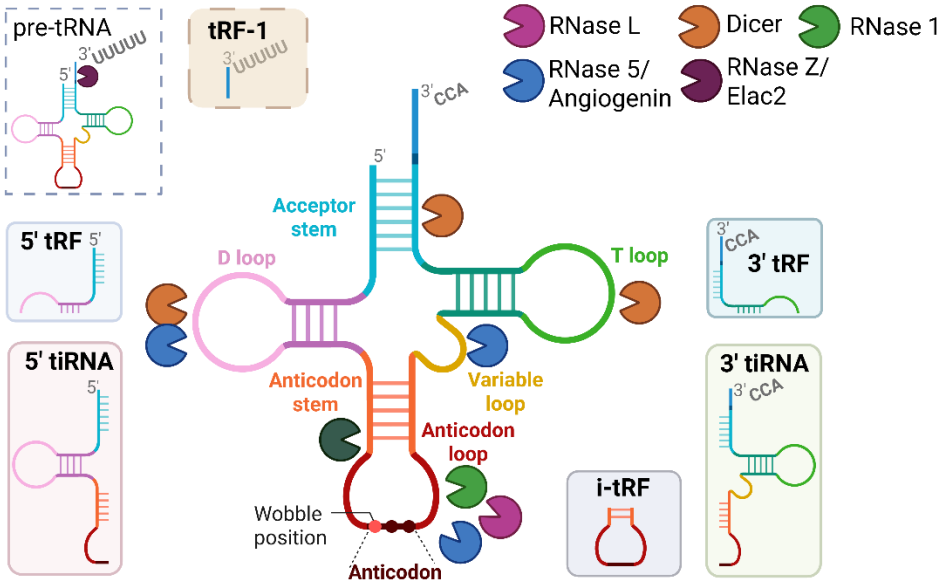

Figure 2. Map of tRNA cleavage sites and resulting tRF classes. The canonical cloverleaf structure is depicted with the following elements: acceptor stem, D loop, anticodon stem/loop, variable loop, T loop; 3′ CCA. The dashed inset highlights pre-tRNA. Major cleavage events are indicated by nuclease icons: (i) anticodon-loop cleavage produces 5′ tiRNA and 3′ tiRNA (~30–35 nt); (ii) D-arm/anticodon-stem cleavages yield tRF-5 (14–30 nt); (iii) T-arm/acceptor-junction cleavages yield tRF-3 (14–30 nt, often containing 3′ CCA); (iv) internal cuts generate i-tRFs; and (v) cleavage of pre-tRNA trailer results in tRF-1.

1.1.2.2. Other classes of RFs

Beyond tRFs, additional sncRNAs originating from non-coding RNAs are increasingly reported. Their biogenesis, prevalence, and mechanisms are, in general, less comprehensively characterized and more system-dependent than for tRFs.

Small nucleolar RNA-derived fragments (sdRNAs, referred to in this thesis as snoRFs) have been observed from both C/D and H/ACA box snoRNAs. In several systems, snoRFs are Dicer-processed into 20–25 nt species that are loaded onto Argonaute and repress targets in a miRNA-like manner. A well-studied example is the Cajal body–associated small Cajal body specific RNA 15 (SCARNA15 or ACA45), which produces an AGO-bound guide that represses Cyclin-dependent kinase 19 (CDC2L6) through seed-like pairing (Ender et al. 2008). Broader surveys and follow-up analyses have identified subsets of snoRNAs that generate AGO-associated sncRNAs

with regulatory potential, although most snoRFs appear to be cell type- and context-specific (Brameier et al. 2011; Taft et al. 2009; Kufel and Grzechnik 2019).

Small nuclear RNA-derived fragments (snDRs, referred to in this thesis as snRFs) have also been detected. Their accumulation and AGO association appear selective rather than global, and proposed functions range from miRNA-like repression to links with splicing programs, with evidence varying by system (Burroughs et al. 2011; Thomson et al. 2015).

Y-RNA-derived fragments (yRFs) are produced from conserved Y RNAs that normally bind the 60 kDa Ro/ Sjögren syndrome antigen (SSA) and La proteins. yRFs are upregulated during stress and apoptosis and have been implicated in RNA quality control and additional regulatory roles (Montoya 2014; Valkov and Das 2020). In vertebrate cells, intact Y RNAs are essential cofactors for initiating chromosomal DNA replication, whereas their cleavage generates yRFs with distinct regulatory roles, so fragments and precursors can coexist but act independently (Shaw et al. 2024; Valkov and Das 2020; Krude et al. 2009).

rRNA-derived fragments (rRFs) are widespread products from 5S, 5.8S, 18S and 28S rRNAs and their mitochondrial counterparts. Large-scale datasets indicate that rRF repertoires can be organized by rRNA of origin, length, and genomic position, and display cohort- and cell-type–dependent patterns consistent with regulated production rather than random degradation. Emerging reports connect rRFs to translation modulation and stress-granule dynamics, the precise mechanisms of their action remain under active investigation (Cherlin et al. 2020; Lambert et al. 2019).

Promoter-associated small RNAs (PASRs) and transcription start site–associated RNAs (TSSa-RNAs) are short RNAs that map around active promoters and often reflect divergent transcription. These species connect sncRNA production to chromatin states and transcription initiation, and in some systems, engage AGO complexes at promoter-proximal regions (D. Yu et al. 2018; Zamudio et al. 2014).

1.1.3. Bioinformatic approaches for sncRNA analysis

Analysis of sncRNA sequencing (sncRNA-seq) data is constrained by practical features of the molecules and of the libraries. sncRNAs are short and lack poly(A) tails, which necessitates the ligation of oligonucleotide adapters during library preparation to enable sequencing. Many sncRNAs originate from highly modified RNA species that are

also represented by numerous nearly identical genomic copies, such as tRNAs and rRNAs. A short read can align to several copies equally well, which is referred to as multi-mapping. These properties lead to characteristic challenges for data analysis: (i) adapter sequence carry-over, (ii) reverse-transcription (RT) stops or misincorporations at modified bases, (iii) and ambiguous assignment of multi-mapping reads. The simplified workflow below explains how each problem is handled in practice (G. Zheng et al. 2015; Behrens and Nedialkova 2022). An overview of the minimal analysis is shown in **Figure 3**.

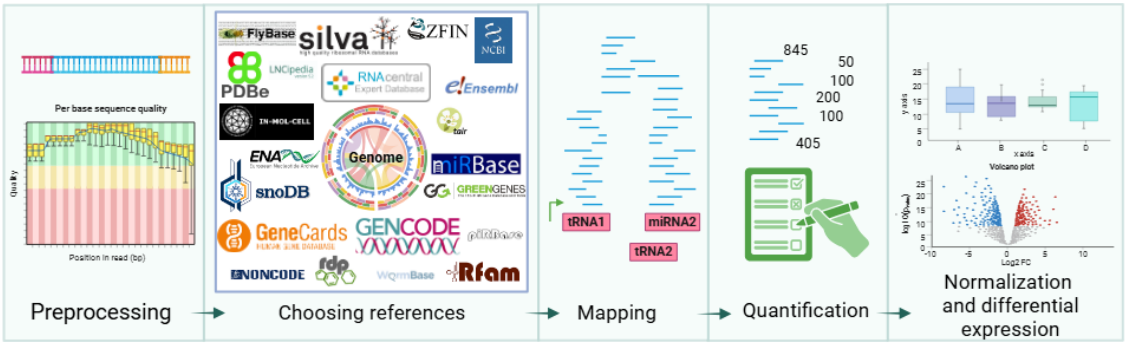

Figure 3. Minimal workflow for sncRNA-seq analysis. The schematic summarizes five core stages: preprocessing (adapter trimming and per-base QC), choosing references (genome and class-specific catalogs for tRNAs, rRNAs, miRNAs), mapping (short-read alignment tolerant of modification signatures), quantification (handling multi-mappers via fractional or family-level counting), and normalization with differential analysis (composition-aware models).

Preprocessing. In sncRNA libraries, the read is often longer than the insert (the sncRNA itself), so the 3′ sequencing adapter appears inside many reads. If not removed, aligners misclassify lengths and mis-map reads. Adapter-aware trimming searches for the known adapter sequence with a few allowed errors and removes it before mapping. This is routine and essential for sncRNA data (Martin 2011).

Choosing references. While mapping to the genome is a standard approach in gene expression profiling, sncRNA analysis must be complemented by the use of class-specific catalogs (databases of RNA sequences) to reduce ambiguity and enable annotation. For tRFs, the use of mature tRNAs and pre-tRNAs including trailers is recommended. For rRFs, references for 5S/5.8S/18S/28S and mitochondrial rRNAs should be included. For miRNAs, a curated mature set should be used. Dedicated resources such as miRbase, MINTbase, tRFdb, and tRFtarget, and broad ncRNA repositories such as RNAcentral,

assist with standardized identifiers and downstream annotation (Kozomara et al. 2019; N. Li et al. 2024; Kumar et al. 2015; Pliatsika et al. 2018; Sweeney et al. 2021). The goal is to reduce the number of false positives from multi-copy families and to report sncRNAs at the appropriate family or locus level.

Mapping. As mentioned above, multi-mapping represents a significant difficulty in sncRNA analyses, and it is therefore essential to determine how such reads should be counted. Possible solutions include assigning a single read to multiple loci that it maps to, or aggregating reads based on sequence similarities within a gene family, rather than assigning them to individual loci. Another challenge arises from the fact that tRNAs and rRNAs carry many base modifications that block reverse transcriptase or cause specific mismatches. Two strategies to address this limitation have been developed. Biochemical approaches, such as Alpha-ketoglutarate-dependent dioxygenase (AlkB)-assisted or demethylase-assisted protocols, including AlkB-facilitated RNA methylation sequencing (ARM-seq), demethylase (DM)-tRNA-seq, or modification-induced misincorporation (mim)-tRNAseq, reduce RT blocks and reveal otherwise hidden reads (Cozen et al. 2015; Clark et al. 2016). These methods, however, are not widely applied, necessitating reliance on bioinformatic solutions. Computational approaches allow limited, position-aware mismatches, so genuine modified sites are not discarded as errors (Behrens and Nedialkova 2022). In practice, careful trimming, stringent short-read parameters, and at least one of these modification-tolerant strategies are combined.

Quantification. sncRNAs often arise from families with many genomic copies, which causes difficulties in their proper quantification. Reasonable options to overcome this problem include fractional assignment of reads across equally good loci, family-level counting when loci are indistinguishable, and sensitivity analyses that test whether the conclusions derived from the data change under alternative reasonable assignments (Selitsky and Sethupathy 2015; Shi et al. 2018; Holmes et al. 2022).

Normalization and differential analysis. Count-based models such as DESeq2 remain standard for sncRNA analysis. However, they rely on the assumption that most genes remain stable across samples. For sncRNAs, where composition can shift between samples, this assumption may be violated. Thus, size-factor normalization (which corrects for differences in library size) should be applied with caution, library composition should be inspected, and any compositional effects that could skew fold-changes should be reported (Love et al. 2014; Evans et al. 2018).

While bulk sncRNA analysis is widely used and well established, single-cell sncRNA profiling remains technically challenging. Most high-throughput single-cell RNA-seq methods are droplet-based and rely on oligo(dT) capture of poly(A) tails, followed by 3′ end counting of long cDNAs. Single-cell datasets are stored as a count matrix. Rows represent features, typically genes, and columns – individual cells. Each entry is the number of molecules from that feature detected in that cell, usually unique molecular identifier counts. Mature miRNAs and many other sncRNAs lack poly(A) tails and are short so they cannot be efficiently captured or converted into the long cDNAs. Thus, none of the leading commercial single-cell platforms offers solutions for sncRNA profiling. Dedicated single-cell sncRNA protocols exist and can profile miRNAs and some RFs, but they remain lower-throughput and are sensitive to ligation and modification biases (Saliba et al. 2014; Hücker et al. 2021; Maji et al. 2025). In practice, two approaches are commonly used. First is based on enrichment or sorting the cell types of interest, followed by preparation of low-RNA-input sequencing libraries. If the cells are sorted one per well of a multiwell plate, each library is actually generated from a single cell, providing the required resolution (Kondratov et al. 2024; Juzenas et al. 2017; Grolmusz et al. 2016). Second, involves indirect inference of miRNA and miRNA-like activity from mRNA expression data using enrichment models, which has been shown to recover cell-type patterns at single-cell resolution even when miRNAs themselves were not directly captured (M. M. Nielsen and Pedersen 2021; Benesova et al. 2021).

1.2. *Schmidtea mediterranea* as a model for regeneration

S. mediterranea is a free-living flatworm (Phylum Platyhelminthes; Class Rhabditophora; Order Tricladida; Family Dugesiidae) within the protostome superclade Lophotrochozoa (Marlétaz et al. 2019) (**Figure 4**). It thereby provides a phylogenetic counterpoint to the widely used *Drosophila melanogaster* and *Caenorhabditis elegans* models, both of which represent Ecdysozoa superclade (Panteleimon Rompolas et al. 2009). Consequently, observations made in *S. mediterranea* are particularly significant because they provide insights that extend beyond Ecdysozoa, offering more universally applicable conclusions across a broader range of protostomes. As a bilaterian, *S. mediterranea* is triploblastic: derivatives of ectoderm, mesoderm, and endoderm are organized into organ systems, including a central nervous system with paired cephalic ganglia and ventral nerve cords, an eversible pharynx that connects to a branched

intestine, protonephridial excretory structures, body-wall muscle, and a ciliated epidermis. The organism is dorsoventrally flattened (approximately 4–8 mm long), with photo- and chemosensory auricles.

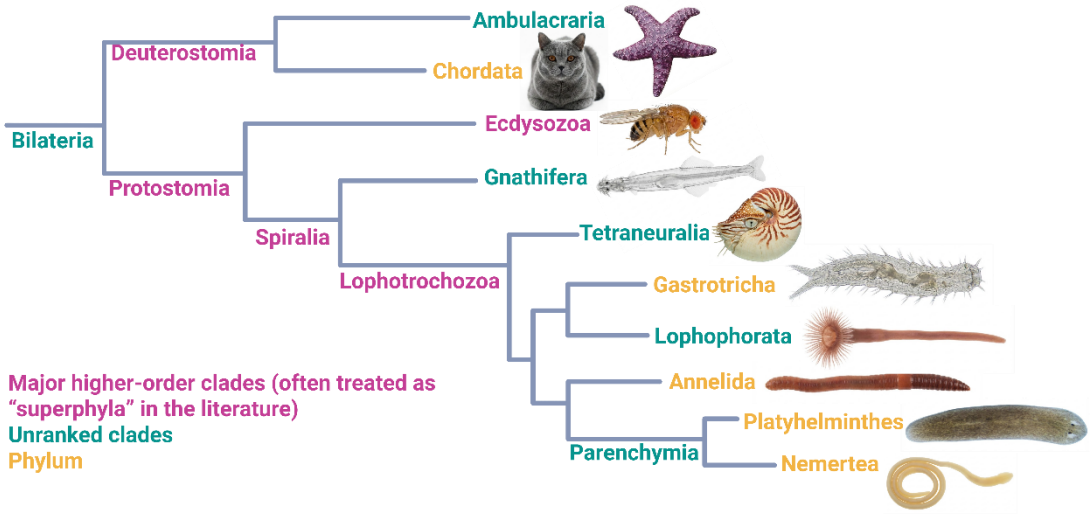

Figure 4. Simplified bilaterian phylogeny highlighting spiralian clades and the position of Platyhelminthes. The schematic tree of Bilateria showing the split into Deuterostomia and Protostomia, with Protostomia further divided into Ecdysozoa and Spiralia. Within Spiralia, major higher-order clades (magenta), unranked clades (teal), and phyla (gold) are indicated, illustrating the placement of Platyhelminthes (planarians) alongside Nemertea, Annelida, Gastrotricha, and lophophorate and tetraneuralian lineages.

In nature, *S. mediterranea* inhabits clean, slow-moving freshwater streams in parts of southern Europe and Mediterranean islands, while in the laboratory it is easily maintained and observed due to its small size and semi-transparent body. Two natural cytogenetically distinct biotypes exist: the sexual strain (S2) with a standard 2n = 8 karyotype, and the asexual strain (CIW4), also typically 2n = 8 but bearing a heteromorphic translocation between chromosomes 1 and 3. The sexual strain comprises cross-fertilising hermaphrodites that develop testes, ovaries and accessory glands. The asexual strain lacks reproductive organs and propagates by transverse fission followed by regeneration. These alternatives facilitate studies of germline specification and clonal propagation within a single species (Lázaro et al. 2011; Chong et al. 2011; Lázaro et al. 2011). Unless stated otherwise, this thesis refers to the asexual CIW4 biotype of *S. mediterranea*, which retains the somatic lineage architecture shared with sexual strains.

S. mediterranea is a cornerstone model for regeneration research. Regeneration is a process of cell replacement, triggered by injury, which initiates an inflammatory response for tissue restoration. Thus, it is distinct from routine cell turnover during homeostasis, which renews cells within intact tissues (Pellettieri and Alvarado 2007). Remarkably, the planarian body can be divided into many fragments, each of which regenerates a complete animal within days. This capacity is sustained by neoblasts, adult pluripotent stem cells distributed throughout the parenchyma. Neoblasts divide and generate immediate progenitors that differentiate into all somatic lineages. Classic irradiation and transplantation experiments established that a single clonogenic neoblast can reconstitute an entire animal, demonstrating true pluripotency (Wagner et al. 2011). Because amputation triggers a synchronous, organism-wide program, regeneration in *S. mediterranea* can be staged reproducibly and aligned with morphological changes such as blastema formation and eye reappearance. **Figure 5** presents a concise overview of this process. An immediate response to injury begins within minutes, followed by a rapid increase in neoblast proliferation. Next, a mass of undifferentiated cells, called blastema, forms at the site of injury. The blastema generates progenitor cells with a limited cell lineage commitment. Subsequently, organs such as the brain, ventral nerve cords, eyes, pharynx, intestine, protonephridia, epidermis, and muscles are rebuilt as a result of differentiation and tissue pattern formation, leading to the restoration of function. The fact that this process occurs in a predictable staged manner, forms the basis of the experimental design in planarian regeneration research.

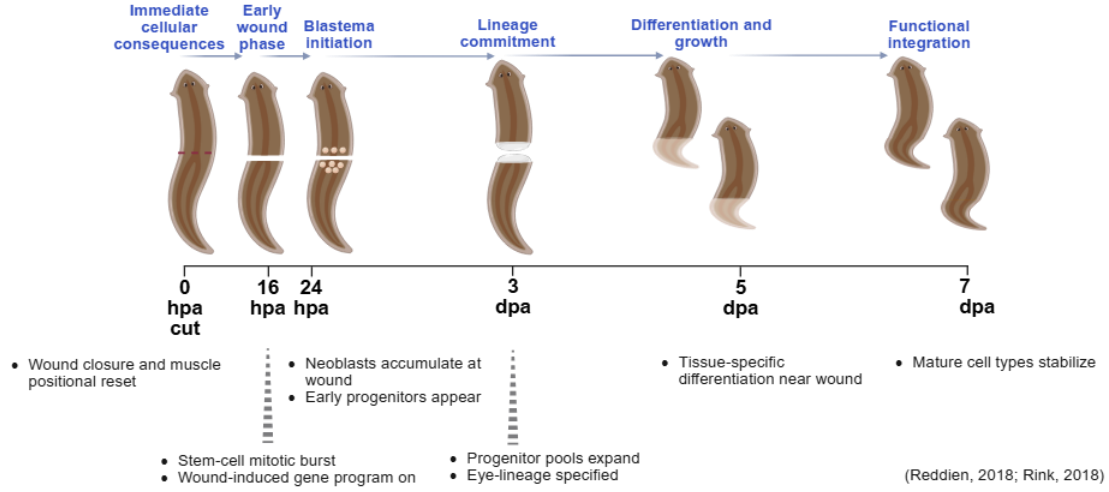

Figure 5. Regeneration timeline in *S. mediterranea.* Schematic of post-amputation phases up to 7 days post amputation (dpa). Immediately after cutting, injury signals trigger a body-wide wound program and a rapid neoblast proliferation. In the early wound phase, neoblasts accumulate

near the wound while muscle-derived positional cues are refined. Blastema then forms a proliferative, undifferentiated cap that seeds lineage commitment. Differentiation and growth expand neuronal, epidermal, gut and other tissues. Functional integration restores organ architecture and behavior.

1.2.1. Cell types

Adult planarians comprise a proliferative stem-cell compartment and a wide spectrum of differentiated tissues. Several studies have applied single-cell RNA sequencing (scRNA-seq) to characterize specific cell types; however, two whole-organism atlases of *S. mediterranea* currently serve as the standard references. These atlases encompass the major cell types and many subtypes, organizing them into lineage trajectories rooted in neoblasts, and capturing wound- and region-specific states (Fincher et al. 2018; Plass et al. 2018). The following chapters provide a brief overview of the planarian cell types, with a focus on their marker genes, which are summarized in **Table 1.** Throughout this thesis, planarian gene symbols are written in lowercase italics (e.g., *smedwi-1*, *foxA*, *wnt11-5*). Proteins are in roman type (e.g., PC2, Laminin B), gene families are in lowercase roman type (e.g., lectins, tektins), bundled or slash-separated gene sets are italicized as a unit (e.g., *gata4/5/6*, *smedwi-1/-2/-3*), "-like" designations and planarian contig identifiers (e.g., *nkx2-like*, *dd_554*) are treated as gene symbols and italicized. Ortholog-style gene symbols (e.g., *DNAI1, HYDIN, PRSS12, FOXJ1*) are used in upper-case in order to avoid inventing new lowercase forms where a planarian-specific gene symbol does not exist. Lastly, by convention in the planarian community, locus-style identifiers such as SMESG0000… are presented in roman type (not italicized), as they function as stable database identifiers rather than standardized gene symbols.

Table 1. Cell types in *S. mediterranea* and their commonly used markers.

Cell type	Cell subtype	Common markers	References
Epidermis	Early epidermal progenitor	*myb-1*, *prog-1*, *prog-2*	(Fincher et al. 2018; Cheng et al. 2018)
	Epidermal secretory gland	lectins, *MAGT1*, *RCN1*	(H. O. King et al. 2024; Zayas et al. 2010)
	Epidermal secretory gland progenitor	*fer3l-2*	(H. O. King et al. 2024)
	Injury-induced epidermis	*equinox*	(Scimone et al. 2022)
	Late epidermal progenitor	*GATM*	(Tu, Cheng, Vu, et al. 2015)
	Mature epidermis (DV-boundary)	*BARHL1*, *HOXC12*, *LAMB1*, *tlx-1*	(Fincher et al. 2018; Wurtzel et al. 2017)

	Epidermis (broad)	*PRSS12*, zpc family genes	(Tu, Cheng, Vu, et al. 2015)
	Epidermis (multiciliated)	*DNAI1, FOXJ1, HYDIN, IFT88, TEKT1, TEKT2, TEKT3, TEKT4*	(Tu et al. 2015; Guemez-Gamboa et al. 2014; Nishie et al. 2023)
Eye lineage	Eye photoreceptor	*ARR2, opsin*	(Lapan and Reddien 2011)
	Eye progenitor	*eya, otxA, SIX2*	(Lapan and Reddien 2011; Vásquez-Doorman and Petersen 2016)
	Pigment cup cell	*DLX1, SP1 (sp6-9)*	(Lapan and Reddien 2011, 2012)
Intestine	Basal cell	*slc22a-6*	(Forsthoefel et al. 2020)
	Goblet cell	*NPC2*	(Forsthoefel et al. 2020)
	Phagocyte (broad)	*A1CF, CTSL, RBM47 (APOB-1)*	(Forsthoefel et al. 2020; Wong et al. 2022)
Muscle	Anterior pole cell	*foxD, ndl-2, notum, zic-2*	(Scimone et al. 2018; Vogg et al. 2014; Witchley et al. 2013)
	BWM (circular)	*nkx1-1*	(Scimone et al. 2017)
	BWM (dorsal midline)	*UNC5A, bambi*	(Schad and Petersen 2022)
	BWM (longitudinal)	*myoD*	(Scimone et al. 2017)
	ECM-producing muscle	*COL4A1, hmcn-1*	(Cote et al. 2019)
	Posterior pole/PCG muscle	*FZD4, wnt11-2, PTK7, wntP-1, wntP-2, wnt11-1*	(Owlarn and Bartscherer 2016; Witchley et al. 2013)
Neoblast	Eye lineage neoblast	*ovo*	(Lapan and Reddien 2012)
	Muscle neoblast	*myoD*	(Scimone et al. 2014)
	Parenchymal neoblast	*GLIPR1L1, nkx6-2*	(H. O. King et al. 2024)
	Protonephridial neoblast	*VVL (POU2/3)*	(Fincher et al. 2018)
	Pharyngeal neoblast	*foxA*	(Adler et al. 2014)
	γ-neoblast (intestinal-fated)	*gata456, hnf4, NKX2-2, PROS*	(Fincher et al. 2018)
	ζ-neoblast (epidermal-fated)	*EGR1, SOXP-3, TRP63, ZFPL1*	(Cheng et al. 2018)
	ν-neoblast (neural-fated)	*neuroD, NKX6-1, SCRO, soxB, STNB, TGS1*	(Molina and Cebrià 2021; King et al. 2024)
	σ-neoblast (broad-lineage)	*AGO3, smedwi-1/-2/-3 (PIWIL1, wi-1, wi-2), SOXP-1, SOXP-2, bruli*	(Barghouth et al. 2019; Niu et al. 2021)
	cNeoblast	*tspan-1, tgs-1*	(Zeng et al. 2018; Pearson 2022)
Nervous system	Brain branch neuron	*Gpas*	(Fincher et al. 2018)
	Catecholaminergic neuron	*dat/slc6a-3, th/tph*	(Fraguas et al. 2012)
	Cholinergic neuron	*CHAT, slc18a-6*	(Roberts-Galbraith et al. 2016)
	GABAergic neuron	*gad, slc6a-1*	(Duba-Kiss et al. 2021; X. Wang et al. 2023)
	Glia	*estrella, calamari (cali)*	(Roberts-Galbraith et al. 2016)
	Glutamatergic neuron	*slc17a-6, slc20a-1*	(Molina and Cebrià 2021)
	Glycinergic neuron	*slc6a-5*	(Hirata et al. 2011;

			Zeilhofer et al. 2018)
	Mechanosensory neuron	*pou4l-1*	(McCubbin et al. 2025)
	Neuropeptidergic neuron	*AMON, CPE, PCSK1*	(Currie et al. 2016; H. Yu et al. 2022)
	PKD^{+} sensory neuron (ciliated)	*PKD1L-2, PKD1L-3, PKD2-1, PKD2-3, PKD2-4*	(Fincher et al. 2018)
	Serotonergic neuron	*AWH, LHX1, LHX2, LHX5, LMX1A, LMX1B.2, PITX3/PTX1, slc6a4/sert, tph*	(Currie and Pearson 2013; Fincher et al. 2018)
	Tyraminergic neuron	*TDC2*	(Pauls et al. 2018)
	otp^{+} neuron	*otp*	(K. G. Ross et al. 2024)
Parenchyma	Abraçada cell	*CTSL2*	(Settles et al. 2025)
	AQP^{+} parenchymal cell	Aquaporins (e.g., *AQP4, AQP6*)	(Plass et al. 2018)
	FER3L-2^{+} *parenchymal progenitor*	*fer3l-3*	(H. O. King et al. 2024)
	GLIPR-1^{+} *parenchymal progenitor*	*GLIPR*	(H. O. King et al. 2024)
	ldlrr-1^{+} parenchymal cell	*ldlrr-1*	(Roberts-Galbraith et al. 2016)
	NKX2^{+} *parenchymal progenitor*	*NKX2*	(H. O. King et al. 2024)
	SSPO^{+} parenchymal cell	*SSPO*	(H. O. King et al. 2024)
	SSPO^{+} parenchymal progenitor	*fer3l-1*	(H. O. King et al. 2024)
	PGRN^{+} parenchymal cell	*PGRN-1*	(Emili et al. 2025)
	PSAP^{+} parenchymal cell	*PSAP*	(Plass et al. 2018)
	ptf^{+} head parenchymal progenitor	*ptf-1, dd_8476*	(H. O. King et al. 2024)
Pharynx	Pharyngeal epithelium	*foxA1*	(Adler et al. 2014)
	Pharyngeal muscle	*DMRT2, foxC1, MEIS1, HTH, musculin, coe, sim, ZicA, ap2*	(King et al. 2024; Fincher et al. 2018; Molina et al. 2023)
	Pharyngeal progenitor	*dd_554*	(Zhu et al. 2015)
Protonephridia	Protonephridial flame cell	*EGFR/ERBB4, MMEL1/dd_5256, dd_2920*	(Fincher et al. 2018)
	Protonephridial tubule cell	*ALPPL2, SIX1, SIX2, dd_9435, cav-1*	(Scimone et al. 2011; Fincher et al. 2018; Rink et al. 2011)
	Protonephridial precursor	*HB, dd_10830*	(Scimone et al. 2011; Fincher et al. 2018)
Pigment	Body pigment cell	*ALAD, ALAS, HMBS*	(Stubenhaus et al. 2016)

1.2.1.1. Neoblasts and their immediate progeny

Neoblasts constitute approximately 20-30% of all cells. Classic irradiation and transplantation experiments demonstrated that a single clonogenic neoblast (cNeoblast) can repopulate an animal *in vivo*, demonstrating true pluripotency (Wagner et al. 2011). Subsequent studies then revealed heterogeneity within the dividing stem cell pool (van Wolfswinkel et al. 2014).
The single-cell atlases resolved cycling neoblasts ($S/G_2/M$) with lineage-biased subgroups that give rise to immediate post-mitotic progenitors (G_0) and then differentiated branches (H. O. King et al. 2024; Fincher et al. 2018; Plass et al. 2018). A working model of the stem-cell population has emerged from these datasets and functional assays. A pluripotent cNeoblast core sustains all somatic lineages, while four recurrent classes of cycling, lineage-biased neoblasts precede many progenitor pools: σ-neoblasts with broad multipotency, ζ-neoblasts biased to epidermis, γ-neoblasts biased to intestine, and ν-neoblasts biased to neural lineages (van Wolfswinkel et al. 2014). These cycling populations produce G_0 "early progenitors" that then differentiate into tissue-specific progenitors and mature cell types. Notably, it was demonstrated that σ-neoblasts possess broad lineage capacity and can give rise to ζ-neoblasts. In contrary, ν-neoblasts and γ-neoblasts are understood as specialized neoblast classes that are derived from pluripotent cNeoblasts (Reddien 2018, 2021a). In parallel with these subclasses labelled with Greek letters, it has been shown that several experimentally defined subgroups of neoblasts initiate specific cell lines. Such subclusters were described for eye, muscle, parenchymal, protonephridial and pharyngeal lineages, each defined by expression of distinct fate-specific transcription factors. Recent clonal and single-cell RNA-seq analyses indicate that many specialized neoblasts are not irreversibly committed but can undergo fate switching, such that an epidermal ζ-neoblast, for example, can generate progeny that adopt intestinal or neural fates (Reddien 2021a). However, the precise hierarchy and connectivity among specialized neoblast classes, beyond the established σ-to-ζ relationship, remain incompletely resolved and will ultimately require direct long-term lineage tracing at single-cell resolution.

Notably, recent single-cell studies indicate that each tissue follows a distinct "fate allocation" program, defined by when along the neoblast → progenitor → mature axis transcriptionally distinct subtypes first appear (**Figure 6**) (H. O. King et al. 2024; van Wolfswinkel et al. 2014; Reddien 2021a). In many lineages, fate allocation is already

evident within the cycling neoblast compartment. In contrast, in the nervous system the first major diversification happens after cell cycle exit: early G_0 progenitors split into multiple, lineage-specific branches with distinct repertoires of fate-determining transcription factors. Lastly, protonephridial neoblasts give rise to post-mitotic progeny that diversify more gradually at mature stage into distal, tubule and flame-cell branches (H. O. King et al. 2024).

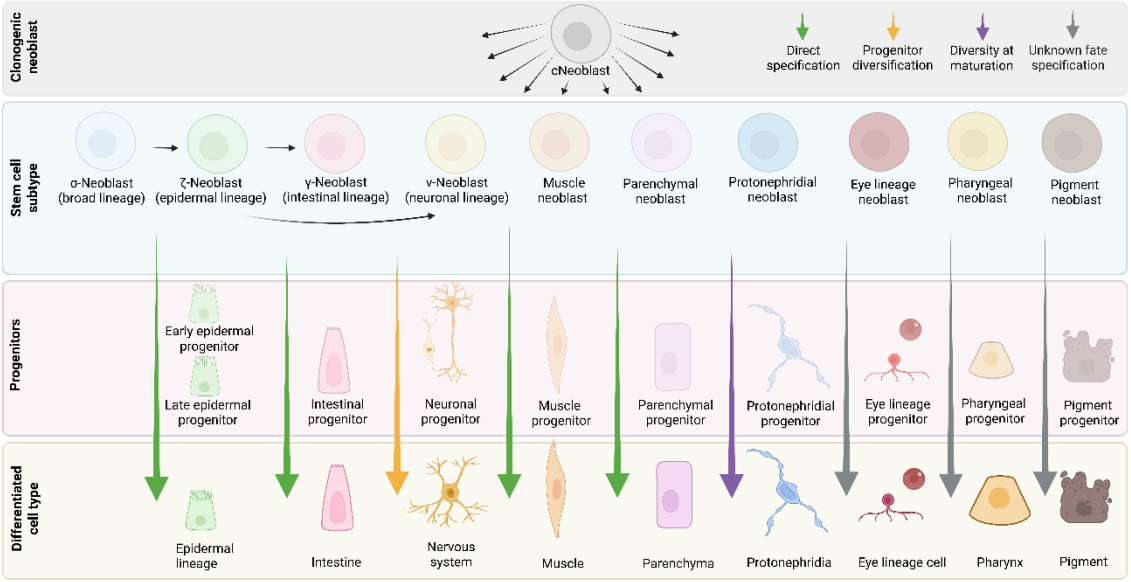

Figure 6. Differentiation of cell lineages in asexual *S. mediterranea*. The scheme summarizes how a single clonogenic neoblast (cNeoblast) seeds multiple lineage-biased neoblasts, which then generate progenitors and terminal cell types. Three fate-allocation programs are indicated across lineages: direct specification within cycling neoblasts, progenitor diversification after cell-cycle exit, and diversity at maturation. Unknown fate specification is indicated with grey arrow.

Distinct neoblast subpopulations are characterized by the expression of specific marker genes. All dividing neoblasts express PIWI-family genes (*smedwi-1/-2/-3*). *tspan-1* (the surface protein Tetraspanin-1) is strongly expressed in clonogenic neoblasts (van Wolfswinkel et al. 2014; Reddien 2018; Dai et al. 2025; X. Chen 2025). Broad lineage σ-neoblasts are characterized by PIWI-high, *soxP-1/-2* and *bruli* expression (van Wolfswinkel et al. 2014). γ-neoblasts, biased to intestinal and secretory trajectories, are characterized by the expression of *hnf4*, *gata4/5/6*, *prox-1* and *nkx2.2* (Fincher et al. 2018). The epidermal ζ-neoblasts, express *egr-1*, *soxP-3* and *zfp-1* (Wurtzel et al. 2017; Fincher et al. 2018). The neural ν-neoblasts show the expression of *nkx6-like*, *soxB*, *tgs-1* and *neuroD-1* (Molina and Cebrià 2021; H. O. King et al. 2024; Fincher et al. 2018). Eye-lineage neoblasts, expressing *ovo* generate the photoreceptors, specialized cells in

the retina that detect light and convert it into electrical signals processed by the brain as an image, and pigment-cup progenitors; loss of *ovo* impairs eye formation (Lapan and Reddien 2012). A *myoD*$^+$ neoblast subset generates body-wall muscle progenitors and is required for myogenesis during regeneration (Reddien 2021a). Across recent single-cell atlases, parenchymal lineages were initially resolved into several mature subtypes (*SSPO*$^+$, Aquaporins$^+$, *ldlrr-1*$^+$, *PGRN-1*$^+$, *PSAP*$^+$). Only subsequently were neoblast and progenitor markers identified for a subset of these lineages, including a *GLIRP1L1*$^+$/*nkx6-2*$^+$/NKX2$^+$ parenchymal neoblast population that gives rise to *SSPO*$^+$, and head parenchymal progenitors expressing *fer3l-1* and *ptf-1*. In contrast, subtype-specific neoblast and progenitor states for the Aquaporins$^+$, *ldlrr-1*$^+$, *PGRN-1*$^+$ and *PSAP*$^+$ parenchymal lineages have not yet been resolved as discrete classes and are viewed as arising from a shared parenchymal neoblast pool (H. O. King et al. 2024). *POU2/3*$^+$ neoblasts give rise to protonephridial tubule and flame-cell lineages (Fincher et al. 2018). Pharynx itself lacks resident neoblasts, but there are pharynx-specialized neoblasts outside the organ, marked by *foxA* expression (Adler et al. 2014; Reddien 2018).

1.2.1.2. Epidermal lineage

The epidermal trajectory is among the best resolved in *S. mediterranea*. Heterogeneous σ-neoblasts generate ζ-neoblasts, which in turn produce early epidermal progenitors (*prog-1*$^+$) and late epidermal progenitors (*AGAT-1*$^+$). These post-mitotic progenitors arise in the subepidermal parenchyma and migrate toward the body surface: early *prog-1*$^+$ cells occupy deeper subepidermal positions, late *AGAT-1*$^+$ cells lie more superficially, and terminally differentiated epidermal cells integrate at the apical surface. The transcription factor-encoding gene *Egr-5* functions post-mitotically to promote terminal epidermal differentiation, and additional regulator-encoding *zpuf-6* marks late maturation (Tu, Cheng, Vu, et al. 2015; Kent et al. 2026) . These markers are used routinely for fate mapping and to define stages of regeneration. Specialized secretory and gland cells differentiate in parallel and extend ducts to epidermal pores (Fincher et al. 2018). Secretory gland cells are distinguished by the expression of *mag-1*, lectins and *RCN1*, and their *fer3l-2*$^+$ progenitors give rise to multiple gland subclasses within the epidermal layer (H. O. King et al. 2024; Zayas et al. 2010). Regionally specialized mature epidermis at the dorsal–ventral (DV) boundary shows *post-2c*, *BARHL1*, *tlx-1* and *LamininB* expression (Fincher et al. 2018; Wurtzel et al. 2017). Broad mature epidermis

is enriched for ZP-domain proteins and the expression of *PRSS12* (Tu, Cheng, Vu, et al. 2015). A transient injury-induced epidermal state that is broadly activated after wounding and declines as regeneration proceeds, is marked by *equinox* (Scimone et al. 2022). The planarian epidermis includes also multiciliated cells required for gliding movement. They carry dense fields of motile cilia, which are hair-like, microtubule-based organelles structurally equivalent to flagella; their coordinated beating moves fluid along the body surface and propels the animal. The core scaffold of each cilium, the 9+2 axoneme, depends on axonemal dynein and central-pair components such as *DNAI1* and *HYDIN*, and is assembled by the intraflagellar transport machinery that includes *IFT88*. The conserved transcription factor FOXJ1 controls the motile-ciliogenesis gene set (Guemez-Gamboa et al. 2014; Nishie et al. 2023; Boehlke et al. 2015).

1.2.1.3. Muscle cells

Body-wall muscle serves both as contractile tissue and as a central source of positional information. Canonical muscle markers include ECM components (collagens) and myogenic regulators. In addition, planarian muscle cells constitutively express PCGs that encode conserved patterning proteins (Wnt/β-catenin, BMP, FGF, Hedgehog) and associated modulators. During regeneration, the spatial domains of PCG mRNAs and proteins are re-established around wounds, restoring axial coordinates (Fincher et al. 2018). Anterior pole at the head tip comprises a small organizer that expresses *notum*, *zic-2*, *foxD* and *ndl-2*; these factors mark and help establish anterior identity during regeneration (Scimone et al. 2018; Vogg et al. 2014; Witchley et al. 2013). Circular body-wall muscle is defined by *nkx1-1* activity and contributes to contractility and patterning (Scimone et al. 2017). A distinct dorsal-midline muscle stripe expresses *UNC5A* and *bambi*, marking a specialized dorsal-ventral (DV) boundary domain implicated in midline patterning (Schad and Petersen 2022). Longitudinal body-wall muscle depends on the expression of *myoD* for specification and maintenance, as loss of *myoD* disrupts longitudinal fibers and impairs regeneration (Scimone et al. 2017). The extracellular matrix (ECM)-producing muscle subset expresses *COL4A1* and *hmcn-1*, and secretes collagens and hemicentin into the ECM, shaping the biomechanical and signaling microenvironment for regeneration (Cote et al. 2019). Posterior pole and PCG-rich posterior muscle express *wnt11-1*, *wnt11-2*, *wnt11-5* (*wntP-2*), *wntP-1*, *FZD4* and *PTK7*,

forming a tail organizer that complements anterior *notum*-positive muscle in setting global axial polarity (Owlarn and Bartscherer 2016; Witchley et al. 2013).

1.2.1.4. Nervous system

The planarian central nervous system (CNS) comprises bilateral cephalic ganglia connected by a commissural bridge and paired ventral nerve cords (VNCs) that extend along the anterior-posterior axis (Cebrià 2008; Kuznetsov et al. 2024; Molinaro and Pearson 2016). Remarkably for an adult CNS, neuronal turnover is continuous: pluripotent neoblasts generate lineage-biased neoblasts that further differentiate into post-mitotic neuronal progenitors, which eventually replace neurons during homeostasis and regeneration. The neuronal cells are marked by the expression of pan-neuronal and neuropeptide genes, including *pc2*, and synaptic components, such as *synapsin*. Prohormone convertase 2 (PC2) is required for neuropeptide maturation, and synapsin is a presynaptic vesicle-associated protein that labels synaptic terminals (Currie et al. 2016). Midline guidance is controlled by the Slit–Robo pathway. Midline cells secrete Slit, which activates Robo on axons to generate repulsion that keeps the ventral nerve cords separated and guides commissural axons to cross only at defined commissures, then turn away to extend along the opposite cord rather than running along the midline (Yuasa-Kawada et al. 2022; Fincher et al. 2018). In planarians there are several types of neurotransmitter-producing neurons, localized in specific domains within the brain lobes and along the cords. The identity of these cells is specified by the expression of conserved transmitter-specific genes: *CHAT* for cholinergic, *tph* for serotonergic, *gad* for GABAergic, *th* for dopaminergic, and *TDC2* for tyraminergic neurons (Fincher et al. 2018). After head or trunk amputation, CNS regeneration begins with rapid wound closure and epithelial sealing, followed by a wound-induced gene program and two mitotic waves that expand neurogenic progenitors. As guidance cues are re-established, photoreceptor and brain primordia then extend axons and refine their connections. This progression can be quantified by measuring, over time, the number of $pc2^{+}$ neurons and the expression of transmitter-specific genes.

1.2.1.5. Eye lineage

Planarian eyes are composed of cup-shaped pigment cells and rhabdomeric photoreceptors whose axons cross at the optic chiasmata before terminating in the brain

lobes. Eye lineage is initiated by *ovo*$^{+}$ neoblasts that give rise to post-mitotic eye progenitors, defined by conserved eye transcription factors *otxA*, *eya* and *six*, and then diverge into photoreceptor and pigment-cup branches (Lapan and Reddien 2012). Terminal photoreceptor identity is marked by opsins and arrestins, while pigment-cup cells are specified by *dlx* and *sp6-9* gene expression and assemble the pigmented optic cup that surrounds photoreceptor rhabdomeres (Yamamoto and Agata 2011; Reddien 2021b; Saad et al. 2025). After decapitation, eye primordia appear and establish projections with highly reproducible timing. Since eye formation is predictable and observable, it provides a reliable indicator of how the experimental manipulations affect the organism.

1.2.1.6. Protonephridia (excretory system)

Planarian protonephridia form a branched epithelial network that filters interstitial fluid via ciliated flame cells and conveys the filtrate through proximal and distal tubules into collecting ducts that open at excretory pores. Ciliary beating drives inflow, and downstream epithelia reabsorb solutes to maintain osmotic balance (Rink et al. 2011; Thi-Kim Vu et al. 2015; Vij et al. 2012). Flame cells are marked by the expression of *dd_2920* and *ECE2* (*dd_5256*) and require *egfr-5* for maintenance and branching (Fincher et al. 2018; H. O. King et al. 2024). Proximal tubules express markers such as *SIX1/2*, *dd_9435* and *ALPPL2* (*dd_8942*), while distal tubules are defined by *cav-1* (H. O. King et al. 2024). A post-mitotic tubule precursor population expresses *HB* (Hunchback) and *dd_10830*. Transcriptional regulators *POU2/3* and *eya* are required for protonephridial regeneration and maintenance (Scimone et al. 2011; H. O. King et al. 2024; Fincher et al. 2018).

1.2.1.7. Intestine cells

The planarian intestine is a branched epithelium that digests, absorbs and distributes nutrients via a lumen lined by highly phagocytic enterocytes. Enterocyte identity is specified by the expression of conserved endodermal genes, including *gata4/5/6*, *nkx2.2*, and *hnf4* (Fincher et al. 2018). Three principal epithelial cell types can be distinguished: (i) phagocytes, expressing genes involved in lysosomal function and lipid metabolism, such as cathepsins and *APOB-1*; (ii) secretory goblet cells, which release proteins into the lumen and are marked by the expression of *NPC2*; and (iii) an outer basal cell layer, marked by *slc22a6*, that interfaces with surrounding parenchyma

(Forsthoefel et al. 2020). Phagocytes increase endo-lysosomal activity after feeding or injury. During regeneration they repopulate the gut in parallel with γ-biased neoblast input, which accelerates phagocytosis and digestion (Fincher et al. 2018; Forsthoefel et al. 2020; Wong et al. 2022).

1.2.1.8. Pharynx

The planarian pharynx is a muscular, ciliated and secretory feeding organ that everts through the ventral mouth to grasp and ingest food, then retracts to its position within the midline cavity (Fincher et al. 2018). This dual contractile–epithelial identity of pharyngeal cells is underlain by the expression of genes encoding myogenic regulators, ion transporters and secretory factors, as well as genes involved in ciliogenesis and intraflagellar transport (IFT) (Fincher et al. 2018; Plass et al. 2018). Pharynx regeneration is rapid and follows a well-defined, temporally ordered program. Selective pharynx amputation induces an early expansion of *foxA*$^{+}$ progenitors arising from surrounding neoblasts, and a dedicated regeneration program. Notably, *foxA* knockdown blocks pharynx formation, establishing *foxA* as a lineage determinant (Adler et al. 2014). Pharyngeal epithelium is marked with the expression of *foxA*; pharyngeal muscle with *foxC1*, *meis*, *musculin*, *ap-2*, *zicA*, *DMRT2*, *sim* and *coe*; and post-mitotic pharyngeal progenitors with *dd_554* (Adler et al. 2014; Fincher et al. 2018; Molina et al. 2023; Zhu et al. 2015).

1.2.1.9. Parenchymal cells

In planarians, parenchyma denotes the mesenchymal, non-epithelial compartment that occupies the space between the epithelia and the organs. It is composed of multiple secretory and support cell types that are neither epidermis, muscle, nor intestine. Parenchymal cells contribute to ECM turnover, secretion (mucus and adhesive factors), intercellular signaling and wound remodeling (Fincher et al. 2018; Plass et al. 2018; Wurtzel et al. 2017). Fate bias is already detectable at the S/G_2/M neoblast stage: transcription factor–encoding genes, including *nkx6-2* and *GLIPR1L1*, appear in S/G2/M neoblast subsets and persist as recognizable signatures in immediate G_0 progenitors. Silencing of *nkx6-like* using RNAi selectively reduces relevant differentiated parenchymal subtypes *in vivo*.

Parenchymal cell types are classified according to their marker gene expression patterns. Specifically, *PSAP* (prosaposin) marks granule-rich secretory parenchyma with high lysosomal activity that often co-expresses genes encoding cathepsin (*ctsb, ctsl*), v-ATPase subunits (proton pump that acidifies lysosomes) and trafficking factors such as *rab7* (Plass et al. 2018). In turn expression of aquaporins is characteristic for fluid-handling secretory and absorptive cell subtypes, which are positioned at tissue interfaces or alongside gland and excretory ducts, meaning the epithelial conduits that carry secretions to surface pores. The expression of *SSPO* (*dd_9342*) marks a differentiated parenchymal cell subtype, which produces a spondin-like secreted ECM glycoprotein, suggesting roles in matrix scaffolding and in supporting nearby neurites. *fer3l-1* expression labels immediate progenitors upstream of *SSPO*$^{+}$ parenchyma that expand during regeneration and supply the *SSPO*$^{+}$ *cells* (H. O. King et al. 2024). The expression of *ldlrr-1* gene indicates another parenchymal subset; its transcript redistributes after injury with puncta in the blastema, and together with related LDLRR family members and F-spondin it has been proposed to supply extracellular cues that support CNS regeneration (Roberts-Galbraith et al. 2016). Next, *PGRN-1* (progranulin) expression marks injury-responsive secretory cells, associated with wound remodeling (Emili et al. 2025). *ptf-1* and *dd_8476* identify a head-enriched parenchymal progenitor set, which gives rise to cephalic parenchymal cells (H. O. King et al. 2024). Lastly, phagocytic parenchymal population marked by *CTSL2*, termed abraçada cells, is frequently localize next to neoblasts. Abraçada cells can physically enwrap neoblasts and have been linked to regeneration through innexin-dependent regulation (Settles et al. 2025).

In practice, during the analysis of the single-cell RNA-seq data, parenchymal identity is assigned to cells that lack epidermal progenitor markers (*prog-1, agat-1, zpc/zp*), muscle structural genes (*mhcA/B, myosin*, troponins), and intestinal absorptive/digestive pathways. Simultaneously, these cells must exhibit the expression of genes encoding secretory granules, ECM/collagen modifiers and lineage-specific TFs that connect S/G_2/M neoblasts through immediate parenchymal progenitors to differentiated states (Fincher et al. 2018; Plass et al. 2018).

1.2.1.10. Pigment cells

Body pigment cells are subepithelial pigment cells distributed across most of the planarian body surface (excluding minor unpigmented regions such as the

photoreceptors). In *S. mediterranea*, these cells contribute to normal body coloration by producing an ommochrome-type pigment and, in addition, they synthesize porphyrins under physiological conditions (Stubenhaus et al. 2016). Transcript localization identified strong pigment-cell-enriched expression of early heme/porphyrin biosynthesis enzymes, including *ALAS*, *ALAD*, and *PBGD* (Stubenhaus et al. 2016). Regarding the functional implications, it was demonstrated that disruption of this pathway resulted in altered pigmentation and repigmentation associated with regeneration, and intense visible light caused loss of pigment cells through porphyrin-dependent photosensitization, establishing pigment cells as a vulnerable indicator of cell-type-specific metabolism and stress response in flatworms (Stubenhaus et al. 2016).

1.2.2. Advantages and limitations of *S. mediterranea* compared with classical model organisms

S. mediterranea offers a distinctive combination of strengths and constraints when compared with long-established laboratory models such as *Drosophila* and *C. elegans.* Its major advantage is the capacity for true whole-body regeneration (Levin et al. 2019). Because each amputated fragment initiates a regeneration program at a well-defined time point (the moment of amputation), animals can be collected at closely spaced intervals after injury (for example, every few hours) and aligned by "hours or days post amputation," which enables densely sampled, time courses of regeneration that are difficult to achieve in many other adult systems (Kao et al. 2013; Molinaro and Pearson 2016). Despite the significant evolutionary distance from vertebrates, *S. mediterranea* harbors deeply conserved signalling pathways, including canonical Wnt/β-catenin, BMP, FGF and Hedgehog, enabling direct comparison of developmental and regenerative mechanisms across Bilateria (De Robertis 2010). Functional genomics is enabled by robust, systemic RNAi, achieved by feeding the animals bacteria engineered to produce target-specific double-stranded RNA (dsRNA) or by directly injecting the dsRNA obtained by *in vitro* transcription (Adler and Alvarado 2018). Finally, as noted in **Section 1.2**, the lophotrochozoan position of *S. mediterranea* complements ecdysozoan and vertebrate models, strengthening comparative analyses.

In addition to these advantages, there are also certain characteristics of this organism that pose limitations for research. The *S. mediterranea* genome is highly polymorphic and rich in repetitive elements, including abundant LTR retrotransposons

(for example Burro elements) and long AT-rich microsatellites. Because of these features, the first viable attempts at generating a genome assembly only became possible with the advent of third-generation sequencing technologies (Grohme et al. 2018; Ivanković et al. 2024). Extensive sequence divergence and large-scale rearrangements relative to other bilaterians further complicate orthology and synteny inference when compared with comparatively compact and structurally stable genomes such as those of *Drosophila* or *C. elegans*, where conserved gene order is easier to trace (Grohme et al. 2018; Ivanković et al. 2024). Robust germline transgenesis and stable fluorescent reporter lines have not been established in *S. mediterranea* and no mutant strains are available. This is due to the fact that neoblasts continuously self-renew and reconstruct cell lineages, that *in vitro* neoblast cultures are inefficient and short-lived, and that in asexual strains, reproduction occurs by fission (Grohme et al. 2018; Ivanković et al. 2024). Consequently, functional studies rely on systemic RNAi, whole-mount *in situ* hybridization and immunostaining of fixed tissue, with more recent protocols enabling only transient somatic expression of fluorescent or luminescent reporters for short-term live imaging (R. S. King and Newmark 2013; Hall et al. 2022; Forsthoefel et al. 2014). In addition, strain divergence introduces practical complexity: the asexual and sexual karyotypes differ structurally, so read mapping, variant interpretation, and CRISPR guide design require careful reference selection.

1.3. Non-coding RNA dynamics in planarian regeneration

Planarian regeneration entails rapid injury responses, proliferation of neoblasts, and coordinated differentiation of multiple lineages. sncRNA pathways are integral to this plasticity. Deep-sequencing studies have defined the rich miRNA and piRNA repertoires involved in these processes in *S. mediterranea*. It was revealed that PIWI/piRNA systems are essential for stem-cell function and that miRNAs display expression dynamics tightly linked to specific stages of regeneration (I. V. Kim et al. 2020; Sasidharan et al. 2013; Friedländer et al. 2009; Palakodeti et al. 2008, 2006). In contrast, our understanding of the involvement of other sncRNAs in planarian regeneration, particularly RFs, is still very limited. Because such molecules modulate regenerative and developmental processes in other metazoans (S. Keam and Hutvagner 2015; Guzzi and Bellodi 2020), it is likely that they also contribute to the regeneration in planarians (Lakshmanan et al. 2021).

1.3.1. miRNAs in regeneration

Deep sequencing in *S. mediterranea* first identified ~70 miRNA genes, including conserved families (let-7, lin-4, miR-1, miR-2, miR-124) (Lu et al. 2009). Later studies roughly doubled this number and, together with more recent "small RNAome" profiling, indicated that the planarian miRNome now comprises >200–250 primary miRNA genes organized in several genomic clusters (Friedländer et al. 2009; Lu et al. 2009; Sasidharan et al. 2013; Y. Li et al. 2017).

To determine the cell-type specificity of individual miRNAs it was first necessary to isolate distinct cell fractions. Fluorescence-activated cell sorting (FACS) served as the primary method, enabling the separation of cells based on DNA content and sensitivity to ionizing radiation, thereby yielding the canonical three major populations: irradiation-sensitive X1 neoblasts, X2 progenitors, and irradiation-insensitive Xins differentiated cells. By combining FACS-based profiling with approaches such as *in situ* hybridization, researchers identified miRNAs enriched in neoblasts or progenitors as well as those characteristic of differentiated cell types. A detailed analysis of the available literature revealed distinct functional groups within the planarian miRNome. The X1/X2 fraction is dominated by miRNAs implicated in stem cell and progenitor maintenance, self-renewal, and the repression of apoptosis. A core set including let-7a/b/d, miR-2a/2d, miR-13, miR-36a/b, miR-71a/b/c, miR-190a, miR-2160, miR-752, and miR-756 is enriched in X1/X2. Notably, the polycistronic cluster containing miR-71b, miR-2d, miR-752, and miR-13 appears to act cooperatively to maintain the neoblast state, potentially by repressing pro-apoptotic factors (e.g., *hid* homologs) and differentiation drivers, analogous to the bantam pathway in *Drosophila* (Lu et al. 2009). Additionally, miR-190a and miR-125a are associated with cell cycle control within this compartment. In contrast, the Xins fraction comprises lineage-specific miRNAs that drive tissue specification. This group includes the miR-1 family members, miR-9, miR-124, miR-277, miR-750, miR-87, and others are preferentially expressed in differentiated tissues such as muscle, intestine, pharynx, and the CNS (Sasidharan et al. 2013; Lu et al. 2009). During regeneration, the expression levels of many of these miRNAs change in a stage-specific manner: several neoblast miRNAs are induced within hours after amputation, whereas others, including miR-2d and the miR-124 family, become enriched in the anterior blastema and brain primordium at 3–7 days post-amputation (dpa) (Sasidharan et al. 2013).

Despite these advances in characterizing the spatiotemporal expression of miRNAs, the mechanistic aspects of their functions remain poorly understood, and their molecular targets are only beginning to be elucidated. One of the best-characterized examples to date is the miR-124 family (Sasidharan et al. 2017). Combined knockdown of miR-124 paralogs disrupted brain and eye regeneration, reduced eye progenitors, and caused mispatterning of the visual system. miR-124 directly or indirectly down-regulated *slit-1* and components of the Notch pathway, linking neuronal differentiation to midline and axon-guidance cues (Sasidharan et al. 2017). Global profiling and network analyses further connect neoblast-enriched miRNAs (e.g.,bantam, miR-13, miR-31, miR-71, miR-190) to gene sets controlling chromatin state, cell-cycle progression, and signaling pathways required for blastema growth and patterning, although most of these interactions remain predictive rather than experimentally validated (Y. Li et al. 2017; Pawlicka et al. 2017).

Uniquely, the *S. mediterranea* genome encodes both lin-4 (the nematode-typical founder miRNA) and its vertebrate ortholog miR-125 (Sasidharan et al. 2013; Palakodeti et al. 2006). In addition, notable discrepancies exist in the enrichment profiles of two miRNAs. Firstly, miR-31a was defined as X1-enriched by Sasidharan et al. in 2013, but as Xins-enriched by Li et al. in 2017. This may result from the differing conditions under which the experiments were conducted and may reflect miR-31a upregulation during the wound response. Secondly, miR-13 was established as a core X1-enriched by multiple early studies (Lu et al. 2009; Friedländer et al. 2009; Sasidharan et al. 2013), yet classified as Xins-enriched by Li et al. 2017. This contradiction may stem from differences in normalization methods or sequencing platforms. **Table 2** summarizes the best-supported cell-type enrichments and candidate functions currently available for commonly discussed *S. mediterranea* miRNAs involved in regeneration; for many entries, functional assignments remain provisional and are based primarily on expression patterns, perturbation in global sncRNA studies, and conservation with other metazoans.

Table 2. Selected *S. mediterranea* miRNAs with reported enrichment and predicted roles in regeneration

Functions, expression and possible targets refer to roles inferred in *S. mediterranea* unless otherwise noted; all functions and possible targets in *S. mediterranea* are still predicted.

miRNA	Function	Expression	Literature	Possible targets / pathways
bantam-a	Regulation of neoblast self-renewal and tissue growth; apoptosis control; Stimulation of tissue growth and control of the cell cycle control in *Drosophila*	X1-enriched ↓ upon amputation	(Brennecke et al. 2003 (*Drosophila*); Friedländer et al. 2009; Li et al 2017)	pro-apoptotic gene *hid*, negative regulators of cell growth or cell proliferation (in *Drosophila*)
let-7a	Neoblast and progenitor maintenance	X1-enriched in Friedländer 2009 study; X1/X2-enriched in Sasidharan 2013 study	(Friedländer et al 2009; Sasidharan et al 2013)	*EAAT2* (gi|116035649), gi|84606196, developmental regulators
let-7b	Neoblast and progenitor maintenance	X1/X2-enriched	(Lu et al 2009; Sasidharan et al 2013)	–
let-7d	Neoblast maintenance; Regeneration control	X1-enriched	(Li et al 2017)	TFs including Pax6A and Runt-like 1
lin-4	Digestive tissue specification	Xins-enriched; ↑ in intestine and pharynx	(Sasidharan et al 2013; R. C. Lee et al. 1993 (*C. elegans*))	Heterochronic pathway (in *C. elegans*)
miR-124b	CNS regeneration and maintenance	Xins-enriched; ↑ in cephalic ganglion (CG); ↑ in anterior blastema	(Sasidharan et al 2013, 2017)	*slit-1*, Notch pathway
miR-124c	CNS regeneration and maintenance	Xins-enriched; ↑ in cephalic ganglion (CG); ↑ in anterior blastema	(Sasidharan et al 2013, 2017)	*slit-1*, Notch pathway
miR-125a	Neoblast and progenitor maintenance	X1/X2-enriched; ↓ upon amputation	(Li et al 2017; Sasidharan et al 2013)	–
miR-125b	Regulation of cellular differentiation	Xins-enriched	(Sasidharan et al 2013)	–
miR-13	Neoblast and progenitor maintenance; apoptosis control	X1/X2-enriched; ↑ upon amputation; part of miR-71b/miR-2d/miR-752/miR-13 polycistronic cluster	(Lu et al 2009; Sasidharan et al 2013; Li et al 2017)	Targets Lectin-like protein (in *S. mediterranea*) ; pro-apoptopic gene *hid* (in *Drosophila*),
miR-190a	Apoptosis regulation and cell cycle control	X1-enriched; ↓ upon amputation	(Li et al 2017)	ECM and TFs regulators including Zinc finger and Collagen alpha-1(IV) chain

miRNA	Function	Expression	Literature	Possible targets / pathways
miR-1a	Muscle differentiation and maintenance (by homology)	Xins-enriched;	(Sasidharan et al 2013; Safa et al. 2020)	Muscle specific genes
miR-1c	Pharyngeal muscle differentiation	Xins-enriched; ↑ in pharynx-	(Sasidharan et al 2013)	–
miR-2150	Neoblast and progenitor maintenance	X1/X2-enriched	(Sasidharan et al 2013)	–
miR-2152	Acute wound response	X1-enriched; ↑ upon amputation	(Sasidharan et al 2013)	
miR-2156a	Regulation of cellular differentiation	Xins-enriched; Moderate across tissues	(Clarke et al. 2025 (miRGeneDB))	–
miR-2156b	Regulation of cellular differentiation	Xins-enriched; Moderate across tissues	(Clarke et al. 2025 (miRGeneDB))	–
miR-2157	Prevention of premature differentiation into eye or neural lineages	Xins – enriched ↑ upon amputation	(Li et al 2017)	TFs including Pax6A and Runt-like 1
miR-2160	Neoblast maintenance	X1-enriched	(Friedländer et al 2009; Sasidharan et al 2013)	–
miR-219	Neural differentiation (by homology)	Xins – enriched; ↓ in neoblast and pharynx /; ↑ in body sections and whole animal (unsorted worm)	(miRGeneDB; (Hudish et al. 2016 (*D.rerio*))	Targets *pard3* and *prkci* mRNAs (in *D. rerio*) to promote differentiation. Linked to NMDA receptor signaling (CaMKIIγ) in humans
miR-277c	Developmental transition	Xins -enriched; ↑ upon amputation	(Sasidharan et al 2013; Protasio et al. 2017 (*Schistosoma mansoni*))	
miR-2a	Neoblast and progenitor maintenance; apoptosis control	X1/X2-enriched; ↑ in posterior blastema	(Lu et al 2009; Sasidharan et al 2013)	Apoptosis / cell-cycle (by homology)
miR-2c	Wound response	Xins-enriched; ↑ in posterior blastema	(Sasidharan et al 2013)	–
miR-2d	Neurogenesis and CNS regeneration	X1-enriched; ↑ in anterior blastema ↑ upon amputation; ↑ CG tissue; part of miR-71b/miR-2d/miR-752/miR-13 polycistronic cluster	(Sasidharan et al 2013)	
miR-31a	Wound response	X1-enriched in Sasidharan 2013 study; Xins-enriched (↑ after irradiation) in Li 2017 study; ↑ upon amputation	(Sasidharan et al 2013; Li et al 2015, 2017)	Lectin-like and secreted factors

miRNA	Function	Expression	Literature	Possible targets / pathways
miR-36a	Neoblast maintenance	X1-enriched; ↑ in anterior blastema	(Sasidharan et al 2013)	–
miR-36b	Neoblast and progenitor maintenance	X1/X2-enriched;	(Friedländer et al 2009; Sasidharan et al 2013)	–
miR-71a	Progenitor control	X2-enriched;	(Sasidharan et al 2013; De Lencastre et al. 2010 (*C. elegans*))	In *C. elegans*, miR-71 promotes longevity and stress resistance by repressing insulin signaling components (*age-1/PI3K, pdk-1*) and mediating DNA damage responses
miR-71b	Neoblast and progenitor maintenance	X1/X2-enriched; part of miR-71b/miR-2d/miR-752/miR-13 polycistronic cluster	(Sasidharan et al 2013; Lu et al 2009; Friedländer et al 2009)	It likely acts cooperatively with mir-2 to repress pro-apoptotic factors or differentiation factors, ensuring stem cell survival and proper lineage progression
miR-71c	Neoblast maintenance	X1-enriched	(Li et al 2017)	MFSDHA1 transporter family (solute carrier family 18)
miR-750	Regulation of cellular differentiation	Xins-enriched	(Sasidharan et al 2013)	
miR-752	Neoblast maintenance	X-enriched; part of miR-71b/miR-2d/miR-752/miR-13 polycistronic cluster	(Friedländer et al 2009; Sasidharan et al 2013)	It likely acts cooperatively with mir-2 to repress pro-apoptotic factors or differentiation factors, ensuring stem cell survival and proper lineage progression
miR-755	Regulation of cellular differentiation	Xins-enriched	(Sasidharan et al 2013)	–
miR-756	Neoblast maintenance	X1-enriched	(Friedländer et al 2009; Sasidharan et al 2013)	–
miR-7b	Neural-fated neoblast regulation (by homology)	X1-enriched;	(Sasidharan et al 2013; Aparicio et al. 2015 (*Drosophila*))	Notch target genes and *EGFR* antagonists (in *Drosophila*)
miR-7c	Neural differentiation	X2- enriched	(Sasidharan et al 2013; Aparicio et al. 2015 (*Drosophila*))	Notch target genes and *EGFR* antagonists (in *Drosophila*)
miR-87b	Regulation of cellular differentiation	Xins-enriched	(Sasidharan et al 2013)	–
miR-87d	Regulation of cellular differentiation	Xins- enriched	(Sasidharan et al 2013)	–

miRNA	Function	Expression	Literature	Possible targets / pathways
miR-8a	Neurogenesis and neural specification	Xins-enriched; ↑in anterior blastema	(Sasidharan et al 2013; Y. Lee et al. 2018 (*Ictidomys tridecemlineatus*))	Conserved neural candidates include *zfp-1* or cadherin homologs (in neuroblastoma cell line)
miR-8b	CNS and eye regeneration (in *D. japonica*)	Xins-enriched; ↑ in CNS- in *D. japonica*	(H. Liu et al. 2021 (*D. japonica*); Friedländer et al. 2009)	Eye and brain genes (*eye53* and *nou-darake*, validated in *D. japonica*)
miR-96a	Regulation of cellular differentiation	Xins-enriched; ↑ in posterior blastema;	(Sasidharan et al 2013)	
miR-96b	Regulation of cellular differentiation	Xins-enriched; ↑ in posterior blastema;	(Sasidharan et al 2013)	–
miR-9a	Neural progenitor control	Xins-enriched; ↑ in CNS in *Drosophila*	(Sasidharan et al 2013; Coolen et al. 2013 (*Drosophila*)))	Proneural / Notch pathway (in *Drosophila*)

1.3.2. piRNAs in regeneration

piRNAs were first characterized in *S. mediterranea* by cloning and deep sequencing of sncRNAs (Palakodeti et al. 2008; Friedländer et al. 2009). These analyses showed that piRNAs comprise both transposon-derived and mRNA-derived molecules, with abundances that differ between neoblasts, where they are strongly enriched, differentiated tissues, and regenerating fragments (Palakodeti et al. 2008; Y. Li et al. 2017; I. V. Kim et al. 2019). Key piRNA–PIWI populations and their inferred roles are summarized in **Table 3**.

