Investigation into gene regulatory networks governing the development and maintenance of the planarian reproductive system

Dissertation
for the award of the degree
“Doctor rerum naturalium”
of the Georg-August-Universität Göttingen

within the doctoral program
“International Max Planck Research School for Genome Science”
of the Georg-August University School of Science (GAUSS)

submitted by
Thomas Brochier
from Brussels
Göttingen 2023

Thesis advisory committee:

Prof. Dr. Jochen Rink (Supervisor and 1st Reviewer)
Max Planck Institute for Multidisciplinary Sciences, Department of Tissue Dynamics and Regeneration

Prof. Dr. Argyris Papantonis (2nd Reviewer)
University Medical Center Göttingen, Institute of Pathology

Dr. Nico Posnien
Johann-Friedrich-Blumenbach-Institute of Zoology and Anthropology, Department of Developmental Biology

Members of the Examination Board:

Dr. Marieke Oudelaar
Max Planck Institute for Multidisciplinary Sciences, Department of Molecular Biology

Dr. Ufuk Günesdogan
Göttingen Center for Molecular Biosciences, Department of Developmental Biology

Prof. Dr. Carsten Lüder
University Medical Centre Göttingen, Department for Medical Microbiology

Date of oral examination: 12th January, 2024

Declaration of independent work

I hereby declare that this doctoral thesis entitled “Investigation into gene regulatory networks governing the development and maintenance of the planarian reproductive system” has been written independently and with no other sources or aids than those referenced within.

Abstract

Regulation of gene expression is a fundamental mechanism allowing the existence of complex living systems. This regulation takes shape in the form of hierarchical gene regulatory networks (GRNs), culminating in the binding of transcription factors (TFs) to regulatory elements (REs). Uncovering the structures of these GRNs has shed light on the mechanistic basis of complex spatiotemporal processes like embryonic development. The reconstruction of such GRN, for any biological process, is therefore the ultimate goal for its understanding. The planarian *Schmidtea mediterranea* is an important model organism for the study of adult stem cell systems, cell differentiation and regeneration. However, very little is known about the structure of GRNs in this organism due to the historical lack of adequate tools to study regulatory elements.

At the beginning of my thesis, I developed a robust Start-seq protocol to study RE activity in *S. mediterranea* by probing for transcription initiation events. Following this, I characterized the identified putative REs on the basis of their distribution, chromatin context as well as motif content. Putative REs identified using Start-seq possess characteristic epigenomic signatures as they are situated within regions of open chromatin and are enriched for active epigenetic marks such as H3K27Ac, H3K4me3 and H3k4me1. I next showed that both planarian putative enhancers and promoter are enriched for core promoter motifs that are mostly situated at the expected position. Interestingly, the DPE motif was enriched at the transcription initiation site in both types of REs instead of its described position at ± +30 nt. Following this, I showed that most of the identified REs showed sign of bidirectional transcription initiation which is considered as a characteristic feature of REs.

After characterizing the REs, I leveraged the existence of two naturally occurring biotypes within *S. mediterranea* (sexual and asexual) to investigate gene regulatory networks (GRNs) responsible for the development and maintenance of the planarian reproductive system. These biotypes are genetically similar, but one fails to develop sexual organs. Comparing their transcription initiation landscapes allowed me to identify multiple transcription factors potentially part of in these GRNs. Notably, the majority of the selected TF candidates exhibited expression in sexual organs, with half of the candidates being specifically expressed in these tissues. Knockdown of three candidates (*tead 1*, *thap* and *cebp 4*) showed dysregulation of many reproductive genes. *In situ* hybridizations on sexual markers after RNAi of the three candidates confirmed the abrogation of sexual tissues in these conditions. The results indicated that *tead 1* produced a severe phenotype, where most sexual organs were absent, except for the ovaries. The *thap* candidate exhibited a defect in shell gland formation and yolk tissue patterning, while the *cebp 4* candidate displayed an impairment in sperm differentiation.

Overall, this study was not able to recreate a GRN governing the development and maintenance of the reproductive tissue in *S. mediterranea.* However, it lays a foundation for further research on GRNs in planaria as it shows that identification of differentially active REs using the established protocol was successful in obtaining important regulators the planarian reproductive system.

Acknowledgments

During my time in Germany, I had the pleasure to meet, work or simply be around multiple persons who made me grow both scientifically as well as personally. I owe them a debt of gratitude as this thesis would not have come to fruition without them.

First and foremost, I would like to thank my advisor Dr. Jochen Rink to for having given me the opportunity to work with him. Your passion and excitement for science is something amazing that inspired me all along my thesis. Your resilience in the face of hardships that came with the lab move and the pandemic is admirable. I would like to thank my TAC members Prof. Dr. Argyris Papantonis and Dr. Nico Posnien for their support and constructive feedback right from the first committee meetings. Your encouragements are something that kept me going.

Many thanks go to past and present lab members of the Rink lab. First, I would like to thank Mario Ivankovic. You acted as a second supervisor to me and helped me both on the wet- and dry-lab side of my work. I will always remember fondly of our late evening 'last' experiments/analysis before calling it a day. I also thank you for your friendship. Goettingen would have been much more boring without you. I thank Dr. Grohme and Dr. Rosanski for their infinite knowledge in molecular biology/biochemistry and bioinformatics respectively. I would like to thank Dr. Vu for her kindness and patience to teach me how to perform whole-mount *in situs*. Furthermore, I thank Dr. Boothe for his impeccable professionalism and indispensable knowledge on microscopy. I thank you also for inspiring me to finally do a marathon. Speaking about running, I thank Dr. Vila-Farré for the many running sessions we did together and which brought me a lot of joy. I also thank you for your expertise in planarian biology, which was critical to me especially during the last part of my thesis. My gratitude goes also to the scientific support staff of the department of tissue dynamics and regeneration. The lab would not be running without you. My special thanks go to Delia and Fruzsina for being great co-workers. I would also like to thank all the present lab members for making the lab a pleasant work space (shout-out to Jun-Ru, Ziheng, Annabel, Ludwik and Jeremias).

I thank my friends from Dresden, with whom I started this adventure and had to leave to continue my research in Göttingen. I thank the friends I made in Göttingen through the beer hours and various other events (shout-out to Jakob and Amrutha). I also thank my friends from Belgium with whom I still feel as close despite the distance.

I would like to thank my family for their support. In particular, I thank my mother for her unconditional dedication to her children. Finally, I would like to thank Sonali, with whom I started this chapter of my life. I am a better person because of you.

Contents

List of Figures

List of Tables

Abbreviation list

abs	absolute
Bp	base pair
CAGE	Cap Analysis of Gene Expression
ChIP	Chromatin Immunoprecipitation
CGI	CG Island
CIP	Calf Intestinal Phosphatase
CNS	Central Nervous System
DBD	DNA Binding Domain
DE	Differentially Expressed
DGEA	Differential Gene Expression Analysis
DHS	DNase I hypersensitive site
DNA	Desoxyribonucleic Acid
DREAA	Differential Regulatory Element Activity Analysis
DTT	Dithiothreitol
ECM	Extracellular Matrix
EMSA	Electrophoretic Mobility Shift Assay
E-P	Enhancer-Promoter
eRNA	enhancer RNA
FGP	Female Germ cell Progenitor
FSTF	Fate Specifying Transcription Factor
GO	Gene Ontology
GTF	General Transcription Factor
GRN	Gene Regualtory Network
GSC	Germline Stem Cell
HDAC	Histone Deacetylase
IDR	Intrinsically Disordered Region
l2fc	Log2 Fold-Change
MPRA	Massively parallel reporter assays
mRNA	messenger RNA
NDR	Nucleosome Depleted Region
NHR	Nuclear Hormone Receptor
NRO	Nuclear Run On
Nt	nucleotide
ODE	Ordinary Differential Equation
Osm	Osmol
padj	adjusted P value
PC	Principal Component
PFM	Position Frequency Matrix
PIC	Pre-initiation Complex
PROMPT	PROMoter uPstream Transcript
PTM	Post Translational Modification
PWM	Position Weight Matrix
RE	Regulatory Element
RNA	Ribonucleic Acid
RNA pol II	RNA polymerase II
RNAi	RNA interference
rRNA	ribosomal RNA
SAGE	Serial Analysis of Gene Expression

TAD	Topologically Associated Domain
TAP	Tobacco Acid Pyrophosphatase
TBP	TATA Binding Protein
TCA	Tricarboxylic Acid
TE	Transposable Element
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
TIC	Transcription Initiation Cluster
TSS	Transcription Start Site
uaRNA	upstream antisense RNA
UTR	Untranslated Region
VCN	Ventral Nerve Chord
WISH	Whole-mount In Situ Hybridization

Chapter 1

Introduction

This introduction aims at providing the broader context of regulation of gene expression and gene regulatory networks in which my study is placed. The first section provides an overview of our current understanding of gene expression and the factors influencing it. The second section explores gene regulatory networks in terms of structure and their importance during development. In the third section, I introduce the planarian model, outline the relevant research on GRNs in this organism and finish by stating some outstanding questions related to this subject. Finally, in the fourth and last section, I outline the primary objectives of my thesis.

1.1 Regulation of gene expression

All cells in multicellular organisms, apart from a few specialized exceptions, possess the same genetic content. However, it would be energetically and functionally counterproductive for a cell to express every gene product at all times. Cells are inherently dynamic. During embryogenesis, they divide and differentiate into cells with specialized functions that will form the organs and tissues of the organism. This intricate process relies on key regulatory signals and their effectors for proper orchestration. Furthermore, cells possess the capacity to integrate external stimuli and respond appropriately. If effectors were always present in cells or active at all times, it would be impossible to effectively implement the appropriate responses. Hence, precise spatiotemporal regulation of gene expression is therefore paramount for organismal life. Development offers a compelling case for how this regulatory information is encoded. The observation that frogs generate embryos that will always develop into frogs leads to the inevitable conclusion that a species possesses heritable instructions governing the precise timing and location of gene expression. This is achieved by an interplay between cis-regulatory elements and transcription factors, all within the broader context of the chromatin environment. In this part of the introduction, I will delve into these three elements and detail how they collaborate to regulate gene expression.

1.1.1 The regulatory genome

At the center of gene expression regulation lies the regulatory genome, a term describing non-coding elements whose functions is to integrate regulatory signals and regulate gene expression (Davidson & Peter, 2015). This section will focus on two primary classes that directly influence transcription: promoters and enhancers. Additional classes of regulatory elements exist, which are involved in governing gene expression within the three-dimensional structure of the nucleus. They will be introduced in a later section.

Promoters drive gene transcription

Promoters are regulatory elements located within the upstream region of genes and serve as platform to initiate gene expression (Figure 1.1). Within this region, the core promoter plays an essential role in transcription initiation. It contains sequences that are bound by general transcription factors (GTFs) and RNA polymerase II, collectively forming the pre-initiation complex (PIC). Further upstream within the promoter is a region that acts as a binding platform for transcription factors and serves to additionally regulate gene expression. This second region is referred to as the proximal promoter and is variable in size (Huminiecki & Horbanczuk, 2017). The definition of a promoter is therefore more arbitrary whereas the core promoter answers to more defined criteria.

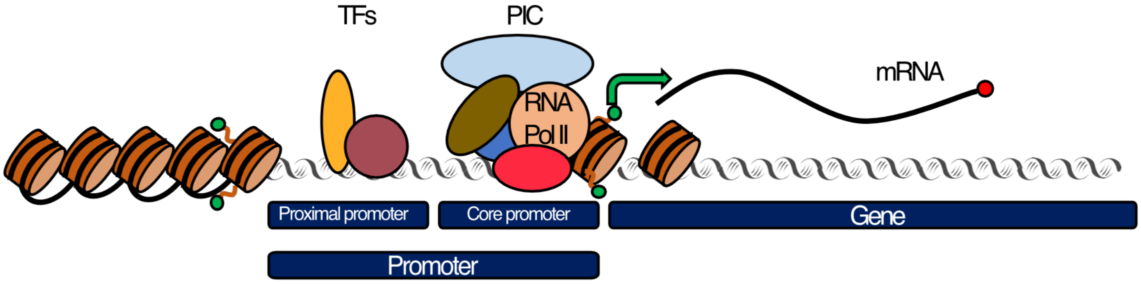


Figure 1.1: Schematic of a promoter. Promoters are located upstream of the 5' end of genes. They are composed of a proximal promoter just upstream of the gene that binds components of the pre-initiation complex. This pre-initiation complex (PIC) is composed of general transcription factors and RNA polymerase II . Promoters also possess another region called the proximal promoter which can harbor transcription factor binding sites for other transcription factors. Nucleosomes flanking the promoter region are enriched for histone modifications represented here as green dots.

Promoter architecture

Core promoters contain sequences affecting transcription The core promoter primarily serves to assemble the transcriptional machinery upstream of its associated gene by containing short DNA sequences known as motifs, to which general transcription factors (GTFs) bind. In the case of the core promoter, these motifs are called core promoter motifs.

Various methods have been used to identify such motifs. For instance, the initiator motif (Inr), located at the transcription start site was identified to be necessary for transcription by so-called promoter bashing: Smale and Baltimore assessed the transcriptional output of successively smaller regions of the terminal deoxynucleotidyltransferase gene in an *in vitro* setting and concluded that a 17-mer was sufficient to drive transcription of the reporter gene (Smale & Baltimore, 1989). Since it was located at the transcription initiation site (TSS), they decided to call it the initiator motif. Comparisons of multiple promoter sequences enabled the identification of the TATA-box (Lifton, Goldberg, Karp, & Hogness, 1978) and the Downstream Promoter Element (DPE) (Burke & Kadonaga, 1996) motifs. TSS profiling techniques greatly helped at further characterizing the core promoter motif content. Using computational tools, researchers were able to identify overrepresented sequences in core promoters in a genome-wide manner (FitzGerald, Sturgill, Shyakhtenko, Oliver, & Vinson, 2006; Hendrix, Hong, Zeitlinger, Rokhsar, & Levine, 2008; Ohler, Liao, Niemann, & Rubin, 2002) (Figure 1.2 A). Many of these motifs have been further characterized and their GTF binding partner identified (Burke & Kadonaga, 1997; Louder et al., 2016) (Figure 1.2 B).

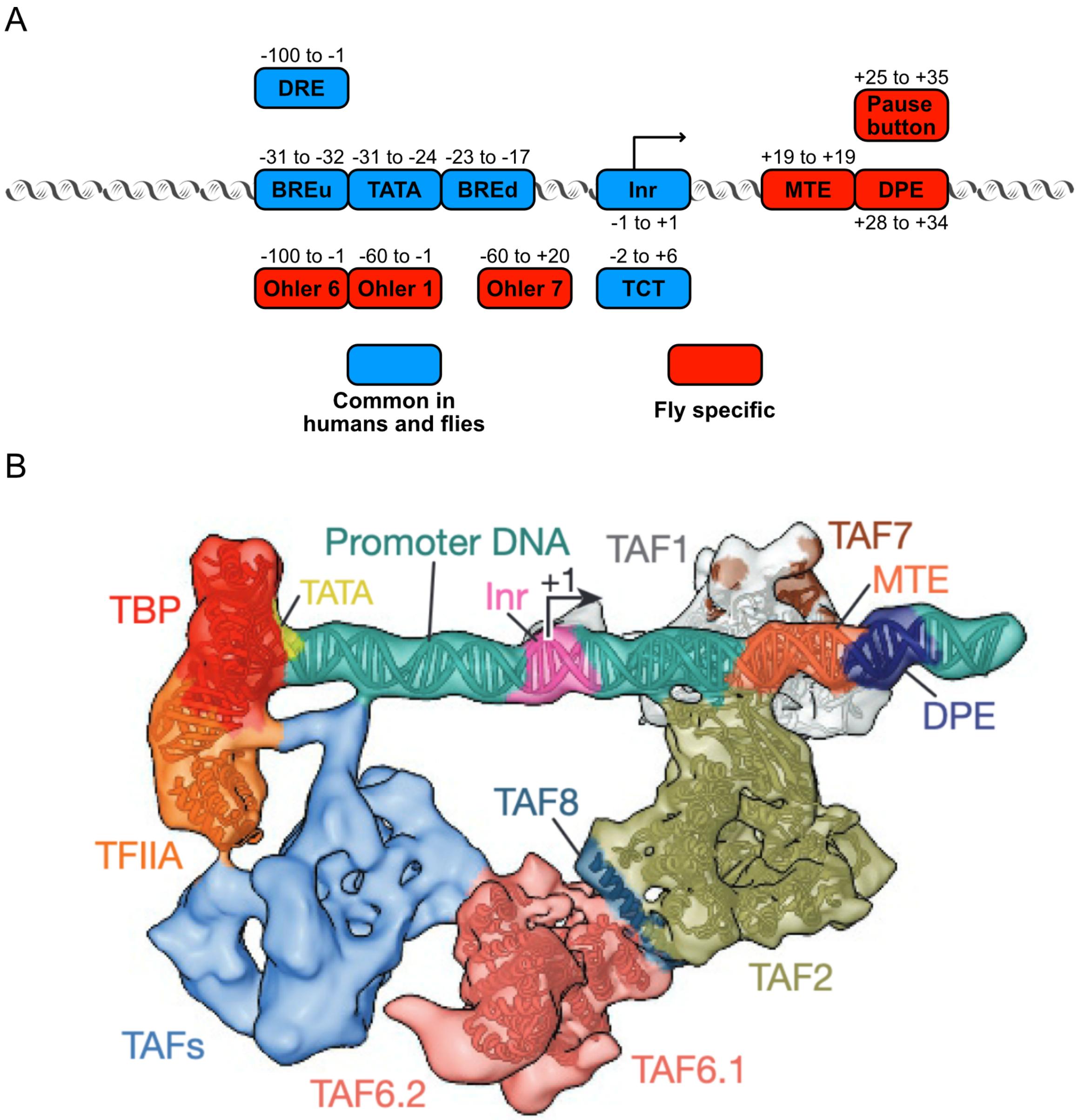


Figure 1.2: The core promoter contains core promoter motifs. A) location of core promoter motifs found in humans and fly. Core promoter motifs are found close to the transcription start site denoted as the +1 nucleotide, adapted from(Haberle & Lenhard, 2016). B) Structure of the pre-initiation complex (PIC) (without TFIIB and RNA polymerase II) bound to their core promoter motifs, adapted from (Louder et al., 2016). TBP: TATA-Binding Protein; TAF: TBP-Associated Factors; Inr: Initiator; MTE: Motif Ten Element; DPE: Downstream Promoter Element; DRE: Dehydration-Responsive Element; BRE: B Recognition Element

An important observation is that not all canonical motifs are present at every core promoter. For example, the TATA-box is only found at about 5% of core promoters in flies (Ohler et al., 2002) and 24% in humans (Yang, Bolotin, Jiang, Sladek, & Martinez, 2007). Some associations between motifs have been identified. Core promoters lacking TATA box in flies often contain a combination of Inr and DPE (Burke & Kadonaga, 1996) and the spacing between the two motifs is thought to play a role in the binding of the TFIID GTF (Louder et al., 2016). Other combinations, like the TATA and DPE, occur only rarely (Kutach & Kadonaga, 2000; Ohler et al., 2002). Beyond specific motifs, other sequence determinants characterize core promoters. One example is that core promoters overlap CG islands (CGIs) (Gardiner-Garden & Frommer, 1987). However, like certain core promoter motifs, CGIs are not present in all organisms. Certain combinations of these sequence determinants have been shown to associate with functionally distinct groups of genes, contributing to the delineation of different classes of promoters (Carninci et al., 2006; Lenhard, Sandelin, & Carninci, 2012; Saxonov, Berg, & Brutlag, 2006). For example, genes active in differentiated tissues are enriched for TATA-box and the Inr motifs (Figure 1.3).

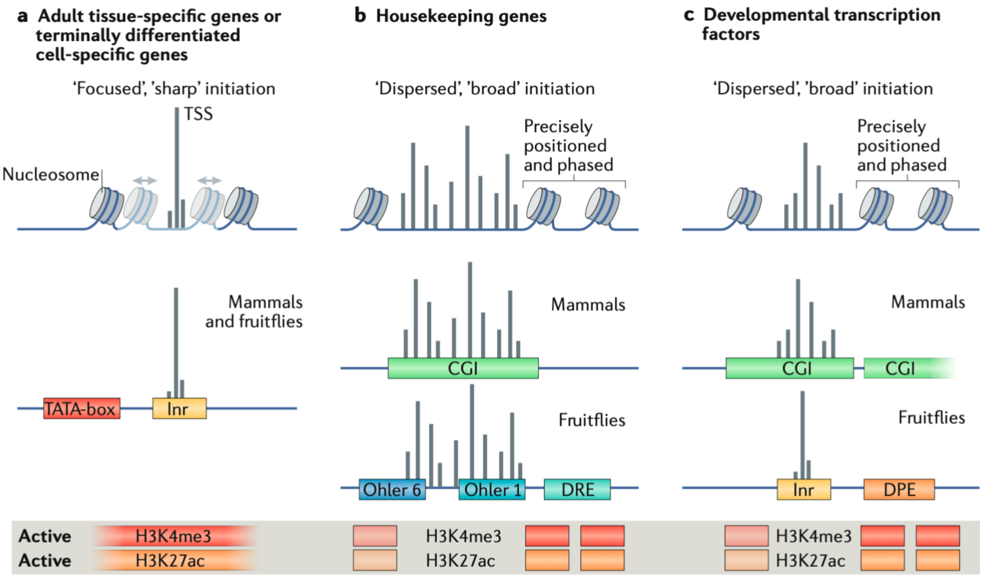

a Adult tissue-specific genes or terminally differentiated cell-specific genes
b Housekeeping genes
c Developmental transcription factors
'Focused', 'sharp' initiation
TSS
Nucleosome
'Dispersed', 'broad' initiation
Precisely positioned and phased
'Dispersed', 'broad' initiation
Precisely positioned and phased
Mammals and fruitflies
TATA-box
Inr
Mammals
CGI
Fruitflies
Ohler 6
Ohler 1
DRE
Mammals
CGI
CGI
Fruitflies
Inr
DPE
Active
Active
H3K4me3
H3K27ac
H3K4me3
H3K27ac
H3K4me3
H3K27ac

Figure 1.3: Architecture of different core promoters in mammals and flies, adapted from (Haberle & Stark, 2018) . A) Tissue specific genes tend to have a focused transcription initiation pattern with less well phased nucleosomes. They are also enriched in both the TATA-box and Inr motifs. B) Housekeeping genes possess a broader transcription initiation pattern with well positioned and phased downstream nucleosomes. In mammals these core promoters are enriched for CG islands while in flies, they are enriched for the Ohler1, Ohler 6 and DRE motifs. C) Important developmental genes possess similar promoter architectures as housekeeping genes in mammals unlike in flies. There they possess a more focused transcription initiation pattern and are enriched for the Inr and DPE motifs. All active core promoters are marked by the H3K27Ac and H3K4me4 marks.

Nucleosome architecture at active promoters For transcription to initiate, DNA must be accessible to the transcription machinery. Active promoters are therefore situated within a region of open chromatin called the nucleosome-depleted region (NDR) (C.-K. Lee, Shibata, Rao, Strahl, & Lieb, 2004) (Figure 1.4). How chromatin becomes accessible will be addressed in a later section of the introduction. Downstream of the NDR, nucleosomes (composed of DNA wrapped around a histone core) are organized in a phased manner, forming a nucleosomal array that extends into the gene body (Figure 1.4). Nucleosome positioning around active promoters has been extensively studied in organisms such as yeast (Yuan et al., 2005), flies (Mavrich et al., 2008) and humans (Barski et al., 2007) and this nucleosomal organization is indicative of active promoters in eukaryotes.

However, similar to core promoter motifs, nucleosome positioning can vary as well depending on the class of promoters (Lenhard et al., 2012). Notably, tissue-specific genes tend to have a less well-positioned nucleosomes flanking the NDR while housekeeping genes have been characterized by a larger NDR with well-phased downstream nucleosomes (Figure 1.3). How nucleosomal positioning is achieved is still rather unclear and beyond the scope of this introduction. However, some sequence determinants such as di-nucleotide periodicity patterns and homopolymeric sequences can affect nucleosome positioning and interaction stability with DNA (Lai & Pugh, 2017).

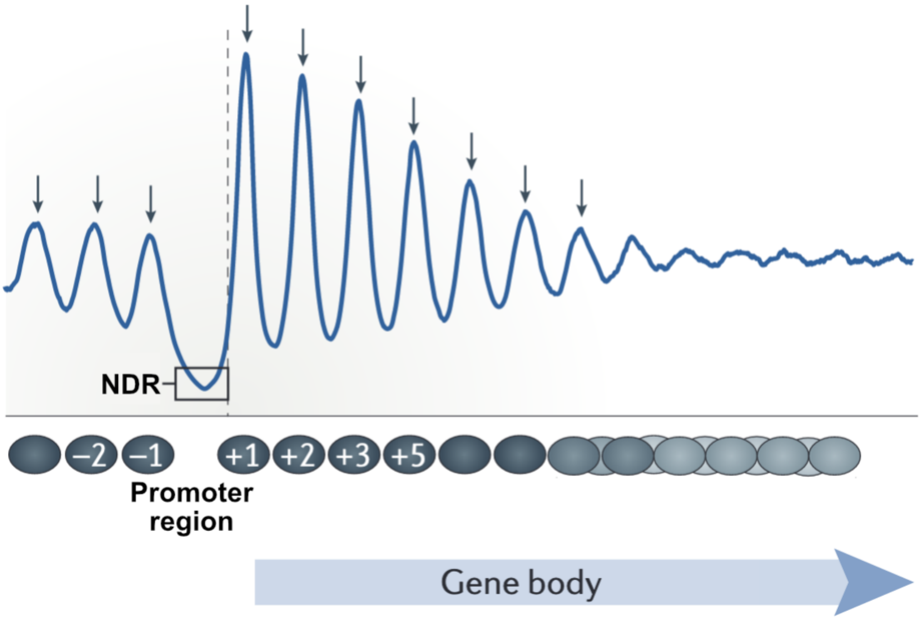


Figure 1.4: Nucleosome organization around active promoters, adapted from (Lai & Pugh, 2017). Promoters are situated in a nucleosome depleted region (NDR). The peaks and valleys represent sited of high and low nucleosome occupancy respectively. Nucleosomes are well phased around the promoter and tend to become less phased or 'fuzzy' further away. The +1 shows the highest phasing.

Histone modifications Another well studied feature of promoters is the presence of post-translational modifications (PTMs) on histones around these regulatory elements (Haberle & Stark, 2018). This section only addresses histone modifications that serve as markers characterizing active promoters. The role of histone PTMs and their link to the transcription regulation will be detailed in a later section (1.1.3).

Two marks, lysine 27 residue acetylation and lysine 4 methylation of histone 3 (H3K27Ac and H3K4me3 respectively) are highly conserved at eukaryotic promoters (Ho et al., 2014; Pokholok et al., 2005) (Figure 1.3). H3K27Ac is primarily found on nucleosomes around the promoter, whereas the H3K4me3 signal peaks at promoters but also extends into actively transcribed gene bodies. These conserved patterns played a crucial role the identification and study of regulatory elements genome-wide across a diverse range of organisms (Bourdareau et al., 2021; Duncan, Chitsazan, Seidel, & Alvarado, 2015; Duttke, Chang, Heinz, & Benner, 2019; ENCODE Project Consortium, 2012; Gerstein et al., 2010; modENCODE Consortium et al., 2010).

Transcription initiation at promoters

Transcription initiates at promoters, extending downstream to generate a mRNA copy of the gene that needs to be expressed. The position of the first transcribed nucleotide is called the transcription start site (TSS). Researches developed genome-wide techniques such as cap analysis of gene expression (CAGE) to study TSS characteristics (Shiraki et al., 2003). These methods rely on the properties of the 5' ends of RNA polymerase II transcripts that possess a modified nucleotide called the 5'cap playing a role in transcript stability (Rottman, Shatkin, & Perry, 1974).

Analyses of transcription initiation profiles using CAGE revealed that promoters do not initiate transcription at a specific position. Instead, multiple transcription initiation sites for a given promoter are organized in clusters called transcription initiation clusters (TICS) (Carninci et al., 2006). Four categories of TICs were mentioned in the original publication describing two main modes of TIC structures (Figure 1.5), namely, broad and sharp peaks. Sharp peaks are defined by their focused transcription initiation pattern around one specific site while broad peaks initiate transcription in different locations within the promoter region. In addition to the nucleosome positioning and presence of specific core promoter motifs, different initiation patterns are associated to specific classes of genes but can vary between organisms. For instance, tissue specific genes tend to have a sharp TIC while housekeeping genes have a broader transcription initiation pattern. However, this can vary from species to species. For example, TIC shape in key developmental genes differ in flies and humans (Figure 1.3).

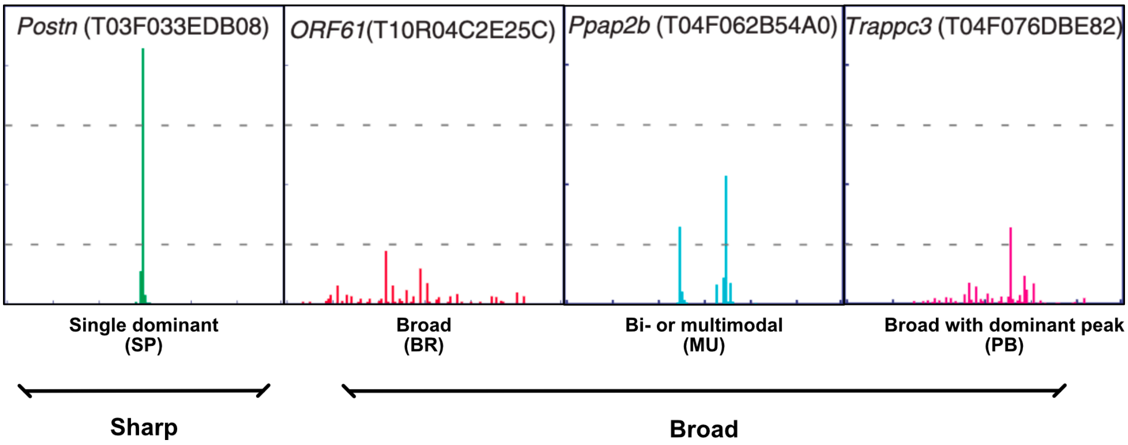


Figure 1.5: Different types of transcription initiation modes in mammalian promoters, adapted from (Carninci et al., 2006). Four classes of promoters exist within mammalian promoters that are divided in two main modes of transcription initiation. Namely promoters with broad or sharp transcription initiation patterns.

Another feature of promoters is that they initiate transcription in a bidirectional manner (Core, Waterfall, & Lis, 2008; Seila et al., 2008) (Figure 1.6). Antisense transcription is thought to arise from a PIC binding at the other edge of the NDR (Scruggs et al., 2015) but arrests prematurely and gives rise to short and unstable transcripts called upstream antisense RNAs (uaRNAs) or promoter-upstream transcripts (PROMPTs). These transcripts are subsequently degraded by the exosome (Preker et al., 2008). One factor contributing to the premature arrest of antisense transcription is the presence of enriched polyadenylation sites upstream of promoters (Ntini et al., 2013).

A major question arising from this observation is the extent to which bidirectional transcription characterizes promoters. An interesting study conducted in yeast suggests that promoter regions may inherently exhibit bidirectionality and have evolved to favor the transcription of coding transcripts while repressing non-coding antisense transcription (Y. Jin, Eser, Struhl, & Churchman, 2017). This conclusion was reached by introducing DNA from foreign yeast species into *S. cerevisae.* In this new context, previously unidirectional promoters lost their directionality and initiated transcription in both orientations. Furthermore, Jin and colleagues investigated the directionality in both conserved and newly evolved promoters in both yeast and humans. In both cases, they concluded that evolutionarily older promoters displayed a stronger bias towards unidirectional transcription, while newly evolved promoters exhibited greater bidirectionality. They finally proposed that, over time, organisms can bias promoter transcription towards their coding transcripts.

Additional insights into the directionality of regulatory elements come from studies in humans and mice (Duttke et al., 2015). There, they argue that promoters are intrinsically unidirectional and that divergent transcription arises from two reverse oriented core promoters. According to this study, bidirectionality would therefore not be an inherent feature of promoters. These results have been a source of debate. Re-analysis of the data published in Duttke and colleague in addition to other experimental data showed that the majority of unidirectional promoters displayed signal in the antisense orientation (Andersson et al., 2015), something that was not detected using only the data used in the original publication. Andersson and colleagues argue that bidirectional transcription is an inherent feature of regulatory elements since the overwhelming majority of promoters exhibit bidirectional transcription initiation. They conclude by stating that sensitivity of the methods used can explain the lack of observable antisense transcription at the remaining fraction of promoters. This aligns with the findings of the study conducted by Jin and colleagues and gives credibility to the inherent nature of bidirectional transcription initiation at promoters.

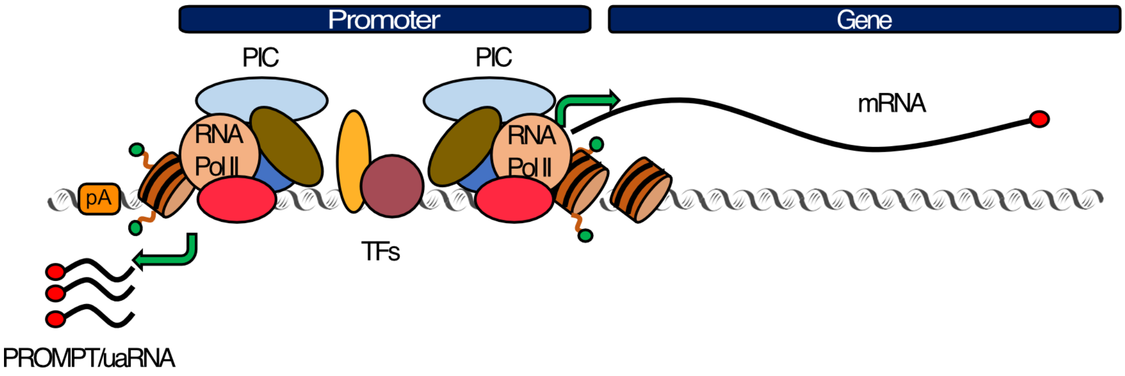


Figure 1.6: Promoters transcribe in a bidirectional fashion. Transcription initiation occurs at both sides of the NDR within promoters by the assembly of two different PICs. RNA polymerase II however only enters into productive elongation downstream of the promoter in the direction of the gene body. Antisense transcription is quickly halted resulting in the production of upstream antisense RNAs (uaRNAs) also called PROMoter uPstream Transcripts (PROMPTs). Early polyadenylation sites upstream of the promoters are thought to play a role in this early transcription termination

Enhancers integrate regulatory signals to modulate gene expression

Enhancers, the other main class of DNA elements that regulate transcription, operate in an orientation-independent manner and can influence the transcription of genes located at distant genomic positions. This class of REs was initially discovered in the SV40 virus genome where deletion of a 72 bp tandem repeat led to reduced expression of a gene necessary for viral replication (Benoist & Chambon, 1981; Gruss, Dhar, & Khoury, 1981). Interestingly, this same repeat was able to increase the transcription of the beta-globin gene in mammalian cells by a factor of 200 even when situated far away (Banerji, Rusconi, & Schaffner, 1981).

Enhancers are now recognized to be an integral part of transcription regulation and are found throughout the tree of life (Banerji, Olson, & Schaffner, 1983; Belitsky & Sonenshein, 1999; Petrascheck et al., 2005; Timko et al., 1985). They are viewed to be responsible for fine tuning gene regulation by integrating regulatory signals through the binding of transcription factors. A great deal of research implicates enhancers as a major contributor for tissue-specific gene expression (Heinz, Romanoski, Benner, & Glass, 2015). Furthermore, enhancers do not only function in differentiated tissues but are also during development. For example, the ZRS enhancer is critical for proper limb development in vertebrates and mutation of a transcription factor binding motif leads to sever limb reduction in mice (Kvon et al., 2016).

Enhancer architecture

Similar to promoters, active enhancers are located within a nucleosome-depleted region (He et al., 2010) and contain transcription factor binding sites (TFBSs) (Heinz et al., 2010) (Figure 1.7). Around the NFR are nucleosomes that are post-translationally modified. The major histone marks found at active enhancers are H3K27Ac (Creyghton et al., 2010) and H3K4me1 (Heintzman et al., 2007). The lysine 4 methylation pattern contrasts with what is found at promoter sites and was suggested to be a characteristic that architecturally differentiates enhancer from promoters. However, later reports have shown nucleosomes around highly active enhancers can also bear the H3K4me3 mark (Henriques et al., 2018; Pekowska et al., 2011).

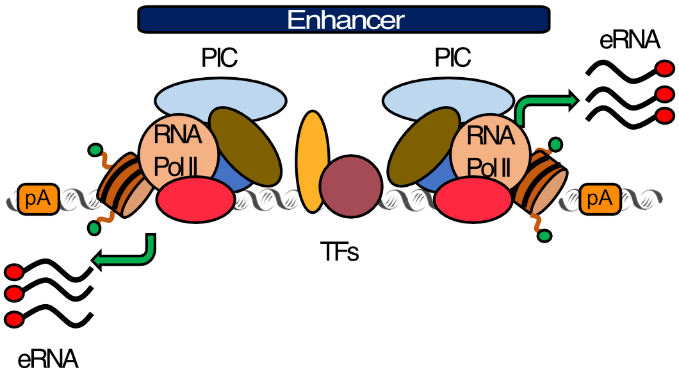


Figure 1.7: Schematic of a classical enhancer. Active enhancers are also situated within a NDR. Nucleosomes flanking the NDR are enriched with histones containing active post-translational modifications. Enhancers contain transcription factor binding sites and are also capable of recruiting the pre-initiation complex to initiate transcription. Transcription occurs in a bidirectional fashion at enhancers but doesn't lead to productive elongation. Early transcription termination leads to the production of short and unstable capped RNAs called enhancer RNAs (eRNAs). In some species, DNA in the vicinity of enhancers is enriched for polyadenylation sites which play a role in early transcription termination.

Enhancers regulate gene expression form distal genomic coordinates

In contrast to promoters, these regulatory elements can be situated further away from the genes they regulate. Enhancers are found at a wide range of distances from their target genes, typically within non-coding regions of the genome (Furlong & Levine, 2018). It is now known that distal enhancers do not exert their regulatory functions from afar but are brought in close proximity to promoters in the 3D nuclear space.

The formation of DNA loops by the cohesin complex has been shown to play a critical role in bringing REs together that are otherwise located far away from each other on the linear genome (Figure 1.8). The process of loop formation starts with the loading of cohesin onto DNA, followed by a process called loop extrusion, which pulls DNA from both sides through the cohesin complex, eventually bringing distal REs in close proximity (Karpinska & Oudelaar, 2023). Another important factor in enhancer-promoter interactions is 'molecular affinities' between factors binding REs. This mechanism is thought mainly to stabilize and maintain RE interactions rather than initiate them. The mediator complex plays a pivotal role in this mechanism by interacting both with RNA pol II and transcription factors (Conaway & Conaway, 2011) and therefore effectively acting as a bridge between REs. A recent study confirmed this by showing that rapid depletion of mediator leads to reduced enhancer-promoter interaction (Ramasamy et al., 2023).

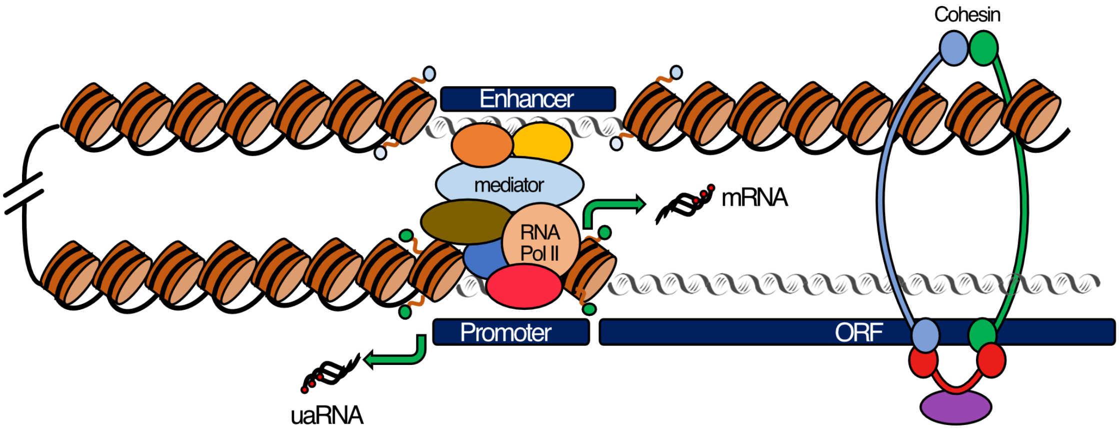


Figure 1.8: Enhancers regulate gene expression by coming in close proximity to their gene promoter. Multiple factors allow enhancers to regulate the activity of gene promoters. Transcription factors bound at enhancers can regulator the PIC activity through the mediator complex. Moreover, enhancers are brought in close proximity to promoters through cohesin-mediated loop formation.

Transcription initiates at enhancers but does not lead to productive elongation

Another common feature shared between promoter and enhancers is the presence of core promoter motifs like the Inr and TATA-box at both types of elements (Andersson et al., 2014). Although these motifs are described to be more degenerate at enhancers, recruitment of GTFs at these sites does happen (Koch et al., 2011). Moreover, numerous reports demonstrate the recruitment of RNA pol II at enhancers where transcription initiation occurs and can be seen as a defining feature of active enhancers (Andersson et al., 2014; De Santa et al., 2010; Kim et al., 2010). Similar to uaRNAs, enhancer RNAs are usually short and unstable. Some organisms also present polyadenylation sites around enhancer NDRs, which are believed to play a role in early transcription termination (Andersson & Sandelin, 2020). Additionally, enhancers are also transcribed in a bidirectional fashion and bidirectional transcription was shown to be a good predictor for enhancer identification

(Andersson et al., 2014).

Revisiting the promoter/enhancer dichotomy

It is now evident that promoters and enhancers share similar features in terms of architecture and transcription initiation potential (Figure 1.9). This prompts a crucial question: how can we best differentiate between these two types of regulatory elements, or is there even a need to make such a distinction? High-throughput reporter assays revealed that many promoters can act as enhancers *in vitro* and vice versa (Nguyen et al., 2016). These findings have also been verified *in vivo*, where some promoters act as enhancers for other promoters (Z. Xu, Wei, Chepelev, Zhao, & Felsenfeld, 2011) and enhancers are used as alternative promoters (Kowalczyk et al., 2012). All of the evidence presented above prompted researchers to revisit the longstanding promoter/enhancer dichotomy. The updated model characterizes regulatory elements possessing properties that are either enhancer-like or promoter-like, with the two categories not being mutually exclusive in (Andersson & Sandelin, 2020).

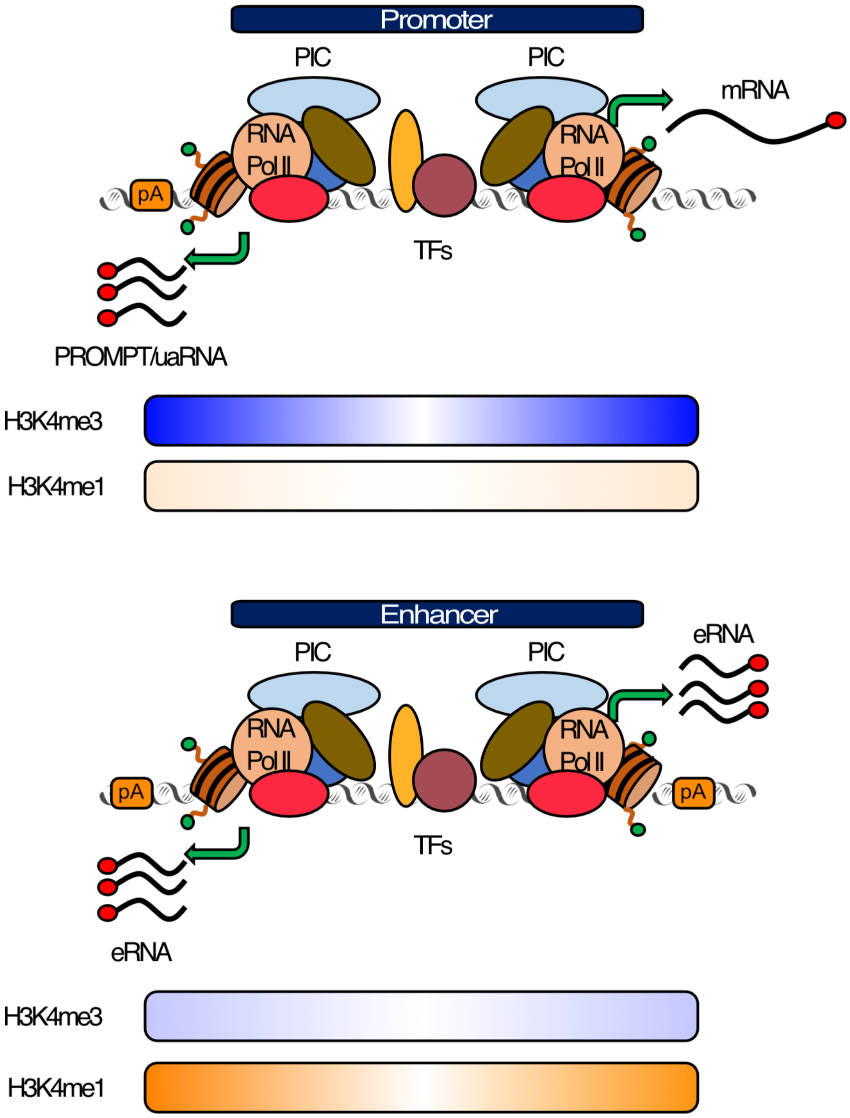


Figure 1.9: Enhancer/Promoter dichotomy. Enhancers and promoters share a lot of common features. They both bind transcription factors, possess active chromatin marks, are capable of PIC recruitment and initiate transcription bidirectionally, adapted from (Andersson & Sandelin, 2020).

1.1.2 Transcription factors

Transcription factors are defined as proteins able to bind DNA in a sequence-specific manner and regulate transcription through repression or activation of transcription (Lambert et al., 2018). They typically bind within regulatory elements through conserved domains called DNA binding domains (DBD). This conservation has been used to classify TFs into families on the basis of their characteristics (Figure 1.10 A) (Wingender, Schoeps, Haubrock, Krull, & Dönitz, 2018). Each DBD binds to certain regions in REs called transcription factor binding sites (TFBS) with a specific sequence composition referred to as a DNA binding motif (Figure 1.10 B).

Importantly, the DNA binding motif of a transcription factor is not a fixed sequence and can be expressed as a position weight matrix (PWM) which contains the probability of finding each nucleotide of a TFBS species at a specific position. These matrices are constructed from sequences that have been experimentally verified to bind to the respective transcription factor or are believed to do so. Different sequences binding a specific transcription factor with a similarly high affinity are called sequence optima. Recent structural studies have shed light on two binding mechanisms that enable transcription factors to interact with various sequence optima. A DBD can make direct contact with the DNA bases or interact indirectly by intermediate of water bridges (Ekaterina Morgunova et al., 2023; Morgunova et al., 2018).

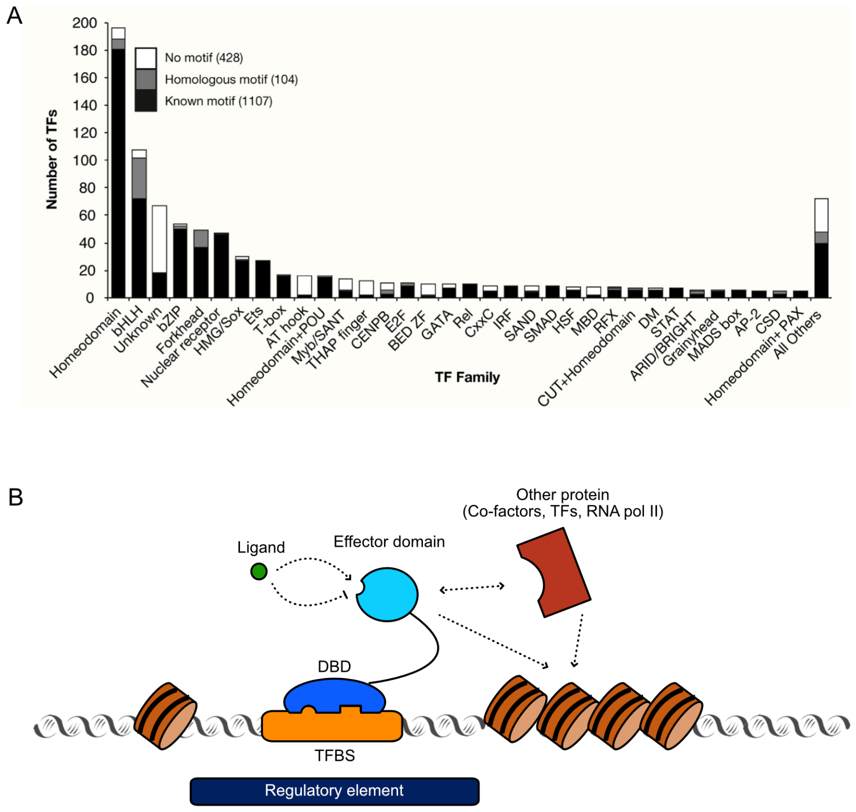


Figure 1.10: Transcription factors, adapted from (Lambert et al., 2018). A) Transcription factors are divided into families on the basis of their DNA binding domain (DBD). B) Transcription factors are composed of different domains. The DBD will recognize a specific transcription factor binding site (TFBS). The effector domain of a TF can regulate gene expression in various ways. It can interact with ligands to modulate its activity. It can have a direct effect on nucleosome organization or interact with other proteins to exert its function.

Transcription factors usually do not work alone but exhibit cooperative behaviours in (Morgunova & Taipale, 2017). TFs can form homo- or heteromers that increase their affinity for DNA. Conversely, DNA binding can facilitate interactions between two transcription factors by inducing conformational changes that promote their interaction. Additionally, DNA can also be the sole factor that mediates TF cooperativity. One TF can induce changes in the DNA that promote the binding of the other without direct protein-protein interaction (Panne, 2008). Another indirect mechanism for TF cooperativity involves pioneering transcription factors. These proteins have the capability to uncover previously inaccessible transcription factor binding sites bound by nucleosomes, allowing their cognate transcription factors lacking this chromatin-opening ability to subsequently bind to them (Mayran & Drouin, 2018).

In addition to their DNA binding domains, typical TFs also contain another conserved domain known as an effector domain, which serves to integrate signals and modulate protein interactions (Figure 1.10 B). For example, the TATA binding protein (TBP) directly recruits RNA polymerase. Nuclear hormone receptors possess a ligand binding domain that will induce conformational changes activating the transcription factor (Rosenfeld, Lunyack, & Glass, 2006). Finally, TFs with pioneering abilities can also directly remodel chromatin and play an important role in cell fate decision during development (Gualdi et al., 1996; Mayran & Drouin, 2018).

The majority of TFs do not possess a pioneering activity or directly recruit RNA polymerase II. Instead, they rely on the co-factors with which they interact to carry out their functions. These co-factors are typically large protein complexes with diverse functions (Reiter, Wienerroither, & Stark, 2017). Many of them possess enzymatic functions and deposit PTMs on surrounding histones, TFs or even RNA pol II. One well-studied co-factor is the p300/CBP that acetylates histones upon recruitment by TFs, thereby destabilizing nucleosomes and allowing other TFs to bind to newly exposed TFBSs (Q. Jin et al., 2011). The mediator complex is another important co-factor. It not only acts as a bridging platform to mediate enhancer-promoter interactions, but also possesses a kinase module capable of regulating the activity of transcription factors and RNA polymerase II (Richter, Nayak, Iwasa, & Taatjes, 2022).

1.1.3 Chromatin environment

Histone modifications

Histones, the proteinaceous components of nucleosomes, are highly conserved proteins in eukaryotes and play an important role in the initial compaction of DNA (Talbert & Henikoff, 2010). They possess a characteristic histone tail at their N-terminus which protrudes out of the nucleosome core and serves as substrate of many enzymes. Since the first evidence that histone tails were post-translationally modified (Allfrey, Faulkner, & Mirsky, 1964), many more have been identified, especially with the development of mass-spectrometry-based methods (Lu, Coradin, Porter, & Garcia, 2021) (Figure 1.11).

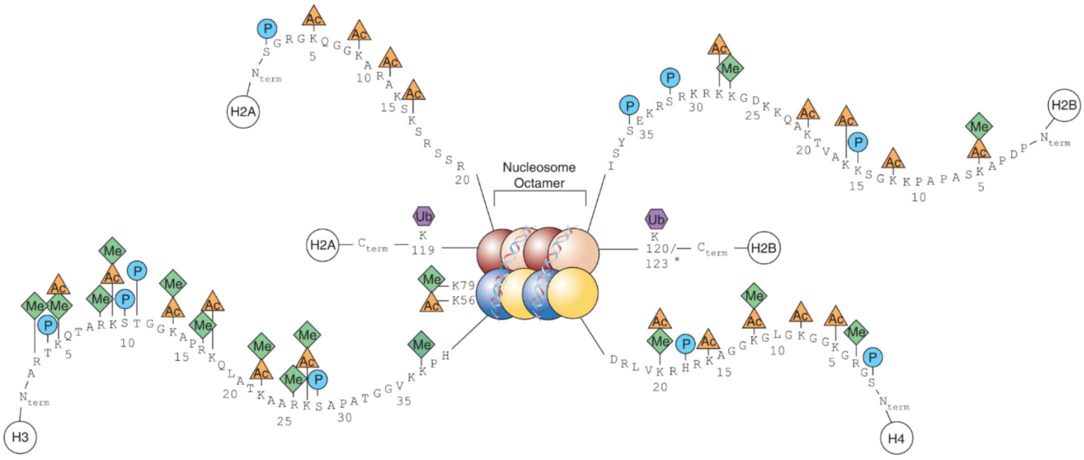


Figure 1.11: Histone modifications, adapted from (Keppler & Archer, 2008). Histones possess long N-terminal tails that serve as a substrate for many enzymes to deposit post-translational modifications (PTMs) such as methyl, acetyl or phosphoryl groups. While the N-terminal histone tails are the main substrate of these enzymes, other parts of the nucleosome can also be modified. Histone PTMs are important actors for gene expression regulation as they can serve as signals for other proteins or affect nucleosome stability.

Histone PTMs, despite being covalently bound, are dynamic in nature. For instance, methylation and acetylation marks on lysines of histone tail are deposited and removed by enzymes classes respectively called writers and erasers. When these modifications are present, they are interpreted by effector proteins referred to as readers (Hyun, Jeon, Park, & Kim, 2017; Musselman, Lalonde, Côté, & Kutateladze, 2012). Many of these modifications have been linked to regulation of gene expression through the development of chromatin immuno-precipitation (ChIP) methods. Some modifications act through these readers (Zippo et al., 2009) while others can be directly bound by PIC components to activate transcription (Vermeulen et al., 2007).

As previously discussed, histone PTMs can also promote transcription by acting on nucleosome stability. For example, the p300-deposited H3K64Ac mark facilitates nucleosome eviction by destabilizing histone-DNA interactions (Di Cerbo et al., 2014). Moreover, the H3K14ac mark is directly recognized by a nucleosome remodeler to render the chromatin accessible (Dann et al., 2017). Histone PTMs are not only associated with transcriptional activation; they can also serve repressive functions. For example, the H3K9me mark is associated with transcriptional repression as it recruits HP1, an important component of heterochromatin (Bannister et al., 2001).

Histone modifications are observed in both discrete patterns around REs and broader patterns delineating chromatin domains. For example, the H3k27Ac usually marks active REs, while H3K27me3 and H3K36me3 delineate larger regions linked to repressed or active chromatin states. These patterns often overlap with regions of heterochromatin and euchromatin (Evans et al., 2016; Valouev et al., 2011; Van Bortle et al., 2012) (Figure 1.12).

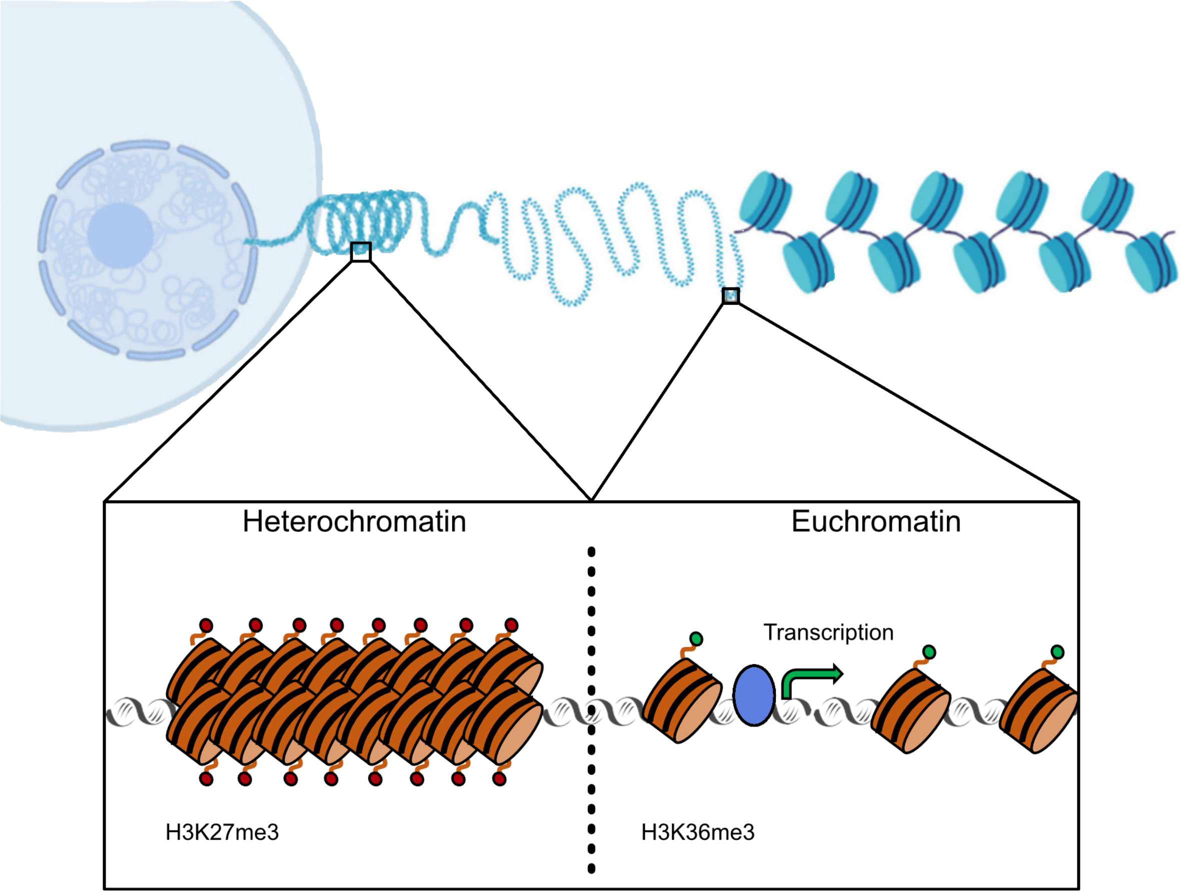


Figure 1.12: Chromatin domains, adapted from (J. Xu & Liu, 2021). DNA is organized in the nucleus such as not all regions of the DNA are evenly accessible. Large region of inaccessible chromatin with high nucleosome compaction is called heterochromatin and is enriched for the H3K27me3 histone PTM. The less compacted regions of the genome are called euchromatin and is permissive for transcription. One histone PTM mostly found in regions of euchromatin is the H3K36me3 mark.

The discovery of various histone PTMs led researchers to hypothesize that different combinations of PTMs could represent a fundamental mechanism of gene regulation, referred to as the histone code (Jenuwein & Allis, 2001). Although many studies have established a link between histone modifications and gene expression, whether a causal relationship between the two exists remains debated. For instance, reports show that the H3K4me3 depositing enzymes are recruited at active genes in a transcription-dependent manner in yeast (Krogan et al., 2003) and mammals (Milne et al., 2005). It was therefore proposed that H3K4me3 serves more as an epigenetic bookmark of recent transcriptional activity to facilitate further transcription at these loci (Ng, Robert, Young, & Struhl, 2003). Similarly, a study showed that mutating canonical H3 at Lys 27 did not result in widespread transcription down-regulation, suggesting that the PTMs found on this residue are not essential for transcription (Pengelly, Copur, Jackle, Herzig, & Müller, 2013). Finally, H3K4me1 was also shown to be dispensable for enhancer activity (Rickels et al., 2017).

Chromatin accessibility

It is now evident that chromatin accessibility is a fundamental facet of gene expression regulation. Active genes and their associated regulatory elements are situated within regions of euchromatin while sequences that need to be repressed like satellite repeats and transposable elements are located in inaccessible heterochromatin (Allshire & Madhani, 2018; Klemm, Shipony, & Greenleaf, 2019) (Figure 1.11). Chromatin compaction is not only needed to repress transcription of undesired sequences but also serves a physical role to fit the entirety of the genome within the limited nuclear space. Furthermore, it is also necessary to accurately distribute genetic information to daughter cells during mitosis or meiosis by organizing DNA in compacted, discrete metaphasic chromosomes. All of these processes require the chromatin compaction levels to be highly dynamic.

Nucleosome occupancy and turnover at specific locations are therefore a valuable metric to assess their potential to bind factors promoting gene expression (Klemm et al., 2019). Active regions, such as strong promoters and enhancers, typically exhibit lower nucleosome occupancy and higher turnover compared to transcriptionally repressed heterochromatin (Figure 1.13). Multiple mechanisms affecting nucleosome dynamics like histone PTMs, nucleosome remodelers and pioneering transcription factors have already been detailed above. Long non-coding RNAs also play a role in influencing gene expression by modulating chromatin accessibility. A notable example is involved in the dosage compensation mechanism during X-chromosome inactivation. During this process, the X-inactive specific transcript (Xist) spreads over one of the two X chromosomes, recruiting a variety of chromatin-modifying enzymes, such as HDACs and PRC2. These enzymes facilitate the establishment of a repressive chromatin state, contributing to the formation of the condensed heterochromatic structure known as the Barr body (Statello, Guo, Chen, & Huarte, 2021).

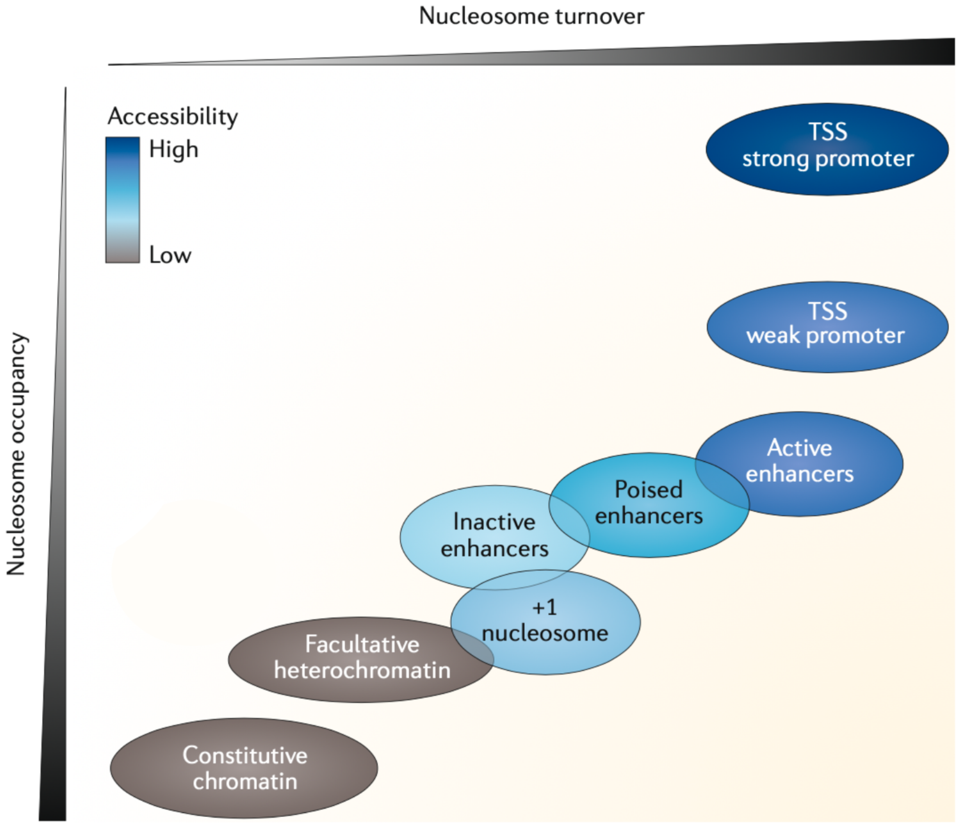


Figure 1.13: Nucleosome occupancy and turnover in different regions of the genome, adapted from (Klemm et al., 2019). Active promoters and enhancers are characterized by a high nucleosome turnover and low occupancy to keep the regions open and allow TFs and PIC components to bind. On the other hand, inactive regions like heterochromatin are constantly bound by nucleosomes and have a low nucleosome turnover. Note that the only element found in active regions that is characterized with a higher nucleosome occupancy and low turnover is the +1 nucleosome.

Nowadays, nucleosome occupancy and chromatin accessibility are studied using a variety of methods relying on the (in)accessibility of DNA to enzymes capable of introducing double strand breaks. These methods rely on the same underlying principle but give each their own characteristic patterns. DNAse-seq is able to cut in regions devoid of nucleosomes called DNAse hypersensitivity sites (DHS) and overlaps with REs. MNase-seq has both an endo- and an exonuclease activity and is able to digest all inter-nucleosomal DNA. This method is therefore used to study nucleosome positioning. Lastly, ATAC-seq relies on a transposase (Tn5) to directly insert sequencing adapters in accessible regions (Buenrostro, Giresi, Zaba, Chang, & Greenleaf, 2013). This method captures information similar to DNAse-seq but can also inform on nucleosome positioning around accessible sites. ATAC-seq has gained particular popularity due to its robustness and minimal input requirements. Consequently, it has enabled researchers to explore accessible regions of the genome in non-traditional model organisms (Magri et al., 2020; Pascual-Carreras et al., 2023).

Although chromatin accessibility is an important factor in transcriptional regulation and correlates with gene activity, a causal relationship between accessibility and active transcription is not always clear. Some studies propose that open chromatin is established first by pioneering factors binding to closed chromatin, thereby recruiting chromatin remodeling enzymes. This process would then enable RNA polymerase II to bind to the newly accessible region and initiate transcription (Fuda et al., 2015). Conversely, transcription has also been shown to regulate chromatin accessibility. For instance, a study shows that eRNAs, and therefore transcription, play a role in maintaining open chromatin for transcription factor binding (Mousavi et al., 2013). Furthermore, open chromatin can also lead to binding of repressors or co-repressors to TFs, which would lead to transcriptional silencing.

1.1.4 The 3D regulatory genome

As previously discussed, enhancers are able to regulate the transcription of their target genes by being brought into close proximity to the gene promoter. This process is also regulated to prevent unwanted E-P interactions, which could result in the mis-regulation of gene expression. To achieve this, chromatin is organized within the nucleus through various mechanisms (Bonev & Cavalli, 2016).

First, interphase chromosomes rarely intermingle and occupy distinct regions of the nucleus called chromosome territories. Within a chromosome territory, chromatin can be divided into two distinct types of compartments named A and B. These compartments rarely interact with each other and are roughly associated with active and inactive regions, respectively (Figure 1.14 A). A study showed that the B compartment is mostly found associated with the nuclear lamina or the nucleolus, a space where predominantly inactive genes have been reported (Xing, Johnson, Moen Jr, McNeil, & Lawrence, 1995) while the A compartment has a more central location (Stevens et al., 2017). Moreover, A and B compartments are also enriched for active (H3K36me3) and repressive (H3K27me3) chromatin marks respectively (Lieberman-Aiden et al., 2009).

Within compartment, other lower-level chromatin structures can be found (Rowley & Corces, 2018). Topologically associated domains (TADs) are one of such structures, defined as regions composed of sequences that preferentially interact together than with other regions of the genome (Figure 1.14 B). The boundaries of TADs are often enriched with regulatory elements known as insulators, which bind to proteins like CTCF. When bound, CTCF limit the process of cohesin-mediated loop extrusion by steric hindrance generating a region of the genome that favors interacting within itself (Figure 1.14 C). Moreover, perturbation of TAD boundaries by mutation of CTCF binding sites can lead to ectopic gene activation by enhancers not found within the initial TAD (Lupiáñez et al., 2015).

In *D. melanogaster*, a novel class of organizational REs have been described, called tethering elements (Batut et al., 2022). Similar to enhancers, these elements are characterized by H3K4me1-modified histones that can bind pioneering transcription factors, but do not display enhancer activity. Batut and colleagues suggest that these REs promote E-P contacts and are particularly useful when rapid gene activation is required.

The final concept addressed in this section relates to the observation that active RNA pol II is not homogeneously dispersed in the nucleus but occurs in foci, initially termed transcription factories. It is hypothesized that these clusters are composed of a high concentration of factors required to initiate transcription that exhibits phase-separated properties through via the presence of proteins with intrinsically disordered regions (IDR) and long non-coding RNAs as scaffold (Rippe & Papantonis, 2021; Statello et al., 2021). However, it's important to note that this model is still a subject of debate and requires *in vivo* validation regarding the role of IDR-mediated phase separation in the regulation of transcription. A more conventional perspective on the formation of transcription factories suggests that they result from classical protein-protein interactions, which lead to the clustering of factors necessary for transcription.

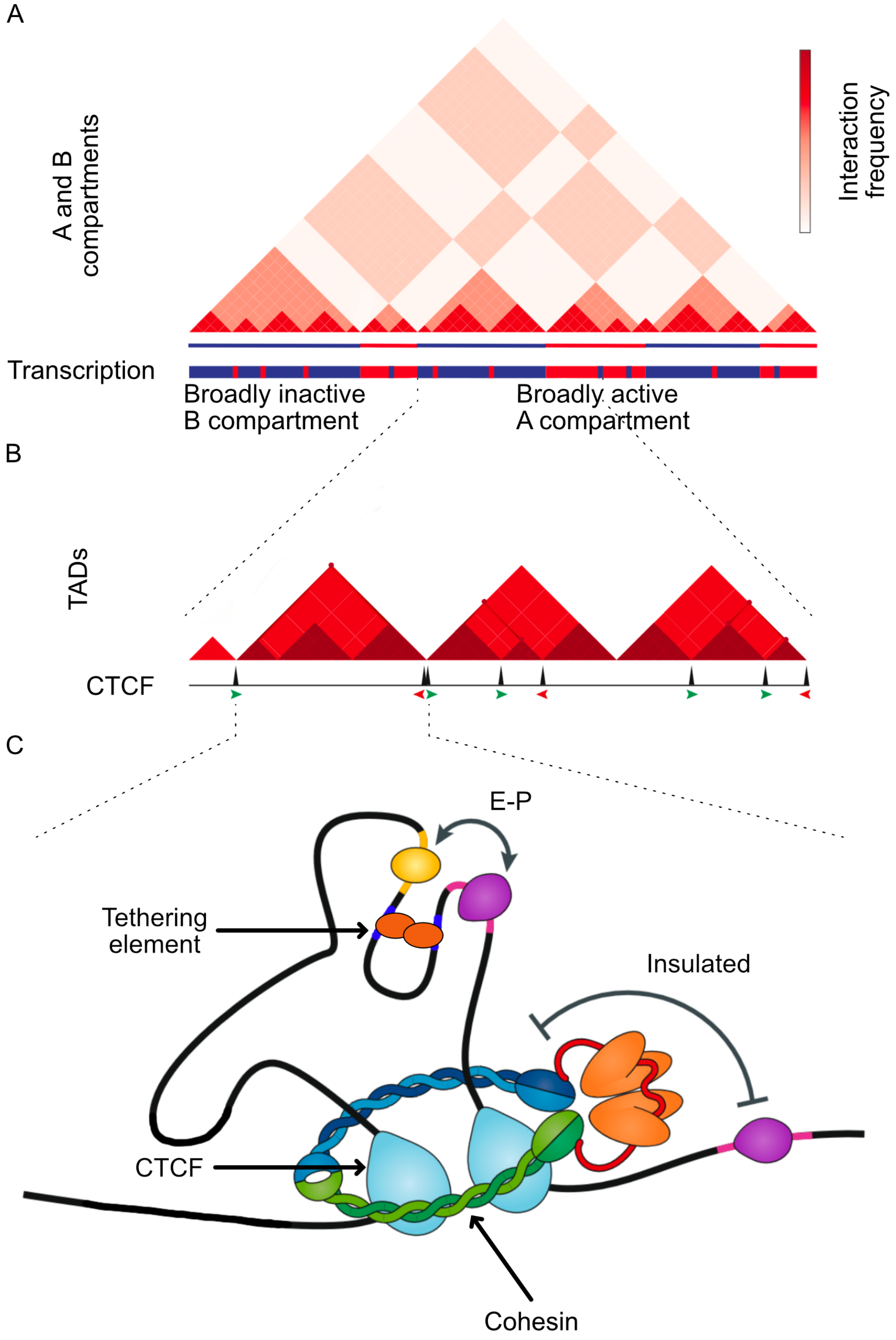


Figure 1.14: The 3D genome, adapted from (Rowley & Corces, 2018). A) The mammalian genome can be viewed as organized in a hierarchical fashion. Large-scale chromatin organization divide DNA from one chromosome in two different compartments (A and B). A compartments are enriched for mostly active genes while B compartments for inactive genes. B) Compartments possess lower levels of organizations called topologically associated domains (TADs) where regions within a same TAD preferentially interact with each other. Borders of TADs are enriched for CTCF binding sites that function as insulators. C) Through the process of cohesin mediated loop extrusion, elements within a TAD will be brought in close proximity favoring their interactions. Tethering elements also play a role in enhancer promoter interaction.

Beyond the regulation of transcription initiation

All the previously discussed mechanisms of gene expression regulation primarily revolve around the recruitment of RNA pol II to a promoter and initiating transcription of the target gene. While this step is undeniably crucial, additional phases during and after transcription are also subject to regulation. After PIC assembly and an initial phase of transcription (±30 to 50 nt), RNA pol II undergoes promoter proximal pausing (Zeitlinger et al., 2007). This pausing is, in part, regulated by two factors (NELF and DSIF). Phosphorylation of these factors by CDK9, a subunit of P-TEFb, is necessary to release RNA pol II from this paused state, enabling it to progress into productive elongation (Yamaguchi, Shibata, & Handa, 2013).

mRNA stability is also a facet of gene expression that can be subject to regulation. The addition of a methylguanosine cap to the 5'end of mRNAs shortly after transcription (± 30 nt) plays an essential role in mRNA stability, splicing, export and translation (Ghosh & Lima, 2010). During transcription termination, a poly-A tail is added at the 3'end of the mRNA which also regulates it stability, localization and translation (Ozsolak et al., 2010). Cells also employ multiple RNA surveillance pathways to destroy defective mRNAs that would potentially to the production of deleterious proteins (X. Wu & Brewer, 2012). This mRNA degradation is known to take place in cytoplasmic membrane-less organelles called P-bodies.

Furthermore, the translation of mRNAs can be repressed by sequestering them in another type of cytoplasmic organelle called stress granules (Buchan & Parker, 2009). Additionally, multiple classes of small RNAs such as microRNAs (miRNAs) and small interfering RNAs (siRNAs), the mediator of the RNA interference process, regulate mRNA stability or transcription by primarily binding to the transcript 3'UTRs. Moreover, siRNAs can also affect gene expression by targeting the endogenous loci of their mRNA targets for epigenetic silencing, a process referred to as RNA-induced transcriptional silencing (RITS) (Bhattacharjee, Roche, & Martienssen, 2019).

1.1.5 Model for gene transcription

A simple example of gene transcription can be formulated using all of the information described above. A pioneering transcription factor enters in competition with nucleosomes for DNA at a gene promoter situated within an A compartment, harboring H3K36me3 histone modifications, permissive to transcription. It displaces the nucleosomes creating an NDR and reveals core promoter motifs. These will in turn be bound by GTFs that initiate PIC formation through the recruitment of RNA pol II.

Through cohesin mediated loop extrusion and the action of tethering elements, a critical enhancer is brought in close proximity to the gene promoter and is bound by additional TFs. The mediator complex interacts with these TFs, comes into contact with the PIC stabilizing it and promotes transcription initiation. Concurrently, p300 is recruited and catalyzes the deposition of H3K27Ac at the target promoter and enhancer. The concentration of many transcription-related factors and the coalescence of other active loci in the vicinity leads to the formation of a transcription factory on basis of protein-protein contacts or through IDR-mediated phase separation.

At the locus of interest, activated PIC opens up the DNA duplex, RNA pol II separates from the GTFs and initiates transcription. After having transcribed 30-50 nt RNA poll II enters the promoter-proximal pausing state. Meanwhile, co-transcriptional capping of the nascent transcript occurs. Through interactions with co-factors recruited at the enhancer, P-TEFb is brought in close proximity with the paused polymerase and its CDK9 kinase subunit promotes RNA pol II pause release into productive elongation.

At the enhancer, another PIC has formed and initiates transcription leading to the synthesis of eRNAs supporting open chromatin maintenance at that locus. Active transcription at the gene locus promotes H3K4me3 deposition at the promoter which creates a more permissive environment for further transcription initiation cycles. During transcription elongation, the spliceosome removes intronic regions of the pre-mRNA molecule to gradually transform it into a mature mRNA. Finally, RNA polymerase enters the termination phase. A poly-A tail is added to the 3'end and splicing finishes to create a mature mRNA. The mRNA is then exported to the cytoplasm where it is finally translated to create the effector protein.

In summary, gene expression is a tightly regulated phenomenon involving a variety of processes happening at different scales throughout the cell. The complexity of the regulatory processes increases exponentially when multiple genes have to be coordinated to effectively implement complex responses, like an immune reaction or, even more dramatically, embryonic development. The following section gives a brief overview about regulatory complexity and their organization in hierarchical gene regulatory networks, specifically in the context of embryonic development.

1.2 Gene regulatory networks

1.2.1 Introduction to gene regulatory networks

In a multicellular organism, every cell contains the genetic information necessary for producing all the proteins found in that organism. However, each cell only produces a subset of these proteins. Selectivity in gene expression patterns is established through the action of complex and hierarchical gene regulatory processes organized into gene regulatory networks (GRNs). During the developmental process, GRNs become activated and gradually specify the identity and functions of all the cells that ultimately form the organism. A definition of "GRN" in the context of development is the following:

> "'Gene regulatory network' (GRN) is shorthand for the system of regulatory genes and their encoded interactions that determines the genetic functions to be expressed in cells of each spatial domain in the organism, at every stage of development."

