Investigating the Regulation of Early Development of the Amphipod Crustacean *Parhyale hawaiensis*

Salha Eissa Rawas

St Cross College

University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Trinity Term 2023

Declaration

I hereby declare that this thesis entitled “Investigating the regulation of early development of the amphipod crustacean *Parhyale hawaiensis*” has been originally carried out by me under the supervision of Professor Aziz Aboobaker. This work has not formed the basis for award of any degree or diploma previously. The particulars given in the thesis are true to the best of my knowledge.

Salha Eissa Rawas

Department of Biology
St Cross college
University of Oxford

Acknowledgements

First and foremost, I praise the Almighty God for blessing me with this precious opportunity and empowering me to progress successfully.

I would like to express my heartfelt gratitude to my supervisor, Professor Aziz Aboobaker, for his invaluable support and guidance throughout my time in his lab. His patience, encouragement, and, above all, the enriching scientific discussions have contributed significantly to my growth and development. I have learned a great deal from him and will forever cherish his guidance and support. Thank you for your profound impact on my journey in this field.

I would like to sincerely express my gratitude to Dr. Manuel Jara Espejo for his invaluable assistance in the field of bioinformatics and the analysis of transcriptomic data. I am immensely grateful for your unwavering generosity in always making time to provide me with your guidance and advice whenever I needed it. The enlightening scientific discussions we have had have profoundly enriched my knowledge and understanding.

I consider myself fortunate to have crossed paths with brilliant colleagues from around the world throughout my journey. I am grateful for their assistance, advice, and engaging scientific discussions.

I owe an immeasurable debt of gratitude to my parents, Dr. Eissa Rawas and Mrs. Feryal Al-Refi, whose unwavering love and support have been instrumental in bringing me to where I am today. Your belief in me has been the driving force behind my accomplishments. I am fully aware that without you, I would not have achieved what I have. Thank you from the bottom of my heart for everything you have done for me.

To my incredible little family -Hamed, Qusai, and Awab- who have been with me every step of the way, I cannot find words to express the depth of my gratitude and how fortunate I feel to have you by my side. Despite the challenges we faced as a family, your unconditional love and unwavering endurance have been the pillars that helped me overcome obstacles and maintain my sanity and happiness. Thank you for always being there for me in the truest sense of the word.

I want to extend special thanks to my siblings Mohammed, Ahmed, Doaa, Yousef, and Rania. Your endless support, comfort, and encouragement have been my guiding light during moments of disappointment and uncertainty. I am forever grateful to have you as my backbone. Thank you for being such an amazing siblings.

Thesis Abstract

The amphipod crustacean *Parhyale hawaiensis* is an ideal model organism for studying development, evolution, and regeneration. Its ease of rearing in the lab, accessibility of embryos at all developmental stages, large broods produced year-round, and the variety of functional experiments that can be conducted on animals and embryos make it a valuable resource for investigating many biological questions. As an out group to insects, *Parhyale* offers a platform for studying biological diversity through comparative studies. Recent research has identified a full repertoire of single-copy genes encoding the machinery associated with DNA methylation in the *Parhyale* genome. This discovery provides an opportunity to explore the role of DNA methylation in early embryogenesis using an invertebrate model system. In this thesis, we present *Parhyale hawaiensis* as a tractable model to study DNA methylation dynamics during embryonic development. Our aim was to investigate the regulation during early embryo development, describe the maternal-to-zygotic transition (MZT) and zygotic-genome-activation (ZGA) in *Parhyale*, and explore the potential function of DNA methylation during embryogenesis while examining its regulatory role in gene expression.

Parhyale possesses a large genome size of approximately 3.6 Gb, which was initially sequenced by Kao et al. in 2016. An update using Dovetail technology enhanced the assembly by generating large scaffolds but left significant gaps. Therefore, our objective was to further improve the existing assembly by integrating PacBio data to close the gaps and correct assembly errors. This effort proved successful, as over 70% of the gaps in the assembly were closed, suggesting further coverage would give further improvement. Subsequently, we performed an expression-driven annotation using a wide range of RNA-seq data from various embryonic stages and different adult conditions. The well-annotated genome served as the foundation for analyzing transcriptomic data from early embryonic stages to detect *de novo* zygotic transcription using intronic RNA signals.

We described the transcriptome of early *Parhyale* embryos and proposed a model for the maternal-to-zygotic-transition (MZT) and zygotic-genome-activation (ZGA) timelines. Our findings demonstrated that zygotic transcription begins as early as 11 hours post-fertilization and occurs in two waves. The minor wave of ZGA in *Parhyale* commences at the 32-cell stage, while the major wave takes place at the start of the blastodisc formation stage. We discovered that the earliest transcribed genes in *Parhyale* are typically short, intron-less or intron-poor, and newly evolved. Furthermore, we validated the presence of DNA methylation mediator genes, with a focus on DNMT1, DNMT3, and MBD2/3. We analyzed the expression pattern of DNA methylation machinery genes during the early stages of *Parhyale* embryogenesis and found that they are provided maternally. To explore the relationship between DNA methylation and gene expression, we

correlated our gene expression datasets with methylseq datasets. Our results revealed a positive correlation between gene-body methylation and gene expression levels.

Lastly, we conducted functional experiments targeting DNMT1, DNMT3, and MBD2/3 to understand their roles during embryonic development. We utilized CRISPR/Cas9 to generate knockout animals for each of the three genes, revealing that the loss of any of these genes is lethal to embryos. Additionally, we performed RNA interference (RNAi) knockdown on MBD2/3, which also resulted in an early embryonic lethality, confirming its essential role in embryogenesis. Profiling the transcriptome of knockdown embryos revealed that the knockdown of MBD2/3 altered the expression of many genes, including developmental transcription factors with low levels of gene-body methylation.

The work presented in this thesis offers a comprehensive understanding of the early embryonic development of *Parhyale* and emphasizes the crucial role of DNA methylation during embryogenesis. In future, we aim to investigate whether MBD2/3 regulates gene expression through its association with the NuRD complex, and if this occurs in a DNA methylation-dependent, independent, or both manners. Furthermore, the improved assembly and annotation presented in this thesis will greatly facilitate more precise analyses, enabling us to address intriguing questions regarding the promising model organism *Parhyale hawaiensis*.

Table of Contents

List of abbreviations:

gbM = gene-body methylation
RINP = Ratio of Intronic read counts to Not covered intronic Positions
TE = Transposable Elements
TF = Transcription Factor
MZT = Maternal-to-zygotic transition
ZGA = Zygotic genome activation
TPM = Transcripts per million
PCA = Principal component analysis
RBP = RNA binding protein
RT = Reverse transcriptase
DNMT = DNA methyltransferase
MBD = Methyl-binding domain
TET= Ten-eleven translocation
EM-seq = Enzymatic methyl-seq
CpG = Phosphate-linked cytosine-guanine pairs
mCpG = methylated CpG
CRISPR = Clustered regularly interspaced short palindromic repeats
5mc = addition of methyl group at the 5th position of cytosine
ESC = embryonic stem cell
NuRD = Nucleosome remodeling and deacetylation
NGS = Next generation sequencing
TGS = Third generation sequencing
CDS = coding sequence
LSGs = Lineage-specific genes

Chapter I

Introduction

Contents

Abstract

Parhyale hawaiensis, an amphipod crustacean, serves as a valuable model for the study of development and regeneration. This chapter provides a comprehensive introduction to *Parhyale* biology, consolidating various studies that have established *Parhyale* as the leading crustacean system currently available. It offers a wide range of transgenic tools and genomic resources, providing year-round access to embryos at all developmental stages. We review the factors involved in the regulation of early development, with a specific focus on DNA methylation as a potential epigenetic mechanism contributing to regulation of early embryogenesis in *Parhyale*. Additionally, this chapter provides an overview of a major process in the embryonic development of all metazoans —the maternal-to-zygotic-transition (MZT). We discuss studies that have investigated the potential role of DNA methylation during the MZT process. Lastly, we outline the objectives of this thesis and present a summary of all the chapters covered.

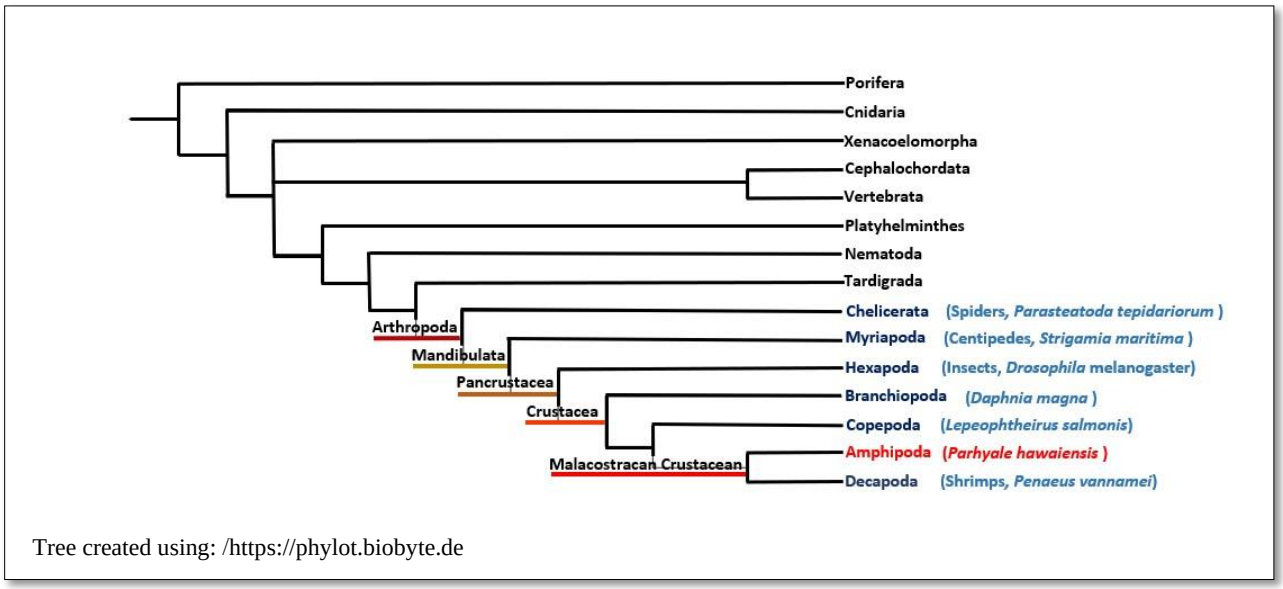
1.1 General *Parhyale hawaiensis* introduction

1.1.1 *Parhyale* phylogeny, taxonomy, and natural habitat

The phylum Arthropoda comprises one of the most diverse and highly abundant groups of animals. Arthropods have been extensively used as model organisms in various areas of biology, including aquaculture, disease research, pollination biology, and perhaps most importantly, genetics and evolution (Giribet and Edgecombe 2019). Within the Arthropoda phylum, the subphylum Crustacea encompasses a diverse array of organisms, such as decapods (e.g., crabs, lobsters, and shrimps) and copepods. Crustaceans occupy various habitats, with some found in freshwater or marine environments, while others are terrestrial, like woodlice, or parasitic, like fish lice. The class Malacostraca represents the largest class of crustaceans, consisting of approximately 30,000 species that inhabit oceans, inland water, even terrestrial habitats (Throp and Rogers 2011).

Within malacostraca, the marine Amphipod *Parhyale hawaiensis* belongs to the superorder Peracarida, which encompasses a significant number of species (Throp and Rogers 2011, Dana, 1853). *Parhyale* is classified under the talitrid superfamily, which includes various species known as scuds, beach hoppers, sand hoppers, or sandfleas (Paris et al., 2021). This marine crustacean, closely related to shrimps, crabs, and lobsters, can be found in tropical coastlines worldwide (Barnard and Karaman 1991, Lindeman 1991, Shoemaker 1956, Barnard 1965). *Parhyale* inhabits intertidal and shallow-water habitats such as mangrove leaves, bays, and estuaries, and it can exist as a species complex (Myers 1985, Barnard 1965, Shoemaker 1956, Poovachiranon et al. 1986). It thrives in rocky and macroalgal fauna aquaria (Barnard 1965). *Parhyale* exhibits a wide range of salinity and temperature tolerance and is known to form large populations, with up to 7,000 individuals per square meter (Poovachiranon et al. 1986). As a detritus feeder, *Parhyale* plays a significant role in ecosystems, particularly in areas such as mangrove forests (Paris et al. 2021, Poovachiranon et al. 1986).

Figure 1.1 Phylogenetic tree illustrating the placement of *Parhyale* within the animal kingdom, as well as its proximity to other major model organisms within arthropod clades, including spiders, centipedes, insects, branchiopods, copepods, amphipods, and decapods.



Tree created using: /https://phylot.biobyte.de

1.1.2 Lifecycle and embryogenesis of *Parhyale*

Parhyale exhibits sexual dimorphism, with mature males differing significantly different from females. Males are larger in size and possess enlarged claws, known as clappers, on their third thoracic appendages (T3, gnathopods) (Rehm et al., 2009) (Figure 1.2). They use these structures to grasp onto females during mating, forming a position known as amplexus (Paris et al., 2021). Females, on the other hand, have specialized oosteogites (visible ovaries) formed on their thoracic appendages (T3-T6) (Browne et al., 2005, Rehm et al., 2009). Mating pairs maintain this position until the female molts, at which point the male deposits sperm into the female's paired oviduct and releases her (Browne et al., 2005). Subsequently, the female begins to release eggs into the newly formed brood pouch or marsupium, where they become fertilized upon release (Paris et al., 2021, Browne et al., 2005). Adult females of *Parhyale* are capable of producing a substantial brood of embryos, with up to 30 eggs. They lay eggs throughout the year, and the embryos can be safely collected from the ventral brood pouch of the mother without causing harm to either of them (Rehm et al. 2009, Kao et al. 2016).

Embryonic development in *Parhyale* is direct, without involving larval stages (Paris et al., 2021). It has been thoroughly described as a series of morphologically distinguishable stages (Browne et al., 2005). The entire process of embryogenesis is completed within approximately 250 hours at a temperature of 26 ºC and is divided into 30 stages. After around 10 days, the embryos hatch into juvenile forms that resemble miniature versions of adult *Parhyale* (Browne et al., 2005) (Figure 1.3).

Embryonic development in *Parhyale* begins with mediolateral holoblastic (total) invariant cleavage. The first division occurs at 4 hours of development, resulting in a 2-cell embryo (Browne et al., 2005). Two additional divisions lead to the formation of an 8-cell embryo (stage 4, S4), which exhibits two distinct regions: macromeres (large cells) and micromeres (small cells) (Browne et al., 2005). The 8-cell stage is significant in *Parhyale* embryogenesis as each blastomere becomes committed to one of the three germ layers or the germ line (Gerberding et al., 2002). By 12 hours, the embryo consists of approximately 100 cells of equal size, giving it a "soccer-ball" appearance (Browne et al., 2005).

After this stage, major cell migration events occur, coinciding with early gastrulation. The cytoplasm of cells separates from the yolk, and cells begin migrating toward the embryo's surface, forming two predominant clusters. The first cluster emerges on the presumptive ventral side of the embryo and includes right, left, and posterior ectoderm macromere descendant cells, which contribute to the development of the ectoderm analgen (Browne et al., 2005, Sun and Patel 2019). The second cluster, referred to as the "rosette", forms through the aggregation of visceral mesoderm and germ descendant cells (Browne et al., 2005). Gastrulation takes place at around 20 hours, with cell aggregation commencing along the anterior ventral region of the embryo. The "rosette" cells migrate beneath the ectoderm analgen, marking the embryonic "germ disc" stage (Awes, Hinchen, & Extavour. 2009; Browne et al., 2005; Chaw & Patel, 1978). Following this stage, between 36 to 60 hours of development, germ disc cells continue to proliferate and recruit additional cells through migration from posterior or lateral positions. This process leads to formation of the germband and head lobes (Browne et al., 2005).

Parhyale undergoes successive molting throughout their lifetime, where they cast off their old cuticle and replace it with a new layer. Upon hatching, the hatchlings experience a molt and invert into a smaller size. However, as they grow and mature, their body size increases from approximately 1 mm to up to 10 mm or even larger in adulthood (Browne et al., 2005).

These animals take around 6-7 weeks at a temperature of 26 ^{o}C to reach sexual maturation. They have the ability to reproduce throughout the year, which makes studies on their embryos easily accessible and available for investigation.

Figure 1.2. Sexual dimorphism in *Parhyale*, showing an adult male and female. The male is larger in size compared to the female and exhibits distinctive features such as an enlarged claw-like structure. This characteristic is used by males to hold onto females during mating, forming a position known as amplexus. On the other hand, adult females possess a ventral brood pouch, which serves as a protective space to house and nurture the developing embryos.

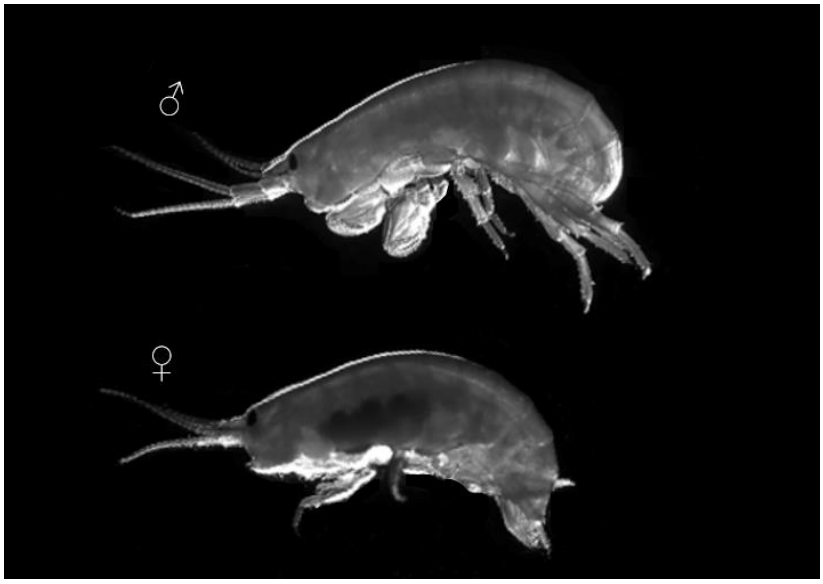
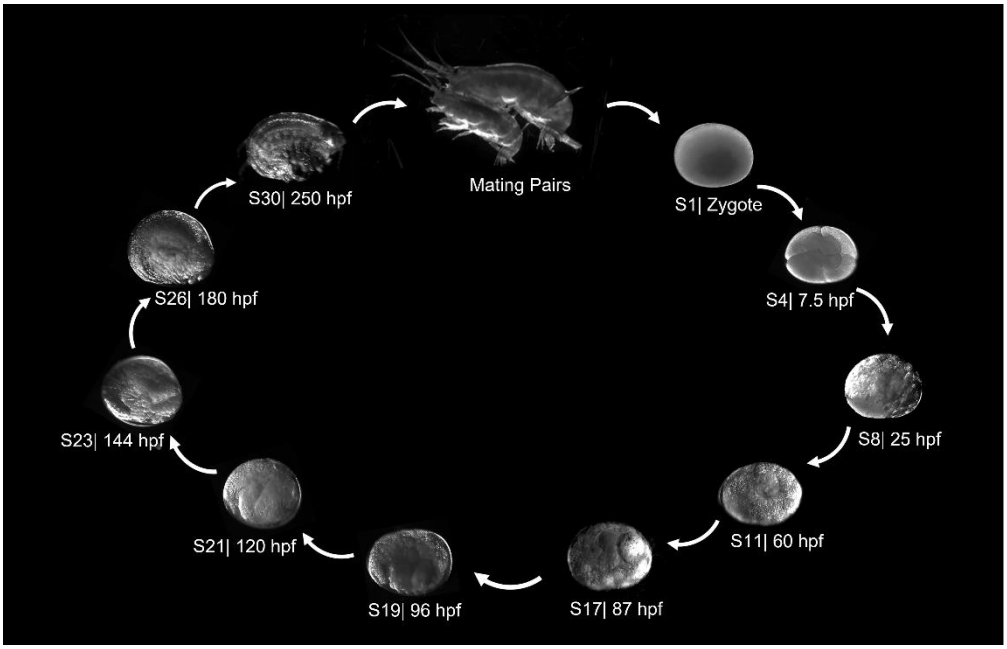


Figure 1.3. Life cycle of *Parhyale hawaiensis*. The embryogenesis of *Parhyale* spans approximately 10 days. After fertilization, the female *Parhyale* keeps the developing embryos within her ventral brooding pouch until they are ready to hatch.



1.1.3 *Parhyale* as a model for developmental and regeneration studies

In recent years, the amphipod crustacean *Parhyale hawaeinsis* has garnered attention from biologists as a valuable model system for investigating development and regeneration (Konstantinides and Averof 2014, Benton et al., 2014, Pavlopoulos et al. 2009, Martin et al. 2015, Chaw and Patel 2012, Gerberding et al. 2002, Stamataki et al. 2016, Price et al. 2010; Sinigaglia et al., 2022). This species possesses several characteristics that contribute to its suitability as a model organism. Unlike research involving vertebrates, there are no restrictions on using arthropods in experimental studies. Additionally, the small size of these animals makes them easily manageable and maintainable in laboratory settings.

In 2014, *Parhyale hawaiensis* was established as a model for limb regeneration. Adults of this species have the remarkable ability to fully regenerate their appendages. Regenerated limb can be seen encapsulated within the exoskeleton of the previously amputated leg. Within 5 to 8 days, new functional regenerated limb is released from the old cuticle after molting (Alwes et al. 2016, Konstantinides and Averof 2014, Grillo et al. 2016). One unique aspect of regeneration in *Parhyale*, which sets it apart from other well-known non-arthropod invertebrate regeneration models such as planarians and cnidarians, is its resemblance to the regeneration process observed in vertebrates (Figure 1.4). This similarity is evident in the presence of lineage-committed progenitors, where ectodermal and mesodermal cells arise from distinct progenitors localized within the amputated limb. In contrast, planarians rely on a common pool of pluripotent stem cells for regeneration (Konstantinides and Averof 2014, Kao et al. 2016, Alwes et al. 2016).

In *Parhyale,* the process of limb regeneration begins shortly after amputation (summarized in figure 1.5). Within a few minutes, wound closure initiates through the adhesion of hemocytes at the amputation site. By the first day, a melanized scab forms to cover the wound surface. Melanization serves as a typical wound response in arthropods, acting as part of the innate immune system to prevent hemolymph loss and pathogen invasion (Alwes et al. 2016, Bilandzija et al. 2017).

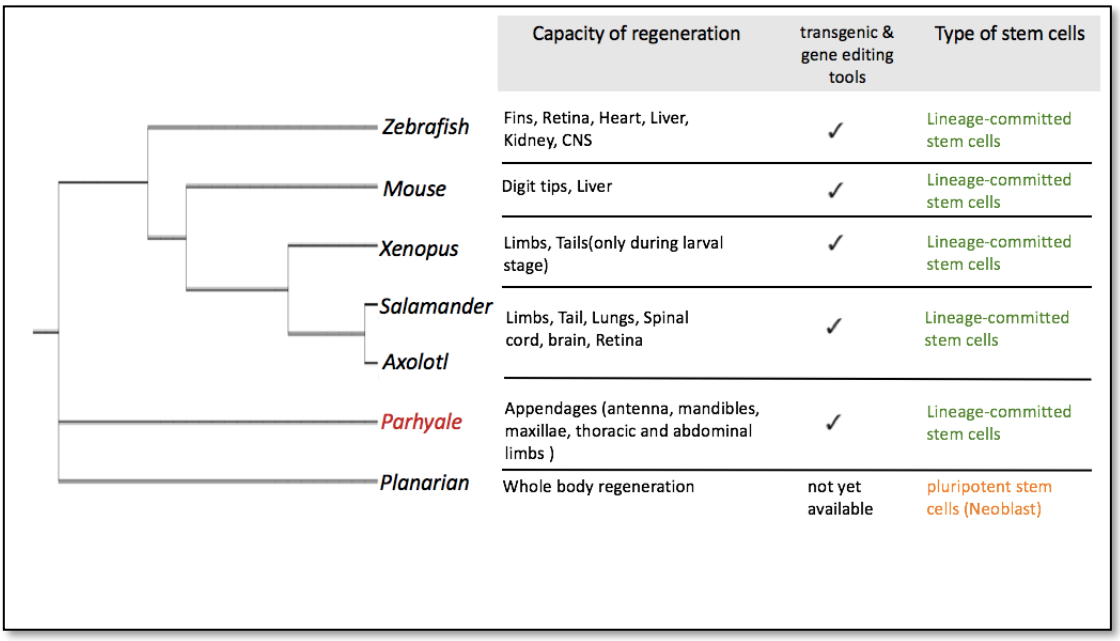
Within 1-2 days post amputation, the blastema starts to form as epidermis cells migrate underneath the

melanized scab, creating a layer of proliferating epithelial cells that act as a barrier between the scab and the inner tissues of the limb. During this stage, the migration of epidermal cells into the wound site occurs at a slow pace, known as the ‘quiescent period’. However, around 2-3 days after amputation, an abrupt transition from quiescence to extensive cell proliferation takes place, accompanied by cell movement and apoptosis.

This proliferation phase leads to the growth of a newly regenerated limb at the distal part of the stump, marking the third phase where actual regeneration occurs. Simultaneously, morphogenesis of the regenerated limb begins as the epidermal cells undergo rearrangement, ultimately resulting in the elongated, segmented structure of the leg. At the distal tip of the stump, a population of mesodermal progenitor cells is observed within the inner space of the wound stump. These progenitors give rise to regenerated muscles, and among them, a specific group called satellite-like cells contribute to the formation of regenerating muscle fibers. This phenomenon is reminiscent of satellite cells in vertebrates, which plays a crucial role in muscle progenitor function during regeneration and growth (Tanaka et al. 2016, Gargioli et al. 2004, Morrison et al. 2006, Alwes et al. 2016, Konstantinides and Averof 2014).

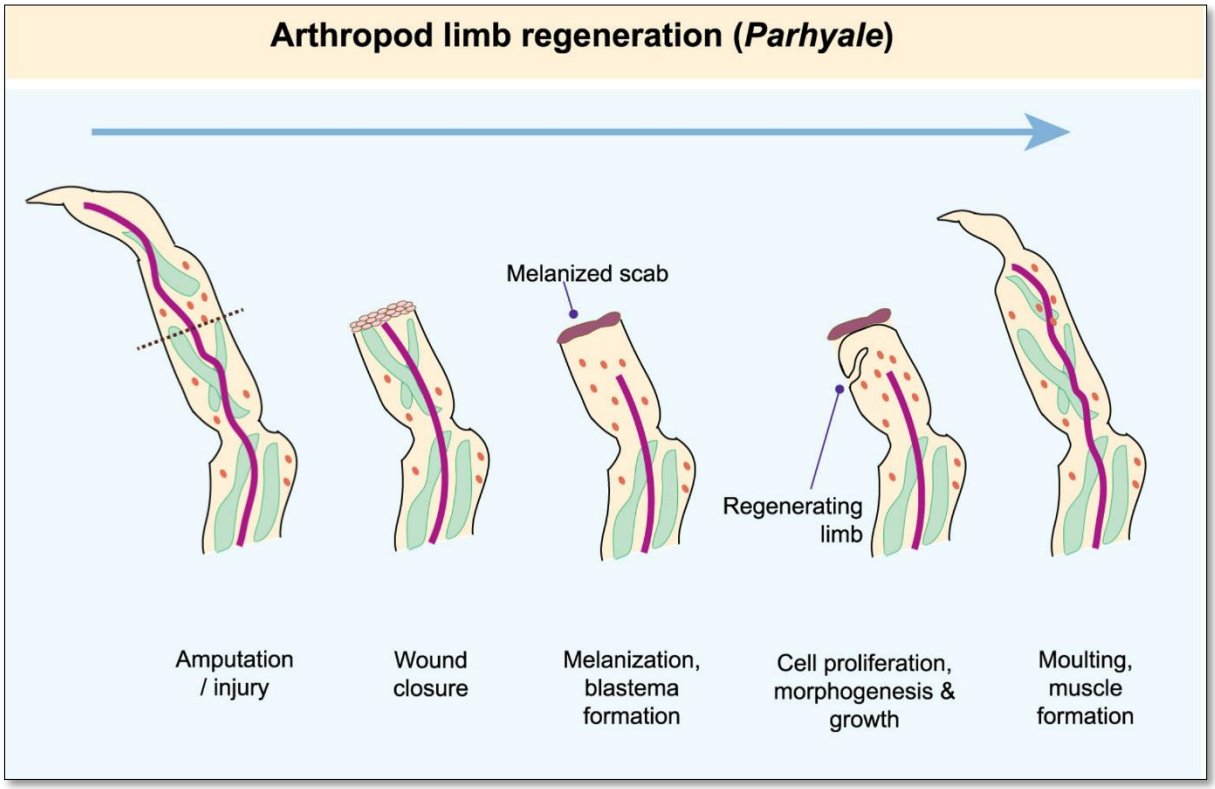
The question of whether these distinct lineage progenitors are a group of pre-existing self-renewing stem cells, akin to satellite cells, that maintain their undifferentiated state after embryogenesis within limb tissues, or if they emerge through de-differentiation, remains an open area of investigation. Furthermore, further research is needed to determine the nature and source of signals that induce the transition from the quiescent phase to more active regeneration.

Figure 1.4. Summary of key vertebrate regenerative models, with planarians and *Parhyale* serving as outgroups to emphasize the similarities between vertebrate regeneration and the regeneration observed in *Parhyale*.



	Capacity of regeneration	transgenic & gene editing tools	Type of stem cells
Zebrafish	Fins, Retina, Heart, Liver, Kidney, CNS	✓	Lineage-committed stem cells
Mouse	Digit tips, Liver	✓	Lineage-committed stem cells
Xenopus	Limbs, Tails(only during larval stage)	✓	Lineage-committed stem cells
Salamander / Axolotl	Limbs, Tail, Lungs, Spinal cord, brain, Retina	✓	Lineage-committed stem cells
Parhyale	Appendages (antenna, mandibles, maxillae, thoracic and abdominal limbs)	✓	Lineage-committed stem cells
Planarian	Whole body regeneration	not yet available	pluripotent stem cells (Neoblast)

Figure 1.5. An adapted figure from Lai and Aboobaker, 2018 that provides a Summary of the limb regeneration process in *Parhyale hawaiensis*. Upon injury or amputation, the regeneration in *Parhyale* begins with wound closure. This is achieved through the formation of a melanized scab around the limb epithelium at the amputation site. Subsequently, a layer of proliferating epithelial cells begins to aggregate beneath the melanized scab, culminating in the formation of the blastema. Nearby the distal limb stump, a population of mesodermal progenitor cells is observed. These cells give rise to regenerated muscles. A subsequent phase of rapid cell proliferation ensues, facilitating the growth of the newly regenerated limb within the limb stump. The process concludes when molting occurs, releasing the regenerated limb.



1.1.4 Functional studies and experimental tools in *Parhyale*:

Parhyale, including embryonic stages, has been extensively utilized in numerous studies involving experimental manipulations. As a result, a wide array of genetic tools and resources have now been established, solidifying *Parhyale's* position as a model organism (Sun & Patel, 2021; Paris et al., 2022).

Several **embryological experiments** have been conducted in *Parhyale* to investigate the roles of specific lineages in later developmental events. For instance, Vargas-Vila et al. (2010) explored the function of ventral midline and the associated gene *single-minded* (*ph-sim*) by using laser ablation to remove midline cells. Other studies used photoablation using FITC-Dextran injection into 4 or 8 cells stages, an entire lineage during gastrulation or at germband stage (Price et al., 2010; Hannibal et al., 2012a). Cell ablation was achieved through RNase and DNase injection (Chaw and Patel 2012) or manual ablation by puncturing a hole and removing the cell contents of the target stage (Alwes et al., 2011). Additionally, individual blastomeres were independently separated and isolated to investigate their developmental functions independently of neighboring cells (Extavour 2005).

In *Parhyale* embryos, *in-situ* hybridization and antibody staining techniques have been well-established. Some antibodies were specifically developed against *Parhyale* protein sequences (Rehm et al., 2009a; Rehm et al., 2009b). Hybridization chain reaction (HCR) *In-situ* methods have also been utilized to probe several transcripts (Bruce and Patel 2020, Bruce et al., 2021).

Loss-of function approaches have been employed in *Parhyale* using various techniques. One such approach is genome editing, specifically knockout and knock-in strategies facilitated via the clustered regularly interspaced short palindromic repeats (CRISPR/Cas9) system. This technique has been applied in *Parhyale* to investigate genes involved in limb patterning (Kao et al. 2016, Martin et al. 2016, Serano et al. 2016). More recently, CRISPR was used to knockout leg patterning genes, revealing the homology between the wing gene network in *Parhyale* and insects. This finding demonstrated that wings are not a novel

structure in insects, but rather they evolved from an ancestral structure (Clark-Hachtel and Tomoyasu 2020; Bruce and Patel 2020).

RNA interference (RNAi) knockdown has also been widely utilized in *Parhyale* studies. This technique involves the injection of Stealth siRNAs to suppress gene expression. Researchers have employed RNAi knockdown in *Parhyale* to investigate various genes, including those involved in limb development (Liubicich et al. 2009; Vargas-Vila et al., 2010; Nestorov et al., 2013). Additionally, RNAi has been performed using *Minos*-based transgene integration to generate a heat-inducible hairpin RNA targeting *abd*-A, enabling temporal control of the knockdown (Martin et al., 2016).

Furthermore, knockdown experiments in *Parhyale* have been conducted by injecting morpholinos to silence *vasa*, resulting in the loss of germ cells (Ozhan-Kizil and Gerberding 2009, Grillo et al. 2016; Nesotorov et al., 2013).

Transgenesis lines have been employed as a foundation for conducting conditional gene expression studies in *Parhyale*. Studies by Rehm et al. (2009a, 2009b), Pavlopoulos et al. (2009), Konstantinides and Averof (2014) and Ramos et al. (2019) have utilized transgenic lines to investigate gene expression patterns under specific conditions.

The versatile transposon *Minos* vectors have been used in *Parhyale* for transgenesis purposes. These vectors have been employed in studies conducted by Kontarakis and Pavlopoulos (2014) and Pavlopoulos and Averof (2005).

In addition, alternative approaches targeted to specific loci have been used to generate transgenesis lines in *Parhyale*. For instance, CRISPR knock-in techniques were employed by Kao et al. (2016) and Serano et al. (2016) to introduce transgenes at specific loci. Another approach involved the use of *phiC31* integrase, as demonstrated by Kontarakis et al. (2011) to produce transgenesis lines.

Live imaging techniques using fluorescent reporters have been employed in *Parhyale* studies. Hannibal et al. (2012a, 2012b), Chaw and Patel (2012), Price et al. (2010), Rehm et al. (2009c), Alwes et al. (2011),

and Price & Patel (2008) utilized live imaging with fluorescent markers to visualize and track specific cellular processes and events in *Parhyale*.

Microinjection of *Parhyale* at the 8-cell stage has been utilized to **label cells** originating from each blastomere. By employing **fluorescent markers,** researchers were able to visualize the contribution of each cell lineage to the regenerating limb (Gerberding et al. 2002, Pavlopoulos and Averof 2005, Konstantinides and Averof 2014, Grillo et al. 2016).

Moreover, the *Parhyale* heat-shock (*PhHS*) *cis*-regulatory element (CREs) derived from a *Parhyale hsp70* gene, cloned by Pavlopoulos et al. (2009), has been used to establish an inducible gene expression system (Alwes et al., 2016; Wolf et al., 2018). This allowed researchers to precisely control the expression of specific genes in response to heat induction.

Additionally, the expression of two types of photoreceptor cells in the *Parhyale* eye has been achieved using *PhOpsin1* and *PhOpsin2* CREs derived from *Parhyale opsin 1* and *opsin 2* genes. This approach was employed by Ramos et al. (2019) to investigate photoreceptor cell development.

1.1.5 Genome assembly and annotation of *Parhyale*

The process of building genomic resources for *Parhyale* began with the generation of bacterial artificial chromosome (BAC) libraries using *Parhyale* genomic DNA, providing approximately 5x coverage (Parchem et al., 2010). Subsequently, in 2016, the whole genome of *Parhyale* was sequenced and annotated for the first time using short-read high-throughput sequencing technology. The sequencing was conducted on a single male from the Chicago iso-Female inbred line, used for functional studies (Kao et al. 2016).

Parhyale's genome is relatively large, with a size of approximately 3.6 Gb, which is comparable to the size of human genome and larger than many other arthropod species, including *Drosophila melanogaster*. The genome exhibits high levels of heterozygosity and is repeat-rich, leading to an initial fragmented assembly. To improve the genome assembly, subsequent efforts utilized Dovetail Genomics by Chicago and Hi-C technologies (Burton et al., 2013; Putnam et al., 2016). These approaches resulted in a larger scaffold but

introduced gaps in the assembly. Integration of long-read sequencing data is expected to further enhance the quality of the assembly.

Additionally, for comprehensive analysis requiring accurate gene model information, functional annotation of the *Parhyale* transcriptome is still necessary. This process involves identifying and characterizing the various transcripts and their associated function in the genome.

Overall, the ongoing efforts to refine the *Parhyale* genome assembly and perform functional annotation will contribute to a more comprehensive understanding of the genetic and genomic aspects of *Parhyale* biology.

1.2Regulation of embryonic development and regeneration

1.2.1 Regulation of differentiation and fate specification

During the process of cell differentiation, stem cells undergo fate determination to give rise to distinct cell types. This crucial process is controlled by complex networks of gene activity. Cell differentiation plays a vital role in embryonic development and regeneration, as embryonic stem cells contribute to the formation of various body organs, while adult stem cells are involved in tissue homeostasis, repair, and regeneration upon injury. Specific gene expression programs are required for each cell type during cellular differentiation, and these programs are regulated by multiple gene regulatory mechanisms. Understanding the mechanisms that governs stem cell differentiation pathways is of great importance in biology and medicine. Although significant progress has been made in studying the mechanisms of differentiation in embryonic development and adulthood, there is still much to uncover to obtain a comprehensive understanding of the molecular pathways that regulate stem cell behaviour.

One of the major regulators of stem cell fate and lineage determination is cell-cell interaction. Small signalling molecules transferred between cells play a recognized role in cell differentiation (Noort et al., 2021; Chen et al., 2007). Cell-cell communication coordinates the activities of individual cells, ultimately leading to the formation of organs and complex organisms. Metabolic changes also contribute to the

regulation of cellular differentiation. For example, molecules like acetyl-CoA and α-ketoglutarate regulate histone methylation and acetylation during stem cell activation and differentiation (Ludikhuize and Colman, 2021; Tyurin-kuzmin et al., 2020; Diamante and Martello, 2022). The localization of particular RNAs is another key process during germ layer specification (Gavis and Lehmann, 1992; Holt et al., 2009). Additionally, epigenetic mechanisms are the major driver of cell-type-specific gene expression programs required for stem cell differentiation.

Several studies have demonstrated significant similarities between embryonic development and regeneration (Nacu et al., 2011; Khan et al., 2002; Satoh et al., 2007; Suzuki et al., 2015, 2007; Muneoka & Bryant, 1982). However, a recent study compared the transcriptional profiles of developing and regenerating legs in *Parhyale* to investigate whether limb regeneration mirrors limb development during embryogenesis (Sinigaglia et al., 2022). The researchers found that a large proportion of genes that exhibit dynamic expression during leg development are also present among the temporally variable genes during regeneration. Interestingly, the transcriptional dynamics of regeneration show stronger variation between individuals, which is correlated with the molting cycle of each individual. The study also highlights that the developmental environment is more stable compared to the regenerative environment, which occurs in complex and physiologically different adults in terms of nutrition, size, and the timing and extent of the injury. The authors accounted for this inter-individual variation and correlated the transcriptomic profiles of development and regeneration. While they observed gene overlap between the two datasets, the temporal order of gene deployment is not the same. In conclusion, these findings indicate that regeneration is not a simple recapitulation of development. However, both processes involve similar sets of regulatory genes and may be regulated by similar mechanisms, albeit in different temporal orders.

1.2.2 Epigenetic regulation of development and regeneration

Epigenetic regulation plays a critical role in controlling and maintaining gene expression during various biological processes, including growth, stem cell differentiation and plasticity, cell cycle regulation, and genome stability (Katsuyama and Paro 2011). It encompasses a range of biological processes that can modify gene expression and bring about heritable changes without altering the DNA sequence itself (Waddington et al., 1957). These processes include histone modifications, pre- and post-transcriptional regulation by small non-coding RNAs, and DNA methylation. The unique epigenetic landscape of a cell determines its gene expression program, which governs the cell's identity and biological function (Srinageshwar et al., 2016; Lunyak and Rosenfeld, 2008; Meissner, 2010).

Different epigenetic mechanisms perform crosstalk to regulate gene expression during cellular differentiation, rather than operating in isolation. For instance, some studies have suggested that DNA methylation can prevent the deposition of repressive histone marks, such as H3K27me3 (Brinkman et al., 2012; Wu et al., 2010; Atlasi & Stunnenberg, 2017). Conversely, loss of TET expression, which is involved in demethylation, can lead to increased DNA methylation levels and loss of H3K4me3 and H3K27me3 marks (Weber et al., 2007; Thomson et al., 2010; Okitsu & Hsieh, 2007). Furthermore, certain histone marks, like H3K9me2, have been found to be positively correlated with DNA methylation. In mouse embryonic stem cells, H3K9me2 is recognized by the DNA methylation component UHRF1, which facilitates the maintenance of DNA methylation after replication (Rothbart et al., 2012).

1.2.3 DNA methylation and its machinery

DNA methylation is a significant epigenetic modification, primarily represented in animals by the addition of methyl group at the 5th position of cytosine (5mC) in the CG sequence context. However, methylation at different positions on cytosine and adenine have also been documented (Jeltsch, 2006; Law and Jacobsen, 2010; Sarkies, 2022). This modification serves a conserved mark that is predominantly associated with epigenetic repression, as observed in genomic imprinting (Paulsen and Ferguson-Smith 2001), X chromosome inactivation (Cotton et al. 2015), and the repression of transposable element transcription

(Suzuki et al. 2007b; Nagamori et al. 2015; Su et al. 2012). Furthermore, DNA methylation plays a crucial role in embryonic development and contributes to the exposure of promotors to transcription factors in many eukaryotes (Tost 2009; Edwards et al. 2017; Jones P. A. 2012).

Patterns of DNA methylation also undergo changes during various diseases, including different types of cancers, diverse cell types and developmental stages, as well as individual variation (Bird, 2002; Jeltsch 2010; Lee et al. 2010; Tost 2009; Jones P. A. 2012). Methylation at CG-rich promoter sequence is often associated with the silencing of relevant genes, while unmethylated promoters are generally associated with active transcription (Brid, 2002). Conversely, when DNA methylation is detected in the gene-bodies of protein-coding genes, it correlates with active transcription. This can be explained by DNA methylation's role in regulating alternative splicing at these sites and preventing the initiation of spurious transcripts (Maunakea et al., 2010; Jones P.A. 2012; Yang et al., 2014).

DNA methylation exhibits significant variability across different animal species (Feng et al. 2010; Suzuki & Bird, 2008; Zemach et al., 2010). It evolves rapidly across eukaryotes, even between closely related species. The genome-wide distribution of DNA methylation shows considerable diversification and, surprisingly, is lost altogether in many lineages. The traditional view that vertebrates and invertebrates have distinct DNA methylation levels is being challenged as more phylogenetically distinct species are examined. Many profiled invertebrate species display dramatically lower levels of DNA methylation compared to vertebrates, which exhibit global DNA methylation distribution at CpG sites, except for CpG islands located at promoters. Methylation of these islands is associated with gene suppression (Feng et al., 2010; Jeltsch, 2010).

Furthermore, many invertebrates, including *Caenorhabditis elegans*, *D. melanogaster*, *Schmidtea mediterranea*, *Saccharomyces cerevisiae*, and *Schizosaccharomyces pombe*, lack detectable levels of DNA methylation (Zemach and Zilberman 2010; Feng et al. 2010; Lee et al. 2010; Jeltsch 2010; Rosic et al. 2018; Ponger and Li, 2005; Jurkowski & Jeltsch, 2011). However, early branching animals such as sponge species (*Amphimedon queenslandica* and *Sycon ciliatum*), the sea anemone *Nematostella vectensis*, and the comb

jelly *Mnemiopsis leidyi* do exhibit DNA methylation levels. Among them, *A. queesnlandica* displays a vertebrate-like methylome, with approximately 80% genome-wide methylation levels and variable promoter methylation. This suggests a conserved role for DNA methylation in regulating gene expression (de Mendoza et al., 2019a; Sarkies, 2022).

Numerous theories have been proposed to explain the patchy distribution of DNA methylation across the animal kingdom. Multicellular organisms tend to retain higher levels of DNA methylation, possibly to regulate the development of different cell types effectively and to suppress transposons more efficiently particularly due to sexual outcrossing (Zemach et al. 2010, Jeltsch 2010). Sexual outcrossing increases the aggressiveness of transposable elements (TE) because the transmission rate of TE is twice that of mendelian genes, and since TE represents a significant source of deleterious mutations. Consequently, sexual organisms are under strong selection pressure to suppress TEs (Bestor 1999, Arkhipova 2005).

The loss of methylation could be explained by the mutagenic effect of 5mC, which is repaired less efficiently compared to unmethylated cytosine. In almost all examined vertebrate species and insect species with a subset of highly methylated genes, the percentage of CG sites is consistently lower than expected (Bird, 1980; Lyko et al., 2010). Methylated cytosine is highly mutagenic, and a high mutation rate could contribute to the loss of DNA methylation in certain lineages (Waters & Swann, 1998; Svilar et al., 2011; Sarkies, 2022).

A recent study investigated the evolution of DNA methylation by comparing nematode species that lack this feature (*C. elegans*) with other nematodes that retain DNA methylation, such as *Romanomermis culicivorax* and *Plectus sambesii* (Rosic et al. 2018). The researchers found that DNMT activity introduces 3-methylcytosine (3mC) lesions as an off-target activity, resulting in alkylation DNA damage. This damage is repaired by ALKB2/3 enzyme, a member of Fe+ -dependent oxygenase DNA repair enzyme family. Enrichment analysis demonstrated that ALKB coevolved with DNMTs. Further analysis revealed that species retaining ALKB2/3 have median levels of 5mC that are ten times higher than species that have lost

ALKB2/3. This suggests that the mutagenic cost associated with DNA methylation could be a compelling explanation for why DNA methylation is lost in many animal lineages (Rosic et al. 2018).

DNA methyltransferases (DNMTs) are a conserved family of enzymes that catalyze the addition of methyl group to the DNA. These enzymes are highly expressed in undifferentiated cells, such as embryonic stem cells (ESC), and are subsequently downregulated in differentiated cells (Tomazou & Meissner, 2010; Jackson et al., 2004). Disruption of DNMTs in human and mice ESC lines has demonstrated the vital role of DNA methylation in embryonic development (Okano et al. 1999, Li et al. 1992, Liao et al. 2015, Tsumura et al. 2006). In mammals, DNA methylation patterns are established by the *de novo* methyltransferase family (DNMT3), which introduces methylation into both strands of previously unmethylated DNA (Jeltsch, 2006; Law & Jacobsen, 2010). DNMT1 is a methyltransferase with a high affinity for hemi-methylated DNA. It is responsible for copying cell-type specific methylation patterns into the new DNA strand during cell division (Holliday, 2006; Law & Jacobsen, 2010). This mechanism of maintaining DNA methylation by DNMT1 is conserved in plants and fungi, suggesting an early evolution in eukaryotic (Sarkies, 2022; Jurkowski & Jeltsch, 2011). Additional DNMTs such as DNMT5, DNMT4 and DNMT6 are found in some fungi, protists, and algae species, but their functions have not yet been determined (Sarkies, 2022; Huff & Zilberman, 2014; Ponger & Li, 2005). Lastly, DNMT2 is a group of enzymes that share sequence similarity with mammalian and bacterial DNMTs, methylate transfer RNA (tRNA) instead of DNA in many species (Jeltsch et al., 2017). DNMT1 knockout resulted in immediate lethality in human ESC (Liao et al. 2015), while DNMT3a knockout mice died a few weeks after birth (Okano et al. 1999, Li et al. 1992, Liao et al. 215). Complete knockout of DNA methyltransferases in mouse ESC lines (DNMT1-/-, DNMT3a-/-, DNMT3b-/-) demonstrated that normal growth and chromosome number can be maintained in the absence of CpG methylation, but differentiation is impaired (Tsumura et al. 2006). This highlights the important role of DNA methylation in differentiation and cell fate determination.

The ten eleven translocation (TET) enzyme family is responsible for the removal of DNA methylation through a stepwise oxidation process, resulting in the formation of 5-methylhydroxycytosine (5hmc), 5-

formylcytosine (5fC) and 5-carboxylcytosine (5caC) (He et al., 2011; Ito et al., 2011). Methyl DNA-binding proteins (MBDs), a conserved family of proteins, are known to 'read' DNA methylation patterns and convert them into relevant functional states. In vertebrates, there are 11 members of MBD-containing protein family, but six members are well-known: MeCP2, MBD1, MBD2, MBD3, and MBD4 (Hendrich & Bird, 1998; Jaenisch & Bird, 2003). All of these proteins are believed to be involved in transcriptional repression, except for MBD4, which possesses DNA N-glycosylase enzymatic activity and functions in DNA damage repair. MBD4 is involved in base excision repair, specifically correcting G/T mismatches resulting from the deamination of 5-methylcytosine (Bird & Wolffe, 1999; Hendrich & Tweedie, 2003; Wade 2001; Hendrich et al., 1999). Additionally, six more MBD proteins have been found in the mammalian genome: MBD5, MBD6, BAZ2A (TIP5) and BAZ2B; SETDB1, and SETDB2. More members of the MBD family are found in plant species (Hendrich & Tweedie, 2003; Roloff et al., 2003; Grafi et al., 2007). In contrast, in invertebrates, only one (or sometimes two) MBD protein is found.

MBD proteins play a critical role in coordinating and mediating the interplay between DNA methylation, histone modifications, and chromatin organization to regulate the gene transcription program (Du et al., 2015). MBD2 and MBD3 are paralog proteins and exhibit high sequence similarity, in addition to sharing the MBD motif (Hendrich & Bird, 1998). MBD2 is capable of binding to methylated DNA, while MBD3 cannot, except for *Xenopus laevis* MBD3, which carries a mutation in its motif (Iwano et al., 2004; Saito & Ishikawa, 2002). MBD2 and MBD3 are part of the nucleosome remodeling and deacetylation (NuRD) complex, independently mediating transcriptional repression and contributing to the epigenetic regulation of cell fate (Le Guezennec et al., 2006; Hendrich et al., 2001; Xue et al., 1998; Zhang et al., 1999; Denslow & Wade, 2007). Knocking out MBD3 was lethal for mice, while MBD2 knockout mice were viable and showed mild phenotypes (Hendrich et al., 2001).

Several invertebrate models, such as Drosophila, *C. elegans*, and *S. mediterranea*, possess a single copy of MBD gene known as MBD2/3. It exhibits orthology to both MBD2 and MBD3 and is considered the ancestral form of the gene (Roder et al., 2000; Gutierrez & Sommer, 2004; Jaber-Hijazi et al., 2013;

Hendrich & Tweedie, 2003). These species possess the MBD2/3 gene despite lacking a DNA methylation system. In Drosophila, it has been shown that MBD2/3 continues to associate with the NuRD silencing complex, although the same confirmation is yet to be made for *S. mediterranea* and *C. elegans* (Marhold et al., 2004; Dattani et al., 2019).

Arthropods, including honeybees and other species, exhibit highly variable levels of DNA methylation. Honeybees, one of the first arthropods found to have DNA methylation, possess a relatively low DNA methylation level, approximately 1% (Wang et al., 2006). In honeybee, only a few genes are highly methylated, while intergenic or repetitive elements have little methylation (Wang et al., 2006). Similarly, the burying beetle *Nicrophorus vespilloides* shows DNA methylation at only around 0.5% of CG sites (Cunningham et al., 2015).