Functional studies have shown that the PIWI–piRNA pathway is indispensable for regeneration. RNAi of *smedwi-2/-3* reduced piRNA levels, depleted neoblasts, abolished regenerative capacity, and caused irradiation-induced stem-cell loss. These data indicate that SMEDWI-2/3 associated with piRNAs are required for stem-cell maintenance and whole-body regeneration (Palakodeti et al. 2008; Shibata et al. 2016). Further it was demonstrated that SMEDWI-2 and SMEDWI-3 cooperated in a ping–pong cycle to degrade active transposons, while SMEDWI-3 also bound exon-derived piRNAs that either trigger mRNA decay or mediate mRNA surveillance depending on piRNA–

mRNA complementarity, thereby reshaping the neoblast transcriptome during regeneration (I. V. Kim et al. 2019, 2020).

Table 3. piRNA–PIWI populations and their inferred roles in *S. mediterranea* regeneration

piRNA / PIWI population	Enrichment	Inferred role in regeneration	Literature
Global neoblast piRNAs	X1-enriched; dynamic between neoblasts and whole body	Dominate neoblast sncRNA landscape; support stem-cell identity and regenerative competence	(Friedländer et al. 2009; Rink 2013)
SMEDWI-1–bound piRNAs	X1-enriched; ubiquitous in proliferative cells	Likely mark pluripotent neoblasts and contribute to general stem cell regulation	(Palakodeti et al. 2008; Reddien et al. 2005)
SMEDWI-2–bound piRNAs	X1-enriched; piRNAs drop after *smedwi-2* RNAi	Primary transposon silencing and maintenance of genome integrity in expanding neoblasts; required for regeneration and homeostasis	(Palakodeti et al. 2008; Friedländer et al. 2009)
SMEDWI-3–bound piRNAs (transposon-directed)	X1-enriched; show ping–pong signatures with transposon transcripts	Cytoplasmic ping–pong degradation of active transposons; cooperate with SMEDWI-2 to protect neoblast genomes during regeneration	(I. V. Kim et al. 2019)
SMEDWI-3–bound piRNAs (mRNA-directed)	X1-enriched	piRNA-guided mRNA decay and surveillance in neoblasts; regulates transcript turnover during the transition from stem cell to committed progenitor states	(I. V. Kim et al. 2019)
piRNA clusters (genomic)	X1-enriched	Provide the sequence reservoir for transposon- and mRNA-directed piRNAs; likely shaped by regenerative selection pressures	(Friedländer et al. 2009)

1.3.3. RNA fragments in regeneration

Beyond miRNAs and piRNAs, *S. mediterranea* accumulates multiple RFs, with tRFs representing the best-characterized class identified to date. The dominating subclasses are: tRF-5s, itRFs, and 5′-tiRNAs (Lakshmanan et al. 2021). Spatial profiling of tRFs across serial transverse sections of the animal showed distinct axial patterns, with subsets of molecules enriched in particular regions. Moreover time-course profiling demonstrated dynamic regulation of all three classes during both anterior and posterior regeneration. It was found that 5′-tiRNAs interacted with all three SMEDWI proteins. In contrast, itRFs displayed sequence signatures, specifically a strong 5' uridine bias, characteristic of Dicer processing and AGO1 loading, suggesting that planarian tRFs

intersect with both piRNA and miRNA-like pathways in neoblasts (Lakshmanan et al. 2021).

In other model systems, including *H. sapiens*, *M. musculus*, *D. melanogaster*, *Tetrahymena thermophila* and *Trypanosoma brucei*, tRFs have emerged as critical regulators of development and stem cell function. For instance, they are known to direct fate decisions in embryonic stem cells, and mediate intergenerational epigenetic inheritance (Ivanov et al. 2011; Krishna et al. 2021; Sharma et al. 2016). Given these conserved functions, the regeneration stage–specific and region-specific accumulation of planarian tRFs suggests analogous roles in coordinating neoblast behavior and differentiation, although this aspect remains largely unexplored (Lakshmanan et al. 2021). The main planarian tRF classes and their inferred roles in regeneration are summarized in **Table 4**.

Table 4. Main classes of tRFs in *S. mediterranea* and their inferred roles in regeneration (based on Lakshmanan et al. 2021)

tRF class	Enrichment	Protein associations	Inferred role in regeneration
tRF-5	*Temporal:* The collective pool is unchanged early post-amputation (3 h); individual tRF-5s vary over regeneration and can be categorized into early vs late waves; *Spatial:* Specific subsets show distinct spatial patterns	No strong dependence on canonical miRNA factors (unaffected by *DICER* or *AGO1/AGO2* knockdown (KD)).	Dynamic accumulation during regeneration suggests a regulatory role; putative miRNA-like mechanism; no validated targets.
itRF	*Sequence:* Strong bias for uridine at the 5' end. *Temporal:* Upregulated during early wound responses. *Spatial:* Specific subsets show distinct spatial patterns.	Likely processed by Dicer and loaded into AGO1 (twofold decrease of the itRF pool in *AGO1* KD). Strong 5'-uridine bias (signature of Dicer/AGO1 loading).	Dependence on AGO1 supports a miRNA-like function.
5′-tiRNA	*Temporal:* The collective pool doubles at 3 h post-amputation (both anterior and posterior), individual 5′-tiRNAs show dynamic changes over regeneration time points; *Spatial:* Specific subsets show distinct spatial pattern.	SMEDWI 1/2/3 (pulldown evidence)	Dynamic accumulation during regeneration suggests a regulatory role; putative translation control mechanisms.
tRF-3 / 3′ tiRNA	No data	No data	No data

1.4. Bioinformatic and genomic resources for *S. mediterranea*

Over the past decade, genomic and transcriptomic resources for *S. mediterranea* have advanced from short-read draft assemblies and *de novo* transcriptome catalogs to highly contiguous reference genomes with curated gene annotations, including chromosome-scale assemblies in which the two homologous chromosome copies in this diploid species are reconstructed separately (Resch and Palakodeti 2012; Plass et al. 2018; Fincher et al. 2018; Cui et al. 2023). In parallel, community datasets now include broad bulk mRNA-seq and sncRNA-seq resources, as well as widely used single-cell and spatial atlases that map gene expression to cell states and anatomical context during regeneration (Y. Li et al. 2017; Fincher et al. 2018; Plass et al. 2018; Cui et al. 2023). These developments enable integrative analyses in which multiple molecular layers (genome, transcriptome, sncRNAs, and cell-state-resolved atlases) are interpreted jointly, rather than treating each layer in isolation.

1.4.1. Genome assemblies

Genome assemblies are commonly described at different levels of contiguity: contigs are continuous reconstructed DNA sequences without gaps, scaffolds link contigs into larger sequences using long-range information (typically leaving gaps of unknown sequence), and chromosome-level assemblies further order and orient scaffolds into full chromosomes (Dyer et al. 2025). For many years, assembling the *S. mediterranea* genome remained technically challenging, and available resources were dominated by fragmented genomic sequences and *de novo* transcriptome catalogs. A major breakthrough came with the dd_Smes_g4 assembly (Grohme et al. 2018), which used long-read sequencing data and produced a highly contiguous, scaffold-level reference. This assembly revealed extensive heterozygosity – high sequence divergence between the two homologous chromosome copies in this diploid species. This feature complicates parental chromosome copy (haplotype) reconstruction and can also impair short-read mapping by increasing multi-mapping and mismatch rates in highly polymorphic regions. High heterozygosity and abundant repeats also increase multi-mapping, in which a sequencing read aligns equally well to multiple genomic loci. This is particularly consequential for sncRNAs reads because they are short, and for analyses focusing on tRNA or tRFs, because tRNA genes form large, highly similar families. To assess assembly completeness, Grohme et al. back-mapped two widely used de novo transcriptome

resources, dd_Smes_v1 (sexual strain) and dd_Smed_v6 (asexual strain), to dd_Smes_g4. Only a small fraction of transcripts failed to map under their criteria, and further inspection suggested that very few of these transcripts represented genuinely missing genes (Grohme et al. 2018). Building on dd_Smes_g4, subsequent work (L. Guo et al. 2022) produced a chromosome-scale scaffolding (schMedS2) and then a chromosome-scale, haplotype-phased assembly (S3h1 and S3h2) for the sexual S2 strain, together with improved gene models and regulatory annotation supported by Assay for Transposase-Accessible Chromatin using Sequencing (ATAC-seq) and Chromatin Immunoprecipitation followed by Sequencing (ChIP-seq) (Ivanković et al. 2024). The same work generated in 2024 chromosome-level assemblies for close relatives (*S. polychroa*, *S. nova*, *S. lugubris*), enabling comparative and regulatory conservation analyses (Ivanković et al. 2024). For the asexual CIW4 strain, historically the community relied on *de novo* transcriptomes and contig-level resources; high-quality phased assemblies for CIW4 are now emerging but not yet on the same level as the S2 resource. Accordingly, many studies analyze CIW4 transcriptome by mapping reads to the S2 reference and projecting gene models and other genomic features between references using whole-genome alignments. As of October 2025, a bioRxiv preprint has reported a high-quality, haplotype-phased asexual genome ("schMedA2") which helps reduce the longstanding disparity in assembly quality between sexual and asexual resources and supports biotype comparisons, structural variant analyses (e.g., inversions, translocations, and copy-number changes), and the integration of allele-resolved sequence variation with gene expression (Brand et al. 2025).

1.4.2. Transcript and gene annotations

Several annotation resources are still used in *S. mediterranea* studies as they were created at different stages of genome exploration. Early resources include *de novo* transcriptomes for asexual (dd_Smed_v6) and sexual (dd_Smes_v1) strains, *ab initio* gene predictions for sexual strains dd_Smes_g4 (SMESG), and a set that merges evidence from multiple sources (Oxford_v1) (Rozanski et al. 2019; Grohme et al. 2018; Ivanković et al. 2024). The newest S3 annotations for sexual strain were built on the chromosome-scale, phased sexual reference using hybrid evidence (transcript data, homology, and prediction). They generally improved completeness and reduced fragmentation relative to older sets, which in practice yielded more reliable read mapping, transcript boundaries

(including UTRs), and cross-strain comparability (Ivanković et al. 2024). Importantly, these resources still have documented gaps: the S3 genome is not yet fully finished, with hundreds of unscaffolded contigs remaining (containing > 500 annotated genes), and the gene annotations still retain some errors, such as chimeric, truncated, or frame-shifted transcripts (Ivanković et al. 2024). For the asexual strain, as of December 16, 2025, the newest schMedA2/CIW4 assembly (Brand et al. 2025) and its accompanying gene annotation are not yet retrievable from a public repository (e.g., NCBI Assembly/GenBank), so routine CIW4 analyses still typically rely on the publicly downloadable sexual S2 reference.

1.4.3. Public bulk and sncRNA sequencing datasets

An extensive collection of public bulk RNA-seq datasets is currently available, originating from experiments that investigate key aspects of planarian biology. They encompass a range of studies, including time-course analyses of regeneration, irradiation-induced depletion of neoblasts, and the functional characterization of specific genes. These datasets are accessible through community resources such as PlanMine and PLANAtools (Hoffman and Wurtzel 2023; Rozanski et al. 2019). Notably, they provide more than raw expression profiles: they supply well-annotated sample metadata, and reference expression patterns that can be reused across studies.

sncRNA resources have been developed in a similar, cumulative manner. The most informative datasets originated from studies explicitly designed to investigate sncRNA populations. More details on this were presented in the previous chapter, in **Section 1.3**. However, because sncRNA profiling is not systematically incorporated as a standard component of transcriptional analyses, comprehensive sncRNA datasets remain limited. Consistent with this, the NCBI Taxonomy Browser for *S. mediterranea* (Taxonomy ID 79327) reports 6,033 SRA records, including 5,066 RNA libraries (accessed December 17, 2025), but only 67 correspond to dedicated sncRNA strategies. This limitation constrains integrative analyses of gene regulatory mechanisms, particularly those involving sncRNA-mediated control of gene expression.

1.4.4. Single-cell and spatial atlases

Since the advent of scRNA-seq, this technology has been extensively applied to the study of planarian biology, including tissue homeostasis, regeneration, cell-state

transitions. Two landmark single-cell atlases (Drop-seq / Smart-seq–based) cover essentially all major *S. mediterranea* cell types and many subtypes (Plass et al. 2018; Fincher et al. 2018). These atlases are routinely used to transfer cell-type annotations between datasets, infer differentiation trajectories, and project bulk differential expression data onto specific cell types to support biological interpretation (Plass et al. 2018; Fincher et al. 2018).

Further technological advances have facilitated the development of a spatiotemporal transcriptomic atlas of planarian regeneration, which integrates single-cell gene expression data with tissue architecture. It provides three-dimensional anatomical context and quantitative maps of spatially patterned gene expression and cell-types (Cui et al. 2023).

1.4.5. Databases and portals

A range of databases and online tools exist, including resources specifically focused on planarians, as well as general platforms used in broader biological research. PlanMine 3.0 is the primary integrative portal, offering genome browsers and aggregating *S. mediterranea* genomes, gene models, transcriptomes and expression tracks (Rozanski et al. 2019). Planosphere/PLANA provides the Planarian Anatomy Ontology and a curated expression knowledge base that standardize anatomical terms and enable cross-dataset interoperability (Nowotarski et al. 2021). PLANAtools offers an interactive compendium for quickly accessing published planarian RNA-seq studies, including injury time courses and RNA interference (RNAi) (Hoffman and Wurtzel 2023). An additional resource is SmedGD, which references older literature and contains early gene identifiers. It remains a historical reference point, enabling the reconciliation of these original identifiers with current genomic and transcriptomic data (Robb et al. 2008). In line with standard practices for other organisms, raw and processed sequencing data from *S. mediterranea* are routinely deposited in NCBI GEO (expression matrices) and SRA (raw reads) (Barrett 2004; Kodama et al. 2012). The European Nucleotide Archive (ENA) mirrors SRA accessions and provides stable, programmatically accessible entry pages (Yuan et al. 2024). Some historical submissions also appear under ArrayExpress, now consolidated within BioStudies/Expression Atlas (Madrigal et al. 2026).

Specialized resources for planarian sncRNAs, such as dedicated databases or web portals, are currently lacking. Thus, researchers must rely on RNAcentral

(https://rnacentral.org/), the main repository for such information, which is a comprehensive database compiling ncRNA sequences from a broad spectrum of species and can be used to generate a project-specific ncRNA catalogue (Sweeney et al. 2021). In addition, miRBase (https://www.mirbase.org/) remains the canonical source for mature miRNA sequences (Kozomara et al. 2019).

1.4.6. Limitations of existing bioinformatics resources

Despite substantial progress, several challenges remain directly relevant to both sncRNA–focused and integrative analyses in *S. mediterranea*. At the level of reference sequences, the combination of a repeat-rich, polymorphic genome and the short length of sncRNA reads increases ambiguity in alignment, particularly for multicopy loci and highly similar gene families. This complicates locus-resolved quantification and interpretation even when high-contiguity references are used (Grohme et al. 2018; Ivanković et al. 2024). In practice, these constraints make analysis choices about multi-mapping handling, reference selection, and coordinate consistency disproportionately consequential for sncRNAs relative to bulk mRNA profiling. At the annotation level, improvements in chromosome-scale assemblies have reduced fragmentation and strengthened cross-study comparability. However, transcript structures (including exon–intron structure and UTR boundaries), still vary across annotation sets and biotypes, which introduces uncertainty when interpreting expression changes or when projecting features across references and biotypes (Ivanković et al. 2024). These limitations negatively impact sncRNA studies because mechanistic interpretation often relies on precise transcript definitions (e.g., UTR-localized miRNA target sites) and on the ability to unambiguously connect short-read signal to a specific genomic origin (Ivanković et al. 2024).

For sncRNA-specific resources, standardized, high-confidence annotations are uneven across portals and studies, and tRNA gene annotation has historically been a key bottleneck for systematic tRF analyses (Lakshmanan et al. 2021). Dedicated planarian tRNA annotation enabled structured classification and quantification of tRF classes, but tRNA family redundancy still requires careful mapping practices to avoid overconfident locus assignments. More broadly, as mentioned earlier, the public sncRNA-seq data landscape remains smaller and more study-specific than bulk RNA-seq, because many projects prioritize mRNA profiling and only a subset are explicitly designed to

characterize sncRNA populations or particular sncRNA classes (Friedländer et al. 2009; Palakodeti et al. 2006).

Finally, while single-cell and spatial atlases provide a powerful framework for assigning gene expression to cell types and anatomical context (Fincher et al. 2018; Plass et al. 2018; Cui et al. 2023), they primarily measure polyadenylated mRNAs and therefore do not directly profile most sncRNAs. As a result, high-throughput cell-resolved sncRNA analyses are currently constrained to indirect strategies as mentioned in **Section 1.1.3**.

Addressing these challenges will directly improve sensitivity and specificity for sncRNA discovery and interpretation.

1.5. *ELAC2* knockdown as a strategy to investigate the effects of sncRNA pool alterations on regeneration process in *S. mediterranea*

As outlined in the preceding sections, our understanding of the role of sncRNAs in planarian regeneration is largely limited to information on miRNAs and piRNAs. Moreover, despite the availability of relatively comprehensive catalogs of miRNAs specific to neoblasts, progenitors and differentiated cells, little is known about the precise molecular mechanisms through which individual molecules exert their functions. In addition, the role of other sncRNAs, particularly RFs, in regeneration remains poorly understood, although studies in other organisms have demonstrated that such molecules regulate key stem cell-related and developmental processes, as mentioned in **Section 1.3.3** (Ivanov et al. 2011; Krishna et al. 2021; Sharma et al. 2016; Krishna et al. 2019). At the Laboratory of Single Cell Analyses, IBCH PAS, we therefore initiated research aimed at characterizing the sncRNAs involved in *S. mediterranea* regeneration. The first stage of this work, carried out by Dr. Annasha Dutta as part of her doctoral studies, focused on developing a strategy to identify these molecules. The approach involved disrupting the sncRNA pool by silencing an RNase and assessing its impact on regeneration through initial phenotypic assessment. Subsequently, comparative analysis of sncRNAs from both RNase knockdown and wild-type worms was intended to identify the sncRNAs implicated in regeneration. The selection of the ribonuclease was guided by three criteria: (i) its human homolog should be associated with Mendelian disease, (ii) it should be directly involved in sncRNA metabolism, and (iii) it should be upregulated in neoblasts. Based on these considerations, a set of candidate ribonucleases was selected for RNAi screening experiments aimed at identifying the enzyme whose knockdown

disrupts regeneration in *S. mediterranea* (Dutta 2025). As a result, the ribonuclease ELAC2 was chosen as the target for this research, which also comprises the studies documented in the present dissertation.

RNase Z catalyzes endonucleolytic removal of 3′ trailers from pre-tRNAs, generating substrates for CCA addition and aminoacylation (Maraia and Lamichhane 2011). In humans, 2 paralogous enzymes exist: a short, primarily cytosolic RNase Z^S (*ELAC1*) and the long form RNase Z^L encoded by *ELAC2*. The latter functions in the nuclear and cytosolic compartment and executes essential 3′ processing for both mitochondrial and nuclear pre-tRNAs (Rossmanith 2011; Xue et al. 2024). Perturbing *ELAC2* shifts the balance between precursor and mature tRNAs and can indirectly modulate the availability and cleavage susceptibility of specific tRNAs into tRFs (Haack et al. 2013; Brzezniak et al. 2011; Y. S. Lee et al. 2009; Siira et al. 2018; Xue et al. 2024; Choi et al. 2021). Notably, the pre-tRNA-derived tRF-1 species tRF-1001 is ELAC2-dependent (Y. S. Lee et al. 2009; Choi et al. 2021).

Biallelic pathogenic variants in human *ELAC2* cause combined oxidative phosphorylation (OXPHOS) deficiency, a disorder characterized by cardiomyopathy and neurological manifestations, with the cardiac phenotype being the most extensively studied. Integration of evidence from patient-derived studies and findings from model organisms reveals a conserved molecular background of the cardiac muscle abnormalities (Dutta et al. 2025). Loss of *ELAC2* function causes defective mitochondrial pre-tRNA 3′ processing, which leads to accumulation of unprocessed mt-pre-tRNAs, impaired mitochondrial translation, and downstream respiratory-chain dysfunction (Haack et al. 2013; Saoura et al. 2019; Siira et al. 2018) . This mechanistic cascade is schematized in **Figure 7** (Dutta et al. 2025).

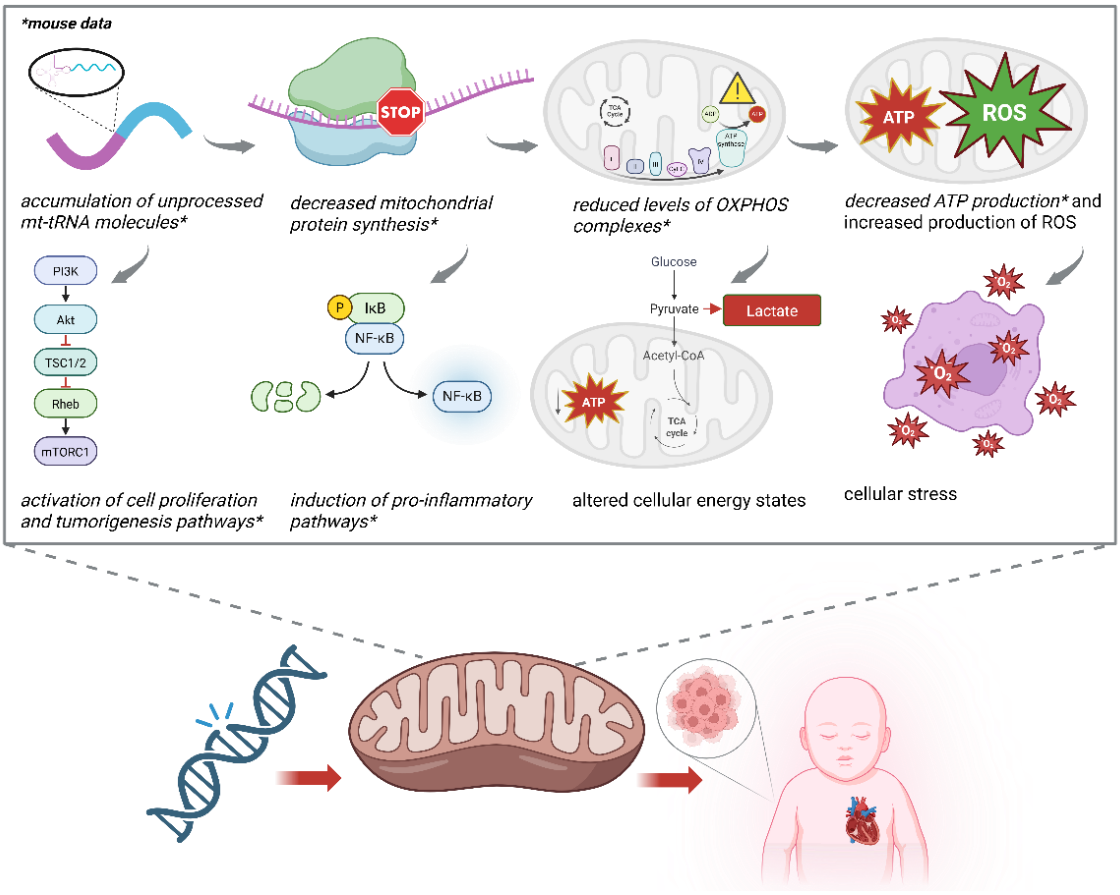

Figure 7. ***ELAC2*** **loss: from mitochondrial pre-tRNA processing defects to pediatric cardiomyopathy** (adapted from Dutta et al. 2025). Pathogenic *ELAC2* variants impair mitochondrial pre-tRNA 3′ end processing, which leads to accumulation of unprocessed RNA precursors, reduced mitochondrial translation and OXPHOS deficiency. The resulting bioenergetic stress and redox imbalance activate compensatory metabolic and stress-response pathways, with the most severe functional consequences in high-energy-demand tissues, especially cardiomyocytes, consistent with infantile hypertrophic cardiomyopathy observed in *ELAC2*-associated disease.

In model invertebrates, ELAC2 is a conserved tRNA maturation enzyme: in *Drosophila*, the RNaseZ (JhI-1) homolog processes 3′ ends of nuclear and mitochondrial pre-tRNAs *in vivo*, and its depletion leads to the accumulation of unprocessed tRNA precursors (Dubrovsky 2004). In *C. elegans*, the ELAC2 homolog HOE-1 (Held et al. 2022) regulates the mitochondrial unfolded protein response (UPRmt) through a tRNA-dependent mechanism. In *S. mediterranea* the functional role of ELAC2 had not been characterized prior to this study.

2. AIMS AND OBJECTIVES

Regulatory sncRNAs have been extensively investigated with respect to their biogenesis and functions; however, many fundamental questions remain, particularly regarding the RFs. Regeneration represents one of the most important processes in which sncRNAs are involved across multiple organisms. While bioinformatic approaches for sncRNA analysis are well established for classical model systems such as human, mouse, and *Drosophila*, extending these analyses across diverse phyla is essential for a comprehensive understanding of sncRNA function and evolutionary context. Consequently, there is a clear need for the development of computational tools that enable systematic sncRNA profiling and functional analysis beyond traditional model organisms. In our research group, a strategy was developed to investigate the impact of alterations in the sncRNA pool on regeneration in *S. mediterranea* through the silencing of *Smed ELAC2*. This experimental framework was adopted for the development and application of bioinformatics tools dedicated to sncRNA analysis.

The aim of this dissertation was to develop and optimize bioinformatic approaches for the analysis of sncRNAs in *S. mediterranea*, and to employ these approaches to elucidate the role of selected sncRNAs in regeneration. Achieving this goal required the completion of the following specific tasks:

1. Development of a comprehensive genome annotation tool for analyzing coding and ncRNAs and its application to the *S. mediterranea* genome;
2. Application of the bulk RNA-seq approach to analyze the impact of *Smed ELAC2* silencing on sncRNA pool and gene expression;
3. *In silico* functional analysis of selected ncRNAs;
4. Optimization of sncRNA-seq bioinformatic tools and their application for the analysis of the impact of *Smed ELAC2* silencing on cell lineages.

3. MATERIAL AND METHODS

3.1. Study design and datasets

A multi-layer design was used to characterize the sncRNAs involved in *S. mediterranea* regeneration. Most of the experimental work of this project was carried out by Dr. Annasha Dutta as part of her doctoral research (Dutta, 2025), with contributions from Dr. Anna Samelak-Czajka (preparation of sequencing libraries) and Magdalena Trybus, MSc (cell sorting). All computational analyses were performed by the author of this dissertation, except for the primary analysis of the scRNA-seq data, which was conducted by Dr. Małgorzata Marszałek-Zeńczak.

All analyses were carried out using *S. mediterranea* CIW4 strain. Three experimental conditions were analyzed throughout the study: *Smed ELAC2* knockdown (KD), GFP mock, and wild type animals (WT), unless specified otherwise.

Datasets were generated for distinct regeneration stages depending on the experiment type. Samples for bulk long RNA-seq (standard bulk RNA-Seq/ gene expression profiling) and standard bulk sncRNA-seq were collected at 3 and 5 days post-amputation (dpa). Throughout this dissertation, the term long RNA-seq refers to sequencing of the bulk long RNA fraction (>200 nt), analyzed alongside the short RNA fraction (<80 nt), and it is distinct from PacBio Iso-seq long-read transcript sequencing. scRNA-seq gene expression (3’ GEX) libraries were generated from animals at 0, 16, 24, and 72 hours post-amputation (hpa). Samples for bulk sncRNA-seq from sorted cell populations were collected at 3 dpa from neoblast-enriched and non-neoblast fractions. A single composite sample for PacBio Iso-seq included only WT animals. It was a pooled material collected at 0, 1, 3, 6, 9, 12, 24, and 48 hpa and from non-regenerating animals.

Dataset-level metadata, including NCBI accessions, where available, treatment and extraction protocols, read configuration, and platform details, are provided in **Supplementary Table S1**.

3.2. Reference genome and ncRNA resources

All analyses were performed using the sexual strain SchMedS2 (L. Guo et al. 2022) nuclear reference genome and the circular 17.2 kb mitochondrial genome (E. Ross et al. 2016). The annotation schMed_Pn, adapted to the asexual strain, comprising 25,264 genes and 41,396 transcripts was generated as part of this dissertation (described in

Section 4.1). The non-coding RNA catalogue was created by combining data retrieved from RNAcentral and 457 planarian tRNA sequences in fasta format (Lakshmanan et al. 2021).

3.3. Gene model construction, functional annotation, and curation

A detailed description of the genome annotation strategy, including RNA-seq pre-processing, alignment, transcriptome assembly, artifact detection, functional annotation, and manual refinement, was provided in Zaremba et al. 2025, which has been included as part of this dissertation. Genome annotation SchMed_Pn generated as part of this dissertation was produced with a SmedAnno pipeline, consisting of an automated, containerized assembly and annotation module (Part I), followed by manual inspection and refinement (Part II). Briefly, short-read RNA-seq data were aligned with STAR (v2.7.11a) (Dobin et al. 2013), whereas hybrid assemblies that integrated short- and long-read evidence were generated using a mixed-read strategy. Transcript models were assembled with StringTie2 (Shumate et al. 2022), using v2.1.1 for short-read assemblies and v2.2.1 for mixed-read assemblies, and were merged across samples to obtain unified reference-guided and *de novo* transcript sets. Predicted transcriptomes were then structurally compared to available reference annotations using gffcompare, and annotation syntax and feature consistency were standardized using AGAT. Open reading frames (ORFs) were inferred using TransDecoder (Haas et al. 2013), homology assignments were generated using BLASTp and BLASTx (Altschul et al. 1990) searches against SwissProt database (Bateman et al. 2023), and conserved domains were identified using hmmscan across the PFAM (Mistry et al. 2021) and InterProScan (Quevillon et al. 2005) databases. Quality control and benchmarking were performed using spike-ins SIRV (Paul et al. 2016), BUSCO v5.7.1 (Manni et al. 2021), rnaQUAST v2.3.0 (Bushmanova et al. 2016), and Transrate v1.0.3 (Smith-Unna et al. 2016). Automated filters and flags were applied to identify likely technical artifacts, including models with unrealistically long exons or genomic spans, reversed duplicates, and loci suggestive of fragmentation or chimerism. Flagged loci were subsequently inspected during the manual refinement stage, and final gene models were deduplicated and cleaned to minimize propagation of assembly artifacts into downstream analyses.

3.4. Bulk long RNA-seq data analysis (gene expression profiling)

Quality control of raw FASTQ files was performed using standard read-level QC reporting (FastQC aggregated with MultiQC (Ewels et al. 2016)). Adapter sequences and low-quality bases with Phred score < 30 were removed using Trim Galore (Krueger et al. 2023), using paired-end mode for long RNA-seq reads; this trimming procedure removed low-quality terminal segments rather than discarding reads solely due to isolated low-quality positions. Trimmed reads were mapped to nuclear and mitochondrial references using STAR v2.7.9a with default parameters. Reads were quantified using StringTie2 against custom asexual strain annotation (**Section 3.3**). Differential expression was performed using DESeq2, using the default size-factor normalization based on the median-of-ratios method, followed by dispersion estimation and negative binomial generalized linear modeling with Wald tests for pairwise comparisons. Benjamini–Hochberg correction was applied to control the false discovery rate across genes. Three pairwise comparisons were evaluated: (A) *Smed ELAC2* KD vs WT, (B) *Smed ELAC2* KD vs GFP and (C) GFP vs WT. A gene was considered differentially regulated in response to *Smed ELAC2* KD if it met the following criteria: (i) significantly altered in *Smed ELAC2* KD vs WT, (ii) unchanged between the 2 control groups ($(A \cap \neg C) \cup (A \cap B \cap C)$). Statistical significance was set at an adjusted $p \leq 0.05$ with $|\log_2 \text{fold-change}| \geq 0.57$ (equal to 1.5-fold change). Gene ontology (GO) enrichment analysis was performed using topGO and both the “elim” and “classic” algorithms on GO terms obtained from PlanMine and homology-derived annotations (detailed in Results, **Section 4.2.6**). Significant GO terms ($p \leq 0.05$) were grouped into custom functional categories to facilitate interpretation.

3.5. Standard sncRNA-seq data analysis

Quality control of raw FASTQ files was performed using MultiQC reporting. Adapter sequences were removed with Cutadapt, and trimming was performed using Trim-Galore. Reads with any Phred score < 30 and very short inserts were removed (minimum insert length ≥15 nt per analysis). An iterative mapping strategy was employed. Trimmed reads were first mapped using Bowtie2 in sensitive local mode (Langmead and Salzberg 2012) against the deduplicated ncRNA catalog (RNAcentral sequences supplemented with additional planarian tRNAs (Lakshmanan et al. 2021) and piRNAs (Palakodeti et al. 2008)). Reads that did not align at this stage were subsequently mapped to the transcriptome, and any remaining reads were finally mapped to the

reference genome. Alignments exceeding 3 mismatches were discarded. tRNAs without amino acid/isoacceptor annotations were assigned using tRNAscan-SE v2.0 (Chan et al. 2021). Following alignment, ncRNA class assignment and multimapper resolution were performed as described in following **Section 3.6**, and tRF-to-parent tRNA assignment was carried out as described in **Section 3.7**. These post-mapping steps were completed prior to generation of count matrices for quantification. Read summarization, multimapping handling, and visualization were performed then in R using custom script (**Supplementary Script 1**). Differential sncRNA accumulation was tested with DESeq2 using median-of-ratios size-factor normalization, dispersion estimation, and negative binomial generalized linear modeling with Wald tests, followed by Benjamini–Hochberg correction. Three pairwise comparisons were evaluated: (A) *Smed ELAC2* KD versus WT, (B) *Smed ELAC2* KD versus GFP mock, and (C) GFP mock versus WT. sncRNAs were retained for testing if they met a minimal abundance requirement (CPM >10 in at least 2 replicates). A sncRNA was considered differentially regulated in response to *Smed ELAC2* KD if it met the following criteria: (i) significantly altered in *Smed ELAC2* KD vs WT, (ii) unchanged between the 2 control groups ($(A \cap \neg C) \cup (A \cap B \cap C)$), using the same specificity logic as applied for long RNA-seq (**Section 3.4**). Statistical significance was set at an adjusted $p \leq 0.05$ with $|\log_2 \text{fold-change}| \geq 2$.

3.6. sncRNA class assignment, multimapper handling and tRNA subclassification

A dedicated classification and multimapper-resolution steps were applied to produce consistent class-level summaries and to generate count tables suitable for differential analysis. Reference names retrieved from RNAcentral database were parsed to recover transcript identifiers and genomic coordinates. A deterministic classifier assigned each alignment to major ncRNA classes (miRNA, tRFs/tiRNA, piRNA, snoRF, snRF, SRP fragments, rRF, lncRNA fragments, mRNA fragments and “other”). When a read mapped to multiple candidates, ambiguity was resolved in 2 steps. First, a per-read tie-breaker was applied that favored (i) the highest fraction of the read aligned (alignment coverage), (ii) longer inserts (when competing alignments differed in effective aligned length), and (iii) a predefined class-priority order to prevent inflation of classes that frequently receive spurious alignments from abundant fragments. Second, when ambiguity remained after deterministic filtering (for example, multiple equally good loci

within a class), read counts were apportioned across the remaining candidates using an expectation–maximization procedure operating on sequence-level equivalence classes. In this step, each unique read sequence was treated as an observation, and its count was distributed across its indistinguishable target set in proportion to the current abundance estimates, iterating until convergence.

3.7. tRF read assignment to parent tRNA

An anticodon tag was extracted from every reference tRNA entry when available. For tRNA records lacking anticodon annotation, tRNAscan-SE was applied to infer anticodon and isoacceptor labels. In unresolved cases, UFold structural folding was used to support anticodon-loop identification. All reads aligning to one or more tRNA loci were grouped by sequence. For each sequence, the set of compatible parent tRNAs was summarized at two levels: (i) isoacceptor (amino acid; e.g., Gly, Asp) and (ii) anticodon (e.g., GCC, TCC). On this basis, reads were classified as: (i) unique anticodon, mapping unambiguously to a single isoacceptor and a single anticodon (e.g., Gly-GCC only), (ii) unique isoacceptor (mapping to multiple anticodons but all corresponding to the same isoacceptor (e.g., Gly-GCC and Gly-TCC)), or (iii) multi-isoacceptors (mapping to tRNAs corresponding to different isoacceptors (e.g., Gly-GCC and Asp-GTC)).

Counts were generated in two complementary forms. For anticodon-level quantification, only unique-anticodon reads were retained and counted per locus (weight = 1). For isoacceptor-level quantification, unique-isoacceptor reads were assigned full weight to the corresponding isoacceptor (weight = 1), whereas for multi-isoacceptor reads, the count was divided equally among all candidate isoacceptors (weight = 1 / n, where n is the number of candidate isoacceptors). The weighted counts were then summed for each sample and anticodon combination to produce the isoacceptor-level count matrix. Then best anticodon and isoacceptor were selected for each read separately.

3.8. *In-silico* target modeling for the 5′ tiRNA-Gly-GCC

3.8.1. Smed ELAC2 guide RNA target modeling

To identify candidate target sites for 5′ tiRNA-Gly-GCC under an ELAC2 sgRNA-cleavage hypothesis, a TRUE-like bipartite pairing strategy was applied to evaluate putative interactions between the sncRNA and target transcripts. In this pathway,

known as TRUE silencing, ELAC2 cleaves target RNAs, guided by complementary 5′ tRFs or structurally related RNAs (Elbarbary, Takaku, Uchiumi, et al. 2009; Nashimoto 2022). A guide RNA forms two separated complementary segments with the target RNA, and the duplex adopts a tRNA-like secondary structure. The scan was applied to the full transcriptome (41,396 transcripts) and searched for composite motifs consisting of (i) a 7-nt region complementary to the acceptor-stem segment of the 5′ tiRNA, (ii) an intervening spacer of 18–30 nt, and (iii) a 3–5 nt region complementary to the anticodon-stem segment, thereby enforcing a bipartite interaction pattern rather than a single continuous antisense match. Motif candidates were generated using sequence complementarity and spacer-length constraints (without thermodynamic scoring) and were then evaluated for structural feasibility by folding the implied tiRNA–target duplex with UFold v1.3 (L. Fu et al. 2022). Candidate transcripts were required to show differential expression inversely correlated with sncRNA levels, such that decreased abundance of 5′ tiRNA-Gly-GCC corresponded to increased target transcript levels.

3.8.2. miRNA-like target modeling

Putative miRNA-like interactions were predicted by scanning transcript sequences for complementarity to 5′ tiRNA-Gly-GCC under pairing regime consistent with miRNA-like recognition. Candidate sites were ranked using a seed-weighted scoring scheme in which guide positions 2–8 were prioritized, while additional pairing outside the seed was treated as supportive but not strictly required. For each candidate site, a duplex representation was generated, and base pairs were classified as Watson–Crick (A–U, G–C), G–U wobble, or mismatch/gap to support downstream filtering (**Supplementary Script 2**). Candidate target transcripts were required to satisfy stringent pairing criteria in the seed and to show differential expression inversely correlated with sncRNA levels, such that decreased abundance of 5′ tiRNA-Gly-GCC corresponded to increased target transcript levels.

3.8.3. piRNA-like target modeling

Putative piRNA-like interactions of 5’tiRNA-Gly-GCC were predicted by integrating three orthogonal evidence layers: SMEDWI immunoprecipitation (IP) enrichment, direct RNA–RNA contact (CLASH), and cleavage signatures of the targets,

supported by degradome data. Public SMEDWI datasets were analyzed (I. V. Kim et al. 2019).

SMEDWI association (IP sncRNA-seq). The exact 32 nt 5′ tiRNA-Gly-GCC sequence was quantified in SMEDWI-1, SMEDWI-2, and SMEDWI-3 IP libraries using normalized abundance (CPM) to compare relative enrichment of this sncRNA across paralogs.

Direct RNA–RNA contacts (SMEDWI-3 CLASH). CLASH data were used to extract ligation-captured chimeric reads containing the 5′ tiRNA-Gly-GCC sequence. In the source workflow, chimera formation was enabled by end-repair steps followed by ligation, producing reads in which the guide and target fragments were covalently joined. Target fragments were mapped to the transcriptome to assign transcript identities and target coordinates. For each chimera, a local alignment between the guide and the reverse-complement target segment was computed to obtain a base-resolved duplex. Duplexes were annotated by pairing class (Watson–Crick, wobble, mismatch, gap), seed complementarity, and pairing across the cleavage-critical region. Seed pairing was evaluated at guide nucleotides 2–8, consistent with the mandatory seed match criterion used in the source analysis, while cleavage-competent pairing was evaluated specifically at guide nucleotides 10 and 11 because continuous base pairing across the scissile phosphate between target nucleotides aligned to positions 10 and 11 was a key determinant of endonucleolytic cleavage competence in Argonaute-family proteins, including PIWI proteins (I. V. Kim et al. 2019).

Predicted cut coordinate. A predicted cut position was defined in transcript as the target nucleotide aligned to guide position 11 (t11), corresponding to the expected 5′ end generated by cleavage between guide nucleotides 10 and 11.

Degradome register profiling. Degradome libraries in the source dataset were prepared by 5′-monophosphate-dependent cloning of mRNA fragments. Degradome data were mapped to planarian transcriptome, then 5′ ends were extracted from BAM alignments as strand-aware 5′ reference coordinates and were counted per transcript position to define putative cleavage sites. For each predicted cut coordinate (t11), a symmetric window around the site was defined from −K to +K nucleotides, where K denoted the flank size used for local background estimation. Fold enrichment profiles were summarized across sites (mean ± SE) and stratified based on pairing at guide positions 10 and 11. The first group comprised chimeras exhibiting base pairing at both

guide positions 10 and 11, whereas the second group included chimeras lacking base pairing at either position 10 or 11.

3.9. Single-cell RNA-seq analysis

3.9.1. Data preprocessing

Pre-processing of the scRNA-seq datasets comprised read mapping and UMI counting, cell- and feature-level quality control, mitochondrial-content filtering, doublet detection and removal, construction of 4 time point-specific Seurat (Seurat_5.3.0) objects (0, 16, 24, 72 hpa) each containing all conditions (WT, GFP mock, *Smed ELAC2* KD), and initial clustering. These pre-processing steps were performed by Dr Małgorzata Marszałek-Zeńczak. All scRNA-seq analyses reported in **Section 4.4** start from these pre-processed Seurat objects.

3.9.2. Assessment of the need for scRNA-seq data integration

Local mixing of labels was evaluated with the Local Inverse Simpson's Index (LISI) computed on k-nearest-neighbor (kNN) neighborhoods in the PCA embedding (Korsunsky et al. 2019) (**Supplementary Script 3**). LISI provides biologically interpretable information about whether cells that should represent the same basic state are locally mixed under different conditions, while maintaining distinct cell identities. For each cell, the labels carried by its nearest neighbors were summarized as an effective number of labels. Values close to 1 indicate that a neighborhood is dominated by a single label (for example, neighbors are almost all from one condition, or almost all from one cell type), whereas higher values indicate that multiple labels are present at comparable frequencies (for example, neighbors include cells from several conditions, or several cell types). Two LISI variants were computed. For integration LISI (iLISI), the label was the sample condition grouping used in the comparison, and higher values indicated stronger local mixing across those condition labels. For cell-type LISI (cLISI), the label was the annotated cell identity, and lower values indicated that neighbors predominantly shared the same cell type. For cell identity, a purity-oriented transform was applied to cLISI so that higher values corresponded to purer cell-type neighborhoods and values closer to 0 corresponded to greater mixing of cell identities within neighborhoods. LISI distributions

were summarized by robust statistics (median and selected quantiles) to capture both central tendency and spread.

The k-nearest neighbor batch effect test (kBET) was applied using the same PCA embedding to test whether local neighborhood composition deviated from global condition proportions (Büttner et al. 2019). For each evaluated neighborhood, a statistical test was performed under the null hypothesis that labels were locally well mixed given their global frequencies, and the mean rejection rate across tested neighborhoods was reported. Lower kBET rejection rates were interpreted as weaker condition-driven separation in local neighborhoods. All metrics were computed with fixed parameters across representations to enable consistent comparison, and the results were visualized as point estimates with interval summaries to facilitate interpretation across the 3 dataset-combination strategies.

3.9.3. Single-cell RNA-seq clustering

The 4 pre-processed per-time point Seurat objects (0, 16, 24, and 72 hpa; WT, GFP mock, *Smed ELAC2* KD) were merged into a single Seurat object for joint clustering. SCTransform normalization was inherited from the first object. Genes with high biological variance (HVGs) (Hafemeister and Satija 2019), selected by ranking genes by the residual variance of Pearson residuals after SCTransform normalization, were used for dimensionality reduction. In total, 3,000 HVGs were retained. PCA was computed on HVGs and the shared-nearest-neighbor (SNN) graph was built in PCA space using the first 25 PCs. The number of PCs was chosen using an elbow criterion on a smoothed variance-explained profile, with the additional requirement that approximately 80% of cumulative variance be retained. Clustering was performed using Louvain graph community detection across a grid of kNN values (k = 5, 8, 10, 15, 20) and resolution parameters. Candidate clustering parameters were evaluated using graph modularity, the fraction of very small clusters (defined as clusters comprising <1.5% of all cells), and deviation from a target granularity of ~60 parent clusters. These criteria were combined into a composite score used to select final parameters. Detailed description of the strategies is provided in Results **Section 4.4.1.3**.

3.9.4. Cell-type annotation framework

A consensus annotation (**Supplementary Script 4**) combined 8 strategies against a curated marker set. (i) average-expression marker voting based on z-scored cluster mean expression across curated markers, (ii) hypergeometric marker-set enrichment using the global expressed background, (iii) DEG-to-marker "majority" overlap voting, (iv) log2 fold-change–weighted marker scoring with optional –log10(FDR) weighting and penalties for negative markers when provided, (v) CellMaNam-based marker occurrence, (vi) a cluster-specific hypergeometric enrichment variant using a cluster-expressed background, (vii) fgsea-based gene-set enrichment on differentially expressed markers-ranked log2 fold changes, and (viii) UCell signature scoring computed on a per-cluster subsample of cells. Detailed description of the strategies is provided in Results **Section 4.4.1.3**. Per-method labels were stored in the Seurat metadata, and a final consensus label was assigned by an equal-weight vote across methods. Clusters below the agreement threshold (the fraction of methods supporting the label) were automatically re-evaluated by within-cluster subclustering followed by a second annotation.

3.9.5. *Smed ELAC2* expression distribution across cell populations

Smed ELAC2 abundance was quantified per cell from the scRNA-seq data using the normalized expression matrix from WT samples only. Cells were classified as *Smed ELAC2*$^{+}$ when the *Smed ELAC2* expression was greater than zero, and *Smed ELAC2*$^{-}$ otherwise. For each time point, the total number of cells, the number and fraction of *Smed ELAC2*$^{+}$ cells, and the mean *Smed ELAC2* expression (across all cells and within *Smed ELAC2*$^{+}$ cells) were calculated for every cell population. To describe which populations contributed most to the *Smed ELAC2*$^{+}$ cells, the share of the *Smed ELAC2*$^{+}$ cell pool was computed as the number of *Smed ELAC2*$^{+}$ cells in a given population divided by the total number of *Smed ELAC2*$^{+}$ cells in the same condition and time point.

Two complementary summaries were generated. First, a stacked bar plot showing the composition of the *Smed ELAC2*$^{+}$ pool across populations. Second, enrichment of *Smed ELAC2*$^{+}$ cells by population was quantified using a binomial regression fitted to aggregated counts of *Smed ELAC2*$^{+}$ compared to *Smed ELAC2*$^{-}$ cells per population, with adjustment for time point.

3.9.6. Cell population composition associated with *Smed ELAC2* knockdown

For each cell population within a given time point, the number of cells contributed by each group was counted and then converted to proportions by dividing by the total number of cells in that population at that time point. These within-population proportions were visualized as a 100% stacked representation, where each bar corresponded to a single cell population and the stacked segments (WT, GFP mock, *Smed ELAC2* KD) summed to 1, allowing direct comparison of condition composition even when absolute population sizes differed. Because biological replicates were not available, the visualization was complemented by descriptive per-time point proportion tests to flag cell populations in which the *Smed ELAC2* KD fraction differed from WT and GFP control. For each cell population, the fraction of that population within each condition (cells of that type divided by all cells in that condition at that time point) was compared using a two-proportion test (Fisher's exact test when expected counts were low), followed by Benjamini–Hochberg correction. To reduce false positives *Smed ELAC2*-specific calls were required to be significant for *Smed ELAC2* KD vs WT, and either not significant for GFP mock vs WT or also significant for *Smed ELAC2* KD vs GFP mock in the same direction.

3.9.7. Single-cell RNA-seq gene expression profiling

Differential expression was computed separately for each regeneration time point (0, 16, 24, 72 hpa) using time point-specific Seurat objects derived from the merged dataset. Within each time point, analyses were performed independently for each curated cell population. For each population and time point, pairwise comparisons were tested when both groups were present and each contained at least 10 cells: *Smed ELAC2* KD vs WT, S*med ELAC2* KD vs GFP mock, and GFP mock vs WT. Differential expression was calculated with Seurat FindMarkers using the Wilcoxon rank-sum test on the SCTransform-normalized SCT assay, requiring testing genes detected in a minimum fraction of cells (min.pct = 0.10 (10%)) and logfc.threshold = 0.25. Multiple testing correction was applied using Benjamini–Hochberg method, and genes were retained at p_val_adj $\leq$ 0.05. To derive *Smed ELAC2*-specific population-level signatures, results were post-filtered per time point and population by retaining genes significant in *Smed ELAC2* KD vs WT that were not simultaneously significant in GFP mock vs WT with the same direction of change. Gene ontology enrichment analysis of scRNA-seq gene sets

was performed using the same workflow and parameters as for bulk long RNA-seq (**Section 3.4**), including topGO with Benjamini–Hochberg correction.

3.9.8. Trajectory inference

Spliced and unspliced count matrices were generated for each scRNA-seq library using the Velocyto workflow on the Cell Ranger pos-sorted genome BAM files and stored in Loom format (La Manno et al. 2018). All Loom files were loaded in blocks (25,000 cells per block) to limit peak memory use.

Cell-type labels and metadata were transferred from the curated Seurat object. The original Seurat UMAP coordinates were retained for reference, and a single joint UMAP embedding was additionally computed in Scanpy and reused for all visualizations (Wolf et al. 2018). To define a consistent neighborhood graph, PCA was performed on log-normalized spliced counts using a highly variable gene set (3,000 genes). Genes expected to dominate variation but not inform lineage movement (for example ribosomal, cell-cycle, and RNA-processing genes) were excluded from the PCA feature set, while a curated set of lineage markers was forced to remain in the feature set. A k-nearest neighbor graph was then built on the PCA space (n_neighbors = 30, n_pcs = 30), and UMAP was computed once from this graph using a fixed random seed.

Trajectory topology was summarized with partition-based graph abstraction (PAGA) (Wolf et al. 2019). PAGA was computed separately for each condition within each time point, but always using the same global kNN graph to ensure comparability. In this framework, each node represents a cell-type cluster and the weight ("connectivity") of each edge quantifies how strongly two clusters are connected in the single-cell kNN graph after normalization by within-cluster connectivity. Uncertainty of PAGA connectivities was quantified by stratified bootstrap resampling within each cell-type cluster (B = 100 resamples; 80% of cells per partition, sampled with replacement). For each bootstrap replicate, PAGA was recomputed and edge weights were summarized by the mean and the 2.5% and 97.5% quantiles, which were treated as an empirical 95% confidence interval. Within each time point, contrasts to WT were computed for *Smed ELAC2* KD and GFP (KD–WT and GFP–WT) and *Smed ELAC2*-specific effects were defined as (KD − WT) − (GFP–WT). *Smed ELAC2*-specific effects were called significant when the bootstrap confidence interval excluded zero. All steps were performed in custom Python script (**Supplementary Script 6**).

3.10. Low input sncRNA-seq data analysis

Low-input sncRNA libraries from FACS-sorted cells (X1 neoblast-enriched and X2/Xins non-neoblast-enriched) were prepared with the GenXPro low-input kit (The GenXPro small RNA sequencing kit (v1.0)), which incorporates unique molecular identifiers (UMIs) to support higher PCR cycle numbers while enabling removal of PCR duplicates during analysis. The bioinformatic workflow recommended by the manufacturer involved identification and removal of PCR duplicates based on the UMI sequences. After adapter/quality trimming and removal of long homopolymer tails, only inserts ≥14 nt were retained. UMIs were extracted after trimming using a custom bash script (**Supplementary Script 7**) that requires an 8-nt 5′ UMI and records a 4-nt 3′ UMI. Reads were deduplicated before alignment by exact identity of the insert plus the UMI pair. Deduplicated reads were aligned using Bowtie2 in sensitive local mode against a curated ncRNA FASTA collection. This reference differed from the whole-animal sncRNA-seq workflow (**Section 3.5**) in that it was restricted to ncRNA sequences supported in *S. mediterranea* (derived from RNAcentral entries that were mapped to the planarian genome and filtered to retain planarian-supported ncRNAs), which reduced off-target assignments and ensured that quantification was performed against a planarian-only ncRNA catalog (**Supplementary Script 7**). Alignments were quantified with featureCounts *(-M --fraction*) in accordance with the manufacturer's recommendations. Because featureCounts require an annotation model, a minimal single-exon transcript GTF was generated from FASTA headers, with 1 transcript feature spanning positions 1 to transcript length for each FASTA record. Each reported alignment of a multi-mapper contributed 1/x of a count, where x represents the number of valid mapping locations reported for that read after alignment. Transcript lengths were used to compute TPM, and counts/TPM matrices were assembled across samples. Only plus-strand alignments were retained for quantification to enforce strand consistency. rRFs were excluded prior to differential testing, and only sncRNAs with sufficient read support were retained (with a total count of at least 10 reads summed across all samples after rRF removal (rowSums ≥ 10)).

3.11. sncRNA target activity inference in single cells

Single-cell transcriptomes were analyzed in R using a preprocessed in previous steps Seurat object that contained normalized gene expression, annotated cell populations,

and UMAP embeddings (**Section 3.9.3**). The analysis was restricted to the 72 hpa time point.

Candidate sncRNAs were defined from bulk sncRNA differential expression results comparing *Smed ELAC2* KD to controls at 3 dpa. The bulk log2 fold change was used to define an expected direction of regulation in the knockdown: when a sncRNA was reduced in *Smed ELAC2* KD, decreased regulatory activity and a corresponding tendency for targets to increase in the knockdown were expected.

Predicted mRNA target sets were generated separately for each sncRNA under several binding models to capture different plausible pairing registers. These models included a canonical miRNA-like seed match (7mer-m8) and 7-mer offset variants (7mer-1A, 8mer-1A, (Oulas et al. 2015)) for miRNA and tiRNA candidates, as well as an extended piRNA-like model for tiRNs and piRNA candidates using a reverse-complement motifs corresponding to guide positions 2–8 (7 nt), 2–11 (10 nt), and 2–12 (11 nt), with assigned weights 3, 2, 2, 1, 2, and 3, respectively. For each model, the appropriate reverse-complement binding motifs were derived from the sncRNA sequence and searched across transcript sequences, prioritizing 3′ UTRs and optionally allowing coding-sequence matches with reduced weight (0.5). Motif occurrences were counted per transcript, converted into weighted target scores, and then aggregated to the gene level using a transcript-to-gene mapping. Genes were retained as putative targets only if they achieved a minimum gene-level targeting score, and the target set was restricted to the top-ranked 500 genes per sncRNA-model combination to limit noise from very large gene lists and to stabilize module scoring across models and sncRNAs.

To quantify sncRNA activity in individual cells, the expression of each predicted target set was compared to expression-matched control genes to account for baseline expression differences across genes. First, mean expression for each gene was computed across all analyzed cells, and genes were assigned to 1 of 24 expression bins based on quantiles of these mean expression values. For each target gene, 20 control genes were randomly selected from the same expression range (excluding the target genes), and then these controls were combined for all target genes. For each cell, a weighted mean expression of the target genes was computed using gene-level targeting scores as weights (normalized to sum to 1), and the mean expression of the matched control set were computed without weights. A per-cell score was defined as: Score = (target mean) − (control mean), and the activity metric was defined as Activity (−Score) = (control mean)

− (target mean). Under this definition, larger (more positive) values indicates that predicted targets are expressed lower than expected given their baseline expression, which is consistent with stronger net repression of the target set. More negative values indicate relatively higher target expression, consistent with weaker repression or de-repression. Random sampling steps were made reproducible by using deterministic pseudo-random seeds derived from the sncRNA and model identifiers.

Activity scores were then summarized at the level of annotated cell populations by computing mean activity per condition within each population at 72 hpa. The primary comparison was the difference in mean activity between *Smed ELAC2* KD and WT within each population (Δ = mean (Activity$_{Smed\ ELAC\ KD}$) − mean (Activity$_{WT}$)), and effect sizes were additionally expressed as Cohen's d (a standard measure of effect size that quantifies the difference between 2 means divided by the combined standard deviation). Only populations with at least 20 cells per group in both WT and *Smed ELAC* KD were considered for statistical testing. For each sncRNA, model, and cell population, the sign of the activity change between *Smed ELAC2* KD and WT was checked against the direction expected from bulk sncRNA-seq (for example, sncRNAs depleted in KD were expected to show reduced activity in KD), and only combinations with matching directions were retained for permutation testing. Statistical significance was assessed by permutation tests (2000 permutations) that repeatedly randomized the the *Smed ELAC2* to WT labels within each population while preserving group sizes, producing both two-sided and expectation-directed p-values. Multiple testing was controlled using Benjamini–Hochberg adjustment across tested combinations.

For visualization, per-cell activity scores were added to the Seurat metadata and projected onto the precomputed UMAP embeddings. To facilitate direct comparison across genotypes within a given population, a centered Δ activity was plotted, defined as: Δ Activity$_{cell}$ = Activity$_{cell}$ − median (Activity$_{WT72\ cells\ in\ the\ same\ population}$). All steps of the workflow were performed in custom R script developed as part of this doctoral research (**Supplementary Script 8**).

4. RESULTS

4.1. Development of a comprehensive genome annotation tool for analyzing coding and ncRNAs and its application to the *S. mediterranea* genome

Existing *S. mediterranea* genome annotations contain inconsistencies in transcript structures, including incomplete isoform reconstruction, fragmentary gene models, and artifacts. Given that these issues may affect all subsequent analyses (gene expression, isoform discovery, and ncRNA mapping), the first goal of this thesis was to create an improved genome annotation for the asexual strain CIW4.

As a coordinate reference, the chromosome-scale nuclear genome assembly of the sexual strain schMedS2 together with the 17.2 kb mitochondrial genome was used (**Section 3.2**). To maximize annotation accuracy for the asexual strain, several complementary evidence types were integrated (hereafter "evidence-integrated"): short-read RNA-seq and long-read transcript data generated in this study across regeneration stages (**Section 3.1**), as well as external protein and domain databases used to assign functional support to predicted transcripts (**Section 3.3**).

A two-stage strategy was implemented. First, an automated, containerized pipeline (SmedAnno) was used for read alignment, transcript assembly, structural standardization, and functional annotation (**Section 3.3**). Second, quality control and artifact suppression were applied to limit propagation of assembly errors. In parallel, a non-coding RNA catalogue was constructed by combining RNAcentral sequences with a curated planarian tRNA set, followed by deduplication and integration with the genome annotation. The workflow was implemented as a modular, containerized pipeline that incorporated complementary data types (Illumina and PacBio transcriptomic data) evidence-based integration (including reference-guided and *de novo* transcript Stringtie2 (Shumate et al. 2022) reconstruction, splice-site refinement with DeepSplice (Abrar et al. 2024), and functional-annotation-driven filtering and quality control.

When applied to the asexual strain resource, an annotation (schMed_Pn) was generated and curated, and the resulting reference was deposited in GEO (accession GSE294569). This annotation comprised 25,265 genes with functional entries, including 4,011 novel genes relative to the reference set, and automated quality-control steps flagged a subset of models for fragmentation or chimerism, a fraction of which was confirmed as artifacts during manual inspection. Comparative evaluation placed

schMed_Pn among the stronger available planarian annotations, with 767 complete BUSCOs out of 954 (426 single-copy, 341 duplicated), 34 fragmented, and 153 missing, consistent with an annotation optimized for isoform richness while retaining high core-gene completeness. In parallel, an expanded ncRNA catalogue was produced, including (among others) 179 miRNA hits across 1,042 genomic positions and 488 tRNA hits across 4,298 positions, and this ncRNA layer was integrated into the final unified annotation used for downstream analyses in this dissertation.

A detailed description of the pipeline design, benchmarking, and full results is provided in the associated published paper incorporated as a component of the dissertation (Zaremba et al. 2025)**.**

These results enabled the creation of an improved genome annotation and updated transcript model basis for further analyses performed during the course of this doctoral research.

4.2. Application of the bulk RNA-seq approach to analyze the impact of *Smed ELAC2* silencing on sncRNA pool and gene expression

4.2.1. Refinement of the *Smed ELAC2* gene model

During construction of the schMed_Pn annotation (**Section 4.1)**, the *Smed ELAC2* locus was reconstructed and the best-supported isoform was defined as MSTRG.12918.1. This locus required additional verification as unambiguous transcript coordinates were essential for expression quantification and for sequence-dependent experimental steps of the study, including RNAi dsRNA and qPCR primer design.

To validate the schMed_Pn *Smed ELAC2* model, MSTRG.12918.1 was compared against existing transcript models (**Figure 8**). Although splice junctions and exon connectivity were supported by the bulk RNA-seq alignments, a consistent discrepancy was detected at a single genomic position. Specifically, RNA-seq reads systematically supported the absence of one nucleotide present in the reference genome, indicating that the reference sequence contained an artifactual insertion at this locus.

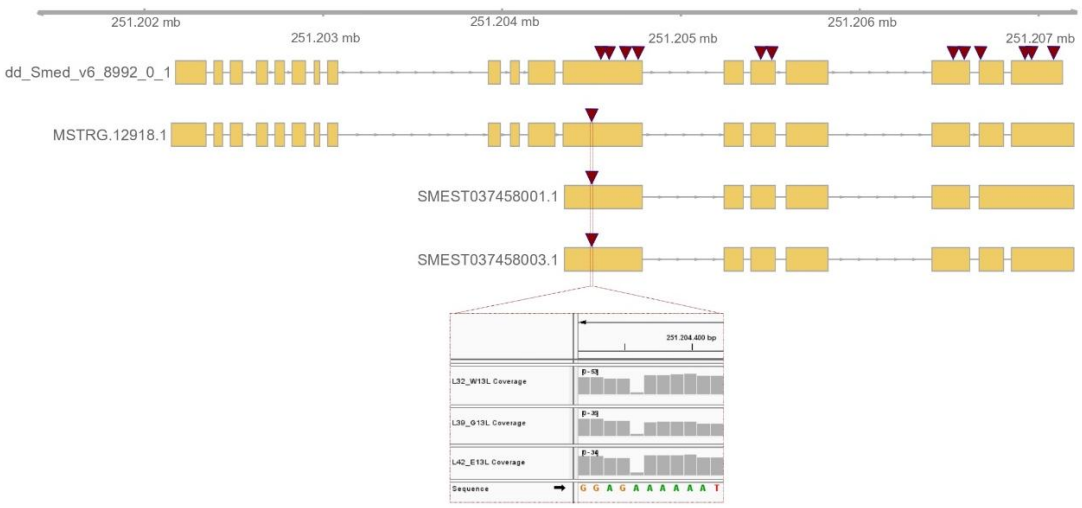

Figure 8. Annotation of *Smed ELAC2*. Transcript MSTRG.12918.1 was assembled using the StringTie2 program and compared with transcripts available in the PlanMine database. Purple triangles indicate nucleotide substitution sites. The graph in the square shows read coverage mapped to the genome at the site of the detected genomic insertion.

Coding potential was then assessed by longest ORF prediction (https://www.ncbi.nlm.nih.gov/orffinder/). After removing artifactual base, the MSTRG.12918.1_corrected model resulted in a full-length Smed ELAC2 ORF with all conserved motifs and an N-terminal mitochondrial targeting sequence, validated using MitoFate web server (Fukasawa et al. 2015). Instead, the dd_Smed_v6 and SMEST-derived models remained defective due to additional N-terminal substitutions or truncation (**Figure 9**). The corrected *Smed ELAC2* transcript model was then incorporated into schMed_Pn annotation and was used as the reference for downstream bulk RNA-seq and sncRNA-seq analyses.

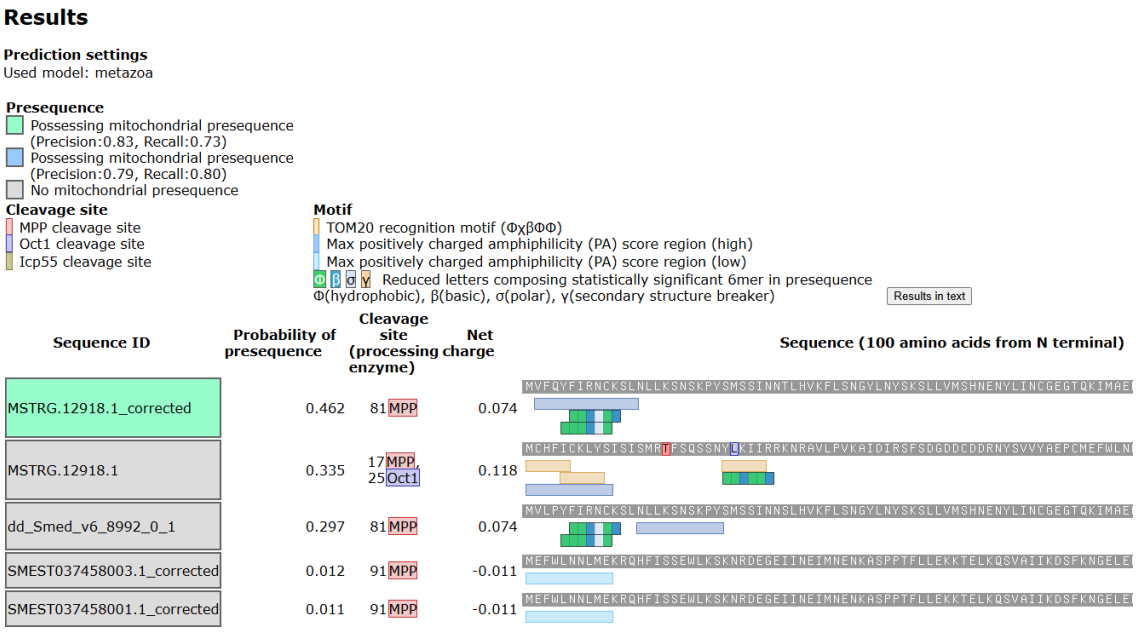

Figure 9. **MitoFates prediction of mitochondrial targeting for Smed ELAC2.** Prediction output from the MitoFates web server comparing the N-terminal mitochondrial presequence probability across *Smed ELAC2* longest ORFs (including the corrected MSTRG.12918.1 model).