Genomic Control Process 2015, Eric H. Davidson and Isabelle S. Peters

At the core of GRNs are regulatory genes, which are defined by Davidson and Peters as genes that encode transcription factors (TFs). TFs are the critical actors of GRNs as they control gene expression and therefore determine the 'genetic functions' of a cell at a certain moment in time. These functions include the expression of other regulatory genes, genes involved in signaling pathways and genes that are involved in the differentiation process or in functions of terminally differentiated cells. This category of genes is referred to as effector genes. Consequently, transcription factors serve a dual role in GRNs. They are on one side regulatory genes and on the other side effector genes of other regulatory genes. Knowing this, one of the most crucial functions of a GRN is to ensure the precise and context-specific expression of regulatory genes in distinct parts of the developing organism.

The interactions within a GRN are those that link the regulatory genes to their effector genes. They are mediated through regulatory elements (REs) and through regulatory genes themselves. REs of effector genes bind TFs and recruit the transcriptional machinery to these genes to mediate their expression. Effector genes that are components of signaling pathways can, in turn, activate transcription factors leading to the activation or repression of their target genes (Figure 1.15), with beta-catenin as the transcriptional effector of the Wnt-pathway as one example. It is through these links that GRNs encode the co-expression of certain genes to orchestrate a coordinated response. Importantly, these links are strictly unidirectional and confer a hierarchical organization to GRNs. The reconstruction of a GRN therefore requires to not only identify the genes belonging to it but also their regulatory interactions.

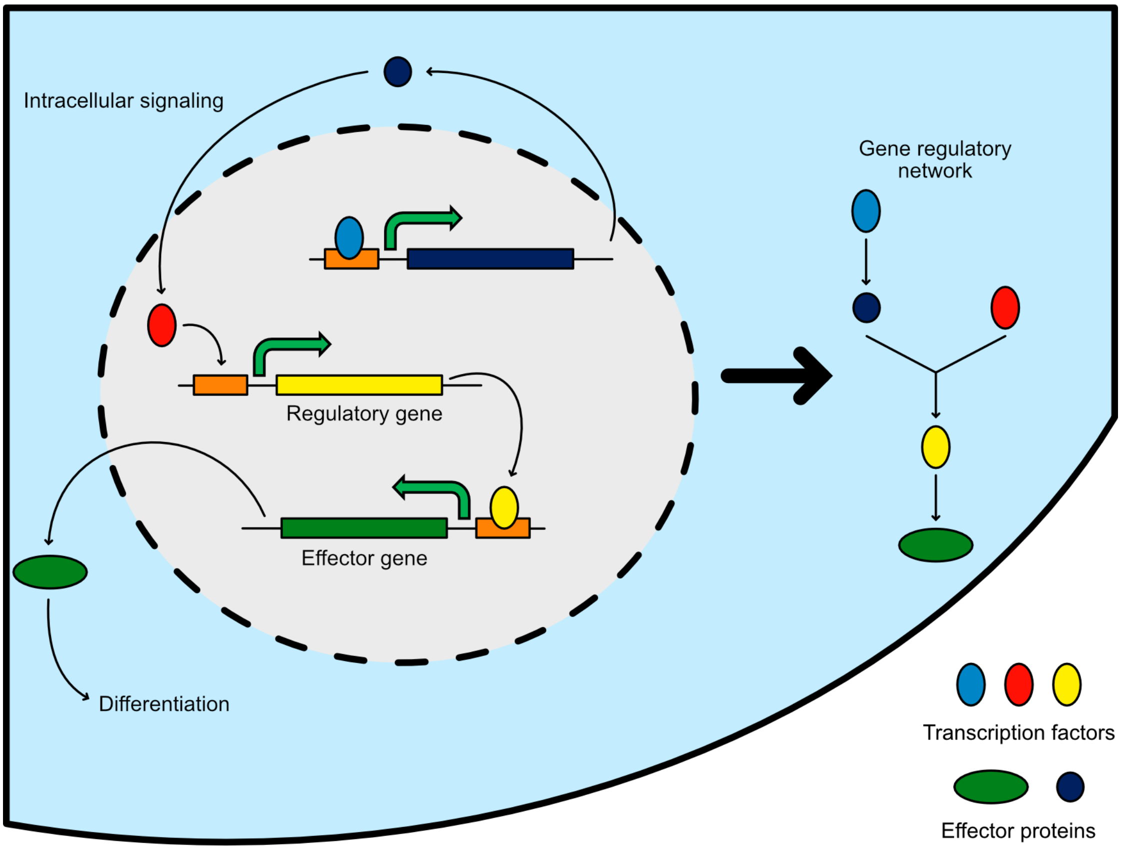


Figure 1.15: Schematic of a simple gene regulatory network (GRN). GRNs are composed of regulatory genes and effector genes. Regulatory genes are transcription factors able to activate the transcription of their effector genes. In turn, these effector genes will implement a certain function or process in the cell such as activating other regulatory genes or contribute to differentiation.

1.2.2 Studying gene regulatory networks

Regulatory genes within a GRN are not organized in a linear fashion but have both multiple inputs and outputs. A well-documented example from embryonic development in *D. melanogaster* is the even-skipped (*eve*) pair-rule gene regulation (Nüsslein-Volhard & Wieschaus, 1980). *eve* expression is under the control of transcriptional repressors (termed gap genes) in combination with transcription factors (termed maternal factors). These maternal factors also regulate the expression of the repressors ultimately defining specific domains of *eve* expression (Mannervik, 2014). Each expression domain is controlled by a specific enhancer that integrates the combination of various maternal factors and the gap gene present/absent at this location (Figure 1.16). In turn, *eve* regulates the expression of multiple targets to pattern the *Drosophila* embryo (Kobayashi, Goldstein, Fujioka, Paroush, & Jaynes, 2001).

Developmental GRNs are extremely complex. To identify their components and regulatory logic, researchers used a variety of techniques and computational methods over the years. This aimed to create GRN models that reflect the experimental data acquired. Additionally, models can also lead to hypotheses about so far unknown functional links between GRN components which can be experimentally tested.

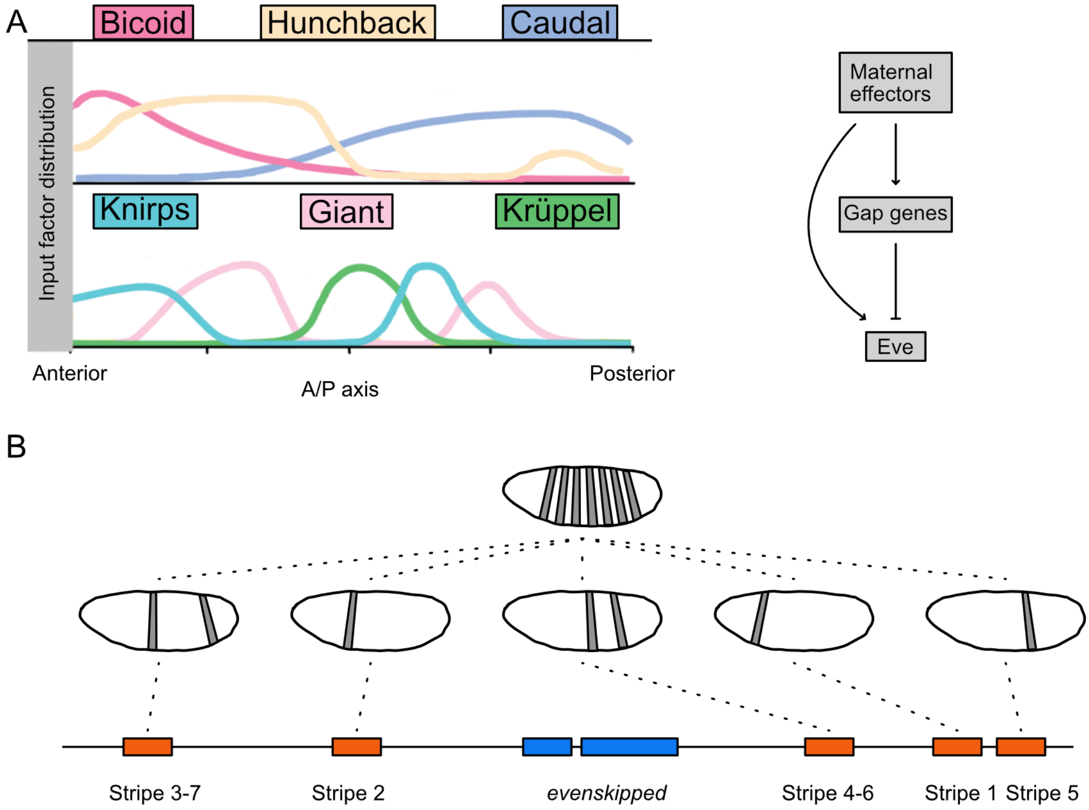


Figure 1.16: Regulation of *eve* expression during *Drosophila* embryogenesis, adapted from (Segal et al., 2008). A) Distribution of maternal effector and gap gene expression along the anteroposterior axis of the *Drosophila* embryo and simple depiction of the regulatory network orchestrating *eve* expression. B) *eve* (blue) transcription depends on a set of enhancers (orange) each responsible for the regulation of *eve* expression in separate striped along the anteroposterior axis.

Experimental methods to study gene regulatory networks

Early approaches relied on a combination of forward genetics and naturally occurring mutants to identify regulators of developmental GRNs (Nüsslein-Volhard & Wieschaus, 1980; Quiring, Walldorf, Kloter, & Gehring, 1994). Links between these regulators were progressively discovered by epistasis experiments involving reporter gene expression and *in situ* hybridizations in different backgrounds (Czerny et al., 1999; Frasch, Warrior, Tugwood, &

Levine, 1988; Harding, Hoey, Warrior, & Levine, 1989; Howard & Ingham, 1986). While these approaches were of limited throughput, they played a pivotal role in unraveling the regulatory principles that underlie developmental GRNs.

The emergence of high-throughput methods such as microarrays to measure gene expression, in combination with system-wide perturbation methods, allowed researchers to understand developmental GRNs in a much more detailed manner (Peter & Davidson, 2011). The completion of reference genome assemblies and characterization of non-coding elements through chromatin immuno-precipitation assays (Gerstein et al., 2010) allowed the identification of yet more GRN components and their regulatory interactions (Lei, Liu, Fukushige, Fire, & Krause, 2009; Van Nostrand & Kim, 2013). Additionally, computational analyses also improved the identification of TFs involved in certain processes through the presence of their motifs in promoters of activated genes (The FANTOM Consortium & Riken Omics Science Center, 2009).

Further refinements of high-throughput techniques and the development of novel methods, such as ATAC-seq, improved the ability to study non-coding elements important in GRN function. These advances in combination with improved computational analyses integrating multiple genome-wide datasets, greatly improved the reconstruction of GRNs (Madsen et al., 2018). It also shifted the focus from the study of TFs to studying REs directly and identifying TF candidates through their associated motifs (W. Wang et al., 2020). This strategy increased the throughput of regulatory links between TFs and REs and did not require the availability of TF-specific antibodies, which are often the salient bottleneck for ChIP approaches. However, follow-up experiments are necessary to confirm the direct action of the TF candidates on their potential targets to reduce false-positive discoveries. Finally, these technological advances also allowed researchers to explore GRNs in less extensively studied organisms (Gehrke et al., 2019; Neiro, Sridhar, Dattani, & Aboobaker, 2022; Pascual-Carreras et al., 2023; Ramirez, Loubet-Senear, & Srivastava, 2020).

Gene regulatory network modeling strategies

Topological models provide graphical representations of GRNs. They are composed of the key actors within the GRN, connected by links representing direct interactions. For instance, a TF will be shown to bind to a regulatory sequence controlling a certain gene and either activating or repressing its expression (Figure 1.17 A). Such models capture the overall structure of a GRN and serve as a basis for other modeling approaches.

As seen with the above-cited *evenskipped* example, regulatory genes controlling spatiotemporal processes such as development often exhibit discrete expression domains (Maduro, 2010; Niwa et al., 2005; Peter & Davidson, 2011). This characteristic is Boolean in nature, enabling researchers to formalize GRNs through logical models, such as Boolean networks (Karlebach & Shamir, 2008). In a Boolean network, each actor is either present (1) or absent (0) at a certain time in a given spatial domain. This is determined by the combination of regulators that are previously present in the same domain. The action of regulators towards their targets is encoded using Boolean logic functions such as AND, OR and NOT. The interactions between the GRN components can be visualized using topological networks while the spatiotemporal aspect of the network components is best represented using a Boolean matrix (Figure 1.17 B).

The two first types of models are inherently qualitative. Additionally, Boolean networks also assume that processes regulated by such networks occur in discrete steps. To better understand the temporal dynamics of regulatory processes and to integrate more quantitative data, continuous models such as Ordinary Differential Equations (ODE) are used (Jaeger et al., 2004; Manu et al., 2009; Perkins, Jaeger, Reinitz, & Glass, 2006). These models are represented by a set of mathematical equations that describe the change of each component as a function of the influences exerted by other network components (Figure 1.18). Such models are deterministic and only depend on the initial conditions of the system, thus not accounting for random variations between individual cells. Single molecule or stochastic models have been developed to incorporate this variability (Karlebach & Shamir, 2008).

Each of these models comes with its own set of advantages and disadvantages. Topological and Boolean networks can incorporate many actors but only give limited insight into the dynamics of the developmental process. In contrast, ODEs and stochastic models incorporate quantitative data and can explain specific processes on a very fine timescale but are computationally expensive. They can therefore model only a subset of the processed present in topological or Boolean networks.

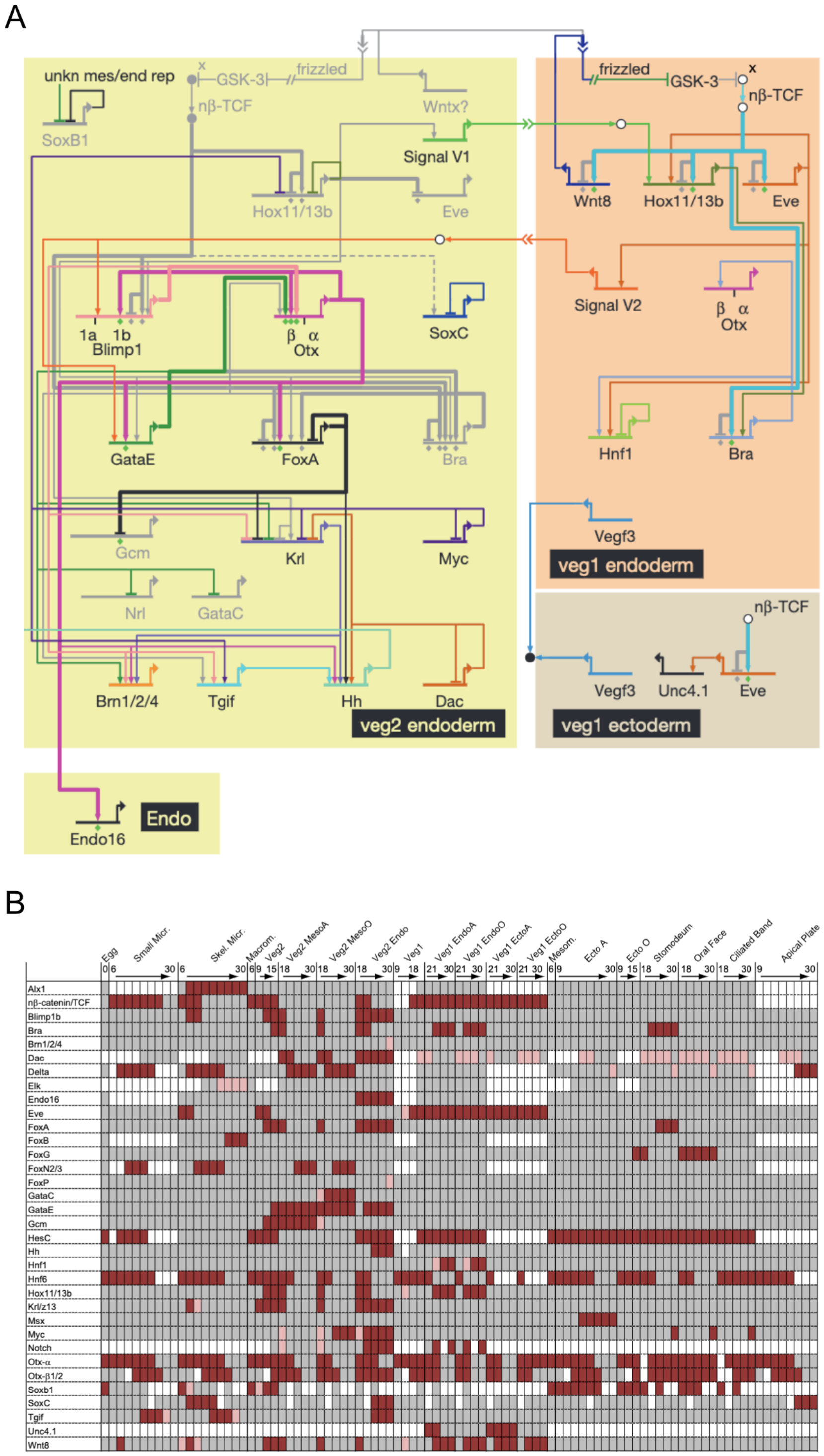


Figure 1.17: Topological and Boolean Gene regulatory network, adapted from (Peter & Davidson, 2011). A) example of a Topological GRN representing the anterior and posterior GRN modules just before gastrulation in the sea urchin embryo. B) Output of the Boolean computational model of the sea urchin endomesoderm GRN. This matrix represents the spatial and temporal expression chart for the important endomesodermal genes in different part of the developing sea urchin embryo.

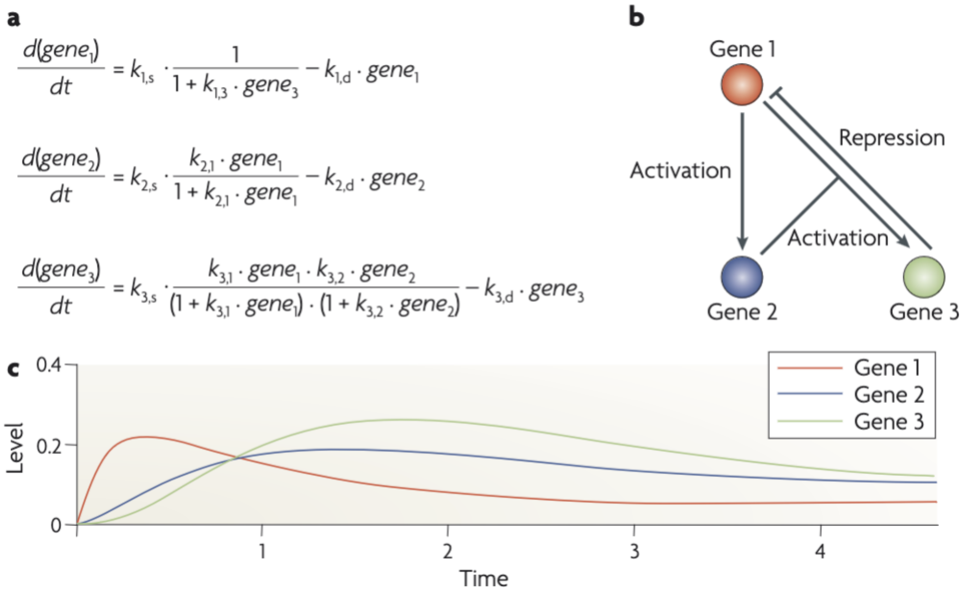

a
$\frac{d(gene_1)}{dt} = k_{1,s} \cdot \frac{1}{1 + k_{1,3} \cdot gene_3} - k_{1,d} \cdot gene_1$
$\frac{d(gene_2)}{dt} = k_{2,s} \cdot \frac{k_{2,1} \cdot gene_1}{1 + k_{2,1} \cdot gene_1} - k_{2,d} \cdot gene_2$
$\frac{d(gene_3)}{dt} = k_{3,s} \cdot \frac{k_{3,1} \cdot gene_1 \cdot k_{3,2} \cdot gene_2}{(1 + k_{3,1} \cdot gene_1) \cdot (1 + k_{3,2} \cdot gene_2)} - k_{3,d} \cdot gene_3$
b
Gene 1
Repression
Activation
Activation
Gene 2
Gene 3
c
0.4
0.2
0
Level
1
2
3
4
Time
Gene 1
Gene 2
Gene 3

Figure 1.18: Example of an ordinary differential equation model representing a simple regulatory network, adapted from (Karlebach & Shamir, 2008). A) Mathematical formulas representing the evolution of each component of the network over time in function of the other network components. B) Graphical representation of the regulatory relations between each network components. C) Evolution over time of each network component starting from an initial condition.

1.2.3 Modularity of gene regulatory networks and their use in different developmental contexts

Gene regulatory network sub-circuits

Development is a continuous process that relies on the precise orchestration of specific regulatory programs at each stage of development and in the correct spatial locations within the developing embryo. Various parts of the GRN are therefore active at different times and locations throughout embryogenesis. The temporal aspect of development confers a hierarchical nature to the GRN, while the spatial distribution of active GRN parts at any given moment organizes it into discrete modules (Davidson & Peter, 2015). These modules are further structured into individual sub-circuits, each responsible for specific functions within the module (Figure 1.19) (Peter & Davidson, 2009).

The characteristic of each sub-circuit is not the genes that compose it but the inherent topology of the circuit itself. Indeed, there are numerous examples where sub-circuits with identical topologies, yet composed of different genes, perform the same developmental functions in various organisms or distinct parts of a developing organism. This stems from the fact that development is orchestrated by the same types of processes, namely: 1) initial transient inputs have to be interpreted, 2) the subsequent regulatory state has to be stabilized and other states need to be repressed and 3) effector genes need to be expressed (Davidson, 2010b). The following section exemplify the similarities in sub-circuit topologies used during development of different organisms

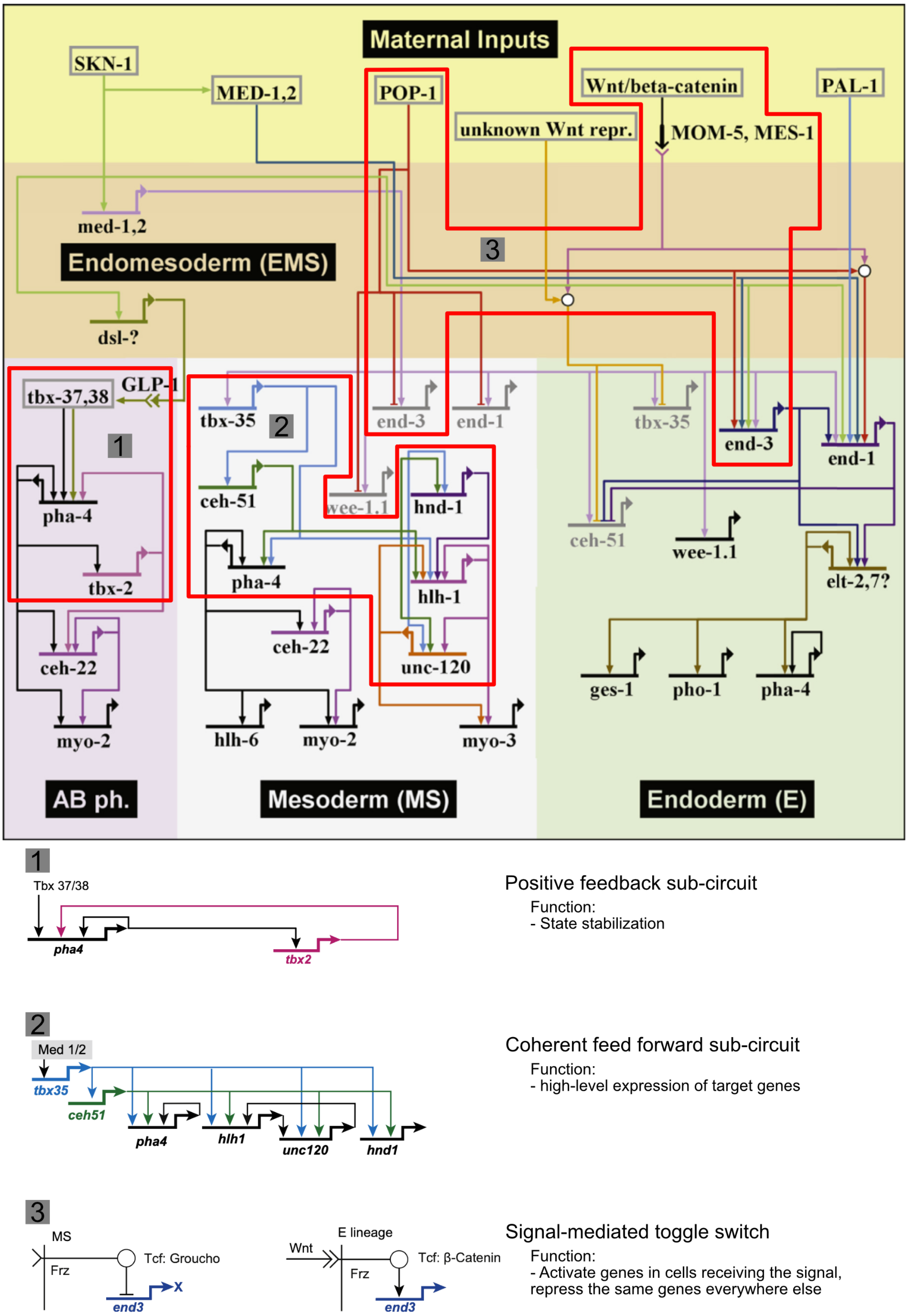


Figure 1.19: Gene regulatory network sub-circuits, adapted from (Davidson & Peter, 2015; Owraghi et al., 2010). GRN for the endomesoderm specification in *C. elegans*. Different types of common sub-circuits have been encircled in red. Their topology and function are depicted below the GRN.

Sub-circuits throughout development

Early embryogenesis

The initial task of an embryo is to establish its axes of symmetry. It does this by defining different spatial domains, each corresponding to a specific regulatory state that influences and restricts the possible fate of descendant cells within that domain. Each regulatory state corresponds to a module of the GRN and is characterized by the activation of certain regulatory genes. To initiate the specification of these special domains, certain inputs are necessary. These are in part already present before fertilization in the oocyte as asymmetrically deposited maternal regulatory molecules such as bicoid in *D. melanogaster* (Driever & Nüsslein-Volhard, 1988). In other organisms, the sperm entry point helps to determine the anteroposterior axis while the dorsoventral axis is established through cleavage asymmetry (Gotta & Ahringer, 2001). This first rough regionalization of the embryo will allow further implementation of different regulatory states driven by topologically conserved sub-circuits of hierarchically lower modules.

Interpretation of the initial transient input One example of such circuits is the double-negative gate, which has been described in early embryogenesis of the sea urchin as well as in *Drosophila*. This circuit facilitates the activation of a GRN module in a given part of the embryo. It accomplishes this by inhibiting the expression of an inhibitor, present throughout the embryo and acts to suppress the regulatory genes required for the activation of this GRN module (Davidson & Levine, 2008).

In the case of sea urchins, regulators are maternally and asymmetrically localized, resulting in the exclusive expression of *pmar1* in the mesodermal micromeres that will eventually give rise to the biomineralized skeleton (Figure 1.20 A). This expression locally represses *hesC*, an inhibitor of the GRN module necessary for skeletogenic mesoderm fate specification that is expressed globally (Oliveri, Carrick, & Davidson, 2002; Revilla-i Domingo, Oliveri, & Davidson, 2007). Similarly in *Drosophila*, Snail represses *tom* transcription in the mesoderm. This in turn allows local notch signaling that will activate gene expression in adjacent cells necessary for ventral midline specification (De Renzis, Yu, Zinzen, & Wieschaus, 2006) (Figure 1.20 A).

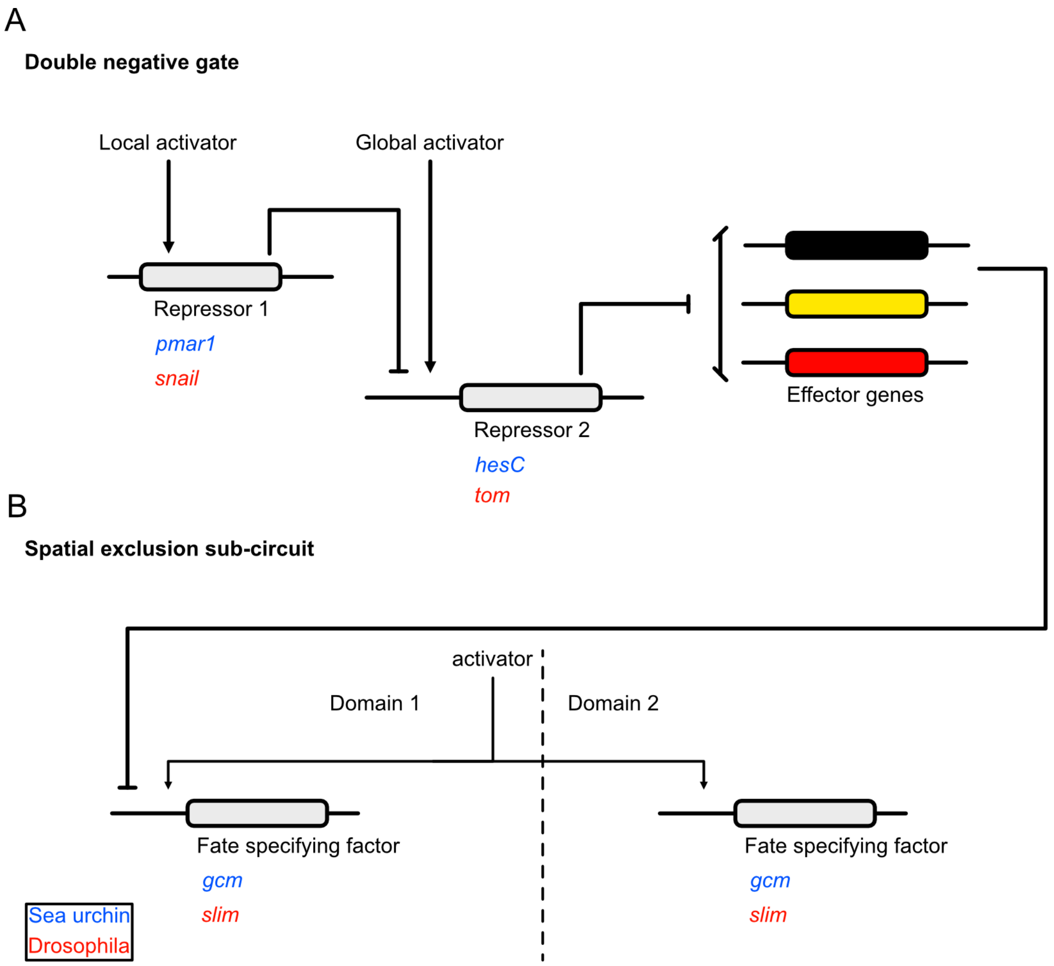


Figure 1.20: Exampled of two common sub-circuits found during embryogenesis, adapted from (Davidson & Levine, 2008). A) The double negative gate allows the expression of a gene in a certain part of the embryo while ensuring it is repressed in other parts. B) The spatial exclusion sub-circuit allows the repression of genes important for a certain fate by regulatory genes important for another cell fate.

Stabilization of the regulatory state Positive feedback loops are usually employed to stabilize a regulatory state once it has been established. This is exemplified in the mesodermal micromeres of the sea urchin embryo (Oliveri, Tu, & Davidson, 2008). The genes activated by *pmar1* activate three genes (*erg*, *hex* and *tgif*), which are organized in a positive feedback circuit to mutually increase their expression (Oliveri et al., 2008). Moreover, these genes activate effector genes that are essential for specifying the fate of skeletogenic mesoderm. A similar phenomenon is observed in the GRN module for pharynx muscle progenitor specification in *Caenorhabditis elegans* (Owraghi et al., 2010) (Figure 1.19).

Concurrently to the establishment of one regulatory state, other GRN modules necessary for the specification of different regions of the embryo must be repressed. Spatial exclusion sub-circuits prevent cells within a particular domain from responding to signals that are essential for specifying another domain. For example, in the skeletogenic mesoderm, *pmar1* leads to transcription of the notch ligand gene *delta*. This ligand then activate transcription of *gcm* in the adjacent mesoderm and leads to the activation of a GRN module for pigment cell fate specification (Yamazaki & Minokawa, 2016). However, in the skeletogenic mesoderm, this gene is repressed through the action of genes that are de-repressed by *pmar1* action, effectively inhibiting pigment cell fate specification in these cells. The same mechanism exists in *Drosophila* where *sim*, necessary for ventral midline specification, is repressed in the surrounding mesoderm expressing *snail* (Figure 1.20 B) (Kasai, Nambu, Lieberman, & Crews, 1992).

Organogenesis and body part formation

Later during embryonic development, parts of the initially regionalized embryo undergo further specification as different GRN modules are activated within these domains, initiating the formation of organs and body parts. The first step of body-part formation is the generation of a progenitor field, a domain containing cells that will subsequently divide and differentiate into a specific body part, such as a limb bud or imaginal disc (Davidson, 2001). Much like earlier stages of embryonic development, this process depends on initial inputs of regulatory genes. In this case, these inputs are established by GRN modules that were active earlier. For instance, previously established *hox* gene expression domains activate *tbx5*, a crucial step in the formation of the forelimb bud (Minguillon et al., 2012). This, in turn, activates the GRN module for forelimb specification which is then stabilized regionally by a positive feedback loop between *fgf 10* and *fgf 8* (Duboc & Logan, 2011). These examples demonstrate that similar processes are used in different developmental contexts and are driven by sub-circuits with similar topologies.

Following this, regionalization of the progenitor field needs to be established, a process that dictates the distinct sub-parts within the body part. One common sub-circuit frequently encountered during this phase is the signal-mediated reciprocal repression system. Its role is to repress the expression of regulatory genes of an adjacent domain upon the reception of a certain signal.

One example can be found in the domain subdivision around the *Drosophila* ocellus (Aguilar-Hidalgo et al., 2013). The ocellar space can be subdivided in 3 separate domains (interocellar, ocellar and periocellar). The formation of the ocelli requires the expression of 2 key genes (*eya* and *so*) which are specifically required in the ocellar domain (Figure 1.21). The interocellar domain expresses *hh* and creates a gradient of hh signaling that extends through the 3 domains. In the interocellar domain, hh is highly present and the *en* gene is transcribed through the action of the signal transducer of hh (ci). This gene will inhibit further hh signaling by repressing the expression of hh receptor *ptc*. A positive feedback loop keeps *en* expression in the interocellar domain. Constitutive Wnt signaling allows the expression of the ocellar specification antagonist (*hth*) in the interocellar domain, thereby repressing the expression of *eya* and *so*.

In the ocellar domain, hh signaling is also active but at a weaker level, preventing the expression of *en*. The hh signal transducer ci activates expression of *eya* and *so* mediating ocellar fate specification in this domain. Moreover, these two genes repress the expression of their antagonist *hth* to further stabilize the GRN module for ocellar fate specification. In the periocellar domain, hh signaling is not active and the ocellar fate specification is repressed by the constitutive expression of *hth* through Wnt signaling. A similar sub-circuit plays a role during the *C. elegans* vulva organogenesis (Ririe, Fernandes, & Sternberg, 2008) (Figure 1.21). In this case, a somatic gonadal cell controls the fate specification of two different vulval cell types by producing a signaling gradient that will be interpreted differently by each cell type on the basis of their position relative to the signaling source.

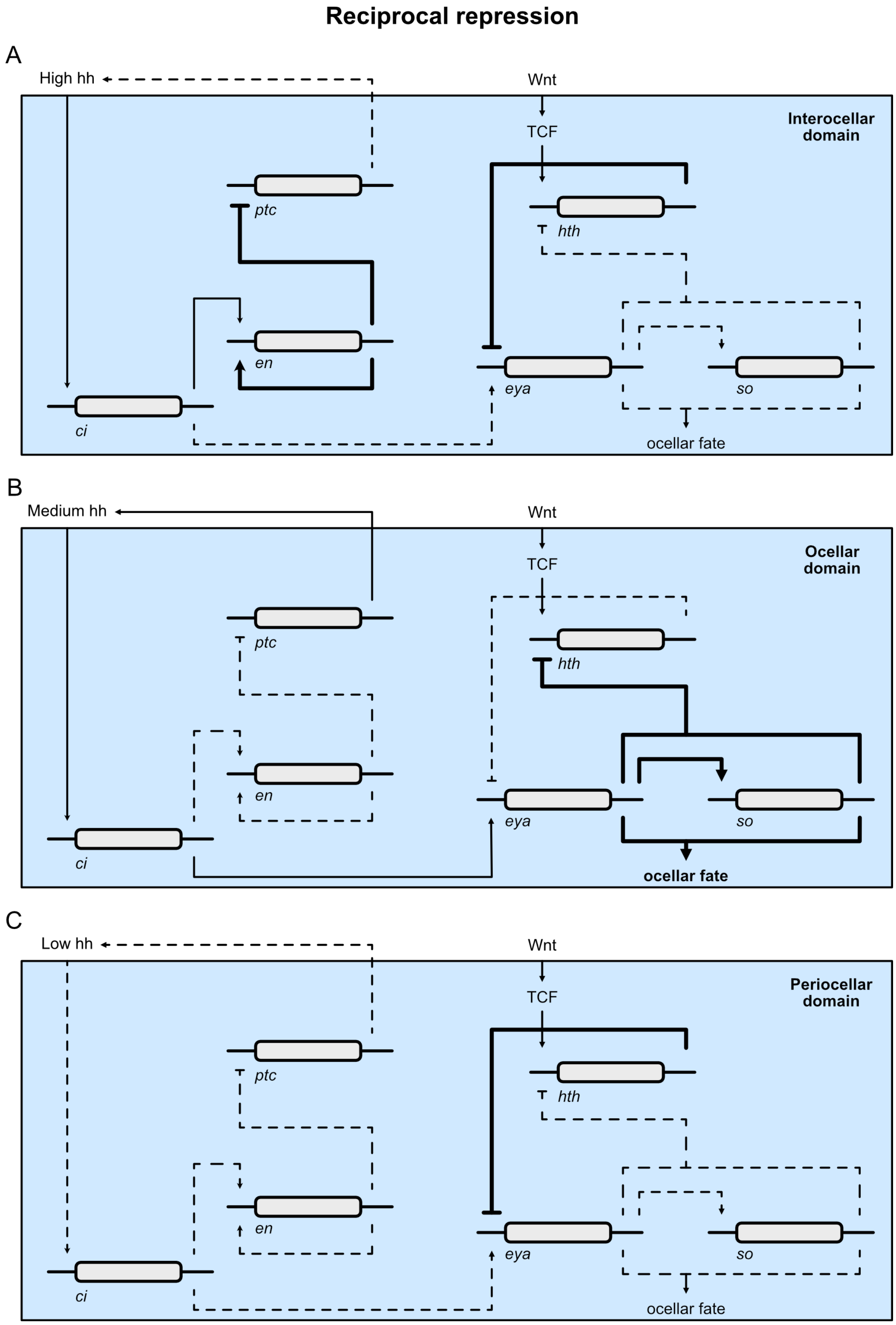


Figure 1.21: Example of the reciprocal repression sub-circuit during *D. melanogaster* ocellus formation, adapted from (Aguilar-Hidalgo et al., 2013). Depending on the intensity of hedgehog signaling, tissue will adopt the interocellar ocellar or periocellar domain fate relying on the repression or activation of two important genes for ocellar fate specification.

Terminal cell fate specification

Terminal differentiation is driven by a cohort of effector genes responsible for defining the structural and functional characteristics of the cell. These effector genes are situated at the bottom of the GRN modules that are activated during the transition of the last progenitor cell to the differentiated cell. The decision which GRN module to activate, i.e. which cell fate choice to make, relies on the activation of one or a few regulatory genes that control the expression of the terminal effector genes.

In cases involving binary cell fate decisions, such as hematopoietic cell differentiation (Graf & Enver, 2009), the choice is determined by the activation of one regulatory gene that is organized in a mutual repression sub-circuit with a regulatory gene determining the other cell fate (Davidson, 2010a). This choice depends on the exogenous signals to which the multipotent cell is exposed. These regulatory genes, in conjunction with others, subsequently activate the effector genes, commonly referred to as the 'differentiation gene battery' (Davidson & Peter, 2015). Terminal effector genes are often wired in a coherent feed-forward sub-circuit or directly under the regulation of the regulatory gene.

An example illustrating this sub-circuit topology has been elucidated in the context of pancreatic beta-cell differentiation (Habener, Kemp, & Thomas, 2005; Jensen, 2004; Servitja & Ferrer, 2004). Activation of the *ngn3* gene in pancreatic precursor cells will lead to *pax4* expression. This gene is organized in a mutual antagonism sub-circuit with the pancreatic alpha-cell differentiation regulatory gene *arx* (Collombat et al., 2003). Activation of *pax4* will therefore inhibit alpha-cell differentiation while also activating the expression of genes required for beta-cell fate such as *nkx6.1.* Moreover, *ngn3* activates the differentiation gene battery (composed of *insulin, iaap* and *gk*) through a coherent feed-forward loop involving *nkx2.2* among others (Figure 1.22).

A similar sub-circuit organization is found during embryonic erythropoiesis in zebrafish (A. T. Chen & Zon, 2009; Davidson & Peter, 2015). In the posterior lateral mesoderm, responsible for generating embryonic erythrocytes, *tif1gamma* is proposed to positively regulates both the erythroid and the myeloid cell fate regulatory gene (*gata1* and *pu.1* respectively). These genes are organized within a mutual repression sub-circuit. However, a positive feedback loop exists between *gata1* and *tif1gamma*, allowing *gata1* to repress *pu.1* and activate the differentiation gene battery for erythrocyte cell fate. In mice, *gata1* also drives the GRN module for differentiation of precursors in erythrocytes where a coherent feed-forward loop between *gata1* and the differentiation gene-battery has been resolved (Swiers, Patient, & Loose, 2006).

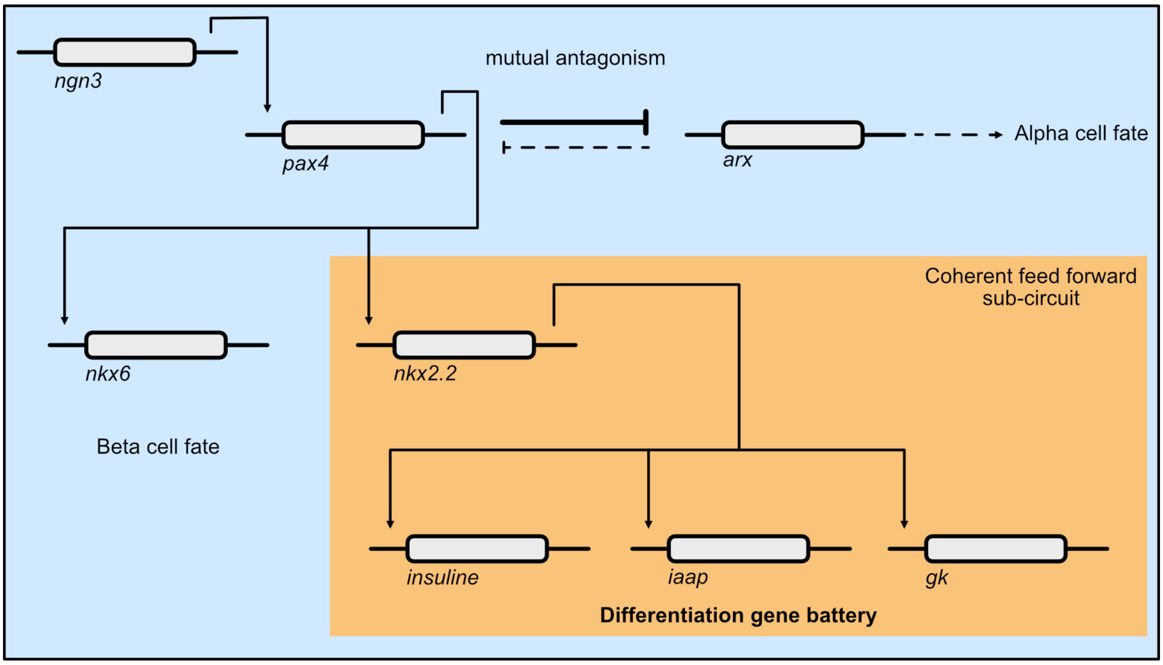


Figure 1.22: Example of two common sub-circuits found in terminal cell fate specification of pancreatic beta cells, adapted from (Davidson, 2010b). Regulatory genes important for terminal cell fate specification are often found in a mutual antagonism sub-circuit with regulatory genes involved in other cell fates. Moreover, terminal effector genes can also be found in coherent feed-forward loops.

1.3 Studying gene regulatory networks in planaria

Planarians are well known for their ability to perform whole body regeneration (Morgan, 1898). This remarkable capability hinges upon their abundant population of adult pluripotent somatic stem cells, situated within the mesenchymal tissue, known as neoblasts. Neoblasts are known to be a heterogeneous cell population with some capable of giving rise to all the differentiated cell types present in planaria (Wagner, Wang, & Reddien, 2011). Furthermore, they are the only division-competent somatic cells and constantly supply differentiated tissues with new cells. This is necessary since all planarian tissues are characterized by high cell turn over (Pellettieri & Alvarado, 2007; Rink, 2013). This distinctive trait necessitates that the progeny of neoblasts undergo differentiation, transitioning from a pluripotent state to a fully committed terminal cell fate within the context of an already fully developed organism. Moreover, the differentiating neoblasts will also need to migrate and be incorporated in the target tissue as they are located within the mesenchyme (Reddien, 2021).

This is vastly different from adult vertebrate multipotent stem cells. These cells are situated within specific niches of their target tissues and directly supplement its target tissues. Moreover, adult vertebrate stem cells are not pluripotent and their fate are restricted to certain lineages. For example, a hematopoietic stem cell is multipotent and can give rise to certain blood cell types. However, it cannot differentiate into a goblet cell found in the gut. This goblet cell originates from another type of adult multipotent stem cells called an intestinal stem cell that is situated within another niche located in the intestine (Santos, Lo, Mah, & Kuo, 2018; Seita & Weissman, 2010).

It is also different from the progressive differentiation of pluripotent stem cells during embryonic development that relies on the hierarchical progression of GRN modules to organize the developing embryo into different domains, generate body parts and organs to finally lead to the activation of differentiation gene batteries for the assignment of a terminal cell fate (see 1.2). Indeed, developmental GRNs that govern the progressive differentiation of pluripotent stem cells rely on some transitory developmental states that do not exist anymore in an adult animal. Planarians therefore emerge as a valuable system for investigating the gene regulatory networks associated with the maintenance and differentiation of adult pluripotent stem cells.

Moreover, planarians possess different reproductive strategies, sometimes utilized within the same species by different biotypes. Asexual reproduction via fission is a characteristic of the asexual biotype, distinguished by the absence of reproductive organs. On the other hand, the sexual biotype is able to generate gametes and perform sexual reproduction. The absence of specific tissues associated with sexual reproduction in the asexual biotype provides a unique opportunity to uncover the gene regulatory network (GRN) dedicated to the development and maintenance of the reproductive system in planarians (Issigonis & Newmark, 2019; Y. Wang, Stary, Wilhelm, & Newmark, 2010).

Finally, the regeneration response also involve complex regulatory programs. Although they certainly share certain features with embryonic GRNs (Johnston et al., 2019), these programs must possess distinctive regulatory mechanisms given the variability associated to the regenerative process. Indeed, since planarians can regenerate from a seemingly random piece of tissue. A certain regulatory logic thus needs to exist to direct the remaining tissue towards the regeneration of only what is missing. Planarians represent therefore another unique opportunity to study GRNs implicated in regeneration (Goldman & Poss, 2020).

In summary, planarians represent a unique model system appropriate the study of GRNs involved in adult stem cell systems, the differentiation and maintenance of specific tissues and the orchestration of regeneration. In the following paragraphs, I will give a brief overview of the planarian model organism. I will then detail its anatomy with a focus on its reproductive system and describe the different modes of planarian reproduction. Finally, I will discuss the initial investigations into GRNs in planaria.

1.3.1 Introduction to planaria

Planarians, also called triclads are a taxonomic group within the phylum platyhelminthes, which includes various organisms such as parasitic flatworms (e.g., schistosomes and tapeworms), macrostomids that are also increasingly studied as biomedical model organisms (Wudarski et al., 2020) and the sometimes flamboyantly coloured marine polyclads (Sluys & Riutort, 2018) (Figure 1.23). Unlike schistosomes and tapeworms, planarians are free living animals found in a wide variety of ecosystems such as marine, fresh water and terrestrial habitats. They are carnivorous by nature, typically preying on live or recently deceased organisms, including insect and crustacean larvae, annelids, mollusks, and even amphibian eggs (Vila-Farré & C Rink, 2018).

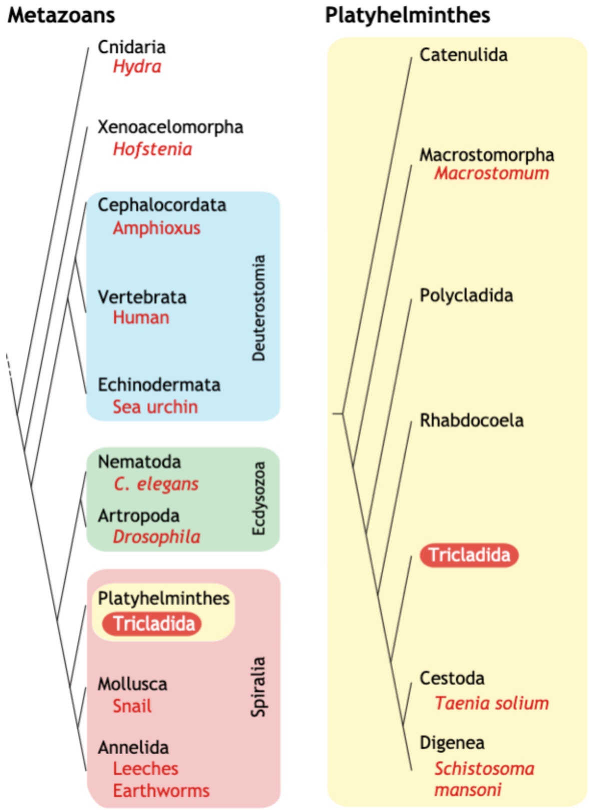


Figure 1.23: Location of planaria (Tricladida) within the larger phylogeny, adapted from (Ivankovic et al., 2019). Left, phylogenetic relationship between Platyhelminthes and other metazoan. The large colored boxes represent the different clades. Right, zoom on the Platyhelminthes phylum contextualizing the place of planaria within it. Important species for each group are denoted in red.

The main model species of planarian research is *Schmidtea mediterranea*, a freshwater species occurring around the Mediterranean basin. It is used as a model species because of its capacity to perform whole body regeneration even from small pieces of tissue (Benazzi, Baguñà, & Ballester, 1970). This species consists of two distinct biotypes, each with its specific reproductive strategy (Benazzi, Baguñà, Ballester, Puccinelli, & Papa, 1975). The asexual biotype lacks any reproductive organs and creates clonal progeny by ripping the tip of its tail off, which subsequently regenerates into a new organism. The asexual biotype is used in most laboratories to study regeneration due to the easy maintenance of large populations of worms in laboratory conditions. In contrast, the sexual biotype is able to generate gametes engages in sexual reproduction through cross-fertilization to generate their F1 progeny (Guo, Zhang, Rubinstein, Ross, & Alvarado, 2016; P. A. Newmark & Alvarado, 2002).

Planarian anatomy

General anatomy

Planarians are triploblastic acoelomates and possess somewhat less complex organ systems than other bilaterians. Their lack of a coelom results in their internal organs not being enclosed within a body cavity but rather surrounded by loosely organized mesodermal tissue known as the parenchyma or mesenchyme (Sluys & Riutort, 2018). Notably, planarians do not possess a dedicated respiratory system and instead rely on diffusion to provide oxygen to their various organs. Despite their relative simplicity, planarians still exhibit remarkable sophistication in their other organ systems.

A distinctive feature of planarians is their three-branched intestine which is the origin of the taxonomic designation tricladida. The three primary branches branch off side branches that again branch hierarchically into trees of side branches extending throughout the planarian body (Forsthoefel, Park, & Newmark, 2011) (Figure 1.24). This network not only functions in food digestion but also covers the role of a circulatory system, by facilitating the distribution of nutrients throughout the worm's entire body. Consequently, it is referred to as the gastrovascular system.

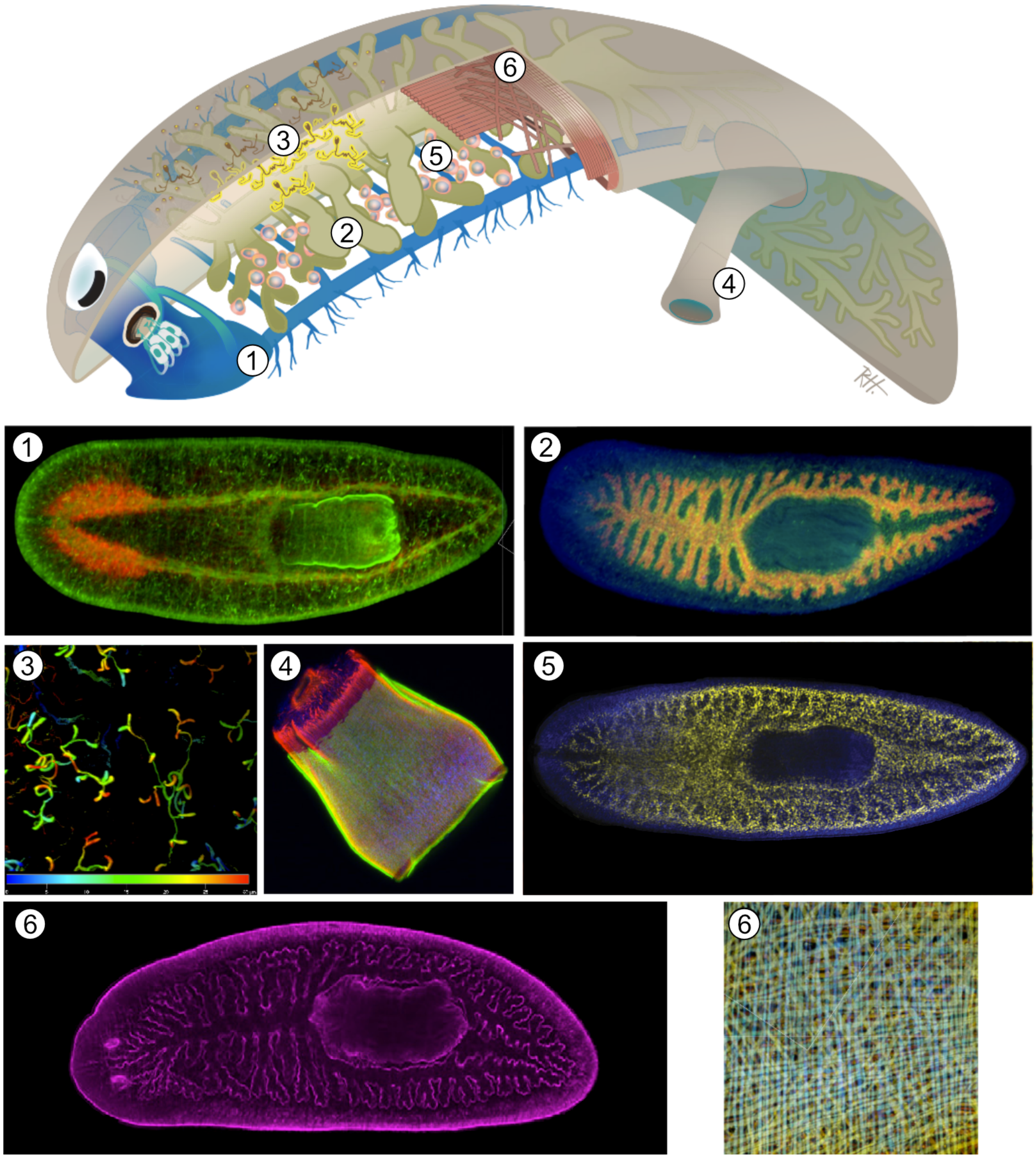


Figure 1.24: The general anatomy of planarians, adapted from (Forsthoefel et al., 2011; Ivankovic et al., 2019). 1. The planarian CNS consisting of the cephalic ganglia (red) and connected to two central nerve chords that run along the anteroposterior axis of the animal. 2. The gastrovascular system composed of three primary branches (one anterior and two posterior) as well as many more higher order branches. 3. Zoom on individual protonephridial units composing the planarian excretory system. 4. The muscular pharynx 5. Neoblasts (yellow) located within the mesenchyme 6. The planarian musculature around the gastrovascular system. The second image shows the body-wall musculature composed of longitudinal, circular and diagonal fibers.

The intestine is connected to the outside by a single opening that serves both as a mouth and anus, extending into a muscular pharynx responsible for nutrient ingestion. Muscles also surround the gastrovascular system to aid in the dispersion of nutrients through peristaltic movements. Planarians also feature other muscles that compose the body wall musculature and are arranged in differently oriented networks of fibers (Roberts-Galbraith & Newmark, 2015; Witchley, Mayer, Wagner, Owen, & Reddien, 2013) (Figure 1.24). Their main role is to coordinate movement but have been shown to also play a critical part in orchestrating regeneration (Reddien, 2018).

The nervous system of planarians includes two cephalic ganglia in the head, which are connected to ventral nerve cords (VNCs) that extend posteriorly to the tail tip, which collectively form the central nervous system (CNS) (Figure 1.24). The VNCs are interconnected by transverse commissures, giving the CNS a ladder-like structure (Sluys & Riutort, 2018). Planarians are negatively phototactic, which is at least in parts mediated by photoreceptive neurons associated with the prominent pigment cups in the planarian head, also known as eyes (Ross, Currie, Pearson, & Zayas, 2017). The peripheral nervous system is composed of sensory neurons and nervous plexuses that innervate all organs and are connected to the CNS via nerve tracts (Monjo & Romero, 2015). Protonephridia embedded in the mesenchyme serve as their excretory system and play an important role in osmoregulation and waste excretion (Rink, Vu, & Alvarado, 2011). They consist of branched tubules that are capped by so-called flame cells, which provide cilia-driven ultrafiltration (Thi-Kim Vu et al., 2015) (Figure 1.24).

Finally, a cell population called neoblasts is situated within the mesenchyme (Figure 1.24). As the only mitotically active cells outside the reproductive system, they are responsible for the constant supply of new cells to all planarian tissues which are characterized by generally high turnover rates (Pellettieri & Alvarado, 2007; Rink, 2013). As previously mentioned, neoblasts are a heterogeneous cell population. Some neoblasts are truly pluripotent stem cells capable of repopulating animals devoid of stem cells (Wagner et al., 2011). Other neoblast classes can be identified by their expression of specific transcription factors and represent already lineage-committed cells (Fincher, Wurtzel, de Hoog, Kravarik, & Reddien, 2018; Plass et al., 2018).

Despite of this heterogeneity, all neoblasts are characterized by the expression of germ line related genes (such as *piwi*, *vasa*, *bruno* and others), with *piwi* family genes being the typical neoblast markers (Reddien, Oviedo, Jennings, Jenkin, & Alvarado, 2005). The specific mechanism of neoblast differentiation is still debated. A hierarchical model of neoblast differentiation has been proposed. There, one specific subgroup of neoblast are the truly pluripotent adult stem cells and give rise to all neoblast progeny (Zeng et al., 2018). However, a more recent report showed evidence that lineage-committed neoblasts could produce progeny with different cell fates through asymmetric divisions (Raz, Wurtzel, & Reddien, 2021) and therefore possibly retain their pluripotency.

Neoblasts continuously proliferate at a basal rate to counteract cell turnover but can also respond to certain stimuli like feeding (Baguñà, 1974; Kang & Alvarado, 2009) or injury (Wenemoser & Reddien, 2010) by increasing their proliferation rate. In response to a feeding stimulus, neoblast proliferation will result in the growth of the animal and the scaling of its internal organs to maintain body proportions (Takeda, Nishimura, & Agata, 2009). Conversely, during prolonged periods of starvation, planarians undergo degrowth, enabling them to survive without sustenance for extended amounts of time. This characteristic also means that planarians do not have a fixed body size (Thommen et al., 2019). In the event of injury where a segment of the worm is missing, neoblasts will be responsible of the regeneration of the missing body part by increasing their proliferation rate and differentiate into the appropriate cell type, all under the guidance of positional cues (Reddien, 2018).

The planarian reproductive system

Sexually reproducing planarians, including the sexual *S. mediterranea* biotype, are typically simultaneous hermaphrodites, meaning they possess both male and female reproductive gonads (Sluys & Riutort, 2018) (Figure 1.25). In addition to these gonads, the reproductive system includes several accessory reproductive organs required for various functions related to sexual reproduction.

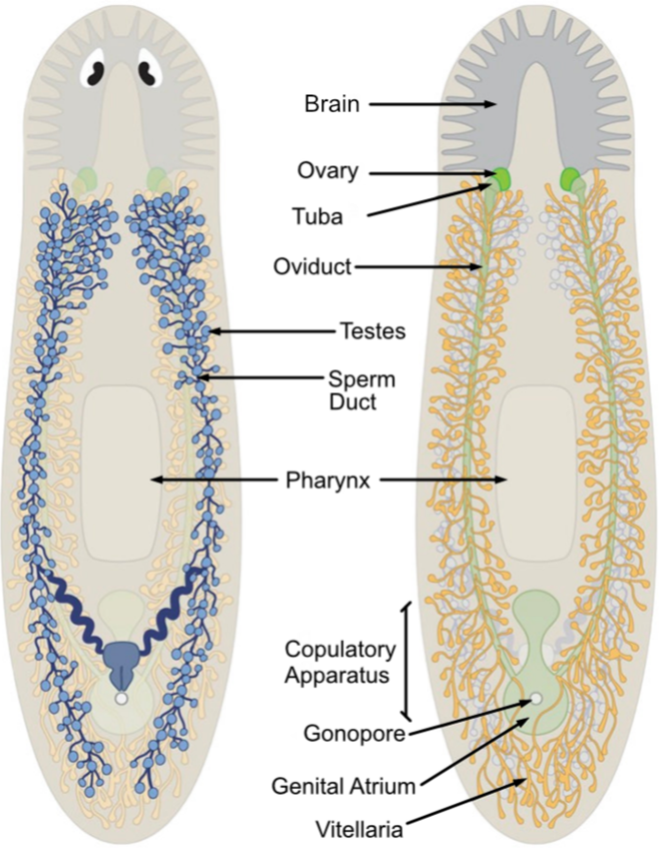


Figure 1.25: The planarian reproductive system, adapted from (Issigonis et al., 2022). Planarians are hermaphrodites and therefore possess both male and female reproductive structures. A pair of ovaries are situated ventrally at the base of the brain and connected to the genital atrium via the oviducts. The male reproductive system is composed of a series of dorsally located testes lobules. Sperm will be transported via the sperm ducts and stored in the seminal vesicle (dark blue). The vitellaria produced yolk cells which will serve as the food source for the developing embryo. The gonopore is the opening through which the copulatory apparatus protrudes during copulation. It is also through which planarians lay their cocoons.

The male reproductive system The testes are distributed dorsolaterally on either side of the midline up to the posterior end of the cephalic ganglia and are organized in individual lobules (Chong, Stary, Wang, & Newmark, 2011). In each testis lobule, spermatogenesis progresses from the outer layer of the lobule towards the lumen (Figure 1.26). The outer layer is composed of spermatogonia arising from presumptive germline stem cells (GSCs), themselves descendants of mesenchymal neoblasts (Issigonis & Newmark, 2019). Spermatogonial cells undergo three rounds of mitosis, resulting in the formation of eight spermatogonial cysts that remain interconnected by intracellular bridges (Issigonis et al., 2022). Subsequently, they differentiate into spermatocytes and undergo meiosis, ultimately producing 32 round spermatids. These cells will elongate during spermiogenesis to form elongated spermatids and finally give rise to mature sperm. Mature sperm will accumulate in the lumen of the lobule and be transported through the sperm ducts and vasa defferentia to be finally stored in the seminal vesicle (Figure 1.27). During copulation the stored sperm is delivered to the partner by passing through the penis papilla and stored in the copulatory bursa (Vila-Farré & C Rink, 2018) (Figure 1.27).

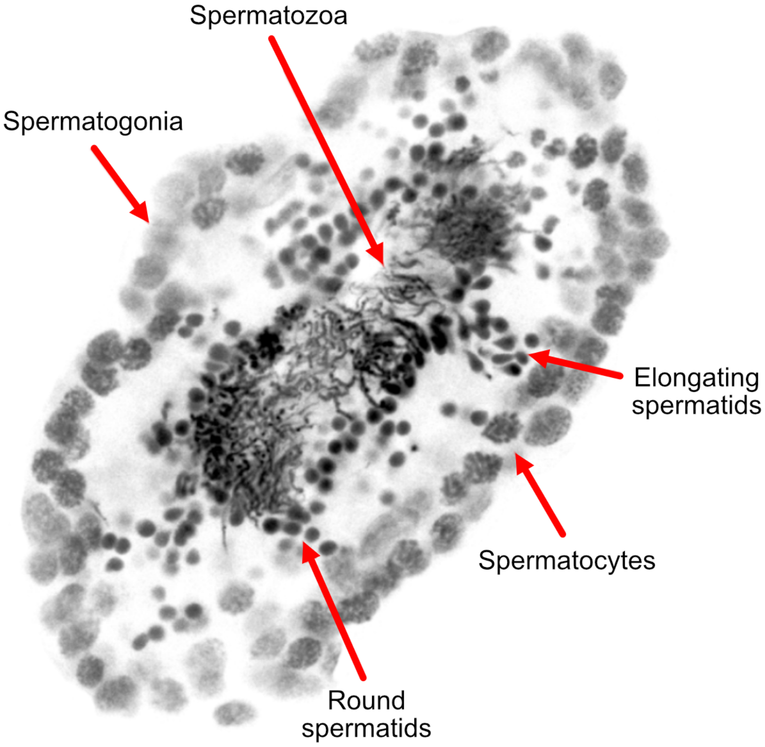


Figure 1.26: The planarian testis, adapted from (Chong et al., 2011). The testes lobule is composed of cells in different stages of spermatogenesis. The outer layer is composed of spermatogonial cells. They will differentiate in spermatocytes. After meiosis, spermatocytes give rise to round spermatids that elongate during spermiogenesis to form elongating spermatids and finally give rise to mature sperm.

Numerous markers associated with distinct stages of spermatogenesis have been identified. The presumptive GSCs are marked by *klf4l*, an ortholog of the KLF4 pluripotency factor (Issigonis et al., 2022), as well as *nanos* (Handberg-Thorsager & Saló, 2007; Y. Wang, Zayas, Guo, & Newmark, 2007), a conserved RNA binding protein found in germ cells across various organisms (Extavour, 2007). As GSCs differentiate into spermatogonia, they gradually lose *klf4l* expression, followed by a decrease in *nanos* expression. Concurrently, spermatogonial markers such as *rap55* and *gapdh* begin to be expressed (Y. Wang et al., 2010).

Spermatocytes are characterized by the expression of *tkn1* while *pp2* and *pka* are expressed specifically in spermatids (Chong et al., 2011). Notably, no distinct marker has been identified for mature sperm; however, their elongated nuclei make them readily distinguishable through DAPI staining. Specification and differentiation of sperm cell

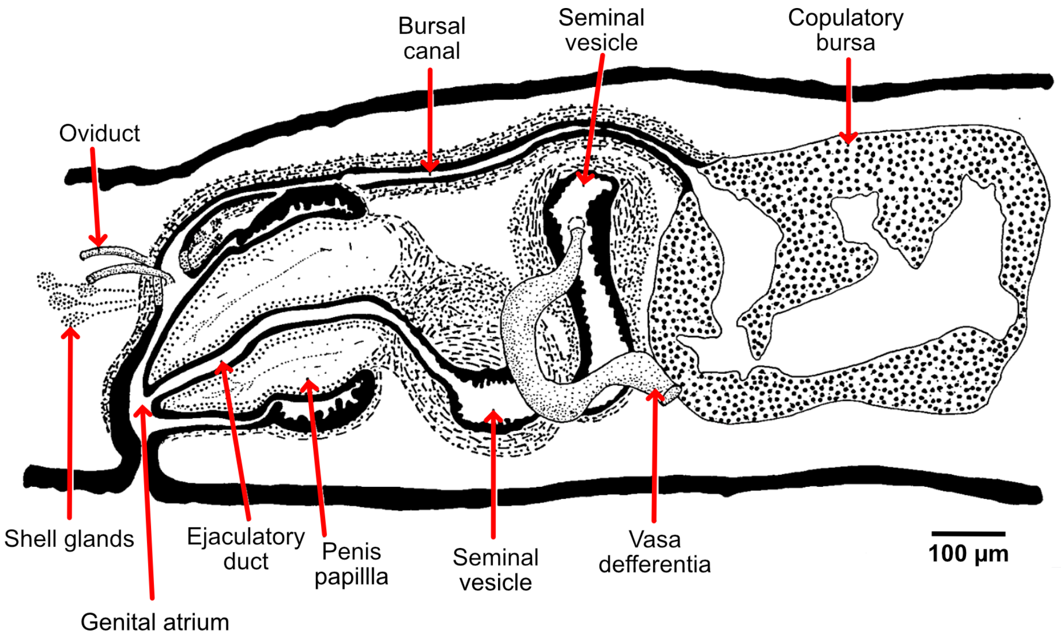


Figure 1.27: The planarian copulatory apparatus, adapted from (Harrath et al., 2004). Sperm arrives through the vasa defferentia and stored in the seminal vesicles. During copulation the penis papilla will be inserted in the genital atrium of the partner and release mature sperm through its ejaculatory duct. Sperm will then be stored in the copulatory bursa. They will then travel through the oviducts to fertilize the oocytes. Zygotes will migrate down the oviducts and be surrounded by yolk cells. In the genital atrium excretions of shell glands will react with yolk protein to form a hard shell around yolk cells and multiple developing embryos, encasing them in a cocoon. The cocoon will finally be laid through the gonopore.

progenitors is supported by somatic cells associated to the testes, collectively known as the somatic gonad (Chong, Collins, Brubacher, Zarkower, & Newmark, 2013). Genes such as *dmd-1* and *ophis* are markers of the male somatic gonad and are necessary for its integrity (Chong et al., 2013; Saberi, Jamal, Beets, Schoofs, & Newmark, 2016). Lastly, the sperm ducts and seminal vesicle are characterized by the expression of *grn* (Chong et al., 2011).

The female reproductive system Sexual *S. mediterranea* also feature a pair of ovaries located at the base of the cephalic ganglia on the ventral side of the animal (Sun, Xie, Sun, Song, & Li, 2012) (Figure 1.25). Similar to the testes, the ovaries display multiple stages of oogenesis. Female germline progenitors (FGPs)/oogonia are specified in the periphery of the ovary. These cells are then internalized and begin to differentiate into oocytes (U. W. Khan & Newmark, 2022) (Figure 1.28). Immature oocytes are mostly found in the distal part of the ovary while bigger mature oocytes are located proximally with respect to the tuba, a specialized part of the oviduct. Following ovulation, the mature oocytes are fertilized by sperm in the tuba and subsequently traverse the ciliated oviduct toward the genital atrium (Issigonis & Newmark, 2019) (Figure 1.27). In addition to germ cells, the planarian ovary is also composed of somatic gonadal cells that help FGP differentiation and oocyte maturation (Figure 1.28).

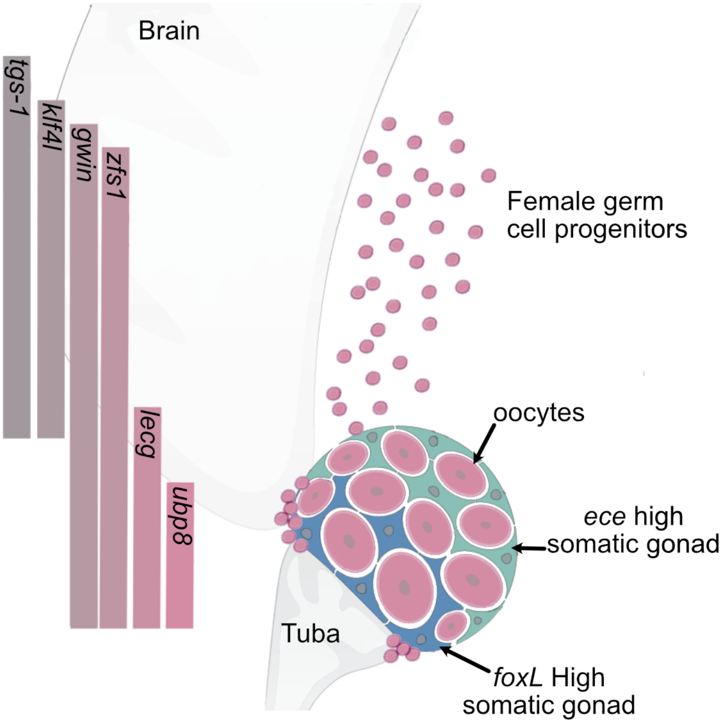


Figure 1.28: Schematic of the planarian ovary, adapted from (U. W. Khan & Newmark, 2022). Female germ cell progenitors are specified outside of the ovaries and will be incorporated subsequently. They will then begin the process of oogenesis. Immature oocytes are found at the distal part of the ovary compared to the entry of the oviduct (the tuba). Multiple genes marking different stages of oocyte development are indicated on the left. The somatic gonad is important for oocyte development and heterogeneous in gene expression patterns.

The study of the planarian female reproductive system has been hampered by the limited abundance of these tissues compared to other reproductive organs like the testes. Recently, a study performed by Khan and colleagues managed to overcome this difficulty by generating gonad-specific transcriptomes using laser-capture microdissection (U. W. Khan & Newmark, 2022). Their study uncovered multiple female germ cell markers spanning different stages of development (Figure 1.28). Like male presumptive GSCs, FGPs are marked by *nanos* (Handberg-Thorsager & Saló, 2007; Y. Wang et al., 2007) and *klf4l* (Issigonis et al., 2022) but also express other markers like *gwin*, *tgs-1* and *zfs1* (U. W. Khan & Newmark, 2022). During oocyte differentiation *nanos, klf4l* and *tgs-1* expression is lost but *zfs1* and *gwin* expression is retained. In addition to this, early oocytes start to express *lecg* and *ubp8* expression becomes visible in mature oocytes located in proximity to the tuba (U. W. Khan & Newmark, 2022). The study also reveals that the ovarian somatic gonad is not uniform as cells proximal and distal to the tuba show different expression patterns. Tuba-distal cells are characterized by a low *fox* and high *ece* expression, while the reverse pattern is observed in the tuba-proximal cells (Figure 1.28). Other genes like *ophis* and *delta3* also mark the female somatic gonad (U. W. Khan & Newmark, 2022; Saberi et al., 2016). Additionally, the oviducts are marked by their expression of an unnamed hypothetical protein conserved in *Schistosoma mansoni* (Rouhana, Tasaki, Saberi, & Newmark, 2017).

Accessory reproductive structures Other tissues outside of the testes and ovaries are necessary for planarian sexual reproduction and together form the accessory reproductive organs.

The vitellaria is the largest and most extensively studied accessory reproductive organ (Figure 1.25). It exists as an extensive network of cells located on the ventral side of the animal, primarily responsible for the production of yolk cells, known as vitellocytes, which serve as the main nutrient source for developing embryos (Benazzi & Gremigni, 1982). Indeed, planarian eggs are characterized by their lack of yolk and these specialized yolk cells assume the vital role of providing essential nutrients to the developing embryos, a phenomenon referred to as ectolecithality (Laumer & Giribet, 2014).

Interestingly, a study performed by Issigonis and colleagues revealed that yolk cell development shares similarities with that of planarian germ cells (Issigonis et al., 2022). Yolk cell progenitors are also marked by their expression of *nanos* and *klf4l* and gradually lose these markers during yolk cell differentiation. As they mature, yolk cells will then acquire other markers like *surfactant-b* (Steiner, Tasaki, & Rouhana, 2016) and *cpeb-1* (Rouhana et al., 2017) and mature vitellocytes are marked by their expression of *mx1* (Rouhana et al., 2017). Furthermore, the vitellaria also comprises a distinct population of cells with high expression of the somatic gonadal marker *ophis* (Saberi et al., 2016). It is hypothesized that these cells play a role akin to the somatic gonad in maintaining and differentiating the yolk cell progenitors (Issigonis et al., 2022).

Various types of shell glands are also found around the genital atrium and are crucial for the formation of a cocoon, also called egg capsule, which serves as the protective enclosure for developing embryos (Figure 1.25 and 1.26) (Sluys & Riutort, 2018). These glands are characterized by the expression of *tsp-1* (Chong et al., 2011) or *tsp66e* (Rouhana et al., 2017).

Shell glands are thought to secrete proteins in the genital atrium. These proteins subsequently interact with the content of yolk cells, leading to the formation of a sclerotin shell in a through a mechanism called 'quinone tanning', encapsulating many yolk cells and multiple zygotes in the process (Gremigni & Domenici, 1974). While the shell glands are integral to egg capsule formation, a majority of the proteins and molecules involved in shell construction are expressed within yolk cells (Rouhana et al., 2017). For instance, genes expressed in yolk like *tan-1*, *synaptotagmin XV* and *surfactant b* are important for capsule formation and their repression lead to capsule defects as well as infertility (Rouhana et al., 2017).

Extrinsic regulation of germ cell maintenance and differentiation

Germ cells are known to rely on extrinsic regulation by somatic tissues for their specification, maintenance and development (Greenspan, De Cuevas, & Matunis, 2015; Steinberger, 1971). Among these somatic tissues, the somatic gonad plays an important role in these processes as it is in direct contact with the germ cells. As mentioned before, both the planarian testes and ovaries as well as potentially the vitellaria partly composed of somatic cells. Genes expressed in these tissues such as *dmd-1, foxL, delta3* and *ophis* have been shown to play an important role in germline biology as their repression affects germ cell specification, maintenance and/or differentiation (Chong et al., 2013; U. W. Khan & Newmark, 2022; Saberi et al., 2016).

Moreover, the somatic gonad, including in the vitellaria, also expresses an enzyme (AADC) necessary for the production of monoamines such as dopamine and serotonin (U. W. Khan & Newmark, 2022). This enzyme is required for the maintenance and regeneration of germ cells and yolk cells. Knocking down *aadc* results in the ablation of ovaries and yolk cell ablation, while leading to hyperplastic testes and an increase in male GSCs that fail to differentiate (U. W. Khan & Newmark, 2022). Interestingly, a derivative peptide hormone of another monoamine is necessary for the induction of female sexual development in schistosomes (R. Chen et al., 2022). It is plausible that the local production of monoamines by AADC in the planarian somatic gonad serves a similar function.