Studies profiling DNA methylation levels across various insects and arthropod species have revealed a wide range of variability (Lewis et al., 2020; Bewick et al., 2017). For instance, the centipede *Spartina maritima* was found to have much higher DNA methylation levels, approximately 30% (de Mendoza et al., 2019b; Lewis et al., 2020). This indicates that DNA methylation levels vary significantly within arthropods.

The variation in DNA methylation levels across arthropods has been linked to the presence or absence of the alkylation repair enzyme (ALKB2). It has been observed that the conservation of DNMT1 and DNMT3, the enzymes responsible for DNA methylation, coincides with the presence of ALKB2 (Rosic et al., 2018). In species with a small subset of genes exhibiting methylation, the overall levels of 3mC would be lower, and therefore, the presence of ALKB2 would be unnecessary (Sarkies, 2022).

Figure 1.6. Illustration of DNA methylation writers, readers, and erasers. **(A).** DNA methyltransferase Enzymes: DNMT1 and DNMT3 are responsible for adding DNA methylation. DNMT1 acts as the maintenance DNA methyltransferase, copying the DNA methylation pattern to the newly synthesized DNA strand. On the other hand, DNMT3 functions as the *de novo* methylation enzyme, introducing DNA methylation to previously unmethylated DNA. **(B)** DNA methylation readers: These are families of proteins that recognize and bind to methylated DNA. The families include MBD proteins, UHRF proteins, and zinc-finger proteins. These proteins recognize and bind methylated DNA, subsequently recruiting modifying enzymes and leading to alterations in gene expression. **(C)** DNA methylation erasers: the TET proteins form a family of enzymes that demethylate or erase DNA methylation through multiple oxidization steps (detailed in Figure 1.7). Subsequently, the base excision repair process restores the cytosine to its unmodified form. It is worth noting that demethylation can also occur independently of enzymatic activity during cell division.

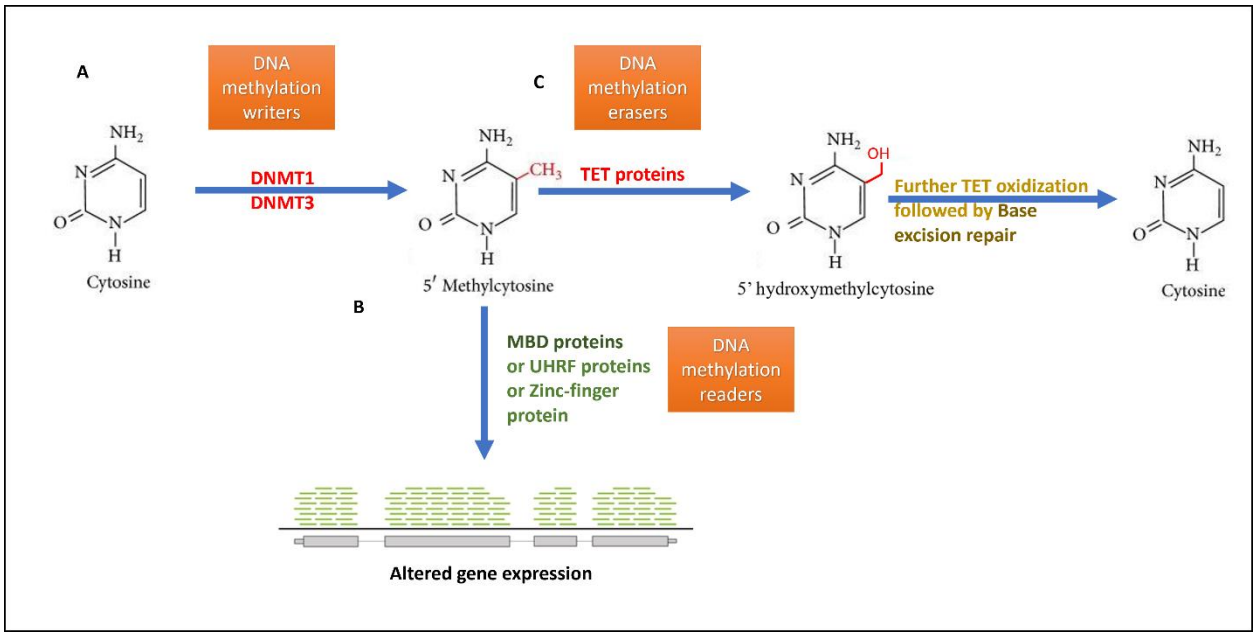
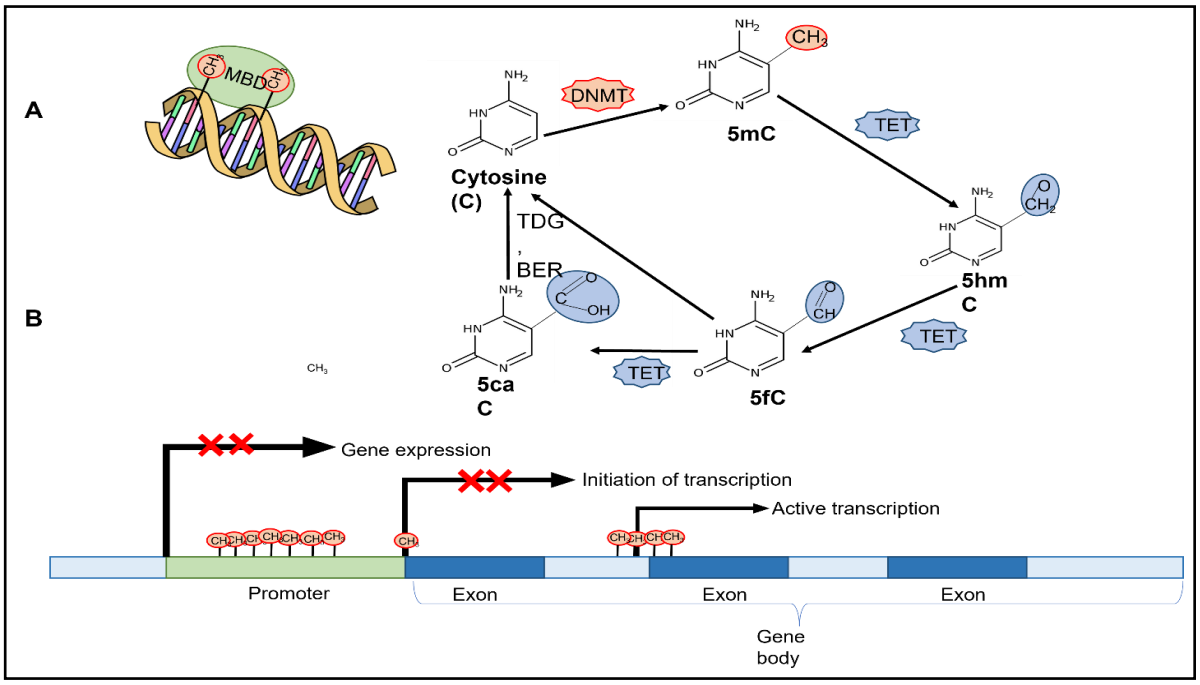


Figure 1.7. DNA methylation and demethylation. **(A)** Cytosine can be methylated by DNMT enzymes to give 5mC. MBD can recognize and bind 5mC, recruiting histone modifying enzymes to eventually give rise to inaccessible chromatin. TET enzymes can oxidize 5mC into 5hmC, and further oxidation yields 5fC and 5caC as a demethylation pathway. These products are then replaced by the thymine-DNA glycosylase enzyme (TDG) and base excision repair (BER) mechanisms to give rise to unmodified cytosine. **(B)** DNA methylation at regulatory regions, such as promoters and transcription start site, often results in silencing of gene expression, while TET activity in demethylation can potentially leads to transcriptional activation. Conversely, pretense of DNA methylation in the gene-body is often associated with active transcription.

Abbreviations: 5mC = 5-methylcytosine; 5hmC = 5-hydroxymethylcytosine; 5fC = 5-formylcytosine; 5caC = 5-carboxylcytosine; TDG = thymine-DNA glycosylase; BER = base excision repair; DNMT = DNA methyltransferase; MBD = methyl-CpG-binding domain protein; TET = ten-eleven translocation.



1.2.4 Maternal-to-zygotic transition and Zygotic-genome activation

When the differentiated sperm and egg unite, the resulting zygote undergoes reprogramming, transitioning into a totipotent state (Schulz and Harrison 2019). In the early stages of development, maternal products present in the egg play a crucial role in ensuring efficient reprogramming (Schulz and Harrison 2019). Consequently, metazoan early development is primarily regulated by proteins and mRNAs supplied by the mother, while the zygotic genome remains inactive. As embryogenesis progresses, the zygotic genome is eventually activated, often leading to the active elimination of a significant protein of maternal products. This transition in control is referred to as the maternal-to-zygote-transition (MZT), which occurs in all animals. However, the timing and dimensions of this transition vary among different organisms. In rapidly developing species such as worms, flies, fish, and frogs, the MZT takes place within the first few hours of development (Pálfy et al., 2017; Hamm and Harrison 2018). On the other hand, in more complex animals like mammalian preimplantation embryos, the MZT can take several days to complete (Jukam et al., 2017; Hamm and Harrison 2018) (Figure 1.8).

When the egg cell is produced, the cell cycle and metabolic activity are suspended. Following fertilization, external factors trigger the activation of the egg or zygote, bringing it out of its inactive state (Tadros and Lipshitz 2009). The fertilized oocyte, in addition to supplying half of the DNA content and nutrients, possesses the capacity to give rise to the remaining cells of the embryo. Hence, the maternal products deposited in the egg play crucial roles in various processes, including meiosis, egg development, fertilization, the initial mitotic divisions, cell fate specification, and implementation of biosynthetic processes (Tadros and Lipshitz 2009). Maternal mRNAs, including transcriptional activators, are among the deposited factors that promote zygotic transcription (Tadros and Lipshitz 2009). During MZT, two phases of genome regulation modes are recognized (Vastenhouw et al., 2019). Initially, there is post-transcriptional and post-translational regulation predominant at the beginning of development, followed by transcriptional regulation and the dominance of the zygotic genome (Vastenhouw et al., 2019). The first event of MZT involve the elimination of maternal products (Tadros and Lipshitz 2009). Maternal mRNAs

loaded into the oocytes of all animals comprise a significant fraction of the total protein-coding genes. For instance, in mouse and *C. elegans*, maternal transcripts account for one third of the transcriptome, while in *D. melanogaster* and zebrafish, they make up three-quarters of the transcriptome (Wang et al., 2004; Baugh et al., 2003; De Renzis et al., 2007 and Aanes et al., 2011).

After implementing fundamental molecular and cellular processes during initial developmental stages, the clearance of maternal products occurs in a phased manner, with the initial phase primarily directed by maternal transcripts, followed by later phase involving expression from zygotic transcriptome (Vastenhouw et al., 2019). The extent of maternal mRNA clearance during MZT varies among different organisms. For instance, in zebrafish, approximately one quarter of maternal mRNA is eliminated (Vastenhouw et al., 2019; Aanes et al., 2011; Bazzini et al., 2012; Mishima and Tomari 2016). In mice and *C. elegans*, around one third of maternal mRNA is cleared (Vastenhouw et al., 2019; Baugh et al., 2003; Hamatani et al., 2010), While in *D. melanogaster*, up to two-thirds of maternal mRNA is eliminated (Vastenhouw et al., 2019; Thomsen et al., 2010).

The second event of the MZT is Zygotic Genome Activation (ZGA), which occurs in both minor and major waves in most studied model organisms (Vastenhouw et al., 2019; Tadros and Lipshitz 2009). However, these waves are not entirely distinct, as transcriptional activation happens gradually over a period that varies among different taxa (Aanes et al., 2011; Collart et al., 2014; Harvey et al., 2013; Heyn et al., 2014; Lott et al., 2011; Owens et al., 2016; Pauli et al., 2012; Sandler and Stathopoulos, 2016; Tan et al., 2013; White et al., 2016).

The minor wave of ZGA typically occurs during the rapid cell cycle or cleavage divisions of the early embryo, where the G1 and G2 gap phases are absent, and no cell growth takes place (Schulz and Harrison 2019). On the other hand, the major wave of ZGA coincides with the first pause in the division cycle in many species (Schulz and Harrison 2019).

In humans, mice, and sea urchin, zygotic transcription is detected early, even at the one-cell stage (Vastenhouw et al., 2019; Abe et al., 2015; Aoki et al., 1997; Gildor and Ben-Tabou de-Leon, 2015; Hamatani et al., 2004; Materna et al., 2010; Yan et al., 2013). However, in other model organisms, zygotic transcription is detected later in subsequent cell cycles, such as the second cell cycle in *C. elegans*, the third cycle in frog, the sixth cycle in zebrafish, and around the eighth cycle in Drosophila (Vastenhouw et al., 2019).

During ZGA, a significant fraction of genes in the zygote are transcribed, although the specific set of activated genes varies among species (Heyn et al., 2014; Vastenhouw et al., 2019). The encoded proteins are often enriched in developmental regulators, transcription factors, and transcripts important for the degradation of maternal mRNAs (Collart et al., 2014; Lee et al., 2013; Bushati et al., 2008; Giraldez et al., 2006; Lund et al., 2009). Some of the genes that are transcribed during ZGA are also maternally loaded and re-expressed, reinforcing their expression (Vastenhouw et al., 2019; Lee et al., 2013). However, in animals like Drosophila and zebrafish, many genes encode different isoforms of mRNA by utilizing variable promoters, splicing, or polyadenylation sites between maternal and zygotic transcripts (Aanes et al., 2013; Atallah and Lott, 2018; Haberle et al., 2014).

The molecular mechanisms and developmental timing of each MZT event vary across species. Furthermore, MZT exhibit differences between somatic cells and germ cells in animals establish primordial germ cells (PGCs) during embryogenesis. For example, in *C. elegans*, zygotic genome activation (ZGA) commences at the four-cell stage in soma, while it remains silenced in PGCs due to global transcriptional repression (Seydoux et al., 1996).

Multiple layers of regulatory mechanisms underlie the precise gene expression program that handover control of embryogenesis during MZT. for example, the stabilization and degradation of maternal transcripts in the oocyte are controlled by mechanisms such as RNA-binding proteins (RBP), small non-coding RNAs, codon optimality, and DNA and RNA modifications, including methylation and histone modification (Vastenhouw et al., 2019).

In frog and zebrafish, the Y-box RBP binds to maternal transcripts to stabilize them and repress their translation (Bouvet and Wolffe, 1994; Matsumoto et al., 1996; Sun et al., 2018). In *C. elegans*, the dead-box helicase Dhh1, CGH-1 and in Drosophila, ME31B function in the stabilization and translational repression of maternal transcripts (Arnold et al., 2014; Boag et al., 2005; Wang et al., 2017). In Drosophila, two RBPs, *Smaug* and *BRAT*, along with *Pumilio*, coordinate the degradation of maternal mRNAs (Chen et al., 2014; Tadros et al., 2007). *Smaug* triggers the degradation of maternal mRNAs by recruiting the CCR4-NOT-deadenylase complex (Semotok et al., 2005). While BRAT directs the degradation of many maternal mRNAs through both maternal and zygotic pathways (Laver et al., 2015a, b). MicroRNAs (miRNAs) are the most well studied small non-coding RNAs that play critical roles in the decay of maternal products via the zygotic pathway in many species. For instance, the *miR-430* family in zebrafish is expressed during the minor wave of ZGA and contributes to translational repression and clearance of approximately 40% of maternal products through deadenylation (Bazzini et al., 2012; Giraldez et al., 2006).

Maternal product translation and post translational regulation are also important components of the MZT process. Poly(A) tail length is a major regulator of translation during transcriptionally silent stages (Subtelny et al., 2014). Cytoplasmic polyadenylation and deadenylation are key regulators of mRNA translation and stability in many species. For instance, the cytoplasmic polyadenylation element-binding protein (CPEB) family coordinates cytoplasmic polyadenylation in the early embryos of frogs and zebrafish, resulting in poly(A) tail lengthening and activation of translation (Collart et al., 2014; Winata et al., 2018). In Drosophila embryos, the Wispy (GLD2) cytoplasmic poly(A) polymerase directs the polyadenylation of maternal transcripts and has been shown to be essential for MZT (Benoit et al., 2008; Cui et al., 2013; Salles et al., 1994).

Indeed, various mechanisms regulate ZGA during MZT, including cell cycle length, transcriptional repressors and activators, and chromatin accessibility, which coordinate the transition from a transcriptionally silent to transcriptionally activated genome (Vastenhouw et al., 2019). Studies have shown that the speed of the cell cycle influences the length and number of transcripts produced in species such as

Drosophila and zebrafish (Edgar and Schubiger, 1986; Hadzhiev et al., 2019; Rothe et al., 1992; Dalle Nogare et al., 2009). Rapid cell cycles during the early stages of development result in expression of short, often intron-less or intron-poor transcripts during the minor wave of ZGA (Heyn et al., 2014; Kwasnieski et al., 2019; Rothe et al., 1992).

Transcriptional repressors and activators play crucial roles in orchestrating ZGA. Histones have been identified as transcriptional repressors that are highly expressed during early developmental stages in species such as frog, zebrafish, and Drosophila (Amodeo et al., 2015; Joseph et al., 2017; Shindo and Amodeo, 2019). They act as blocking agents, preventing transcriptional activators from binding to DNA (Workman and Kingston, 1998). The increase in DNA content and changes in the cytoplasm-to-nuclei ratio during ZGA in zebrafish lead to a decrease in nuclear histone concentration, allowing transcriptional activators to bind DNA (Joseph et al., 2017).

Examples of transcription factors involved in ZGA include *Zelda* and *Clamp* in Drosophila (Liang et al., 2008; Harrison et al., 2011; Duan et al., 2021). In Zebrafish, transcription factors such as *Pou5f3, Nanog*, and *Sox19b* are involved (Lee et al., 2013), while in humans, *Oct4* and *Dux4* are among the transcriptional activators (Gao et al., 2018). These transcription factors are deposited maternally and expressed during the early stages of development to promote activation of zygotic genome (Blythe and Wieschaus, 2016).

Figure 1.8. Overview of MZT events in multiple species across the animal kingdom. The table displays the variation between species in the timepoint and developmental stage of zygotic genome activation, the percentage of transcripts deposited as maternal genes, the number of genes detected at the minor and major waves of ZGA in each of the presented species. Additionally, the table compares the timepoint and stage of gastrulation initiation, which is typically prepared for during the major wave of ZGA. (Information in the figure collected from Vastenhouw et al., 2019 and Tadros and Lipshitz, 2009).

	Maternal transcripts (% of transcriptome)	Cell cycle of first zygotic transcription detected	Zygotically expressed genes (% of genes)	Gastrulation timepoint	Minor wave onset	Number of zygotic genes in minor wave	Major wave onset	Number of zygotic genes in major wave
C.elegans	68%	Cell cycle 2	~10%	~ 2.5 hpf (26-cell stage)	 0.8 hours	undefined	 1.6 hours	undefined
D.melanogaster	65%	Nuclear cycle 8	~35%	3 hpf (13 mitotic cycles)	 1 hour	59	 2 hours	1158
D.rario	45%	Cell cycle 6	~25%	5.25 hpf (50% epiboly stage)	 2.3 hours	125	 3.7 hours	1967
X.tropicalis	undefined	Cell cycle 3	~5%	~9 hpf (Stage 10.25)	 2 hours	278	 4 hours	1899
M.musculus	40%	Cell cycle 1	~20%	From day E6.26 to E9.5	 10 hours	50	 24 hours	1733

1.2.5 Role of DNA methylation during MZT

Gene expression is controlled at multiple molecular layers, including pre-transcriptional or post transcriptional regulation. Pre-transcriptional regulation involves DNA methylation, which plays a crucial role during development in many animals, including the process of MZT (Santos et al., 2002; Yartseva & Giraldez, 2015). The dynamics of DNA methylation have been studied in various vertebrate and invertebrate species, revealing species-specific differences. For example, in mice and humans, there is a global DNA demethylation event after fertilization, coinciding with the restoration of totipotency in the zygote, along with chromatin remodelling and transcriptional changes (Guo et al., 2014; Li et al., 2018; Peat et al., 2014; Santos et al., 2002; Shen et al., 2014; Eckersley-Maslin et al., 2018). Methylation levels are then re-established around the gastrulation stage (Eckersley-Maslin et al., 2018; Lee et al., 2014; Hackett et al., 2013). DNA demethylation can occur passively during DNA replication in the absence of DNMT enzymes or actively through mechanisms involving ten-eleven translocation (TET) enzymes or DNA repair pathways (Eckersley-Maslin et al., 2018). Notably, maternal and paternal genomes undergo DNA demethylation via different pathways (Howell et al., 2001; Smith et al., 2012; Wang et al., 2014; Eckersley-Maslin et al., 2018). Although the mechanisms of removing DNA methylation was investigated, the precise interplay between DNA methylation and other regulators and its role in reprogramming the zygote during early embryogenesis are still not fully understood. However, considering the MZT process, DNA methylation is likely to be involved in ZGA, as it affects gene expression at the pre-transcriptional level. Despite the presence of dynamic DNA methylation and chromatin landscapes during ZGA, it remains unclear whether changes in transcription are triggered by changes in DNA methylation levels and chromatin accessibility or if transcription itself leads to an open chromatin structure. One study proposed several models that potentially lead to the major wave of ZGA. In all of these models, changes in DNA methylation, along with DNA replication and factors present in the gametes, are identified as the main drivers of the minor wave of ZGA and subsequent activation of the major wave of zygotic genes (Eckersley-Maslin et al., 2018).

In zebrafish, unlike mammals, DNA methylation signatures inherited from the gametes are stable and not cleared after fertilization. The high levels of methylation inherited from the sperm in zebrafish facilitate the binding of transcription factors with DNA methylation affinity, such as *Pou5f3* and *Nanog,* which in turn prime the establishment of accessible chromatin during genome activation (Liu et al., 2018).

1.3Thesis aim and outline

The primary objective of this thesis was to investigate the potential function of DNA methylation in the Amphipod crustacean *Parhyale hawaiensis* and explore its regulatory role in gene expression. In *Parhyale*, global DNA methylation levels were found to be higher compared to other arthropods that retained DNA methylation, measuring at approximately 6-8%. As mentioned earlier, DNA methylation plays a crucial role in processes involving cellular differentiation, such as embryonic development and regeneration. Given *Parhyale's* amenability as a model organism, various functional techniques have been successfully applied to both its embryos and adults.

Based on these factors, I propose *Parhyale* as an intriguing regenerative invertebrate model for studying the role of DNA methylation in the context of development and regeneration. I hypothesize that DNA methylation will serve an essential function during *Parhyale's* embryonic development. To establish an informative foundation of *Parhyale's* early development, multiple approaches were employed.

Chapter I is divided into two parts. The first section provides a general introduction to *Parhyale hawaiensis* as an emerging model for developmental and regeneration studies. It highlights the significance of *Parhyale* as a tractable organism and its potential for investigating various biological processes. The second section offers an overview of regulation of cellular differentiation and fate specification, with a particular focus on the process of Maternal-to-Zygotic Transition (MZT). It delves into the mechanisms that regulate differentiation, including a detailed introduction to DNA methylation and previous studies exploring its role during MZT.

In **Chapter II,** we present our work on improving the genome assembly of *Parhyale hawaiensis*. This involved integrating long-read PacBio data to achieve a more contiguous assembly. Additionally, we generated an expression-based annotation of the updated assembly and introduced the DNA methylation machinery genes present in *Parhyale*.

The updated assembly and annotation generated in chapter II served as the basis for the work conducted in **Chapter III**. In this chapter, we performed a time course RNA sequencing of early embryonic stages in *Parhyale* to identify MZT and ZGA events. We also analyzed the expression patterns of DNA methylation mediator genes during these stages.

Chapter IV provides an outline of the results obtained from the functional experiments performed on *Parhyale* embryos. The focus of the experiments was on targeting the DNA methylation machinery genes responsible for installing DNA methylation. The goal was to gain insights into the role of DNA methylation during *Parhyale's* embryogenesis. The chapter begins by summarizing the outcomes of CRISPR/Cas9 knockout experiments for DNMT1, DNMT3 and MBD2/3 genes. It reveals that the absence of any of these genes led to embryonic lethality, as the embryos were found dead by the fourth day of development. The chapter also discusses the attempts made to perform a CRISPR/Cas9 knock-in by constructing specific DNA constructs and delivering them with CRISPR mix into one-cell embryos. Furthermore, an RNAi knockdown experiment targeting MBD2/3 is described, where RNA samples were collected from treated embryos to investigate the transcriptional response to the knockdown of MBD2/3.

In Chapter V we discuss the findings of this thesis and propose possible future directions to enhance our understanding of the role of DNA methylation in embryonic development and other intriguing processes like regeneration. **Chapter VI** provides a comprehensive description of the methods and materials employed throughout the project, covering all the presented results chapters.

Chapter II

Improving the genome assembly and genome annotation of *Parhyale hawaiensis*

Contents

Abstract

The amphipod crustacean *Parhyale hawaiensis* is a promising model system for studying development, regeneration, and other biological questions. Its genome size is large (approximately 3.6 Giga-base), it was first sequenced in 2016 (Kao et al., 2016). With a genome of this size, further improvements and sequencing are required to achieve a coherent and contiguous assembly that can facilitate various bioinformatic and experimental analyses.

The genome assembly has been updated using Dovetail technology (Putnam et al., 2016). This update upgraded the assembly into large scaffolds, some approaching chromosomal size. However, the assembly process has also led to several large gaps. Furthermore, the original genome annotation released in 2016 was based on relatively sparse and restricted RNA libraries from specific life-history stages. Additionally, the Dovetail updated assembly was not annotated.

Therefore, in this chapter, we utilized long-read PacBio sequencing to improve the assembly quality by closing the gaps and correcting assembly errors in the existing assembly. We also generated a new genome annotation using a wider range of RNA-seq data, including samples from various embryonic stages and different adult conditions. An accurately annotated genome is crucial for our subsequent results chapter, where we analyse RNA-seq data from early embryonic stages and utilize intronic RNA signals to detect nascent zygotic transcription.

Our aim was to produce a more reliable genome assembly for bioinformatic analysis. The existing assembly, lacking annotation and riddled with gaps, poses challenges for analyses requiring accurate intron annotation. Therefore, we hypothesized that using PacBio sequencing for the first time would fill the assembly gaps, leading to more accurate annotations suitable for detecting nascent transcription from intronic signals. Additionally, validating PacBio sequencing in *Parhyale* could pave the way for profiling DNA methylation without bisulfite conversion in the future. A crucial step in this experiment was extracting high-quality, high molecular weight genomic DNA and using the right analysis tools. We anticipated a

significant reduction in the number of gaps and achieved it. The resultant well-annotated genome can be used in subsequent analyses, ensuring the best outcome from the data.

2.1 Introduction

The genome of *Parhyale* comprises 23 pairs of chromosomes and has a size of approximately 3.6 Giga-base (Gb) (Kao et al., 2016). It is currently the second largest reported arthropod genome, surpassed only by the locust genome (6.5 Gb) (Parchem et al., 2010; Wang et al., 2014). In 2016, the genome was sequenced for the first-time using DNA isolated from a single male taken from the "Chicago-F" line, which originated from a single female (referred to as the first assembly, *Phaw_3.0*) (Kao et al., 2016). *Parhyale* exhibits an elevated level of heterozygosity, approximately ten times higher than that observed in the human genome. This level of heterozygosity is similar to that observed in the oyster *Crassostrea gigas*, which has a highly polymorphic genome sequenced from multiple individuals (Zhang et al., 2012; Kao et al., 2016).

However, the repeat content of *Parhyale's* genome is four times higher than that reported in both human and oyster genomes. This suggests that the large size of *Parhyale's* genome could be attributed to the expansion of repetitive sequences (Kao et a., 2016). These two characteristics, high heterozygosity and repeat content, pose challenges in achieving a complete and coherent assembly.

The *Parhyale* genome assembly was later updated by Dovetail Genomics using Chicago and Hi-C technologies to improve its accuracy. This update resulted in an assembly with increased range sequence scaffolds by utilizing *in vitro* reconstructed chromatin (Burton et al., 2013; Putnam et al., 2016). The process involved combining high-molecular weight DNA with purified histone and chromatin assembly factors. Subsequently, the reconstituted DNA was crosslinked using fixative agents, resulting in condensed and globular chromatin instead of the previous long, linear DNA strands. This approach enhances the quality of the genome assembly for two reasons. First, it brings together linearly distant parts of DNA in close spatial proximity, proving linkage between otherwise separate contigs, through the relationship between ligated read pairs. Second, the *in vitro* assembled chromatin exhibits lower background noise compared to *in vivo* obtained chromatin, reducing confounding biological noise.

The updated assembly of *Parhyale*, named *Phaw_5.0*, consists of 283867 scaffolds, with a total length of 2,755,851,339 Gb (Table 2.1). The N50 length of the scaffolds is 53,694,927 base-pair (bp). The assembly is publicly available at https://www.ncbi.nlm.nih.gov/assembly/ GCA_001587735.2/.

Scaffolds are created by chaining contigs together based on their relative positions in the genome, utilizing information from *in vitro* reconstructed chromatin. However, these scaffolds are separated by gaps of unknown sequence. The assembly update using Dovetail Genomics has resulted in many large scaffolds, some approaching chromosomal size. Unfortunately, this assembly process has introduced many large gaps, which may indicate potential mis-assembly. Consequently, further improvements are required to close these large gaps in the assembly.

In this chapter, we aimed to produce a new annotation for the assembly using a wide range of RNA-seq libraries. Additionally, we will employ long-read sequencing technologies in an attempt to reduce the number of gaps in the assembly.

Table 2.1. Assembly statistics for *Phaw_5.0* assembly.

entry	Scaffolds number	N50	N70	N90	Scaffolds length	#Ns
Scaffolds	283,876	53,694,927	34,722,307	234,048	2,755,851,339	558,775,491
Top 100 scaffolds	100	58,195,049	41,044,879	25,226,633	2,482,780,020	500,578,679

2.2 improving the genome assembly using PacBio sequencing

Analyses that rely on genome assemblies are critically affected by the completeness, contiguity, and accuracy of those assemblies. The process of generating a continuous sequence using short reads is often hindered by heterozygous or repeat-rich regions of the genome. As previously mentioned, *Parhyale* exhibits high heterozygosity, with a repeat content that constitutes for 57% of the assembly (Kao et al., 2016). Consequently, addressing assembly errors resulting from these characteristics, such as erroneous duplications, mis-joins, and collapses, is crucial to achieving a coherent chromosome-scale assembly in *Parhyale*.

In recent years, significant advancements in sequencing technologies have emerged to overcome challenges in generating high-quality de-novo genome assemblies. These platforms include **linked reads** (Wang et al., 2019), **optical maps** (Mendelowitz et al., 2014), **Hi-C data** (Putnam et al., 2016), and **long reads** (Logsdon et al., 2020). Long-read sequencing produces reads ranging from kilobases to mega-bases, in contrast to short-read sequencing platforms (also known as next-generation sequencing, NGS) such as Illumina's Novaseq and HiSeq (Amarasinghe et al., 2020; Bently et al., 2008; Goodwin et al., 2016), which generate 150-600 bp reads (Coombe et ls., 2021).

While short-read sequencing is cost effective, accurate, and supported by a various pipelines and analysis tools, the sequencing of short, amplified fragments complicates the task of reconstructing the original molecules (Amarasinghe et al., 2020; Heather et al.,2016). On the other hand, long reads can enhance *de novo* assembly, mapping certainty, detection of structural variants, and identification of transcripts isoforms. Moreover, long reads eliminate the bias introduced by amplification since they sequence single native molecule only (Coombe et al.,2021; Amarasinghe et al., 2020). However, long reads have higher error rates and cost compared to typical short-read technologies (Coombe et al., 2021). Nevertheless, read accuracy is continuously improving with advancements in base-calling algorithms, with current read accuracies between 87 and 98% (Coombe et al., 2021; Logsdon et al., 2020). Additionally, the throughput

and cost of long read sequencing are also improving, making it a viable option for a wide range of applications in genomics (Amarasinghe et al., 2020; Burgess 2018; Yuan et al., 2017).

Two dominant long-read sequencing technologies, known as third generation sequencing, are Pacific Biosciences of California, Inc. (PacBio) single-molecule real-time (SMRT) sequencing and Oxford Nanopore Technologies plc. (ONT, Oxford UK) (Eid et al., 2009; Jain et al., 2016). These platforms were commercially released in 2011 and 2014, respectively, and have since become the preferred choices for an increasing number of applications (Coombe et al., 2020; Amarasinghe et al., 2020).

SMRT sequencers, including RSII, Sequel, and Sequel II, employ a sequence-by-synthesis technology. In this approach, nucleotides are fluorescently tagged as they synthesize along the individual DNA template molecule using DNA polymerase to drive the reaction (Roberts et al., 2013; Amarasinghe et al., 2020). Real-time imaging is utilized to detect the fluorescent signal. Since the sequencing is performed on a single template molecule, there is no degradation over time, and the sequencing reaction only concludes when the template and polymerase dissociate (Roberts et al., 2013; Amarasinghe et al., 2020).

We conducted a pilot experiment of PacBio sequencing on *Parhyale's* genome, which to our knowledge, represents the first PacBio dataset for *Parhyale*. Therefore, we generated only one library of PacBio data, resulting in relatively low coverage of approximately only 8.5x. Utilizing long reads exclusively for *de novo* assembly necessitates sufficient sequencing coverage, along with the high computational costs associated with error correction of the assembly (Xu et al., 2020; Ou et al., 2020).

Error correction is crucial because raw PacBio reads exhibit lower accuracy compared to NGS platforms, and base-calling errors can introduce frameshifts and alterations in protein-coding or regulatory regions, leading to inaccurate genomic information (Xu et al., 2020; Watson & Warr, 2019). To address these challenges, numerous hybrid assemblers have been developed to leverage both TGS and NGS read data in constructing a final assembly (Ye et al., 2016; Boetzer et al., 2014; Zimin et al., 2013; Luo et al., 2019). Additionally, several tools have been designed to close gaps in existing NGS-based assemblies using long-reads data. These algorithms reduce computational complexity and cost by improving only in the missing regions while preserving information from the majority of the existing assembly. Various tools have been

developed for gap filling, employing different strategies (Xu et al., 2020; English et al., 2012; Piro et al., 2014; Kosugi et a., 2015; McGinnis et al., 2004; Warren et al., 2016; Xu et al., 2019). The efficiencies of most of these tools relies on the quality coverage depth of the long reads, and some may require prior fragmentation of the long reads into short fragments, leading to loss of valuable information (Xu et al., 2020; English et al., 2012; Walker et al., 2014).

We employed a software tool called TGS-GapCloser (Xu et al., 2020) to leverage low-coverage, error-prone long reads for improving the quality of large genome assemblies and subsequently enhancing gene annotation. TGS-GapCloser provides a more efficient and accurate way to close gaps compared to other tools, leading to higher-quality downstream analyses of any type of data.

Another task that can be accomplished using PacBio data is scaffolding or re-scaffolding contigs from an existing assembly, which enhances the assembly's contiguity and links smaller contigs into larger scaffolds (Bashir et al., 2012; Zimin and Salzberg, 2022). Several tools have been developed for this purpose, although most are designed for bacterial and other small-sized genomes (Qin et al., 2019; Warren et al., 2015; Bashir et al., 2012; Boetzer et al.2014; Zimin and Salzberg, 2022). However, SAMBA (Zimin and Salzberg, 2022) and LRScaf (Qin et al., 2019) tools are capable of handling large-sized genomes, such as mammalian genomes, which is applicable to *Parhyale's* genome. Both tools can process both low-coverage and higher-coverage long-read data. Therefore, we applied both SAMBA and LRScaf to our PacBio dataset in an attempt to link smaller scaffolds within the assembly into larger ones.

Data

We utilized the PacBio Sequel II SMRT sequencing platform to generate HiFi reads, which are significantly longer compared to the reads produced by second-generation Illumina platform (Lee et al., 2014). In total, we generated 30.642 Gb of read data with an average read length of 17,480 bp (Table 2.2). The increased read length offers the advantage of spanning more repetitive elements, thereby facilitating the achievement of a more contiguous assembly, potentially even at the level of complete chromosomes. Further information regarding sample preparation and sequencing can be found in the materials and methods chapter.

Table2.2. PacBio data statistics. HiFi reads are produced using circular consensus sequencing (CCS) mode on PacBio system, while the reads are shorter than the raw long read data, they yield higher accuracy.

	Read base	**Read number**	**Average read length**	**N50**
Raw data	516,288,342,433	6,572,239	78,555	163,488
HiFi reads	30.642 G	1,752,885	17,480	17,555

Scaffolding

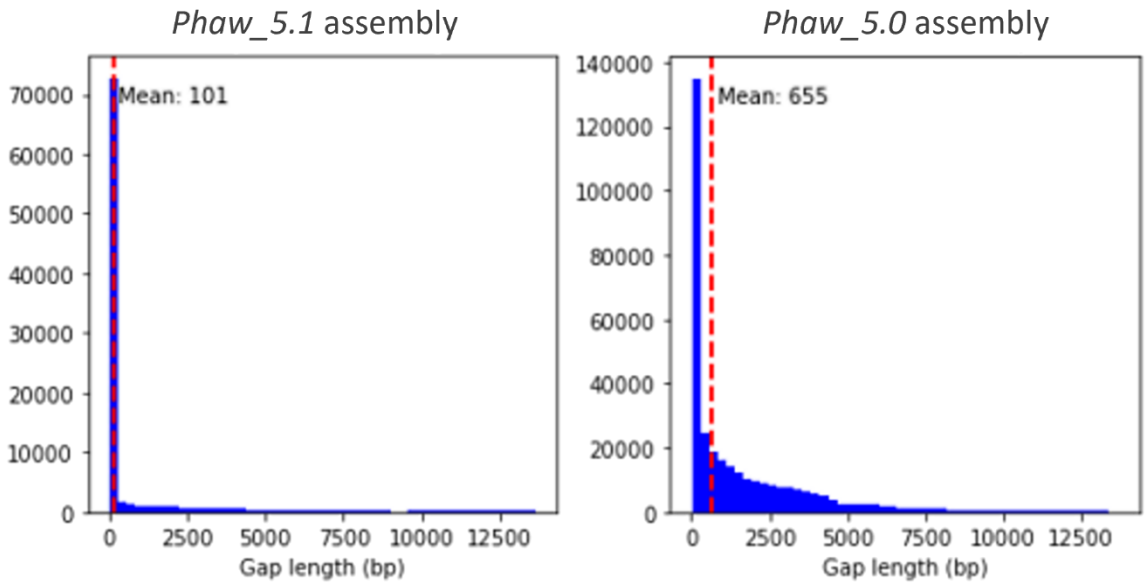
To begin with, we employed the SAMBA tool (Zimin and Salzberg, 2022) to scaffold the genome using HiFi reads. We evaluated the assembly metrics using QUAST 5.0 (Gurevich et al., 2013). The application of SAMBA resulted in a reduction in the number of scaffolds to 280585, which corresponds to a decrease of 1.16% in the total number of scaffolds. However, there was no change in the N50 value after the application of SAMBA (Table 3). Overall, the scaffolding process using the available sequencing depth (approximately 8x coverage) did not significantly reduce the number of scaffolds in the assembly.

Gap filling

Following the scaffolding of the assembly, we utilized TGS-GapCloser (Xu et al., 2020) to bridge the gap regions between contigs within the scaffolds. The improvements in the assembly after gap closure are illustrated in Figure 1. Remarkably, TGS-GapCloser successfully closed 73.4% of the total 558,775,491 gaps in the assembly using PacBio HiFi reads. As a result, the N50 value of the assembly increased to 54784699 (a 2.03% increase) after the combined steps of scaffolding and gap closure (table 2.3).

The updated assembly now has a larger size compared to the reference *Phaw_5.0* assembly, with sizes of 2.82 Gb and 2.76 Gb, respectively. Additionally, the remaining gaps in the new assembly, denoted as *Phaw_5.1*, were generally shorter in length compared to the gaps in the *Phaw_5.0* assembly. The median length of these remaining gaps in the *Phaw_5.1* assembly was 101 bp, whereas it was 655 bp in the *Phaw_5.0* assembly.

Figure 2.1. Comparison of gap content before and after gap closure using PacBio HiFi reads. The figure demonstrates a significant reduction of 73.4% in the number of gaps after incorporating PacBio data, resulting in the closure of a total of 300,225 gaps.



The overall genome fraction, when compared to the reference assembly, has experienced a notable increase of 1.69%. Notably, the completeness of numerous scaffolds exhibited significant improvement following the implementation of gap closure, as depicted in Figure 2.2.

Next, an expression-based annotation was generated for the updated assembly. The details of the annotation will be described in the following section.

Figure 2.2. Comparison of gap fraction in each scaffold (top100 presented) before and after gap closure. The top panel represents the reference assembly (*Phaw_5.0*), displaying the percentage of genomic gap (Ns) in each scaffold. The bottom panel illustrates the results after applying TGS-GapCloser to fill the assembly gaps (*Phaw_5.1*), showcasing the corresponding gap fractions in each scaffold.

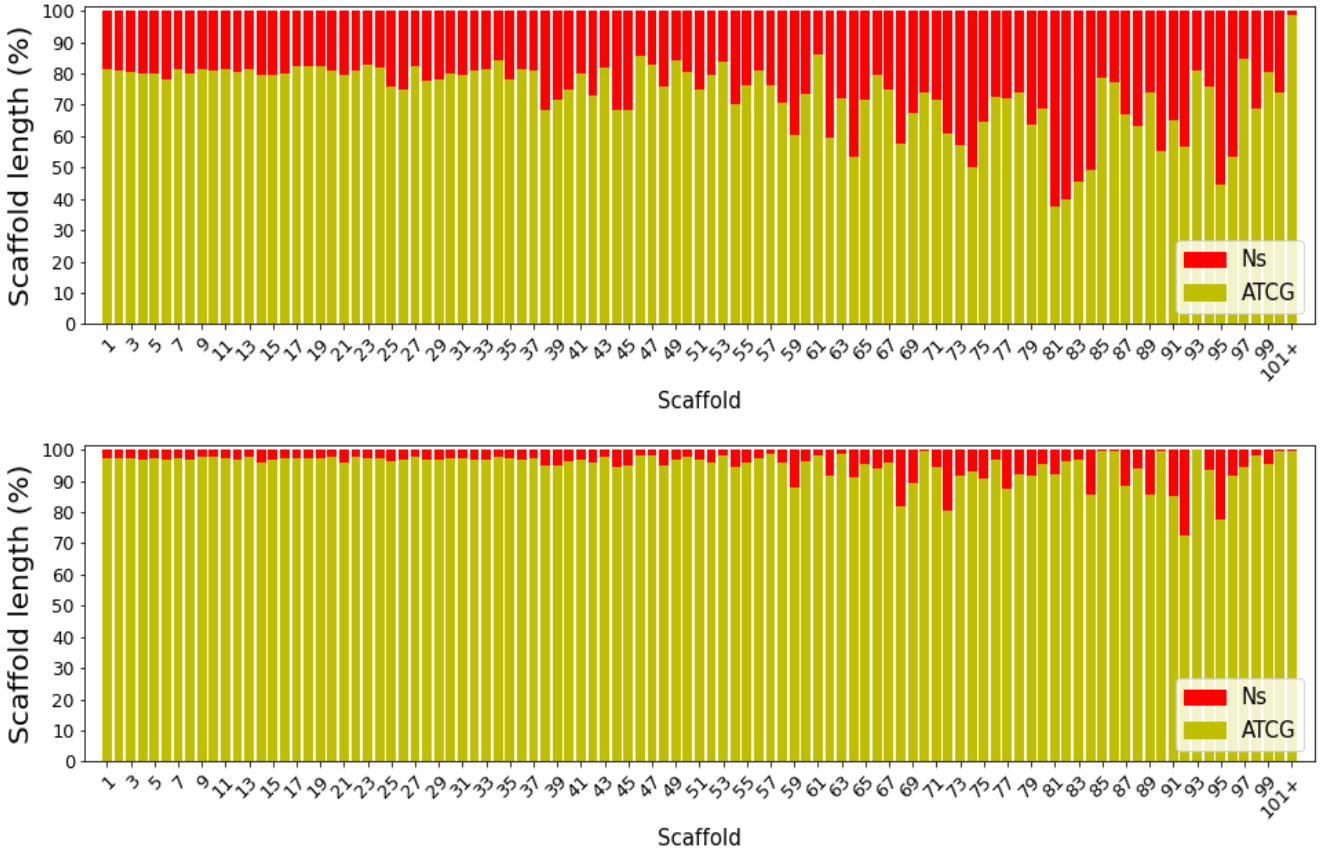


Table 2.3. Assembly statistics before and after scaffolding and gap filling.

	Phaw_5.0	**TGS-GapCloser**	**Change**	**SAMBA**	**Change**	**SAMBA + TGS-GapCloser (*Phaw_5.1*)**	**Change**
Length	2,755,851,339	2,809,677,865	+1.95%	2,769,876,797	+0.5%	2,823,710,058	+1.69%
Scaffold number	283,876	283,876	NA	280,585	-1.16%	280,585	-1.16%
Scaffold mean length	9,707.94	9,897.55	+1.95%	9,871.79	+1.65%	10,063.65	+3.08%
Scaffold max length	111,408,412	113,994,380	+2.32%	111,408,412	NA	114,006,073	+2.28%
Scaffold min length	200	199	-0.5%	200	NA	199	-0.5%
N_count	558,775,491	92,700,848	-83.4%	558,769,713	-0.001%	92,739,792	-83.40%
N50	53,694,927	54,786,424	+1.99%	53,694,927	NA	54,784,699	+2.03%
N70	34,722,307	35,465,393	+2.09%	34,866,082	+0.41%	35,596,768	+2.52%
N90	234,048	298,080	+21.74%	240,551	-39.94%	253,979	+34.21%
Gaps	408,799	108,587	-73.44%	408,781	0.004%	108,574	-73.44%
Gap mean length	1,678	990	-41.%	1,678	NA	991	-41.%
Gap median length	655	101	-84.58%	655	NA	101	-84.58%

2.3 Pipeline for expression-driven annotation of *Parhyale hawaiensis* genome

We sought to produce an expression-based annotation of all transcribed loci on the *Parhyale hawaiensis* genome. To achieve this, we utilized a combination of the *de novo* assembled transcriptome from Kao et al. (2016) and 85 independent RNA-seq datasets covering a wide range of conditions. By including this diverse range of datasets, we aimed to improve the overall representation of the transcriptome compared to the previous annotation, which relied on relatively sparse and life-history stage restricted RNA-seq data. This method might be useful for discovering lowly expressed protein-coding genes and non-coding RNAs that may not have been discovered using homology-based annotation processes like MAKER (Cantarel et al., 2008). Consequently, our aim was to annotate all protein-coding and transcribed loci in the genome, leveraging a broad range of transcriptomic information.

To carry out the expression-based annotation, we obtained RNA-seq data generated in our lab, including samples from female and male adults, as well as early embryonic stages. Additionally, we incorporated the *de novo* transcriptome assembly from Kao et al. (2016), as well as embryonic time course RNA-seq data from Nipam Patel (unpublished) and Calvo et al. (2022). All these sequences were mapped to the *Phaw_5.1* assembly, with an average mapping rate of 84% (maximum: 87.56% and minimum:79.7%). The mapped sequences were then merged and consolidated, resulting in an assembly of 131,116 transcripts using Stringtie2 (Kovaka et al., 2019) (Figure 2.4).

To reduce redundancy of the assembled transcripts, we employed CD-HIT (Li et al., 2006), which clusters transcripts based on their sequence similarity. After the first round of CD-HIT, the number of transcripts was reduced into 74,420 non-redundant transcripts (identity < 0.95). Subsequently, we ran Transdecoder v.5.0.2 [http://transdecoder.github.io] to predict open reading frames (ORFs). After running Transdecoder predict and performing a second round of CD-HIT, we obtained a final annotation of 22,756 transcripts. In predicting the coding regions, we incorporated information from both BLASTp (Altschul et al., 1990) and pfam UniProt (Coudert et al., 2023) hits.

The mean coding gene size is 1,583 bp, with a median of 1,062 bp, which is larger than in the previous annotation (Table 2.4). Notably, *Parhyale's* gene length is longer than that of *Caenorhabditis elegans*, *Daphnia pulex*, and *D. melanogaster* but similar to *Homo sapiens* (Kao et al., 2016). Moreover, the mean intron length in *Parhyale* is 9.1 kilo-bases (kb), which is higher than that of humans. Previous studies have reported that *Parhyle's* intron size is similar to that of *H. sapiens* (5.9 kb) and dramatically longer than invertebrate models like *D. pulex, D. meanogaster* and *C. elegans* (0.3 – 1kb) (Kao et al., 2016). We annotated the final proteome dataset using Pfam and performed BLAST against Uniprot. The dataset was then clustered using OrthoFinder2 (Emms and Kelly, 2019) to provide phylogenetic inference of orthologs and identify possible gene duplication events (Figure 2.6). BLAST analysis revealed that the highest conservation was observed within the Malacostracan clade, which includes species such as the amphipod *Hyalella azteca* and the decapod *Penaeus vannamei* (Figure 2.6A). This finding indicates that most of the predicted genes have putative orthologs in the genomes of closely related species, highlighting the high quality of the predicted gene models in *Parhyale*.

Furthermore, we performed Orthofinder analysis across a wide range of species, including non-bilaterian species and numerous metazoans. The analysis identified orthologous and paralogous protein groups across 27 species. Overall, 85.6% of the analysed genes (n=499,961) were found in orthogroups (n= 40,199). Specifically, 13,789 orthologous groups contained proteins found exclusively in Panarthropoda, while 13,050 orthologous groups contained proteins found only in Arthropoda (Figure 2.6B). Additionally, we found 7,076 ortho-groups that were shared exclusively among Crustacea. It worth noting that 3,286 of *Parhyale's* genes were not assigned to any orthologous group, potentially this group is lineage specific proteins.

The following table presents the differences observed in the annotation before and after incorporating PacBio data into the assembly.

Table 2.4. Statistics comparing old and new annotation.

Entry	*Phaw_3.0* annotation	*Phaw_5.0* annotation	*Phaw_5.1* annotation
Number of genes	28,666 (26,826 non-redundant)	23,304	22,756
single exon genes	5,617	3,498	2,750
Exon mean length (bp)	400	661	810
Exon median length (bp)	164	166	167
Intron mean length (bp)	8,026	8,884	9,052
Intron median length (bp)	1,934	2,140	2,180
Completeness	12,738 (47%)	13,820 (59.3%)	15,434 (67.8%)
5' Partial	4,108 (15%)	3,537 (15%)	2,946 (13%)
3' Partial	3,940 (15%)	2,485 (11%)	1,780 (8%)
Internal (no UTRs)	6,169 (23%)	3,462 (15%)	2,596 (11%)
CDS mean length (bp)	983	1,436	1,583
CDS median length (bp)	609	948	1,062
Exon per transcript mean	4	5	6
Exon per transcript median	3	4	4

Figure 2.3A. Histogram comparing the *Phaw_5.0* reference assembly and the updated *Phaw_5.1* assembly. The values have been normalised to illustrate relative differences. The most prominent disparities between the two assemblies are observed in N_counts and gap length. In the *Phaw_5.1* assembly, there is an increase in the total length of the assembly and the mean length of scaffolds compared to the *Phaw_5.0* assembly.