4.2.2. Regeneration phenotype associated with *Smed ELAC2* knockdown

A general phenotypic assessment was performed to determine whether knockdown of *Smed ELAC2* caused a regenerative defect that could be further linked to molecular profiles. In this study, all laboratory procedures were performed by Dr. Annasha Dutta as part of her doctoral research. To evaluate regeneration outcomes, RNAi-mediated knockdown of *Smed ELAC2* was performed by targeted dsRNA microinjections, after which animals were amputated into head, trunk, and tail fragments and followed across standard post-amputation time points (0, 3, 5, 7, 10, and 14 dpa). Two controls were processed in parallel, consisting of GFP mock dsRNA-injected animals and uninjected WT animals.

No significant differences in locomotor speed or coordination were recorded after *Smed ELAC2* KD. In contrast, a prominent regeneration defect was observed by approximately 5 dpa: nearly 70% of tail-derived regenerates exhibited abnormal eye outcomes, most frequently an eyeless phenotype or a cyclopean (single-eye) phenotype, whereas bilateral eyespot development occurred as expected in both control groups at the same time point (**Figure 10**). Eye abnormalities in trunk fragments were reported as inconsistent and were not statistically supported in that dataset. Posterior regeneration

proceeded normally across all experimental groups. Notably, the eye phenotype resolved over time, with normalization reported by 8–10 dpa, supporting interpretation as a delay in regeneration rather than a complete regeneration failure.

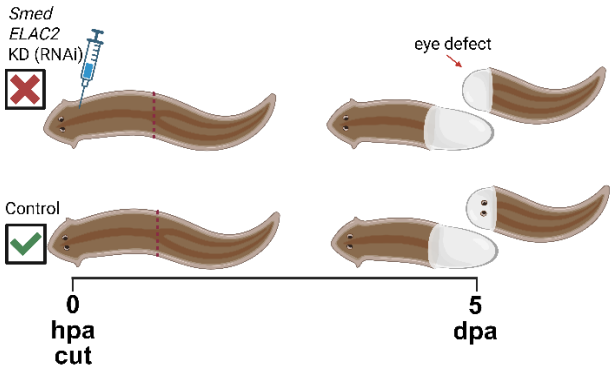

Figure 10. Schematic of *Smed ELAC2* RNAi and representative delayed eye regeneration phenotype. A schematic of the dsRNA microinjection and transverse amputation strategy used for regeneration assays. White caps indicated blastemas. A representative *Smed ELAC2* knockdown outcome in a tail-derived regenerate was illustrated by a blastema lacking visible eye pigmentation at the early scoring time point (5 dpa), while bilateral eyes were observed in control worms (WT and GFP mock).

4.2.3. Characterization of sncRNA landscape upon *Smed ELAC2* knockdown

4.2.3.1. sncRNA pool composition

To establish a baseline for interpreting *Smed ELAC2*-dependent changes, the overall composition of the sncRNA pool was first characterized across conditions (*Smed ELAC2* KD, GFP mock and WT) and regeneration stages (3 and 5 dpa). sncRNA reads were processed using the workflow described in **Sections 3.5–3.7**, which was designed to support sequence-resolved quantification across libraries while minimizing ambiguity from short-read multi-mapping. Alignments were integrated across reference layers (ncRNA catalog, reference transcriptome and genome). RNA-biotype labels were assigned by rule-based parsing of reference annotations to support class-level summaries and downstream characterization.

Using this framework, it was demonstrated that the sncRNA pool comprised a mixture of fragments derived from multiple RNA biotypes (miRNAs, piRNAs, tRFs, snoRFs, snRFs, rRFs, mRNA fragments, lncRNA fragments and other minor fragments) (**Figure 11**). rRFs showed broad length distributions consistent with turnover products

and were therefore not considered further. Across conditions, miRNAs constituted the largest fraction of the sncRNA pool, followed by tRFs, and the overall class proportions did not shift strongly enough to indicate a global collapse or expansion of any single class.

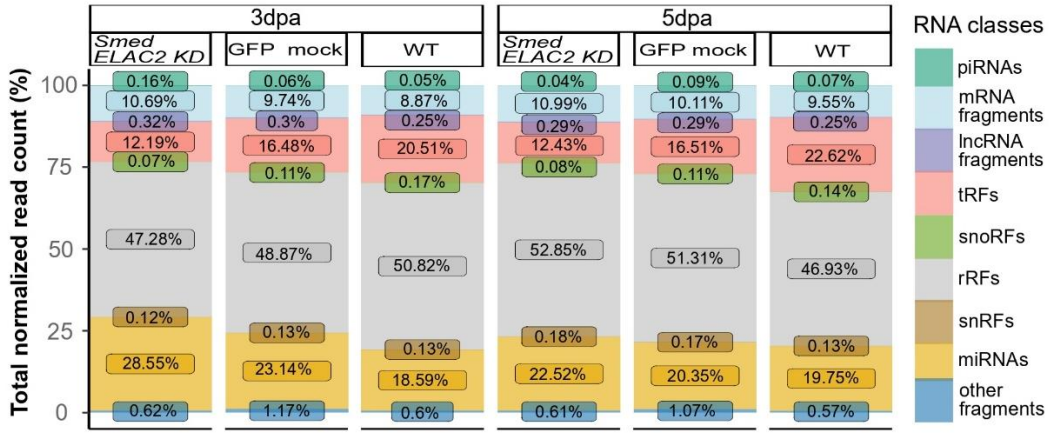

Figure 11. sncRNA class composition in *Smed ELAC2* KD and control worms. Stacked bar plots show the fraction of total normalized sncRNA-seq reads assigned to each RNA class in *Smed ELAC2* KD, GFP mock, and WT worms at 3 dpa and 5 dpa.

4.2.3.2. General differential sncRNA accumulation

To identify which sncRNA species were affected by *Smed ELAC2* silencing and to guide subsequent analyses, differential accumulation was next evaluated between *Smed ELAC2* KD and controls. Differential accumulation was assessed for sncRNA species, defined here as unique nucleotide sequences observed in the libraries. Differential accumulation was tested as described in **Section 3.5** using DESeq2 with Benjamini–Hochberg correction; significance was defined as adjusted $p \leq 0.05$ with $|\log2 \text{ fold change}| \geq 2$, and sncRNAs were required to meet minimal abundance criteria (CPM > 10 in at least two replicates). To reduce potential confounding from nonspecific RNA or injection, species showing control-specific changes (GFP mock versus WT) were excluded, leaving only *Smed ELAC2* KD-dependent alterations (**Supplementary Table S2**).

At 3 dpa, 59 species decreased and 43 increased under *Smed ELAC2* KD, with deregulated species dominated by tRFs (both directions) and miRNAs (predominantly increased), and a smaller contribution from other classes. The latter included 5 C/D box-derived species (snoRFs) with decreased accumulation, 4 lncRNA fragments with increased accumulation, 1 H/ACA box-derived species (snoRF) with increased

accumulation, and reciprocal single-species changes in the Metazoan SRP category (in the **Figure 12** depicted “other fragments”, 1 decreased and 1 increased). At 5 dpa, the deregulated set was smaller and was driven primarily by tRFs (29 species) with a notable tendency toward decreased accumulation. In addition, 1 Metazoan SRP-derived species and 1 H/ACA box-derived species also displayed decreased accumulation, while 2 miRNAs were increased.

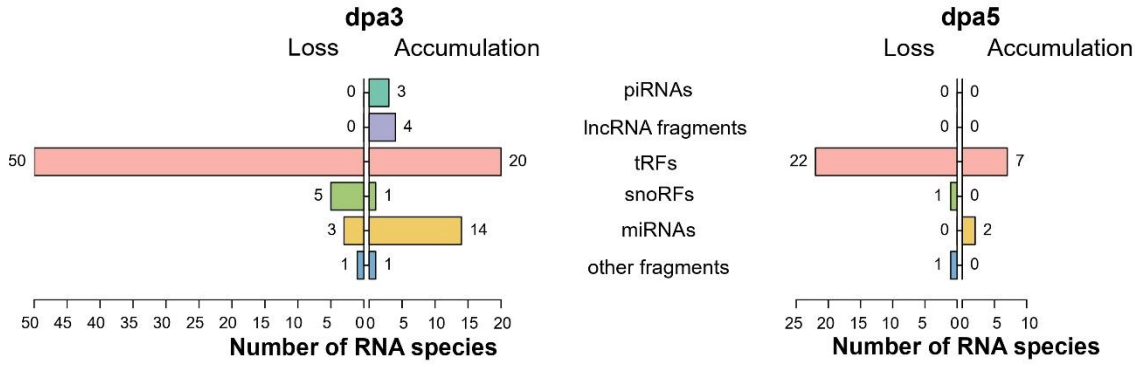

Figure 12. Differentially accumulated sncRNAs by RNA class upon *Smed ELAC2* KD. Horizontal bar plots summarize the number of significantly decreased (loss) and increased (accumulation) unique sncRNA species at 3 dpa and 5 dpa, stratified by RNA class.

Within this broader set of low-frequency classes, species annotated as piRNA represented a very small fraction of the differential signal. At 3 dpa, 3 piRNA species met the differential accumulation thresholds, with no significant piRNA changes detected at 5 dpa (**Figure 12**). Because piRNA annotations in planarian reference sets can include short repetitive sequences and fragment-like reads, these candidates were additionally examined by genome mapping. Overall, differential changes were concentrated in tRFs rather than distributed evenly across RNA classes.

Together, the developed processing and analytical framework enabled comprehensive characterization of the impact of *Smed ELAC2* silencing on the sncRNA pool, motivating the detailed miRNA- and tRF-focused analyses that follow.

4.2.3.3. Differentially accumulated miRNAs

To determine which specific miRNA were affected by *Smed ELAC2* knockdown, their differential accumulation was examined at the level of unique annotated miRNA species (**Figure 13**). In general, miRNA differential accumulation was mostly detectable at 3 dpa and weaker at 5 dpa, supporting the interpretation that *Smed ELAC2* KD affected

specific miRNAs in a time-dependent manner rather than globally shifting the miRNA fraction.

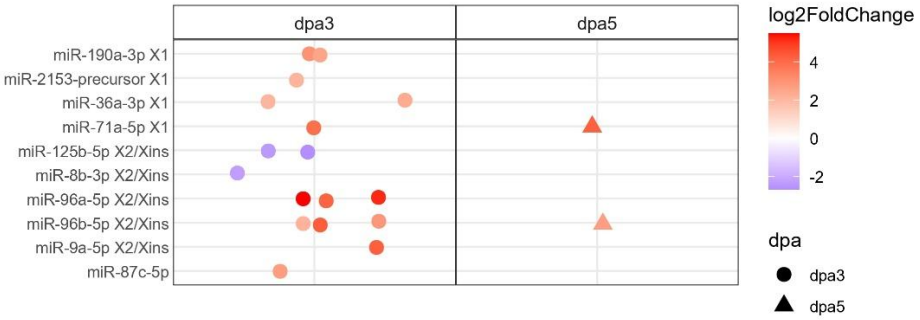

Figure 13. *Smed ELAC2* KD-dependent miRNA changes. Dot plots show log2 fold change for significantly deregulated miRNAs in *Smed ELAC2* KD worms relative to controls at 3 dpa (circles) and 5 dpa (triangles), with color indicating the direction and magnitude of change. Each symbol represents miRNA species. The entries are additionally annotated by published enrichment category (X1, neoblast-enriched; X2/Xins, non-neoblast-enriched), where available.

To enable interpretation of miRNA changes with respect to commonly used cell-population categories, deregulated miRNAs were additionally annotated using published cell-population enrichment labels (X1, neoblast-enriched; X2/Xins, non-neoblast-enriched, detailed in **Table 2**). At 3 dpa increased miRNA species included X1-enriched miR-190a-3p, miR-2153 precursor, miR-36a-3p, miR-71a-5p, as well as X2/Xins-enriched miR-96a-5p, miR-96b-5p, and miR-9a-5p (**Figure 13**). In contrast, decreased species included only X2/Xins-enriched entries, miR-125b-5p and miR-8b-3p. At 5 dpa, significant changes were restricted to a smaller subset, with increased X1-enriched miR-71a-5p and X2/Xins-enriched miR-96b-5p, consistent with a weaker miRNA signal at this stage.

4.2.3.4. Differentially accumulated tRFs

As described in **Section 4.2.3.2**, tRFs constituted the most abundant class of differentially accumulated sncRNAs. To characterize these changes, the parental tRNA isoacceptors from which the fragments originated and the fragment positions within the parent tRNA molecules were analyzed. tRF molecules were classified into subtypes according to the nomenclature described in **Section 1.1.2.1**. This positional interpretation enabled systematic classification into 5′ tiRNAs, 3′ tiRNAs, 5′ tRFs, 3′ tRFs, and internal fragments (i-tRFs), and supported isoacceptor-level aggregation of changes across

regeneration stages. The largest group comprised 5′ tiRNAs, followed by shorter tRF subtypes. Most fragments originated from nuclear-encoded tRNAs, although fragments derived from mitochondrial tRNAs were also detected.

The deregulated fragments originated from multiple parental tRNA isoacceptors, and a clear directional asymmetry was observed. Fragments that increased tended to originate from 3′ regions of parental tRNAs, with 3′-tRFs being prominent among increased species, whereas fragments that decreased were enriched for 5′ tiRNAs (**Figure 14A**).

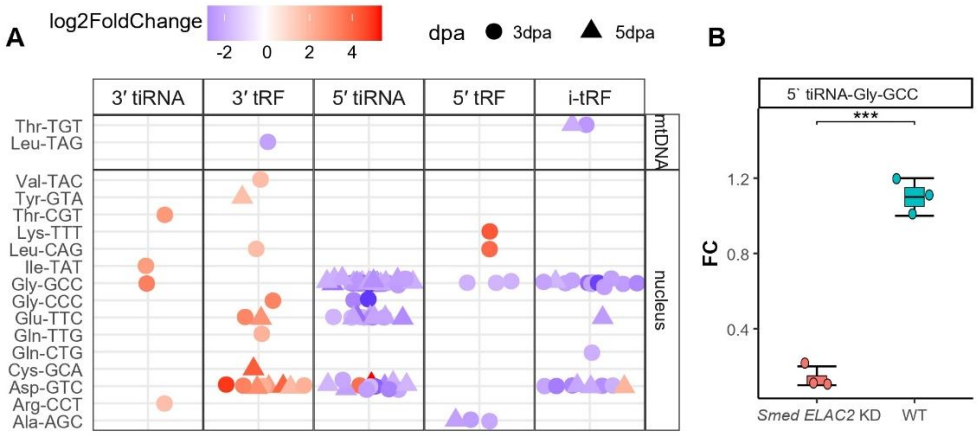

Figure 14. *Smed ELAC2* KD-dependent remodeling of tRFs subtypes across parental tRNAs. (A) Dot plot summarizes differentially accumulated tRFs upon *Smed ELAC2* KD at 3 dpa (circles) and 5 dpa (triangles), grouped by inferred subcellular origin (mitochondrial and nuclear tRNAs), stratified by fragment subtype (3′ tiRNA, 3′ tRF, 5′ tiRNA, 5′ tRF, and i-tRF) and parental tRNA isoacceptor. Color denotes log2 fold change (*Smed ELAC2* KD vs control). (B) qPCR quantification of 5′ tiRNAGly-GCC in *Smed ELAC2* KD and WT animals at 3 dpa. Statistical significance was determined using two-sided t-test: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$; ns, not significant.

Decreased 5′ tiRNAs were predominantly derived from specific parental tRNAs, including tRNA-Gly-GCC, tRNA-Glu-TTC, and tRNA-Asp-GTC, with additional decreased internal fragments from tRNA-Gly-GCC and tRNA-Asp-GTC. The down-regulation of 5′ tiRNA-Gly-GCC was experimentally validated using qRT-PCR (**Figure 14B)**.

Together, these results show that the tRF response to *Smed ELAC2* silencing was structured: increases preferentially involved 3′-derived fragments, whereas decreases were dominated by 5′ tiRNAs from a restricted set of parental tRNA isoacceptors.

In summary, by integrating established bioinformatics tools with custom-developed analytical approaches, the impact of *Smed ELAC2* silencing on the composition of the sncRNA pool was systematically assessed. Subsequently, the resulting alterations were comprehensively characterized at the level of individual miRNA and tRF molecules. Among the sncRNAs whose abundance was affected by *Smed ELAC2* silencing, 5′tiRNA-Gly-GCC was particularly notable due to its markedly reduced accumulation level.

4.2.4. Differential gene expression

4.2.4.1. Identification of differentially expressed genes

To relate sncRNA changes to co-occurring effects at the mRNA level, long RNA-seq data from the same regenerative time points were analyzed in parallel. Differential gene expression testing was performed using a conservative strategy intended to attribute changes specifically to *Smed ELAC2* KD rather than injection-related, nonspecific RNA or background effects. Genes were required to be significantly altered in *Smed ELAC2* KD relative to wild type while remaining stable between the 2 control groups, and additional abundance filtering and a minimum effect-size threshold ($|\log2 \text{ fold change}| \geq 0.57$ were applied (**Section 3.4**). Under these criteria, the number of differentially expressed genes (DEGs) was low, with 19 upregulated and 27 downregulated genes at 3 dpa, increasing to 25 upregulated and 58 downregulated genes at 5 dpa (**Supplementary Table S3**).

4.2.4.2. Gene ontology annotation and enrichment

A standard step in transcriptomic analysis is functional summarization of deregulated genes by Gene Ontology (GO) assignment, covering biological processes, molecular functions, and cellular components, followed by enrichment analysis. In *S. mediterranea,* GO coverage is incomplete for many loci, and therefore a combined GO annotation set was constructed to support enrichment analysis. First, GO assignments available in PlanMine were retrieved for genes with existing planarian annotations. For genes lacking PlanMine GO terms, additional GO assignments were transferred using homology-based evidence generated during annotation, including best-supported SwissProt matches and domain-based mappings derived from InterProScan and Pfam,

and these sources were merged into a unified gene-to-GO table (**Supplementary Table S3**) used for enrichment testing (**Sections 3.3** and **3.4**). Using this approach, GO terms were identified for 18,522 of the 25,264 genes.

GO enrichment analysis was then performed separately for 4 sets of genes: overexpressed at 3 dpa, overexpressed at 5 dpa, underexpressed at 3 dpa, and underexpressed at 5 dpa. Enriched GO terms were then post-processed and summarized: GO were assigned into broader process themes via curated custom rule-based categorization (**Supplementary Table S3**). At 3 dpa, enriched terms among upregulated genes emphasized regulatory and remodeling themes (including annotations related to proliferation regulation and extracellular or vesicle-associated categories), whereas downregulated genes were enriched for metabolic and immune-associated categories (including cholesterol-related and phagocytosis or inflammatory-response annotations) (**Figure 15**). At 5 dpa, broader disruption was observed, with upregulated genes involved in regulation of cell motility, migration and proliferation, as well as carboxylic acid and glucose metabolism, and downregulated genes remained enriched for lipid and cholesterol metabolism and immune or defense-response categories and also included membrane transport and ion-channel regulation terms, including calcium-channel regulation.

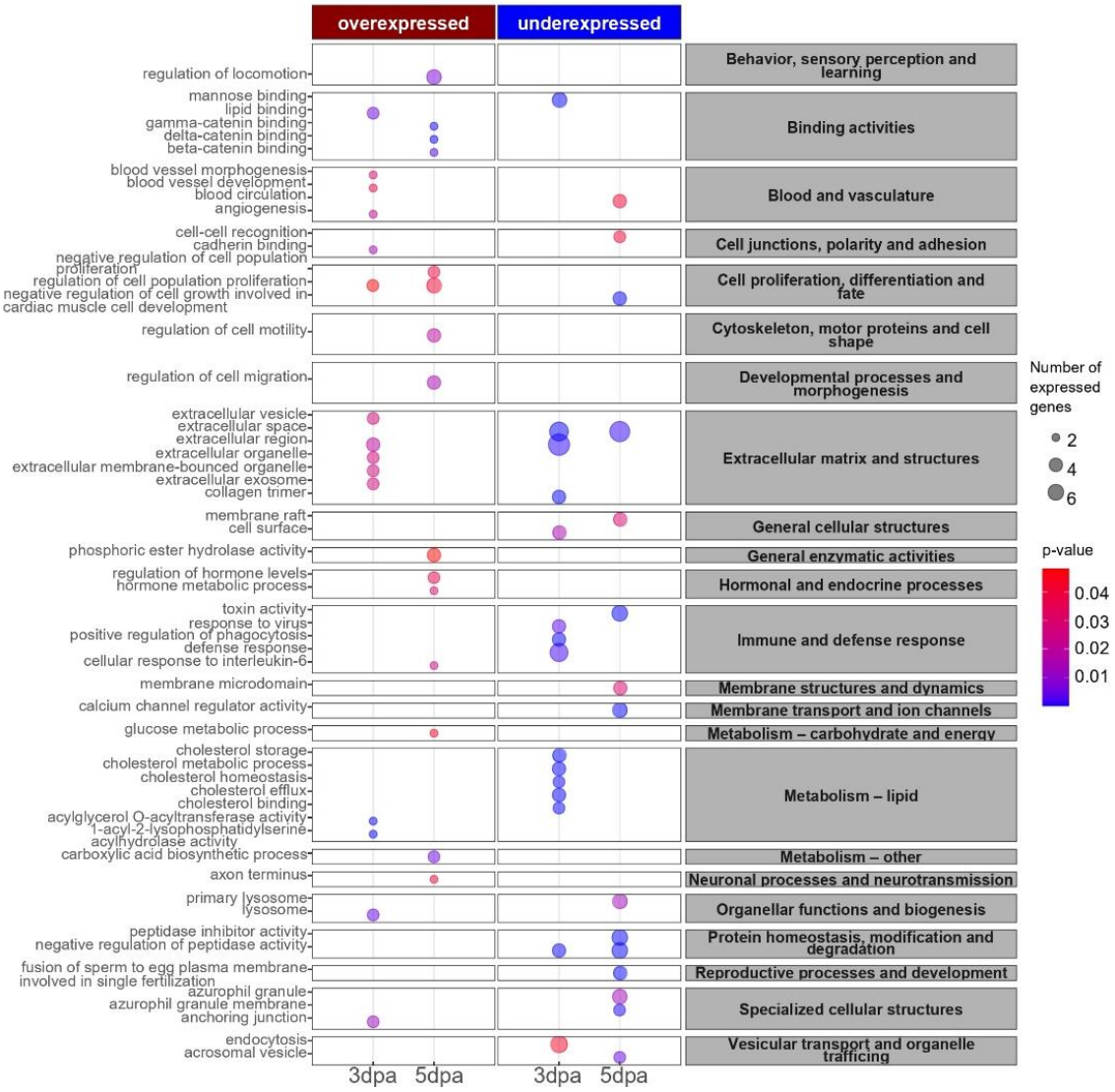

Figure 15. Functional annotation of gene expression changes after *Smed ELAC2* KD. (A) GO enrichment analysis of genes overexpressed and underexpressed after *Smed ELAC2* silencing at 3 days post-amputation (dpa) and 5 dpa. GO terms are listed on the left. The size of the dot reflects the number of genes associated with each GO term, and the color intensity indicates statistical significance (p-value). Custom functional categories manually assigned to GO terms are displayed on the right.

These results established a background for interpreting the sncRNA phenotype by indicating which mRNA programs co-varied with the observed miRNA and tRF changes.

4.2.4.3. Cell-type context for differentially expressed genes

GO enrichment analysis enabled to characterize major transcriptomic changes; however, due to the lack of phylum-specific annotations, its informative value was

limited. To provide additional functional insight in the context of planarians, DEGs were classified based on the cell types in which they are primarily expressed by analyzing publicly available scRNA-seq datasets (Fincher et al. 2018; Plass et al. 2018). A clear pattern emerged at 3 dpa, where upregulated genes were characteristic of parenchymal cells, while downregulated genes were primarily associated with intestine cells (**Figure 16**). However, by 5 dpa, as regeneration progressed, both overexpressed and underexpressed genes became more diverse in terms of the cell types in which they are mainly expressed. At 5 dpa, the set of overexpressed genes included not only those associated with parenchymal cells, but also those expressed in nervous system, epidermal, and intestinal cells. In contrast, no clear pattern was observed for underexpressed genes, as this group contained genes representing a wider range of cell types, including neoblasts, nervous system, epidermal, intestinal, and protonephridial cells.

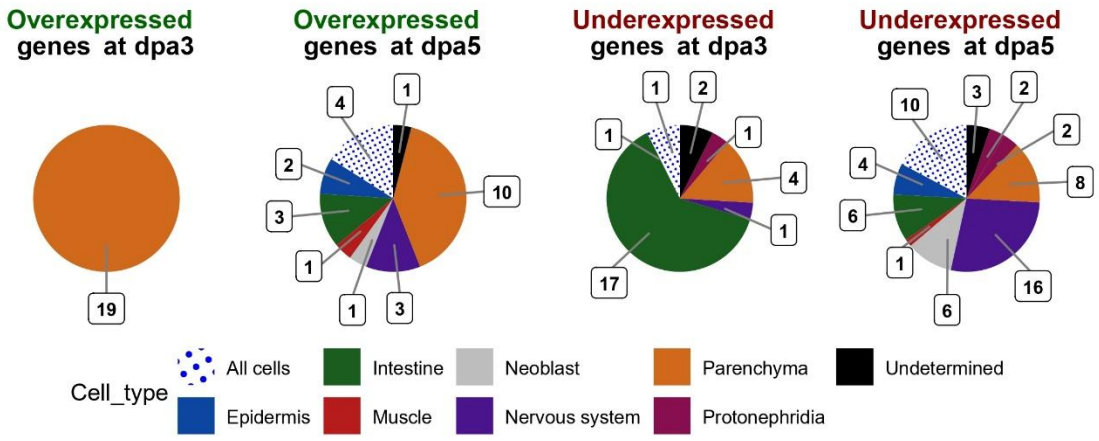

Figure 16. Cell type specific context of bulk gene expression changes after *Smed ELAC2* KD. Assignment of genes overexpressed and underexpressed at 3 dpa and 5 dpa to specific cell types, based on their typical expression patterns obtained from single-cell transcriptomic data.

Overall, the application of standard bioinformatics tools enabled the identification of genes displaying altered expression upon *Smed ELAC2* knockdown. In addition, development of a custom GO assignment approach and incorporation of data from publicly available single-cell atlases enabled the interpretion of gene expression changes in a planarian-specific context.

4.3. *In silico* functional analysis of selected sncRNA

The next stage of the study involved an *in silico* functional analysis of selected sncRNAs to provide insight into their putative roles in regeneration. Given the strong downregulation of 5′ tiRNA-Gly-GCC following Smed *ELAC2* silencing and prior work in other systems supporting regulatory roles for 5′ tiRNAs (mentioned in **Section 1.1.2.1** and **Section 1.3.3**), the analyses focused on this molecule. Based on literature, three potential mechanisms of 5′ tiRNA-Gly-GCC action were considered: (i) guiding Smed ELAC2 cleavage, (ii) miRNA-like targeting via linear antisense base-pairing to transcripts, and (iii) PIWI-associated, piRNA-like targeting.

4.3.1. 5’ tiRNA-Gly-GCC as a small guide RNA for Smed ELAC2

The first considered mechanism was motivated by the observation that ELAC2 can cleave RNA targets in a small guide RNA (sgRNA)-dependent manner (Elbarbary, Takaku, Uchiumi, et al. 2009). In this pathway, known as TRUE silencing, an sgRNA (including 5′-tiRNAs and related fragments) guides ELAC2 to complementary target transcripts, enabling the cleavage (Elbarbary, Takaku, Uchiumi, et al. 2009).

Operationally, this mechanism implied the following structural constraint on the interaction: productive targeting should not be purely linear antisense duplex, but rather should support a duplex architecture that resembles major elements of a tRNA-like fold (a TRUE-like configuration). To accommodate these constraints, a two-stage target search was implemented (**Section 3.8.1**). First, the full planarian transcriptome (41,396 transcripts) was scanned for a composite sequence motif designed to match the expected interaction architecture: a 7-nt segment complementary to the acceptor-stem region of 5′ tiRNA-Gly-GCC, an 18–30 nt spacer, and a 3–5 nt complementary to the anticodon-stem region (**Figure 17A**). This first-pass scan applied sequence complementarity and spacing rules, without thermodynamic scoring.

Because sequence complementarity alone does not ensure that a tRNA-like fold is structurally feasible, all motif hits were then evaluated by secondary-structure prediction to test whether the inferred tiRNA–target duplex could adopt the expected tRNA-like geometry (**Figure 17B**). UFold v1.3 was used for this step because it is a deep-learning-based secondary-structure predictor trained on annotated structures and base-pairing rules, and it supports rapid inference, including prediction of pseudoknots (L. Fu et al. 2022). After motif filtering, 7 motif hits were obtained and 5 candidates were

retained after UFold-based folding assessment (**Figure 17A**). None of the structurally retained candidates overlapped with differentially expressed transcripts in bulk RNA-seq following *Smed ELAC2* KD (0 overlaps).

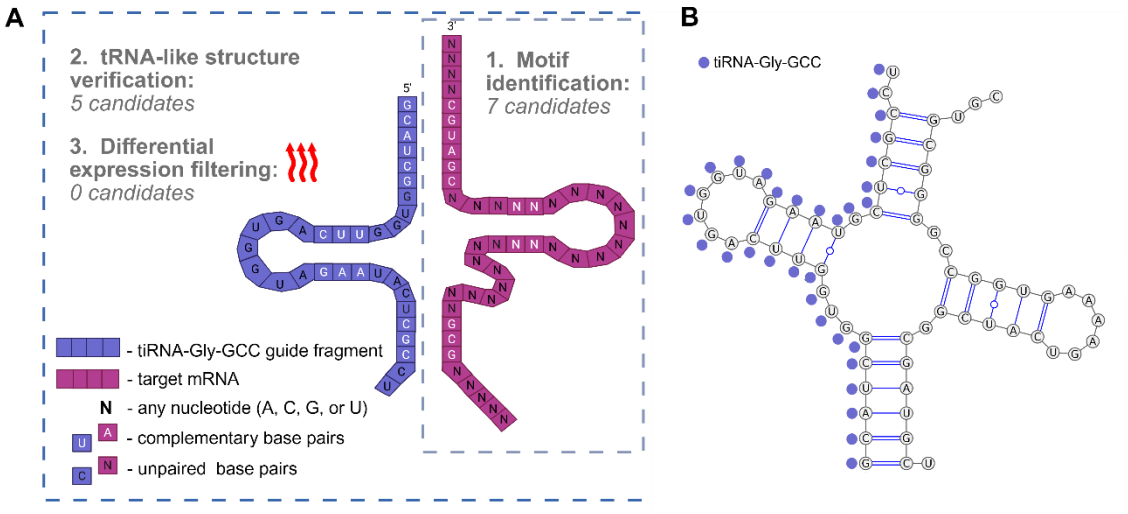

Figure 17. tRNA-like target search for 5′ tiRNA-Gly-GCC. (A) Schematic overview of the target-discovery pipeline: motif scan (7 hits) followed by tRNA-like fold verification by secondary-structure prediction (5 retained). Overlap with *Smed ELAC2* KD DE transcripts eliminated all remaining candidates (0) in bulk RNA-seq. Blue tiles denote the 5′ tiRNA-Gly-GCC guide fragment, magenta tiles denote the transcript segment scanned as a putative target site (letters shown in white). N indicates any nucleotide (A, C, G, or U). Nucleotides written in white font indicate positions that are expected to base-pair between guide and target (complementary, motif-constrained segments), whereas black font indicate positions not constrained to pair (unpaired or unconstrained positions, including spacer/loop regions). (B) UFold-predicted tRNA-like secondary structure of a representative candidate that passed structural filtering. Blue dots denote the 5′ tiRNA-Gly-GCC guide fragment, and the remaining sequence constitutes the target mRNA fragment.

Thus, it was concluded that the 5' tiRNA-Gly-GCC is unlikely to act as a guide RNA for Smed ELAC2, at least in a way that could be tracted using the available transcriptomic data.

4.3.2. 5′ tiRNA-Gly-GCC as a miRNA-like regulator

The second putative mechanism of 5′ tiRNA-Gly-GCC action was motivated by the evidence that some tRFs can act as non-canonicam miRNAs and repress transcripts via seed-like pairing and, in some cases, Argonaute association (Kuscu et al. 2018; Guan

et al. 2020). For a given molecule to function as a miRNA, it must diplay the characteristic features of miRNAs. miRNAs are summarized as: (i) typically falling in characteristic length ranges (with many AGO-bound scnRNAs in the ~20–25 nt range, and broader AGO-associated sncRNAs reported up to ~30 nt) (Bartel 2009; Juvvuna et al. 2012), (ii) recognizing targets via a seed region (positions 2–8) (Bartel 2009; Sheu-Gruttadauria et al. 2019), (iii) sometimes using supplementary pairing toward the 3′ end (including around positions 18–21) to stabilize targeting, and (iv) often showing a bias toward U at the first position (Juvvuna et al. 2012; C. B. Nielsen et al. 2007). The 5′ tiRNA-Gly-GCC analyzed here is a defined 32-nt species, which is longer than canonical miRNAs and did not match the common 5′ U/A bias because it initiated with G. However, it still shared miRNA-relevant features that justified testing a miRNA-like model. Specifically, it showed a well-defined 5′ start, supporting a consistent position for seed-based matching), and its length remained within the broader size range reported for Argonaute-associated tRFs (14–32 nt) (Kumar et al. 2014). miRNA-like targeting model made a different, and weaker, set of assumptions than the guide RNA model: it prioritized local complementarity (especially in a seed region) and allowed targeting outside canonical 3′ UTR sites, including in coding regions (**Section 3.8.2**). miRanda (Enright et al. 2003) was used to identify potential transcript matches because it is a widely used algorithm for miRNA target discovery. In miRanda, miRNA-like recognition is dependent on two main determinants. The first is sequence complementarity between a sncRNA and a transcript region. Imperfect pairing is allowed, as is typical for miRNA targeting. The second determinant is thermodynamic stability of the predicted duplex. This is reported as free energy of binding. To reduce false positives expected in a genome-scale scan, the standard miRanda scoring scheme was used with stringent thresholds (score ≥ 140; binding energy ≤ −20 kcal/mol). Predicted sites were not restricted to 3′ UTRs as UTR annotations can be incomplete in non-classical model systems. In addition, miRNA-like interactions and other sncRNA regulatory modes have been reported in coding regions in some contexts.

Large candidate sets were expected even under strict cutoffs. For that reason, a second constraint involving a *Smed ELAC2* KD-dependent perturbation was included. If 5′ tiRNA-Gly-GCC normally represses transcripts, then its reduction should lead to increased abundance of its targets. Predicted targets were therefore filtered to only keep transcripts that were upregulated under *Smed ELAC2* KD.

Standard not strict miRanda output does not explicitly quantify seed quality in the manner commonly used for miRNA interpretation. Seed-pairing parameters were therefore computed with a custom script (X.sh). The number of paired positions in the canonical 2–8 seed window was counted, including both Watson-Crick and G:U wobble pairs. The remaining candidates were then ranked using two criteria: binding energy and seed-pairing quality. This ranking was used to select a set for downstream analysis.

Application of miRanda under the defined stringency cutoffs (miRanda alignment score ≥ 140 and predicted duplex free energy ≤ −20 kcal/mol) yielded 2,480 predicted interactions. (**Figure 18**). Filtering based on the *Smed ELAC2* KD transcriptomic response reduced this set to 25 candidates by requiring inverse directionality (reduced 5′-tiRNA-Gly-GCC abundance together with increased candidate gene abundance after knockdown). Candidates were then prioritized using a combined ranking based on duplex free energy and miRanda score, together with a seed-window annotation. Specifically, pairing patterns in the seed-like window (guide positions 2–8) were summarized and any mismatches, G:U wobbles, or gaps/bulges within this interval were flagged to distinguish canonical from non-canonical seed configurations. This distinction was important because imperfect pairing in the seed-like window did not exclude functional repression in either miRNAs or tRFs. In miRNAs, non-canonical sites that contained bulges or G:U wobbles in the seed region remained functional when they were supported by additional pairing outside the seed, including 3′-compensatory pairing (Seok et al. 2016). Consistently, tRF-mediated repression required a 5′ seed match but tolerated mutations in the middle of the duplex consistent with a bulge, and was strenghtened by additional complementarity downstream of the seed (Kuscu et al. 2018). Under the mentioned ranking scheme, 3 top candidates were identified. Among these, SMESG000048842.1, annotated as a protein tyrosine phosphatase (PTP), and upregulated at 5 dpa, was selected for experimental testing despite the presence of a single gap within the seed-like region because the interaction still achieved strong overall alignment and energy (miRanda score 149; −27.88 kcal/mol). Importantly, the predicted interaction was non-canonical with respect to location, mapping to the coding sequence region.

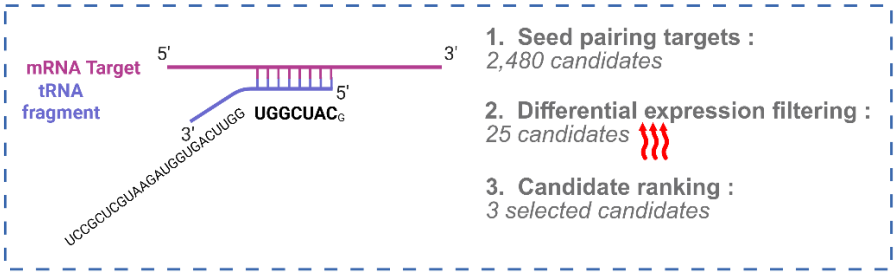

Figure 18. miRNA-like target prediction workflow for 5′ tiRNA-Gly-GCC. Putative transcript interactions were predicted with miRanda using stringent score and binding-energy cutoffs, yielding an initial seed-pairing candidate set (2,480). Candidates were then filtered to keep those upregulated upon *Smed ELAC2* KD (25), consistent with derepression when 5′ tiRNA-Gly-GCC is reduced, and the remaining sites were ranked by binding energy and seed-pairing quality to prioritize a set for downstream analysis (3).

To evaluate this predicted target, *in vivo* experiments were conducted. To this end, 5′ tiRNA-Gly-GCC levels were restored in *Smed ELAC2* KD worms using a synthetic mimic. Soaking *Smed ELAC2* KD worms in a solution of the synthetic 5′ tiRNA-Gly-GCC effectively restored SMESG000048842.1 expression to WT levels. In contrast, treating *Smed ELAC2* KD worms with scrambled RNA or no RNA solution had no effect on the SMESG000048842.1 level. (**Figure 19**).

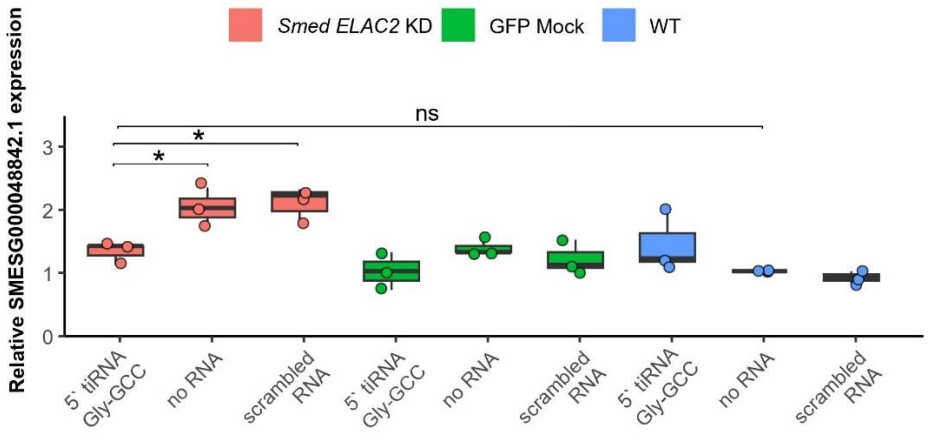

Figure 19. 5′ tiRNA-Gly-GCC downregulates the expression of protein tyrosine phosphatase *in vivo*. The level of SMESG000048842.1 was determined by qPCR, using samples collected from *Smed ELAC2* KD, GFP mock, and WT animals incubated with synthetic 5′ tiRNA-Gly-GCC, scrambled RNA, or no RNA, at 5 dpa. Statistical significance was determined using a

linear mixed-effects model followed by Tukey-adjusted pairwise contrasts: *p < 0.05; ns – not significant.

This result supported the statement that 5′ tiRNA-Gly-GCC acts as a sequence-directed repressor of at least one transcript in *S. mediterranea*.

4.3.3. 5′ tiRNA-Gly-GCC as a piRNA-like regulator

The third considered mechanism of 5′ tiRNA-Gly-GCC action assumed a PIWI-centered targeting. The rationale was that PIWI proteins can be guided by sncRNAs to recognize complementary transcripts, and that such interactions can be captured biochemically as direct RNA–RNA contacts. Importantly, as in animals piRNAs are typically 21-35 nt in lenght (Ozata et al. 2019), the 32 nt long 5′ tiRNA-Gly-GCC would fall within the canonical size range of piRNAs. In the planarian PIWI system, there are 3 SMEDWI proteins with non-redundant functions (**Table 3)**. First, SMEDWI IP sncRNA-seq dataset was used to quantify whether 5′ tiRNA-Gly-GCC was detectably associated with SMEDWI-1, SMEDWI-2, or SMEDWI-3 (**Section 3.8.3**). The 5′ tiRNA-Gly-GCC sequence was detected across all SMEDWI IP libraries, with the highest abundance observed in SMEDWI-2 IP and lower abundance observed in SMEDWI-3 IP (**Figure 20**).

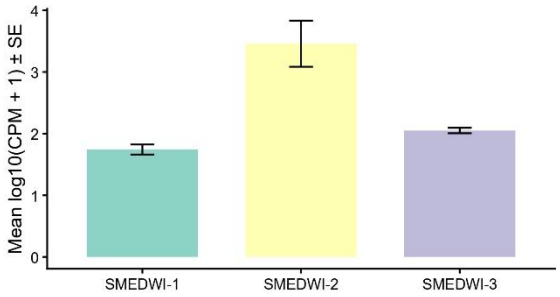

Figure 20. Abundance of the 32-nt 5′ tiRNA-Gly-GCC sequence in SMEDWI IP libraries. Bars show mean log10(CPM + 1) for the 5′ tiRNA-Gly-GCC sequence in SMEDWI-1, SMEDWI-2, and SMEDWI-3 IP sncRNA libraries. Error bars indicate variability across replicates.

The next step was to evaluate whether this association reflected direct guide–target contacts that could nominate candidate mRNA targets. This required a ligation-based interaction assay, so CLASH data were used for target discovery. However, currently, there are no SMEDWI-1 or SMEDWI-2 CLASH data publicly available. Thus,

SMEDWI-3 CLASH dataset was used as the only source of sequence-resolved RNA–RNA contacts (**Section 3.8.3**). In this assay, the sncRNA associated with SMEDWI-3 is ligated to its physically associated target RNA, yielding chimeras. As a result, this method enables the capture of sncRNAs together with their bona fide target molecules, reflecting interactions that occurred *in vivo*. Candidate chimeras carrying the 5′ tiRNA-Gly-GCC sequence were extracted, and target segments were mapped to the transcriptome to define transcript identities and coordinates. A stepwise processing pipeline was applied, and the number of reads retained after key steps was summarized (**Figure 21**). First, anchor reads were found, defined as candidate chimera reads that contained the guide "anchor" (5′ tiRNA-Gly-GCC sequence). Next, the target segments of the reads (excluding anchor sequences) were mapped to the transcriptome. Unique read identifiers for which at least one target segment was successfully mapped to a transcript model were defined as target-mapped reads. Finally, a subset of these target-mapped reads was retained as duplex-resolved chimeras, defined as reads for which a base-resolved guide–target duplex could be reconstructed by local alignment of the guide sequence to the reverse-complement of the mapped target segment. This additional filtering step was required because transcript mapping alone did not guarantee that the mapped target segment contained a reconstructable pairing site in the expected orientation or with sufficient overlap to generate a stable, position-resolved duplex annotation; for example, some mapped fragments were too short, mapped ambiguously, or did not yield a confident local alignment between guide and target segments. For the 5′ tiRNA-Gly-GCC guide, a large pool of anchor reads was retained, whereas only a smaller subset of read identifiers was retained after transcript mapping and final duplex construction (**Figure 21**).

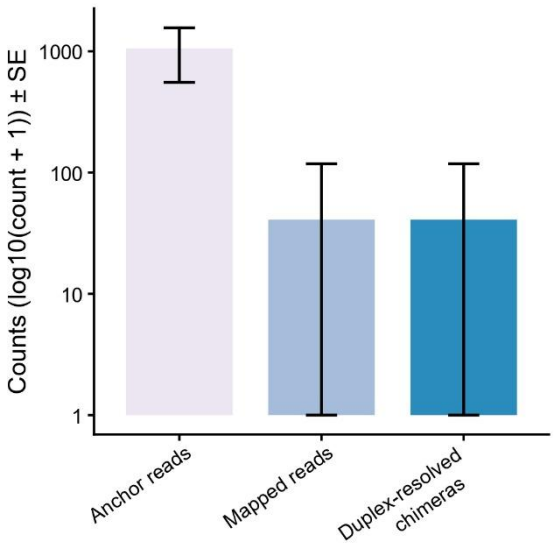

Figure 21. SMEDWI3 CLASH-supported targets of 5′ tiRNA-Gly-GCC. "Anchor reads" denote initial candidate chimera reads carrying the guide anchor. "Mapped reads" denote unique read identifiers with at least 1 target segment mapped to a transcript model. "Paired reads" denote unique read identifiers retained in the final pairing table with an explicit guide–target duplex alignment. Bars show mean log10(count + 1) across libraries. Error bars indicate between-library variability.

Next it was assessed whether CLASH-supported targets were represented among genes whose expression was altered following *Smed ELAC2* KD. Only one fragment matched a locus that overlapped with a bulk DEGs, namely the region corresponding to SMESG000030475.1 (sterile alpha motif, SAM domain). However, this CLASH fragment was only 19 nt in length, and it aligned with equal validity to several additional genomic position. Notably, the target SMESG000048842.1 predicted by miRanda (**Section 4.3.2**) was not found among the SMEDWI-3 CLASH chimeras, supporting the notion that the regulation of SMESG000048842.1 by 5' tiRNA-Gly-GCC does not involve SMEDWI-3 pathway.

In planarians, SMEDWI-3 was reported to repress mRNAs in a complementarity-dependent manner: targets with high piRNA–mRNA pairing were preferentially sliced, whereas targets with weaker pairing were bound without a detectable slicing signature (I. V. Kim et al. 2019). A candidate duplex between 5′ tiRNA-Gly-GCC and a CLASH-supported putative target segment from SMEST030475001.1 (SMESG000030475.1 DEG) was therefore visualized with seed emphasis (guide positions 2–8) and with the guide 10–11 boundary marked as the putative cleavage site (**Figure 22A**). It was found

that this interaction contained imperfect complementarity, which was more consistent with a binding-only outcome than with efficient SMEDWI-3–mediated cleavage.

Even so, a degradome-seq dataset (**Section 3.8.3**) was used independently to test whether an enrichment of 5′ ends was present at the cleavage site predicted from CLASH data, even though the illustrated duplex itself did not suggest near-perfect pairing. The degradome libraries in the source dataset were prepared by 5′-monophosphate-dependent cloning of mRNA fragments, which specifically captures 5′ ends of RNA fragments compatible with endonucleolytic cleavage products. The predicted cut coordinate was defined as the target position aligned to guide position 11 (t11), which corresponds to the expected 5′ end produced by cleavage of the target opposite to guide nucleotides 10 and 11. Degradome signal was then summarized around t11 for CLASH-supported sites, stratified by whether pairing at guide positions 10 and 11 was present. Sites with both positions paired were treated as “putative cleavage-site-compatible”, whereas all other sites were grouped as “other.” In this analysis, the “putative cleavage-site-compatible” class did not show a reproducible enrichment at t11 = 0, while sharp local spikes were observed primarily in the “other” class (**Figure 22B**). Because the “other” group, by definition, lacked the pairing pattern associated with slicing, these peaks were more consistent with heterogeneous RNA ends generated by background turnover of mRNA nucleases or unrelated endonucleolytic events, rather than PIWI cleavage. Therefore, the dataset indeed did not support a slicing-like signature for CLASH-supported sites(I. V. Kim et al. 2019).

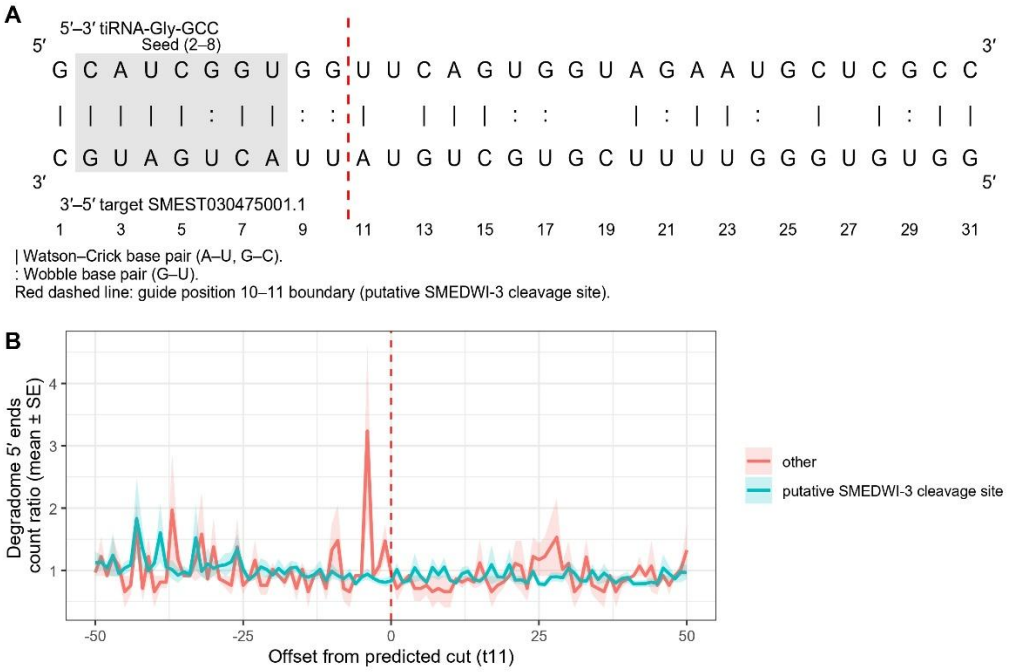

Figure 22. Representative candidate duplex and degradome profiling at the predicted slicing site for 5′ tiRNA-Gly-GCC. (A) Representative candidate duplex between 5′ tiRNA-Gly-GCC (top strand, 5′→3′) and a target segment from SMEST030475001.1 (bottom strand shown 3′→5′ to depict antiparallel pairing). The seed region (guide positions 2–8) is highlighted. Pairing symbols indicate Watson–Crick pairs (|) and G–U wobble pairs (:). The red dashed line marks the guide 10–11 boundary, corresponding to the predicted cleavage position (t11). (B) Degradome 5′-end signal around the predicted cut coordinate (t11), plotted as fold over flanks (mean ± SE) across CLASH-supported sites. Sites were stratified by whether pairing at guide positions 10 and 11 was present ("putative SMEDWI-2 cleavage site") or absent ("other"). The vertical dashed line marks offset 0 (t11).

Taken together, these analyses revealed that the 5′ tiRNA-Gly-GCC associates with SMEDWI-3, supporting its potential piRNA-like function. These findings provide a strong foundation for future experimental validation in our laboratory.

Overall, *in silico* functional analysis of 5′ tiRNA-Gly-GCC enabled the proposal of at least two potential modes of action. The first, based on the prediction strategy applied, was considered miRNA-like, whereas the second incorporated its association with SMEDWI-3 and suggested a piRNA-like mechanism. The former was experimentally validated, thereby supporting the validity of the adopted approach.

4.4. Optimization of sncRNA-seq bioinformatics tools and their application to analyze the impact of *Smed ELAC2* silencing on cell lines

Bulk RNA-seq could detect strong whole-worm transcriptional responses, but it reported an average signal from mixed cell types and could obscure lineage-restricted changes, particularly when the relevant populations were rare (X. Li and Wang 2021). Thus, in the subsequent stage of the study, the impact of *Smed ELAC2* silencing was examined at the single-cell resolution. In this analysis, datasets from WT, GFP mock, and *Smed ELAC2* KD worms were used, encompassing samples collected at early stages of regeneration (0, 16, 24, and 72 hpa) (**Supplementary Table S1**). The 72 hpa time point corresponded to one of the bulk RNA-seq sampling points (3 dpa). The generation of single-cell resolution gene expression data was intended to be subsequently integrated with information on sncRNA abundance and predicted functions, in order to gain insight into their roles in regulating cell lineages. Pre-processing included read mapping, cell and feature QC, mitochondrial content filtering, doublet control, construction of 4 per-time point separate Seurat objects, containing all 3 conditions, and initial clustering. These steps were completed by Dr Małgorzata Marszałek-Zeńczak, and all analyses reported below start from these pre-processed objects (**Section 3.9**).

Basic quality-control metrics were broadly comparable across conditions within each time point (**Figure 22**). After filtering, the number of retained cells ranged from 4,098 (WT 24 hpa) to 9,074 (*Smed ELAC2* KD 72 hpa) (**Figure 22A**). Distributions of library size per cell (unique molecular identifier, UMI, counts; log10(UMI + 1)), detected genes per cell (log10(genes + 1)), and mitochondrial read fractions overlapped across conditions (**Figure 22B–D**), confirmed the high quality of the data, arguing against a dominant influence of sequencing depth or cell quality on the downstream lineage comparisons. Together, these analyses indicated that subsequent differences in lineage composition and inferred differentiation trajectories are unlikely to be driven by gross technical biases in input data quality.

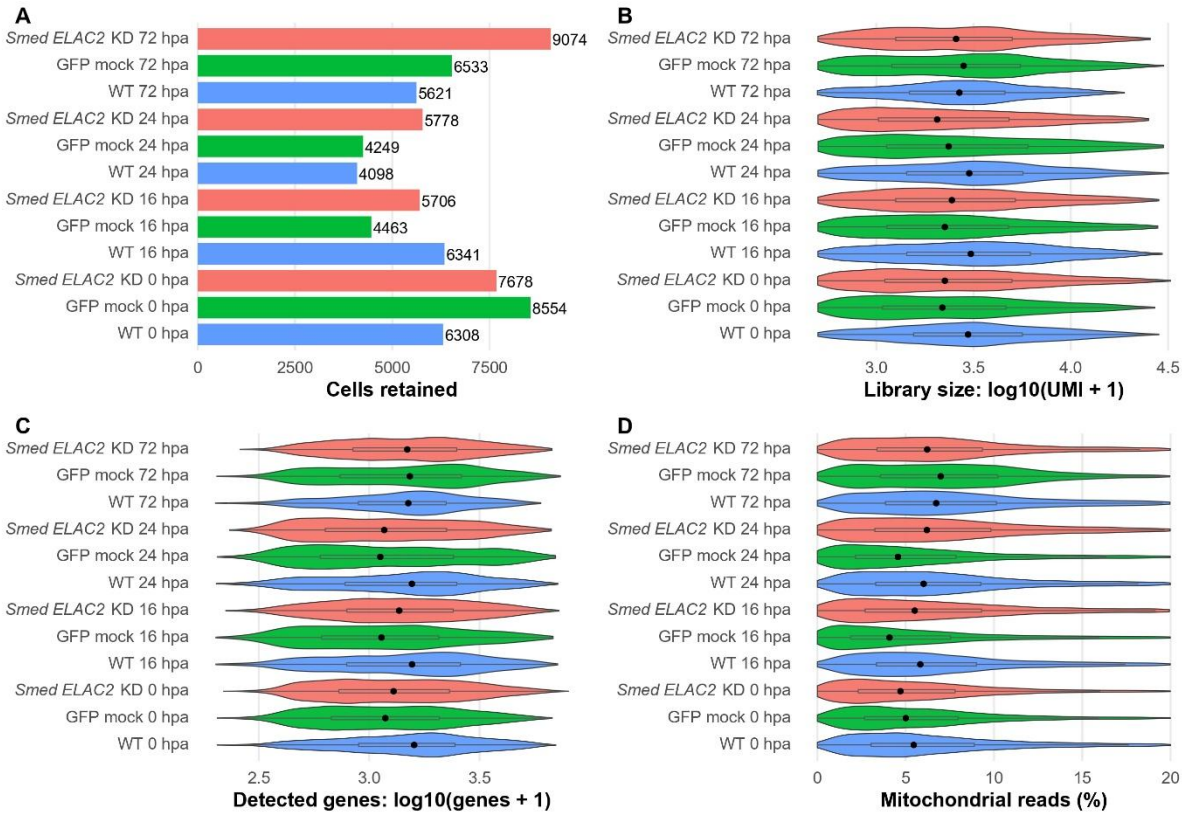

Figure 22. Quality control overview for the scRNA-seq datasets across conditions and regeneration time points. (A) Number of cells retained after filtering. (B) Distribution of per-cell library size expressed as log10(UMI + 1), where UMI denotes unique molecular identifier counts. (C) Distribution of per-cell detected genes expressed as log10(genes + 1). (D) Fraction of mitochondrial reads per cell.

4.4.1. Single-cell atlas construction and cell-type annotation

4.4.1.1. Assessment of the need for data integration

When multiple scRNA-seq datasets are compared, the data can be combined by direct merging or by computational integration. Integration is intended to reduce unwanted technical differences between datasets (known as batch effects) so that the same cell types align across samples. However, integration can also remove real biological structure when the variable being "corrected" is itself biological. In this thesis, regeneration time point and experimental condition (*Smed ELAC2* KD, GFP mock, WT) were biological factors of interest. Therefore, the need for integration was evaluated before downstream analyses were performed, and the risk of overcorrection was considered. Three alternative representations were constructed and evaluated as described

in **Section 3.9.2**. An integration restricted to WT samples was performed first, using Harmony algorithm as a representative method, to verify that the integration procedure preserved established cell identities. Next, a full integration was performed across all time points and conditions using the same Harmony algorithm. Finally, a merged representation was created by concatenating all samples and analyzing them jointly without integration.

Two complementary metrics, LISI and kBET, were used to evaluate the generated representations. LISI was used to quantify local label diversity within each cell's k-nearest-neighbor neighborhood. Conceptually, LISI converts the frequency distribution of labels (for example conditions or cell types) among a cell's neighbors into an effective number of categories, such that a neighborhood consisting of 1 label has an effective diversity of 1, and a neighborhood evenly split across multiple labels has a higher effective diversity. Because raw LISI values depend on the number of possible labels, LISI values were normalized to the [0, 1] range. For iLISI (condition mixing), the normalization iLISI_norm = (iLISI − 1)/(B − 1) was used, where B is the number of condition levels present in the evaluated representation. Under this scaling, 0 indicates minimal local mixing by condition and 1 indicates maximal mixing given the available condition labels. For cell identity, a purity-oriented transform derived from cLISI was used so that larger values corresponded to purer neighborhoods (that is, stronger coherence by cell type) and values closer to 0 corresponded to greater cell-type mixing. In addition, kBET was applied as a direct hypothesis test of neighborhood composition. In this setting, kBET evaluates whether the local neighborhood composition deviates from the global label proportions. In this evaluation, high cell-type purity was desirable (cell identities remained coherent), whereas condition mixing (iLISI_norm and kBET) should be interpreted cautiously because condition corresponded to *Smed ELA2 KD*, GFP mock and WT. Therefore, the preferred representation was the one that preserved high cell-type purity while showing the highest condition-driven neighborhood mixing (lower kBET rejection rate and iLISI_norm).

Across all 3 representations cell-type neighborhoods remained highly coherent (cell-type purity close to 1), indicating that annotated cell identities were preserved regardless of whether Harmony integration was applied (**Figure 23**). The main differences were observed for neighborhood behavior with respect to the condition label. This required cautious interpretation because condition and time point are biological

variables in this study, rather than nuisance batches, and excessive condition mixing could reflect overcorrection and loss of genuine time-resolved regeneration structure.

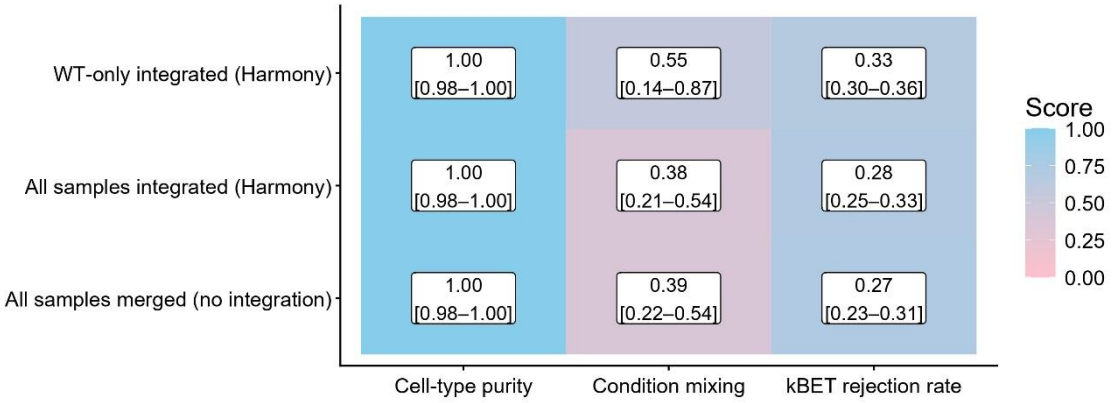

Figure 23. Assessment of dataset coherence for alternative dataset-combination strategies. Three representations were compared: merged without integration, Harmony integration of all samples, and Harmony integration of the WT subset. Left panel shows cell-type neighborhood coherence as a cLISI-derived purity score (normalized to 0–1; higher indicates purer neighborhoods by cell identity). Middle panel shows condition mixing as normalized iLISI, computed as (iLISI − 1)/(B − 1), where B is the number of condition levels (0 indicates minimal mixing by condition and 1 indicates maximal mixing given B). Right panel shows kBET rejection rate by condition (lower indicates weaker condition-driven separation in local neighborhoods). Values in each box indicate central tendency (median for LISI-based metrics; mean for kBET), and bracketed intervals indicate dispersion.

Consistent with this interpretation, integration did not provide a clear advantage. The merged dataset without integration showed the lowest kBET mean rejection rate (0.27), compared with the fully integrated dataset (0.28) and the WT-only integrated dataset (0.32), while cell-type purity was already maximal in all cases (**Figure 23**). The normalized iLISI values differed among approaches (median 0.55 for WT-only integration; 0.38 for full integration; 0.39 for merging), but higher condition mixing was not was not desirable because the condition label represented a biological variable. Therefore, downstream analyses were performed without integration, using the pre-processed per-time point Seurat objects, preserving time-specific organization while controlling condition effects analytically.

4.4.1.2. Curated marker reference assembly

Cluster annotation, i.e., assigning biological identities to clusters, is generally challenging and becomes particularly difficult in non-classical model organisms, due to the lack of dedicated reference databases. To facilitate this process, a comprehensive and manually curated marker reference had to be constructed. To this end, a planarian-specific marker reference table (**Supplementary Table S4**) was prepared by collecting lineage and subtype markers from published planarian scRNA-seq atlases and lineage studies, recording the original gene names and sources, and then mapping each marker to the gene identifiers used in this dissertation (the schMed_Pn annotation). For each entry, a general lineage label and a detailed population label were assigned, together with both a human-readable marker name and the corresponding SMESG gene ID used in the pre-processed Seurat object. Duplicates and naming variants were harmonized so that each marker was represented consistently across analyses, and supporting references were retained in the table to enable manual traceability.

The final reference contained 203 curated entries corresponding to 199 unique marker gene IDs distributed across 10 general lineages and 60 cell populations. This reference provided a standardized bridge between published marker nomenclature and the transcript models used in the current dataset.

4.4.1.3. Joint clustering and automated cell-type assignment

A single joint atlas was then constructed to support cross-time point comparisons, quantify lineage composition across conditions, and provide a unified coordinate system for downstream analyses linking gene expression profiles with sncRNA accumulations. To this end, all pre-processed per-time point SCT-transformed objects (0, 16, 24, and 72 hpa across WT, GFP mock, and *Smed ELAC2* KD) were merged into a single Seurat object (**Section 3.9.3-3.9.4**). Clustering was performed in PCA space using genes with high biological variance after normalization (HVGs). The dimensionality used for graph construction was selected from the PCA of the SCTransform-normalized joint object (**Figure 24A**). The number of informative PCs was selected using an elbow criterion on a smoothed variance-explained profile, with an additional constraint that approximately 80% cumulative variance should be retained. In the final model, 25 PCs were selected. Clustering parameters were then chosen by a grid search over k-nearest-neighbor (kNN) parameters and community-detection resolutions (Louvain and related variants), where

each line corresponded to a different k value used to build the shared-nearest-neighbor graph (k = 5, 8, 10, 15, 20; **Figure 24B–E**). Candidate parameters were scored to favor high modularity (well-separated graph communities), penalize over-fragmentation (excessive tiny clusters, defined globally as clusters smaller than 1.5% of all cells), and remain close to a target granularity of approximately 60 parent clusters. The target of ~60 was used to ensure that global clustering remained interpretable and conservative, while finer structure could be recovered only when supported by marker evidence during subclustering (**Section 4.4.2.1**). As expected, increasing resolution increased the number of clusters (**Figure 24B**). Graph modularity remained high overall but tended to decline at higher resolutions, consistent with fragmentation of communities (**Figure 24C**). The fraction of very small clusters increased with resolution, indicating progressive over-splitting (**Figure 24D**). A composite score that rewarded high modularity, penalized over-fragmentation by tiny clusters, and favored solutions near the target granularity highlighted an optimal region around intermediate resolution (**Figure 24E**). The selected parameters (k = 10 with resolution = 2.085 (dashed line) and modularity = 0.913), yielded60 parent clusters, matching the intended conservative global granularity while leaving finer structure to be recovered by targeted subclustering supported by marker evidence.