Germ cell regulation in vertebrates also involves the central nervous system (CNS) through the hypothalamic–pituitary–gonadal (HPG) axis, utilizing neuropeptides as signaling molecules (Gołyszny, Obuchowicz, & Zieliński, 2022). Investigating neuronal regulation of planarian gonads has revealed that a neuropeptide as well as its receptor (*npy-8* and *npyr-1* respectively) regulates male reproductive structures as their knockdown lead to the loss of copulatory organs as well as testes regression (Collins III et al., 2010; Saberi et al., 2016). It's noteworthy that both the neuropeptide and its receptor are exclusively expressed in the CNS, and their downstream mediators governing male reproductive functions remain to be discovered.

Finally, nuclear hormone receptors (NHRs) are recognized for their roles in reproduction in both mammals (R.-S. Wang, Yeh, Tzeng, & Chang, 2009) and invertebrates (Asahina et al., 2000). In planarians, the nuclear hormone receptor gene *nhr-1* exhibits expression in the male and female reproductive structures such as the oviducts, sperm ducts and seminal vesicle, but not in the testes or ovaries (Tharp, Collins III, & Newmark, 2014). Intriguingly, knocking down this gene not only leads to the loss of these structures but also disrupts germ cell differentiation, suggesting the existence of a feedback regulatory mechanism between the planarian gonads and their accessory reproductive organs.

Origins of asexuality in *Schmidtea mediterranea*

Evidence suggests that the asexual strain originated from a now extinct population of sexual planarians on the Iberian Peninsula, while the remaining sexual populations have survived mainly in islands of the Mediterranean Sea (Lázaro et al., 2011). Despite the discovery of numerous genes vital for sexual reproduction, the genetic basis of asexuality remains elusive. However, the fact that sexual and asexual *S. mediterranea* have been proven to be the same species (Lázaro et al., 2011) suggests that differences in regulation of gene expression might be the cause for the emergence of asexuality. For instance, even though the asexual do not possess sexual organs, gonadal primordia containing *nanos* positive GSCs are still visible but fail to develop (Handberg-Thorsager & Saló, 2007).

A potential contributing factor to the failure of sexual organ development in asexual *S. mediterranea* is the presence of a heteromorphic translocation between 1st and 3rd chromosomes (Baguñà et al., 1999). Interestingly, a recent study put forward arguments suggesting that chromosome 1 functions as a sex-primed autosome (Guo et al., 2022). This chromosome is enriched for genes with important functions in sexual reproduction like *nanos, nhr-1, npy-8* and *npyr-1.* It is also characterized by inbreeding resistant heterozygosity and shown to be incapable of recombination (Guo et al., 2022, 2016), a feature commonly associated with sex chromosomes (Bergero & Charlesworth, 2009). This translocation, might therefore disrupt the chromatin organization necessary for the spatiotemporal regulation of genes with important functions in gonadal development and lead to a premature developmental arrest of the reproductive tissues.

Furthermore, recent findings highlighting the differential Wnt signaling activity between sexual and asexual *S. mediterranea*, along with the enrichment of pathway components in sexual tissues, suggest the involvement of heightened Wnt signaling in sexual reproduction (Vila-Farré et al., 2023). Notably, the knockdown of *beta-catenin-1* in sexual planarians resulted in the loss of vitellaria and shell glands. A similar result is also seen in the seen in the sister species *Schmidtea polychroa* where *beta-catenin* RNAi leads to the loss of sexual organs like the testes (Sureda-Gomez, Martin-Duran, & Adell, 2016).

1.3.2 Previous studies of gene regulatory networks in planaria

Planarians are important model organisms so study complex spatiotemporal processes such as regeneration and adult stem cell differentiation. These processes rely on the sequential activation of specific genes thought to be organized into and regulated by GRNs. Understanding the structure of these GRNs would therefore reveal the mechanistic basis of such processes.

Many tools to study GRNs have been historically lacking in planarians. For instance, RNA interference is the only functional assay to study specific genes as transgenesis still hasn't been developed in planarians. The lack of transgenesis also means that no fluorescent reporter assays are available to study the differentiation process of specific lineages. Only recently did a sufficiently contiguous genome assembly for *S. mediterranea* become available (Grohme et al., 2018). This is also the case for methods to study regulatory elements such as ChIP-seq, CUT & TAG and ATAC-seq (Duncan et al., 2015; Ivankovic et al., 2023; Mihaylova et al., 2018; Neiro et al., 2022; Pascual-Carreras et al., 2023; Poulet, Kratkiewicz, Li, & van Wolfswinkel, 2023). However, no ChIP-grade planarian antibody exists for TFs, limiting the use of protocols using antibodies to generic histone modification marks. Moreover, the annotation of REs in planarians is still in its infancy and a complete annotation of REs has yet to be published. Finally, methods to study transcription initiation in planarians are also non-existent.

Gene regulatory network involved in neoblast maintenance and differentiation

Many studies have uncovered GRN components involved in neoblast fate specification (Molina & Cebrià, 2021) (Figure 1.29). Moreover, multiple models of neoblast fate specification have been put forth (Adler & Alvarado, 2015; Raz et al., 2021; Zeng et al., 2018). However, the topology of the different fate specification modules as well as the cues to maintain pluripotency are still poorly understood.

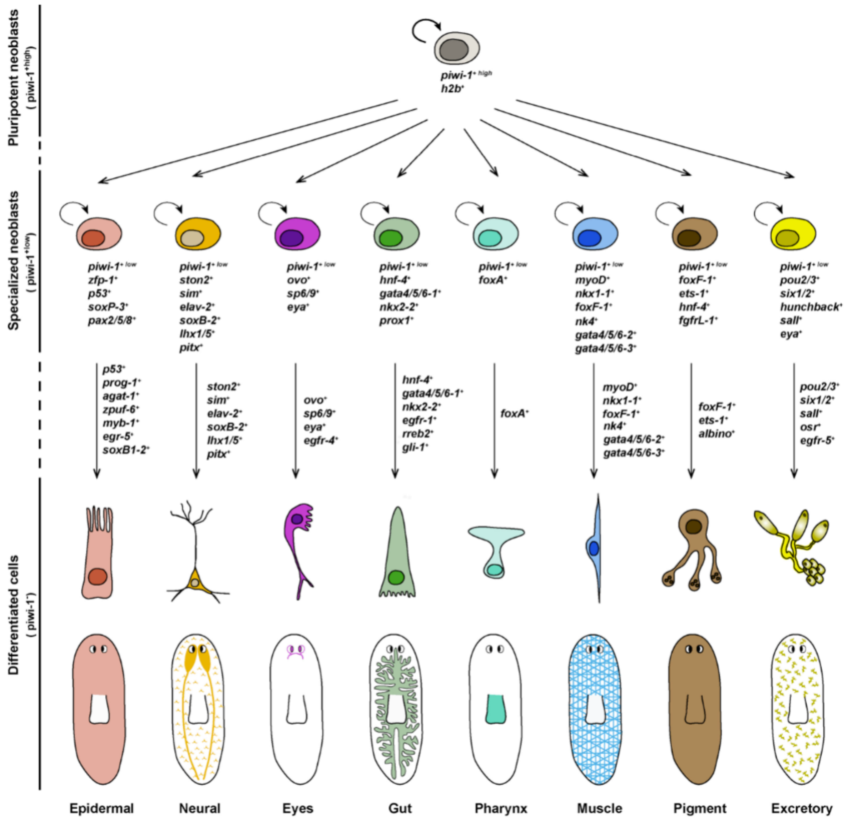


Figure 1.29: Known markers of neoblast differentiation and their subsequent progenitor cells, adapted from (Molina & Cebrià, 2021)

A recent study performed by Neiro and colleagues sought to identify regulatory links between different TFs (i.e., regulatory genes) driving neoblast differentiation. To achieve this, they identified motifs within ATAC-seq footprints in enhancer-like regions associated with each of these TFs. This enabled them to recreate a putative GRNs involving fate-specifying TFs (FSTFs) important for neoblast differentiation (Neiro et al., 2022) (Figure 1.30 A).

Their investigation revealed numerous interactions between TFs involved in different fate specifications, leading to the hypothesis that inhibitory binding between these TFs might contribute to stabilizing specific fate specifications. Additionally, they attempted to validate their model by demonstrating that certain interactions within their network align with functional studies conducted on specific TFs. For instance, their GRN predicts that a *coe*, a COE TF family member, regulates another transcription factor called *pou4-1*, two TFs involved in neuronal fate specification. Furthermore, *coe* RNAi leads to down-regulation of *pou4-1* (Cowles, Omuro, Stanley, Quintanilla, & Zayas, 2014) and knock-down of both genes lead to a similar neuronal defect phenotype. However, it is important to note that many of these connections still require further verification.

Additionally, a recent report from Poulet gave insight into the possible mechanism behind the maintenance of neoblast pluripotency (Poulet et al., 2023). Their findings reveal that, unlike at tissue-specific gene promoters, neoblast specific genes possessed very little transcription factor binding motifs but were enriched in homopolymeric AT stretches that promote nucleosome eviction by chromatin remodelers (Lorch, Maier-Davis, & Kornberg, 2014). Moreover, they demonstrated that knockdown of two neoblast-enriched chromatin remodeler, ISWI and SNF2, reduced neoblast-specific gene expression as well as their capability to proliferate. They conclude that, unlike in vertebrates where pluripotent stem cell identity is regulated by important TFs like OCT4, planarian neoblast identity might rely on the absence of specific TF expression and rely on other mechanisms for their maintenance.

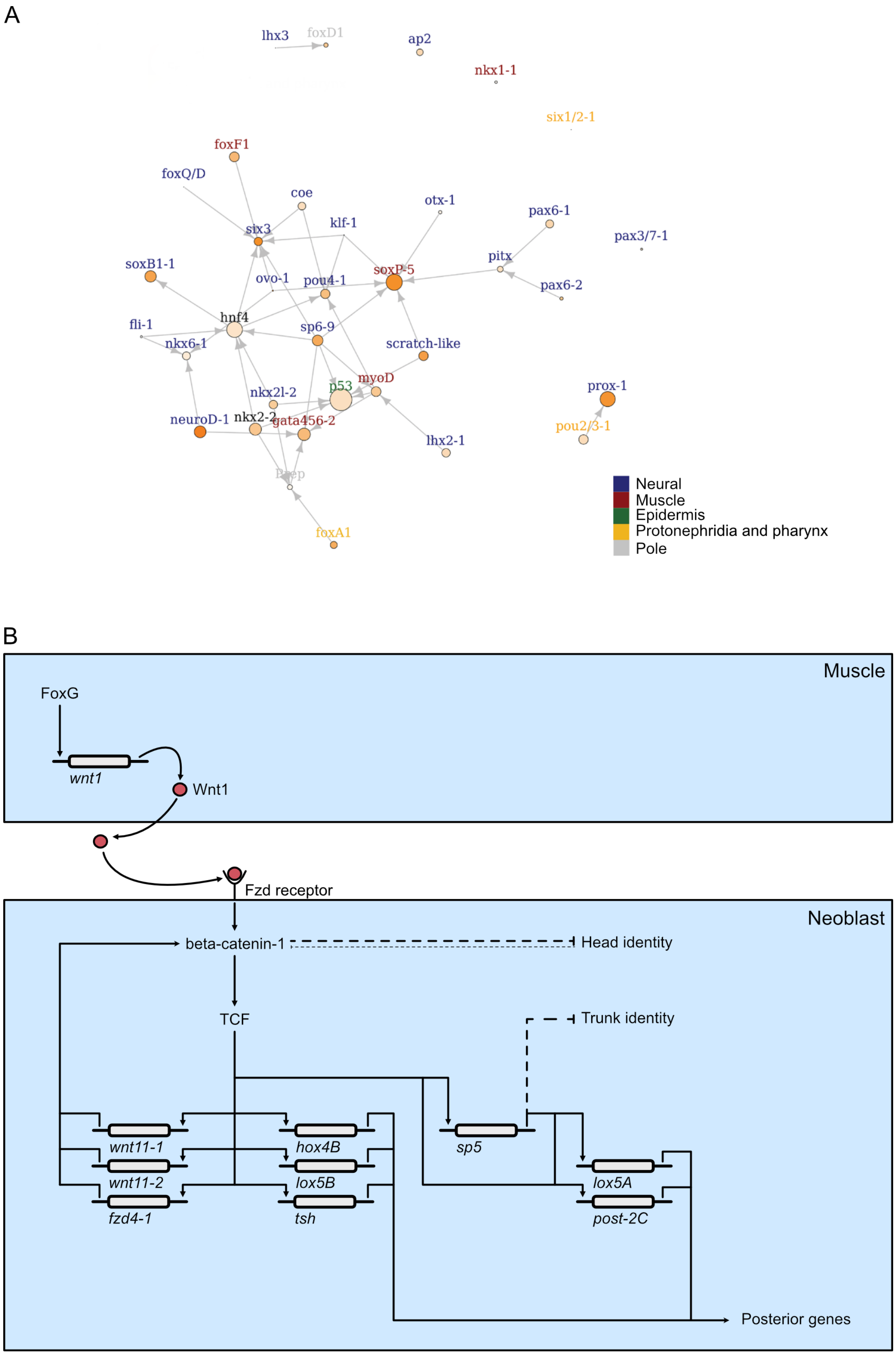


Figure 1.30: Caption on the next page

Figure 1.30: Gene regulatory networks in planaria. A) putative GRN active in planarian stem cells adapted from (Neiro et al., 2022). Genes present in this GRN are known fate specifying transcription factors. Multiple regulatory links between TFs involved in different fates might suggest some regulatory interactions necessary for a specific fate commitment. Note that the arrows in this graph only denote that a motif of the TF has been found in a RE associated with gene it is supposedly regulating and does not necessarily mean activation. The color of the TFs represent the fate they are associated to. The node size represents the absolute expression (in TPM) of the TF within neoblasts. B) GRN governing the anterior fate specification of neoblast after injury, reconstructed from (Neiro et al., 2022; Pascual-Carreras et al., 2023) and references within. Wound induced Wnt1 expression is dependent on the FoxG transcription factor. The Wnt 1 ligand will then activate the cWnt pathway in neoblasts. This initial input is stabilized by the Wnt depended expression of Wnt pathway components organized in a positive feedback loop which increased the intracellular beta-catenin-1 content. Increased beta-catenin-1 is also responsible for anterior fate repression through an unknown mechanism. Anterior fate specification was also shown to repress posterior fate specification, organizing the effectors of both fates in a mutual antagonistic sub-circuit. Additionally, posterior fate effector genes are expressed in a Wnt-dependent manner, some of them organized in a coherent feed-forward loop. Finally, trunk identity has been shown to be repressed by the SP5 transcription factor.

Posterior fate specification during regeneration

A major area of planarian research focuses on the question of regeneration polarity. How does a worm decide what tissue needs to be regenerated after a specific injury and what are the mechanisms behind this decision (Reddien, 2018)? The best studied regeneration paradigm is the head versus tail decision that the worm is confronted to after transversal amputation. Over the years, a multitude of components contributing to this decision-making process have been identified using techniques such as RNAi, *in situ* hybridization and RNA-seq (Adell, Salo, Boutros, & Bartscherer, 2009; Gurley, Rink, & Alvarado, 2008; Iglesias, Gomez-Skarmeta, Saló, & Adell, 2008; Owlarn et al., 2017; Petersen & Reddien, 2011a, 2011b; Reuter et al., 2015; Scimone, Lapan, & Reddien, 2014; Stückemann et al., 2017; Tewari, Owen, Petersen, Wagner, & Reddien, 2019; Vogg et al., 2014) with most of them belonging to the specification of the posterior/tail fate.

As in other organisms the canonical Wnt signaling pathway was found to be a major contributor in the anteroposterior axis specification and re-establishment of appropriate Wnt expression was shown to be critical for the regeneration of the right part. Many Wnt pathway components are expressed in the posterior part of the worm, while Wnt antagonists are mostly anteriorly expressed. This spatial distribution results in the head and tail of the worm being characterized by low and high Wnt signaling environments, respectively (Rink, 2018).

Furthermore, a significant number of genes known to have an instructive role in neoblast differentiation during regeneration and tissue turnover are expressed in a layer of subepidermal muscle cells (Witchley et al., 2013). Genomic indications on the topology of the GRN module involved in tail fate specification came more recently (Neiro et al., 2022; Pascual-Carreras et al., 2023). Although certainly incomplete, a GRN for the implementation of the tail fate decision can be reconstructed using the above information (Figure 1.30 B).

The initial cue triggering posterior fate specification remains unknown. Nevertheless, *wnt1* expression from muscle cells has been shown to be a critical input to start initialize tail specification. Indeed, its repression leads to the regeneration of heads at posterior-facing wounds or the inability to regenerate tails leading to a 'tailless' phenotype (Adell et al., 2009; Petersen & Reddien, 2009). Pascual-Carreras and colleagues identified a transcription factor (*foxG*) that possesses TFBSs in intronic *wnt1* enhancers and phenocopies the *wnt1* regeneration defects (Pascual-Carreras et al., 2023).

When Wnt1 is released into the mesenchyme, it activates the canonical Wnt signaling pathway in neoblasts, resulting in the stabilization of beta-catenin-1. Due to the presence of TCF binding sites at Wnt signaling agonist genes, a positive feedback loop stabilizes the initial Wnt1 input and lead to a commitment to the posterior fate (Pascual-Carreras et al., 2023; Stückemann et al., 2017). Moreover, Wnt signaling activity was also shown to repress anterior fate specification and vice versa, reminiscent of the reciprocal repression system in Figure 1.21 (Stückemann et al., 2017). Other TFs important for posterior fate specification are directly under the control of the Wnt signaling pathway or organized in a coherent feed forward loop (see Figure 1.22) (Neiro et al., 2022). These factors then activate the expression of posterior genes enriched in Hox TFBSs (Pascual-Carreras et al., 2023). Finally, posterior TF (*sp5*) that is also regulated by TCF (Pascual-Carreras et al., 2023) has been shown to repress the more anteriorly expressed trunk genes (Tewari et al., 2019).

Although the direct interactions between TFs and genes depicted in this putative GRN are supported by motifs, it is important to note that definitive proof of binding will require the development of new tools such as transgenesis or TF ChIP. Moreover, many more actors are known to play a role in this process but their location within the GRN is unknown. Further studies are needed to uncover the cue(s) necessary for the initiation of the tail GRN.

1.4 Outstanding questions and scope of the thesis

Planarians are a fascinating model organism to study cellular processes like adult stem cell systems, cell differentiation and regeneration. These processes rely on the activation of specific GRNs driven by key transcription factors. While significant progress has been made in uncovering genes important for these various mechanisms, the historical lack of available methods to identify regulatory elements in planaria has impeded the identification of causal regulatory links between them. Consequently, our understanding of the GRNs governing different facets of planarian biology remains limited, with only a few recent studies beginning to address this subject. Moreover, the recognition that transcription initiation can serve as a good predictor of both enhancer and promoter activity represents a promising avenue to sensitively probe for RE activity in different conditions. This approach holds the potential to identify active parts of GRNs in those conditions.

During my thesis, I aimed at studying gene regulatory elements in *S. mediterranea* through the lens of transcription initiation as this aspect of REs had not yet been explored in this organism. My objective was to uncover the key transcription factors (TFs) responsible for governing the gene regulatory networks (GRNs) active in the planarian reproductive system. To accomplish this, I leveraged the naturally occurring biotypes in *S. mediterranea*, which exhibit variations in their reproductive strategies. My research was structured around four main aims:

The first section of the results describes the establishment of a robust protocol to study transcription initiation in *S. mediterranea* with as aim to identify active REs. The second section pertains to the extensive characterization done on the identified REs composing the planarian transcription initiation landscape. It involved evaluating their distribution, chromatin landscape, motif content, and the occurrence of bidirectional transcription initiation. In the third section, I compare the sexual and asexual transcription initiation landscape and identify motifs, located within the identified REs, that display significant variability between the two biotypes. This allowed me to identify TF candidates that potentially play a role in GRNs governing the development and maintenance of the planarian reproductive system. In the fourth and last section of the results, I try to validate the function of the candidate TFs in the reproductive system. This was done by conducting a comprehensive analysis of the expression patterns of potential TF candidates and performing functional tests to assess their significance in the maintenance and regeneration of the planarian reproductive system.

Chapter 2

Material and Methods

2.1 Experimental methods

2.1.1 Animal husbandry

The sexual and asexual strain of *Schmidtea mediterranea*, S2F2 and CIW4 respectively, were housed in custom re-circulation culture systems in 1x Montjuïc salts (Cebrià & Newmark, 2005) in a temperature-controlled environment at 20°C. The animals were fed homogenized calf liver prepared as described in (Merryman, Sánchez Alvarado, & Jenkin, 2018). Worm cultures were expanded by amputation followed by regeneration. Prior to experiments, planarians were transferred to 20°C stationary cultures of 1x Montjuïc Salts (also called Planarian Water) supplemented with Gentamicin sulfate (Santa Cruz Biotechnology, Cat n°: sc-203334F) if not stated otherwise. Worms used for experiments were starved for at least one week.

2.1.2 Protein extraction and Western Blot

Whole animal tissue fixation and protein extraction

Mucus was stripped by bathing the 2 7-mm worms in 0,5% pH neutral N-acetyl-cysteine (NAC) solution (0,5% (w/v) NAC (Sigma-Aldrich, Cat n°: A7250-100G), 20 mM HEPES-NaOH pH 7,25, 0,1% phenol red (Sigma-Aldrich, Cat n°: P0290-100ml)) (Pearson et al., 2009) for 10 min at room temperature followed by two washes in deionized water. Animal tissue was fixed by incubating the worms in a zinc-based fixation solution (0.5% ZnF3Ac (AlfaAesar, Cat n°: 18686), 0,5% ZnCl2 (Fluka, Cat n°: 96469), 0.05% CaAC, 0.1M Tris-NaCl (Roth, Cat n°: 9090.3) pH 6,7) (Lykidis et al., 2007) for 10 minutes in a 60mm petri dish. The worms were then lysed by mechanical homogenization in Urea lysis buffer (9M

Urea (Merck, Cat n°:1.08487.1000), 100mM NaH2PO4 (Sigma-Aldrich, Cat n°: S7907-500G), 10mM Tris-NaCl pH8, 2% SDS (Serva, Cat n°: 20765,03), 130mM DTT (Thermo Scientific, Cat n°: R0862), 1mM MgCl2, 1x Halt protease inhibitor cocktail (Thermo Scientific, Cat n°: 78429)) (Hall et al., 2022) with the help of a pellet mixer (VWR, cat n°: 431-100) and incubated for 10 minutes at room temperature. The worm lysates were subsequently spun down for 10 min at 12000g to remove debris. Protein concentration of was quantified on the basis of the absorbance at 280nm (Thermo Scientific, Cat n°: ND-1000) and normalized to a 2,4 µg/µl and aliquoted per 33,34 µl. The aliquots were topped up to 40 µl by addition of 6,66 µl of 6x LDS loading buffer (12% LDS (Acros, Cat n°: 413300250), 0.06% Bromophenol Blue (Sigma-Aldrich, Cat n°: B5525-5G), 50% glycerol (Thermo Scientific, Cat n°: 17904), 600 mM DTT, 60 mM Tris-NaCl pH 6,8) (Hall et al., 2022). Aliquots were finally denatured for 10 min at 65°C on a heating block (Eppendorf,Cat n°: 5382000015) before use.

Nuclear protein extraction

The nuclear pellet obtained after nuclei purification (see 2.1.6) was resuspended in hot urea lysis buffer and proteins were denatured for 10 minutes at 65°C. The lysates were then spun for 10 minutes at 12000g after cooling down to room temperature in order to remove cell debris. The protein content was then measured by 280 nm absorbance reading, normalized to 2,4 µg/µl and aliquoted per 33,34 µl. Finally, 6,66µl of 6x LDS loading buffer was added to the aliquots.

SDS-PAGE

Proteins were separated by size using SDS-PAGE (Laemmli, 1970). NuPAGE™ 4-12% Bis-Tris Protein Gels (Thermo Scientific, Cat n°: NP0321BOX) were pre-run for 10 minutes at 100V in 1x NuPAGE™ MES SDS Running Buffer (Thermo Scientific, Cat n°: NP0002) using the XCell SureLock™ Mini-Cell electrophoresis system (Thermo Scientific, Cat n°: EI0001). A total of 10 µg of proteins per sample were loaded on the gel and together with the PageRuler™ Plus Pre-stained Protein Ladder (Thermo Scientific, Cat n°: 26619) were run for ± 1h45 until the migration front has reached the bottom of the gel.

Western blot

Following SDS-PAGE, gels were removed from their cast and rinsed in ddH2O to prepare for Western Blot (Towbin, Staehelin, & Gordon, 1979). The blotting sandwich was carefully assembled in the XCell IITM Blot module (Thermo Scientific, Cat n°: EI9051) using the gel, WhatmanTM papers (GE Healthcare, Cat n°: 10427806) and AsheramTM ProtranTM 0,2 µM nitrocellulose blotting membrane (GE Healthcare, Cat n°: 1060004) and soaked in Transfer buffer (1x NuPAGETM transfer buffer (Thermo Scientific, Cat n°: NP00061), 20% Methanol). Blotting was done at 4°C for 2 hours at a constant voltage of 20V. To ensure proper transfer of the proteins of the desired size, gels were subsequently stained with Blue Silver Coomassie staining solution (5% (w/v) aluminium sulfate, 0.02% (w/v) Coomassie Brillant Blue G250, 10% (v/v) ethanol (96%), 2% (w/v) ortho-phosphoric acid) (Candiano et al., 2004) for one hour, rinsed twice in ddH2O and destained using destaining solution (10% ethanol (v/v), 2% (v/v) ortho-phosphoric acid) until sufficiently cleared.

In order to quantify the total protein amount in each sample, membranes were placed in flat trays, first rinsed twice in ddH2O and stained for 20 minutes with Revert™ 700 Total Protein Stain (LI-COR, Cat n°: 926-11011). Excess stain was then removed away by two subsequent washes of 30 seconds using the total protein stain washing solution (30% methanol (v/v); 6.7% acetic acid(v/v)). Following this, membranes were imaged using the Amersham Typhoon instrument by exciting them using a 685 nm wavelength in combination with a 720/20 bandpass filter (IRshort).

Membranes were afterwards incubated in blocking solution (5% Soy protein powder (Powerstar food, Art n°: psf-1139) in PBS) for 1 hour at room temperature with agitation followed by 2 washes in PBS. To detect both targets, ef1alpha (50kDa) and Histone3 (15kDa), membranes were cut between the 25 and 30 kDa mark according to the ladder and parts of the membrane were incubated with their respective primary antibody overnight at 4°C (anti-H3 (Ab1791, lot: GR3237728-1) and anti-Ef1alpha (Home-made, clone CP21), both diluted 1:100.000 in 1% Soy protein powder 0,1% PBS Tween 20 (Sigma-Aldrich, Cat n°: P9416-100ml) (PBSt)). Next, membranes were washed three times 10 minutes in PBSt and incubated with the appropriate secondary antibody in the dark for one hour at room temperature (ef1alpha; CF® 770 Goat Anti-Mouse IgG (Biotium, Cat n°: 20077-1), H3: CF® 770 Donkey Anti-Rabbit IgG (Biotium, Cat n°: 20484-250µl)). Following secondary antibody incubation, membranes were washed three times 10 minutes in PBSt followed by three additional washes in PBS. Imaging was performed on the Amersham Typhoon instrument by excitation using a 785 nm wavelength in combination with a 825/30 bandpass filter (IRlong).

Quantitative Western blot analysis

All Western blot images were quantified using the ImageStudioLite software (LICOR). Total protein signal was quantified for each sample by drawing a rectangle along the entire length of the lane. Background signal was subtracted using the 'user defined' setting, corresponding to another rectangle on the membrane away from the lanes. Target proteins (H3 and Ef1alpha) were measured for each sample by drawing rectangles, identical in size within a membrane, around each band. Background was also subtracted using the 'median' setting (3 pixels border width, top/bottom segments). For each replicate, H3 and Ef1alpha signal was normalized to the corresponding total protein signal. To assess the significance in differences in both proteins between conditions, a Student's t-test was performed on the normalized protein signal values.

2.1.3 Molecular biology

Uncapped RNA degradation assay

This assay was based on the protocol published by Chiron and Jais (Chiron & Jais, 2017). One kilobase DNA templates for *in vitro* transcription was prepared as outlined in the riboprobe synthesis section. Next the T7 mScript™ Standard mRNA Production System (Biozym, Cat n°: C-MSC11610 and C-MSC100625) was used to generate capped and uncapped ssRNA for the degradation reaction. One microgram of template DNA was combined with 2 µl of 10X mScript T7 Transcription Buffer, 7,2 µl NTP solution, 2 µl 100 mM DTT, 0,5 µl ScriptGuard RNase Inhibitor and topped with RNase-free water. Next 2 µl of mScript T7 Enzyme Solution was added to the reaction and the sample was incubated for 30 min at 37°C. Next, 1 µl of RNAse-free DNase I was added and the sample was again incubated at 37 °C for 15 min. The ssRNA product was next purified by first bringing the reaction to 200 µl of RNase-free water and adding next 200 µl of Phenol:chloroform:Isoamyl Alcohol (PCI) (Thermo Scientific, Cat n°: 327115000). The solution was vortexed for 10 seconds and spun down at 16000g for 5 min on a table-top centrifuge. The aqueous phase was transferred to a DNA LoBind eppendorf tube (Eppendorf, Cat n°: 0030108051), supplemented with 200 µl of 5 M ammonium acetate and mixed thoroughly by pipetting up and down. The reaction was left to incubate 15 min on ice before a centrifugation step of 15 min at 16000g and 4°C. The supernatant was removed next and the pellet washed with 70% ethanol. Afterwards, the ethanol was removed and the pellet left to air-dry for 5 min. Finally, the pellet was resuspended in

50 µl of RNase-free water. For capping, 60 µg of the ssRNA was taken and the volume adjusted to 72 µl of RNase-free water. The RNA was then heat-denatured for 10 min at 65°C and transferred on ice. In a separate tube, the capping 'cocktail' was prepared by mixing 10 µl of 10X ScriptCap Capping Buffer with 5 µl 20 mM GTP, 2,5 µl 20 mM SAM, 2,5 µl ScriptGuard RNase Inhibitor and 4 µl RNase-free water. Next the capping 'cocktail' was added to the denatured RNA and the whole was supplemented with 4 µl of ScriptCap Capping Enzyme (10 U/µl). The sample was left to incubate 30 min at 37°C next. Finally, the capped RNA was purified as explained above and resuspended in 50 µl of RNAse-free water.

For the degradation assay, about 6 µg of capped and uncapped RNA was used as initial input in a volume of 17 µl of RNase-free water. To this, 2 µl of RNA 5' Polyphosphatase 10x reaction buffer and 1 µl of RNA 5' Polyphosphatase (Lucigen, Art n°: 136120) was added and the reaction was incubated for 30 min at 37°C. To proceed with precipitation, the reaction volume was brought to 100 µl by addition of TE buffer (10 mM Tris·Cl, pH 8.0. 1 mM EDTA) and supplemented with 100 µl of PCI. The samples were vortexed and spun down for 5 min at 4°C with 16000g. The aqueous phase was transferred to a new eppendorf and supplemented with 100 µl of chloroform (Merck, Cat n°: 1.02445.2500) before being vortexed and spun down again for 5 min at 4°C with 16000g. The aqueous phase was again transferred to a new eppendorf and precipitated using the ammonium acetate/ethanol precipitation method (Osterburg, Allen, & Finch, 1975). Sixty-six microliters of 7,5 M NH4OAc (Sigma-Aldrich, Cat n°: A1542-250G), 100 µl of ice-cold ethanol and 1 µl of GlycoBlue coprecipitant (Invitrogen, Cat n°: AM9515) was added to the samples and incubated for 30 min at -20°C. Following this, a 20 min centrifugation step at 4°C with 16000g was performed. The supernatant was next removed and washed with 70% ethanol and spun down again for 5 min at 4°C with 16000 g. Finally, the ethanol was removed, the pellet air-dried for and resuspended in 17 µl of RNase-free water. Half of the volume was set aside and put on ice and the rest was used for the second enzymatic reaction.

The 8,5 µl of capped and uncapped RNA were supplemented with 8,5 µl of RNase-free water, 2 µl of Terminator Exonuclease buffer A and 1 µl of Terminator 5'Phosphate-Dependent exonuclease (Biozym, Cat n°: TER51020). Afterwards, reactions were incubated for 1 hour at 30°C. Following this, both the samples that had been subjected to the terminator exonuclease reaction and the samples that weren't were re-precipitated as explained before to account for potential losses during RNA extraction and precipitation. All samples were then resuspended in 20 µl of RNase-free water. Finally, 1 µl of each sample was ran on a 1% agarose-TBE gel.

Decapping assay

Capped and uncapped RNA were synthesized as mentioned in the uncapped RNA degradation assay section. About 6 µg of capped RNA was used as input in 17 µl of RNase-free water. Next, the de-capping reaction was performed by adding 2 µl of Cap-Clip™ Acid Pyrophosphatase (Biozym, Cat n°: 187005) 10x buffer and 0,5 µl of Cap-Clip enzyme. Samples were incubated for 1 hour at 37°C before the RNA was purified by phenol chloroform extraction and precipitation as mentioned above. RNA was then resuspended in 17 µl. Half of the volume was set aside and put on ice and the rest was used for the second enzymatic reaction. The RNA used for the second reaction was supplemented with 8,5 µl of RNAse-free water before adding 2 µl of Terminator Exonuclease buffer A and 1 µl of Terminator 5'Phosphate-Dependent exonuclease. The reaction was then incubated for 1 hour at 30°C. Both the RNA that was set on the side and the RNA subjected to the nuclease treatment were then re-precipitated as mentioned before and resuspended in a final volume of 20 µl of RNAse-free water. Finally, 1 µl of each sample was ran on a 1% agarose-TBE gel.

Total RNA isolation

Two to three 7mm worms were placed in an eppendorf and planarian water was removed. One milliliter of TRI-reagent (Thermo Scientific, Cat n°: AM9738) (Chomczynski & Sacchi, 1987) was added to the eppendorf followed by 3-4 2,3mm metallic beads (BSP, Cat n°: 11079123ss). Samples were placed in a tissue homogenizer for 2 minutes at a frequency of 30Hz at 4°C. Following homogenization, the lysate was transferred in a new eppendorf and 100µl of 1-Bromo-3-Chloropropane (Merck, Cat n° B9673-200ML) was added. Samples were vortexed (Scientific Industries, SKU: SI-0236) for 10 seconds and centrifuged on a table-top centrifuge (Eppendorf, Cat n°: 5406000712) at 4°C for 20 minutes at 16000g. The aqueous phase was transferred to a new eppendorf on ice and 250µl of isopropanol (Merck, Cat n°: 1.09634.2511) and high-salt precipitation solution (0.8 M Sodium-citrate (Sigma-Aldrich, Cat n°: W302600-1KG-K), 1.2 M NaCl) was added. Samples were vortexed, spun down and transferred to a Zymo-Spin IIICG Columns (Zymo, Cat n°: C1006-50-G). Samples were washed once with RNA-wash buffer (Zymo, Cat n°: C1006-50-G) and contaminating DNA was digested by incubating the column in a DNA digestion buffer composed of DNA digestion buffer and DNAse I (0,6 U/µl) (Zymo, Cat n°: E1010) for 15 minutes at room temperature. One wash in RNA pre-wash buffer (Zymo, Cat n°: R1020-2-100) followed by two washes in RNA wash buffer were then performed and samples were eluted in 50µl of nuclease-free water (Thermo Scientific, Cat n°: AM9932). RNA concentration was measured by spectrophotometry and integrity was assessed by microcapillary electrophoresis (Agilent, Cat n°: G2939BA) or regular agarose electrophoresis (Sambrook, Fritsch, & Maniatis, 1989).

cDNA and first strand synthesis

To clone ORFs of interest, total RNA was used to generate a library of cDNA (Lassner, 1995) using the ProtoScript® II First Strand cDNA Synthesis Kit (NEB, Cat n°: E6560S). Two microliters of 50µM oligo-dT primer were added to 1µg of total RNA having passed the integrity check and brought to a total of 8µl using nuclease-free water. The mix was heat-denatured for 5 minutes at 65°C in a thermal cycler and placed directly on ice afterwards. Ten microliters of ProtoScript II Reaction Mix and 2µl of ProtoScript II Enzyme Mix was added to the reaction to bring the volume up to a total of 20 µl and incubated for 1 hour at 42°C followed by 5 minutes at 80°C to inactivate the enzymatic reaction. The final cDNA library was transferred to a 1,5 ml DNA LoBind eppendorf tube and stored at -20°C.

T4PCR and DNA purification

Primers were designed in Geneious Prime® (Version 2019.2.3) with a size between 16 and 29 nt, a melting temperature between 52,9°C and 65,3°C and a GC% between 20% and 80% with optima at 20nt, 60°C and 50% respectively. Forward and reverse primers were designed with a Gibson-homology compatible overhangs (called T4P overhangs) with sequences 'CCAATTCTACCCGCACAGTC' and 'CCAATTCTACCCGCACAGTC' respectively (Gibson et al., 2009). A list of primers can be found in Table 2.1. PCR reactions were set up on ice (1x Q5 Reaction buffer, 200µM dNTPs (Sigma-Aldrich, Cat n°: D7295), 0,5µM forward primer, 0,5µM reverse primer, 2µl cDNA, 0,02 U/µl Q5 DNA polymerase (NEB; Cat n°: M0491S)) and Touchdown PCR (Don, Cox, Wainwright, Baker, & Mattick, 1991) was performed with the following program: Initial denaturation at 95°C for 30 seconds, 10 touchdown cycles composed of an initial denaturation step of 10 seconds at 95°C, an annealing step of 30 seconds starting at 8°C above the melting temperature of the primer with the lowest melting temperature and decreasing by 1°C during each cycle and an elongation step of 30 seconds/kilobase of expected product at 72°C . Following the

Touchdown cycles comes 30 regular PCR cycles composed of a denaturation step of 10 seconds at 95°C, an annealing step at 2°C below the melting temperature of the primer with the lowest melting temperature and an elongation step of 30 seconds/kilobase of expected product at 72°C. Finally comes a final elongation step of 5 minutes at 72°C and a hold step at 10°C. Next, 5 µl of the PCR reaction were mixed with 1 µl of self-made 6X loading dye (10 mM Tris-NaCl pH 7,6, Glycerol 60% (v/v), Cresol-red 0,02% (w/v) (Avantor, Cat n°: 0500-5G), EDTA 60 mM (Sigma-Aldrich, Cat n°: E5134-500G), Tartrazine 0,12% (w/v) (Alfa Aesar, Cat n°: A17682)) and ran for 40 minutes at 80V on a 1% Agarose (Carl Roth, Cat n°: 3810.3) TAE (40 mM Tris base (Sigma-Aldrich, Cat n°: T1503-1KG), 20 mM acetic acid (Merck, Cat n°: 1.00063.2511), 1 mM EDTA) gel prestained with SYBR Safe (Thermo Scientific, Cat n°: S33102) at a 1:10000 dilution. Then, a clean-up of the PCR product was performed using the QIAquick PCR purification kit (Qiagen, Cat n°: 28106) using QIAquick spin columns and following the manufactures protocol with the exceptions of the following steps. Ten microliters of 3M sodium acetate pH 5,5 (Sigma-Aldrich, Cat n°: S2889-1KG) were always added to the clean-up reaction after addition of the binding buffer (PB) and without the use of the pH indicator dye. Two washes with an incubation time of one minute with the wash buffer (PE) were done instead of one wash without incubation time. The final elution volume was 15 µl of EB buffer instead of 50 µl. Finally, the concentration of the PCR product was measured by spectrophotometry.

Table 2.1: Primers and plasmids used in this study

Primer name	Sequence	Source
AA18	CCACCGGTTCCATGGCTAGC	Adler and Alvarado (2018)
PR244	GAGGCCCCAAGGGGTTATGTG	Adler and Alvarado (2018)
T7 AA18	GAAATTAATACGACTCACTATAGGGAGCCACCGGTTCCATGGCTAGC	Adler and Alvarado (2018)
CEBP1_F	CATTACCATCCCGCACTATGAACCTCCAATGGTTCCGCTT	This study
CEBP1_R	CCAATTCTACCCGCACAGTCTCTGCAACTACGCGGTTCAT	This study
CEBP2_F	CATTACCATCCCGCACTATGACATCGAAAGCGTGGAAGTTG	This study
CEBP2_R	CCAATTCTACCCGCACAGTCACATCGCTTCATCAGTGTTGAC	This study
CEBP3_F	CATTACCATCCCGCACTATGAGCCAATGAAGATTATGCAAAGCT	This study
CEBP3_R	CCAATTCTACCCGCACAGTCACATTGAATTATGTGCGAAATTTTGT	This study
CEBP4_F	CATTACCATCCCGCACTATGTGCATAGACCATAGTCCAGT	This study
CEBP4_R	CCAATTCTACCCGCACAGTCACAACTGTGAAAACGTTGAGA	This study
GATA1_F	CATTACCATCCCGCACTATGAGCACTGTATCGAAAGCCTTCT	This study
GATA1_R	CCAATTCTACCCGCACAGTCTGCCTTCTGAAGCCATTCCA	This study
GATA2_F	CATTACCATCCCGCACTATGCGTGAGACGATTGAAATGCCG	This study
GATA2_R	CCAATTCTACCCGCACAGTCACATGATACCATCCAATTCGCA	This study
SNAIL1_F	CATTACCATCCCGCACTATGTGCGATGAGACTGTACCAGC	This study
SNAIL1_R	CCAATTCTACCCGCACAGTCTACTCCGTGATCCCTGCTCA	This study
SNAIL2_F	CATTACCATCCCGCACTATGTGTGCCCGTTTTGCAAAGAT	This study
SNAIL2_R	CCAATTCTACCCGCACAGTCCGTAAGTCCACGAAACTCCAGA	This study
SNAIL3_F	CATTACCATCCCGCACTATGCCCAGCAAATACCGCTACCA	This study
SNAIL3_R	CCAATTCTACCCGCACAGTCCGGGTGTTCTCGACAGACAA	This study
SNAIL4_F	CATTACCATCCCGCACTATGATACGGGGCTCCAAACGTTT	This study
SNAIL4_R	CCAATTCTACCCGCACAGTCTTACCAGGCGAGCCCTTTTT	This study
SNAIL5_F	CATTACCATCCCGCACTATGTCAAATCACAGGGTCGCTGC	This study
SNAIL5_R	CCAATTCTACCCGCACAGTCTGATTCTGGTCGCATCGGAG	This study
TEAD1_F	CATTACCATCCCGCACTATGGCAAACCGGCAAAAGTCGAT	This study
TEAD1_R	CCAATTCTACCCGCACAGTCCTCCAACACTTGACGACCGA	This study
TEAD2_F	CATTACCATCCCGCACTATGTACCGGTAAAACGCGAACCA	This study
TEAD2_R	CCAATTCTACCCGCACAGTCCCCCCGTTTCCACTATCTCG	This study
THAP_F	CATTACCATCCCGCACTATGCCAATTATTATTCGATTACCATTCTCA	This study
THAP_R	CCAATTCTACCCGCACAGTCGCTTTTTCCCGACTTCCTTGT	This study
NFK1_F	CATTACCATCCCGCACTATGGACTGAAGGAAGTCGTGGGG	This study
NFK1_R	CCAATTCTACCCGCACAGTCGCATTCCGATTGGTCGCTTC	This study
NFK2_F	CATTACCATCCCGCACTATGGCAGTGGACATTCAGCGTTG	This study
NFK2_R	CCAATTCTACCCGCACAGTCGTCTCACCAGCTGCTGAAGT	This study
NFK3_F	CATTACCATCCCGCACTATGTTGTGCCCCAACCAAGAGAG	This study
NFK3_R	CCAATTCTACCCGCACAGTCGTTTCCATTCGCCCATCAGC	This study
NFK4_F	CATTACCATCCCGCACTATGCCTGAAAATATTGGTGGTGGAACA	This study
NFK4_R	CCAATTCTACCCGCACAGTCTCCGTTAGGCTTTGCTTGCT	This study
TSP66e_F	CATTACCATCCCGCACTATGTGTGTTTCGGTGAATCAATCTCG	Rouhana et al. (2017)
TSP66e_R	CCAATTCTACCCGCACAGTCAGCGAACACGATGCAGAGAA	Rouhana et al. (2017)
GWIN_F	CATTACCATCCCGCACTATGCTCCGCTGATCAATCACCGA	U. W. Khan and Newmark (2022)
GWIN_R	CCAATTCTACCCGCACAGTCAACAAATTTATTCAACATTAAAGTGTTAC	U. W. Khan and Newmark (2022)
PLASTIN_F	CATTACCATCCCGCACTATGAAGCGGTCTCAGACACATGG	Chong et al. (2011)
PLASTIN_R	CCAATTCTACCCGCACAGTCTGAACGATCCAGACAACCGG	Chong et al. (2011)
TSP1	Published in Vila-Farré et al. (2023)	Vila-Farré et al. (2023)
FERRITIN2	Published in Vila-Farré et al. (2023)	Vila-Farré et al. (2023)

Gibson assembly and transformation

The pPRT4P vector backbone was linearized by a digestion reaction using the SmaI restriction enzyme (Adler & Alvarado, 2018) (1x CutSmart® Buffer (NEB, Cat. No.: B7204S), 0,8 U/µl SmaI (NEB, Cat No.: R0141S), 2 µg of pPRT4P plasmid, total volume 50 µl) for 3 hours at room temperature. The reaction was then purified using the QIAquick PCR purification kit and linearization efficiency was assessed by agarose gel electrophoresis. The Gibson assembly reaction was prepared next by combining 0,05 pmol of linearized vector backbone with 0,1 pmol of purified PCR product containing the T4P overhangs in a total of 6 µl. To this, 6 µl of 2x NEBuilder Hifi DNA Assembly Master Mix (NEB, Cat.No.: E2621) was added. The reaction was incubated for 15 minutes at 50°C and stored at -20°C. An aliquot of DH5 alpha competent E. coli (NEB, Cat n°: C2987H) was thawed on ice for 10 minutes before adding the Gibson assembly reaction mix to the cells and incubate it on ice for 20 minutes. Following incubation, cells were subjected to a heat shock by placing the eppendorf in a 42°C water bath for 30 seconds before putting it back on ice for 2 minutes (Inoue, Nojima, & Okayama, 1990). Following the heat shock, 500µl of LB medium was added and the cells were left to recuperate for 45 minutes at 37°C with agitation. Finally, 50 − 100 µl of the solution was plated on LB agar supplemented with Kanamycin (5 mg/ml, Sigma-Aldrich, Cat n°: K4000-5g) with the help of glass beads. The plates were then placed at 37°C overnight to grow.

Colony PCR and sequencing

Plates of transformed bacteria were tested for proper integration of the sequence of interest into the chosen vector by colony PCR (Bergkessel & Guthrie, 2013). A PCR master mix was set up (1x Q5 Reaction buffer, 200µM dNTPs (Sigma-Aldrich, Cat n°: D7295), 0,5µM forward primer (AA18), 0,5µM reverse primer (PR244), 0,02 U/µl Q5 DNA polymerase (NEB; Cat n°: M0491S))(Adler & Alvarado, 2018) and aliquoted per 20 µl in PCR strips (Sarstedt, Cat n°: 72.991.002) on ice. Individual colonies were picked with the help of a pipette tip, stroked on a new LB agar plate with kanamycin and dipped in the corresponding PCR tube. After this, the PCR strips were placed in a thermal cycler and PCR was performed using the following program. First, an initial denaturation step of 5 minutes at 95°C was performed. Following this, 25 cycles of PCR amplification were done (Denaturation: 10 seconds at 95°C, annealing: 30 seconds at 68°C and elongation: 30 seconds/ kilobase of expected product at 72°C). A final elongation step of 5 minutes at 72°C and a hold step at 10°C followed to finish the PCR.

Four microliters of home-made 6X loading dye were added to each PCR tube and the results were analyzed by agarose gel electrophoresis. One or more positive clone for each construct was put in liquid culture and grown at 37°C over-night. Plasmids were extracted the next day by alkaline lysis with SDS (Sambrook & Russell, 2006) using the QIAprep Spin Miniprep Kit (QIAgen, Cat n°: 27104) according to the manufacturer's instruction with the following modifications. After the addition of the neutralizing N3 buffer and the subsequent centrifugation step, the supernatant was passed twice through the column instead of once. Furthermore, like in the PCR purification procedure, two washes with an incubation time of one minute with the wash buffer (PE) were done instead of one wash without incubation. The eluted plasmid concentrations were measured by spectrophotometry and sent to sanger sequencing (Mycrosynth seqlab, single tube sequencing) with the forward (AA18) primer premixed as per the required specifications. Sequencing results were analyzed in Geneious by mapping the resulting electropherogram to the in-silico constructed plasmid.

2.1.4 Whole-mount RNA *in situ* hybridization

Riboprobe synthesis

Riboprobe template DNA was generated by setting up a PCR reaction (1x Q5 Reaction buffer, 200µM dNTPs (Sigma-Aldrich, Cat n°: D7295), 0,5µM forward primer (AA18), 0,5µM reverse primer (PR244), 0,02 U/µl Q5 DNA polymerase (NEB; Cat n°: M0491S)) (Adler & Alvarado, 2018) using the following program. First, an initial denaturation step of 30 seconds at 95°C was performed. Following this, 30 cycles of PCR amplification were done (Denaturation: 10 seconds at 95°C, annealing: 30 seconds at 68°C and elongation: 30 seconds/ kilobase of expected product at 72°C). A final elongation step of 5 minutes at 72°C and a hold step at 10°C followed to finish the PCR. The product was then run on a 1% Agarose TAE gel to assess proper amplification and purified using the QIAquick PCR purification (see "Gibson assembly and transformation" section) and samples were stored at -20°C. Next, an *in vitro* transcription reaction was set up to generate labelled riboprobes for the ORFs of interest (King & Newmark, 2018). In a tube were mixed 1 µg of DNA template, 2 µl of 10x Transcription buffer (TriLink Biotechnologies, 400 mM Tris-NaCl pH8, 100 mM DTT, 20 mM Spermidine (Sigma-Aldrich, Cat n°: S2626-5G), 0,2% Triton X-100 (Sigma-Aldrich, Cat n°: T8787-50ML), 165 mM MgAc (fisher scientific, Cat n°: 15637920)), 2 µl of 10x RNA labelling mix DIG or Fluresceine (Sigma-Aldrich, Cat n°: 11277073910 or 11685619910 respectively) 0,5 µl of RiboLock RNAse inhibitor

(Thermo Scientific, Cat n°: 24X2500 U), 0,3 µl of inorganic pyrophosphatase (Thermo Scientific, Cat n°: EF0221), and 2 µl of T7 RNA polymerase (Thermo Scientific, Cat n°: EP0111) in a volume of 20 µl. The reaction was left at 37°C overnight to incubate.

The next day, 0,5 µl of DNAse I (NEB, Cat n°: M0303S) was added to the reaction and the sample was incubated for another 45 minutes at 37°C to degrade the DNA template. Samples then were brought to a volume of 100 µl with nuclease-free water before precipitation. Half a volume of 7,5 M NH4OAc was added to the sample followed by two volumes of ice-cold 100% Ethanol (Sigma-Aldrich, Cat n°: 1.00983.1011). Samples were spun for 30 minutes at 4°C at 21000g on a bench-top centrifuge followed by two wash-spin cycles in 70% EtOH. The supernatant was then removed and the pellet was left to air-dry for about 5 minutes. The pellet was then finally resuspended in 100µl of deionized formamide (AppliChem, Cat n°: A2156,0500). One microliter of product was mixed with 5 µl of formamide and 1 µl of self-made 6x loading dye and ran for 40 minutes at 80V on a 1% Agarose TTE (89 mM Tris base, 28.5 mM Taurine (Roth, Cat n°: 4721.2), 0.05 mM EDTA) (Ganguly, Rock, & Prockop, 1993) gel prestained with SYBR Safe. Riboprobes were stored at -70°C.

Large sexual *Schmidtea mediterranea* fixation

This protocol was adapted from two seminal publications in the planarian field authored by King and Newmark (King & Newmark, 2013) and Pearson and colleagues (Pearson et al., 2009). Worms were processed by batches of 30 animals with a size between 1 cm to 1,5 cm. Animals were picked from stationary cultures, washed once with fresh Planarian water (PW) and placed a 50 ml falcon tube (Sarstedt, Cat n°: 62.547.254). Planarian water was replaced with a buffered 0,5% NAC solution and incubated for 10 minutes on a rocker in order to strip the external mucus layer. The animals were then killed by replacing the stripping solution with a concentrated acidic NAC solution (7,5% (w/v) NAC, 1x PBS, 0,0005% Tween 20) for 10 minutes with agitation. Worms were then fixed by rinsing them first once in 4% formaldehyde solution (4% (v/v) formaldehyde (EMS, Cat n°: 15710), 0,5x PBS, 0,15% Triton X-100) followed by a 110 minutes incubation time with occasional rocking every 10 minutes. The fixation solution was then removed and replaced by a quenching solution (0,125M Glycine (Merck, Cat n°: 1.04201.1000), 1x PBS, 0,3% Triton X-100). Worms were then washed twice 10 minutes in a PBStx solution (1x PBS, 0,3% Triton-X 100). Following this, PBStx was replaced with pre-warmed reduction solution (1% (w/v) Igepal CA-630 (Sigma-Aldrich, Cat n°: I8896-100ml), 0,5% (w/v) SDS, 50 mM

DTT, 1x PBS) and incubated for 15 minutes at 37°C with occasional swirling. Worms were then washed twice 10 minutes in PBStx with agitation before incubation in 50% methanol (50% methanol (Merck, Cat n°: 1.06009.2511), 50% PBStx) for 10 minutes with agitation. Finally, two washes of 10 minutes in 100% methanol were performed before placing the worms at -20°C for long-term storage.

Bleaching and riboprobe hybridization

Samples were allowed to equilibrate to room temperature and worms were distributed in 12 well plates (Thermo Scientific, Cat n°: 150200) to have between 5 to 10 worms/condition assayed. Methanol was replaced with a 50% methanol-PBStx solution and incubated for 10 minutes with agitation. Worms were then completely rehydrated by washing them once in PBStx for 10 minutes. PBStx was exchanged for 1X SSC solution (1X SSC (Sigma-Aldrich, Cat n°, 1.06009.2511), 0,1% Tween 20) and the worms were incubated for 10 minutes with agitation. The SSC solution was removed and replaced with a bleaching solution (5% formamide, 1,2% H2O2 (Sigma-Aldrich, Cat n°: H1009-5ML), 0,5x SSC) and incubated under direct light for 2 - 2,5 hours. The bleaching solution was replaced once during the incubation to increase bleaching efficiency.

Following this, Worms were rinse once for 10 minutes in 1X SSC followed by two 10 minutes washes in PBStx. A digestion step followed after by incubating the worms in a proteinase K solution (0,1% SDS, 2µg/ml Proteinase K (NEB, Cat n°: P8107S), 1X PBS) for 30 minutes with agitation. The worms were then fixed by incubating them 10 minutes in a 4% formaldehyde PBStx solution. The fixation solution was replaced by a 1:1 PBStx:PreHybe (50% formamide, 5X SSC, 1X Denhardts (Invitrogen, Cat n°: 750018), 100 µg/µl Heparin (Sigma, Cat n°: H3393-1MU), 1% Tween 20, 0,25 mg/ml Torula yeast RNA (Sigma-Aldrich, Cat n°: R6625-25G), 50 mM DTT) solution and incubated for 10 minutes with agitation. The solution was replaced afterwards with PreHybe and incubated for 2 hours at 58°C with agitation. Probes were diluted 1:2000 in Hybe buffer (50% formamide, 5X SSC, 1X Denhardts, 100 µg/µl Heparin, 1% Tween 20, 0,25 mg/ml yeast RNA (Roche, Cat n°: 10109223001), 50 mM DTT,5% Dextran sulfate (Sigma-Aldrich, Cat n°: D8906-100G)) and heat denatured for 3 minutes at 70°C before being put at 58°C. After the 2-hour incubation time, PreHybe was replaced with the riboprobe mix and the samples were incubated for 16 hours at 58°C with agitation. To avoid evaporation, empty wells were filled with ddH2O and the plates were sealed with parafilm and enveloped in aluminium foil. Samples were subjected to a series of washes the following day. First, two 30-minute washes at 58°C with agitation were done with WashHybe (50% formamide, 5X SSC, 1X Denhardts, 1% Tween 20). Next, two 30-minute washes at 58°C with agitation were done with a 1:1 ratio of WashHybe:2xSSC solution (2x SSC, 0,1% Tween 20). Following these washes came 3 times 3 30-minute washes at 58°C with agitation of 2X SSX solution, 0,2x SSC solution and 0,05X SSC solution respectively. Finally, two washes at room temperature were performed using a TNTx buffer (100 mM Tris-NaCl pH 7,5, 150 mM NaCl (VWR, Cat n°: 0241-5KG), 0,3% (v/v) Triton X-100).

Colorimetric NBT/BCIP development

Samples were blocked for 1 hour at room temperature with agitation in a blocking solution (5% horse serum (Sigma-Aldrich, Cat n°: H1270-100ML), 0,5% Roche western blocking solution (Roche, Cat n°: 11921673001), PBStx). Samples were then incubated overnight with rocking at 4°C with an anti-DIG-AP antibody (Roche, Cat n°: 11093274910, Batch n°: 32871921) diluted 1:3000 in blocking solution. The next day, samples were subjected to 6 20-minute washes in TNTx at room temperature. Following the washes, samples were incubated 10 minutes in AP buffer (100 mM Tris-NaCl pH 9,5, 100 mM NaCl, 50 mM MgCl2 (Supelco, Cat n°: 1.05833.1000), 1% Tween 20). AP buffer was removed and worms were incubated in EQ buffer (100 mM Tris-NaCl pH 9,5, 100 mM NaCl, 50 mM MgCl2, 1% Tween 20, 5% PVA (Sigma-Aldrich, Cat n°: P8136-250G)) for 10 minutes. Samples were next placed in the DEV buffer (100 mM Tris-NaCl pH 9,5, 100 mM NaCl, 50 mM MgCl2, 1% Tween 20, 7,8% PVA, 266 µg/ml NBT (Roche, Cat n°: 11383213001), 266 µg/ml BCIP (Roche, Cat n°: 11383221001)) and observed every 30 minutes until a clear signal was visible. The reaction was then stopped by rinsing the worms 3 times in PBStx followed by an overnight incubation in PBStx at 4°C with agitation. Worms were fixed the next day in 4% formaldehyde solution for 45 minutes at room temperature and rinsed once for 10 minutes in PBStx. PBStx was replaced with 100% EtOH to clear the sample and exchanged with fresh ethanol every 10 minutes until it stayed clear. Worms were then washed once for 5 minutes in a 1:1 solution of EtOH:PBStx followed by a wash in PBStx until the worms were completely rehydrated. Two additional PBStx washes were then performed before proceeding to sample mounting in Scale 2A (2M Urea, 75% Glycerol). Slides were imaged on a Zeiss Stemi 508 equipped with an Axiocam 208 Color camera.

Fluorescent development and nuclear co-stain

Samples were blocked for 2 hours at room temperature with agitation in a blocking solution (5% horse serum, 0,5% Roche western blocking solution, TNTx). Depending on the labelling method, samples were incubated overnight with an anti-DIG-POD (Roche, Cat n°: 11207733910) or anti-Fluoresceine-POD (Merck, Cat n°: 11426346910) antibody diluted in a blocking solution (5% horse serum, 0,5 or 1% Roche western blocking solution respectively, TNTx) for 48 hours at 4°C with agitation. Next, samples were washed 6 times 20 minutes in TNTx before proceeding with the tyramide amplification procedure. TNTx was removed and a tyramide solution was diluted 1:2000 in TSA buffer (2M NaCl, 0,1M Boric acid (Merck, Cat n°: 1.00165.5000) pH 8,5, 0,02% (v/v) H2O2, 20 µg/µl 4-IPBA) and incubated with the samples for 45 minutes at room temperature with agitation. Samples were then washed 6 times for 20 minutes in TNTx before incubating them in a DAPI staining solution (2,5µg/ml DAPI (Thermo Scientific, Cat n°: 62247), TNTx) overnight at 4°C. Samples were rinsed twice in TNTx before being mounted between two coverslips in ScaleS4 (10% (v/v) Glycerol, 15% (v/v) DMSO (Sigma-Aldrich, Cat n°: 276855-100ML), 40% (w/v) Sorbitol (Sigma-Aldrich, Cat n°: S1876-1KG, 4M Urea, 0,1% Triton X-100)).

Whole-mount *in situ* imaging and image processing

Fluorescent images for the characterization of the candidate transcription factors' expression were acquired on an Olympus IX83 spinning disc system with a Yokogawa CSUW1-T2S scan head. The stand was equipped with Hamamatsu Orca Flash 4.0 v3 monochrome sCMOS cameras for image acquisition. An Olympus UPLXAPO20X (NA 0.8 WD 0.6mm) air objective was used to perform imaging. For DAPI imaging, a 405 nm laser in combination with a 477/50 bandpass emission filter was used. Rhodamine was excited with a 561 nm laser and emission was collected with a 617/73 bandpass filter. Imaging data was imported as a hyperstack in FIJI (Schindelin et al., 2012) using the Bio-Formats Importer plugin. Maximum Z projection of selected stacks was then performed to visualize the representative expression pattern of the tested genes.

Images for the characterization of the expression pattern of sexual markers in RNAi animals were acquired using an Olympus VS200 widefield slide scanner equipped with a monochrome Hamamatsu Orca Fusion camera. An Olympus UPLFLN4X (NA 0.13 WD 17mm) air objective was used to perform the imaging.

For DAPI imaging, an excitation band of 378/26 nm was used in combination with a 434/18 bandpass emission filter. Rhodamine was excited with an 554/12 excitation band and emission was collected with a 685/20 bandpass filter. Worms stained for the same marker were imaged using the same exposure time determined on the negative control (egfp). Extended focus imaging was used for image acquisition. Images were then imported into Fiji and worms stained for the same marker were set to the same brightness and contrast to allow comparison between RNAi condition.

2.1.5 RNA interference

dsRNA synthesis and purification

DNA template was generated by setting up a PCR reaction similar to the PCR described in the “riboprobe synthesis” section with the following differences. The primer pair is composed of the forward T7AA18 and the reverse PR244 primers and 31 cycles of amplification were done instead of 30 cycles. Next, an *in vitro* transcription reaction was set up to generate dsRNA targeting the genes of interest (Rouhana et al., 2013). In a tube were mixed order: 124 µl of nuclease-free water, 120 µl of 25 mM rNTPS (Thermo Scientific, Cat n°: R1481), 40 µl of 10x Transcription buffer, 10 µl of RiboLock RNAse inhibitor, 6 µl of inorganic pyrophosphatase, 68 µl of T7 RNA polymerase and 32 µl of DNA template at a concentration of 250 ng/µl. Samples were homogenized using a wide-bore p200 pipet tip and divided in two DNA LoBind 1,5 ml eppendorf tube. The reaction was left at 37°C overnight to incubate. The next day, samples were heated for 3 minutes at 95°C on a heat block with agitation at 900 rpm. The heating block was turned off and the samples were let to cool slowly until they reached room temperature. Afterwards, one volume of PEG-NaCl precipitation solution (2,5 M NaCl, 20% (w/v) PEG-8000 (Roth, Cat n°: BP233-1), 10 mM Tris-NaCl pH 8) was added to the reaction and the tube was vortexed for twice for 5 seconds. dsRNA was then precipitated by centrifugation on a table-top centrifuge at 4°C for 20 minutes at 16000g. The supernatant was then discarded by aspiration and the pellet was dislodged and rinsed in 70% ethanol, incubated 1 minute on ice followed by centrifugation at 4°C for 5 minutes at 16000g. The supernatant was aspirated and pellets corresponding to the same gene of interest were combined in a single eppendorf by placing both openings of the eppendorf against each other and hitting the eppendorfs against the bench until one pellet fell in the other eppendorf. The combined pellets were then washed once more in 70% ethanol, incubated 1 minute on ice and centrifuged down at 4°C for 5 minutes at 16000g. Finally, the pellet was air-dried for 5 minutes at 37°C before being resuspended in 400 µl of nuclease-free water. A 1:50 dilution in nuclease-free water was made from 1 µl of dsRNA solution for each sample and their concentration was measured by spectrophotometry using an extinction coefficient value of 45 (Nwokeoji, Kilby, Portwood, & Dickman, 2017). The integrity of the dsRNA was also assessed by electrophoresis using a 1% agarose TTE gel. Samples were then diluted to 4µg/µl and stored at -70°C.

RNAi food preparation and feeding protocol

Homogenized calf liver was thawed at room temperature and passed through a sieve to remove any liver chunks still present and stored in 5 ml eppendorf tubes (Eppendorf, Cat n°: 0030119401) at -20°C. Before preparing RNAi food, 5 ml aliquots of sieved liver were thawed, aliquoted per 80 µl in PCR strips, spun down and placed at -20°C to freeze. Diluted dsRNA was then thawed on ice and 80 µl of dsRNA was pipetted on top of the frozen liver and stored at -70°C. Before feeding, worms were transferred to a new petri dish with clean planarian water supplemented with antibiotics. RNAi food aliquots were taken out of the -70°C freezer and thawed on ice. The tip (± 0,5 cm) of p200 pipet tips were cut with the help of a scalpel and used to homogenize the RNAi food mixture upon thawing. dsRNA food was then aspirated using the same pipet tip while taking care to not aspirate any air bubbles and placed in a line at the bottom of the petri dish. Worms were left to feed for 2 hours in the dark before transferring them to a new petri dish with clean planarian water supplemented with antibiotics. Worms were washed again the next day. Worms were fed twice a week, on Mondays and Thursdays, until the required number of feedings had been performed.

High-salt edema suppression

Before animal fixation for whole-mount *in situ* hybridization of animals with edemas, planarian water was supplemented with Tropic Marin Sea Salt (Tropic Marin, Art n°: 10134) in 10 mM increments until reaching a final concentration of 75 mM (A. Y. Lin & Pearson, 2014). The supplementation of salts started when animals began showing edemas.

2.1.6 Start-seq protocol

This method was first described by Nechaev and colleagues (Nechaev et al., 2010). The protocol developed during this PhD was mainly inspired by the following references (R. A.-J. Chen et al., 2013; Corces et al., 2017; Duttke et al., 2019; Larke et al., 2021; Nechaev et al., 2010) and adapted to *Schmidtea mediterranea*.

Nuclei extraction

Tissue pulverization and storage

Each replicate is composed of two samples that will be pooled together after RNA extraction following nuclei isolation. This was done to reach a sufficient amount of RNA before proceeding with the Start-seq library preparation protocol. About 20 7-8 mm asexual CIW4 or 10 10-15 mm S2F2 worms were placed in a small petri dish with planarian water. Planarian water was removed and replaced with in 0,5% pH neutral NAC stripping solution for 10 minutes in the dark. Worms were then rinsed twice in ddH2O and placed in a TT05MXT tissue tube (Covaris, SKU: 520140). The tissuetube was closed with a 2 ml milliTUBE (Covaris, SKU: 520132) while taking care that air could still pass through the joint and dipped for 30 seconds in liquid nitrogen with the help of the tissueTUBE TT05 Insertion tool (Covaris, SKU: 500231). After 30 seconds, the tissuetube was placed in the CP02 cryoPREP Automated Dry Pulverizer (Covaris, SKU: 500001) and pre-crushed at the power level 1. The tissueTUBE was then placed back in liquid nitrogen for 30 seconds before being crushed once more with the cryoPREP at power level 3. Next, the sample was then placed back in liquid nitrogen for 30 seconds before being place in dry ice and stored at -70°C.

Tissue dissociation and gradient centrifugation

For the nuclei extraction, the two samples corresponding to one replicate were processed at the same time. All materials and buffers were prepared on ice before taking samples out of the freezer. A swinging bucket centrifuge (Eppendorf, Cat n°: 5804R) was put at 4°C, two 2 ml KIMBLE dounce homogenizers (Sigma-Aldrich, Cat n°: D8938-1SET) were washed first with deionized water, then with a 20% (w/v) SDS solution, rinsed with deionized water again and finally washed in 70% ethanol before being air dried and placed on ice. Ten milliliters of Homogenization buffer (HB) 1 (40 mM Sucrose (Sigma-Aldrich, Cat n°: S9378), 20 mM KCl (Merck, Cat n°: 1.04936.1000), 10 mM MgCl2 (Supelco, Cat n°: 1.05833.1000), 20 mM Hepes-KOH pH 7 (Sigma-Aldrich, Cat n°: RES6008H-A701X), 10 mM DTT, 0,05% Igepal CA-630, 0,5 mM spermidine (Sigma-Aldrich: Cat n°: S2626-5G), 0,25 mM spermine (Sigma-Aldrich, Cat n°: 85590-5g), 0,2 U/µl RiboLock RNAse inhibitor, 1x Halt Protease Inhibitor) and HB2 (40 mM Sucrose, 20 mM KCl, 10 mM MgCl2, 20 mM Hepes-KOH pH 7, 1 mM DTT, 0,05% Igepal CA-630, 0,5 mM spermidine, 0,25 mM spermine, 0,2 U/µl RiboLock RNAse inhibitor, 1x Halt Protease Inhibitor) were prepared with the exception that the RNAse and Protease inhibitors were not yet added for HB2. Next, 14,4 ml of 50% iodixanol (Sigma-Aldrich, Cat n°: D1556-250ml) solutions was prepared by mixing 60% iodixanol with a 6x osmolarity buffer (90 mM KCl, 12 mM MgCl2, 60 mM Hepes-KOH) in a 1:5 ratio supplemented with 7,2 µl of 1M spermidine, 36 µl of 100 mM spermine and 72 µl of RiboLock RNAse inhibitor. Following this, 4 ml of 40% and 12 ml of 30% iodixanol solutions were made by mixing the 50% iodixanol solution with HB1 in a 4:1 and 3:2 ratio respectively. The step density gradient was assembled in 15 ml falcons by layering 5 ml of 30% iodixanol solution above 1,4 ml 40% iodixanol solution on ice. Finally, the samples were taken out of the freezer and transported on the bench on dry ice.

The top of the tissueTUBE was cut with the help of a scalpel and the sample was placed on ice. The crushed worms were quickly resuspended in 2 times 900 µl HB1 with the help of a p1000 wide-bore pipette tip and transferred to the cold dounce homogenizer. The lysate was dounced 10 times with pestle B while taking care of not generating foam or cavitation bubbles. Afterwards, the lysate was filtered through a prewet 20 µm celltricks filter (Sysmex, Cat n°: 04-004-2325) into a 5 ml eppendorf containing 1,8 ml of 50% iodixanol. The second sample was processed in a similar fashion. the 25% iodixanol containing lysate was then carefully layered on top of the 30% iodixanol. Both samples were put in the swinging bucket centrifuge and spun at 4°C for 30 minutes at 3000g with low acceleration/deceleration.

At the end of the centrifugation step, HB2 was supplemented with RNAse and protease inhibitor and the samples were taken out of the centrifuge and placed on ice. The supernatant was removed until the air-liquid interface was close to the nuclei band situated between the 30 and 40% iodixanol layers. The nuclei were then carefully aspirated with the help of a p200 pipette in a total of 400 µl and transferred to a 5 ml eppendorf. The nuclei solution was then topped up to 4 ml with HB2 and homogenized slowly with a p1000 pipette by up-and-down. A 1:10 dilution was made (50 µl nuclei solution and 450 µl HB2), stained with DAPI (0,5µg/µl final concentration) and the nuclei amount and integrity were determined using a hemocytometer (Incyto, Cat n°: DHC-N01). Nuclei were spun down for 7 minutes, 900g at 4°C and resuspended in 1 ml of TRI-reagent for short-capped RNA extraction or hot urea lysis buffer for protein extraction (see nuclear protein extraction).

Short capped RNA library preparation

RNA extraction

After resuspension in TRI-reagent, the solution was incubated for 2 minutes at room temperature before being spun down 10 minutes at 4°C and 12000g to pellet any remaining debris. The supernatant was then transferred to a new 1,5 ml DNA LoBind eppendorf tube and stored at -80°C. To proceed with the protocol, samples were thawed on ice. Phase separation was achieved by addition of 100 µl of 1-Bromo-3-Chloropropane to the samples before thorough vortexing. Samples were then centrifuged for 15 minutes at 4°C and 18000g and the aqueous phase was transferred to a new 1,5 ml DNA LoBind eppendorf tube. Precipitation followed by the addition of 2,5 volumes of 100% ethanol, 0,1 volumes of 7,5M ammonium acetate and GlycoBlue Coprecipitant at a dilution of 1:20000 with a minimum of 1 µl. The solution was then incubated for 1 hour at -20°C and centrifuged for 20 minutes at 4°C and 21000g.

The supernatant was removed, the pellet was rinsed with 1ml of 70% ethanol and spun down for 5 minutes at 4°C and 21000g. The rinsing step was repeated, the supernatant was again removed and the pellet was left to dry at room temperature. Finally, 15 µl nuclease-free water was added to resuspend the pellet. Contaminating DNA was removed by adding 20 µl of DNAse I Buffer and 5 µl of DNAse I (6U/µl) and the solution was incubated for 15 minutes at room temperature. Nuclease-free water was used to top up the reaction to 200 µl and 200 µl of Phenol:chloroform:Isoamyl Alcohol (PCI) was added afterwards. The samples were vortexed thoroughly and spun down for 10 minutes at 4°C and 18000g to separate the phases. The aqueous phase was then extracted by pipetting and placed in a new 1,5 ml DNA LoBind eppendorf. A volume of 200 µl of chloroform was added to the aqueous phase and the sample was vortexed thoroughly before being spun down for 5 minutes at 4°C and 18000g. The aqueous phase was collected in a new eppendorf, RNA was precipitated as explained above and resuspended in 7,5 µl of nuclease-free water. RNA from nuclei isolation preps corresponding to the same replicates were pooled at this step to reach a volume of 15 µl and stored at -80°C.

Size selection and gel extraction

To proceed with size selection, a 15% acrylamide Urea-TBE gel (Thermo Scientific, Cat n°: EC6885BOX), was pre-run in 1x TBE (0,1M Tris base, 0,1M Boric acid, 2mM EDTA) for 30 minutes at 200V in an XCell SureLock™ Mini-Cell electrophoresis system after taking care of flushing the wells with TBE buffer. RNA samples were thawed on ice and 1 volume of 2x RNA loading dye (NEB, Cat n°: B0363S) was added before heat denaturation for 5 minutes at 75°C on a heating block. Samples were then immediately placed on ice. Wells of the gel were again flushed with TBE buffer and samples were loaded with care of leaving one empty lane between each of them. Low Range ssRNA Ladder (NEB, Cat n°: N0364S) was used for size referencing. The gel was run for about 50 minutes at 200V until the bromophenol blue dye reached the 3/4 of the gel. The gel was then cracked open and post stained in 1x TBE with 1x SYBR Gold (Life Technologies, Cat n°: S11494) for 30 minutes. Next, the gel was placed on a Safe Imager™ blue-light transilluminator (Invitrogen, Cat n°: G6600) and RNA from 20 nt until right below the 5s rRNA band was excised and placed in a 2 ml DNA LoBind eppendorf tube (Eppendorf, Cat n°: 0030108078).

The gel piece was then shredded with the help of a RNAse-free pestle (fisher scientific, Cat n°: 12-141-364) and a pellet mixer in 300 µl of GEB (0.4M NaOAc pH 5.5, 10mM Tris-Cl pH 7.5, 1 mM EDTA, 0.05 % Tween 20) (Duttke et al., 2019) supplemented with RiboLock RNAse inhibitor (1U/µl) and eluted for 2 hours under gentle agitation at room temperature. The mixture was then transferred to a spin filter column (Sigma-Aldrich, Cat n°: CLS8162)) placed in a 1,5 ml DNA LoBind eppendorf and spun for 2 minutes at 1000g. the eluate was then supplemented with 1,5 µl of Glycogen (Thermo Scientific, Cat n°: R0551) and 3 volumes of 100% ethanol before being placed at -80°C for 1 hour. RNA was then pelleted by centrifugation for 30 minutes at 4°C with 21000g. One wash with 75% ethanol was then performed before leaving the pellet to air-dry for 2-3 minutes and resuspend it in 10 µl of nuclease-free water. The RNA concentration was measured on the nanodrop and 12,25 µl of nuclease-free water to the remaining 9 µl (21,25 µl total) before transferring 20% (4,25 µl) in a new DNA LoBind eppendorf containing already 3,75 µl of nuclease-free water (8 µl total). All eppendorfs were stored on ice. The tube containing 20% of the small RNAs will be used to make the input libraries while the remaining 80% will serve to make the actual Start-seq library and will therefore be called input and Start-seq sample respectively.

Cap selection

RNA intended for Start-seq library preparation was subjected to cap selection by treatment with Terminator 5' Phosphate-Dependent Exonuclease. RNA was first heat-denatured for 2 minutes at 75°C and placed on ice for 2 minutes afterwards. To the 17 µl remaining was added 2 µl of 10x buffer A and 1 µl of Terminator exonuclease (Biozym, Cat n°: TER51020) before incubating the reaction for 1 hour at 30°C. The reaction was then topped up to

200 µl with nuclease-free water before adding one volume of PCI. The sample was then vortexed and spun down for 10 minutes at 4°C with 18000g. The aqueous phase was then transferred to a new eppendorf and 200 µl of chloroform was added to it. The sample was vortexed and spun down for 5 minutes at 4°C with 18000g. The Aqueous phase was again transferred to a new eppendorf and supplemented with 2,5 volumes of 100% ethanol, 0,1 volume of ammonium acetate, 1 µl of glycogen and precipitated for 1 hour at -20°C. RNA was pelleted by centrifugation at 4°C with 21000g for 20 minutes, the supernatant was removed and the pellet washed with 75% ethanol. The sample was then vortexed and spun down again for 5 minutes at 4°C with 21000g. The ethanol wash was repeated once before removing the supernatant and air-dry the pellet and the RNA was resuspended in 8 µl of nuclease-free water.

3' Adapter ligation

The library preparation kit used in this protocol is the NEBNext® Multiplex Small RNA Library Prep Set for Illumina® (NEB, Cat n°: E7300S). Both the input and Start-seq library concentrations were measured on the nanodrop using 2 µl of sample and the remaining 6 µl were transferred to a PCR tube. If the remaining total RNA amount was closer to 100 ng than 1µg, the 3' SR adaptor was diluted 1:2 before 1 µl was added to the sample. Otherwise, 1 µl of 3' SR adaptor was added bringing the total sample volume to 7 µl. The samples were then incubated 2 minutes at 70°C in a pre-heated thermal cycler and transferred on ice. To each sample was added 3 µl of NEBNext 3´ Ligation Reaction Buffer and 3 µl of NEBNext 3´ Ligation Enzyme Mix and the reaction was incubated for 12 hours at 18°C on a thermal cycler. The reaction was topped up to 200 µl of nuclease-free water, transferred to a 1,5 ml DNA LoBind eppendorf tube. RNA was extracted and precipitated as described in the "Cap selection" section and resuspended in 17 µl of TE'T (10 mM Tris NaCl pH 7,5, 0.1 mM EDTA, 0,05% Tween 20) (Duttke et al., 2019) for Start-seq samples and 16,5 µl of TE'T for input samples. The input samples were stored at -20°C for the duration of the Star-seq specific steps.