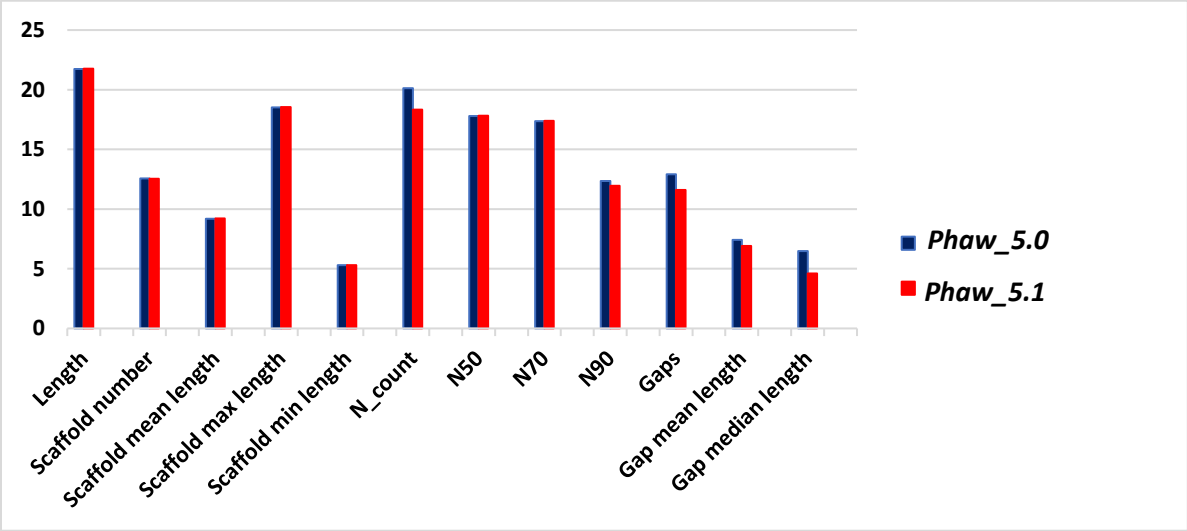


Figure 2.3B. Histogram comparing the annotations of three *Parhyale* assemblies. The values have been normalised to illustrate relative differences. The annotation statistics of the *Phaw_3.0* assembly (the initial *Parhyale* assembly from Kao et al., 2016), the Dovetail *Phaw_5.0* assembly, and the most recent *Phaw_5.1*assembly after integrating PacBio data are included in the comparison.

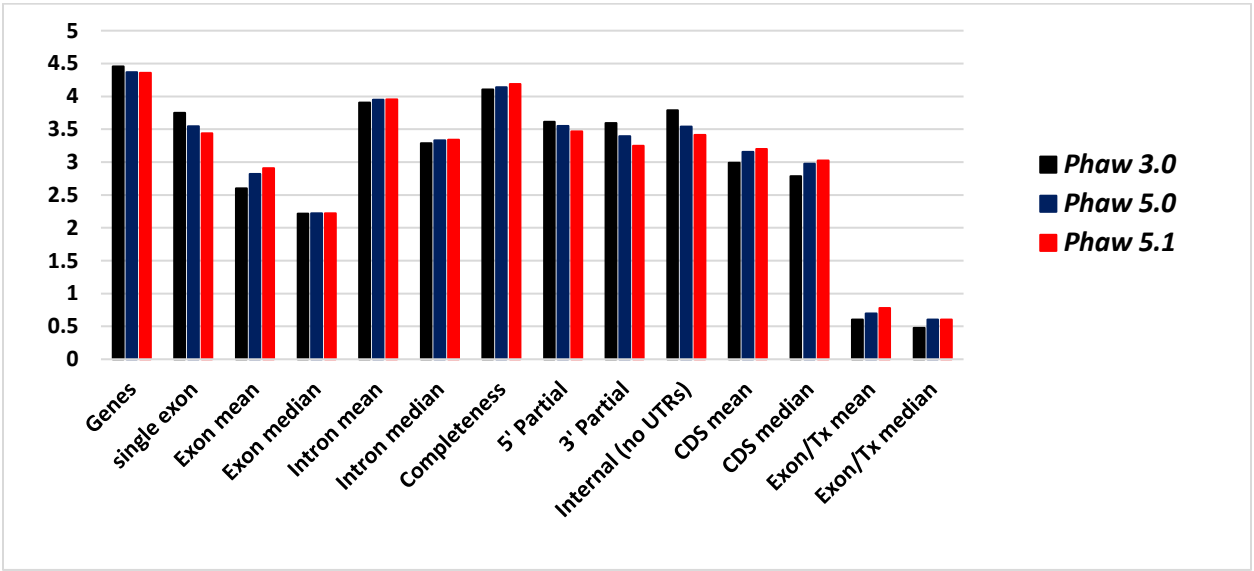


Figure 2.4. Genome annotation pipeline of protein-coding genes.

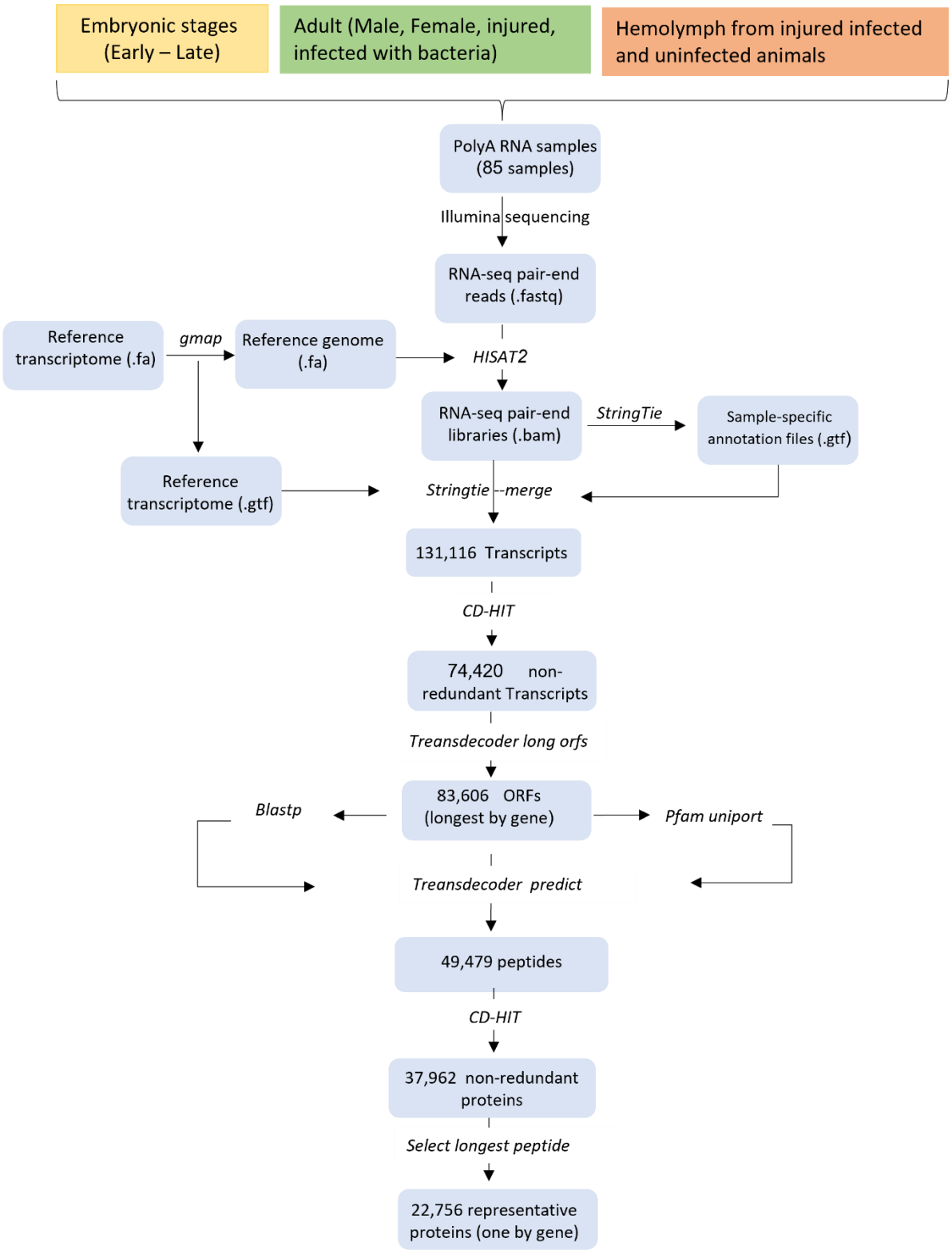
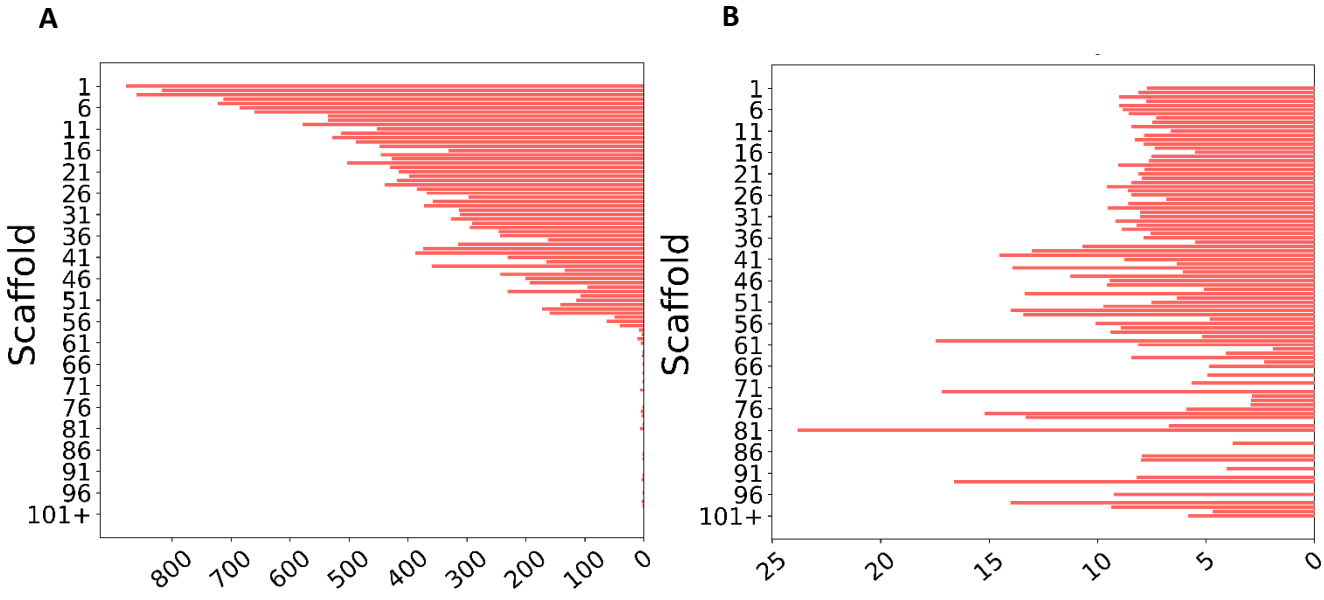


Figure 2.5. Number of protein-coding genes per scaffold after integrating PacBio data. The bar chart on the right **(A)** represents the row number of genes per scaffold, and the bar chart on the left **(B)** represents the number of genes per scaffold, normalised based on the length of each scaffold. Given that scaffold numbers are in decreasing order with respect to scaffold size, it is evident that the longest scaffolds tend to contain a higher number of genes when we look at **(A)**, suggesting a positive correlation between scaffold size and gene count. However, when examining **(B)**, where the values are normalised with scaffold length, a different pattern emerges. The relationship between gene density and scaffold becomes less strict. It is no longer the case that the largest scaffolds have the highest gene density. Instead, smaller scaffolds can exhibit a higher gene density than larger ones.



In summary, Figure 2.5 emphasizes that there is no correlation between scaffold size and the density of genes per scaffold. While larger scaffolds may indeed contain a greater overall number of genes, the distribution of genes within each scaffold does not follow a consistent pattern based on scaffold size. Other factors likely play a role in determining the gene density per scaffold in the *Parhyale* genome.

Figure 2.6A. BLAST hits between *P.hawaiensis* and the proteomes of indicated species. Comparison between *Parhyale* and proteomes of 27 species from diverse animal taxa. The comparison aims to identify the number of hits at different thresholds, indicating the level of sequence similarity between *Parhyale* and other species. The highest number of hits is observed in the proteomes of the most closely related species to *Parhyale*, specifically within the Malacostracan clade.

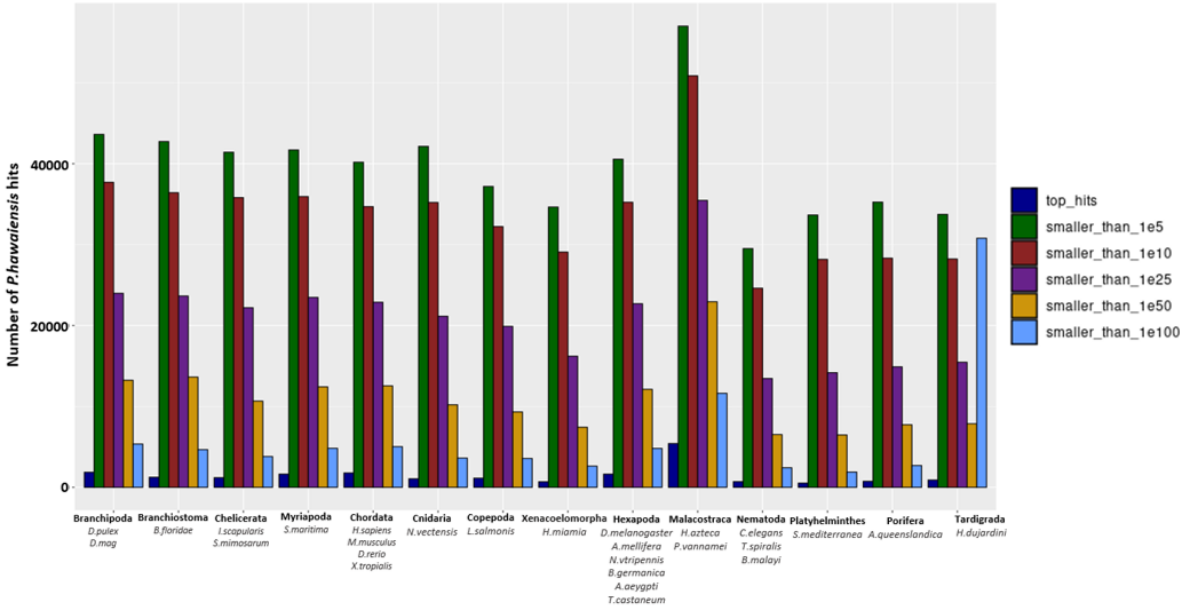
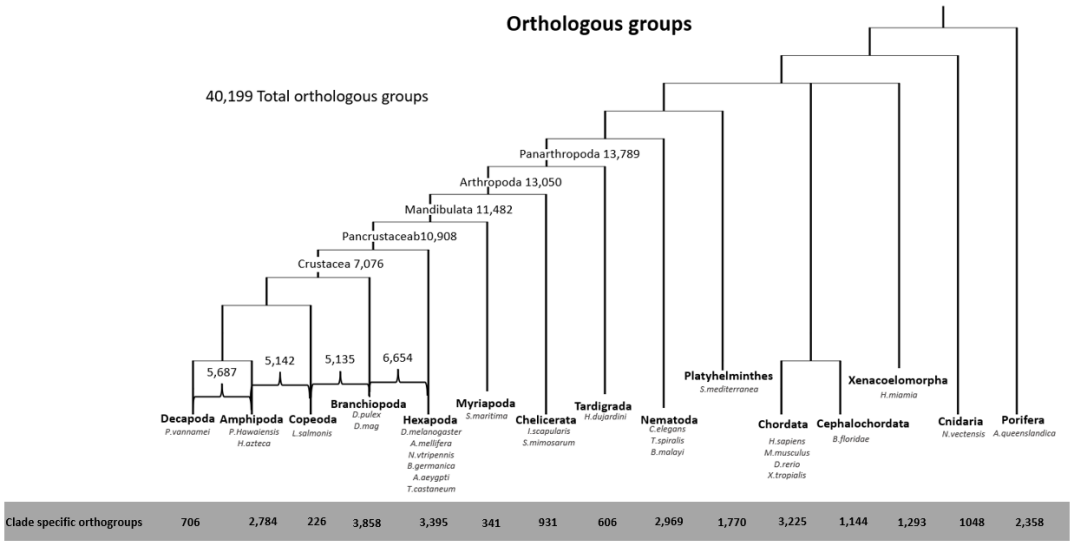


Figure 2.6B. Cladogram illustrating the shared orthologous protein groups at various taxonomic levels. Number of clade-specific groups is also indicated, highlighting the count of unique protein groups present within specific taxonomic clades. A total of 40,199 orthogroups were identified using Orthofinder across 27 proteomes included in the analysis.



2.4 Examples of closed gaps in individual genes

To further evaluate the quality of the updated assembly using PacBio reads, we aimed to examine certain genes that were previously known to have a gap in their structure. Our goal was to determine if these gaps have now been resolved and if a complete sequence is present after integrating long reads into the assembly. One of the most extensively studied gene groups in *Parhyale* is the Hox genes, which comprise a highly conserved set of homeodomain transcription factors (Pavlopoulos et al., 2009; Liubicich et al., 2009; Serano et al., 2016; Albertstat et al.,2022). Typically, these genes are organized within the genome in a conserved cluster (Pourquie,2009; Pace et al., 2016; Serano et al., 2016). Comparative data indicates that the ancestral arthropod Hox cluster consisted of ten genes arranged in a specific orientation, often positioned close to each other on the same chromosome (Ferrier et a, 1996; Powers et al., 2000; Pace et al., 2016).

Studies conducted in *Parhyale* have revealed the presence of 9 canonical Hox genes (Serano et al., 2016; Kao et al., 2016). Early chromosome walking experiments indicated that the hox genes *labial (lab)* and *proboscipedia (pb)*are linked together. Additionally, *Deformed (Dfd), Sex combs reduced (Scr), Antennapedia (Antp),* and *Ultrabithorax (Ubx)* were found to be located contiguously (Serano et al., 2016). However, it remained uncertain whether *Abdominal-A (Abd-A), Abdominal-B (Abd-B), and Hox3* are adjacent to each other or to the aforementioned hox genes. In the initial genome sequencing (Kao et l., 2016), double fluorescent in situ hybridization (FISH) using intronic sequences demonstrated linkage between all the hox genes except *Hox3*. Nevertheless, genome sequencing alone failed to confirm the present of all hox genes within a single cluster.

After updating the genome assembly using Dovetail genomics, all 9 hox genes were found to be present as a single cluster in the genome, spanning approximately 3.6 Mb (figure 2.7). In the Dovetail assembly (*Phaw_5.0*), gaps accounted 0.5% (250 gaps) of the cluster region, which mostly covered the genes in the cluster (detailed figures can be found in Appendix A). However, after integrating PacBio data to update the assembly, the percentage of gaps was reduced to only 0.06 % (46 gaps) of the cluster region. All gaps in *Abd-B, Abd-A, Dfd, Pb, and Lab* have been completely closed in *Phaw_5.1* assembly.

The dovetail assembly revealed the presence of an additional copy of *Antp* hox gene, which was also confirmed after integrating PacBio data and updating the annotation. The two copies share 70% of the coding sequence. The first copy consists of five exons, while the other copy has only two exons, resulting in the generation of two different cDNAs and consequently two distinct proteins. This observation was previously reported by Serano et al. in 2016.

Furthermore, we discovered another gene adjacent to *Pb,* which was not previously reported. This gene consists of a single exon. Blast search analysis revealed that this gene is an orthologue to *Pb* hox genes in various arthropod and other species. It also clusters with the rest of the Hox genes in the Orthofinder analysis output. However, MAFFT alignment analysis indicated no sequence similarity between this gene and the *Pb* gene. It is worth noting that this single exon copy did not appear in the pfam analysis output, indicating the absence of a homeodomain. Further investigation is required to determine if this exon is a part of the original transcript and was erroneously annotated as a separate gene due to miss-assembly or an annotation error, or if it represents a genuine duplication of the Pb Hox gene. Since it originates from a separate transcript, it cannot be considered another isoform or variant of *Pb*. Additional research is necessary to clarify its nature.

We discovered three new genes in the cluster, in addition to the previously known ones. One of these genes is situated between *SCR* and *DFD* in the *Parhyale* genome. Earlier research identified the microRNA mir-10, which is consistently found in both vertebrate and invertebrate Hox clusters, specifically between *Hoxb5/Scr* and *Hoxb4/Dfd* (Kao et al., 2016; Enright et al., 2003). Given this, we investigated whether this transcript was mir-10. In our examination of the *Phaw_5.1* assembly, we confirmed the location of mir-10 between *SCR* and *Dfd* in *Parhyale*. However, the transcript we identified turned out not to be mir-10, but rather a fully structured gene with three exons and both 5' and 3' UTRs.

Additionally, two other transcripts were found: one between *Antp-Ubx* and *Scr*, and the other one between *Dfd* and *Hox3*. These transcripts did not exhibit any significant matches in the database. Furthermore, the Orthofinder analysis did not cluster these genes with the hox genes or any other orthogroup. Sequence alignment with other hox genes indicated that these three transcripts share only 40% or less of their

sequence identity with the known hox genes. Consequently, it is possible that these transcripts are specific to *Parhyale*. Further investigations are necessary to elucidate their functions and understand their significance within the *Parhyal*e genome.

After integrating PacBio data to the dovetail assembly, numerous gaps in different transcripts in the genome were successfully closed. One example is the zinc-finger transcription factor odd-paired (*Opa*), which initially had gaps accounting for 20% of the transcript, including gaps within its coding sequence (Figure 2.8). An exonic gap with an estimated length of 173 nucleotides was filled with a 236 bp sequence fragment after incorporating PacBio data. The accuracy of the inserted fragment was confirmed through PCR amplification of this region, showing a 92% sequence identity between the PCR amplicon and the inserted region (see Appendix B). As a result, the gaps in the *Opa* transcript were reduced to only 0.4% of its length. Furthermore, the total length of the transcript increased by 1% and the coding sequence length expanded from 1188 bp to 1305 bp.

Figure 2.7. The complete Hox gene cluster in the genome assembly. IGV screenshot of the complete Hox gene cluster in the *Parhyale* genome. The top panel displays the gaps in the cluster in the *Phaw_5.0*, while the bottom panel shows the reduced gaps in the *Phaw_5.1* assembly. The number of gaps in the cluster has been significantly reduced from 250 to 46 gaps only.

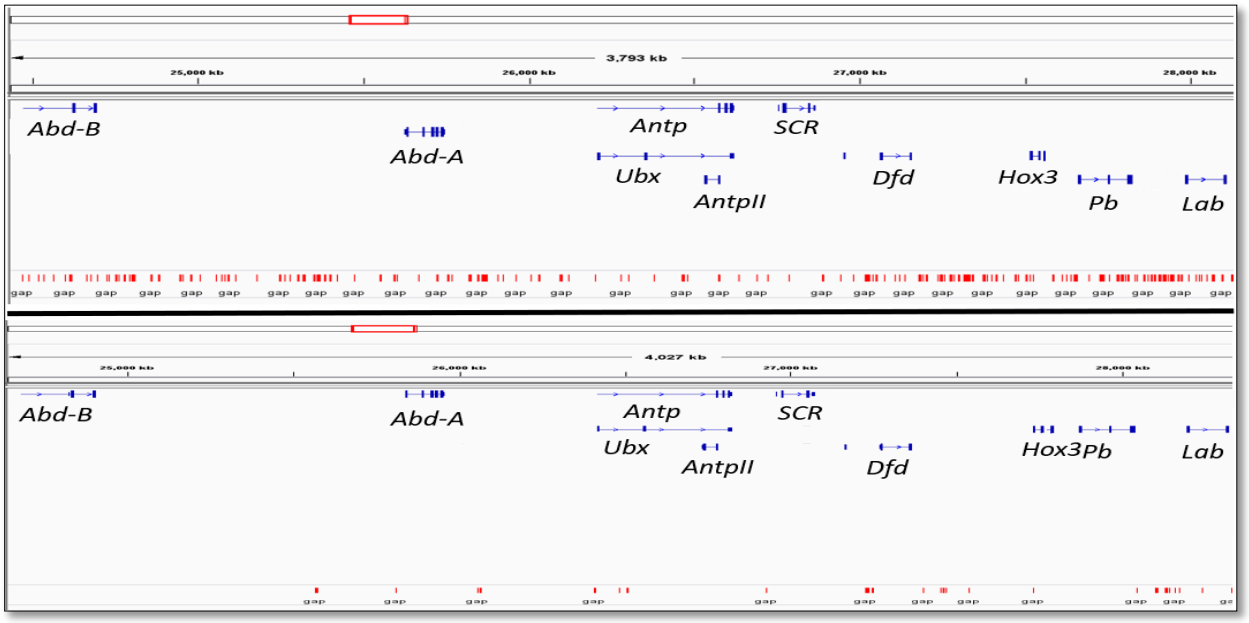
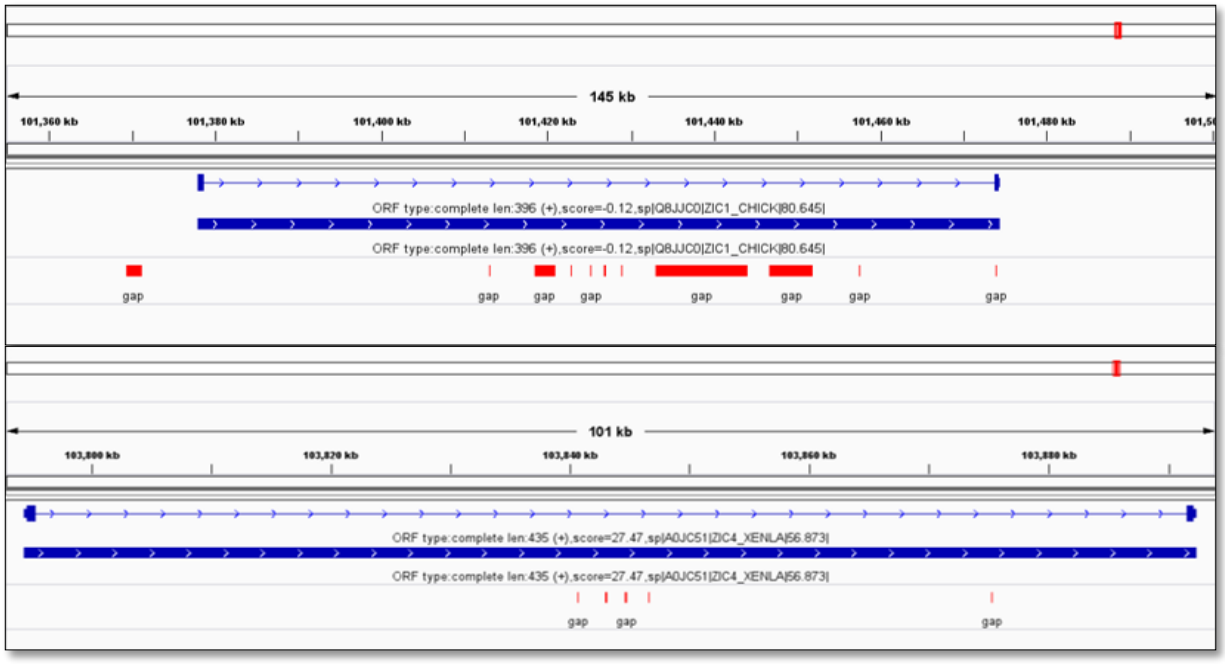


Figure 2.8. Closed gap in Odd-paired transcription factor. IGV screenshot of the *Parhyale Odd-paired* ortholog, highlighting the resolved gaps, including gaps within exonic regions. After integrating PacBio data, the total length of the transcript increased by 1%.



2.5 Genes encoding epigenetic regulation in *Parhyale hawaiensis*

Epigenetic regulation of gene expression involves various mechanisms such as DNA methylation, histone modification, and non-coding RNAs. Among these, post-translational modification of histones is a well-conserved mechanism in eukaryotic genomes. In the case of *Parhyale,* its genome encodes approximately 100 histone modifying enzymes, as reported in Kao et al. in 2016. Appendix C provides updated ids on these enzymes in the new assembly.

Furthermore, the *Parhyale* genome encodes the full repertoire of enzymes involved in establishing and erasing DNA methylation. It includes representatives from all three families of DNA methyltransferases (DNMTs). DNMT1, which is responsible of maintenance of methylation, by copying cell-type specific methylation patterns into the new DNA strand during cell division. DNMT3, which performs *de novo* methylation by introducing methylation to previously unmethylated DNA; and DNMT2, which methylates transfer RNA (tRNA) in many species.

Additionally, the *Parhylae* genome encodes two orthologues of methyl-binding domain (MBD) Proteins (MBD2/3, MBD4). These proteins bind to methylated DNA and recruit histone modifying enzymes, ultimately leading to chromatin inaccessibility. Furthermore, *Parhyale* possesses a single orthologue of Ten-Eleven-Translocation (TET) proteins, namely TET2. TET2 functions in DNA de-methylation by oxidizing 5-methylcytosine into 5- hydroxymethylcytosine (5hmC) and catalysing further oxidation products, such as 5-formylcytosine (5fC) and 5-carboxylcytosine (5caC), as reported by Kao et al. in 2016 and Leoni et al. in 2015.

Parhyale has a single copy of DNMT1. However, after updating the assembly using dovetail technology, it was found that DNMT1 has been split into two transcripts, likely due to a miss-assembly error. Despite this, both transcripts exhibit identical expression patterns in all RNA-seq data. Further, cloning sequences have verified that the coding sequences (CDS) within these two transcripts are continuous and originate from a single mRNA. Unfortunately, this problem has not been resolved after

PacBio data integration as it arises from errors introduced during the genome re-assembly process, a common issue when transitioning from contigs to scaffolds.

We performed BLAST analysis against a selection of species known to have DNA methylation, including *H. sapiens, Mus musculus, Danio rerio,* and *D. pulex*, to confirm the identity of each DNA methylation-related gene in *Parhyale*. The transcript of each gene was annotated in the genome and located on a single scaffold. To further investigate the evolutionary relationships, a phylogenetic analysis was conducted using available protein sequences from various species, including *H. sapiens, M. musculus, Gallus gallus, D. rerio, D. pulex, Apis mellifera, Tribolium castaneum, Bombyx mori, D. melanogaster, S. mediterranea,* and *P. hawaiensis*. Maximum likelihood phylogenies were constructed for each protein family, as shown in figures 2.10, 2.11, and 2.12.

Many species have undergone multiple rounds of duplications and diversification that led to the existence of multiple copies of DNMT, MBD, and TET genes in some species, while other species have experienced losses in certain genes of these families (Jeltsch 2010, Jurkowski and Jeltsch 2011, Firmino et al. 2017). In some invertebrate species, only one or two of these genes were identified, particularly in species where DNA methylation is lost but some of its modulators of methylation are still conserved (e.g., MBD2/3 in planarians and Drosophila). Conversely, due to the two rounds of whole genome duplication during vertebrate evolution, we observe multiple copies of genes such as DNMT3 and MBD.

All *Parhyale* protein sequences clustered appropriately within the respective gene families. In the case of DNMT3, invertebrates clustered with DNMT3B of the vertebrate species (Figure 2.10). Notably, there is one copy of MBD2/3 in invertebrate species, which is believed to be the invertebrate ancestor of MBD2 and MBD3 proteins in vertebrates, as they share approximately 80% sequence homology (Menafra and Stunnenberg 2014, Sarda et al.2012). MBD4, which is lost in many invertebrates, is conserved in *Parhyale* and it clusters together with MBD4 of zebrafish (Figure 2.11). However, while MBD4 can bind methylated DNA, its function appears to be primarily involved in DNA repair rather than transcriptional repression (Du et al. 2015). lastly, for the TET family proteins, a TET protein from

the plant *Arabidopsis thaliana* was used as an outgroup (Figure 2.12). TET2 in all invertebrate species, including *Parhyale*, clustered together with the corresponding protein from vertebrate species.

Figure 2.9. Diagrams illustrating the exon-intron structure of of DNA methylation genes. Synonyms single nucleotide polymorphisms (SNPs) identified within specific exons of each gene in *Parhyale's* DNA methylation machinery are marked by red arrows. Prior to initiating genome editing experiments using CRISPR/Cas9 technique, identifying these SNPs was crucial to ensure guide RNAs were not designed within these loci, thus optimizing the likelihood of successful genome editing. SNPs were identified in DNMT1, DNMT3 and MBD2/3 only. DNMT2, MBD4 and TET2 were not cloned.

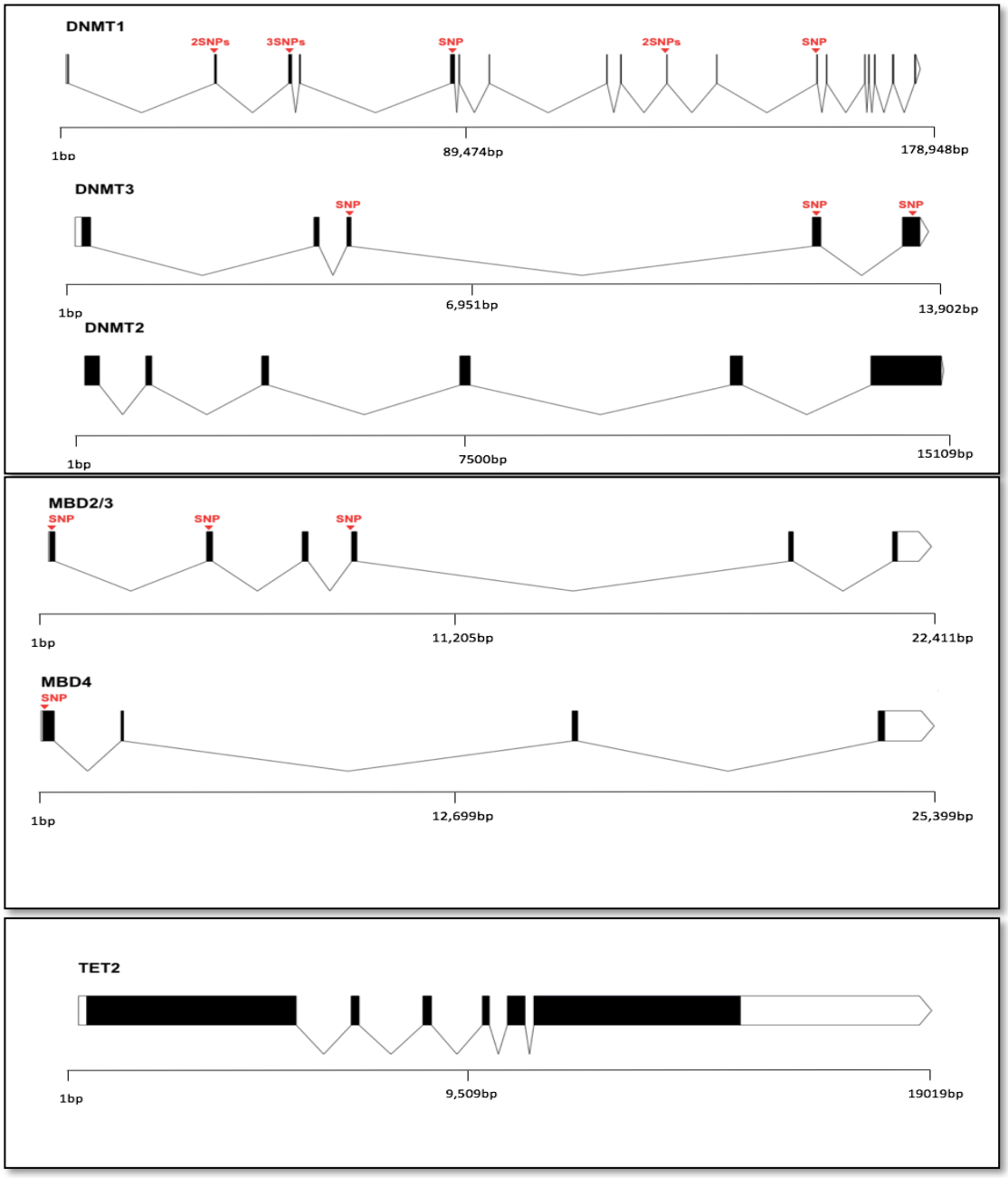


Figure 2.10. Phylogenetic tree of DNMTS. A maximum likelihood phylogeny was constructed using animal species from various taxa to represent each protein family. The tree demonstrates that DNMTs cluster within their designated protein families. Numbers at the tree nodes indicate the frequency of bootstrap values; bootstrapping set at 1,000 replications.

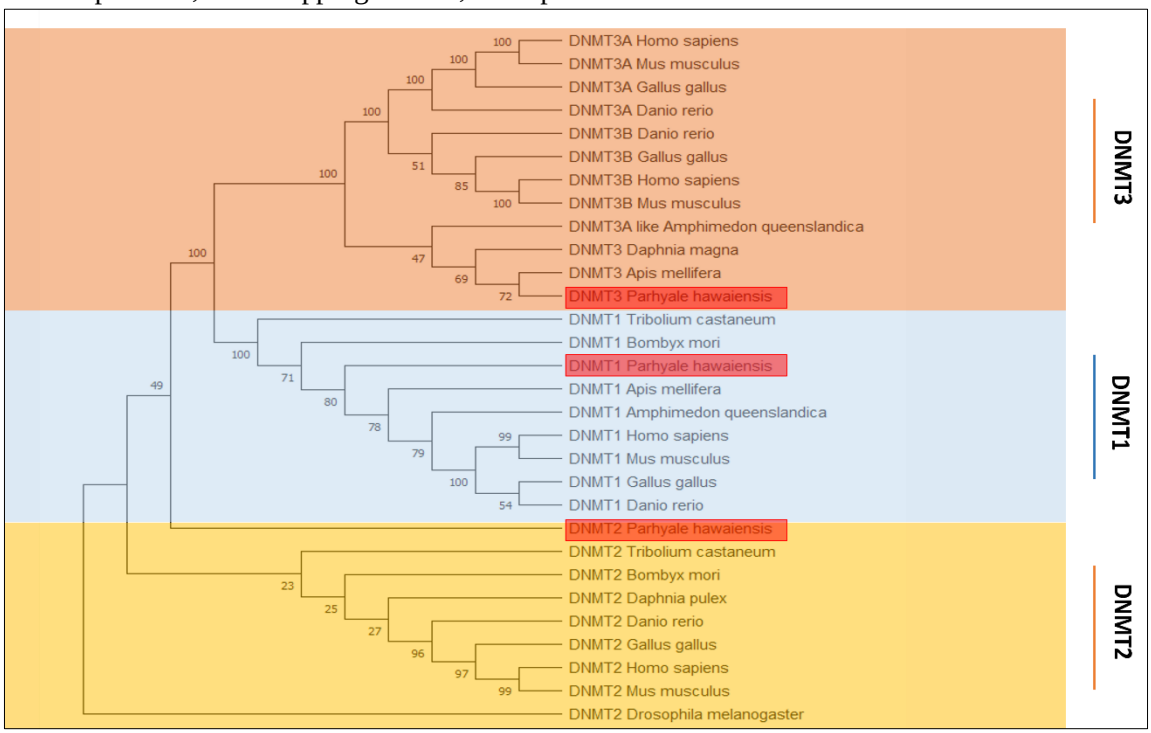


Figure 2.11. Phylogenetic tree of MBDs. The tree represents the evolutionary relationships among MBD proteins. In invertebrate species, a single copy of MBD2/3 is observed, which is believed to be the ancestral counterpart of MBD2 and MBD3 in vertebrates following two rounds of duplications.

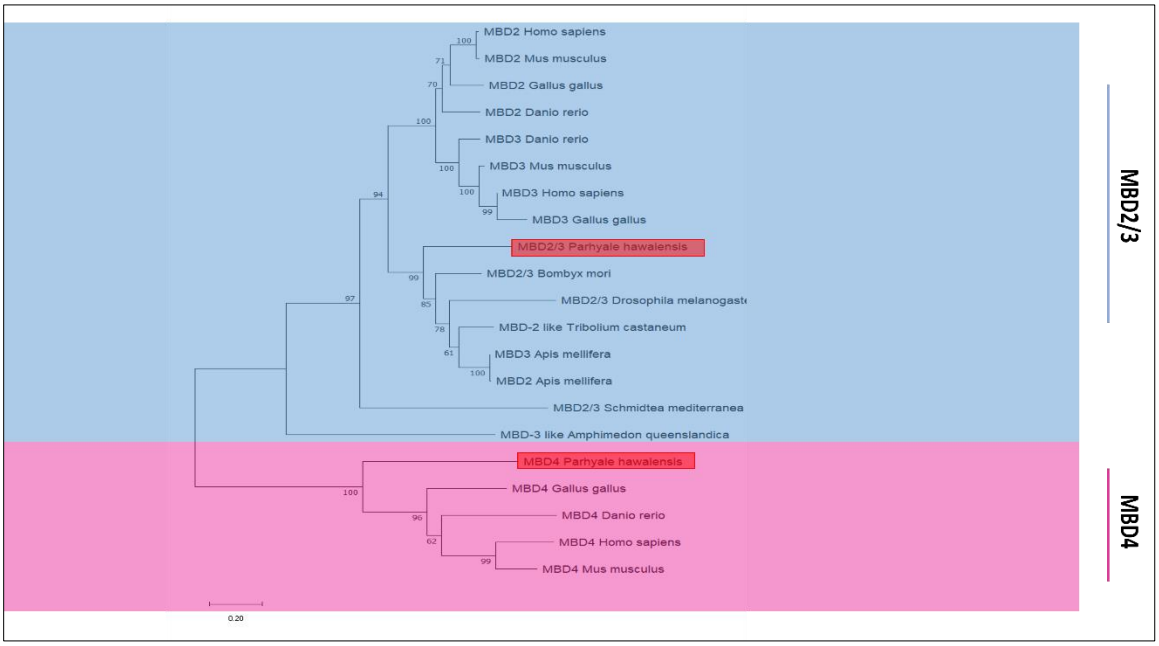
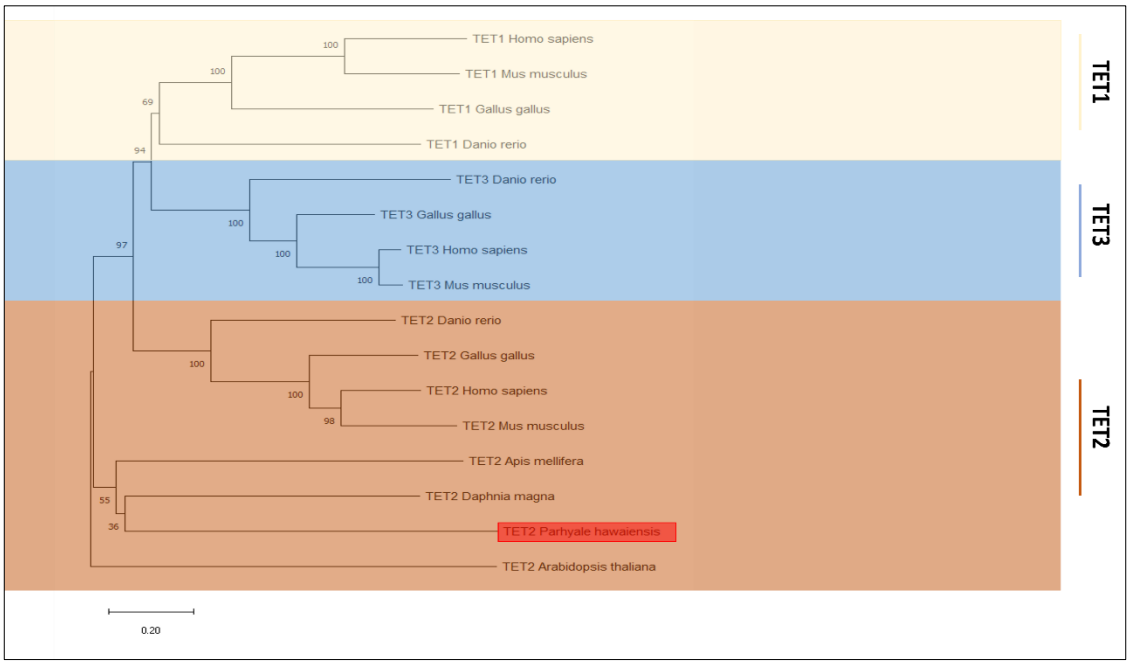


Figure 2.12. Phylogenetic tree of TETs. The tree represents the evolutionary relationships among TET proteins. The *Parhyale* genome encodes a single TET protein (TET2), which cluster together with the corresponding TET2 protein from other species. In contrast, vertebrate species typically possess two or three copies of TET protein.



Cloning and sequencing were conducted to obtain the full-length coding sequences (CDS) of DNMT1, DNMT3, and MBD2/3, aiming to confirm the sequences and identify any potential polymorphisms. Table5 provides the alignment percentage between the CDS obtained from Kao et al. 2016 and various cloned samples. Differences observed primarily corresponded to polymorphisms, as illustrated in Figure 9.

Table 2.5. Cloning results of coding sequence

Gene symbol	CDS size	Sample1	Sample2	Sample3	Sample4	Sample5
DNMT1	5 kb	99.66%	99.7%	99.7%	99.56%	99.68%
DNMT3	1212bp	99.7%	99.9%	99.59%	99.78%	99.66%
MBD2/3	792bp	99.5%	96.7%	99.6%	99.6%	99.6%

2.6 Discussion

In this chapter, we have introduced the utilization of PacBio data, a novel approach for a crustacean genome like *Parhyale*, to improve the existing genome assembly that was rich with gaps. Additionally, we have performed an expression-based annotation of the updated genome. These efforts have resulted in more complete genome by closing over 70% of the gaps in the large *Parhyale* genome assembly and have facilitated more efficient annotation process.

By employing PacBio long-read sequencing in *Parhyale*, we have also opened up the possibility of DNA methylation profiling without the need for bisulfite conversion. Previous studies have demonstrated that PacBio circular consensus sequencing (CCS) can detect DNA 5-methylcytosine (5mc) (Flusberg et al., 2010; Tse et al., 2021; Ni et al., 2022). This advancement offers promising prospects for investigating DNA methylation in *Parhyale* and further enhances our understanding of epigenetic regulation in this species.

The genome of *Parhyale* is considered one of the largest genomes sequenced, estimated to be around 3.6 Gb in size. It is characterized by a high abundance of repetitive sequences and exhibits significant levels of heterozygosity. Achieving a satisfactory level of completeness in the genome assembly was crucial to facilitate comprehensive analysis.

The large genome size of *Parhyale* can be attributed, in part, to the expansion of intron size within its genome. To improve the initial genome assembly, Dovetail technology was employed, which resulted in the generation of larger scaffolds instead of contigs. However, this transition introduced a considerable number of gaps that needed to be resolved in order to obtain a well annotated genome, particularly in the genic regions. Resolving these gaps was essential for conducting detailed analysis and interpreting the functional elements within the *Parhyale* genome.

In the next chapter, we employ our comprehensive genome annotation as template for analysing embryonic time-course data. Our goal is to detect the onset of zygotic transcription and identify the maternal to zygotic transition (MZT) in *Parhyale*. This analysis specifically relies on the availability of well-annotated introns, which serves as signals for the start of transcription.

Chapter III

Transcriptome analysis of early embryonic stages to describe MZT and ZGA

Contents

Abstract

Parhyale is a promising amphipod crustacean model for embryonic development studies. Foundational work has been described for *Parhyale's* embryogenesis, yet no detailed characterization of MZT/ZGA processes have been performed in this model organism. In this chapter, we describe the transcriptome of *Parhyale* embryos during early embryogenesis and propose a timeline for MZT events. We used RNA sequencing for multiple developmental timepoints to identify maternal mRNAs, the onset of ZGA, and relative stage at which the minor and major waves of ZGA occur. The MZT process in *Parhyale* begins a few hours after the start of embryogenesis, and the earliest zygotic transcripts were detected around the 32-cell stage, in agreement with observations from immunostaining against RNA polymerase II performed in *Parhyale* embryos (Nestorov et al., 2013). We observe two peaks of activation of the zygotic genome, the first around the 100-cell stage and the other at the start of blastodisc formation. Early transcripts of *Parhyale* are typically short, intron-less or intron-poor, newly evolved, and abortive, which is consistent with characteristics of early zygotic transcripts in other organisms such as Drosophila and zebrafish.

Next, we analyzed the expression patterns of DNA methylation machinery genes during MZT/ZGA of *Parhyale* and found that all the genes required for installing and erasing DNA methylation are provided maternally. Among the highly expressed genes during the early stages of embryogenesis are DNMTs and MBD2/3. However, as the embryo develops, they are downregulated to very low/no expression until the end of embryogenesis. In contrast, the key demethylation gene TET2 has a more fluctuating expression, with the highest levels of expression observed during the later stages of embryogenesis.

Finally, we correlated gene expression levels and DNA methylation levels before and after ZGA, using EM-seq data generated for embryos at multiple timepoints matching those used for gene expression analysis. We found a positive correlation between gene-body methylation and gene expression, with levels of DNA methylation reducing as embryonic development progresses.

We hypothesized that DNA methylation plays a crucial role in the embryogenesis of *Parhyale.* To investigate this, it's imperative to understand the initial stages of embryonic development thoroughly. This chapter, therefore, offers an in-depth description of the MZT/ZGA process, a pivotal phase of early embryonic development. The specific events of this process have not been meticulously studied in this organism before. To gain comprehensive insights, we harvested RNA from various early embryonic stages using two distinct RNA sequencing techniques, ensuring no interference with the normal transcriptome. This approach allowed us to detect *de novo* transcription using intron signal as a marker and to gather information about adenylation state of maternal transcripts. Beyond detailing maternal transcripts, the onset of ZGA, and the timeline of both minor and major ZGA waves, we explored the expression patterns of all DNA methylation machinery genes. Additionally, the data collated here served as input for further analysis, focusing on the overlap between gene expression and DNA methylation patterns. In summary, this chapter furnishes vital insights into *Parhyale's* development and establishes a positive correlation between gene expression and DNA methylation. This suggests that DNA methylation has a regulatory role in gene expression and *Parhyale's* embryogenesis. It further underscores *Parhyale's* potential as an ideal model to study the influence of DNA methylation on gene expression regulation during embryonic development.

3.1 Introduction

All animals load their oocytes with mRNA depicting a significant fraction of their protein-coding genes, to facilitate the beginning of embryonic development and trigger zygotic genome activation (ZGA) at the maternal-to-zygotic transition (MZT). This universal transition coincides with major molecular process such as epigenetic reprogramming, changes in chromatin organization and accessibility, dynamic regulation of coding and non-coding RNA, and specification of cell fates. MZT is characterized by multiple events, starting with elimination of maternal products and ending with global zygotic genome activation. This series of molecular events, although largely conserved among animals, can vary greatly in timing and can last from few hours in rapidly developing species, like *C. elegans* and *D. melanogaster*, to days in more complex species, such as humans and mice (Vastenhouw and Lipshitz, 2019). Moreover, in animals that set aside primordial germ cells (PGCs), there is an additional distinction in the spatial and temporal control of MZT between PGCs and the somatic cells (Vastenhouw and Lipshitz, 2019).

MZT prepares embryos for cell differentiation and later complex developmental processes. Therefore, the early embryonic gene expression program is primarily responsible for initiating gastrulation and specification of germ layers (Jukam et al., 2017). Maternal pioneer factors initiate the activation of zygotic genome gradually. This activation occurs simultaneously with the degradation of maternal mRNAs that controlled earlier embryonic development. Therefore, the regulation of stability and translation of maternal mRNAs, transcriptional activation, and cell-cycle length all need to be precisely coordinated during early development (Schulz and Harrison, 2019).

Although a significant fraction of protein-coding genes are loaded maternally into the zygote (estimates range from 40% in the mouse (Wang et al., 2004), 65% in *D. melanogaster* (Tadros et al., 2007), to 75% in *S.purpuratus* (Wei et al., 2006)), the first event in MZT is the elimination of those transcripts after a few hours of embryonic development (Figure 3.1). The elimination of maternal transcripts is achieved by two pathways. The first pathway is exclusively mediated by maternal factors, while the second pathway depends

on zygotic activity that destabilizes and clear additional maternal mRNAs. This observation is clear in *D. melanogaster* where 20% of the maternal products are cleared by maternal mRNAs activity only, zygotic transcription then clears an additional 15% of maternal mRNAs (Tadros et al., 2007). In zebrafish and *C.elegans*, three waves of maternal mRNAs clearance can be observed before, during, and after the onset of ZGA (Ferg et al., 2007; Mathavan et al., 2005; Baugh et al., 2003).