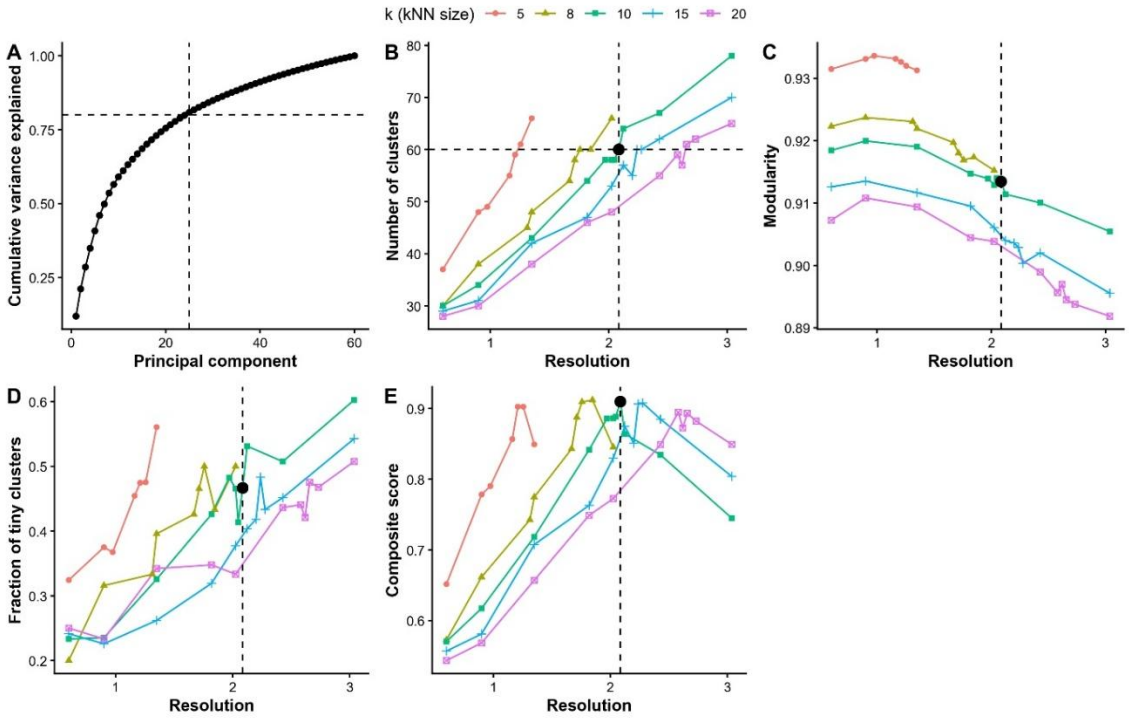

Figure 24. Selection of PCA dimensionality and graph-clustering parameters for the joint atlas. (A) Cumulative variance explained by principal components; dashed lines mark the retained dimensionality (25 PCs) and the cumulative variance level used as a guide for PC selection. (B–E) Results of a grid search over clustering resolution and k-nearest-neighbor graph size (k = 5, 8, 10, 15, 20; colored lines): the number of clusters (B), modularity (C), fraction of very small clusters (D), and the composite score combining these criteria (E). The black point and vertical dashed line indicate the selected solution (k = 10, resolution = 2.085), which yielded ~60 parent clusters with high modularity while limiting over-fragmentation.

For automated cell-type assignment, parent clusters were annotated against the generated curated *S. mediterranea* marker reference using a combination of 8 complementary strategies (**Table 5; Section 3.9.4)**. This workflow was implemented as a dedicated R script (X.R), which ran all scoring methods and combined their outputs into a consensus label (population name). The first method was an average-expression vote that compared z-scored cluster mean expression of marker genes across candidate populations (Stuart et al. 2019), and it performed best when many markers were expressed at moderate levels (e.g., multiciliated epidermis). The second method was a classical hypergeometric test (Rivals et al. 2007) of cluster DEGs against marker sets. Third, a majority-mapping custom rule assigned the label with the most positive DEGs, which worked well for clearly separated types. Fourth, a log fold-change–weighted score (Stuart et al. 2019) combined effect size and and statistical support from cluster DEG testing and included small penalties for negative markers (genes that were established markers of other lineages and were expected to be absent in the candidate population). Fifth, a cluster-specific-background hypergeometric test (Young et al. 2010) repeated enrichment with, per cluster, only genes detectably expressed as background, reducing false positives in sparse or small clusters. Sixth, UCell signature scoring (Andreatta and Carmona 2021) provided rank-based, per-cell scores that were summarized by cluster medians; this was particularly effective when ≥ 3 markers existed and expression was heterogeneous (e.g., parenchymal subsets). Seventh, Fast Gene Set Enrichment Analysis (FGSEA; (Korotkevich et al. 2016)) detected coordinated but subtle shifts spread across many markers. Eighth, a marker-occurrence–based transfer method (CellMaNam; https://github.com/jkubis96/CellMaNam) used the fraction of cells in each cluster expressing each marker above a detection threshold, to support labels in cases with partially redundant or hierarchical marker patterns. Each method returned a best-

supported label (population name) and an internal confidence score, and equal votes were aggregated into a consensus call. Disagreement across methods was treated as evidence of potential heterogeneity rather than being forced into a single label, and such clusters were flagged for subclustering.

Table 5. Summary of eight annotation strategies used in automatic cell type assignment

Method	Core input	What it measures	Major advantage	Performs best for	Limitations/ constrains
Average-expression vote	Cluster mean expression of marker genes	Marker strength at cluster level (z-scored)	Simple, auditable intensity-based call	Broad, well-powered populations with multiple moderate markers	Biased by broadly expressed markers and redundant signatures
Hypergeometric enrichment	Cluster marker list vs marker sets	Overrepresentation of marker sets among cluster markers	Detects "set-level" support independent of effect size	Marker sets of moderate size with adequate marker coverage	Sensitive to marker-set size and background definition
Majority mapping	Cluster marker list vs marker sets	Largest marker overlap count	Robust baseline for obvious identities	Well-separated types with unique markers	Hindered when related types share many markers
LogFC-weighted scoring	Marker list with effect sizes and significance	Specificity driven by a few strong markers	Prioritizes decisive markers	Populations defined by a small number of high-specificity markers	Noisy for small clusters; unstable if fold-changes are variable
Cluster-specific-background enrichment	Marker list; background restricted to genes expressed in cluster	Overrepresentation with cluster-aware background	Reduces false positives in sparse clusters	Small or sparse clusters	Can lose power if the background becomes too small
UCell scoring	Per-cell rank-based signature scoring	Median per-cell marker-set score per cluster	Robust to scale and heterogeneity	Heterogeneous clusters; parenchymal subsets	Weak for very small signatures (requires ≥3 markers in a script)
FGSEA	Ranked gene list for cluster	Coordinated enrichment of marker sets	Detects diffuse signatures across many genes	Subtle programs without single dominant markers	Unstable if ranked lists are short or sparse
CellMaNam transfer	Marker occurrence patterns across types	Occupancy-based support for hierarchical signatures	Helps when markers are partially redundant	Hierarchical or overlapping marker definitions	Conservative; may return weak / ambiguous calls when evidence is limited

Next, within each flagged parent cluster, HVGs were re-identified, PCA and the shared-neighbor graph were recomputed, and a small resolution grid was evaluated to identify stable substructure. Candidate splits were screened for graph coherence using modularity and, critically, were accepted only when minimum size constraints were satisfied to avoid artifactual over-splitting. Specifically, subclustering was performed only for parent clusters containing at least 30 cells, and each retained child cluster was required to contain at least 10 cells and at least 0.5% of the parent cluster. Child clusters failing to meet these criteria were merged to the nearest centroid in PCA space to prevent artifactual over-splitting. When a parent cluster remained excessively large to support interpretable marker-based labeling, an additional pass was applied with adjusted resolution bounds to recover biologically meaningful substructure. In this dataset, 48 of 60 parent clusters were subclustered, producing 223 subclusters.

4.4.1.4. Manual verification and final label curation

In general, cell annotation is sensitive to marker redundancy and to mixed clusters that reflect transitional states or technical mixing. A final verification step was therefore required to validate lineage logic and to ensure that rare but biologically important populations were not mislabeled. Here, “lineage logic” meaned that labels were accepted when a cell group showed the expected combination of lineage markers (e.g., *smedwi-1/-2/-3*$^{+}$ plus intestinal markers for intestinal-fated neoblasts or *smedwi-1/-2/-3*$^{+}$ plus *STNB*$^{+}$ for neural neoblasts). Automated labels were reviewed using two complementary summaries implemented in custom R script (X.R). The first one contained the mean expression level and the proportion of cells expressing each curated marker detected in the dataset. The second comprised of a cluster-level list of curated markers expressed in at least 20% of cells in the cluster. These summaries were inspected together with feature plots generated for canonical lineage markers and known negative markers. When evidence supported a revised identity, a verified cluster-to-label mapping table was produced and applied back to all cells. Manual verification resolved remaining ambiguous calls, particularly within neoblast subclasses, neuronal subtypes, and closely related parenchymal cell subtypes, and ensured that final populations were supported by explicit marker evidence rather than by clustering alone.

This approach yielded a final atlas of 74,403 cells and 36,912 detected genes, resolving 57 biologically coherent populations spanning multiple neoblast states (σ, ζ, γ, ν and lineage-biased subclasses), epidermal progenitors and mature epidermal layers (including multiciliated states), intestinal phagocytes and goblet cells, protonephridial flame and tubule lineages with precursors, ECM-producing and pole/PCG muscle states, glia, diverse neuronal classes (including cholinergic, glutamatergic, catecholaminergic, neuropeptidergic, and sensory subtypes), eye lineages, and multiple parenchymal subtypes (**Supplementary Table S5**; **Figure 25**).

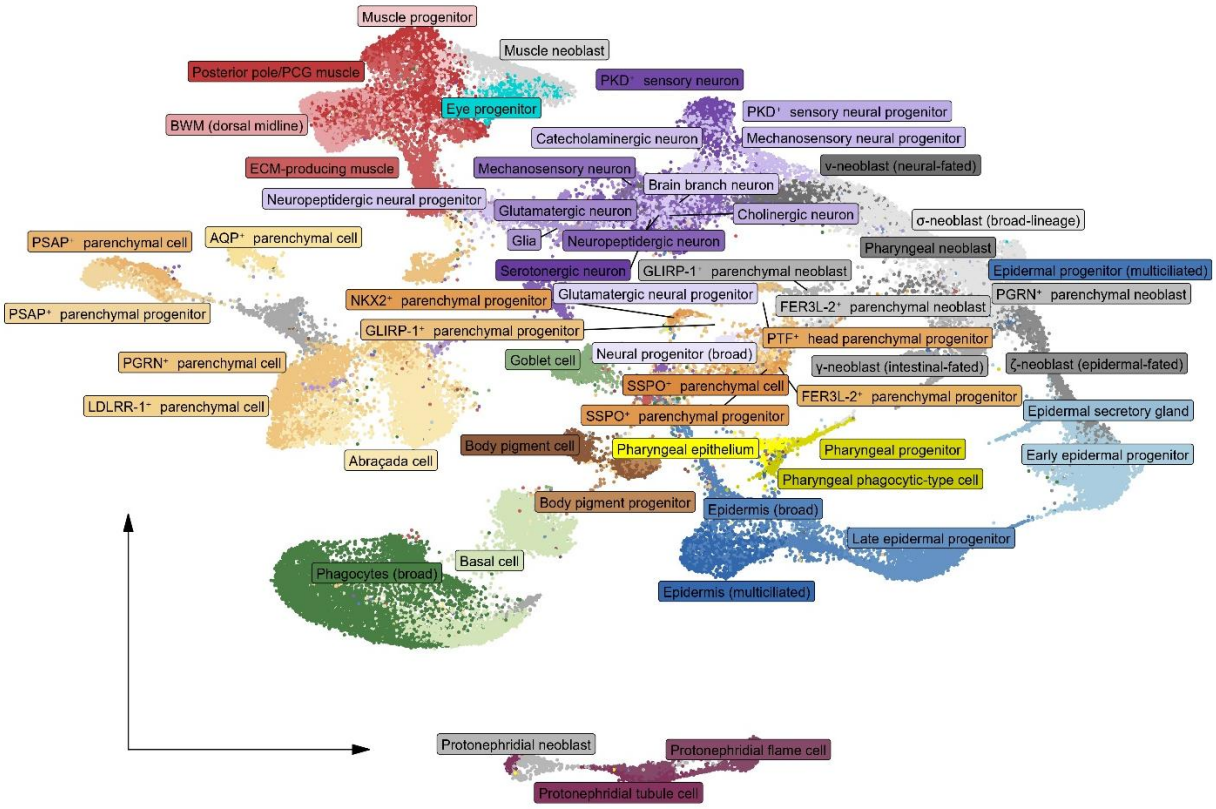

Figure 25. UMAP of the integrated scRNA-seq atlas across regeneration time points and conditions. scRNA-Seq profiling from regenerating *Smed ELAC2* KD, GFP mock, and WT worms at 0, 16, 24, and 72 hpa are embedded in two-dimensional UMAP space and colored by annotated cell population; labels denote cell lineage subtypes resolved in the joint atlas.

4.4.2. Analysis of the impact of *Smed ELAC2* silencing on cell lineages

4.4.2.1. Cell-type distribution of *Smed ELAC2* expression

To interpret lineage-level consequences of *Smed ELAC2* KD, a distribution of *Smed ELAC2* expression across cells was first established. This provided a starting point

for assessment of which populations were most likely to be directly affected by knockdown.

As described in **Section 3.9.5**, *Smed ELAC2* expression data were extracted from the scRNA-seq gene expression matrix and cells were classified as *Smed ELAC2*$^{+}$ when expression was detected (> 0). Next, the contribution of particular cell populations to the *Smed ELAC2*$^{+}$ pool in WT worms was established for all considered regeneration time points.

The composition plot (**Figure 26**) showed that the *Smed ELAC2*$^{+}$ pool was dominated across time points by neoblast-associated populations and their closely related progenitor states, with σ-neoblasts contributing the largest share and additional contributions from lineage-biased neoblast subclasses. Differentiated tissue populations contributed to smaller fractions, indicating that detected *Smed ELAC2* expression was concentrated in the stem-cell pool and immediate descendants rather than being uniformly distributed across mature tissues.

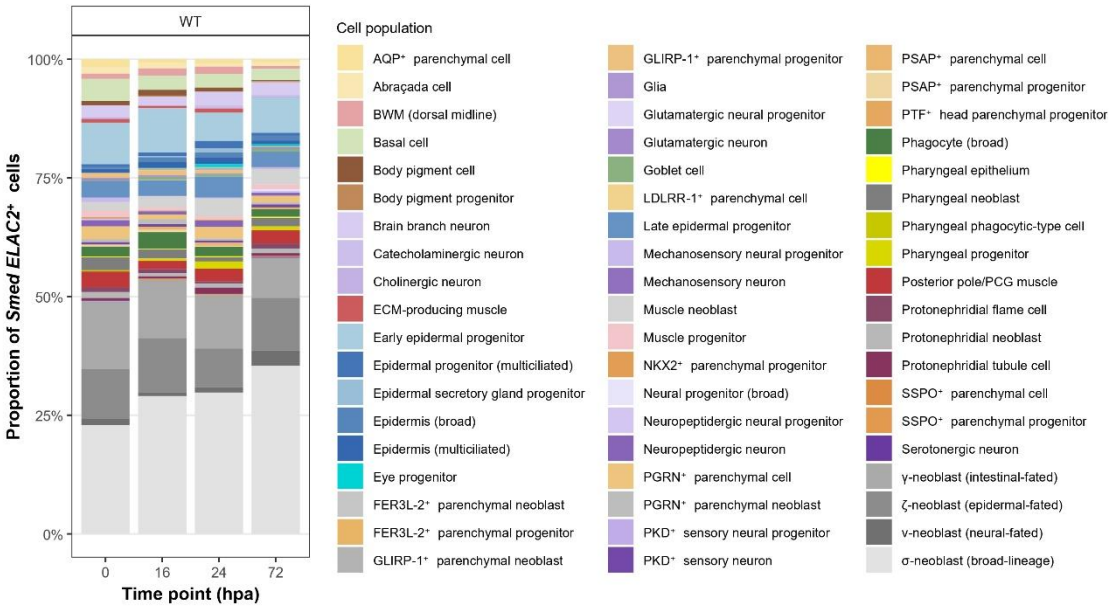

Figure 26. Cell-population composition of the *Smed ELAC2*$^{+}$ pool across regeneration time points in WT. Stacked bar plots showing for each regeneration time point the fraction of all *Smed ELAC2*$^{+}$ cells contributed by each annotated cell population.

However, the composition plot alone could not distinguish whether a population contributed strongly because it was large or because *Smed ELAC2*$^{+}$ pool was genuinely enriched within that population. This was addressed by the odds-ratio analysis (**Figure 27**), which quantified the relative probability of *Smed ELAC2*$^{+}$ cells in each population

compared with σ-neoblasts at each time point. Most differentiated populations showed odds ratios below 1, consistent with lower *Smed ELAC2* detection frequency than in σ-neoblasts, whereas a limited set of progenitor and regeneration-associated populations approached or exceeded the reference level at specific time points. At 0 hpa, only a small subset of populations showed odds ratios (ORs) above 1 relative to σ-neoblasts, meaning that *Smed ELAC2*$^{+}$ cells were detected slightly more often in those groups than in σ-neoblasts. The populations with point estimates > 1 and visibly on the right from the threshold line were epidermal progenitors (multiciliated), pharyngeal epithelium, pharyngeal phagocytic-type cells, and *PGRN*$^{+}$ parenchymal neoblasts. At 72 hpa, epidermal progenitors (multiciliated) remained above 1, indicating that this population remained relatively enriched for *Smed ELAC2*$^{+}$ cells across regeneration. In addition, γ-neoblasts (intestinal-fated) and ζ-neoblasts (epidermal-fated) showed ORs > 1 at 72 hpa, indicating that *Smed ELAC2*$^{+}$ became relatively more frequent in these lineage-biased neoblast subclasses at the late time point. In contrast, the pharyngeal phagocytic-type cells and *PGRN*$^{+}$ parenchymal neoblasts, which were > 1 at 0 hpa, shifted to below 1 at 72 hpa, indicating that their early enrichment for *Smed ELAC2*$^{+}$ cells was not maintained later in regeneration.

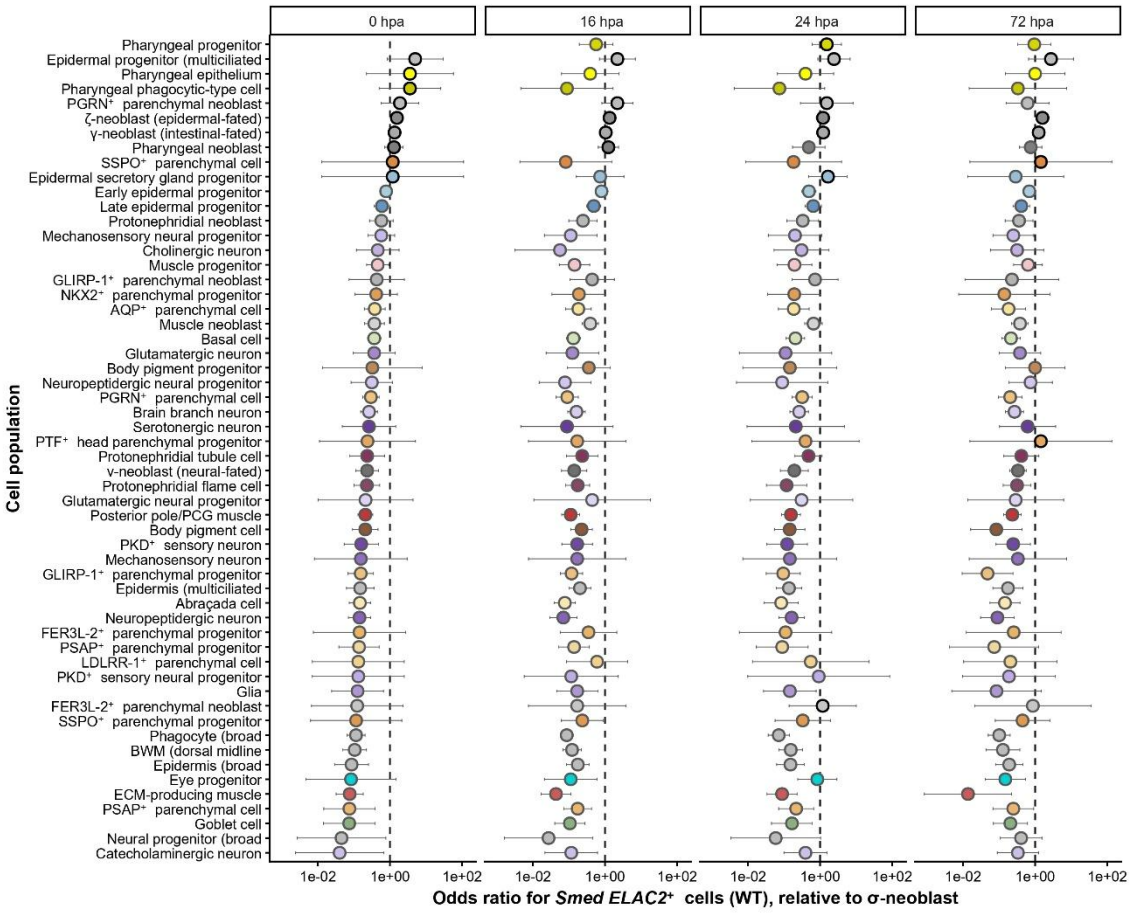

Figure 27. Enrichment of *Smed ELAC2*$^{+}$ pool by cell population across regeneration time points in WT. Odds ratios were estimated for the probability of being *Smed ELAC2*$^{+}$ in each cell population relative to σ-neoblasts within each time point. Dots indicated odds ratios and horizontal lines indicated 95% confidence intervals. The vertical dotted line at odds ratio = 1 marked the reference level (no difference from σ-neoblasts); values to the right indicated higher odds of *Smed ELAC2*$^{+}$ than σ-neoblasts, whereas values to the left indicated lower odds.

Smed ELAC2 expression was detected across multiple cell populations throughout regeneration, with the highest and most consistent enrichment in neoblast-associated fraction, including σ-neoblasts, γ-neoblasts, ζ-neoblasts, and *PGRN*$^{+}$ parenchymal neoblasts, but also in multiciliated epidermal progenitor and differentiated pharyngeal epithelium, pharyngeal phagocytic-type cells. This distribution supported the interpretation that *Smed ELAC2* KD could act both directly, through effects on stem and early progenitor states, and indirectly, through downstream consequences in differentiated lineages.

4.4.2.2. Changes in cell-population abundance over regeneration time points

The silencing of *Smed ELAC2* was expected to potentially impact on two levels: (i) the relative contribution of individual cell populations to the overall composition of the organism, and (ii) gene expression changes within particular cell populations. Regarding the first level, it was predicted that some cell populations should be gained or lost over time, rather than all cell states shifting uniformly. Therefore, cell-type composition was compared across conditions at each regeneration time point to identify which populations deviated most strongly in *Smed ELAC2* KD and when those deviations emerged.

Within each time point (0, 16, 24, 72 hpa), cells were grouped by the final curated population label and by condition. For visualization, a stacked representation was used so that, for each cell population, the bar heigth summed to 100% and showed how the cells of that population were distributed across conditions at that time point (**Figure 28**). At the same time, potential changes in the composition associated with *Smed ELAC2* were tested at each time point, as described in **Section 3.9.6**. Significant changes were highlighted in **Figure 28**.

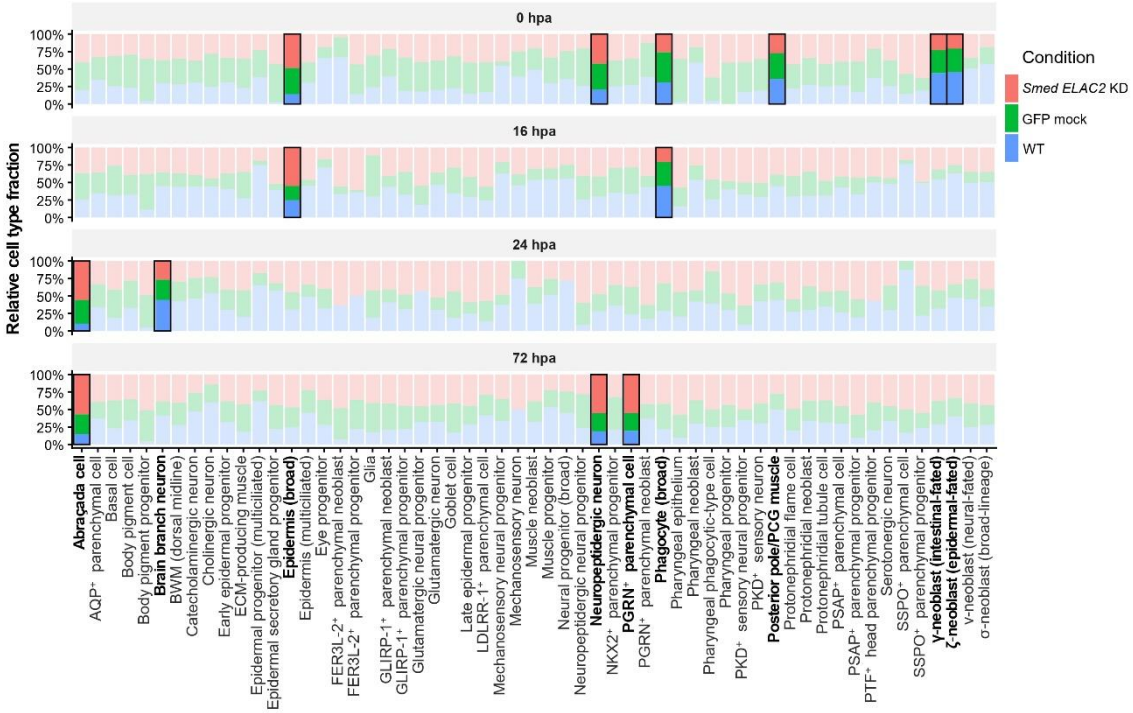

Figure 28. Cell-population composition across conditions with *Smed ELAC2* KD-specific changes highlighted. For each regeneration time point, each bar corresponds to one cell

population and is normalized to 100%, showing the fraction of cells contributed by WT, GFP mock, and *Smed ELAC2* KD within that population. Black outlines indicate populations flagged as *Smed ELAC2* KD-specific composition shifts (two-proportion testing with the Benjamini-Hochberg correction).

Across time points, the contrubutions of particular cell population to the total cell pool were broadly similar in *Smed ELAC2* KD and controls, however specific changes were observed (**Figure 28**). At 0 hpa, *Smed ELAC2* KD resulted in a relative expansion of broad epidermis and neuropeptidergic neurons, accompanied by a concominant reduction in the proportion of phagocytes, γ-neoblast, ζ-neoblast, and posterior pole/PCG muscle, relative to WT. At 16 hpa, the strongest discrepancy persisted for broad epidermis and phagocytes. At 24 hpa, *Smed ELAC2* KD caused a strong overrepresentation of abracada cells, and an opposite tendency for brain branch neurons. Finally, at 72 hpa, the relative expansion of abracada cells, $PGRN^{+}$ parenchymal cells, and neuropeptidergic neurons was observed upon *Smed ELAC2* KD.

Overall, *Smed ELAC2* KD-associated abundance shifts overlapped with WT *Smed ELAC2*$^{+}$ populations primarily in γ-neoblasts and ζ-neoblasts (**Section 4.4.2.1**). Both these populations were decreased. Notably, not all *Smed ELAC2*$^{+}$ cell populations were affected by the knockdown of this gene. Conversely, changes in the abundance of several non-*Smed ELAC2*$^{+}$ cell types, including neurons, were observed. These findings demonstrated that the consequences of *Smed ELAC2* silencing extended beyond cells in which this gene is predominantly expressed.

4.4.2.3. Differential gene expression

Next, gene expression changes induced by *Smed ELAC2* silencing within particular cell populations were evaluated. Bulk RNA-seq captured the responses of the entire worm (**Section 4.2.4**), but its resolution was limited in heterogeneous tissues. Therefore, differential expression was quantified at matched time point (72 hpa) and cell population resolution to identify which lineages mounted the strongest *Smed ELAC2* KD response, and which biological processes were altered within those lineages.

Cell population-resolved DE was computed using the atlas labels described in **Section 4.4.2.1** and the full details were provided in **Section 3.9.7**. Bulk-style replicate-based modeling (for example DESeq2) was not applicable here because scRNA-seq was available as a single biological replicate per condition and time point, and cell-level

testing could be biased when between-replicate variation cannot be estimated (Squair et al. 2021). Accordingly, results were generated as within-sample, population-resolved signatures supported by effect size and by concordance across the set of time-matched control samples (**Supplementary Table S6**). A gene was treated as differentially expressed in response to *Smed ELAC2* silencing in a given population and time point if it was significant in *Smed ELAC2* KD vs WT and was also not significant in GFP mock vs WT in the same direction (**Section 3.9.7**).

The bulk DEG set at 3 dpa contained only a small number of genes compared with the population-resolved scRNA-seq signatures (**Figure 29**). This was consistent with the expectation that lineage-restricted programs could be masked in whole-worm gene expression averages when different cell types respond in different directions. In fact, many genes appeared as both up- and downregulated in the scRNA-seq in different cell populations.

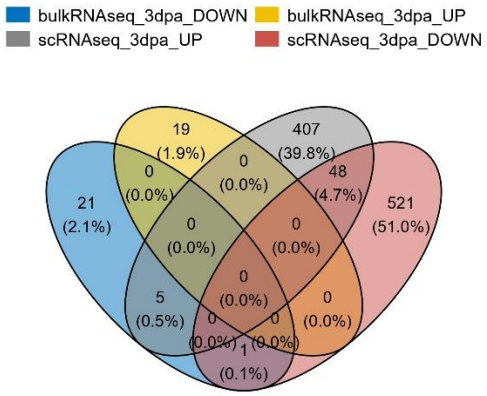

Figure 29. Overlap between bulk and scRNA-seq DEG sets at 3 dpa (72 hpa for scRNA-seq). The 4 sets are bulk UP (upregulated), bulk DOWN (downregulated), scRNA-seq UP, and scRNA-seq DOWN. Counts in intersections show how many genes were shared.

At 3 dpa, concordance between bulkRNA-seq and scRNA-seq DE was very limited, with only a small overlap (restricted to a 6 bulk downregulated genes) and no overlap between bulk upregulated genes and any scRNA-seq DEG set. The scRNA-seq DEGs were subsequently analyzed at all four time points (0, 16, 24 and 72 hpa). They were not uniformly distributed across all cell populations but were concentrated in specific lineages (**Figure 30**).

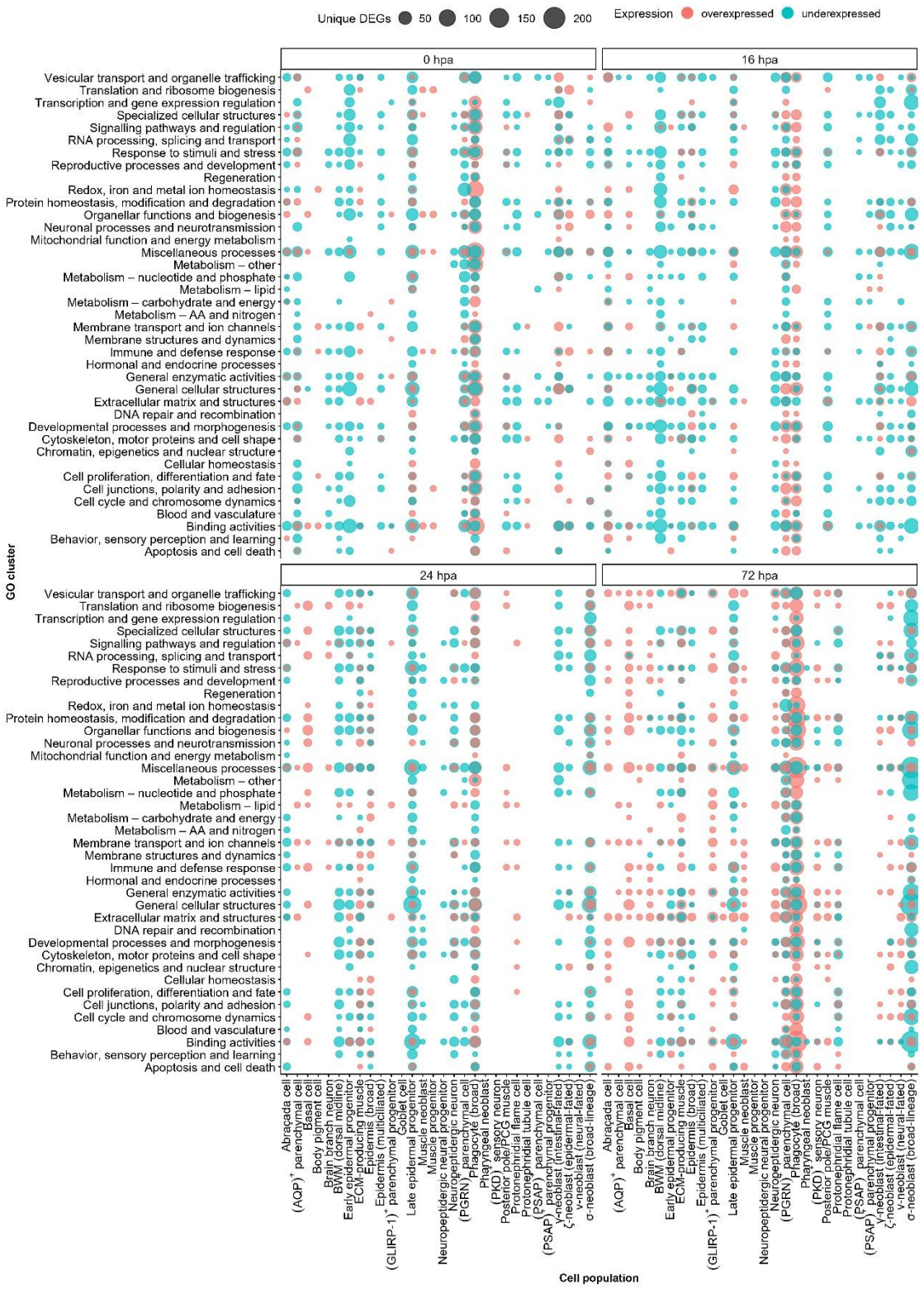

Figure 30. Functional annotation of gene expression changes after *Smed ELAC2* KD at single-cell resolution. Dot plot summarizing DEGs detected in the scRNA-seq experiment across regeneration time points (0, 16, 24, and 72 hours post-amputation, hpa) and annotated cell clusters. The y-axis groups DEGs into custom functional categories), and the x-axis lists cell populations. Dot color indicates direction of change (overexpressed vs underexpressed), and dot

size is proportional to the number of unique DEGs assigned to a given GO cluster within a given cell cluster at that time point.

Across time points, the qualitative pattern shifted from underexpression dominating at 0–24 hpa to increasingly bidirectional remodelling at 72 hpa. Upon *Smed ELAC2* KD at 0 hpa, most DEGs were concentrated in phagocytes (broad) and late epidermal progenitors. Underexpressed genes dominated and were enriched in core biosynthetic and gene-expression GO categories (notably translation and ribosome biogenesis, RNA processing/splicing/transport, transcription and gene expression regulation, mitochondrial function and energy metabolism, and broad metabolism).

At 16 hpa, enriched GO categories were more dominated by underexpressed genes, with DEGs particularly apparent in phagocytes (broad) and *PGRN*$^{+}$ parenchymal cells, alongside contributions from BWM (dorsal midline) and σ-neoblasts (broad-lineage). Underexpressed genes again represented the same GO categories (translation/ribosome, RNA processing, mitochondrial/energy metabolism, and metabolism).

At 24 hpa, underexpression was again the clearest global signature, with the corresponding deregulated genes concentrated in phagocytes (broad), late epidermal progenitors and σ-neoblasts (broad-lineage). These genes were enriched in GO clusters related to core gene expression and metabolism. Genes demonstrating overexpression upon *Smed ELAC2* KD were predominantly observed in phagocytes (broad) and late epidermal progenitors associated with remodeling and response GO categories (for example membrane structures/dynamics, vesicular transport/trafficking, signaling/response to stimuli).

At 72 hpa, transcriptional remodeling was the strongest and most diverse, with differentially expressed genes concentrated in phagocytes (broad) and *PGRN*$^{+}$ parenchymal cells, and secondary contributions from σ-neoblasts (broad-lineage) and late epidermal progenitors. The expression changes were bidirectional, with the balance of positive compared to negative shifts varying by lineage. Overexpressed genes were associated with GO categories consistent with altered cell interactions and tissue remodeling (including extracellular matrix and structures, membrane structures/dynamics, vesicular transport/organelle trafficking, signaling/response to stimuli, and related structural categories). Underexpressed genes were enriched in core biosynthetic and gene-expression-related GO categories (especially translation/ribosome,

RNA processing, mitochondrial/energy metabolism, and in some contexts chromatin/cell-cycle-associated categories).

4.4.2.4. Candidate CLASH-supported target fragments in single-cell DEGs

As described in **Section 4.3.3**, the analysis of CLASH dataset provided direct evidence that the 5′ tiRNA-Gly-GCC bound by SMEDWI-3 can bind specific mRNAs. The expression of transcripts predicted to be targets of 5′ tiRNA-Gly-GCC was therefore examined at single-cell resolution. In addition, changes in their expression relative to the levels of 5′ tiRNA-Gly-GCC were assessed. If the tiRNA contributed to repression *in vivo*, reduced tiRNA abundance would be expected to coincide with higher expression of at least some contacted transcripts, although indirect effects were also possible. These analyses were conducted using the 72 hpa dataset, corresponding to 3 dpa, to allow direct comparison with the bulk RNA-seq dataset obtained at the same time point for gene expression and sncRNA.

The set of gene models that contained CLASH-supported candidate target fragments for the 5′ tiRNA-Gly-GCC guide was intersected with the single-cell differential expression results (**Section 4.4.2.3**), which were computed per time point and cell population. For visualization, the log2 fold change values for the intersecting genes were plotted as a heatmap (**Figure 31**) across the cell population by time point combinations in which each gene was called differentially expressed.

Only a small subset of CLASH-linked candidates showed upregulation at 72 hpa, and the effect was strongly cell-type-specific. Three genes met the criterion of being upregulated at 72 hpa in at least one population, and these upshifts were concentrated in two lineages. Specifically, SMESG000036051.1 (annotated as E3 SUMO-protein transferase) and SMESG000056326.1 (Myo-inositol transporter) were upregulated in phagocytes (broad) at 72 hpa, while SMESG000015210.1 (annotated as retrotransposon) was upregulated in brain branch neurons at 72 hpa. Outside 72 hpa, additional candidate genes displayed cell-type-specific changes at earlier time points, but these time points do not align with the bulk sncRNA time window used to quantify 5’ tiRNA-Gly-GCC abundance. The putative target candidate from bulk RNA-seq SMESG000030475.1 (**Section 4.3.3**) showed no differences in gene expression at single-cell resolution.

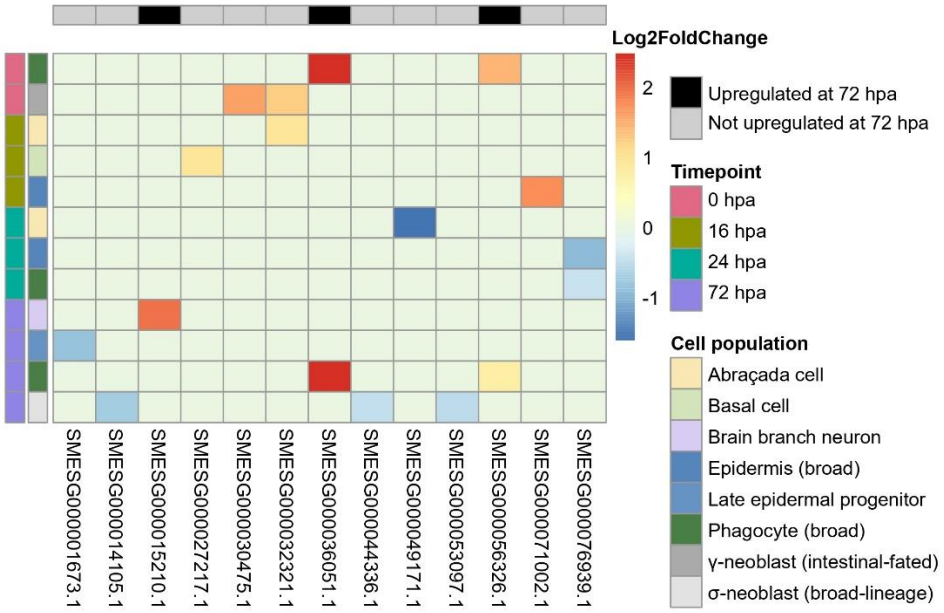

Figure 31. Cell-type and time point-specific differential expression among SMEDWI-3 CLASH-supported targets of 5′ tiRNA-Gly-GCC. Targets were defined from SMEDWI-3–5′ tiRNA-Gly-GCC CLASH chimeras and intersected with single-cell differential expression results. The heatmap shows average log2 fold-change (*Smed ELAC2* KD vs control) for CLASH-supported targets across cell clusters and time points after amputation. Columns with upregulated genes at 72 hpa are marked with black rectangle.

4.4.2.5. Impact of *Smed ELAC2* silencing on cell trajectories

Planarian regeneration depends on continuous cell-type transitions from neoblast into lineage-committed progenitors and differentiated tissues. Cluster-level comparisons could show which cell populations were affteced by *Smed ELAC2* silencing, but they did not directly address the dynamic changes at the level of cell lineages. For this reason, cell lineage trajectories were analyzed by quantifying how strong the particular cell populations of a lineage are connected, using the approach described in **Section 3.9.8**

To this end, the concept of connectivity was used to summarize lineage relationships. In this context, connectivity is a number that measures how often cells from two populations appear as close neighbors in gene expression space. If two populations have high connectivity, many cells from one population are among the nearest neighbors of cells from the other. This is consistent with a continuum of intermediate transcriptional states between them, which is expected when one state transitions into the other during differentiation or regeneration. Connectivity was quantified using PAGA, which summarizes the kNN graph as a weighted graph of cell population nodes, where each

edge weight approximates the fraction of kNN links that connect the two cell populations. In practical terms, higher edge weights indicate that many kNN relationships link the two groups, consistent with a continuum of transcriptional states between them.

To make comparisons meaningful across conditions within a given time point, all conditions (*Smed ELAC2* KD, GFP mock, WT) were evaluated on the same underlying cell-to-cell neighbor definition (**Section 3.9.8**). Essentially, the set of rules defining what constitutes a close neighbor remained unchanged, and the analysis was then repeated within each condition, using only the cells belonging to that condition. This prevented technical differences in neighborhood construction from being misinterpreted as biological differences.

Connectivity was estabiliched at two levels. First, a broad lineage level was considered, pooling together related cells into families. In this representation, the cell families were as follows: σ-neoblast, ν-neoblast, ζ-neoblast, γ-neoblast, eye neoblast, muscle neoblast, parenchymal neoblast, protonephridial neoblast, epidermal progenitor, eye progenitor, pharyngeal progenitor, pharyngeal neoblast, parenchymal progenitor, protonephridial progenitor, pigment progenitor, epidermis, eye, nervous system, pigment, pharynx, muscle, protonephridia, intestine, and parenchyma (**Figure 32**). Second, the same analysis was also performed using detailed cell-type labels (**Supplementary figures 1-4**), as presented in the atlas (**Figure 25**). .

To identify *Smed ELAC2* KD-specific effects, a Δ (delta) connectivity measure was used between *Smed ELAC2* KD and controls per edge and time point. For each edge, Δ can be positive or negative. A positive Δ means the connection was stronger in *Smed ELAC2* KD than expected from controls, while a negative Δ means it was weaker. Uncertainty was assessed by repeated resampling of cells, which provided confidence intervals for Δ (**Section 3.9.8**; **Supplementary Table S7**). For the family-level visualization, edges with very small absolute changes were ignored, specifically, changes with $|\Delta| < 0.003$ were treated as negligible for the summary plot (**Figure 32**).

At 0 hpa, the most prominent increases in Δ connectivity were restricted to a limited set of lineage relationships. The clearest strengthened connections were observed within the epidermal axis (epidermal progenitor–epidermis) and within the neural axis (neural progenitor–nervous system). Increased Δ connectivity was also observed for intestine-related links, most visibly between intestine and parenchyma. The strongest decreases of Δ connectivity were concentrated in the muscle module, where muscle was

connected weaker to parenchymal progenitor, and in connection between parenchymal progenitor and intestine (**Figure 32, Supplementary Figures 5-6**). This suggested that right after amputation the *Smed ELAC2* KD worm is reorganizing its regenerative program so that epidermal and neural lineages appear to progress more, while transitions that involve muscle and the parenchymal progenitor–intestine axis exhibit delayed progression.

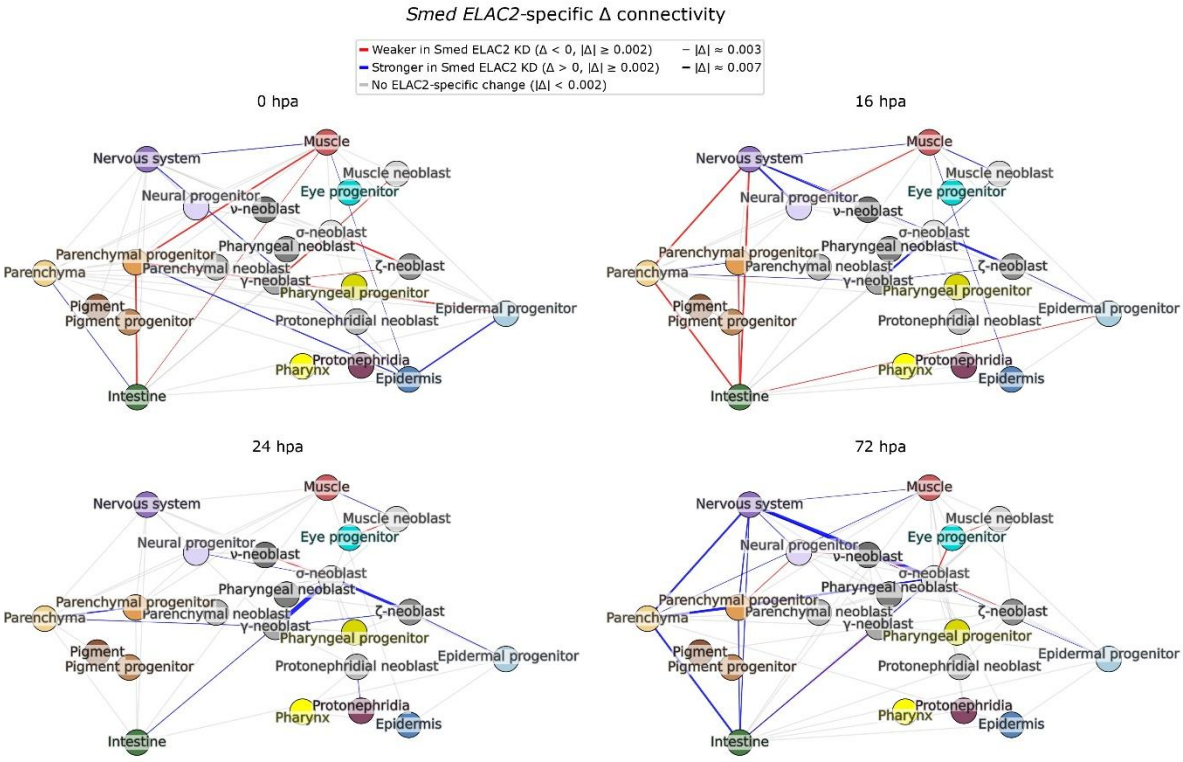

Figure 32. *Smed ELAC2*–specific change in family-level PAGA connectivity. Nodes represent cell families; edges represent PAGA connectivities. Colored edges show the *Smed-ELAC2*–specific delta connectivity, defined as Δ = connectivity (*Smed ELAC2* KD) – connectivity (controls), at the indicated time point. Red edges denote weaker connectivity in *Smed ELAC2* KD (Δ < 0), blue edges denote stronger connectivity in *Smed ELAC2* KD (Δ > 0); only edges with |Δ| ≥ 0.003 are highlighted. Edge width is proportional to |Δ|, and gray edges show the underlying family connectivity network for context.

At 16 hpa, a shift toward reduced intestine-centered continuity was observed. The strongest increases in Δ connectivity were again within the neural axis (neural neoblasts–nervous system, neural progenitor–nervous system) and within the neoblast axis (broad-

lineage neoblast–intestinal neoblast, broad-lineage neoblast–epidermal neoblast). In contrast, multiple intestine-associated edges were weakened, including links from intestine to parenchymal progenitor, parenchyma and nervous system, where the connectivity between the later and parenchyma was as well weakend (**Figure 32, Supplementary Figures 7-8**). This pattern may suggest that *Smed ELAC2* KD was associated with a progress within the stem cells and nervous lineages and with an impairment of gut and parenchymal differentiation.

At 24 hpa, the *Smed ELAC2*-specific changes were dominated by a strengthened connectivity, and the clearest single effect was a pronounced increase in Δ connectivity in neoblast axis (broad-lineage neoblast–intestinal neoblast, broad-lineage neoblast–epidermal neoblast). This edge was the strongest change at this time points, indicating that these two stem cell-associated families had shared substantially more intermediate transcriptional states in *Smed ELAC2* KD than in controls. Additional strengthened edges were observed in parenchymal axis (**Figure 32, Supplementary Figures 9-10**). This may indicate that worms upon *Smed ELAC2* KD have less clearly separated stem cell populations in the intestine and epidermis, with stem cells exhibiting stronger coordination within parenchymal tissue.

At 72 hpa, the strongest changes were again characterized by increased Δ connectivity. The most prominent increase was observed between broad-lineage neoblast and nervous system. Stronger connectivity was also observed between five families in all possible combinations (parenchyma, parenchymal progenitor, intestine, broad-lineage neoblast and nervous system). In contrast, only a small number of edges were weakened at this stage, most clearly involving eye progenitor (**Figure 32, Supplementary Figures 11-12**). This suggested that at 72 hpa, the *Smed ELAC2* KD worms showed an overall increase in coordination between stem cells, intestine, and nervous system, while the eye cell line differentiation was impaired.

Next, to relate the Δ-connectivity patterns to specific stages of regeneration, a subset of lineage links was then tracked across time points (**Figure 33**). These links (pair of cell families) were selected for visualization from the Δ-connectivity analysis as representative edges that showed clear *Smed ELAC2* KD-associated shifts within at least one time point (that is, non-negligible |Δ| and consistent direction relative to controls). In *Smed ELAC2* KD, σ-neoblast–eye progenitor connectivity was reduced, indicating a weaker association (weaker transcriptional continuity) with the eye-progenitor route. At

72 hpa, connectivity was increased for σ-neoblast–nervous system, σ-neoblast–intestine, and σ-neoblast–parenchyma, indicating strengthened late links between broad neoblast states and neural, intestinal, and parenchymal lineages. Edges involving parenchymal development, parenchymal progenitor–nervous system and parenchymal progenitor–parenchyma, displayed elevated connectivity at 72 hpa, indicating reinforced late connection between parenchymal progenitors and their connected lineages.

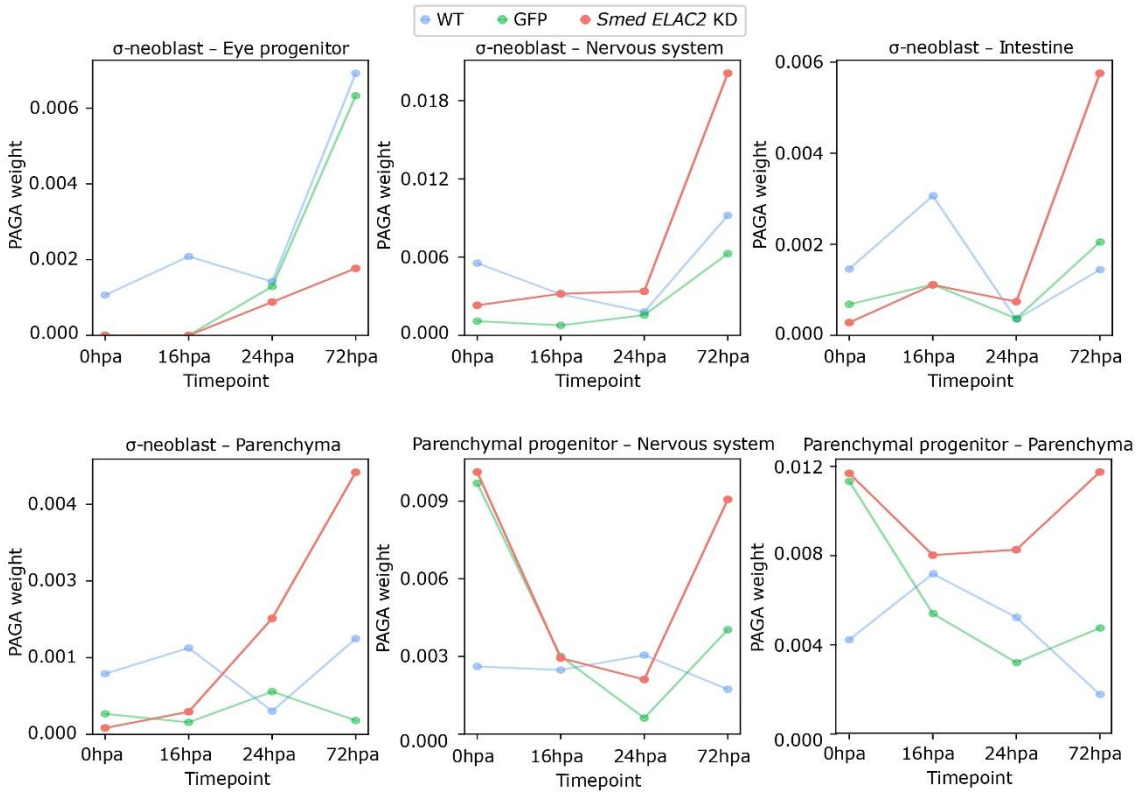

Figure 33. Time-resolved PAGA connectivity between selected lineages. Line plot showing the PAGA edge weight (family-level connectivity) between σ-neoblast–eye progenitor, σ-neoblast–nervous system, σ-neoblast–intestine, σ-neoblast–parenchyma, parenchymal progenitor–nervous system, and parenchymal progenitor–parenchyma, at 0, 16, 24, and 72 hpa for WT (blue), GFP mock (green), and *Smed ELAC2* KD (red). Higher values indicate stronger neighborhood continuity between the two lineages in the kNN manifold.

Overall, these analyses demonstrated that *Smed ELAC2* knockdown reshaped regeneration-stage-specific lineage continuity. Δ-connectivity identified which lineage transitions were selectively strengthened or weakened at each time point. Tracking the most affected links across 0–72 hpa showed that early effects were limited and axis-specific, whereas late (72 hpa) changes converged on broadly increased continuity

between σ-neoblasts and multiple lineages, with a relative weakening of the eye-associated route.

4.4.3. Integration of sncRNA analyses with cell-population context

4.4.3.1. sncRNA profiling in FACS-sorted cell populations

Because the single-cell transcriptome data captures polyadenylated mRNA, lineage-resolved sncRNA profiles could not be obtained directly at single-cell resolution. To provide some initial insight into a cell-population context for the sncRNA-seq, these molecules were therefore profiled in two major FACS-sorted populations: the stem-cell population (X1; neoblast-enriched) and the non-neoblast population (X2/Xins; differentiated and progenitor-enriched) at 72 hpa (**Supplementary Table S1**).

A low-input sncRNA library (GenXPro) was analyzed from the X1 and X2/Xins cells for WT, GFP mock, and *Smed ELAC2* KD worms as described in **Section 3.10**. Reads were first assigned to broad RNA classes to evaluate overall library composition (**Figure 34**). Because rRFs reads dominated the libraries, they were excluded prior to differential testing. sncRNAs were retained for downstream comparisons only if they had at least 10 total reads summed across all samples (**Supplementary Table S8**). Given the lower information content of the fractionated libraries compared with the bulk sncRNA-seq experiment (**Section 4.2.5**), tRF reads were summarized at the tRNA isoacceptors level rather than at the finer resolution of parental tRNA subregions.

The overall RNA-class composition indicated that the libraries were strongly dominated by rRF reads in all conditions and both fractions, with only a small residual fraction contributed by Metazoan SRP fragments, snoRFs, tRFs, miRNAs, and other minor classes (**Figure 34**). When the tRNA-Gly-GCC-assigned reads were additionally inspected at the sequence level, most of them were found to match the canonical 5′ tiRNA-Gly-GCC, reflecting the same molecule that was tracked in the whole-animal experiment.

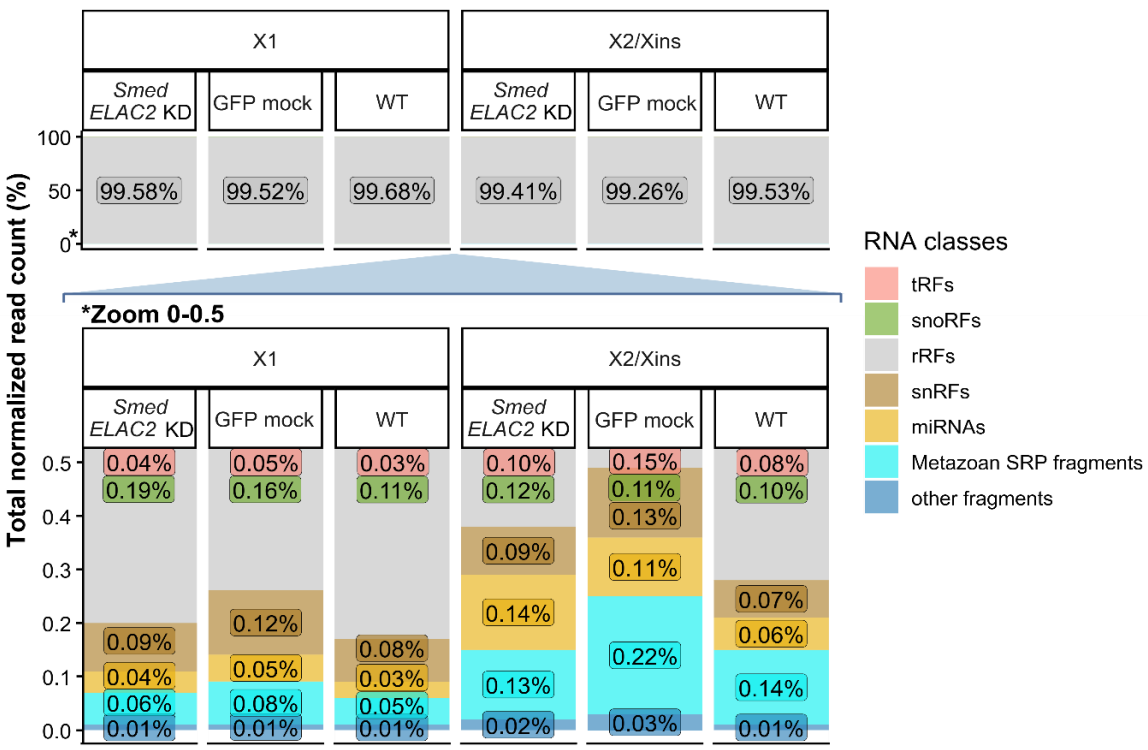

Figure 34. RNA-class composition of low-input sncRNA libraries from FACS-sorted fractions at 72 hpa. Stacked bar plots show the proportion of reads assigned to major RNA classes in GenXPro low-input sncRNA libraries prepared from the neoblast-enriched X1 fraction and the non-neoblast X2/Xins fraction under WT, GFP mock, and *Smed ELAC2* KD conditions (3 dpa, 72 hpa). The upper panel summarizes the full distribution and highlights that rRFs dominated the libraries. The lower panel shows a zoomed view of the remaining minor classes to visualize fraction- and condition-dependent differences.

Subsequently, a differential expression analysis was planned to identify sncRNAs exhibiting altered accumulation levels in the major cell populations (X1 or X2/Xins), in response to *Smed ELAC2* silencing. However, following the removal of rRF-derived reads, the remaining read counts were insufficient to obtain reliable results. Despite this limitation, the dataset was suitable for determining in which of the major cell populations (X1 or X2/Xins) the crucial sncRNAs were detected. For this purpose, samples derived from WT worms were used.

In total, 9 sncRNAs differed significantly between X1 and X2/Xins (**Figure 35A**). As expected, the canonical neoblast-associated miRNA let-7a was enriched in X1, whereas 5′ tiRNA-Gly-GCC showed the opposite pattern, being enriched in X2/Xins.

Additional differences included enrichment of several snRFs in X1 (U2, U4, U5, U3) and enrichment of selected sncRNAs in X2/Xins (including tRF Pro-CGG and miR-2153).

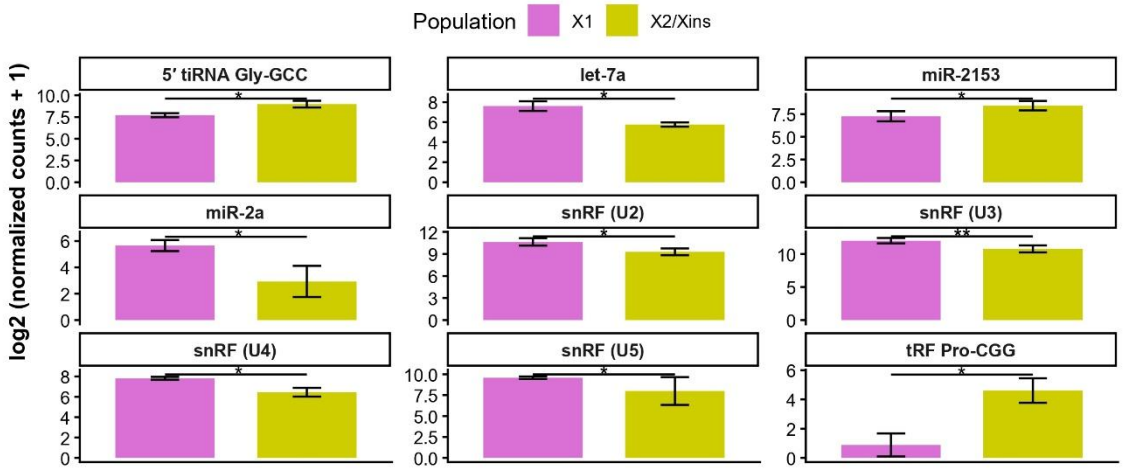

Figure 35. Enrichment of selected sncRNAs between neoblast (X1) and X2/Xins fractions in WT animals at 72 hpa. Bar plots show DESeq2-normalized abundance (log2 [normalized counts + 1]) for sncRNA species that differed significantly between X1 and X2/Xins fractions in WT. ncRNAs with significant baseline enrichment in WT where X1 was compared to X2/Xins after multiple-testing correction. Differential expression was tested with DESeq2. Error bars indicate dispersion across biological replicates. Significance codes: * padj < 0.05; ** padj < 0.01; *** padj < 0.001.

Overall, low-input sncRNA profiling of FACS-sorted cell populations revealed that 5′ tiRNA Gly-GCC was preferentially accumulated in X2/Xins, while the canonical neoblast-associated miRNA let-7a was enriched in X1, consistent with prior reports of let-7 enrichment in neoblast fractions (**Table 2**). Because the let-7a signal recapitulated the expected fraction specificity, it also served as an internal positive control for the sorting and low-input workflow, supporting the credibility of the observed X2/Xins enrichment of 5′ tiRNA-Gly-GCC.

4.4.3.2. scRNA-seq guided inference of sncRNA function

The final step was to link the sncRNA changes observed upon *Smed ELAC2* knockdown with gene expression alterations at the single-cell level. This required the development of a custom approach capable of integrating multiple layers of information, concerning: (i) the changes in abundance of a given sncRNA induced by *Smed ELAC2* silencing, (ii) the major cell populations in which this sncRNA is present, (iii) the

potential target transcripts of the sncRNA, and (iv) the cell types in which these targets are expressed. This integrative framework would allow for the formulation of hypotheses regarding the potential involvement of sncRNAs in regulating gene expression processes at single-cell resolution, even though it is not possible to directly characterize the sncRNA pool within individual cells. The approach based on an assumption that if a sncRNA was functional in a given population, its predicted target genes should show a coordinated expression shift consistent with repression or de-repression. Therefore, the scRNA-seq atlas was used for the evaluation of changes in gene expression, hypothetically exerted by the functional sncRNA.

All miRNAs, piRNAs and tRFs that were significantly affected by *Smed ELAC2* silencing in bulk RNA-seq (**Section 4.2.3.2**) were eligible for analysis (**Supplementary Table S9**). Among these, 5′ tiRNA-Gly-GCC (decreased accumulation upon *Smed ELAC2* KD) was selected for further characterization based on the preceding analyses, whereas 3 species with increased accumulation, miR-190a-3p, miR-9a-5p and representative piRNA, were chosen to facilitate validation of the adopted analytical workflow and the obtained results. As described in **Section 3.11**, candidate target sets were defined for each sncRNA using sequence-based interaction models (miRNA-like canonical targeting, 7-mer offset seed models, or an extended piRNA-like pairing model). For each cell, an activity score was computed by comparing the average expression of the predicted target set to expression-matched control genes. The score was then sign-flipped so that higher values corresponded to stronger inferred target repression. In this representation, values closer to the top of the scale indicated stronger predicted repression of targets (higher inferred sncRNA activity), whereas more negative values indicated relative target de-repression (lower inferred sncRNA activity). Importantly, this activity is a continuous quantity and does not imply a binary “active/inactive” state.

The workflow explicitly separated (i) target identification from (ii) cross-condition evaluation. Target sets were identified once from transcript sequences using the sncRNA sequence and the selected pairing model, and the same target set was then scored in all three genotypes (WT, GFP mock, and *Smed ELAC2* KD) at 72 hpa. Consistency with the sncRNA level changes expected base on the bulk sncRNA profiling (**Section 4.2.3**) was evaluated by comparing the population-level activity shift in *Smed ELAC2* KD versus WT (and verifying that GFP mock behaved similarly to WT). Only sncRNA–model–population combinations whose direction agreed with the expectation from bulk

differential accumulation of the sncRNA were treated as consistent and used for downstream visualization and interpretation. For visualization on UMAP, per-cell activity values were centered within each population relative to WT, using: $\Delta\ \text{Activity}_{\text{cell}} = \text{Activity}_{\text{cell}} - \text{median}\ (\text{Activity}_{\text{WT72 cells in the same population}})$. Under this definition, positive Δ Activity indicated stronger inferred repression than the WT median for that population, whereas negative Δ Activity indicated weaker inferred repression (relative de-repression) than the WT median.