Antarctic phosphatase treatment

Start-seq samples were heat-denatured for 2 minutes at 75°C and placed on ice. Antarctic phosphatase treatment was performed by addition of 2 µl of 10x AnP buffer followed by 1 µl of AnP (NEB, Cat n°: M0289S) and incubation for 30 minutes at 37°C. RNA was then extracted and precipitated as described above and resuspended in 17 µl of TE'T buffer.

A second AnP treatment was done followed by an extraction and precipitation step as described before. Finally, RNA from the Start-seq samples was resuspended in 16,5 µl of TE'T.

Cap removal

Input samples were thawed on ice and cap removal was performed on both sample types. Samples were denatured for 2 minutes at 75°C and placed on ice. Next, 2 µl of 10x Cap-Clip buffer and 1,5 µl of Cap-Clip™ Acid Pyrophosphatase was added and the samples were incubated for 90 minutes at 37°C. RNA was then extracted and precipitated as described before and resuspended in 14,5 µl of nuclease-free water with 0,05% Tween 20 and transferred back to a PCR tube.

First strand cDNA synthesis

The RT primer (SR RT) was diluted 1:2 if the 3'SR adaptor had also been diluted for this sample. Afterwards, 10 µl of 3'adapter ligation buffer (NEB, Cat n°: E7301AA) was added to the PCR tube followed by 1 µl of (diluted) SR RT primer. The samples were placed in a thermal cycler with the lid set to $>$ 85°C and the following program was run: 5 minutes at 75°C, 15 minutes at 37°C, 15 minutes at 25°C and a hold step at 4°C. In the meantime, the 5' SR adapter was diluted 1:2 for the samples for which the 3' SR adapter was also diluted. The 5' SR adapter was then heat-denatured for 2 minutes at 70°C and placed on ice. Within 30 minutes of the adapter denaturation, 1 µl of the (diluted) 5' SR adapter, 1 µl of the 10x 5' ligation reaction buffer and 2,5 µl of the 5' ligation enzyme mix were added to the sample bringing the total volume to 30 µl. The samples were then incubated for 1 hour at 25°C. Following the incubation step, 8 µl of First Strand Synthesis Reaction, 1 µl of Murine RNAse Inhibitor and 1 µl of ProtoScript II Reverse Transcriptase was added to each sample and incubated for 1 hour at 50°C followed by a heat-inactivation step of 15 minutes at 70°C. Samples were finally stored at -20°C.

Cycle number quantitation and indexing

A qPCR experiment was performed to assess the number of PCR cycles needed to amplify the libraries (Ford, 2012). A qPCR master mix was set up for the n samples to be tested composed of n times (with 10% buffer volume) the following reagents: 1 µl of SR primer for Illumina, 1 µl NEBNext index primer 1 for Illumina (NEB, Cat n°: E7335S), 10 µl of 2x KAPA SYBR FAST master mix (Roche, Cat n°: KK4600) and 4 µl of nuclease-free water.

The master mix was kept covered and on ice while the libraries thawed on ice. Each library was tested in triplicate and 3 reactions with 1 µl of nuclease-free water were also done and served as non-template control (NTC). To each well of a 96-well plate (Roche, Cat n°: 04 729 692 001) was added: 3 µl of Tris-NaCl pH 8 0,5% Tween 20, 1 µl of cDNA library and 16 µl of the master mix previously made. The plate was sealed with sealing foil (Roche, Cat n°: 04 729 757 001), centrifuged for 1 minute at 1500g and placed in a LightCycler 480 instrument II (Roche, Cat n°: 05 015 278 001). The following PCR program was run: First, an initial denaturation step of 3 minutes at 95°C was performed. Following this, 20 cycles of PCR amplification were done (Denaturation: 30 seconds at 95°C, annealing: 30 seconds at 62°C and elongation: 30 seconds at 72°C with data acquisition during this step). Afterwards, a melting curve step was performed (65°C for 10 seconds, 95°C with continuous ramp rate of 0,11°C/seconds and 5 acquisitions/°C, hold step at 37°C). The optimal cycle number for a specific sample was identified as the one closest to half of the maximal fluorescence value obtained for that particular sample.

Next, a PCR reaction was set up to generate the seconds strand and index the libraries. A master mix was set up for the n samples to be tested composed of n times (with 10% buffer volume) the following reagents: 50 µl of LongAmp TAQ 2X master mix (NEB, Cat n°: M0287S), 2,5 µl of SR primer for Illumina and 5 µl of nuclease-free water. To each first strand synthesis reaction was added: 57,5 µl of the previously made master mix and 2,5 µl of the appropriate index primer (NEB, Cat n°: E7335S). The indices were chosen such that at every index position, there would be at least one library in the library pool with an A and one library with a T or C.

The following PCR program was run: First, an initial denaturation step of 30 seconds at 94°C was performed. Following this, n cycles of PCR amplification for each library were done in accordance with the amount of PCR cycles determined by qPCR (Denaturation: 15 seconds at 94°C, annealing: 30 seconds at 62°C and elongation: 15 seconds at 70°C). A final elongation step of 5 minutes at 70°C followed by a hold step at 4°C were then done to finalize the PCR program.

DNA was then purified using the Monarch PCR & DNA Cleanup Kit (NEB, Cat n°: T1030S) following the manufacturers guidelines. The buffer to sample ratio used during this purification process was 7:1. At the end of the purification process, the samples were eluted in 27,5 µl of nuclease free water. One microliter of eluate per library was used for diagnostics by microcapillary electrophoresis using the DNA 1000 kit (Agilent, Cat n°: 5067-1504).

Library size selection and purification

The purified libraries were supplemented with 5 µl of 6x gel loading dye part of the library preparation kit and loaded per 15 µl in wells of a 6% PAGE TBE (Life Technologies, Cat n°: EC6265BOX) gel placed in a XCell SureLock™ Mini-Cell electrophoresis system. Five microliters of Quick-Load pBR322 DNA-MspI was used as ladder and the gel was run for 1 hour at 120 V in TBE buffer. The gel was then taken out of its cast and stained for 3 minutes in a clean container with 1x SYBR Gold TBE buffer. The gel was rinsed once with TBE buffer and placed on a Safe Imager™ blue-light transilluminator. For each library, the gel area corresponding to sizes between 160 and 210 nt (insert size between 40 and 110 bp) were cut out and placed in a DNA LoBind eppendorf tube and 300 µl of DNA gel elution buffer was added. The gel slices were then crushed with a RNAse-free pestle and incubated for 2 hours at room temperature with end-to-end rotation. The eppendorfs content was then transferred to a spin filter column (Sigma-Aldrich, Cat n°: CLS8162) and spun for 2 minutes at 13200g. The eluate was next transferred to a DNA LoBind eppendorf and supplemented with 900 µl of 100% ethanol, 30 µl of 3 M sodium acetate pH 5,5 and 1 µl of Linear acrylamide provided in the library preparation kit. The samples were vortexed and incubated for 30 minutes at -80°C. Following this, DNA was pelleted by centrifugation for 30 minutes at 4°C and 21000g. The supernatant was next removed and the pellet was washed with 80% ethanol. An additional centrifugation step followed to pellet the precipitate at 4°C for 5 minutes with 21000g. The supernatant was removed again and the pellet was air-dried for 3 minutes. Finally, the libraries were resuspended in 12 µl of TE buffer and 1 µl was loaded on a DNA 1000 chip and ran on the bioanalyzer to obtain the average library size and concentration.

Sequencing

The sequencing was performed on a Nextseq 550 instrument (Illumina, Cat n°: SY-415-1002) using the Nextseq 500/550 High output kit v2.5 (Illumina, Cat n°: 20024906). Libraries were initially diluted to 4 nM and pooled together as decided when indexing the libraries. Libraries were denatured by mixing 5 µl of the pooled 4 nM library solution with 5 µl of 0,2 N NaOH in a DNA LoBind eppendorf (Illumina 2018). The sample was then vortexed, spun down for 1 minute at 280g and incubated 5 minutes at room temperature. Next, 5 µl of 200 mM Tris-NaCl pH7 was added to the tube followed by a vortexing step and centrifugation for 1 minute at 280g. The library pool was further diluted to 20 pM by adding 985 µl of pre-chilled HT1 buffer. The sample was then vortexed, centrifuged for 1 minute at 280 and placed on ice. Libraries were further diluted to 2,5 pM with HT1 and 1,3 ml of the final dilution was loaded onto the reagent cartridge. The sequencing layout was 86 cycles of single-end sequencing and 6 additional cycles for index reading. Libraries were sequenced at a depth of ± 60M reads each. All libraries were sequenced in 2 batches.

2.2 Computational methods

All the computational analyses have been performed by myself unless specified otherwise.

2.2.1 Start-seq data processing

A custom workflow using the Snakemake workflow management system (Mölder et al., 2021) was built to perform QC, trim and align reads and call peaks on individual samples. FastQ files were validated using `biopet-validatefastq=0.1.1` (Peter van 't Hof, Vorderman, & Cats, 2018) with parameter `--log\_level info`. Reads were trimmed using `trimmomatic=0.39` (Bolger, Lohse, & Usadel, 2014) in single-end (SE) mode with arguments `ILLUMINACLIP:./workflows/resources/illumina\_adapters.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:20` and checked for contaminants using `kraken2=2.1.2` (Wood, Lu, & Langmead, 2019) using the 'Standard' kraken2 database (Version k2_standard_20201202) and parameters `--confidence 0.5 --minimum-base-quality 20 --use-names --gzip-compressed`. FastQC was performed on both the full-length and trimmed reads using `fastqc=0.11.9` (Andrews, 2010) and results were aggregated with the help of `multiqc=1.6` (Ewels, Magnusson, Lundin, & Käller, 2016). The SchMedS3_h1 genome assembly recently assembled by the lab was used as reference (Ivankovic et al., 2023). It was indexed using `star=2.7.8a` (Dobin et al., 2013) using the parameters `--runMode genomeGenerate --genomeSAindexNbases 13`. Reads were then aligned to the genome using `star=2.7.8a` with as following parameters `--readFilesCommand gunzip -c --outFilterType BySJout --outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverLmax 0.1 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --outSAMattributes NH HI NM MD --outSAMtype BAM SortedByCoordinate` and the resulting bamfiles were filtered for reads with a MAPQ value greater or equal to 20 and indexed afterwards using `samtools view -hb -q20` and `samtools index` respectively (H. Li et al., 2009). Bam files were then used to generate Tag Directories, an alignment format compatible with the HOMER tools suite (Heinz et al., 2010), using the `batchMakeTagDirectory.pl` command with arguments `-checkGC -`

`single -fragLength 85`. Following this, transcription initiation clusters were called on Start-seq library Tag directories with `findcsRNATSS.pl` (Duttke et al., 2019) using their respective input samples Tag directory as baseline and the new gene models generated in collaboration with the Pandolfini lab (Ivankovic et al., 2023). The latter was first converted to a GTF format using `gffread=0.12.7` (Pertea & Pertea, 2020) with the `-T` argument. The resulting peak file was then used to generate bed files for each sample using the `pos2bed.pl` command with -bed -color strand arguments. Finally, strand-specific BedGraph files were generated for each sequenced library using the `makeUCSCfile` command with arguments `-style tss -strand +` or`-` Both peak files and BedGraph files were loaded onto an internal instance of the UCSC genome browser for visualization.

2.2.2 RNA-seq data processing

A custom workflow using the Snakemake workflow management system (Mölder et al., 2021) was built to perform QC, trim and align reads on individual samples. FastQ files were first validated using `biopet-validatefastq=0.1.1` with parameter `--log\_level info` for proper formatting and for proper read pairing in case of paired-end data. Reads were then trimmed using `trimmomatic=0.39` in single-end (SE) or paired-end (PE) mode using the following arguments `ILLUMINACLIP:./workflow/resources/illumina\_adapters.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:35`. Contamination was evaluated as in the "Start-seq data processing" section with the exception of the use of the `--paired` parameter in the case of paired-end data. FastQC was performed as explained above and reports were aggregated using MultiQC. Salmon was used to quantify transcript abundances (Patro, Duggal, Love, Irizarry, & Kingsford, 2017). Therefore, a decoy aware transcriptome, or gentrome, was first built. The transcriptome was extracted from the genome using `gffread=0.12.7` with the `-w` argument and the new genome annotations. Then the genomic scaffolds as well as the mitochondrial genome (GenBank accession number: KM821047.2) were concatenated to the transcriptome to generate the gentrome. The gentrome was then indexed using `salmon=1.10.0` in index mode where the names of the decoy scaffolds were mentioned after the `-d` argument. Finally trimmed reads were used to quantify transcript abundance by using `salmon=1.10.0` in `quant` mode with `--libType` set to `A`.

2.2.3 ATAC-seq and ChIP-seq data processing

These samples were not generated by me. The H3K4me1 ChIP-seq data comes from a study performed by Mihaylova and colleagues (Mihaylova et al., 2018) (SRA accession number SRR4089775) while the ATAC-seq, H3K4me3 and H3K27Ac ChIP-seq was performed by my colleague Mario Ivankovic (Ivankovic et al., 2023).

For both sets, I generated a custom workflow using Snakemake detailed below. The H3K4me1 dataset was first downloaded using the `prefetch` command from the sra-tools suite (SRA, 2014). The downloaded file was then transfomed in paired-end fastq files using the `fastq-dump` command from the same suite with the `--gzip --defline-qual '+'` arguments. Following this, reads were validated, trimmed and checked for contaminants as in the case of paired-end data in the "RNA-seq data processing" section. FastQC and MultiQC were performed in a similar manner.

The inhouse datasets were bam files aligned on the previous genome assembly (Grohme et al., 2018) and were first converted back in FastQ paired-end format using `bedtools bamtofastq` (Quinlan & Hall, 2010). Since only high-quality alignments had been retained previously, the reads obtained after the `bamtofastq` command were not submitted to the initial part of the workflow and used directly for mapping. The SchMedS3_h1 genome assembly was indexed using the `bowtie2-build` command from `bowtie2=2.5.0` (Langmead & Salzberg, 2012). Reads were then aligned using `bowtie2` with arguments `-I 0 -X 1000 --no-unal --very-sensitive -S`. Low-quality alignments with MAPQ values lesser than 20 were discarded using `samtools view -hb -q20`. Duplicate alignments were discarded by first sorting the alignments by name using `samtools sort -n`, then filling in the mate coordinates with `samtools fixmate -m` and finally removing duplicates using `samtools markdup -r`. Next, only reads aligned in proper pairs were kept by running the command `samtools view` with arguments `-b -f 3`. Finally, alignments were sorted by coordinate and the bam file was indexed using `samtools sort --write-index`. Before removing duplicates, additional QC steps were taken. For all data sets, the library complexity was assessed by running the `estimateLibComplexity(readsDupFreq(bamfile))` command form the `ATACseqQC 1.22.0` R package (Ou et al., 2018). For the ATAC-seq dataset, the fragment size distribution was also assessed using the `fragSizeDist` command form the same package.

2.2.4 Regulatory element annotation and characterization

This work was done on the asexual *Schmidtea mediterranea* Start-seq datasets unless stated otherwise.

Peak merging and replicate concordance assessment

First, for each biological replicate, a certain set of peaks/transcription initiation clusters were called (see “Start-seq data processing” section). These peak sets were then merged together to generate a final peak set following an Iterative Merging approach (Grandi, Modi, Kampman, & Corces, 2022). For this, a custom script was built to perform this in parallel over all the different genomic scaffolds while taking peak orientation into account. Briefly, peaks from all biological replicates were placed in one file, ordered by genomic coordinate and then divided per scaffold and peak orientation. Following this, for each file, the peaks were ranked by their peak score, a metric obtained when the peak calling was performed. The peak with the highest score is retained while any overlapping peak is excluded. Afterwards, the second most significant peak is retained and any peak overlapping it is again excluded. This process is performed iteratively until no peak overlapped another. Finally, all remaining peaks for each scaffold and strand were combined in one file and sorted by genomic coordinate to generate the final peak set.

Concordance between biological replicates was then assessed by counting tags found in each peak per biological replicate using the `annotatePeaks.pl` from the HOMER tools suite with arguments `-strand + -fragLength 1 -raw`. The tag counts were then Rlog-transformed using the `rlog` command part of the `DESeq2` R package (Love, Huber, & Anders, 2014) and replicate concordance was assessed by creating pairwise scatterplots and calculating pairwise Pearson correlation coefficients. This was done by running the `ggpairs` command from the `GGally 2.1.2` R package (Schloerke et al., 2011) on the R log-transformed data.

Regulatory element distribution around gene models

First, a custom TxDB object was created for the SchMedS3_h1 gene models generated by the lab using the `makeTxDbFromGFF` command form the `GenomicFeatures 1.50.4` R package (Lawrence et al., 2013). The merged peak file was then transformed in a BED format and loaded in R with the `readBed` command from the `genomation 1.30.0` R package (Akalin, Franke, Vlahoviček, Mason, & Schübeler, 2015). All peaks were finally characterized following their position with regards to the gene models using the `annotatePeak` command from the `ChIPseeker 1.34.1` R package (Yu, Wang, & He, 2015). The TSS region was set to 1 kb around the start of the gene models `tssRegion = c(-500,500)` and the orientation of the peak and gene model had to be identical `sameStrand = TRUE`.

Annotation of regulatory elements

The `findcsRNATSS.pl` command automatically assigns an identity to identified putative regulatory elements on the basis of their location with regards to gene models. From these annotations were built the promoters and enhancers sets used in further analyses. All putative regulatory elements with the 'other' label were defined as enhancer. For the promoters were taken, all the putative regulatory elements with the 'TSS' label and elements with the 'firstExon' or 'singleExon' label if the gene model did not have a putative regulatory element with the 'TSS' label associated to it. Bidirectionality of regulatory elements was also determined by `findcsRNATSS.pl`. If 2 or more reads per 10 milion are found in the antisense direction between -500 and + 100 nt relative to the primary called TIC, the cluster is considered to be bidirectional.

Bidirectional transcription initiation at promoters and enhancers

The Start-seq fastq files were first concatenated in one, trimmed and aligned to the SchMedS3_h1 genome as detailed in the "Start-seq data processing" section. Strand-specific bigwig files were then made using the `bamCoverage` command from the `deepTools =3.5.1` suite (Ramírez et al., 2016) with arguments `--filterRNAstrand reverse` or `forward`. Next, an AWK script was written to filter out all gene models with a negative-strand orientation to keep only gene models with a positive-strand orientation. A matrix of scores for Start-seq signal on the forward and reverse strand was built using `computeMatrix` for all promoters and enhancers separately `--skipZeros --regionBodyLength 6000 --upstream 2000 --downstream 2000 --smartLabels` . Finally, strand-specific heatmaps and summary plots were generated for Start-seq signal around promoters and enhancers with a positive-strand orientation using the `plotHeatmap` command with standard parameters.

Epigenomic signature and core promoter motifs at Regulatory Elements

Tag Directories for ATAC seq and ChIP-seq data was generated from bam files using the `makeTagDirectory` command with standard parameters. Average signal of ATAC-, ChIP- and Start-seq signal over all putative regulatory elements, only enhancers or only promoters was computed using the `annotatePeaks.pl` with parameters `-hist 10 -tbp 3 -size 2000`. Plots were generated with `ggplot2 3.4.2` (Wickham, 2011) and combined using `ggarrange` form the `ggpubr 0.6.0` R package (Kassambara, 2020). Core promoters motifs (n = 10) gathered from Haberle and Stark (Haberle & Stark, 2018) and part of the homer tools suite were probed for enrichment at putative regulatory elements using the `annotatePeaks.pl` with the `-size 600 -hist 1 -m lib/motifs/core\_promoters .motifs` arguments, where the `-m` argument refers to a file containing the position probability matrices (PPMs) of all the investigated motifs. Probed motifs that did not have a PPM in the home tools suite had one made using the `seq2profile.pl` command with standard arguments. The initial in input sequence was found in the Haberle and Stark review. Motif occurrence was visualized using ggplot2.

Assay sensitivity assessment

Subsampling

Previously trimmed reads from asexual samples (see "Start-seq data processing" section) were merged by condition to generate one FASTQ file per condition. Next, a sub sample of the total Start-seq read number was taken using the `seqtk sample` command (H. Li, 2023). All the subsamples were then aligned to the reference genome and peaks were called as described in the "Start-seq data processing" section with as input control the totality of the input reads for even background. The number of high-quality alignments (MAPQ > 20) from the subsampled reads were plotted against the number of peaks called by findcsRNATSS.pl for all subsamples and both conditions.

Missing promoter analysis

For the asexual dataset, genes with and without an annotated promoter as defined in the "Annotation of regulatory elements " section were divided into two groups and compared against one another in terms of normalized RNA-seq read counts. The RNA-seq data used was previously published by Davies and colleagues (Davies et al., 2017). This data, comprising sexual and asexual RNA-seq samples, was downloaded from the Sequence Read Archive (SRA accession numbers SRR3629944 - SRR3629952) and processed as in the "RNA-seq data processing" section. Count data was imported in R using `tximport 1.26.1` (Soneson, Love, & Robinson, 2015) and transformed in a `DESeqDataSet` object using `DESeqDataSetFromTximport` from the `DESeq2 1.38.3` R package. The reference condition was set to be the asexual samples with `relevel` and differential gene expression analysis was done using the `DESeq` command. Finally, normalized counts for the asexual condition for each gene were extracted from the baseMean column of the results. The normalized read count distributions of genes with and without an annotated promoter were compared against each other with the help of a box plot and statistical significance for a difference between the two groups was assessed with a t-test.

2.2.5 Sexual versus asexual comparison

Differential Regulatory Element Activity analysis

Peaks identified in all biological replicates from the both the sexual and asexual conditions were merged together by an iterative merging approach explained in the "Peak merging and replicate concordance assessment" section. Pairwise Spearman correlations for sexual and asexual samples were then computed. First, an intermediate `multiBamSummary` object (`deepTools=3.5.1`) was computed using the Start-seq Bam files of each replicate and the merged peak set previously generated in BED format. Pairwise correlations were visualized using the `plotCorrelation` command with `--corMethod spearman --whatToPlot heatmap --skipZeros --colorMap RdYlBu --plotNumbers` arguments. Additionally, a clustered heatmap of the Euclidean distances between samples was generated using `pheatmap` function from `pheatmap 1.0.12`. (Kolde, 2018). Next, raw counts per peak for each replicate was calculated using the `annotatePeaks.pl` command with the `-raw -strand + -fragLength 1 -cpu 20 -gtf` arguments where the `-gtf` argument pointed to the genome annotation file in GTF format previously made in section "Start-seq data processing". Differential Regulatory Element Activity (DREA) between the two conditions was calculated using `getDiffExpression.pl` from the homer tools suite with the using the `-DESeq2` argument. Batch correction was applied at the same time with the `-batch` argument. Batches represented the sequencing batches explained in the "Sequencing" section. Results were then loaded in R for further analysis. First, a principal component analysis (PCA) was performed using the `prcomp` command from the `stats 4.2.2` package and visualized using the `autoplot` command from the `ggfortify 0.4.16` package.

Next, regulatory elements were defined as differentially active if they had an absolute log2 fold change greater or equal to 2 and and adjusted pvalue smaller than 0.01. The top 100 most upregulated promoters, were next selected and the sequence of their associated gene was BLAST'd (Altschul, Gish, Miller, Myers, & Lipman, 1990) against the nr/nt Nucleotide collection using blastn. Genes with no BLAST hit were discarded and the top 25 genes with a BLAST hit were researched for links to reproductive functions. Results from the BLAST analysis and the general differential analysis were brought together and visualised in a volcano plot and elements of interest were annotated using `ggrepel 0.9.3` (Slowikowski, 2023).

Gene expression and promoter activity correlation analysis

The `DESeqDataset` object originating from the RNA-seq data from Davies and colleagues (see "missing promoter analysis") filtered for genes with at least 10 reads shared across the 8 samples to remove lowly expressed genes. A PCA of the rlog'd counts was made and visualized using `plotPCA` from the `DESeq` R package. DGEA was done using the `DESeq` command and the result table was extracted using the results command with `alpha = 0.05, filterFun = ihw` arguments where independent filtering was done using the `IHW 1.26.0` R package (Ignatiadis, Klaus, Zaugg, & Huber, 2016).Next, genes sharing both a differentially (padj $<$ 0.05) regulated promoter and transcript were taken and the Start-seq log2 fold change (l2fc) was plotted against the RNA-seq l2fc. Correlation between promoter activity and gene expression was assessed using `stat\_cor(method=" pearson")` from the `ggpubr` package. Finally, the overlap between RNA-seq DGEA and Start-seq results was assessed by looking genes having a differentially active transcript (padj $<$ 0.05) and/or regulatory element with upset plots (Lex 2014) using `UpSetR 1.4.0` (Conway, 2019).

Gene ontology enrichment analysis

Mapping of GO-terms to gene models was done by Dr. Elham Bavafaye. She used Eggnog (Huerta-Cepas et al., 2019) and interproscan (Jones et al., 2014), to provided GO terms based on the known functional annotations in different species. Genes with a differentially upregulated promoter (l2fc $>$ 2 and adjusted p-value $<$ 0.01) were tested for enrichment for certain GO-terms using the `enricher` function from the `clusterProfiler 4.6.2` R package (T. Wu et al., 2021). Only terms belonging to the Biological Processes category were tested and terms with Benjamini-Hochberg-corrected p-values less than 0.05 were kept. Results were then visualized using the `dotplot` and `cnetplot` function form the same package.

2.2.6 Motif variability analysis

Forging a BSgenome Package

First, a custom BSgenome package (Pagès, 2017) for the SchMedS3_h1 assembly was forged. This was done by converting the assembly in a 2bit format using `faToTwoBit`, generating a custom DESCRIPTION file and a chromosome-size file. The package was then forged in R using the `forgeBSgenomeDataPkg` from the `BSgenome 1.66.3` package. Next, the source package was built using the command `R CMD BUILD`, checked with `R CMD CHECK` and finally installed in the R session using `R CMD INSTALL`.

chromVAR analysis

The motif variability analysis script was adapted from a script originally designed by Mario Ivankovic. The combined peak set from sexual and asexual samples generated in the "Differential Regulatory Element Activity analysis" section was loaded in R using the `getPeaks` command from the `chromVAR 1.20.2` package (A. N. Schep, Wu, Buenrostro, & Greenleaf, 2017) with arguments `sort\_peaks = TRUE`. Peaks were then resized using the `resize(peaks, width = 150, fix = "center")` command from the `IRanges 2.32.0` package (Lawrence et al., 2013). Counts per condition for each peak was calculated using the `getCounts` command from the chromVAR package. The data used to extract counts were the high-quality alignments (MAPQ $>$ 20) obtained after aligning the combined Start-seq biological replicates to the genome. GC bias was then corrected using the `addGCBias` command. Next the motif position frequency matrices (PFMs) from the PHYLOFACTS database part of the `JASPAR2018 1.1.1` R package (A. Khan et al., 2018) were with the `getMatrixSet` command form the `TFBSTools 1.36.0` R package. Motifs were assigned to each peak using the `matchMotifs` from `motifmatchr 1.20.0` (A. Schep, 2020) with as arguments the set of motif PFMs, the counts per peak per condition and the schMedS3_h1 genome in a BSgenome format. Background peaks were then generated using the `getBackgroundPeaks` command, deviations were next computed with the `computeDeviations` command and finally, motif variability was computed with the help of `computeVariability`.

Motif identification and variability plotting

Variable motifs sequences with an adjusted p-value lower than 0.05 were entered in the TOMTOM motif comparison tool (Gupta, Stamatoyannopoulos, Bailey, & Noble, 2007) and the program was ran with the following arguments `-no-ssc -oc . -verbosity 1 -min-overlap 5 -dist pearson -evalue -thresh 0.5 query\_motifs db/JASPAR/JASPAR2022 \_CORE\_vertebrates\_non-redundant\_v2.meme`. Transcription factor family names of each hit were retained. Finally, the top 100 most variable motifs were plotted using ggplot2 with significant motifs annotated with their family names.

2.2.7 *Schmidtea mediterranea* transcription factor database

Database generation

This part was done by Dr. Rozanksi. Rozanski used the DNA binding domain multiple sequence alignments (MSA) available on TFclass (`http://tfclass.bioinf.med.uni-goettingen.de/`) (Wingender et al., 2018) to classify all *Schmidtea mediterranea* gene models possessing one into transcription factor families. First previously generated SMEST gene models (Rozanski et al., 2019) were translated using transdecoder (Haas, 2019). Next, DNA binding domains MSA's were used to make hidden Markov models (HMMs) with HMMER's command `hmmbuild` (`http://hmmer.org/`). These hmms were finally used to assign gene models models to each DNA binding domain with hmmsearch, classifying them in transcription factor families. The same analysis was performed on the 'dd_smed_v6' transcriptome assembly (Rozanski et al., 2019).

Transcription factor candidate selection

Gene IDs belonging to TF families with a variable motif (see "Motif variability analysis" section) were selected and analyzed for their differential expression between sexual and asexual worms using the previously processed data from Davies and colleagues. Log2 fold changes from genes from TF families of interest were extracted and several genes were selected for downstream analysis using the following criteria. If many genes belonged to the TF family, only significant genes with a high positive l2fc were selected (5-6 genes). If there were only few genes in the TF family, criteria were relaxed and genes with a lower l2fc or a non-significant adjusted p-value were selected. If no genes had a positive l2fc, the same analysis was performed on the best SMEST blast hit for the genes in the dd_smed_v6 TF database. Finally, if no additional gene were found in the dd_smed_v6 TF database for a specific TF family, genes with a negative, non-significant l2fc were selected.

2.2.8 Differential gene expression analysis of transcription factor knockdown samples

Counts for each condition (2 biological replicates per condition) were imported in R using `DESeqDataSetFromTximport` and genes with less than 10 counts across all samples were discarded. The EGFP condition was next set as reference using relevel. To visualize similarities between samples, a PCA using `plotPCA` on the rlog'd counts was performed. Additionally, a clustered heatmap of the Euclidean distances between samples was done using `pheatmap` function from `pheatmap 1.0.12`. The normalized counts of several marker genes for sexual tissues identified in the literature (Chong et al., 2011; Issigonis et al., 2022; U. W. Khan & Newmark, 2022; Rouhana et al., 2017; Steiner et al., 2016; Vila-Farré et al., 2023; Y. Wang et al., 2010, 2007; Zayas et al., 2005) were then assessed for each condition and compared to the asexual, *ophis* and *egfp* controls using boxplots in ggplot2. Next differential gene expression analysis was done using the DESeq command and results for each condition were extracted using the following command `results(deseq\_dataset, contrast = c("CONDITION", "target"," EGFP"),alpha = 0.05, filterFun = ihw)` where “target” is the condition being compared to the EGFP reference, “alpha” is the significance threshold and independent filtering was done using the `IHW` R package. All differentially expressed genes for each condition were selected and intersection between sets was assessed with upset plots using `UpSetR 1.4.0`. Significantly downregulated genes (adjusted p-value < 0.01, l2fc < 0) were next taken and processed as in the “Gene ontology enrichment analysis” section.

Chapter 3

Results

3.1 Development of a planarian-compatible method to study transcription initiation

When I decided to profile transcription initiation in *S. mediterranea*, no protocol existed that was adapted for the isolation of short-capped RNA in said species. Existing methodologies for nuclear isolation at the beginning of my thesis were primarily designed for ChIP-seq applications, relying on a formaldehyde fixation step (Dattani et al., 2018; Duncan et al., 2015; Mihaylova et al., 2018) which are incompatible for RNA extraction. In this chapter I will present the work that was done at the beginning of my thesis where I developed a working protocol to isolate a pure nuclear fraction with intact RNA. Later on, several other native nuclei isolation methods were developed (Ivankovic et al., 2023; Neiro et al., 2022; Pascual-Carreras et al., 2023) focusing on the identification of open chromatin regions by ATAC-seq.

3.1.1 Existing transcription initiation assays published in other organisms

Several methods have been developed to study regulatory elements with regards to transcription initiation by enriching libraries with capped RNAs and sequencing from the 5' end. Earlier methods such as CAGE and 5'-end SAGE focused solely on the identification of transcription start sites rather than all regulatory elements, see table 3.1 (Hashimoto et al., 2004; Shiraki et al., 2003) but were instrumental as initial efforts for mapping of promoters genome-wide (Carninci et al., 2005). The CIP-TAP cloning method (Project, 2009) performed an enrichment of short RNAs by size selection before 5' cap enrichment and revealed that a large fraction of short capped RNAs were located in intergenic regions. Later, the Start-seq protocol added a nuclei isolation step before size selection to further enrich for initiating transcripts increasing the specificity of the assay (Nechaev et al., 2010). Finally, the most sensitive methods developed subsequently (5' GRO-seq, GRO-cap and PRO-cap) made use of nuclear run-on (NRO) reactions to incorporated labelled nucleotides into nascent RNA before 5' cap selection to refine transcription initiation profiling even more (Kruesi, Core, Waters, Lis, & Meyer, 2013; Kwak, Fuda, Core, & Lis, 2013; Lam et al., 2013) respectively.

Table 3.1: 5' seq techniques considered for the identification of regulatory elements in *S. mediterranea*, *Hsap*: *Homo sapiens*, *Dmel*: *Drosophila melanogaster*, *Mmus*: *Mus musculus*, *Cel*: *Caenorhabditis elegans*

Method	Ref	Org	Input	Cap selection method
CAGE	Shiraki et al. (2003)	*Hsap*	Total RNA	5' CAP chemical crosslinking to Biotin
5'end SAGE	Hashimoto et al. (2004)	*Hsap*	Total RNA	Uncapped RNA dephosphorylation
CIP TAP Cloning	Project (2009)	*Hsap*	Total RNA	Uncapped RNA dephosphorylation
Start-seq	Nechaev et al. (2010)	*Dmel*	small Nuclear RNA	Uncapped RNA degradation and dephosphorylation
	Scruggs et al. (2015)	*Mmus*	small Nuclear RNA	
CapSeq	Gu et al. (2012)	*Cel*	Total RNA	Uncapped RNA degradation and dephosphorylation
5' Gro-seq	Lam et al. (2013)	*Mmus*	Native nuclei	Uncapped nascent RNA dephosphorylation
Gro-cap	Kruesi et al. (2013)	*Cel*	Native nuclei	Uncapped nascent RNA degradation and dephosphorylation
Pro-cap	Kwak et al. (2013)	*Dmel*	Native nuclei	Uncapped nascent RNA dephosphorylation

One hurdle that all the aforementioned protocols have to overcome is to enrich their samples for capped RNA-species. With the exception of the CAGE protocol, all methods made the enzymatic reactions for cap selection to achieve this goal. A phosphatase step to dephosphorylate 5' ends of uncapped RNAs, rendering 5' adapter ligation impossible on these RNAs, is a crucial step of the 5' cap selection process. Some protocols, such as Start-seq, CapSeq (Gu et al., 2012) and GRO-cap also make use of a 5' RNA exonuclease prior to the dephosphorylation step in order to degrade uncapped RNAs, reducing background signal further.

Taking into account the famed instability of planarian cell content upon lysis, I decided not to opt for a run-on approach to profile transcription initiation. Instead, I directed my attention towards Start-seq. This protocol has only one nuclei isolation step before RNA extraction, limiting the time for nuclear content degradation.

3.1.2 Optimizing a planarian-compatible native nuclei isolation protocol

The protocol workflow is illustrated in Figure 3.1 A. The first step consists of dissociating the planarian tissue to free cells from the connective tissue they reside in. Next mechanical lysis is performed to disrupt the cell membrane in order to isolate nuclei by applying mechanical forces followed by gradient centrifugation. Afterwards the nuclei are collected and RNA is extracted by acid phenol chloroform (Chomczynski & Sacchi, 1987) purification and purified by precipitation. Small RNAs are then isolated by electrophoresis and gel extracted. Next, uncapped RNAs are degraded using a 5' monophosphate specific RNA exonuclease, leaving small capped RNAs intact. 3' adapter ligation ensued. This was done before the phosphatase treatment to exclude degradation products with a 3' phosphate group have a 3' adaptor. The 5' cap removal step followed after to allow the 5' ends of short capped RNAs to be ligated to the 5' adapter. Finally, reverse transcription was performed for first strand synthesis.

***In vitro* assessment of enzymatic efficiency**

The critical enzymatic steps detailed in Figure 3.1 B were tested first in an *in vitro* setting. First, *in vitro* synthesized RNA, both capped and uncapped, were subjected to a 5' RNA exonuclease treatment (Figure 3.1 C). Prior to the exonuclease treatment, both capped and uncapped RNA species were observable after gel electrophoresis. However, post-treatment, only the capped RNA species remained detectable. To assess decapping efficiency, capped RNAs were first decapped by treatment with CapClip acid pyrophosphatase and then subjected to an exonuclease treatment as before. Gel electrophoresis revealed that negligible amounts of RNA was left after exonuclease treatment compared to prior the exonuclease treatment (Figure 3.1 C). These results demonstrate the efficacy of both the 5' RNA exonuclease and the CapClip acid pyrophosphatase in degrading uncapped RNA species and uncapping capped RNA species respectively.

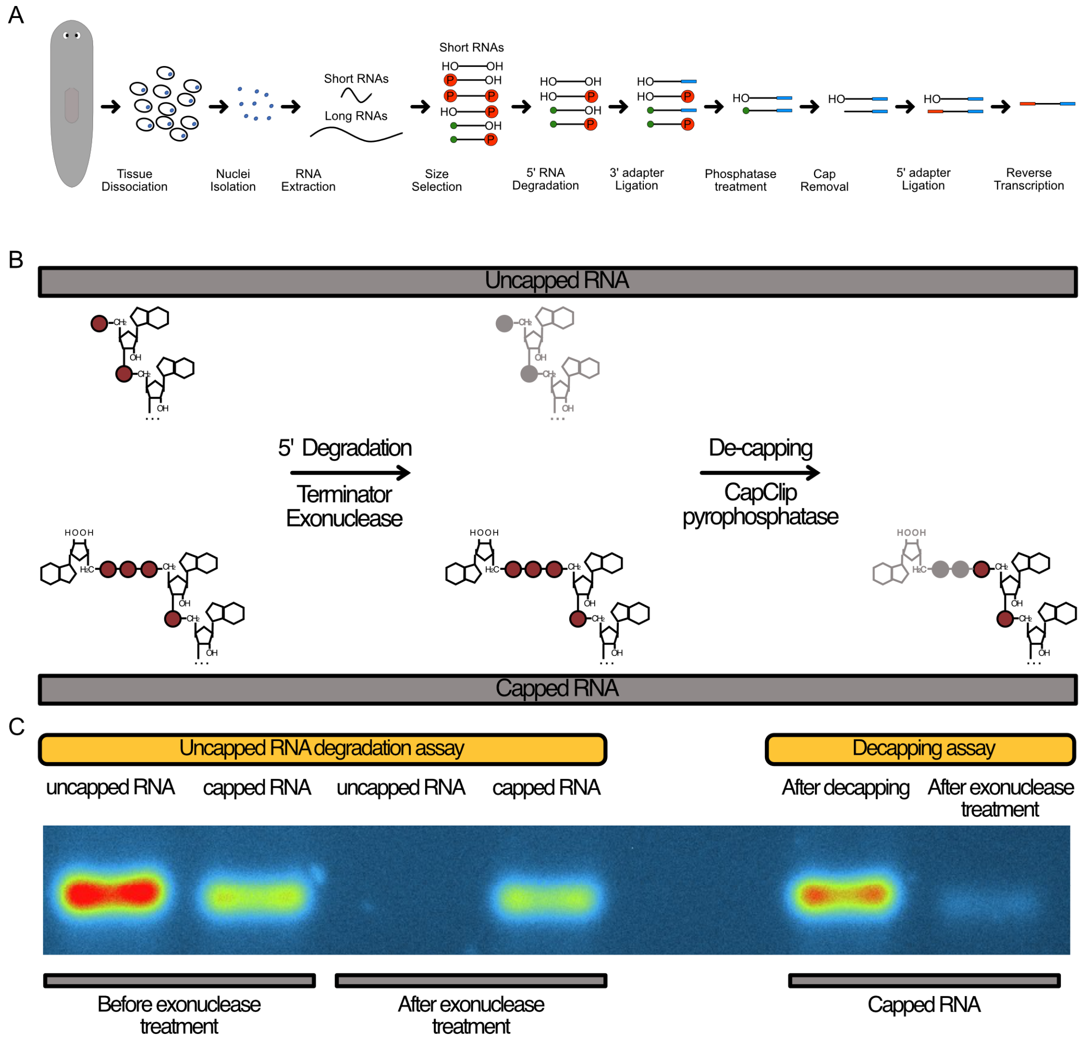


Figure 3.1: The Start-seq workflow and its cap selection procedure. A) Schematic representation of the Start-seq procedure showing the key steps for the isolation of nuclear short capped small RNAs. Whole animals are first dissociated to break down the connective tissue holding the cells together. Nuclei are then isolated by mechanical lysis and isotonic step density gradient centrifugation. Nuclear RNA is then extracted using the acid phenol chloroform procedure. Small RNAs are subsequently isolated by polyacrylamide gel electrophoresis. Uncapped small RNAs possessing a 5' phosphate group are degraded using the 5' phosphate-dependent exonuclease. To avoid cloning of degradation products possessing a 3'phosphate group, the 3' library adaptor is next ligated. Potential undegraded uncapped small RNAs possessing a 5'phosphate are rendered unclonable by Antarctic phosphatase treatment. Next, capped RNA s are de-capped using a de-capping enzyme before proceeding with 5' library adapter ligation and reverse transcription. B) Schematic representation of the key enzymatic steps for cap selection. Uncapped RNAs do not possess a methylguanosine cap and are therefore sensitive to degradation by the 5' phosphate-dependent exonuclease while capped RNAs are left intact. Using the decapping enzyme next remove the methylguanosine cap and renders the 5'RNA cloneable for adaptor ligation. C) Proof of principle experiments showing the efficiency of the key enzymes used in the Start-seq method. Capped and uncapped in vitro synthesized RNAs were subjected to the 5' phosphate-dependent exonuclease. For the de-capping assay, in vitro synthesized capped RNA was first de-capped, half of the RNA was treated with the 5' phosphate-dependent exonuclease while the other was not. All samples were run on a 1% agarose TBE gel.

Tissue dissociation optimization

The first step towards nuclei isolation from whole tissues is to disrupt it by mechanical lysis or enzymatic digestion to facilitate the release of cells from the extracellular matrix (ECM). Enzymatic digestion of the ECM using trypsin has already shown some promises in the planarian field (Plass et al., 2018) but has multiple drawbacks. First, it has to be done on fresh tissue, demanding immediate processing of the samples. Second, since tissue requires to be incubated for a certain amount of time for it to be digested, it gives a window of opportunity to the cellular content to degrade. On the other hand, mechanical lysis by mortar and pestle can be performed on frozen tissue at temperatures where enzymatic reactions are inhibited. It also allows a researcher to prepare multiple samples and store them at -80°C for long periods of time before proceeding with nuclei isolation. I therefore decided to proceed with the second option.

In the earlier trials, I tried to snap freeze worms in liquid nitrogen in a similar fashion as the ones used in the *C. elegans* research where nematodes grown in liquid culture are dripped directly into liquid nitrogen and form frozen drops referred to as 'popcorn' (Jänes et al., 2018). Unfortunately, the sheer size difference between the two organisms made the dripping part of the procedure quite impractical for planaria and I did not manage to make satisfactory 'popcorns'. Snap frozen planarian tissue was also not very amenable to mortar and pestle grinding due to their soft-bodied nature which tends to not easily form fine powder and thaws quickly to become a viscous sludge. I then decided to use the cryoPREP CP02 Dry Pulverizer, a method much more adapted to the processing of soft tissue in small quantities. This device is able to crush snap frozen tissue to a powder while keeping without letting its temperature reach above -80°C. This choice was instrumental in obtaining a fine tissue powder satisfactory for downstream processing, even from very small amounts of input material. RNA integrity was assessed after tissue pulverization (Figure 3.2). A typical electrophoretic profile for *S. mediterranea* was obtained after microcapillary electrophoresis, confirming that RNA was indeed intact after tissue dissociation.

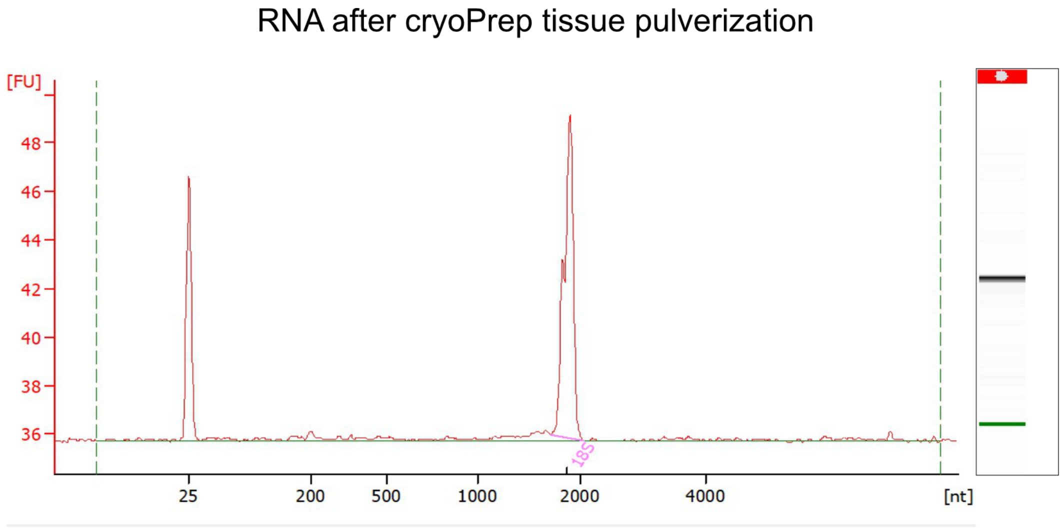


Figure 3.2: Assessment of RNA integrity after tissue dissociation using the CP02 cryoPREP Automated Dry Pulverizer. Worms were placed in a TT05MXT tissuetube and submerged in liquid nitrogen. The sample was then pre-crushed at power level 1 before being crushed a second time at power level 2. RNA was then extracted using the acid phenol chloroform procedure and ran on an Agilent Bioanalyzer instrument to assess RNA integrity.

Optimization of nuclei extraction procedure by rational buffer design

Planarians are renowned for their incompatibility with standard protocols developed for human, mouse or other model organisms. This arises from their very different body composition and inherent biology. They possess pigments that co-precipitate with nucleic acids, are coated in a protective mucopolysaccharide layer interfering with downstream sample processing and possess highly potent nucleases (Grohme et al., 2018). I therefore decided to rationally design a robust nuclei isolation protocol and associated buffers to address these obstacles. When designing the nuclei isolation buffers, I attempted to recreate a medium with similar properties as the planarian cellular environment in terms of osmolarity, ion content, pH and redox status. Planarian osmolality had been investigated in the past and was determined to be around 130 mOsm/l (Prusch, 1976; Schürmann & Peter, 2001). I therefore used this value as a reference instead of the 350 mosm/l usually found in nuclei isolation protocol. Intracellular potassium ion content was also higher than sodium ion content (Prusch, 1976) which led me to choose potassium as the positive counterion over sodium. Magnesium divalent cations have been shown to have stabilizing effect on chromatin (Hartwig, 2001) and was therefore also included. Another abundant ion is the chloride anion. I therefore chose to use KCl and $MgCl_2$ as the main salts to address the ion content of my buffers. I chose sucrose to reach the desired osmolarity, as is commonly done in nuclei isolation protocols.

Since my focus was to isolate short RNAs, I decided to use a pH of 7,4 to avoid any base-catalyzed hydrolysis with MOPS as buffering molecule due to its appropriate pKa of 7,2 and the fact that it does not interact with metal ions. The next components that I used in my buffer design served to stabilize the nuclear contents during extraction. Protective properties of polyamines such as spermidine and spermine against DNA damage have been described in the literature were thus included (Christensson & Lewan, 1974; Di Luccia et al., 2009; A. U. Khan, Mei, & Wilson, 1992). I also added protease and RNAse inhibitors to counteract the degradative activities of such enzyme classes. Finally, since the cytoplasm is primarily a reductive environment (López-Mirabal & Winther, 2008), I also added DTT as a reducing agent to counteract any potential oxidation that might happen during the nuclei extraction process. Lastly, I used a mild non-ionic detergent, IGEPAL-CA 630, to reduce surface tension and adsorption of biomolecules to plastics during nuclei isolation. It also helps with the solubilization of cell membranes for better release of nuclei

To separate nuclei from other organelles and cellular debris I decided to proceed as follows: After tissue disruption by crushing snap-frozen worms, I would resuspend the powder in the nuclei isolation buffer I designed and use a dounce homogenizer to break cell membranes to release the nuclei. I will then use a nylon mesh to filter the lysate and remove large debris from the solution before performing gradient centrifugation where nuclei would settle at a single point away from other organelles and debris. One common way to perform this density gradient is to use a dense sucrose solution (Nabbi & Riabowol, 2015; Widnell & Tata, 1964). However, this method subjects to nuclei to a highly hypertonic environment. I therefore decided to go for another protocol using Iodixanol, where osmolarity can be tuned according to ones needs (Corces et al., 2017). Finally, one last step I decide to take was to reduce the DTT concentration of the medium after the gradient centrifugation step. Many nuclei extraction protocols only add about 1 mM of DTT during isolation. I decided to align myself with this number only after separation of the nuclei from the rest of the cellular content and use a higher concentration before to ensure a reductive environment during the nuclei isolation process. A list of the main variations of the nuclei isolation buffers tried during this study can be found in Table 3.2.

In terms of input material, I decided to start with 100 7mm worms, equivaling to about 100.000.000 cells (Thommen et al., 2019) and perform the mechanical lysis in 10 ml of extraction buffer. In initial trials, I decided to use a step density gradient ranging from 25% to 40% iodixanol using 3% intervals. I then observed that nuclei were settling in bands at multiple interfaces, namely between 31 and 34, 34 and 37 as well as 37 and 40%

Table 3.2: Evolution of the buffer components and concentration during buffer optimization for the native nuclei extraction protocol

		Version 1		Version 2		Version 3	
		Buffer 1	Buffer 2	Buffer 1	Buffer 2	Buffer 1	Buffer 2
Buffer	MOPS-KOH pH 7,4	10 mM	10 mM				
	HEPES-KOH pH 7			20 mM	20 mM	20 mM	20 mM
Ions	KCl	20 mM	20 mM	20 mM	20 mM	20 mM	20 mM
	$MgCl_2$	3mM	3mM	10 mM	10 mM	10 mM	10 mM
Detergents	Igepal CA-630	0,5%	0,5%	0,05%	0,05%	0,05%	0,05%
Osmolarity modulators	Sucrose	40 mM	60 mM	40 mM	60 mM	40 mM	40 mM
Stabilizers	DTT	20 mM	1 mM	20 mM	1 mM	10 mM	1 mM
	RNAse inhibitor	0,4 U/µl	0,4 U/µl	0,1 U/µl	0,1 U/µl	0,2 U/µl	0,2 U/µl
	Protease inhibitor	1x	1x	1x	1x	1x	1x
	Spermidine	0,5 mM	0,5 mM	0,5 mM	0,5 mM	0,5 mM	0,5 mM
	Spermine			0,25 mM	0,25 mM	0,25 mM	0,25 mM

(Figure 3.3 A). Since all of these bands contained nuclei, similar in shape and RNA profile (not shown), I decided to simplify the step density gradient to a 4-step gradient and then finally a 3-step gradient to have all the nuclei collect at one spot. This had also the benefice of drastically decreasing the density gradient preparation time. I also tested whether the density gradient was a necessary step to recover a pure nuclear fraction by western blot. For this, I compared the abundance of a cytoplasmic marker and a nuclear marker (Ef1alpha and Histone 3 respectively) in protein extracts from a crude nuclear extract without gradient centrifugation against protein extracts from several nuclear preps after gradient centrifugation (Figure 3.3 A). The crude nuclear extract possessed an enriched nuclear maker signal over the control similar to nuclear preps having been through gradient centrifugation. On the other hand, the cytoplasmic marker was still clearly visible in the crude nuclear extract confirming the need for the density gradient centrifugation step.

After many trials, I managed to obtain a stable protocol with a yield of about 20-25%. I noticed that additional lower bands were present when comparing electrophoretic profiles from nuclear RNA to RNA extracted from whole tissue (Figure 3.3 C & D). I concluded that they were signs of RNA degradation. Namely, an initial Start-seq experiment showed lots of reads in gene bodies and a high proportion of multimapping reads were mapping to rRNA genes. I therefore decided to modify the extraction buffer composition. After careful research I noticed that MOPS had been described to interact with membranes and decided to replace it with HEPES, that has a similar pKa. I also reduced the detergent concentration tenfold to place a lower stress on the nuclear envelope. In order to increase nuclear content stability, I decided to increase magnesium concentration and added spermine to the buffer. I also decreased the buffer pH to 7 to lower the probability of base catalyzed RNA hydrolysis. With the view of reducing cost of the nuclei isolation procedure, I lowered the RNAse inhibitor concentration, thinking that the other changes made to the buffer would counteract this choice and maintain RNA integrity. Lastly, I reduced the tissue input from 100 to 30 7 mm worms with the idea of “diluting” the lysate and therefore reducing concentration of “lytic material” able to adversely affect the RNA integrity.

The second buffer version improved RNA integrity drastically since no undesirable bands were visible anymore after gel electrophoresis (Figure 3.3 E). Additionally, the nuclei extraction yield increased to 35%, making the protocol better suitable for lower input samples. In order to increase RNA stability and make the protocol compatible with even lower inputs, I decided to perform some final modifications. I decided to work with 20 7 mm worms as input material and perform the extraction in a volume of 2 ml. This allowed me to increase the RNAse inhibitor concentration while still reducing the total amount used. I also decreased DTT concentration of the extraction buffer to 10 mM, thinking it would still be sufficient to maintain a reducing environment. The final extraction procedure had a similar yield of 35% nuclei recovery and produced better electrophoretic profiles compared to previous iterations (Figure 3.3 F).

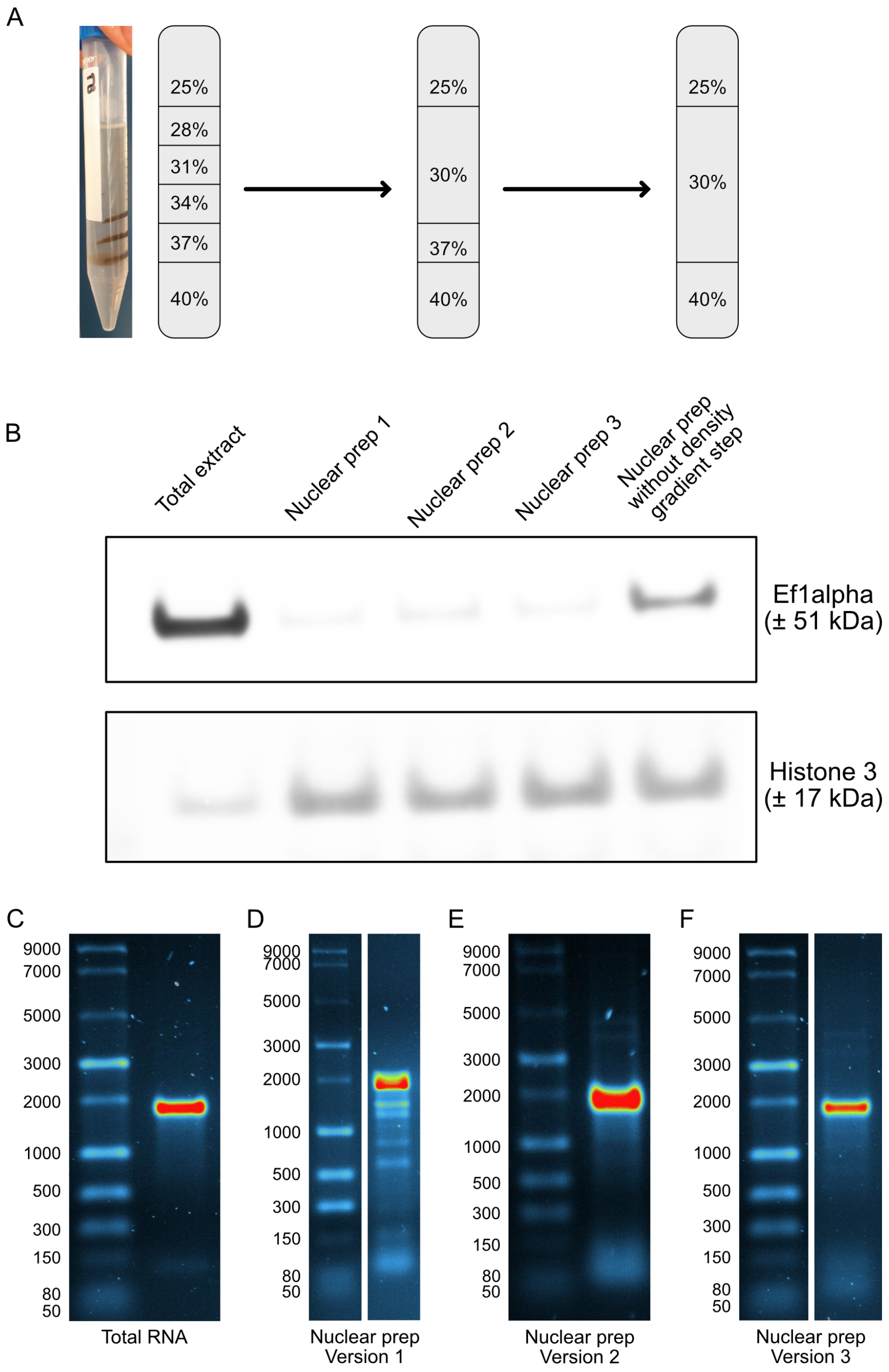


Figure 3.3: Development of a native nuclei isolation protocol for RNA extraction. A) Refinement of the step density gradient centrifugation step. Nuclei pelleted at different densities. Multiple densities are subsequently removed from the procedure to allow pelleting of all the nuclei at the interface between the 30% and 40% iodixanol layers. B) Western blot on a cytoplasmic (Ef1alpha) and nuclear marker (Histone 3) to verify the need of density gradient centrifugation. C) Electrophoretic profile of nuclear RNA from different versions of the nuclei isolation protocol in comparison with total RNA

I then performed western blots to quantify changes in nuclear and cytoplasmic marker abundance between protein lysates from nuclei and whole tissue. The results show a significant decrease in the cytoplasmic marker abundance in the nuclear sample compared to whole tissue with near to no signal observed in the nuclear lysate. On the other hand, a significant increase (t-test, p=5.1$e-9$) of the nuclear marker was observed in the nuclear lysate compared to whole tissue (Figure 3.4 A & B). Finally, I used the more accurate microcapillary gel electrophoresis method to confirm the integrity of RNA on the improved protocol (Figure 3.4 C). The profile showed, as expected, two peaks around the migration size of the 18 s rRNA coming from the 18s rRNA and the two subunits of the 28s rRNA (Sun et al., 2012). Very little signal was seen at lower sizes except for the very small RNAs representing the 5s rRNA and small RNAs.

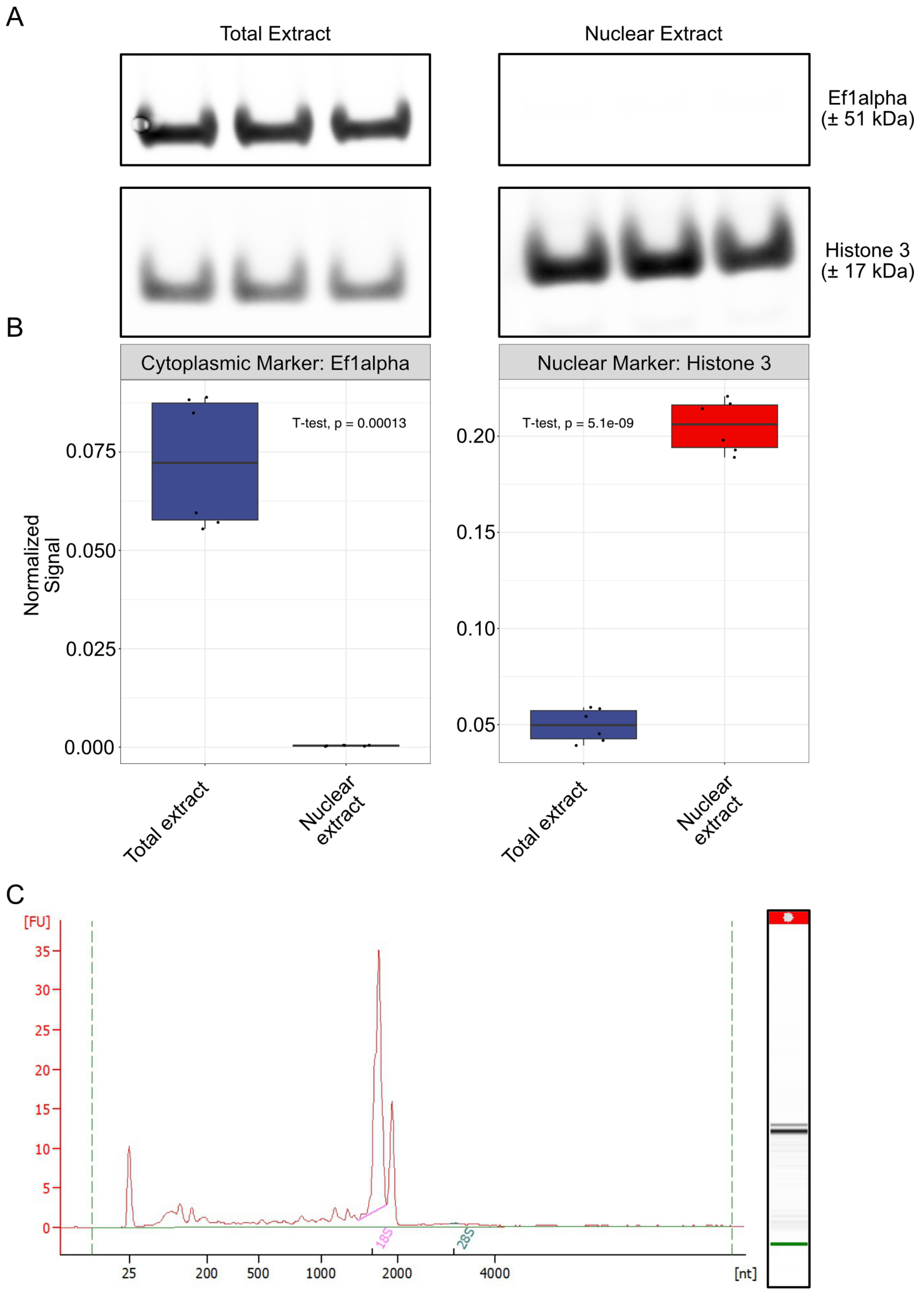


Figure 3.4: Nuclear purity and RNA integrity assessment of the final Start-seq protocol. A) Western blot comparison against the cytoplasmic (Ef1alpha) and nuclear marker (Histone 3) between the total protein extract and nuclear extract to visualize differences in band intensity between the two conditions. Each condition was done in three biological replicates. B) Quantification and statistical analysis of nuclear and cytoplasmic marker intensity between the total protein extract and nuclear extract. Each condition was done in sextuplicate. C) RNA electrophoretic profile of nuclear RNA after nuclear extraction and RNA precipitation using ammonium acetate. Enrichment for small RNAs can be seen with the increased signal at small sizes.

3.1.3 Data processing pipeline design for Start-seq experiments

After having verified the efficiency of the key enzymes and developed a robust nuclei isolation protocol, I then prepared Start-seq samples in triplicate for *Schmidtea mediterranea* asexual biotype (see Material and Methods). Each sample is composed of two libraries and between 7-14 Mio nuclei were used to prepare each sample. The first library was subjected to cap selection and contains the actual information about transcription initiation sites. The second library was not subjected to cap selection and represents background signal similar as the IP control in ChIP-seq experiments.

To process the sequencing data, I built a computational pipeline using snakemake as a reproducible workflow manager (Mölder et al., 2021) and depicted in (Figure 3.5). Both input and Start-seq Fastq files are first checked for proper FASTQ formatting, and then trimmed. Then both trimmed and untrimmed files are analyzed with FastQC. The trimmed files are also checked for contaminants and all the reports are aggregated in a single file using multiQC. After the genome has been indexed, the trimmed files are then aligned on the genome of choice and bam files are used to create TAG directories compatible with downstream processing (Duttke et al., 2019). Stranded Bedgraph files are made for visualization purposes and transcription initiation peaks are called using both the Start-seq data and input data. Finally, bed files are made containing the coordinates of the called peaks.

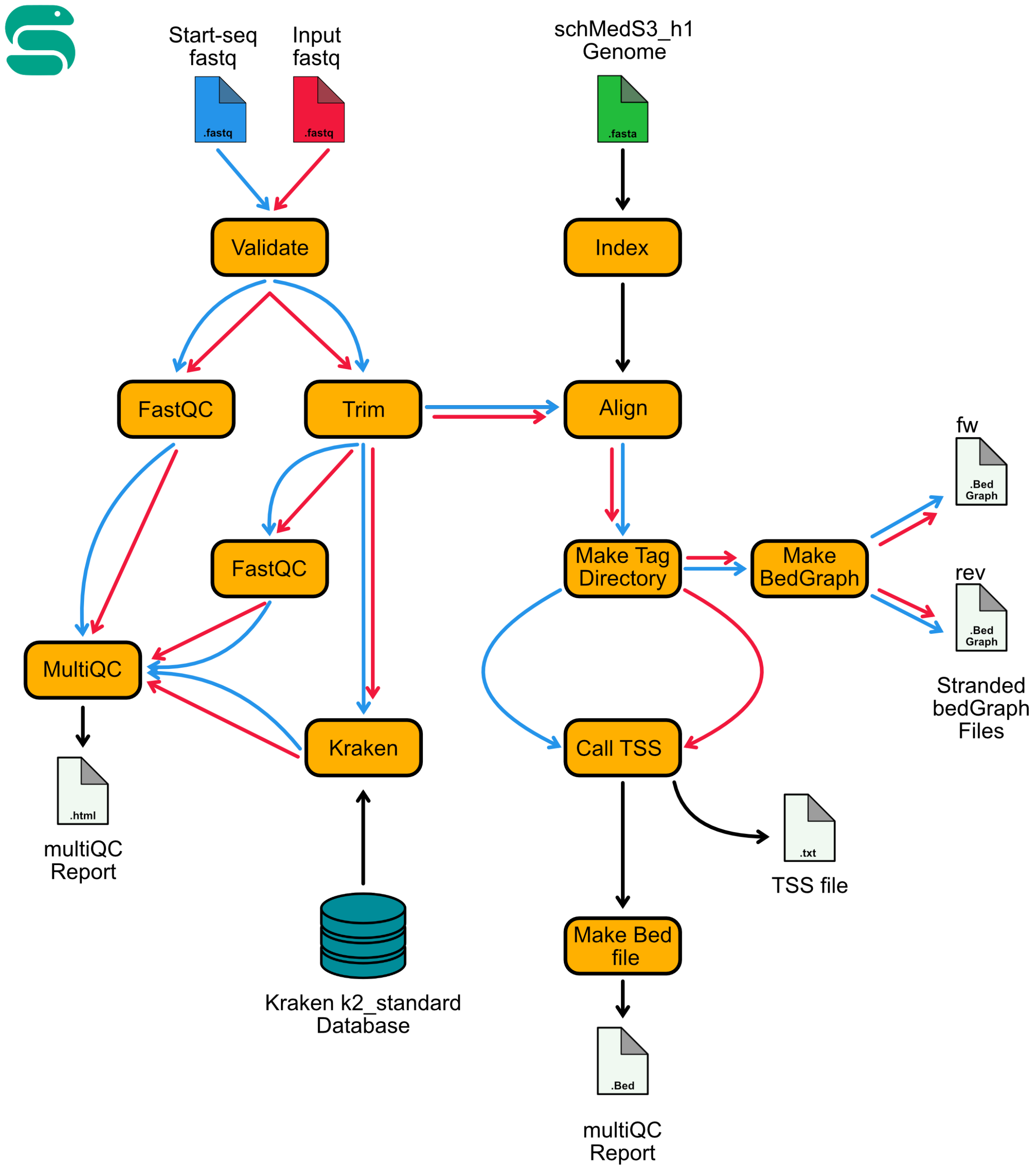


Figure 3.5: Custom workflow for the processing of Start-seq libraries until TSS calling. The blue and red arrows represent the successive actions taken on the Start-seq and input .fastq files respectively

Each library was sequenced at a depth of about 60 Mio reads (table 3.3). The average mapped read size for each library was 58 nt. The unique mapping rate of the Start-seq libraries was fairly low, primarily due to the read size, and was on average 37%. The input libraries had an average unique mapping rate of 29%.

Table 3.3: Sequencing and mapping statistics for the Start-seq and Input libaries prepared for asexual *S. mediterranea*

	Input 1	Input 2	Input 3	Start 1	Start 2	Start 3
Input reads	85165429	78990716	54313501	62246398	60167758	60213906
Read lenght (nt)	59	64	54	54	58	59
Uniquely mapped (%)	30,66	26,84	32,36	40,86	34,74	36,59
Multi-mapping (%)	43,7	52,15	43,2	18,15	15,47	13,91
Unmapped (%)	24,61	20,44	22,62	40,43	49,24	49,04

3.2 Characterization of the transcription initiation landscape in *S. mediterranea*

After having verified the reproducibility of the protocol, I decided to characterize the transcription initiation clusters. My aim was to argue that identified TICs are putative REs by showing that they are mostly found within non-coding regions (Davidson, 2010b), possess identical chromatin features (Buenrostro et al., 2013; Creyghton et al., 2010; Local et al., 2018; Santos-Rosa et al., 2002; Z. Wang et al., 2008), contain core promoter motifs (Andersson et al., 2014; Lenhard et al., 2012) and transcribe in a bidirectional fashion (Core et al., 2014; Kim et al., 2010).

3.2.1 Quality control of transcription initiation data

To arrive to the final set of transcription initiation clusters, I performed an iterative peak merging approach (Grandi et al., 2022) on the different biological replicates (see methods). Using this method, I identified 55213 transcription initiation clusters (TICs) in asexual *S. mediterranea*. Next, I assessed pairwise correlation between biological replicates to ensure replicability of the assay. All replicates showed high concordance (Figure 3.6 A). To estimate the appropriate sequencing depth for this assay, I merged reads from all biological replicates together and then sub-sampled reads corresponding to 1, 2, 5 ,10 ,25, 50, 75 and 100% of the total amount of reads. I then re-ran my data processing pipeline and plotted the number of peaks called for each sub-sample (Figure 3.6 B). Results show that the number of peaks called saturates at around 20 Mio uniquely mappable reads. The distribution of peaks around gene models was then investigated. Most peaks (54,94%) are found in intergenic regions which correspond to enhancers (Figure 3.6 C). The second largest category were promoters (25,77%) followed by intronic peaks (14,26%). Only few peaks were found in protein coding regions.

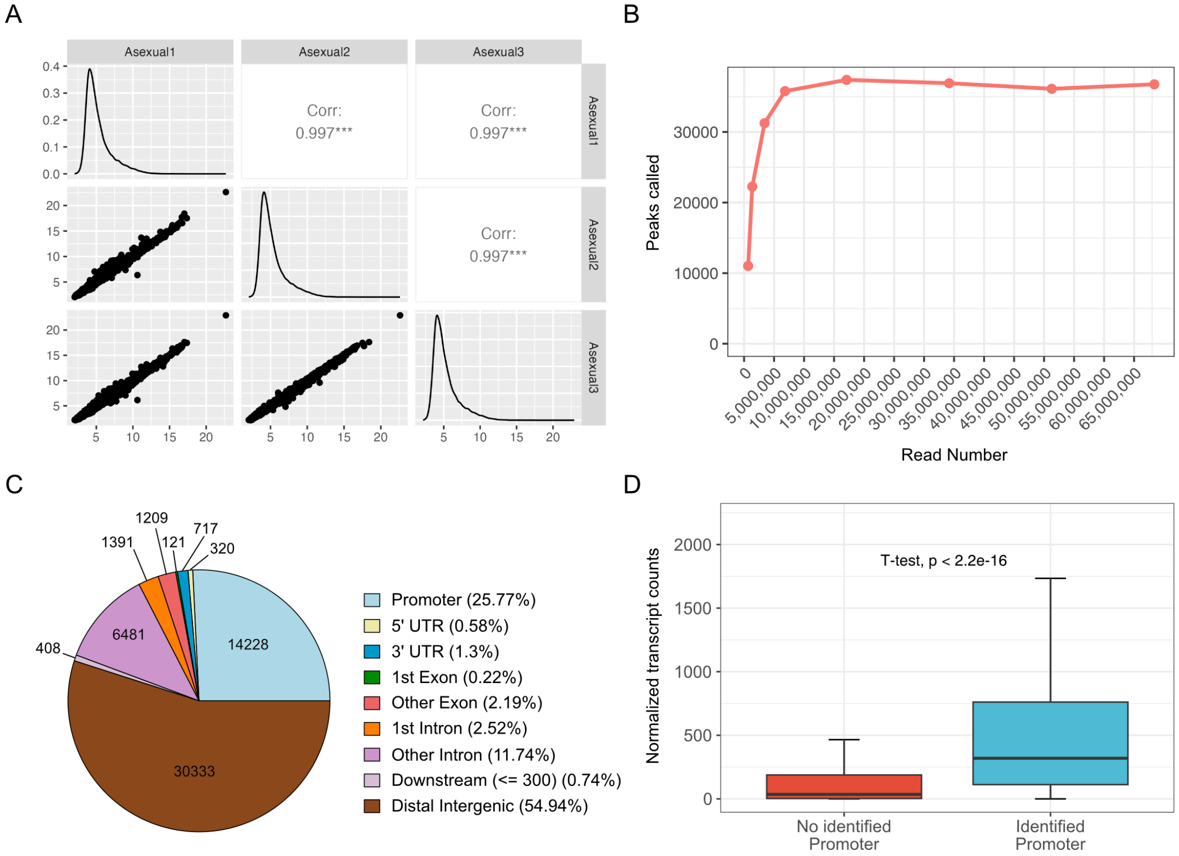


Figure 3.6: Quality control and initial characterization of TICs identified by Start-seq in *S. mediterranea.* A) generalized pairs plot for Start-seq replicate concordance assessment. Line plots are the density plots representing the distribution of TIC in relation to regularized log transformed read number for each TIC. Scatter plots represent the pairwise Rlog transformed read number found for individual TICs. Pearson correlation coefficients represent the pairwise correlation between two biological replicates. B) Subsampling analysis for the identification of TIC number at 1%, 2%, 5%, 10%, 25%, 50%, 75% and 100% of Start-seq reads. C) Pie chart of the distribution of identified TICs around genes in asexual*S. mediterranea.* D) Quantification of transcript abundance differences using RNA-seq data between genes with and without an identified promoter TIC after peak calling using Start-seq data.

I noticed that only about half of annotated genes possessed an annotated promoter in my assay. I therefore used RNA-seq data from comparable samples published by Davies and colleagues (Davies et al., 2017) to investigate the expression level of such genes. When dividing genes in two sets depending on whether a Start-seq promoter was called or not, I could show that the genes without an identified promoter were expressed significantly lower (t-test, pval $< 2.2e-16$) with a median close to 0 (Figure 3.6 D). This points towards the fact the majority of genes with no identified Start-seq promoter are not expressed in whole asexual *S. mediterranea*.

3.2.2 Visualization of transcription initiation data

I also uploaded the Start-seq data on the lab's instance of the UCSC genome browser to look at TICs in their genomic context. Several other genome wide assays such as ATAC-seq and ChIP-seq had been already performed on *S. mediterranea* and can give an idea of whether identified TICs are located withing their expected environment. Active REs are surrounded by active histone marks, such as H3K4me3 and H3k27Ac, and open chromatin. I therefore expected to find TICs surrounded by such environment. I also looked at whether the identified TICs showed the different types of promoters and enhancers described in the literature with regards to directionality of transcription initiation and focus (Figure 3.7).

Genes had often two divergent TICs close to the beginning of the annotated transcript start site that could be overlapping (Figure 3.7 A & B) or not (Figure 3.7 C). Interestingly, convergent transcription (Henriques et al., 2018) where TICs on opposite strands were facing each other could also be seen (Figure 3.7 C). I could also see sharp (Figure 3.7 A) and broad TICs (Figure 3.7 C) where most of the reads were focused on a few nucleotides or spread out over the whole TIC length respectively. TICs within introns or in intergenic regions, were indicative of putative enhancers. Enhancers also had reads mapping in opposite direction. I could also find many examples where two divergent TICs were called (Figure 3.7 C & D). They could, as promoter TICs, be focused (Figure 3.7 A) or broad (Figure 3.7 D).

Overall, the overlap with open chromatin and active histone marks was very visible. TICs were often located within a region of open chromatin (Figure 3.7 A & D). Promoter TICs were usually surrounded by both H3K4me3 and H3k27Ac marks (Figure 3.7 A & C) while enhancer TICs lacked the H3K4me3 mark (Figure 3.7 D).