ZGA occurs gradually and increases in successive waves, usually a minor and a major wave of activation. Mouse and sea urchin embryos begin ZGA the earliest in terms of embryonic mitosis, with transcriptional onset at the early 2-cell stage (Aoki et al., 1997; Tadros and Lipshitz., 2009). *D. melanogaster,* on the other hand, initiates its minor wave of zygotic activation during the 8^{th} cleavage cycle. However, with regard to absolute time, very rapid cleavage cycles in early embryogenesis of *D. melanogaster* means that ZGA occurs several hours in the fly embryo rather than in the mouse where it takes over 24 hours for the first division (Li et al., 2022; Tadros and Lipshitz., 2009). Zygotic genes are found to be divided into two categories: strictly zygotic mRNAs and other mRNAs that are loaded maternally and re-expressed in early embryos, as seen in *D. melanogaster*, sea urchins and *C. elegans* (Wei et al., 2006; De Renzis et al., 2007; and Baugh et al., 2003). The set of strictly zygotic mRNAs is usually enriched for transcription factors, expressed during the minor and major wave, as observed in *D. melanogaster* and sea urchin (Tadros and Lipshitz., 2009).

Transcriptional inhibitors were applied to early embryos in species like *D. melanogaster, C. elegans,* and mouse to assess the developmental role of ZGA (Braude et al., 1979; Edgar and Datar., 1996; Seydoux and Fire, 1994; Seydoux et al., 1996). All of these studies concluded that in the absence of zygotic transcription, development continues to proceed normally until the time of major ZGA, after which embryos begin to experience failure in progressing through embryogenesis. This implicates the importance of zygotic transcription in priming the activation of genes required for gastrulation and establishment of the body plan (Tadros and Lipshitz, 2009).

As seen above, the scale and dynamics of MZT phases vary across species with respect to developmental stage and time (Figure 3.1). While a couple of studies have explored the MZT process in *Parhyale* broadly (Nestorov et al., 2013; Calvo et al., 2022), an accurate characterization of MZT/ZGA timeline and dynamics has not yet been performed in this species.

RNA processing

Transcription and translation are the two core processes that interpret the genetic code into functional proteins. Transcription, is the initial step in decoding genetic information, involves copying a segment of the DNA sequence into an RNA transcript via RNA polymerase II (Pol II) (Hampsey, M. 1998). This process begins at a specific section of DNA called the transcription start site (TSS), located at the 5' end of the gene within the core promoter region. The transcription machinery, comprising Pol II and general transcription factors, binds to the TSS in core promoters to kickstart transcription. additional factors like enhancers, transcription factors, and chromatin organization also play a role in enhancing transcription (Banerji et al., 1981).

Transcription culminates in a primary transcript, a single-stranded RNA, which subsequently undergoes processing to yield various mature RNA products. Precursor mRNA (pre-mRNA) is the type of primary transcript that evolves into messenger RNA (mRNA), which is eventually translated into a protein molecule. The transformation of pre-mRNA into mature mRNA involves several modifications, such as splicing - the removal of non-coding regions (introns) and the subsequent rejoining of coding regions (exons). The splicing process is facilitated by a large RNA-protein complex known as the spliceosome, which acts at conserved splice sites at the 5' and 3' ends of introns (Lamond A. I., 993).

Additionally, pre-mRNA processing encompasses capping and the incorporation of a polyA tail. After these steps, the mature mRNA transitions from the nucleus to the cytoplasm for translation. mRNA capping, which adds a 7-methyl guanosine at the mRNA 5' end, is one of the earliest processing steps, occurring simultaneously with transcription as the first 25-30 nucleotides of the emerging transcript are synthesized

(Shatkin & Manley, 2000). This capping not only plays a pivotal role in initiation of protein synthesis but also aids in recruiting factors linked to splicing, polyadenylation, and nuclear export. Moreover, it safeguards mRNA from degradation and shields it from the innate immune system's recognition.

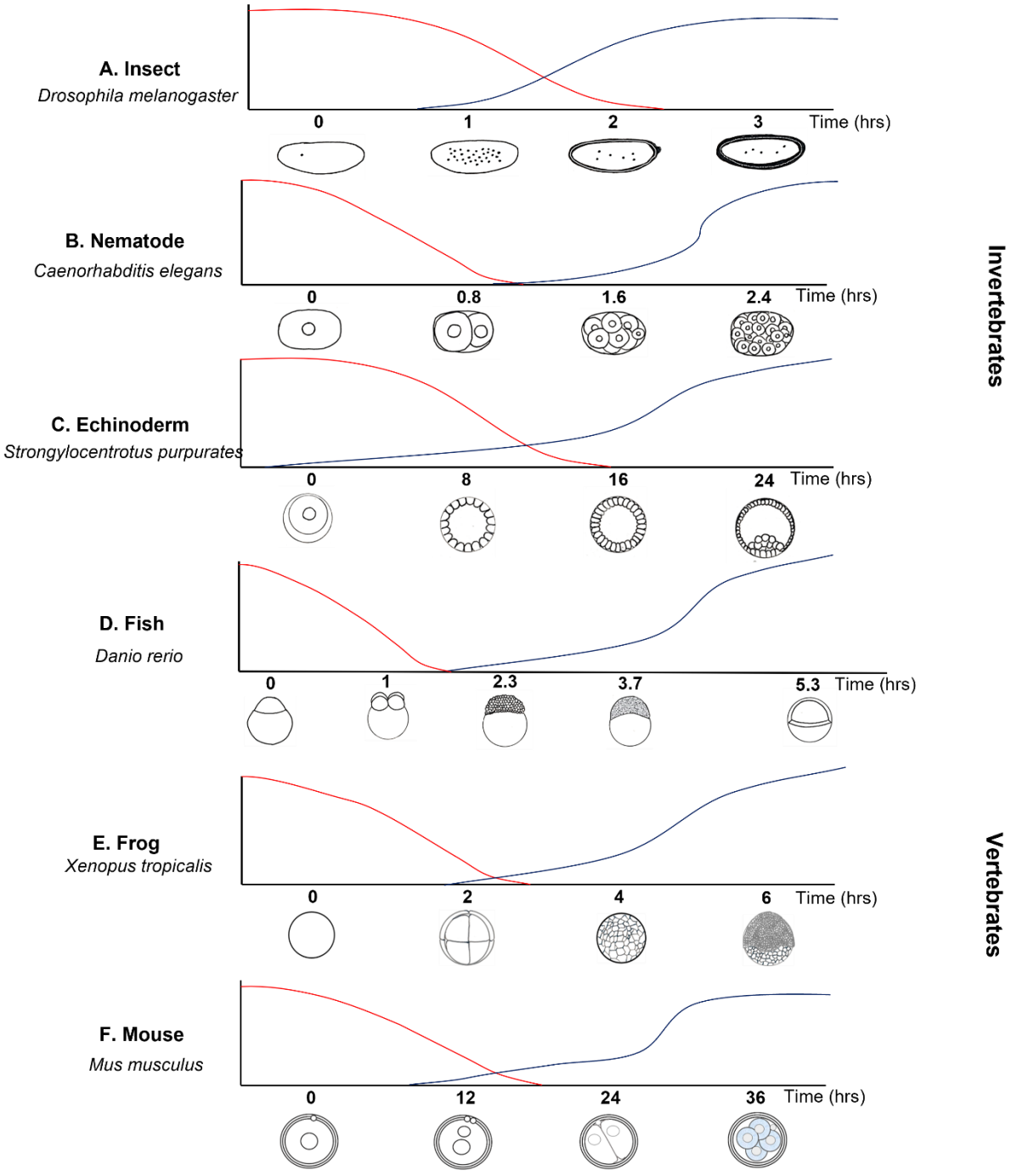
Polyadenylation entails adding a polyA tail to mRNA. This tail comprises only adenine bases. With a few exceptions like some mammalian histone transcripts, nearly all eukaryotic mRNAs possess a polyA tail (Passmore and Coller, 2021). These tails, essential for mRNA export from the nucleus to the cytoplasm, influence mRNA stability and translation. They act as crucial post-transcriptional gene expression regulators; a shorter polyA tail correlates with diminished transcript translation (Passmore and Coller, 2021). For instance, maternal mRNAs in sea urchins, which typically have shorter polyA tails, are translationally silent. Their expression is triggered when their polyA tails are elongated, underscoring the tail's influence on mRNA translation (Rosenthal et al., 1983).

Lastly, translation takes place, resulting in the synthesis of proteins from mRNA. This event transpires within ribosomes and comprises three phases: initiation, elongation, and termination. Initiation commences with the binding of initiation factors to the ribosome's smaller unit, forming a pre-initiation complex. This complex then attaches to the mRNA and the transfer RNA (tRNA) that carries methionine, marking the start codon. During elongation, amino acids are sequentially incorporated into the growing peptide chain as directed by the corresponding codons. Termination unfolds once all the mRNA codons have been interpreted by the tRNA molecule.

To define the factors that govern MZT and ZGA, we aimed to characterize the early transcriptome of embryos, differentiating between maternal mRNAs and the earliest transcribed genes. We integrated polyA+ and total RNA sequencing across multiple time points during early embryogenesis to describe how the gene expression program of *Parhyale* embryos is shaped during MZT.

We theorized that since maternal mRNAs would primarily be spliced during oogenesis, assessing introns from total RNA would help quantify *de novo* transcription. Furthermore, amalgamating polyA+ and total RNA sequencing would provide insights into the adenylation state of the maternal mRNAs.

Figure 3.1. Displays the MZT and ZGA timeline in different model organisms. For each model organism, time after fertilization is displayed with the relevant developmental stage. The red line represents the degradation of maternal products, and the blue line illustrates the timeline of zygotic genome activation from minor to major wave in each species. As shown, the window of MZT/ZGA is variable between species with respect to time and embryonic developmental progression, highlighting the diversity and complexity of MZT/ZGA across different model organisms.



3.2 Experimental design to study MZT and ZGA in *Parhyale hawaiensis*

To gain a comprehensive understanding of MZT and ZGA in *Pahyale*, we selected embryos from various key time-points that covered the early stages of *Parhyale's* development (Figure 3.2 E). Our selection was based on observations made by Nestorov et al. in 2013, who conducted an antibody staining against the Ser2-phosphorylated RNAPII to detect early transcription in *Parhyale* embryos (Figure 3.2 A). The first signal of transcription was detected in some cells of the 32-cell stage embryos. By the 100-cell stage, a uniform signal was detected throughout the embryos. The authors proposed that MZT takes place around the 32-cell stage, and the zygotic genome starts controlling embryogenesis by the 100-cell stage of development.

We followed the developmental staging table of the Extavour lab to accurately stage the embryos we collected. This table, created by multiple *Parhyale* scientists including Matthias Gerberding, Bill Browne, Nipam Patel, and Casandra Extavour, provides precise timing for significant developmental events, allowing us to collect embryos at specific time points.

Our first time point covers the 1 to 16 cell stages, corresponding to the developmental stages of *Parhyale* from S1 to the start of S5 as described in Browne et al., 2005. Based on Nestorov et al. (2013) staining observations, we assume that the zygote is still transcriptionally inactive at this time and exclusively contains maternally deposited RNA products (denoted 0-9). The second time point comprises 32-cell and 64-cell stages (both time points in middle of S5), as mentioned earlier, this time point was suggested to be the starting point of MZT (denoted 11). The third time point is exclusively 12 hours of development (S6), where genes from the minor wave of ZGA are expected to be detected (denoted 12).

The next time point, denoted 13-14, covers 13 and 14 hours of development (between S6 and S7). During this time, significant events include the migration of blastomeres. We assume that capturing more of the earliest minor wave genes would be possible at this stage. The fifth time point, denoted as 16, covered S7, during which major cell migration events take place, marking the start of blastodisc formation. The next

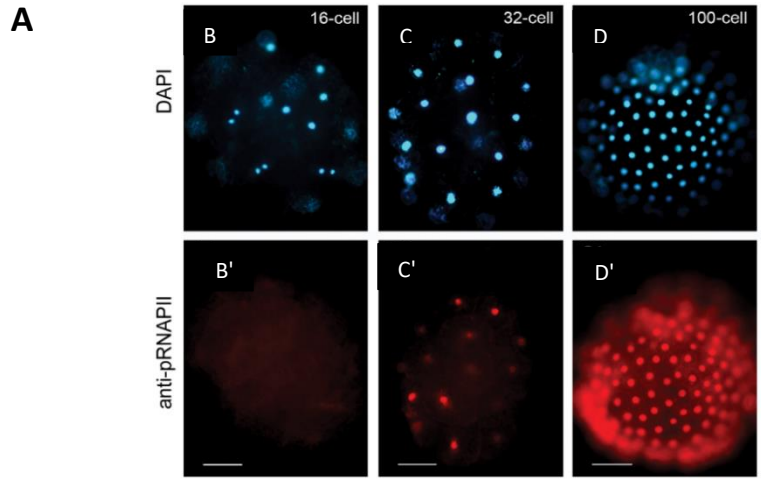
two time points were between S7 and S8, during which the onset of gastrulation occurs at 20 hours and 16 minutes. By 24 hours of embryogenesis, the germ-disc condenses and becomes visible (denoted 20, and 24, respectively). Finally, the last time point spans from 24 to 60 hours of development (~S8 – S11). This time period begins with gastrulation, and several events emerge, including the migration and segregation of embryonic cells from the yolk cell. Subsequently, the formation of germband begins. This stage is denoted 24-60. A detailed description of the collected stages is provided in Table 3.1 below.

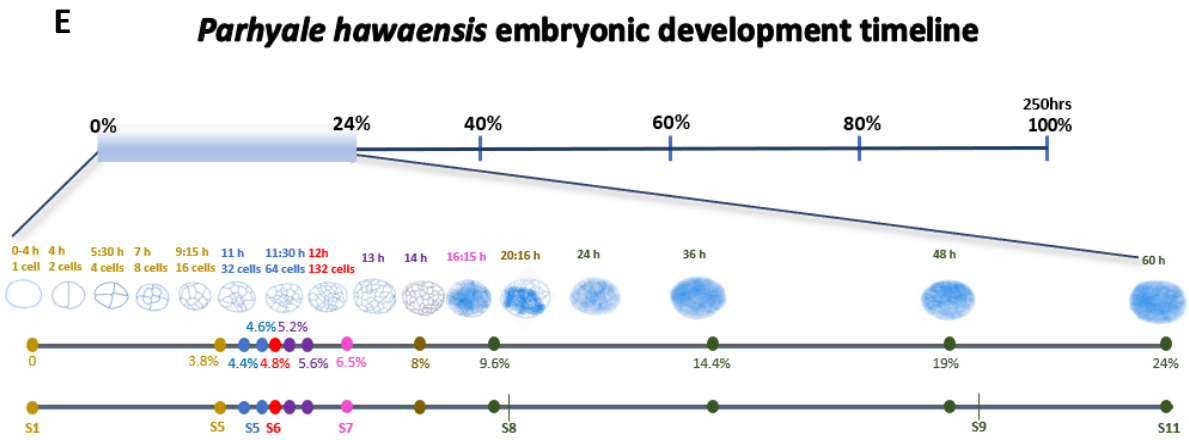
For each timepoint, we generated RNA-seq libraries using both polyA+ and total RNA sequencing strategies. Total RNA sequencing allowed us to detect all transcript species, while polyA+ RNA sequencing enriches transcripts with poly(A) tails likely indicative of mRNAs. Additionally, total RNA sequencing allows for higher detection of intron derived RNA, which can be used as a marker for *de novo* transcription and therefore new zygotic transcripts (Lee et al., 2013; Graf et al., 2014). Integrating data from polyA+ and total RNA sequencing will help provide information about poly(A) tail length, a crucial component of the MZT. In total, we generated 42 RNA-seq libraries library sizes ranging between 19 and 45 million reads. The average mapping rate was 83%, with approximately 63% unique reads and ~8% of multi-mapped reads. The number of multi-mapped reads was higher in total-RNA than in polyA+ samples and was also positively correlated with developmental time (refer to Appendix F for more information).

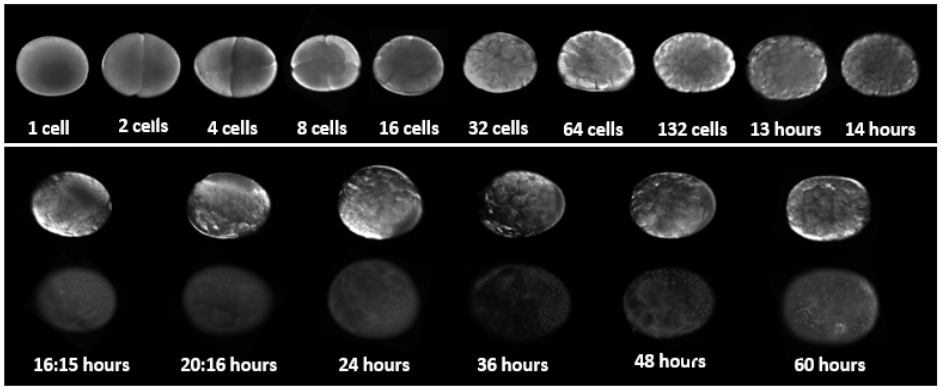
Table 3.1. Summary of early embryogenesis stages used for transcriptome sequencing.

Sequenced order	Developmental stage	Biological definition	Hours post fertilization	Percentage of embryogensis	Stage label
First	S1	1 cell	0-4h	0-1.6%	0-9
	S2	2 cells	4 h	1.6-2.4%	
	S3	4 cells	5.30 h	2.4-3%	
	S4	8 cells	7h	3-3.6%	
	Early S5	16 cells	9.15 h	3.8%	
Second	Mid S5	32 cells	11 h	4.4%	11
	Mid S5	64 cells	11.30h	4.6%	
Third	S6	132 cells (Soccer ball)	12h	4.8%	12
Fourth	Between S6 &S7	Blastomere migration	13h	5.2%	13-14
	between S6 &S7	Blastomere migration	14h	5.6%	
Fifth	S7	Start of blastodisc formation (Rossette stage)	16h & 15 minutes	6.5%	
Sixth	Between S7&S8	Gastrulation onset	20h & 16 minutes	8%	20
seventh	1h before S8	Gastrulation is visible	24h	9.6%	24
Eighth	1h before S8	Gastrulation is visible	24h	9.6%	24-60
	Between S8&S9	Germ disc tightening	36h	14.4%	
	2h before S9	Prior to head lobes appearance	48h	19%	
	S11	Germband formation	60h	24%	

Figure 3.2. Experimental design used to investigate the MZT/ZGA in *Parhayle hawaiensis* embryos. **(A)** Adapted figure from Nestorov et al., 2013, showing Immunodetection of early embryonic transcription using immunostaining against active RNAPII. **(B-D)** DAPI-stained nuclei of embryos at 16, 32 and 100-cell stages. The first sign of active transcription was detected at the 32-cell stage **(C')** with full detection in all cells at the 100-cell stage embryo **(D')**. Panel **(E)** Provides an overview of *Parhyale's* embryonic development timeline and highlights the stages during early embryogenesis used for transcriptome sequencing.







3.3 Data Description

Based on Pearson correlation of the samples with different developmental stages, we observed high reproducibility between replicates of each stage, as shown in Figure 3.3 A and B. The highest correlation was observed between consecutive stages at the early developmental stages. However, as embryogenesis progressed, this correlation decreased. The lowest correlation between consecutive stages was observed between stages (13-14 and 16) and between stages (24 and 24-60), indicating significant time-based transcriptional changes during these stages of embryogenesis. The correlation between replicates and stages was similar in both library types, but it was slightly higher in total RNA-seq, except for correlation between stages (12 and 13-14), where the polyA+ RNA-seq correlation coefficient is slightly higher than that of the total libraries. This lower correlation for polyA+ at this time could be due to dynamic and rapid changes in maternal polyadenylation states.

Principal components analysis (PCA) confirmed high similarity between replicates and showed that the main source of expression variation was the developmental stage (PC1, 86%), with greater similarity observed among the first four stages. Based on the PCA plot, we can identify three distinct expression subsets: early (0-14 hours), intermediate (16-20 hours), and late (24-60 hours). The PCA plots for both library types resulted in the same sample separation of expression data (as shown in Figure 3.4 A and B).

We quantified gene expressions as Transcripts Per Million (TPM) for exons and introns separately. Intron-based mean expression values showed a clear increase as the embryo developed, with a more visible increase in the later stages (16 hours and later) in the total RNA-seq libraries (as depicted in Figure 3.5 B). On the other hand, exon counts had a similar distribution across all stages in both libraries. However, in the polyA+ libraries, stages (11, 12 and 13-14 hours) exhibited a higher inter-quartile range with slightly lower mean expression than the other stages (as shown in Figure 3.5 A).

We examined the correlation of gene expression between polyA+ and total RNA-seq libraries for each timepoint by comparing TPM values (Figure 3.6). In terms of exon-based expression correlation, we found that transcripts from stages 0-9 to 13-14 hours exhibited variable TPM values, with some showing higher values in the total RNA-seq library and others showing higher values in the polyA+ library (Figure 3.6 A). This variability may be attributed to changes in the adenylation state of maternal transcripts or the presence of new transcripts that are still in the pre-mRNA stage, which occurs during stages when transcription is activated. However, starting from stage 16, there is a more linear correlation between polyA+ and total RNA-seq libraries. This indicates that there are more consistent TPM values for a greater number of genes across the two libraries. In contrast, intron-based expressions exhibit a lower correlation coefficient. Furthermore, intron reads are more frequently detected in the total RNA library from stage 12 onwards, which is likely attributed to the initiation of zygotic transcription at that point (as shown in Figure 3.6 B).

Figure 3.3. Heatmap showing the correlations between all replicates across all sequenced stages in (A) PolyA+ RNA-seq and **(B)** Total RNA-seq datasets. The replicates of each stage in each sample display a high level of correlation. Moreover, the correlation between the early stages is higher when compared to the later sequenced stages.

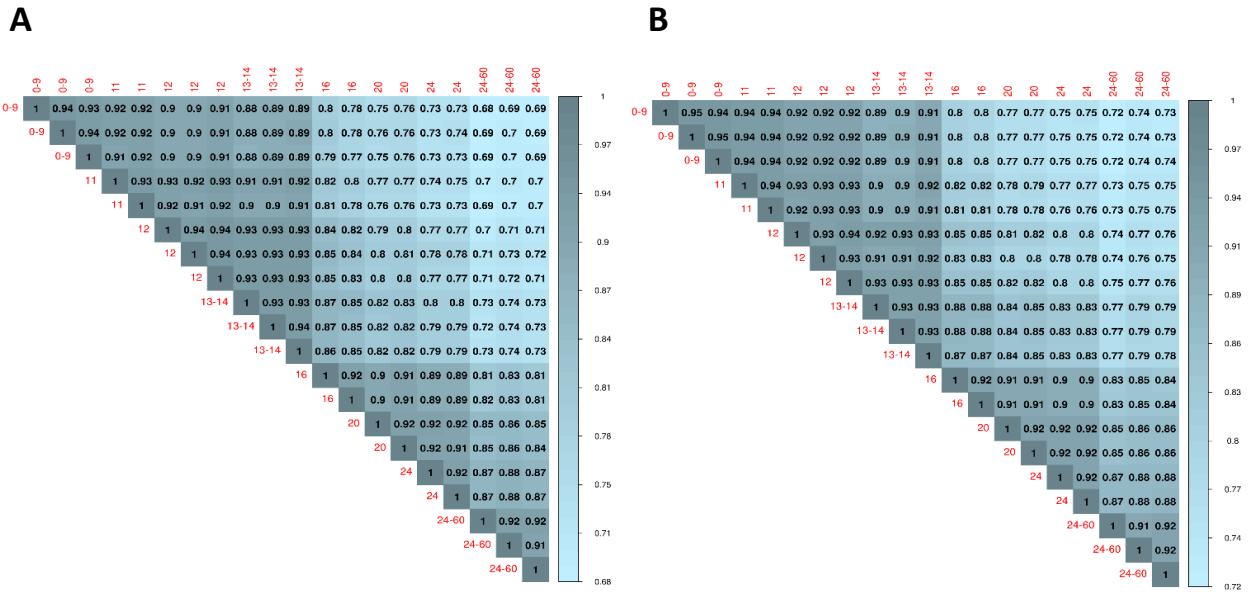


Figure 3.4. Principal component analysis (PCA) plot of all studied stages generated using the top one thousand most variable genes in **(A)** PolyA+ RNA-seq and **(B)** Total RNA-seq datasets. The same pattern is observed in both datasets, and the samples can be classified into early (0-14 hours), middle (16-24 hours), and late (24-60 hours) expression subsets.

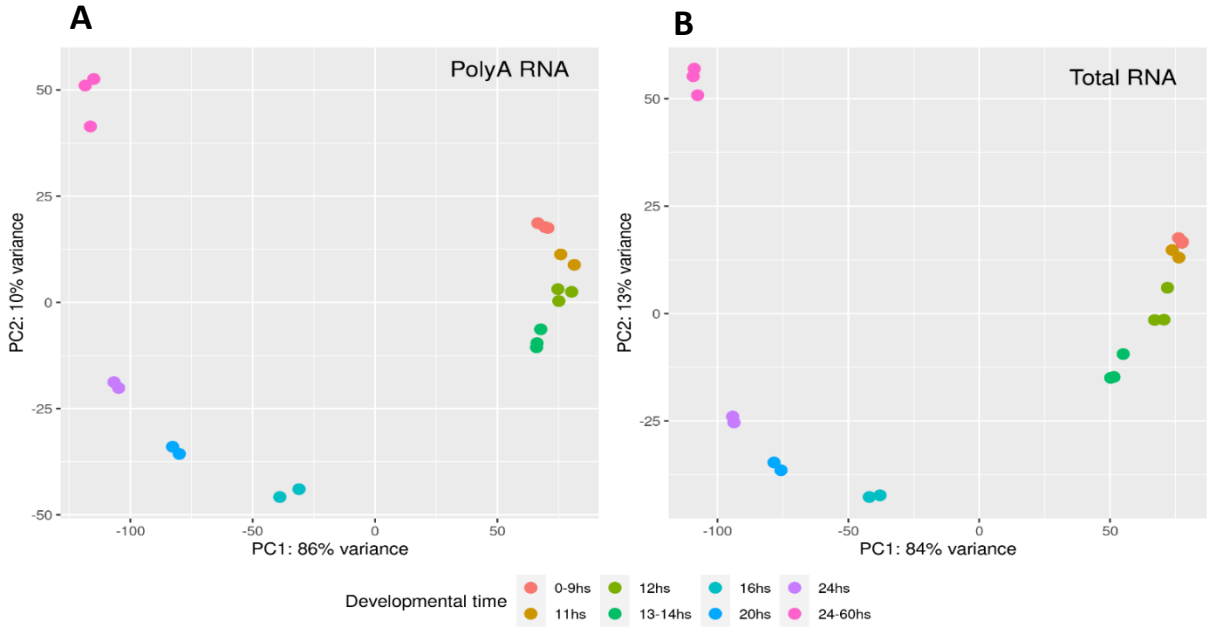
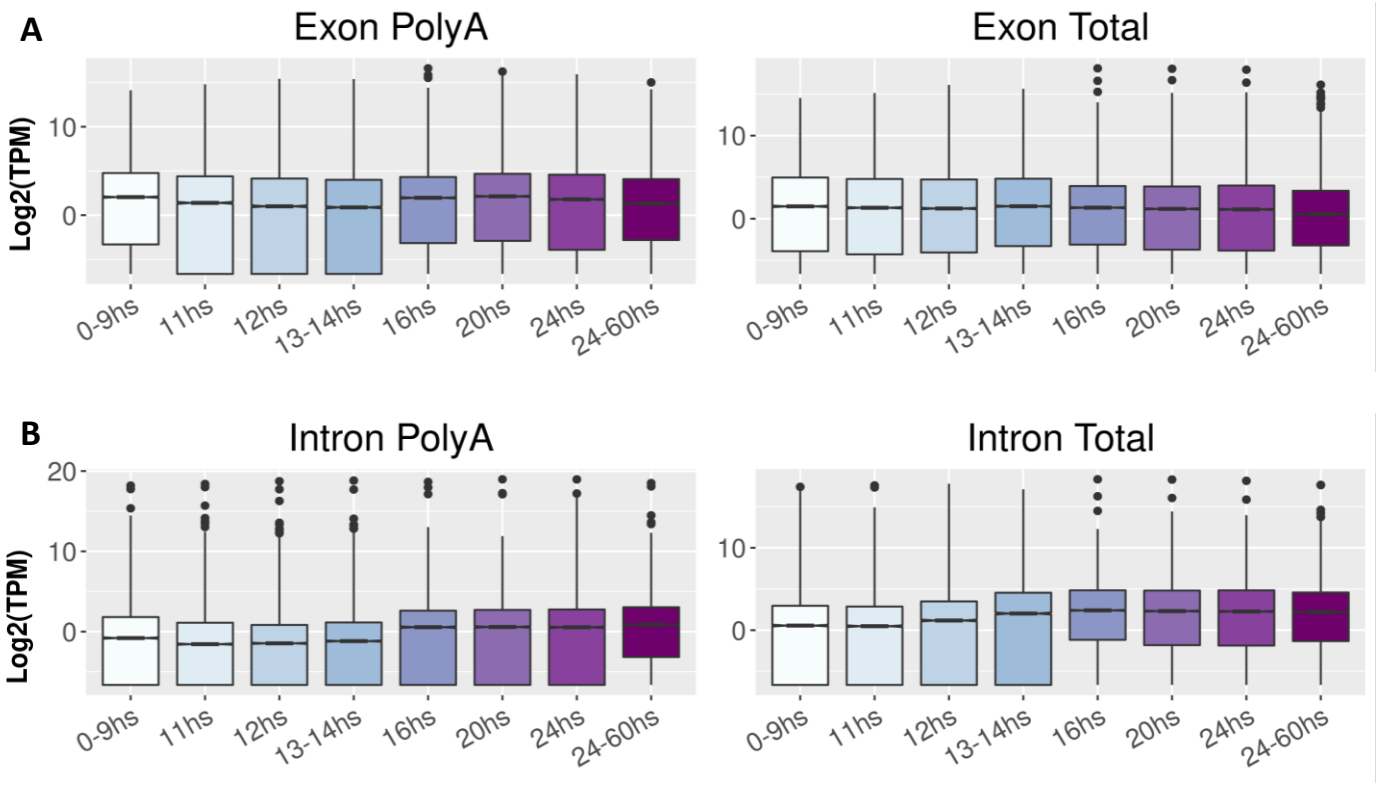
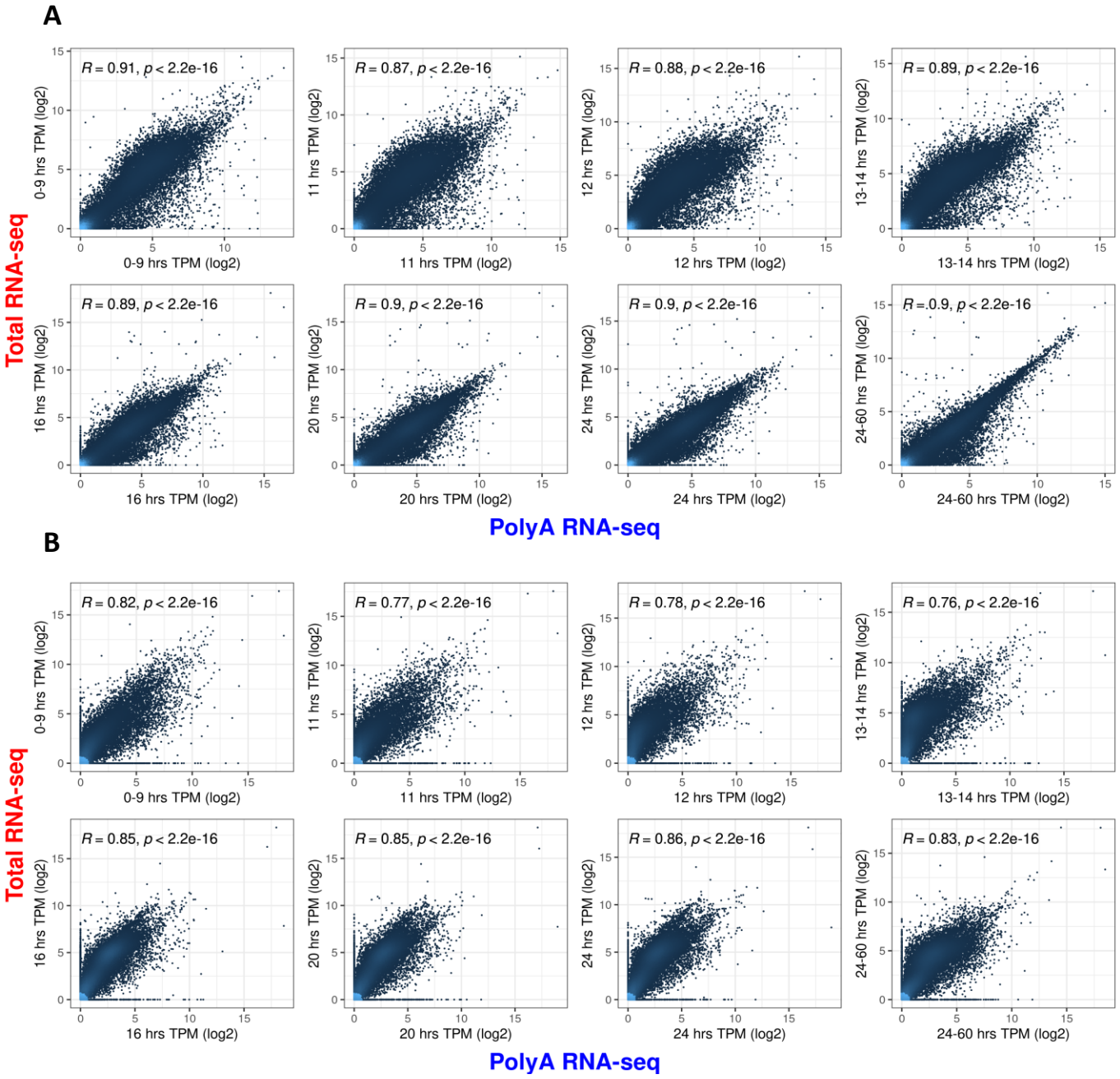


Figure 3.5. TPM distribution across all sequenced developmental stages, including both exon-based **(A)** and intron-based **(B)** distributions in both library types. The polyA+ count distribution is shown on the right. Exon-polyA+ count distribution shows a drop between stages 11-14 hours. This drop may be due to the clearance of maternal products, as maternal mRNAs are translated during the first stage (0-9 hours). However, the later increase in distribution from 16 hours onwards indicates activation of transcription. Some of the maternal genes may get activated zygotically after ZGA. The total RNA dataset's exon-based TPM distribution is more similar between all stages, while intron-based count distribution is lower in general compared to the exon-based count distribution. The difference between early and later stages is more pronounced in the total RNA dataset than in the polyA+ dataset. The increase in intron-count distribution from stage 16 hours onwards is likely due to zygotic transcription activation. Overall, the TPM distribution suggests that maternal mRNAs are cleared during early development stages, with a coupling of gradual zygotic transcription activation.



<u>Figure 3.6. Dot-plot of the expression correlation between polyA+ and total RNA-seq data at each timepoint</u>, with both exon-based **(panel A)** and intron-based **(panel B)** TPM data presented. The correlation coefficient was calculated using the Spearman test. The exon-based TPM data exhibits a higher correlation than the intron-based TPM data. However, it is important to note that many genes have higher TPM values in the total RNA-seq dataset due to its higher coverage of all RNA species, rather than just polyA-enriched mRNAs in polyA+ dataset. The most consistent correlation is observed for stages from 16 to 24-60 hours in exon based TPM values, where zygotic transcription is likely fully activated.



3.4 Identification of maternal mRNAs

Various biological process taking place before ZGA are largely controlled by maternal mRNAs and proteins (Liu et al., 2021; Eckersley-Maslin et al., 2018; Hendrickson et al., 2017). Since this stage lacks active transcription, maternal mRNAs are mainly regulated by post-transcriptional mechanisms such as polyadenylation and deadenylation that controls polyA tail length (Liu et al., 2021; Eckmann et al., 2011). The polyA tail is an essential regulator of translation and stability of mature mRNA transcript (Liu et al., 2021; Eckmann et al., 2011; Weill et al., 2012). Therefore, the fate of maternal mRNAs is tightly deteremned by polyA tail length, and so we expect to see polyA tail dynamics during early stages of development.

To identify maternal transcripts in *Parhyale* embryo, we integrated both polyA+ and total RNA-seq datasets. We set a filtering criteria for transcripts to define as maternally expressed if they have 5 TPM or higher detected in stage (0-9 hours) (Appendix G). We identified 11,007 transcripts as maternal and potentially maternal-zygotic genes as they could be provided maternally and transcribed zygotically later in the early embryo of *Parhyale* (transcripts id listed in Appendix G).

PolyA+ libraries highly enrich for polyadenylated transcripts, and therefore may miss non-polyA transcripts, such as immature pre-mRNAs or long non-coding RNAs. In contrast, total RNA-seq libraries rely on ribosomal RNA depletion followed by random hexamer priming during reverse transcription and thus provide more information on both polyA and non-polyA transcripts. Therefore, we integrated both datasets by calculating the polyA+/total RNA average expression ratios (Figure 3.7A) to estimate the relative adenylation state of maternal mRNAs. Our aim was to confirm whether differences in expression values detected for some genes between library types (Figure 3.6) corresponded to regulatory signatures of polyA tail. For example, transcripts that are expressed in total RNA library but not polyA+ library at the same timepoint are likely to lack a polyA tail due to deadenylation or not yet being polyadenylated. On the other hand, the PolyA+ library always has higher coverage for polyadenylated protein-coding genes.

We found a gradual decrease in the genome-wide average polyA+/total RNA ratios (Figure 3.7), indicating a widespread deadenylation of maternal transcripts during stages between 0-9 hours and 13-14 hours. In contrast, during stage 16 and 20 hours, the polyA+/total RNA ratio increased, suggesting an increase in polyadenylated transcripts. After the 20 hours stage, the polyA+/total RNA ratio approached balance, in which gastrulation start to take place, and the polyA tail regulation becomes less intense as development is controlled mainly by zygotic transcription. This finding is in agreement with Subtelny et al.'s (2014) observation that polyA tail length was coupled with translational efficiency in early embryos of zebrafish and frog, and this strong coupling was diminshed by gastrulation time, indicating a switch in the translational control nature.

We categroized transcripts based on their polyA+/total RNA ratios, distingushing between polyA+ biased (polyA+/total RNA ratio ≥ 1.5) and total RNA biased (polyA+/total RNA ratio ≤ 0.5) to thourougly examine the polyA tail state (Figure 3.7 B). At the first stage (0-9 hours), there were 3,168 polyA+ biased genes and 2,343 total RNA biased genes. This indicates the presence of both polyadenylated and deadenylated or not yet polyadenylated transcripts at this developmental stage. Between stage 11 and 13-14 hours, more genes showed total RNA bias compared to polyA+ bias genes (Figure 3.7 B). This suggests a deadenylation process that may contribute to the clearance of maternal mRNAs, likely paired with the activation of new zygotic transcripts. Starting from 16 hours of development, polyA+ biased genes dominte until the 24-60 hours stage (an average of 4,223 polyA+ biased genes compared to only 354 total RNA biased genes). These observation indicate dynamic polyA tail changes regulating maternal mRNAs during the early embryonic stages of *Parhyale*.

To examine maternal genes with dynamic polyA tail changes, we conducted K-means clustering on maternal genes using their polyA+/total RNA ratios for each gene at stages between 0-9 and 16 hours (Figure 3.8 A). This allowed us to identify six clusters of maternal transcripts based on their adenylation patterns. Notably, we observed a significant prevalence of deadenylated transcripts during the 13-14 hour stage, followed by an increase in polyadenylation at the 16-hour stage.

Cluster 1 (n=1552) and 4 (n=1888) consisted of genes polyadenylated at 0-9 hours that subsequntly underwent deadenylation. In cluter 1, these genes remained deadenylated. Whereas in cluster 4, genes were polyadenylated at a later stages (16 hours). The polyadenylation of these genes at early stages suggests their involvment in early embryogenesis.

Cluster 2 (n=1353) and 3 (n=1080) contained genes that exhibited polyadenylation at stages 11 and 12 hours, respectively, indicating their likely involvment in promoting ZGA.

Transcripts in cluster 5 (n=3155) mostly exhibted a deadenylation state from 0 – 14 hours stage, followed by polyadenylation at stage 16, suggesting a later activation of these transcripts. Notably, DNMTs and MBD2/3 were among the transcripts included in cluster 5, indicating a regulatory role during early *Parhyale* embryogenesis.

Finally, transcripts found in cluster 6 (n= 1091) underwent an exchange between deadenylation and polyadenylation during early stages, suggesting that this set of genes was translated maternally, degraded, and transcribed zygotically later at stage 16.

We compared our annotated list of maternal transcripts with the list of maternal transcripts in *D. melanogaster* identified by Lott et al. (2011) to assess their similarities. We found 1,756 common genes between the two species, including well-known maternal genes such as *nanos, smaug, brat, Bicaudal, Bicoid* and *pumilio,* which have previously been described in *D. melanogaster* (Bashirullah et al., 2001; Tadros et al., 2007; Laver et al., 2015; Suter et al., 1989; Lehmann & Nüsslein-Volhard, 1987; Berleth et al., 1988). These genes are expressed during the earliest stages of development (0-12 hours). Interestingly, MBD2/3 and DNMT2 were also common maternal genes between *Parhyale* and *D. melanogaster*, suggesting a conserved role of these genes during embryogenesis.

A group of the maternal genes are enriched for genes involved in RNA metabolism, such as many spliceosome components (*SF1, Prp19, Prp8, Prp18, SF2*), which are also found in Drosophila's maternal

transcripts list (Lott et al., 2011; Lécuyer et al., 2007). Moreover, the Staufen ortholog, one of the best characterized RNA-binding proteins (RBPs) that was first identified in *Drosophila*, was found on the maternal gene list of *Parhyale* in cluster 2 (Schüpbach & Wieschaus, 1986). This suggests that Staufen is an evolutionary conserved RBP across many species with a central role in the localization, transportation, and translation of mRNA (St Johnston et al., 1991; Kiebler et al., 1999; LeGendre et al., 2013).

Cluster 2, which consists of genes that are mainly polyadenylated at 11 hours followed by deadenylation in later stages, is enriched for genes involved in transcription, such as the *Parhyale zelda* ortholog, the poineer transcrition factor during ZGA, and the *Sox2* ortholog, which are among the upregulated genes at 11 hours (Duan et al., 2021; Lee et al., 2021). Among the common genes between *Parhyale's* and Drosophila's maternal transcripts are HDAC4 and HP1b, which are involved in transcriptional repression (Zeremski et al., 2002; Zenk et al., 2021). These genes are found in cluster 1, where genes are polyadenylated during first stage (0-9 hours), during which the zygote genome is still inactive, and then deadenylated during following stages.

In addition, we conducted a correlation analysis between TPM expression values of genes in each cluster at stage 0-9 and stage 11 hours, for each library type (as shown in Figure 3.8 B and C). The total RNA-seq data showed strong correlation (as seen in Figure 3.8 B), whereas the polyA+ data exhibted correlation patterns that matched the adenylation state of transcripts in each cluster. For instance, genes in cluster 1 and 4 had higher TPM expression values at the 0-9 hours stage, during which there is an upregulation of polyadenylation, followed by deadenylation in the 11-hours stage (as depicted in Figure 3.8 C). These findings suggest that changes in polyA tail length play a role in regulating gene expression.

The analyses performed in this section demonestrate that maternal transcripts undergo dynamic adenylation state changes and that the polyA tail is essential for the regulation of maternal mRNAs. However, it is also likely that nascent transcription is a source of upregulated polyA signal, particularly during stages 12 hours and beyond, where zygotic transcription starts to activate.

Figure 3.7. Changes in polyA tail length of maternal transcripts during early development. (A) Boxplot of polyA+/total RNA ratio for the average expression of all maternal genes detected in each stage. **(B)** Histogram showing the number of genes with a polyA+ bias (polyA+/total RNA ratio ≥ 1.5) or total RNA bias (polyA+/total RNA ratio ≤ 0.5) in each stage.

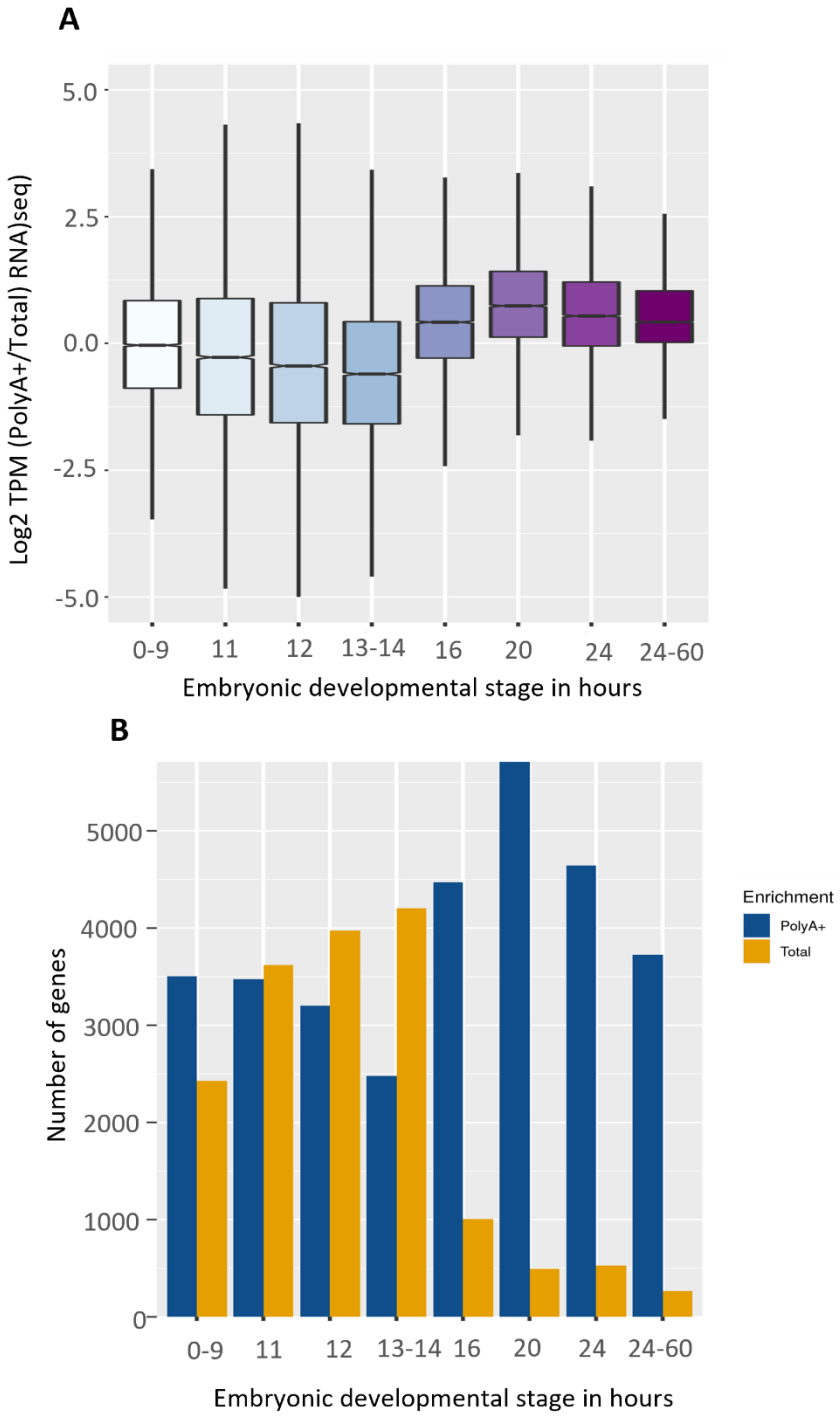
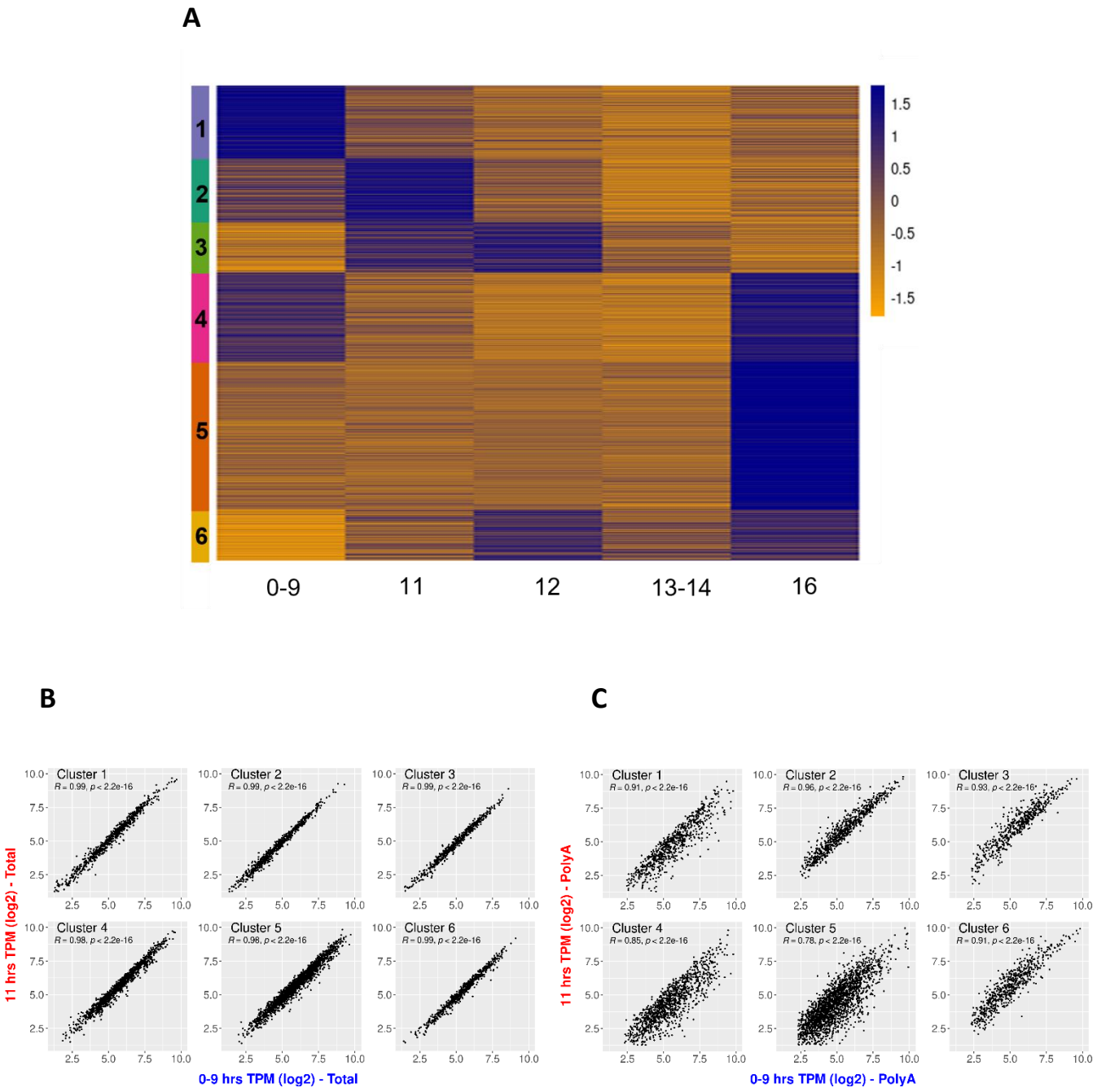


Figure 3.8. Changes in polyA tail length of maternal transcripts in *Parhyale*. **(A)** Heatmap showing maternal expression clusters based on the polyA+/total RNA ratio for each developmental stage. **(B)** and **(C)** Scatter plot showing the correlation of gene expression between 0-9 hours and 11 hours stage embryos using polyA+ or total RNA exon counts, respectively. Spearman correlation coefficients are shown for each cluster comparison.