Across 4 representative sncRNAs, the inferred activity patterns were highly cell-type selective and aligned with expected lineages (**Figure 36**). For 5′-tiRNA-Gly-GCC, 4 independent miRNA-like models highlighted *PGRN*$^{+}$ parenchymal cells, whereas the extended piRNA-like model additionally nominated phagocytes (broad). (**Figure 36A**). These results aligned with earlier observations that the *Smed ELAC2* KD-dependent response was concentrated in these immune-associated and parenchymal populations at the 72 hpa (**Figure 30**), whereas low-input sncRNA profiling described in **Section 4.4.3.1** showed enrichment in X2/Xins in WT worms, supporting these populations as plausible sites of tiRNA-linked regulation. Within these populations, the *Smed ELAC2* KD showed a shift toward lower activity values relative to controls (more blue), consistent with reduced repression of the predicted target set.

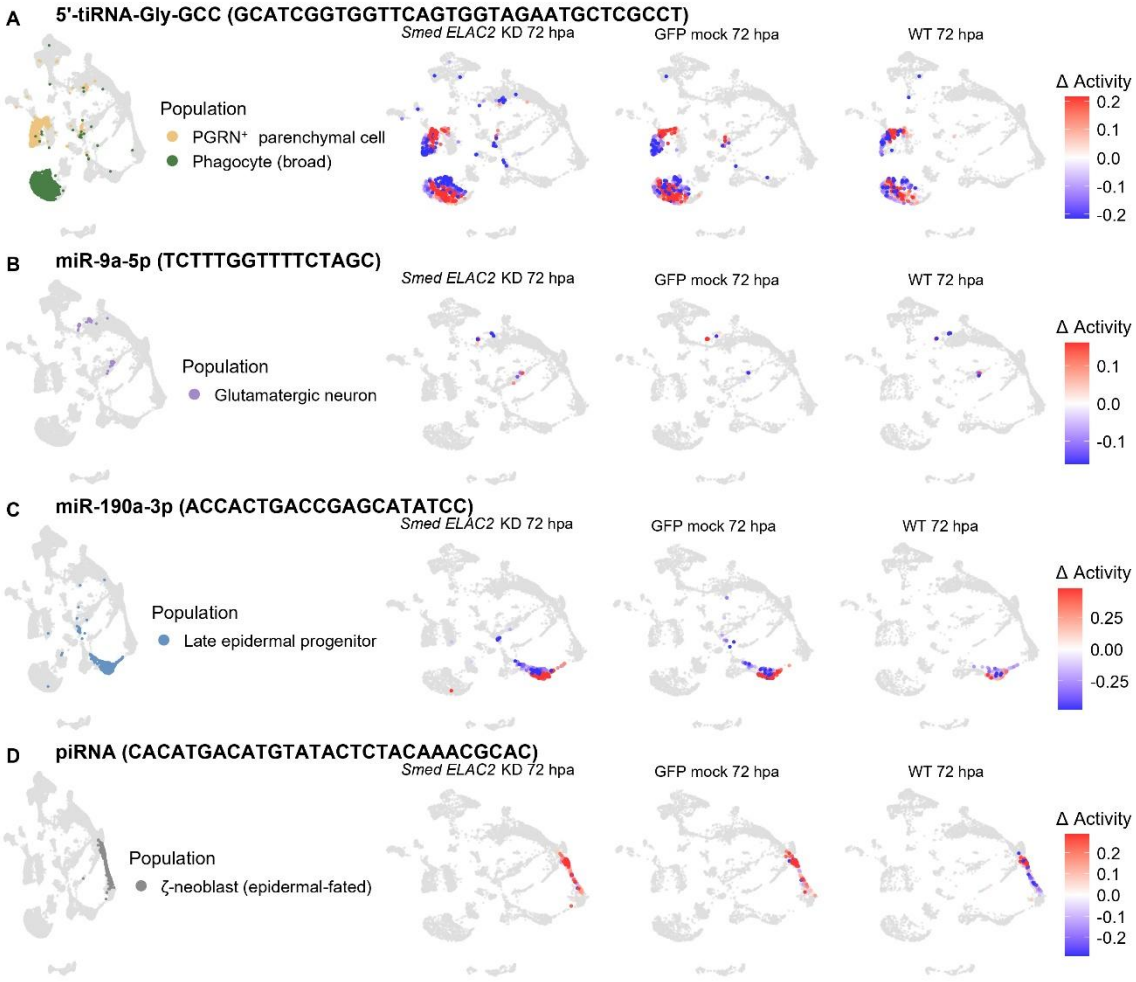

Figure 36. scRNA-seq guided inference of sncRNA activity at 72 hpa. (A) 5′-tiRNA-Gly-GCC, with activity in phagocytes (broad) and *PGRN*$^{+}$ parenchymal cells. (B) sme-miR-9a-5p showing activity in glutamatergic neurons. (C) sme-miR-190a-3p with activity in late epidermal progenitors. (D) A representative piRNA, showing activity in ζ-neoblasts (epidermal-fated). In each panel, the left UMAP highlights the relevant population(s) in their atlas colors, and the right UMAPs show the inferred activity score in the same cells across *Smed ELAC2* KD, GFP mock, and WT at 72 hpa. The activity scale is not intended for direct comparison across sncRNAs because each sncRNA uses a distinct target set and dynamic range; within each panel, higher values indicate stronger inferred repression of predicted targets, while lower (more negative) values indicate relative de-repression.

For sme-miR-9a-5p, only the off1_7mer model yielded a consistent activity pattern, which was most apparent in glutamatergic neurons (more red in *Smed ELAC2* KD in **Figure 36B**). The activity range was narrow compared with the other examples, which is expected when targets are modestly expressed and when the predicted effect size

is small within a single mature neuronal subtype. For sme-miR-190a-3p, also scored using the off1_7mer model, activity was concentrated in late epidermal progenitors (more red in *Smed ELAC2* KD in **Figure 36C**). For the miRNAs, the inferred footprints were broadly consistent with the cell-type enrichments summarized in **Table 2**, with miR-9a-5p mapping to a differentiated neuronal context that matches its predicted enrichment, while miR-190a-3p showed an epidermal-progenitor footprint that could be compatible with its reported X1 enrichment. Finally, the selected piRNA sequence, scored using the piRNA_extended model, showed a restricted signal within the ζ-neoblast (epidermal-fated) population (more red in *Smed ELAC2* KD in **Figure 36D**). Because piRNA function is tightly linked to stem and progenitor cells in many systems, the confinement of the inferred activity to a neoblast subclass was consistent with the expected cellular context for piRNA-like regulation, validating the developed approach.

Together, these results contributed to establishing and optimizing the process of inferring regulatory effects associated with sncRNA in scRNA-seq data by integrating sncRNA profiling from FACS-sorted fractions (X1 and X2/Xins) with mRNA expression patterns at single-cell resolution in the 72 hpa atlas. This approach was then validated on known, diverse examples by recovering activity signals restricted to specific cell types consistent with known biological contexts, including miR-9a-5p in glutamatergic neurons, miR-190a-3p in late epidermal progenitors, and a representative piRNA in the neoblast subclass. When applied to 5′-tiRNA-Gly-GCC, multiple seed-based targeting models converged on a consistent set of candidate populations, with the strongest and most reproducible activity patterns detected in *PGRN*$^{+}$ parenchymal cells and phagocytes. Thus, these specific cellular locations could be prioritized for further functional analyses of 5′-tiRNA-Gly-GCC.

5. DISCUSSION

sncRNAs are central regulators of gene expression, but the interpretation of sncRNA data depends strongly on bioinformatic assumptions. Widely used discovery and annotation workflows for miRNAs and other sncRNAs were largely developed and tested in organisms with well-defined genome resources and relatively stable ncRNA catalogues (Friedländer et al. 2012; Axtell 2013). In contrast, in *S. mediterranea* the available resources are improving but remain less standardized across annotations and databases (Rozanski et al. 2019; L. Guo et al. 2022). Therefore, the first task of this dissertation was to generate a genome and ncRNA annotation that provided a consistent framework for downstream analyses. The second and third tasks then focused on the impact of *Smed ELAC2* silencing on the sncRNA pool and bulk gene expression, and on the functional implications of selected sncRNA. This step was important because abundant sncRNAs include RNA fragments that do not follow miRNA biogenesis rules, so classification and target modeling must account for multi-copy origins and non-canonical mechanisms (Haussecker et al. 2010; Telonis et al. 2016). Finally, the last task addressed cellular context. To connect bulk sncRNA changes to cell lineages, indirect inference was employed. It was based on testing whether the gene sets predicted as targets for sncRNA showed repression-like behavior in specific cell populations and time points (M. M. Nielsen and Pedersen 2021; J. Li et al. 2025; Maji et al. 2025; Olgun et al. 2022). For these reasons, the Discussion follows the task order: annotation framework (**Section 5.1**), bulk effects of *Smed ELAC2* silencing (**Section 5.2**), *in silico* functional analysis of selected ncRNAs (**Section 5.3**), lineage-resolved inference using single-cell data (**Section 5.4**) and summary of the developed tools (**Section 5.5**).

5.1. Reference resources in non-classical models

The reference annotation defines which reads can be assigned to genes, which isoforms exist, and which loci are considered real. If the annotation contains fragmented gene models, chimeric transcripts, or missing loci, then downstream analyses inherit these errors and can turn technical artifacts into false biological insights. This is particularly relevant for *S. mediterranea*, where limited and heterogeneous genomic resources, strain differences, and incomplete standardization across data portals made annotation choice unusually consequential (Zaremba et al. 2025). For this reason, the first task of this research was to develop an annotation workflow and apply it to *S. mediterranea* genome.

A detailed discussion of the aspects related to this topic is presented in the publication incorporated as part of this doctoral thesis (Zaremba et al. 2025). In summary, this approach combined short- and long-read sequencing, integrated reference-based and *de novo* assembly, utilized deep learning tool predictions to improve splicing site detection, and applied functional annotations to filter low-confidence transcripts. The workflow also included automatic artifact detection and clear quality control annotations but still required individual verification of annotated loci during curation, which ensured transparency of the process but meant that some subjectivity remained unavoidable (Zaremba et al. 2025).

This reference has been sucesfully applied to accomplish further steps of the doctoral research.

5.2. sncRNAs in *Schmidtea mediterranea*

After establishing an improved reference framework and annotation, the second task of this research was to use bulk RNA-seq to quantify how *Smed ELAC2* knockdown alters the sncRNA pool. Planarians employ several sncRNA–mediated regulatory systems, most prominently miRNAs and PIWI-associated piRNAs, and additionally generate diverse RFs derived from multiple RNA classess.

5.2.1. Dysregulation of a specific group of sncRNAs upon *Smed ELAC2* knockdown

In bulk sncRNA data, the overall class composition remained broadly stable between *Smed ELAC2* KD and control animals across regeneration time points, which clearly demonstrated lack of a global failure of the core RNA silencing regulatory pathways. Instead, a selective remodeling of specific sncRNA subclasses, primarily miRNAs and tRFs was observed.

5.2.1.1. miRNAs

For miRNAs (**Sections 4.2.3.2–4.2.3.3**), the accumulation changes were limited to a small set and were strongest at 3 dpa. Deregulated miRNAs were annotated using published FACS-based enrichment labels (X1, neoblast-enriched; X2/Xins, non-neoblast-enriched), which are widely used in planarian miRNA studies but are not fully consistent across publications for all miRNAs (Lu et al. 2009; Sasidharan et al. 2013).

No changes in miRNA accumulation levels were detected at either 3 dpa or 5 dpa that could be attributed to the specific cell populations for which these miRNAs are characteristic. This pattern suggested that the miRNA response did not correspond to a failure in a single cellular fraction.

Currently, the literature provides relatively detailed information on planarian miRNAs; however, these data are largely limited to identifying the specific cell populations in which individual miRNAs are enriched. Consequently, there is a lack of information that would provide mechanistic insight into the specific roles of miRNAs. The same applies to miRNAs that displayed altered accumulation. For example based on the enrichment of miR-36 and miR-71 family members in neoblasts, and of miR-9a, miR-96a/96b, and miR-125b in differentiated cells, their involvement in regenerative processes has been postulated (Sasidharan et al. 2013). Given that miRNAs are evolutionarily conserved, insights into their potential roles can be inferred from studies conducted in other animal models. One of such regulators, increased upon *Smed ELAC2* silencing, is miR-9a, a representative of a broadly conserved neurogenic miRNA family. In vertebrate neural progenitors, miR-9 participates in a double-negative feedback with Hes1, a transcriptional repressor, and modulates Hes1 mRNA stability, thereby influencing the transition from proliferative progenitors toward differentiation (Bonev et al. 2012). Another example is miR-71a. The miR-71 family was strongly linked to stress physiology and survival in *C. elegans*, where mir-71 acts in the nervous system to promote longevity and stress adaptation (Boulias and Horvitz 2012; X. Zhang et al. 2011). miR-96a/miR-96b could be related to the vertebrate miR-183/96/182 sensory cluster, in which miR-96 seed mutations cause progressive hearing loss and the cluster contributes to sensory tissue differentiation and maintenance (Mencía et al. 2009; Lewis et al. 2009). miR-125b is a potent regulator of stem and progenitor programs in vertebrate hematopoiesis, where it expands hematopoietic stem cells and can drive leukemogenic phenotypes when overexpressed (Ooi et al. 2010; Bousquet et al. 2010). Finally, miR-8 family members map to the conserved miR-8/miR-200 axis implicated in growth and insulin/PI3K-linked signaling, with demonstrated conservation between *Drosophila* miR-8 and human miR-200 pathway wiring (Hyun et al. 2009).

The mechanism by which *Smed ELAC2* silencing influences miRNA levels remains unclear, although it is likely to be indirect. Alterations in miRNA accumulation may either reflect disruptions in cell lineage differentiation during delayed regeneration,

or serve as a compensatory response to such disturbances. This issue requires additional investigation.

5.2.1.2. tRFs

Because tRFs constituted the most abundant class of differentially accumulated sncRNAs after *Smed ELAC2* knockdown, they were characterized is detail. The analsis focused on tRF subtypes showed a clear directional asymmetry: fragments that increased tended to originate from the 3′ region of parental tRNAs, with 3′-tRFs being prominent among increased species, whereas fragments that decreased were enriched for 5′ tiRNAs. These observations clearly demonstrated that the changes were specific.

Decreased 5′ tiRNAs were predominantly derived from a restricted set of parental tRNAs, including tRNA-Gly-GCC, tRNA-Glu-TTC, and tRNA-Asp-GTC, and decreased internal fragments were also observed for tRNA-Gly-GCC and tRNA-Asp-GTC (**Sections 4.2.3.2–4.2.3.4**). The strongest and most reproducible example was the depletion of 5′-tiRNA-Gly-GCC, which was independently validated by qRT-PCR (**Figure 14B**). These isoacceptor-specific changes were consistent with prior literature data showing that Gly- and Asp-derived tRFs were selectively produced and could participate in defined regulatory complexes. In human breast cancer cells, hypoxia-induced tiRNAs derived from tRNA-Gly and tRNA-Asp were reported to bind the RNA-binding protein YBX1 and to repress pro-metastatic transcripts (Goodarzi et al. 2015). In addition, 28-nt sperm 5′-tiRNA derived from tRNA-Gly-GCC were reported to repress MERVL-associated gene expression in early mouse embryos (Sharma et al. 2016), and subsequent mechanistic work showed that same tiRNA regulated noncoding RNA production and modulated chromatin accessibility (Boskovic et al. 2020). Together, these reports supported the idea that specific tRNA isoacceptors, including Gly and Asp, could give rise to regulatory fragments with well-defined functions. Beyond the sperm and early embryo context, the 5′ tiRNA-Gly-GCC has also been implicated in regulation of translation. In mammals, angiogenin-induced 5′-tiRNAs derived from tRNA-Cys and tRNA-Ala, particularly those bearing a 5′-terminal oligoguanine (5′-TOG) motif that enables G-quadruplex–like assembly, were shown to act as bona fide regulatory molecules by inhibiting translation initiation through displacement of eIF4G and eIF4A and by promoting stress granule assembly (Ivanov et al. 2011; Emara et al. 2010).

Currently the identity of the nuclease that generated planarian 5′-tiRNAs is unclear. In vertebrates, angiogenin cleaved mature tRNAs in the anticodon loop under stress, producing tiRNAs, but sequence-homology surveys demonstrated that *S. mediterranea* lacked an angiogenin homolog, implying that planarian 5′-tiRNAs were produced by an angiogenin-independent route (Lakshmanan et al. 2021). Evidence from mammals also supported the broader concept that angiogenin was not the only enzyme capable of producing tiRNAs, because angiogenin knockout left most stress-induced tiRNAs unchanged, implying the existence of other RNases that generated tiRNAs (Su et al. 2019). In non-vertebrate systems that lacked RNase A family members, RNase T2 family enzymes were implicated in anticodon-loop cleavage of mature tRNAs, for example the RNase T2 member Rny1p in yeast and multiple RNase T2 enzymes in *Tetrahymena* (Thompson and Parker 2009; Andersen and Collins 2012), and plant RNase T2 enzymes (RNS) were shown to be major contributors to tRF biogenesis (Megel et al. 2019). Importantly, a recent mechanistic study identified an angiogenin-independent route, involving cleavage mediated by IRE1α that selectively generated the 5′ tiRNA-Gly-GCC during endoplasmic reticulum stress. The resulting fragment was shown to exert regulatory function via binding specific hnRNP proteins (Jin et al. 2024). Considering the substantial decrease in the level of 5' tiRNA-Gly-GCC after *Smed ELAC2* knockdown, our research team undertook an investigation into the biogenesis of this sncRNA, specifically examining whether Smed ELAC2 may be directly involved in this process. This hypothesis is supported by earlier findings implicating ELAC2 in the generation of 5′ tiRNA-Gln-CTG (Choi et al. 2021). These studies are currently ongoing.

A remaining question was why depletion of a tRNA-processing enzyme did not produce a uniform decrease of all tRFs. The key point is that *Smed ELAC2* KD was expected to affect only part of the tRF landscape directly. ELAC2 in metazoans cleaved the 3′ trailers of pre-tRNAs and generated the pre-tRNA trailer-derived class (tRF-1), but most other tRF subclasses, including tiRNAs, were generated from mature or near-mature tRNAs by other endonucleases and in a context-dependent manner (Y. S. Lee et al. 2009; Ivanov et al. 2011). Obviously, *Smed ELAC2* knockdown could have been expected to affect the availability of mature tRNAs that give rise to tRFs. This is because ELAC2 is essential for the maturation of all tRNAs. However, it is plausible that, due to the interplay of various factors – such as gene expression levels of specific tRNA isotypes or differences in their stability – *Smed ELAC2* silencing selectively impairs the maturation

of only certain isotypes, such as tRNA-Gly-GCC. This notion is supported by NGS data from ELAC2-deficient mouse models, showing reduced levels of only a few mature tRNAs, including tRNA-Gly-GCC (Siira et al. 2018). It now remains to be determined whether the reduced level of 5′tiRNA-Gly-GCC results from a selective decrease in tRNA-Gly-GCC abundance or from a disruption in the biogenesis of 5′tiRNA-Gly-GCC itself.

5.2.2. Bulk transcriptome context for interpreting *Smed ELAC2*-dependent sncRNA changes

The bulk mRNA-seq results were then included as a necessary context layer for evaluating the functional potential of sncRNA (**Sections 4.2.4.1–4.2.4.3**). It was assumed that if sncRNAs exert regulatory functions, changes in their abundance would be accompanied by alterations in the expression levels of their target genes. Accordingly, the most common mechanism of sncRNA action - namely post-transcriptional control of mRNAs - was taken into consideration. However, higher-order regulation that reshaped transcriptional states indirectly cannot be excluded. Differential gene expression analysis after *Smed ELAC2* KD revealed a modest number of DEGs. This low count was not interpreted as an absence of transcriptional impact, but as a consequence of a deliberately conservative strategy designed to minimize false positives and to attribute changes specifically to *Smed ELAC2* silencing rather than non-specific RNA, injection-related or background effects. Genes were required to differ between *Smed ELAC2* KD and WT while remaining stable between the two control groups, and additional abundance filtering and an effect-size threshold were applied (**Section 3.4**). A practical implication was that additional, weaker, or cell-restricted transcriptional effects could have been present but remained undetected because whole-worm averaging diluted lineage-specific signals and the analysis prioritized specificity over sensitivity. At 3 dpa, the transcriptional response was dominated by genes associated with regulatory or tissue remodeling processes among the upregulated set, while the downregulated set was enriched for metabolism- and organellar-linked programs and included reduced immune/defense-response GO terms, including positive regulation of phagocytosis (**Figure 15**). By 5 dpa, additional downregulated genes were involved in membrane transport and ion-channel regulation, including calcium-channel regulation (**Figure 15**). This stage dependence was consistent with established regeneration course in planarians,

where early generic injury programs transition into later tissue- and function-specific programs (Wurtzel et al. 2015).

The majority of DEGs identified upon *Smed ELAC2* knockdown, had been previously reported as DEGs in other whole-animal experiments. The strongest direction-consistent overlaps occurred under *yki* RNAi (Lin and Pearson 2017) and *hippo* RNAi (De Sousa et al. 2018), with additional overlaps for *COE* (Cowles et al. 2014), *CHD4* (Tu, Cheng, Tk Vu, et al. 2015), *mex3-1* (Tu, Cheng, Tk Vu, et al. 2015), *p53* (Tu, Cheng, Tk Vu, et al. 2015), and *zfp-1* (Cheng et al. 2018; van Wolfswinkel et al. 2014) silencing. Both *hippo* and *yki* encode planarian homologs of proteins that function in the conserved Hippo signaling pathway (De Sousa et al. 2018; Lin and Pearson 2017). Accordingly, *hippo* encodes a core kinase that maintains cell-cycle control and stable differentiated cell state, while *yki* encodes the downstream transcriptional effector that normally restricts early injury responses (Lin and Pearson 2014; De Sousa et al. 2018). This indicated that the DEG list described in this thesis included genes that were deregulated under diverse conditions that affected tissue turnover and lineage-state balance.

Mapping the same bulk DEGs onto public single-cell atlases provided first insight into which cell populations are affected by *Smed ELAC2* silencing. Particularly at 3 dpa the changes were restricted to a very narrow cell subset – upregulated genes were characteristic for parenchymal-associated populations, whereas the downregulated genes mapped to intestinal populations (**Figure 16**). Planarian parenchyma has long been proposed to contribute to blastema formation by providing a structural and contact-guidance environment for migrating regenerative cells (Hori 1991). Recent spatial data showed that fate specification in neoblasts is intermingled and that specialized neoblast states occur in close physical association with parenchymal cells (C. Park et al. 2023). In parallel, the intestine is the major metabolic and absorptive organ and comprises transcriptionally distinct cell types with regional specialization, making it a plausible source of the metabolic and organellar signatures detected in bulk DEGs (Forsthoefel et al. 2020).

5.3. Functional analysis of 5′-tiRNA-Gly-GCC

Next task of this research was to conduct *in silico* functional analysis of selected sncRNAa, and the efforts were focused on the most strongly depleted species, 5′-tiRNA-Gly-GCC. Several hypothetical mechanisms underlying its activity were examined. This

approach was justified because tRFs had been reported to act through multiple effectors. In human cells, subsets of tRFs associated with Argonaute proteins and mediated sequence-directed repression in a miRNA-like manner (Kumar et al. 2014; Kuscu et al. 2018). In other systems, tRFs have been shown to associate with PIWI proteins (S. P. Keam et al. 2014). In planarians specifically, 5′-tiRNAs were reported to interact with all three SMEDWI proteins and AGO1 was implicated in maintaining an i-tRF population (Lakshmanan et al. 2021).

5.3.1. 5′-tiRNA-Gly-GCC as a Smed ELAC2 guide RNA (TRUE-like model)

The first evaluated hypothesis was a TRUE-like mechanism, in which a sncRNA guides cytosolic ELAC2 to cleave a complementary RNA substrate when the guide–target complex adopts an appropriate tRNA-like geometry (Elbarbary, Takaku, Uchiumi, et al. 2009). This principle underlies ELAC2-utilizing efficacious (TRUE) gene silencing (Nashimoto 2022), a technology developed in mammalian systems, where short artificial sgRNAs, including 5′-tiRNA-like guides, can redirect ELAC2 to cleave chosen RNAs in living cells. Beyond cell culture applications (Elbarbary, Takaku, Uchiumi, et al. 2009; Elbarbary, Takaku, Tamura, et al. 2009; Habu 2005), TRUE gene silencing has also been demonstrated in an *in vivo* context: Nakashima and colleagues reported inhibition of gene expression in the livers of postnatal mice using an only 7-nt sgRNA, supporting feasibility of sgRNA-guided ELAC2 repression in a whole-animal setting (Nakashima et al. 2007). Because ELAC is broadly conserved and occur across bacteria, archaea, and eukaryotes, it was reasonable to test whether an analogous guide-dependent cleavage mode might also be operational in planarians, even though most functional TRUE gene silencing demonstrations have been performed in mammalian systems (Schiffer 2002). Given that Smed ELAC2 itself is the catalytic factor in this pathway, *Smed ELAC2* KD would be expected to make the entire route less efficient, which would tend to reduce cleavage-dependent downregulation of targets. Therefore, in the *Smed ELAC2* KD condition, the most plausible transcript-level signature of a TRUE-like pathway would have been target derepression (higher target abundance), because both reduced enzyme abundance and reduced guide abundance would have acted in the same direction, limiting effective cleavage (Elbarbary, Takaku, Uchiumi, et al. 2009; Brzezniak et al. 2011). No targets that were both geometrically plausible and consistent with the knockdown-responsive transcriptome signal were found (**Section 4.3.1**). Importantly, this observation did not

imply that TRUE-like cleavage could never occur in planarians; rather, it indicated that the data did not support it as the notable driver of the transcriptomic changes occurring under *Smed ELAC2* KD.

5.3.2. miRNA-like repression model

The second model treated 5′-tiRNA-Gly-GCC as a sequence-directed repressor of specific mRNAs. This model was described as "miRNA-like" because it used miRNA-style pairing rules and prediction tools to generate candidates, not because 5′-tiRNA-Gly-GCC was assumed to be a canonical miRNA. Canonical animal miRNAs are typically ~22 nt (Bartel 2018), whereas tiRNAs are commonly ~30–35 nt (Lakshmanan et al. 2021). Despite this size discrepancy, miRNA-like repression remained biologically plausible for tRNA-derived RNAs. For ~30–35 nt tiRNAs, the best-established mode of action is Argonaute-independent, involving inhibition of translation initiation and stress granule biology rather than RISC-like targeting (Ivanov et al. 2011). However, there are some well-documented cases where Argonaute-family proteins bind guide RNAs longer than canonical ~22-nt miRNAs. *C. elegans* Argonautes ALG-3 and ALG-4 acted with 26G-RNAs (26-nt sncRNAs) in spermatogenesis, and loss of ALG-3/4 eliminated this 26G-RNA subclass (Conine et al. 2010). Short excised introns of ~80–100 nt, termed agotrons, were found to be associated with and stabilized by Argonaute proteins in the cytoplasm, with long (≥30-nt) reads prominent in Ago2 HITS-CLIP, and some agotrons capable of seed-like repression (Hansen et al. 2016). In addition, there is evidence that at least some tiRNAs can interact with Argonaute-family proteins: in an EBV-immortalized B lymphoblastic cell model, a lactate-induced 5′-tiRNA-His-GUG was reported to bind AGO2, while 5′-tiRNA-Gly-GCC did not bind AGO2 or AGO1 (Mo et al. 2020). Another mechanistic possibility compatible with a 32-nt guide is that longer sncRNA precursors can be loaded and then trimmed to a mature length within Argonaute-centered pathways, as shown for Dicer-independent priRNAs/siRNAs in fission yeast (Marasovic et al. 2013), which would imply that a longer tiRNA could, in principle, serve as a precursor to a shorter, AGO-compatible effector species.

The predicted binding site for the selected candidate lay outside the 3′UTR. This also did not invalidate miRNA-like regulation, because miRNA binding sites in coding sequences had been documented and were shown to mediate regulatory effects, although on average they caused weaker repression than 3′UTR sites (Fang and Rajewsky 2011)

The identity of the validated target also supported biological plausibility. Protein tyrosine phosphatases are a large family of enzymes involved in the regulation of key cellular processes, including cell division, migration and differentiation (Tautz et al. 2013; Tonks 2006). In planarians, disruption of a non-receptor protein tyrosine phosphatase (Djptpn11) was reported to interfere with regeneration and to affect wound response and Wnt pathway gene expression (Q. Wang et al. 2022). Since gene annotation in planarians relies on homology-based approaches, current data are insufficient to determine precisely which PTP is encoded by SMESG000048842.1. Functional characterization of this protein will require dedicated studies. Nonetheless, given the known roles of PTPs in controlling cellular signaling and fate decisions, it is plausible that the protein is involved in the regulation of regeneration.

5.3.3. PIWI-centered, piRNA-like engagement model

A third hypothesis considered that 5′-tiRNA-Gly-GCC entered PIWI complexes and acted through PIWI-guided targeting rules. This hypothesis was plausible because in planarians SMEDWI proteins are central to neoblast biology, and planarian piRNAs are typically ~32 nt, which matches the length of 5′-tiRNA-Gly-GCC (Reddien et al. 2005; Palakodeti et al. 2008). The three planarian SMEDWI paralogs display distinct functional and mechanistic characteristics. SMEDWI-2 and SMEDWI-3, but not SMEDWI-1, were shown to be required for piRNA biogenesis and regenerative processes (Palakodeti et al. 2008). They also differ in their localization, both with respect to cell populations and subcellular compartments. Accordingly, SMEDWI-2 was found to be enriched in nuclei of neoblasts and differentiated cells, while SMEDWI-1 and SMEDWI-3 were predominantly cytoplasmic and enriched in neoblasts (I. V. Kim et al. 2019). Their diverse distribution determines their different functions. Accordingly, the nuclear SMEDWI-2 is involved in transposon control through chromatin remodelling. Depletion of this protein caused transposon derepression and decrease in the transcription of lineage-specific genes, disrupting the differentiation process (D. Li et al. 2021). SMEDWI-1 was reported to contribute to transcript quality control by associating with ribosomes during the pioneer round of translation and processing poorly translated transcripts into piRNAs (Allikka Parambil et al. 2024). SMEDWI-3 cooperates with SMEDWI-2 to degrade active transposons. In addition, it is involved in mRNA surveillance. Depending on the degree of complementarity between the target mRNA and

SMEDWI-3-bound sncRNA, the targets are either sliced or bound without degradation (I. V. Kim et al. 2019).

The results described in this dissertation demonstrated that 5′-tiRNA-Gly-GCC was present in pulldown pools for all three SMEDWI proteins, with the strongest enrichment in SMEDWI-2 (**Section 4.3.3**). This finding supported the plausibility of PIWI-complex entry and was consistent with the independent report that planarian 5′-tiRNAs interacted with all SMEDWI proteins (Lakshmanan et al. 2021). Unfortunately, the potential significance of 5′ tiRNA-Gly-GCC for the SMEDWI-2-dependent regulation could not be assessed based on the available NGS data. Considering that SMEDWI-2 is directly targeted to repetitive regions in the genome and is implicated in H3K9 methylation (D. Li et al. 2021), dedicated studies, focused on epigenetics rather than transcriptomics or genomics, are required to dissect this mechanism. Notably, the involvement of 5′ tiRNA-Gly-GCC in this process is plausible, as tRFs, typically shorter than tiRNAs, were reported to participate in chromatine remodeling, primarily via incorporation into PIWI regulatory pathway (J. Park et al. 2020). Using the SMEDWI-3 CLASH datasets, direct evidence was provided that 5′-tiRNA-Gly-GCC interacted with RNA targets *in vivo*. Most of them did not diplay changes in abundance upon depletion of 5′-tiRNA-Gly-GCC. The absence of a relevant degradome footprint argued against a dominant SMEDWI-3 slicing mechanism of the targets. This outcome remained compatible with alternative PIWI-centered scenarios, including SMEDWI-3 binding without degradation (I. V. Kim et al. 2019).

In summary, the identified putative mechanisms of 5′ tiRNA-Gly-GCC action are consistent with the broader context of what is known about the roles of tRFs. These findings will provide a foundation for further experimental functional analyses. Currently, ongoing studies in our research group aim to determine whether alterations in 5′tiRNA levels directly contribute to the observed regeneration delay phenotype.

5.4. Cell-lineage consequences of *Smed ELAC2* knockdown and sncRNA-associated regulation

The final step of this research was to place the *Smed ELAC2* KD-dependent sncRNA changes into a cell-lineage context, because regeneration phenotypes emerge from coordinated behavior of specific cell populations. Last task therefore focused on optimizing sncRNA-seq bioinformatic tools and integrating them with a curated scRNA-

seq atlas to identify which lineages were most affected by *Smed ELAC2* knockdown and where sncRNA-associated regulatory signals, including those linked to 5′-tiRNA-Gly-GCC, were most plausibly deployed.

5.4.1. A curated 72 hpa single-cell atlas as a reference resource

scRNA-seq atlases have become essential in planarian biology because they provide an explicit map from transcriptional states to cell identities, including stem cells, progenitors, and differentiated tissues. Widely used *S. mediterranea* atlases established broad coverage of adult cell types and marker sets and demonstrated that transcriptome-based clustering can recover both major tissues and finer subtypes in a whole animal (Fincher et al. 2018; Plass et al. 2018). Earlier work already showed that wound responses are partly generic and partly cell-type-specific, implying that regeneration experiments affect multiple state changes and therefore obviously benefit from cell-resolved analysis (Wurtzel et al. 2015). More recent resources have expanded the atlas concept by adding specialized layers, such as transcription factor programs that define fate trajectories (H. O. King et al. 2024), spatiotemporal maps of regeneration (Cui et al. 2023), and multimodal integration of transcriptomes with chromatin accessibility to infer gene-regulatory logic during differentiation (Pérez-Posada et al. 2025). Together, these studies make a clear methodological point: planarian regeneration is now routinely studied through atlases because they reduce whole-worm averaging and allow perturbations to be interpreted at the level at which biological processes actually occur, namely lineages and cell states (García-Castro and Solana 2022).

The atlas developed in this study was optimized for a different goal than reference atlases built from unperturbed animals. Its purpose was to support direct comparison of WT, GFP mock, and *Smed ELAC2* KD animals during regeneration in a shared coordinate system. Cross-study atlas reuse is powerful but can be confounded by technical differences in dissociation, sequencing depth, and sampling, which can introduce shifts that resemble biological effects. Therefore, by building a dedicated joint atlas in which all samples of interest were processed under a single analytical framework, the interpretational ambiguities were reduced, making *Smed ELAC2* silencing effects easier to attribute.

A core novelty of this atlas was the annotation strategy. In planarians, annotation is not hindered by a shortage of markers, but by redundancy of markers, partial overlap

between transitional and terminal states, and inconsistent naming across literature and datasets (García-Castro and Solana 2022). To address this, a curated marker reference was assembled in the present study (199 unique marker genes across 10 broad lineages and 60 populations) and coupled to multi-method consensus annotation workflow. Eight complementary methods were applied at the parent-cluster level with disagreement treated as evidence of heterogeneity that justified targeted subclustering without forcing inconsistent labeling. This approach well aligned with how the planarian research commmunity has used atlas-derived marker logic to resolve complex lineages such as epidermal differentiation and neural progenitors (Wurtzel et al. 2017; Molinaro and Pearson 2016). In practical terms, the novelty was a reproducible annotation pipeline that made it easier to assign stable labels in a non-classical model organism.

In terms of biological coverage, the final atlas resolved 57 coherent populations spanning the cell types most relevant to regeneration-stage interpretation. These comprised multiple neoblast populations (σ, ζ, γ, ν and lineage-biased subclasses), epidermal progenitors and mature epidermal layers including multiciliated states, intestinal cell types including phagocytes and goblet cells, protonephridial flame and tubule lineages with precursors, ECM-producing muscle and pole or PCG muscle populations, glia, diverse neuronal classes (including cholinergic, glutamatergic, catecholaminergic, neuropeptidergic, and sensory subtypes), eye lineages, and multiple parenchymal subtypes. This breadth matches what was expected from modern planarian atlases, where parenchymal diversity and immune-like or phagocytic populations were now considered functionally important components and not just residual space (Fincher et al. 2018; García-Castro and Solana 2022). The presence of populations such as *CTSL2*-associated (phagocytic abraçada and *PGRN*$^{+}$ parenchymal cells) was also notable because these cells just recently have been proposed to form a local niche-like microenvironment for pluripotent stem cells (Settles et al. 2025). While that specific niche model has remained under active development in the literature, it illustrated why capturing and consistently annotating parenchymal and phagocyte-adjacent states was increasingly important for mechanistic regeneration studies.

For the planarian community, the atlas contributes in two complementary ways. First, it provided a regeneration-stage, perturbation-aware reference in which cell labels were curated and consistent across sample types, which was directly usable for interpreting knockdown effects that alter lineage composition. Second, it provided a

practical annotation toolkit: a curated marker reference and a consensus labeling workflow that can be reused to annotate new planarian scRNA-seq datasets with less manual ambiguity, especially in lineages where marker redundancy and transitional states complicate single-method assignments.

5.4.2. Cell- and lineage-specific consequences of *Smed ELAC2* knockdown

To interpret lineage-level consequences of *Smed ELAC2* knockdown, next step was to establish where *Smed ELAC2* mRNA was detected in the scRNA-seq atlas (**Section 4.4.2.1; Figures 26–27**). In WT animals, the *Smed ELAC2*$^{+}$ pool was dominated across regeneration time points by neoblast-associated populations and closely related progenitor states, with σ-neoblasts contributing the largest share and additional contributions from lineage-biased neoblast subclasses. This pattern matched the fundamental characteristics of planarians, where neoblasts are the major proliferative compartment responsible for tissue turnover and regeneration, and are transcriptionally and functionally heterogeneous (Wagner et al. 2011; van Wolfswinkel et al. 2014). Consistent with this, an odds-ratio analysis relative to σ-neoblasts indicated that most differentiated populations had lower *Smed ELAC2* detection frequency than σ-neoblasts, whereas only a limited set of progenitor or regeneration-associated populations reached or exceeded the σ-neoblast baseline at specific stages, including *PGRN*$^{+}$ parenchymal neoblasts at 0 hpa (**Section 4.4.2.1**; **Figure 27**). This distribution suggested that *Smed ELAC2* KD most plausibly exerted direct effects in stem and early progenitor states, while changes in differentiated populations were expected to be indirect consequences of altered lineage output, altered tissue remodeling, or altered niche-like interactions. Cell-population composition analyses supported this distinction (**Section 4.4.2.2**; **Figure 28**), as several of the affected populations were not those that typically display the highest *Smed ELAC2* expression. Moreover, there was a temporal discrepancy between this expression and cell type composition: although *Smed ELAC2* expression exceeded baseline σ -neoblasts in *PGRN*$^{+}$ parenchymal neoblasts at 0 hpa, this population did not yet show a detectable change in abundance at 0 hpa, while the population of *PGRN*$^{+}$ parenchymal cells underwent relative expansion at 72 hpa in the knockdown. This was consistent with an early intracellular disturbance in the stem cell compartment, which spread over time, causing a change or maintenance of the cell line, and thus triggered delayed, indirect effects in related populations.

The intestine-adjacent and parenchymal changes correspond well with current models in which regeneration depends on coordinated responses across stem cell layers and includes specialized, regeneration-activated states in post-mitotic tissues. A construction of a large regeneration atlas allowed to identify transient regeneration-activated cell states (TRACS) in muscle, epidermis, and intestine and to show that genes expressed in these states were required for proper regeneration (Benham-Pyle et al. 2021). In addition, laser-capture microdissection RNA-seq showed that the planarian intestine is transcriptionally regionalized (medial versus lateral branches), particularly among goblet-cells, and identified intestine-enriched regulators linked to gut cell maintenance and regeneration (Forsthoefel et al. 2020). In the present dataset, the persistence of intestinal phagocyte shifts (0 hpa and 16 hpa) together with the overrepresentation of *CTSL2*$^{+}$ phagocytic-like parenchymal populations, namely abraçada cells (24 hpa and 72 hpa) and *PGRN*$^{+}$ parenchymal cells (72 hpa), suggested that *Smed ELAC2* KD altered tissue remodeling and/or immune-like supportive transcriptional programs spanning both the gut epithelium and the surrounding compartment. This intestine-versus-parenchyma split is consistent with the bulk RNA-seq cell-type mapping at 3 dpa (**Section 4.2.4.3** and **5.2.2**), where upregulated genes were characteristic of parenchymal cells. A recent study proposed that *CTSL2*$^{+}$ phagocytic cells physically associate with stem cells and can regulate stem cell function and regeneration through innexin-dependent mechanisms, highlighting a plausible niche-like coupling between cathepsin-positive phagocytic cells and neoblast behavior (Settles et al. 2025). Under this model, the *CTSL2*$^{+}$ expansion observed here in abraçada cells and in *CTSL2*-coexpressing *PGRN*$^{+}$ parenchymal cells could have been either compensatory (responding to altered stem or progenitor physiology) or causal (reflecting a shifted niche environment) effect, that feeds back onto lineage allocation, thereby linking a neoblast-centered *Smed ELAC2* expression pattern to reduced eye-progenitor connectivity and delayed anterior regeneration.

The epidermal-related changes were consistent with established properties of the planarian epidermal lineage. ζ-neoblasts and downstream epidermal progenitors represent a defined trajectory from neoblasts to epidermal differentiation, with multiple intermediate states and spatially patterned progenitor identities (van Wolfswinkel et al. 2014; Tu, Cheng, Tk Vu, et al. 2015; Wurtzel et al. 2017). Therefore, an early increase in the epidermis fraction together with altered epidermal-progenitor connectivity (**Section 4.4.2.5**; **Figure 32**) can be interpreted as a shift in the balance between epidermal

progenitor production, differentiation. Importantly, the PAGA analysis identified increased connectivity within the epidermal axis (epidermal progenitor–epidermis) already at 0 hpa, which is exactly the type of lineage-continuity change expected if intermediate transcriptional states accumulated or if transitions along the lineage became more gradual (**Section 4.4.2.5**; **Figure 32**). In PAGA, each edge weight between two clusters is a connectivity statistic computed from the single-cell kNN graph. Specifically, PAGA connectivity is defined as the ratio of the observed number of inter-edges between two populations (epidermal progenitor and epidermis) to the number expected under a random null model (Wolf et al. 2019). When this ratio increases, it means that more cells in one population have nearest neighbors in the other population than would be expected by chance, so the two populations behave more like a continuous region of the data manifold rather than two well-separated groups (Wolf et al. 2019). This type of effect is compatible with a primary perturbation in RNA metabolism in proliferative cells that then propagates into differentiation dynamics.

The early reduction of posterior pole/PCG muscles in *Smed ELAC2* KD provided an additional route by which indirect effects could emerge. Planarian body-wall muscle expresses PCGs that provide instructive patterning information for regeneration and tissue maintenance (Witchley et al. 2013; Reddien 2018). A transient change in the abundance or state of PCG-expressing muscle near the posterior pole could therefore alter the signaling environment experienced by neoblasts and their progeny, without requiring high *Smed ELAC2* expression in muscle itself. In this framework, composition shifts in epidermis and later changes in lineage connectivity could reflect altered positional cues, altered wound-response programs, or both.

Finally, the nervous system-associated changes were notable because mature neural populations did not appear among the *Smed ELAC2*$^{+}$cells, yet neuropeptidergic neurons expanded at 0 hpa and 72 hpa and neural-axis connectivity increased across time points (**Sections 4.4.2.2** and **4.4.2.5**; **Figures 28** and **32–33**). This discrepancy supported an indirect interpretation: *Smed ELAC2* KD most likely altered neural lineage flow and intermediate states rather than acting primarily within terminal neurons. Planarian neurogenesis proceeds through defined neural stem/progenitor populations (Molinaro and Pearson 2016). In addition, planarian peptidergic signaling can operate via long-range volumetric transmission and can also act in a neuroendocrine-like manner; therefore, changes in neuropeptidergic neuron abundance could plausibly influence regeneration

physiology at the whole-animal level, not only within the nervous system (Bray et al. 2024; Collins et al. 2010). In this context, the increased σ-neoblast–nervous system connectivity at 72 hpa (**Section 4.4.2.5**; **Figure 33**) was consistent with strengthened late continuity between broad neoblast states and neural programs, while the reduced σ-neoblast–eye progenitor connectivity suggested a relative weakening of the eye-progenitor route. Because eye regeneration depends on neoblast-derived eye progenitors and the transcription factor *ovo* is required early in eye regeneration (Lapan and Reddien 2012), these connectivity shifts provided a lineage-level context in which delayed eye regeneration could arise from altered allocation of neoblast production and altered coupling among parenchymal, intestinal, and neural compartments and simply from a single tissue-specific transcriptional defect.

5.4.3. Modeling 5′-tiRNA-Gly-GCC activity in single cells

Standard droplet scRNA-seq captures mainly polyadenylated mRNAs and therefore could not quantify sncRNAs such as 5′-tiRNA-Gly-GCC at single-cell resolution. Dedicated single-cell sncRNA methods exist, but they require specialized library preparation and have been lower-throughput than mainstream mRNA scRNA-seq. For example, Small-seq3 enabled ligation-based capture and molecular counting of sncRNAs from up to 192 libraries prepared from individual cells, including miRNAs and tRFs (Hagemann-Jensen et al. 2018). A large method-comparison study evaluated 19 single-cell sncRNA protocols and demonstrated feasibility across cell lines and circulating tumor cells, but also highlighted that sensitivity and the number of miRNAs detected per cell depend strongly on the protocol (Hücker et al. 2021). More recent improvements addressed efficiency and throughput constraints, such as PSCSR-seq for parallel single-cell sncRNA profiling (J. Li et al. 2023), while comparative analyses showed substantial heterogeneity in isomiR composition across single-cell sncRNA datasets, emphasizing that isomiR-aware processing can be important for interpretation (Smith and Hutvagner 2022). Consistent with this rapid method development, a high-throughput droplet-based approach for simultaneous miRNA and mRNA capture has been proposed in the droplet-small-seq project, but this capability is not yet a routine component of atlas-scale workflows (European Commission CORDIS, droplet-small-seq project https://cordis.europa.eu/project/id/101030265).

Multiple computational approaches have been proposed to extract information about miRNA regulation from scRNA-seq, but most were designed around canonical miRNAs and vertebrate-grade transcript annotations. Motif-enrichment activity methods infered miRNA activity by combining seed-site information with the expectation that miRNAs reduced target mRNA levels, as implemented for example in miReact and miTEA-HiRes (M. M. Nielsen and Pedersen 2021; Herbst et al. 2025). Machine-learning approaches such as miRSCAPE instead infered miRNA expression from mRNA profiles using models trained on paired datasets (Olgun et al. 2022). In parallel, dedicated toolkits, such as SingmiR, have improved preprocessing and quality control for single-cell miRNA sequencing itself (Engel et al. 2024). However, recent reviews highlighted that those methods can be informative but depend strongly on accurate UTR annotations, well-defined miRNA repertoires, and matched training resources, which are often not available outside a small set of model organisms (Maji et al. 2025). These assumptions did not fit the specific constraints of the present study, where the molecule of interest was a 32-nt 5′ tiRNA and where it was necessary to consider many alternative pairing rules sets in addition to relying on a single canonical set of rules for miRNAs.

The integration strategy developed in this doctoral research addressed that gap by combining two complementary approaches: (i) bulk and low-input sncRNA profiling, which established in which major FACS fractions sncRNAs were enriched, and (ii) scRNA-seq-based approach, which localized where the downstream mRNA consequences of a given sncRNA were most consistent with repression or de-repression across annotated cell populations at 72 hpa. This is a central methodological achievement of the thesis because it converted sncRNA changes measured in sorted populations into lineage-resolved hypotheses that could be evaluated using the atlas. Several interaction models were assessed (canonical miRNA-like, offset seed miRNA-like variants, and an extended piRNA-like model), and only sncRNA–model–population combinations that were directionally consistent with the bulk-derived expectation were carried considered.

Two observations supported the practical validity of this approach. First, low-input fraction profiling at 72 hpa showed strong cell-state specificity of dominant sncRNAs: 5′-tiRNA-Gly-GCC was enriched in X2/Xins, whereas let-7a was enriched in X1, with the latter observation being consistent with the literature (**Table 2**). Second, the scRNA-seq-based approach recovered cell-type-restricted signals for three sncRNAs

used as methodological controls (miR-9a-5p, miR-190a-3p, and a representative piRNA, **Section 4.4.3.2**).

Across multiple seed-based miRNA-like and piRNA-like models, the most reproducible inferred 5′-tiRNA-Gly-GCC activity localized to phagocytes (broad) and *PGRN*$^{+}$ parenchymal cells at 72 hpa. In these populations, *Smed ELAC2* knockdown shifted the activity distribution toward lower values relative to WT, consistent with de-repression of the predicted target sets when the tiRNA was depleted (**Section 4.4.3.2**; **Figure 36**).

These cellular assignments also helped reconcile an apparent discrepancy between where *Smed ELAC2* expression was enriched and where tiRNA-linked consequences were most visible. First, *Smed ELAC2* expression was concentrated in neoblast-associated compartments and early progenitors, including *PGRN*$^{+}$ parenchymal neoblasts (**Section 5.4.2**). It can be thus hypothesized that silencing of *Smed ELAC2* disrupted the 5' tiRNA-Gly-GCC biogenesis in these cells. Second, while the activity metric did not measure tiRNA presence per se, it measured whether a predicted set of target set behaved as if they were collectively repressed. Therefore, a strong activity signal was expected to occur where three elements coincide: the tiRNA is available, suitable effector proteins are present, and a sufficient fraction of predicted targets is expressed at detectable levels. Even if *Smed ELAC2*-dependent 5' tiRNA-Gly-GCC generation occurred upstream in broad neoblasts or *PGRN*$^{+}$ parenchymal neoblasts, the most visible repression footprint could emerge downstream in *PGRN*$^{+}$ parenchymal populations (and possibly phagocytes) if those populations provided the right combination of targets and cofactors for functional implementation. Third, the temporal ordering supported a coherent lineage-coupling interpretation. *Smed ELAC2* enrichment in *PGRN*$^{+}$ parenchymal (phagocytic type) neoblasts at early stages provided a plausible upstream entry point for the perturbation, while the later expansion and altered regulatory signatures in mature *PGRN*$^{+}$ parenchymal cells and phagocytes provided a plausible downstream shift. In this stepwise model, *Smed ELAC2* knockdown altered the 5' tiRNA-Gly-GCC biogenesis in stem cells, and the functional consequences became most detectable in niche-associated parenchymal and immune-like states where target repression at the mRNA level could occur. This framing emphasized consistency across layers of evidence, because it connected *Smed ELAC2* localization, tiRNA depletion, and lineage-restricted target-set behavior into a single mechanistic narrative.

PGRN^{+} parenchymal and abraçada cells were extremely similar and adjacent in transcriptional space. Both populations consistently showed *PGRN* and *CTSL2* expression among their common markers, and the difference largely reflected relative detection frequency: *PGRN* was detected in a similar fraction of cells across both groups (means ~81–82%; **Supplementary Table S5**) during manual verification of final cell annotation labeling, whereas *CTSL2* was detected in a smaller fraction of cells in *PGRN*$^{+}$ parenchymal clusters (mean ~48%) than in abraçada clusters (mean ~82%), consistent with a *CTSL2*-centered definition for abraçada cells (Settles et al. 2025). In addition, phagocyte clusters were distinguished by near-uniform *PSAP* detection (mean ~87%) together with additional co-markers that were less frequent or less dominant in the *PGRN*$^{+}$ parenchymal and abraçada cells. Thus, *PSAP* primarily reported a shared lysosome-rich program across these populations, but phagocytes in this dataset were best defined by *PSAP* in combination with the broader phagocyte marker set rather than by *PSAP* alone.This interpretation was mechanistically plausible because, in other animals, both progranulin (encoded by *PGRN*) and prosaposin (encoded by *PSAP*) are lysosome-directed proteins that physically interact and support lysosomal trafficking, and progranulin can be proteolytically processed in lysosomes by cathepsin L (Zhou et al. 2015; C. W. Lee et al. 2017).

As an additional layer, intersecting the CLASH-supported candidate target list for the 5′-tiRNA-Gly-GCC guide with the time point resolved scRNA-seq differential expression results yielded only a small number of candidates that changed concordantly at 72 hpa. These signals were strongly lineage-restricted: SMESG000036051.1 (annotated E3 SUMO-protein transferase) and SMESG000056326.1 (annotated myo-inositol transporter) were upregulated specifically in phagocytes (broad), while SMESG000015210.1 (annotated retrotransposon) was upregulated in brain branch neurons. SUMO pathway function is regeneration-relevant in planarians, because *Ubc9* (the SUMO E2) is required for normal neoblast proliferation and blastema formation, and its knockdown impairs regeneration (Thiruvalluvan et al. 2018). The myo-inositol transporter candidate is not itself a validated regeneration gene, but its upregulation is consistent with the established requirement of phosphoinositide signaling for regeneration and tissue maintenance, as shown by regeneration defects upon PI3K inhibition or knockdown and by the essential role of Akt signaling for regeneration-associated proliferation and cell death dynamics (H. Zheng et al. 2021; Peiris et al. 2016).

Lastly, the regulation of retrotransposon reflects the well-established requirement for PIWI-mediated transposon silencing in planarian stem cell differentiation and regeneration (Reddien et al. 2005; Shibata et al. 2016).

A recurring pattern in regeneration biology is that regulatory signals act at sites where stem cells interact with supportive neighbor tissues and immune-like cells (Witchley et al. 2013; Wurtzel et al. 2015; Peiris et al. 2014; Settles et al. 2025). In this doctoral research, the strongest inferred 5′-tiRNA-Gly-GCC activity was shown to be localized in *PGRN*$^{+}$ parenchymal cells (expressing also *CTSL2*) and phagocytes, so the key mechanistic question is whether the effects of tiRNA-mediated regulation (or tiRNA itself) can be propagated to adjacent neoblasts and nearby intestinal lineages via intercellular transfer or secondary signaling. One plausible route is direct cell–to-cell contact signaling via gap-junction-like channels. In vertebrates, sncRNAs including miRNAs can move through connexin-based gap junctions and remain functional in recipient cells (Zong et al. 2016). Planarians lack connexins, but they do express innexins, which form gap junctions and are required for regeneration and stem cell homeostasis (Nogi and Levin 2005; Oviedo and Levin 2007; Oviedo et al. 2010). Importantly, the *CTSL2*$^{+}$ cells niche itself was reported to require two innexin-encoding genes for pluripotent stem cell function and regeneration (Settles et al. 2025). However, because the demonstrated RNA permeability was for connexin channels, not innexins, innexin dependence here should be treated as evidence for contact-dependent coupling in general, not as proof of direct tiRNA passage; the relevant transmitted entities could instead be ions, metabolites, or electrical states that reprogram RNA regulation downstream. A second plausible route is extracellular vesicle (EV) transport. Exosomes originate from endosomal multivesicular bodies and their secretion competes with lysosomal degradation, making endolysosomal trafficking a determinant of whether vesicle cargo is released or destroyed (Théry et al. 2002; Colombo et al. 2014). Planarians produce abundant EVs, and EVs can carry sncRNAs with organism-wide regulatory consequences, including systemic RNAi, with EV release modulated by stress and regeneration-relevant contexts (Sasidharan et al. 2026; Avalos et al. 2024). In *C. elegans*, EV cargo profiling indicates that EVs harbor RNA and include components consistent with extracellular RNA communication (Nikonorova et al. 2022). In *Drosophila*, exosome-like vesicles have been implicated as carriers for morphogens (for example Hedgehog-containing exovesicles required for graded signaling and secretion machinery

dependence), and Wingless can be present on exosome-like vesicles in cultured cells(Gradilla et al. 2014; Beckett et al. 2013). Therefore, stable tiRNAs can be deposited in EVs and delivered to recipient cells, supporting a general model in which tiRNAs can act as intercellular signals under defined conditions (Gámbaro et al. 2020). Although no apparent activity of 5’ tiRNA-Gly-GCC was detected outside of *PGRN*$^{+}$ parenchymal cells and phagocytes, arguing against transport of the tiRNA itself, propagation of the effects of its action remains plausible. The strong lysosomal signature shared by *PGRN*/*CTSL2*/*PSAP*-marked populations suggests high flux through late endosome–lysosome pathways that can support both EV uptake/processing and, depending on MVB routing, EV secretion (Zhou et al. 2015; C. W. Lee et al. 2017).

In summary, the multilayer results of the present study allowed to propose a stepwise model in which the initial perturbation of sncRNA pool is introduced by silencing the *Smed ELAC2* gene, while the most visible regulatory impact is observed in the intestinal and parenchymal populations (**Figure 37**).

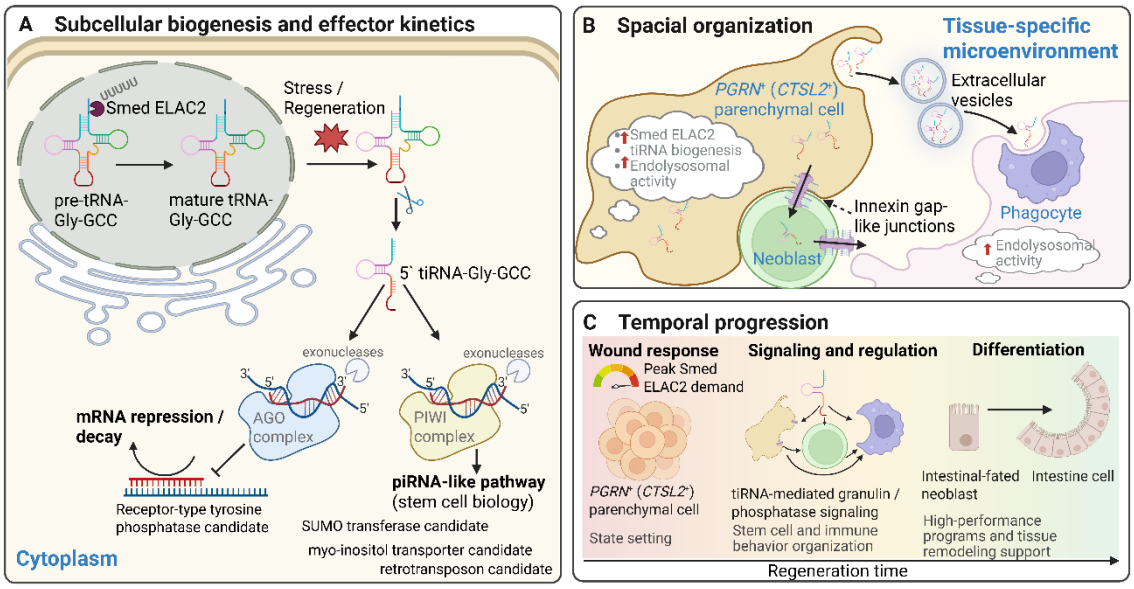

Figure 37. Mechanistic, spatial, and temporal model of Smed ELAC2 control of 5′-tiRNA-Gly-GCC during planarian regeneration. (A) Subcellular model in which Smed ELAC2 supports pre-tRNA-Gly-GCC maturation, enabling production or stability of 5′-tiRNA-Gly-GCC and subsequent engagement with Argonaute or PIWI effector complexes to promote post-transcriptional repression of candidate targets. (B) Tissue microenvironment model highlighting *PGRN*$^{+}$ (co-expressed with *CTSL2*) parenchymal cells as niche-like neighbors of neoblasts and as plausible sites for tiRNA production and exchange, with transfer hypotheses via direct-contact channels and extracellular vesicles, and with endolysosomal activity as a shared

feature of *PGRN*/*CTSL2*/*PSAP*-marked populations. (C) Temporal model linking early wound response, intermediate regulatory signaling, and late differentiation to shifting Smed ELAC2 demand and tiRNA-mediated regulation during regeneration. Created with BioRender.com

First, *Smed ELAC2* expression was concentrated in neoblast-associated populations, including early enrichment in *PGRN*$^{+}$ parenchymal neoblasts, defining a plausible entry point where the knockdown could shift regeneration programs. Second, the knockdown of *Smed ELAC2* was accompanied by a strong depletion of 5′-tiRNA-Gly-GCC in bulk profiling, suggesting the specific sncRNA change that must be propagated through the system. Third, at the single-cell level, the strongest inferred 5′-tiRNA-Gly-GCC activity was localized to *PGRN*$^{+}$ parenchymal cells (expressing also *CTSL2*) and intestinal phagocytes, and the latter one becoming compositionally overrepresented late in regeneration, consistent with a niche-like response that reorganizes tissue remodeling and immune-like support functions. Finally, because regenerative control commonly concentrates at interfaces where stem cells interact with supportive and immune-like neighbors, impairments in these *CTSL2*$^{+}$ and phagocyte populations provided a plausible route by which a stem-cell-biased perturbation can bias lineage outputs indirectly, aligning with the observed transient delay in anterior regeneration (eye formation) after *Smed ELAC2* KD.

5.5. Developed bioinformatic resources and analytical tools

To directly address the resource gaps summarized in **Section 1.4.6**, a cohesive set of reference resources and analysis tools was developed in the framework of this doctoral research. They enabled reproducible, planarian-specific interpretation of both long RNA-seq and sncRNA-seq, and the integration of bulk RNA-Seq results with single-cell atlases. These developments ensured coordinate consistency across datasets, improved functional interpretability for planarian data, and enabled explicit, mechanism-aware hypothesis testing for the depleted 5′-tiRNA-Gly-GCC in the context of *Smed ELAC2* knockdown. The major resources and tools developed in this study are publicly available or will be made accessible upon publication of the results to facilitate their use by the scientific community.

First the foundational task of creating a comprehensive genome annotation tool was successfully completed. Applying this workflow generated and curated the schMed_Pn annotation, which was deposited in GEO (GSE294569) and provided an

updated coordinate-consistent basis for subsequent bulk and single-cell analyses, including locus-level refinement of the *Smed ELAC2* model for accurate quantification and sequence-dependent experimental design. Next, a sequence-resolved sncRNA processing strategy was implemented that integrates alignments across reference layers and assigns RNA-biotype labels by rule-based parsing to enable robust class-level summaries and differential accumulation testing. Furthermore, dedicated tools for the functional analysis of sncRNAs were developed. They include planarian-appropriate Gene Ontology coverage for enrichment analysis and a workflow for sncRNA target identification. Lastly, sncRNA-informed analysis was optimized by building a reproducible single-cell annotation and inference stack that compensates for the core limitation that scRNA-seq measures polyadenylated mRNAs, not sncRNAs. A curated marker reference was assembled and harmonized into the schMed_Pn identifier space, providing a standardized bridge between published marker nomenclature and the transcript models used in this project. Building on this, an automated cell-type assignment workflow was implemented as a dedicated script that runs eight complementary scoring strategies and combines them into consensus population labels, while flagging disagreement as potential heterogeneity rather than forcing overly confident calls. The marker reference and annotation tool will be released as a shiny application. To connect bulk-identified sncRNA shifts to lineage-specific transcriptional programs, a custom scRNA-seq guided activity-inference framework was then developed that integrates (i) the direction of sncRNA change under knockdown, (ii) candidate target sets from sequence models, and (iii) population-resolved expression shifts expected under repression or de-repression, yielding continuous per-cell activity scores that can be compared across genotypes within populations. These tools are available in https://norreanea.github.io/AZ-PhD-THESIS-Supplementary/.

6. CONCLUSIONS

The results obtained in this doctoral dissertation led to the following conclusions:

1. The developed genome annotation for *S. mediterranea* displays high structural consistency and core-gene completeness, and integrates an expanded ncRNA catalogue suitable for sncRNA-focused studies.
2. *Smed ELAC2* knockdown triggers a selective, time-dependent response in individual sncRNAs, with miRNA and tRF being the most affected classes.
3. The most prominent miRNA changes at 3 dpa involve increased: miR-190a-3p, miR-2153 precursor, miR-36a-3p, miR-71a-5p, miR-96a-5p, miR-96b-5p, and miR-9a-5p, and decreased: miR-125b-5p and miR-8b-3p. Deregulated miRNAs do not preferentially originate from any specific cell population (X1, X2, or Xins).
4. The most strongly decreased tRFs are: 5′ tiRNAs derived from a restricted set of parental tRNAs (tRNA-Gly-GCC, tRNA-Glu-TTC, and tRNA-Asp-GTC), internal fragments from tRNA-Gly-GCC and tRNA-Asp-GTC; among all tRFs, the most prominently reduced is 5′ tiRNA-Gly-GCC. Increased tRFs are enriched for 3′-derived fragments (notably 3′ tRFs).
5. Genes upregulated upon *Smed ELAC2* silencing at 3 dpa are characteristic of parenchymal cells, while downregulated are primarily associated with intestine cells. By 5 dpa, both upregulated and downregulated sets become more cell-type diverse.
6. The mechanism whereby 5′ tiRNA-Gly-GCC functions as a guide RNA for Smed ELAC2 is unlikely.
7. SMESG000048842.1 was identified as a target of 5′tiRNA-Gly-GCC based on a miRNA-like regulatory model and was subsequently validated experimentally.
8. 5′ tiRNA-Gly-GCC associates with SMEDWI-1, SMEDWI-2, and SMEDWI-3, as demonstrated by analyses of SMEDWI immunoprecipitation datasets.
9. 5′ tiRNA-Gly-GCC in complex with SMEDWI-3 interacts with target mRNAs, as evidenced by analyses of SMEDWI-3 CLASH datasets, supporting potential PIWI-associated engagement *in vivo*.
10. In scRNA-seq, transcriptional responses to *Smed ELAC2* KD at 72 hpa showed the strongest and most bidirectional remodeling, with DEGs concentrated mainly in phagocytes (broad) and *PGRN*$^{+}$ parenchymal cells.

11. *Smed ELAC2* silencing causes cell-composition shifts: at 0 hpa, relative expansion of broad epidermis and neuropeptidergic neurons with concomitant reductions in phagocytes, γ-neoblasts, ζ-neoblasts, and posterior pole/PCG muscle; at 16 hpa, the strongest discrepancies persisted for broad epidermis and phagocytes; at 24 hpa, marked overrepresentation of abraçada cells with an opposite tendency for brain branch neurons;by 72 hpa, the largest compositional deviations involve Abracada cells, $PGRN^{+}$ parenchymal cells, and neuropeptidergic neurons, consistent with a late-stage shift toward specific parenchymal and neuronal states.
12. 5′ tiRNA-Gly-GCC is strongly enriched in the X2/Xins cell population.
13. 5′ tiRNA-Gly-GCC displays predicted activity in phagocytes and *PGRN*$^{+}$ parenchymal cells.
14. *Smed ELAC2* silencing, which causes regeneration delay, alters inferred lineage connectivity: σ-neoblast–eye progenitor connectivity was reduced at 0 hpa (consistent with a weakened eye-progenitor route), while by 72 hpa σ-neoblast connectivity increased toward nervous system, intestine, and parenchyma, indicating strengthened late continuity between broad neoblast states and multiple differentiated/regeneration-associated compartments.

SUPPLEMENTARY FILES

Supplementary files, including the interactive UMAP browser for exploring predicted small RNA activity, as well as supplementary figures, tables, and analysis scripts, are available at: https://norreanea.github.io/AZ-PhD-THESIS-Supplementary.

BIBLIOGRAPHY

Abrar, Mohammad, Didar Hussain, Izaz Ahmad Khan, et al. 2024. “DeepSplice: A Deep Learning Approach for Accurate Prediction of Alternative Splicing Events in the Human Genome.” *Frontiers in Genetics* 15 (June). https://doi.org/10.3389/fgene.2024.1349546.