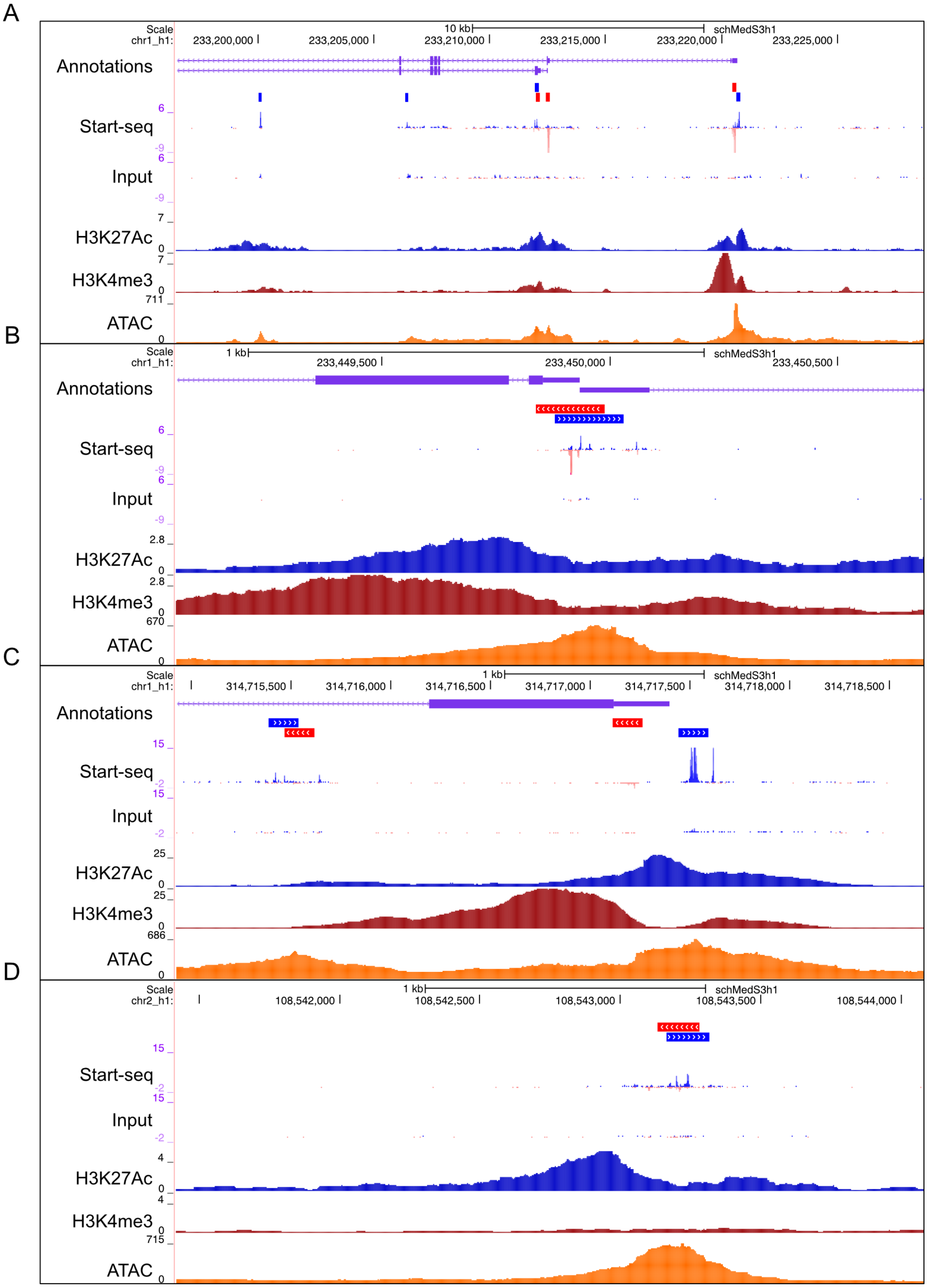


Figure 3.7: Start-seq identifies different promoter and enhancer architectures and their location in the chromatin environment. Example of bidirectional and unidirectional putative promoters and unidirectional intronic putative enhancers. A) Example of bidirectional overlapping putative promoters with genes in opposite orientation. B) Example of sharp and broad TICs and a putative enhancer with convergent transcription. C) Example of a putative bidirectional enhancer situated in an open chromatin region and surrounded by H3K27Ac signal but with no visible enrichment for H3K4me3

3.2.3 Characterization of the chromatin landscape around transcription initiation clusters

To better characterize TICs, I next decided to systematically quantify the chromatin landscape around them. I therefore used previously generated ChIP-seq and ATAC-seq data from the lab and other sources (Ivankovic et al., 2023; Mihaylova et al., 2018).

I generated a custom pipeline to map ATAC-seq, H3K4me3, H3k4me1 and H3K27ac on the newly generated *S. mediterranea* genome (Ivankovic et al., 2023) (see Material and Methods). I then averaged the signal of previously mentioned data as well as the Start-seq data over all TICs (Figure 3.8 A), TICs classified as promoters (Figure 3.8 B) and enhancers (Figure 3.8 C) (see Material and Methods).

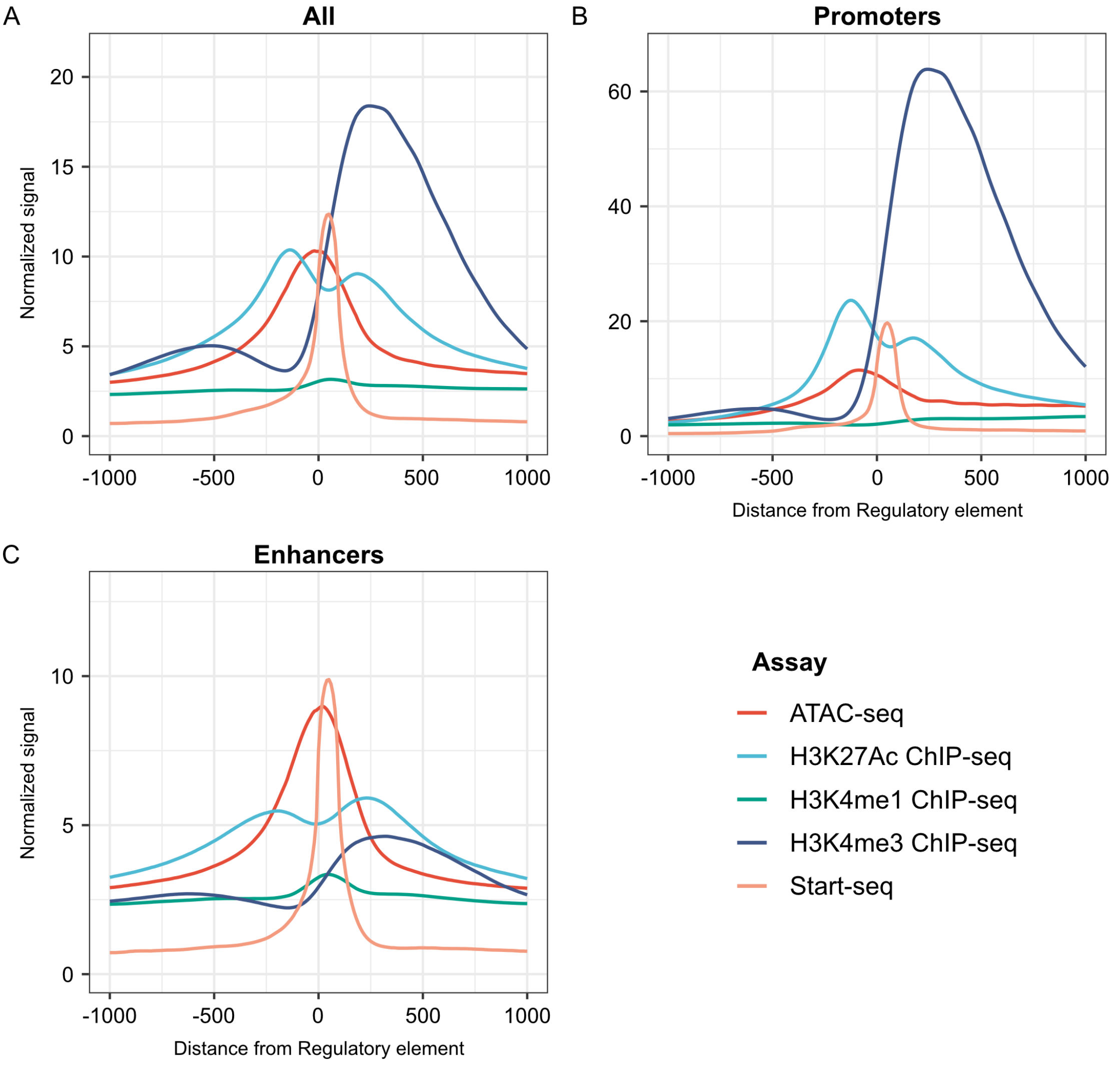


Figure 3.8: Characterization of the chromatin environment around identified TICs. Average representation of the chromatin environment around (A) all identified TICs, (B) putative promoter TICs and (C) putative enhancer TICs. Normalized signal for open chromatin (ATAC-seq), transcription initiation (Start-seq) and active chromatin marks (H3K4me3 and H3K27Ac) are represented using line plots.

When looking at the whole set of putative REs (Figure 3.8 A), Start-seq TICs were situated within a region of open chromatin and surrounded by the active chromatin marks H3K4me3 and H3K27ac. A very faint but visible increase in H3K4me1 was also visible. The H3K27ac mark showed a clear bimodal distribution around the ATAC-seq summit as well as the Start-seq signal. The upstream H3K27ac peak was slightly higher than the downstream one. The H3K4me3 signal showed a bimodal distribution as well but had a much higher enrichment downstream of the TIC in the sense of transcription. The upstream H3K4me3 peak was also broader compared to the downstream peak and positioned upstream of the first H3K27ac peak.

When only analyzing promoter TICs (Figure 3.8 B), the H3K4me3 mark became much more enriched but with a similar profile. Higher H3K27ac signal could also be observed but showed also a similar pattern, with the upstream peak being higher than the downstream peak. The ATAC-seq signal was also more broadly distributed. A slight increase of Start-seq signal was visible at around -500 bp, hinting towards upstream antisense transcription at promoters (sometimes referred to as PROMPTs or uaRNAs).

Performing the same analysis on enhancer TICs (Figure 3.8 C) showed that the H3K4me3 was highly reduced compared to the promoter TICs. The H3K27ac signal still showed a bimodal distribution around the Start-seq signal as well as the ATAC-seq signal. I also noted that the H3K27ac peaks were more even in enhancers compared to promoters. Finally, the H3K4me1 signal, showed a single peak at enhancer TICs that was located close to the center of the TIC and interestingly, nearly perfectly aligned with the Start-seq summit.

3.2.4 Assessing bidirectional transcription initiation using Start-seq

Since regulatory elements have been described to initiate transcription bidirectionally (Core et al., 2014; Kim et al., 2010), I assessed whether Star-seq data showed evidence of bidirectional transcription both at the promoters as well as enhancers. Transcriptional bias in the sense orientation, i.e. the orientation of the called peak, was clearly visible at promoters (Figure 3.9 A). However, I found that there was also antisense transcription initiation. This signal was more dispersed and much weaker than the sense signal. Interestingly, strong promoter TICs showed more antisense transcription than weaker promoter TICs. Similarly at enhancer TICs, Start-seq data showed a clear bias towards the sense strand but also showed transcription initiation signal in the antisense orientation. Interestingly, the directional bias was less defined in enhancer TICs compared to promoter

TICs. Moreover, enhancer TICs with stronger signal in the sense direction had a more diffuse signal in the antisense orientation that tended to move away from the sense signal (Figure 3.9 B). This was not observed for promoter TICs where all antisense transcription was diffuse. In total, the peak calling algorithm classified 52,7% of promoter TICs as bidirectional whereas 59,6% of enhancer TICs passed the threshold to be classified as bidirectional.

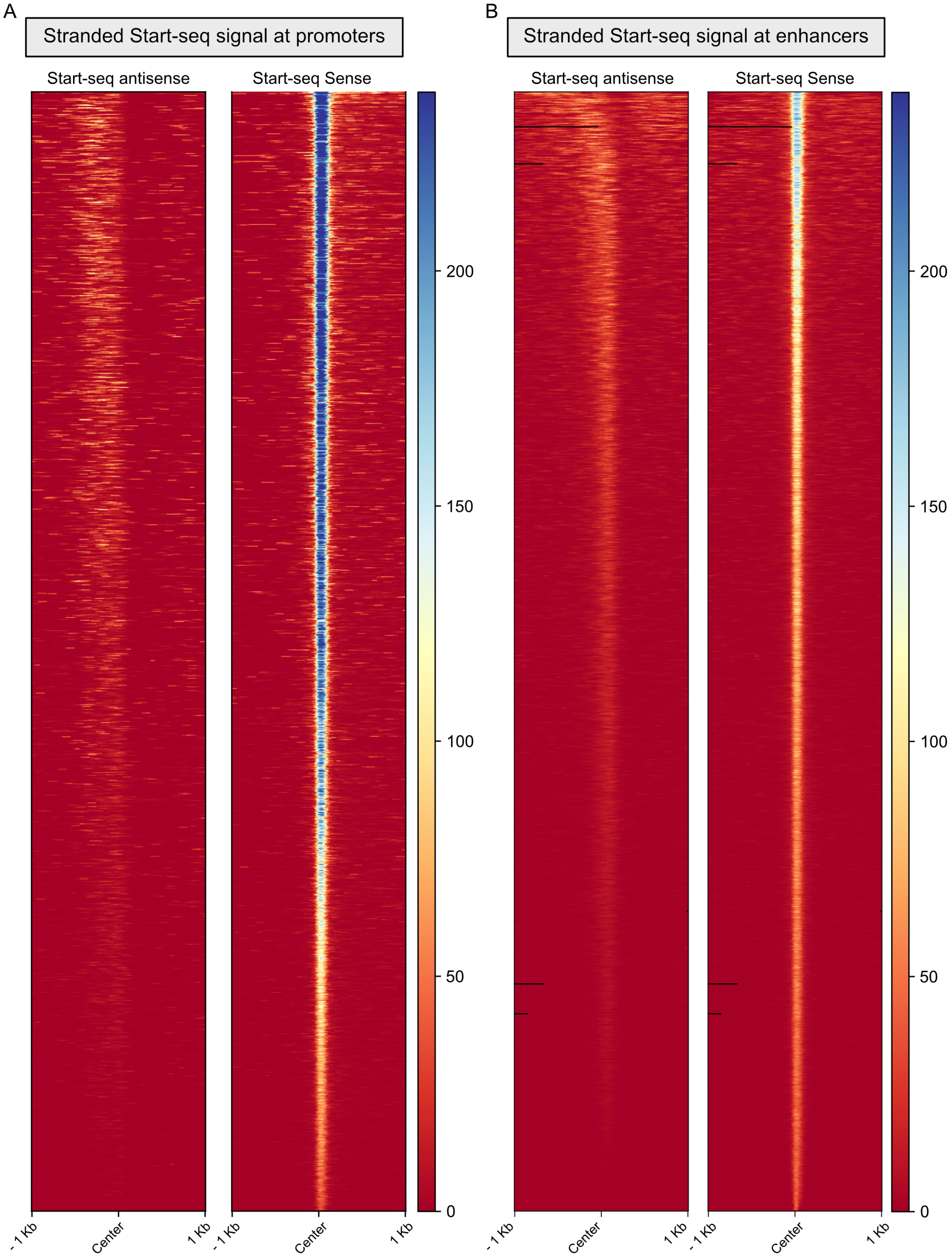


Figure 3.9: Start-seq identifies widespread bidirectional transcription at *S. mediterranea* putative regulatory elements. Bidirectional transcription initiation visualization at (A) putative promoter and (B) enhancer TICs. The heatmaps display forward-stranded putative promoters and enhancers ranked by forwards Start-seq signal.

3.2.5 Motif content of transcription initiation clusters

I next investigated whether core promoter motifs and widely present motifs other different eukaryotes such as the SP1 (Wierstra, 2008) motif and the CCAAT-box (Nardone, Chaves-Sanjuan, & Nardini, 2017) were present in or around the identified TICs. I tested 10 motifs in total (Figure 3.10 C). Many of the motifs were not enriched around the TICs except for the Downstream Promoter Element (DPE) (Burke & Kadonaga, 1997), the Initiator motif (Inr) (Smale & Baltimore, 1989), the CCAAT-box (Benoist, O'hare, Breathnach, & Chambon, 1980) and the TATA-box (Lifton et al., 1978) (Figure 3.10 A).

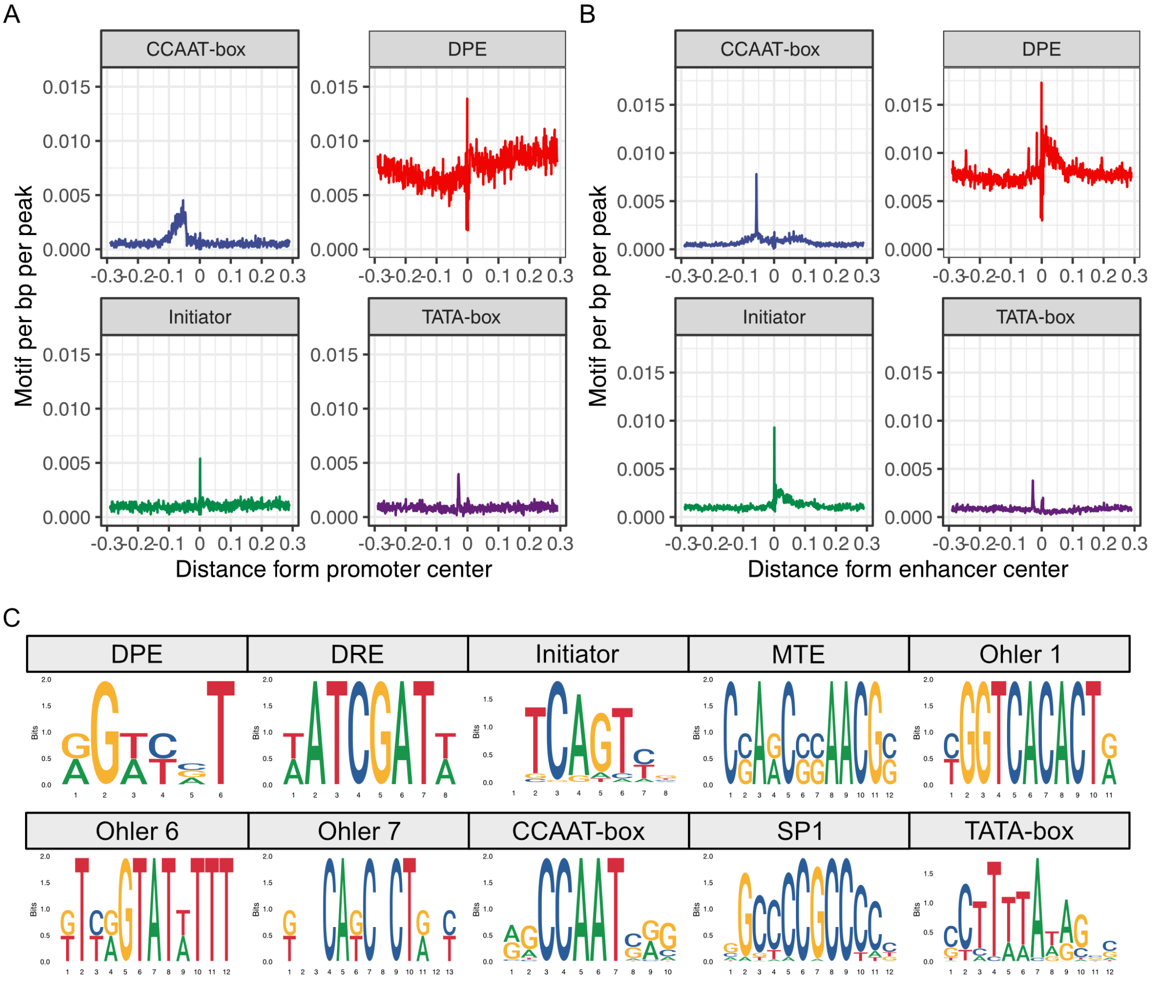


Figure 3.10: Characterization of the motif content at putative regulatory elements in *S. mediterranea*. Line plots showing the abundance of the CCAAT-box, DPE, Initiator and TATA-box motifs relative to the TIC center distance in (A) putative promoter and (B) enhancer TICs. (C) Position frequency matrices of all probed motifs used for identification represented as LOGO plots.

I first looked at the promoter TICs. Interestingly, I found that the DPE motif, usually found downstream relative to the transcription initiation site (± +30 nt), was most enriched right at the center of the TIC (-1 nt) (Figure 3.10 A). The initiator motif was found properly positioned at the transcription initiation site (+1 nt). The CCAAT-box was enriched between -60 and -100 nt as expected (Nardone et al., 2017). Finally, the same could be said for the TATA-box, that was enriched at ± -30 nt. When performing the same analysis on the enhancer TICs, I could see some motif distribution differences compared to the promoter set (Figure 3.10 B). In addition to what I described for the promoters, the results showed that the Inr motif was also enriched within the first 25 nt after the TIC. The CCAAT-box displayed a high enrichment downstream of TICs (-57 nt) compared to a more dispersed distribution in promoter TICs. Finally, the TATA-box showed an increase of signal at -29 nt but also at the center of TICs (0±4 nt).

I next investigated whether these enriched promoters had a certain orientation preference relative to TICs (Figure 3.11). Interestingly, the CCAAT-box showed a strong orientation preference in enhancers (Figure 3.11 B) but not promoters (Figure 3.11 A). The DPE motif showed a clear sense orientation in both promoter and enhancers at the center of TICs. The Initiator motif was also found mostly on the sense strand except for the downstream portion enriched in the enhancer TICs where it showed a reverse orientation. Finally, the TATA-box was mostly found both in promoter and enhancer TICs in the antisense direction compared to the TIC orientation.

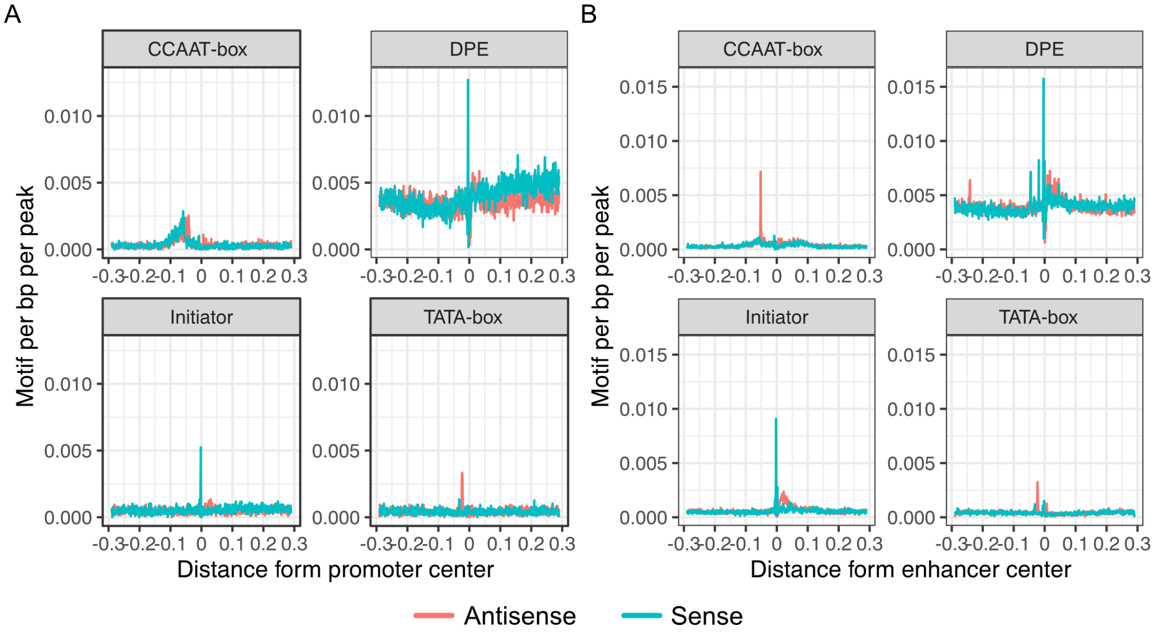


Figure 3.11: orientation preference of the enriched motifs at putative regulatory elements in *S. mediterranea*. Line plots showing the abundance and orientation of the CCAAT-box, DPE, Initiator and TATA-box motifs relative to the TIC center distance in (A) putative promoter and (B) enhancer TICs.

In conclusion, Star-seq TICs possess similar characteristics as regulatory elements in terms of distribution, chromatin environment and motif content.

3.3 Differential regulatory element activity analysis between sexual and asexual biotypes

Schmidtea mediterranea comes in two different biotypes (Figure 3.12 A). The asexual biotype only reproduces by fission while the sexual biotype is able to perform sexual reproduction. For this, it possesses a variety of organs and tissues that enable sexual reproduction and are only found in this biotype. I hypothesized that the development and maintenance of these tissues rely on the activation of gene regulatory networks specific to these tissues (Figure 3.12 B). My aim was to uncover key element of these hypothetical GRNs by comparing the activity of REs in those two biotypes. REs part of these GRNs should only be active in the sexual biotype and could be identified by differential regulatory element activity analysis.

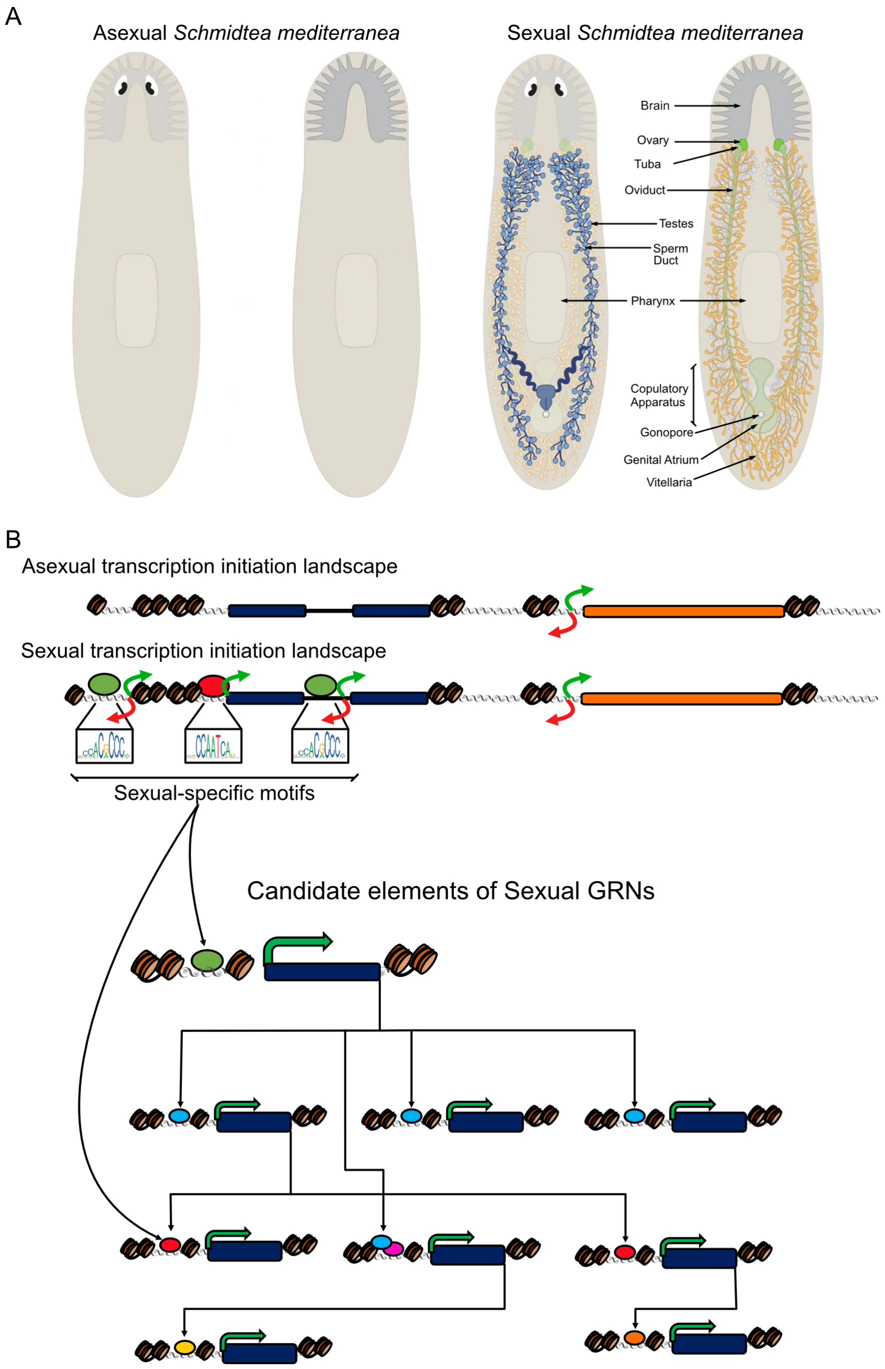

A
Asexual Schmidtea mediterranea
Sexual Schmidtea mediterranea
Brain
Ovary
Tuba
Oviduct
Testes
Sperm Duct
Pharynx
Copulatory Apparatus
Gonopore
Genital Atrium
Vitellaria
B
Asexual transcription initiation landscape
Sexual transcription initiation landscape
Sexual-specific motifs
Candidate elements of Sexual GRNs

Figure 3.12: Identification of gonad specific GRN components by determining the enriched motif content of differentially active regulatory elements. A) Illustration of the two *S. mediterranea* biotypes adapted from (Issigonis et al., 2022). The sexual biotype possesses gamete producing gonads consisting of the testes and ovaries. In addition, it also has a series of accessory reproductive organs such as the yolk glands (or vitellaria). The asexual biotype lacks all these organs B) Depiction of the experimental paradigm for the identification of important GRN components in the planarian gonad. Identification of differentially active regulatory elements between sexual and asexual *S. mediterranea* informs about the motifs enriched in the transcription initiation landscape of sexual *S. mediterranea*. In turn, specific motifs allow for the identification of important transcription factors that act on gene regulatory networks required for the development and maintenance of the planarian gonad.

3.3.1 Correlation analysis of transcription initiation data and transcriptomic data

I generated sexual *S. mediterranea* Start-seq libraries and processed them similarly to the asexual datasets. In order to get a peak set encompassing both sexual and asexual TICs, I again performed an iterative merging approach to arrive to a final peak set and quantified the read number per peak for each replicate. In total, I identified 75963 non-overlapping stranded TICs shared between the 6 samples.

I then looked at the concordance between the replicates of both biological conditions by making a clustered heatmap of Euclidean distances between samples and also by PCA (Figure 3.13 A & C). In both analyses, samples separated clearly by biological condition. I compared my results with results of the same analysis form previously used RNA-seq data on sexual versus asexual worms form Davies and colleagues (Figure 3.13 B & D). A similar trend was found there as well, where samples from the same biological condition clustered clearly together. Principal component 1 showed a clear sexual-asexual axis in both cases but explained more of the variance in the RNA-seq dataset compared to Start-seq (Figure 3.13 C & D).

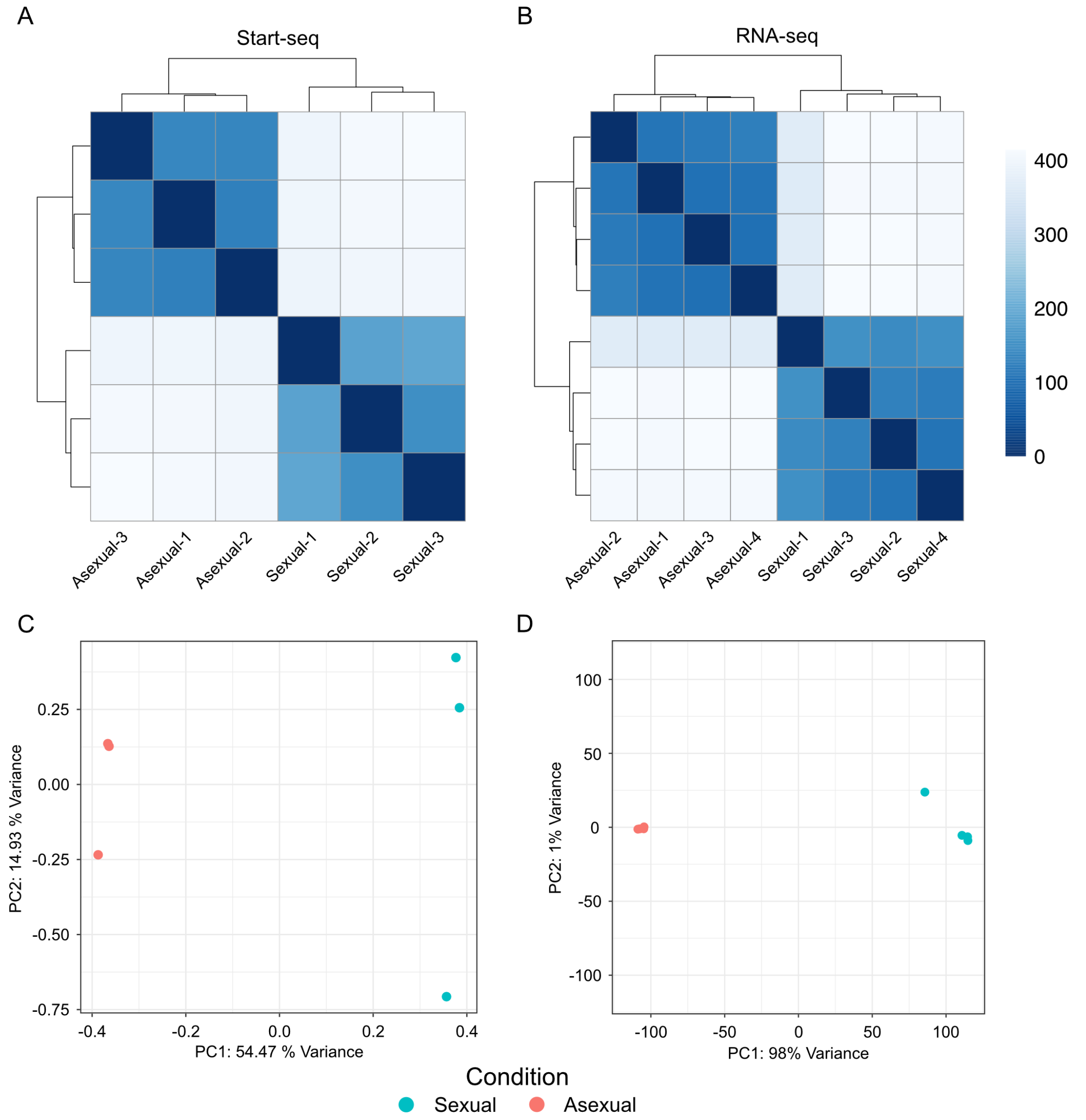


Figure 3.13: Quality control on Start-seq and RNA-seq samples for *S. mediterranea* biotypes. A) Clustered heatmap of Start-seq samples for the sexual and asexual biotypes. The scale is shared with (B) and represents the pairwise Euclidean distance between the samples. B) Clustered heatmap of RNA-seq samples for the sexual and asexual biotypes originally published in (Davies et al., 2017). C) Principal component analysis plot of the Start-seq samples for the sexual and asexual biotypes. D) Principal component analysis plot of the RNA-seq samples for the sexual and asexual biotypes originally published in (Davies et al., 2017).

I next wanted to explore whether differences in transcript expression would be visible in Start-seq data by looking at differential promoter activity. I therefore plotted the log2 fold-change between sexual and asexual worms of all significantly differentially expressed transcripts (padj < 0.05) versus the log2 fold-change of significantly differentially expressed promoters (padj < 0.05) associated to a gene present in both datasets (Figure 3.14 A). I found that there was a significant positive correlation between Start-seq and RNA-seq results ($\mathrm{R} = 0.64$, $\mathrm{p} < 2.2 * 10^{-16}$). Most of the genes were situated in the first and third quadrant (upper right and lower left), indicating log2 fold-changes with the same sign. From those, many were situated in the first quadrant as expected since they represent upregulated genes in the sexual biotype, most likely belonging to the genes only expressed in tissues involved in sexual reproduction. For the genes with opposite log2 fold-change signs, the majority were situated very close to the origin and had both a small transcript and promoter log2 fold-change.

Next, I investigated the overlap between differentially regulated transcripts and REs associated to the same gene (Figure 3.14 B). In total, 16821 genes were found to have a differentially expressed transcript and/or at least a differentially active RE. The largest set (43%) contained the genes present in both data sets. Additionally, most of the genes (65.4%) with a differentially regulated transcript had differentially regulated RE associated to it. There were however many genes that had differentially regulated REs but showed no differentially regulated transcript.

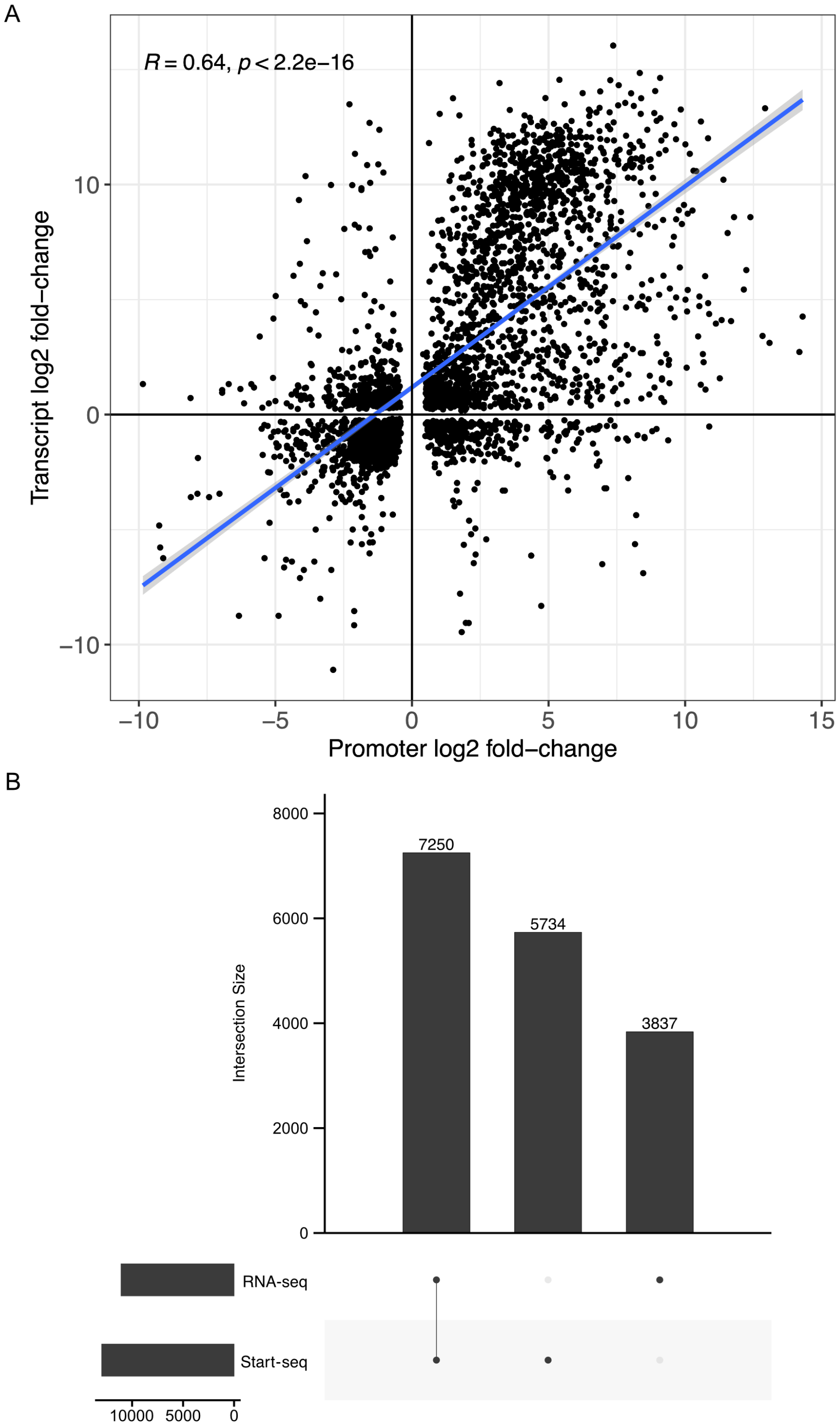


Figure 3.14: Concordance analysis between RNA-seq and Start-seq for differential expression in *S. mediterranea* biotypes. The RNA-seq data was originally published by (Davies et al., 2017) A) Scatter plot showing the significantly differentially expressed genes having a differentially expressed promoter in the Start-seq data. Blue line represents the linear model best representing the relationship between promoter and transcript log2 fold-change. The confidence interval is represented in gray. B) Upset plot showing the overlaps between differentially regulated genes and regulatory elements associated to genes.

3.3.2 Differential regulatory element activity analysis

To visualize the overall number of significantly up- and downregulated REs in sexual *S. mediterranea*, I generated a volcano plot from the differential expression analysis results (Figure 3.15). Out of the 75963 REs tested, 13612 REs were significantly differentially active (abs(log2 fold-change) $>$ 2, adjusted p-value $<$ 0.01) among which 10292 (75,6%) and 3320 (24,4%) were significantly up- and downregulated respectively. Using these thresholds, I found 469 downregulated and 1753 upregulated promoters whereas the rest of significant REs were classified as enhancers. As a sanity control, I then investigated whether upregulated promoters were linked to genes with known functions related to sexual reproduction. I performed a BLAST search on the genes of the top 100 most upregulated promoters. Many genes (n =43) did not have an annotated BLAST hit. From the 57 remaining, I could readily assign 15 genes to reproduction-related functions. Other genes could be attributed to metabolism, immune-related functions and transport proteins. Interestingly, promoters associated to yolk marker genes were highly present in the most upregulated fraction of promoters.

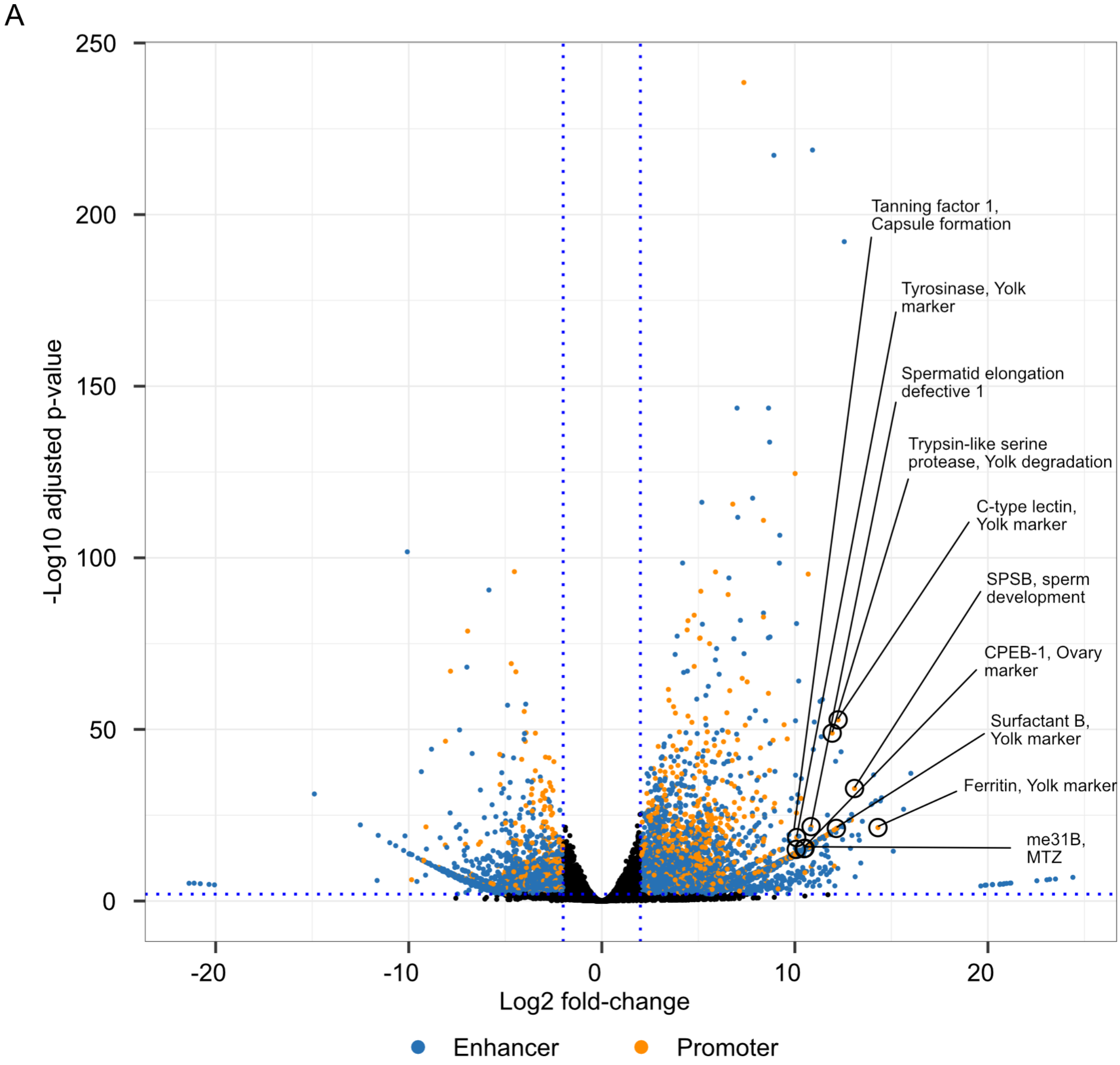


Figure 3.15: Volcano plot of the differential regulatory element analysis between the sexual and asexual biotypes in *S. mediterranea*. Significant enhancers are colored in blue while significant promoters are in orange. Several upregulated promoters linked to genes with a described function in sexual reproduction are annotated. The threshold for significance was set to padj < 0.01 and absolute l2fc > 2.

To generalize my findings to all genes with an upregulated promoter, I performed a Gene-Ontology enrichment analysis (Figure 3.16). A dot-plot representation of the results showed that the 'Microtubule-based movement' GO-term, likely referring to sperm motility, had the most upregulated genes (Figure 3.16). I found also many terms related to the transport of several molecules. Finally, some terms referring to metabolism and immunity were also found, similarly to the BLAST analysis performed before. Visualizing the results in the form of a cnet plot showed the associations between GO terms and upregulated genes (Figure 3.16 B). I found that all the transport-associated terms clustered together since many of them shared the same genes. A similar clustering of GO-terms could also be found for the immune-related terms and metabolic terms. Only the microtubule-based movement term did not share any genes with other terms. I also investigated whether the significantly downregulated promoters showed any enrichment for certain processed but none could be found (not shown.)

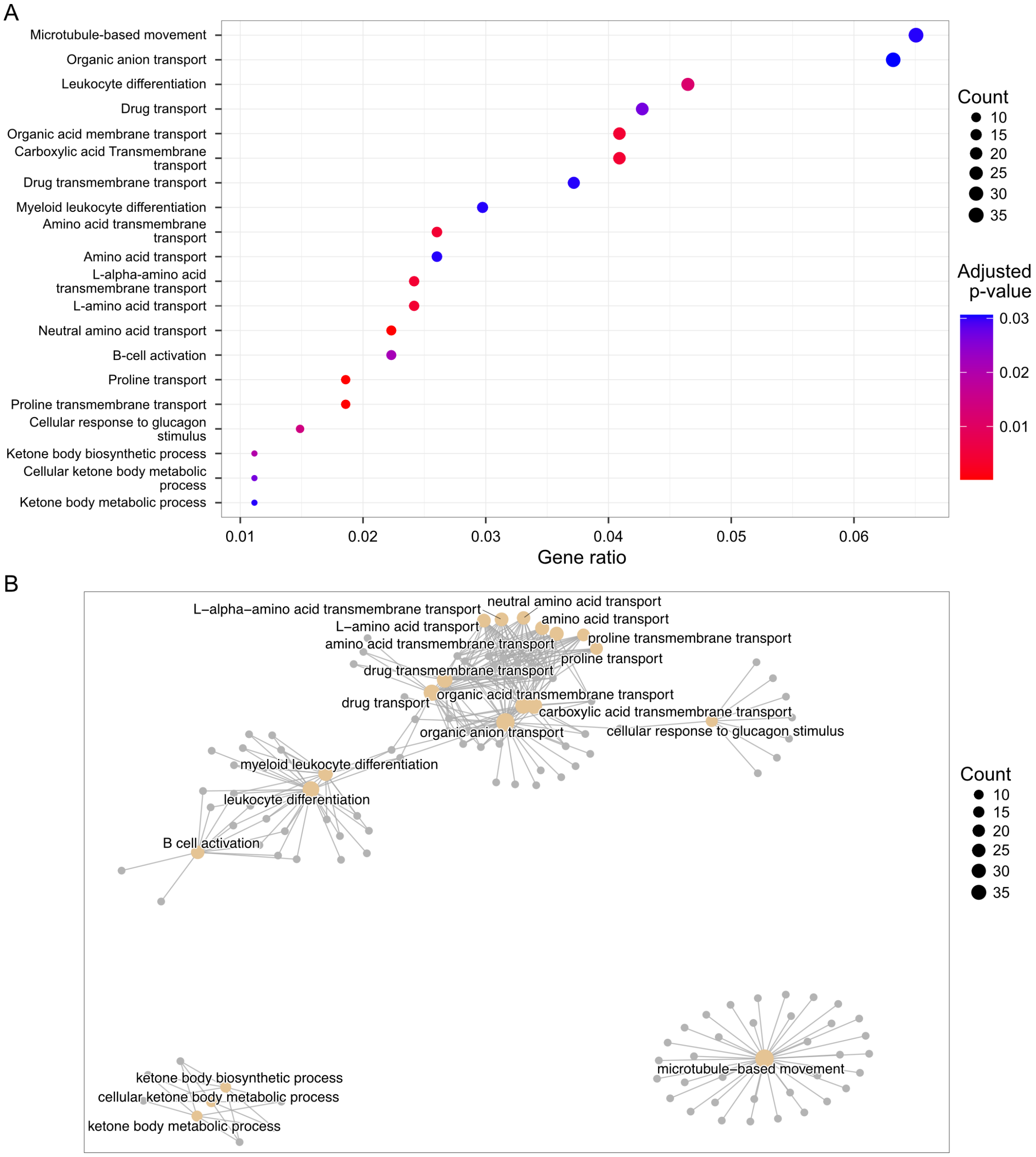


Figure 3.16: GO enrichment analysis for genes of significantly upregulated promoters in sexual *S. mediterranea*. A) Dot plot representation of the top 20 significant terms associated to of significantly upregulated promoters in sexual *S. mediterranea*. B) Cnet plot representation of the top 20 significant terms (brown dots) associated to genes (gray dots) with significantly upregulated promoters in in sexual *S. mediterranea*.

3.3.3 Motif variability analysis

Having established that upregulated REs are associated to genes with sexual reproduction-related functions, I next sought to identify the transcription factors involved in their regulation. These candidate TFs would therefore be putative elements of GRNs involved in the development and maintenance of the planarian reproductive apparatus. I performed this analysis using the chromVAR R package (A. N. Schep et al., 2017). I chose the JASPAR 2018 PHYLOFACTS as transcription factor database since, although smaller, it only contains evolutionarily conserved motifs (Vlieghe et al., 2006). I rationalized that this would be preferable to identify TF motifs in a distant species such as *S. mediterranea*.

The analysis revealed 10 significantly variable motifs belonging to 9 different TF families (Figure 3.17 A). The only TF family represented twice in the significant motifs was the NFY family. Interestingly, this family as well as the FOX and KLF family, had already been described to play an important role in planarian sexual reproduction (Figure 3.17 B-D). Planarian NF-YB was shown to be important for planarian spermatogonial stem cell proliferation (Iyer, Collins III, & Newmark, 2016). Moreover, a *foxL* homolog is important for oocyte differentiation (U. W. Khan & Newmark, 2022) and a Kruppel-like factor, *klf4l*, is expressed in primordial germ cells and yolk cell progenitors and regulates their survival (Issigonis et al., 2022). The other families with significantly variable motifs were the, THAP, CEBP-related, NF-κB-related, C4-GATA-related, JUN:FOS-related and EBF families. Additionally, motifs associated to the TEF-1-related (comprising the TEAD TFs) and More-than-3-adjacent-zinc-fingers (more specifically from the Snail-like subfamily), TF families were significantly variable in a similar analysis done on the previous *S. mediterranea* genome (Grohme et al., 2018) (not shown). Unfortunately, the last two motifs were not found significantly variable using the newer version of the assembly. However, I still decided to use these TF families in the downstream analyses. Finally, I decided to focus on the CEBP-related (referred to as CEBP), NF-κB-related (referred to as NFK), C4-GATA-related (referred to as GATA), THAP, Snail-like (referred to as SNAIL) and TEF-1-related (referred to as TEAD) families.

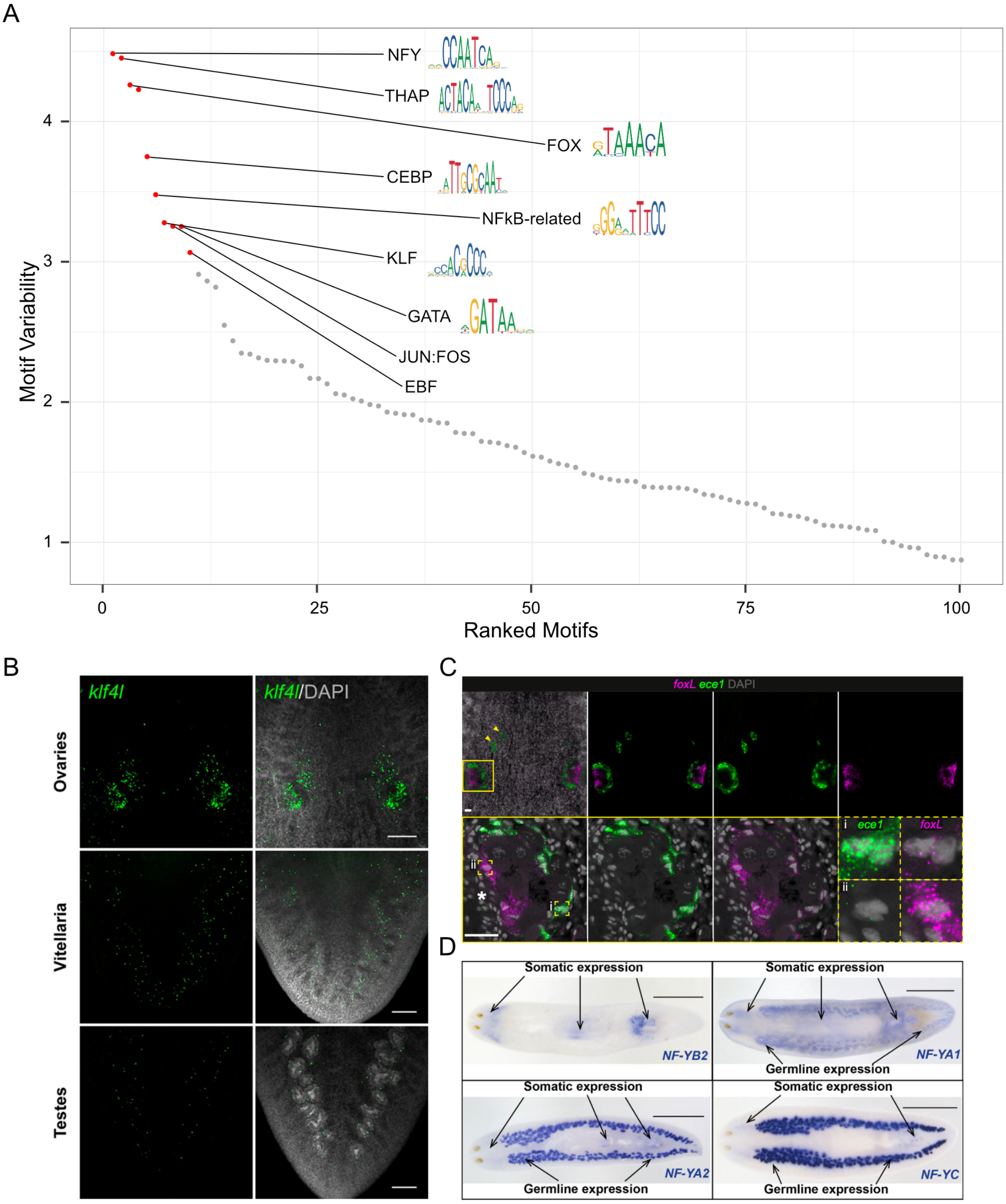
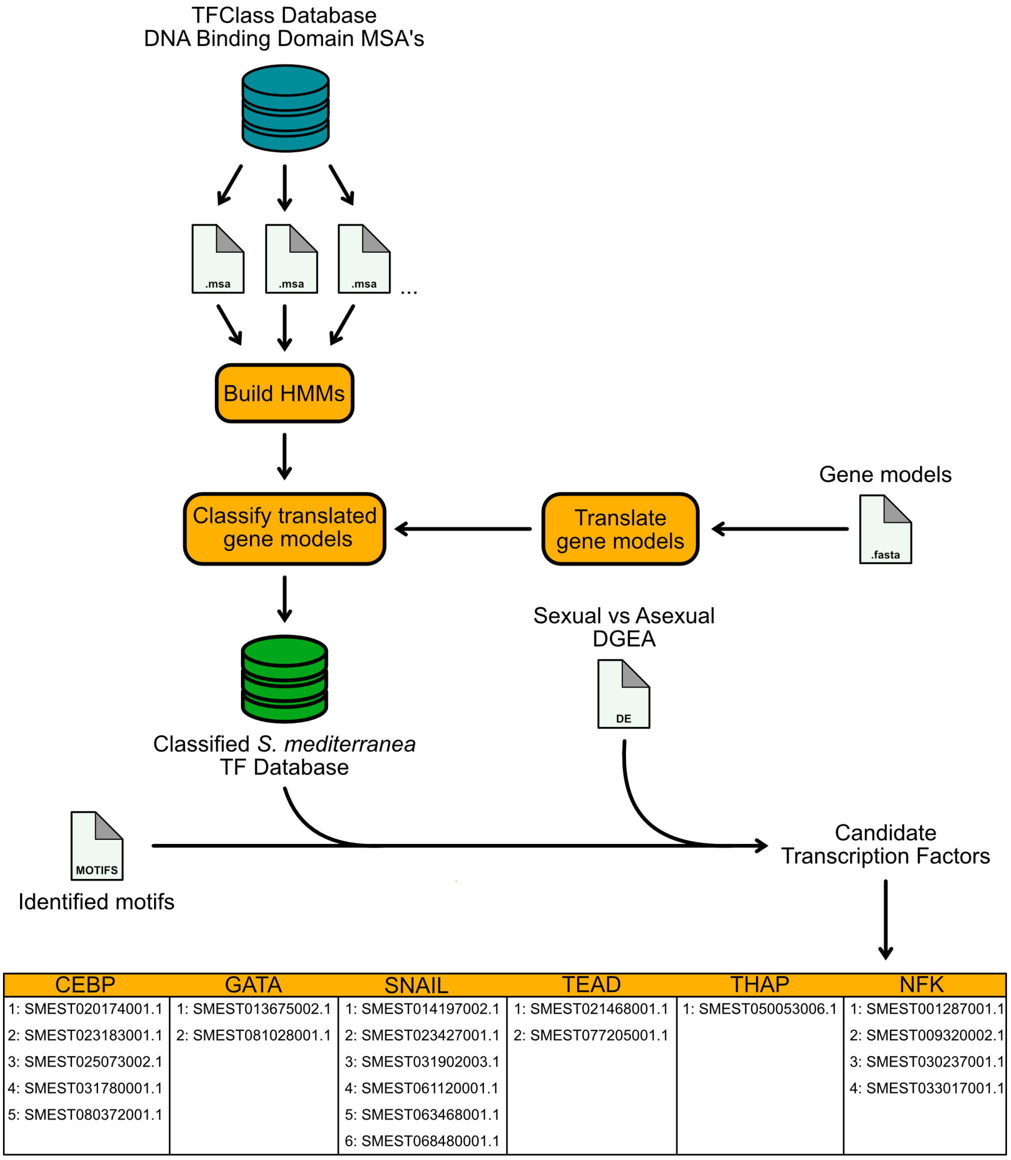


Figure 3.17: Motif variability analysis between sexual and asexual *S. mediterranea* biotypes. A) chromVAR results showing the top 100 most variable motifs between *S. mediterranea* biotypes. The significantly variable motifs are shown in red and annotated with their transcription factor family name by TOMTOM. B) Result from (Issigonis et al., 2022) showing that a KLF family transcription factor, *klf4l*, is expressed in the presumptive germline stem cells and yolk cell progenitors of Sexual *S. mediterranea*. C) Original result from (U. W. Khan & Newmark, 2022) showing that a FOX family transcription factor, foxL, is expressed in the somatic ovary of Sexual *S. mediterranea*. D) Original result from (Iyer et al., 2016) showing that several NFY family transcription factor members are expressed in the testes sexual *S. mediterranea*

In order to arrive to a set of candidate transcription factors, I teamed up with Dr. Andrei Rozanski to generate a planarian transcription factor database (see Material and Methods) (Figure 3.18). I then used DGEA results from the sexual versus asexual data from Davies and colleagues to refine the selection of my candidates (see Material and Methods). In total, 20 candidates were selected across the 6 TF families.



CEBP	GATA	SNAIL	TEAD	THAP	NFK
1: SMEST020174001.1	1: SMEST013675002.1	1: SMEST014197002.1	1: SMEST021468001.1	1: SMEST050053006.1	1: SMEST001287001.1
2: SMEST023183001.1	2: SMEST081028001.1	2: SMEST023427001.1	2: SMEST077205001.1		2: SMEST009320002.1
3: SMEST025073002.1		3: SMEST031902003.1			3: SMEST030237001.1
4: SMEST031780001.1		4: SMEST061120001.1			4: SMEST033017001.1
5: SMEST080372001.1		5: SMEST063468001.1			
		6: SMEST068480001.1			

Figure 3.18: Workflow for the identification of candidate transcription factors potentially involved in the development and maintenance of the planarian reproductive system. A *S. mediterranea* transcription factor database was built using hidden Markov models of classified DNA binding domains obtained in TFClass (Wingender et al., 2018). Genes belonging to transcription factor families identified in the motif variability analysis were considered. Using differential gene expression data, some members of the transcription factor families were selected for further investigation.

Overall, Start-seq is capable of identifying differentially active regulatory elements between two conditions. It shows a significantly positive correlation with RNA-seq data from identical conditions although differences exist. Many differentially upregulated promoters identified by Start-seq can be readily attributed to genes with a described function in sexual reproduction and GO analysis supports those claims. Finally, Start-seq is able to identify relevant motif families within differentially regulated regulatory with published functions in planarian sexual reproduction.

3.4 Characterization of putative sexual reproduction gene-regulatory network components

To characterize putative regulators of reproductive organ development and maintenance, I decided to proceed with the following workflow (Figure 3.19). First, I investigated their expression pattern in sexual worms by whole-mount *in situ* hybridization (WISH). Following this, I knocked down each transcription factor by RNA interference and perform RNA-seq to visualize global transcriptomic changes compared to a negative control. I also looked at the relative changes in the abundance of published markers for reproductive tissues like testes, ovaries, yolk and shell glands. Finally, I assessed the expression pattern and abundance of select markers of sexual tissues under RNAi conditions of TF candidates showing positive results in the DGEA analysis.

Two candidates (*cebp* 5 and *snail* 6) were unable to be characterized due to cloning difficulties. I also did not perform the knock-down experiment on *cebp* 1, *snail* 1, *gata* 2 and *nfk* 1-4 since I only managed to clone them after the RNAi experiment had been performed. In total, I characterized 18 out of 20 candidates with respect to their expression pattern and 11 out of 20 candidates were functionally tested by RNAi.

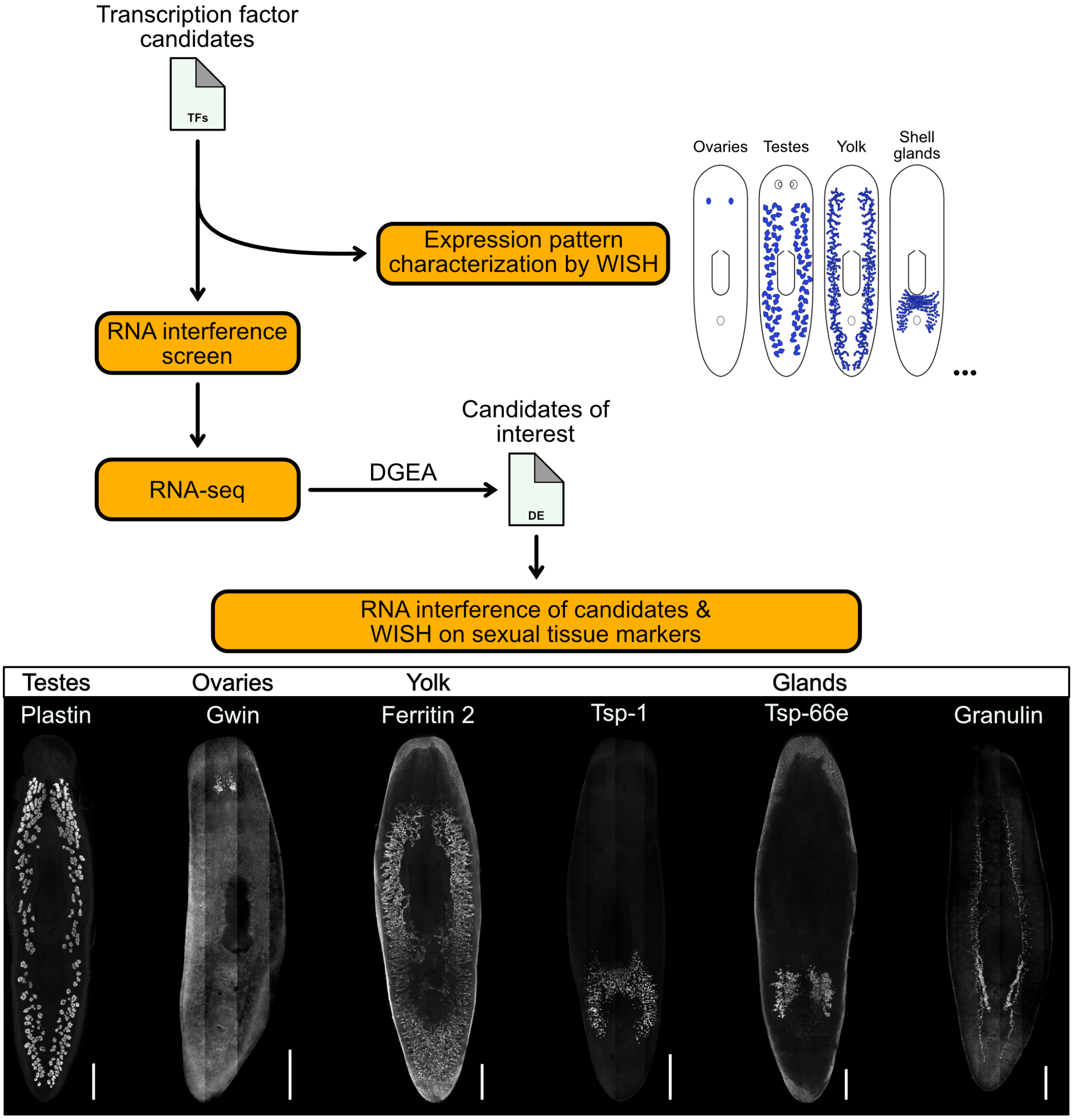


Figure 3.19: Workflow for the characterization of candidate TFs potentially involved in the development and maintenance of the planarian reproductive system. The expression pattern of candidate TFs is assessed by whole-mount *in situ* hybridization (WISH). Functional relevance of each TF for the planarian gonad is assessed by RNA interference followed by RNA-seq and interesting candidates are selected for further analysis. The expression pattern of several reproduction-associated tissues is assessed by WISH on previously selected candidates.

3.4.1 Visualizing the expression pattern of candidate transcription factors

I performed both colorimetric and fluorescent WISH on sexual *Schmidtea mediterranea* with probes targeting the TF candidates (Figures 3.20 until 3.26). Overall, almost all candidates tested showed expression in organs associated with sexual reproduction.

Candidates belonging to the CEBP TF family showed exclusive expression in the testes but each showed a different expression pattern (Figure 3.20). *cebp* 1 showed expression throughout the stages of sperm development, from the spermatogonia stage until the elongating spermatids (Chong et al., 2011). However, it was more expressed in the round spermatids compared to all other stages. *cebp* 2 was highly expressed in all stages of sperm development with no clear preference for a particular stage. *cebp* 3 had a similar expression pattern as *cebp* 1, with higher expression in the round spermatids, but the signal was overall less restricted to that compartment. Finally, *cebp* 4 showed very low signal throughout the stages of sperm development with a slight preference for the spermatogonia stage.

The two GATA candidates had a very different expression pattern (Figure 3.21). *gata* 1 signal could be found ubiquitously with increased expression in the eyes and gut. However, the signal was most visible in the testes. A high expression could be seen in spermatogonia with a relative decrease in signal intensity in later stages. Interestingly, *gata* 2 was found nearly exclusively expressed in oocytes with a higher expression in oocytes located distally compared to the tuba. Sparse signal could also be seen around the ovaries, potentially marking female germ cell progenitors (U. W. Khan & Newmark, 2022).

The *thap* candidate showed a wider expression pattern (Figure 3.21). However, some areas showed increased expression such as the nervous system, around the ovaries and around the copulatory apparatus. Interestingly, expression pattern above the ovaries was found similar as to the *gata* 2 candidate. Therefore, this candidate could also be another TF potentially important the female germ line development. Signal in the ovaries did not resemble an oocyte expression pattern and looked to be restricted on the distal part of the somatic ovary (U. W. Khan & Newmark, 2022). Additionally, sparse labeling could be found in an arc around the top of the pharynx, possibly labeling uncharacterized glands.

The *tead* 1 candidate was ubiquitously expressed similar to *tead* 2 but at a higher level (Figure 3.22). It showed in both cases enrichment in the central nervous system and around the copulatory apparatus. *tead* 1 also clearly marked the copulatory bursa while signal was found around the penis papilla in both cases. Additionally, the staining around the pharynx was also clearly visible for the *tead* 1 candidate.

All members of the SNAIL family, except for *snail* 2, showed exclusive expression in sexual organs (Figure 3.23 & 3.24). *snail* 1 showed low expression throughout the stages of sperm development with higher staining in the outer layer of the testes lobules where the spermatogonia and spermatocytes reside (Figure 3.23). Signal in the copulatory apparatus around the penis papilla was also visible as well as ventrally around the pharynx in the form of an arc similar to the *tead* 1 and *thap* candidates. The *snail* 2 candidate showed high enrichment in testes, especially in the spermatids stages. Ventrally, signal was also visible around the pharynx and the copulatory apparatus. Additionally, the oviducts were strongly stained as well as some cells around the ovaries. Dorsally, the *snail* 3 candidate showed expression in the outer layer of the testis lobule (Figure 3.24). Additionally, *snail* 3 also marked oocytes. Sparse signal could also be seen again around the pharynx. *snail* 4 signal was only found dorsally and specifically marked the outer layer of the testes lobules. Finally, the *snail* 5 candidate was found in the testes where it showed a heterogeneous signal. Only a sub-population of male germ cells from the spermatogonia to the round spermatid stage expressed this transcription factor.

The *nfk* candidate transcription factors showed overall a more widespread expression pattern with enrichment in some sexual organs (Figure 3.25 & 3.26). *nfk* 1 showed a strong expression in non-sexual organs such as the gut (Figure 3.25). Additionally, the ovarian somatic cells showed enriched staining. Very sparse labeling in the testes, reminiscent of the *snail* 5 labeling but with even less cells labeled, could be seen as well. The *nfk* 2 candidate was ubiquitously expressed with some areas such as cells resembling shell glands and elements of the copulatory apparatus showing a stronger signal. Enrichment of the signal in the testes was also found, especially in the outer layer of the lobule. The *nfk* 3 candidate also showed somatic expression with enrichment in the brain (Figure 3.26). Additionally, expression in sexual tissues such as the testes and copulatory apparatus could be seen. In the testes, signal could be found throughout the different stages but the outer layer of the lobule had again a higher signal. Finally, no real staining, neither somatic nor germline expression could be seen for *nfk* 4.

A summary of the expression patterns can be found in Figure 3.27.

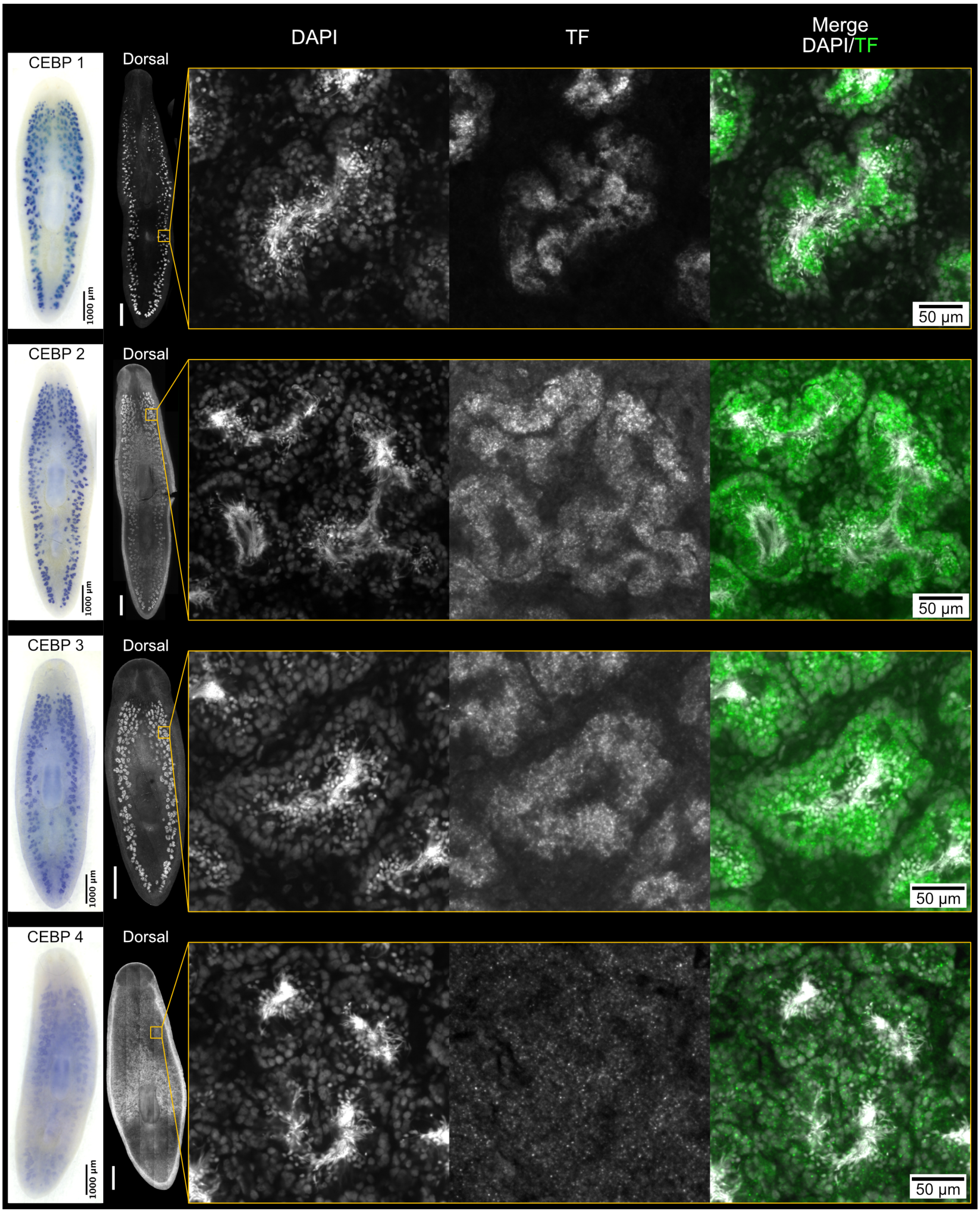


Figure 3.20: Expression pattern of the CEBP transcription factor family candidates. From left to right, colorimetric WISH of the TF candidate. Fluorescent WISH overview of the expression pattern of the TF candidate obtained by maximum intensity projections of confocal sections. Panel detailing relevant expression patterns observed in the fluorescent WISH by maximum intensity projections of confocal sections. Unlabeled scale bars represent 1 mm.

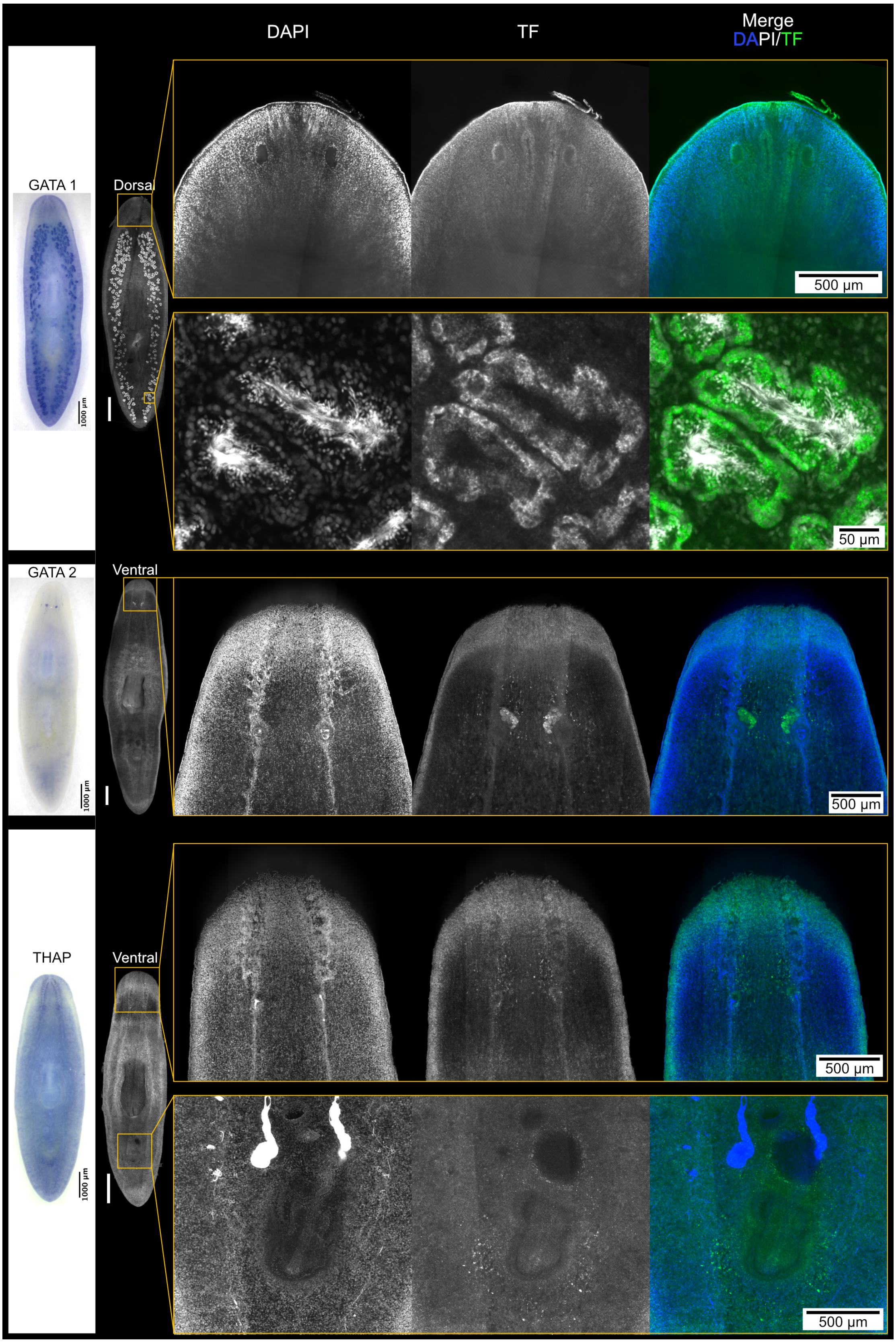


Figure 3.21: Expression pattern of the GATA and THAP transcription factor family candidates. From left to right, colorimetric WISH of the TF candidate. Fluorescent WISH overview of the expression pattern of the TF candidate obtained by maximum intensity projections of confocal sections. Panel detailing relevant expression patterns observed in the fluorescent WISH by maximum intensity projections of confocal sections. Unlabeled scale bars represent 1 mm.