3.5 Zygotic genome activation in *Parhyale*

3.5.1 Exon-polyA K-means clustering analysis

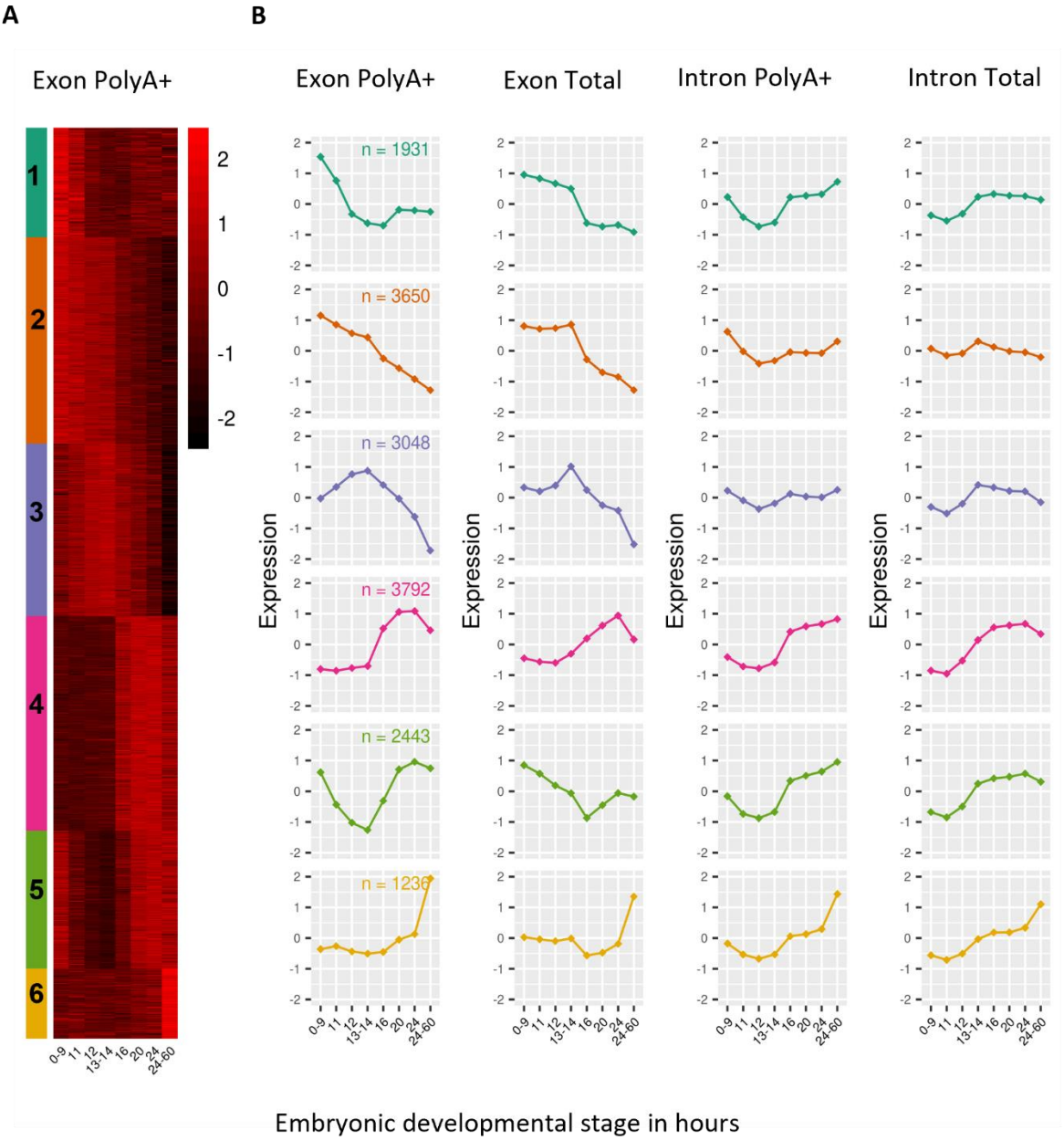
To analyze the transcriptomic landscapes during the early embryogenesis of *Parhyale* and identify factors that mediate zygotic transcriptional activation, we performed k-means clustering to partition genes based on their exon polyA+ expression patterns in all the sequenced stages (Figure 3.9 A). Only transcripts with an average TPM ≥ 3 across three consecutive developmental stages were included in the clustering analysis to exclude lowly expressed genes while preserving early zygotic transcription. Although an increase in exon signal could indicate zygotic transcriptional activation, polyadenylation of maternal transcripts could also contribute to the increase in exon polyA+ signal.

Our analysis identified six expression clusters. Clusters 3 and 4 exhibited mRNAs with an increase in exon signal at 12 and 16 hours, respectively, suggesting a potential minor and a major transcriptional onset during these timepoints (Figure 3.9 A). In cluster 3, genes were expressed from early stages and gradually upregulated until stage 13-14 hours of development, followed by downregulation in subsequent stages. In contrast, genes in cluster 4 were switched on starting from stage 16 hours, indicating that they were only zygotic and not maternally provided, whereas genes in cluster 3 were among the maternal transcripts that activated a few hours later zygotically. The timing of activation of these genes is likely the result of the MZT process. Cluster 1 and 2 contained genes highly expressed during early stages (0-14 hours) that downregulated in subsequent stages, suggesting a role for these transcripts during early embryogenesis. Genes in cluster 1 might be expressed later zygotically but at substantially lower levels, while genes in cluster 2 showed no signal of activation in later stages. Cluster 5 contained genes that were expressed during 0-9 hours, downregulated, and then reactivated around stage 20 hours at a comparable level to their expression during early stages. This group of genes could be maternally provided transcripts that were deadenylated or cleared after 9 hours of development and then activated zygotically later. Finally, cluster 6 comprised genes that were switched off or lowly expressed until around 24 hours, where they were strongly upregulated. These genes are likely required for advanced developmental events.

To gain a more comprehensive understanding of the transcriptomic dynamics during MZT/ZGA, we compared expression patterns of the clusters in both polyA+ and total RNA datasets for exons and introns in each library type (see Figure 3.9 B). Comparing exon versus intron signal, we found that the intron signal was better correlated between polyA+ and total RNA data than exon signal for all clusters. This slight variability in exon signal between the two library types could be explained by the enrichment of polyA transcripts in the polyA+ data, which makes changes in the polyA tail more obvious in the total RNA data. Cluster 3 showed a gradual increase in exon-polyA+ signal from 0-14 hours, while in the intron-total signal, the increase began from stage 12, peaked at stage 13-14 hours, and then dropped. Cluster 1 and 5 had a clearer distinction between exon and intron signal, where the signal increased in the early stages, dropped gradually until 14 or 16 hours, and then increased again in the exon signal. In the intron signal, the increase started only at stage 13-14 hours. This indicates adenylation changes during the early stages and transcription onset around stage 13-14 hours.

Although the exon-signal-based clustering analysis provided an indication about the fate of maternal mRNAs and the time points of potential transcription onset occurrence, it remains challenging to differentiate actual transcriptional activity from polyadenylation changes by relying on polyA+ data alone. To identify genes transcribed at the onset of ZGA, we used two different methodologies (Graf et al., 2014). The first strategy involved searching for the earliest new transcripts absent from the maternal total and polyA RNA to detect newly activated genes. The second strategy was using upregulation of intronic signal as a marker of *de novo* transcription.

Figure 3.9. K-means clustering of exon-polyA+ TPM counts using all expressed genes in at least one of the sequenced developmental time points. (A) Heatmap showing the clusters defined by the analysis. **(B)** Mean of scaled expression values in each time point for each cluster defined using exon-polyA+ counts. The expression behavior of genes in each cluster represented in both library types for each genomic feature (exon-intron).



3.5.2 Detection of first zygotic transcripts based on timepoint-specific expression

Maternal mRNAs were defined as transcripts where five reads or more were detected during the first timepoint (0-9 hours). Genes that were not maternally loaded were considered newly transcribed in the subsequent stages. For a gene to be classified as expressed for the first time in the embryo, it had to have fewer than five reads during the 0-9 hours stage and at least 20 reads in one of the stages after 0-9 hours. Therefore, the transcript abundance had to be differentially upregulated (FC >1) for the analyzed time point to be designated as first expressed. These genes were considered purely zygotic and not provided maternally. The number of activated genes detected was calculated for each stage between 11 to 24 hours using polyA+ and total RNA libraries. Figure 3.10 displays the number of activated genes in each library type, in addition to the number of genes detected in both libraries.

The analysis revealed that there were 38 activated genes at the 12-hour stage and 126 genes activated at 13–14-hour stage in the total RNA dataset only. In contrast, using polyA+ library, only one and 19 transcripts were detected at 12- and 13-14-hour stages, respectively. These transcripts likely represent the beginning of the minor wave of ZGA. The higher number of genes detected using the total RNA library compared to the polyA+ library can be explained by the fact that nascent RNAs may not be fully polyadenylated during the early stages of development, which makes them undetectable using the polyA+ library. This indicates that these transcripts are being transcribed but are polyadenylated later.

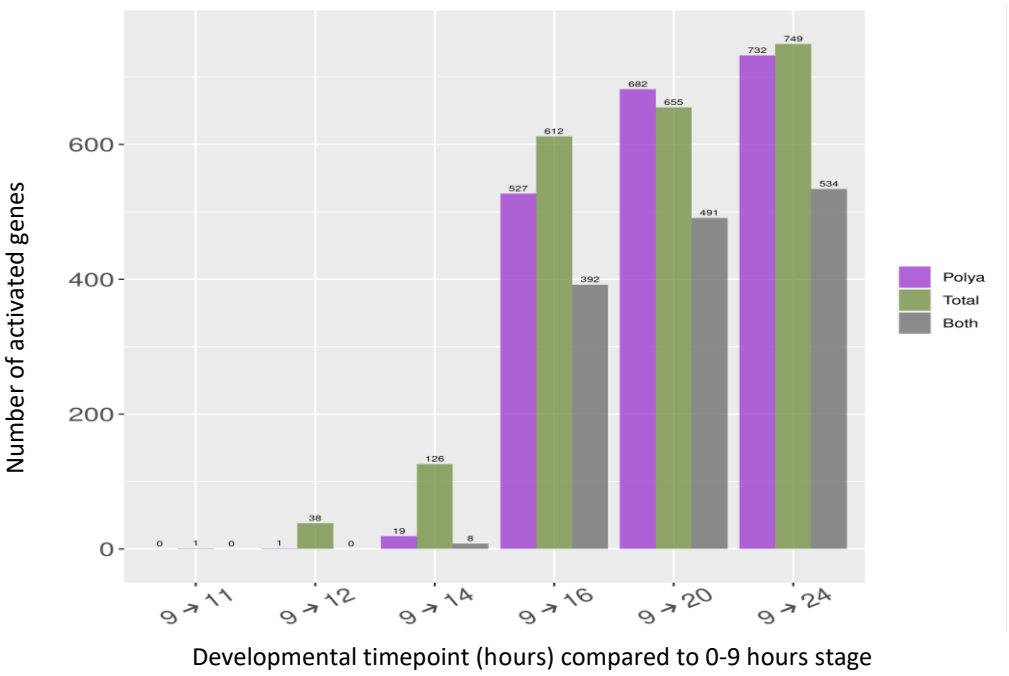
By looking at the identities of the genes activated during the 12- and 13-14-hour stages through blasting, we found that 40% of them did not have a significant hit in any other organism, suggesting that they could be *Parhyale*-specific genes. Of these, 22% had significant hit only in *Hyalella azteca*, suggesting that they could be amphipod-specific genes. This observation suggests that maternally provided *Parhyale* genes are mostly evolutionary conserved, while the earliest activated purely zygotic genes are likely species-specific genes. This phenomenon has also been observed in other species such as Drosophila, mouse, and zebrafish (Heyn et al., 2014).

It is interesting to note that a fraction of the earliest activated genes in *Parhyale* is enriched with reverse-transcriptase (RT) domain-containing proteins, which account for 27% of activated genes during 12–14-hour stages. These proteins belong to a group of enzymes capable of synthesizing DNA using retrotransposon RNA as a template. This observation suggests that retroelements may play a role during early development in *Parhyale*. This phenomenon has also been observed in mouse embryos (Sciamanna et al., 2011), further highlighting the potentially conserved importance of retroelements in early embryonic development across species. The paper showed that Inhibition of highly expressed LINE-1-encoded RT led to developmental arrest at the two- or four-stage mouse embryo. In addition to this group of activated genes, eight zinc-finger transcription factors (TF) were activated during 12 -14-hour stages, four of them are CCHC-type domain-containing TF, three RING-type domain-containing TF, and a single PHD-type domain-containing TF.

At the 16-hour stage, the number of activated genes increased significantly by 612, almost five times more than the genes activated at the 13–14-hour stage. Subsequently, the number of newly transcribed genes detected at 20 and 24 hours showed a slight increase of 1.1 times. Amon the genes detected from 16 hours onwards were twelve TF domain-containing genes involved in growth control and body patterning, such as orthologs for *FoxP, labial, gooseberry (gsb), cubitus interruptus (ci), aristaless*, and two zinc-finger *Parhyale* or amphipod specific. Moreover, the number of activated genes detected using polyA+ and total RNA seq libraries started to be highly overlap (around 70%) from the 16-hour stage, indicating sufficient production of full-length functional transcripts during the later stages.

Overall, these findings suggest that purely zygotic genes are transcribed in two waves, with the first wave consisting of approximately 120 genes at the 12-hour stage and the second wave starting from the 16-hour stage and involving around 600 genes.

Figure 3.10. Timepoint-specific detection of first zygotic transcripts. The expression of each gene at each timepoint was compared to the stage 0-9 hours reference. The datasets used were polyA+ and total RNA seq libraries. Starting from 16 hours, the number of activated genes detected using both library types become highly overlapping (~70%), indicating sufficient production of full-length functional transcripts.



3.5.3 Detection of gene activation by identifying precursor mRNAs

As previously mentioned, we employed another strategy to identify nascent zygotic transcripts. This method relied on detection of intronic sequences generated from incompletely spliced transcripts (Ameur et al., 2011; Graf et al., 2014). To distinguish between intronic reads arising from primary gene transcripts and those coming from independently transcribed repetitive regions, we defined a parameter called RINP as measure for the coverage of all the intronic sequences in each transcript (see Figure 3.11). RINP is an indicator of the **R**atio of **I**ntronic read counts to **N**ot-covered intronic **P**ositions. We measured RINP for all the sequenced timepoints in both libraries (Figure 3.12 A), a fold change ≥ 2.5 between consecutive stages was considered as a signal of nascent transcription. We defined the 75^{th} percentile of RINP in the 0-9 hours stage of total RNA data as the base background of intronic coverage.

This analysis revealed the activation of new zygotic genes as early as the 11-hour stage of development. Moreover, the activation of zygotic genes across stages was gradual and continuous until stage 16 hours, at blastodisc formation. Using the total RNA intron signal, we found 112 zygotically activated genes at 11-hour stage, 451 genes at 12 hours, and 1712 genes at 13–14-hour stage embryos (Figure 3.12 B). The number of detected genes using polyA+ data intron signal was always lower for these stages (11- 14 hours); however, at the 16-hour stage, the number of zygotically activated genes based on polyA+ intron signal increased dramatically and reached a peak with 3,853 genes activated (3,854 genes using total RNA intron signal). These results indicate that ZGA consists of a period where transcription is gradually activated, which has been observed in many model organisms, as mentioned in the introduction (Vastenhouw et al., 2019). Moreover, since this method relied on intronic reads as a proxy for zygotic activation, we were able to distinguish between transcripts that are maternal only, zygotically activated only, or maternally loaded and zygotically transcribed.

Among the genes activated at 11 hours is the *Parhyale* ortholog for *brat*, a transcriptional regulator that is also provided maternally. In agreement with the previous method, around 50% of genes activated at this stage have no ortholog in other organisms, except for *H. azteca* for few of them. This highlights the

observation that the earliest activated genes of *Parhyale* are evolutionary young and species-specific or amphipod-specific genes. The *Rotund* ortholog was one of the earliest TF transcribed at 12 hours, in addition to many developmental regulators that were activated during the minor wave of ZGA, like *Parhyale's* ortholog of *Zelda* which is also deposited maternally but activated zygotically at 13–14-hour stage. In addition, gap genes orthologs *ocelliless, buttonhead* and *even-skipped*, the pair-ruled ortholog were activated during minor wave ZGA. Some genes were maternally loaded and activated at 16 hours during major ZGA wave, such as the pair-rule TFs *odd-paired* and *sloppy paired 2*, while TFs like *odd-skipped* and *runt* had no maternal or zygotic expression detected.

The analysis also identified 84 TFs that were activated during the major ZGA wave, including orthologs of *Sox21b, shuttle craft, Tox3 and cubitus interruptus (ci)*, which have been found to be required for embryonic development in other species (Kelberman et al., 2008; Roshina et al., 2014; Sahu et al., 2016; Schwartz et al., 1995). All these TFs were provided maternally, apart from *ci*, highlighting the interplay between zygotic and maternal programs in mediating MZT/ZGA process (Hamm and Harrison, 2018; Tadros and Lipshitz, 2009).

The number of detected activated genes dropped in both datasets at later stages (20 and 24 hours), with 370 in total RNA and 1012 in polyA+ during 20 hours, and 529 in total RNA and 167 in polyA+ during 24 hours timepoints. This observation supports the proposed timepoint of major wave of ZGA during the 16-hour stage, it's also possible that the expression of zygotic transcripts is dynamic at later stages. For example, transcription of *brat* starts at 11-hour stage, but the strongest signal of activation was detected during the transition from 16 to 20 hours.

We also observed that orthologs of *GATA* genes, which are known for their role in gut formation (Aronson et al., 2014; Hernández de Madrid and Casanova, 2018), were activated during different timepoints. A *GATA* ortholog was provided maternally and activated by 16 hours of development onwards, while another *GATA* ortholog and *grain* ortholog were activated only zygotically after 24 hours and stayed expressed in 24-60 hours embryos.

Figure 3.11. Illustration of potential source of intronic reads. **(A)** Diagram of gene structure with little or no intron reads **(B)** Diagram of a gene structure with uniform intron reads covering most intronic position, which occurs in the case of nascent transcription. **(C)** Diagram of a gene structure with high count of intron reads covering specific intronic region, which could arise from repetitive sequence expression rather than real transcriptional activation of the related gene. The RINP parameter aims to detect intronic reads coming from real nascent transcription by measuring ratio of covered intronic reads to non-covered intronic position. In the case of real transcription of a given gene, the RINP score will be significantly higher than that of maternal stage RINP score (transcriptionally silent stage) or than the RINP score of previous stage.

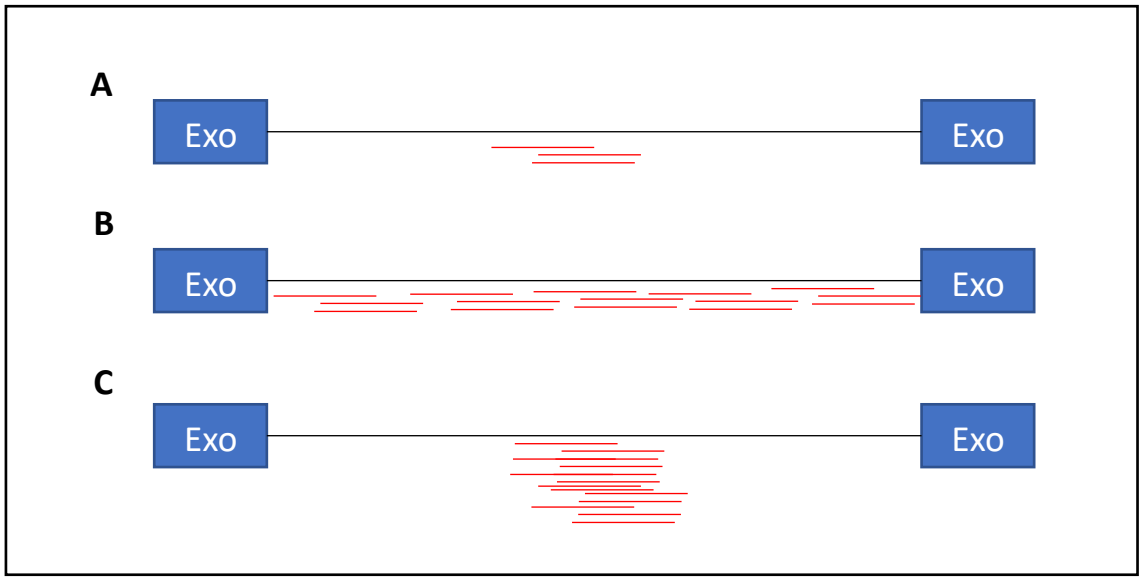
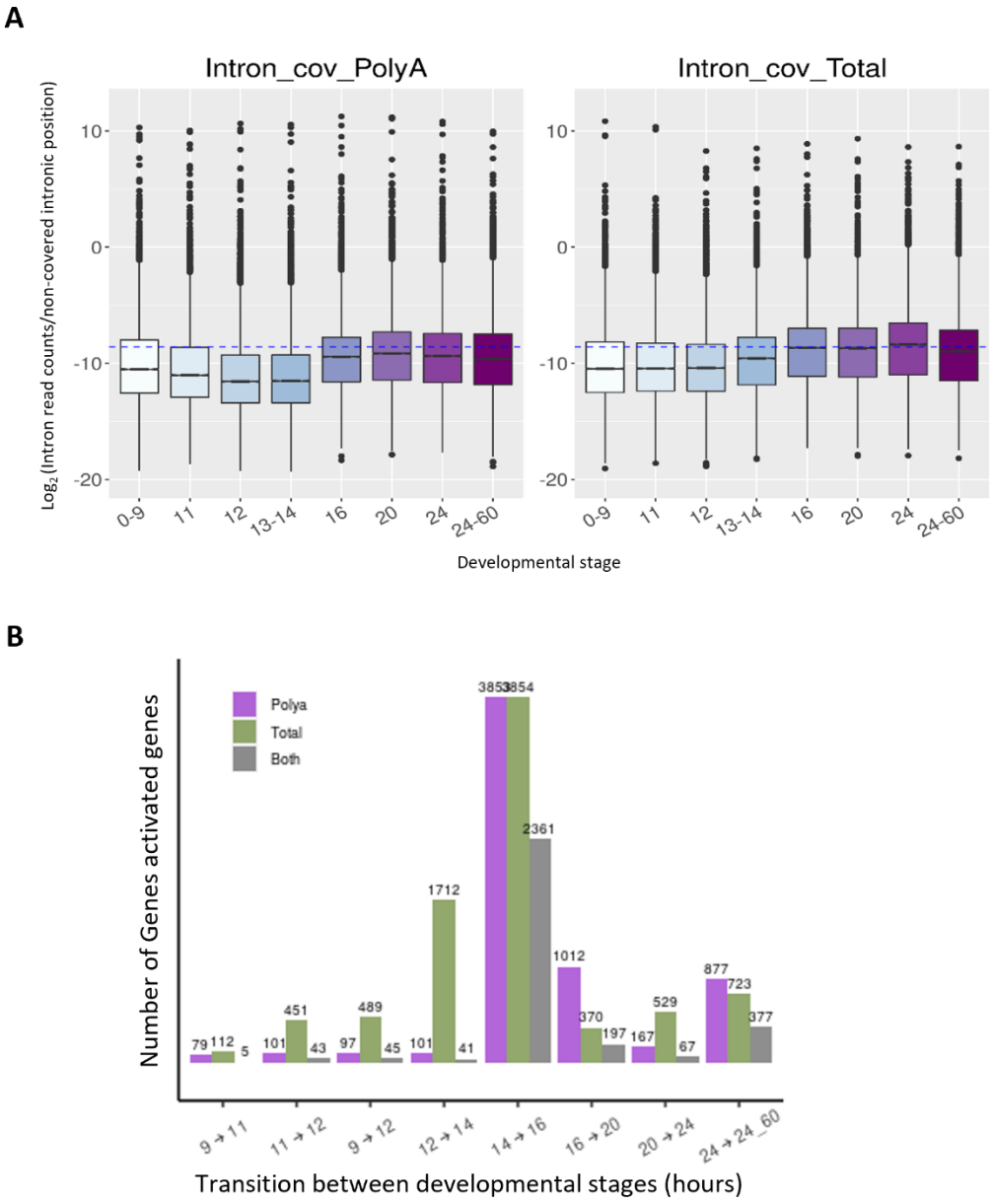


Figure 3.12. Detection of gene activation by identifying precursor mRNAs. (A) Boxplot showing the dynamics of intron coverage based on RINP score calculated for each timepoint in polyA+ and total RNA dataset. The log2 of the mean RINP score for each developmental stage is plotted, with the 75th percentile of RINP in the 0-9 hours stage of total RNA data considered as the baseline background of intronic coverage. **(B)** Zygotic genes detected using a significant increase in RINP score (coverage of intronic counts) between consecutive stages, using polyA+ and total RNA libraries. The overlap of genes detected in both libraries is also displayed. A Peak of activation is noticed during the 16-hour stage, after a gradual activation between the 11 to 14 hour stages.



3.5.4 Earliest zygotic genes are short and specific to *Parhyale*

The set of genes transcribed during the beginning of ZGA varies among species and is primarily composed of young, newly evolved genes (Heyn et al., 2014). To determine if this pattern is observed in *Parhyale's* early ZGA genes, we compared *Parhyale*'s early zygotic genes with orthologs in established model organisms. Overall, only 25% of *Parhyale's* early zygotic genes had orthologs in Chordate species, such as humans and mice, or insects like *Drosophila* (see Figure 2.6 in chapter 2). In contrast, 40% of the examined genes had orthologs in the amphipod *H. azteca*. To further investigate how conserved *Parhyale's* early ZGA genes are among different species, we performed a reciprocal blast top-hit analysis between *Parhyale* and *Drosophila* or zebrafish. We found only a few genes that were commonly activated early in both *Parhyale* and these species, with just two genes in *Drosophila* and six genes in zebrafish. We could not compare *Parhyale's* early ZGA genes with those of other amphipod species as no such list of genes was available in the literature. These findings support the notion that *Parhyale*-specific genes are predominantly activated during early ZGA, lending further support the idea of evolution of young species-specific transcripts during the onset of the minor wave of ZGA in animals (Heyn et al., 2014; Ali-Murthy et al., 2013; Kwasnieski et al., 2019). Previous studies have demonstrated that genes transcribed during early ZGA tend to be short, presumably because newly evolved genes are usually short (Heyn et al., 2014; Ali-Murthy et al., 2013; Kwasnieski et al., 2019). Moreover, rapid cell-cycle divisions during early development can limit the length of potential productively transcribed genes. Therefore, we tested if this theory applies to *Parhyale's* earliest transcripts.

We observed that the primary-transcripts of 807 early transcribed genes in 11-14 hours embryos were significantly shorter than the maternal genes mRNAs (with a median of 9.1 kb compared to 38 kb, respectively; Wilcoxon test *p-value* < 2.2e-16) (see Figure 3.13 A). We also found that these early transcribed zygotic minor wave genes had significantly fewer introns compared to maternal transcripts (Wilcoxon test *p-value* < 2.2e-16). In addition, 40% (57/139) of the zygotic-only and 5.5% (511/9244) of the maternal-zygotic minor wave genes lacked introns (see Figure 3.13 B). These results in *Parhyale* are

consistent with evidence from other organisms that early zygotic genes tend to be short, intron-poor or intron-less, and newly evolved genes (Atallah and Lott, 2018; De Renzis et al., 2007; Artieri and Frase, 2014; Heyn et al., 2014; Kwasnieski et al., 2019).

One study on fruit fly embryogenesis has suggested that truncated transcripts are prevalent during early stages of development, as the rapid cell divisions interfere with RNA polymerase II and lead to aborted transcripts (Kwasnieski et al., 2019). We therefore investigated whether this phenomenon was also observed in *Parhyale*. To do this, we used the 5' read density in total RNA data as an indication of transcription truncation, as polyA+ selection would deplete any aborted transcripts. Each transcript was divided into 20 windows, and we calculated the adjusted coverage using local-log coverage normalized by maximum log-coverage for each gene. We then plotted the averaged adjusted coverage across the gene length of maternal, maternal-zygotic, and zygotic only genes (Figure 3.13 E and F). We analyzed exon and intron reads separately and found that exon-based coverage analysis showed a uniform gene body coverage pattern between maternal genes and maternal-zygotic genes with a slight bias toward the 5' end (Figure 3.13 E). This similarity in the coverage pattern between maternal only and maternal-zygotic genes is likely due to truncated transcription happening simultaneously with degradation of maternal genes. We therefore added the purely zygotic activated genes (139 genes expressed in 11-14 hours using first expression detection method) as a third category to distinguish maternal clearance from zygotic aborted transcription. We observed a higher 5' bias in the zygotic-only minor wave genes compared to the other two categories (Figure 3.13 E). Furthermore, the intron-based coverage plot showed a greater 5' bias for zygotic minor wave transcripts than for maternal transcripts that had no intronic signal detected (Figure 3.13 F).

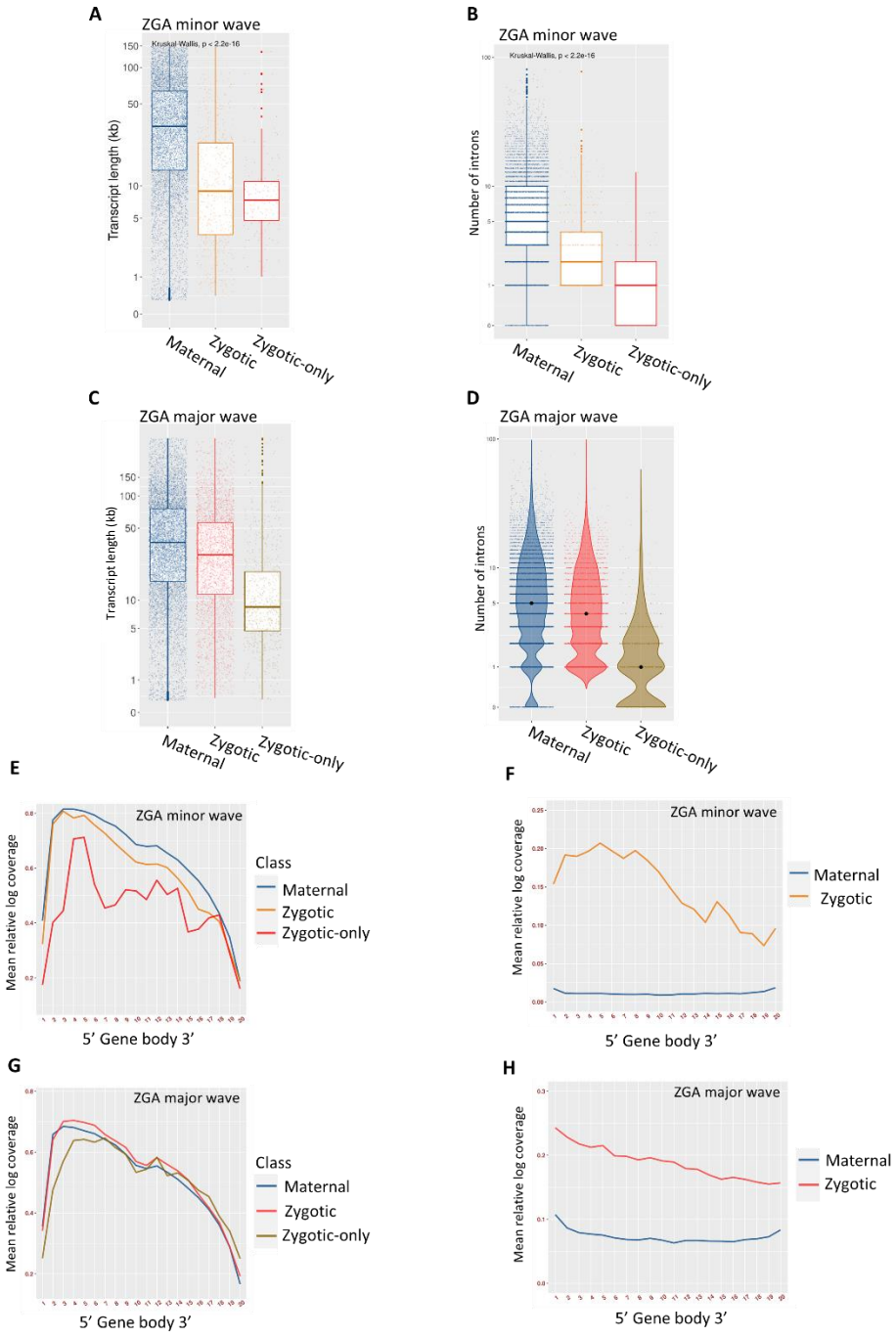
During the major wave of ZGA, the median primary-transcript length of newly transcribed genes was longer compared to those transcribed during the minor wave (with a median length of 28.25 kb versus 9.1 kb, respectively). Additionally, major wave ZGA transcripts had a smaller length difference (fold change = 1.4) compared to maternal transcripts, as opposed to the larger difference observed between maternal genes and minor wave ZGA genes. These major wave genes also had a similar intron number (Figure 3.13 C and

D). In contrast, zygotic-only genes detected in 16-hour stage embryos had the shortest length (median = 8.6 kb) and mostly intron-less.

Exon based reads showed similar 5' coverage between maternal and major wave zygotic genes, with a persistent 5' bias, while the zygotic-only genes had slightly reduced 5' coverage (Figure 3.13 G). Intron-based coverage was more uniform for the zygotic gene-bodies, indicating the synthesis of full-length pre-mRNAs of major wave genes (Figure 3.13 H). Interestingly, an intronic signal was detected for some maternal genes, suggesting that our approach may have missed some genes that are activated during this period. This indicated that zygotic activation for a limited number of genes could occur earlier than the 11-hour stage, or that these genes could be unspliced maternal candidates, deposited as pre-mRNAs.

Overall, these findings suggest that early minor wave transcripts in *Parhyale* tend to be short, intron-less or intron-poor, newly evolved, and aborted. These observations are consistent with those made in previous studies on *Drosophila* and zebrafish embryos (Heyn et al., 2014; Kwasnieski et al., 2019; Sandler al., 2018). Further investigation is required to determine if this pattern in *Parhyale* is also linked to cell-cycle length and to understand the potential impact of these short species-specific genes on early transcription.

<u>Figure 3.13. Characterization of minor and major zygotic-genome-activation transcripts.</u> (A) and **(C)** Primary-transcript length comparison between maternal, zygotic (all zygotic including maternally provided ones), and zygotic-only genes during minor and major wave of ZGA, respectively. **(B)** and **(D)** Comparing number of introns in maternal genes, all zygotic and zygotic-only genes during minor and major wave of ZGA, respectively. Exon-based gene-body coverage patterns of the same gene groups in minor **(E)** and major **(G)** wave of ZGA. Intron-based gene-body coverage of maternal and zygotic genes is shown in **(F)** for minor wave and in **(H)** for major wave of ZGA.



3.6 Dynamics of DNA methylation mediator genes during MZT/ZGA in *Parhyale*

We analyzed the expression of DNA methylation mediator genes (DNMT1, DNMT3, MBD2/3, and TET2) to elucidate their potential necessity for MZT/ZGA in *Parhyale*. We found that all the genes responsible for adding, maintaining, and erasing DNA methylation are maternally deposited. Overall, the expression of all the genes tends to be higher in the early stages, followed by a downregulation in the later stages of our sequenced time points. DNMT1 and MBD2/3 exhibit the highest expression values, although their expression decreases as development progresses, they consistently fall within the range of moderately to highly expressed genes (Figure 3.14 A). On the other hand, DNMT3, the *de novo* methylation gene, shows low and stable expression, which begins to downregulate at the 20-hour stage until it becomes barely expressed during the 24-60-hour stage (Figure 3.14 B). TET2, the gene responsible for demethylation, exhibits the most fluctuating expression (Figure 3.14 B). Its expression starts at a moderate to low level, then continues to upregulate until it reaches a peak during 13-14-hour stage before downregulating again until 24-hour stage. A slight increase in expression is observed at the 24-60-hour stage. DNMT1 reaches its highest expression value during 16-hour stage, while MBD2/3 shows two peaks of expression at 11-hour and 13–14-hour stages, indicating a potential role of these genes during ZGA of *Parhyale*. DNMTs reach the highest peak of polyA+/total RNA ratio during the 16-hour stage, while MBD2/3 have the highest ratio during 20-hour stage, suggesting a role during the major wave of ZGA. DNMT1 and MBD2/3 maintain a polyA+/total RNA ratio higher than 1.5 even after their expression downregulation, while this ratio decreases for DNMT3 to 0.27, indicating a potential degradation of DNMT3 after MZT. On the other hand, TET2 experiences the highest ratio of polyA+/total RNA during the 11 -hour stage, indicating a role during the onset of ZGA. The ratio fluctuates between consecutive stages, but it remains above 1.5.

None of the DNA methylation mediator genes were detected in any of the methods we used to identify activated zygotic genes in the minor or major wave. However, the coverage profiles of exon-intron signals

across the gene-body show an increased and uniform, although not strong, intronic signal starting from the 16-hour stage for DNMT1 and DNMT3 (Figure 3.15) (RINP score fold change of 1.5 and 1.9, respectively). In the case of MBD2/3, no obvious intronic signal was detected, and there was no increase in the RINP score in any stage, suggesting a later activation of MBD2/3. For TET2, the intronic signal begins to appear at 24 hours (0.4-fold change in RINP score) and becomes stronger at 24–60-hour stage (Figure 3.15 A). The coverage patterns were predominantly similar between the polyA+ and total RNA seq libraries. However, for DNMT1, the exonic signal was significantly higher in the polyA+ library from stage 16 to 60 hours. Additionally, TET2 showed a stronger exonic signal in the polyA+ dataset from 0-14 hours, displaying similar signal in both datasets after that point, likely due to the enrichment of maternal polyadenylated mRNAs in the polyA+ data. The uniform intronic signal indicates zygotic activation of all genes except for MBD2/3. Low levels of intron counts have prevented the detection of these genes using the RINP approach.

Furthermore, we generated the same coverage plots using RNA-seq data generated by Calvo et al., 2022, which covers a wider range of embryogenesis from the first to the last stage. The coverage pattern using this dataset shows a downregulation in the expression of all the genes starting from 25 to 60 hours of development, and they maintain the same level of expression from that point onward until the last stage of embryogenesis, which further support zygotic activation of these genes (Figure 3.15 B).

Altogether, the findings suggest that DNA methylation machinery genes play an important role in the early stages of development and during the MZT/ZGA process. These genes show a tendency to downregulate after gastrulation and maintain a stable expression until the end of embryogenesis. This result indicates the significance of DNA methylation during the early embryogenesis of *Parhyale*. Furthermore, it suggests a potential decrease in DNA methylation as embryos progress in development, as supported by the observed downregulation of DNMT1 and upregulation of TET2 during development. Further investigation of DNA methylation levels in embryos during the early stages will provide us with additional insights into the dynamics of methylation during *Parhyale* MZT and embryogenesis.

Figure 3.14. Expression pattern of DNA methylation mediator genes. **(A)** Histograms of log2 polyA+ Exon-based TPM distribution in each sequenced stage, illustrating the expression levels of DNA methylation machinery genes. **(B)** Dot plot of expression values across all sequenced developmental timepoints for each gene. DNMTs and MBD2/3 possess high expression before and during ZGA, at later stages they experience a drop in expression, with DNMT3 showing the lowest levels of expression overall (next page).

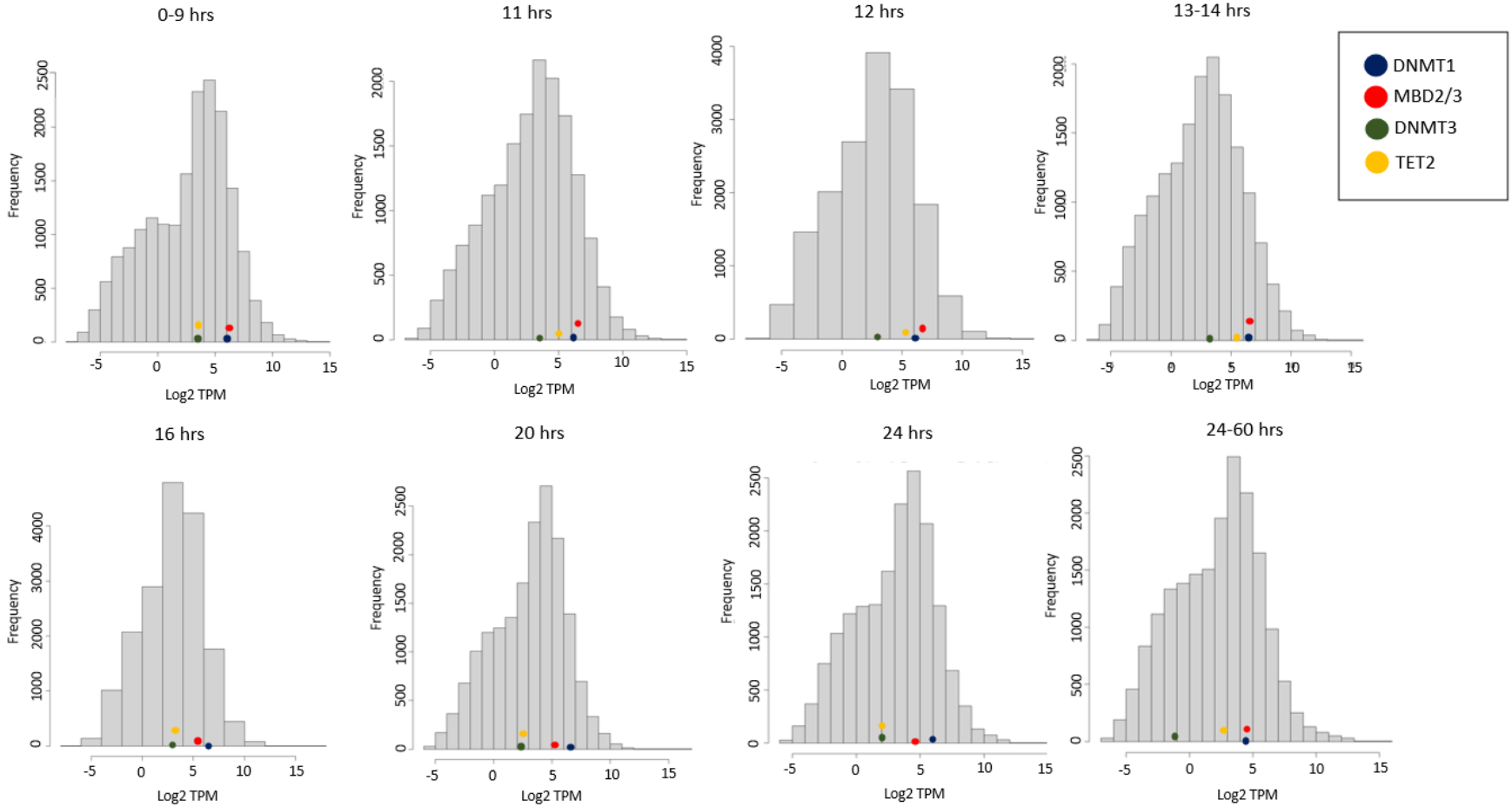
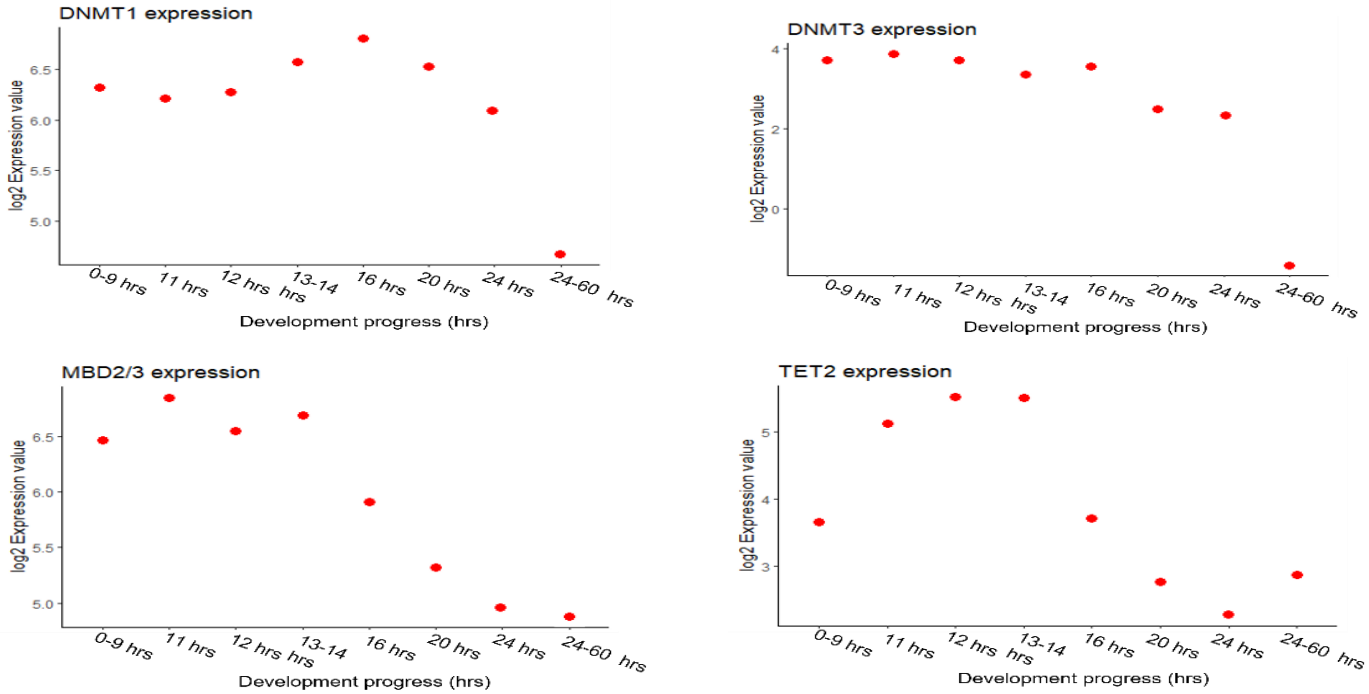


Figure 3.14B. Expression pattern of DNA methylation mediator genes (B) Dot plot of expression values across all sequenced developmental timepoints for each gene. DNMTs and MBD2/3 possess high expression before and during ZGA, at later stages they experience a drop in expression, with DNMT3 showing the lowest levels of expression overall.

B

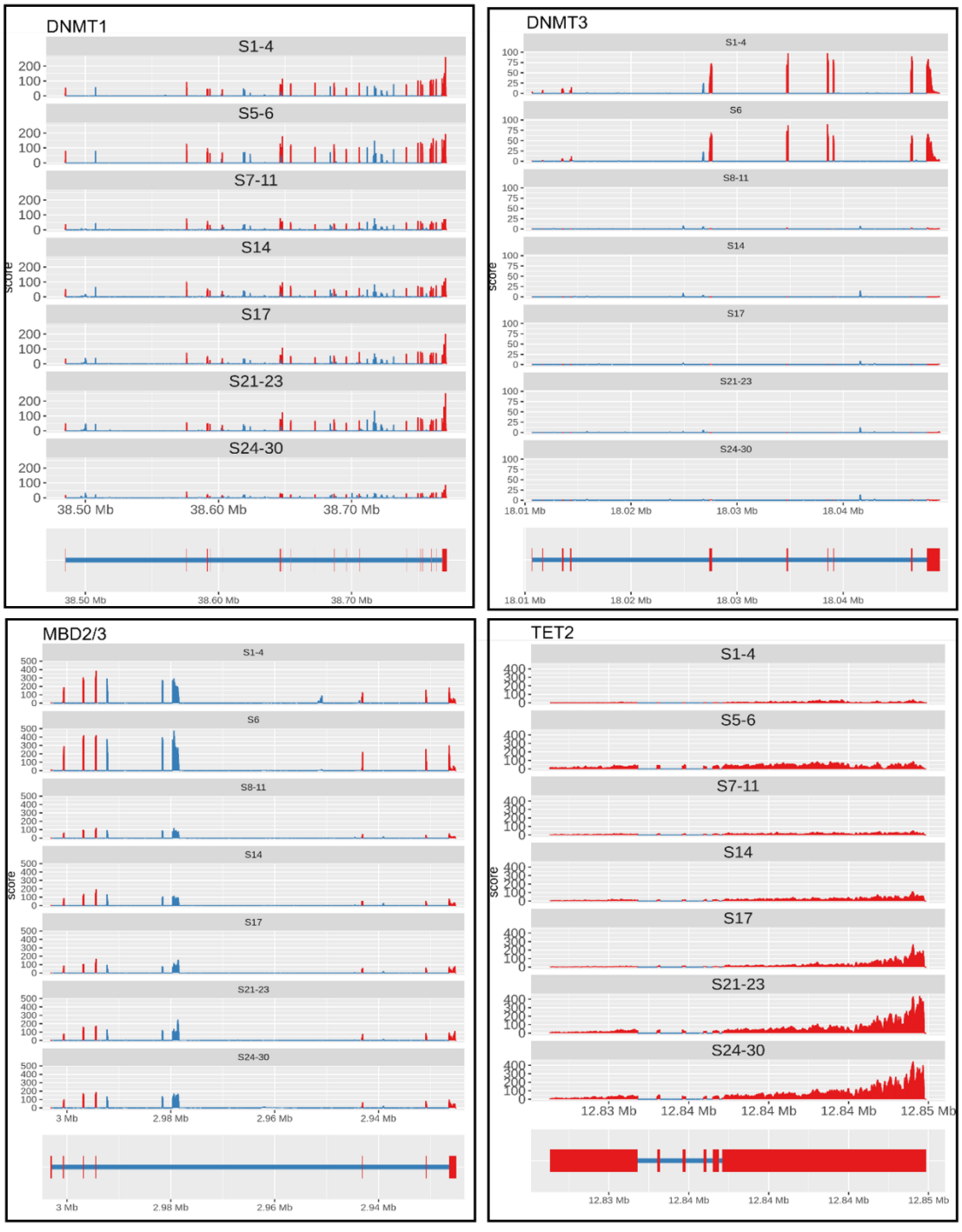


<u>Figure 3.15. Coverage plots for DNA methylation mediator genes.</u> (A) Coverage pattern of exon regions (green) and intron regions (orange) in the polyA+ (left) and total RNA seq (right) libraries are displayed for DNMT1, DNMT3, MBD2/3, and TET2. DNMTs and MBD2/3 exhibit the highest expression levels during early stages, followed by a decrease in expression as development progresses. TET2 shows a more fluctuating expression pattern, with an initial gradual increase until reaching a peak around 11-14 hours, followed by a downregulation. At 60 hours of development, the expression starts to increase again. **(B)** coverage plots using data from Calvo et al., 2022 (details provided on the next page).

A

Figure 3.15. Coverage plots for DNA methylation mediator genes. (B) Coverage pattern of exon regions (red) and intron regions (blue) in polyA+ libraries generated in (Calvo et al., 2022) for timepoints spanning the entire embryogenesis of *Parhyale* are displayed for DNMT1, DNMT3, MBD2/3, and TET2. Expression levels detected for DNMT1, DNMT3 and MBD2/3 are higher at the beginning of embryogenesis, and subsequently downregulate, maintaining low expression levels until the end of embryogenesis. In contrast, TET2 exhibits variable expression levels between consecutive timepoint, with the highest levels of expression detected towards the end of embryogenesis.