Adler, Carolyn E., and Alejandro Sánchez Alvarado. 2018. “Systemic RNA Interference in Planarians by Feeding of dsRNA Containing Bacteria.” In *Planarian Regeneration*, edited by Jochen C. Rink, vol. 1774. Methods in Molecular Biology. Springer New York. https://doi.org/10.1007/978-1-4939-7802-1_17.

Adler, Carolyn E., Chris W. Seidel, Sean A. McKinney, and Alejandro Sánchez Alvarado. 2014. “Selective Amputation of the Pharynx Identifies a FoxA-Dependent Regeneration Program in Planaria.” *eLife* 3 (April). https://doi.org/10.7554/eLife.02238.

Allikka Parambil, Sudheesh, Danyan Li, Michael Zelko, Axel Poulet, and Josien C. van Wolfswinkel. 2024. “piRNA Generation Is Associated with the Pioneer Round of Translation in Stem Cells.” *Nucleic Acids Research* 52 (5): 2590–608. https://doi.org/10.1093/nar/gkad1212.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Andersen, Kasper L., and Kathleen Collins. 2012. “Several RNase T2 Enzymes Function in Induced tRNA and rRNA Turnover in the Ciliate *Tetrahymena*.” *Molecular Biology of the Cell* 23 (1): 36–44. https://doi.org/10.1091/mbc.e11-08-0689.

Andreatta, Massimo, and Santiago J. Carmona. 2021. “UCell: Robust and Scalable Single-Cell Gene Signature Scoring.” *Computational and Structural Biotechnology Journal* 19: 3796–98. https://doi.org/10.1016/j.csbj.2021.06.043.

Aparicio, Ricardo, Carolina J. Simoes Da Silva, and Ana Busturia. 2015. “MicroRNA *miR-7* Contributes to the Control of *Drosophila* Wing Growth.” *Developmental Dynamics* 244 (1): 21–30. https://doi.org/10.1002/dvdy.24214.

Asanuma, Takahiro, Soichi Inagaki, Tetsuji Kakutani, Hiroyuki Aburatani, and Yota Murakami. 2022. “Tandemly Repeated Genes Promote RNAi-Mediated Heterochromatin Formation via an Antisilencing Factor, Epe1, in Fission Yeast.” *Genes & Development* 36 (21–24): 1145–59. https://doi.org/10.1101/gad.350129.122.

Avalos, Priscilla N., Lily L. Wong, and David J. Forsthoefel. 2024. “Extracellular Vesicles Promote Proliferation in an Animal Model of Regeneration.” Preprint, Cell Biology, March 27. https://doi.org/10.1101/2024.03.22.586206.

Axtell, Michael J. 2013. “ShortStack: Comprehensive Annotation and Quantification of Small RNA Genes.” *RNA* 19 (6): 740–51. https://doi.org/10.1261/rna.035279.112.

Barghouth, Paul G., Manish Thiruvalluvan, Melanie LeGro, and Néstor J. Oviedo. 2019. "DNA Damage and Tissue Repair: What We Can Learn from Planaria." *Seminars in Cell & Developmental Biology* 87 (March): 145–59. https://doi.org/10.1016/j.semcdb.2018.04.013.

Barrett, T. 2004. "NCBI GEO: Mining Millions of Expression Profiles--Database and Tools." *Nucleic Acids Research* 33 (Database issue): D562–66. https://doi.org/10.1093/nar/gki022.

Bartel, David P. 2009. "MicroRNAs: Target Recognition and Regulatory Functions." *Cell* 136 (2): 215–33. https://doi.org/10.1016/j.cell.2009.01.002.

Bartel, David P. 2018. "Metazoan MicroRNAs." *Cell* 173 (1): 20–51. https://doi.org/10.1016/j.cell.2018.03.006.

Bateman, Alex, Maria-Jesus Martin, Sandra Orchard, et al. 2023. "UniProt: The Universal Protein Knowledgebase in 2023." *Nucleic Acids Research* 51 (D1): D523–31. https://doi.org/10.1093/nar/gkac1052.

Baulcombe, David. 2004. "RNA Silencing in Plants." *Nature* 431 (7006): 356–63. https://doi.org/10.1038/nature02874.

Beckett, Karen, Solange Monier, Lucy Palmer, et al. 2013. "Drosophila S2 Cells Secrete Wingless on Exosome-Like Vesicles but the Wingless Gradient Forms Independently of Exosomes." *Traffic* 14 (1): 82–96. https://doi.org/10.1111/tra.12016.

Behrens, Andrew, and Danny D. Nedialkova. 2022. "Experimental and Computational Workflow for the Analysis of tRNA Pools from Eukaryotic Cells by Mim-tRNAseq." *STAR Protocols* 3 (3): 101579. https://doi.org/10.1016/j.xpro.2022.101579.

Benesova, Sarka, Mikael Kubista, and Lukas Valihrach. 2021. "Small RNA-Sequencing: Approaches and Considerations for miRNA Analysis." *Diagnostics* 11 (6): 964. https://doi.org/10.3390/diagnostics11060964.

Benham-Pyle, Blair W., Carolyn E. Brewster, Aubrey M. Kent, 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): 939–52. https://doi.org/10.1038/s41556-021-00734-6.

Bernstein, Emily, Amy A. Caudy, Scott M. Hammond, and Gregory J. Hannon. 2001. "Role for a Bidentate Ribonuclease in the Initiation Step of RNA Interference." *Nature* 409 (6818): 363–66. https://doi.org/10.1038/35053110.

Boehlke, Christopher, Heike Janusch, Christoph Hamann, et al. 2015. "A Cilia Independent Role of Ift88/Polaris during Cell Migration." *PLOS ONE* 10 (10): e0140378. https://doi.org/10.1371/journal.pone.0140378.

Bonev, Boyan, Peter Stanley, and Nancy Papalopulu. 2012. “MicroRNA-9 Modulates Hes1 Ultradian Oscillations by Forming a Double-Negative Feedback Loop.” *Cell Reports* 2 (1): 10–18. https://doi.org/10.1016/j.celrep.2012.05.017.

Boskovic, Ana, Xin Yang Bing, Ebru Kaymak, and Oliver J. Rando. 2020. “Control of Noncoding RNA Production and Histone Levels by a 5′ tRNA Fragment.” *Genes & Development* 34 (1–2): 118–31. https://doi.org/10.1101/gad.332783.119.

Boulias, Konstantinos, and H. Robert Horvitz. 2012. “The C. Elegans MicroRNA Mir-71 Acts in Neurons to Promote Germline-Mediated Longevity through Regulation of DAF-16/FOXO.” *Cell Metabolism* 15 (4): 439–50. https://doi.org/10.1016/j.cmet.2012.02.014.

Bousquet, Marina, Marian H. Harris, Beiyan Zhou, and Harvey F. Lodish. 2010. “MicroRNA miR-125b Causes Leukemia.” *Proceedings of the National Academy of Sciences* 107 (50): 21558–63. https://doi.org/10.1073/pnas.1016611107.

Brameier, Markus, Astrid Herwig, Richard Reinhardt, Lutz Walter, and Jens Gruber. 2011. “Human Box C/D snoRNAs with miRNA like Functions: Expanding the Range of Regulatory RNAs.” *Nucleic Acids Research* 39 (2): 675–86. https://doi.org/10.1093/nar/gkq776.

Brand, Jeremias N., Ajinkya Bharatraj Patil, Luca Pandolfini, et al. 2025. “Stepwise Emergence of Recombination Suppression Precedes Fissiparous Asexuality in the Planarian *Schmidtea Mediterranea*.” Preprint, Evolutionary Biology, September 16. https://doi.org/10.1101/2025.09.12.675808.

Bray, Samuel R., Livia S. Wyss, Chew Chai, Maria E. Lozada, and Bo Wang. 2024. “Adaptive Robustness through Incoherent Signaling Mechanisms in a Regenerative Brain.” *Cell Reports* 43 (8): 114580. https://doi.org/10.1016/j.celrep.2024.114580.

Brennecke, Julius, David R. Hipfner, Alexander Stark, Robert B. Russell, and Stephen M. Cohen. 2003. “Bantam Encodes a Developmentally Regulated microRNA That Controls Cell Proliferation and Regulates the Proapoptotic Gene Hid in Drosophila.” *Cell* 113 (1): 25–36. https://doi.org/10.1016/S0092-8674(03)00231-9.

Brzezniak, Lien K., Monika Bijata, Roman J. Szczesny, and Piotr P. Stepien. 2011. “Involvement of Human ELAC2 Gene Product in 3’ End Processing of Mitochondrial tRNAs.” *RNA Biology* 8 (4): 616–26. https://doi.org/10.4161/rna.8.4.15393.

Burroughs, Alexander Maxwell, Yoshinari Ando, Michiel Laurens de Hoon, et al. 2011. “Deep-Sequencing of Human Argonaute-Associated Small RNAs Provides Insight into miRNA Sorting and Reveals Argonaute Association with RNA Fragments of Diverse Origin.” *RNA Biology* 8 (1): 158–77. https://doi.org/10.4161/rna.8.1.14300.

Bushmanova, Elena, Dmitry Antipov, Alla Lapidus, Vladimir Suvorov, and Andrey D. Prjibelski. 2016. “RnaQUAST: A Quality Assessment Tool for de Novo

Transcriptome Assemblies." *Bioinformatics* 32 (14): 2210–12. https://doi.org/10.1093/bioinformatics/btw218.

Büttner, Maren, Zhichao Miao, F. Alexander Wolf, Sarah A. Teichmann, and Fabian J. Theis. 2019. "A Test Metric for Assessing Single-Cell RNA-Seq Batch Correction." *Nature Methods* 16 (1): 43–49. https://doi.org/10.1038/s41592-018-0254-1.

Carthew, Richard W., and Erik J. Sontheimer. 2009a. "Origins and Mechanisms of miRNAs and siRNAs." *Cell* 136 (4): 642–55. https://doi.org/10.1016/j.cell.2009.01.035.

Carthew, Richard W., and Erik J. Sontheimer. 2009b. "Origins and Mechanisms of miRNAs and siRNAs." *Cell* 136 (4): 642–55. https://doi.org/10.1016/j.cell.2009.01.035.

Cebrià, Francesc. 2008. "Organization of the Nervous System in the Model Planarian Schmidtea Mediterranea: An Immunocytochemical Study." *Neuroscience Research* 61 (4): 375–84. https://doi.org/10.1016/j.neures.2008.04.005.

Chan, Patricia P., Brian Y. Lin, Allysia J. Mak, and Todd M. Lowe. 2021. "tRNAscan-SE 2.0: Improved Detection and Functional Classification of Transfer RNA Genes." *Nucleic Acids Research* 49 (16): 9077–96. https://doi.org/10.1093/nar/gkab688.

Chen, Ling-Ling, and V. Narry Kim. 2024. "Small and Long Non-Coding RNAs: Past, Present, and Future." *Cell* 187 (23): 6451–85. https://doi.org/10.1016/j.cell.2024.10.024.

Chen, Qi, and Tong Zhou. 2023. "Emerging Functional Principles of tRNA-Derived Small RNAs and Other Regulatory Small RNAs." *Journal of Biological Chemistry* 299 (10): 105225. https://doi.org/10.1016/j.jbc.2023.105225.

Chen, Xuhui. 2025. "Stem Cells (Neoblasts) and Positional Information Jointly Dominate Regeneration in Planarians." *Heliyon* 11 (2): e41833. https://doi.org/10.1016/j.heliyon.2025.e41833.

Cheng, Li-Chun, Kimberly C. Tu, Chris W. Seidel, Sofia M. C. Robb, Fengli Guo, and Alejandro Sánchez Alvarado. 2018. "Cellular, Ultrastructural and Molecular Analyses of Epidermal Cell Development in the Planarian Schmidtea Mediterranea." *Developmental Biology* 433 (2): 357–73. https://doi.org/10.1016/j.ydbio.2017.08.030.

Cherlin, Tess, Rogan Magee, Yi Jing, Venetia Pliatsika, Phillipe Loher, and Isidore Rigoutsos. 2020. "Ribosomal RNA Fragmentation into Short RNAs (rRFs) Is Modulated in a Sex- and Population of Origin-Specific Manner." *BMC Biology* 18 (1): 38. https://doi.org/10.1186/s12915-020-0763-0.

Choi, Eun-Jin, Wenzhe Wu, Ke Zhang, et al. 2021. "ELAC2, an Enzyme for tRNA Maturation, Plays a Role in the Cleavage of a Mature tRNA to Produce a tRNA-Derived RNA Fragment During Respiratory Syncytial Virus Infection." *Frontiers*

in Molecular Biosciences 7 (November). https://doi.org/10.3389/fmolb.2020.609732.

Chong, Tracy, Joel M. Stary, Yuying Wang, and Phillip A. Newmark. 2011. "Molecular Markers to Characterize the Hermaphroditic Reproductive System of the Planarian Schmidtea Mediterranea." *BMC Developmental Biology* 11 (1): 69. https://doi.org/10.1186/1471-213X-11-69.

Clark, Wesley C., Molly E. Evans, Dan Dominissini, Guanqun Zheng, and Tao Pan. 2016. "tRNA Base Methylation Identification and Quantification via High-Throughput Sequencing." *RNA* 22 (11): 1771–84. https://doi.org/10.1261/rna.056531.116.

Clarke, Alexander W., Eirik Høye, Anju Angelina Hembrom, et al. 2025. "MirGeneDB 3.0: Improved Taxonomic Sampling, Uniform Nomenclature of Novel Conserved microRNA Families and Updated Covariance Models." *Nucleic Acids Research* 53 (D1): D116–28. https://doi.org/10.1093/nar/gkae1094.

Collins, James J., Xiaowen Hou, Elena V. Romanova, et al. 2010. "Genome-Wide Analyses Reveal a Role for Peptide Hormones in Planarian Germline Development." *PLoS Biology* 8 (10): e1000509. https://doi.org/10.1371/journal.pbio.1000509.

Colombo, Marina, Graça Raposo, and Clotilde Théry. 2014. "Biogenesis, Secretion, and Intercellular Interactions of Exosomes and Other Extracellular Vesicles." *Annual Review of Cell and Developmental Biology* 30 (1): 255–89. https://doi.org/10.1146/annurev-cellbio-101512-122326.

Conine, Colin C., Pedro J. Batista, Weifeng Gu, et al. 2010. "Argonautes ALG-3 and ALG-4 Are Required for Spermatogenesis-Specific 26G-RNAs and Thermotolerant Sperm in *Caenorhabditis Elegans*." *Proceedings of the National Academy of Sciences* 107 (8): 3588–93. https://doi.org/10.1073/pnas.0911685107.

Coolen, Marion, Shauna Katz, and Laure Bally-Cuif. 2013. "miR-9: A Versatile Regulator of Neurogenesis." *Frontiers in Cellular Neuroscience* 7. https://doi.org/10.3389/fncel.2013.00220.

Cote, Lauren E., Eric Simental, and Peter W. Reddien. 2019. "Muscle Functions as a Connective Tissue and Source of Extracellular Matrix in Planarians." *Nature Communications* 10 (1): 1592. https://doi.org/10.1038/s41467-019-09539-6.

Cowles, Martis W., Kerilyn C. Omuro, Brianna N. Stanley, Carlo G. Quintanilla, and Ricardo M. Zayas. 2014. "COE Loss-of-Function Analysis Reveals a Genetic Program Underlying Maintenance and Regeneration of the Nervous System in Planarians." *PLoS Genetics* 10 (10): e1004746. https://doi.org/10.1371/journal.pgen.1004746.

Cozen, Aaron E., Erin Quartley, Andrew D. Holmes, Eva Hrabeta-Robinson, Eric M. Phizicky, and Todd M. Lowe. 2015. "ARM-Seq: AlkB-Facilitated RNA Methylation Sequencing Reveals a Complex Landscape of Modified tRNA Fragments." *Nature Methods* 12 (9): 879–84. https://doi.org/10.1038/nmeth.3508.

Cui, Guanshen, Kangning Dong, Jia-Yi Zhou, et al. 2023. “Spatiotemporal Transcriptomic Atlas Reveals the Dynamic Characteristics and Key Regulators of Planarian Regeneration.” *Nature Communications* 14 (1): 3205. https://doi.org/10.1038/s41467-023-39016-0.

Currie, Ko W., Alyssa M. Molinaro, and Bret J. Pearson. 2016. “Neuronal Sources of Hedgehog Modulate Neurogenesis in the Adult Planarian Brain.” *eLife* 5 (November). https://doi.org/10.7554/eLife.19735.

Currie, Ko W., and Bret J. Pearson. 2013. “Transcription Factors *Lhx1/5-1* and *Pitx* Are Required for the Maintenance and Regeneration of Serotonergic Neurons in Planarians.” *Development* 140 (17): 3577–88. https://doi.org/10.1242/dev.098590.

Czech, Benjamin, and Gregory J. Hannon. 2016. “One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing.” *Trends in Biochemical Sciences* 41 (4): 324–37. https://doi.org/10.1016/j.tibs.2015.12.008.

Dai, Xiaoting, Xinghua Li, Alexander Tyshkovskiy, et al. 2025. “Regeneration Leads to Global Tissue Rejuvenation in Aging Sexual Planarians.” *Nature Aging* 5 (5): 780–98. https://doi.org/10.1038/s43587-025-00847-9.

De Lencastre, Alexandre, Zachary Pincus, Katherine Zhou, Masaomi Kato, Siu Sylvia Lee, and Frank J. Slack. 2010. “MicroRNAs Both Promote and Antagonize Longevity in C. Elegans.” *Current Biology* 20 (24): 2159–68. https://doi.org/10.1016/j.cub.2010.11.015.

De Robertis, Edward M. 2010. “Wnt Signaling in Axial Patterning and Regeneration: Lessons from Planaria.” *Science Signaling* 3 (127). https://doi.org/10.1126/scisignal.3127pe21.

De Sousa, Nídia, Gustavo Rodríguez-Esteban, Jose Ignacio Rojo-Laguna, Emili Saló, and Teresa Adell. 2018. “Hippo Signaling Controls Cell Cycle and Restricts Cell Plasticity in Planarians.” *PLOS Biology* 16 (1): e2002399. https://doi.org/10.1371/journal.pbio.2002399.

Dhahbi, Joseph M., Stephen R. Spindler, Hani Atamna, et al. 2013. “5′ tRNA Halves Are Present as Abundant Complexes in Serum, Concentrated in Blood Cells, and Modulated by Aging and Calorie Restriction.” *BMC Genomics* 14 (1): 298. https://doi.org/10.1186/1471-2164-14-298.

Dhahbi, Joseph Mohsen. 2015. “5â€^{2} tRNA Halves: The Next Generation of Immune Signaling Molecules.” *Frontiers in Immunology* 6 (February). https://doi.org/10.3389/fimmu.2015.00074.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, et al. 2013. “STAR: Ultrafast Universal RNA-Seq Aligner.” *Bioinformatics* 29 (1): 15–21. https://doi.org/10.1093/bioinformatics/bts635.

Du, Li, Wei Chen, Dong Zhang, Yinghong Cui, and Zuping He. 2024. “The Functions and Mechanisms of piRNAs in Mediating Mammalian Spermatogenesis and Their

Applications in Reproductive Medicine." *Cellular and Molecular Life Sciences* 81 (1): 379. https://doi.org/10.1007/s00018-024-05399-6.

Duba-Kiss, Robert, Yosuke Niibori, and David R. Hampson. 2021. "GABAergic Gene Regulatory Elements Used in Adeno-Associated Viral Vectors." *Frontiers in Neurology* 12 (October). https://doi.org/10.3389/fneur.2021.745159.

Dubrovsky, E. B. 2004. "Drosophila RNase Z Processes Mitochondrial and Nuclear Pre-tRNA 3' Ends in Vivo." *Nucleic Acids Research* 32 (1): 255–62. https://doi.org/10.1093/nar/gkh182.

Dutta, Annasha. 2025. "Small Non-Coding RNAs in Regeneration of Schmidtea Mediterranea." Doctoral dissertation, Institute of Bioorganic Chemistry, Polish Academy of Sciences.

Dutta, Annasha, Anastasiia Zaremba, and Paulina Jackowiak. 2025. "Ribonucleases in Mendelian Disease: Characterization and Insight from Model Organisms." *Genes & Diseases* 12 (5): 101613. https://doi.org/10.1016/j.gendis.2025.101613.

Dyer, Sarah C., Olanrewaju Austine-Orimoloye, Andrey G. Azov, et al. 2025. "Ensembl 2025." *Nucleic Acids Research* 53 (D1): D948–57. https://doi.org/10.1093/nar/gkae1071.

Eichhorn, Stephen W., Huili Guo, Sean E. McGeary, et al. 2014. "mRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues." *Molecular Cell* 56 (1): 104–15. https://doi.org/10.1016/j.molcel.2014.08.028.

Elbarbary, Reyad A., Hiroaki Takaku, Masato Tamura, and Masayuki Nashimoto. 2009. "Inhibition of Vascular Endothelial Growth Factor Expression by TRUE Gene Silencing." *Biochemical and Biophysical Research Communications* 379 (4): 924–27. https://doi.org/10.1016/j.bbrc.2008.12.173.

Elbarbary, Reyad A., Hiroaki Takaku, Naoto Uchiumi, et al. 2009. "Modulation of Gene Expression by Human Cytosolic tRNase ZL through 5′-Half-tRNA." *PLoS ONE* 4 (6): e5908. https://doi.org/10.1371/journal.pone.0005908.

Elbashir, Sayda M., Jens Harborth, Winfried Lendeckel, Abdullah Yalcin, Klaus Weber, and Thomas Tuschl. 2001. "Duplexes of 21-Nucleotide RNAs Mediate RNA Interference in Cultured Mammalian Cells." *Nature* 411 (6836): 494–98. https://doi.org/10.1038/35078107.

Elder, Jessica J. H., Ry Papadopoulos, Cassandra K. Hayne, and Robin E. Stanley. 2024. "The Making and Breaking of tRNAs by Ribonucleases." *Trends in Genetics* 40 (6): 511–25. https://doi.org/10.1016/j.tig.2024.03.007.

Emara, Mohamed M., Pavel Ivanov, Tyler Hickman, et al. 2010. "Angiogenin-Induced tRNA-Derived Stress-Induced RNAs Promote Stress-Induced Stress Granule Assembly." *Journal of Biological Chemistry* 285 (14): 10959–68. https://doi.org/10.1074/jbc.M109.077560.

Emili, Elena, Alberto Pérez-Posada, Virginia Vanni, et al. 2025. “Allometry of Cell Types in Planarians by Single-Cell Transcriptomics.” *Science Advances* 11 (19). https://doi.org/10.1126/sciadv.adm7042.

Ender, Christine, Azra Krek, Marc R. Friedländer, et al. 2008. “A Human snoRNA with MicroRNA-Like Functions.” *Molecular Cell* 32 (4): 519–28. https://doi.org/10.1016/j.molcel.2008.10.017.

Engel, Annika, Shusruto Rishik, Pascal Hirsch, et al. 2024. “SingmiR: A Single-Cell miRNA Alignment and Analysis Tool.” *Nucleic Acids Research* 52 (W1): W374–80. https://doi.org/10.1093/nar/gkae225.

Enright, Anton J., Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora S. Marks. 2003. “MicroRNA Targets in Drosophila.” *Genome Biology* 5 (1): R1. https://doi.org/10.1186/gb-2003-5-1-r1.

Evans, Ciaran, Johanna Hardin, and Daniel M. Stoebel. 2018. “Selecting Between-Sample RNA-Seq Normalization Methods from the Perspective of Their Assumptions.” *Briefings in Bioinformatics* 19 (5): 776–92. https://doi.org/10.1093/bib/bbx008.

Ewels, Philip, Måns Magnusson, Sverker Lundin, and Max Käller. 2016. “MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report.” *Bioinformatics* 32 (19): 3047–48. https://doi.org/10.1093/bioinformatics/btw354.

Fang, Zhuo, and Nikolaus Rajewsky. 2011. “The Impact of miRNA Target Sites in Coding Sequences and in 3′UTRs.” *PLoS ONE* 6 (3): e18067. https://doi.org/10.1371/journal.pone.0018067.

Fincher, Christopher T., Omri Wurtzel, Thom de Hoog, Kellie M. Kravarik, and Peter W. Reddien. 2018. “Cell Type Transcriptome Atlas for the Planarian *Schmidtea Mediterranea*.” *Science* 360 (6391). https://doi.org/10.1126/science.aaq1736.

Fire, Andrew, SiQun Xu, Mary K. Montgomery, Steven A. Kostas, Samuel E. Driver, and Craig C. Mello. 1998. “Potent and Specific Genetic Interference by Double-Stranded RNA in Caenorhabditis Elegans.” *Nature* 391 (6669): 806–11. https://doi.org/10.1038/35888.

Forsthoefel, David J., Nicholas I. Cejda, Umair W. Khan, and Phillip A. Newmark. 2020. “Cell-Type Diversity and Regionalized Gene Expression in the Planarian Intestine.” *eLife* 9 (April). https://doi.org/10.7554/eLife.52613.

Forsthoefel, David J., Forrest A. Waters, and Phillip A. Newmark. 2014. “Generation of Cell Type-Specific Monoclonal Antibodies for the Planarian and Optimization of Sample Processing for Immunolabeling.” *BMC Developmental Biology* 14 (1): 45. https://doi.org/10.1186/s12861-014-0045-6.

Fraguas, Susanna, Sara Barberán, Begoña Ibarra, Linda Stöger, and Francesc Cebri. 2012. “Regeneration of Neuronal Cell Types in Schmidtea Mediterranea: An Immunohistochemical and Expression Study.” *The International Journal of*

Developmental Biology 56 (1-2–3): 143–53. https://doi.org/10.1387/ijdb.113428sf.

Friedländer, Marc R., Catherine Adamidi, Ting Han, et al. 2009. "High-Resolution Profiling and Discovery of Planarian Small RNAs." *Proceedings of the National Academy of Sciences* 106 (28): 11546–51. https://doi.org/10.1073/pnas.0905222106.

Friedländer, Marc R., Sebastian D. Mackowiak, Na Li, Wei Chen, and Nikolaus Rajewsky. 2012. "miRDeep2 Accurately Identifies Known and Hundreds of Novel microRNA Genes in Seven Animal Clades." *Nucleic Acids Research* 40 (1): 37–52. https://doi.org/10.1093/nar/gkr688.

Fu, Bi-Fei, and Chao-Yang Xu. 2022. "Transfer RNA-Derived Small RNAs: Novel Regulators and Biomarkers of Cancers." *Frontiers in Oncology* 12 (April). https://doi.org/10.3389/fonc.2022.843598.

Fu, Laiyi, Yingxin Cao, Jie Wu, Qinke Peng, Qing Nie, and Xiaohui Xie. 2022. "UFold: Fast and Accurate RNA Secondary Structure Prediction with Deep Learning." *Nucleic Acids Research* 50 (3): e14–e14. https://doi.org/10.1093/nar/gkab1074.

Fukasawa, Yoshinori, Junko Tsuji, Szu-Chin Fu, Kentaro Tomii, Paul Horton, and Kenichiro Imai. 2015. "MitoFates: Improved Prediction of Mitochondrial Targeting Sequences and Their Cleavage Sites*." *Molecular & Cellular Proteomics* 14 (4): 1113–26. https://doi.org/10.1074/mcp.M114.043083.

Gámbaro, Fabiana, Marco Li Calzi, Pablo Fagúndez, et al. 2020. "Stable tRNA Halves Can Be Sorted into Extracellular Vesicles and Delivered to Recipient Cells in a Concentration-Dependent Manner." *RNA Biology* 17 (8): 1168–82. https://doi.org/10.1080/15476286.2019.1708548.

García-Castro, Helena, and Jordi Solana. 2022. "Single-Cell Transcriptomics in Planaria: New Tools Allow New Insights into Cellular and Evolutionary Features." *Biochemical Society Transactions* 50 (5): 1237–46. https://doi.org/10.1042/BST20210825.

Gaylord, Abigail, Elizabeth A. Holzhausen, Bridget Chalifour, et al. 2024. "tRNA-Derived RNAs in Human Milk Extracellular Vesicles and Associations with Breastfeeding Variables and Maternal Diet." *Epigenomics* 16 (23–24): 1429–41. https://doi.org/10.1080/17501911.2024.2430943.

George, Tina P., Suja Subramanian, and M. H. Supriya. 2024. "A Brief Review of Noncoding RNA." *Egyptian Journal of Medical Human Genetics* 25 (1): 98. https://doi.org/10.1186/s43042-024-00553-y.

Godoy, Paula M., Nirav R. Bhakta, Andrea J. Barczak, et al. 2018. "Large Differences in Small RNA Composition Between Human Biofluids." *Cell Reports* 25 (5): 1346–58. https://doi.org/10.1016/j.celrep.2018.10.014.

Goh, Wee Siong Sho, Ilaria Falciatori, Oliver H. Tam, et al. 2015. “piRNA-Directed Cleavage of Meiotic Transcripts Regulates Spermatogenesis.” *Genes & Development* 29 (10): 1032–44. https://doi.org/10.1101/gad.260455.115.

Goodarzi, Hani, Xuhang Liu, Hoang C. B. Nguyen, Steven Zhang, Lisa Fish, and Sohail F. Tavazoie. 2015. “Endogenous tRNA-Derived Fragments Suppress Breast Cancer Progression via YBX1 Displacement.” *Cell* 161 (4): 790–802. https://doi.org/10.1016/j.cell.2015.02.053.

Gou, Lan-Tao, Qifan Zhu, and Mo-Fang Liu. 2023. “Small RNAs: An Expanding World with Therapeutic Promises.” *Fundamental Research* 3 (5): 676–82. https://doi.org/10.1016/j.fmre.2023.03.003.

Gradilla, Ana-Citlali, Esperanza González, Irene Seijo, et al. 2014. “Exosomes as Hedgehog Carriers in Cytoneme-Mediated Transport and Secretion.” *Nature Communications* 5 (1): 5649. https://doi.org/10.1038/ncomms6649.

Grohme, Markus Alexander, Siegfried Schloissnig, Andrei Rozanski, et al. 2018. “The Genome of Schmidtea Mediterranea and the Evolution of Core Cellular Mechanisms.” *Nature* 554 (7690): 56–61. https://doi.org/10.1038/nature25473.

Grolmusz, Vince Kornél, Eszter Angéla Tóth, Kornélia Baghy, et al. 2016. “Fluorescence Activated Cell Sorting Followed by Small RNA Sequencing Reveals Stable microRNA Expression during Cell Cycle Progression.” *BMC Genomics* 17 (1): 412. https://doi.org/10.1186/s12864-016-2747-6.

Guan, Lingyu, Spyros Karaiskos, and Andrey Grigoriev. 2020. “Inferring Targeting Modes of Argonaute-Loaded tRNA Fragments.” *RNA Biology* 17 (8): 1070–80. https://doi.org/10.1080/15476286.2019.1676633.

Guemez-Gamboa, Alicia, Nicole G. Coufal, and Joseph G. Gleeson. 2014. “Primary Cilia in the Developing and Mature Brain.” *Neuron* 82 (3): 511–21. https://doi.org/10.1016/j.neuron.2014.04.024.

Guo, Huili, Nicholas T. Ingolia, Jonathan S. Weissman, and David P. Bartel. 2010. “Mammalian microRNAs Predominantly Act to Decrease Target mRNA Levels.” *Nature* 466 (7308): 835–40. https://doi.org/10.1038/nature09267.

Guo, Longhua, Joshua S. Bloom, Daniel Dols-Serrate, et al. 2022. “Island-Specific Evolution of a Sex-Primed Autosome in a Sexual Planarian.” *Nature* 606 (7913): 329–34. https://doi.org/10.1038/s41586-022-04757-3.

Gurtan, Allan M., and Phillip A. Sharp. 2013. “The Role of miRNAs in Regulating Gene Expression Networks.” *Journal of Molecular Biology* 425 (19): 3582–600. https://doi.org/10.1016/j.jmb.2013.03.007.

Guzzi, Nicola, and Cristian Bellodi. 2020. “Novel Insights into the Emerging Roles of tRNA-Derived Fragments in Mammalian Development.” *RNA Biology* 17 (8): 1214–22. https://doi.org/10.1080/15476286.2020.1732694.

Ha, Minju, and V. Narry Kim. 2014. “Regulation of microRNA Biogenesis.” *Nature Reviews Molecular Cell Biology* 15 (8): 509–24. https://doi.org/10.1038/nrm3838.

Haack, Tobias B., Robert Kopajtich, Peter Freisinger, et al. 2013. “ELAC2 Mutations Cause a Mitochondrial RNA Processing Defect Associated with Hypertrophic Cardiomyopathy.” *The American Journal of Human Genetics* 93 (2): 211–23. https://doi.org/10.1016/j.ajhg.2013.06.006.

Haas, Brian J., Alexie Papanicolaou, Moran Yassour, et al. 2013. “De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis.” *Nature Protocols* 8 (8): 1494–512. https://doi.org/10.1038/nprot.2013.084.

Habu, Y. 2005. “Inhibition of HIV-1 Gene Expression by Retroviral Vector-Mediated Small-Guide RNAs That Direct Specific RNA Cleavage by tRNase ZL.” *Nucleic Acids Research* 33 (1): 235–43. https://doi.org/10.1093/nar/gki164.

Hafemeister, Christoph, and Rahul Satija. 2019. “Normalization and Variance Stabilization of Single-Cell RNA-Seq Data Using Regularized Negative Binomial Regression.” *Genome Biology* 20 (1): 296. https://doi.org/10.1186/s13059-019-1874-1.

Hagemann-Jensen, Michael, Ilgar Abdullayev, Rickard Sandberg, and Omid R. Faridani. 2018. “Small-Seq for Single-Cell Small-RNA Sequencing.” *Nature Protocols* 13 (10): 2407–24. https://doi.org/10.1038/s41596-018-0049-y.

Halic, Mario, and Danesh Moazed. 2009. “Transposon Silencing by piRNAs.” *Cell* 138 (6): 1058–60. https://doi.org/10.1016/j.cell.2009.08.030.

Hall, Richard Nelson, Uri Weill, Leonard Drees, et al. 2022. “Heterologous Reporter Expression in the Planarian Schmidtea Mediterranea through Somatic mRNA Transfection.” *Cell Reports Methods* 2 (10): 100298. https://doi.org/10.1016/j.crmeth.2022.100298.

Hammond, Scott M., Emily Bernstein, David Beach, and Gregory J. Hannon. 2000. “An RNA-Directed Nuclease Mediates Post-Transcriptional Gene Silencing in Drosophila Cells.” *Nature* 404 (6775): 293–96. https://doi.org/10.1038/35005107.

Hansen, Thomas B., Morten T. Venø, Trine I. Jensen, Anne Schaefer, Christian K. Damgaard, and Jørgen Kjems. 2016. “Argonaute-Associated Short Introns Are a Novel Class of Gene Regulators.” *Nature Communications* 7 (1): 11538. https://doi.org/10.1038/ncomms11538.

Haussecker, Dirk, Yong Huang, Ashley Lau, Poornima Parameswaran, Andrew Z. Fire, and Mark A. Kay. 2010. “Human tRNA-Derived Small RNAs in the Global Regulation of RNA Silencing.” *RNA* 16 (4): 673–95. https://doi.org/10.1261/rna.2000810.

Held, James P., Gaomin Feng, Benjamin R. Saunders, Claudia V. Pereira, Kristopher Burkewitz, and Maulik R. Patel. 2022. “A tRNA Processing Enzyme Is a Key Regulator of the Mitochondrial Unfolded Protein Response.” *eLife* 11 (April): e71634. https://doi.org/10.7554/eLife.71634.

Herbst, Efrat, Yael Mandel-Gutfreund, Zohar Yakhini, and Hadas Biran. 2025. “Inferring Single-Cell and Spatial microRNA Activity from Transcriptomics Data.” *Communications Biology* 8 (1): 87. https://doi.org/10.1038/s42003-025-07454-9.

Hirata, Hiromi, Megumi Takahashi, Kenta Yamada, and Kazutoyo Ogino. 2011. “The Biological Role of the Glycinergic Synapse in Early Zebrafish Motility.” *Neuroscience Research* 71 (1): 1–11. https://doi.org/10.1016/j.neures.2011.06.003.

Hoffman, Michael, and Omri Wurtzel. 2023. “PLANAtools—An Interactive Gene Expression Repository for the Planarian Schmidtea Mediterranea.” *Frontiers in Cell and Developmental Biology* 11 (March). https://doi.org/10.3389/fcell.2023.1149537.

Holmes, Andrew D., Jonathan M. Howard, Patricia P. Chan, and Todd M. Lowe. 2022. “tRNA Analysis of eXpression (tRAX): A Tool for Integrating Analysis of tRNAs, tRNA-Derived Small RNAs, and tRNA Modifications.” July. https://doi.org/10.1101/2022.07.02.498565.

Hori, I. 1991. “Role of Fixed Parenchyma Cells in Blastema Formation of the Planarian Dugesia Japonica.” *The International Journal of Developmental Biology* 35 (2): 101–8.

Hücker, Sarah M., Tobias Fehlmann, Christian Werno, et al. 2021. “Single-Cell microRNA Sequencing Method Comparison and Application to Cell Lines and Circulating Lung Tumor Cells.” *Nature Communications* 12 (1): 4316. https://doi.org/10.1038/s41467-021-24611-w.

Hudish, Laura I., Domenico F. Galati, Andrew M. Ravanelli, Chad G. Pearson, Peng Huang, and Bruce Appel. 2016. “*miR-219* Regulates Neural Progenitors by Dampening Apical Par Protein-Dependent Hedgehog Signaling.” *Development*, January 1, dev.137844. https://doi.org/10.1242/dev.137844.

Hyun, Seogang, Jung Hyun Lee, Hua Jin, et al. 2009. “Conserved MicroRNA miR-8/miR-200 and Its Target USH/FOG2 Control Growth by Regulating PI3K.” *Cell* 139 (6): 1096–108. https://doi.org/10.1016/j.cell.2009.11.020.

Ivanković, Mario, Jeremias N. Brand, Luca Pandolfini, et al. 2024. “A Comparative Analysis of Planarian Genomes Reveals Regulatory Conservation in the Face of Rapid Structural Divergence.” *Nature Communications* 15 (1): 8215. https://doi.org/10.1038/s41467-024-52380-9.

Ivanov, Pavel, Mohamed M. Emara, Judit Villen, Steven P. Gygi, and Paul Anderson. 2011. “Angiogenin-Induced tRNA Fragments Inhibit Translation Initiation.” *Molecular Cell* 43 (4): 613–23. https://doi.org/10.1016/j.molcel.2011.06.022.

Jehn, Julia, Jana Treml, Svenja Wulsch, et al. 2020. “5′ tRNA Halves Are Highly Expressed in the Primate Hippocampus and Might Sequence-Specifically Regulate Gene Expression.” *RNA* 26 (6): 694–707. https://doi.org/10.1261/rna.073395.119.

Jiang, Qin, Yan Ma, Ya Zhao, et al. 2022. “tRNA-Derived Fragment tRF-1001: A Novel Anti-Angiogenic Factor in Pathological Ocular Angiogenesis.” *Molecular Therapy - Nucleic Acids* 30 (December): 407–20. https://doi.org/10.1016/j.omtn.2022.10.016.

Jin, Hanyong, Ji-Hyun Yeom, Eunkyoung Shin, et al. 2024. “5′-tRNAGly(GCC) Halves Generated by IRE1α Are Linked to the ER Stress Response.” *Nature Communications* 15 (1): 9273. https://doi.org/10.1038/s41467-024-53624-4.

Jonas, Stefanie, and Elisa Izaurralde. 2015. “Towards a Molecular Understanding of microRNA-Mediated Gene Silencing.” *Nature Reviews Genetics* 16 (7): 421–33. https://doi.org/10.1038/nrg3965.

Juvvuna, Prasanna Kumar, Piyush Khandelia, Li Ming Lee, and Eugene V. Makeyev. 2012. “Argonaute Identity Defines the Length of Mature Mammalian microRNAs.” *Nucleic Acids Research* 40 (14): 6808–20. https://doi.org/10.1093/nar/gks293.

Juzenas, Simonas, Geetha Venkatesh, Matthias Hübenthal, et al. 2017. “A Comprehensive, Cell Specific microRNA Catalogue of Human Peripheral Blood.” *Nucleic Acids Research* 45 (16): 9290–301. https://doi.org/10.1093/nar/gkx706.

Kao, Damian, Daniel Felix, and Aziz Aboobaker. 2013. “The Planarian Regeneration Transcriptome Reveals a Shared but Temporally Shifted Regulatory Program between Opposing Head and Tail Scenarios.” *BMC Genomics* 14 (1): 797. https://doi.org/10.1186/1471-2164-14-797.

Keam, Simon, and Gyorgy Hutvagner. 2015. “tRNA-Derived Fragments (tRFs): Emerging New Roles for an Ancient RNA in the Regulation of Gene Expression.” *Life* 5 (4): 1638–51. https://doi.org/10.3390/life5041638.

Keam, Simon P., Paul E. Young, Alexandra L. McCorkindale, et al. 2014. “The Human Piwi Protein Hiwi2 Associates with tRNA-Derived piRNAs in Somatic Cells.” *Nucleic Acids Research* 42 (14): 8984–95. https://doi.org/10.1093/nar/gku620.

Kent, Aubrey M., Carlos Guerrero-Hernández, Carolyn Brewster, et al. 2026. “Metabolites Produced by Agat+ Cells Support Regeneration in the Planarian Schmidtea Mediterranea.” *Developmental Biology* 529 (January): 106–20. https://doi.org/10.1016/j.ydbio.2025.10.001.

Kim, Iana V., Elizabeth M. Duncan, Eric J. Ross, et al. 2019. “Planarians Recruit piRNAs for mRNA Turnover in Adult Stem Cells.” *Genes & Development* 33 (21–22): 1575–90. https://doi.org/10.1101/gad.322776.118.

Kim, Iana V., Sebastian Riedelbauch, and Claus-D. Kuhn. 2020. “The piRNA Pathway in Planarian Flatworms: New Model, New Insights.” *Biological Chemistry* 401 (10): 1123–41. https://doi.org/10.1515/hsz-2019-0445.

Kim, Sung-Il, Haomin Lyu, Dinesh S. Pujara, et al. 2025. “A Nuclear tRNA-Derived Fragment Triggers Immunity in Arabidopsis.” *Communications Biology* 8 (1): 533. https://doi.org/10.1038/s42003-025-07737-1.

King, Hunter O., Kwadwo E. Owusu-Boaitey, Christopher T. Fincher, and Peter W. Reddien. 2024. “A Transcription Factor Atlas of Stem Cell Fate in Planarians.” *Cell Reports* 43 (3): 113843. https://doi.org/10.1016/j.celrep.2024.113843.

King, Ryan S., and Phillip A. Newmark. 2013. “In Situ Hybridization Protocol for Enhanced Detection of Gene Expression in the Planarian Schmidtea Mediterranea.” *BMC Developmental Biology* 13 (1): 8. https://doi.org/10.1186/1471-213X-13-8.

Kodama, Y., M. Shumway, R. Leinonen, and on behalf of the International Nucleotide Sequence Database Collaboration. 2012. “The Sequence Read Archive: Explosive Growth of Sequencing Data.” *Nucleic Acids Research* 40 (D1): D54–56. https://doi.org/10.1093/nar/gkr854.

Kondratov, Kirill A., Alexander A. Artamonov, Yuri V. Nikitin, et al. 2024. “Revealing Differential Expression Patterns of piRNA in FACS Blood Cells of SARS-CoV−2 Infected Patients.” *BMC Medical Genomics* 17 (1): 212. https://doi.org/10.1186/s12920-024-01982-9.

Korotkevich, Gennady, Vladimir Sukhov, Nikolay Budin, Boris Shpak, Maxim N. Artyomov, and Alexey Sergushichev. 2016. “Fast Gene Set Enrichment Analysis.” Preprint, Bioinformatics, June 20. https://doi.org/10.1101/060012.

Korsunsky, Ilya, Nghia Millard, Jean Fan, et al. 2019. “Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony.” *Nature Methods* 16 (12): 1289–96. https://doi.org/10.1038/s41592-019-0619-0.

Kozomara, Ana, Maria Birgaoanu, and Sam Griffiths-Jones. 2019. “miRBase: From microRNA Sequences to Function.” *Nucleic Acids Research* 47 (D1): D155–62. https://doi.org/10.1093/nar/gky1141.

Krishna, Srikar, Srikala Raghavan, Ramanuj DasGupta, and Dasaradhi Palakodeti. 2021. “tRNA-Derived Fragments (tRFs): Establishing Their Turf in Post-Transcriptional Gene Regulation.” *Cellular and Molecular Life Sciences* 78 (6): 2607–19. https://doi.org/10.1007/s00018-020-03720-7.

Krishna, Srikar, Daniel Gr Yim, Vairavan Lakshmanan, et al. 2019. “Dynamic Expression of tRNA-derived Small RNAs Define Cellular States.” *EMBO Reports* 20 (7): e47789. https://doi.org/10.15252/embr.201947789.

Krude, Torsten, Christo P. Christov, Olivier Hyrien, and Kathrin Marheineke. 2009. “Y RNA Functions at the Initiation Step of Mammalian Chromosomal DNA

Replication." *Journal of Cell Science* 122 (16): 2836–45. https://doi.org/10.1242/jcs.047563.

Krueger, Felix, Frankie James, Phil Ewels, et al. 2023. *FelixKrueger/TrimGalore: V0.6.10 - Add Default Decompression Path*. V. 0.6.10. Zenodo, released February 2. https://doi.org/10.5281/ZENODO.5127898.

Kufel, Joanna, and Pawel Grzechnik. 2019. "Small Nucleolar RNAs Tell a Different Tale." *Trends in Genetics* 35 (2): 104–17. https://doi.org/10.1016/j.tig.2018.11.005.

Kumar, Pankaj, Jordan Anaya, Suresh B. Mudunuri, and Anindya Dutta. 2014. "Meta-Analysis of tRNA Derived RNA Fragments Reveals That They Are Evolutionarily Conserved and Associate with AGO Proteins to Recognize Specific RNA Targets." *BMC Biology* 12 (1): 78. https://doi.org/10.1186/s12915-014-0078-0.

Kumar, Pankaj, Suresh B. Mudunuri, Jordan Anaya, and Anindya Dutta. 2015. "tRFdb: A Database for Transfer RNA Fragments." *Nucleic Acids Research* 43 (D1): D141–45. https://doi.org/10.1093/nar/gku1138.

Kuscu, Canan, Pankaj Kumar, Manjari Kiran, Zhangli Su, Asrar Malik, and Anindya Dutta. 2018. "tRNA Fragments (tRFs) Guide Ago to Regulate Gene Expression Post-Transcriptionally in a Dicer-Independent Manner." *RNA* 24 (8): 1093–105. https://doi.org/10.1261/rna.066126.118.

Kuznetsov, G. V., D. E. Mitkovskii, and N. D. Kreshchenko. 2024. "Morphometric Analysis of Serotoninergic Structures in the Nervous System of the Planarian Schmidtea Mediterranea." *Biophysics* 69 (1): 63–74. https://doi.org/10.1134/S0006350924700088.

La Manno, Gioele, Ruslan Soldatov, Amit Zeisel, et al. 2018. "RNA Velocity of Single Cells." *Nature* 560 (7719): 494–98. https://doi.org/10.1038/s41586-018-0414-6.

Lakshmanan, Vairavan, T. N. Sujith, Dhiru Bansal, Padubidri V. Shivaprasad, Dasaradhi Palakodeti, and Srikar Krishna. 2021. "Comprehensive Annotation and Characterization of Planarian tRNA and tRNA-Derived Fragments (tRFs)." *RNA* 27 (4): 477–95. https://doi.org/10.1261/rna.077701.120.

Lambert, Marine, Abderrahim Benmoussa, and Patrick Provost. 2019. "Small Non-Coding RNAs Derived from Eukaryotic Ribosomal RNA." *Non-Coding RNA* 5 (1): 16. https://doi.org/10.3390/ncrna5010016.

Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59. https://doi.org/10.1038/nmeth.1923.

Lapan, Sylvain W., and Peter W. Reddien. 2011. "Dlx and Sp6-9 Control Optic Cup Regeneration in a Prototypic Eye." *PLoS Genetics* 7 (8): e1002226. https://doi.org/10.1371/journal.pgen.1002226.

Lapan, Sylvain W., and Peter W. Reddien. 2012. “Transcriptome Analysis of the Planarian Eye Identifies Ovo as a Specific Regulator of Eye Regeneration.” *Cell Reports* 2 (2): 294–307. https://doi.org/10.1016/j.celrep.2012.06.018.

Lázaro, Eva M., Abdul Harrath, Giacinta A. Stocchino, Maria Pala, Jaume Bagũ, and Marta Riutort. 2011. “Schmidtea Mediterranea Phylogeography: An Old Species Surviving on a Few Mediterranean Islands?” *BMC Evolutionary Biology* 11 (1). https://doi.org/10.1186/1471-2148-11-274.

Lee, Chris W., Jeannette N. Stankowski, Jeannie Chew, et al. 2017. “The Lysosomal Protein Cathepsin L Is a Progranulin Protease.” *Molecular Neurodegeneration* 12 (1): 55. https://doi.org/10.1186/s13024-017-0196-6.

Lee, Rosalind C., Rhonda L. Feinbaum, and Victor Ambros. 1993. “The C. Elegans Heterochronic Gene Lin-4 Encodes Small RNAs with Antisense Complementarity to Lin-14.” *Cell* 75 (5): 843–54. https://doi.org/10.1016/0092-8674(93)90529-Y.

Lee, Yang-ja, Joshua D. Bernstock, Dace Klimanis, and John M. Hallenbeck. 2018. “Akt Protein Kinase, miR-200/miR-182 Expression and Epithelial-Mesenchymal Transition Proteins in Hibernating Ground Squirrels.” *Frontiers in Molecular Neuroscience* 11 (January): 22. https://doi.org/10.3389/fnmol.2018.00022.

Lee, Yong Sun, Yoshiyuki Shibata, Ankit Malhotra, and Anindya Dutta. 2009. “A Novel Class of Small RNAs: tRNA-Derived RNA Fragments (tRFs).” *Genes & Development* 23 (22): 2639–49. https://doi.org/10.1101/gad.1837609.

Levin, Michael, Alexis M. Pietak, and Johanna Bischof. 2019. “Planarian Regeneration as a Model of Anatomical Homeostasis: Recent Progress in Biophysical and Computational Approaches.” *Seminars in Cell & Developmental Biology* 87 (March): 125–44. https://doi.org/10.1016/j.semcdb.2018.04.003.

Lewis, Morag A., Elizabeth Quint, Anne M. Glazier, et al. 2009. “An ENU-Induced Mutation of miR-96 Associated with Progressive Hearing Loss in Mice.” *Nature Genetics* 41 (5): 614–18. https://doi.org/10.1038/ng.369.

Li, Danyan, David H. Taylor, and Josien C. Van Wolfswinkel. 2021. “PIWI-Mediated Control of Tissue-Specific Transposons Is Essential for Somatic Cell Differentiation.” *Cell Reports* 37 (1): 109776. https://doi.org/10.1016/j.celrep.2021.109776.

Li, Jia, Jing Tian, and Tao Cai. 2025. “Integrated Analysis of miRNAs and mRNAs in Thousands of Single Cells.” *Scientific Reports* 15 (1): 1636. https://doi.org/10.1038/s41598-025-85612-z.

Li, Jia, Zhirong Zhang, Yinghua Zhuang, Fengchao Wang, and Tao Cai. 2023. “Small RNA Transcriptome Analysis Using Parallel Single-Cell Small RNA Sequencing.” *Scientific Reports* 13 (1): 7501. https://doi.org/10.1038/s41598-023-34390-7.

Li, Ningshan, Siqiong Yao, Guangjun Yu, Lingeng Lu, and Zuoheng Wang. 2024. "tRFtarget 2.0: Expanding the Targetome Landscape of Transfer RNA-Derived Fragments." *Nucleic Acids Research* 52 (D1): D345–50. https://doi.org/10.1093/nar/gkad815.

Li, Xinmin, and Cun-Yu Wang. 2021. "From Bulk, Single-Cell to Spatial RNA Sequencing." *International Journal of Oral Science* 13 (1): 36. https://doi.org/10.1038/s41368-021-00146-0.

Li, Yue, An Zeng, Xiao-Shuan Han, et al. 2017. "Small RNAome Sequencing Delineates the Small RNA Landscape of Pluripotent Adult Stem Cells in the Planarian Schmidtea Mediterranea." *Genomics Data* 14 (December): 114–25. https://doi.org/10.1016/j.gdata.2017.10.004.

Lin, Alexander Y. T., and Bret J. Pearson. 2014. "Planarian *Yorkie/YAP* Functions to Integrate Adult Stem Cell Proliferation, Organ Homeostasis and Maintenance of Axial Patterning." *Development* 141 (6): 1197–208. https://doi.org/10.1242/dev.101915.

Lin, Alexander Y. T., and Bret J. Pearson. 2017. "Yorkie Is Required to Restrict the Injury Responses in Planarians." *PLOS Genetics* 13 (7): e1006874. https://doi.org/10.1371/journal.pgen.1006874.

Liu, Hongjin, Qian Song, Hui Zhen, Hongkuan Deng, Bosheng Zhao, and Zhonghong Cao. 2021. "miR-8b Is Involved in Brain and Eye Regeneration of Dugesia Japonica in Head Regeneration." *Biology Open* 10 (6): bio058538. https://doi.org/10.1242/bio.058538.

Liu, Jidong, Michelle A. Carmell, Fabiola V. Rivas, et al. 2004. "Argonaute2 Is the Catalytic Engine of Mammalian RNAi." *Science* 305 (5689): 1437–41. https://doi.org/10.1126/science.1102513.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550. https://doi.org/10.1186/s13059-014-0550-8.

Lu, Yi-Chien, Magda Smielewska, Dasaradhi Palakodeti, et al. 2009. "Deep Sequencing Identifies New and Regulated microRNAs in *Schmidtea Mediterranea*." *RNA* 15 (8): 1483–91. https://doi.org/10.1261/rna.1702009.

Madrigal, Pedro, Anil S. Thanki, Silvie Fexova, et al. 2026. "Expression Atlas in 2026: Enabling FAIR and Open Expression Data through Community Collaboration and Integration." *Nucleic Acids Research* 54 (D1): D147–57. https://doi.org/10.1093/nar/gkaf1238.

Maji, Ranjan K., Matthias S. Leisegang, Reinier A. Boon, and Marcel H. Schulz. 2025. "Revealing microRNA Regulation in Single Cells." *Trends in Genetics* 41 (6): 522–36. https://doi.org/10.1016/j.tig.2024.12.009.

Manni, Mosè, Matthew R. Berkeley, Mathieu Seppey, and Evgeny M. Zdobnov. 2021. "BUSCO: Assessing Genomic Data Quality and Beyond." *Current Protocols* 1 (12). https://doi.org/10.1002/cpz1.323.

Maraia, Richard J., and Tek N. Lamichhane. 2011. "3′ Processing of Eukaryotic Precursor tRNAs." *WIREs RNA* 2 (3): 362–75. https://doi.org/10.1002/wrna.64.

Marasovic, Mirela, Manuel Zocco, and Mario Halic. 2013. "Argonaute and Triman Generate Dicer-Independent priRNAs and Mature siRNAs to Initiate Heterochromatin Formation." *Molecular Cell* 52 (2): 173–83. https://doi.org/10.1016/j.molcel.2013.08.046.

Marlétaz, Ferdinand, Katja T. C. A. Peijnenburg, Taichiro Goto, Noriyuki Satoh, and Daniel S. Rokhsar. 2019. "A New Spiralian Phylogeny Places the Enigmatic Arrow Worms among Gnathiferans." *Current Biology* 29 (2): 312-318.e3. https://doi.org/10.1016/j.cub.2018.11.042.

Martin, Marcel. 2011. "Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads." *EMBnet.Journal* 17 (1): 10. https://doi.org/10.14806/ej.17.1.200.

Matranga, Christian, Yukihide Tomari, Chanseok Shin, David P. Bartel, and Phillip D. Zamore. 2005. "Passenger-Strand Cleavage Facilitates Assembly of siRNA into Ago2-Containing RNAi Enzyme Complexes." *Cell* 123 (4): 607–20. https://doi.org/10.1016/j.cell.2005.08.044.

McCubbin, Ryan A., Mohammad A. Auwal, Shengzhou Wang, et al. 2025. "*Smed-Pou4-2* Regulates Mechanosensory Neuron Regeneration and Function in Planarians." May. https://doi.org/10.1101/2025.05.15.654132.

Megel, Cyrille, Guillaume Hummel, Stéphanie Lalande, et al. 2019. "Plant RNases T2, but Not Dicer-like Proteins, Are Major Players of tRNA-Derived Fragments Biogenesis." *Nucleic Acids Research* 47 (2): 941–52. https://doi.org/10.1093/nar/gky1156.

Mencía, Ángeles, Silvia Modamio-Høybjør, Nick Redshaw, et al. 2009. "Mutations in the Seed Region of Human miR-96 Are Responsible for Nonsyndromic Progressive Hearing Loss." *Nature Genetics* 41 (5): 609–13. https://doi.org/10.1038/ng.355.

Michlewski, Gracjan, and Javier F. Cáceres. 2019. "Post-Transcriptional Control of miRNA Biogenesis." *RNA* 25 (1): 1–16. https://doi.org/10.1261/rna.068692.118.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, et al. 2021. "Pfam: The Protein Families Database in 2021." *Nucleic Acids Research* 49 (D1): D412–19. https://doi.org/10.1093/nar/gkaa913.

Mleczko, Anna M., Piotr Celichowski, and Kamilla Bąkowska-Żywicka. 2018. "Transfer RNA-Derived Fragments Target and Regulate Ribosome-Associated Aminoacyl-Transfer RNA Synthetases." *Biochimica et Biophysica Acta (BBA) - Gene*

Regulatory Mechanisms 1861 (7): 647–56. https://doi.org/10.1016/j.bbagrm.2018.06.001.

Mo, Xiaohui, Shujuan Du, Xiaoting Chen, et al. 2020. "Lactate Induces Production of the tRNAHis Half to Promote B-Lymphoblastic Cell Proliferation." *Molecular Therapy* 28 (11): 2442–57. https://doi.org/10.1016/j.ymthe.2020.09.010.

Molina, M. Dolores, Dema Abduljabbar, Anna Guixeras, Susanna Fraguas, and Francesc Cebrià. 2023. "LIM-HD Transcription Factors Control Axial Patterning and Specify Distinct Neuronal and Intestinal Cell Identities in Planarians." *Open Biology* 13 (12). https://doi.org/10.1098/rsob.230327.

Molina, M. Dolores, and Francesc Cebrià. 2021. "Decoding Stem Cells: An Overview on Planarian Stem Cell Heterogeneity and Lineage Progression." *Biomolecules* 11 (10): 1532. https://doi.org/10.3390/biom11101532.

Molinaro, Alyssa M., and Bret J. Pearson. 2016. "In Silico Lineage Tracing through Single Cell Transcriptomics Identifies a Neural Stem Cell Population in Planarians." *Genome Biology* 17 (1): 87. https://doi.org/10.1186/s13059-016-0937-9.

Montoya, Michelle. 2014. "TALlying Lipid-Protein Interactions." *Nature Structural & Molecular Biology* 21 (1): 19–19. https://doi.org/10.1038/nsmb.2761.

Nakashima, A., H. Takaku, H. S. Shibata, et al. 2007. "Gene Silencing by the tRNA Maturase tRNase ZL under the Direction of Small-Guide RNA." *Gene Therapy* 14 (1): 78–85. https://doi.org/10.1038/sj.gt.3302841.

Nashimoto, Masayuki. 2022. "TRUE Gene Silencing." *International Journal of Molecular Sciences* 23 (10): 5387. https://doi.org/10.3390/ijms23105387.

Nielsen, Cydney B., Noam Shomron, Rickard Sandberg, Eran Hornstein, Jacob Kitzman, and Christopher B. Burge. 2007. "Determinants of Targeting by Endogenous and Exogenous microRNAs and siRNAs." *RNA* 13 (11): 1894–910. https://doi.org/10.1261/rna.768207.

Nielsen, Morten Muhlig, and Jakob Skou Pedersen. 2021. "miRNA Activity Inferred from Single Cell mRNA Expression." *Scientific Reports* 11 (1): 9170. https://doi.org/10.1038/s41598-021-88480-5.

Nikonorova, Inna A., Juan Wang, Alexander L. Cope, et al. 2022. "Isolation, Profiling, and Tracking of Extracellular Vesicle Cargo in Caenorhabditis Elegans." *Current Biology* 32 (9): 1924-1936.e6. https://doi.org/10.1016/j.cub.2022.03.005.

Nishie, Tomomi, Yoshio Ohta, Emi Shirai, et al. 2023. "Identification of *TEKTIN1* -expressing Multiciliated Cells during Spontaneous Differentiation of Non-human Primate Embryonic Stem Cells." *Genes to Cells* 28 (7): 516–25. https://doi.org/10.1111/gtc.13031.

Niu, Kaimeng, Hao Xu, Yuanyi Zhou Xiong, et al. 2021. “Canonical and Early Lineage-Specific Stem Cell Types Identified in Planarian SirNeoblasts.” *Cell Regeneration* 10 (1): 15. https://doi.org/10.1186/s13619-021-00076-6.

Nogi, Taisaku, and Michael Levin. 2005. “Characterization of Innexin Gene Expression and Functional Roles of Gap-Junctional Communication in Planarian Regeneration.” *Developmental Biology* 287 (2): 314–35. https://doi.org/10.1016/j.ydbio.2005.09.002.

Nowotarski, Stephanie H., Erin L. Davies, Sofia M. C. Robb, et al. 2021. “Planarian Anatomy Ontology: A Resource to Connect Data within and across Experimental Platforms.” *Development* 148 (15). https://doi.org/10.1242/dev.196097.

Okamura, Katsutomo, Wei-Jen Chung, J. Graham Ruby, Huili Guo, David P. Bartel, and Eric C. Lai. 2008. “The Drosophila Hairpin RNA Pathway Generates Endogenous Short Interfering RNAs.” *Nature* 453 (7196): 803–6. https://doi.org/10.1038/nature07015.

Olgun, Gulden, Vishaka Gopalan, and Sridhar Hannenhalli. 2022. “miRSCAPE - Inferring miRNA Expression from scRNA-Seq Data.” *iScience* 25 (9): 104962. https://doi.org/10.1016/j.isci.2022.104962.

Onishi, Ryo, Soichiro Yamanaka, and Mikiko C. Siomi. 2021. “piRNA- and siRNA-mediated Transcriptional Repression in *Drosophila* , Mice, and Yeast: New Insights and Biodiversity.” *EMBO Reports* 22 (10). https://doi.org/10.15252/embr.202153062.

Ooi, A. G. Lisa, Debashis Sahoo, Maddalena Adorno, Yulei Wang, Irving L. Weissman, and Christopher Y. Park. 2010. “MicroRNA-125b Expands Hematopoietic Stem Cells and Enriches for the Lymphoid-Balanced and Lymphoid-Biased Subsets.” *Proceedings of the National Academy of Sciences* 107 (50): 21505–10. https://doi.org/10.1073/pnas.1016218107.

Oulas, Anastasis, Nestoras Karathanasis, Annita Louloupi, et al. 2015. *Prediction of miRNA Targets*. https://doi.org/10.1007/978-1-4939-2291-8_13.

Oviedo, Néstor J., and Michael Levin. 2007. “*Smedinx-11* Is a Planarian Stem Cell Gap Junction Gene Required for Regeneration and Homeostasis.” *Development* 134 (17): 3121–31. https://doi.org/10.1242/dev.006635.

Oviedo, Néstor J., Junji Morokuma, Peter Walentek, et al. 2010. “Long-Range Neural and Gap Junction Protein-Mediated Cues Control Polarity during Planarian Regeneration.” *Developmental Biology* 339 (1): 188–99. https://doi.org/10.1016/j.ydbio.2009.12.012.

Owlarn, Suthira, and Kerstin Bartscherer. 2016. “Go Ahead, Grow a Head! A Planarian’s Guide to Anterior Regeneration.” *Regeneration* 3 (3): 139–55. https://doi.org/10.1002/reg2.56.

Ozata, Deniz M., Ildar Gainetdinov, Ansgar Zoch, Dónal O’Carroll, and Phillip D. Zamore. 2019. “PIWI-Interacting RNAs: Small RNAs with Big Functions.”

Nature Reviews Genetics 20 (2): 89–108. https://doi.org/10.1038/s41576-018-0073-3.

Palakodeti, Dasaradhi, Magda Smielewska, and Brenton R. Graveley. 2006. "MicroRNAs from the Planarian *Schmidtea Mediterranea* : A Model System for Stem Cell Biology." *RNA* 12 (9): 1640–49. https://doi.org/10.1261/rna.117206.

Palakodeti, Dasaradhi, Magda Smielewska, Yi-Chien Lu, Gene W. Yeo, and Brenton R. Graveley. 2008. "The PIWI Proteins SMEDWI-2 and SMEDWI-3 Are Required for Stem Cell Function and piRNA Expression in Planarians." *RNA* 14 (6): 1174–86. https://doi.org/10.1261/rna.1085008.

Panteleimon Rompolas, Ramila S. Patel-King, and Stephen M. King. 2009. "Schmidtea Mediterranea." In *Methods in Cell Biology*, vol. 93. Elsevier. https://doi.org/10.1016/S0091-679X(08)93004-1.

Park, Chanyoung, Kwadwo E. Owusu-Boaitey, Giselle M. Valdes, and Peter W. Reddien. 2023. "Fate Specification Is Spatially Intermingled across Planarian Stem Cells." *Nature Communications* 14 (1): 7422. https://doi.org/10.1038/s41467-023-43267-2.

Park, Joonhyeong, Se Hee Ahn, Myung Geun Shin, Hak Kyun Kim, and Suhwan Chang. 2020. "tRNA-Derived Small RNAs: Novel Epigenetic Regulators." *Cancers* 12 (10): 2773. https://doi.org/10.3390/cancers12102773.

Paul, Lukas, Petra Kubala, Gudrun Horner, et al. 2016. "SIRVs: Spike-In RNA Variants as External Isoform Controls in RNA-Sequencing." Preprint, Molecular Biology, October 13. https://doi.org/10.1101/080747.

Pauls, Dennis, Christine Blechschmidt, Felix Frantzmann, Basil el Jundi, and Mareike Selcho. 2018. "A Comprehensive Anatomical Map of the Peripheral Octopaminergic/Tyraminergic System of Drosophila Melanogaster." *Scientific Reports* 8 (1): 15314. https://doi.org/10.1038/s41598-018-33686-3.

Pawlicka, Kamila, Patrick McDevitt Perrigue, and Jan Barciszewski. 2017. "Regulatory RNAs in Planarians." *Acta Biochimica Polonica* 63 (4). https://doi.org/10.18388/abp.2016_1355.

Pearson, Bret J. 2022. "Finding the Potency in Planarians." *Communications Biology* 5 (1): 970. https://doi.org/10.1038/s42003-022-03905-9.