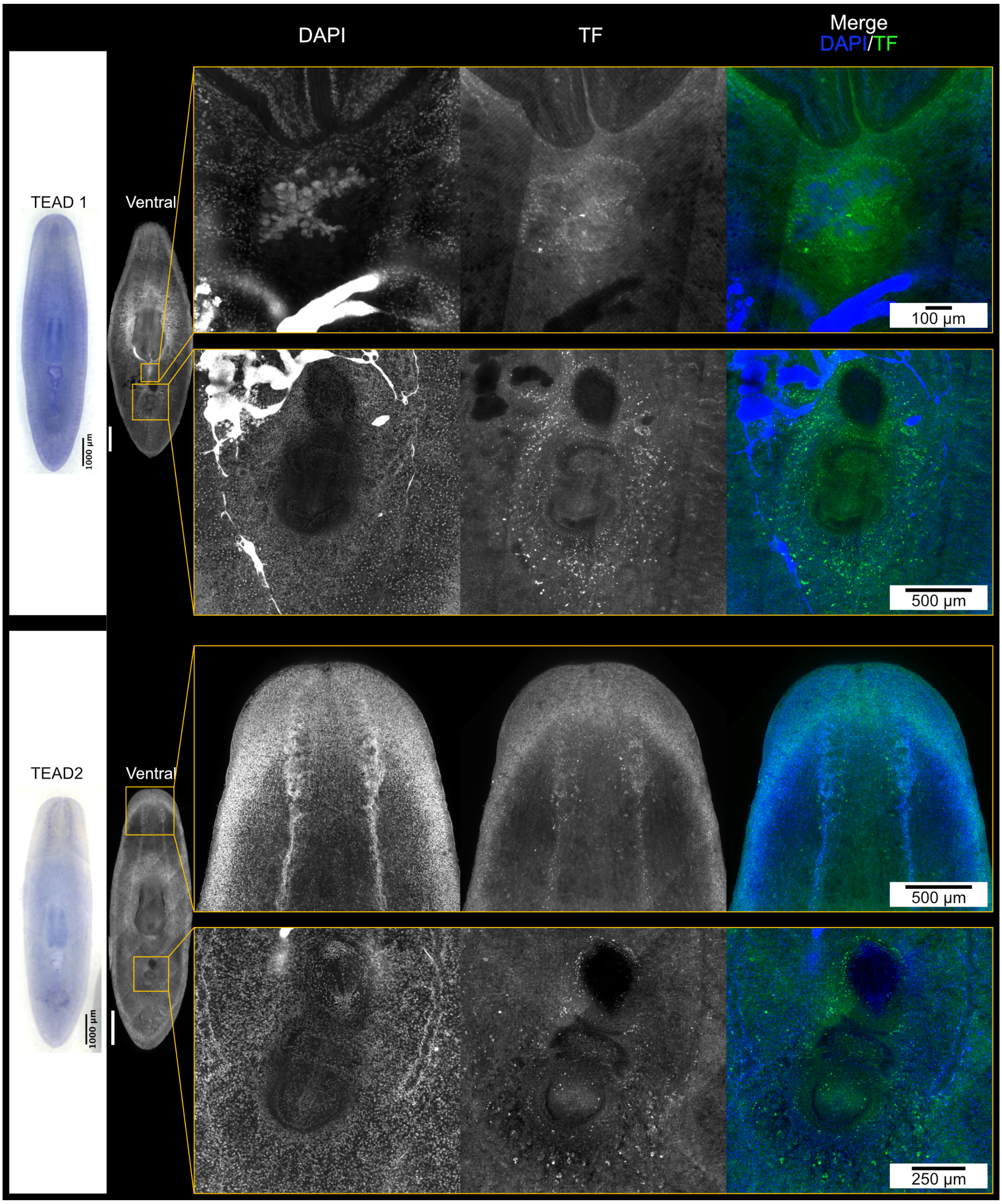


Figure 3.22: Expression pattern of the TEAD transcription factor family candidates. From left to right, colorimetric WISH of the TF candidate. Fluorescent WISH overview of the expression pattern of the TF candidate obtained by maximum intensity projections of confocal sections. Panel detailing relevant expression patterns observed in the fluorescent WISH by maximum intensity projections of confocal sections. Unlabeled scale bars represent 1 mm.

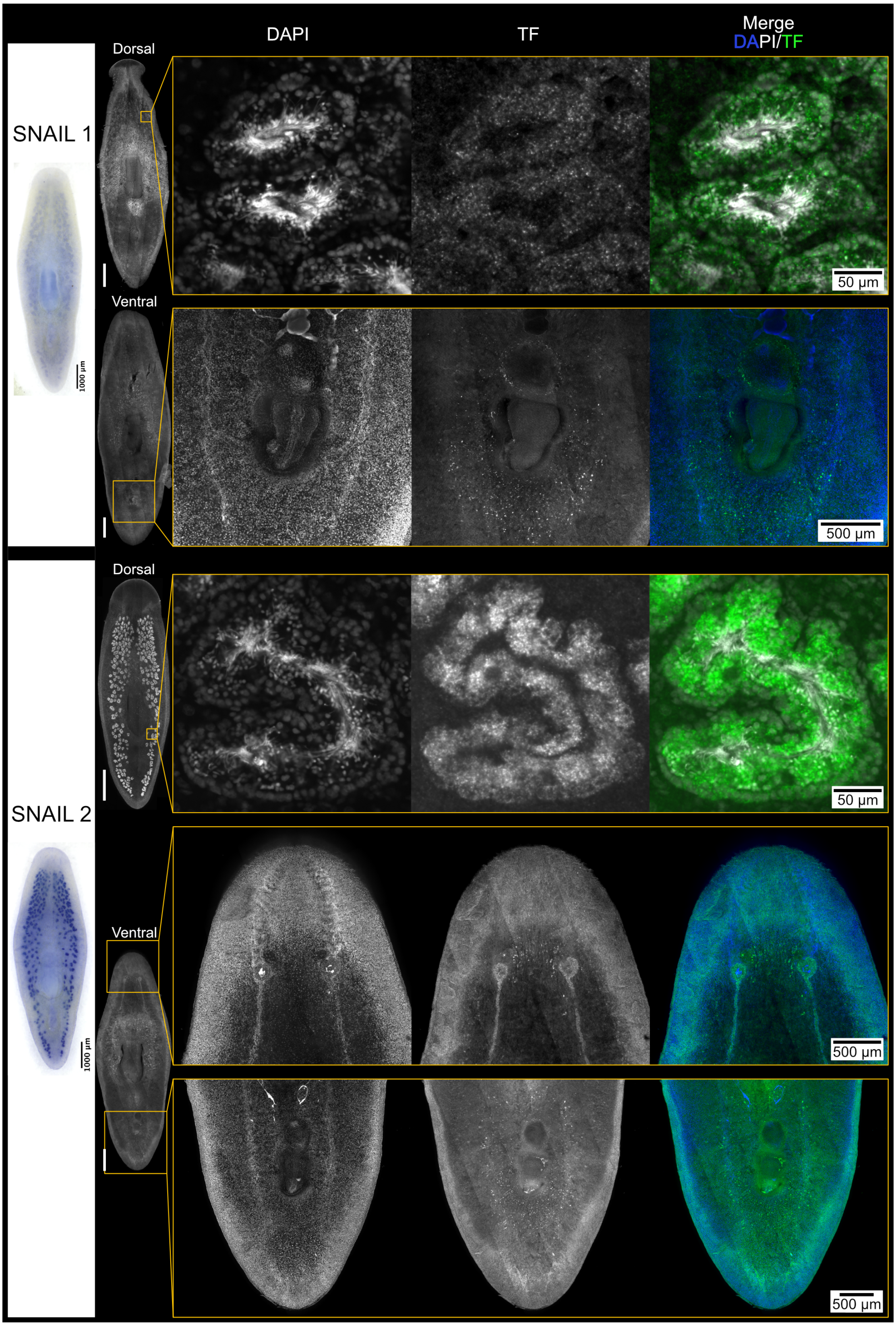


Figure 3.23: Expression pattern of the *snail* 1 and 2 transcription factor family candidates. From left to right, colorimetric WISH of the TF candidate. Fluorescent WISH overview of the expression pattern of the TF candidate obtained by maximum intensity projections of confocal sections. Panel detailing relevant expression patterns observed in the fluorescent WISH by maximum intensity projections of confocal sections. Unlabeled scale bars represent 1 mm.

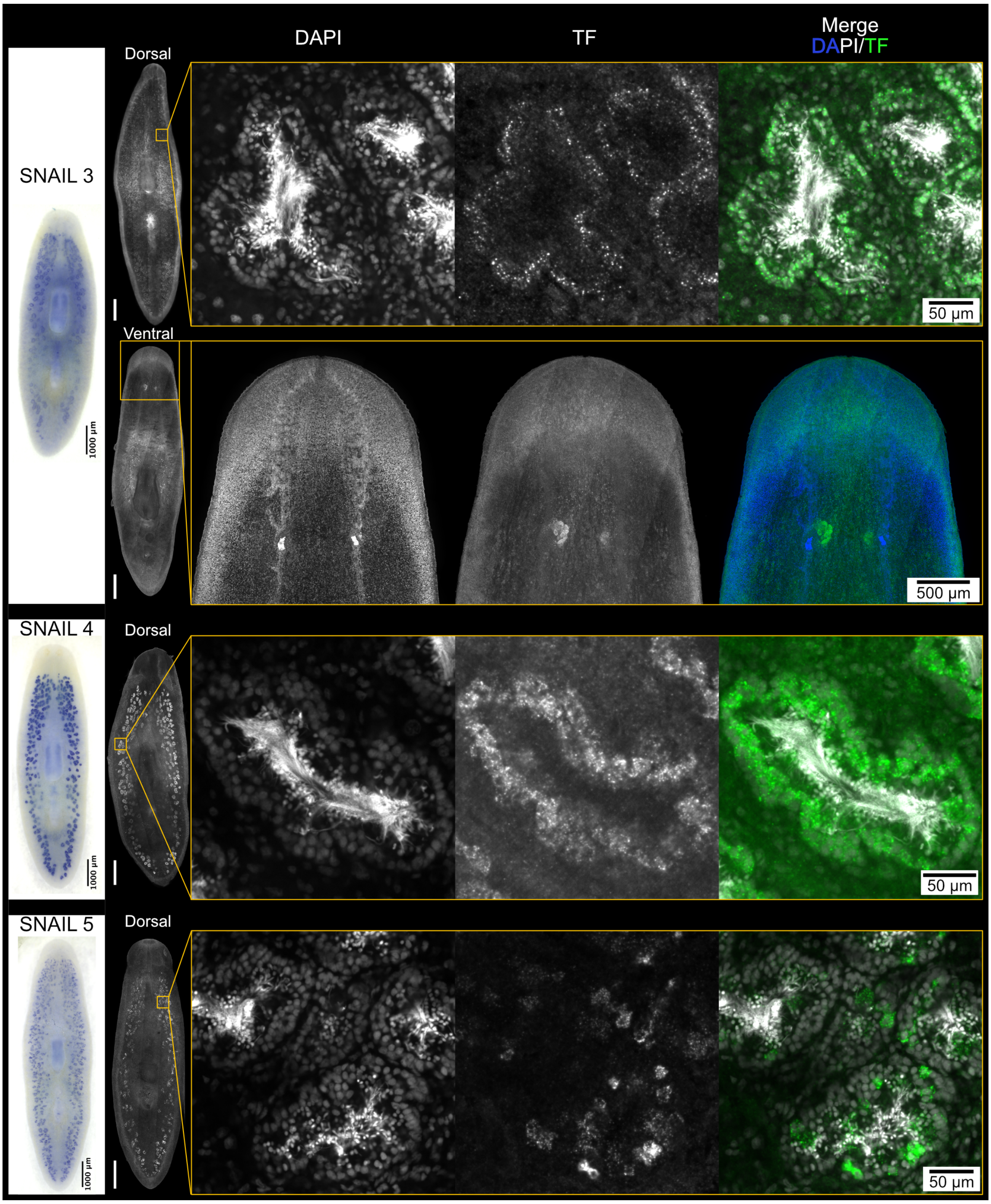


Figure 3.24: Expression pattern of the *snail* 3 to 5 transcription factor family candidates. From left to right, colorimetric WISH of the TF candidate. Fluorescent WISH overview of the expression pattern of the TF candidate obtained by maximum intensity projections of confocal sections. Panel detailing relevant expression patterns observed in the fluorescent WISH by maximum intensity projections of confocal sections. Unlabeled scale bars represent 1 mm.

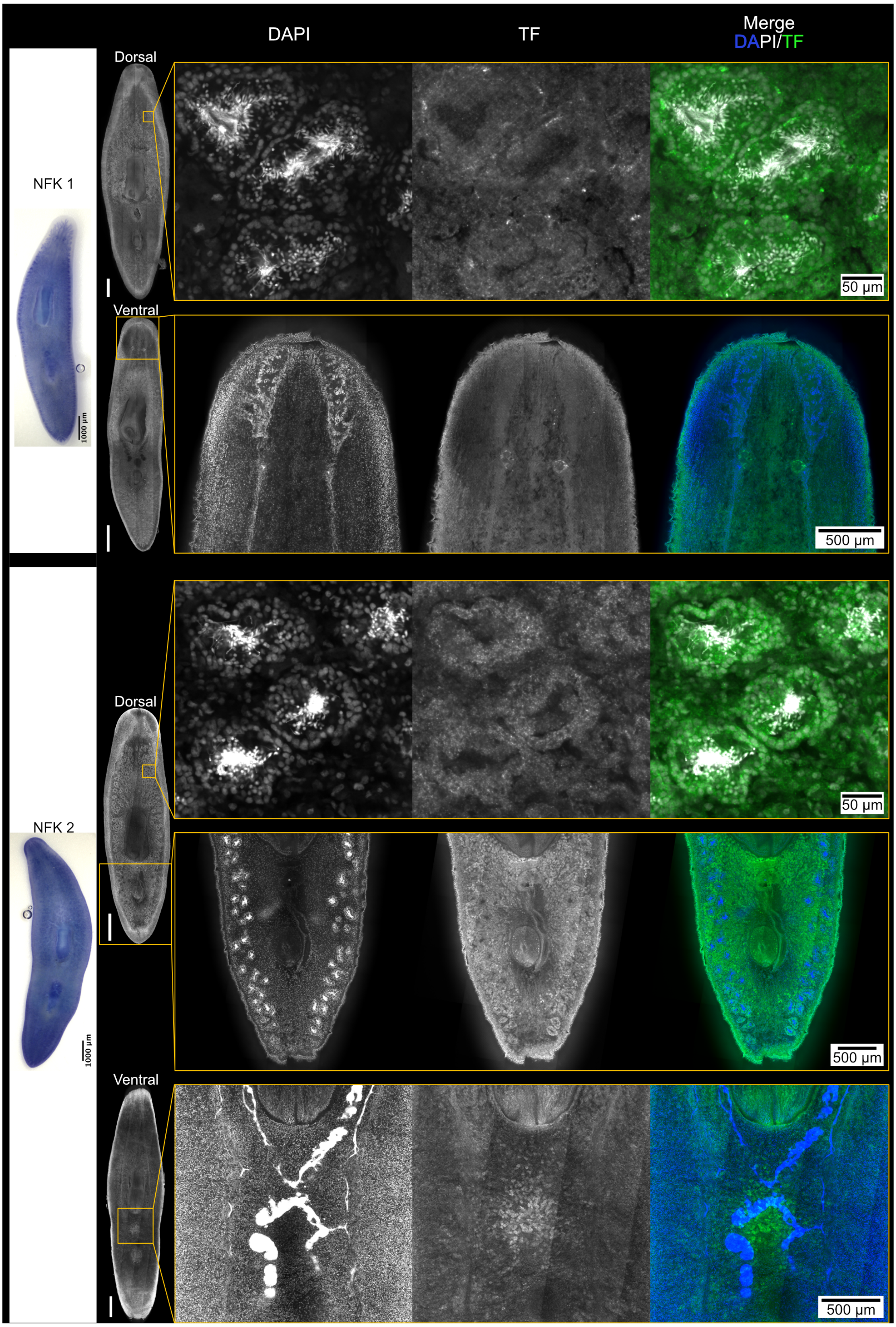


Figure 3.25: Expression pattern of the NF-κB-related 1 and 2 transcription factor family candidates. From left to right, colorimetric WISH of the TF candidate. Fluorescent WISH overview of the expression pattern of the TF candidate obtained by maximum intensity projections of confocal sections. Panel detailing relevant expression patterns observed in the fluorescent WISH by maximum intensity projections of confocal sections. Unlabeled scale bars represent 1 mm.

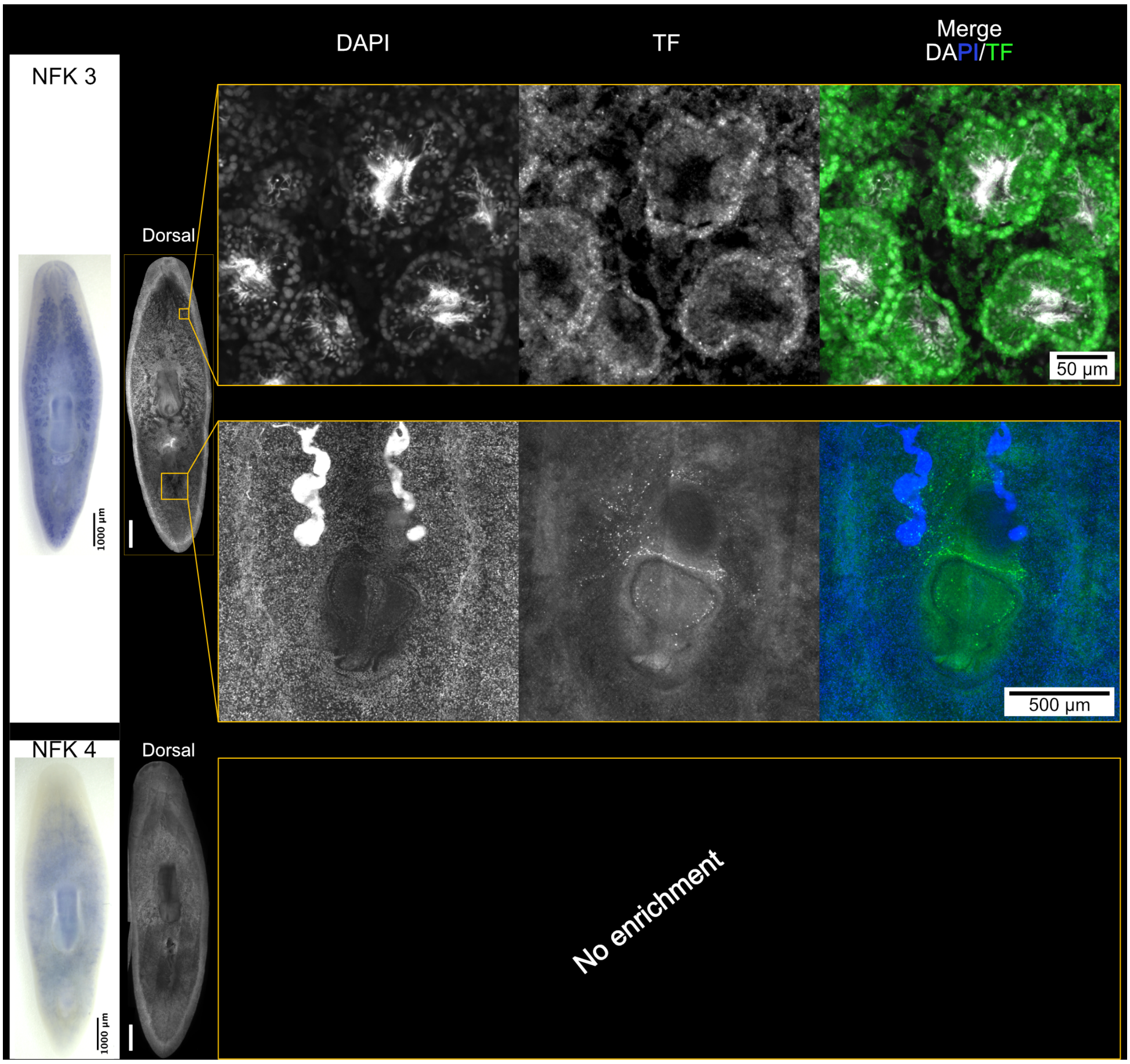
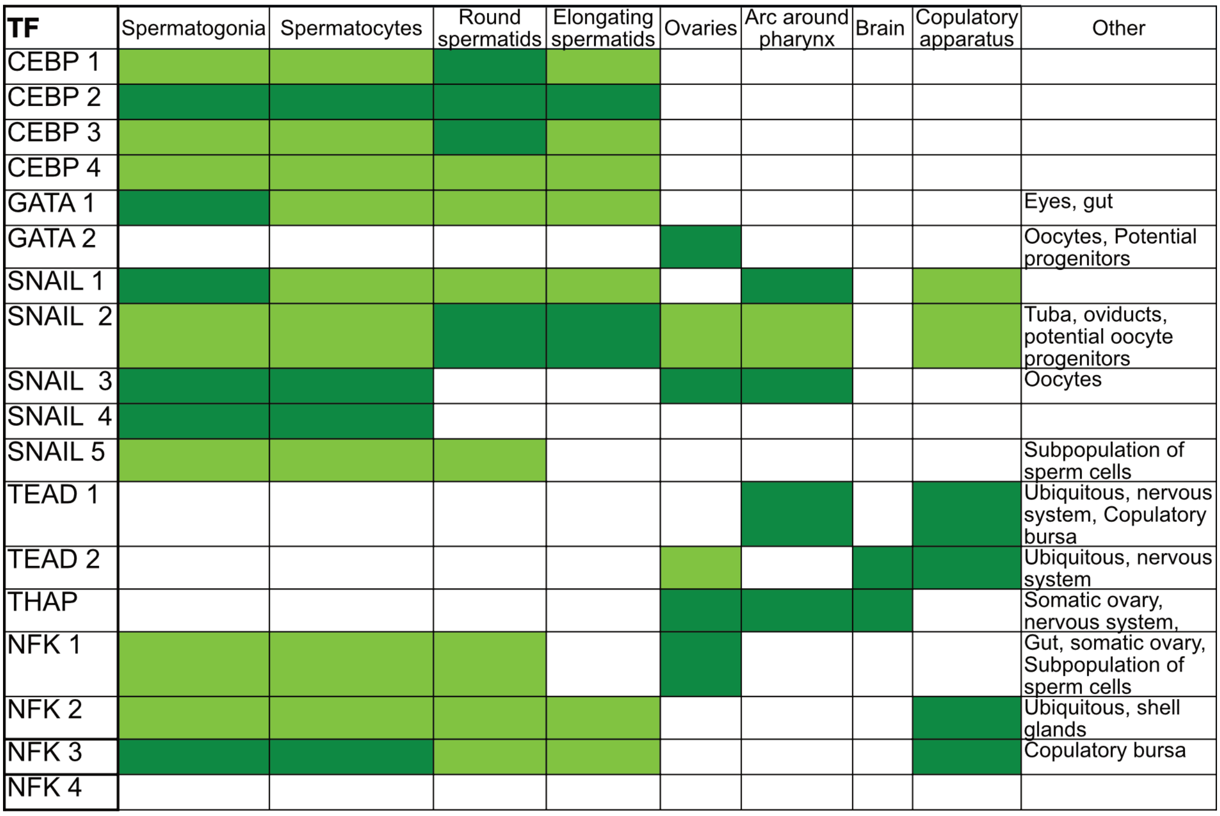


Figure 3.26: Expression pattern of the NF-κB-related 3 and 4transcription factor family candidates. From left to right, colorimetric WISH of the TF candidate. Fluorescent WISH overview of the expression pattern of the TF candidate obtained by maximum intensity projections of confocal sections. Panel detailing relevant expression patterns observed in the fluorescent WISH by maximum intensity projections of confocal sections. Unlabeled scale bars represent 1 mm.



TF	Spermatogonia	Spermatocytes	Round spermatids	Elongating spermatids	Ovaries	Arc around pharynx	Brain	Copulatory apparatus	Other
CEBP 1	🟩	🟩	■	🟩					
CEBP 2	■	■	■	■					
CEBP 3	🟩	🟩	■	🟩					
CEBP 4	🟩	🟩	🟩	🟩					
GATA 1	■	🟩	🟩	🟩					Eyes, gut
GATA 2					■				Oocytes, Potential progenitors
SNAIL 1	■	🟩	🟩	🟩		■		🟩	
SNAIL 2	🟩	🟩	■	■	🟩	🟩		🟩	Tuba, oviducts, potential oocyte progenitors
SNAIL 3	■	■			■	■			Oocytes
SNAIL 4	■	■							
SNAIL 5	🟩	🟩	🟩						Subpopulation of sperm cells
TEAD 1						■		■	Ubiquitous, nervous system, Copulatory bursa
TEAD 2					🟩		■	■	Ubiquitous, nervous system
THAP					■	■	■		Somatic ovary, nervous system,
NFK 1	🟩	🟩	🟩		■				Gut, somatic ovary, Subpopulation of sperm cells
NFK 2	🟩	🟩	🟩	🟩				■	Ubiquitous, shell glands
NFK 3	■	■	🟩	🟩				■	Copulatory bursa
NFK 4									

Figure 3.27: Summary of the expression pattern for all the tested candidate transcription factors. The color represents relative expression strength within one condition with light green representing low expression and dark green high expression.

3.5 Validation of transcription factor function by RNA interference

I next investigated the effect of TF knockdown on the sexual organs by RNA interference followed by RNA-seq (Figure 3.28). I first let worms regenerate from a piece without reproductive organs proceeded with 8 dsRNA feedings of each of the 11 candidate TFs individually (Figure 3.28 A). After a week of starving, I extracted RNA and sent it to sequencing. In addition to the 11 TFs conditions, I also added a negative control (*egfp*) and a positive control (*ophis*) that fails to develop any germ line (Saberi et al., 2016). All conditions were sent in duplicates.

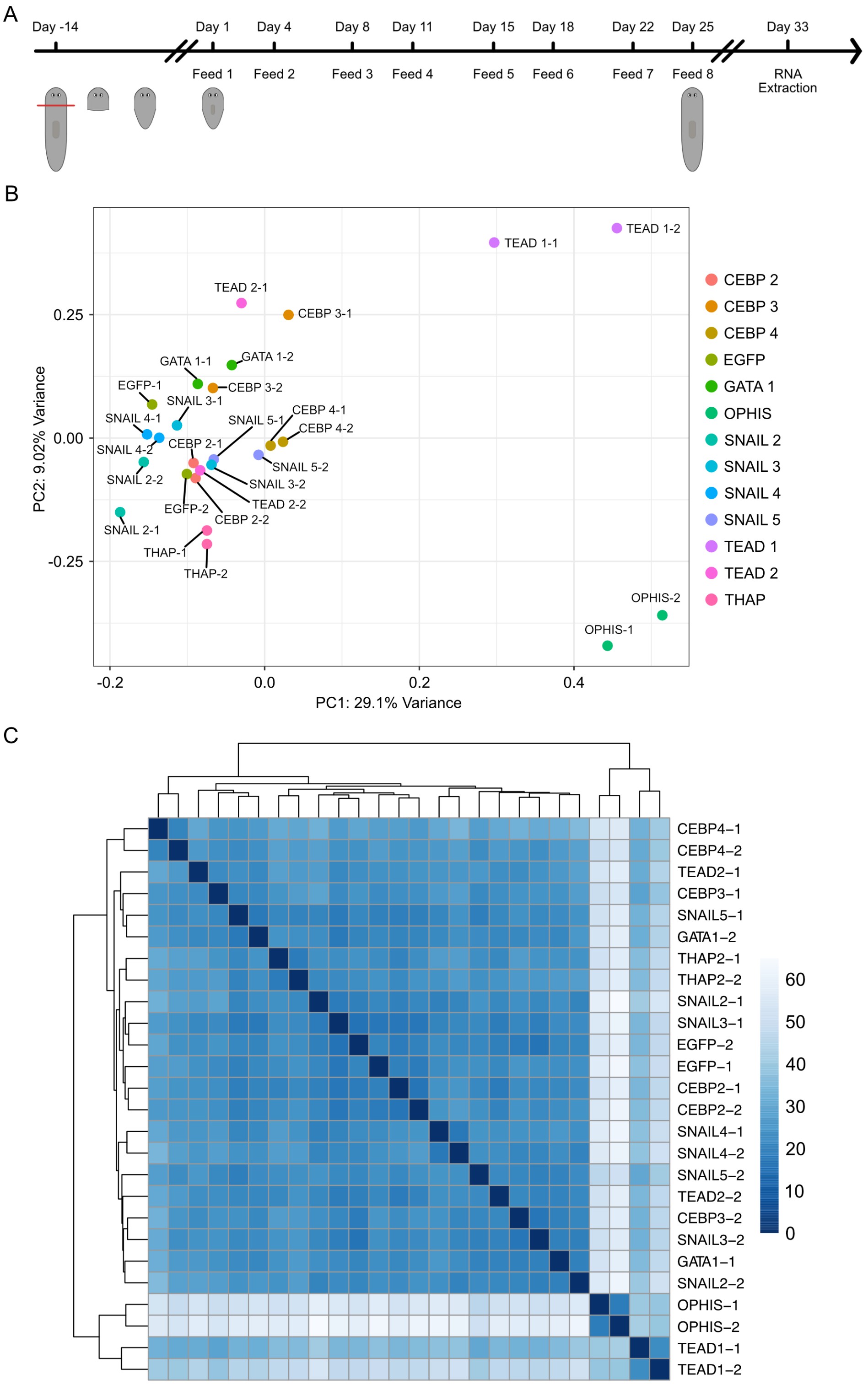


Figure 3.28: RNA interference screen of TF candidates potentially involved in the development and maintenance of the planarian reproductive tissues. A) Workflow used for the RNAi screen. Sexually mature worms are cut and heads devoid of reproductive organs are left to regenerate for 2 weeks. Regenerated worms are fed 8 times RNAi liver over a period of 4 weeks to grow to a sexually mature size. Worms are then left to starve for 1 week and RNA is extracted for RNA-seq. B) Principal component analysis of the RNA interference screen samples. C) Clustered heatmap of RNA-seq samples from the RNAi screen. The scale represents the pairwise Euclidean distances between the samples.

3.5.1 Differential gene expression analysis of candidate transcription factor knockdown

A principal component analysis revealed that many of the transcription factors were clustered around the negative control (Figure 3.28 B). Interestingly, the *tead* 1 samples were situated at approximately the same coordinate on the x-axis as the *ophis* samples but were on opposite sides of the Y-axis. The variances explained by both PC1 and PC2 were quite low suggesting that the samples could differ in ways not visualized by the first two principal components. A clustered heatmap of euclidean distances between samples showed that the *ophis* and *tead* 1 samples clearly clustered separately from the rest of the conditions (Figure 3.28 C). Interestingly, the *cebp* 4 samples were the second closest to the *ophis* and *tead* conditions.

After DGEA versus the negative control (*egfp*), I decided to investigate which samples shared the most differentially regulated genes (adjusted p-value < 0.05) with the positive control (*ophis*) (Figure 3.29). For this, I generated an upset plot comparing all conditions. Unsurprisingly, the condition that shared by far the most differentially expressed genes (DE) with *ophis* exclusively was the *tead* 1 candidate (3274 shared genes). After that, the *cebp* 4 candidate shared about 1000 DE genes with both *ophis* and *tead* 1. Finally, the *thap* candidate also had about 300 DE genes shared with *tead* 1 and *ophis.* All the other conditions had a very low overlap with the positive control. I next generated an upset plot exclusively with the 3 candidates mentioned previously to facilitate the visualization of the different overlaps (Figure 3.30). Interestingly, there was not much overlap between the *cebp* 4 and *thap* conditions hinting that they could potentially regulate different aspects of sexual reproduction

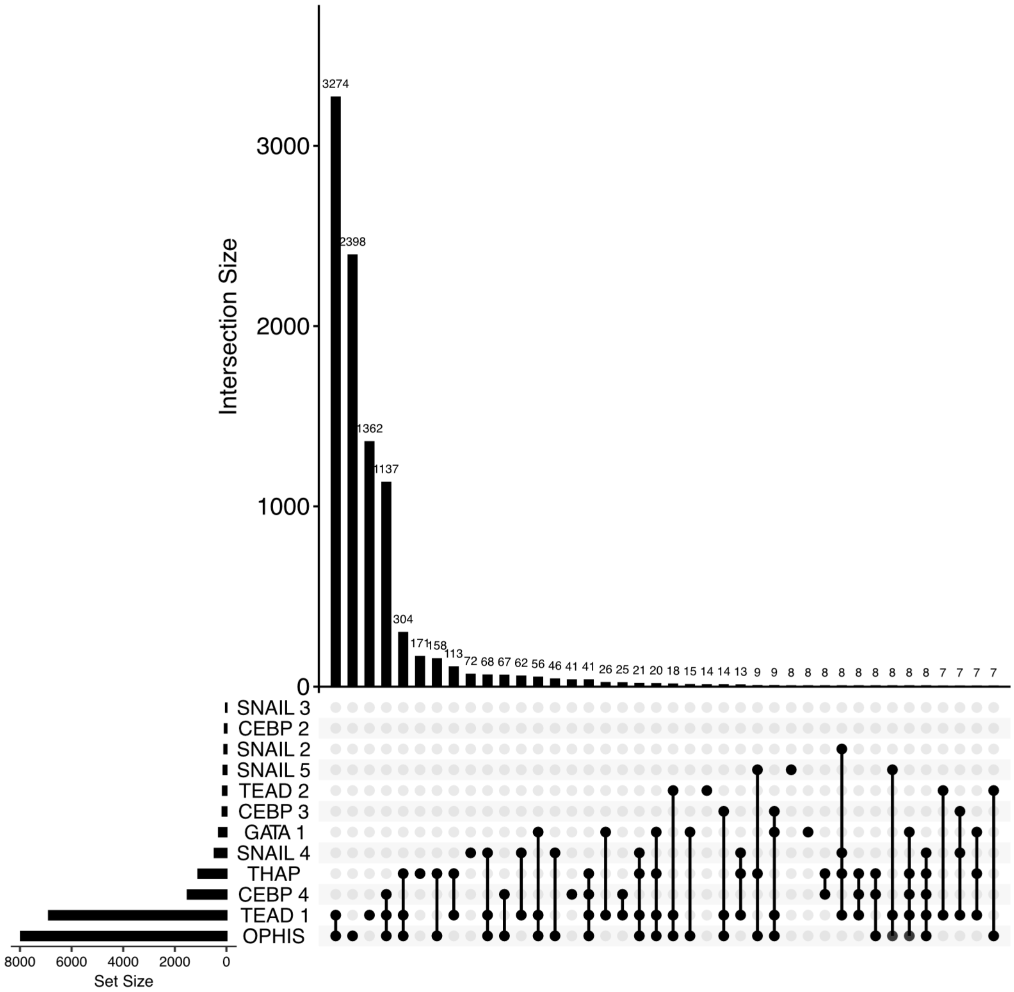


Figure 3.29: Overlap of differentially regulated genes (adjusted p-value < 0.05), compared to the negative control, between the different conditions in the RNAi screen.

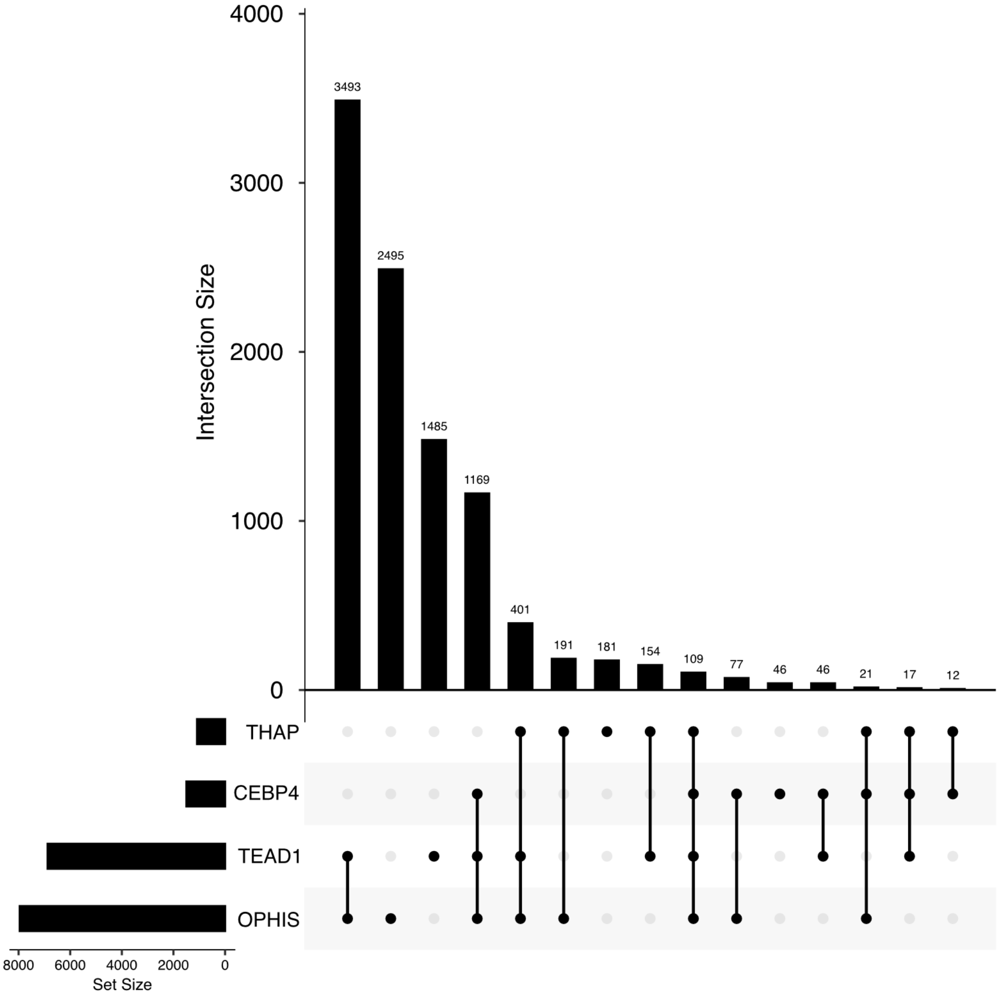


Figure 3.30: Overlap of differentially regulated genes (adjusted p-value < 0.05), compared to the negative control, between the positive control *ophis* and interesting candidates (*tead* 1, *cebp* 4 and *thap*).

3.5.2 Sexual tissue marker analysis

I next tried to assess whether any specific reproductive organ was affected by knock-down of a candidate TF by looking at the expression level of several published markers for each type of reproductive organs (Chong et al., 2011; Issigonis et al., 2022; U. W. Khan & Newmark, 2022; Rouhana et al., 2017; Steiner et al., 2016; Vila-Farré et al., 2023; Y. Wang et al., 2010, 2007; Zayas et al., 2005) (Figure 3.31 until 3.33). A list of the markers and their expression pattern can be found in table 3.4.

Table 3.4: Published markers used in this study for the various tissues involved in sexual reproduction in *S. mediterranea*

Name	General expression	Specific expression	Source
msy4	Testes	all except fully mature spermatozoa	Chong et al. (2011)
tplh	Testes	spermatocytes and spermatids	Chong et al. (2011)
cpeb-2	Testes	planarian brain and testes (spermatogonia and spermatocytes)	Rouhana et al. (2017)
cathepsin-L	Testes	all except fully mature spermatozoa	Zayas et al. (2005)
pde	Testes	all except fully mature spermatozoa	Chong et al. (2011)
plastin	Testes	all except fully mature spermatozoa	Chong et al. (2011)
pp2	Testes	spermatids	Chong et al. (2011)
pka	Testes	spermatids	Chong et al. (2011)
thmg-1	Testes	all except fully mature spermatozoa	Chong et al. (2011)
tkn-1	Testes	spermatocytes	Chong et al. (2011)
tkn-2	Testes	all except fully mature spermatozoa	Chong et al. (2011)
cct-1	Testes	N.A.	Rouhana et al. (2017)
ferritin-1	Yolk	yolk glands	Vila-Farré et al. (2023)
ferritin-2	Yolk	yolk glands	Vila-Farré et al. (2023)
cpeb-1	Ovaries/yolk	ovaries (in oocytes) and yolk glands	Rouhana et al. (2017)
surfactant B	Yolk	yolk glands	Steiner et al. (2016)
tanning factor-1	yolk	yolk glands	Rouhana et al. (2017)
tyrosinase	yolk	yolk glands and posterior to copulatory apparatus	Rouhana et al. (2017)
c-type lectin	yolk	yolk glands	Rouhana et al. (2017)
klf4l	Precursors	early germ line and yolk precursors	Issigonis et al. (2022)
nanos	Precursors	germ line and yolk precursor	Z. Wang et al. (2008)
tsp-1	Shell glands	Shell glands	Chong et al. (2011)
tetraspanin 66e	Shell glands	Shell glands	Rouhana et al. (2017)
Hypothetical (S. mansoni)	accessory reproductive organs	oviducts	Rouhana et al. (2017)
granulin	accessory reproductive organs	sperm duct and seminal vesicles	Chong et al. (2011)
zfs1	Ovaries	early germ cells	U. W. Khan and Newmark (2022)
lecg	Ovaries	oocytes	U. W. Khan and Newmark (2022)
ubp8	Ovaries	oocytes	U. W. Khan and Newmark (2022)

Most of the testes markers showed similar changes in expression depending on the condition (Figure 3.31). Interestingly, some of the testes markers such as *pka* and *tkn-1* were not downregulated in the *ophis* condition. A possible explanation could be that these genes are enriched in testes but are also expressed in other somatic tissues and was not reported in the literature (Chong et al., 2011). Additionally, the *cathepsin-L* marker was significantly upregulated in many conditions and will require additional investigation. In total, 9/12 testes markers were significantly downregulated in the *ophis* conditions The *cebp* 4 candidate showed downregulation of 5/12 markers, suggesting an importance on sperm development. Finally, the *tead* 1 candidate also showed a strong downregulation of testes markers (9/12), similar to *ophis*. All the other candidates did not show significant changes in testes marker expression. Interestingly, RNAi of the *snail* 2 candidate showed a consistent upregulation of testes markers (9/12) but did not reach the threshold set for significance.

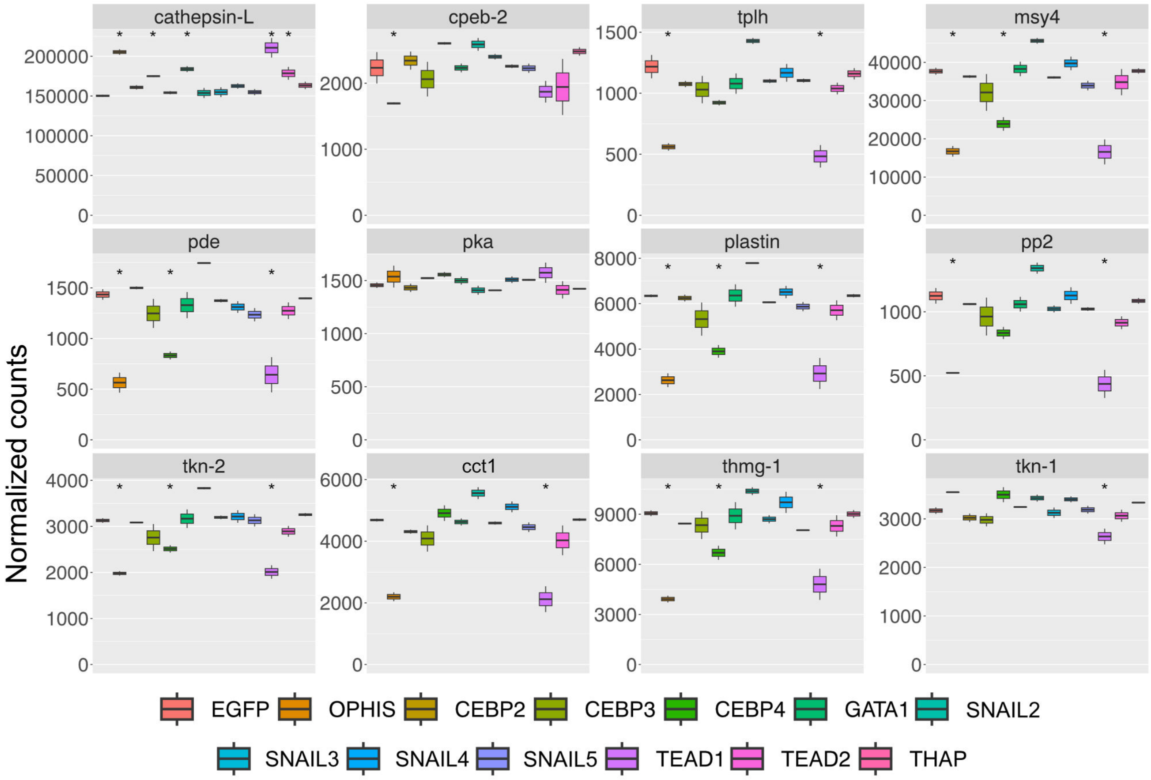


Figure 3.31: Comparison of normalized counts for testes markers between the different RNAi conditions. Asterisks represent a significance of $p< 0.05$ compared to the *egfp* control using the Wald test.

I next investigated the yolk markers (Figure 3.32). All the markers showed significant downregulation in *ophis* RNAi. Additionally, only *tead* 1 RNAi also showed significant downregulation of several yolk markers (2/7) while the other markers were also less expressed but did reach the significance threshold. These results suggest that only the *tead* 1 candidate affects the yolk tissue.

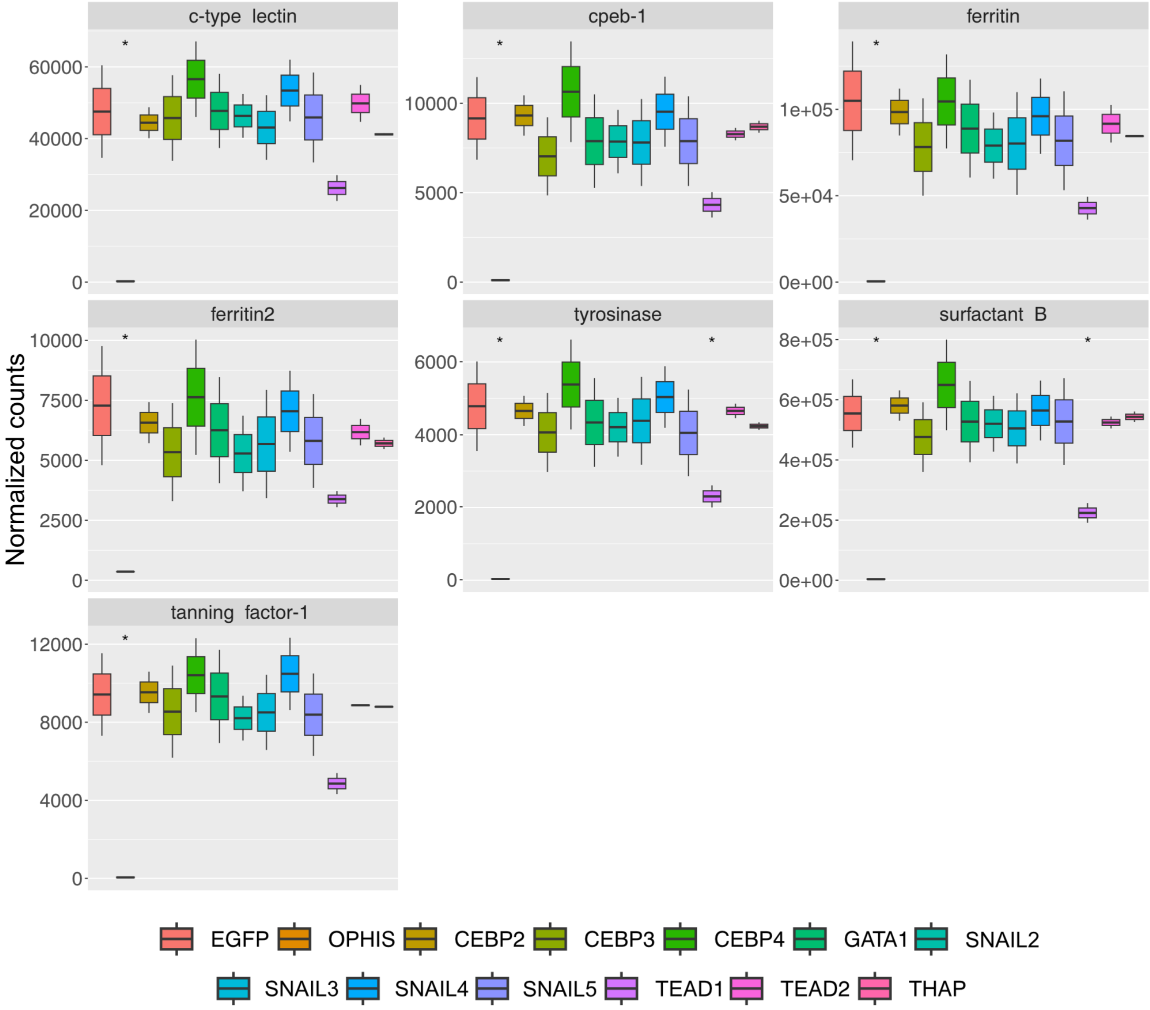


Figure 3.32: Comparison of normalized counts for yolk markers between the different RNAi conditions. Asterisks represent a significance of $p<0.05$ compared to the *egfp* control using the Wald test.

Next, I looked at markers related to germ cell precursors, ovaries as well as markers of accessory reproductive organs such as shell glands, oviducts and sperm ducts (Figure 3.33). Interestingly, the *nanos* germ cell precursor marker showed a significant increase in the *ophis* and *tead* 1 conditions. On the other hand, the other germ cell precursor marker, *klf4l*, did not show any significant change. Two of the ovary markers were lowly expressed. *zfs1* did not show any significant changes while *ubp8* was only down regulated in *ophis* RNAi. *ubp8* counts were also down in *thap* RNAi although not significantly. The last marker *lecg* was generally more highly expressed compared to *zfs1* and showed significant down regulation in *ophis* and *tead* 1 RNAi.

Regarding the accessory reproductive organs, the oviduct marker was significantly decreased in *ophis* and *tead* 1 RNAi. On the other hand, the sperm duct marker was upregulated in both *thap* and *tead* 1 condition while being down regulated in *ophis* RNAi. Interestingly, a similar pattern of gene marker upregulation was observed again in the *snail* 2 RNAi condition. Both the oocyte (*lecg*) and oviduct marker were more highly expressed in this condition. Finally, the shell gland markers *tsp-1* and *tsp 66e* showed significant downregulation in *tead* 1 and *ophis*. Additionally, *thap* RNAi also affected their expression level with *tsp 66e* being significantly downregulated. *tsp-1* was also down regulated in *thap* RNAi although not significantly.

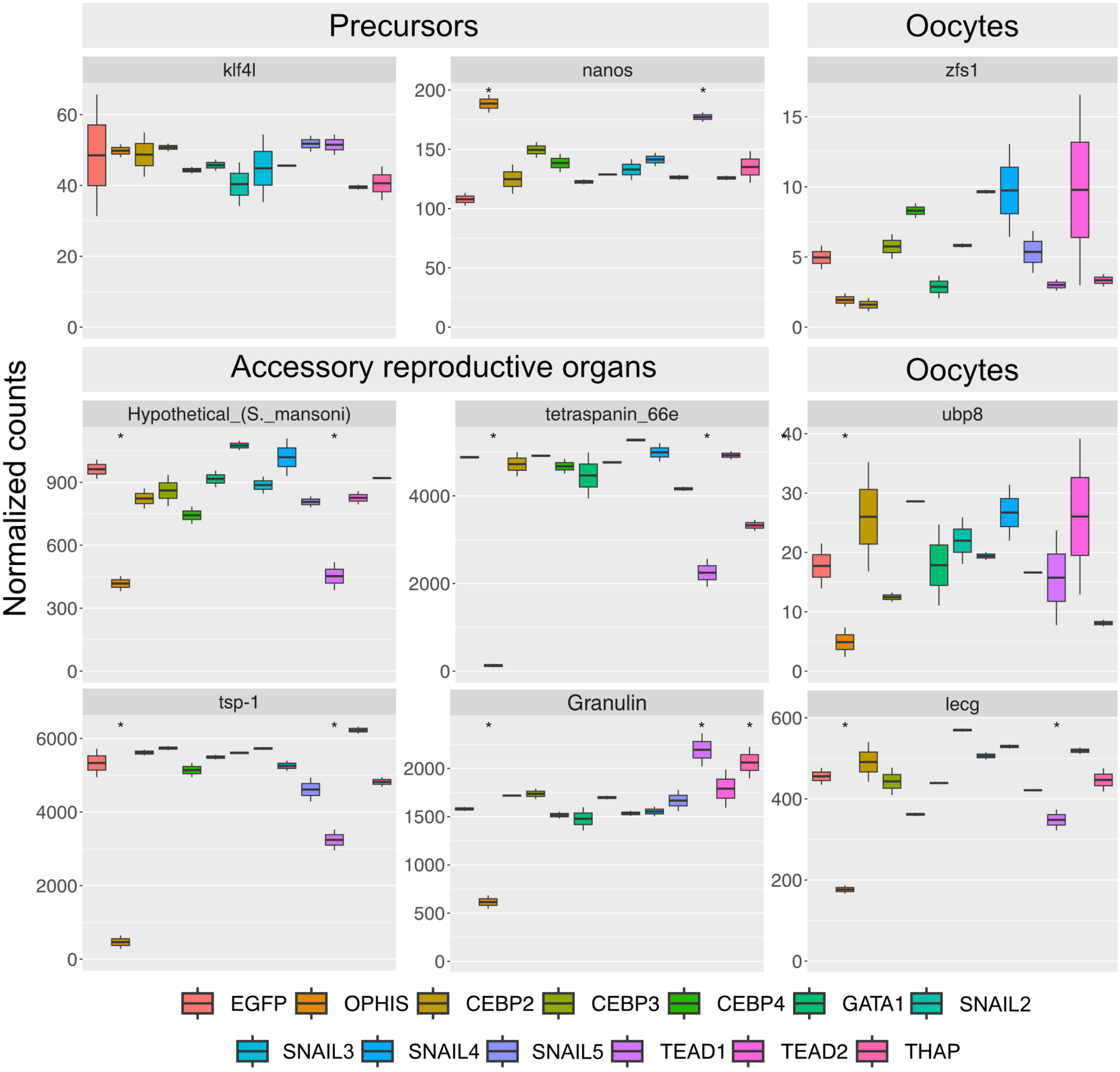


Figure 3.33: Comparison of normalized counts for oocyte, progenitors and accessory glands markers between the different RNAi conditions. Asterisks represent a significance of p< 0.05 compared to the *egfp* control using the Wald test.

Altogether, the three candidates (*tead 1*,*thap* and *cebp 4*) sharing many differentially expressed genes with the positive control showed downregulation of different aspects of *ophis* RNAi. The *tead* 1 phenotype resembled *ophis* RNAi the most, with downregulation of many markers of different reproductive tissues. *cebp* 4 RNAi had a clear testes phenotype, with only testes markers being downregulated while *thap* RNAi affected only shell gland markers.

3.5.3 Gene ontology enrichment analysis of candidates of interest

I next proceeded with the identification of potential functions associated with the genes being regulated by the three candidates. For this, I performed a GO enrichment analysis on significantly downregulated genes (adjusted p-value < 0.01) in each condition. I compared the results with those obtained for *ophis* RNAi.

Genes downregulated in *ophis* RNAi showed enrichment for two main classes of terms (Figure 3.34 A & B). The first class related to the assembly and function of cilia and can be visualized easily using a cnet plot (Figure 3.34 B). This is reminiscent of sperm development and function. The other category related to the transport of various organic molecules. Interestingly, many of these terms were also found during the GO enrichment analysis of the upregulated genes in the sexual Start-seq promoters (Figure 3.16), confirming its results. Downregulated genes in *thap* RNAi are involved in more basal functions related to anabolism and catabolism with significant terms such as translation or cellular amino acid catabolic process (Figure 3.35 A & B). For the *cebp* 4 RNAi, I found that post-translational protein modifications such as phosphorylation and microtubule related functions were most prevalent (Figure 3.36 A & B). Another group of terms were linked to a cluster of genes related to transport of intermediates of the TCA cycle. Finally, I observed that the *tead* 1 RNAi results of the GO enrichment analysis had a lot of overlap with terms found in *ophis* RNAi. I found many terms related to the assembly and function of cilia (Figure 3.37 A & B).

Overall, both the DGEA as well as the GO enrichment analysis show that *thap*, *tead* 1 and *cebp* 4 RNAi possess each similarities with *ophis* RNAi but do not regulate the same set of genes. *tead* 1 has a more global effect on reproductive tissues while the *cebp* 4 and *thap* candidates are more focused on specific tissues.

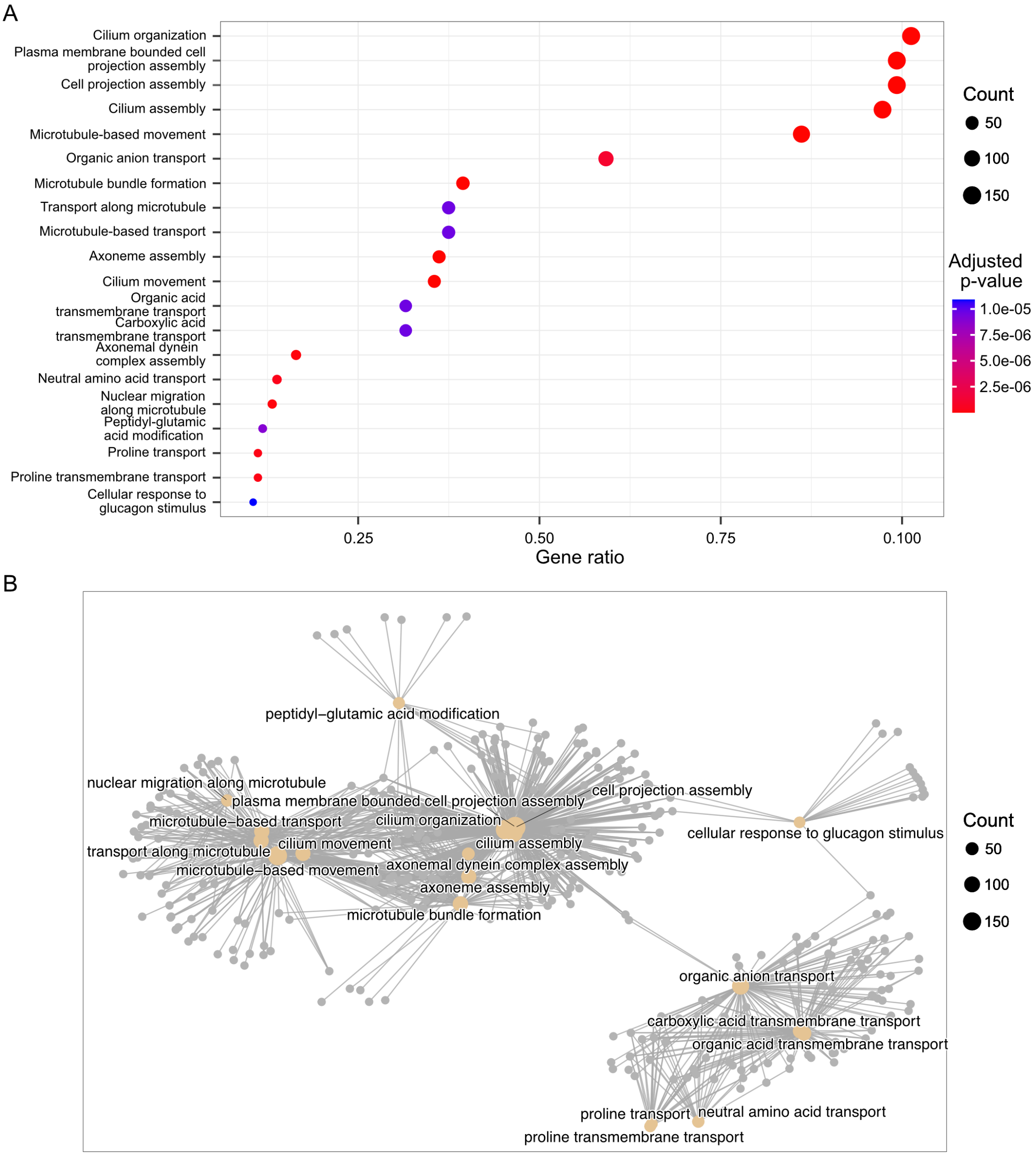


Figure 3.34: GO enrichment analysis for significantly downregulated transcripts in the positive control (*ophis*) knock-down. A) Dot plot representation of the top 20 significant terms associated to significantly downregulated transcripts in the positive control (*ophis*) knock-down. B) Cnet plot representation of the top 20 significant terms (brown dots) associated to significantly downregulated transcripts (gray dots) in the positive control (*ophis*) knock-down.

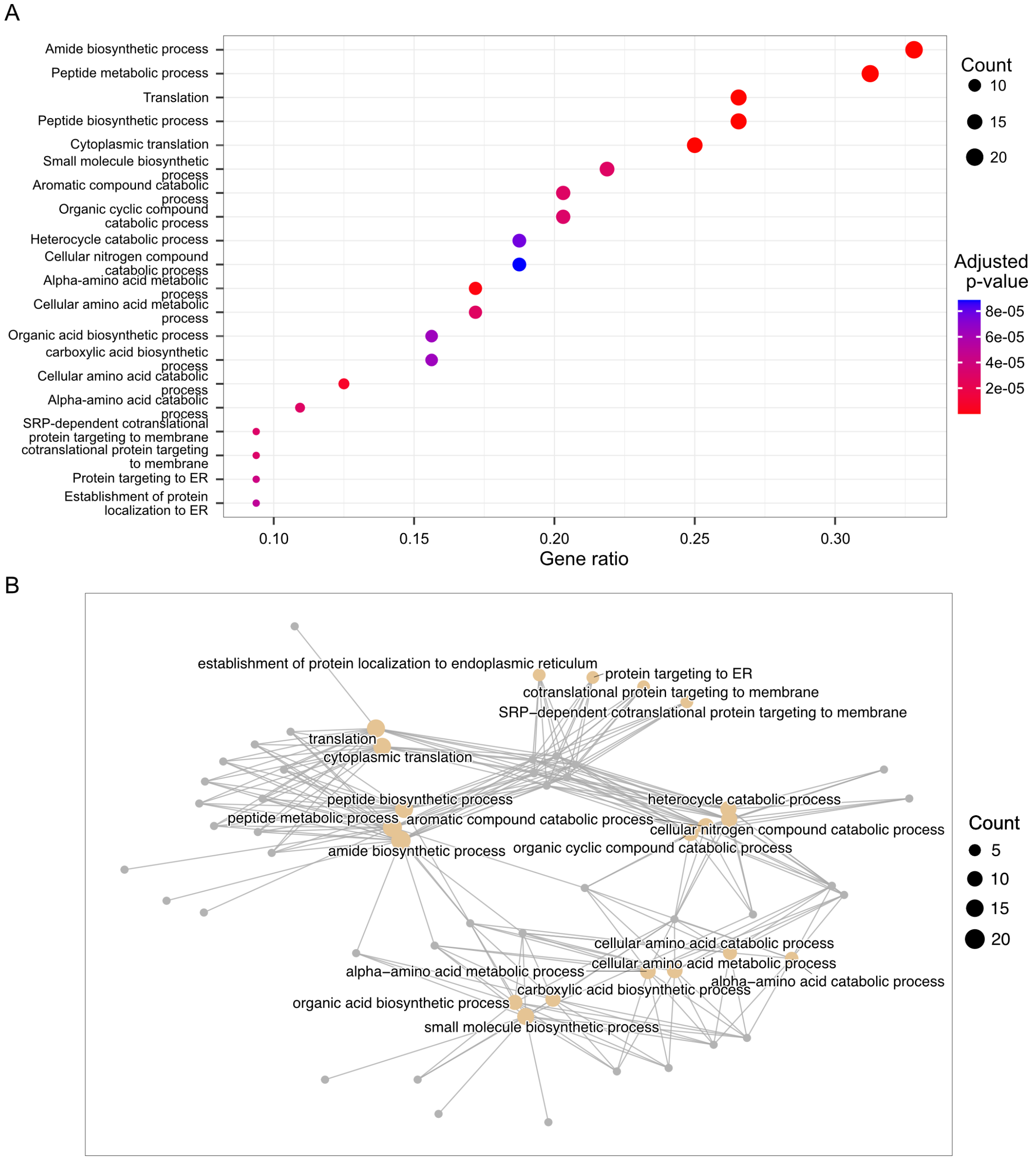


Figure 3.35: GO enrichment analysis for significantly downregulated transcripts in the *thap* RNAi condition. A) Dot plot representation of the top 20 significant terms associated to significantly downregulated transcripts in the *thap* RNAi condition. B) Cnet plot representation of the top 20 significant terms (brown dots) associated to significantly downregulated transcripts (gray dots) in *thap* RNAi condition.

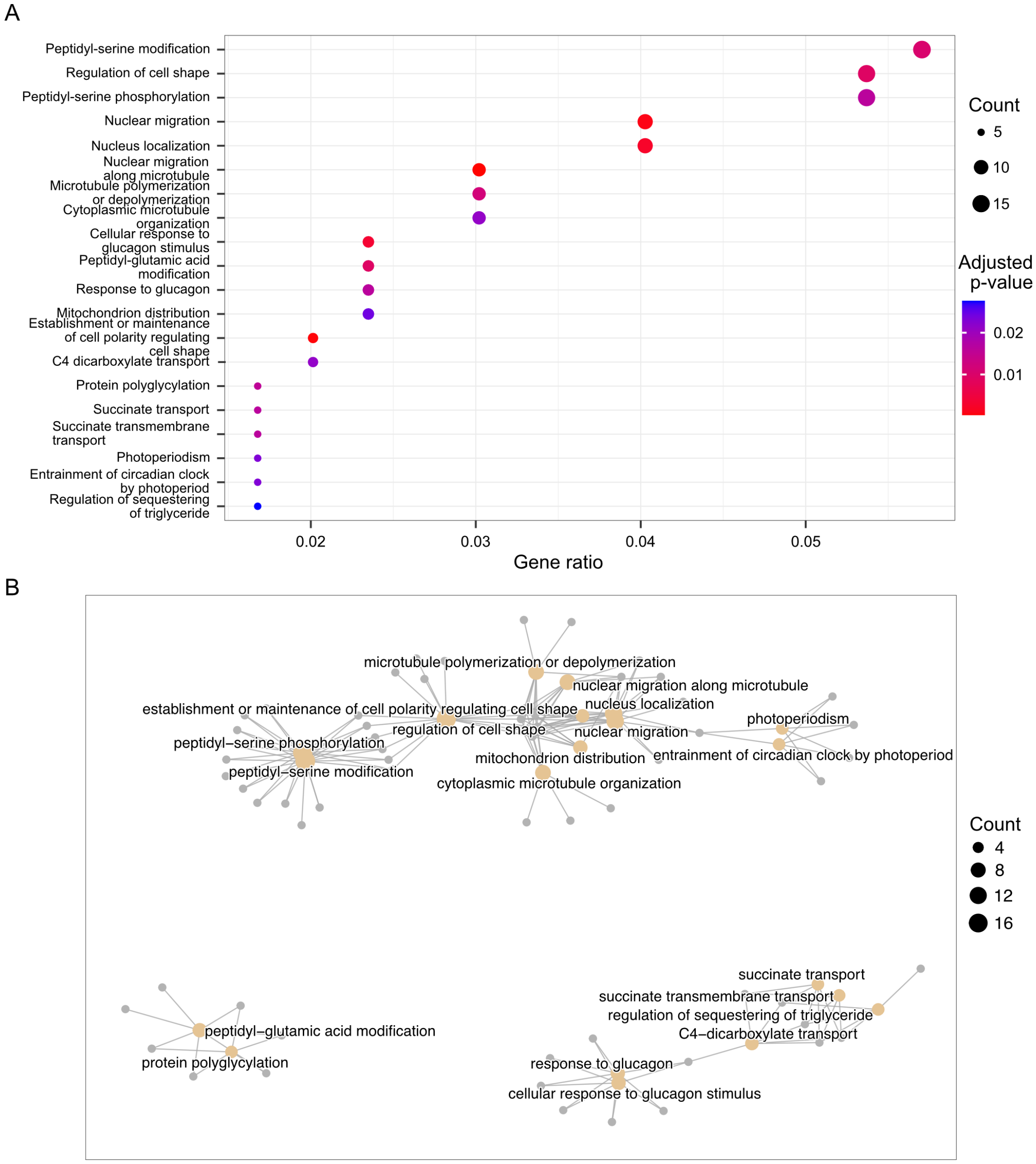


Figure 3.36: GO enrichment analysis for significantly downregulated transcripts in the *cebp* 4 RNAi condition. A) Dot plot representation of the top 20 significant terms associated to significantly downregulated transcripts in the *cebp* 4 RNAi condition. B) Cnet plot representation of the top 20 significant terms (brown dots) associated to significantly downregulated transcripts (gray dots) in the *cebp* 4 RNAi condition.

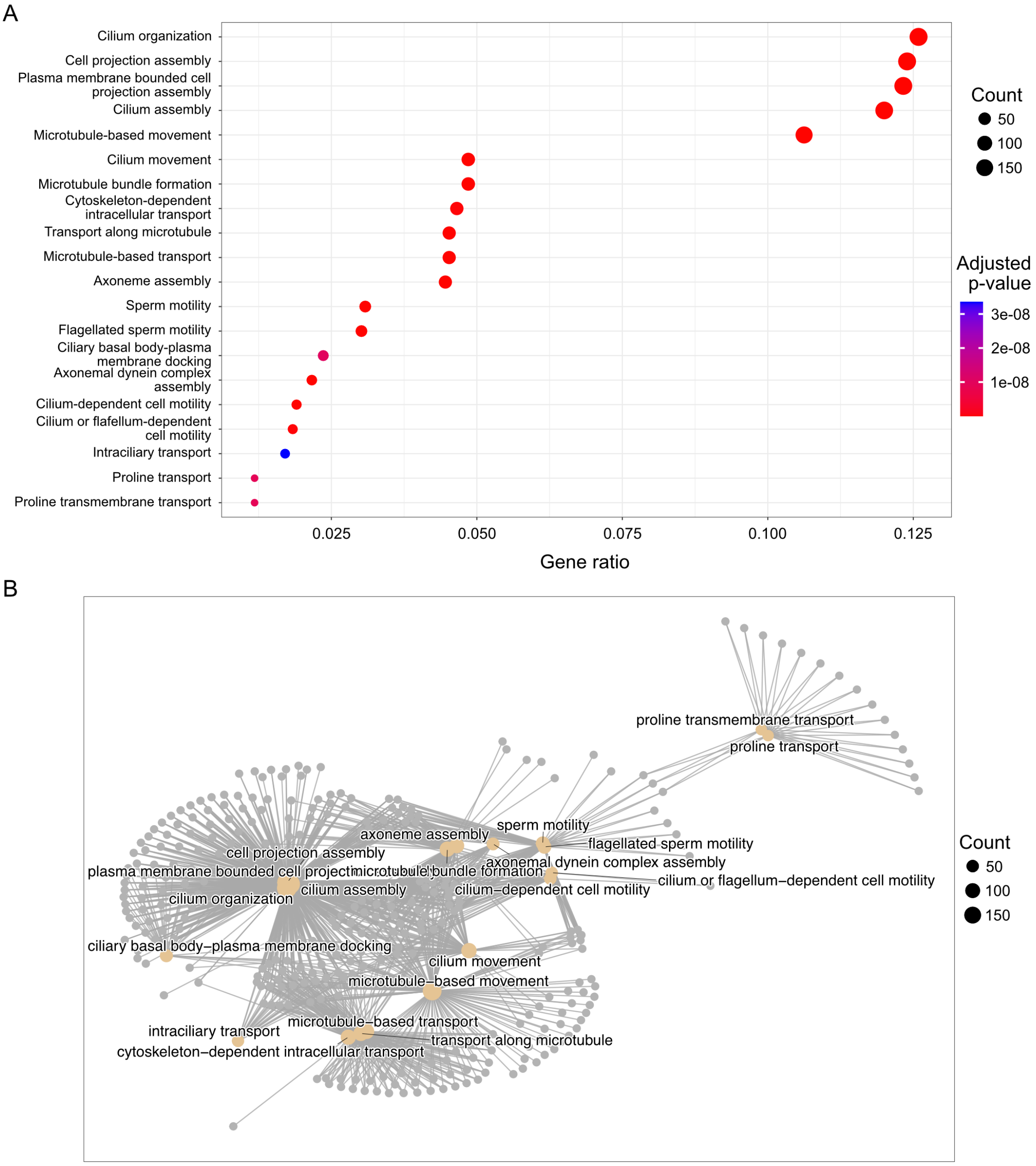


Figure 3.37: GO enrichment analysis for significantly downregulated transcripts in the *tead 1* RNAi condition. A) Dot plot representation of the top 20 significant terms associated to significantly downregulated transcripts in the *tead 1* RNAi condition. B) Cnet plot representation of the top 20 significant terms (brown dots) associated to significantly downregulated transcripts (gray dots) in the *tead 1* RNAi condition.

3.5.4 Visualization of sexual markers after RNAi of TF candidates

With the results of the DGEA in mind, I set out to confirm the state of sexual tissues by visualizing the expression pattern and intensity of markers from these tissues. I set up an RNAi experiment as depicted in Figure 3.38 A.

I used *ferritin* 2 as a yolk marker (Figure 3.38 B). Both *egfp* and *cebp* 4 showed a normal expression pattern. *ophis* and *tead* 1 RNAi worms did not show any signal. Interestingly, the *thap* RNAi condition showed a *ferritin* 2 signal but the expression pattern was less reticulated compared to *egfp* and *cebp* 4. The testes marker, *plastin*, was absent in both *tead* 1 and *ophis* RNAi conditions. It was also more weakly expressed in *cebp 4* RNAi worms, as expected. Interestingly, the *cebp* 4 condition still possessed testes lobules, compared to *tead* 1 and *ophis* RNAi worms, but I observed no progression of the developing sperm past the spermatocyte stage (Fig 3.39). Interestingly, the number of spermatogonia and spermatocytes per testes lobule seemed to be higher than the control, indicating a differentiation defect. Signal for the oocyte marker, *gwin*, was drastically reduced in *thap* RNAi animals compared to *egfp* (Figure 3.38). *cebp* 4 and *tead* 1 RNAi animals still showed oocyte signal although the staining was weaker. The sperm duct marker, *granulin*, showed similar expression pattern in all conditions except *ophis*, where it was absent. The shell gland markers, *tsp1* and *tsp 66e* had similar expression patterns. *cebp* 4 RNAi worms showed no decreased expression whereas *tead 1* RNAi had the most drastic decrease in signal intensity besides the positive control. Additionally, *thap* RNAi animals also showed decreased expression of the shell gland marker, especially for *tsp 66e*.

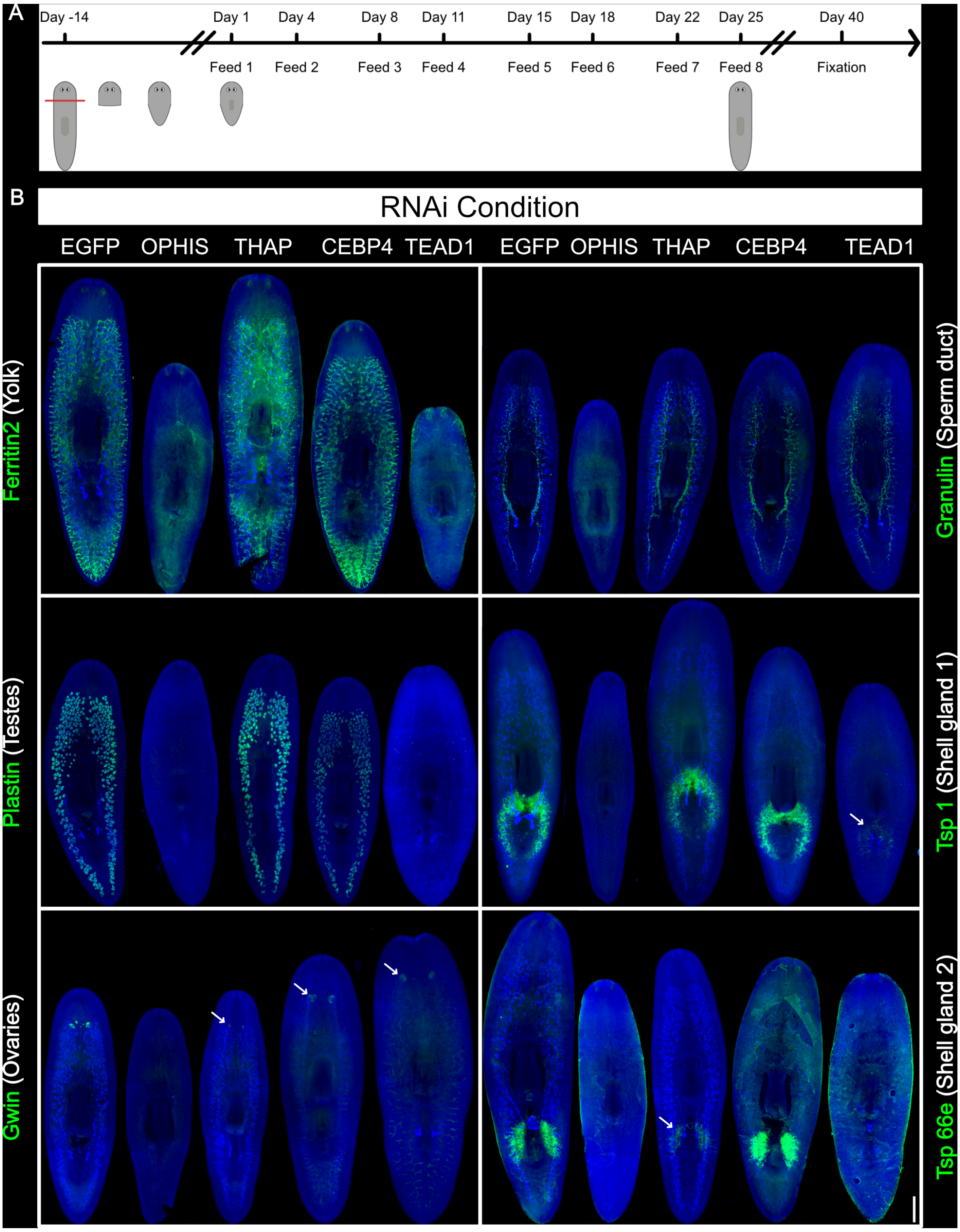


Figure 3.38: Integrity assessment of sexual reproduction-related tissues after RNA interference of TF candidates. A) Workflow used for the RNAi experiment. Sexually mature worms are cut and heads devoid of reproductive organs are left to regenerate for 2 weeks. Regenerated worms are fed 8 times RNAi liver over a period of 4 weeks to grow to a sexually mature size. Worms are then left to starve for 2 weeks and fixed for whole mount *in situ* hybridization (WISH). B) Results of the WISH on markers for sexual reproduction-related tissues after RNAi for TF candidates as well as a positive (*ophis*) and negative (*egfp*) control. The scale bar represents 1 mm.