B



3.7 Correlation between gene expression and DNA methylation during ZGA

A whole-genome enzymatic methyl-seq (EM-seq) was performed in the Aboobaker lab to analyze the MZT/ZGA in *Parhyale* embryos matching 3 of the stages we sequenced. We focused on three specific stages of early embryogenesis (0-9.5 hours, 11-12 hours and 24-60 hours) for DNA methylation profiling. The average methylation levels across all CpG sites in the *Parhyale* genome were found to be 4.45%, 3.70%, and 3.37% for each of the sequenced stages, respectively (refer to Figure 3.16 A).

When considering only the effectively covered CpG sites (with a minimum coverage of 5X), the percentage of methylated CpG (mCpGs) for each stage was 8.9%, 7%, and 6.3%, respectively. However, there was no significant difference observed in the overall genome-wide methylation level between embryos before and after ZGA.

Additionally, we investigated the methylation levels in non-coding intergenic regions and repetitive elements in *Paryale*. It was found that these regions displayed the highest methylation levels, with LTR elements exhibiting the highest levels (~10%) compared to other annotated transposable elements (~5%) (see Figure 3.16 B). Overall, the average methylation levels of genes in *Parhyale* displayed a bimodal distribution, with the majority of genes exhibiting very low methylation levels (Figure 3.16 C).

We performed an integration analysis of our RNA-seq data with the EM-seq data to investigate potential interaction between DNA methylation and gene expression during *Parhyale* MZT process. Overall, we observed a consistent positive correlation between gene body methylation and expression levels (Figure 3.16 D).

We found that highly expressed genes (≥ 5TPM) tended to exhibit higher levels of gene body methylation compared to genes with low expression or those that were unexpressed. Additionally, we observed that at least 50% of expressed genes showed enrichment of mCpG.

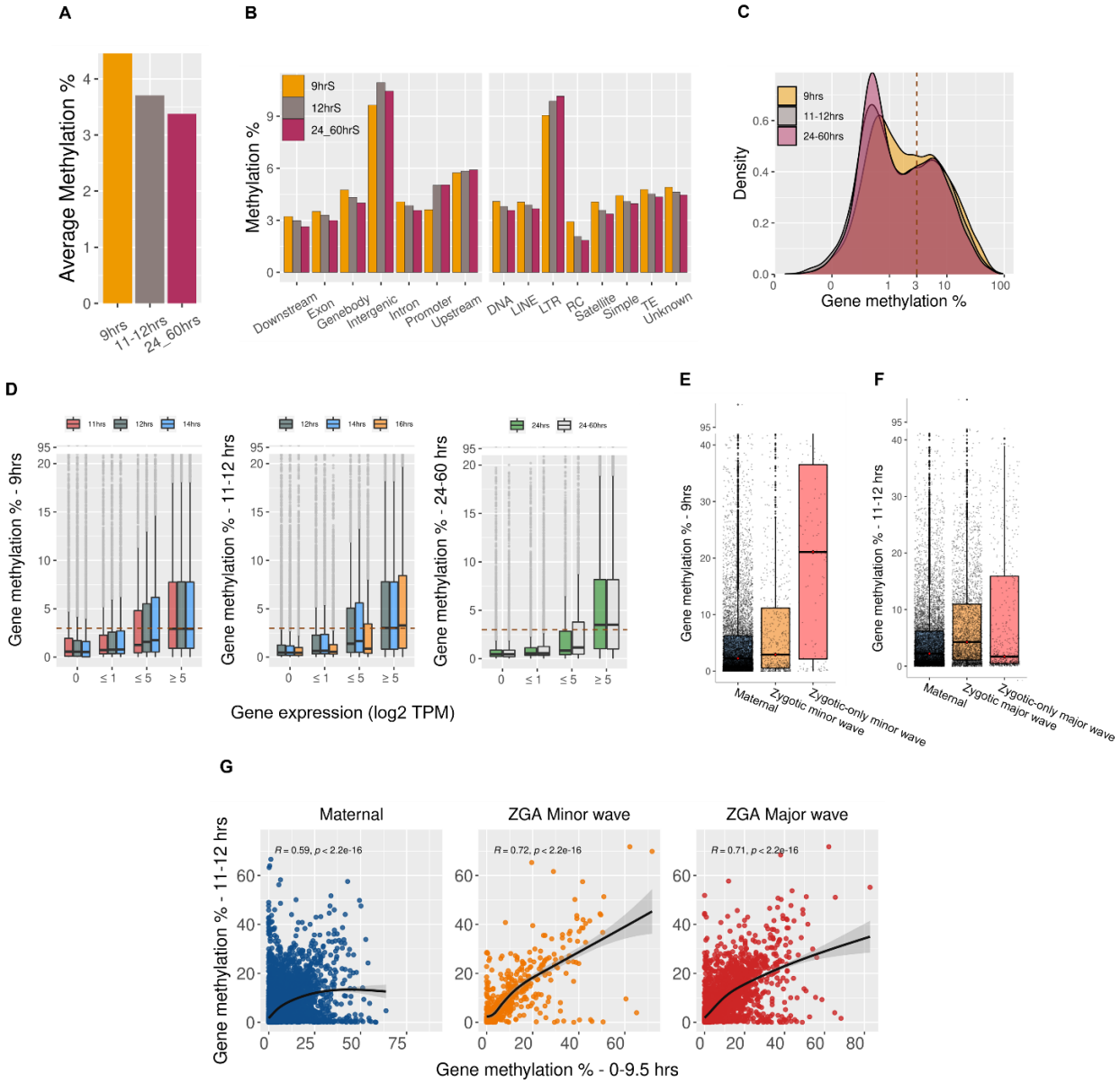
To examine the association between DNA methylation and gene expression at different stages, we compared each of the three DNA methylation stages with the temporally closest timepoint in the RNA-seq dataset (Figure 3.16). We consistently observed a positive correlation between the level of DNA methylation and gene expression across all datasets. This suggests a potential regulatory interaction between DNA methylation in transcriptional activity during *Parhyale* embryogenesis.

Furthermore, zygotic-only genes activated during the minor wave of ZGA (11-14 hours) exhibit significantly higher methylation levels compared to maternal genes and all minor wave ZGA genes (Wilcoxon test *p-value*= 6.9e-12) (Figure 3.16 E). Additionally, these genes maintain a similar methylation level before and after onset of ZGA (Figure 3.16 G). On the other hand, zygotic-only genes of major ZGA display comparable DNA methylation levels to maternal genes (Figure 3.16 F), with no significant difference in their gene-body methylation levels before or after ZGA (Figure 3.16 G). Some maternal genes tend to exhibit higher methylation levels after ZGA onset. Upon investigation, it was found that some of these genes are maternally provided and activated during ZGA. However, a subset of maternal genes also undergoes an increase in DNA methylation levels after ZGA, indicating potential activation during advanced stages of development. Overall, these findings suggest that DNA methylation is required for promoting gene activation during *Parhyale's* embryogenesis and potentially contributes to the regulation of MZT/ZGA. However, further investigations are needed to elucidate the underlying mechanisms of this contribution.

In conclusion, our findings suggest a positive correlation between gene expression and DNA methylation during *Parhyale* embryogenesis. The decrease in methylation levels during later stages aligns with the downregulation of DNA methylation mediators observed in the previous section. The maternal deposition of genes responsible for DNA methylation, along with the positive correlation between highly expressed genes in the early stages, indicates a regulatory role of DNA methylation during the early embryogenesis of *Parhyale*.

Figure 3.16. Correlation between DNA methylation and gene expression during *Parhyale* MZT/ZGA.

(A) Average methylation levels of all covered CpG sites at three timepoints during early embryogenesis. **(B)** Fractional methylation levels across all covered CpG sites mapped to different genomic features. **(C)** Distribution of average methylation levels for all protein-coding genes. **(D)** Methylation levels of genes at different TPM thresholds. Methylation levels for each screened timepoint (y-axis) plotted against temporal-associated gene expression levels. Gene methylation levels of maternal, all zygotic, and zygotic-only genes during **(E)** The minor wave of ZGA and **(F)** The major wave of ZGA. **(G)** Correlation between DNA methylation levels before ZGA onset (0-9.5 hours) and after ZGA onset (11-12 hours) of maternal genes, minor wave ZGA genes, and major wave ZGA genes.



3.8 Discussion

In this chapter, we focused on exploring the dynamics of MZT and ZGA during the early embryogenesis of *Parhyale*, an area that had not been extensively studied before. We carefully selected critical time points to examine and proposed a timeline for the minor and major waves of ZGA. Additionally, we delved into the behavior of maternal transcripts during this process. Through our comprehensive transcriptome analysis during MZT, we have gained significant insights into the embryonic development of *Parhyale* and the dynamics of DNA methylation during this major developmental process.

We utilized two datasets, consisting of polyA+ RNA sequencing and total RNA sequencing of the same samples, to accurately characterize early transcriptome of *Parhyale*. This approach enabled us to distinguish between maternal factors *and de novo* transcription. We used changes in polyA tail length as a proxy for the adenylation state of transcripts. Our observations revealed a genome-wide polyadenylation of maternal transcripts during the initial stages of development, followed by a subsequent deadenylation process occurring between 11-14 hours. This deadenylation phase likely signifies the clearance of maternal mRNAs during this period of embryogenesis.

Concurrently, we observed the onset of ZGA, with zygotic genes being activated as early as 11 hours of development. The activation of zygotic genes occurred gradually, reaching a peak during 13-14 hours and a subsequent activation phase during 16 hours, suggesting the existence of both a minor and a major wave of zygotic genome activation (Figure 3.17).

The maternal mRNAs expressed during the early stages of embryogenesis encompass genes involved in guiding early development, including those associated with translation regulation, mRNA processing, and silencing. Notably, we identified the pioneer transcription factor Zelda, recognized for its significant role in transcriptional regulation and promoting ZGA in various arthropod species (Ribeiro et al., 2017; Duan et al., 2021). Our findings demonstrate that Zelda is maternally loaded and among the earliest activated zygotic genes in *Parhyale*.

To explore the transcriptomic landscape of ZGA, we employed multiple approaches. Initially, we performed k-means clustering on maternal and zygotic genes to identify distinct expression patterns. This analysis enabled us to classify genes into different groups, including maternal, minor ZGA, and major ZGA potential genes, based on their expression behavior across the sequenced stages.

During our investigation, we observed that some maternal genes underwent degradation after the onset of ZGA and were subsequently activated either at a similar level to maternal transcripts or at a lower level. Additionally, we utilized two additional strategies to identify newly activated zygotic genes. These approaches allowed us to distinguish between zygotic genes that are also provided maternally and zygotic genes that are exclusively activated by the zygotic genome.

Interestingly, while maternally deposited genes were predominantly conserved, the earliest genes transcribed from the zygote exhibited enrichment of species-specific transcripts, a phenomenon observed in other organisms as well (Heyn et al., 2014). Furthermore, through detection using intronic signals, we uncovered the activation of several developmentally important genes during the early stages of *Parhyale* embryogenesis (11-14 hours). Noteworthy examples include the TF Zelda, the transcriptional regulator *brat*, and orthologs of gap genes like *buttonhead*, which are known to be expressed during the initial hours of *Drosophila* embryogenesis (Papatsenko and Levine, 2011; Estella et al., 2003).

Moreover, we observed the activation of numerous TFs during the major wave of ZGA, particularly starting from the blastodisc formation stage. Over 80 TFs, including *Sox21b*, *FoxP*, *labial* and *cubitus*, were among the genes activated during this stage. These TFs play crucial roles in body patterning and growth control.

Furthermore, we conducted an analysis of the expression patterns of DNA methylation machinery genes throughout the developmental stages we sequenced. We found that all the genes required to install and erase DNA methylation, as well as MBD2/3, were loaded maternally. The highest levels of expression of DNMTs and MBD2/3 were observed during the early stages of development. Subsequently, their expression levels

exhibited a downregulation during later stages, and they maintained relatively low expression levels until the end of embryogenesis.

In contrast, the demethyltion gene TET2 displayed a more diverse expression pattern. It exhibited low expression levels during the early stages, which gradually increased and peaked at the 13–14-hour stage, likely coinciding with the zygotic activation of the gene. During later stages of development, TET2 expression showed a further increase, persisting until the last stage of development.

Lastly, we performed an integration analysis of our transcriptomic dataset with methyl-seq dataset encompassing three early developmental timepoints before and after ZGA. The key discovery from this analysis was the positive correlation observed between gene-body methylation levels and gene expression during *Parhyale's* embryogenesis. Highly expressed genes tended to exhibit higher levels of methylation, while non-methylated genes predominantly displayed low or no expression.

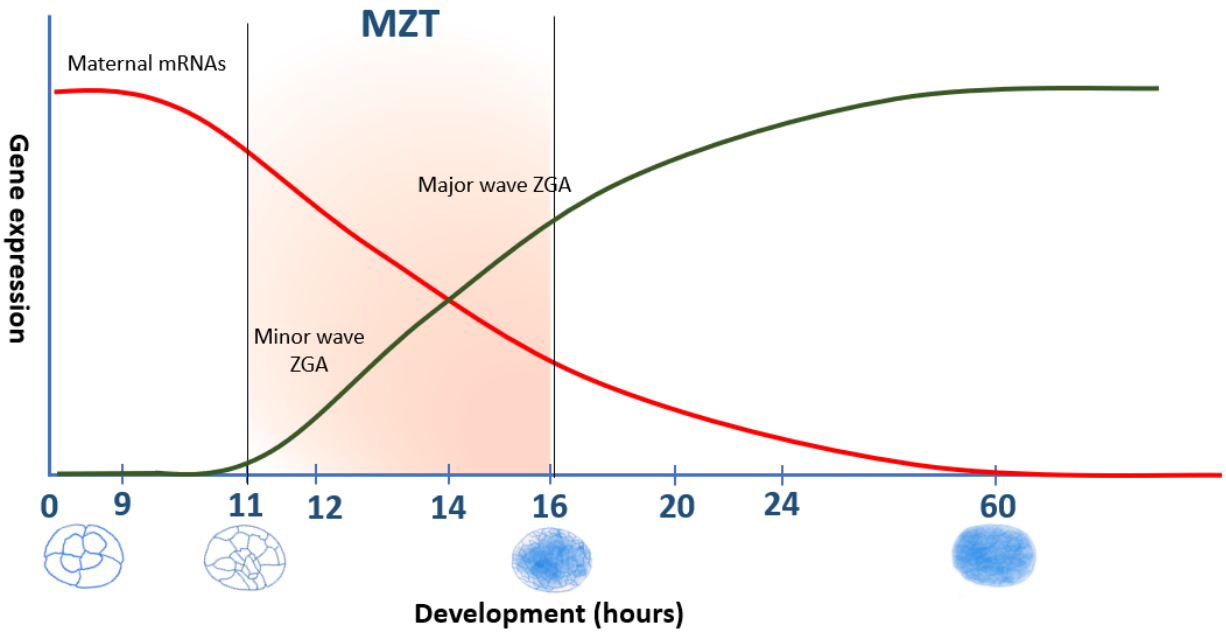
Furthermore, when examining the DNA methylation profiles across the three timepoints, we observed that the early stages of development exhibited the highest level of methylation. As embryogenesis progressed, there was a noticeable reduction in methylation levels, corresponding to the downregulation of genes responsible for DNA methylation.

Taken together, our findings indicate a regulatory role for DNA methylation during the early embryogenesis of *Parhyale*, specifically during MZT/ZGA. Moreover, the observed reduction in gene-body methylation as embryonic development proceeds suggests a subsequent removal or remodeling of DNA methylation marks.

In conclusion, our study has contributed to the understanding of *Parhyale's* embryonic development by proposing a timeline for key events and shedding light on the behavior of DNA methylation during early embryogenesis. This comprehensive analysis provides a solid foundation for future research on *Parhyale* as invertebrate and an arthropod model for studying embryonic development and regeneration. Additionally, our findings regarding the role of DNA methylation in regulating these processes in such a

model organism open up avenues for further investigations in this area. Overall, our work paves the way for deeper exploration of *Parhyale's* developmental mechanisms and the involvement of DNA methylation in their regulation.

Figure 3.17. Summery model depicting Maternal-to-zygotic transition and zygotic-genome-activation in *Parhyale*. During initial stages of development, embryogenesis in *Parhyale* is primarily regulated at a post-transcriptional level. A significant fraction of maternal transcripts is highly expressed highly during 0-100 cell stage (0-12 hours) and undergoes deadenylation in subsequent stages, indicating a clearance of these maternal mRNAs. The activation of zygotic genome begins around 32-cell stage, and the number of activated genes gradually increases until it reaches a peak during blastomere migration time (13-14 hours) and around blastodisc formation (16 hours), indicating the occurrence of two distinct waves of zygotic activation.



Chapter IV

Functional experiments on DNA methylation machinery genes during embryogenesis

Contents

Abstract

The role of DNA methylation has been reportedly associated with processes that require cell differentiation, such as embryonic development. However, the level and the genomic context of DNA methylation varies widely between and within species, which suggests significant functional differences. In this chapter, we focus on three DNA methylation mediator genes encoded in *Parhyale's* genome. We used CRISPR/Cas9 mediated genome editing to generate knockout animals and observe the effect of these genes on embryogenesis. Initial knockout experiments indicate that losing any of the three genes (DNMT1, DNMT3 and MBD2/3) is lethal to embryos. RNA-seq analysis of early embryonic stages, as presented in the previous chapter, showed that all three genes are maternally provided. Therefore, knocking out the gene in the zygotic genome may not affect early embryogenesis, as all three genes are expressed maternally at their highest levels. Additionally, persistence of maternal mRNA might compensate for the loss of the zygotic transcripts after ZGA.

We performed an RNAi knockdown experiment with a focus on MBD2/3 to degrade mRNA of the gene, with the possibility of impacting maternal expression. This also resulted in an early embryonic lethal phenotype. MBD2/3 is an integral component of the NuRD complex, which could function independently of DNA methylation in species like Drosophila. We profiled the transcriptome of MBD2/3 knockdown embryos to understand the effect of MBD2/3 on gene expression and determine if it regulates gene expression in a methylation-dependent manner or not. The knockdown led to upregulation of many transcription factors that are known to play important roles in development. Interestingly, these genes have low levels of gene-body methylation (0.03 -16.9%), and some also have promoter methylation.

The objective of this chapter is to determine the necessity of DNA methylation for embryonic development in *Parhyale* and to understand its potential role in gene expression regulation. To achieve this, we targeted genes that mediate DNA methylation, specifically DNMT1, DNMT3, and MBD2/3. Disrupting any of these genes should influence DNA methylation levels, providing insights into their importance for

embryogenesis. However, MBD2/3 may operate independently of DNA methylation. Our initial approach involved genome editing, with the added intention of tagging these genes to monitor their expression. Subsequently, we employed RNAi knockdown, aiming to affect both the maternal and zygotic copies of the gene, with a particular emphasis on MBD2/3. Based on the premise that DNA methylation is essential for normal embryogenesis in *Parhyale,* we postulated that the absence of any of these genes would be detrimental to embryos, and this is what we observed as mentioned earlier. Given the potential of MBD2/3 to function independently of DNA methylation, we profiled the RNA of MBD2/3 knockdown embryos and examined the methylation status of differentially expressed (DE) genes. We observed that most DE genes exhibited gene body methylation, albeit at low levels.

In summary, the lethality of embryos upon loss of these genes underscores the significance of DNA methylation in *Parhyale's* embryogenesis. Furthermore, these findings highlight the potential of *Parhyale* as a promising model system for further exploring DNA methylation and the NuRD complex in an arthropod system.

4.1 Introduction

DNA methylation is a major epigenetic mark. However, patterns of DNA vary greatly between different species. Differences in abundance and the genomic target of methylation in different animals result in functional differences in DNA methylation and its contribution to gene regulation. Until recently, DNA methylation patterns were seen as distinct between vertebrates and invertebrates. In vertebrates, DNA methylation is globally distributed in the genome, mainly targeted at CpG dinucleotides, except for CpG islands located at promoter regions. In invertebrates, DNA methylation is generally considered to be at lower levels, has a mosaic distribution, and is targeted to non-coding regions proximal to genes (Lister et al., 2009; Feng et al., 2010; Hon et al., 2013). However, this general view has started to change as more genomes from non-vertebrate species are being profiled for DNA methylation (de Mendoza et al., 2019a).

The 5-methylcytosine (5mc) modification is primarily targeted to CpG dinucleotides in animals (Suzuki et al., 2008), and serves many critical functions, including genomic imprinting, X-chromosome inactivation and transposon silencing (Suzuki et al., 2008; Bird et al., 2002; Jones et al., 2012; Xu et al., 2019). However, the traditional model of the association between methylation and transcriptional silencing has been challenged recently with findings from several studies (Rauluseviciute et al.,2020; Héberlé et al., 2019). Transcriptional activation and repression have both been found to be associated with DNA methylation in a context-specific manner.

In many well studied vertebrate species, 5mc of regulatory regions such as promoters and enhancers are linked to transcriptional silencing of the downstream genes (Lou et al., 2014; Mendizabal and Yi 2016). Additionally, DNA methylation of transposable elements (TE) is also associated with silencing (Yoder et al., 1997). In contrast, 5mc of gene bodies is often associated with active transcription of relevant genes, and higher levels of gene-body methylation (gbM) are often associated with elevated transcription (Feng et al., 2010; Zemach et al., 2010; Jjingo et al., 2012).

As mentioned earlier, DNA methylation in invertebrates is targeted at gene bodies of protein-coding genes, and promoter methylation has not been observed in these species (Zemach et al., 2010; de Mendoza et al., 2019a; Xu et al., 2019). However, little evidence exists about the role of gbM in direct regulation of transcription (Dixon and Matz 2022). Studies that profiled 5mc in several invertebrate species failed to detect the differences in gbM in many examined cell types and developmental stages, despite the transcriptional variation in these methylated genes (Gatzmann et al.,2018; Harris et al., 2019; de Mendoza et al., 2019a; Dixon and Matz 2022). This indicates that the correlation between gbM and transcription remains unclear.

A recent study analyzed methylomic and transcriptomic data from Anthozoa and Hexapoda to examine the correlation between gbM and transcriptional variation under different conditions in multiple datasets (Dixon and Matz 2022). They found that changes in gbM are not necessarily required to induce changes in transcription, as differences in gbM across different conditions did not consistently show an association with differences in transcription, which was consistent across all examined datasets. Another study quantified DNA methylation and gene expression in the sea urchin *Strongylocentrotus purpuratus* larvae exposed to various ecological conditions during gametogenesis or embryogenesis (Bogan et al., 2020). Differential expression and splicing were modeled as a function of DNA methylation, genic feature type and chromatin accessibility. They found that the effect of gbM on differential gene expression was conditional upon chromatin accessibility and the type of methylated genetic feature. Together, these observations indicate that DNA methylation in invertebrates does not have a simple linear relationship with transcription, but rather is a part of more complex mechanisms that may involve interactions with other epigenetic modifications.

DNA methylation is a reversable, enzyme-mediated modification. One approach to understand DNA methylation is to study the function of the mediator genes of DNA methylation. The most effective approach to study gene function is to inactivate the gene through knockout or knockdown techniques and observe phenotypic effects in the organism.

Loss of DNA methyltransferase (DNMT) enzymes or methyl-binding domain (MBD) proteins, either through knockout or knockdown, has been performed in many species. In vertebrates, deletion of DNMT enzymes results in embryonic lethality (the case of DNMT1 and DNMT3b) or postnatal lethality (the case of deleting DNMT3a), indicating they are essential for normal development (Li et al., 1992; Okano et al., 1999; Liao et al., 2015).

In human embryonic stem cells line (ESC), homozygous knockouts of DNMT1 caused immediate lethality (Liao et al., 2015; Sen et al., 2010; Barra et al., 2012). Whereas DNMT3 knockout, either single knockout for DNMT3a or DNMT3b and double knockout for both genes, resulted in morphologically normal human ESCs that were able to differentiate into three germ layers (Liao et al., 2015). This observation was consistent with previous studies in mouse ESCs, although double knockouts in mouse ESCs were only able to generate teratomas (Okano et al., 1999; Chen et al., 2003).

These findings suggest that function of DNMT3 might arise after embryogenesis, as DNMT3a knockout mice develop normally but die a few weeks after birth (Liao et al., 2015; Okano et al., 1999). Loss of DNMT1 in frogs triggers an apoptotic response induced by hypomethylation, leading to death of the embryos (Stancheva et al., 2001). Consistent with findings in mice and frogs, DNMT1 morpholino-injected embryos of zebrafish have a lethal phenotype during gastrulation (Rai et al., 2006).

Zebrafish has a single copy of DNMT3, which is an ortholog to DNMT3b in mammals. DNMT3 knockdown embryos display multiple abnormalities and eventually die by 96 hours post-fertilization (Rai et al., 2010). It's important to note that depleting DNMT1 is always paired with hypomethylation in all examined vertebrate species.

Depletion of DNMT1 through RNA interference (RNAi) in the German cockroach *Blatella germanica* impaired blastoderm formation, which confirmed the requirement of DNA methylation for normal development (Ventos-Alfonso et al., 2020). Similar results were observed in the hymenopteran *Nasonia vitripennis*, where DNMT1 RNAi led to failure to complete embryogenesis and affected embryos died at

the onset of gastrulation (Zwier et al., 2012). RNAi of DNMT was also performed in the flour beetle *Tribolium castaneum*, even though this species lacks CpG methylation, the offspring of injected females were arrested after the first few cleavage cycles (Schulz et al., 2018). In the large milkweed bug *Oncopeltus fasciatus*, knockdown of DNMT1 in females' ovaries produced inviable eggs, demonstrating that DNMT1 is required for reproduction in this species (Bewick et al., 2019). However, in some insects, such as milkweed bugs, switching off DNMT1 does not always alter the expression of methylated genes (Bewick et al., 2019). In addition, the function of DNMT1 could be unrelated to DNA methylation, as observed in *T.castaneum* (Schulz et al., 2018).

In the annelid *Platynereis dumerilii*, high levels of 5mC methylation (more than 80% detected at some developmental stages) were observed. Inhibition of DNMT1 using two different drugs during first through third day post-fertilization did not block development but it led to morphological defects such as appendage formation (Planques et al., 2021). When the hypomethylating agent was applied to amputated animals (from 0 to 5 days post amputation), regeneration was delayed (Planques et al., 2021). Similarly, in the oyster, the DNMT1 inhibitor 5-aza-cytidine was used to decrease DNA methylation levels during early stages of development. Treated embryos suffered from dramatic morphological alterations and died within 24 to 48 hours of development (Riviere et al., 2013).

The *de novo* methyltransferase enzyme DNMT3 is absent in several invertebrate species that have DNA methylation. Examples include *Ixodes scapularis, Planococcus citri, Polistes canadensis, Bombyx mori and Heliconius melpomene* (Lewis et al., 2020; Albalat et al., 2012). In the invertebrate species where DNMT3 is conserved, results regarding its essentiality for normal embryonic development vary. In the honeybee *Apis mellifera,* silencing DNMT3 is associated with lethality of embryos, indicating its crucial role during embryonic development (Kucharski et al., 2008). On the other hand, DNMT3 in the hymenopteran *N. vitripennis* is not essential for embryogenesis, as no phenotype after RNAi-mediated knockdown was observed during development (Zwier et al., 2012). The detected levels of DNMT3 expression were found to be very low in multiple species, such as the milkweed bug *O. fasciatus* (<1 FPKM)

and the cockroach *B. germanica*, which could be the reason for unsuccessful knockdown (Bewick et al., 2019; Ventós-Alfonso et al., 2020).

Altogether, these studies show that when DNA methylation is conserved in an invertebrate, it seems to be essential for normal embryogenesis. However, its exact function and the networks it is involved in are still poorly understood.

The methyl binding domain protein MBD2/3 is conserved in many invertebrate species that have lost DNA methylation, suggesting a DNA methylation-independent function (Marhold et al., 2004a; Jaber-Hijazi et al., 2013; Gutierrez et al., 2007). This protein shares high homology with two mammalian methyl-binding proteins, MBD2 and MBD3 (Tweedie et al., 1999; Wade et al., 1999). The vertebrates MBD2 and MBD3 genes are likely the result of whole genome duplication from the MBD2/3 common ancestor (Hendrich and Tweedie 2003). MBD2/3 is an integral component of the Nucleosome Remodeling and Deacetylase (NuRD) complex, a major chromatin remodeling which is mainly known for its transcriptional silencing activity (Ng et al., 1999; Zhang et al., 1999; Guezennec et al., 2006). MBD proteins selectively recognize methylated DNA and recruit the associated NuRD complex.

In mammals, MBD2 can bind methylated DNA and mediate transcriptional repression as part of the MeCP1 histone deacetylase complex. It is also associated with the NuRD complex (Hendrich et al., 2001; Feng and Zhang et al., 2001). In contrast, MBD3 does not bind methylated DNA and is a component of the Mi-2/NuRD repressor complex (Hendrich et al., 2001). A recent study found that MBD2/NuRD and MBD3/NuRD bind to the same genomic loci, indicating that they are interdependent, and both form a part of a transcriptional silencing regulatory loop (Hainer et al., 2016).

Knockout experiments of these two genes in mice showed that although MBD2 and MBD3 share over 70% amino acid identity, mutant mice of each of them displayed different phenotypes (Hendrich et al., 2001; Hendrich et al., 1999; Hendrich and Bird 1998). Mice lacking MBD3 died during embryogenesis, while MBD2 mutant mice were viable and fertile (Hendrich et al., 2001). However, the maternal behavior of

MBD2 knockout mice was defective, as their average litter size was half of that found in wild-type mice. Therefore, lacking MBD2 affects the mother's ability to carry or deliver viable offsprings (Hendrich et al., 2001).

In the fruit fly *Drosophila melanogaster*, which lacks DNA methylation, an ortholog of MBD2/3 is present in the genome (Marhold et al., 2004a). Similar to observations in the mammalian genome, there is a suggested association between MBD2/3 and the NuRD complex (Ballestar et al., 2001; Tweedie et al., 1999). The drosophila MBD2/3 gene is specifically expressed during embryogenesis in two isoforms that are developmentally regulated (Ballestar et al., 2001; Marhold et al., 2004a; Tweedie et al., 1999). However, MBD2/3 homozygous mutant flies were found to be viable and fertile, indicating that MBD2/3 is not essential for Drosophila embryogenesis (Marhold et al., 2004b). The mutant animals, despite being viable, show chromosome segregation defects, indicating a potential role for MBD2/3 in the stability of pericentric heterochromatin (Marhold et al., 2004b).

Another example of an invertebrate that lacks a DNA methylation system but has conserved MBD2/3 is the flatworm *Schmidtea mediterranea*. RNAi knockdown showed that this gene is required for appropriate differentiation of planarian adult stem cells (pASC) during both regeneration and tissue homeostasis (Jaber-Hijazi et al., 2013).

Observations from planarian and Drosophila suggest that the function of MBD2/3 could be independent of DNA methylation, most likely it is mainly involved in transcriptional silencing through the NuRD complex (Marhold et al., 2004a, b; Kaji et al., 2007; Dattani et al., 2019). Investigating the role of MBD2/3 in more species, where CpG methylation is present or absent, is still required to understand its function in the context of DNA methylation and/or methylation-independent function as a component of the NuRD silencing complex.

4.2 CRISPR/Cas 9 knockout in embryos

Parhyale has emerged as model for developmental genetic studies (Calvo et al., 2022; Tserevelakis et al., 2022; Pavlopoulos and Wolff 2020; Bruce and Patel 2020; Price et al., 2010; Hannibal et al., 2012a; Ozhan-kizil et al., 2009; Gupta and Extavour 2013). The utility of Clustered regularly interspaced short palindromic repeats with Cas9 (CRISPR/Cas9) system has already been established in *Parhyale* and used to explore the role of different genes involved in limb development and diversity (Martin et al. 2016, Serano et al. 2016, Kao et al. 2016; Bruce and Patel et al., 2020; Clark-Hachtel and Tomoyasu 2020; Alberstat *et al.*, 2022; Sun *et al.*, 2022). In these studies, CRISPR was successfully used in zygotic injections to produce somatic mutagenesis and analyze the function of targeted genes, as well as CRISPR knock-in approach for fluorescent tagging of genes. CRISPR is a type II bacterial adaptive immune system that has been appropriated for genome engineering (Cong et al., 2013). The system is adapted from *Streptococcus pyogenes* to facilitate site-specific DNA cleavage using a complex of single guide RNA (sgRNA) and non-specific Cas9 endonuclease. The formation of the Cas9/gRNA complex allows the gRNA to guide Cas9 to genomic DNA target complementary to the gRNA sequence, where Cas9 induces a double-strand break (DSB). DSBs are repaired through two different mechanisms: non-homologous end-joining (NHEJ), or homology-directed repair (HDR) (Iliakis et l., 2004). NHEJ is an error-prone mechanism that modifies DNA broken ends by ligating them together regardless of the homology between them, generating deletions or insertions (indels) mutations. On the other hand, HDR applies a DNA template with a sequence homologous to the DSB location to seal it in an error-free manner (Iliakis et l., 2004; Mao et al., 2008).

The sgRNA is divided into two parts, a scaffold to which Cas9 binds to and a twenty-nucleotides protospacer that is complementary to the target genomic DNA sequence to be edited (Bassett et al., 2014; Jinek et al., 2012). The Cas9 protein recognizes a protospacer adjacent motif (PAM) sequence, which is a three nucleotides sequence downstream of the 20 bp unique protospacer sequence. The *S.pyogenes* system uses "NGG" PAM (Jinek et al., 2012; Ran et al., 2013; Liu et al., 2015).

To assess the potential role of DNA methylation through the mediator genes during embryonic development, we used CRISPR/Cas9 mediated gene knockout to generate loss of function mutations.

4.2.1 Experimental design

In our experiment, coding sequences are targeted to induce a loss of function mutation. Guide RNA (gRNA) and Cas9 protein were injected into 1-cell embryos, both cells of the 2-cell embryos, or 1 cell of the 2-cell embryos to allow comparison of wild-type versus mutant phenotype in the same animal, given that the 2-cell stage gives rise to the right and left axis of the *Parhyale* body.

Guide RNAs(gRNAs) were designed near to the 5' end to ensure a high likelihood of creating frameshift mutation by the NHEJ repair mechanism and eventual disruption of protein function. At least two guides were designed for each gene to increase the chance of successful knockout. Off-target effects might arise by the CRISPR/Cas9 system, which refers to creation of nonspecific and unintended mutations outside the target site (Zhang et al., 2015). Therefore, potential off-target effects were predicted by multiple tools. BLAST was used to blast the sgRNA sequence against the *Parhyale* genome assembly to confirm the gRNA sequence we chose is specific to the target site (Altschul et al., 1990). In addition, off-target loci prediction was available on the same tools used to design sgRNAs like ZiFit (Sander et al., 2007; Sander et al., 2010) and Synthego (ICE Analysis. 2019.v3.0. Synthego).

Because *Parhyale* has high heterozygosity, cloning full-length coding sequence of each gene from multiple individuals and sequencing was performed to confirm the DNA sequence and to identify single-nucleotide polymorphisms (SNPs) in each gene (as presented in chapter 2). This was important during gRNA designing process, as regions with many SNPs should be avoided because the existence of SNPs in the guide sequence might cause a failure of the CRISPR/Cas9 system to recognize and cut the endogenous gene.

In vitro digestion using a plasmid with an insertion containing the relevant coding sequence as a template was carried out to test the efficiency of each guide (Figure 4.1). As shown in the gel images (Figure 4.1),

each *in vitro* assay is comprised of five reactions. A negative control showing the circular plasmid with no gRNA, restriction enzyme or Cas9 is added. The second reaction contains a restriction enzyme that has a restriction site in the plasmid and the gRNA only added to the plasmid. In this reaction, the plasmid is linearized by the restriction enzyme. In the third reaction Cas9 protein is added with the same restriction enzyme and the gRNA. The plasmid here should be linearized by the restriction enzyme first then cut by the Cas9/gRNA complex. The third reaction confirms that the gRNA is cutting efficiently. The fourth reaction is a negative control using only gRNA with the plasmid and should look the same as the first reaction. Finally, the fifth reaction tests the ability of the Cas9/gRNA complex to linearize the plasmid in the same way as the restriction enzyme, therefore, it should appear similar to the second reaction.

We examined three DNA methylation mediator genes in *Parhyale*: DNMT1, DNMT3 and MBD2/3. A guide targeting *Distalless* gene that was used in a previous study (Kao et al.2016) that produced animals with truncated legs was used as a positive control to ensure quality of the injections. The following sections summarize the results of these experiments.

***Distallesss* knockout positive control:**

Distalless (*PhDll-e*) is a highly conserved leg patterning gene with a specific role in animal limb development, In Kao et al., 2016, a knockout (KO) for *PhDll-e* was performed using CRISPR/Cas9. The resulting mutants exhibited truncated limbs. To establish a positive control for our own CRISPR experiment, we used the same gRNA (denoted DII) reported in Kao et al., 2016 as the most efficient. As shown in Figure 4.2, we were able to generate mutants with truncated limbs using the DII2 gRNA. Details of the injections are reported in table 4.1. Multiple concentrations of the Cas9/gRNA complex were tested to choose the optimal concentration, with a constant 2:1 ratio between Cas9 to gRNA, as recommended in most CRISPR KO studies. We found that a concentration of 500 ng/µl Cas9: 250 ng/µl gRNA consistently produced mutant animals and was therefore used in subsequent injections.

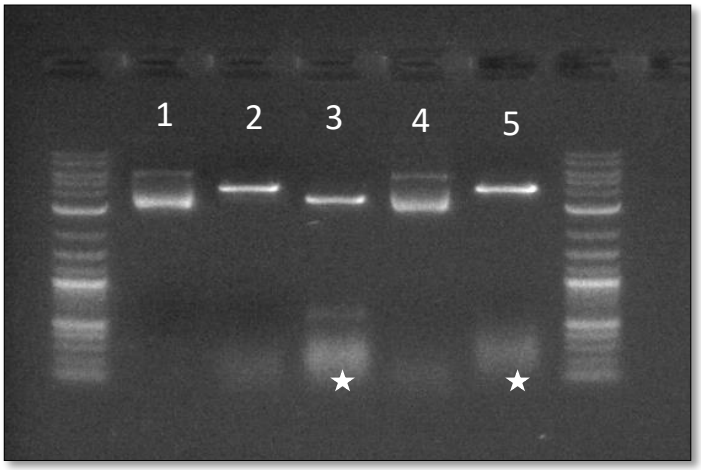
Our positive control injections, using the same gRNA, yielded mutants with a penetrance percentage that matched certain batches reported by Kao et al. (see Table 4.1). It is worth noting that the extent of inbreeding within each laboratory culture may have a substantial impact on variations in heterozygosity, consequently influencing the number of mutants.

Table 4.1. Summary of Positive control injections of PhDll-e knockout using CRISPR-Ca9 based targeted genome editing.

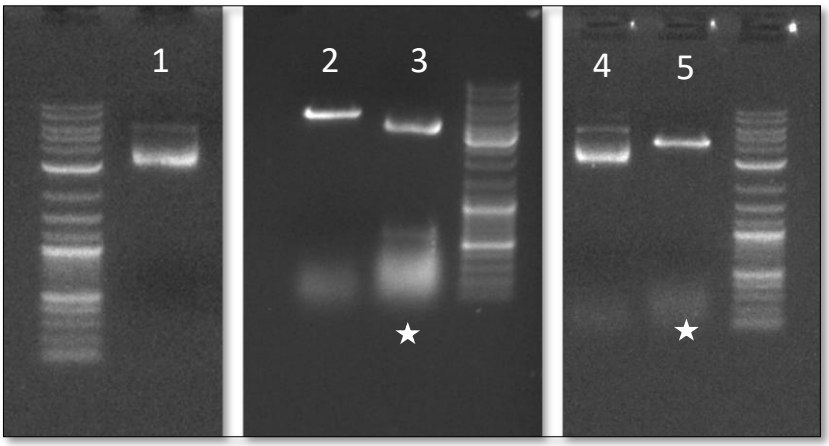
Cas9 (ng/µl)	**sgRNA (ng/µl)**	**No. 1-cell injected**	**No. wild-type embryos**	**No. mutant embryos**	**Survival**
400	200	90	62	5	74%
400	200	32	16	0	50%
400	200	57	40	1	72%
600	300	145	100	0	69%
500	250	50	32	1	64%
500	250	150	96	2	65%
500	250	100	63	4	67%

Figure 4.1. *In vitro* digestion assay for all gRNAs used in this study. *In vitro* digestions are displayed for **(A)** DNMT1-1 gRNA. **(B)** DNMT1-2 gRNA. **(C-G in the next page)** DNMT3-1 gRNA, DNMT3-2 gRNA, MBD2/3-1, MBD2/3-2 gRNA, and MBD2/3-22 gRNA. Gel images depict five reactions for each gRNA *in vitro* digestion assay. Reaction number **1** contains only the plasmid containing the relevant CDS, and it is always circular. Reaction number **2** is the linearized plasmid through restriction enzyme digestion (*Dra III, Sca I or NcoI*). Reaction number **3** depicts a double-digested plasmid through restriction enzyme digestion and CRISPR/Cas9 complex cleavage. In the **4**th reaction, the plasmid is combined with gRNA only and it remains circular as it's a negative control. Finally, reaction **5** shows the ability of CRISPR-Ca9 complex to cleave target DNA and linearize the plasmid the same way as done through restriction enzyme digestion. Star sign indicate the cas9/gRNA complex.

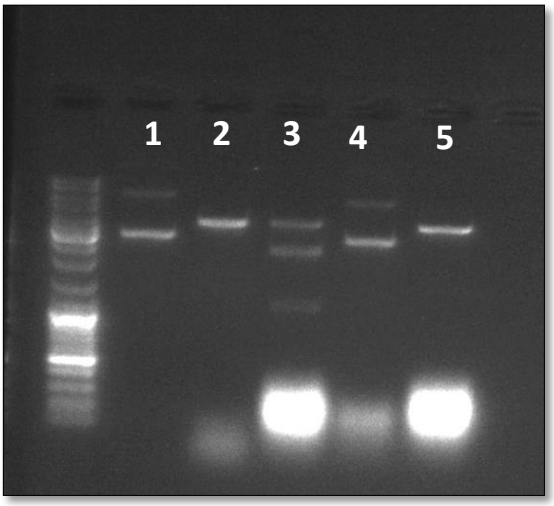
A



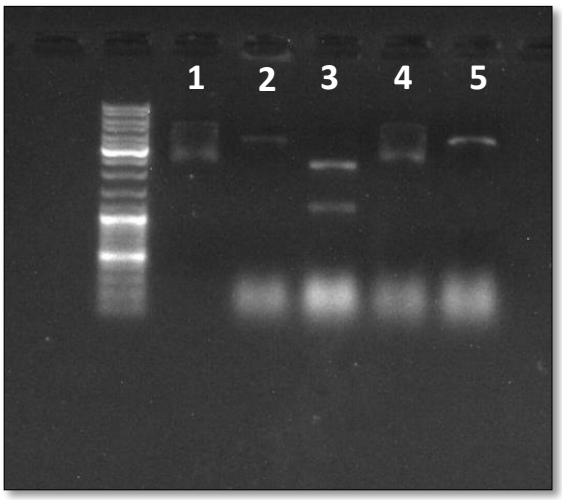
B



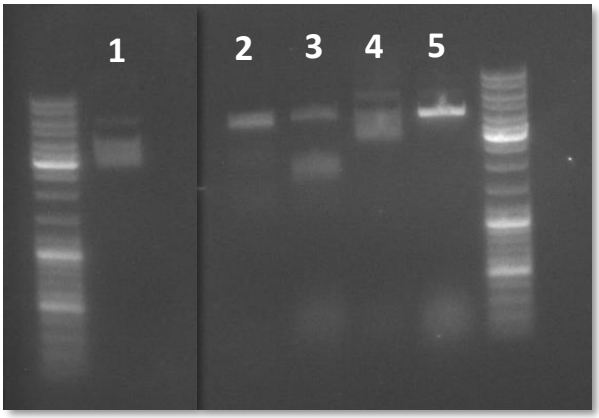
C



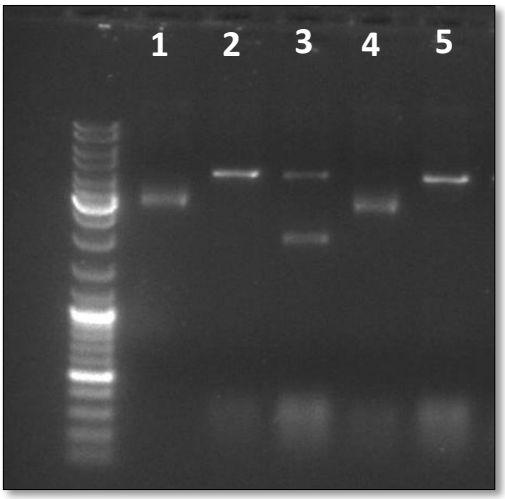
D



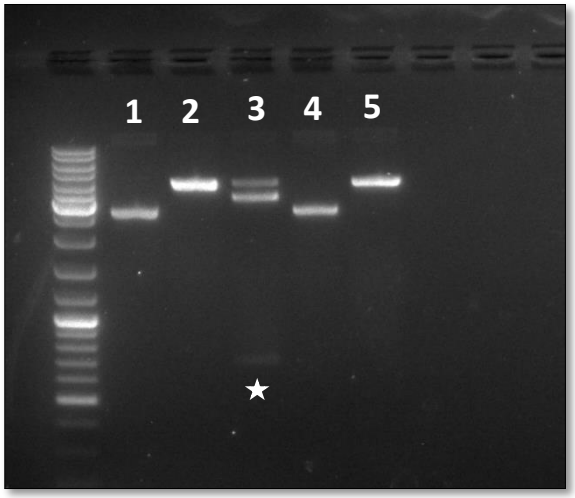
E



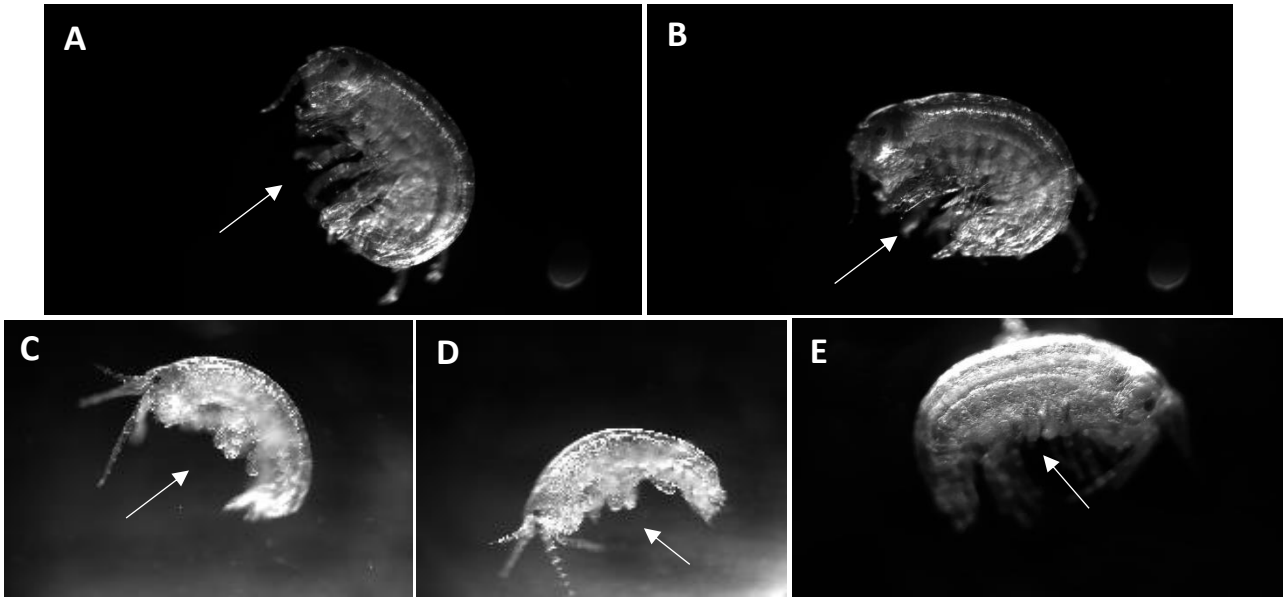
F



G



<u>Figure 4.2. Shows a CRISPR/Cas9 experiment using *Distallesss* knockout as positive control.</u> (A) and **(B)** Are images of wildtype embryos. **(C-E)** Depict mutant *Distalless* Ko embryos, generated using the DII 2 gRNA from Kao et al., 2016 to create animals with truncated limbs. The same gRNA was used in this study as positive control, and the truncated limbs phenotype reported in Kao et al. was successfully replicated. Animals in **(C)** and **(D)** show mutants with completely truncated limbs, while the hatchling in **(E)** has truncated limbs on the right axis of the animal body and normally developed limbs on the left axis. This result is due to the injection of the CRISPR/Cas9 mix into 1-cell of the 2-cell embryo. Arrows in C, D and E indicate the truncated limbs in mutant animals, as opposed to wild-type animals in A and B with normally developing limbs.