Peiris, T. Harshani, Katrina K. Hoyer, and Néstor J. Oviedo. 2014. "Innate Immune System and Tissue Regeneration in Planarians: An Area Ripe for Exploration." *Seminars in Immunology* 26 (4): 295–302. https://doi.org/10.1016/j.smim.2014.06.005.

Peiris, T. Harshani, Daniel Ramirez, Paul G. Barghouth, and Néstor J. Oviedo. 2016. "The Akt Signaling Pathway Is Required for Tissue Maintenance and Regeneration in Planarians." *BMC Developmental Biology* 16 (1): 7. https://doi.org/10.1186/s12861-016-0107-z.

Pellettieri, Jason, and Alejandro Sánchez Alvarado. 2007. “Cell Turnover and Adult Tissue Homeostasis: From Humans to Planarians.” *Annual Review of Genetics* 41 (1): 83–105. https://doi.org/10.1146/annurev.genet.41.110306.130244.

Pérez-Posada, Alberto, Helena García-Castro, Elena Emili, et al. 2025. “Multimodal Single Cell Analyses Reveal Gene Networks of Planarian Stem Cell Differentiation.” *Nature Communications* 16 (1): 10683. https://doi.org/10.1038/s41467-025-65712-0.

Plass, Mireya, Jordi Solana, F. Alexander Wolf, et al. 2018. “Cell Type Atlas and Lineage Tree of a Whole Complex Animal by Single-Cell Transcriptomics.” *Science* 360 (6391). https://doi.org/10.1126/science.aaq1723.

Pliatsika, Venetia, Phillipe Loher, Rogan Magee, et al. 2018. “MINTbase v2.0: A Comprehensive Database for tRNA-Derived Fragments That Includes Nuclear and Mitochondrial Fragments from All The Cancer Genome Atlas Projects.” *Nucleic Acids Research* 46 (D1): D152–59. https://doi.org/10.1093/nar/gkx1075.

Protasio, Anna V., Stijn Van Dongen, Julie Collins, et al. 2017. “MiR-277/4989 Regulate Transcriptional Landscape during Juvenile to Adult Transition in the Parasitic Helminth Schistosoma Mansoni.” *PLOS Neglected Tropical Diseases* 11 (5): e0005559. https://doi.org/10.1371/journal.pntd.0005559.

Quevillon, E., V. Silventoinen, S. Pillai, et al. 2005. “InterProScan: Protein Domains Identifier.” *Nucleic Acids Research* 33 (Web Server): W116–20. https://doi.org/10.1093/nar/gki442.

Rand, Tim A., Sean Petersen, Fenghe Du, and Xiaodong Wang. 2005. “Argonaute2 Cleaves the Anti-Guide Strand of siRNA during RISC Activation.” *Cell* 123 (4): 621–29. https://doi.org/10.1016/j.cell.2005.10.020.

Reddien, Peter W. 2018. “The Cellular and Molecular Basis for Planarian Regeneration.” In *Cell*, vol. 175. no. 2. Cell Press, October. https://doi.org/10.1016/j.cell.2018.09.021.

Reddien, Peter W. 2021a. “Positional Information and Stem Cells Combine to Result in Planarian Regeneration.” *Cold Spring Harbor Perspectives in Biology*, September, a040717. https://doi.org/10.1101/cshperspect.a040717.

Reddien, Peter W. 2021b. “Principles of Regeneration Revealed by the Planarian Eye.” *Current Opinion in Cell Biology* 73 (December): 19–25. https://doi.org/10.1016/j.ceb.2021.05.001.

Reddien, Peter W., Néstor J. Oviedo, Joya R. Jennings, James C. Jenkin, and Alejandro Sánchez Alvarado. 2005. “SMEDWI-2 Is a PIWI-Like Protein That Regulates Planarian Stem Cells.” *Science* 310 (5752): 1327–30. https://doi.org/10.1126/science.1116110.

Resch, Alissa M., and Dasaradhi Palakodeti. 2012. “Small RNA Pathways in Schmidtea Mediterranea.” *The International Journal of Developmental Biology* 56 (1-2–3): 67–74. https://doi.org/10.1387/ijdb.113436ar.

Rink, Jochen C. 2013. “Stem Cell Systems and Regeneration in Planaria.” *Development Genes and Evolution* 223 (1–2): 67–84. https://doi.org/10.1007/s00427-012-0426-4.

Rink, Jochen C., Hanh Thi-Kim Vu, and Alejandro Sánchez Alvarado. 2011. “The Maintenance and Regeneration of the Planarian Excretory System Are Regulated by EGFR Signaling.” *Development* 138 (17): 3769–80. https://doi.org/10.1242/dev.066852.

Riquelme, Ismael, Pablo Pérez-Moreno, Pablo Letelier, Priscilla Brebi, and Juan Carlos Roa. 2021. “The Emerging Role of PIWI-Interacting RNAs (piRNAs) in Gastrointestinal Cancers: An Updated Perspective.” *Cancers* 14 (1): 202. https://doi.org/10.3390/cancers14010202.

Rivals, Isabelle, Léon Personnaz, Lieng Taing, and Marie-Claude Potier. 2007. “Enrichment or Depletion of a GO Category within a Class of Genes: Which Test?” *Bioinformatics* 23 (4): 401–7. https://doi.org/10.1093/bioinformatics/btl633.

Robb, Sofia M. C., Eric Ross, and Alejandro Sánchez Alvarado. 2008. “SmedGD: The Schmidtea Mediterranea Genome Database.” *Nucleic Acids Research* 36 (SUPPL. 1). https://doi.org/10.1093/nar/gkm684.

Roberts-Galbraith, Rachel H., John L. Brubacher, and Phillip A. Newmark. 2016. “A Functional Genomics Screen in Planarians Reveals Regulators of Whole-Brain Regeneration.” *eLife* 5 (September). https://doi.org/10.7554/eLife.17002.

Ross, Eric, David Blair, Carlos Guerrero-Hernández, and Alejandro Sánchez Alvarado. 2016. “Comparative and Transcriptome Analyses Uncover Key Aspects of Coding- and Long Noncoding RNAs in Flatworm Mitochondrial Genomes.” *G3 Genes|Genomes|Genetics* 6 (5): 1191–200. https://doi.org/10.1534/g3.116.028175.

Ross, Kelly G., Sarai Alvarez Zepeda, Mohammad A. Auwal, Audrey K. Garces, Sydney Roman, and Ricardo M. Zayas. 2024. “The Role of *Polycystic Kidney Disease-like* Homologs in Planarian Nervous System Regeneration and Function.” July. https://doi.org/10.1101/2024.07.17.603829.

Rossmanith, Walter. 2011. “Localization of Human RNase Z Isoforms: Dual Nuclear/Mitochondrial Targeting of the ELAC2 Gene Product by Alternative Translation Initiation.” *PLoS ONE* 6 (4): e19152. https://doi.org/10.1371/journal.pone.0019152.

Rozanski, Andrei, Hong Kee Moon, Holger Brandl, et al. 2019. “PlanMine 3.0 - Improvements to a Mineable Resource of Flatworm Biology and Biodiversity.” *Nucleic Acids Research* 47 (D1): D812–20. https://doi.org/10.1093/nar/gky1070.

Saad, Luiza O., Thomas F. Cooke, Kutay D. Atabay, Peter W. Reddien, and Federico D. Brown. 2025. “Reduced Adult Stem Cell Fate Specification Led to Eye Reduction in Cave Planarians.” *Nature Communications* 16 (1): 304. https://doi.org/10.1038/s41467-024-54478-6.

Safa, Amin, Zahra Bahroudi, Hamed Shoorei, et al. 2020. “miR-1: A Comprehensive Review of Its Role in Normal Development and Diverse Disorders.” *Biomedicine & Pharmacotherapy* 132 (December): 110903. https://doi.org/10.1016/j.biopha.2020.110903.

Saliba, Antoine-Emmanuel, Alexander J. Westermann, Stanislaw A. Gorski, and Jörg Vogel. 2014. “Single-Cell RNA-Seq: Advances and Future Challenges.” *Nucleic Acids Research* 42 (14): 8845–60. https://doi.org/10.1093/nar/gku555.

Saoura, Makenzie, Christopher A. Powell, Robert Kopajtich, et al. 2019. “Mutations in ELAC2 Associated with Hypertrophic Cardiomyopathy Impair Mitochondrial tRNA 3′-end Processing.” *Human Mutation* 40 (10): 1731–48. https://doi.org/10.1002/humu.23777.

Sasidharan, Vidyanand, Laura Ancellotti, Viraj Doddihal, et al. 2026. “Extracellular Vesicles Mediate Stem Cell Signaling and Systemic RNAi in Planarians.” *Science Advances* 12 (6): eady1461. https://doi.org/10.1126/sciadv.ady1461.

Sasidharan, Vidyanand, Yi-Chien Lu, Dhiru Bansal, et al. 2013. “Identification of Neoblast- and Regeneration-Specific miRNAs in the Planarian *Schmidtea Mediterranea*.” *RNA* 19 (10): 1394–404. https://doi.org/10.1261/rna.038653.113.

Sasidharan, Vidyanand, Srujan Marepally, Sarah A. Elliott, et al. 2017. “The *miR-124* Family of microRNAs Is Crucial for Regeneration of the Brain and Visual System in the Planarian *Schmidtea Mediterranea*.” *Development* 144 (18): 3211–23. https://doi.org/10.1242/dev.144758.

Schad, Erik G., and Christian P. Petersen. 2022. “Planarian Dorsoventral Netrins Control a Muscle Midline Signaling Center and Regulate Blastema Formation.” September. https://doi.org/10.1101/2022.08.31.506052.

Schiffer, S. 2002. “Assigning a Function to a Conserved Group of Proteins: The tRNA 3’-Processing Enzymes.” *The EMBO Journal* 21 (11): 2769–77. https://doi.org/10.1093/emboj/21.11.2769.

Schorn, Andrea J., Michael J. Gutbrod, Chantal LeBlanc, and Rob Martienssen. 2017. “LTR-Retrotransposon Control by tRNA-Derived Small RNAs.” *Cell* 170 (1): 61-71.e11. https://doi.org/10.1016/j.cell.2017.06.013.

Scimone, M. Lucila, Jennifer K. Cloutier, Chloe L. Maybrun, and Peter W. Reddien. 2022. “The Planarian Wound Epidermis Gene Equinox Is Required for Blastema Formation in Regeneration.” *Nature Communications* 13 (1): 2726. https://doi.org/10.1038/s41467-022-30412-6.

Scimone, M. Lucila, Lauren E. Cote, and Peter W. Reddien. 2017. “Orthogonal Muscle Fibres Have Different Instructive Roles in Planarian Regeneration.” *Nature* 551 (7682): 623–28. https://doi.org/10.1038/nature24660.

Scimone, M. Lucila, Kellie M. Kravarik, Sylvain W. Lapan, and Peter W. Reddien. 2014. “Neoblast Specialization in Regeneration of the Planarian Schmidtea

Mediterranea." *Stem Cell Reports* 3 (2): 339–52. https://doi.org/10.1016/j.stemcr.2014.06.001.

Scimone, M. Lucila, Mansi Srivastava, George W. Bell, and Peter W. Reddien. 2011. "A Regulatory Program for Excretory System Regeneration in Planarians." *Development* 138 (20): 4387–98. https://doi.org/10.1242/dev.068098.

Scimone, M. Lucila, Omri Wurtzel, Kathryn Malecek, et al. 2018. "foxF-1 Controls Specification of Non-Body Wall Muscle and Phagocytic Cells in Planarians." *Current Biology* 28 (23): 3787-3801.e6. https://doi.org/10.1016/j.cub.2018.10.030.

Selitsky, Sara R., and Praveen Sethupathy. 2015. "tDRmapper: Challenges and Solutions to Mapping, Naming, and Quantifying tRNA-Derived RNAs from Human Small RNA-Sequencing Data." *BMC Bioinformatics* 16 (1): 354. https://doi.org/10.1186/s12859-015-0800-0.

Seok, Heeyoung, Juyoung Ham, Eun-Sook Jang, and Sung Wook Chi. 2016. "MicroRNA Target Recognition: Insights from Transcriptome-Wide Non-Canonical Interactions." *Molecules and Cells* 39 (5): 375–81. https://doi.org/10.14348/molcells.2016.0013.

Settles, Skylar E., Kathleen E. Miller, and Rachel H. Roberts-Galbraith. 2025. "A Parenchymal Niche Regulates Pluripotent Stem Cell Function in Planarians." Preprint, Cell Biology, August 2. https://doi.org/10.1101/2025.08.01.668211.

Sharma, Upasna, Colin C. Conine, Jeremy M. Shea, et al. 2016. "Biogenesis and Function of tRNA Fragments during Sperm Maturation and Fertilization in Mammals." *Science* 351 (6271): 391–96. https://doi.org/10.1126/science.aad6780.

Shaw, Annabelle, Kamal Ajit, and Monika Gullerova. 2024. "Y RNA-Derived Fragments in a Complex with YBX1 Modulate PARP1 Residency at DNA Double Strand Breaks." June. https://doi.org/10.1101/2024.06.18.599493.

Sheu-Gruttadauria, Jessica, Yao Xiao, Luca Fr Gebert, and Ian J. MacRae. 2019. "Beyond the Seed: Structural Basis for Supplementary Micro RNA Targeting by Human Argonaute2." *The EMBO Journal* 38 (13): e101153. https://doi.org/10.15252/embj.2018101153.

Shi, Junchao, Eun-A. Ko, Kenton M. Sanders, Qi Chen, and Tong Zhou. 2018. "SPORTS1.0: A Tool for Annotating and Profiling Non-Coding RNAs Optimized for rRNA- and tRNA-Derived Small RNAs." *Genomics, Proteomics & Bioinformatics* 16 (2): 144–51. https://doi.org/10.1016/j.gpb.2018.04.004.

Shibata, Norito, Makoto Kashima, Taisuke Ishiko, et al. 2016. "Inheritance of a Nuclear PIWI from Pluripotent Stem Cells by Somatic Descendants Ensures Differentiation by Silencing Transposons in Planarian." *Developmental Cell* 37 (3): 226–37. https://doi.org/10.1016/j.devcel.2016.04.009.

Shumate, Alaina, Brandon Wong, Geo Pertea, and Mihaela Pertea. 2022. "Improved Transcriptome Assembly Using a Hybrid of Long and Short Reads with

StringTie." *PLoS Computational Biology* 18 (6). https://doi.org/10.1371/journal.pcbi.1009730.

Siira, Stefan J., Giulia Rossetti, Tara R. Richman, et al. 2018. "Concerted Regulation of Mitochondrial and Nuclear Non-coding RNA by a Dual-targeted RNase Z." *EMBO Reports* 19 (10). https://doi.org/10.15252/embr.201846198.

Smith, Christopher Michael, and Gyorgy Hutvagner. 2022. "A Comparative Analysis of Single Cell Small RNA Sequencing Data Reveals Heterogeneous isomiR Expression and Regulation." *Scientific Reports* 12 (1): 2834. https://doi.org/10.1038/s41598-022-06876-3.

Smith-Unna, Richard, Chris Boursnell, Rob Patro, Julian M. Hibberd, and Steven Kelly. 2016. "TransRate: Reference-Free Quality Assessment of de Novo Transcriptome Assemblies." *Genome Research* 26 (8): 1134–44. https://doi.org/10.1101/gr.196469.115.

Squair, Jordan W., Matthieu Gautier, Claudia Kathe, et al. 2021. "Confronting False Discoveries in Single-Cell Differential Expression." *Nature Communications* 12 (1): 5692. https://doi.org/10.1038/s41467-021-25960-2.

Stuart, Tim, Andrew Butler, Paul Hoffman, et al. 2019. "Comprehensive Integration of Single-Cell Data." *Cell* 177 (7): 1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.

Stubenhaus, Bradford M., John P. Dustin, Emily R. Neverett, et al. 2016. "Light-Induced Depigmentation in Planarians Models the Pathophysiology of Acute Porphyrias." *eLife* 5 (May): e14175. https://doi.org/10.7554/eLife.14175.

Su, Zhangli, Canan Kuscu, Asrar Malik, Etsuko Shibata, and Anindya Dutta. 2019. "Angiogenin Generates Specific Stress-Induced tRNA Halves and Is Not Involved in tRF-3–Mediated Gene Silencing." *Journal of Biological Chemistry* 294 (45): 16930–41. https://doi.org/10.1074/jbc.RA119.009272.

Su, Zhangli, Briana Wilson, Pankaj Kumar, and Anindya Dutta. 2020. "Noncanonical Roles of tRNAs: tRNA Fragments and Beyond." *Annual Review of Genetics* 54 (1): 47–69. https://doi.org/10.1146/annurev-genet-022620-101840.

Sun, Yu H., Brent Lee, and Xin Zhiguo Li. 2022. "The Birth of piRNAs: How Mammalian piRNAs Are Produced, Originated, and Evolved." *Mammalian Genome* 33 (2): 293–311. https://doi.org/10.1007/s00335-021-09927-8.

Sweeney, Blake A., Anton I. Petrov, Carlos E. Ribas, et al. 2021. "RNAcentral 2021: Secondary Structure Integration, Improved Sequence Search and New Member Databases." *Nucleic Acids Research* 49 (D1): D212–20. https://doi.org/10.1093/nar/gkaa921.

Taft, Ryan J., Evgeny A. Glazov, Timo Lassmann, Yoshihide Hayashizaki, Piero Carninci, and John S. Mattick. 2009. "Small RNAs Derived from snoRNAs." *RNA* 15 (7): 1233–40. https://doi.org/10.1261/rna.1528909.

Takenaka, Yoshika, Asuka Yamada, Yoshihisa Tomioka, Yasutoshi Akiyama, and Pavel Ivanov. 2025. “RNase L Produces tRNA-Derived RNAs That Contribute to Translation Inhibition.” *RNA* 31 (7): 961–72. https://doi.org/10.1261/rna.080419.125.

Tautz, Lutz, David A. Critton, and Stefan Grotegut. 2013. “Protein Tyrosine Phosphatases: Structure, Function, and Implication in Human Disease.” In *Phosphatase Modulators*, edited by José Luis Millán, vol. 1053. Methods in Molecular Biology. Humana Press. https://doi.org/10.1007/978-1-62703-562-0_13.

Telonis, Aristeidis G., Phillipe Loher, Yohei Kirino, and Isidore Rigoutsos. 2016. “Consequential Considerations When Mapping tRNA Fragments.” *BMC Bioinformatics* 17 (1): 123. https://doi.org/10.1186/s12859-016-0921-0.

Théry, Clotilde, Laurence Zitvogel, and Sebastian Amigorena. 2002. “Exosomes: Composition, Biogenesis and Function.” *Nature Reviews Immunology* 2 (8): 569–79. https://doi.org/10.1038/nri855.

Thi-Kim Vu, Hanh, Jochen C. Rink, Sean A. McKinney, et al. 2015. “Stem Cells and Fluid Flow Drive Cyst Formation in an Invertebrate Excretory Organ.” *eLife* 4 (June): e07405. https://doi.org/10.7554/eLife.07405.

Thiruvalluvan, Manish, Paul G. Barghouth, Assaf Tsur, Limor Broday, and Néstor J. Oviedo. 2018. “SUMOylation Controls Stem Cell Proliferation and Regional Cell Death through Hedgehog Signaling in Planarians.” *Cellular and Molecular Life Sciences* 75 (7): 1285–301. https://doi.org/10.1007/s00018-017-2697-4.

Thompson, Debrah M., and Roy Parker. 2009. “The RNase Rny1p Cleaves tRNAs and Promotes Cell Death during Oxidative Stress in *Saccharomyces Cerevisiae*.” *Journal of Cell Biology* 185 (1): 43–50. https://doi.org/10.1083/jcb.200811119.

Thomson, Daniel W., Katherine A. Pillman, Matthew L. Anderson, et al. 2015. “Assessing the Gene Regulatory Properties of Argonaute-Bound Small RNAs of Diverse Genomic Origin.” *Nucleic Acids Research* 43 (1): 470–81. https://doi.org/10.1093/nar/gku1242.

Tonks, Nicholas K. 2006. “Protein Tyrosine Phosphatases: From Genes, to Function, to Disease.” *Nature Reviews Molecular Cell Biology* 7 (11): 833–46. https://doi.org/10.1038/nrm2039.

Tu, Kimberly C., Li-Chun Cheng, Hanh Tk Vu, et al. 2015. “Egr-5 Is a Post-Mitotic Regulator of Planarian Epidermal Differentiation.” *eLife* 4 (October): e10501. https://doi.org/10.7554/eLife.10501.

Tu, Kimberly C., Li-Chun Cheng, Hanh TK Vu, et al. 2015. “Egr-5 Is a Post-Mitotic Regulator of Planarian Epidermal Differentiation.” *eLife* 4 (October). https://doi.org/10.7554/eLife.10501.

Valkov, Nedyalka, and Saumya Das. 2020. *Y RNAs: Biogenesis, Function and Implications for the Cardiovascular System*. https://doi.org/10.1007/978-981-15-1671-9_20.

van Wolfswinkel, Josien C., Daniel E. Wagner, and Peter W. Reddien. 2014. "Single-Cell Analysis Reveals Functionally Distinct Classes within the Planarian Stem Cell Compartment." *Cell Stem Cell* 15 (3): 326–39. https://doi.org/10.1016/j.stem.2014.06.007.

Vásquez-Doorman, Constanza, and Christian P. Petersen. 2016. "The NuRD Complex Component *P66* Suppresses Photoreceptor Neuron Regeneration in Planarians." *Regeneration* 3 (3): 168–78. https://doi.org/10.1002/reg2.58.

Vij, Shubha, Jochen C. Rink, Hao Kee Ho, et al. 2012. "Evolutionarily Ancient Association of the FoxJ1 Transcription Factor with the Motile Ciliogenic Program." *PLoS Genetics* 8 (11): e1003019. https://doi.org/10.1371/journal.pgen.1003019.

Vogg, Matthias C., Suthira Owlarn, Yuvia A. Pérez Rico, et al. 2014. "Stem Cell-Dependent Formation of a Functional Anterior Regeneration Pole in Planarians Requires Zic and Forkhead Transcription Factors." *Developmental Biology* 390 (2): 136–48. https://doi.org/10.1016/j.ydbio.2014.03.016.

Wagner, Daniel E., Irving E. Wang, and Peter W. Reddien. 2011. "Clonogenic Neoblasts Are Pluripotent Adult Stem Cells That Underlie Planarian Regeneration." *Science* 332 (6031): 811–16. https://doi.org/10.1126/science.1203983.

Wajahat, Maliha, Cameron Peter Bracken, and Ayla Orang. 2021. "Emerging Functions for snoRNAs and snoRNA-Derived Fragments." *International Journal of Molecular Sciences* 22 (19): 10193. https://doi.org/10.3390/ijms221910193.

Wang, Qinghua, Xinxin Sun, Jing Xiao, et al. 2022. "Djptpn11 Is Indispensable for Planarian Regeneration by Affecting Early Wound Response Genes Expression and the Wnt Pathway." *Biochimie* 201 (October): 184–95. https://doi.org/10.1016/j.biochi.2022.07.007.

Wang, Xiaomeng, Xiaotong Wu, Hao Wu, et al. 2023. "Neural Adaption in Midbrain GABAergic Cells Contributes to High-Fat Diet–Induced Obesity." *Science Advances* 9 (44). https://doi.org/10.1126/sciadv.adh2884.

Wang, Yu, Hongxia Li, Qixin Sun, and Yingyin Yao. 2016. "Characterization of Small RNAs Derived from tRNAs, rRNAs and snoRNAs and Their Response to Heat Stress in Wheat Seedlings." *PLOS ONE* 11 (3): e0150933. https://doi.org/10.1371/journal.pone.0150933.

Watanabe, Toshiaki, Yasushi Totoki, Atsushi Toyoda, et al. 2008. "Endogenous siRNAs from Naturally Formed dsRNAs Regulate Transcripts in Mouse Oocytes." *Nature* 453 (7194): 539–43. https://doi.org/10.1038/nature06908.

Weick, Eva-Maria, and Eric A. Miska. 2014. "piRNAs: From Biogenesis to Function." *Development* 141 (18): 3458–71. https://doi.org/10.1242/dev.094037.

Witchley, Jessica N., Mirjam Mayer, Daniel E. Wagner, Jared H. Owen, and Peter W. Reddien. 2013. “Muscle Cells Provide Instructions for Planarian Regeneration.” *Cell Reports* 4 (4): 633–41. https://doi.org/10.1016/j.celrep.2013.07.022.

Wolf, F. Alexander, Philipp Angerer, and Fabian J. Theis. 2018. “SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis.” *Genome Biology* 19 (1): 15. https://doi.org/10.1186/s13059-017-1382-0.

Wolf, F. Alexander, Fiona K. Hamey, Mireya Plass, et al. 2019. “PAGA: Graph Abstraction Reconciles Clustering with Trajectory Inference through a Topology Preserving Map of Single Cells.” *Genome Biology* 20 (1): 59. https://doi.org/10.1186/s13059-019-1663-x.

Wong, Lily L., Christina G. Bruxvoort, Nicholas I. Cejda, Matthew R. Delaney, Jannette Rodriguez Otero, and David J. Forsthoefel. 2022. “Intestine-Enriched Apolipoprotein b Orthologs Are Required for Stem Cell Progeny Differentiation and Regeneration in Planarians.” *Nature Communications* 13 (1): 3803. https://doi.org/10.1038/s41467-022-31385-2.

Wurtzel, Omri, Lauren E. Cote, Amber Poirier, Rahul Satija, Aviv Regev, and Peter W. Reddien. 2015. “A Generic and Cell-Type-Specific Wound Response Precedes Regeneration in Planarians.” *Developmental Cell* 35 (5): 632–45. https://doi.org/10.1016/j.devcel.2015.11.004.

Wurtzel, Omri, Isaac M. Oderberg, and Peter W. Reddien. 2017. “Planarian Epidermal Stem Cells Respond to Positional Cues to Promote Cell-Type Diversity.” *Developmental Cell* 40 (5): 491-504.e5. https://doi.org/10.1016/j.devcel.2017.02.008.

Xue, Chenyang, Junshan Tian, Yanhong Chen, and Zhongmin Liu. 2024. “Structural Insights into Human ELAC2 as a tRNA 3′ Processing Enzyme.” *Nucleic Acids Research* 52 (21): 13434–46. https://doi.org/10.1093/nar/gkae1014.

Yamamoto, Hiroshi, and Kiyokazu Agata. 2011. “Optic Chiasm Formation in Planarian I: Cooperative *Netrin*- and *Robo* -mediated Signals Are Required for the Early Stage of Optic Chiasm Formation.” *Development, Growth & Differentiation* 53 (3): 300–311. https://doi.org/10.1111/j.1440-169X.2010.01234.x.

Yang, Na, Ruijun Li, Ruai Liu, et al. 2024. “The Emerging Function and Promise of tRNA-Derived Small RNAs in Cancer.” *Journal of Cancer* 15 (6): 1642–56. https://doi.org/10.7150/jca.89219.

Young, Matthew D., Matthew J. Wakefield, Gordon K. Smyth, and Alicia Oshlack. 2010. “Gene Ontology Analysis for RNA-Seq: Accounting for Selection Bias.” *Genome Biology* 11 (2): R14. https://doi.org/10.1186/gb-2010-11-2-r14.

Yu, Dongliang, Xiaoxia Ma, Ziwei Zuo, Huizhong Wang, and Yijun Meng. 2018. “Classification of Transcription Boundary-Associated RNAs (TBARs) in Animals and Plants.” *Frontiers in Genetics* 9 (May). https://doi.org/10.3389/fgene.2018.00168.

Yu, Hui, Marcelo Rubinstein, and Malcolm J. Low. 2022. “Developmental Single-Cell Transcriptomics of Hypothalamic POMC Neurons Reveal the Genetic Trajectories of Multiple Neuropeptidergic Phenotypes.” *eLife* 11 (January). https://doi.org/10.7554/eLife.72883.

Yu, Jie, Natsuko Izumi, Yukihide Tomari, and Keisuke Shoji. 2024. “Autonomous Shaping of the piRNA Sequence Repertoire by Competition between Adjacent Ping-Pong Sites.” May. https://doi.org/10.1101/2024.05.22.595280.

Yuan, David, Alisha Ahamed, Josephine Burgin, et al. 2024. “The European Nucleotide Archive in 2023.” *Nucleic Acids Research* 52 (D1): D92–97. https://doi.org/10.1093/nar/gkad1067.

Yuasa-Kawada, Junichi, Mariko Kinoshita-Kawada, Yoshio Tsuboi, and Jane Y. Wu. 2022. “Neuronal Guidance Genes in Health and Diseases.” *Protein & Cell*, July 15, pwac030. https://doi.org/10.1093/procel/pwac030.

Zamudio, Jesse R., Timothy J. Kelly, and Phillip A. Sharp. 2014. “Argonaute-Bound Small RNAs from Promoter-Proximal RNA Polymerase II.” *Cell* 156 (5): 920–34. https://doi.org/10.1016/j.cell.2014.01.041.

Zaremba, Anastasiia, Małgorzata Marszałek-Zeńczak, Annasha Dutta, Anna Samelak-Czajka, and Paulina Jackowiak. 2025. “A Modular Pipeline for Evidence-Integrated Genome Annotation across Species: A Case Study on Schmidtea Mediterranea.” *Genomics* 117 (6): 111104. https://doi.org/10.1016/j.ygeno.2025.111104.

Zayas, Ricardo M., Francesc Cebrià, Tingxia Guo, Junjie Feng, and Phillip A. Newmark. 2010. “The Use of Lectins as Markers for Differentiated Secretory Cells in Planarians.” *Developmental Dynamics* 239 (11): 2888–97. https://doi.org/10.1002/dvdy.22427.

Zeilhofer, Hanns Ulrich, Mario A. Acuña, Jacinthe Gingras, and Gonzalo E. Yévenes. 2018. “Glycine Receptors and Glycine Transporters: Targets for Novel Analgesics?” *Cellular and Molecular Life Sciences* 75 (3): 447–65. https://doi.org/10.1007/s00018-017-2622-x.

Zeng, An, Hua Li, Longhua Guo, et al. 2018. “Prospectively Isolated Tetraspanin+ Neoblasts Are Adult Pluripotent Stem Cells Underlying Planaria Regeneration.” *Cell* 173 (7): 1593-1608.e20. https://doi.org/10.1016/j.cell.2018.05.006.

Zhang, Peijing, Wenyi Wu, Qi Chen, and Ming Chen. 2019. “Non-Coding RNAs and Their Integrated Networks.” *Journal of Integrative Bioinformatics* 16 (3). https://doi.org/10.1515/jib-2019-0027.

Zhang, Xiaochang, Rebecca Zabinsky, Yudong Teng, Mingxue Cui, and Min Han. 2011. “microRNAs Play Critical Roles in the Survival and Recovery of *Caenorhabditis Elegans* from Starvation-Induced L1 Diapause.” *Proceedings of the National Academy of Sciences* 108 (44): 17997–8002. https://doi.org/10.1073/pnas.1105982108.

Zheng, Guanqun, Yidan Qin, Wesley C. Clark, et al. 2015. “Efficient and Quantitative High-Throughput tRNA Sequencing.” *Nature Methods* 12 (9): 835–37. https://doi.org/10.1038/nmeth.3478.

Zheng, Hanxue, Hongbo Liu, Qian Xu, et al. 2021. “PI3K Plays an Essential Role in Planarian Regeneration and Tissue Maintenance.” *Frontiers in Cell and Developmental Biology* 9 (August): 649656. https://doi.org/10.3389/fcell.2021.649656.

Zhou, Xiaolai, Lirong Sun, Francisco Bastos De Oliveira, et al. 2015. “Prosaposin Facilitates Sortilin-Independent Lysosomal Trafficking of Progranulin.” *Journal of Cell Biology* 210 (6): 991–1002. https://doi.org/10.1083/jcb.201502029.

Zhu, Shu Jun, Stephanie E. Hallows, Ko W. Currie, ChangJiang Xu, and Bret J. Pearson. 2015. “A Mex3 Homolog Is Required for Differentiation during Planarian Stem Cell Lineage Development.” *eLife* 4 (June). https://doi.org/10.7554/eLife.07025.

Zong, Liang, Yan Zhu, Ruqiang Liang, and Hong-Bo Zhao. 2016. “Gap Junction Mediated miRNA Intercellular Transfer and Gene Regulation: A Novel Mechanism for Intercellular Genetic Communication.” *Scientific Reports* 6 (1): 19884. https://doi.org/10.1038/srep19884.

APPENDICES

INSTITUTE OF BIOORGANIC CHEMISTRY
Polish Academy of Sciences

Noskowskiego 12/14, 61-704 Poznań
tel.: +48 61 852 85 03, secretariat +48 61 852 89 19
fax: +48 61 852 05 32, e-mail: ibch@ibch.poznan.pl
REGON 000849327 VAT no. PL 7770002062
http://www.ibch.poznan.pl

MSc. Eng. Anastasiia Zaremba
Laboratory of Single Cell Analyses
Institute of Bioorganic Chemistry Polish Academy of Sciences

Statement of the doctoral candidate specifying contributions to the scientific paper included in the doctoral dissertation

Publication title: A modular pipeline for evidence-integrated genome annotation across species: A case study on *Schmidtea mediterranea*
Authors: Anastasiia Zaremba, Małgorzata Marszałek-Zenczak, Annasha Dutta, Anna Samelak-Czajka, Paulina Jackowiak (corresponding author)

Journal: Genomics
Year: 2025

I declare that my contribution to this publication included: (i) conceptualization and methodological design of the annotation strategy; (ii) implementation of the workflow; (iii) development and execution of automated post-processing and quality-control steps; (iv) preparation of validation and benchmarking analyses; (v) preparation of visualizations presenting the workflow and performance; and (vi) drafting the original version of the manuscript and contributing to subsequent revisions and editing.

ANASTASIIA ZAREMBA

INSTITUTE OF BIOORGANIC CHEMISTRY
Polish Academy of Sciences

Noskowskiego 12/14, 61-704 Poznań
tel.: +48 61 852 85 03, secretariat +48 61 852 89 19
fax: +48 61 852 05 32, e-mail: ibch@ibch.poznan.pl
REGON 000849327 VAT no. PL 7770002062
http://www.ibch.poznan.pl

Dr. habil. Paulina Jackowiak, Assoc. Prof.
Laboratory of Single Cell Analyses
Institute of Bioorganic Chemistry Polish Academy of Sciences

Statement of the corresponding author specifying contributions of the doctoral candidate to the scientific paper included in the doctoral dissertation

Publication title: A modular pipeline for evidence-integrated genome annotation across species: A case study on *Schmidtea mediterranea*
Authors: Anastasiia Zaremba, Małgorzata Marszałek-Zenczak, Annasha Dutta, Anna Samelak-Czajka, Paulina Jackowiak (corresponding author)

Journal: Genomics
Year: 2025

I declare that Ms. Anastasiia Zaremba's contribution to this publication included: (i) conceptualization and methodological design of the annotation strategy; (ii) implementation of the workflow; (iii) development and execution of automated post-processing and quality-control steps; (iv) preparation of validation and benchmarking analyses; (v) preparation of visualizations presenting the workflow and performance; and (vi) drafting the original version of the manuscript and contributing to subsequent revisions and editing.

Paulina Jackowiak

Genomics 117 (2025) 111104

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Research Paper

A modular pipeline for evidence-integrated genome annotation across species: A case study on *Schmidtea mediterranea*

Anastasiia Zaremba, Małgorzata Marszałek-Zeńczak, Annasha Dutta, Anna Samelak-Czajka, Paulina Jackowiak *

Laboratory of Single Cell Analyses, Institute of Bioorganic Chemistry Polish Academy of Sciences, Zygmunta Noskowskiego str. 12/14, 61-704 Poznań, Poland

ARTICLE INFO

Keywords:
Genome annotation
Transcriptome assembly
Schmidtea mediterranea
Non-model organisms
Functional annotation

ABSTRACT

Despite advancements in genome annotation tools, challenges persist for non-classical model organisms with limited genomic resources, such as *Schmidtea mediterranea*. To address these challenges, we developed a flexible and scalable genome annotation pipeline that integrates short-read (Illumina) and long-read (PacBio) sequencing technologies. The pipeline combines reference-based and *de novo* assembly methods, effectively handling genomic variability and alternative splicing events. To improve splice site detection accuracy, DeepSplice deep learning predictions are used. Functional annotation is conducted to filter out low-confidence transcripts and ensure biological relevance. Applying this pipeline to the asexual strain of *S. mediterranea* revealed thousands of previously undescribed putative genes and transcripts, and improved the existing gene models, highlighting its utility in annotating complex, underexplored genomes. The modularity and comprehensiveness of our pipeline ensure its adaptability for genome annotation across diverse species, making it a valuable tool for annotating genomes of non-model organisms and supporting broader genomic research. The source code and implementation details are available at https://github.com/Norreanea/SmedAnno.

1. Background

Understanding the genetic makeup of organisms is crucial for unraveling the complexities of their biology, making genome annotation a foundational aspect of molecular and genomic research. By precisely identifying coding regions, non-coding elements, and regulatory sequences, genome annotation transforms raw genomic data into a practical blueprint that opens up possibilities for studying living organisms at the molecular level.

Existing genome annotation methods can be broadly categorized into *ab initio* [1], homology-based [2], evidence-based (reference-based) [3], and hybrid approaches [4–6], each leveraging different strategies and tools to predict and validate genomic features. Each category has its unique advantages and challenges. *Ab initio* methods, which predict genes solely from intrinsic sequence signals without external evidence, are capable of identifying novel genes; however, they tend to produce a higher rate of false positives and inaccurate gene models—such as overpredicted exons, missed genes, and incorrect exon–intron boundaries—especially when applied to complex or poorly characterized genomes. Homology-based methods use evolutionary conservation to achieve high accuracy for well-conserved genes but often miss novel or rapidly evolving genes that lack close homologs. Evidence-based approaches utilize transcriptomic or proteomic data to support gene models, thereby enhancing annotation quality for expressed genes, yet they are inherently limited to regions where experimental data exist and are as robust as the underlying evidence. Hybrid approaches combine intrinsic signals with external evidence in an effort to balance sensitivity and specificity.

To streamline the annotation process, several genome annotation platforms and pipelines have been developed. For instance, the RefSeq eukaryotic annotation pipeline at NCBI [7] employs highly curated alignment-based evidence and gene prediction models, while another

Abbreviations: BP, base pairs; NT, nucleotide; RNA, ribonucleic acid; rRNA, ribosomal RNA; ncRNA, non-coding RNA; SIRV, Spike-In RNA Variants; ORF, open reading frame; GTF, Gene Transfer Format; BLAST, Basic Local Alignment Search Tool; STAR, Spliced Transcripts Alignment to a Reference; BUSCO, Benchmarking Universal Single-Copy Orthologs; GO, Gene Ontology; cDNA, complementary DNA; DNA, deoxyribonucleic acid; RIN, RNA Integrity Number; GGA, Galaxy Genome Annotation.

* Corresponding author.
E-mail addresses: azaremba@ibch.poznan.pl (A. Zaremba), adutta@ibch.poznan.pl (A. Dutta), asczajka@ibch.poznan.pl (A. Samelak-Czajka), paulinaj@ibch.poznan.pl (P. Jackowiak).

https://doi.org/10.1016/j.ygeno.2025.111104
Received 5 May 2025; Received in revised form 5 August 2025; Accepted 6 September 2025
Available online 8 September 2025

recently reported pipeline TAGADA [8] aims to integrate evidence from multiple data sources for eukaryotic genome annotation. Web-based platforms such as Galaxy Genome Annotation (GGA) [9] provide an accessible interface for constructing annotation workflows, and tools like Prokka [10] offer rapid prokaryotic annotation that automates the integration of gene prediction and functional annotation, generating standardized output formats compatible with GenBank. Additionally, PAGIT (Post Assembly Genome Improvement Toolkit) [11], the community-driven Nextflow-based approach nf-core GenomeAnnotator [12], and other pipelines [13–15] expand the available toolkit.

Despite these advancements, applying genome annotation methods to non-model or non-classical model organisms presents significant challenges. Due to limited reference data, tools must rely more heavily on *ab initio* methods, which—as noted above— are prone to errors, particularly for complex genomes, and often lack sufficient training data. Repetitive sequences and the presence of multiple isoforms complicate the annotation process. Genetic diversity and evolutionary distance from well-studied species can impede the identification of orthologous genes and conserved pathways, while limited experimental data often constrain the ability to validate or refine gene models. Such constraints are further compounded by annotation artifacts—including fragmented and chimeric transcripts—that can obscure true biological functions and introduce errors into downstream analyses.

These challenges are particularly relevant to research on the sexual (S2F2 clonal line) and asexual (CIW4) strains of *S. mediterranea*, a highly regenerative flatworm [16,17]. Currently, reference genome sequences and comprehensive annotations are available primarily for the sexual strain, including contig-level genome assemblies such as SmedGD_c1.3 [18,19], dd_Smes_g4 [20,21], chromosome-scale genome assemblies schMedS2 [22] and recent schMedS3 [23], as well as corresponding annotations like SmedGD_c1.3, high-confidence SMESG, schMedS1, schMedS2, schMedS3_1 and schMedS3_2. Chromosome-scale assemblies of the sexual strain indicate that sexual and asexual *S. mediterranea* share high genomic similarity—they are both diploid with four pairs of chromosomes. Although the asexual strain possesses a chromosomal translocation between chromosomes 1 and 3, these reference sequences and annotations offer a strong starting point for transcriptome analysis. Nonetheless, given potential strain-specific transcripts and structural variations, further refinement of the asexual strain's transcript set is essential [24]. Additionally, the asexual strain exhibits a simplified reproductive strategy [25–28] and it has been widely used in numerous studies, with extensive data deposited in public repositories. As of January 2, 2025, there are 122 DNA and 1694 RNA datasets on NCBI [29] for this strain, the majority of which are short-read data. In comparison, for the sexual strain, there are only 18 DNA and 87 RNA datasets. The predominance of RNA-seq data highlights the extensive transcriptomic resources available for the asexual strain, which offers a strong foundation for gene expression and functional studies and underscores the potential for leveraging RNA sequencing to enhance genome annotation efforts. However, limited detailed reference resources for asexual strain remain a significant challenge. Currently, the scientific community must rely mostly on reference transcriptomic data—such as dd_Smed_v6.pcf.contigs.fasta [21] (a transcriptome assembly generated from RNA-Seq data) and a contig-based genome assembly and annotation SmedAsxl_genome_v1.1 [19]. Beyond that, existing transcriptome assemblies are fragmented and incomplete, lacking the high-resolution insights offered by newer sequencing technologies. To bridge this gap, our study aims to implement a sophisticated bioinformatics pipeline to improve annotated genes and transcripts in asexual *S. mediterranea* using the existing chromosome-scale genome of the sexual strain (schMedS2) as a reference and in-house generated comprehensive collection of transcriptomic data for asexual worms. Our dataset includes both short-read and long-read data, providing a significant advantage in resolving complex isoforms. Addressing these challenges requires the integration of various sequencing technologies and advanced bioinformatics tools. Short-read sequencing platforms, such as Illumina, offer high accuracy and depth of coverage, a low error rate, which is essential for identifying single-nucleotide polymorphisms, small indels, and importantly, splice sites—ensuring accurate distinction of exon-intron boundaries [30]. Conversely, long-read technologies like PacBio provide valuable information on structural variants and complex transcript isoforms, enhancing the resolution of genome assemblies [31,32]. Combining these technologies benefits from their complementary strengths, enabling a more comprehensive and accurate genome annotation [4].

Here we present a universal, modular, and adaptable pipeline for genome annotation to process and interpret the vast amounts of sequencing data generated, that integrates both short- and long-read sequencing data, employs state-of-the-art bioinformatics tools for transcriptome assembly and genome characterization, and incorporates robust methods for functional and homology annotation. By doing so, we aim to advance genome annotation efforts within *S. mediterranea* and to provide a scalable framework applicable to other non-model organisms, fostering broader genomic research and discovery.

2. Methods

2.1. Planarian culturing and sample collection

Asexual *S. mediterranea* were cultured as described in [33] and fed chicken liver twice weekly, and were used either in wild-type form or after RNAi treatment, as detailed in Table S1. Animals were collected, immediately flash-frozen in liquid nitrogen to preserve RNA integrity, and stored at −80 °C until RNA isolation.

2.2. Dataset details

Full *S. mediterranea* dataset-specific details including experimental treatments, RNA extraction, quality control, are provided in Table S1 and in the associated GEO submissions. Table S1 lists transcriptomic cDNA libraries from the asexual CIW4 strain of *S. mediterranea*: 3 Illumina short-read runs (2 on NextSeq 550, 1 on NovaSeq 6000) and 3 PacBio long-read runs (1 on Sequel II, 2 on Sequel IIe). Publicly available short-read Illumina RNA-seq data for *Arabidopsis thaliana* (PRJNA477336) and *Caenorhabditis elegans* (CRA008476) were retrieved from https://www.ncbi.nlm.nih.gov/ and https://ngdc.cncb.ac.cn/gsa, respectively.

2.3. Validation with simulated data

A synthetic genome was created using a custom Python script, which constructed a single chromosome spanning 2 M base pairs (bp). This genome included 1000 genes randomly distributed along the chromosome, each comprising between 2 and 20 exons with lengths ranging from 100 to 1000 nucleotides (nt) and introns of 100 to 1000 nt. Paired-end short-read RNA-seq data were simulated using ART Illumina tool, generating 100 bp reads at 10× coverage across six samples. Additionally, long-read sequencing data were produced using PBSIM tool with 3 kb reads in whole-genome mode at 20× depth. To evaluate the pipeline's ability to handle incomplete annotations, a random subset of genes was removed from the annotation file using the custom bash script, resulting in a GTF file with missing gene information.

2.4. Reference annotation polishing

The reference genome schMedS2 [22] was utilized for this study. High-confidence annotations for this genome were obtained from two sources: SmedSxl smed_chr_ref_v1 UCLA (SIMRBase (https://simrbase.stowers.org)) and schMedS2_PlanMine_high-confidence_liftover (PlanMine [21]). Both annotations are compatible with the genome sequence; however, due to minor differences between them, the genes were first unified, and the annotation was carefully reviewed and corrected.

Through detailed analysis and comparison, 27,676 SMEST transcripts and 21,332 SMESG genes were identified as the most confident and consistent across both sources. Next, 169 transcripts with identical exon structures using the AGAT tool [34] were removed to eliminate redundancy and enhance annotation quality.

2.5. Benchmarking against BRAKER3

2.5.1. Input data

Illumina RNA-seq reads trimmed with TrimGalore were aligned to reference genome with STAR v2.7.11a [3]. For *A. thaliana* and *C. elegans*, the following reference genomes and annotations were used: GCF_000001735.4_TAIR10.1 [35,36] and PRJNA13758/ WBcel235 (WBPS19) [37,38], respectively.

2.5.2. BRAKER3 test

BRAKER3 v2.1.8 [39] was launched in "RNA-seq only" mode: braker.pl –genome genome.fa –bam star.bam –softmasking –threads 8.

2.5.3. SmedAnno test

Four runs were produced: no guide and no TransDecoder [40] (-TD) annotation added, no guide but +TD, DeepSplice [41] guidance and -TD, DeepSplice guidance and +TD. All used StringTie2 v2.1.1 and omitted long-read input.

2.5.4. Structural accuracy

Predictions were compared to references with gffcompare v0.12.6 [42]. Sensitivity, precision and F1 score were extracted at exon, intron and transcript level.

2.5.5. Completeness

BUSCO v5.7.1 [43] was run in transcriptome mode (−m tran) with the following lineages: Embryophyta_odb10 (*A. thaliana*), Nematoda_odb10 (*C. elegans*) and Metazoa_odb10 (*S. mediterranea*).

2.6. StringTie2 parameter optimization

The transcriptome assembly was evaluated using the SIRV-set 4 RNA spike-in controls to assess its accuracy, with specific StringTie2 parameters (−c, the minimum reads per bp coverage to consider for multi-exon transcript, and -f, the minimum isoform fraction) being tested (Table S2). Five combinations of StringTie2 parameters were used: default settings (−c 1; −f 0.01), a stricter isoform fraction filter alone (−f 0.02), two higher coverage thresholds (−c 2 and -c 5), and a mixed setting (−c 1.5, −f 0.02).

2.7. Assembly evaluation and comparative analysis

Completeness was assessed with BUSCO [43] (v5.7.1) with a Metazoa benchmark of 954 conserved genes, while assembly quality metrics were obtained using rnaQUAST [44] (v2.3.0) and Transrate [45] (v1.0.3). Benchmarking procedures were conducted by comparing the performance of the pipeline against existing annotations, with metrics such as the number of correctly identified transcripts, assembly scores, and completeness indicators being analyzed.

2.8. ncRNA identification

All ncRNA sequences were downloaded from RNAcentral [46] and supplemented with 457 tRNA sequences from an external publication [47]. The sequences were linearized and deduplicated using seqkit rmdup. The initial dataset contained 27,390,440 sequences, which was reduced to 27,317,511 after deduplication. A BLASTn [2] search was conducted to identify homologous sequences within the *S. mediterranea* nuclear and mitochondrial genomes. The following parameters were used: -query: 27,317,511 ncRNA sequences; −db: *S. mediterranea* genome (nuclear + mtDNA), −evalue: 1e-20, −word_size: 15, −qcov_hsp_perc: 90, −dust: no. The output was then processed to extract the original sequence length of hits for further filtering. A custom Python script (Table S3) was used to classify ncRNA types based on query ID patterns. Hits were retained if alignment coverage met predefined thresholds (95 % for tRNA, 90 % for miRNA, 70 % for lncRNA, 90 % for rRNA, 80 % for other ncRNA). Then another custom R script (Table S3) was used for converting to gtf format, refining and merging overlapping regions to ensure only non-redundant ncRNA annotations were retained.

3. Analyses

3.1. Selection and validation of transcriptome assembly approach

One of the key steps in genome annotation is the assembly of the transcriptome. In our efforts to enhance genome characterization for the asexual strain of *S. mediterranea*, we employed StringTie2 [4], a widely recognized tool for transcriptome assembly. StringTie2 is versatile, supporting both short-read and long-read sequencing datasets, which makes it well-suited for comprehensive transcriptome analysis. However, StringTie2 has multiple versions, each offering distinct features and performance enhancements. Therefore, before integrating StringTie2 into our pipeline, we thoroughly optimized its parameters to ensure the highest accuracy and reliability in transcript predictions. To this end, we employed two complementary strategies.

In the first strategy, we generated an *in-silico* dataset to simulate genomic sequences with corresponding full and incomplete annotations and short-read RNA-seq data (see Methods section). This synthetic data allowed us to create scenarios with known simple transcript structures, enabling a comprehensive evaluation of StringTie2 performance. Reads were aligned to the *in-silico* reference genome using STAR. We then compared two commonly available versions of StringTie2, 2.1.1 and 2.2.1, which allowed us to assess version-specific performance improvements. We evaluated the results of the assembly step by analyzing transcript identification rates for both StringTie2 versions at base, exon, locus and transcript levels using gffcompare. Our results indicated that StringTie2 version 2.1.1 provided a higher identification rate of short-read data in reference-based mode than its descendant (Fig. 1A), while there were no differences in *de novo* mode. Although the StringTie2 documentation does not explicitly state that version 2.1.1 is more sensitive for short-read data, the combination of release note details and our experimental observation suggested that less stringent filtering of version 2.1.1 results in a higher count of assembled isoforms. Meanwhile, version 2.2.1 has been optimized for hybrid-read assembly (with the –mix option) and incorporates more conservative measures to reduce false positives—which can lower the total transcript count when only short-read data are used. This trade-off between sensitivity and precision is common in transcript assemblers. When maximizing the number of detected isoforms from short reads is a priority, 2.1.1 may be advantageous, whereas 2.2.1's improvements (including its ability to process mixed-read data) make it preferable for datasets that include long reads. It should be noted that these version choices are based on their prevalence at the time of our study, and although our benchmarking indicates reliable performance, further testing in diverse datasets may be beneficial.

In the second strategy, we used Spike-In RNA Variants (SIRVs), a commercially available collection of RNA spike-in controls, which were added to the RNA isolated from planarians (see Table S1) prior to the preparation of sequencing libraries. These SIRVs included three distinct modules — ERCC Mix 1 (a set of 92 synthetic RNA molecules ranging between 250 and 2000 nt in length to resemble natural eukaryotic transcripts), Isoform Mix E0 (referred here as multi-exon SIRVs), and long SIRVs (referred here as mono-exon SIRVs) — each designed to mimic specific aspects of transcriptome complexity, including transcript abundance, isoform diversity, and transcript length. RNA-seq reads were aligned to the reference SIRVs sequences using STAR for short reads and

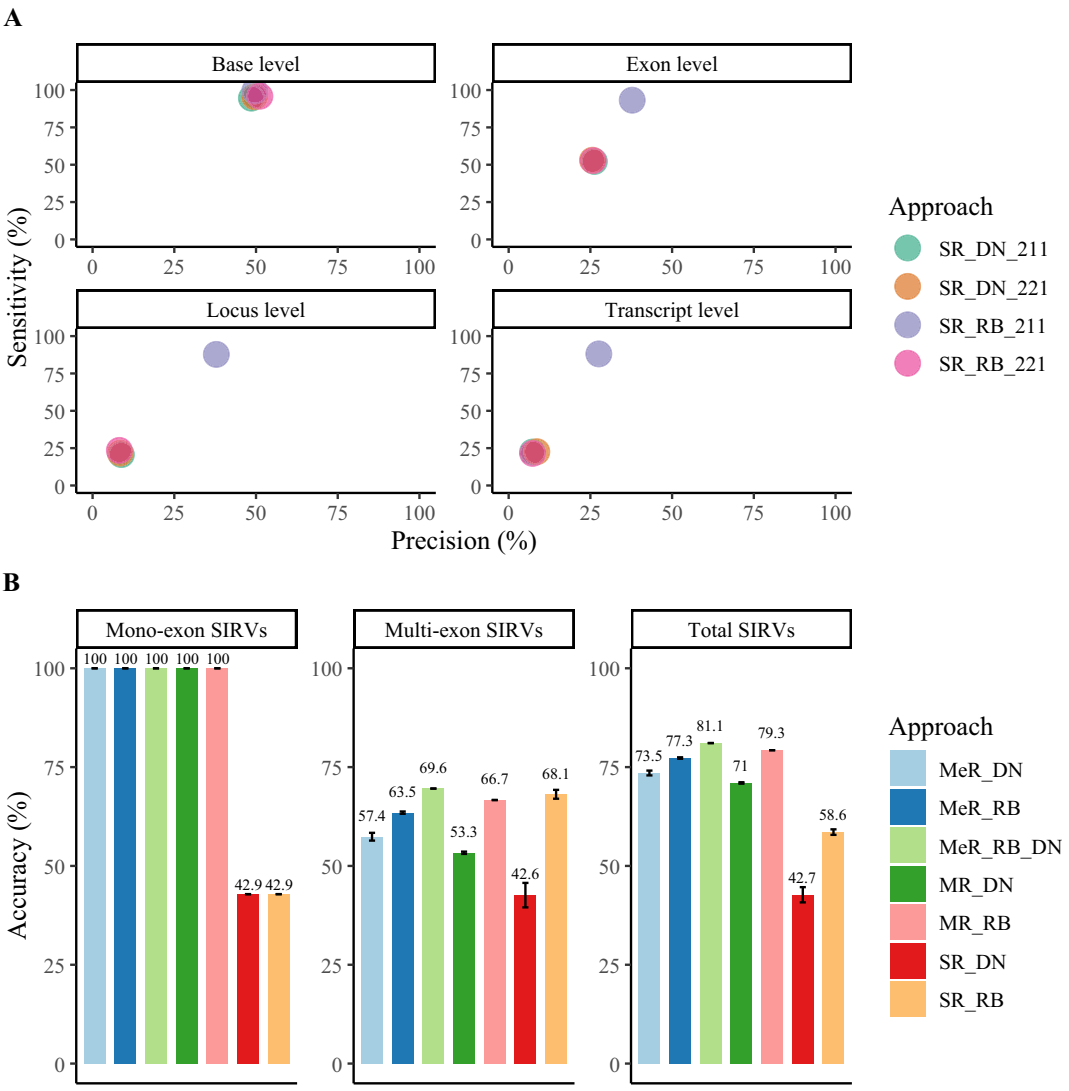

Fig. 1. Identification rates and accuracy metrics for *in-silico* and SIRVs datasets.
(A) - Transcript identification rates using *in-silico* data. Circles depict the four approaches tested for short read (SR) data, using *de novo* (DN) or reference-based (RB) assembly and StringTie2 version 2.1.1 (211) or version 2.2.1 (221). (B) - Accuracy metrics of transcript isoform identification using SIRVs on SR, mixed reads (MR) and merged reads (MeR) approaches. Whiskers on the bar plots represent the variability in accuracy across five different StringTie2 parameter configurations tested.

Minimap2 for long reads. We then employed StringTie2 using the "insufficient annotation" as the reference provided by the manufacturer of SIRVs (SIRV_I), where some SIRVs that are actually present in the mixes are not annotated, and compared the prediction to the correct annotation of all SIRVs (SIRV_C). Such a comparative analysis demonstrated a proper reconstruction of missing spike-in transcripts. Then, focusing on the mixture of synthetic RNA isoforms Iso Mix E0 category, we systematically assessed the performances of seven assembly approaches for transcript variant detection, implementing different combinations of read types—short and mixed—and assembly modes—reference-based and *de novo*. These approaches included short reads in reference-based mode (SR_RB), short reads in *de novo* mode (SR_DN), mixed reads in reference-based mode (MR_RB), mixed reads in *de novo* mode (MR_DN), merged reads in *de novo* mode (MeR_DN), merged reads in reference-based mode (MeR_RB), and merged reads in both modes (MeR_RB_DN). Mixed reads refer to the combined use of short and long reads within a single assembly process, whereas merged reads indicate separate assemblies of short reads and mixed reads, followed by merging the resulting assemblies. For each approach, StringTie2 performance (Table S2) was quantified in terms of accuracy metrics, reflecting the ability to correctly identify transcript isoforms (Fig. 1B). SR_DN approach achieved the lowest accuracy, ranging from 35.1 % to 45.1 % with its maximum at -c 1.5 and -f 0.02. Surprisingly, generally, the highest accuracy rate of 81.1 % was obtained when predictions from both short and mixed reads in both modes were merged (MeR_RB_DN). This outcome underscores the synergistic benefits of integrating multiple read types and assembly strategies.

Consequently, our pipeline employs a dual-version strategy—by default, StringTie2 version 2.1.1 is used for assembling short-read data and version 2.2.1 for mixed-read data, ensuring optimal performance across a range of sequencing data types. Users should note that while these versions provided robust performance in our tests, alternative versions may be considered depending on specific experimental contexts and can be adjusted *via* command-line. We set -c 1.5 and -f 0.02 based on pilot experiments with SIRVs, which balanced sensitivity and precision in isoform detection (see Table S2). After completing individual assemblies, we incorporated the MeR_RB_DN approach to merge reference-based and *de novo* transcripts, resulting in a comprehensive and unified transcriptome annotation that captures a diverse array of transcript isoforms.

3.2. Identification of artifactual transcripts

In our tests with *in-silico* data, we identified several types of artifactual transcripts, including those with unusually long exons exceeding 100,000 nt, reversed duplicates (transcripts overlapping the same genomic regions as the reference ones but on the opposite strand), as well as fragmented and chimeric RNA. Fragmented transcripts are individual RNAs that have been split into multiple predicted gene models due to incomplete transcript reconstruction or assembly errors. Chimeric transcripts, on the other hand, erroneously combine sequences from multiple distinct genes. In eukaryotic genomes, exons generally range in size from approximately 50 to 5000 nt, with the majority falling between 100 and 1000 nt [48]. Instances of exons exceeding 10,000 nt are exceptionally rare and are usually associated with specific organismal features, such as immune-related genes or tandem duplications. Even in these rare cases, such lengths remain an order of magnitude below 100,000 nt [49]. While highly effective in many contexts, StringTie2 may incorrectly model transcripts, especially in regions with overlapping genes or incomplete stop codons. Such inaccuracies can extend exon boundaries far beyond biologically realistic lengths [50].

To address the identified challenges and eliminate artifacts, we added a custom function to the pipeline that ensured that transcripts with exons exceeding 10,000 nt and/or transcripts that surpassed 100,000 nt of the genomic region were removed, thus refining the annotation and preventing the propagation of erroneous data. Additionally, we included a function to identify and filter out reversed duplicates without inadvertently removing legitimate transcripts. We also implemented criteria to flag potential fragmented and chimeric RNA based on exon length, genomic proximity, and functional annotations. In our pipeline, indicators for fragmentation include multiple predicted genes in close genomic proximity (within 1000 nt) on the same strand and adjacent or overlapping genomic coordinates. The exon length and genomic proximity thresholds are based on general eukaryotic genome characteristics [49]. These parameters should be adjusted as needed based on the specific organism or dataset characteristics. Transcripts and genes are flagged as chimeric if they meet at least one of the following criteria: (i) they display structural inconsistencies, such as widely spaced or non-contiguous exons; (ii) they span multiple distinct genomic regions; or, in the case of genes, (iii) they possess multiple, unrelated functional annotations. As part of our automated pipeline, we only added information indicating whether a gene or transcript is possibly fragmented or chimeric. The decision to remove corresponding genes or transcripts should be made by the user based on specialized evaluation. Incorporating functional annotations into the pipeline as a part of manual curation can significantly improve the accuracy of detecting artifacts.

3.3. Pipeline overview and workflow

Having resolved the challenges outlined above, we developed a versatile genome annotation pipeline and demonstrated its performance by applying it to the asexual strain of *S. mediterranea.* This pipeline integrates multiple sequencing technologies and bioinformatics tools to achieve comprehensive and accurate gene and transcript annotations across different organisms. The pipeline is divided into two main parts, each structured into distinct sequential steps.

3.3.1. Part I: pre-processing and assembly (SmedAnno)

The first part of the pipeline is container-orchestrated, fully automated, and encompasses pre-processing, alignment, assembly, and functional annotation. The input data for the pipeline can be either short-read or mixed, the latter including both short and long reads. While even a single sample can be processed when only short reads are available, incorporating multiple samples is recommended to achieve more reliable and comprehensive results. Mixed-read datasets require at least one sample with short reads and one with long reads. Additionally, the automated pipeline is optimized for paired-end reads in short-read mode. Mapping and transcript assembly are performed independently for each sample. The workflow begins by pre-processing the RNA-seq data, performing adapter and quality trimming with TrimGalore and Filtlong for short- and long reads, respectively, followed by optional length and quality filtering. Next, the pipeline focuses in particular on the removal of reads corresponding to ribosomal RNA (rRNA), based on their alignment to user-provided rRNA reference sequences (Fig. 2). This step was found to be very important in the case of organisms with unusually high A/T content, such as *S. mediterranea*. Library preparation methods typically involve rRNA depletion or polyA enrichment to facilitate the study of protein-coding gene expression. However, commercially available rRNA depletion kits are mostly optimized for model species, and polyA enrichment is not effective in organisms with a high A/T content. Consequently, reads corresponding to rRNA can be abundant in sequencing data for non-model organisms. In scenarios where rRNA reference sequences are unavailable (not provided by the user), the pipeline bypasses the rRNA removal step, continuing with the whole RNA-seq dataset. The pipeline then aligns the processed RNA-seq reads to the reference genome using STAR for short reads and Minimap2 for long reads. After completing this step, splice junctions can be optionally identified using a deep-learning-based predictor DeepSplice [41]. It uses convolutional neural network architectures (OpenSpliceAI-like) and supports CUDA acceleration for efficient computation. DeepSplice infers splicing donor and acceptor motifs directly from the reference genome, based on an integrated model. Currently the user can select a model, according to phylogenetic proximity to the target organism, from five available ones (human, mouse, zebrafish, honeybee, and thalecress). Next, the transcript assembly phase uses StringTie2 – by default, version 2.1.1 is applied to short-read data, while version 2.2.1 is used for mixed-read data (which integrates both short and long reads). Users can specify alternative versions if desired. After assembly, the pipeline merges transcripts across all samples using StringTie2 to create unified sets of transcripts. This merging process is performed separately for reference-based (if reference annotation was provided) and *de novo* assemblies before combining them into the final complete annotation. This dual approach by performing both reference-based and *de novo* annotation, provides comprehensive assembly, thus capturing a wide range of transcript isoforms. DeepSplice evidence is then merged with StringTie2 predictions. Subsequently, transcripts with excessively long exons or that cover unusually large genomic regions are filtered out. Next, the pipeline uses gffcompare to compare the predicted annotations against the reference annotations (if provided), assessing the completeness and accuracy of the assembled transcripts. This comparison generates a report that includes data summaries and accuracy estimations. Additionally, it enriches the predicted annotations with detailed information about their overlaps with known reference transcripts and genes. However, the generated annotations may contain issues that can compromise the quality and usability of the annotation files. To address these issues, the pipeline employs AGAT (Another GTF/GFF Tool), which systematically identifies and corrects common syntax errors, ensuring that essential attributes such as gene_id and transcript_id are present for each feature, preventing incomplete annotations, correcting improper use of delimiters in attribute fields (such as replacing spaces with a semicolon), detecting and removing duplicate entries, and ensuring correct key-value pair syntax. This correction phase is crucial for maintaining high-quality genome annotation. The obtained set of transcripts is then subjected to functional annotation in a process that involves the sequential application of several tools, and which includes homology analysis and the identification of protein domains. To this end, first, the open reading frames (ORFs) are predicted using TransDecoder. Following that, BLASTp and BLASTx are used for homology search through the SwissProt database. Finally, protein domains are identified using hmmscan through the PFAM database and InterProScan, which aggregates multiple domain resources (SMART, Panther, TIGRFAM, *etc.*) and enriches predictions with Gene Ontology

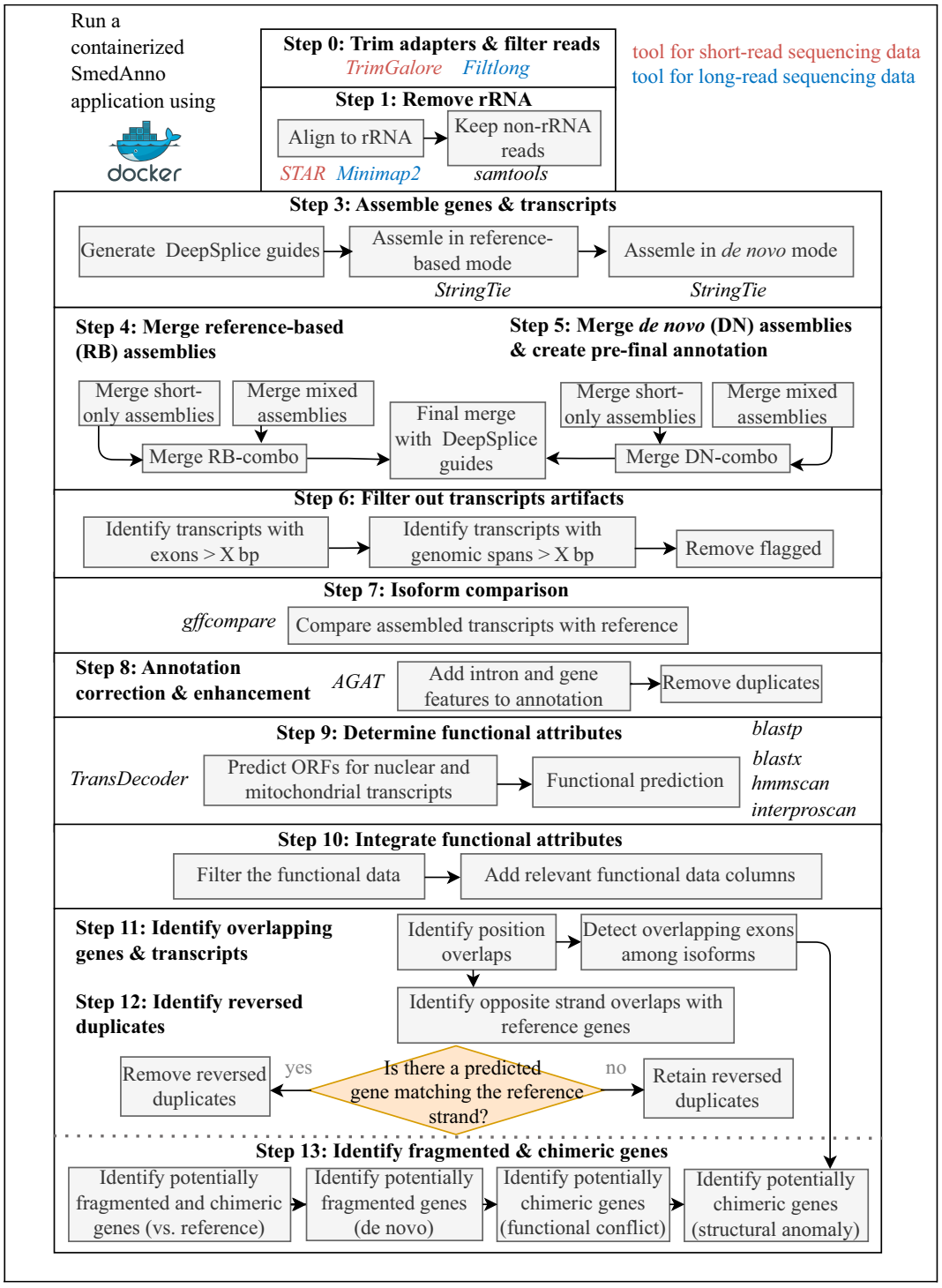

Fig. 2. Workflow diagram for RNA-seq pre-processing, transcript assembly, and functional annotation.