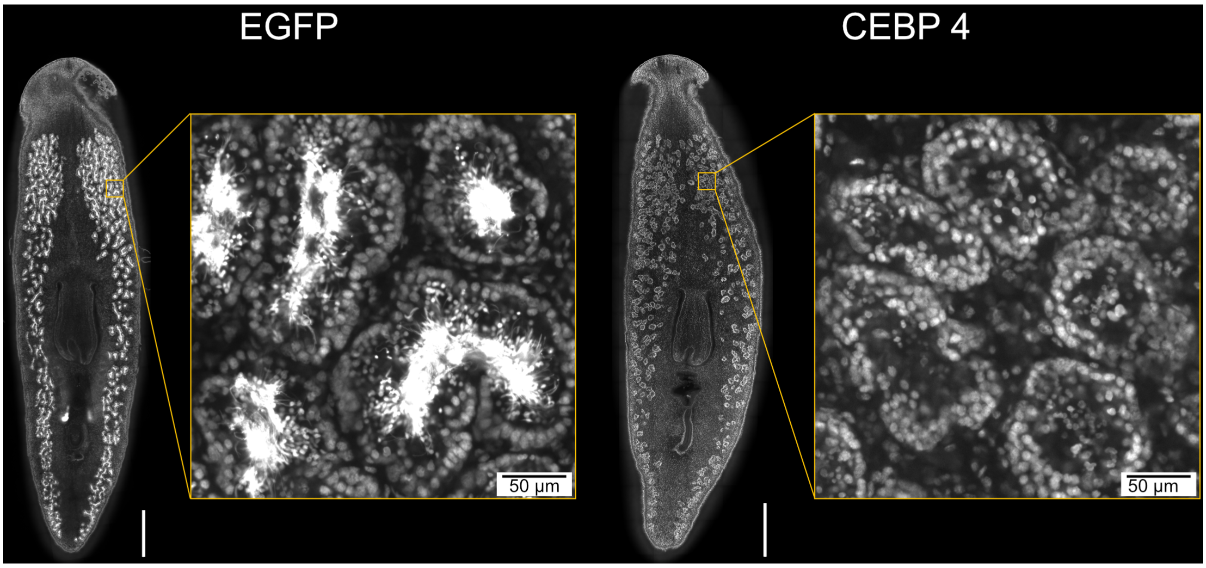


Figure 3.39: The *cebp* 4 candidate shows defects in sperm development. DAPI staining on *egfp* RNAi (negative control) and *cebp* 4 RNAi worms. Overview images were obtained by maximum intensity projections of confocal sections. The scale bar represents 1 mm. The zoom on planarian testes is composed of a single confocal section.

Overall, the results of the WISH experiment were consistent with what I observed in RNA-seq and showed that the three candidates, *thap*, *cebp* 4 and *tead* 1 possess a phenotype where one or more sexual tissue is absent or atrophied.

Chapter 4

Discussion

4.1 Development of a Start-seq protocol in planaria

There are many ways to identify and study regulatory elements, each focusing on a one of its specific characteristics. ChIP-seq targeting histone modifications like H3k4me3, H3k4me1 or H3K27ac relies on the abundant presence of these marks at these sites to identify promoters and enhancers. Open chromatin, assayed by ATAC-seq is also a widely used technique due to its low input requirement and applicability to many different species. Transcription initiation assays were first used to study RNA pol II initiation and pausing at promoters but researchers soon realized that transcription initiation was also a hallmark of both promoters and enhancers.

If ones aim is to identify differentially active REs during a developmental process or simply between two conditions, profiling the transcription initiation landscape provides multiple advantages over alternative methods such as looking at histone modifications or probing for open chromatin regions. Indeed, promoter RNA II pol binding and transcription initiation represent the initial stages of active transcription. Therefore, assessing promoter activity by looking at transcription initiation is a more direct way than to profile open chromatin or histone modifications. Transcription initiation not only shown to be a good predictor of promoter activity but also of enhancers (Andersson et al., 2014; Kim et al., 2010; Mikhaylichenko et al., 2018). Notably, a study performed by Duttke and colleagues showed that, upon stimulation of bone marrow-derived mouse macrophages, transcription initiation profiling showed the best correlation with active transcription compared to ATAC-seq and H3K27Ac ChIP-seq (Duttke et al., 2019). This technique has the advantage to be an RNA-based method and therefore benefit from a higher dynamic range than DNA-based sequencing methods such as ChIP or ATAC-seq which can maximally yield only one read per genomic locus. Additionally, it has also the ability to uncover sites of transcription initiation at the nucleotide resolution, allowing for more precise delineation of REs compared to the two other methods. Finally, transcription initiation also has the potential to identify the true transcript start sites of genes undergoing trans splicing (R. A.-J. Chen et al., 2013), a process also documented in planaria (Rossi, Ross, Jack, & Alvarado, 2014; Zayas et al., 2005). These arguments underscore the potential advantages of developing a transcription initiation assay in planaria.

In the first section of the results, I outlined the development of a method for isolating pure native nuclei, a crucial prerequisite for obtaining short capped nuclear RNA. This process presented several challenges that that I had to overcome in order to isolate intact short nuclear RNA.

One of the main hurdles in making this protocol usable was the initially poor yield and input requirements for sufficient short RNA extraction. The main factor leading to material loss was tissue aggregation during mechanical lysis. To address this issue, I reduced the input amount, resulting in a higher relative yield of nuclei. Modification of the step density gradient also allowed all the nuclei to pellet at the same height and be collected with reduced loss. Another challenge encountered was the presence of planarian pigments, which are known to co-precipitate with nucleic acids and interfere with library preparation (Grohme et al., 2018). The reduction of input material greatly helped with the RNA eluate color, indicating reduced pigment contamination. Furthermore, the isolation of short RNAs by gel extraction following size selection likely eliminated any remaining pigments since a clear eluate was obtained following this procedure. The most critical challenge was to stabilize the nuclear RNA content during the extraction process. Initially, I observed unstable RNA content which manifested as a ladder-like pattern after gel electrophoresis. To address this issue, I made modifications to the extraction buffer, including the addition of RNAse inhibitors in sufficient quantities and switching from a MOPS to HEPES as a buffering system. These changes greatly improved RNA stability throughout the extraction process.

Start-seq is one of the multiple transcription initiation assays that have been developed throughout the years. It stands out as a relatively simple protocol, in contrast to others like 5' GRO-seq, PRO-cap and GRO-cap, which necessitates a run-on reaction. This simplicity provides a significant advantage when working on whole organisms such as *S. mediterranea*. Indeed, the addition of a run-on reaction would require a very high nuclear integrity, a condition that was not guaranteed to be achieved at the time of choosing which method to use. However, I believe that the developed extraction protocol could be suitable for such protocol and could be used as a basis to explore this option should the need arise.

One limiting factor for efficient application of such methods remains the input requirements. A literature review revealed that the 3 aforementioned techniques employ a wide range of input material. Notably, the original 5' GRO-seq protocol uses the lowest number of nuclei (5 Mio) while PRO-cap (20 Mio) and GRO-cap (100 Mio) requires larger quantities. Additionally, it is worth noting that only the GRO-cap protocol was performed on a whole animal, namely *C. elegans*, while the other two techniques applied on cell lines. The same study also used 100 Mio nuclei for GRO-seq in *C. elegans* (Kruesi et al., 2013) which hints at the increased difficulty of using such type of techniques in whole organismal research. In comparison, the Start-seq libraries generated in this study used between 7 and 14 Mio nuclei.

This works presents the first nuclei extraction protocol proven RNA integrity in *S. mediterranea*. It could also serve as a foundation for other methods enabling scientists to investigate a variety of research questions. For instance, it could be used to perform single nuclei sequencing experiments if current single cell RNA sequencing approaches are not suitable. Additionally, it could be the basis for other methods such as TT-seq (Schwalb et al., 2016) to maps active transcription and would be able to profile both RE activity and mRNA synthesis rates at the same time (Michel et al., 2017).

The developed protocol does have some caveats, especially after RNA extraction. It remains an input-intensive and lengthy protocol. Transcription initiation profiling in single cells is therefore not yet applicable. However, several improvements could address these problems. First, optimization of RNA precipitation could be done using magnetic beads. If minimal losses are observed, this modification could significantly shorten the protocol length. Second, magnetic beads could also be used to optimize the size selection step after RNA purification, a step known to be prone to material loss. Successfully integrating beads-based size selection would substantially lower the initial input requirements. Third, it may be worthwhile to investigate whether certain enzymes can function effectively in the buffers used during earlier enzymatic steps. This approach is already employed in in commercially available kits like the NEBnext small RNA library preparation from NEB. Notably if the 3' adapter ligation, phosphatase treatment and cap removal steps all could work without the need for RNA cleanup steps in between, it would again greatly reduce sample losses and overall protocol length. Lastly, once all these suggestions have been implemented, a titration experiment could help determine the minimum number of nuclei needed to generate high-quality Start-seq libraries.

Although this method does come with certain limitations and could benefit from further optimization to lower input requirement, its use along other epigenomic profiling methods will certainly be beneficial to research important biological questions. Future research should explore whether this method can effectively be applied to whole body regeneration studies, one of the main features that make planarians an attractive model.

4.2 Transcription initiation landscape of *Schmidtea mediterranea*

The small size of Start-seq reads poses a limitation as only a fraction of reads can be uniquely assigned and used in downstream analysis. A preliminary analysis hinted at the fact that many of the multimapping reads were found in repetitive regions and could therefore be the sign of active transposable elements (TE) or from TE-derived regulatory elements (Chuong, Elde, & Feschotte, 2017). Despite the low mapping rate, I calculated that about 50 Mio reads was necessary to call all active REs detected in this study (Figure 3.6 B). This quantity of reads is still reasonable, especially since sequencing costs keep getting lower.

In this part of the thesis, I aimed at characterizing the identified TICs and establishing a case for calling them putative regulatory elements. This study demonstrates that the identified TICs share similar features to those described for REs including their localization, chromatin environment, bidirectional transcription initiation and motif content.

The chromatin environment surrounding TICs closely resembled patterns described in other species. Specifically, I observed a pronounced peak of H3K4me3 at promoters of active genes (Santos-Rosa et al., 2002), the presence of H3K27Ac at active enhancers (Creyghton et al., 2010) and promoters (Z. Wang et al., 2008) as well as the presence of the H3K4me1 mark at enhancers (Heintzman et al., 2007) (Figure 3.8). The presence of H3K4me3 signal in enhancers TICs could be due to two reasons. Firstly, it is possible that some genes may not be included in the current version of our genome annotations, leading to a misclassification of REs with promoter-like characteristics as enhancers and increase the H3K4me3 signal in this subset of REs. Additionally, highly active enhancers have been reported to possess tri-methylated lysine of H3 in *Drosophila* and mice (Henriques et al., 2018).

As shown in other species such as yeast (Neil et al., 2009), human (Core et al., 2008), *Drosophila* (Mikhaylichenko et al., 2018), *C. elegans* (R. A.-J. Chen et al., 2013), Zebrafish (Baranasic et al., 2022) and even rice (Duttke et al., 2019), bidirectional transcription at regulatory elements is prevalent in *S. mediterranea.* However, as approximately 50% of identified promoters and 40% of enhancers in *S. mediterranea* were classified as unidirectional, we could therefore conclude that this characteristic is not an inherent feature of REs in this species. However, given the information detailed in the introduction, I believe that the lack of widespread bidirectional transcription initiation observed at REs can be attributed to the sensitivity of the method as well as the threshold applied to call bidirectional transcription initiation.

Another interesting observation was that highly active enhancers showed antisense transcription further away from the sense strand, in contrast to less expressed enhancers (Figure 3.9). This finding aligns with the work of Scruggs and colleagues, who demonstrated that more active enhancers tend to have more distal antisense transcription and could possibly be explained that these more active enhancers have a larger NDR (Scruggs et al., 2015). Like in that study, it would be interesting to group enhancers and promoters by distance between sense and antisense transcription and see with mononucleosomal reads from ATAC-seq if the NDR is larger.

In terms of motif content, I demonstrated the presence of core promoter motifs such as the TATA box, Inr, and DPE in *S. mediterranea.* Additionally, the CCAAT-box is also found enriched at REs. The CCAAT-box is highly conserved in eukaryotes and is bound by the nuclear factor Y (NF-Y), which consist of a heterotrimer composed of NF-YA, NF-YB and NF-YC (X.-Y. Li et al., 1992; Maity & De Crombrugghe, 1998). This complex is involved in regulation of housekeeping genes but also has functions in stem cell identity by promoting open chromatin for TF binding (Oldfield et al., 2014).

In planarians, multiple paralogs of NF-Y proteins exist, including two for NF-YA and NF-YB and one for NF-YC (Iyer et al., 2016). A study in asexual *S. mediterranea* showed that, *nf-yA1, B2* and *C* are ubiquitously expressed with enrichment in the cephalic ganglia (Rodríguez-Esteban, González-Sastre, Rojo-Laguna, Saló, & Abril, 2015). These proteins were shown to play a critical role in neoblasts, as knockdown of these members resulted in a neoblast depletion phenotype. In sexual planarians, *nf-ya 1, nf-ya 2, nf-yb 2* and *nf-yc* show enrichment in testes but are also expressed somatically (Iyer et al., 2016). Similar to the asexual strain RNAi of *nf-ya1, b2* and *c* show a stem cell phenotype. Notably, the paralog NF-YB has a testes-specific function by regulating spermatogonial stem cell renewal (Iyer et al., 2016; Y. Wang et al., 2010). NF-Y complexes made of different subunits may therefore have distinct targets in planaria, which could explain the differences in phenotypes observed following knock-down.

The DPE motif exhibited a high signal over the whole assayed interval with an enrichment at the center of the both the promoter and enhancer TICs (Figure 3.10). This 'background signal' could be attributed to the PFM used to probe for this motif. The use of a different PFM could help to mitigate this background signal. Indeed, multiple motifs have been described for DPE (Haberle & Stark, 2018). An intriguing observation is the disparity between this study and existing literature regarding the DPE motif's location. Typically, the DPE motif is found around +30 nucleotides relative to the transcription initiation site. However, in this study, it was identified at the transcription initiation site itself (-1 nucleotide). Interestingly, a structural study of the PIC bound at promoters showed that TAF1, a protein part of the TFIID general transcription factor and usually binding to the Inr motif (Chalkley & Verrijzer, 1999), was found to interact with the DPE as well (Louder et al., 2016). Therefore, one could imagine that DPE could replace the function of the Inr motif by binding to TAF 1 at the transcription initiation site. Nevertheless, it's essential to acknowledge that this hypothesis currently lacks robust supporting evidence, and further research is needed to substantiate it conclusively.

In enhancers, a second TATA-box motif was found to be enriched both at its expected location (- 30nt) as well as at the transcription initiation site. This suggest that some enhancers might have a mis-positioned TATA-box and further investigation is needed to explain this localization.

Core promoter motifs were found both in enhancer and promoter TICs, exhibiting similar patterns. This observation is not in accordance with the literature where core promoter motifs are indeed found at both promoters and enhancers but observed that enhancers contain weaker or more degenerate core promoter sequences (Haberle & Stark, 2018). Only enhancers with promoter-like characteristics contained motifs that were closer to the consensus (Mikhaylichenko et al., 2018).

Regarding motif orientation, some motifs exhibited a preferred orientation such as the TATA-box, DPE and the initiator motif. Interestingly in enhancer, the CCAAT-box also shows an orientation preference that is not present in promoters. The enrichment of initiator motif upstream of the transcription site in enhancers is intriguing and may suggest the presence of a reverse-oriented core promoter. This orientation preference could potentially contribute to bidirectional transcription in a subset of enhancers, as mentioned earlier. It is however interesting that this enrichment is observed upstream of the sense motif rather than downstream since bidirectional transcription is thought to occur more in a divergent fashion at both ends of a NDR (Duttke et al., 2015; Scruggs et al., 2015). Nevertheless, convergent transcription has also been described in the literature (Hobson, Wei, Steinmetz, & Svejstrup, 2012) and could also be seen in *S. mediterranea* (Figure 3.7 C). Further analysis is required to understand if this core promoter orientation is more prevalent than the more classical divergent orientation. Additionally, it is possible that divergent transcription my not be visible by analysis Inr motif orientations if this process does not occur consistently in a defined distance from the sense transcription initiation site. This could explain the lack of visible motif enrichment.

Overall, the frequency of core promoter motifs in *S. mediterranea* was observed to relatively low. This could be attributed to the possibility that the correct position frequency matrices (PFMs) were not employed to accurately identify the true core promoter motifs specific to this species. For example, the Inr motif in flies and humans are very different (Haberle & Stark, 2018). It is important to note that this analysis represents an initial exploration of motif content within regulatory elements in *S. mediterranea*. Further investigations are necessary to identify and better characterize core promoter motifs in this organism. One possible avenue would be to do a *de novo* motif search on promoters to identify enriched sequenced within these REs. In the process of writing the thesis, a study performed by Poulet and colleagues performed such analysis (Poulet et al., 2023). There, they propose that a motif enriched at the TSS could be a candidate for the Inr motif in planaria.

In conclusion, planarian TICs share features of REs in terms of location, chromatin environment and motif content. These shared features provide a basis for designating the identified TICs as putative REs. To prove that they actually are REs would require experiments which show that they in fact do regulate gene transcription. One could think of using massively parallel reporter assays (MPRAs)like STARR-seq (Arnold et al., 2013) to assess enhancer functions. Other MPRA setups where the putative RE is placed in front of the reporter gene can be used to test for promoter activity.

4.3 Differential regulatory element activity analysis for the identification of gonadal TFs

Identification of motifs in differentially regulated REs has been successfully carried out in the past in *S. mediterranea* using various techniques, including ChIP-seq, ATAC-seq, or a combination of both (Neiro et al., 2022; Pascual-Carreras et al., 2023). Start-seq has been used in other model organism (R. A.-J. Chen et al., 2013; Nechaev et al., 2010) but it's use has been restricted to identify REs characterize their properties such as bidirectional transcription and study RNA pol II pausing (R. A.-J. Chen et al., 2013; Henriques et al., 2013, 2018; Nechaev et al., 2010). Similar transcription initiation methods have been used in other model systems to study more cellular and developmental processes. For example, Duttke and colleagues used csRNA-seq during BMDM activation using TLR4 agonist Kdo2-lipid A (KLA) (Duttke et al., 2019). Employing this transcription initiation method, they identified changes in motif prevalence in differentially activated REs upon KLA stimulation. These REs were enriched for motifs bound by the major drivers of KLA response mediation (NF-κB and AP-1). I used the same reasoning to identify motifs important for the planarian germ line.

My results reveal that many putative REs exhibit differential activity between the two *S. mediterranea* biotypes, and these putative REs possess variable motifs that are associated to TF families with known functions in planarian gonadal functions. Moreover, I identify motifs belonging to other TFs families that have not been previously characterized in the context of planarian sexual reproduction and identify candidate TFs that could potentially bind them.

Prior to this, I investigated the overall agreement between Start-seq differential regulatory element activity analysis (DREAA) and RNA-seq differential gene expression analysis (DGEA) results. This comparison aimed to provide support for the notion that differences in regulatory element (RE) activity indeed translated into differences in gene expression. The fact that there was a good overlap between DGEA and DREAA gave more confidence in the results.

However, I observed that many significantly differentially expressed genes had a l2fc with an opposite sign compared their promoter l2fc. One caveat of this analysis is that I compared differential promoter activity to the differential abundance of mature transcripts. Therefore, transcript stability could be an additional variable explaining why some genes show opposite l2fcs in my analysis. A better way to perform this analysis would be to look at productive elongation, as was done in the study performed by Larke and colleagues (Larke et al., 2021). By analyzing transcription initiation data together with the investigation of transcripts originating from RNA pol II in productive elongation, Larke and colleagues were able to show a strong correlation between transcription initiation and productive elongation. This reinforces the argument put forth earlier.

Differential RE activity analysis revealed that many REs are both up and down regulated in sexual *S. mediterranea* compared to the asexual biotype, with the majority being upregulated (75,6%). This outcome was expected since the planarian gonad is exclusively present in the sexual biotype. Consequently, REs associated to GRNs controlling the development and maintenance of these tissues are expected to be only active in the sexual biotype. One prominent organ in sexual *S. mediterranea*, the vitellaria (also known as the yolk glands), was well represented among the most upregulated promoters (Figure 3.15). However, this was not the case in the GO terms found enriched for the sexual biotype (Figure 3.16). I suspect that it is because planarian yolk glands and ectolecithality in general is an evolutionarily derived feature of a subgroup of platyhelminths (neoophora) (Laumer & Giribet, 2014; Martín-Durán & Egger, 2012) and would therefore rely mostly on genes not present in other animals. Since the assignment of GO terms to planarian genes is based on homology with genes with a functional annotation in other species, many planarian-specific genes were not annotated and therefore absent in the analysis. This observation underscores the need for the flatworm research community to intensify efforts aimed at characterizing flatworm-specific genes. For example, *S. mansoni*, a parasitic flatworm causing significant mortality in many impoverished regions, relies on a vitellaria for reproduction (J. Wang, Chen, & Collins III, 2019). Improving our knowledge about the characteristics of tissues specific to certain groups of organisms would bring new therapies to specifically target diseases they inflict.

In an attempt to identify important regulators of planarian gonadal development and maintenance, I sought to identify motifs located within putative regulatory elements that were variable between sexually and asexually reproducing worms. Several TFs had already been shown be required in various steps of germ line development and maintenance. A study from Issigonis and colleagues identified a Krüpple-like factor expressed in what is thought to be the gonadal stem cells. Furthermore, knock down of this factor resulted in the abrogation of the whole germ line as well as the vitellaria (Issigonis et al., 2022). A previously mentioned study identified one paralog of NF-YB to be required for the maintenance of spermatogonial stem cells (Y. Wang et al., 2010) while Iyer and colleagues showed that other NF-Y factors are highly enriched in the testes (Iyer et al., 2016). Finally, Khan and colleagues identified by laser capture microdissection a *foxL* homolog expressed in the somatic ovarian cells (U. W. Khan & Newmark, 2022). Knockdown of this gene leads oocyte differentiation defects.

Motifs associated to each of the TF families mentioned above were found to be significantly variable between sexual and asexual planarians, confirming the potential of Start-seq for TF discovery by differential regulatory element activity analysis. Novel motifs identified in this study belong to TFs from the GATA, THAP, C/EBP, NF-κB related, EBF and JUN:FOS families. These TF families have been extensively studied in other species. In the following paragraphs, I describe their known functions and attempt to link them to sexual reproduction related functions.

GATA: The GATA family of transcription factors are important conserved regulators with described roles in cell differentiation, organ morphogenesis and development (Flores, Oviedo, & Sage, 2016; Lentjes et al., 2016; Scazzocchio, 2000). Among its many roles GATA factors were shown to be expressed in the gonads of many species such as fruit flies (Lossky & Wensink, 1995), humans (Ketola et al., 2000), mice (Ito et al., 1993) snakes and birds (Singh, Wadhwa, Naidu, Nagaraj, & Ganesan, 1994). Their essentiality for proper gonadal development has also been shown. For example, in mice, absence of GATA 4 during development abrogates the formation of the genital ridge, which differentiates into the testes or ovary (Y.-C. Hu, Okumura, & Page, 2013).

THAP: THAP TFs are characterized by their conserved THAP domain comprising a C2H2 zinc finger domain (Roussigne, Cayrol, Clouaire, Amalric, & Girard, 2003). Interestingly this domain shares a striking resemblance to the DNA binding domain of the *Drosophila* P-element transposase and are thought to have derived their DBD from this transposable element (Quesneville, Nouaud, & Anxolabehere, 2005). They have been shown to be involved in many different processes. In humans, THAP proteins were described to control cell proliferation (Cayrol et al., 2007) and regulation of apoptosis (Roussigne et al., 2003). In mice, a THAP protein called RONIN regulates embryonic development by controlling embryonic stem cell proliferation and differentiation (Dejosez et al., 2008). It does so by repressing target genes such as *gata 4* and *gata 6* which are involved in the development and differentiation of endoderm and mesoderm-derived tissues such as the cardiovascular tissue (Kuo et al., 1997). Interestingly, *ronin* is highly expressed in the mouse oocyte and may be involved in oocyte maturation (Dejosez et al., 2008). In *C. elegans*, the THAP domain containing TF, LIN-15B, is maternally inherited and regulates primordial germ-cell development (C.-Y. S. Lee, Lu, & Seydoux, 2017). Moreover, HIM-17, another THAP containing TF has been shown to directly regulate many germline genes by binding to a transposable element-derived motif (Carelli et al., 2022).

C/EBP: Members of the C/EBP transcription factor families have been associated to a variety of functions such as the liver homeostasis (Grøntved et al., 2013), adipose tissue differentiation (Darlington, Ross, & MacDougald, 1998) or granulopoiesis (Hirai et al., 2006). One Study in *Drosophila* also shows that a C/EBP TF member, Slbo, is essential for sexual reproduction by controlling cell migrations leading to the formation of the micropyle (Rørth, Szabo, & Texido, 2000), a passage through which sperm can fertilize the oocytes. As part of the b-zip class of TFs, C/EBP TFs form dimers to bind to DNA. Interestingly, C/EBP do not only form homodimers or heterodimers with other C/EBP family members but have been shown to interact with members of the JUN and FOS TFs (Ubeda, Vallejo, & Habener, 1999) or CREB/ATF families (Vallejo, Ron, Miller, & Habener, 1993). Furthermore, interactions with non-bzip TFs such as members of the NF-kb-related family like REL A (Chumakov, Silla, Williamson, & Koeffler, 2007) and NF-κB 1 (LeClair, Blanar, & Sharp, 1992). Motifs associated to TFs of members of both the NF-κB-related, JUN:FOS families were also found highly variable in planarians suggesting a possible interaction between these TF families in the planarian gonad.

NF-κB-related: The NF-κB family of transcription factors is characterized by their DNA binding and dimerization domain both situated in the Rel Homology Region (RHR). It is a very well described family of TFs with their main function being the regulation of the innate immune system in a variety of organisms as well as the adaptive immune system in vertebrates (Zhang, Lenardo, & Baltimore, 2017). Additionally, NF-κB has been implicated in ovarian development in Zebrafish (Pradhan et al., 2012). This is done by inhibition of apoptosis in the structure developing into the testes, the juvenile ovary. By inhibiting apoptosis as well as suppressing the expression of genes important for testes development, NF-κB is able direct the development of the zebrafish reproductive system towards female structures.

EBF: Early B-Cell factor members have been studied in mice for their role in B-cell development (Hagman, Belanger, Travis, Turck, & Grosschedl, 1993), neuronal development (S. S. Wang, Tsai, & Reed, 1997) and adipogenesis (Jimenez, Åkerblad, Sigvardsson, & Rosen, 2007). They are part of the wider COE transcription factor family and have been found throughout metazoans (Daburon et al., 2008). There, they have been implicated for example in *Drosophila* early development (Crozatier, Valle, Dubois, Ibnsouda, &

Vincent, 1996) as well as neuronal differentiation in *C. elegans* (Prasad et al., 1998). In *S. mediterranea*, a COE family member is required for brain maintenance and regeneration (Cowles et al., 2014). No direct link to sexual reproduction or gonadal development could be found for this family of TFs. Interestingly, orthologs of genes involved in B-cell differentiation in vertebrates had a significantly differentially upregulated promoter in sexual planaria (Figure 3.16). This could potentially mean that some part of the GRN involved in vertebrate B-cell differentiation is used in planaria to serve other functions related to sexual reproduction.

JUN FOS: JUN and FOS TFs, together with ATF protein members are the main components of the AP-1 transcription factor complex. Depending on the heterodimer composition of AP-1, it can have very different effects on gene regulation (Karin, Liu, & Zandi, 1997). It is involved in processes such as cell proliferation (Karin et al., 1997), differentiation (Madrigal et al., 2023) and apoptosis (Ameyar, Wisniewska, & Weitzman, 2003). It does this by integrating a plethora of external stimuli such as cytokines, growth factors and other stress signals (Hess, Angel, & Schorpp-Kistner, 2004). AP-1 has also been implicated in gonad morphogenesis in *D. melanogaster* where it regulates ensheathment of germ cells by somatic gonadal precursor cells (Jemc, Milutinovich, Weyers, Takeda, & Van Doren, 2012). Since planaria constitutive generate progenitors using their pool of adult stem cells, a similar function of AP-1 could be envisaged in adult worms.

In addition to these families, I also decided to include the TEAD and SNAIL TF families since previous a previous chromVAR analysis showed motifs of these TF families to be significantly variable.

TEAD: TEAD family of TFs is best known to be the effectors of the Hippo signaling pathway (K. C. Lin, Park, & Guan, 2017). This evolutionarily conserved pathway regulates cell growth, proliferation and homeostasis and plays a critical role in stem cell functions, growth control and organ patterning during development in flies and vertebrates (Dong et al., 2007; Halder & Johnson, 2011; Lian et al., 2010). In planaria, two TEAD transcription factors exist and have been shown to regulate the homeostatic maintenance and regeneration of protonephridia as well as restrict neoblast proliferation (A. Y. Lin & Pearson, 2014). This study was done on the asexual strain of *S. mediterranea* so any potential implications of TEAD TFs in reproductive functions have not been explored. Interestingly, Scalloped (*sd*), the TEAD homolog in *Drosophila*, has been shown to regulate germ cell proliferation in the fly ovary and knockdown of the TF resulted in a significant reduction of germ cell number in the ovary (Sarikaya & Extavour, 2015).

SNAIL: The SNAIL TF family regulate embryonic development and are involved in processes requiring large-scale movements such as gastrulation or neural crest formation (Barrallo-Gimeno & Nieto, 2005). This is done by their regulation of genes necessary for epithelial-mesenchymal transition (EMT) (Carver, Jiang, Lan, Oram, & Gridley, 2001) and seems to be a conserved function of this TF family (Lespinet et al., 2002). They are considered to be repressors since they exert their function through repressing their target gene expression (Cano et al., 2000; Mayor, Guerrero, Young, Gomez-Skarmeta, & Cuellar, 2000). ESG, a Snail family member in *D. melanogaster* has also been reported to be expressed in male germ cells (Kiger, White-Cooper, & Fuller, 2000) as well as in the fly's embryonic somatic gonad (Streit, Bernasconi, Sergeev, Cruz, & Steinmann-Zwicky, 2002) and that its expression is required for male germ-line stem cell maintenance (Voog et al., 2014).

I decided to work with motif families instead of the specific TFs given in the output of the TOMTOM motif comparison tool for multiple reasons. First, orthology between *S. mediterranea* TFs and TFs of species binding the motifs present in the database used for motif matching has not been established. Moreover, performing this orthology assignment is out of my skillset. Given this limitation, I reasoned that expanding my search to encompass members of the identified TF families would address this issue and simultaneously broaden my initial list of candidates. This is the second reason for working with TF families instead. By extending the search to a whole family, the odds increase of finding regulators of gonadal development. This comes with the drawback that the association between the motif identified as variable and the potential candidates becomes weaker. Nonetheless, it's important to note that TFs within the same families tend to share similar DNA binding motifs (Ambrosini et al., 2020; Sielemann, Wulf, Schmidt, & Bräutigam, 2021). Additionally, the TFs binding these motifs in *S. mediterranea* might not have identical DNA binding domains to those associated to the motifs in database used in the motif comparison tool and could all be potential binders of the identified motifs. A definitive answer on whether these TF candidates bind the identified DNA motifs is beyond the scope of this thesis and would require the development of new resources such as ChIP-seq grade antibodies for TFs of interest or binding assays such as EMSA or transAM.

To increase the chances of finding TFs with important functions in gonadal development, I decided to focus on 6/8 TF families and increase the number of candidates for each selected family. Some families like the CEBP and SNAIL families had many more candidates that could have been investigated which could be a future research avenue.

One candidate of interest was the THAP family candidate. In addition to its THAP DNA binding domain, it harbored a Tesmin domain in its third exon. Interestingly, Tesmin domain-containing proteins have been shown to play a role in mouse spermatogenesis (Oji et al., 2020). More specifically, they accumulate in the cytoplasm in the pachytene stage of meiosis and then translocate to the nucleus just before meiotic division (Sutou et al., 2003) where it is thought to play a role in meiotic cell cycle regulation. Tesmin-domain containing proteins have also been described in other species such as *D. melanogaster* and *A. thaliana* to be important for the male or male and female fertility respectively (Andersen et al., 2007; Jiang, Benson, Bausek, Doggett, & White-Cooper, 2007).

4.4 Characterization and functional validation of TF candidates for sexual reproduction-related functions

Overall, the candidate TF selection method proved to be successful since 17 out of the 18 tested candidates showed enriched expression in reproduction-related tissues. Among these candidates, 9 were exclusively expressed in these tissues (Figures 3.20 until 3.26). Among these tissues, the testes were the most frequently represented organ. This aligns with expectations due to their abundant nature in sexually mature animals. As whole animals were used as input for this experiment, the proportion of nuclei originating from the testes was significant in comparison to other organs, such as the ovaries. It was therefore surprising to obtain an oocyte-specific TF like *gata 2* but could be explained since I decided to work with TF families instead of the specific TFs assigned to the identified motifs in the chromVAR analysis. *gata 1* had for example an expression pattern with high enrichment in testes and would therefore be more likely the TF associated to the identified motif.

Signal outside of clearly defined sexual structures like around the ovaries and around the pharynx will require more investigation. The expression pattern around the ovaries in *gata 2*, *thap* and *snail 2* (Figures 3.21, 3.23, 3.27) is reminiscent of the *klf4l* + female germ line progenitors (Issigonis et al., 2022). A double *in situ* hybridization with each TF and *klf4l* would give an answer to this observation. As for the pattern observed around the pharynx, there is currently no existing literature that aligns with this specific pattern. One possible experiment to try to determine if these cells are part of the reproductive system would be to perform *in situ* hybridization of the TFs showing this expression pattern in the asexual biotype. If no expression can be found there, it would indicate that they could play a role in sexual reproduction.

One clear observation was the lack of candidates TF expressed in yolk cells or shell glands, even though yolk is a prominent tissue in sexually mature planarians. Therefore, it was expected to find candidates expressed in this tissue at a similar frequency than the testes. Furthermore, the most upregulated promoters found in the differential regulatory element activity analysis belonged to yolk marker genes, proving that REs belonging to this tissue were present in the RE set used for the motif variability analysis. Multiple explanations can be put forth to explain this discrepancy. First, only a subset of TF candidates was tested with many more to be investigated. One could therefore think that some other untested candidates would be expressed in this tissue. Second, since ectolecithality is a specific feature to a subgroup of platyhelminths, it is plausible that the regulatory network responsible for the development and maintenance rely on TFs and motifs absent in other animals. Therefore, these motifs would not have been present in the motif database used to identify variable motifs and could explain why none of the TF candidates tested in this study showed expression in yolk cells. A similar reasoning could be used for the shell glands.

One unexpected observation from the RNA-seq results was that many TF candidate knockdowns had very little differentially expressed genes (Figure 3.29) and suggest some sort of functional redundancy between TFs. This redundancy is further supported by the similarity in expression patterns within TF families, where TFs of the same family exhibit nearly identical expression patterns. For example, *cebp 2* and *3* show a very similar expression pattern (Figure 3.22) and showed very little DE genes after knock-down (Figure 3.29).

A similar scenario was described in yeast. Hu and colleagues (Z. Hu, Killion, & Iyer, 2007) performed microarray experiments for 263 TF knockout strains and compared their differential expression results with sets of genes verified by ChIP-seq to be bound by TFs performed in a previous study (Harbison et al., 2004). Surprisingly, a very small overlap between each dataset was found with 3-6% agreement depending on the performed analysis (W.-S. Wu & Lai, 2015). Further analysis put forth arguments for transcription factor redundancy as an explanation for this discrepancy (Gitter et al., 2009).

Another factor that most likely played a role in the low amount of DE genes obtained after RNAi was the low replicate number used (n=2). For example, the *snail 2* RNAi condition shows a clear trend of upregulation in both testes markers (Figure 3.31) but stays below the threshold of significance. A power analysis followed by the adequate replicate number for RNA-seq would surely uncover many more differentially expressed genes in every condition.

As mentioned above the *snail 2* candidate showed an upregulation of many tested markers (Figure 3.31,3.33). This is consistent with the described role of this TF's family to be transcriptional repressors (as discussed previously). Interestingly, a similar upregulation of the *lecg* oocyte marker was observed, even though this TF is not expressed in oocytes. This could be an indirect effect of the knockdown, possibly resulting from misregulation in the female progenitor germ cell, where *snail 2* is believed to be expressed (Figure 3.23).

Sparse testes expression patterns such as the ones found for *snail 5* and *nfk 1* are also of interest and would require further investigation. Conducting knockdown experiments of these target genes and evaluating testes morphology could be an initial step in uncovering their function. The distribution and sparsity of *snail 5* signal (Figure 3.24) could potentially represent cells that are actively engaged in a process of meiotic cell division. It is known that Histone 3 phosphorylation is specifically regulated during mitosis and meiosis. A *snail 5 in situ* hybridization coupled with an immunostaining targeting phosphorylated Histone 3 would be a suitable experiment to verify this hypothesis. Furthermore, the sparse distribution on the outer layer of the testes lobule in *nfk 1* animals was also reminiscent of the *klf4l* or *dmd-1* expression pattern in the testes, marking the putative germline progenitors and the male somatic gonad respectively (Issigonis et al., 2022). A double *in situ* hybridization on these targets should in combination with *nfk 1* should be performed.

In the last part of the discussion, I will delve deeper into the TF candidates that showed an observable reproduction-related phenotype. I will discuss their phenotype in detail and attempt to uncover their potential function in sexual reproduction.

4.4.1 *cebp 4*

The *cebp 4* RNAi phenotype showed a clear deficiency for sperm development in sexual *S. mediterranea*. Although the testes were still present, they appeared empty, with a notable lack of mature sperm (Figure 3.39). Additionally, *in situ* hybridization results showed a reduction in the testes marker signal. This phenotype likely indicates a meiotic arrest before meiosis II since only very few cells with small round-shaped nuclei (designating round spermatids) were visible in the testes lobules with many having odd shapes. Another observation is that the outer layers of the testes lobules seemed more populated than the control. This could be attributed to an arrest in meiotic progression and the accumulation of earlier stages of sperm development.

GO enrichment analysis on downregulated genes in *cebp 4* RNAi shows they are involved in post-translational modifications, especially serine phosphorylation. This may suggest that *cebp 4*regulates genes involved in signal transduction pathways such as the MAPK and PI3K/AKT/mTOR signaling pathways, both of which are known to be crucial for spermatogenesis. PI3K signaling is important for spermatogonial entry into meiosis (Blume-Jensen et al., 2000) and inhibition of mTORC1 in mice leads to the accumulation of undifferentiated spermatogonia (Busada, Niedenberger, Velte, Keiper, & Geyer, 2015). Additionally, both the PI3K and MAPK signaling pathways play a role in the activation of the cell cycle in spermatogonia (Suzuki, McCarrey, & Hermann, 2021). Furthermore, the MAPK pathway has also been described to play a role in chromosome condensation and is essential for meiotic progression (Di Agostino, Botti, Di Carlo, Sette, & Geremia, 2004). The MAPK and PI3K pathways are also regulators of the cytoskeleton and GO terms associated to cytoskeleton function and organization were also well represented in the enrichment analysis. Of course, further analysis of downstream genes of *cebp 4* would be needed to confirm the link between this transcription factor and these signaling pathways by, for example, investigating in detail the downstream targets of *cebp 4* and look for members of these pathways.

Surprisingly, the *cebp 4* candidate showed also homology to cyclin B1 interacting protein 1 (CCNB1IP1), an E3 SUMO ligase and shown to be necessary for chiasmata formation during mice spermatogenesis and mutations in this gene leads to meiotic arrest (Strong & Schimenti, 2010; Ward et al., 2007). This protein has also been described to regulate cell cycle progression by interacting with cyclin B and promote its degradation (Toby, Gherraby, Coleman, & Golemis, 2003). I confirmed that the *cebp 4* candidate possessed the same E3 ubiquitin protein ligase domain as CCNB1IP1. This could mean that the sperm development arrest phenotype observed in *cebp 4* RNAi could not be mediated by its DNA binding domain but rather by its potential ubiquitin ligase activity. Disruption of the bzip DNA binding domain by genetic editing could bring a definitive answer to this question but would require stable transgenesis in planaria, a method that has yet to be established.

4.4.2 *thap*

As described in mice (Dejosez et al., 2008), the *thap* candidate seems to affect oocyte development since the oocyte marker showed very little signal in *thap* RNAi animals compared to controls. This contrasts with the RNA-seq results obtained previously and could potentially be due to incomplete knockdown of *thap* during that experiment. Unfortunately, *in situ* hybridization did not show expression of this marker in the testes nor did the testis look underdeveloped in *thap* RNAi compared to the control, ruling out the proposed function of the Tesmin domain found in this gene. The other aspects of the *thap* RNAi phenotype were more unexpected given its expression pattern in WT animals. One type of shell gland marker was significantly downregulated after RNAi and this result was confirmed by *in situ* hybridization (Figure 3.38).

Additionally, *thap* knock-down appeared to disrupt yolk tissue patterning, leading to a more chaotic tissue organization compared to the control animals. However, *thap* knock-down did not seem be essential for yolk development since markers for this tissue were not differentially expressed in this condition (Figure 3.32) and the yolk marker was still visible by *in situ* hybridization (Figure 3.38). Issigonis and colleagues describe that yolk cells originate from germ cells similar to those found in the testes and ovaries, requiring somatic support cells for proper development (Issigonis et al., 2022). Unfortunately, no available literature exists on yolk tissue morphogenesis. Since the *thap* candidate is not specifically expressed in or around yolk cells, it is likely that this effect on yolk tissue morphogenesis is indirect.

A tempting hypothesis would be that, given its neuronal expression, the *thap* candidate regulates the transcription of some unknown extrinsic factors important for shell gland and yolk development. Investigating the differential expression of neuropeptide pro-hormones and neuropeptide receptors in *thap* RNAi compared to *egfp* RNAi could indicate if such hypothesis is true (Collins III et al., 2010; Saberi et al., 2016). Concerning the genes affected by *thap* RNAi, they primarily appear to be associated with amino acid metabolism (Figure 3.35) but I could find no direct link to the observed phenotype. Very generic terms such as translation or peptide metabolic process were among the most significant. However, the genes associated to such terms have to be redundant or only necessary for these functions in very specific conditions since the *thap* knock-down did not show a lethal phenotype.

4.4.3 *tead 1*

The candidates belonging to the TEAD family had been previously characterized in the asexual (A. Y. Lin & Pearson, 2014) but not in the sexual biotype. The *tead 1* candidate (*sd-2*) was ubiquitously expressed similar to *tead 2* (*sd-1*) but at a higher level consistent with the literature (Figure 3.22). They have been implicated in the regulation of the protonephridia in the asexual biotype, and knock-down of the pair of TFs resulted in oedema formation (A. Y. Lin & Pearson, 2014). Here, I obtained a similar phenotype by only knocking down *tead 1* (not shown). However, Lin and colleagues had a different RNAi setup where only 3 feeds of dsRNA liver were done and where the phenotype appeared 15 days after the first feed. In this study, I performed eight feedings over the course of four weeks where the oedema phenotype appeared after the seventh feed. Given that *tead 1* his more abundant than *tead 2* it is possible that they have redundant functions where *tead 2* could complement, but not completely replace, *tead 1*. Therefore, even in the presence of TEAD 2, *tead 1* RNAi would still show the observed phenotype.

Protonephridia and sperm both contain axonemal structures (cilia and flagella). Considering this similarity and the fact that *tead 1* RNAi abrogates protonephridial functions, it is plausible that *tead 1* would be necessary for sperm development by regulating genes necessary for axoneme formation. This is supported by the GO analysis done on *tead 1* RNAi samples, where nearly all enriched GO terms relate to axonemal structures (Figure 3.37). You would therefore expect that cells containing flagella or cilia, like mature sperm and protonephridia would not be present in sexual *S. mediterranea*. This is true but this does not explain the whole phenotype observed in this knock down. Not only is mature sperm but the whole testes were absent from these animals.

An interesting observation is that the germline stem cell marker *nanos*, is upregulated in *ophis* and *tead 1* RNAi. The fact that *nanos* positive cells were still present in *ophis* RNAi had already been described in its initial publication (Saberi et al., 2016) but the increase in expression was, to my knowledge, not commented on further. This same increase in *nanos* expression is observed in *tead 1* RNAi animals. Considering that these animals lack testes, it is plausible that both *ophis* and *tead 1* RNAi conditions arrest the development of male germ line due to the incapacity of the gonadal niche to facilitate the differentiation of these male germline stem cells. A similar reasoning can be put forward for the lack of yolk structures in *tead 1* RNAi (Figure 3.38), although replication of the *in situ* should be performed on slightly larger worms to rule out any possible effect of worm size on yolk tissue maturation. Nevertheless, RNA-seq of *tead 1* RNAi animals point in the direction of yolk defects in this condition with most of the tested markers being significantly downregulated (Figure 3.32).

One clear difference between the two conditions is that oocytes were still present in *tead 1* RNAi, suggesting that *tead 1* does not regulate directly the gonadal niche like *ophis* (Figure 3.38). Another explanation for the absence of testes and yolk tissue in *tead 1* RNAi could be that it is coming from an indirect effect of this TF knock-down. It is known that upon injury or prolonged starvation, planarians do resorb their gonadal structures (P. Newmark, Wang, & Chong, 2008). Therefore, it is conceivable that the osmotic stress induced by the disruption of the planarian excretion system, caused by *tead 1* RNAi, could trigger a similar response. In my opinion, it is likely that the osmotic stress is responsible for the absence of yolk and testes but does not negate the argument that *tead 1* could still be important for flagellar assembly in the testis. To further explore this possibility, the identification of *tead 1* binding sites in the *S. mediterranea* genome could provide valuable insights into whether this transcription factor indeed regulates genes essential for flagellar assembly.

Chapter 5

Conclusions and outlook

The primary objective of my thesis was to study gene regulatory networks governing the development and maintenance of the planarian reproductive system. To achieve this, I developed a robust, planarian compatible Start-seq protocol to identify active regulatory elements by probing for transcription initiation events in a genome-wide fashion. The nuclei isolation part of the Start-seq protocol is also a versatile tool that could be used as a basis for other protocols. Therefore, the developed Start-seq protocol represents an important contribution to the planarian community that will facilitate the study of gene expression regulation in this model organism. It also has the potential to be applicable in other planarians or soft-bodied species with highly unstable nuclear content.

Characterization of the planarian transcription initiation landscape showed that the identified putative enhancers and promoters had similar characteristics as to what is described in other model organisms. I showed that the identified transcription initiation clusters (TICs) were mostly found in non-coding regions of the genome. Moreover, TICs were situated within regions of open chromatin and flanked by nucleosomes enriched for the H3K27Ac mark. Additionally, TICs assigned as putative promoters had a characteristic H3K4me3 distribution pattern that extended towards the gene body whereas enhancer TICs were enriched in H3k4me1.

Study of motifs revealed that both enhancers and promoter TICs showed proper positioning of the TATA-box, Inr and CCAAT-box motifs whereas the DPE motif was found highly enriched at the TSS instead of its characterized +30 nt position. The low frequency of motif occurrence in addition to the recent report of a potential planarian-specific Inr motif (Poulet et al., 2023) suggests that planarians might have different core-promoter motifs. Therefore, additional research through de novo motif identification should be performed. Finally, promoter and enhancer TICs showed pervasive signs of bidirectional transcription initiation. Altogether, these results point towards the fact that the identified TICs are in fact regulatory elements. A definite answer will need the development of transgenic reporter assays to determine the enhancer and promoter potential of the identified REs.

Comparison of the sexual and asexual transcription initiation landscape showed many differentially active REs between the two biotypes. Start-seq and RNA-seq data comparison between the two conditions showed a positive correlation between promoter activity and gene expression confirming that RE activity analysis by Start-seq is a good predictor of gene expression. Moreover, BLAST and GO enrichment analysis showed that upregulated promoters in the sexual biotype were associated to genes involved in sexual reproduction.

Analysis of the conserved motifs between differentially active REs in the *S. mediterranea* biotypes uncovered different families of transcription factors that could potentially play a role in the development and maintenance of the planarian reproductive system. The creation of a structured TF database based on DNA binding domains of known TF families and various selection criteria yielded a total of 20 candidates across 6 different TF families.

Interestingly, nearly all of the selected TF were successfully cloned showed expression in the reproductive system with 50% being specifically expressed in these tissues. A major observation was that none of the TFs were expressed in the abundant vitellaria/yolk cells suggesting that other, maybe non-conserved motif families are responsible for yolk development and maintenance. Indeed, this accessory reproductive organ is specific to a subclass of flatworms, including parasitic species like *S. mansoni*, that are characterized by their yolk-less eggs and rely on specialized yolk cells as a nutrient source for the developing embryos. Understanding how this organ develops and is maintained could not only inform us about the interesting flatworm biology but also advance our knowledge on how to combat diseases by specifically targeting organs required for the reproduction of certain parasitic species.

Functional assessment of the importance of 11 candidates in the development and maintenance of the planarian reproduction revealed three candidates that affected the expression of many reproductive genes. The *tead 1* candidate was previously shown to be important for the maintenance of the excretory system and osmoregulation in asexual planaria. It is therefore not unlikely that the observed phenotype could be partially due to dysregulation of osmoregulatory processes. However, the downregulation of many ciliary genes after RNAi could suggest a role for *tead 1* in spermatogenesis at least. The *thap* candidate affected shell gland function as well as yolk morphogenesis. Interestingly, its expression in the CNS and not in these tissues could suggest an indirect role in accessory reproductive organ function through the expression of soluble molecules like neuropeptides. Finally, the *cebp 4* candidate specifically affected spermatogenesis potentially due to meiotic defects.

Overall, no GRNs involved in the development and maintenance of the planarian reproductive system was created in this study. However, it showed that Start-seq is a valid way to 1) identify active regulatory elements 2) REs differentially active between different conditions and 3) transcription factor binding motifs and their putative targets that play a role in the biological process of interest. The reconstruction of GRNs will require the identification of the regulatory links between the identified TFs and their effector genes. This can be done by specifically identifying TFBSs at the target genes of a specific TF. Actual binding could be detected via chromatin foot printing or performing ChIP-seq experiments against TFs of interest.

References

Adell, T., Salo, E., Boutros, M., & Bartscherer, K. (2009, 03). Smed-Evi/Wntless is required for β-catenin-dependent and-independent processes during planarian regeneration. *Development*, *136*(6), 905-910. doi: 10.1242/dev.033761

Adler, C. E., & Alvarado, A. S. (2015). Types or states? Cellular dynamics and regenerative potential. *Trends in Cell Biology*, *25*(11), 687–696. (Publisher: Elsevier)

Adler, C. E., & Alvarado, A. S. (2018). Systemic RNA Interference in Planarians by Feeding of dsRNA Containing Bacteria. In J. C. Rink (Ed.), *Planarian Regeneration: Methods and Protocols* (pp. 445–454). New York, NY: Springer New York. Retrieved from https://doi.org/10.1007/978-1-4939-7802-1_17 doi: 10.1007/978-1-4939-7802-1_17

Aguilar-Hidalgo, D., Domínguez-Cejudo, M. A., Amore, G., Brockmann, A., Lemos, M. C., Córdoba, A., & Casares, F. (2013). A Hh-driven gene network controls specification, pattern and size of the Drosophila simple eyes. *Development*, *140*(1), 82–92. (Publisher: Company of Biologists)

Akalin, A., Franke, V., Vlahoviček, K., Mason, C. E., & Schübeler, D. (2015). Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics*, *31*(7), 1127–1129. (Publisher: Oxford University Press)

Allfrey, V. G., Faulkner, R., & Mirsky, A. E. (1964, May). ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS*. *Proceedings of the National Academy of Sciences*, *51*(5), 786–794. Retrieved 2023-10-09, from https://doi.org/10.1073/pnas.51.5.786 (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.51.5.786

Allshire, R. C., & Madhani, H. D. (2018, April). Ten principles of heterochromatin formation and function. *Nature Reviews Molecular Cell Biology*, *19*(4), 229–244. Retrieved from https://doi.org/10.1038/nrm.2017.119 doi: 10.1038/nrm.2017.119

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403–410. (Publisher: Elsevier)

Ambrosini, G., Vorontsov, I., Penzar, D., Groux, R., Fornes, O., Nikolaeva, D. D., ... Makeev, V. (2020). Insights gained from a comprehensive all-against-all transcription factor binding motif benchmarking study. *Genome biology*, *21*(1), 1–18. (Publisher: BioMed Central)

Ameyar, M., Wisniewska, M., & Weitzman, J. (2003). A role for AP-1 in apoptosis: the case for and against. *Biochimie*, *85*(8), 747–752. (Publisher: Elsevier)

Andersen, S. U., Algreen-Petersen, R. G., Hoedl, M., Jurkiewicz, A., Cvitanich, C., Braunschweig, U., ... Jensen, E. O. (2007). The conserved cysteine-rich domain of a tesmin/TSO1-like protein binds zinc in vitro and TSO1 is required for both male and female fertility in Arabidopsis thaliana. *Journal of Experimental Botany*, *58*(13), 3657–3670. (Publisher: Oxford University Press)

Andersson, R., Chen, Y., Core, L., Lis, J. T., Sandelin, A., & Jensen, T. H. (2015). Human gene promoters are intrinsically bidirectional. *Molecular cell*, *60*(3), 346–347. (Publisher: Elsevier)

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., ... Suzuki, T. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, *507*(7493), 455–461. (Publisher: Nature Publishing Group UK London)

Andersson, R., & Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. *Nature Reviews Genetics*, *21*(2), 71–87. (Publisher: Nature Publishing Group UK London)

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data.* (Publisher: Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom)

Arnold, C. D., Gerlach, D., Stelzer, C., Boryn, L. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, *339*(6123), 1074–1077. (Publisher: American Association for the Advancement of Science)

Asahina, M., Ishihara, T., Jindra, M., Kohara, Y., Katsura, I., & Hirose, S. (2000). The conserved nuclear receptor Ftz-F1 is required for embryogenesis, moulting and reproduction in Caenorhabditis elegans. *Genes to Cells*, *5*(9), 711–723. (Publisher: Wiley Online Library)

Baguñà, J. (1974). Dramatic mitotic response in planarians after feeding, and a hypothesis for the control mechanism. *Journal of Experimental Zoology*, *190*(1), 117–122. (Publisher: Wiley Online Library)

Baguñà, J., Carranza, S., Pala, M., Ribera, C., Giribet, G., Arnedo, M. A., … Riutort, M. (1999). From morphology and karyology to molecules. New methods for taxonomical identification of asexual populations of freshwater planarians. A tribute to Professor Mario Benazzi. *Italian Journal of Zoology*. (Publisher: Taylor & Francis)

Banerji, J., Olson, L., & Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell*, *33*(3), 729–740. (Publisher: Cell Press)

Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, *27*(2), 299–308. (Publisher: Cell Press)

Bannister, A. J., Zegerman, P., Partridge, J. F., Miska, E. A., Thomas, J. O., Allshire, R. C., & Kouzarides, T. (2001, March). Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature*, *410*(6824), 120–124. Retrieved from https://doi.org/10.1038/35065138 doi: 10.1038/35065138

Baranasic, D., Hörtenhuber, M., Balwierz, P. J., Zehnder, T., Mukarram, A. K., Nepal, C., … Li, N. (2022). Multiomic atlas with functional stratification and developmental dynamics of zebrafish cis-regulatory elements. *Nature genetics*, *54*(7), 1037–1050. (Publisher: Nature Publishing Group US New York)

Barrallo-Gimeno, A., & Nieto, M. A. (2005). The Snail genes as inducers of cell movement and survival: implications in development and cancer. *Development*. (Publisher: Oxford University Press for The Company of Biologists Limited)

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., … Zhao, K. (2007, May). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, *129*(4), 823–837. Retrieved 2023-10-10, from https://doi.org/10.1016/j.cell.2007.05.009 (Publisher: Elsevier) doi: 10.1016/j.cell.2007.05.009

Batut, P. J., Bing, X. Y., Sisco, Z., Raimundo, J., Levo, M., & Levine, M. S. (2022). Genome organization controls transcriptional dynamics during development. *Science*, *375*(6580), 566–570.

Belitsky, B. R., & Sonenshein, A. L. (1999). An enhancer element located downstream of the major glutamate dehydrogenase gene of Bacillus subtilis. *Proceedings of the National Academy of Sciences*, *96*(18), 10290–10295. (ISBN: 0027-8424 Publisher: National Acad Sciences)

Benazzi, M., Baguñà, J., & Ballester, R. (1970). First report on an asexual form of the planarian dugesia lugubris. *Atti della Accademia Nazionale dei Lincei*, *48*, 282–284.

Benazzi, M., Baguñà, J., Ballester, R., Puccinelli, I., & Papa, R. D. (1975). Further contribution to the taxonomy of the dugesia lugubris polychroa group with description of dugesia mediterranea n. sp. (tricladida, paludicola). *Italian Journal of Zoology*, *42*(1), 81–89. (Publisher: Taylor & Francis)

Benazzi, M., & Gremigni, V. (1982). Developmental biology of triclad turbellarians (Planaria). *Developmental biology of freshwater invertebrates. Liss, New York*, 151–211.

Benoist, C., & Chambon, P. (1981). In vivo sequence requirements of the SV40 early promoter region. *Nature*, *290*(5804), 304–310. (Publisher: Nature Publishing Group UK London)

Benoist, C., O'hare, K., Breathnach, R., & Chambon, P. (1980). The ovalbumin gene-sequence of putative control regions. *Nucleic Acids Research*, *8*(1), 127–142. (Publisher: Oxford University Press)

Bergero, R., & Charlesworth, D. (2009). The evolution of restricted recombination in sex chromosomes. *Trends in ecology & evolution*, *24*(2), 94–102. (Publisher: Elsevier)

Bergkessel, M., & Guthrie, C. (2013, January). Chapter Twenty Five - Colony PCR. In J. Lorsch (Ed.), *Methods in Enzymology* (Vol. 529, pp. 299–309). Academic Press. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780124186873000252 doi: 10.1016/B978-0-12-418687-3.00025-2

Bhattacharjee, S., Roche, B., & Martienssen, R. A. (2019). RNA-induced initiation of transcriptional silencing (RITS) complex structure and function. *RNA biology*, *16*(9), 1133–1146. (Publisher: Taylor & Francis)

Blume-Jensen, P., Jiang, G., Hyman, R., Lee, K.-F., O'Gorman, S., & Hunter, T. (2000). Kit/stem cell factor receptor-induced activation of phosphatidylinositol 3'-kinase is essential for male fertility. *Nature genetics*, *24*(2), 157–162. (Publisher: Nature Publishing Group)

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. (Publisher: Oxford University Press)

Bonev, B., & Cavalli, G. (2016). Organization and function of the 3D genome. *Nature Reviews Genetics*, *17*(11), 661–678. (Publisher: Nature Publishing Group UK London)

Bourdareau, S., Tirichine, L., Lombard, B., Loew, D., Scornet, D., Wu, Y., … Cock, J. M. (2021). Histone modifications during the life cycle of the brown alga Ectocarpus. *Genome Biology*, *22*(1), 1–27. (Publisher: BioMed Central)

Buchan, J. R., & Parker, R. (2009, December). Eukaryotic Stress Granules: The Ins and Outs of Translation. *Molecular Cell*, *36*(6), 932–941. Retrieved 2023-10-16, from https://doi.org/10.1016/j.molcel.2009.11.020 (Publisher: Elsevier) doi: 10.1016/j.molcel.2009.11.020

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, *10*(12), 1213–1218. (Publisher: Nature Publishing Group US New York)

Burke, T. W., & Kadonaga, J. T. (1996). Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes & development*, *10*(6), 711–724. (Publisher: Cold Spring Harbor Lab)

Burke, T. W., & Kadonaga, J. T. (1997). The downstream core promoter element, DPE, is conserved fromDrosophila to humans and is recognized by TAFII60 of Drosophila. *Genes & development*, *11*(22), 3020–3031. (Publisher: Cold Spring Harbor Lab)

Busada, J. T., Niedenberger, B. A., Velte, E. K., Keiper, B. D., & Geyer, C. B. (2015). Mammalian target of rapamycin complex 1 (mTORC1) Is required for mouse spermatogonial differentiation in vivo. *Developmental biology*, *407*(1), 90–102. (Publisher: Elsevier)

Candiano, G., Bruschi, M., Musante, L., Santucci, L., Ghiggeri, G. M., Carnemolla, B., … Righetti, P. G. (2004). Blue silver: a very sensitive colloidal Coomassie G-250 staining for proteome analysis. *Electrophoresis*, *25*(9), 1327–1333. (Publisher: Wiley Online Library)

Cano, A., Pérez-Moreno, M. A., Rodrigo, I., Locascio, A., Blanco, M. J., del Barrio, M. G., … Nieto, M. A. (2000). The transcription factor snail controls epithelial–mesenchymal transitions by repressing E-cadherin expression. *Nature cell biology*, *2*(2), 76–83. (Publisher: Nature Publishing Group)

Carelli, F. N., Cerrato, C., Dong, Y., Appert, A., Dernburg, A., & Ahringer, J. (2022). Widespread transposon co-option in the caenorhabditis germline regulatory network. *Science Advances*, *8*(50), eabo4082.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M., Maeda, N., … Wells,

C. (2005). The transcriptional landscape of the mammalian genome. *science*, *309*(5740), 1559–1563. (Publisher: American Association for the Advancement of Science)

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., ... Frith, M. C. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature genetics*, *38*(6), 626–635. (Publisher: Nature Publishing Group US New York)

Carver, E. A., Jiang, R., Lan, Y., Oram, K. F., & Gridley, T. (2001). The mouse snail gene encodes a key regulator of the epithelial-mesenchymal transition. *Molecular and cellular biology*, *21*(23), 8184–8188. (Publisher: Taylor & Francis)

Cayrol, C., Lacroix, C., Mathe, C., Ecochard, V., Ceribelli, M., Loreau, E., ... Aguilar, L. (2007). The THAP–zinc finger protein THAP1 regulates endothelial cell proliferation through modulation of pRB/E2F cell-cycle target genes. *Blood*, *109*(2), 584–594. (Publisher: American Society of Hematology)

Cebrià, F., & Newmark, P. A. (2005). Planarian homologs of netrin and netrin receptor are required for proper regeneration of the central nervous system and the maintenance of nervous system architecture. *Development*. (Publisher: Oxford University Press for The Company of Biologists Limited)

Chalkley, G. E., & Verrijzer, C. P. (1999). DNA binding site selection by RNA polymerase II TAFs: a TAFII250–TAFII150 complex recognizes the initiator. *The EMBO journal*, *18*(17), 4835–4845. (Publisher: John Wiley & Sons, Ltd)

Chen, A. T., & Zon, L. I. (2009). Zebrafish blood stem cells. *Journal of cellular biochemistry*, *108*(1), 35–42. (Publisher: Wiley Online Library)

Chen, R., Wang, J., Gradinaru, I., Vu, H. S., Geboers, S., Naidoo, J., ... Ross, E. M. (2022). A male-derived nonribosomal peptide pheromone controls female schistosome development. *Cell*, *185*(9), 1506–1520. (Publisher: Elsevier)

Chen, R. A.-J., Down, T. A., Stempor, P., Chen, Q. B., Egelhofer, T. A., Hillier, L. W., ... Ahringer, J. (2013). The landscape of RNA polymerase II transcription initiation in C. elegans reveals promoter and enhancer architectures. *Genome research*, *23*(8), 1339–1347. (Publisher: Cold Spring Harbor Lab)

Chiron, S., & Jais, P. H. (2017). Non-radioactive monitoring assay for capping of messenger RNA. *Transl Genet Genom*, *1*, 46–49.

Chomczynski, P., & Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical biochemistry*, *162*(1), 156–159. (Publisher: Elsevier)

Chong, T., Collins, J. J., Brubacher, J. L., Zarkower, D., & Newmark, P. A. (2013, May). A sex-specific transcription factor controls male identity in a simultaneous hermaphrodite. *Nature Communications*, *4*(1), 1814. Retrieved from https://doi.org/10.1038/ncomms2811 doi: 10.1038/ncomms2811

Chong, T., Stary, J. M., Wang, Y., & Newmark, P. A. (2011). Molecular markers to characterize the hermaphroditic reproductive system of the planarian Schmidtea mediterranea. *BMC developmental biology*, *11*, 1–13. (Publisher: Springer)

Christensson, E., & Lewan, L. (1974). The use of spermidine for the isolation of nuclei from mouse liver. Studies of purity and yield during different physiological conditions. *Zeitschrift für Naturforschung C*, *29*(5-6), 267–271. (Publisher: Verlag der Zeitschrift für Naturforschung)

Chumakov, A. M., Silla, A., Williamson, E. A., & Koeffler, H. P. (2007). Modulation of DNA binding properties of CCAAT/enhancer binding protein epsilon by heterodimer formation and interactions with NFkappaB pathway. *Blood*, *109*(10), 4209–4219. (Publisher: American Society of Hematology)

Chuong, E. B., Elde, N. C., & Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics*, *18*(2), 71–86. (Publisher: Nature Publishing Group UK London)

Collins III, J. J., Hou, X., Romanova, E. V., Lambrus, B. G., Miller, C. M., Saberi, A., ... Newmark, P. A. (2010). Genome-wide analyses reveal a role for peptide hormones in planarian germline development. *PLoS biology*, *8*(10), e1000509. (Publisher: Public Library of Science San Francisco, USA)

Collombat, P., Mansouri, A., Hecksher-Sørensen, J., Serup, P., Krull, J., Gradwohl, G., & Gruss, P. (2003). Opposing actions of Arx and Pax4 in endocrine pancreas development. *Genes & development*, *17*(20), 2591–2603. (Publisher: Cold Spring Harbor Lab)

Conaway, R. C., & Conaway, J. W. (2011). Function and regulation of the Mediator complex. *Current opinion in genetics & development*, *21*(2), 225–230. (Publisher: Elsevier)

Conway, J. (2019, May). *UpsetR*. Retrieved from https://github.com/hms-dbmi/UpSetR

Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., ... Wu, B. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nature methods*, *14*(10), 959–962. (Publisher: Nature Publishing Group UK London)

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., & Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature genetics*, *46*(12), 1311–1320. (Publisher: Nature Publishing Group US New York)

Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, *322*(5909), 1845–1848. (Publisher: American Association for the Advancement of Science)

Cowles, M. W., Omuro, K. C., Stanley, B. N., Quintanilla, C. G., & Zayas, R. M. (2014). COE loss-of-function analysis reveals a genetic program underlying maintenance and regeneration of the nervous system in planarians. *PLoS Genetics*, *10*(10), e1004746. (Publisher: Public Library of Science San Francisco, USA)

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., . . . Sharp, P. A. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, *107*(50), 21931–21936. (Publisher: National Acad Sciences)

Crozatier, M., Valle, D., Dubois, L., Ibnsouda, S., & Vincent, A. (1996). Collier, a novel regulator of Drosophila head development, is expressed in a single mitotic domain. *Current Biology*, *6*(6), 707–718. (Publisher: Elsevier)

Czerny, T., Halder, G., Kloter, U., Souabni, A., Gehring, W. J., & Busslinger, M. (1999). twin of eyeless, a second Pax-6 gene of Drosophila, acts upstream of eyeless in the control of eye development. *Molecular cell*, *3*(3), 297–307. (Publisher: Elsevier)

Daburon, V., Mella, S., Plouhinec, J.-L., Mazan, S., Crozatier, M., & Vincent, A. (2008). The metazoan history of the COE transcription factors. Selection of a variant HLH motif by mandatory inclusion of a duplicated exon in vertebrates. *BMC evolutionary biology*, *8*(1), 1–13. (Publisher: BioMed Central)

Dann, G. P., Liszczak, G. P., Bagert, J. D., Müller, M. M., Nguyen, U. T. T., Wojcik, F., . . . Muir, T. W. (2017, August). ISWI chromatin remodellers sense nucleosome modifications to determine substrate preference. *Nature*, *548*(7669), 607–611. Retrieved from https://doi.org/10.1038/nature23671 doi: 10.1038/nature23671

Darlington, G. J., Ross, S. E., & MacDougald, O. A. (1998). The role of C/EBP genes in adipocyte differentiation. *Journal of Biological Chemistry*, *273*(46), 30057–30060.

(Publisher: ASBMB)

Dattani, A., Kao, D., Mihaylova, Y., Abnave, P., Hughes, S., Lai, A., … Aboobaker, A. A. (2018). Epigenetic analyses of planarian stem cells demonstrate conservation of bivalent histone modifications in animal stem cells. *Genome research*, *28*(10), 1543–1554. (Publisher: Cold Spring Harbor Lab)

Davidson, E. H. (2001). *Genomic regulatory systems: in development and evolution.* Elsevier.

Davidson, E. H. (2010a). Emerging properties of animal gene regulatory networks. *Nature*, *468*(7326), 911–920. (Publisher: Nature Publishing Group UK London)

Davidson, E. H. (2010b). *The regulatory genome: gene regulatory networks in development and evolution.* Elsevier.

Davidson, E. H., & Levine, M. S. (2008). Properties of developmental gene regulatory networks. *Proceedings of the National Academy of Sciences*, *105*(51), 20063–20066. (Publisher: National Acad Sciences)

Davidson, E. H., & Peter, I. S. (2015). *Genomic control process.* Elsevier.

Davies, E. L., Lei, K., Seidel, C. W., Kroesen, A. E., McKinney, S. A., Guo, L., … Alvarado, A. S. (2017). Embryonic origin of adult stem cells required for tissue homeostasis and regeneration. *Elife*, *6*, e21052. (Publisher: eLife Sciences Publications, Ltd)

Dejosez, M., Krumenacker, J. S., Zitur, L. J., Passeri, M., Chu, L.-F., Songyang, Z., … Zwaka, T. P. (2008). Ronin is essential for embryogenesis and the pluripotency of mouse embryonic stem cells. *Cell*, *133*(7), 1162–1174. (Publisher: Elsevier)

De Renzis, S., Yu, J., Zinzen, R., & Wieschaus, E. (2006). Dorsal-ventral pattern of Delta trafficking is established by a Snail-Tom-Neuralized pathway. *Developmental cell*, *10*(2), 257–264. (Publisher: Elsevier)

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B. K., … Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS biology*, *8*(5), e1000384. (Publisher: Public Library of Science San Francisco, USA)

Di Agostino, S., Botti, F., Di Carlo, A., Sette, C., & Geremia, R. (2004). Meiotic progression of isolated mouse spermatocytes under simulated microgravity. *Reproduction*, *128*(1), 25–32. (Publisher: Society for Reproduction and Fertility)

Di Cerbo, V., Mohn, F., Ryan, D. P., Montellier, E., Kacem, S., Tropberger, P., … Schneider, R. (2014, March). Acetylation of histone H3 at lysine 64 regulates nucleosome dynamics and facilitates transcription. *eLife*, *3*, e01632. Retrieved from https://doi.org/10.7554/eLife.01632 (Publisher: eLife Sciences Publications, Ltd) doi: 10.7554/eLife.01632

Di Luccia, A., Picariello, G., Iacomino, G., Formisano, A., Paduano, L., & D'Agostino, L. (2009). The in vitro nuclear aggregates of polyamines. *The FEBS Journal*, *276*(8), 2324–2335. (Publisher: Wiley Online Library)

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. (Publisher: Oxford University Press)

Don, R., Cox, P., Wainwright, B., Baker, K., & Mattick, J. (1991). 'Touchdown'PCR to circumvent spurious priming during gene amplification. *Nucleic acids research*, *19*(14), 4008. (Publisher: Oxford University Press)

Dong, J., Feldmann, G., Huang, J., Wu, S., Zhang, N., Comerford, S. A., ... Pan, D. (2007). Elucidation of a universal size-control mechanism in Drosophila and mammals. *Cell*, *130*(6), 1120–1133. (Publisher: Elsevier)

Driever, W., & Nüsslein-Volhard, C. (1988). A gradient of bicoid protein in Drosophila embryos. *Cell*, *54*(1), 83–93. (Publisher: Cell Press)

Duboc, V., & Logan, M. P. (2011). Regulation of limb bud initiation and limb-type morphology. *Developmental Dynamics*, *240*(5), 1017–1027. (Publisher: Wiley Online Library)

Duncan, E. M., Chitsazan, A. D., Seidel, C. W., & Alvarado, A. S. (2015). Set1 and MLL1/2 target distinct sets of functionally different genomic loci in vivo. *Cell reports*, *13*(12), 2741–2755. (Publisher: Elsevier)

Duttke, S. H., Chang, M. W., Heinz, S., & Benner, C. (2019). Identification and dynamic quantification of regulatory elements using total RNA. *Genome Research*, *29*(11), 1836–1846. (Publisher: Cold Spring Harbor Lab)

Duttke, S. H., Lacadie, S. A., Ibrahim, M. M., Glass, C. K., Corcoran, D. L., Benner, C., ... Ohler, U. (2015). Human promoters are intrinsically directional. *Molecular cell*, *57*(4), 674–684. (Publisher: Elsevier)

Ekaterina Morgunova, Yimeng Yin, Fangjie Zhu, Tianyi Xiao, Ilya Sokolov, Alexander Popov, ... Jussi Taipale (2023, January). Interfacial water confers transcription factors with dinucleotide specificity. *bioRxiv*, 2023.10.03.560647. Retrieved from http://biorxiv.org/content/early/2023/10/03/2023.10.03.560647.abstract doi: 10.1101/2023.10.03.560647

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57. (Publisher: NIH Public Access)

Evans, K. J., Huang, N., Stempor, P., Chesney, M. A., Down, T. A., & Ahringer, J. (2016, November). Stable Caenorhabditis elegans chromatin domains separate broadly expressed and developmentally regulated genes. *Proceedings of the National Academy of Sciences*, *113*(45), E7020–E7029. Retrieved 2023-10-13, from https://doi.org/10.1073/pnas.1608162113 (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.1608162113

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. (Publisher: Oxford University Press)

Extavour, C. G. (2007). Evolution of the bilaterian germ line: lineage origin and modulation of specification mechanisms. *Integrative and Comparative Biology*, *47*(5), 770–785. (Publisher: Oxford University Press)

Fincher, C. T., Wurtzel, O., de Hoog, T., Kravarik, K. M., & Reddien, P. W. (2018). Cell type transcriptome atlas for the planarian schmidtea mediterranea. *Science*, *360*(6391), eaaq1736.

FitzGerald, P. C., Sturgill, D., Shyakhtenko, A., Oliver, B., & Vinson, C. (2006). Comparative genomics of Drosophila and human core promoters. *Genome biology*, *7*, 1–22. (Publisher: Springer)

Flores, N. M., Oviedo, N. J., & Sage, J. (2016). Essential role for the planarian intestinal GATA transcription factor in stem cells and regeneration. *Developmental biology*, *418*(1), 179–188. (Publisher: Elsevier)

Ford, E. (2012, February). *NGS PCR Cycle Quantitation.* Retrieved from https://ethanomics.files.wordpress.com/2012/02/cycle_quantitation4.pdf

Forsthoefel, D. J., Park, A. E., & Newmark, P. A. (2011). Stem cell-based growth, regeneration, and remodeling of the planarian intestine. *Developmental biology*, *356*(2), 445–459. (Publisher: Elsevier)

Frasch, M., Warrior, R., Tugwood, J., & Levine, M. (1988). Molecular analysis of even-skipped mutants in Drosophila development. *Genes & development*, *2*(12b), 1824–1838. (Publisher: Cold Spring Harbor Lab)

Fuda, N. J., Guertin, M. J., Sharma, S., Danko, C. G., Martins, A. L., Siepel, A., & Lis, J. T. (2015). GAGA factor maintains nucleosome-free regions and has a role in RNA polymerase II recruitment to promoters. *PLoS genetics*, *11*(3), e1005108.

(Publisher: Public Library of Science San Francisco, CA USA)

Furlong, E. E., & Levine, M. (2018). Developmental enhancers and chromosome topology. *Science*, *361*(6409), 1341–1345. (Publisher: American Association for the Advancement of Science)

Ganguly, A., Rock, M. J., & Prockop, D. J. (1993). Conformation-sensitive gel electrophoresis for rapid detection of single-base differences in double-stranded pcr products and dna fragments: evidence for solvent-induced bends in dna heteroduplexes. *Proceedings of the National Academy of Sciences*, *90*(21), 10325–10329.