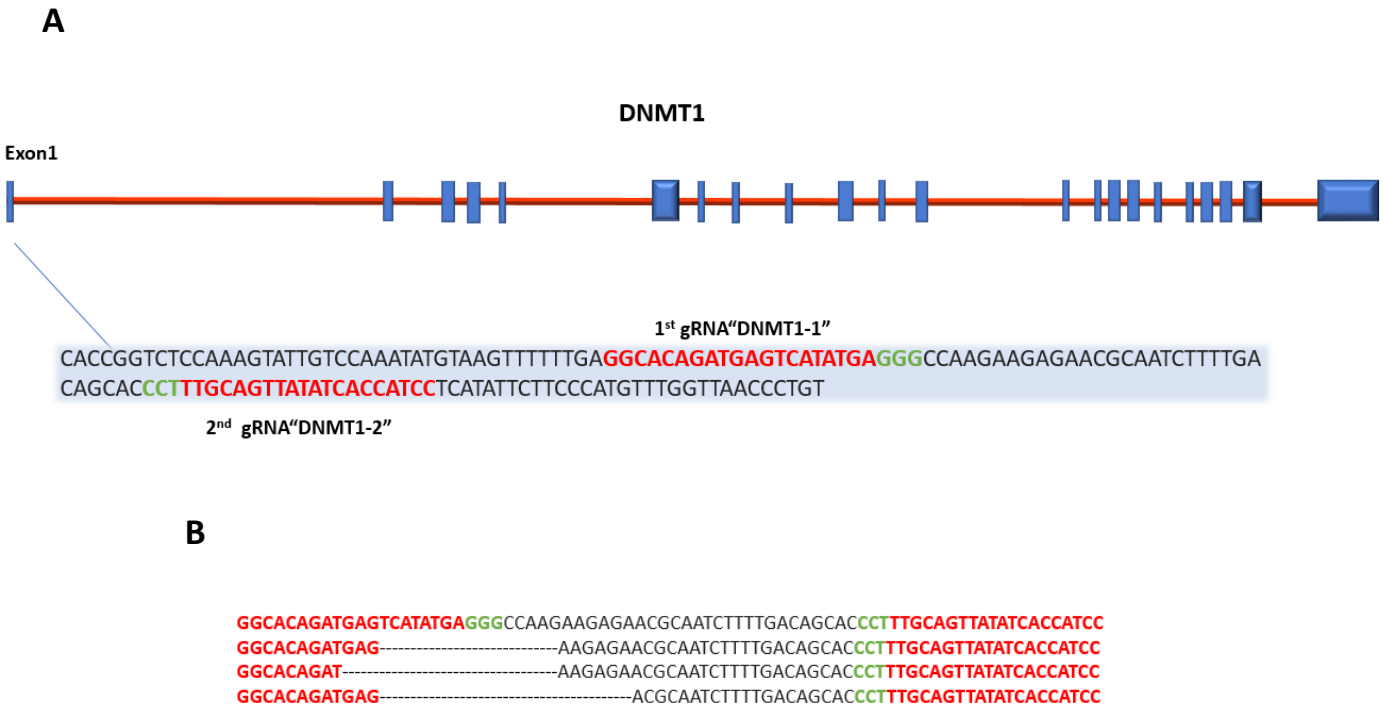


4.2.2 DNMT1 is essential for embryogenesis

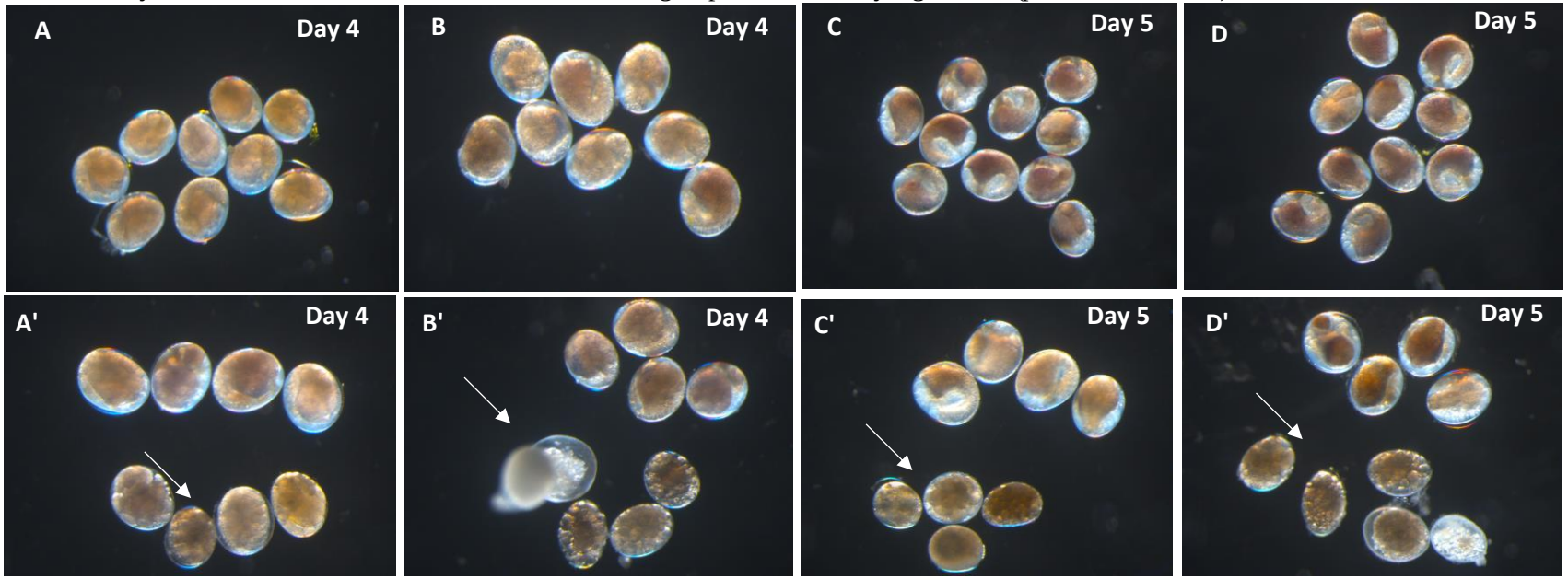
We used two gRNA located in the first exon (Figure 4.3), which were very close to each other. Injections were performed using both gRNAs at the same time to increase the chance of creating DSBs, after testing each one of them separately. Injections with only gRNA at the same concentration were used as a negative control injection. Additionally, injections using only Cas9 protein and injections with different dyes like Rhodamine-B, Phenol-red and Dextran were used as negative controls to assess mortality rate in comparison with CRISPR injections (results shown in Table 4.2). The results of DNMT1 KO are summarized below in Table 4.3 and Figure 4.4 The mean survival rate between KO embryos and control embryos is significantly different; a statistical comparison was performed via unpaired 2-tailed *t*-test (*p*-value < 0.05, " 3.078e-12"). CRISPR/Cas9 injected embryos failed to complete embryogenesis and were confirmed to be dead by day 4 or 5 of development. Earlier stages were similar between the knockout embryos and control group, but the effect probably started earlier than the time point where mutants have already begun to degenerate (Figure 4.4). Given that the gene is provided maternally, as was seen from the analysis in the third chapter, it is expected to start to see the effect at a later stage after the activation of the zygotic genome. Mutant embryos never seemed to reach stage 14 of development, where the germband starts to expand posteriorly to the full length of the egg, and the bilateral anlagen of the midgut forms two well-defined circular structures (Browne et al., 2005).

We extracted genomic DNA from knockout and control embryos to screen for mutations. Notabley, we identified deletions proximal to the DNMT1-1 gRNA location using Sanger sequencing as illustrated in Figure 4.3 B.

Figure 4.3. DNMT1 CRISPR/Cas9 knockout. **(A)** Diagram of DNMT1 gene structure with a zoom into exon 1 showing gRNAs locations. **(B)** Mutations introduced by CRISPR/Cas9 KO of DNMT1 in the target site shown by Sanger sequencing in the region flanking both gRNAs.



<u>Figure 4.4. DNMT1 KO phenotype</u>. (A to D) Control embryos developing normally, imaged during day 4 and 5 of development. **(A' to D')** Show DNMT1 KO embryos at the same time point. Arrows point to dead embryos that have begun to degenerate. **(E)** Boxplot comparing the number of hatched embryos; the difference between the KO and control groups is statistically significant (*p*-value 3.078e-12).



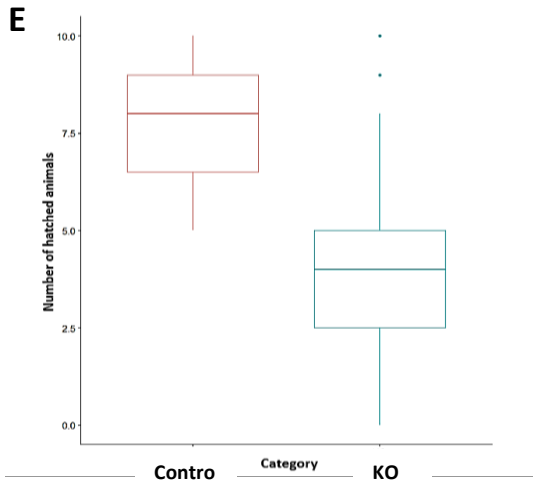


Table 4.2. Injection records of different types of negative controls for CRISPR/Cas9 experiment.

Type of Injection	Original No. of Embryos	Hatched No. of Embryos	Percentage	Average
Plain (No injection)	11	9	82%	90%
	11	10	90%	
	10	10	100%	
	10	10	100%	
	10	9	90%	
	10	8	80%	
Cas9 only	10	8	80%	77%
	10	8	80%	
	10	7	70%	
	10	9	90%	
	10	7	70%	
	10	8	80%	
	10	6	60%	
	10	6	60%	
	10	7	70%	
	10	8	80%	
	10	9	90%	
	10	7	70%	
	10	8	80%	
	10	9	90%	
	10	8	80%	
	10	8	80%	
	10	8	80%	

Table 4.2. continued. Injection records of different types of negative controls for CRISPR/Cas9

Type of Injection	Original No. of Embryos	Hatched No. of Embryos	Percentage	
			individual	average
Rhodamine-B only	10	9	90%	79%
	10	8	80%	
	10	8	80%	
	10	7	70%	
	10	6	60%	
	10	9	90%	
	10	8	80%	
	10	8	80%	
	10	6	60%	
	10	8	80%	
	10	8	80%	
	10	9	90%	
	10	9	90%	
	10	9	90%	
	10	7	70%	
	10	8	80%	
Dextran only	23	20	87%	80%
	27	24	89%	
	23	18	78%	
	21	15	71%	
	25	19	76%	
	20	16	80%	
	22	19	86%	
	20	17	85%	
	25	17	68%	
	24	20	83%	
	23	18	78%	
	25	20	80%	
	20	16	80%	
	23	19	83%	
	21	16	76%	
	25	21	84%	
	27	20	74%	
	30	24	80%	
	23	20	87%	

Table 4.3. CRISPR/Cas9 injections of DNMT1 knockout

Negative control (guide only)					CRISPR/Cas9				
Guide	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**	**Guide**	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**
DNMT1-1	10	7	7	70%	DNMT1-1	10	3	3	30%
	10	6	6	60%		10	0	0	0%
	10	5	5	50%		10	2	2	20%
	10	5	5	50%		10	3	3	30%
	10	5	5	50%		10	2	2	20%
	10	7	7	70%		10	2	2	20%
	10	10	10	100%		10	7	7	70%
	10	9	9	90%		10	3	3	30%
	10	8	8	80%		10	5	5	50%
	10	7	7	70%		10	4	4	40%
	10	8	8	80%		10	4	4	40%
	10	7	7	70%		10	10	10	100%
	10	7	7	70%		10	9	9	90%
	10	9	9	90%		10	8	8	80%
	10	8	8	80%		10	3	3	30%
	10	7	7	70%		10	3	3	30%
	10	6	6	60%		10	0	0	0%
	10	5	5	50%		10	2	2	20%
	10	5	5	50%		10	3	3	30%
	10	5	5	50%		10	2	2	20%
	10	7	7	70%		10	2	2	20%
	10	10	10	100%		10	7	7	70%
	10	9	9	90%		10	3	3	30%
	10	8	8	80%		10	5	5	50%
Average	70.8%					40%			
DNMT1-2	10	7	7	70%	DNMT1-2	10	4	4	40%
	10	7	7	70%		10	3	3	30%
	10	6	6	60%		10	8	8	80%
	10	8	8	80%		10	2	2	20%
Average	70%					42.5%			

Table 4.3 continued. CRISPR/Cas9 injections of DNMT1 knockout

Negative control (guide only)					CRISPR/Cas9				
Guide	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**	**Guide**	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**
DNMT1-1 & DNMT1-2	10	8	8	80%	DNMT1-1 & DNMT1-2	10	5	5	50%
	10	10	10	100%		10	5	5	50%
	10	9	9	90%		10	2	2	20%
	10	9	9	90%		10	4	4	40%
	10	5	5	50%		10	6	6	60%
	10	9	9	90%		10	3	3	30%
	10	9	9	90%		10	4	4	40%
	10	8	8	80%		10	4	4	40%
	10	9	9	90%		10	4	4	40%
	10	8	8	80%		10	4	4	40%
	10	6	6	60%		10	2	2	20%
	10	7	7	70%		10	6	6	60%
	10	8	8	80%		10	4	4	40%
	10	9	9	90%		10	6	6	60%
	10	8	8	80%		10	7	7	70%
Average	81%					44%			

The data presented in Table 4.3 demonstrate that the average survival rate observed when using two gRNAs is comparable to that obtained when using only one gRNA.

4.2.3 DNMT3 KO embryos have the highest survival rate

As mentioned in the introduction, many invertebrate species have lost the *de novo* methyltransferase DNMT3. Even in the species that conserved DNMT3, in some of them, losing DNMT3 has no obvious effect on the embryonic development (Zwier et al., 2012, Bewick et al., 2019). Moreover, other studies have shown that the existence of DNMT1 is always positively correlated with DNA methylation, whereas the same is not observed for DNMT3 (Bewick et al., 2017). Another study demonstrated that DNMT1 has the potential to act as *de novo* methyltransferase. This data suggests that DNMT1 might be able to play both *de novo* and maintenance functions of DNA methylation, while DNMT3 could have a minor role that is redundant compared to DNMT1 (Fatemi et al., 2002; Ventos-Alfonso et al., 2020).

In our experiment, gRNA for DNMT3 were designed at the first and second exons of CDS, away from single nucleotide polymorphism sites (Figure 4.5). There was only one gRNA available in the first exon. Moreover, the gRNA designed in an overlapping site with the gRNA in the second exon did not cut in the *in vitro* cleavage assay. Therefore, DNMT3 injections were always performed using single gRNA. Unlike DNMT1 KO, survival rate of DNMT3 KO embryos was closer to the control group (results shown in Table 4.4 and figure 4.6). Despite the significance of *t*-test *p*-value (1.333e-11), the difference between the mean of the two groups is smaller than what was observed in the DNMT1-KO case, as can be seen in Figure 4.6 below. The high survival rate for many injection batches suggests that DNMT3 might not be essential for embryogenesis in *Parhyale*. The observation reported in the previous chapter about DNMT3 expression pattern, which consistently exhibited the lowest expression values among both maternal and zygotic transcripts, provides compelling support for this theory.

Figure 4.5. DNMT3 CRISPR/Cas9 knockout. Diagram of DNMT3 gene structure, with a zoom into exon 1 and exon 2, highlighting the locations of the guide RNAs (gRNAs) used in the knockout experiment.

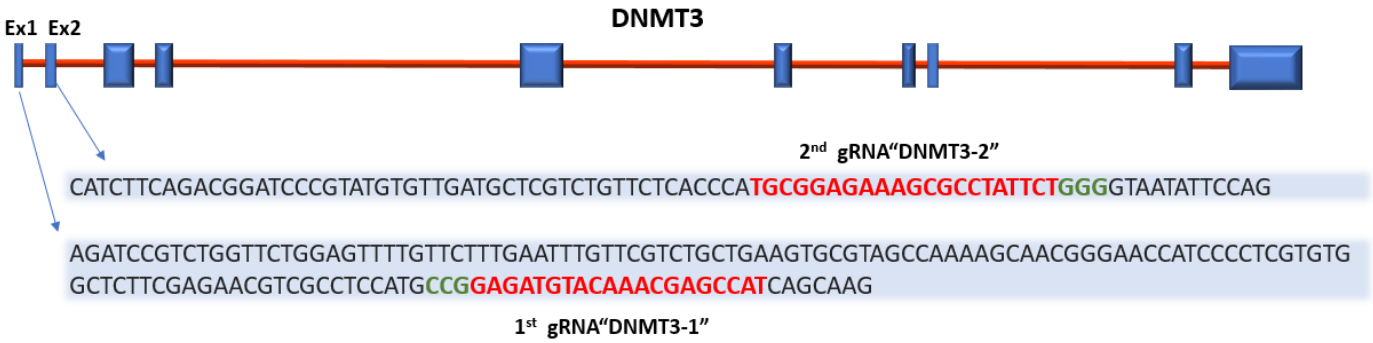
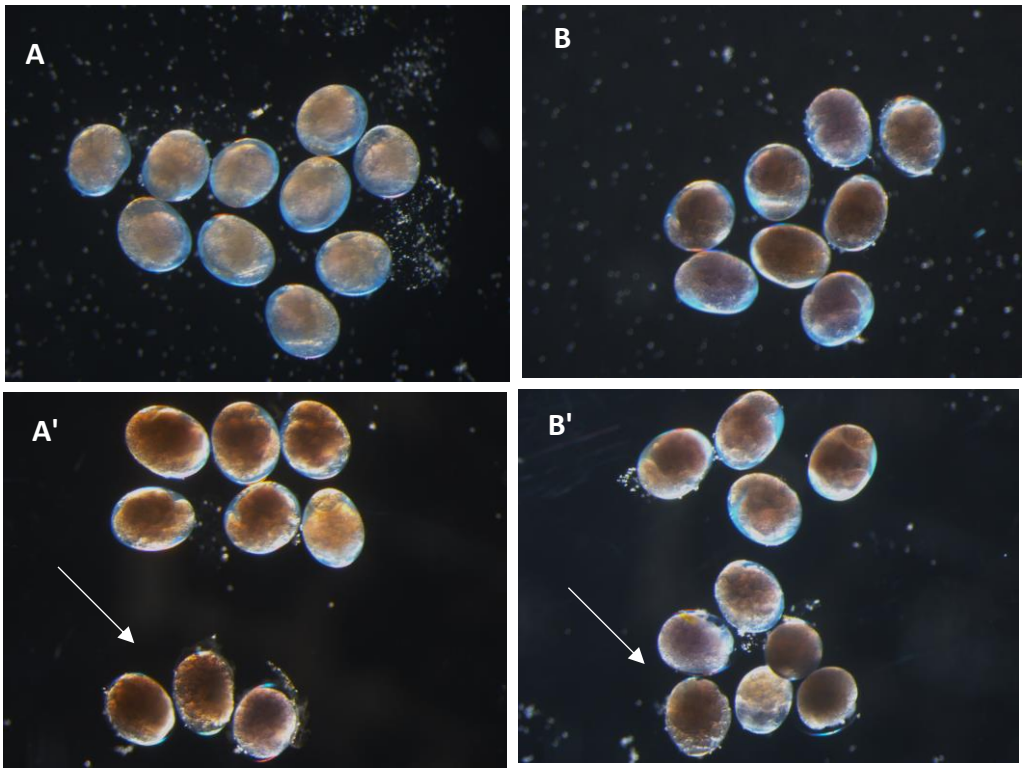


Figure 4.6. DNMT3 KO phenotype. **(A)** and **(B)** show control embryos developing normally at day 4 of development (around stage19), while **(A')** and **(B')** depict DNMT3-KO embryos with an arrow indicating dead embryos. **(C)** Presents a boxplot comparing difference in survival rates between KO and control group, with a statistically significant difference observed (*p*-value 1.333e-11).



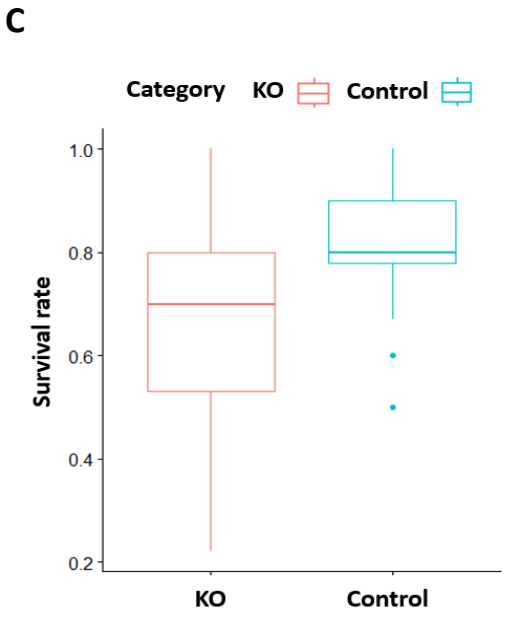


Table 4.4. summary of CRISPR/Cas9 injections of DNMT3knockout

Negative control (guide only)					CRISPR/Cas9				
Guide	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**	**Guide**	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**
DNMT3-2	10	10	10	100%	DNMT3-2	9	6	6	66%
	10	8	8	80%		9	4	4	44%
	10	9	9	90%		10	3	3	30%
	10	8	8	80%		10	5	5	50%
	10	10	10	100%		5	5	5	100%
	11	10	10	91%		6	6	6	100%
	7	6	6	86%		10	6	6	60%
	9	7	7	78%		10	5	5	50%
	9	7	7	78%		7	3	3	43%
	8	7	7	88%		10	5	5	50%
	10	9	9	90%		8	8	8	100%
	10	10	10	100%		7	7	7	100%
	10	9	9	90%		9	8	8	89%
	10	10	10	100%		9	8	8	89%
	9	7	7	78%		10	5	5	50%
	10	8	8	80%		10	7	7	70%
	10	10	10	100%		10	8	8	80%
	10	6	6	60%		10	6	6	60%
	28	22	22	79%		7	6	6	86%
	10	9	9	90%		10	7	7	70%
	10	7	7	70%		9	8	8	89%
	10	8	8	80%		4	3	3	75%
	10	7	7	70%		10	5	5	50%
	10	7	7	70%		6	3	3	50%
	10	9	9	90%		7	5	5	71%
	10	9	9	90%		9	5	5	56%
	10	8	8	80%		9	8	8	89%
	10	9	9	90%		12	8	8	66%
	15	13	13	87%		16	14	14	88%

Table 4.4 continued. summary of CRISPR/Cas9 injections of DNMT3knockout

Negative control (guide only)					CRISPR/Cas9				
Guide	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**	**Guide**	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**
DNMT3-2	20	20	20	100%	DNMT3-2	14	12	12	86%
	15	15	14	93%		9	5	5	55%
	13	11	11	85%		10	7	7	70%
	10	8	8	80%		9	6	6	64%
	13	11	11	85%		9	8	8	89%
	12	10	10	83%		9	7	7	78%
	15	13	13	87%		8	6	6	75%
	18	15	15	83%		10	7	7	70%
	15	15	15	100%		9	8	8	88%
	15	12	12	80%		11	8	8	73%
	14	13	13	93%		13	7	7	54%
	15	12	12	80%		14	10	10	71%
	22	18	18	82%		10	8	8	80%
	14	12	12	86%		8	6	6	75%
	NA	NA	NA	NA		7	5	5	71%
	NA	NA	NA	NA		9	8	8	89%
	10	8	8	80%		10	10	10	100%
	10	8	8	80%		10	10	10	100%
	10	6	60	60%		10	8	8	80%
	10	8	8	80%		10	8	8	80%
	10	10	10	100%		10	7	7	70%
	10	9	9	90%		10	7	7	70%
	10	8	8	80%		10	7	7	70%
	10	7	7	70%		10	6	6	60%
	10	7	7	70%		10	8	8	80%
	10	6	6	60%		10	8	8	80%
	10	6	6	60%		10	8	8	80%
	10	10	10	100%		10	7	7	70%
	10	9	9	90%		10	8	8	80%
	10	8	8	80%		10	8	8	80%
	10	7	7	70%		10	8	8	80%
	10	7	7	70%		10	7	7	70%
	10	6	6	60%		10	7	7	70%
	10	6	6	60%		10	7	7	70%
	10	6	6	60%		10	5	5	50%
	10	8	8	80%		10	7	7	70%
	10	7	7	70%		10	7	7	70%
	10	10	10	100%		10	7	7	70%
	10	6	6	60%		10	7	7	70%
	10	6	6	60%		10	8	8	80%

Table 4.4 continued. summary of CRISPR/Cas9 injections of DNMT3knockout

Negative control (guide only)					CRISPR/Cas9				
Guide	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**	**Guide**	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**
DNMT3-2	10	7	7	70%	DNMT3-2	10	7	7	70%
	10	8	8	80%		10	7	7	70%
	6	4	4	67%		10	9	9	90%
	NA	NA	NA	NA		10	9	9	90%
	NA	NA	NA	NA		10	6	6	60%
	NA	NA	NA	NA		10	7	7	70%
	NA	NA	NA	NA		10	8	8	80%
	NA	NA	NA	NA		10	7	7	70%
Average	81.2%					74%			
DNMT3-1	10	8	8	80%	DNMT3-1	8	5	5	63%
	10	8	8	80%		9	5	5	55%
	10	8	8	80%		8	4	4	50%
	9	8	8	89%		10	4	4	40%
	10	9	9	90%		10	5	5	50%
	9	7	7	78%		10	5	5	50%
	10	9	9	90%		10	6	6	60%
	10	7	7	70%		10	6	6	60%
	8	7	7	88%		6	5	5	83%
	7	6	6	86%		9	6	6	66%
	10	6	6	60%		8	5	5	63%
	10	8	8	80%		10	8	8	80%
	10	9	9	90%		8	7	7	88%
	10	8	8	80%		9	7	7	78%
	10	10	10	100%		10	8	8	80%
	10	8	8	80%		10	6	6	60%
	10	9	9	90%		10	5	5	50%
	10	10	10	100%		8	5	5	63%
	10	9	9	90%		8	4	4	50%
	10	9	9	90%		10	5	5	50%
	10	8	8	80%		10	7	7	70%
	10	9	9	90%		10	7	7	70%
	10	8	8	80%		10	7	7	70%
	10	8	8	80%		10	7	7	70%
	10	8	8	80%		10	8	8	80%
	10	7	7	70%		10	8	8	80%
	10	8	8	80%		10	10	10	100%
	12	10	10	83%		10	6	6	60%
Average	89%					66%			

4.2.4 MBD2/3 KO fail to complete embryogenesis

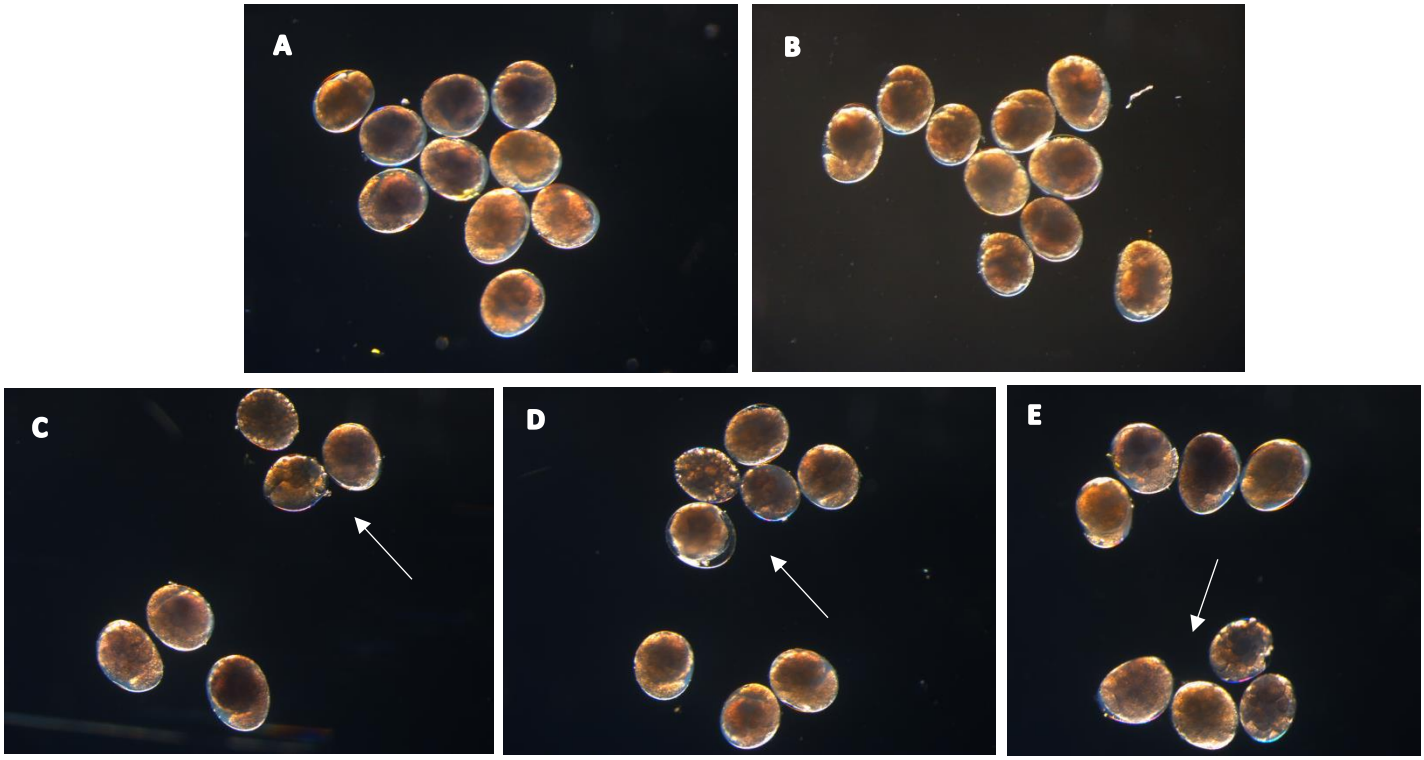
To increase the chance of successful knockout, gRNAs were designed at the first and second exon near the 5' end, ensuring that they were away from any identified SNPs. Two guides close to each other in exon 1 or overlapping in exon 2 were designed for MBD2/3 (shown in Figure 4.7 below). Injections were performed using either single gRNAs or double gRNAs at the same exon.

The knockout experiment of MBD2/3 in *Parhyale* resulted in lethal phenotype at a stage similar to what was observed in the DNMTs. The KO embryos were also arrested on the fourth day of development, as summarized in Table 4.5 and Figure 4.8 The difference in survival rate between the KO and the control groups were significant (unpaired two-tiled *t*-test *p*-value < 2.2e-16). As seen in Figure 4.8, the mean survival rate of the KO group was almost half that of the control group. This result indicates that MBD2/3 is important for embryonic development. Additional experiments are necessary to determine whether the function of MBD2/3 is related to DNA methylation machinery, and the NuRD complex or if it operates independently of DNA methylation.

CRIRPS/Cas9 experiments were performed in parallel while collecting RNA from early embryonic stages to investigate MZT dynamics. We discovered that all three genes targeted were maternally provided. Therefore, we decided that knockout experiments in embryos are not ideal for studying the effect of these genes. Even if we switch off the gene in the embryo, it will still be present maternally, and the effect we would see would only be after the maternal product was degraded, and the embryo is completely reliant on the zygotic version of the gene. As a result, in the second part of the chapter, we switched to a different type of experiment, RNA interference (RNAi) knockdown. We proceeded with a focus on MBD2/3 to investigate its role in the regulatory mechanism in *Parhyale*.

Figure 4.7. MBD2/3 CRISPR/Cas9 knockout Diagram of MBD2/3 gene structure, with a zoom into exon 1 and exon 2, highlighting the locations of the guide RNAs (gRNAs) used in this knockout experiment.

Figure 4.8. MBD2/3-KO phenotype (A & B) shows control embryos developing normally at day 4 of development (around stage19 of development) and **(C to E)** are the MBD2/3-KO embryos, arrow points to dead embryos. **(F)** boxplot comparing difference in survival rates between KO and control group, the difference between KO and control group is statistically significant (*p*-value < 2.2e-16).



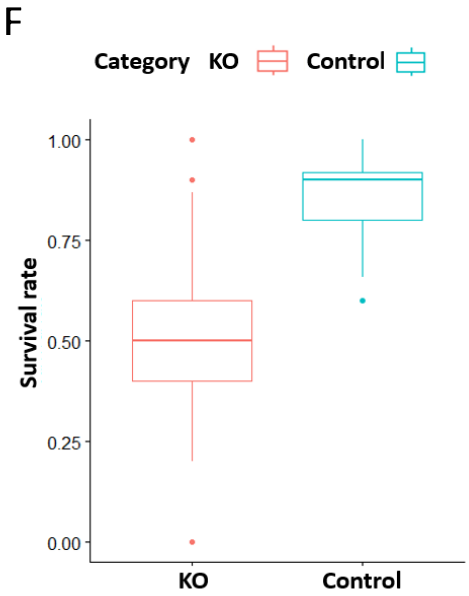


Table 4.5. summary of CRISPR/Cas9 injections of MBD2/3 knockout

Negative control (guide only)					CRISPR/Cas9				
Guide	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**	**Guide**	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**
MBD2/3-2	10	10	10	100%	MBD2/3-2	10	8	8	80%
	10	10	10	100%		10	5	5	50%
	10	10	10	100%		8	6	6	75%
	10	10	10	100%		7	2	2	29%
	10	9	9	90%		10	5	5	50%
	10	9	9	90%		10	0	0	0%
	10	10	10	100%		9	6	6	66%
	10	9	9	90%		10	5	5	50%
	8	6	6	75%		7	4	4	57%
	10	9	9	90%		10	5	5	50%
	9	9	9	100%		10	5	5	50%
	10	9	9	90%		10	8	8	80%
	10	8	8	80%		10	3	3	30%
	10	9	9	90%		10	6	6	60%
	10	7	7	70%		10	5	5	50%
	9	8	8	89%		10	6	6	60%
	10	10	10	100%		10	6	6	60%
	10	10	10	100%		10	8	8	80%
	10	10	10	100%		10	9	9	90%
	9	8	8	89%		10	3	3	30%
	10	10	10	100%		8	7	7	87%
	10	10	10	100%		8	0	0	0%
	10	10	10	100%		10	7	7	70%
	10	8	8	80%		10	3	3	30%
	10	9	9	90%		10	3	3	30%
	10	9	9	90%		10	2	2	20%
	8	7	7	88%		10	4	4	40%
	10	9	9	90%		6	6	6	100%
	10	6	6	60%		10	8	8	80%
Average	91%					54%			

Table 4.5 continued. summary of CRISPR/Cas9 injections of MBD2/3 knockout

Negative control (guide only)					**CRISPR/Cas9**				
Guide	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**	**Guide**	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**
MBD2/3-1	8	8	8	80%	MBD2/3-1	10	6	6	60%
	9	6	6	76%		10	4	4	40%
	10	10	10	100%		10	5	5	50%
	10	8	8	80%		10	6	6	60%
	10	8	8	80%		10	7	7	70%
	9	6	6	76%		9	5	5	55%
	10	6	6	60%		10	3	3	30%
	11	9	9	82%		12	8	8	66%
	12	10	10	83%		10	5	5	50%
	10	8	8	80%		10	5	5	50%
	11	9	9	82%		10	6	6	60%
	13	13	13	100%		7	2	2	29%
	9	7	7	78%		12	7	7	58%
	8	8	8	100%		12	6	6	50%
	13	12	12	92%		10	4	4	40%
	9	6	6	66%		10	5	5	50%
	9	7	7	78%		10	6	6	60%
	10	9	9	90%		10	5	5	50%
	10	8	8	80%		10	6	6	60%
	10	8	8	80%		12	7	7	58%
	9	7	7	78%		10	8	8	80%
	10	9	9	90%		11	6	6	55%
	13	12	12	92%		11	6	6	55%
	10	9	9	90%		13	7	7	54%
	10	9	9	90%		10	7	7	70%
	10	9	9	90%		10	4	4	40%
	13	12	12	92%		10	2	2	20%
	4	3	3	75%		10	3	3	30%

Table 4.5 continued. summary of CRISPR/Cas9 injections of MBD2/3 knockout

Negative control (guide only)					CRISPR/Cas9				
Guide	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**	**Guide**	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**
MBD2/3-1	8	8	8	80%	MBD2/3-1	10	6	6	60%
	9	6	6	76%		10	4	4	40%
	10	10	10	100%		10	5	5	50%
	10	8	8	80%		10	6	6	60%
	10	8	8	80%		10	7	7	70%
	9	6	6	76%		9	5	5	55%
	10	6	6	60%		10	3	3	30%
	11	9	9	82%		12	8	8	66%
	12	10	10	83%		10	5	5	50%
	10	8	8	80%		10	5	5	50%
	11	9	9	82%		10	6	6	60%
	13	13	13	100%		7	2	2	29%
	9	7	7	78%		12	7	7	58%
	8	8	8	100%		12	6	6	50%
	13	12	12	92%		10	4	4	40%
	9	6	6	66%		10	5	5	50%
	9	7	7	78%		10	6	6	60%
	10	9	9	90%		10	5	5	50%
	10	8	8	80%		10	6	6	60%
	10	8	8	80%		12	7	7	58%
	9	7	7	78%		10	8	8	80%
	10	9	9	90%		11	6	6	55%
	13	12	12	92%		11	6	6	55%
	10	9	9	90%		13	7	7	54%
	10	9	9	90%		10	7	7	70%
	10	9	9	90%		10	4	4	40%
	13	12	12	92%		10	2	2	20%
	4	3	3	75%		10	3	3	30%
Average	84%					52%			

Table 4.5 continued. summary of CRISPR/Cas9 injections of MBD2/3 knockout

Negative control (guide only)					CRISPR/Cas9				
Guide	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**	**Guide**	**Successfully injected day1**	**Day 5**	**Hatched**	**Survival rate**
MBD2/3-2 & MBD2/3-22	9	8	8	89%	MBD2/3-2 & MBD2/3-22	10	6	6	60%
	10	10	10	100%		10	4	4	40%
	12	9	9	75%		10	4	4	40%
	13	12	12	92%		13	6	6	46%
	14	10	10	71%		12	4	4	33%
	17	12	12	71%		7	5	5	71%
	12	11	11	92%		11	6	6	55%
	10	9	9	90%		10	7	7	70%
	10	9	9	90%		10	9	9	90%
	10	8	8	80%		10	4	4	40%
	10	10	10	100%		10	3	3	30%
	10	8	8	80%		10	7	7	70%
	10	7	7	70%		10	5	5	50%
	10	8	8	80%		10	7	7	70%
	11	11	11	100%		10	3	3	30%
	11	9	9	81%		10	5	5	50%
	11	10	10	90%		11	5	5	45%
	11	10	10	90%		10	7	7	70%
	10	8	8	80%		11	6	6	54%
	10	8	8	80%		10	7	7	70%
	10	6	6	60%		10	3	3	30%
	13	11	11	85%		11	3	3	27%
Average	84%					52%			

4.3 CRISPR/Cas9 knock-in project

After conducting knockout experiments to test gRNAs for targeted genes, we aimed to fluorescently tag animals using homology-independent CRISPR/Cas9 knock-in to fluorescently tag animals and visualize patterns of expression of DNA mediator genes throughout the life history of *Parhyale*. We started by building the donor plasmid from a plasmid that had been previously constructed in Kao et al., 2016, which was designed to tag any genomic locus because using the non-homologous-end-joining (NHEJ) repair mechanism. The construct included the following components (illustrated in figure 4.9):

- The gRNA Sequence of the relevant gene.
- The coding sequence (CDS) of the targeted gene (for each of DNMT3 and MBD2/3) to rescue the endogenous gene.
- Histone *H2B* and Ruby2 red fluorescent reporter.
- Self-cleavage *T2A* peptide fused in frame with the tagging sequence and the CDS and the fluorescent reporter (Ruby) to allow co-expression of the functional protein and the fluorescent tag.
- The *3'UTR* sequence of the same gene.
- pGEM-T easy vector backbone.

Upon injection and cleavage by Cas9 protein, the endogenous coding sequence would be restored from the tagging plasmid in a bicistronic mRNA that expresses the functional DNA methylation mediator protein and the fluorescent reporter.

We generated constructs for DNMT3 and MBD2/3 (see Appendix E), as their CDS length is much shorter than that of DNMT1. *In vitro* cleavage assay showed that the construct was linearized by the Cas9/gRNA complex for both genes. However, we did not obtain any fluorescently tagged embryos after injections. While Kao et al., (2016) reported a tagging efficiency of only 6.6%, our experiment, as well as our colleague's experiment on a different gene, did not yield any successfully tagged embryos.

One of the batches injected with MBD2/3 knock-in construct exhibited 100% fluorescence, which was observed throughout the embryos and persisted until later stages of development before starting to weaken. However, the signal from this batch was unlikely to be due to a successful knock-in, as it is possible that the fluorescent reporter was being expressed independently of the relevant gene (Figure 4.10). To investigate this possibility, we raised these embryos to later stages and collected genomic DNA to sequence the target site, but the sequencing results did not reveal any changes in the wild-type sequence. We performed more control injections using only the plasmid without CRISPR mix to test this hypothesis, and some embryos were fluorescent. Therefore, the combination of the low efficiency of the knock-in method and the possibility of random fluorescence signal generation creates several challenges in using this approach to tag embryos. As a result, other tagging techniques that may be more efficient should be considered in the future.

Figure 4.9. CRISPR/Cas9 knock-in approach. This schematic shows the construct used for knock-in experiments, using the MBD2/3 construct as an illustration. The tagging plasmid carries a copy of the coding sequence of the relevant gene (shown in yellow), the *T2A* self-cleaving peptide (in green), a fusion of the *Parhyal*e histone *H2B* (in purple) with the *Ruby2* monomeric red fluorescent protein (in red), and the *3'UTR* sequence (in purple). The tagging plasmid and CRISPR/Cas9 mix are injected into 1-cell *Parhyale* embryos. The endogenous gene would be cut by the CRISPR/Cas9 mixed and repaired by the sequence from the tagging plasmid. If the tagging is successful, it would result in a functional protein fused with a nuclear fluorescent reporter.

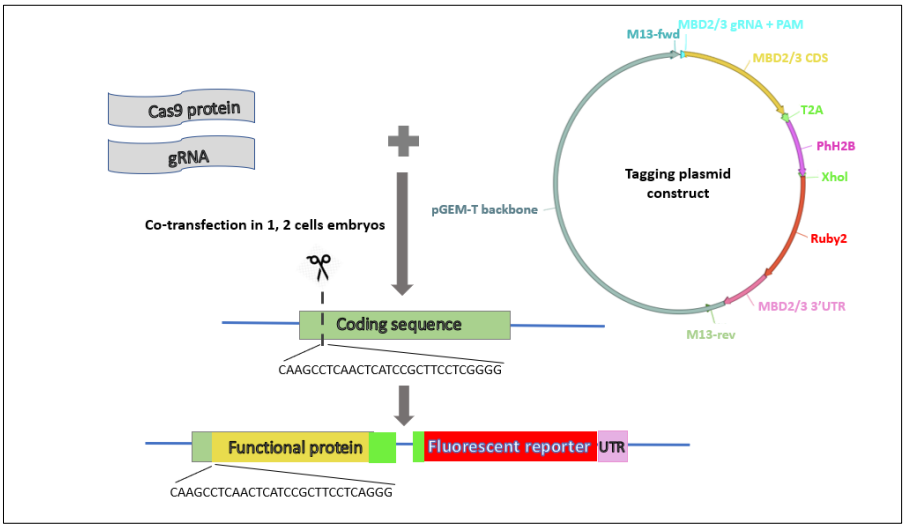
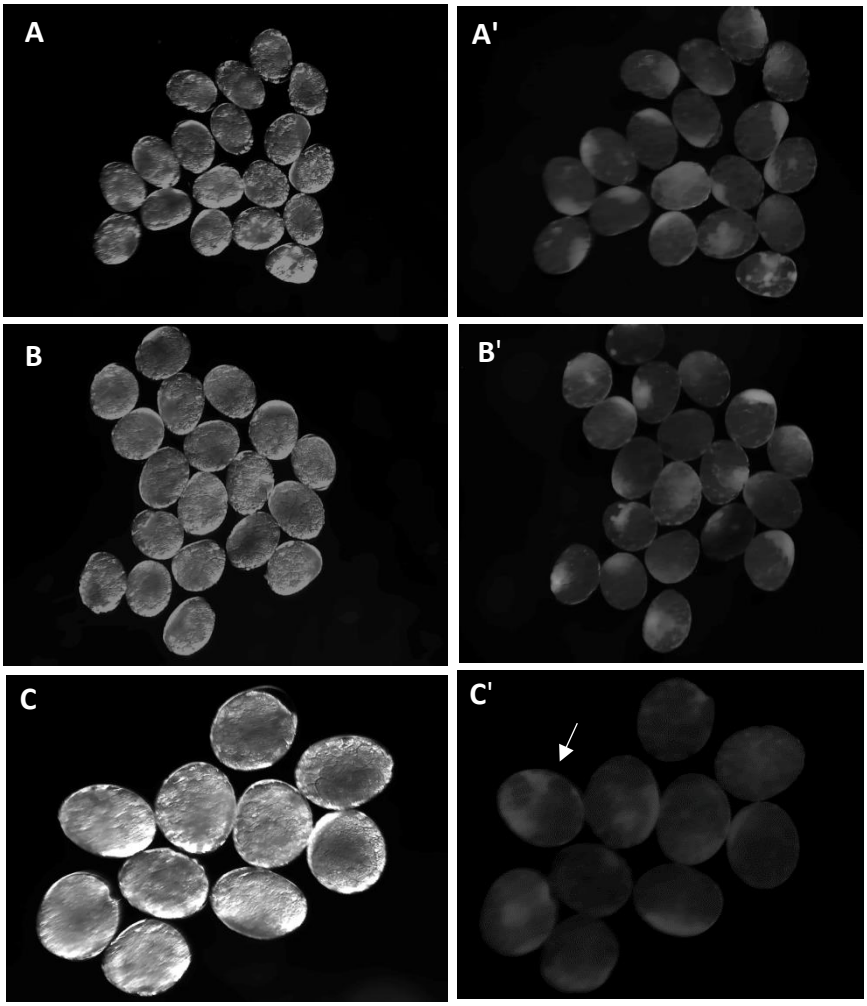


Figure 4.10. Results of CRISPR/Cas9 knock-in injection of MDB2/3 gene. Images of 48-hour embryos injected with CRISPR/Cas9 mix and tagging constructs. **(A and B)** Bright filed images for knock-in injected embryos. **(A' and B')** The same embryos in A and B imaged using the red fluorescent channel. Fluorescent signal is detected in one plate of the knock-in injection batch, likely a random signal coming from the tagging plasmid independent from MBD2/3 expression. **(C and C')** Bright field and fluorescent image of embryos injected with the tagging plasmid only. Some embryos display a weak fluorescent signal (indicated by an arrow in the figure).



4.4 Knockdown of MBD2/3 in embryos

Invertebrates, such as the platyhelminth *Schistosoma japonicum* and the sponge *Ephydatia muelleri*, encode MBD3-like proteins with a bona fide MBD, suggesting that the role of these proteins in DNA methylation or in methylation-dependent binding function is ancestral character of the protein (Cramer et al., 2017; Hendrich & Tweedie, 2003). This observation has led to the speculation that the MBD-like proteins found in species such as the nematodes *C. elegans*, *Caenorhabditis briggsae* and *Pristionchus pacificus* and the fruit fly *D. melanogaster* may represent a loss of the MBD protein domain and a diversification in its function (Gutierrez and Sommer, 2007). Studies investigating the role of MBD proteins are limited in the literature, and more work is needed to understand how it contributes to gene regulation, whether in methylation dependent or independent manner.

Knockout experiments presented earlier in this chapter showed that MBD2/3 in *Parhyale* is essential for normal embryogenesis. In this section, we performed RNAi knockdown targeting MBD2/3 in the hope of affecting maternal MBD2/3 and producing a stronger phenotype. We aimed to look at transcriptional changes in response to loss of MBD2/3.

4.4.1 Experimental design

Previous studies in *Parhyale* have utilized the stealth RNAi or antisense Morpholino techniques to perform knockdown experiments (Özhan-Kizil et al., 2009; Liubicich et al., 2009; Nestorov et al., 2013; Martin et al., 2016). In this project, we employed RNA interference (RNAi) through double-stranded (dsRNA) to effectively silence the expression of MBD2/3, wheather it is maternal or zygotic.

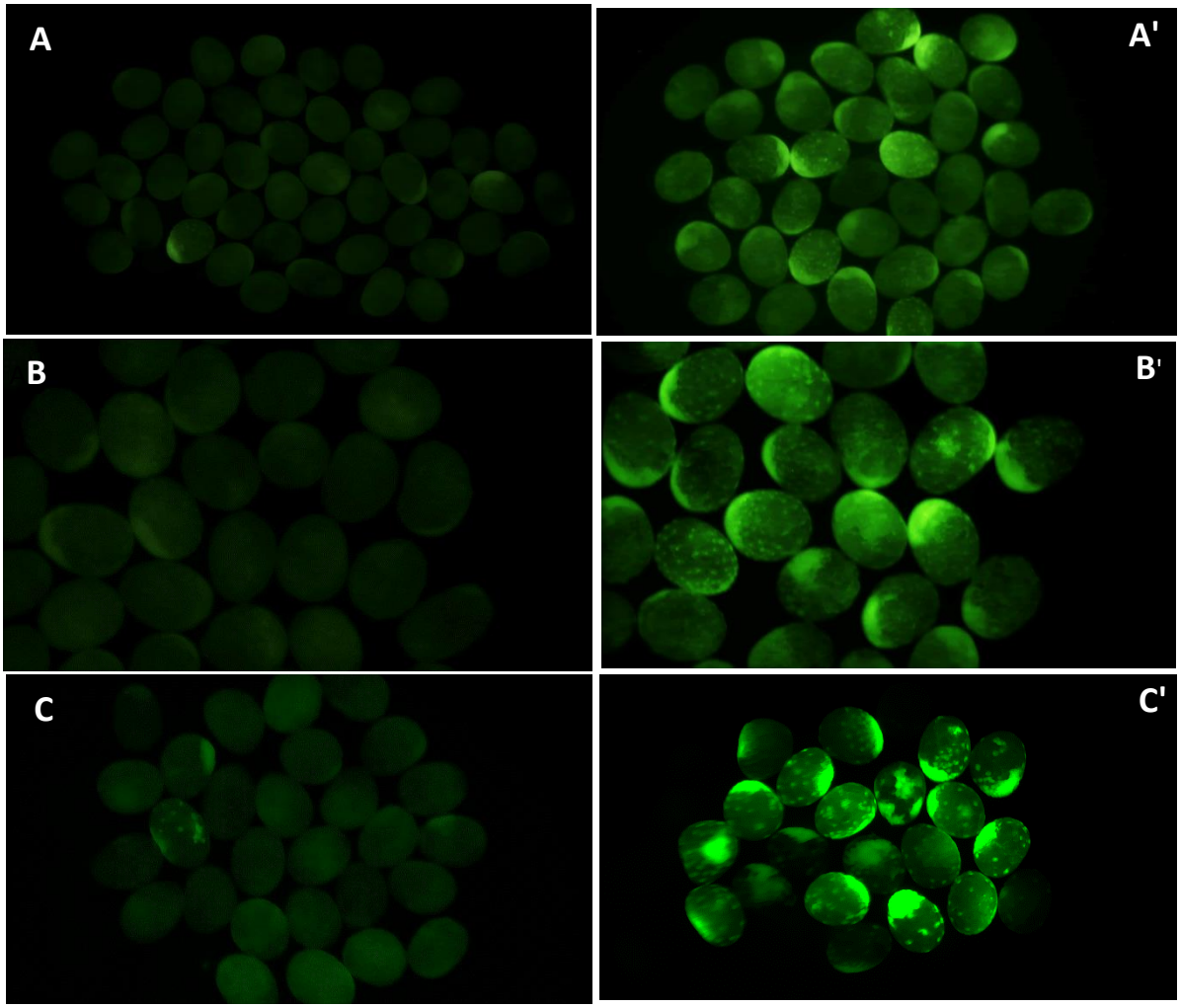
Since dsRNA has not been used before in *Parhyale*, we conducted a positive control experiment to test the efficiency of this system in *Parhyale*. Specifically, 1-cell embryos were co-injected with EGFP mRNA and dsRNA against GFP, while the control group consisted of embryos injected with EGFP mRNA but with dsRNA against MBD2/3. The embryos were monitored for fluorescent signal, and the results presented in Table 4.6 and figure 4.11 showed that gfp RNAi embryos had a very low percentage of fluorescent embryos

(0-10%) compared to the high fluorescence rate in the control group (76-100%). Following this, we proceeded to inject MBD2/3 RNAi into 1-cell embryos, using either gfp or Ruby2 dsRNA as a negative control RNAi at matching concentration.

Table 4.6. RNAi positive control experiment

Concentration of dsRNA	GFP RNAi			MBD RNAi		
	Injected embryos	Fluorescent embryos	Percentage	Injected embryos	Fluorescent embryos	Percentage
200 ng/µl	52	5	10%	34	26	76%
200 ng/µl	47	2	4%	32	25	78%
700 ng/ µl	28	2	7%	22	21	96%
700 ng/ µl	10	0	0%	10	10	100%

Figure 4.11. RNAi positive control experiment on P*arhyale* embryos. **(A-C)** Embryos injected with GFP RNAi and EGFP at 48 hours of development, with no detectable fluorescent signal in most embryos**.** **(A'-C')** Embryos injected with MBD RNAi and EGFP mRNA of the same stage (48 hours), showing a strong and evident fluorescent signal in all injected embryos.



4.4.2 MBD2/3 knockdown is lethal to embryos

Parhyale's MBD encoding gene has the conserved C-terminus domain of the methyl-CpG binding protein (pfam accession pfam14048), which is lost in some MBD-like proteins found in some invertebrates such as planarian and nematodes (Jaber-Hijazi et al., 2013; Gutierrez and Sommer, 2007).

To confirm that MBD2/3 is required for normal embryogenesis in *Parhyale*, we performed knockdown of MBD2/3 using RNAi. One-cell embryos were injected with dsRNA corresponding to MBD2/3 or Ruby (as control). Embryos were then incubated at 26 ºC to monitor development, and changes in knockdown embryos were observed starting at stage 11 (60 hours) of development. Microinjection and plate incubation can sometimes affect the speed of embryogenesis, and some embryos arrived at S11 a few hours later than 60 hours of development. A summary of phenotypic progression during embryonic development is presented in Table 4.7 and Figure 4.12, 4.13, 4.14 and 4.15 below. Similar to knockout phenotype, MBD2/3 knockdown embryos also exhibit lethality by day 4 of development. Embryos were arrested at Stage 11 of development (Figure 4.13). Imaging of embryos showed that at stage 11 of development, knockdown embryos fail to develop the dorsal organ or the midgut anlagen that can be seen as an aggregation of cells in normally developing embryos (Figure 4.14). Knockdown embryos never developed the ovoid-shaped midgut anlagen that can be seen at S14. Even if they appeared normal at S11, they started to degenerate and failed to complete embryogenesis beyond this point.