(GO). This multi-faceted approach filters out low-confidence annotations, retaining only the transcripts supported by multiple lines of evidence. We used the SwissProt, PFAM and InterProScan databases as a part of the automated pipeline specifically for their low to medium memory requirements. However, we recommend to use additional resources, for example eggNOG, NCBI, UniRef, KEGG, RefSeq, COG, Reactome, and BioCyc for conducting in-depth homology analyses, as these databases provide complementing information, evolutionary relationships, and diverse protein characteristics. Next, the functional annotations are integrated into the GTF file. This integration applies default thresholds: for PFAM predictions, a bit score greater than 20, a conditional e-value below 1e-5, and coverage exceeding 50 %; for SwissProt predictions, a BLAST e-value threshold of 1e-5 and an identity threshold of 25 %; InterPro hits are retained if they meet the e-value threshold of <1e-5 in at least one of the member databases. These criteria ensure that low-confidence functional annotations are discarded. However, all filtering thresholds can be configured by the user. Then, the pipeline adds new columns to the GTF annotation file, including has_ORF (TRUE/FALSE), best_SwissProt_blastp_hit, best_SwissProt_blastx_hit, best_Pfam_hit, and best_InterPro_hit domain information. The next step is the identification of overlapping genes and transcripts, which detects genes that share genomic regions and checks if

each transcript in a gene has at least one overlapping exonic region with any other transcript in the same gene. Lastly, the pipeline identifies the aforementioned reversed duplicates and addresses fragmented and chimeric genes and transcripts. At the completion of functional annotation, the pipeline emits an interactive summary-and-advice block. Six boolean quality control flags (*overlapped_predicted_gene*, *is_potentially_fragmented_reference*, *is_potentially_chimeric_reference*, *is_potentially_fragmented_de_novo*, *is_potentially_chimeric_structure_de_novo*, and *is_potentially_chimeric_function_de_novo*) are printed together with plain-language recommendations (*e.g.* "merge fragmented loci", "split structural chimeras"). Users are expected to review each flagged locus individually in IGV, confirm read support, and decide case-by-case whether to merge, split, or retain the model unchanged. All these steps (Fig. 2) are implemented in the SmedAnno tool, forming Part I of the entire pipeline. The source code can be accessed on GitHub (https://github.com/Norreanea/SmedAnno).

3.3.2. Part II: manual curation and refinement

The second part of the pipeline involves user-driven inspection and refinement of the annotated data. Despite the implementation of several automated steps aimed at optimizing transcript assembly and obtaining biologically relevant transcripts that encode functional protein domains, resolving overlapping or fragmented transcripts remains a significant challenge that requires specialized approaches. While the primary focus is on transcript assembly, gene features are subsequently derived from the assembled transcripts. Fragmented and chimeric transcripts, which primarily affect RNA-level annotations, are addressed by prioritizing the longest isoform to define each gene, thereby reducing redundancy and improving annotation accuracy. To accurately determine whether genes should be merged after being marked as potentially fragmented and/or overlapped, promoter and terminator sequences should be analyzed using Promoter 2.0 and ARNold tools (Fig. 3). This analysis verifies the coherence of regulatory regions, ensuring that merged gene models maintain consistent promoter and terminator elements. These tools are accessible *via* their respective web pages and are not included in the automated pipeline. Users must extract sequences between inspected genes to check for the presence of promoters and terminators. After potential merges, manual adjustments are advised to finalize gene models, ensuring accurate gene length prediction and overall annotation reliability.

While genome annotation typically focuses on protein-coding genes, our pipeline also includes a step for annotating non-coding RNAs (ncRNAs), which have significant functional relevance. This step is optional, as it requires downloading the extensive (>20 GB) RNAcentral database. The ncRNA sequences are first linearized and deduplicated using seqkit rmdup to remove redundant entries. Homologous regions are then identified in the reference genome using BLASTn. After the identification step, ncRNA types are classified based on the sequence identifier. The matching results are then converted to GTF format. A multi-step filtering and refinement process is used to improve the accuracy of the annotations. This process addresses several common issues, including multiple alignments at identical genomic positions, inconsistent naming conventions, and clustering.

The final step integrates all gathered information into a unified genome annotation file, ready for downstream applications, including both coding and non-coding elements with comprehensive functional insights.

The bioinformatics tools used in each pipeline step, along with their respective versions are summarized in Table 1.

3.4. SmedAnno benchmarking

To benchmark the automated Part I of our pipeline (SmedAnno), we compared its performance against BRAKER3, a widely used genome annotation tool that integrates GeneMark-ET and AUGUSTUS gene prediction, guided by short-read RNA-seq data [39]. The comparison was carried out using three genomes: two classic models – *A. thaliana* (AT) and *C. elegans* (CE) – and *S. mediterranea* (SM). Since BRAKER3 utilizes *ab initio* approach and short-read data only, we ran a scaled-down version of SmedAnno without reference-based guidance, exclusively on Illumina reads, to ensure a valid comparison. For *A. thaliana* and *C. elegans*, publicly available data were used, while for *S. mediterranea* we employed in-house generated datasets (Table S1). For DeepSplice, which is integrated into our pipeline, we employed thalecress model for *A. thaliana*, and honebee model for *C. elegans* and *S. mediterranea*. SmedAnno was run in four parameter configurations (with/without DeepSplice and TransDecoder). After assembling transcripts for each genome, we evaluated the resulting transcriptomes using BUSCO, which assesses the presence of conserved orthologs, by comparing a transcriptome of interest to those within the Metazoa lineage, and gffcompare.

BRAKER3 reached slightly higher BUSCO completeness in AT (98.9

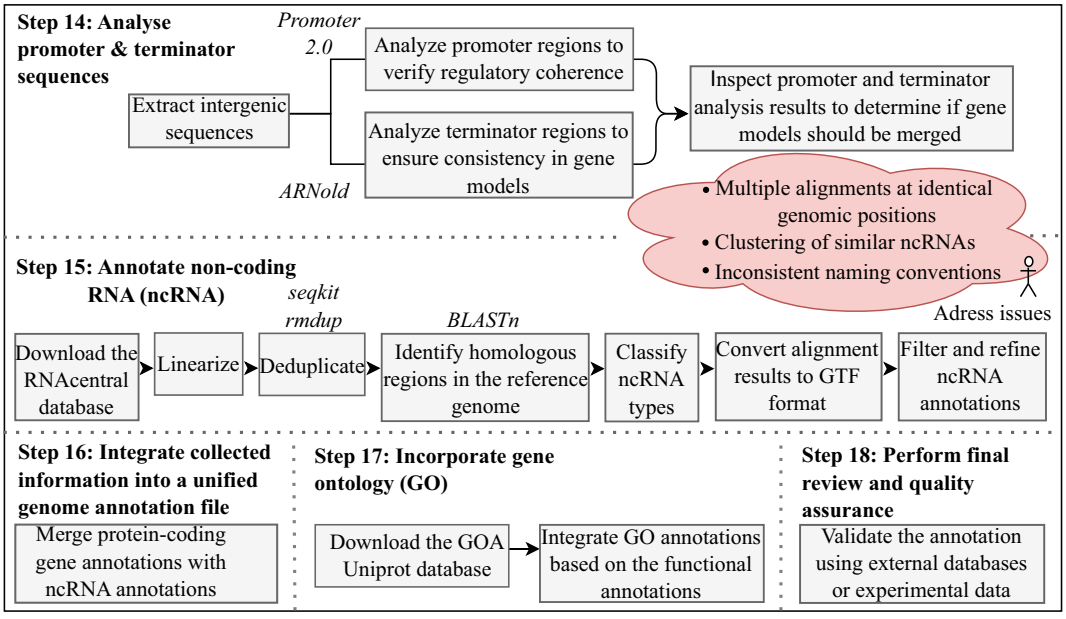

Fig. 3. Workflow diagram for user-driven inspection and refinement.

Table 1
Bioinformatics tools and pipeline steps.

Step	Tool	Version
Step 0: Trim adapters & filter reads	TrimGalore	0.7.1
	Filtlong	0.2.1
Step 1: Remove rRNA	STAR	2.7.9a
	Minimap2	
	samtools	
Step 2: Align reads to the reference genome	STAR	2.7.9a
	Minimap2	
Step 3: Assemble genes & transcripts	StringTie2	2.1.1 (short reads)
	DeepSplice	2.2.1 (mixed reads)
		0.10
Step 4: Merge reference-based assemblies	StringTie2	2.2.1
Step 5: Merge *de novo* assemblies & create pre-final annotation	StringTie2	2.2.1
Step 6: Filter out transcripts artifacts	bash	
Step 7: Compare isoforms	gffcompare	0.12.6
Step 8: Correct & enhance annotation	AGAT	1.4.1
Step 9: Determine functional attributes	TransDecoder	5.7.1
	blastp	2.15.0+
	blastx	2.15.0+
	hmmscan	3.3.2
	InterProScan	5.67
Step 10: Integrate functional attributes	R	4.4.1
Step 11: Identify overlapping genes & transcripts	R	4.4.1
Step 12: Identify reversed duplicates	R	4.4.1
Step 13: Identify fragmented & chimeric genes	R	4.4.1
Step 14: Analyse promoter & terminator sequences	Promoter 2.0	No official version number
	ARNold	No official version number
Step 15: Annotate non-coding RNA (ncRNA)	blastn	2.15.1+
	seqkit	2.8.2
Step 16: Integrate collected information into a unified genome annotation file	R	4.4.1
Step 17: Incorporate gene ontology (GO)	R	4.4.1
Step 18: Perform final review and quality assurance	R	4.4.1

%) and CE (98.6 %) than the most comparable SmedAnno runs (~95 % and 88.8 %, respectively), but fell behind in SM (76.9 % *vs* 78.6 % for both DeepSplice –TD and noGuide –TD) (Fig. 4A). Notably, SmedAnno yielded increased fraction of duplicated BUSCOs for all tested genomes (Fig. 4A). We found that SmedAnno configurations without TransDecoder generated more transcripts being identified as single-copy BUSCOs. Next, the annotations obtained with BRAKER3 and SmedAnno were compared with corresponding reference genome annotations using gffcompare. The proportion of true positive predictions was assessed at transcript, exon and intron levels, and depicted as F1 scores, where the higher value better reflects the reference. As expected, BRAKER3 performed better in CE. BRAKER3's advantage in *C. elegans* is consistent with its AUGUSTUS species-specific parameter set, which was trained on that very genome. SmedAnno+DeepSplice (-TD) outperformed BRAKER3 in tests on AT and SM, displaying higher F1 score at the exon and transcript levels in AT, and at all levels in SM (Fig. 4B).

3.5. Genome annotation of the asexual strain of S. mediterranea

After demonstrating the efficiency of SmedAnno, we applied the whole pipeline to generate a genome annotation of the asexual strain of *S. mediterranea*. To this end, we used the available reference genome annotation and both Illumina and PacBio sequencing datasets (Table S1). The SmedAnno output was subjected to manual curation and refinement, which included identification of artifactual transcripts and detection of missing genes. The annotation, schMed_Pn, is available in GEO repository under accession GSE294569.

3.6. Identifying fragmented and chimeric artifacts

In the context of genomic annotation, artifacts are erroneous gene or transcript models that arise from limitations in the annotation process rather than representing true biological entities. Specifically, we refer here to fragmented artifacts as those that occur when a single gene is incorrectly split into multiple separate gene models. This fragmentation can result from incomplete transcript assembly or sequencing errors, leading to discontinuous or truncated representations of what is biologically a single gene. On the other hand, chimerism typically arises from misassembly during transcriptome reconstruction, especially in regions with high gene density or overlapping gene structures.

SmedAnno employs flags such as: *is_potentially_fragmented_de_novo*, *is_potentially_chimeric_structure_de_novo*, and *is_potentially_chimeric_function_de_novo* to identify these artifacts based on functional and structural criteria. Below, we discuss three illustrative examples from schMed_Pn (depicted in Fig. 5) that highlight the complexities and implications of these annotation challenges. SMESGs here are reference genes and MSTRGs are putative genes.

3.6.1. Prediction of gene fragments

In one genomic region, three genes—SMESG000004200.1, SMESG000004201.1, and MSTRG.967—were located in close proximity on chromosome 1 (smed_chr1) (Fig. 5A). Specifically, gene SMESG000004200.1 spanned position 14,973,263 to 15,003,452 on the positive strand, while SMESG000004201.1 extended from 15,003,453 to 15,011,292 on the same strand. StringTie2, predicted an additional transcript, MSTRG.967, which encompassed the regions of both genes (14,973,256 to 15,011,292). Both SMESG000004200.1 and SMESG000004201.1 were independently annotated as "Epidermal growth factor receptors." However, the prediction by StringTie2 that merged these two genes into a single transcript was more biologically plausible. This is because the close genomic proximity and shared functional annotations suggested a possible complex gene structure that is better represented by a single, contiguous transcript. Flagging these genes as potentially fragmented (*is_potentially_fragmented_de_novo*) highlighted the necessity to consider merged annotations in densely packed genomic regions to avoid misinterpretation of gene boundaries.

3.6.2. Chimeric transcript prediction due to excessive intron length

The gene SMESG000065897.1 initially comprised a single transcript, SMEST065897001.1, spanning positions 325,241,809 to 325,242,781 on chromosome 1 (Fig. 5B). However, StringTie2 predicted an additional transcript, MSTRG.16892.1, which extended dramatically from 325,241,809 to 325,347,749, introducing an intron exceeding 100,000 nt. This unusually large intron length is atypical and suggested a misassembly, classifying MSTRG.16892.1 as potentially chimeric (*is_potentially_chimeric_structure_de_novo*).

The presence of such a large intron likely indicated that MSTRG.16892.1 erroneously combined sequences from multiple genomic loci or failed to accurately represent the true transcript structure of SMESG000065897.1. Consequently, the chimeric transcript should be disregarded, and the gene localization should remain unchanged, basing solely on the original transcript SMEST065897001.1.

3.7. Functionally chimeric gene due to divergent transcript annotations

The gene SMESG000077853.1 exhibited a complex transcriptional landscape (Fig. 5C). While most transcripts, such as MSTRG.31576.1, MSTRG.31576.5, MSTRG.31576.6, and all reference SMEST transcripts were annotated as "ankyrin repeat proteins," one transcript, MSTRG.31576.11, was annotated as "Squamous cell carcinoma antigen recognized by T-cells 3 (SART-3)". This divergence in functional annotation across transcripts resulted in the gene being flagged as chimeric (*is_potentially_chimeric_function_de_novo*). These inconsistent functional annotations suggested that MSTRG.31576.11 may have been incorrectly

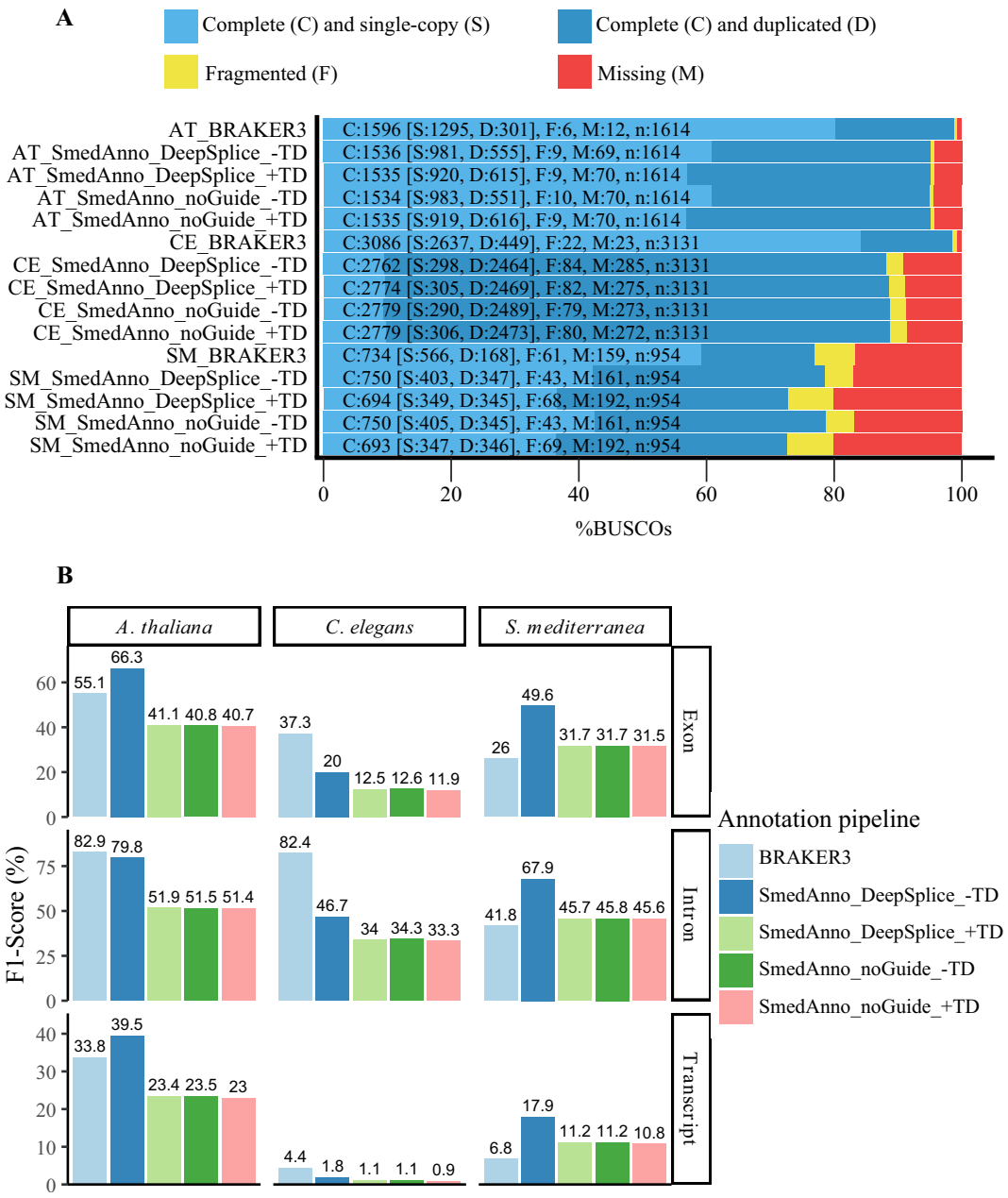

Fig. 4. Benchmark of SmedAnno *versus* BRAKER3 on short-read data.
(A) - BUSCO completeness for annotations of *A. thaliana* (AT), *C. elegans* (CE) and *S. mediterranea* (SM) transcriptomes. Stacked bars depict the proportion of: complete single-copy, complete duplicated, fragmented and missing BUSCOs for each pipeline: BRAKER3; SmedAnno with DeepSplice junction hints merged but without TransDecoder (SmedAnno_DeepSplice_-TD); the same plus TransDecoder (+TD); and SmedAnno runs without DeepSplice guidance (noGuide_ ± TD). Numbers inside the bars correspond to raw BUSCO counts. (B) - Structural accuracy. Bars show F1 scores (%, harmonic mean of sensitivity and precision) at exon, intron and transcript level for the same pipelines. Colors denote pipelines, panels group the three species.

assembled, possibly incorporating sequences from a different gene or representing a mispredicted splice variant. Given that "SART-3" functions are unrelated to "ankyrin repeat proteins," it is evident that MSTRG.31576.11 does not belong to SMESG000077853.1. Consequently, MSTRG.31576.11 was filtered out.

In total, the automated run flagged 5.1 % of all gene models, of which 1.2 % were confirmed as artifacts upon manual inspection.

3.8. Detecting missing genes through functional annotation and homology assignments

Functional annotations and homology assignments were carried out to explain the biological function of the assembled transcripts and identify orthologous genes by using established databases and computational tools. This approach facilitated the detection of missing genes, thus highlighting potential gaps and guiding further refinement of the transcriptome assembly.

To incorporate functional data into the newly developed *S. mediterranea* genome annotation (schMed_Pn), we used the following databases: Pfam, SwissProt, and non-redundant (nr) protein sequences from NCBI. Our annotation comprised 25,265 genes, of which 4011 were novel and 21,254 were identical to the reference genes. Each gene possessed at least one function-related entry, either through a database hit or by being classified as having an ORF. For 18,552 genes, we successfully identified Gene Ontology (GO) terms using homology-based strategies, which included the pfam2go package, UniProt GOA

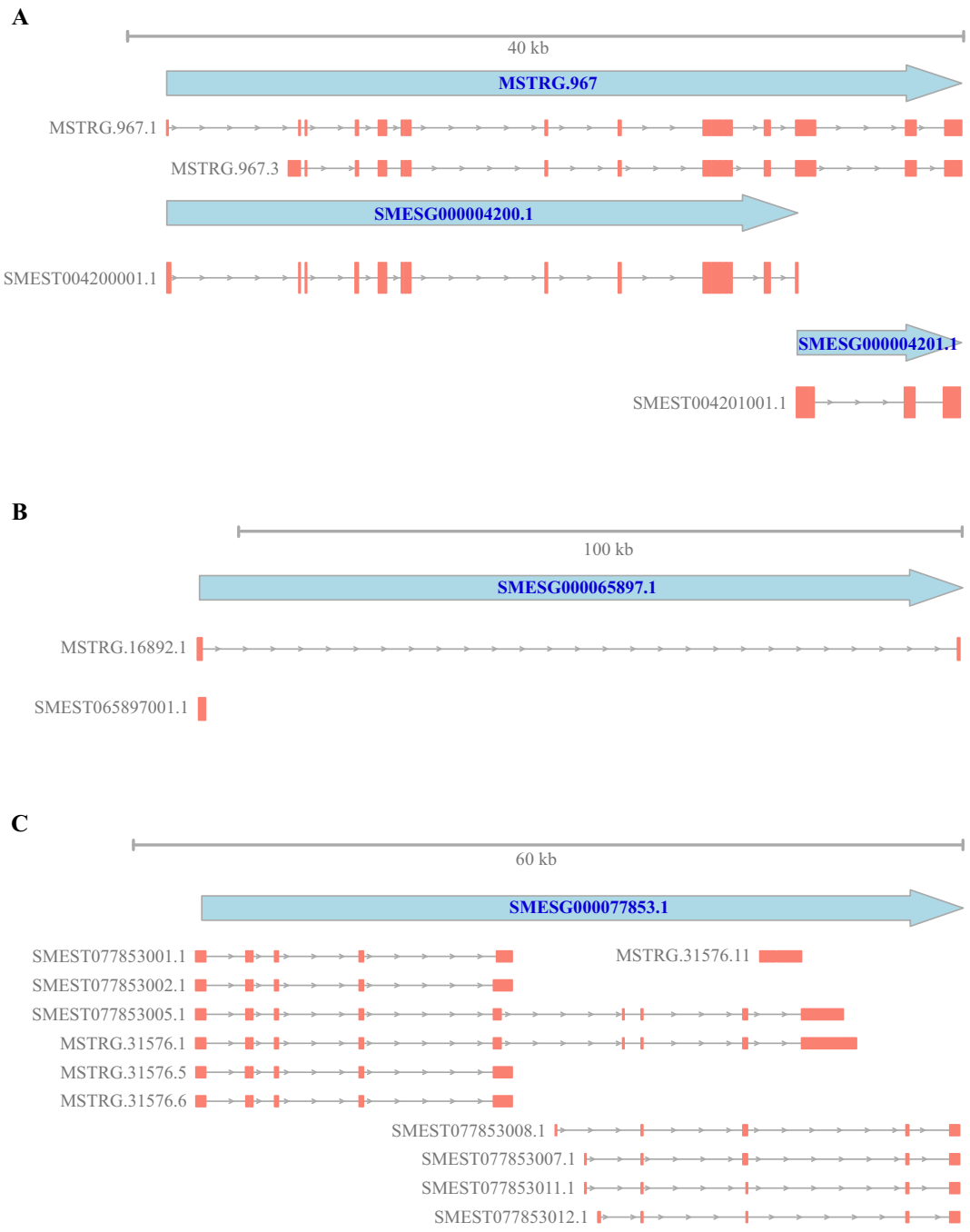

Fig. 5. Representative annotation artifacts generated by SmedAnno pipeline.
(A) - Fragmented gene. (B) Chimeric transcript. (C) - Functionally chimeric gene with divergent transcript annotations. Genes are depicted with cyan arrows, while transcripts are shown as combinations of gray lines and orange rectangles, denoting introns and exons, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

conversion, and a custom R script (Table S3) to retrieve GO annotations based on NCBI IDs. Notably, 2427 of these GO-annotated genes were putative. To evaluate the similarity of GO terms between novel and reference genes, Venn diagrams were constructed, highlighting shared and unique functional categories across the gene sets. We assigned a total of 23,909 GO terms to genes within our annotation. Specifically, 18,586 GO terms were assigned to SMESG (reference) genes, and 5323 GO terms were assigned to MSTRG (putative) genes. Of these, 5180 GO terms were shared between both gene categories, while 143 GO terms were uniquely assigned to MSTRG genes (Fig. 6A). To visualize the functional distribution of the gene sets, we generated word clouds representing the most frequent unique GO terms for each group: MSTRG (Fig. 6D) and SMESG (Fig. 6B), as well as those that were common for two groups (Fig. 6C).

3.9. Insights into planarian ncRNA diversity

Beyond refining the annotations of protein-coding genes, we also performed a preliminary assessment of non-coding RNAs (ncRNAs) in *S. mediterranea*. Although not an exhaustive characterization, our analysis revealed a diverse repertoire of regulatory RNAs, highlighting significant differences between publicly available annotations and newly generated data. Publicly available datasets from RNAcentral assigned a total of 154 rRNAs, 578 miRNAs (including 269 precursors), 61 tRNAs, 2

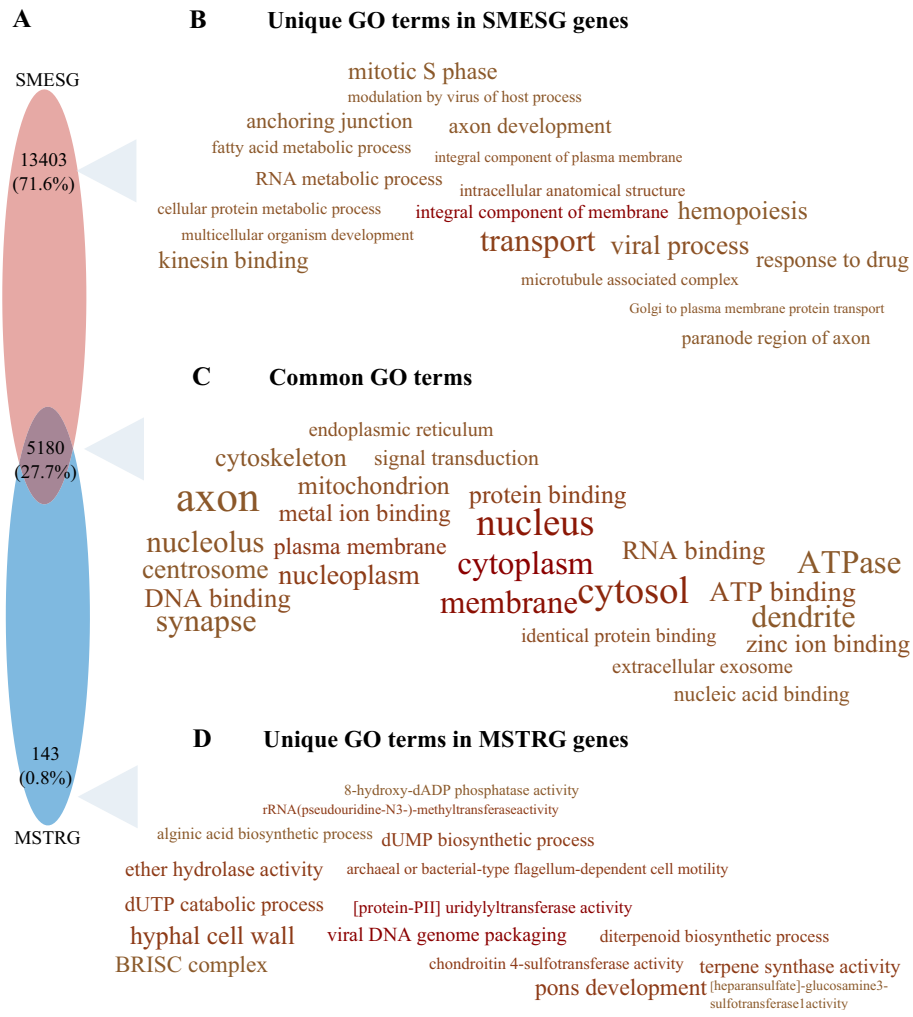

Fig. 6. Venn diagram and a word cloud of GO terms assigned for novel and reference genes for the *S. mediterranea* asexual strain.
(A) - Venn diagram of GO term overlaps between MSTRG (novel genes) and SMESG (reference genes). (B) - Word cloud of unique GO terms assigned to SMESG (reference) genes. (C) - Word cloud of common GO terms assigned to both SMESG and MSTRG genes. (D) - Word cloud of unique GO terms assigned to MSTRG (novel) genes.

snoRNAs, 12 snRNAs, and 2 SRP RNA hits to *S. mediterranea*. After deduplication, these numbers were refined to 129 rRNA, 571 miRNAs (with 156 precursors), 60 tRNAs, 2 snoRNAs, 12 snRNAs, and 2 SRP RNA hits. Additionally, according to recent literature [47], *Schmidtea* genome contains 457 tRNA hits, which were reduced to 434 after deduplication. In contrast, using our annotation pipeline, we identified 154 rRNA hits across 2287 genomic positions, 179 miRNA hits across 1042 genomic positions, 488 tRNA hits distributed over 4298 genomic positions, 2 snoRNA hits across 69 genomic positions, 18 snRNA hits across 141 genomic positions, 6 miscellaneous RNAs spread across 4752 genomic positions, 18 RNase P ncRNA components across 408 genomic positions, 21 lncRNA hits across 185 genomic positions and 2 sRNA hits across 187 genomic positions. The discrepancy between published records assigned to *S. mediterranea* and our annotated RNAs highlights the ongoing need for careful re-annotation and experimental validation in underexplored organisms like planarians. In addition, we uncovered 40 copies of regions encoding signal recognition particle RNAs (SRP_RNAs), localized on chromosome 3 and flanked by the genes SMESG000020110.1 (BLAST-predicted as a lysine-specific demethylase) and MSTRG.33721 (BLAST-predicted as containing a thrombospondin type 3 repeat). Moreover, we identified 10 copies of TERC_RNA located on smed_chr4, while *TERT* gene (SMESG000061912.1) was located on smed_chr3.

3.10. Comparative analysis of selected S. mediterranea genome annotations

To ensure the reliability of schMed_Pn genome annotation, we compared it with commonly used genome annotations of *S. mediterranea*. First, we analyzed transcript length distributions. It revealed that all annotations predominantly consisted of relatively short transcripts, as reflected by a sharp decline in cumulative transcript counts with increasing transcript length (Fig. 7A). Specifically, dd_Smed_v6 showed the steepest decline in cumulative transcript counts with increasing length, suggesting a higher fragmentation rate; however, caution is advised in interpreting these distributions, as there is no "standard" shape for planarian transcript length. In contrast, schMed_Pn demonstrated a more gradual decline, which indicated a broader distribution of transcript lengths. Assemblies such as schMedS1, schMedS2, schMedS3_1, and schMedS3_2 exhibited moderate slopes with slight differences, indicating varying transcript length distributions among them. The SmedAsxl genome_v1.1 displayed a narrower range of longer transcripts. In contrast, SmedGD_c1 and SMESG.1, similar to schMed_Pn, demonstrated relatively smooth distributions, signifying balanced transcript coverage with a significant proportion of longer transcripts.

The completeness of the transcriptome assemblies was evaluated using BUSCO (Fig. 7B). The dd_Smed_v6 and SMESG.1 annotations achieved the highest single-copy ortholog recovery—783 and 756

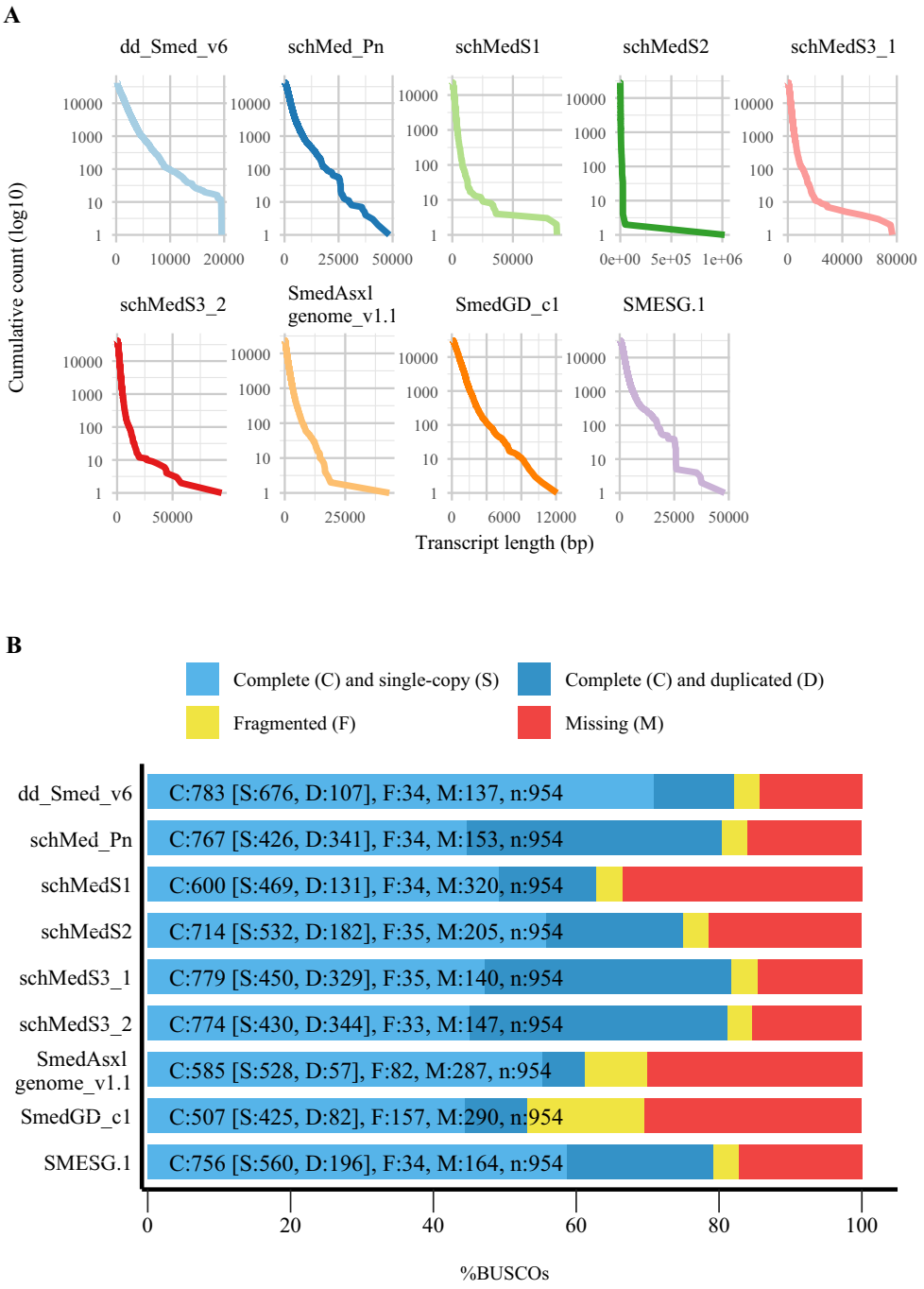

Fig. 7. Comparative analysis of selected *S. mediterranea* genome annotations.
(A) - Cumulative distribution of transcript lengths in selected annotations. The y-axis (logarithmic scale) represents the cumulative count of transcripts/isoforms, indicating how many transcripts exceed each length threshold on the x-axis (linear scale). (B) - BUSCO completeness of selected *S. mediterranea* annotations.

complete BUSCOs out of 954, respectively—with low duplication rates (107 and 196 duplicated BUSCOs), indicative of very accurate core gene models but at the cost of fewer total transcript isoforms. In contrast, schMed_Pn attained a balanced profile of completeness and isoform richness, with 767 complete BUSCOs (426 single-copy, 341 duplicated), coupled with minimal fragmentation (34) and moderate missing counts (153). This reflected our pipeline's emphasis on maximizing isoform discovery without substantially sacrificing core-gene completeness. Other assemblies (schMedS1–3, SmedAsxl genome_v1.1) showed lower complete counts and higher fragmentation or missing rates, consistent with greater redundancy or gaps. Based on BUSCO analysis, both SMESG.1 and dd_Smed_v6 demonstrated robust performance (albeit with different trade-offs in duplication and missing orthologs). Notably, schMed_Pn also ranked among the stronger annotations, demonstrating moderate duplication and missing ortholog levels, while retaining high completeness.

In addition, we employed rnaQUAST and Transrate to generate detailed assembly quality metrics (Table 2). Notably, schMed_Pn exhibited a high number of transcripts over 1000 nt and maintained a competitive N50 value (the length of the shortest contig or scaffold covering 50 % of the total assembly), reflecting effective transcript coverage and solid assembly quality.

Performance benchmarking enabled the identification of improvements and areas needing enhancement, while also positioning

Table 2
Comparative metrics for selected transcriptome assemblies of *S. mediterranea.*

Metric	dd_Smed_v6	schMed_Pn	schMedS1	schMedS2	schMedS3_1	schMedS3_2	SmedAsxl_v1.1	SmedGD_c1.3	SMESG.1
No. of Transcripts	41,745	41,396	22,348	27,790	41,358	41,036	22,855	29,850	31,102
Min transcript length (nt)	201	200	267	267	145	131	75	87	267
Max transcript length (nt)	19,524	48,466	86,373	1,022,601	76,733	94,201	43,226	12,11	48,121
Total Bases (M)	53.23	97.09	40.12	57.93	72.60	71.81	25.55	23.15	65.55
Mean transcript length (nt)	1275	2345	1795	2084	1755	1750	1114	770	2108
No. of transcripts > 1 k	18,487	31,199	16,082	20,727	29,402	28,932	8610	7218	23,648
No. of transcripts > 10 k	94	702	67	293	116	119	47	2	346
No. of transcripts with ORF	20,461	32,798	18,694	23,854	31,503	31,147	12,937	17,567	26,665
N50 (nt)	1931	3070	2159	2535	2133	2141	1653	1011	2570
GC (%)	0.34	0.33	0.33	0.33	0.33	0.33	0.34	0.36	0.33

schMed_Pn within the context of current resources. These analyses collectively demonstrated that our approach offers a balanced and complete annotation, effectively capturing a wide range of transcript lengths and maintaining high completeness.

4. Discussion

Numerous pipelines and tools have been developed to tackle genome annotation across diverse organisms. For well-studied species—such as *Drosophila melanogaster* or *Arabidopsis thaliana*—community-maintained platforms (FlyBase [51] and TAIR [52,53] respectively) provide robust resources that streamline annotation. In prokaryotes, automated tools like Prokka [10] efficiently handle smaller, less complex genomes, while eukaryotic model organisms often benefit from RefSeq [7] or large consortia efforts, which incorporate extensive experimental validation. However, non-model eukaryotes, including many invertebrates and plants, typically face an incomplete reference, novel repetitive elements, and limited training data for *ab initio* predictions. Pipelines such as MAKER [5,54], BRAKER [55,56], GGA [9], nf-core GenomeAnnotator [12], and others can automate workflow portions—especially when reference data or orthologous gene sets are available. However, in certain species with high genomic complexity or less functional annotation data, these tools may still underperform or require extensive manual curation.

In this study, we developed and validated a comprehensive genome annotation pipeline designed to address the challenges of working with non-classical model organisms, while generating high-quality annotations applicable to broader genomic research. Our pipeline uniquely combines: (i) the integration of multiple data types, (ii) automated artifact detection, (iii) dual approach assembly —reference-based plus *de novo*, and (iv) broad functional integration.

The BUSCO completeness of SmedAnno (the automated part of the pipeline) was within 4 % of BRAKER3 in *A. thaliana*, surpassed it by approximately 2 % in *S. mediterranea*, and lagged behind by approximately 10 % in *C. elegans*. Notably, SmedAnno runs consistently doubled the proportion of duplicated BUSCO genes. Because BUSCO counts every alternatively spliced isoform as a duplicate, such inflation is widely regarded as a proxy for improved isoform recovery rather than assembly noise [57]. In practical terms, this means that the apparent redundancy in SmedAnno is biologically informative when transcript diversity, rather than the most parsimonious gene count, is the primary research goal.

Leveraging this enhanced isoform recovery, our analysis identified 4011 novel genes and multiple previously uncharacterized isoforms and ncRNAs in the asexual strain, enriching the genomic landscape of *S. mediterranea*. Notably, the GO term analysis of putative MSTRG genes revealed enrichment in diverse biological functions. The occurrence of these genes necessitates further investigation to better understand their implications and potential significance. While our transcriptomic data offer strain-specific insights, it should be considered that the annotation relies on the genome of the sexual strain as the reference. Consequently, differences between strains, such as chromosomal translocation and other genetic variations, may compromise the accuracy of gene prediction and transcript assembly, leading to potential inconsistencies in annotation quality.

The existing, planarian-specific repositories—such as PlanMine [21], PLANAtools [58], WormBase [59], PlanoSphere [60], PlanNET [61], and various planaria single-cell datasets [62–65]—provide essential resources for gene model comparisons and functional information in *S. mediterranea*. Despite their value, discrepancies in naming conventions, annotation updates, and differences between sexual and asexual strains can lead to inconsistencies. These challenges highlight the importance of robust, well-documented pipelines that can reconcile or supplement these external databases, thus ensuring more accurate gene models.

A detailed inspection of the curated annotation uncovered two intriguing features of the genomic architecture: genes located in close proximity and non-contiguous transcripts. Although these observations could potentially represent annotation artifacts, it is crucial to carefully assess them in the context of established genomic knowledge. Genes located in close proximity can reflect various biological scenarios. One possibility is the presence of functionally independent genes that happen to be closely spaced, as observed in other genomes where gene density is high [66,67]. For example, two planarian genes—SMESG000043996.1 and SMESG000043998.1—that encode kynureninase and synaptotagmin XVII, respectively, are unrelated but positioned close to each other. Alternatively, such arrangements may arise from evolutionary events such as gene duplication followed by functional divergence, where both copies remain in close physical proximity. Another scenario includes operon-like gene structures commonly seen in prokaryotes but also identified in some eukaryotic contexts [68]. These structures allow the co-regulation of adjacent genes through shared promoters or other regulatory elements. Bidirectional promoters, enabling two genes to be transcribed in opposite directions from the same regulatory region, represent another mechanism driving close gene arrangements [69]. On the other hand non-contiguous transcripts, where exons are unusually spaced or appear fragmented, often arise from structural variations, alternative splicing, or sequencing artifacts. Biologically, large introns or dispersed exonic regions can be linked to long-range regulatory elements or structural genome features such as transposon insertions. Moreover, such patterns may indicate complex gene models, as seen in genes encoding for proteins with repetitive domains [70,71] or in cases where alternative promoters and polyadenylation sites lead to diverse transcript forms [72,73]. Consequently, accurately capturing these intricate gene architectures often requires more comprehensive strategies than those employed for simpler genomes, underscoring the need for user scrutiny and flexibility in annotation pipelines. The primary advantage of our pipeline lies in its capacity to address these complexities. By combining reference-based and *de novo* methods and incorporating both coding and non-coding RNA annotations it addresses the challenges posed by scarcely studied genomes. However, the pipeline's reliance on user-defined thresholds and manual curation introduces subjectivity, potentially impacting reproducibility.

While the automated steps of SmedAnno effectively identify

fragmented and chimeric transcripts, resolving these artifacts requires expertise and thus is not fully automated. Another limitation concerns functional annotations, which can lead to false positives in identifying chimeric genes. For instance, the same protein from different species or proteins representing different subunits of a complex, may be annotated as distinct entries in functional databases. This can lead to genes being incorrectly flagged as chimeric if the functional entries do not exactly match. Accordingly, development of refined algorithms that account for near-identical annotations and minor terminological differences is vital.

Future enhancements could include the integration of machine learning algorithms to automate artifact resolution and improve annotation consistency. The present container infrastructure already simplifies the implementation of graphics processing units (GPUs) and central processing units (CPUs). Expanding the pipeline to support multi-omics data integration, such as proteomics and epigenomics, would further enhance its utility. Community-driven efforts to standardize annotation protocols and databases would ensure broader applicability and reproducibility. Finally, experimental validation of novel genes and isoforms identified through this pipeline will be critical to confirm their biological relevance.

Compared to pipelines that excel in either prokaryotic or model vertebrate annotation –where reference data is abundant and well-curated – our approach fills the gap for non-classical model systems with more limited resources. In doing so, it advances genome annotation efforts for *S. mediterranea* and provides a robust framework for addressing similar challenges in other species, fostering broader genomic discoveries.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2025.111104.

Availability of source code and requirements

Project name: SmedAnno.
Project homepage: https://github.com/Norreanea/SmedAnno
Operating system: Linux.
Container image: Docker ≥20.10, docker-compose ≥2.5.
Programming languages: shell, R, Python, awk.
License: MIT License.

Ethics in publishing statement

This research presents an accurate account of the work performed, all data presented are accurate and methodologies detailed enough to permit others to replicate the work.

This manuscript represents entirely original works and or if work and/or words of others have been used, that this has been appropriately cited or quoted and permission has been obtained where necessary.

This material has not been published in whole or in part elsewhere.

The manuscript is not currently being considered for publication in another journal.

That generative AI and AI-assisted technologies have not been utilized in the writing process or if used, disclosed in the manuscript the use of AI and AI-assisted technologies and a statement will appear in the published work.

That generative AI and AI-assisted technologies have not been used to create or alter images unless specifically used as part of the research design where such use must be described in a reproducible manner in the methods section.

All authors have been personally and actively involved in substantive work leading to the manuscript and will hold themselves jointly and individually responsible for its content.

CRediT authorship contribution statement

Anastasiia Zaremba: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Conceptualization. **Małgorzata Marszałek-Zeńczak:** Writing – review & editing, Methodology, Conceptualization. **Annasha Dutta:** Writing – review & editing, Investigation. **Anna Samelak-Czajka:** Writing – review & editing, Investigation. **Paulina Jackowiak:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used the free version of OpenAI-ChatGPT c2022 (v.GPT-o3-mini, February 2025) in order to improve manuscript readability and language. After using this tool, the authors thoroughly reviewed and modified the manuscript and took full responsibility for the content of the publication. The literature survey was done only by the authors, with no input from the AI tools.

Funding

This work was financed by the National Science Centre, Poland, through a grant no. 2019/35/B/NZ2/02658 to PJ.

Declaration of competing interest

Nothing to declare.

Data availability

The *in-silico* datasets supporting the results and all other code used in this article are available in the GitHub repository archived in Zenodo https://doi.org/10.5281/zenodo.15130394. RNA-seq data can be accessed from the following GEO repositories: GSE293999 and GSE293997 from NCBI BioProject PRJNA1243063; and GSE294569 from NCBI BioProject PRJNA1246370. The final genome annotation file is available in GSE294569.

References

[1] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, B. Morgenstern, AUGUSTUS: ab initio prediction of alternative transcripts, Nucleic Acids Res. (2006), https://doi.org/10.1093/nar/gkl200.
[2] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, et al., BLAST+: architecture and applications, BMC Bioinformatics (2009), https://doi.org/10.1186/1471-2105-10-421.
[3] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, et al., STAR: ultrafast universal RNA-seq aligner, Bioinformatics (2013), https://doi.org/10.1093/bioinformatics/bts635.
[4] A. Shumate, B. Wong, G. Pertea, M. Pertea, Improved transcriptome assembly using a hybrid of long and short reads with StringTie, PLoS Comput. Biol. (2022), https://doi.org/10.1371/journal.pcbi.1009730. Public Library of Science.
[5] B.L. Cantarel, I. Korf, S.M.C. Robb, G. Parra, E. Ross, B. Moore, et al., MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes, Genome Res. (2008), https://doi.org/10.1101/gr.6743907.
[6] B.J. Haas, S.L. Salzberg, W. Zhu, M. Pertea, J.E. Allen, J. Orvis, et al., Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments, Genome Biol. (2008), https://doi.org/10.1186/gb-2008-9-1-r7.
[7] N.A. O'Leary, M.W. Wright, J.R. Brister, S. Ciufo, D. Haddad, R. McVeigh, et al., Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, Nucleic Acids Res. (2016), https://doi.org/10.1093/nar/gkv1189. Oxford University Press.
[8] C. Kurylo, C. Guyomar, S. Foissac, S. Djebali, TAGADA: a scalable pipeline to improve genome annotations with RNA-seq data, NAR Genom. Bioinform. (2023), https://doi.org/10.1093/nargab/lqad089. Oxford University Press.
[9] S. Hiltemann, H. Rasche, S. Gladman, H.R. Hotz, D. Larivière, D. Blankenberg, et al., Galaxy training: a powerful framework for teaching!, PLoS Comput. Biol. (2023), https://doi.org/10.1371/journal.pcbi.1010752. Public Library of Science.
[10] T. Seemann, Prokka: rapid prokaryotic genome annotation, Bioinformatics (2014), https://doi.org/10.1093/bioinformatics/btu153. Oxford University Press.
[11] M.T. Swain, I.J. Tsai, S.A. Assefa, C. Newbold, M. Berriman, T.D. Otto, A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs, Nat. Protoc. (2012), https://doi.org/10.1038/nprot.2012.068.
[12] P.A. Ewels, A. Peltzer, S. Fillinger, H. Patel, J. Alneberg, A. Wilm, et al., The nf-core framework for community-curated bioinformatics pipelines, Nat. Biotechnol. (2020), https://doi.org/10.1038/s41587-020-0439-x.

[13] S. Wu, S. Yuan, Z. Su, User-Friendly Genome Assembly and Gene Annotation Pipelines for Vertebrates, bioRxiv (2024), https://doi.org/10.1101/2024.05.21.595213.

[14] G.A. Maia, V.B. Filho, E.K. Kawagoe, T.A. Teixeira Soratto, R.S. Moreira, E. C. Grisard, et al., AnnotaPipeline: an integrated tool to annotate eukaryotic proteins using multi-omics data, Front. Genet. (2022), https://doi.org/10.3389/fgene.2022.1020100.

[15] Humann J.L., Lee T., Ficklin S., Main D. Structural and Functional Annotation of Eukaryotic Genomes with GenSAS.

[16] A. Sánchez Alvarado, The freshwater planarian *Schmidtea mediterranea*: embryogenesis, stem cells and regeneration, in: Curr. Opin. 13, Elsevier Ltd, Genet. Dev, 2003, pp. 438–444, https://doi.org/10.1016/s0959-437x(03)00082-0.

[17] J. Baguñà, S. Carranza, M. Pala, C. Ribera, G. Giribet, M.A. Arnedo, et al., From morphology and karyology to molecules. New methods for taxonomical identification of asexual populations of freshwater planarians. A tribute to professor Mario Benazzi, Ital. J. Zool. (1999), https://doi.org/10.1080/11250009909356258. Societa Paleontologica Italiana.

[18] S.M.C. Robb, E. Ross, A.S. Alvarado, SmedGD: the *Schmidtea mediterranea* genome database, Nucleic Acids Res. (2008), https://doi.org/10.1093/nar/gkm684.

[19] S.M.C. Robb, K. Gotting, E. Ross, Alvarado A. Sánchez, SmedGD 2.0: the *Schmidtea mediterranea* genome database, Genesis (2015), https://doi.org/10.1002/dvg.22872. John Wiley and Sons Inc.

[20] M.A. Grohme, S. Schloissnig, A. Rozanski, M. Pippel, G.R. Young, S. Winkler, et al., The genome of *Schmidtea mediterranea* and the evolution of core cellular mechanisms, Nature (2018), https://doi.org/10.1038/nature25473. Nature Publishing Group.

[21] A. Rozanski, H.K. Moon, H. Brandl, J.M. Martín-Durán, M.A. Grohme, K. Hüttner, et al., PlanMine 3.0 - improvements to a mineable resource of flatworm biology and biodiversity, Nucleic Acids Res. (2019), https://doi.org/10.1093/nar/gky1070. Oxford University Press.

[22] L. Guo, J.S. Bloom, D. Dols-Serrate, J. Boocock, E. Ben-David, O.T. Schubert, et al., Island-specific evolution of a sex-primed autosome in a sexual planarian, Nature (2022), https://doi.org/10.1038/s41586-022-04757-3. Nature Research.

[23] M. Ivanković, J.N. Brand, L. Pandolfini, T. Brown, M. Pippel, A. Rozanski, et al., A comparative analysis of planarian genomes reveals regulatory conservation in the face of rapid structural divergence, Nat. Commun. (2024), https://doi.org/10.1038/s41467-024-52380-9.

[24] A. Rabiasz, E. Ziętkiewicz, Schmidtea mediterranea as a model organism to study the molecular background of human motile ciliopathies, Int. J. Mol. Sci. Multidisciplinary Digital Publishing Institute (MDPI) 24 (5) (2023) 4472, https://doi.org/10.3390/ijms24054472.

[25] S. Quinodoz, M.A. Thomas, J. Dunkel, E.M. Schötz, The more the merrier?: entropy and statistics of asexual reproduction in freshwater planarians, J. Stat. Phys. (2011), https://doi.org/10.1007/s10955-011-0157-3. Springer New York LLC.

[26] P.W. Reddien, Cell 175 (2018) 327–345, https://doi.org/10.1016/j.cell.2018.09.021.

[27] P.W. Reddien, A. Sánchez Alvarado, Fundamentals of planarian regeneration, Annu. Rev. Cell Dev. Biol. 20 (1) (2004) 725–757, https://doi.org/10.1146/annurev.cellbio.20.010403.095114.

[28] B.W. Benham-Pyle, F.G. Mann, C.E. Brewster, E.R. Dewars, D.M. Vuu, S. H. Nowotarski, et al., Planarians Employ Diverse and Dynamic Stem Cell Microenvironments to Support Whole-Body Regeneration, 2022.

[29] E.W. Sayers, E.E. Bolton, J.R. Brister, K. Canese, J. Chan, D.C. Comeau, et al., Database resources of the national center for biotechnology information, Nucleic Acids Res. (2022), https://doi.org/10.1093/nar/gkab1112. Oxford University Press.

[30] N. Palmieri, C. Schlötterer, Mapping accuracy of short reads from massively parallel sequencing and the implications for quantitative expression profiling, PLoS One (2009), https://doi.org/10.1371/journal.pone.0006323.

[31] F.J. Sedlazeck, H. Lee, C.A. Darby, M.C. Schatz, Piercing the dark matter: bioinformatics of long-range sequencing and mapping, Nat. Rev. Genet. (2018), https://doi.org/10.1038/s41576-018-0003-4.

[32] Z. Chen, N.U. Ain, Q. Zhao, X. Zhang. From tradition to innovation: conventional and deep learning frameworks in genome annotation, Brief. Bioinform 25, Oxford University Press, 2024, https://doi.org/10.1093/bib/bbae138 bbae138.

[33] P.A. Newmark, A. Sánchez Alvarado, Bromodeoxyuridine specifically labels the regenerative stem cells of planarians, Dev. Biol. (2000), https://doi.org/10.1006/dbio.2000.9645. Academic Press Inc.

[34] J. Dainat, D. Hereñú, K.D. Murray, E. Davis, I. Ugrin, K. Crouch, et al., NBISweden/AGAT: AGAT-v1.4.1 (v1.4.1), Zenodo (2024), https://doi.org/10.5281/zenodo.13799920.

[35] P. Lamesch, T.Z. Berardini, D. Li, D. Swarbreck, C. Wilks, R. Sasidharan, et al., The arabidopsis information resource (TAIR): improved gene annotation and new tools, Nucleic Acids Res. (2012), https://doi.org/10.1093/nar/gkr1090.

[36] Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*, Nature (2000), https://doi.org/10.1038/35048692.

[37] K.L. Howe, B.J. Bolt, M. Shafie, P. Kersey, M. Berriman, WormBase ParaSite − a comprehensive resource for helminth genomics, Mol. Biochem. Parasitol. (2017), https://doi.org/10.1016/j.molbiopara.2016.11.005.

[38] Genome sequence of the nematode *C. elegans*: a platform for investigating biology, Science (1979) (1998), https://doi.org/10.1126/science.282.5396.2012.

[39] L. Gabriel, T. Brůna, K.J. Hoff, M. Ebel, A. Lomsadze, M. Borodovsky, et al., BRAKER3: Fully Automated Genome Annotation using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA, Genome Res. 34 (5) (2024) 769–777, https://doi.org/10.1101/gr.278090.123.

[40] B.J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, et al., De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis, Nat. Protoc. (2013), https://doi.org/10.1038/nprot.2013.084.

[41] M. Abrar, D. Hussain, I.A. Khan, F. Ullah, M.A. Haq, M.A. Aleisa, et al., DeepSplice: a deep learning approach for accurate prediction of alternative splicing events in the human genome, Front. Genet. (2024), https://doi.org/10.3389/fgene.2024.1349546.

[42] M. Pertea, G. Pertea, GFF utilities: GffRead and GffCompare, F1000Res (2020), https://doi.org/10.12688/f1000research.23297.1. F1000 Research Ltd.

[43] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics (2015), https://doi.org/10.1093/bioinformatics/btv351. Oxford University Press.

[44] E. Bushmanova, D. Antipov, A. Lapidus, V. Suvorov, A.D. Prjibelski, RnaQUAST: a quality assessment tool for de novo transcriptome assemblies, Bioinformatics (2016), https://doi.org/10.1093/bioinformatics/btw218. Oxford University Press.

[45] R. Smith-Unna, C. Boursnell, R. Patro, J.M. Hibberd, S. Kelly, TransRate: reference-free quality assessment of de novo transcriptome assemblies, Genome Res. (2016), https://doi.org/10.1101/gr.196469.115. Cold Spring Harbor Laboratory Press.

[46] B.A. Sweeney, A.I. Petrov, C.E. Ribas, R.D. Finn, A. Bateman, M. Szymanski, et al., RNAcentral 2021: secondary structure integration, improved sequence search and new member databases, Nucleic Acids Res. (2021), https://doi.org/10.1093/nar/gkaa921.

[47] V. Lakshmanan, T.N. Sujith, D. Bansal, P.V. Shivaprasad, D. Palakodeti, S. Krishna, Comprehensive annotation and characterization of planarian tRNA and tRNA-derived fragments (tRFs), RNA (2021), https://doi.org/10.1261/rna.077701.120.

[48] W.R. Francis, G. Wörheide, Similar ratios of introns to intergenic sequence across animal genomes, Genome Biol. Evol. (2017), https://doi.org/10.1093/gbe/evx103. Oxford University Press.

[49] S. Gudlaugsdottir, D.R. Boswell, G.R. Wood, J. Ma, Exon size distribution and the origin of introns, Genetica (2007), https://doi.org/10.1007/s10709-007-9139-4.

[50] A. Bhasi, P. Philip, V. Manikandan, P. Senapathy, ExDom: an integrated database for comparative analysis of the exon-intron structures of protein domains in eukaryotes, Nucleic Acids Res. (2009), https://doi.org/10.1093/nar/gkn746.

[51] A. Öztürk-Çolak, S.J. Marygold, G. Antonazzo, H. Attrill, D. Goutte-Gattat, V. K. Jenkins, et al., FlyBase: updates to the *Drosophila* genes and genomes database, Genetics (2024), https://doi.org/10.1093/genetics/iyad211.

[52] T.Z. Berardini, L. Reiser, D. Li, Y. Mezheritsky, R. Muller, E. Strait, et al., The arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome, Genesis (2015), https://doi.org/10.1002/dvg.22877.

[53] C. Cheng, V. Krishnakumar, A.P. Chan, F. Thibaud-Nissen, S. Schobel, C.D. Town, Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome, Plant J. (2017), https://doi.org/10.1111/tpj.13415.

[54] C. Holt, M. Yandell, MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects, BMC Bioinformatics (2011), https://doi.org/10.1186/1471-2105-12-491.

[55] K.J. Hoff, A. Lomsadze, M. Borodovsky, M. Stanke, Whole-Genome Annotation with BRAKER, Methods Mol. Biol. (2019), https://doi.org/10.1007/978-1-4939-9173-0_5.

[56] T. Brůna, K.J. Hoff, A. Lomsadze, M. Stanke, M. Borodovsky, BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database, NAR Genom. Bioinform. (2021), https://doi.org/10.1093/nargab/lqaa108.

[57] M. Manni, M.R. Berkeley, M. Seppey, E.M. Zdobnov, BUSCO: assessing genomic data quality and beyond, Curr. Protoc. (2021), https://doi.org/10.1002/cpz1.323.

[58] M. Hoffman, O. Wurtzel, PLANAtools—an interactive gene expression repository for the planarian *Schmidtea mediterranea*, Front. Cell Dev. Biol. (2023), https://doi.org/10.3389/fcell.2023.1149537.

[59] P.W. Sternberg, K. Van Auken, Q. Wang, A. Wright, K. Yook, M. Zarowiecki, et al., WormBase 2024: status and transitioning to alliance infrastructure, Genetics (2024), https://doi.org/10.1093/genetics/iyae050.

[60] S.H. Nowotarski, E.L. Davies, S.M.C. Robb, E.J. Ross, N. Matentzoglu, V. Doddihal, et al., Planarian anatomy ontology: a resource to connect data within and across experimental platforms, Development (2021), https://doi.org/10.1242/dev.196097.

[61] S. Castillo-Lara, J.F. Abril, PlanNET: homology-based predicted interactome for multiple planarian transcriptomes, Bioinformatics (2018), https://doi.org/10.1093/bioinformatics/btx738.

[62] G. Cui, K. Dong, J.-Y. Zhou, S. Li, Y. Wu, Q. Han, et al., Spatiotemporal transcriptomic atlas reveals the dynamic characteristics and key regulators of planarian regeneration, Nat. Commun. (2023), https://doi.org/10.1038/s41467-023-39016-0.

[63] H.O. King, K.E. Owusu-Boaitey, C.T. Fincher, P.W. Reddien, A transcription factor atlas of stem cell fate in planarians, Cell Rep. (2024), https://doi.org/10.1016/j.celrep.2024.113843.

[64] C.T. Fincher, O. Wurtzel, T. de Hoog, K.M. Kravarik, P.W. Reddien, Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*, Science (1979) (2018), https://doi.org/10.1126/science.aaq1736.

[65] M. Plass, J. Solana, F.A. Wolf, S. Ayoub, A. Misios, P. Glažar, et al., Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics, Science (1979) (2018), https://doi.org/10.1126/science.aaq1723.

[66] J.T. Arnone, A. Robbins-Pianka, J.R. Arace, S. Kass-Gergi, M.A. McAlear, The adjacent positioning of co-regulated gene pairs is widely conserved across eukaryotes, BMC Genomics (2012), https://doi.org/10.1186/1471-2164-13-546.

[67] A. Gherman, R. Wang, D. Avramopoulos, Orientation, distance, regulation and function of neighbouring genes, Hum. Genomics (2009), https://doi.org/10.1186/1479-7364-3-2-143.
[68] J. Kominek, D.T. Doering, D.A. Opulente, X.X. Shen, X. Zhou, J. DeVirgilio, et al., Eukaryotic acquisition of a bacterial operon, Cell (2019), https://doi.org/10.1016/j.cell.2019.01.034. Cell Press.
[69] S.S. Ahmad, N.S.N. Samia, A.S. Khan, R.R. Turjya, M.A.A.K. Khan, Bidirectional promoters: an enigmatic genome architecture and their roles in cancers, in: Mol. Biol. Rep. 48, Springer Science and Business Media B.V, 2021, pp. 6637–6644, https://doi.org/10.1007/s11033-021-06612-6.
[70] T.C. Hoock, L.L. Peters, S.E. Lux, Isoforms of ankyrin-3 that lack the NH2-terminal repeats associate with mouse macrophage lysosomes, J. Cell Biol. (1997), https://doi.org/10.1083/jcb.136.5.1059.
[71] G. Yamankurt, H.C. Wu, M. McCarthy, S.R. Cunha, Exon organization and novel alternative splicing of Ank3 in mouse heart, PLoS One (2015), https://doi.org/10.1371/journal.pone.0128177.
[72] I. Aísa-Marín, R. García-Arroyo, S. Mirra, G. Marfany, The alter retina: alternative splicing of retinal genes in health and disease, Int. J. Mol. Sci. (2021), https://doi.org/10.3390/ijms22041855.
[73] S. Pal, R. Gupta, H. Kim, P. Wickramasinghe, V. Baubet, L.C. Showe, et al., Alternative transcription exceeds alternative splicing in generating the transcriptome diversity of cerebellar development, Genome Res. (2011), https://doi.org/10.1101/gr.120535.111.