Gardiner-Garden, M., & Frommer, M. (1987). CpG islands in vertebrate genomes. *Journal of molecular biology*, *196*(2), 261–282. (Publisher: Elsevier)

Gehrke, A. R., Neverett, E., Luo, Y.-J., Brandt, A., Ricci, L., Hulett, R. E., ... Reddien, P. W. (2019). Acoel genome reveals the regulatory landscape of whole-body regeneration. *Science*, *363*(6432), eaau6173. (Publisher: American Association for the Advancement of Science)

Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., ... Ikegami, K. (2010). Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. *Science*, *330*(6012), 1775–1787. (Publisher: American Association for the Advancement of Science)

Ghosh, A., & Lima, C. D. (2010). Enzymology of RNA cap synthesis. *Wiley Interdisciplinary Reviews: RNA*, *1*(1), 152–172. (Publisher: Wiley Online Library)

Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A., & Smith, H. O. (2009, May). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods*, *6*(5), 343–345. Retrieved from https://doi.org/10.1038/nmeth.1318 doi: 10.1038/nmeth.1318

Gitter, A., Siegfried, Z., Klutstein, M., Fornes, O., Oliva, B., Simon, I., & Bar-Joseph, Z. (2009). Backup in gene regulatory networks explains differences between binding and knockout results. *Molecular systems biology*, *5*(1), 276. (Publisher: John Wiley & Sons, Ltd Chichester, UK)

Goldman, J. A., & Poss, K. D. (2020). Gene regulatory programmes of tissue regeneration. *Nature Reviews Genetics*, *21*(9), 511–525. (Publisher: Nature Publishing Group UK London)

Gotta, M., & Ahringer, J. (2001). Axis determination in C. elegans: initiating and transducingpolarity. *Current opinion in genetics & development*, *11*(4), 367–373. (Publisher: Elsevier)

Gołyszny, M., Obuchowicz, E., & Zieliński, M. (2022). Neuropeptides as regulators of the hypothalamus-pituitary-gonadal (HPG) axis activity and their putative roles in stress-induced fertility disorders. *Neuropeptides*, *91*, 102216. (Publisher: Elsevier)

Graf, T., & Enver, T. (2009). Forcing cells to change lineages. *Nature*, *462*(7273), 587–594. (Publisher: Nature Publishing Group UK London)

Grandi, F. C., Modi, H., Kampman, L., & Corces, M. R. (2022). Chromatin accessibility profiling by ATAC-seq. *Nature protocols*, *17*(6), 1518–1552. (Publisher: Nature Publishing Group UK London)

Greenspan, L. J., De Cuevas, M., & Matunis, E. (2015). Genetics of gonadal stem cell renewal. *Annual review of cell and developmental biology*, *31*, 291–315. (Publisher: Annual Reviews)

Gremigni, V., & Domenici, L. (1974). Electron microscopical and cytochemical study of vitelline cells in the fresh water triclad Dugesia lugubris sl: I. Origin and morphogenesis of cocoon-shell globules. *Cell and tissue research*, *150*(2), 261–270. (Publisher: Springer)

Grohme, M. A., Schloissnig, S., Rozanski, A., Pippel, M., Young, G. R., Winkler, S., ... Powell, S. (2018). The genome of Schmidtea mediterranea and the evolution of core cellular mechanisms. *Nature*, *554*(7690), 56–61. (Publisher: Nature Publishing Group UK London)

Gruss, P., Dhar, R., & Khoury, G. (1981). Simian virus 40 tandem repeated sequences as an element of the early promoter. *Proceedings of the National Academy of Sciences*, *78*(2), 943–947. (Publisher: National Acad Sciences)

Grøntved, L., John, S., Baek, S., Liu, Y., Buckley, J. R., Vinson, C., ... Hager, G. L. (2013). C/EBP maintains chromatin accessibility in liver and facilitates glucocorticoid receptor recruitment to steroid response elements. *The EMBO journal*, *32*(11), 1568–1583. (Publisher: John Wiley & Sons, Ltd Chichester, UK)

Gu, W., Lee, H.-C., Chaves, D., Youngman, E. M., Pazour, G. J., Conte, D., & Mello, C. C. (2012). CapSeq and CIP-TAP identify Pol II start sites and reveal capped small RNAs as C. elegans piRNA precursors. *Cell*, *151*(7), 1488–1500. (Publisher: Elsevier)

Gualdi, R., Bossard, P., Zheng, M., Hamada, Y., Coleman, J. R., & Zaret, K. S. (1996). Hepatic specification of the gut endoderm in vitro: cell signaling and transcriptional control. *Genes & development*, *10*(13), 1670–1682. (Publisher: Cold Spring Harbor Lab)

Guo, L., Bloom, J. S., Dols-Serrate, D., Boocock, J., Ben-David, E., Schubert, O. T., … Chui, C. (2022). Island-specific evolution of a sex-primed autosome in a sexual planarian. *Nature*, *606*(7913), 329–334. (Publisher: Nature Publishing Group UK London)

Guo, L., Zhang, S., Rubinstein, B., Ross, E., & Alvarado, A. S. (2016). Widespread maintenance of genome heterozygosity in Schmidtea mediterranea. *Nature ecology & evolution*, *1*(1), 0019. (Publisher: Nature Publishing Group UK London)

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., & Noble, W. S. (2007). Quantifying similarity between motifs. *Genome biology*, *8*(2), 1–9. (Publisher: BioMed Central)

Gurley, K. A., Rink, J. C., & Alvarado, A. S. (2008). β-catenin defines head versus tail identity during planarian regeneration and homeostasis. *Science*, *319*(5861), 323–327. (Publisher: American Association for the Advancement of Science)

Haas, B. (2019). *TransDecoder (Find Coding Regions Within Transcripts).* (https://github.com/TransDecoder/TransDecoder/wiki)

Habener, J. F., Kemp, D. M., & Thomas, M. K. (2005). Minireview: transcriptional regulation in pancreatic development. *Endocrinology*, *146*(3), 1025–1034. (Publisher: Endocrine Society)

Haberle, V., & Lenhard, B. (2016). Promoter architectures and developmental gene regulation. In *Seminars in cell & developmental biology* (Vol. 57, pp. 11–23).

Haberle, V., & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nature reviews Molecular cell biology*, *19*(10), 621–637. (Publisher: Nature Publishing Group UK London)

Hagman, J., Belanger, C., Travis, A., Turck, C. W., & Grosschedl, R. (1993). Cloning and functional characterization of early B-cell factor, a regulator of lymphocyte-specific gene expression. *Genes & development*, *7*(5), 760–773. (Publisher: Cold Spring Harbor Lab)

Halder, G., & Johnson, R. L. (2011). Hippo signaling: growth control and beyond. *Development*, *138*(1), 9–22. (Publisher: Company of Biologists)

Hall, R. N., Weill, U., Drees, L., Leal-Ortiz, S., Li, H., Khariton, M., … Quake, S. R. (2022). Heterologous reporter expression in the planarian Schmidtea mediterranea through somatic mRNA transfection. *Cell Reports Methods*, *2*(10), 100298. (Publisher: Elsevier)

Handberg-Thorsager, M., & Saló, E. (2007). The planarian nanos-like gene Smednos is expressed in germline and eye precursor cells during development and regeneration.

Development genes and evolution, *217*, 403–411. (Publisher: Springer)

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., … Yoo, J. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, *431*(7004), 99–104. (Publisher: Nature Publishing Group UK London)

Harding, K., Hoey, T., Warrior, R., & Levine, M. (1989). Autoregulatory and gap gene response elements of the even-skipped promoter of Drosophila. *The EMBO journal*, *8*(4), 1205–1212.

Harrath, A. H., Charni, M., Sluys, R., Zghal, F., & Tekaya, S. (2004). Ecology and distribution of the freshwater planarian Schmidtea mediterranea in Tunisia. *Italian Journal of Zoology*, *71*(3), 233–236. (Publisher: Taylor & Francis)

Hartwig, A. (2001). Role of magnesium in genomic stability. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *475*(1-2), 113–121. (Publisher: Elsevier)

Hashimoto, S.-i., Suzuki, Y., Kasai, Y., Morohoshi, K., Yamada, T., Sese, J., … Matsushima, K. (2004). 5'-end SAGE for the analysis of transcriptional start sites. *Nature biotechnology*, *22*(9), 1146–1149. (Publisher: Nature Publishing Group UK London)

He, H. H., Meyer, C. A., Shin, H., Bailey, S. T., Wei, G., Wang, Q., … Lupien, M. (2010). Nucleosome dynamics define transcriptional enhancers. *Nature genetics*, *42*(4), 343–347. (Publisher: Nature Publishing Group US New York)

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., … Ching, K. A. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, *39*(3), 311–318. (Publisher: Nature Publishing Group US New York)

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., … Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell*, *38*(4), 576–589. (Publisher: Elsevier)

Heinz, S., Romanoski, C. E., Benner, C., & Glass, C. K. (2015). The selection and function of cell type-specific enhancers. *Nature reviews Molecular cell biology*, *16*(3), 144–154. (Publisher: Nature Publishing Group UK London)

Hendrix, D. A., Hong, J.-W., Zeitlinger, J., Rokhsar, D. S., & Levine, M. S. (2008). Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. *Proceedings of the National Academy of Sciences*, *105*(22), 7762–7767. (Publisher:

National Acad Sciences)

Henriques, T., Gilchrist, D. A., Nechaev, S., Bern, M., Muse, G. W., Burkholder, A., ... Adelman, K. (2013). Stable pausing by rna polymerase ii provides an opportunity to target and integrate regulatory signals. *Molecular cell*, *52*(4), 517–528.

Henriques, T., Scruggs, B. S., Inouye, M. O., Muse, G. W., Williams, L. H., Burkholder, A. B., ... Adelman, K. (2018). Widespread transcriptional pausing and elongation control at enhancers. *Genes & development*, *32*(1), 26–41. (Publisher: Cold Spring Harbor Lab)

Hess, J., Angel, P., & Schorpp-Kistner, M. (2004). AP-1 subunits: quarrel and harmony among siblings. *Journal of cell science*, *117*(25), 5965–5973. (Publisher: Company of Biologists)

Hirai, H., Zhang, P., Dayaram, T., Hetherington, C. J., Mizuno, S.-i., Imanishi, J., ... Tenen, D. G. (2006). C/EBPβ is required for'emergency'granulopoiesis. *Nature immunology*, *7*(7), 732–739. (Publisher: Nature Publishing Group US New York)

Ho, J. W., Jung, Y. L., Liu, T., Alver, B. H., Lee, S., Ikegami, K., ... Appert, A. (2014). Comparative analysis of metazoan chromatin organization. *Nature*, *512*(7515), 449–452. (Publisher: Nature Publishing Group UK London)

Hobson, D. J., Wei, W., Steinmetz, L. M., & Svejstrup, J. Q. (2012). RNA polymerase II collision interrupts convergent transcription. *Molecular cell*, *48*(3), 365–374. (Publisher: Elsevier)

Howard, K., & Ingham, P. (1986). Regulatory interactions between the segmentation genes fushi tarazu, hairy, and engrailed in the Drosophila blastoderm. *Cell*, *44*(6), 949–957. (Publisher: Cell Press)

Hu, Y.-C., Okumura, L. M., & Page, D. C. (2013). Gata4 is required for formation of the genital ridge in mice. *PLoS genetics*, *9*(7), e1003629. (Publisher: Public Library of Science San Francisco, USA)

Hu, Z., Killion, P. J., & Iyer, V. R. (2007). Genetic reconstruction of a functional transcriptional regulatory network. *Nature genetics*, *39*(5), 683–687. (Publisher: Nature Publishing Group US New York)

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., ... Jensen, L. J. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research*, *47*(D1), D309–D314. (Publisher: Oxford University Press)

Huminiecki, L., & Horbanczuk, J. (2017). Can we predict gene expression by understanding proximal promoter architecture? *Trends in Biotechnology*, *35*(6), 530–546. (Publisher: Elsevier)

Hyun, K., Jeon, J., Park, K., & Kim, J. (2017, April). Writing, erasing and reading histone lysine methylations. *Experimental & Molecular Medicine*, *49*(4), e324–e324. Retrieved from https://doi.org/10.1038/emm.2017.11 doi: 10.1038/emm.2017.11

Iglesias, M., Gomez-Skarmeta, J. L., Saló, E., & Adell, T. (2008). Silencing of Smed-βcatenin1 generates radial-like hypercephalized planarians. *Development*. (Publisher: Oxford University Press for The Company of Biologists Limited)

Ignatiadis, N., Klaus, B., Zaugg, J. B., & Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, *13*(7), 577–580. (Publisher: Nature Publishing Group US New York)

Inoue, H., Nojima, H., & Okayama, H. (1990). High efficiency transformation of Escherichia coli with plasmids. *Gene*, *96*(1), 23–28. (Publisher: Elsevier)

Issigonis, M., & Newmark, P. A. (2019). From worm to germ: germ cell development and regeneration in planarians. *Current Topics in Developmental Biology*, *135*, 127–153. (Publisher: Elsevier)

Issigonis, M., Redkar, A. B., Rozario, T., Khan, U. W., Mejia-Sanchez, R., Lapan, S. W., ... Newmark, P. A. (2022). A Krüppel-like factor is required for development and regeneration of germline and yolk cells from somatic stem cells in planarians. *PLoS Biology*, *20*(7), e3001472. (Publisher: Public Library of Science San Francisco, CA USA)

Ito, E., Toki, T., Ishihara, H., Ohtani, H., Gu, L., Yokoyama, M., ... Yamamoto, M. (1993). Erythroid transcription factor GATA-1 is abundantly transcribed in mouse testis. *Nature*, *362*(6419), 466–468. (Publisher: Nature Publishing Group UK London)

Ivankovic, M., Brand, J. N., Pandolfini, L., Brown, T., Pippel, M., Rozanski, A., ... others (2023). A comparative analysis of planarian genomes reveals regulatory conservation in the face of rapid structural divergence. *bioRxiv*, 2023–12.

Ivankovic, M., Haneckova, R., Thommen, A., Grohme, M. A., Vila-Farré, M., Werner, S., & Rink, J. C. (2019). Model systems for regeneration: planarians. *Development*, *146*(17), dev167684.

Iyer, H., Collins III, J. J., & Newmark, P. A. (2016). NF-YB regulates spermatogonial stem cell self-renewal and proliferation in the planarian Schmidtea mediterranea. *PLoS* *genetics*, *12*(6), e1006109. (Publisher: Public Library of Science San Francisco, CA USA)

Jaeger, J., Blagov, M., Kosman, D., Kozlov, K. N., Manu, Myasnikova, E., … Sharp, D. H. (2004). Dynamical analysis of regulatory interactions in the gap gene system of Drosophila melanogaster. *Genetics*, *167*(4), 1721–1737. (Publisher: Oxford University Press)

Jemc, J. C., Milutinovich, A. B., Weyers, J. J., Takeda, Y., & Van Doren, M. (2012). raw Functions through JNK signaling and cadherin-based adhesion to regulate Drosophila gonad morphogenesis. *Developmental biology*, *367*(2), 114–125. (Publisher: Elsevier)

Jensen, J. (2004). Gene regulatory factors in pancreatic development. *Developmental dynamics: an official publication of the American Association of Anatomists*, *229*(1), 176–200. (Publisher: Wiley Online Library)

Jenuwein, T., & Allis, C. D. (2001, August). Translating the Histone Code. *Science*, *293*(5532), 1074–1080. Retrieved 2023-10-13, from https://doi.org/10.1126/science.1063127 (Publisher: American Association for the Advancement of Science) doi: 10.1126/science.1063127

Jiang, J., Benson, E., Bausek, N., Doggett, K., & White-Cooper, H. (2007). Tombola, a tesmin/TSO1-family protein, regulates transcriptional activation in the Drosophila male germline and physically interacts with always early. *Development*. (Publisher: Oxford University Press for The Company of Biologists Limited)

Jimenez, M. A., Åkerblad, P., Sigvardsson, M., & Rosen, E. D. (2007). Critical role for Ebf1 and Ebf2 in the adipogenic transcriptional cascade. *Molecular and cellular biology*, *27*(2), 743–757. (Publisher: Taylor & Francis)

Jin, Q., Yu, L., Wang, L., Zhang, Z., Kasper, L. H., Lee, J., … Ge, K. (2011). Distinct roles of GCN5/PCAF-mediated H3K9ac and CBP/p300-mediated H3K18/27ac in nuclear receptor transactivation. *The EMBO journal*, *30*(2), 249–262. (Publisher: John Wiley & Sons, Ltd Chichester, UK)

Jin, Y., Eser, U., Struhl, K., & Churchman, L. S. (2017). The ground state and evolution of promoter region directionality. *Cell*, *170*(5), 889–898. (Publisher: Elsevier)

Johnston, H., Warner, J. F., Amiel, A. R., Nedoncelle, K., Carvalho, J. E., & Röttinger, E. (2019). Whole body regeneration deploys a rewired embryonic gene regulatory network logic. *bioRxiv*, 658930. (Publisher: Cold Spring Harbor Laboratory)

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., … Nuka, G. (2014).

InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. (Publisher: Oxford University Press)

Jänes, J., Dong, Y., Schoof, M., Serizay, J., Appert, A., Cerrato, C., ... Huang, N. (2018). Chromatin accessibility dynamics across C. elegans development and ageing. *Elife*, *7*, e37344. (Publisher: eLife Sciences Publications, Ltd)

Kang, H., & Alvarado, A. S. (2009). Flow cytometry methods for the study of cell-cycle parameters of planarian stem cells. *Developmental dynamics: an official publication of the American Association of Anatomists*, *238*(5), 1111–1117. (Publisher: Wiley Online Library)

Karin, M., Liu, Z.-g., & Zandi, E. (1997, April). AP-1 function and regulation. *Current Opinion in Cell Biology*, *9*(2), 240–246. Retrieved from https://www.sciencedirect.com/science/article/pii/S0955067497800683 doi: 10.1016/S0955-0674(97)80068-3

Karlebach, G., & Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nature reviews Molecular cell biology*, *9*(10), 770–780. (Publisher: Nature Publishing Group UK London)

Karpinska, M. A., & Oudelaar, A. M. (2023). The role of loop extrusion in enhancer-mediated gene activation. *Current Opinion in Genetics & Development*, *79*, 102022. (Publisher: Elsevier)

Kasai, Y., Nambu, J. R., Lieberman, P. M., & Crews, S. T. (1992, April). Dorsal-ventral patterning in Drosophila: DNA binding of snail protein to the single-minded gene. *Proceedings of the National Academy of Sciences*, *89*(8), 3414–3418. Retrieved 2023-10-21, from https://doi.org/10.1073/pnas.89.8.3414 (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.89.8.3414

Kassambara, A. (2020). ggpubr:"ggplot2" based publication ready plots. *R package version 0.4. 0*, *438*.

Keppler, B. R., & Archer, T. K. (2008, October). Chromatin-modifying enzymes as therapeutic targets – Part 1. *Expert Opinion on Therapeutic Targets*, *12*(10), 1301–1312. Retrieved from https://doi.org/10.1517/14728222.12.10.1301 (Publisher: Taylor & Francis) doi: 10.1517/14728222.12.10.1301

Ketola, I., Pentikäinen, V., Vaskivuo, T., Ilvesmäki, V., Herva, R., Dunkel, L., ... Heikinheimo, M. (2000). Expression of transcription factor GATA-4 during human testicular development and disease. *The Journal of Clinical Endocrinology & Metabolism*, *85*(10), 3925–3931. (Publisher: Oxford University Press)

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., Van Der Lee, R., … Tan, G. (2018). JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic acids research*, *46*(D1), D260–D266. (Publisher: Oxford University Press)

Khan, A. U., Mei, Y.-H., & Wilson, T. (1992). A proposed function for spermine and spermidine: protection of replicating DNA against damage by singlet oxygen. *Proceedings of the National Academy of Sciences*, *89*(23), 11426–11427. (Publisher: National Acad Sciences)

Khan, U. W., & Newmark, P. A. (2022). Somatic regulation of female germ cell regeneration and development in planarians. *Cell reports*, *38*(11). (Publisher: Elsevier)

Kiger, A. A., White-Cooper, H., & Fuller, M. T. (2000). Somatic support cells restrict germline stem cell self-renewal and promote differentiation. *Nature*, *407*(6805), 750–754.

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., … Kuersten, S. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, *465*(7295), 182–187. (Publisher: Nature Publishing Group UK London)

King, R. S., & Newmark, P. A. (2013). In situ hybridization protocol for enhanced detection of gene expression in the planarian Schmidtea mediterranea. *BMC developmental biology*, *13*, 1–16. (Publisher: Springer)

King, R. S., & Newmark, P. A. (2018). Whole-mount in situ hybridization of planarians. *Planarian Regeneration: Methods and Protocols*, 379–392. (Publisher: Springer)

Klemm, S. L., Shipony, Z., & Greenleaf, W. J. (2019, April). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, *20*(4), 207–220. Retrieved from https://doi.org/10.1038/s41576-018-0089-8 doi: 10.1038/s41576-018-0089-8

Kobayashi, M., Goldstein, R. E., Fujioka, M., Paroush, Z., & Jaynes, J. B. (2001). Groucho augments the repression of multiple Even skipped target genes in establishing parasegment boundaries. *Development*, *128*(10), 1805–1815. (Publisher: The Company of Biologists Ltd)

Koch, F., Fenouil, R., Gut, M., Cauchy, P., Albert, T. K., Zacarias-Cabeza, J., … Hintermair, C. (2011). Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nature structural & molecular biology*, *18*(8), 956–963. (Publisher: Nature Publishing Group US New York)

Kolde, R. (2018, December). *Pheatmap: A function to draw clustered heatmaps.* Retrieved from https://github.com/raivokolde/pheatmap

Kowalczyk, M. S., Hughes, J. R., Garrick, D., Lynch, M. D., Sharpe, J. A., Sloane-Stanley, J. A., … Higgs, D. R. (2012, February). Intragenic enhancers act as alternative promoters. *Molecular cell*, *45*(4), 447–458. (Place: United States) doi: 10.1016/j.molcel.2011.12.021

Krogan, N. J., Dover, J., Wood, A., Schneider, J., Heidt, J., Boateng, M. A., … Johnston, M. (2003). The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: linking transcriptional elongation to histone methylation. *Molecular cell*, *11*(3), 721–729. (Publisher: Elsevier)

Kruesi, W. S., Core, L. J., Waters, C. T., Lis, J. T., & Meyer, B. J. (2013). Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation. *elife*, *2*, e00808. (Publisher: eLife Sciences Publications, Ltd)

Kuo, C. T., Morrisey, E. E., Anandappa, R., Sigrist, K., Lu, M. M., Parmacek, M. S., … Leiden, J. M. (1997). GATA4 transcription factor is required for ventral morphogenesis and heart tube formation. *Genes & development*, *11*(8), 1048–1060. (Publisher: Cold Spring Harbor Lab)

Kutach, A. K., & Kadonaga, J. T. (2000). The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. *Molecular and cellular biology*, *20*(13), 4754–4764. (Publisher: Taylor & Francis)

Kvon, E. Z., Kamneva, O. K., Melo, U. S., Barozzi, I., Osterwalder, M., Mannion, B. J., … Lee, E. A. (2016). Progressive loss of function in a limb enhancer during snake evolution. *Cell*, *167*(3), 633–642. (Publisher: Elsevier)

Kwak, H., Fuda, N. J., Core, L. J., & Lis, J. T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*, *339*(6122), 950–953. (Publisher: American Association for the Advancement of Science)

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *nature*, *227*(5259), 680–685. (Publisher: Nature Publishing Group UK London)

Lai, W. K., & Pugh, B. F. (2017). Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nature reviews Molecular cell biology*, *18*(9), 548–562. (Publisher: Nature Publishing Group UK London)

Lam, M. T., Cho, H., Lesch, H. P., Gosselin, D., Heinz, S., Tanaka-Oishi, Y., … Kosaka, M. (2013). Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature*, *498*(7455), 511–515. (Publisher: Nature Publishing Group UK London)

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., … Weirauch, M. T. (2018). The human transcription factors. *Cell*, *172*(4), 650–665. (Publisher: Elsevier)

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature methods*, *9*(4), 357–359. (Publisher: Nature Publishing Group US New York)

Larke, M. S., Schwessinger, R., Nojima, T., Telenius, J., Beagrie, R. A., Downes, D. J., … Bender, M. (2021). Enhancers predominantly regulate gene expression during differentiation via transcription initiation. *Molecular cell*, *81*(5), 983–997. (Publisher: Elsevier)

Lassner, D. (1995). Synthesis of cDNA. In T. Köhler, D. Laßner, H. Remke, A.-K. Rost, B. Thamm, & B. Pustowoit (Eds.), *Quantitation of mRNA by Polymerase Chain Reaction: Nonradioactive PCR Methods* (pp. 65–70). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007/978-3-642-79712-5_6 doi: 10.1007/978-3-642-79712-5_6

Laumer, C. E., & Giribet, G. (2014). Inclusive taxon sampling suggests a single, stepwise origin of ectolecithality in Platyhelminthes. *Biological journal of the Linnean Society*, *111*(3), 570–588. (Publisher: Oxford University Press)

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., … Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS computational biology*, *9*(8), e1003118. (Publisher: Public Library of Science San Francisco, USA)

LeClair, K. P., Blanar, M. A., & Sharp, P. A. (1992). The p50 subunit of NF-kappa B associates with the NF-IL6 transcription factor. *Proceedings of the National Academy of Sciences*, *89*(17), 8145–8149. (Publisher: National Acad Sciences)

Lee, C.-K., Shibata, Y., Rao, B., Strahl, B. D., & Lieb, J. D. (2004). Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature genetics*, *36*(8), 900–905. (Publisher: Nature Publishing Group US New York)

Lee, C.-Y. S., Lu, T., & Seydoux, G. (2017, November). Nanos promotes epigenetic reprograming of the germline by down-regulation of the THAP transcription factor LIN-15B. *eLife*, *6*, e30201. Retrieved from https://doi.org/10.7554/eLife.30201 (Publisher: eLife Sciences Publications, Ltd) doi: 10.7554/eLife.30201

Lei, H., Liu, J., Fukushige, T., Fire, A., & Krause, M. (2009). Caudal-like PAL-1 directly activates the bodywall muscle module regulator hlh-1 in C. elegans to initiate the embryonic muscle gene regulatory network. *Development*. (Publisher: Oxford

University Press for The Company of Biologists Limited)

Lenhard, B., Sandelin, A., & Carninci, P. (2012). Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics*, *13*(4), 233–245. (Publisher: Nature Publishing Group UK London)

Lentjes, M. H., Niessen, H. E., Akiyama, Y., De Bruine, A. P., Melotte, V., & Van Engeland, M. (2016). The emerging role of GATA transcription factors in development and disease. *Expert reviews in molecular medicine*, *18*, e3. (Publisher: Cambridge University Press)

Lespinet, O., Nederbragt, A. J., Cassan, M., Dictus, W. J., van Loon, A. E., & Adoutte, A. (2002). Characterisation of two snail genes in the gastropod mollusc Patella vulgata. Implications for understanding the ancestral function of the snail-related genes in Bilateria. *Development genes and evolution*, *212*, 186–195. (Publisher: Springer)

Li, H. (2023, May). *Seqtk.* Retrieved from https://github.com/lh3/seqtk

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. *bioinformatics*, *25*(16), 2078–2079. (Publisher: Oxford University Press)

Li, X.-Y., Mantovani, R., Hooft van Huijsduijnen, R., Andre, I., Benoist, C., & Mathis, D. (1992). Evolutionary variation of the CCAAT-binding transcription factor NF-Y. *Nucleic acids research*, *20*(5), 1087–1091. (Publisher: Oxford University Press)

Lian, I., Kim, J., Okazawa, H., Zhao, J., Zhao, B., Yu, J., . . . Abujarour, R. (2010). The role of YAP transcription coactivator in regulating stem cell self-renewal and differentiation. *Genes & development*, *24*(11), 1106–1118. (Publisher: Cold Spring Harbor Lab)

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., . . . Dorschner, M. O. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, *326*(5950), 289–293. (Publisher: American Association for the Advancement of Science)

Lifton, R., Goldberg, M., Karp, R., & Hogness, D. (1978). The organization of the histone genes in drosophila melanogaster: functional and evolutionary implications. In *Cold spring harbor symposia on quantitative biology* (Vol. 42, pp. 1047–1051).

Lin, A. Y., & Pearson, B. J. (2014). Planarian yorkie/YAP functions to integrate adult stem cell proliferation, organ homeostasis and maintenance of axial patterning. *Development*, *141*(6), 1197–1208. (Publisher: Company of Biologists)

Lin, K. C., Park, H. W., & Guan, K.-L. (2017). Regulation of the Hippo pathway transcription factor TEAD. *Trends in biochemical sciences*, *42*(11), 862–872. (Publisher: Elsevier)

Local, A., Huang, H., Albuquerque, C. P., Singh, N., Lee, A. Y., Wang, W., ... Ge, K. (2018). Identification of H3K4me1-associated proteins at mammalian enhancers. *Nature genetics*, *50*(1), 73–82. (Publisher: Nature Publishing Group US New York)

Lorch, Y., Maier-Davis, B., & Kornberg, R. D. (2014). Role of DNA sequence in chromatin remodeling and the formation of nucleosome-free regions. *Genes & development*, *28*(22), 2492–2497. (Publisher: Cold Spring Harbor Lab)

Lossky, M., & Wensink, P. C. (1995). Regulation of Drosophila yolk protein genes by an ovary-specific GATA factor. *Molecular and cellular biology*, *15*(12), 6943–6952. (Publisher: Taylor & Francis)

Louder, R. K., He, Y., López-Blanco, J. R., Fang, J., Chacón, P., & Nogales, E. (2016). Structure of promoter-bound TFIID and model of human pre-initiation complex assembly. *Nature*, *531*(7596), 604–609. (Publisher: Nature Publishing Group UK London)

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 1–21. (Publisher: BioMed Central)

Lu, C., Coradin, M., Porter, E. G., & Garcia, B. A. (2021, January). Accelerating the Field of Epigenetic Histone Modification Through Mass Spectrometry–Based Approaches. *Molecular & Cellular Proteomics*, *20*. Retrieved 2023-10-13, from https://doi.org/10.1074/mcp.R120.002257 (Publisher: Elsevier) doi: 10.1074/mcp.R120.002257

Lupiáñez, D., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., ... Mundlos, S. (2015, May). Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, *161*(5), 1012–1025. Retrieved 2023-10-15, from https://doi.org/10.1016/j.cell.2015.04.004 (Publisher: Elsevier) doi: 10.1016/j.cell.2015.04.004

Lykidis, D., Van Noorden, S., Armstrong, A., Spencer-Dene, B., Li, J., Zhuang, Z., & Stamp, G. W. (2007). Novel zinc-based fixative for high quality DNA, RNA and protein analysis. *Nucleic acids research*, *35*(12), e85. (Publisher: Oxford University Press)

Lázaro, E. M., Harrath, A. H., Stocchino, G. A., Pala, M., Baguñà, J., & Riutort, M.

(2011). Schmidtea mediterranea phylogeography: an old species surviving on a few Mediterranean islands? *BMC evolutionary biology*, *11*, 1–15. (Publisher: Springer)

López-Mirabal, H. R., & Winther, J. R. (2008). Redox characteristics of the eukaryotic cytosol. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, *1783*(4), 629–640. (Publisher: Elsevier)

Madrigal, P., Deng, S., Feng, Y., Militi, S., Goh, K. J., Nibhani, R., … Brown, S. (2023). Epigenetic and transcriptional regulations prime cell fate before division during human pluripotent stem cell differentiation. *Nature Communications*, *14*(1), 405. (Publisher: Nature Publishing Group UK London)

Madsen, J. G. S., Rauch, A., Van Hauwaert, E. L., Schmidt, S. F., Winnefeld, M., & Mandrup, S. (2018). Integrated analysis of motif activity and gene expression changes of transcription factors. *Genome research*, *28*(2), 243–255. (Publisher: Cold Spring Harbor Lab)

Maduro, M. F. (2010). Cell fate specification in the C. elegans embryo. *Developmental dynamics: an official publication of the American Association of Anatomists*, *239*(5), 1315–1329. (Publisher: Wiley Online Library)

Magri, M. S., Jiménez-Gancedo, S., Bertrand, S., Madgwick, A., Escrivà, H., Lemaire, P., & Gómez-Skarmeta, J. L. (2020). Assaying chromatin accessibility using ATAC-seq in invertebrate chordate embryos. *Frontiers in Cell and Developmental Biology*, *7*, 372. (Publisher: Frontiers Media SA)

Maity, S. N., & De Crombrugghe, B. (1998). Role of the CCAAT-binding protein CBF/NF-Y in transcription. *Trends in biochemical sciences*, *23*(5), 174–178. (Publisher: Elsevier)

Mannervik, M. (2014, February). Control of Drosophila embryo patterning by transcriptional co-regulators. *Developmental Biology*, *321*(1), 47–57. Retrieved from https://www.sciencedirect.com/science/article/pii/S0014482713004394 doi: 10.1016/j.yexcr.2013.10.010

Manu, Surkova, S., Spirov, A. V., Gursky, V. V., Janssens, H., Kim, A.-R., … Samsonova, M. (2009). Canalization of gene expression in the Drosophila blastoderm by gap gene cross regulation. *PLoS biology*, *7*(3), e1000049. (Publisher: Public Library of Science San Francisco, USA)

Martín-Durán, J. M., & Egger, B. (2012). Developmental diversity in free-living flatworms. *EvoDevo*, *3*(1), 1–23. (Publisher: BioMed Central)

Mavrich, T. N., Jiang, C., Ioshikhes, I. P., Li, X., Venters, B. J., Zanton, S. J., …

Schuster, S. C. (2008). Nucleosome organization in the Drosophila genome. *Nature*, *453*(7193), 358–362. (Publisher: Nature Publishing Group UK London)

Mayor, R., Guerrero, N., Young, R., Gomez-Skarmeta, J., & Cuellar, C. (2000). A novel function for the Xslug gene: control of dorsal mesendoderm development by repressing BMP-4. *Mechanisms of development*, *97*(1-2), 47–56. (Publisher: Elsevier)

Mayran, A., & Drouin, J. (2018). Pioneer transcription factors shape the epigenetic landscape. *Journal of Biological Chemistry*, *293*(36), 13795–13804. (Publisher: ASBMB)

Merryman, M. S., Sánchez Alvarado, A., & Jenkin, J. C. (2018). Culturing Planarians in the Laboratory. *Planarian regeneration: Methods and protocols*, 241–258. (Publisher: Springer)

Michel, M., Demel, C., Zacher, B., Schwalb, B., Krebs, S., Blum, H., … Cramer, P. (2017). TT-seq captures enhancer landscapes immediately after T-cell stimulation. *Molecular systems biology*, *13*(3), 920.

Mihaylova, Y., Abnave, P., Kao, D., Hughes, S., Lai, A., Jaber-Hijazi, F., … Aboobaker, A. A. (2018). Conservation of epigenetic regulation by the MLL3/4 tumour suppressor in planarian pluripotent stem cells. *Nature communications*, *9*(1), 3633. (Publisher: Nature Publishing Group UK London)

Mikhaylichenko, O., Bondarenko, V., Harnett, D., Schor, I. E., Males, M., Viales, R. R., & Furlong, E. E. (2018). The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes & development*, *32*(1), 42–57. (Publisher: Cold Spring Harbor Lab)

Milne, T. A., Dou, Y., Martin, M. E., Brock, H. W., Roeder, R. G., & Hess, J. L. (2005). MLL associates specifically with a subset of transcriptionally active target genes. *Proceedings of the National Academy of Sciences*, *102*(41), 14765–14770. (Publisher: National Acad Sciences)

Minguillon, C., Nishimoto, S., Wood, S., Vendrell, E., Gibson-Brown, J. J., & Logan, M. P. (2012). Hox genes regulate the onset of Tbx5 expression in the forelimb. *Development*, *139*(17), 3180–3188. (Publisher: Company of Biologists)

modENCODE Consortium, Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., … Ma, L. (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, *330*(6012), 1787–1797. (Publisher: American Association for the Advancement of Science)

Molina, M. D., & Cebrià, F. (2021). Decoding stem cells: An overview on planarian stem cell heterogeneity and lineage progression. *Biomolecules*, *11*(10), 1532. (Publisher: MDPI)

Monjo, F., & Romero, R. (2015). Embryonic development of the nervous system in the planarian Schmidtea polychroa. *Developmental Biology*, *397*(2), 305–319. (Publisher: Elsevier)

Morgan, T. H. (1898). *Experimental studies of the regeneration of Planaria maculata* (Vol. 2). W. Engelmann.

Morgunova, E., & Taipale, J. (2017, December). Structural perspective of cooperative transcription factor binding. *Protein–nucleic acid interactions • Catalysis and regulation*, *47*, 1–8. Retrieved from https://www.sciencedirect.com/science/article/pii/S0959440X17300088 doi: 10.1016/j.sbi.2017.03.006

Morgunova, E., Yin, Y., Das, P. K., Jolma, A., Zhu, F., Popov, A., . . . Taipale, J. (2018, April). Two distinct DNA sequences recognized by transcription factors represent enthalpy and entropy optima. *eLife*, *7*, e32963. Retrieved from https://doi.org/10.7554/eLife.32963 (Publisher: eLife Sciences Publications, Ltd) doi: 10.7554/eLife.32963

Mousavi, K., Zare, H., Dell'Orso, S., Grontved, L., Gutierrez-Cruz, G., Derfoul, A., . . . Sartorelli, V. (2013). eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Molecular cell*, *51*(5), 606–617. (Publisher: Elsevier)

Musselman, C. A., Lalonde, M.-E., Côté, J., & Kutateladze, T. G. (2012, December). Perceiving the epigenetic landscape through histone readers. *Nature Structural & Molecular Biology*, *19*(12), 1218–1227. Retrieved from https://doi.org/10.1038/nsmb.2436 doi: 10.1038/nsmb.2436

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., . . . Kanitz, A. (2021). Sustainable data analysis with Snakemake. *F1000Research*, *10*. (Publisher: Faculty of 1000 Ltd)

Nabbi, A., & Riabowol, K. (2015). Isolation of Pure Nuclei Using a Sucrose Method. *Cold Spring Harbor protocols*, *2015*(8), 773–776.

Nardone, V., Chaves-Sanjuan, A., & Nardini, M. (2017). Structural determinants for NF-Y/DNA interaction at the CCAAT box. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, *1860*(5), 571–580. (Publisher: Elsevier)

Nechaev, S., Fargo, D. C., dos Santos, G., Liu, L., Gao, Y., & Adelman, K. (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science*, *327*(5963), 335–338. (Publisher: American

Association for the Advancement of Science)

Neil, H., Malabat, C., d'Aubenton Carafa, Y., Xu, Z., Steinmetz, L. M., & Jacquier, A. (2009). Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature*, *457*(7232), 1038–1042. (Publisher: Nature Publishing Group UK London)

Neiro, J., Sridhar, D., Dattani, A., & Aboobaker, A. (2022). Identification of putative enhancer-like elements predicts regulatory networks active in planarian adult stem cells. *Elife*, *11*, e79675. (Publisher: eLife Sciences Publications Limited)

Newmark, P., Wang, Y., & Chong, T. (2008). Germ cell specification and regeneration in planarians. In (Vol. 73, pp. 573–581). Cold Spring Harbor Laboratory Press.

Newmark, P. A., & Alvarado, A. S. (2002, March). Not your father's planarian: a classic model enters the era of functional genomics. *Nature Reviews Genetics*, *3*(3), 210–219. Retrieved from https://doi.org/10.1038/nrg759 doi: 10.1038/nrg759

Ng, H. H., Robert, F., Young, R. A., & Struhl, K. (2003). Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Molecular cell*, *11*(3), 709–719. (Publisher: Elsevier)

Nguyen, T. A., Jones, R. D., Snavely, A. R., Pfenning, A. R., Kirchner, R., Hemberg, M., & Gray, J. M. (2016). High-throughput functional comparison of promoter and enhancer activities. *Genome research*, *26*(8), 1023–1033.

Niwa, H., Toyooka, Y., Shimosato, D., Strumpf, D., Takahashi, K., Yagi, R., & Rossant, J. (2005). Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation. *Cell*, *123*(5), 917–929. (Publisher: Elsevier)

Ntini, E., Järvelin, A. I., Bornholdt, J., Chen, Y., Boyd, M., Jørgensen, M., ... Jensen, T. H. (2013, August). Polyadenylation site–induced decay of upstream transcripts enforces promoter directionality. *Nature Structural & Molecular Biology*, *20*(8), 923–928. Retrieved from https://doi.org/10.1038/nsmb.2640 doi: 10.1038/nsmb.2640

Nwokeoji, A. O., Kilby, P. M., Portwood, D. E., & Dickman, M. J. (2017). Accurate quantification of nucleic acids using hypochromicity measurements in conjunction with UV spectrophotometry. *Analytical chemistry*, *89*(24), 13567–13574. (Publisher: ACS Publications)

Nüsslein-Volhard, C., & Wieschaus, E. (1980, October). Mutations affecting segment number and polarity in Drosophila. *Nature*, *287*(5785), 795–801. Retrieved from https://doi.org/10.1038/287795a0 doi: 10.1038/287795a0

Ohler, U., Liao, G.-c., Niemann, H., & Rubin, G. M. (2002). Computational analysis of core promoters in the Drosophila genome. *Genome biology*, *3*, 1–12. (Publisher: Springer)

Oji, A., Isotani, A., Fujihara, Y., Castaneda, J. M., Oura, S., & Ikawa, M. (2020). TESMIN, METALLOTHIONEIN-LIKE 5, is Required for Spermatogenesis in Mice. *Biology of reproduction*, *102*(4), 975–983. (Publisher: Oxford University Press)

Oldfield, A. J., Yang, P., Conway, A. E., Cinghu, S., Freudenberg, J. M., Yellaboina, S., & Jothi, R. (2014). Histone-fold domain protein NF-Y promotes chromatin accessibility for cell type-specific master transcription factors. *Molecular cell*, *55*(5), 708–722. (Publisher: Elsevier)

Oliveri, P., Carrick, D. M., & Davidson, E. H. (2002). A regulatory gene network that directs micromere specification in the sea urchin embryo. *Developmental biology*, *246*(1), 209–228. (Publisher: Elsevier)

Oliveri, P., Tu, Q., & Davidson, E. H. (2008). Global regulatory logic for specification of an embryonic cell lineage. *Proceedings of the National Academy of Sciences*, *105*(16), 5955–5962. (Publisher: National Acad Sciences)

Osterburg, H. H., Allen, J., & Finch, C. E. (1975). The use of ammonium acetate in the precipitation of ribonucleic acid. *Biochemical Journal*, *147*(2), 367. (Publisher: Portland Press Ltd)

Ou, J., Liu, H., Yu, J., Kelliher, M. A., Castilla, L. H., Lawson, N. D., & Zhu, L. J. (2018). ATACseqQC: a Bioconductor package for post-alignment quality assessment of ATAC-seq data. *BMC genomics*, *19*, 1–13. (Publisher: Springer)

Owlarn, S., Klenner, F., Schmidt, D., Rabert, F., Tomasso, A., Reuter, H., . . . Weidinger, G. (2017). Generic wound signals initiate regeneration in missing-tissue contexts. *Nature communications*, *8*(1), 2282. (Publisher: Nature Publishing Group UK London)

Owraghi, M., Broitman-Maduro, G., Luu, T., Roberson, H., & Maduro, M. F. (2010, April). Roles of the Wnt effector POP-1/TCF in the C. elegans endomesoderm specification gene network. *Special Section: Gene Regulatory Networks for Development*, *340*(2), 209–221. Retrieved from https://www.sciencedirect.com/science/article/pii/S0012160609012421 doi: 10.1016/j.ydbio.2009.09.042

Ozsolak, F., Kapranov, P., Foissac, S., Kim, S. W., Fishilevich, E., Monaghan, A. P., . . . Milos, P. M. (2010, December). Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation. *Cell*, *143*(6), 1018–

1029. Retrieved 2023-10-16, from https://doi.org/10.1016/j.cell.2010.11.020 (Publisher: Elsevier) doi: 10.1016/j.cell.2010.11.020

Pagès, H. (2017). BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs. *R package version*, *1* (0), 10–18129.

Panne, D. (2008, April). The enhanceosome. *Theory and simulation / Macromolecular assemblages*, *18* (2), 236–242. Retrieved from https://www.sciencedirect.com/science/article/pii/S0959440X07002023 doi: 10.1016/j.sbi.2007.12.002

Pascual-Carreras, E., Marín-Barba, M., Castillo-Lara, S., Coronel-Córdoba, P., Magri, M. S., Wheeler, G. N., ... Adell, T. (2023). Wnt/β-catenin signalling is required for pole-specific chromatin remodeling during planarian regeneration. *Nature Communications*, *14* (1), 298. (Publisher: Nature Publishing Group UK London)

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, *14* (4), 417–419. (Publisher: Nature Publishing Group US New York)

Pearson, B. J., Eisenhoffer, G. T., Gurley, K. A., Rink, J. C., Miller, D. E., & Sánchez Alvarado, A. (2009). Formaldehyde-based whole-mount in situ hybridization method for planarians. *Developmental dynamics*, *238* (2), 443–450. (Publisher: Wiley Online Library)

Pekowska, A., Benoukraf, T., Zacarias-Cabeza, J., Belhocine, M., Koch, F., Holota, H., ... Spicuglia, S. (2011). H3K4 tri-methylation provides an epigenetic signature of active enhancers. *The EMBO journal*, *30* (20), 4198–4210. (Publisher: John Wiley & Sons, Ltd Chichester, UK)

Pellettieri, J., & Alvarado, A. S. (2007). Cell turnover and adult tissue homeostasis: from humans to planarians. *Annu. Rev. Genet.*, *41* , 83–105. (Publisher: Annual Reviews)

Pengelly, A. R., Copur, O., Jackle, H., Herzig, A., & Müller, J. (2013). A histone mutant reproduces the phenotype caused by loss of histone-modifying factor Polycomb. *Science*, *339* (6120), 698–699. (Publisher: American Association for the Advancement of Science)

Perkins, T. J., Jaeger, J., Reinitz, J., & Glass, L. (2006). Reverse engineering the gap gene network of Drosophila melanogaster. *PLoS computational biology*, *2* (5), e51. (Publisher: Public Library of Science San Francisco, USA)

Pertea, G., & Pertea, M. (2020). GFF utilities: GffRead and GffCompare. *F1000Research*, *9* . (Publisher: Faculty of 1000 Ltd)

Peter, I. S., & Davidson, E. H. (2009). Modularity and design principles in the sea urchin embryo gene regulatory network. *FEBS letters*, *583*(24), 3948–3958. (Publisher: Elsevier)

Peter, I. S., & Davidson, E. H. (2011). A gene regulatory network controlling the embryonic specification of endoderm. *Nature*, *474*(7353), 635–639. (Publisher: Nature Publishing Group UK London)

Peter van 't Hof, Vorderman, R., & Cats, D. (2018, February). *Validatefastq.* GitHub. Retrieved from https://github.com/biopet/validatefastq

Petersen, C. P., & Reddien, P. W. (2009). A wound-induced Wnt expression program controls planarian regeneration polarity. *Proceedings of the National Academy of Sciences*, *106*(40), 17061–17066. (Publisher: National Acad Sciences)

Petersen, C. P., & Reddien, P. W. (2011a). Polarized notum activation at wounds inhibits Wnt function to promote planarian head regeneration. *Science*, *332*(6031), 852–855. (Publisher: American Association for the Advancement of Science)

Petersen, C. P., & Reddien, P. W. (2011b). Polarized notum activation at wounds inhibits wnt function to promote planarian head regeneration. *Science*, *332*(6031), 852–855.

Petrascheck, M., Escher, D., Mahmoudi, T., Verrijzer, C. P., Schaffner, W., & Barberis, A. (2005). DNA looping induced by a transcriptional enhancer in vivo. *Nucleic acids research*, *33*(12), 3743–3750. (Publisher: Oxford University Press)

Plass, M., Solana, J., Wolf, F. A., Ayoub, S., Misios, A., Glažar, P., … Rajewsky, N. (2018). Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, *360*(6391), eaaq1723. (Publisher: American Association for the Advancement of Science)

Pokholok, D. K., Harbison, C. T., Levine, S., Cole, M., Hannett, N. M., Lee, T. I., … Herbolsheimer, E. (2005). Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, *122*(4), 517–527. (Publisher: Elsevier)

Poulet, A., Kratkiewicz, A. J., Li, D., & van Wolfswinkel, J. C. (2023). Chromatin analysis of adult pluripotent stem cells reveals a unique stemness maintenance strategy. *Science Advances*, *9*(40), eadh4887.

Pradhan, A., Khalaf, H., Ochsner, S. A., Sreenivasan, R., Koskinen, J., Karlsson, M., … Olsson, P.-E. (2012). Activation of nf-kb protein prevents the transition from juvenile ovary to testis and promotes ovarian development in zebrafish. *Journal of Biological Chemistry*, *287*(45), 37926–37938. (Publisher: ASBMB)

Prasad, B. C., Ye, B., Zackhary, R., Schrader, K., Seydoux, G., & Reed, R. R. (1998).

unc-3, a gene required for axonal guidance in Caenorhabditis elegans, encodes a member of the O/E family of transcription factors. *Development*, *125*(8), 1561–1568. (Publisher: The Company of Biologists Ltd)

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M. S., Mapendano, C. K., ... Jensen, T. H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. *Science*, *322*(5909), 1851–1854. (Publisher: American Association for the Advancement of Science)

Project, A. S. H. L. E. T. (2009). Post-transcriptional processing generates a diversity of 5'-modified long and short rnas. *Nature*, *457*(7232), 1028–1032.

Prusch, R. D. (1976). Osmotic and ionic relationships in the fresh-water flatworm, Dugesia dorotocephala. *Comparative Biochemistry and Physiology Part A: Physiology*, *54*(3), 287–290. (Publisher: Elsevier)

Quesneville, H., Nouaud, D., & Anxolabehere, D. (2005, March). Recurrent Recruitment of the THAP DNA-Binding Domain and Molecular Domestication of the P-Transposable Element. *Molecular Biology and Evolution*, *22*(3), 741–746. Retrieved 2023-09-24, from https://doi.org/10.1093/molbev/msi064 doi: 10.1093/molbev/msi064

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. (Publisher: Oxford University Press)

Quiring, R., Walldorf, U., Kloter, U., & Gehring, W. J. (1994). Homology of the eyeless gene of Drosophila to the Small eye gene in mice and Aniridia in humans. *Science*, *265*(5173), 785–789. (Publisher: American Association for the Advancement of Science)

Ramasamy, S., Aljahani, A., Karpinska, M. A., Cao, T. B. N., Velychko, T., Cruz, J. N., ... Oudelaar, A. M. (2023, July). The Mediator complex regulates enhancer-promoter interactions. *Nature Structural & Molecular Biology*, *30*(7), 991–1000. Retrieved from https://doi.org/10.1038/s41594-023-01027-2 doi: 10.1038/s41594-023-01027-2

Ramirez, A. N., Loubet-Senear, K., & Srivastava, M. (2020). A regulatory program for initiation of Wnt signaling during posterior regeneration. *Cell reports*, *32*(9). (Publisher: Elsevier)

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., ... Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic acids research*, *44*(W1), W160–W165. (Publisher: Oxford University Press)

Raz, A. A., Wurtzel, O., & Reddien, P. W. (2021). Planarian stem cells specify fate yet retain potency during the cell cycle. *Cell Stem Cell*, *28*(7), 1307–1322. (Publisher: Elsevier)

Reddien, P. W. (2018). The cellular and molecular basis for planarian regeneration. *Cell*, *175*(2), 327–345. (Publisher: Elsevier)

Reddien, P. W. (2021). Principles of regeneration revealed by the planarian eye. *Current Opinion in Cell Biology*, *73*, 19–25.

Reddien, P. W., Oviedo, N. J., Jennings, J. R., Jenkin, J. C., & Alvarado, A. S. (2005). Smedwi-2 is a piwi-like protein that regulates planarian stem cells. *Science*, *310*(5752), 1327–1330.

Reiter, F., Wienerroither, S., & Stark, A. (2017, April). Combinatorial function of transcription factors and cofactors. *Genome architecture and expression*, *43*, 73–81. Retrieved from https://www.sciencedirect.com/science/article/pii/S0959437X17300059 doi: 10.1016/j.gde.2016.12.007

Reuter, H., März, M., Vogg, M. C., Eccles, D., Grífol-Boldú, L., Wehner, D., ... Bartscherer, K. (2015). β-catenin-dependent control of positional information along the AP body axis in planarians involves a teashirt family member. *Cell reports*, *10*(2), 253–265. (Publisher: Elsevier)

Revilla-i Domingo, R., Oliveri, P., & Davidson, E. H. (2007). A missing link in the sea urchin embryo gene regulatory network: hesC and the double-negative specification of micromeres. *Proceedings of the National Academy of Sciences*, *104*(30), 12383–12388. (Publisher: National Acad Sciences)

Richter, W. F., Nayak, S., Iwasa, J., & Taatjes, D. J. (2022). The Mediator complex as a master regulator of transcription by RNA polymerase II. *Nature Reviews Molecular Cell Biology*, *23*(11), 732–749. (Publisher: Nature Publishing Group UK London)

Rickels, R., Herz, H.-M., Sze, C. C., Cao, K., Morgan, M. A., Collings, C. K., ... Rendleman, E. J. (2017). Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. *Nature genetics*, *49*(11), 1647–1653. (Publisher: Nature Publishing Group US New York)

Rink, J. C. (2013). Stem cell systems and regeneration in planaria. *Development genes and evolution*, *223*, 67–84. (Publisher: Springer)

Rink, J. C. (2018). Stem cells, patterning and regeneration in planarians: self-organization at the organismal scale. *Planarian regeneration: methods and protocols*, 57–172. (Publisher: Springer)

Rink, J. C., Vu, H. T.-K., & Alvarado, A. S. (2011). The maintenance and regeneration of the planarian excretory system are regulated by EGFR signaling. *Development*, *138*(17), 3769–3780. (Publisher: Company of Biologists)

Rippe, K., & Papantonis, A. (2021). RNA polymerase II transcription compartments: from multivalent chromatin binding to liquid droplet formation? *Nature Reviews Molecular Cell Biology*, *22*(10), 645–646. (Publisher: Nature Publishing Group UK London)

Ririe, T. O., Fernandes, J. S., & Sternberg, P. W. (2008, December). The Caenorhabditis elegans vulva: A post-embryonic gene regulatory network controlling organogenesis. *Proceedings of the National Academy of Sciences*, *105*(51), 20095–20099. Retrieved 2023-10-21, from https://doi.org/10.1073/pnas.0806377105 (Publisher: Proceedings of the National Academy of Sciences) doi: 10.1073/pnas.0806377105

Roberts-Galbraith, R. H., & Newmark, P. A. (2015). On the organ trail: insights into organ regeneration in the planarian. *Current opinion in genetics & development*, *32*, 37–46. (Publisher: Elsevier)

Rodríguez-Esteban, G., González-Sastre, A., Rojo-Laguna, J. I., Saló, E., & Abril, J. F. (2015). Digital gene expression approach over multiple RNA-Seq data sets to detect neoblast transcriptional changes in Schmidtea mediterranea. *BMC genomics*, *16*(1), 1–23. (Publisher: BioMed Central)

Rosenfeld, M. G., Lunyack, V. V., & Glass, C. K. (2006). Sensors and signals: a coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response. *Genes & development*, *20*(11).

Ross, K. G., Currie, K. W., Pearson, B. J., & Zayas, R. M. (2017). Nervous system development and regeneration in freshwater planarians. *Wiley Interdisciplinary Reviews: Developmental Biology*, *6*(3), e266. (Publisher: Wiley Online Library)

Rossi, A., Ross, E. J., Jack, A., & Alvarado, A. S. (2014). Molecular cloning and characterization of SL3: a stem cell-specific SL RNA from the planarian Schmidtea mediterranea. *Gene*, *533*(1), 156–167. (Publisher: Elsevier)

Rottman, F., Shatkin, A. J., & Perry, R. P. (1974). Sequences containing methylated nucleotides at the 5' termini of messenger RNAs: possible implications for processing. *Cell*, *3*(3), 197–199. (Publisher: Elsevier)

Rouhana, L., Tasaki, J., Saberi, A., & Newmark, P. A. (2017). Genetic dissection of the planarian reproductive system through characterization of Schmidtea mediterranea CPEB homologs. *Developmental biology*, *426*(1), 43–55. (Publisher: Elsevier)

Rouhana, L., Weiss, J. A., Forsthoefel, D. J., Lee, H., King, R. S., Inoue, T., ... Newmark, P. A. (2013). RNA interference by feeding in vitro–synthesized double-stranded RNA to planarians: Methodology and dynamics. *Developmental dynamics*, *242*(6), 718–730. (Publisher: Wiley Online Library)

Roussigne, M., Cayrol, C., Clouaire, T., Amalric, F., & Girard, J.-P. (2003). THAP1 is a nuclear proapoptotic factor that links prostate-apoptosis-response-4 (Par-4) to PML nuclear bodies. *Oncogene*, *22*(16), 2432–2442. (Publisher: Nature Publishing Group)

Rowley, M. J., & Corces, V. G. (2018). Organizational principles of 3D genome architecture. *Nature Reviews Genetics*, *19*(12), 789–800. (Publisher: Nature Publishing Group UK London)

Rozanski, A., Moon, H., Brandl, H., Martín-Durán, J. M., Grohme, M. A., Hüttner, K., ... Rink, J. C. (2019). PlanMine 3.0—improvements to a mineable resource of flatworm biology and biodiversity. *Nucleic acids research*, *47*(D1), D812–D820. (Publisher: Oxford University Press)

Rørth, P., Szabo, K., & Texido, G. (2000). The level of C/EBP protein is critical for cell migration during Drosophila oogenesis and is tightly controlled by regulated degradation. *Molecular cell*, *6*(1), 23–30. (Publisher: Elsevier)

Saberi, A., Jamal, A., Beets, I., Schoofs, L., & Newmark, P. A. (2016). GPCRs direct germline development and somatic gonad function in planarians. *PLoS biology*, *14*(5), e1002457. (Publisher: Public Library of Science San Francisco, CA USA)

Sambrook, J., Fritsch, E. F., & Maniatis, T. (1989). Chapter 5 Gel Electrophoresis of DNA and Pulsed-Field Agarose. In *Molecular cloning: a laboratory manual.* (pp. 444–450). Cold spring harbor laboratory press.

Sambrook, J., & Russell, D. W. (2006). Preparation of plasmid DNA by alkaline lysis with SDS: minipreparation. *Cold Spring Harbor Protocols*, *2006*(1), pdb–prot4084. (Publisher: Cold Spring Harbor Laboratory Press)

Santos, A. J., Lo, Y.-H., Mah, A. T., & Kuo, C. J. (2018). The intestinal stem cell niche: homeostasis and adaptations. *Trends in cell biology*, *28*(12), 1062–1078. (Publisher: Elsevier)

Santos-Rosa, H., Schneider, R., Bannister, A. J., Sherriff, J., Bernstein, B. E., Emre,

N. T., … Kouzarides, T. (2002). Active genes are tri-methylated at K4 of histone H3. *Nature*, *419*(6905), 407–411. (Publisher: Nature Publishing Group UK London)

Sarikaya, D. P., & Extavour, C. G. (2015). The Hippo pathway regulates homeostatic growth of stem cell niche precursors in the Drosophila ovary. *PLoS genetics*, *11*(2), e1004962. (Publisher: Public Library of Science San Francisco, CA USA)

Saxonov, S., Berg, P., & Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*, *103*(5), 1412–1417. (Publisher: National Acad Sciences)

Scazzocchio, C. (2000). The fungal GATA factors. *Current opinion in microbiology*, *3*(2), 126–131. (Publisher: Elsevier)

Schep, A. (2020). motifmatchr: fast motif matching in R. *R package version*, *1*(0).

Schep, A. N., Wu, B., Buenrostro, J. D., & Greenleaf, W. J. (2017). chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nature methods*, *14*(10), 975–978. (Publisher: Nature Publishing Group UK London)

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., … Schmid, B. (2012). Fiji: an open-source platform for biological-image analysis. *Nature methods*, *9*(7), 676–682. (Publisher: Nature Publishing Group US New York)

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., … Crowley, J. (2011). *Ggally: Extension to ggplot2. r package version 3.2. 5.*

Schwalb, B., Michel, M., Zacher, B., Frühauf, K., Demel, C., Tresch, A., … Cramer, P. (2016). TT-seq maps the human transient transcriptome. *Science*, *352*(6290), 1225–1228. (Publisher: American Association for the Advancement of Science)

Schürmann, W., & Peter, R. (2001). Planarian cell culture: a comparative review of methods and an improved protocol for primary cultures of neoblasts. *Belg. J. Zool*, *131*(Suppl 1), 123–130.

Scimone, M. L., Lapan, S. W., & Reddien, P. W. (2014). A forkhead transcription factor is wound-induced at the planarian midline and required for anterior pole regeneration. *PLoS genetics*, *10*(1), e1003999.

Scruggs, B. S., Gilchrist, D. A., Nechaev, S., Muse, G. W., Burkholder, A., Fargo, D. C., & Adelman, K. (2015). Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Molecular cell*, *58*(6), 1101–1112. (Publisher: Elsevier)

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., & Gaul, U. (2008). Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, *451*(7178), 535–540. (Publisher: Nature Publishing Group UK London)

Seila, A. C., Calabrese, J. M., Levine, S. S., Yeo, G. W., Rahl, P. B., Flynn, R. A., . . . Sharp, P. A. (2008). Divergent transcription from active promoters. *science*, *322*(5909), 1849–1851. (Publisher: American Association for the Advancement of Science)

Seita, J., & Weissman, I. L. (2010). Hematopoietic stem cell: self-renewal versus differentiation. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, *2*(6), 640–653. (Publisher: Wiley Online Library)

Servitja, J., & Ferrer, J. (2004). Transcriptional networks controlling pancreatic development and beta cell function. *Diabetologia*, *47*, 597–613. (Publisher: Springer)

Shiraki, T., Kondo, S., Katayama, S., Waki, K., Kasukawa, T., Kawaji, H., . . . Arakawa, T. (2003). Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, *100*(26), 15776–15781. (Publisher: National Acad Sciences)

Sielemann, J., Wulf, D., Schmidt, R., & Bräutigam, A. (2021). Local DNA shape is a general principle of transcription factor binding specificity in Arabidopsis thaliana. *Nature communications*, *12*(1), 6549. (Publisher: Nature Publishing Group UK London)

Singh, L., Wadhwa, R., Naidu, S., Nagaraj, R., & Ganesan, M. (1994). Sex-and tissue-specific Bkm (GATA)-binding protein in the germ cells of heterogametic sex. *Journal of Biological Chemistry*, *269*(41), 25321–25327. (Publisher: Elsevier)

Slowikowski, K. (2023). *ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'.* Retrieved from https://github.com/slowkow/ggrepel

Sluys, R., & Riutort, M. (2018). Planarian diversity and phylogeny. *Planarian regeneration: methods and protocols*, 1–56. (Publisher: Springer)

Smale, S. T., & Baltimore, D. (1989). The "initiator" as a transcription control element. *Cell*, *57*(1), 103–113. (Publisher: Cell Press)

Soneson, C., Love, M. I., & Robinson, M. D. (2015). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, *4*. (Publisher: Faculty of 1000 Ltd)

Statello, L., Guo, C.-J., Chen, L.-L., & Huarte, M. (2021, February). Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular*

Cell Biology, *22*(2), 96–118. Retrieved from https://doi.org/10.1038/s41580-020-00315-9 doi: 10.1038/s41580-020-00315-9

Steinberger, E. (1971). Hormonal control of mammalian spermatogenesis. *Physiological Reviews*, *51*(1), 1–22.

Steiner, J. K., Tasaki, J., & Rouhana, L. (2016). Germline defects caused by Smed-boule RNA-interference reveal that egg capsule deposition occurs independently of fertilization, ovulation, mating, or the presence of gametes in planarian flatworms. *PLoS Genetics*, *12*(5), e1006030. (Publisher: Public Library of Science San Francisco, CA USA)

Stevens, T. J., Lando, D., Basu, S., Atkinson, L. P., Cao, Y., Lee, S. F., ... O'Shaughnessy-Kirwan, A. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, *544*(7648), 59–64. (Publisher: Nature Publishing Group UK London)

Streit, A., Bernasconi, L., Sergeev, P., Cruz, A., & Steinmann-Zwicky, M. (2002). mgm 1, the earliest sex-specific germline marker in Drosophila, reflects expression of the gene esg in male stem cells. *International Journal of Developmental Biology*, *46*(1), 159–166. (Publisher: University of the Basque Country Press (UBC Press))

Strong, E. R., & Schimenti, J. C. (2010). Evidence implicating CCNB1IP1, a RING domain-containing protein required for meiotic crossing over in mice, as an E3 SUMO ligase. *Genes*, *1*(3), 440–451. (Publisher: MDPI)

Stückemann, T., Cleland, J. P., Werner, S., Vu, H. T.-K., Bayersdorf, R., Liu, S.-Y., ... Rink, J. C. (2017). Antagonistic self-organizing patterning systems control maintenance and regeneration of the anteroposterior axis in planarians. *Developmental cell*, *40*(3), 248–263. (Publisher: Elsevier)

Sun, S., Xie, H., Sun, Y., Song, J., & Li, Z. (2012). Molecular characterization of gap region in 28S rRNA molecules in brine shrimp Artemia parthenogenetica and planarian Dugesia japonica. *Biochemistry (Moscow)*, *77*, 411–417. (Publisher: Springer)

Sureda-Gomez, M., Martin-Duran, J. M., & Adell, T. (2016). Localization of planarian β-CATENIN-1 reveals multiple roles during anterior-posterior regeneration and organogenesis. *Development*, *143*(22), 4149–4160. (Publisher: The Company of Biologists Ltd)

Sutou, S., Miwa, K., Matsuura, T., Kawasaki, Y., Ohinata, Y., & Mitsui, Y. (2003). Native tesmin is a 60-kilodalton protein that undergoes dynamic changes in its localization during spermatogenesis in mice. *Biology of reproduction*, *68*(5), 1861–1869. (Publisher: Oxford University Press)

Suzuki, S., McCarrey, J. R., & Hermann, B. P. (2021). An mTORC1-dependent switch orchestrates the transition between mouse spermatogonial stem cells and clones of progenitor spermatogonia. *Cell reports*, *34*(7). (Publisher: Elsevier)

Swiers, G., Patient, R., & Loose, M. (2006). Genetic regulatory networks programming hematopoietic stem cells and erythroid lineage specification. *Developmental biology*, *294*(2), 525–540. (Publisher: Elsevier)

Takeda, H., Nishimura, K., & Agata, K. (2009). Planarians maintain a constant ratio of different cell types during changes in body size by using the stem cell system. *Zoological science*, *26*(12), 805–813. (Publisher: BioOne)

Talbert, P. B., & Henikoff, S. (2010, April). Histone variants — ancient wrap artists of the epigenome. *Nature Reviews Molecular Cell Biology*, *11*(4), 264–275. Retrieved from https://doi.org/10.1038/nrm2861 doi: 10.1038/nrm2861

Tewari, A. G., Owen, J. H., Petersen, C. P., Wagner, D. E., & Reddien, P. W. (2019). A small set of conserved genes, including sp5 and Hox, are activated by Wnt signaling in the posterior of planarians and acoels. *PLoS genetics*, *15*(10), e1008401. (Publisher: Public Library of Science San Francisco, CA USA)

Tharp, M. E., Collins III, J. J., & Newmark, P. A. (2014). A lophotrochozoan-specific nuclear hormone receptor is required for reproductive system development in the planarian. *Developmental biology*, *396*(1), 150–157. (Publisher: Elsevier)

The FANTOM Consortium, & Riken Omics Science Center. (2009, May). The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nature Genetics*, *41*(5), 553–562. Retrieved from https://doi.org/10.1038/ng.375 doi: 10.1038/ng.375

Thi-Kim Vu, H., Rink, J. C., McKinney, S. A., McClain, M., Lakshmanaperumal, N., Alexander, R., & Sánchez Alvarado, A. (2015). Stem cells and fluid flow drive cyst formation in an invertebrate excretory organ. *Elife*, *4*, e07405. (Publisher: eLife Sciences Publications, Ltd)

Thommen, A., Werner, S., Frank, O., Philipp, J., Knittelfelder, O., Quek, Y., . . . Jülicher, F. (2019). Body size-dependent energy storage causes Kleiber's law scaling of the metabolic rate in planarians. *Elife*, *8*, e38187. (Publisher: eLife Sciences Publications, Ltd)

Timko, M. P., Kausch, A. P., Castresana, C., Fassler, J., Herrera-Estrella, L., Van den Broeck, G., . . . Cashmore, A. R. (1985). Light regulation of plant gene expression by an upstream enhancer-like element. *Nature*, *318*(6046), 579–582. (Publisher: Nature Publishing Group UK London)

Toby, G. G., Gherraby, W., Coleman, T. R., & Golemis, E. A. (2003). A novel RING finger protein, human enhancer of invasion 10, alters mitotic progression through regulation of cyclin B levels. *Molecular and cellular biology*, *23*(6), 2109–2122. (Publisher: Taylor & Francis)

Towbin, H., Staehelin, T., & Gordon, J. (1979). Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proceedings of the national academy of sciences*, *76*(9), 4350–4354. (Publisher: National Acad Sciences)

Ubeda, M., Vallejo, M., & Habener, J. F. (1999). CHOP enhancement of gene transcription by interactions with Jun/Fos AP-1 complex proteins. *Molecular and cellular biology*, *19*(11), 7589–7599. (Publisher: Taylor & Francis)

Vallejo, M., Ron, D., Miller, C. P., & Habener, J. F. (1993). C/ATF, a member of the activating transcription factor family of DNA-binding proteins, dimerizes with CAAT/enhancer-binding proteins and directs their binding to cAMP response elements. *Proceedings of the National Academy of Sciences*, *90*(10), 4679–4683. (Publisher: National Acad Sciences)

Valouev, A., Johnson, S. M., Boyd, S. D., Smith, C. L., Fire, A. Z., & Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. *Nature*, *474*(7352), 516–520. (Publisher: Nature Publishing Group UK London)

Van Bortle, K., Ramos, E., Takenaka, N., Yang, J., Wahi, J. E., & Corces, V. G. (2012). Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. *Genome research*, *22*(11), 2176–2187. (Publisher: Cold Spring Harbor Lab)

Van Nostrand, E. L., & Kim, S. K. (2013). Integrative analysis of C. elegans modENCODE ChIP-seq data sets to infer gene regulatory interactions. *Genome Research*, *23*(6), 941–953. (Publisher: Cold Spring Harbor Lab)

Vermeulen, M., Mulder, K. W., Denissov, S., Pijnappel, W., van Schaik, F. M., Varier, R. A., … Timmers, H. (2007, October). Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4. *Cell*, *131*(1), 58–69. Retrieved 2023-10-13, from https://doi.org/10.1016/j.cell.2007.08.016 (Publisher: Elsevier) doi: 10.1016/j.cell.2007.08.016

Vila-Farré, M., & C Rink, J. (2018). The ecology of freshwater planarians. *Planarian* *regeneration: Methods and protocols*, 173–205. (Publisher: Springer)

Vila-Farré, M., Rozanski, A., Ivanković, M., Cleland, J., Brand, J. N., Thalen, F., ... Vu, H. T.-K. (2023). Evolutionary dynamics of whole-body regeneration across planarian flatworms. *Nature Ecology & Evolution*, 1–17. (Publisher: Nature Publishing Group UK London)

Vlieghe, D., Sandelin, A., De Bleser, P. J., Vleminckx, K., Wasserman, W. W., Van Roy, F., & Lenhard, B. (2006). A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic acids research*, *34*(suppl_1), D95–D97. (Publisher: Oxford University Press)

Vogg, M. C., Owlarn, S., Rico, Y. A. P., Xie, J., Suzuki, Y., Gentile, L., ... Bartscherer, K. (2014). Stem cell-dependent formation of a functional anterior regeneration pole in planarians requires zic and forkhead transcription factors. *Developmental Biology*, *390*(2), 136–148.

Voog, J., Sandall, S. L., Hime, G. R., Resende, L. P. F., Loza-Coll, M., Aslanian, A., ... Jones, D. L. (2014). Escargot restricts niche cell to stem cell conversion in the Drosophila testis. *Cell reports*, *7*(3), 722–734. (Publisher: Elsevier)

Wagner, D. E., Wang, I. E., & Reddien, P. W. (2011). Clonogenic neoblasts are pluripotent adult stem cells that underlie planarian regeneration. *Science*, *332*(6031), 811–816. (Publisher: American Association for the Advancement of Science)

Wang, J., Chen, R., & Collins III, J. J. (2019). Systematically improved in vitro culture conditions reveal new insights into the reproductive biology of the human parasite Schistosoma mansoni. *PLoS biology*, *17*(5), e3000254. (Publisher: Public Library of Science San Francisco, CA USA)

Wang, R.-S., Yeh, S., Tzeng, C.-R., & Chang, C. (2009, April). Androgen Receptor Roles in Spermatogenesis and Fertility: Lessons from Testicular Cell-Specific Androgen Receptor Knockout Mice. *Endocrine Reviews*, *30*(2), 119–132. Retrieved 2023-10-31, from https://doi.org/10.1210/er.2008-0025 doi: 10.1210/er.2008-0025

Wang, S. S., Tsai, R. Y., & Reed, R. R. (1997). The characterization of the Olf-1/EBF-like HLH transcription factor family: implications in olfactory gene regulation and neuronal development. *Journal of Neuroscience*, *17*(11), 4149–4158. (Publisher: Soc Neuroscience)

Wang, W., Hu, C.-K., Zeng, A., Alegre, D., Hu, D., Gotting, K., ... Schnittker, R. (2020). Changes in regeneration-responsive enhancers shape regenerative capacities in vertebrates. *Science*, *369*(6508), eaaz3090. (Publisher: American Association for the Advancement of Science)

Wang, Y., Stary, J. M., Wilhelm, J. E., & Newmark, P. A. (2010). A functional genomic screen in planarians identifies novel regulators of germ cell development. *Genes & development*, *24*(18), 2081–2092. (Publisher: Cold Spring Harbor Lab)

Wang, Y., Zayas, R. M., Guo, T., & Newmark, P. A. (2007). Nanos function is essential for development and regeneration of planarian germ cells. *Proceedings of the National Academy of Sciences*, *104*(14), 5901–5906. (Publisher: National Acad Sciences)

Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., ... Zhang, M. Q. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature genetics*, *40*(7), 897–903. (Publisher: Nature Publishing Group US New York)

Ward, J. O., Reinholdt, L. G., Motley, W. W., Niswander, L. M., Deacon, D. C., Griffin, L. B., ... O'Brien, M. J. (2007). Mutation in mouse hei10, an e3 ubiquitin ligase, disrupts meiotic crossing over. *PLoS genetics*, *3*(8), e139. (Publisher: Public Library of Science San Francisco, USA)

Wenemoser, D., & Reddien, P. W. (2010). Planarian regeneration involves distinct stem cell responses to wounds and tissue absence. *Developmental biology*, *344*(2), 979–991. (Publisher: Elsevier)

Wickham, H. (2011). ggplot2. *Wiley interdisciplinary reviews: computational statistics*, *3*(2), 180–185. (Publisher: Wiley Online Library)

Widnell, C., & Tata, J. (1964). A procedure for the isolation of enzymically active rat-liver nuclei. *Biochemical Journal*, *92*(2), 313. (Publisher: Portland Press Ltd)

Wierstra, I. (2008). Sp1: emerging roles-beyond constitutive activation of TATA-less housekeeping genes. *Biochem Biophys Res Commun*, *372*, 1–13.

Wingender, E., Schoeps, T., Haubrock, M., Krull, M., & Dönitz, J. (2018). TFClass: expanding the classification of human transcription factors to their mammalian orthologs. *Nucleic acids research*, *46*(D1), D343–D347. (Publisher: Oxford University Press)

Witchley, J. N., Mayer, M., Wagner, D. E., Owen, J. H., & Reddien, P. W. (2013). Muscle cells provide instructions for planarian regeneration. *Cell reports*, *4*(4), 633–641. (Publisher: Elsevier)

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome biology*, *20*, 1–13. (Publisher: Springer)

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., ... Zhan, L. (2021). clusterProfiler

4.0: A universal enrichment tool for interpreting omics data. *The innovation*, *2*(3). (Publisher: Elsevier)

Wu, W.-S., & Lai, F.-J. (2015). Functional redundancy of transcription factors explains why most binding targets of a transcription factor are not affected when the transcription factor is knocked out. *BMC systems biology*, *9*(6), 1–9. (Publisher: BioMed Central)

Wu, X., & Brewer, G. (2012). The regulation of mRNA stability in mammalian cells: 2.0. *Gene*, *500*(1), 10–21. (Publisher: Elsevier)

Wudarski, J., Egger, B., Ramm, S. A., Schärer, L., Ladurner, P., Zadesenets, K. S., ... Berezikov, E. (2020). The free-living flatworm macrostomum lignano. *EvoDevo*, *11*, 1–8.

Xing, Y., Johnson, C. V., Moen Jr, P. T., McNeil, J. A., & Lawrence, J. (1995). Nonrandom gene organization: structural arrangements of specific pre-mRNA transcription and splicing with SC-35 domains. *The Journal of cell biology*, *131*(6), 1635–1647.

Xu, J., & Liu, Y. (2021). Probing chromatin compaction and its epigenetic states in situ with single-molecule localization-based super-resolution microscopy. *Frontiers in Cell and Developmental Biology*, *9*, 653077. (Publisher: Frontiers Media SA)

Xu, Z., Wei, G., Chepelev, I., Zhao, K., & Felsenfeld, G. (2011, March). Mapping of INS promoter interactions reveals its role in long-range regulation of SYT8 transcription. *Nature Structural & Molecular Biology*, *18*(3), 372–378. Retrieved from https://doi.org/10.1038/nsmb.1993 doi: 10.1038/nsmb.1993

Yamaguchi, Y., Shibata, H., & Handa, H. (2013). Transcription elongation factors DSIF and NELF: promoter-proximal pausing and beyond. *Biochimica Et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, *1829*(1), 98–104. (Publisher: Elsevier)

Yamazaki, A., & Minokawa, T. (2016). Roles of hesC and gcm in echinoid larval mesenchyme cell development. *Development, Growth & Differentiation*, *58*(3), 315–326. (Publisher: Wiley Online Library)

Yang, C., Bolotin, E., Jiang, T., Sladek, F. M., & Martinez, E. (2007). Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, *389*(1), 52–65. (Publisher: Elsevier)

Yu, G., Wang, L.-G., & He, Q.-Y. (2015). ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, *31*(14), 2382–2383. (Publisher: Oxford University Press)

Yuan, G.-C., Liu, Y.-J., Dion, M. F., Slack, M. D., Wu, L. F., Altschuler, S. J., & Rando, O. J. (2005, July). Genome-Scale Identification of Nucleosome Positions in S. cerevisiae. *Science*, *309*(5734), 626–630. Retrieved 2023-10-10, from https://doi.org/10.1126/science.1112178 (Publisher: American Association for the Advancement of Science) doi: 10.1126/science.1112178

Zayas, R. M., Hernández, A., Habermann, B., Wang, Y., Stary, J. M., & Newmark, P. A. (2005). The planarian Schmidtea mediterranea as a model for epigenetic germ cell specification: analysis of ESTs from the hermaphroditic strain. *Proceedings of the National Academy of Sciences*, *102*(51), 18491–18496. (Publisher: National Acad Sciences)

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.-W., Nechaev, S., Adelman, K., ... Young, R. A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nature genetics*, *39*(12), 1512–1516. (Publisher: Nature Publishing Group US New York)

Zeng, A., Li, H., Guo, L., Gao, X., McKinney, S., Wang, Y., ... Ross, E. (2018). Prospectively isolated tetraspanin+ neoblasts are adult pluripotent stem cells underlying planaria regeneration. *Cell*, *173*(7), 1593–1608. (Publisher: Elsevier)

Zhang, Q., Lenardo, M. J., & Baltimore, D. (2017). 30 years of nf-kb: a blossoming of relevance to human pathobiology. *Cell*, *168*(1), 37–57. (Publisher: Elsevier)

Zippo, A., Serafini, R., Rocchigiani, M., Pennacchini, S., Krepelova, A., & Oliviero, S. (2009, September). Histone Crosstalk between H3S10ph and H4K16ac Generates a Histone Code that Mediates Transcription Elongation. *Cell*, *138*(6), 1122–1136. Retrieved 2023-10-13, from https://doi.org/10.1016/j.cell.2009.07.031 (Publisher: Elsevier) doi: 10.1016/j.cell.2009.07.031