The survival rate was significantly different between the control group (mean of 78%) and the MBD2/3 RNAi knockdown group (mean of 44%).

In the following section, we collected RNA samples from MBD2/3 RNAi embryos to perform RNA-sequencing and investigate the transcriptional changes in response to MBD2/3 knockdown.

Figure 4.12. Results of MBD2/3 RNAi knockdown injection. The boxplot shows a comparison of the overall survival rate between the control group (injected with GFP or Ruby dsRNA) and the group injected with MBD2/3 dsRNA. The difference between the two groups is significant (unpaired two tailed *t*-test *p*-value < 2.2e-16).

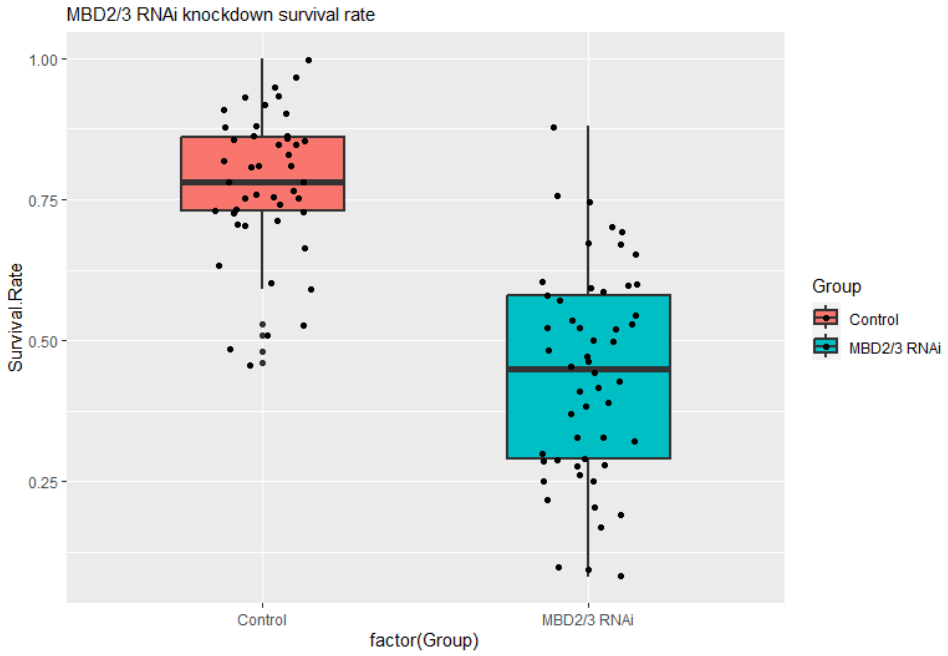


Figure 4.13. The results of injecting dsRNA at multiple dosages. The charts display the number of surviving embryos (Y-axis) during their development from day 1 (24 hours) until day 6 (x-axis), using various concentrations of dsRNA. The number of live embryos remained stable after day 6 until hatching. We found that increasing the dsRNA concentration above 300 ng/µl did not result in a larger effect, so we chose to proceed with this concentration to collect samples for RNA sequencing.

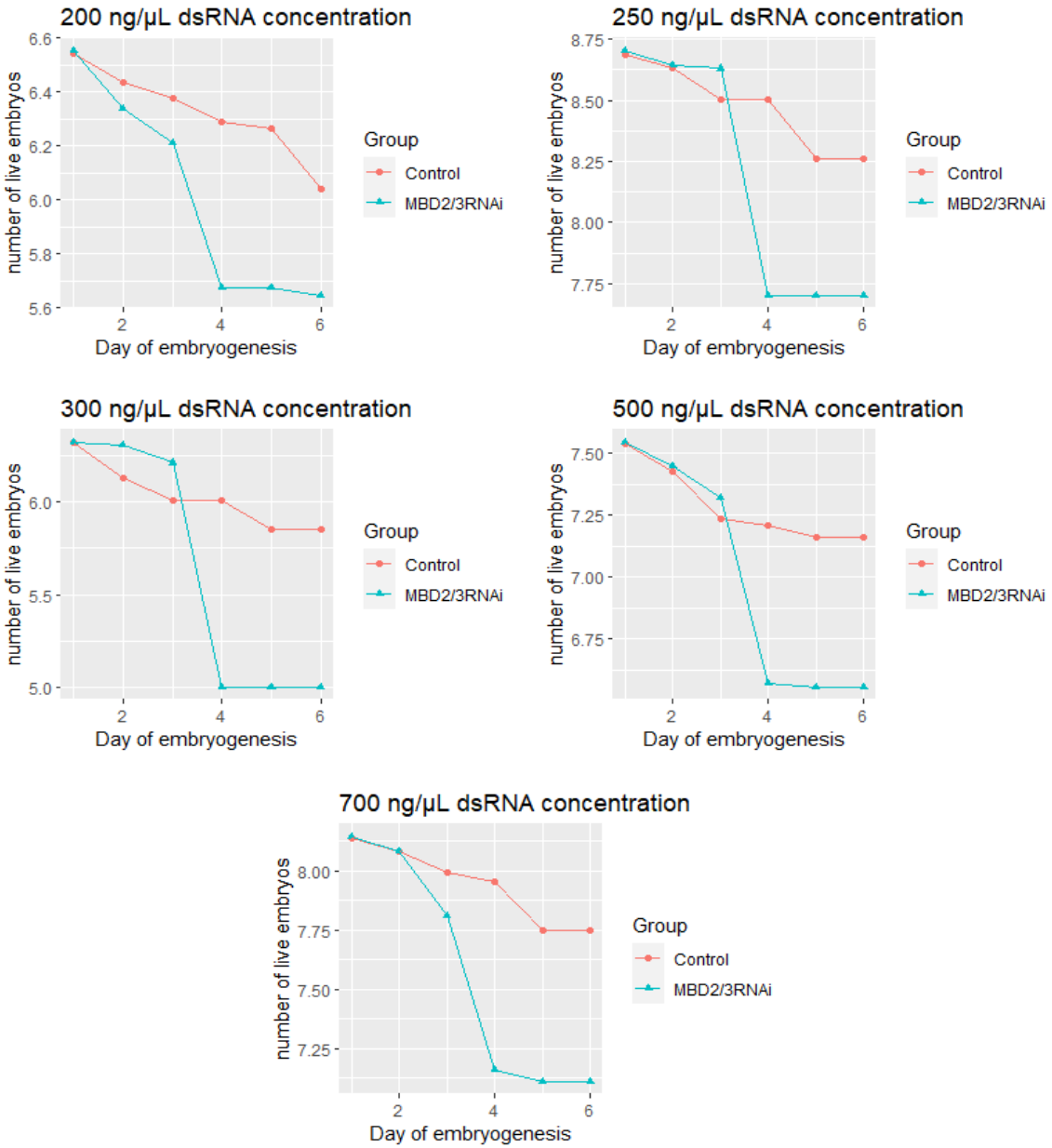


Figure 4.14 A. Ruby RNAi embryos imaged throughout their course of development. The figure includes bright-field images of normally developing embryos from day 1 (24 hours) until day 5 (120 hours) and corresponding fluorescence microscopy images of EGFP mRNA-injected embryos **(A-G and A'-G', respectively)**. Additionally, fluorescence microscopy images of Hoechst-stained embryos taken between 60 hours to 120 hours of development are shown **(C''-G'')**. These images correspond to the same embryos in the upper images. At 60 hours of development, an aggregation of cells representing the starts of midgut anlagen can be seen **(C'', white arrow)**, along with another aggregation of cells forming the dorsal organ **(C'', red arrow)**. At 77 hours (Stage 14), the ovoid shape of midgut anlagen becomes very clear **(D' and D'')**. At 96 hours (Stage 19) **(F-F'')** and 120 hours (Stage 21) **(G-G''),** embryos continue to develop normally, with the posterior part of the animal starting to widen and limb buds becoming visible.

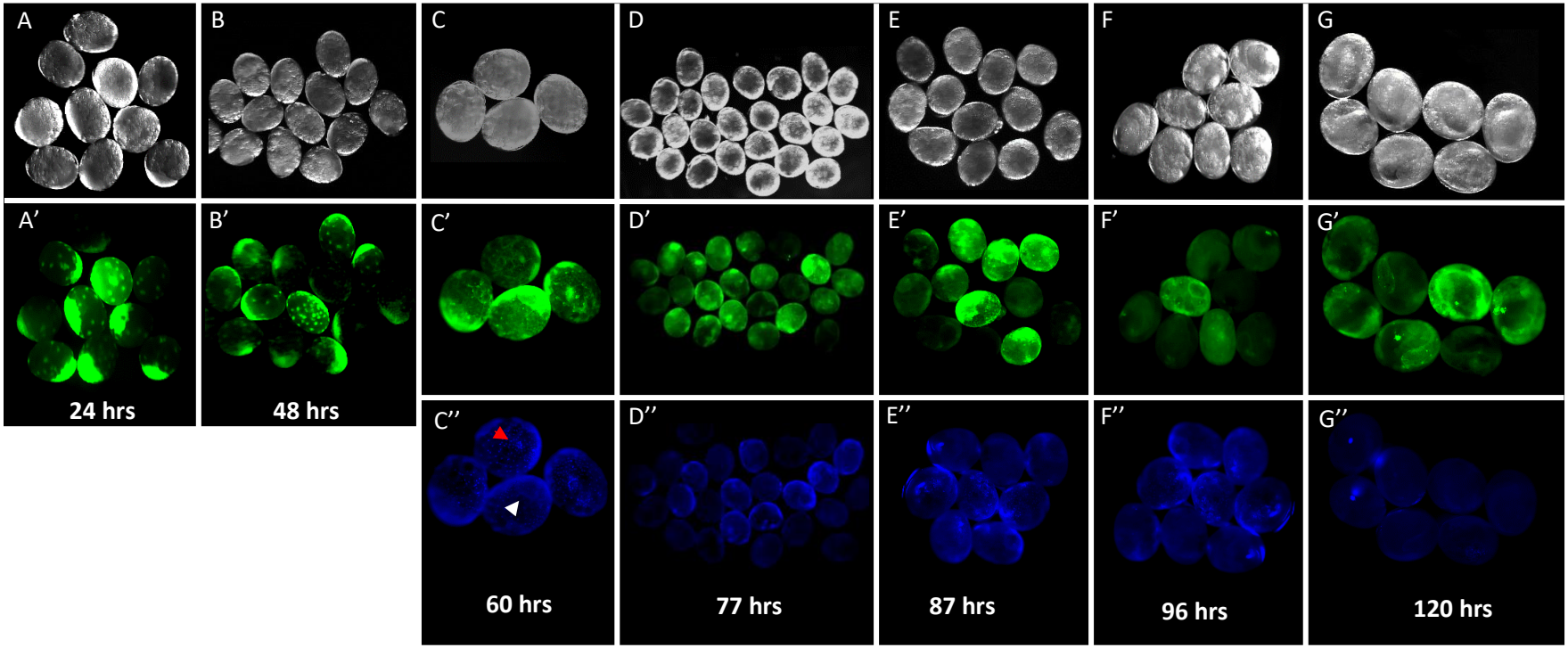


Figure 4.14 B. MBD2/3 RNAi embryos imaged throughout their course of development. The figure depicts bright-field images of MBD2/3 knockdown embryos from day 1 (24 hours) until day 5 (120 hours) and corresponding fluorescence microscopy images of EGFP mRNA-injected embryos **(A-G and A'-G', respectively)**. Additionally, fluorescence microscopy images of Hoechst-stained embryos taken between 60 hours to 120 hours of development are shown **(C''-G'')**, which correspond to the same embryos in the upper images. At 60 hours of development **(C' and C'')**, contrary to normally developing embryos in Figure 4.14 A, the formation of dorsal organ and midgut anlagen is not detected as expected at this stage of development (Stage 11). In the following hours **(E-E'' through G-G'')**, the embryos stop developing and start to degenerate until they eventually die.

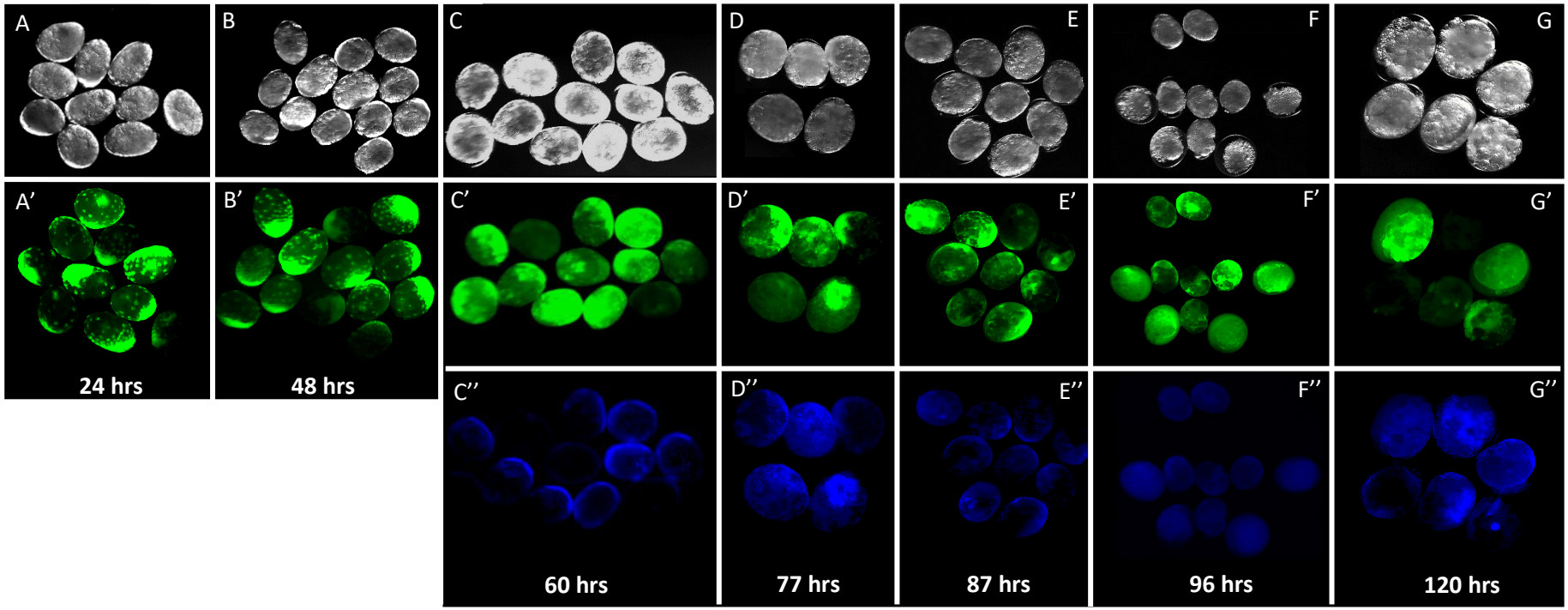
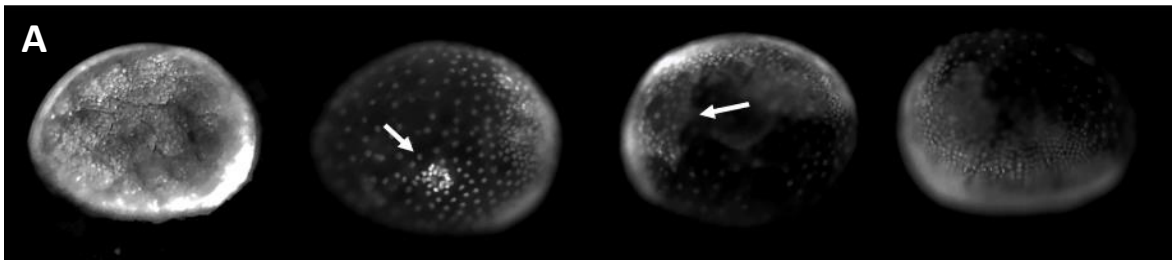


Figure 4.15. Control and MBD2/3 RNAi embryos at Stage 11 (S11) of development. Panel **(A)** shows normal embryos at S11, imaged in bright-field and Hoechst-stained. The same embryos are imaged at various orientations, second image from the left showing a dorsal view highlighting aggregated cells that form the dorsal organ (indicated by white arrow), and the third image showing the aggregation of cells to form midgut anlagen (indicated by the white arrows). Panel **(B)** shows MBD2/3 knockdown embryos at the same stage, imaged using bright-field microscopy, and fluorescence microscopy of EGFP-mRNA injection, and Hoechst staining of the same embryo. Stage 11 (60 hours) is the earliest stage at which we believe the effect of MBD2/3 RNAi can be detected. Most knockdown embryos do not reach this stage, or they get arrested at this stage of development.

.



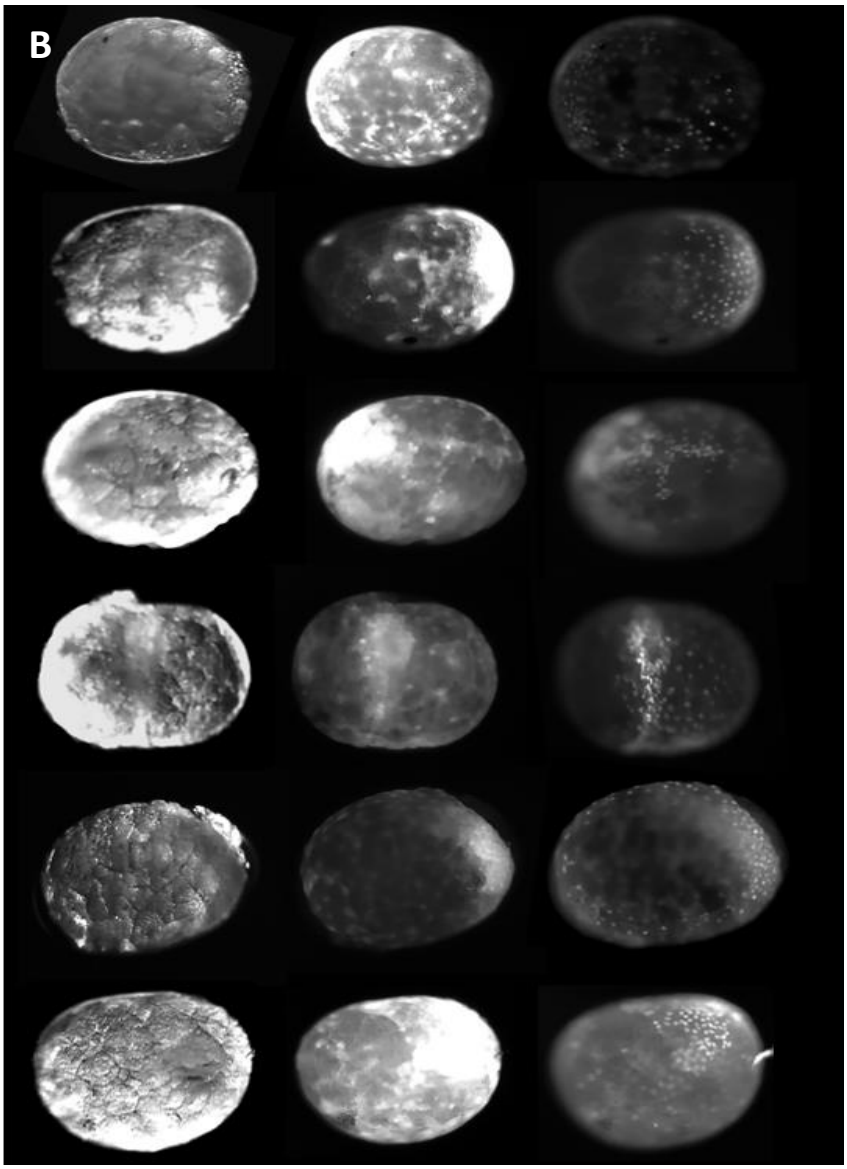


Figure 4.16. Single embryo images from control and MBD2/3 knockdown groups. The progression of development is shown from 24 h (as indicated in the figure) until hatching of normal embryos (on the next page). Up to 48 hours of development, both groups appear normal. However, at stage 11 (60 hours), MBD2/3 knockdown embryos can be seen falling behind in development with fewer cells compared to control embryos. As time progresses, embryonic lethality becomes more visible in the knockdown group. Bright-field and fluorescence-microscopy images were taken for each embryo. Embryos were injected with EGFP-mRNA (green) and Dextran (red), and then they were stained with Hoechst (blue).

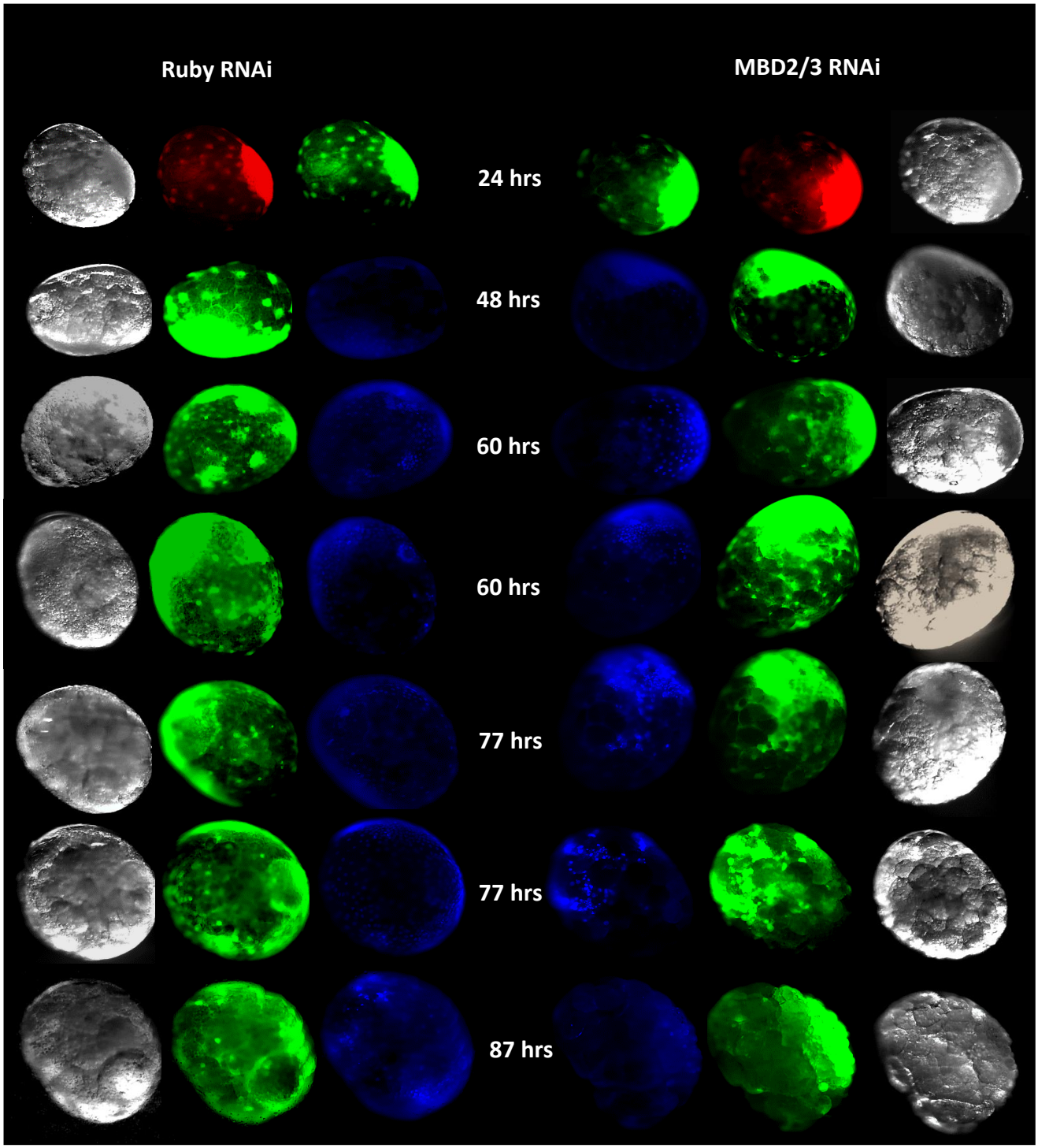


Figure 4.16 continued. Single embryo images from control and MBD2/3 knockdown groups. In this panel, the embryos in the control group continue to develop normally until hatching, while the knockdown embryos are dead and beginning to decompose.

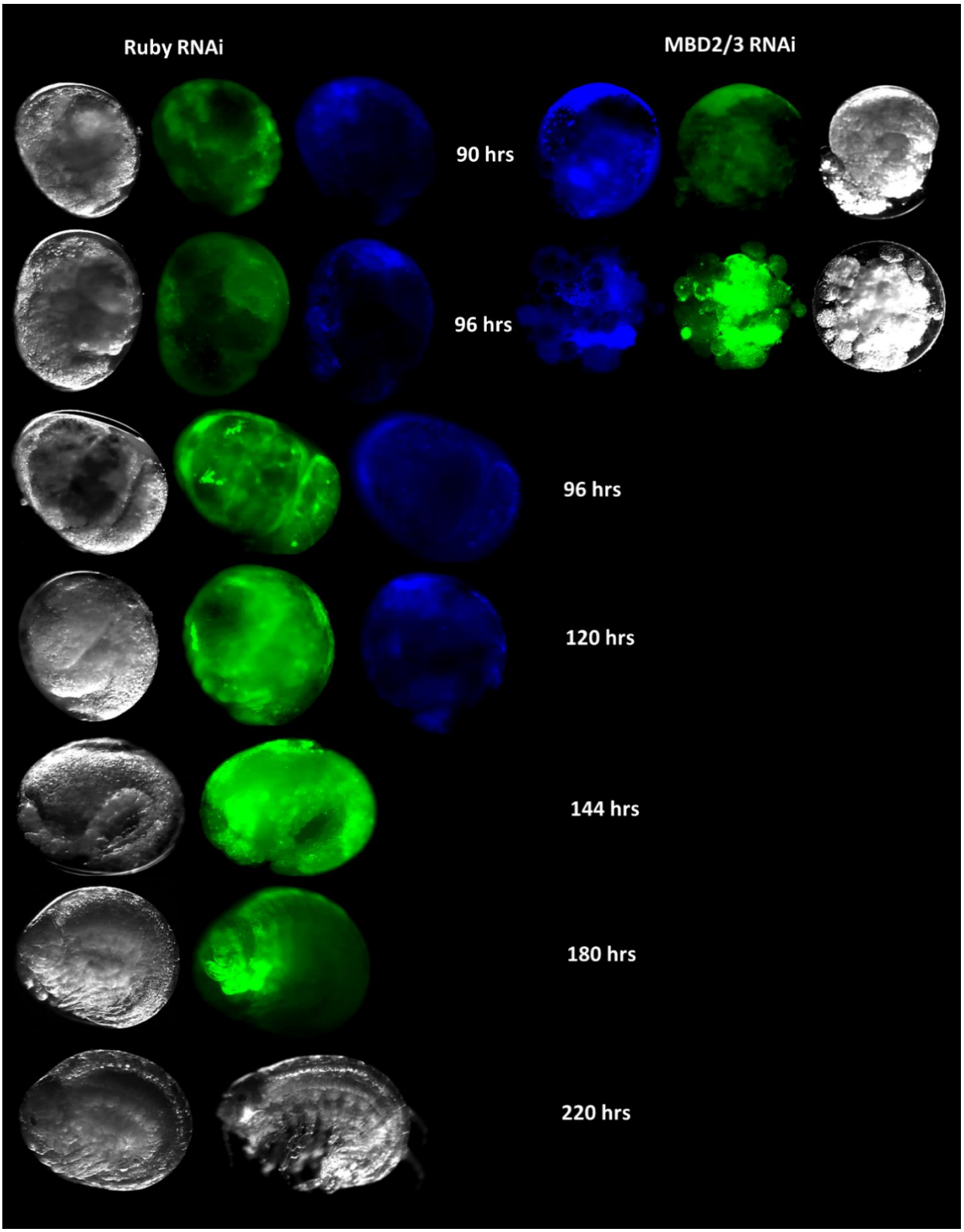


Table 4.7. MBD2/3 RNAi injection results using different concentrations of dsRNA.

MBD dsRNA concentration	24 h	48 h	72 h	96 h	d5	d6	Survival percentage
200ng/µl	23	18	16	9	9	9	39%
200ng/µl	25	19	18	13	13	13	52%
200ng/µl	17	16	15	13	13	13	76%
200 ng/µl	14	13	12	8	8	8	57%
200 ng/µl	15	15	13	8	8	7	53%
250ng/µl	NA	47	NA	NA	4	4	9%
250ng/µl	NA	28	NA	NA	8	8	29%
250ng/µl	NA	19	NA	NA	7	7	37%
250ng/µl	39	NA	NA	NA	10	10	26%
250ng/µl	40	NA	NA	NA	10	10	25%
250ng/µl	34	NA	NA	NA	11	11	32%
250ng/µl	46	NA	NA	NA	27	27	59%
250ng/µl	44	NA	NA	NA	23	23	52%
250ng/µl	37	NA	NA	NA	15	15	41%
250ng/µl	33	NA	NA	NA	17	17	52%
250ng/µl	32	NA	NA	NA	9	9	28%
250ng/µl	42	NA	NA	NA	20	20	48%
250ng/µl	46	NA	NA	NA	30	30	65%
250ng/µl	24	NA	NA	NA	17	17	75%
300ng/µl	24	NA	22	NA	7	7	29%
300ng/µl	21	NA	18	NA	4	4	19%
300ng/µl	35	NA	34	NA	21	21	60%
300ng/µl	36	NA	NA	3	NA	3	8%
300ng/µl	37	NA	NA	20	NA	20	54%
300ng/µl	32	NA	NA	15	NA	15	47%

Table 4.7 continued. MBD2/3 RNAi injection results using different concentrations of dsRNA.

MBD dsRNA concentration	24 h	48 h	72 h	96 h	d5	d6	Survival percentage
500 ng/µl	20	20	13	6	5	5	30%
500 ng/µl	4	4	2	1	1	1	25%
500 ng/µl	10	3	NA	1	1	1	10%
500 ng/µl	18	15	NA	4	4	4	22%
500 ng/µl	32	31	NA	15	15	14	44%
500 ng/µl	30	29	NA	23	23	21	70%
500 ng/µl	NA	10	NA	NA	5	5	50%
500 ng/µl	NA	11	NA	NA	5	5	45%
500 ng/µl	NA	16	NA	NA	14	14	88%
500 ng/µl	NA	13	NA	NA	7	7	54%
500 ng/µl	NA	13	10	9	9	9	69%
500 ng/µl	NA	10	7	5	5	5	50%
700 ng/µl	46	45	37	15	15	15	33%
700 ng/µl	40	35	28	11	11	11	28%
700 ng/µl	14	13	12	4	4	4	29%
700 ng/µl	15	11	NA	9	9	9	60%
700 ng/µl	22	20	NA	13	13	13	59%
700 ng/µl	49	49	NA	38	38	33	67%
700 ng/µl	18	NA	5	NA	3	3	17%
700 ng/µl	12	NA	5	NA	5	5	42%
700 ng/µl	26	NA	19	NA	12	12	46%
700 ng/µl	13	12	10	7	5	5	38%
700 ng/µl	15	15	12	9	9	9	60%
700 ng/µl	12	11	8	8	8	8	67%
700 ng/µl	14	14	8	7	6	6	43%
						AVERAGE	44%

Table 4.7 continued. Negative control RNAi injection results using different concentrations of dsRNA.

GFP or Ruby dsRNA concentration	24 h	48 h	72 h	96 h	d5	d6	Survival percentage
200 ng/µl	NA	23	20	18	18	18	78%
200 ng/µl	NA	16	15	15	14	14	88%
200 ng/µl	NA	20	18	18	18	18	90%
200 ng/µl	NA	13	13	13	13	13	100%
200 ng/µl	37	30	30	28	27	27	73%
200 ng/µl	25	23	23	22	22	22	88%
200 ng/µl	19	17	17	14	14	14	74%
250ng/µl	45	NA	43	NA	35	35	78%
250ng/µl	32	NA	30	NA	26	26	81%
250ng/µl	21	NA	15	NA	15	15	71%
250ng/µl	27	NA	22	NA	17	17	63%
300ng/µl	21	NA	21	NA	16	12	76%
300ng/µl	8	NA	8	NA	6	4	75%
300ng/µl	40	NA	NA	19	NA	6	48%
300ng/µl	30	NA	NA	22	NA	16	73%
300ng/µl	28	NA	NA	13	NA	8	46%
300ng/µl	47	NA	NA	39	NA	37	83%
300ng/µl	32	NA	NA	19	NA	19	59%
500 ng/µl	25	21	18	15	13	13	60%
500 ng/µl	33	30	30	28	26	26	85%
500 ng/µl	20	19	NA	16	14	14	70%
500 ng/µl	35	30	NA	30	30	30	86%
500 ng/µl	48	48	NA	37	35	35	73%
500 ng/µl	20	17	NA	NA	15	15	75%
500 ng/µl	13	13	NA	NA	12	12	92%
500 ng/µl	26	23	NA	NA	21	21	81%
500 ng/µl	7	7	NA	NA	6	6	86%
500 ng/µl	17	14	NA	NA	12	12	71%
500 ng/µl	NA	14	12	12	12	12	86%
500 ng/µl	NA	20	20	17	17	17	85%

Table 4.7 continued. Negative control RNAi injection results using different concentrations of dsRNA.

GFP or Ruby dsRNA concentration	24 h	48 h	72 h	96 h	d5	d6	Survival percentage
700 ng/µl	35	32	27	NA	18	18	51%
700 ng/µl	30	30	26	NA	16	16	53%
700 ng/µl	22	21	21	NA	17	17	77%
700 ng/µl	28	28	NA	26	26	26	93%
700 ng/µl	29	29	NA	28	28	28	97%
700 ng/µl	32	28	NA	25	21	21	66%
700 ng/µl	14	NA	14	13	13	13	93%
700 ng/µl	20	NA	19	19	19	19	95%
700 ng/µl	21	21	2	20	17	17	81%
700 ng/µl	11	11	11	9	9	9	82%
700 ng/µl	13	13	13	11	11	11	85%
700 ng/µl	8	7	7	7	6	6	75%
700 ng/µl	22	20	20	16	16	16	73%
						Average	78%

4.4.3 Transcriptional response of embryogenesis to loss of MBD2/3

To identify transcriptional changes caused by the silencing of *Parhyale's* MBD2/3, we performed transcriptional profiling of embryos after knockdown of MBD2/3 at two developmental timepoints. We chose to collect embryos at 77 hours post fertilization (hpf). Normally, embryos would be mostly at S14, but due to the fact that injection slows down embryogenesis, embryos at this time point would be a mixture of stages ranging from S11 to S14, while knockdown embryos would be arrested before S11. The other time point was 48 hpf. We chose this time point because, although no visible defects can be detected at this point, transcription of genes regulated by MBD2/3 might have been altered by this point. Given that this gene is highly expressed maternally during early stages, it is possible to find changes at molecular level.

We noticed that some embryos were able to escape knockdown and develop normally, most likely due to the variations in the size of injected droplet, which was difficult to control during the injection process. Therefore, for the 77-hour timepoint, we followed a collection procedure to ensure we only sequenced affected embryos (explained in Figure 4.17 below). We also collected some of the normally developing MBD2/3 RNAi embryos to compare their similarities and differences with the dead RNAi embryos. At the 48-hour timepoint, we collected all the injected embryos, as the phenotype was not distinguishable at this point (detailed images of collected embryos in Figure 4.18).

In this study, we generated a total of 14 polyA RNA libraries, including three replicates for each group from control and MBD2/3 RNAi embryos at 48 and 77 hours of development. Additionally, we generated two replicates for MBD2/3 RNAi embryos that escaped knockdown and developed normally at 77 hours. We used HISAT2 (Kim et al., 2015) to generate read mapping, and the mean overall mapping rate for all libraries is 83.7%, with a range of 81.81% to 86.66%.

Figure 4.19 A displays a PCA plot that includes all 14 sequenced datasets. At both timepoints, the knockdown libraries are not clustering as tight as the control libraries. This may be due to variation in the amount of injected dsRNA in the embryos or in RNAi level. Nevertheless, the knockdown libraries are still

distinct from the control group. At 77 hours, the live knockdown libraries, which represent normally developing embryos, cluster more closely with the control group. In figure 4.19 B, we clustered the dataset from the 48 hours timepoint separately to determine whether there is a difference between them, as they cluster more tightly when all datasets are included. As shown, we found that the control and the knockdown group are clustering separately, although the three replicates are not closely clustered in either condition. Overall, the PCA plot suggests that even if there is no visible phenotype at 48 hours of development, there may be a molecular-level effect of MBD2/3 RNAi.

To assess the quality of the libraries generated in this study, we plotted count distribution and library sizes (Figure 4.20). All libraries exhibit similar sizes and count distributions. However, the average count values for all the libraries, particularly for embryos at these stages, were relatively low. This could be due to the use of dsRNA, as this is the first time RNAi knockdown has been performed in *Parhyale* in this project. Furthermore, previous studies used stealth RNAi or Morpholino techniques for knockdown in *Parhyale* did not collect RNA from injected embryos. Therefore, more RNA-seq data from knockdown embryos are needed to confirm whether use of dsRNA or any knockdown technique affects the embryo's transcriptome. These low counts resulted in a reduced number of genes that could be analyzed for differential expression, leading to fewer differentially expressed genes. (Table 4.8). Nonetheless, our analysis identified differentially expressed genes between the control and the knockdown groups.

It is worth noting that negative control injected embryos developed normally and hatched, indicating that any effect on the transcriptome from the injection or dsRNA was temporary. However, MBD RNAi embryos died due to the specificity of dsRNA against MBD2/3, which is the only difference between the control and knockdown groups. Overall, our findings suggest that the use of dsRNA for knockdown in *Parhyale* requires further investigation to determine its effects on the embryo's transcriptome.

We used FeatureCount (Liao et al., 2014) to count reads aligned to the genome assembly we improved and annotated in chapter 2. For differential gene expression analysis, we used the DESEq2 package (Love et al., 2014). Analysis of gene expression datasets at 48 hours after knockdown revealed only 24 differentially

expressed genes (DEG). Of these, 5 genes were downregulated (adjusted *p*-value <0.05 and Fold change 1) and 19 genes were upregulated (Figure 4.21 A). While analysis of the gene expression datasets at 77 hours after knockdown revealed 158 DEG, with 73 genes downregulated, and 85 genes upregulated (Figure 4.21 B).

The *Parhyale* genome transcriptome contains one transcript that codes for MBD2/3 gene. Our analysis showed that MBD2/3 was successfully knocked-down at both 48h and 77h timepoints, with a greater than 5 and 4 log-fold change, respectively (Figure 4. 22). This demonstrates the effectiveness of the knockdown. In contrast, MBD2/3 expression was not significantly downregulated in embryos that escaped knockdown, and no DEG were found when comparing this group with the control group. This explains why the embryos that escaped knockdown were able to continue developing normally.

At the 48h RNAi timepoint, among the upregulated genes are three Zinc-Finger transcription factors (TFs), including *Blimp-1*, a C2H2 type domain containing TF, and a PHD Zinc-Finger domain TF. All the upregulated genes have low gene-body methylation (gbM) levels (0.1% - 9.5%), with few of them having promoter methylation. The C2H2 TF has an intermediate level of promoter methylation (26.9%). *Blimp-1* a tumor suppressor, is one of the master regulators during development (Wang et al., 2019). Two more genes with gene ontology (GO) terms related to transcription were also upregulated, one of which has an ortholog to *Spt-5*, a component of RNA polymerase II (RNAP II) elongation complex and one of the few upregulated genes with promoter methylation (4.1% promoter methylation level) (Crickard et al., 2017). Two of the upregulated genes are involved in ribosome binding and translation, *RRBP1* and *RBD* domain-containing protein. We also found genes with GO terms related to proteolysis, a process involved in protein degradation that may be necessary to produce an active protein (Ciechanover 2005). Six of the upregulated genes did not have any orthologs in other organisms, which suggests that they could be specific to *Parhyale*. However, one of these six gene had an ortholog in the Amphipod *Hyalella Azteca*, indicating that it could be an Amphipod-specific gene. The remaining four upregulated genes were associated with GO terms related to biological processes, such as chitin binding and protein localization.

At the 48-hour RNAi timepoint, there were only a few downregulated genes. Apart from MBD2/3, there were only four other downregulated genes, with low levels of gbM ranging from 0.61% to 4.2%, and no promoter methylation in any of them. Of these four genes, two had an orthology to *Cp1* gene, which is involved in protein degradation in lysosome or proteolysis (Tryselius and Hultmark 1997).

At the 77-hour RNAi timepoint, 15 transcription factors (TFs) were upregulated. Seven of these TFs were uncharacterized ring-type domain Zinc-Finger (ZF) TFs that only had an orthologs in the Amphipod *Hyalella azteca*. Another upregulated TF that appeared to be Amphipod-specific had a bZIP (basic-region leucine zipper) domain. Another upregulated TF with a bZIP domain was b-ATF, which is known to be involved in heterochromatin assembly and regulation of transcriptional activation (Hai et al., 1989). The rest of the upregulated TF include genes with orthology to *Erg*, *RFX*, *Pax6*, *Prg* and *tll* TFs.

In addition, 38 upregulated genes did not have orthologs in any other organism, indicating that they could be specific to *Parhyale*, except for 8 genes that had orthologs in *H. azteca* only, suggesting that they could be Amphipod-specific genes. Furthermore, 12 other upregulated genes are enriched with GO terms like protein catabolic process, ATP and metal ion binding, ribosome biogenesis, translation and RNA binding, positive regulation of transcription by RNAP II, and transcription coactivation activity (see gene IDs and related GO terms in Table 4.9). All the upregulated genes had low levels of gbM (0.03 – 16.9%), and only 16 upregulated genes had low levels of promoter methylation as well (0.03 – 7%). Only five upregulated genes have no methylated cytosine at all.

At the 77h RNAi timepoint, the knockdown of MBD2/3 resulted in the downregulation of six transcription factors, including two C2H2-type domain Zinc-Finger TF, one is an ortholog to *Sp1* TF, which regulates transcription by RNAP II (Kaczynski et al., 2003). The third TF is a Zinc-Finger domain containing TF that has an ortholog only in *H. azteca*. Another downregulated TF is a homeodomain-containing TF with orthology to *NKX2-2* TF,which plays a role in neuronal development and neuroendocrine differentiation in vertebrates (Briscoe et al., 1999). Additionally, three downregulated genes have GO terms related to nervous system, namely *ACRC*, which is involved in neuropeptide hormone activity, *NTM*, which mediates

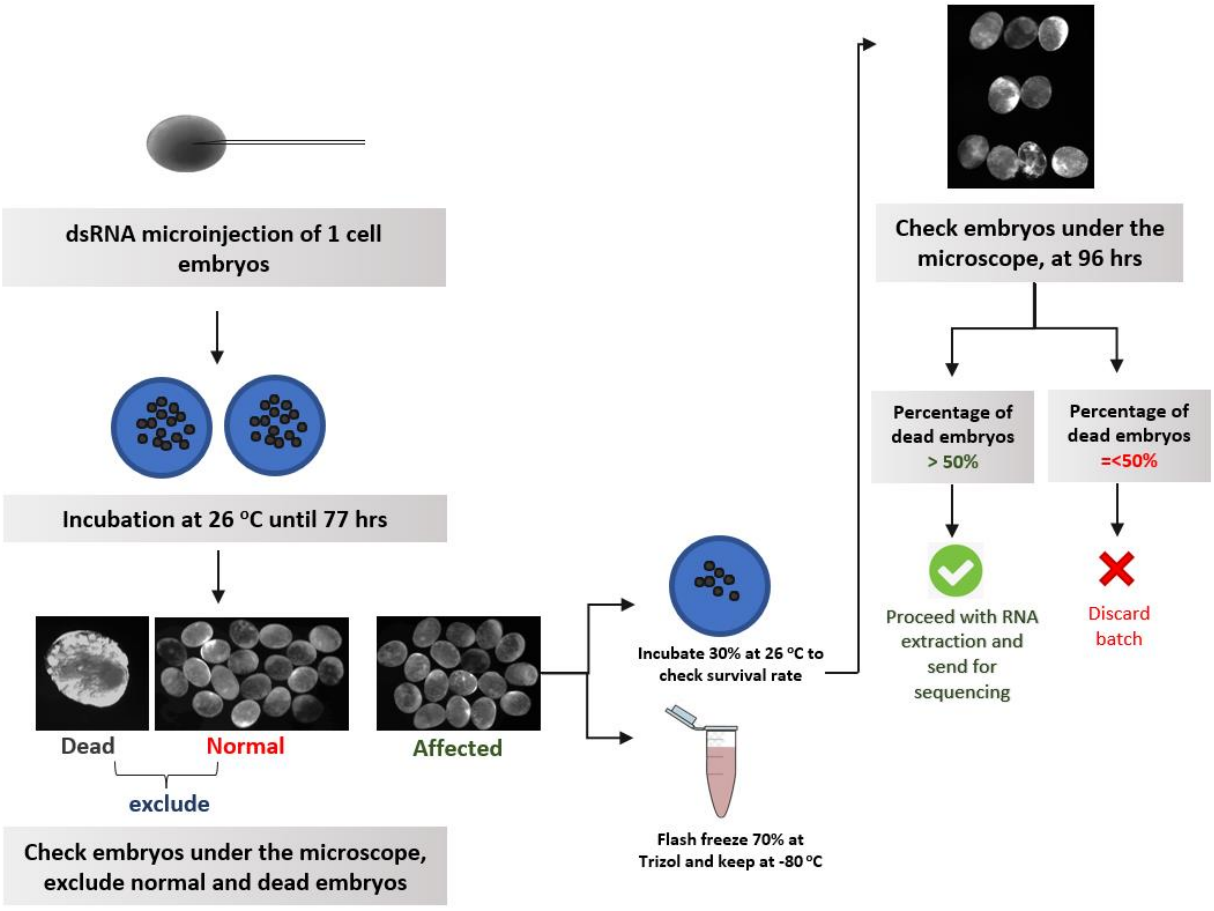
neurite outgrowth, and a SEA-domain containing protein (Wolf et al., 2017; Gil et al., 1998). One of the downregulated TF is *BTF3*, initiates transcription by promoter element binding like TATA box and CAAT box region (Cavallini et al., 1988; Kanno et al., 1992).

Several other downregulated genes have GO terms related to protein metabolic process, nucleic acid binding and RNA binding protein (detail of genes id and related GO terms in table 4.9). Thirteen of the downregulated genes are enriched with proteolysis and protein catabolic process GO terms, like *Ctsc, Ctsb, rdx, CTRB1, MME, ACE, trypsin-1* and *PRSS1*.

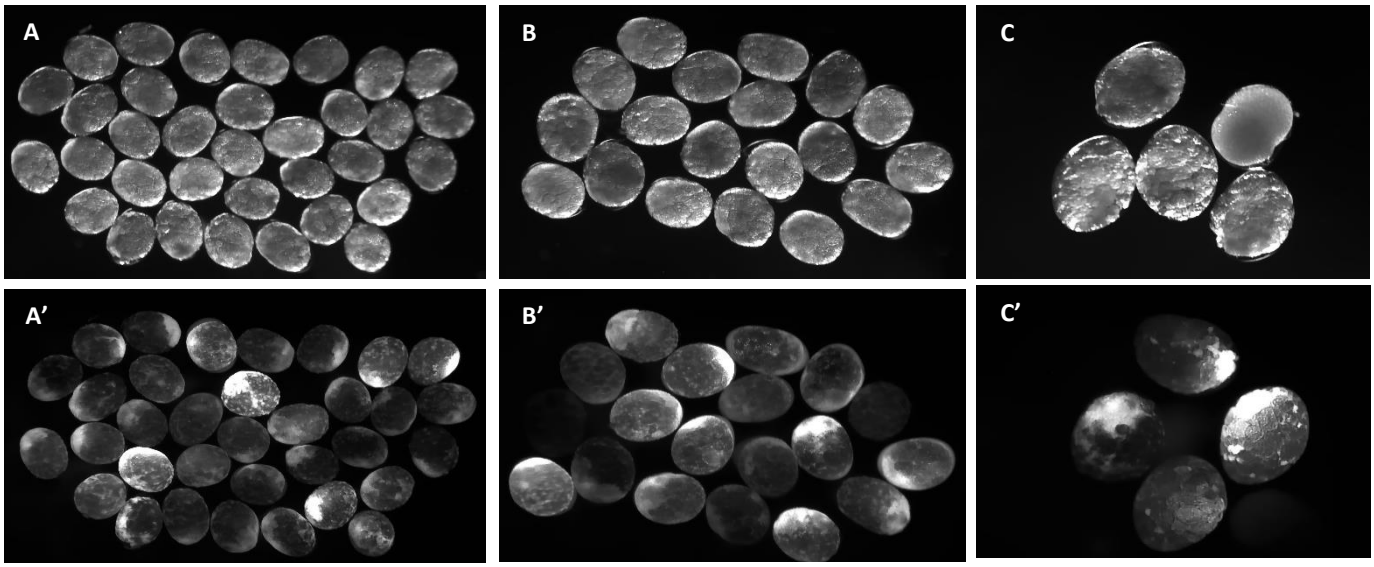
Again, genes with no orthologs in any organisms are also among the downregulated genes. suggesting that knockdown of MBD2/3 is affecting the expression of potentially unique genes specific to *Parhyale*. Additionally, all downregulated genes also exhibit low levels of gbM (0.1 – 11.3%), with only two genes lacking any methylation at all. A few genes display promoter methylation in addition to gbM.

Overall, the knockdown experiment revealed that MBD2/3 is essential for normal embryogenesis. Losing MBD2/3 altered the expression of many genes, including transcription factors. Our data indicates that MBD2/3 contributes to the regulation of transcription. The presence of gene-body methylation in most of the differentially expressed genes suggests a methylation-dependent function of MBD2/3. However, further investigations are required to understand the mechanism underlying the function of MBD2/3 and how it associates with the NuRD complex activity.

Figure 4.17. Illustration of the RNA collection procedure used for MBD2/3 RNAi knockdown embryos at 77 hours timepoint. This procedure was designed to prevent the collection of RNA from dead samples or from embryos that escaped RNAi and developed normally, ensuring that only successfully knocked down embryos were collected.



<u>Figure 4.18 A. RNAi injections for RNA-seq data collection</u>. Images show one batch of MBD2/3 RNAi at 77 hours collection point. The top panel displays a bright-field image, and the bottom panel displays a fluorescent image with red Dextran injection. **(A and A')** Embryos arrested at S11 of development, 70% of which were frozen in Trizol, and 30% kept in an incubator for observation. **(B and B')** Embryos that escaped knockdown and developed normally, those were excluded from the collection, and some were collected in Trizol for separate RNA extraction. Dead embryos were discarded, as shown in **(C and C')**. (Figure 4.18 B on the next page shows the progression of each group).



<u>Figure 4.18 B. RNAi injections for RNA-seq data collection</u>. Images of the same batch as in Figure 4.18A of MBD2/3 RNAi at 96 hours. The top panel displays a bright-field image, and the bottom displays a fluorescent image with red dextran injection. **(A and A')** Embryos were kept for observation, and only 33% of the embryos on the plate developed normally, while the rest were dead. **(B and B')** Embryos that were excluded from the study continued to develop normally. **(C and C')** same embryos at A and A' at day 5, mostly dead embryos. **(D and D')** same embryos at B and B' at day 5, mostly normally developing embryos.

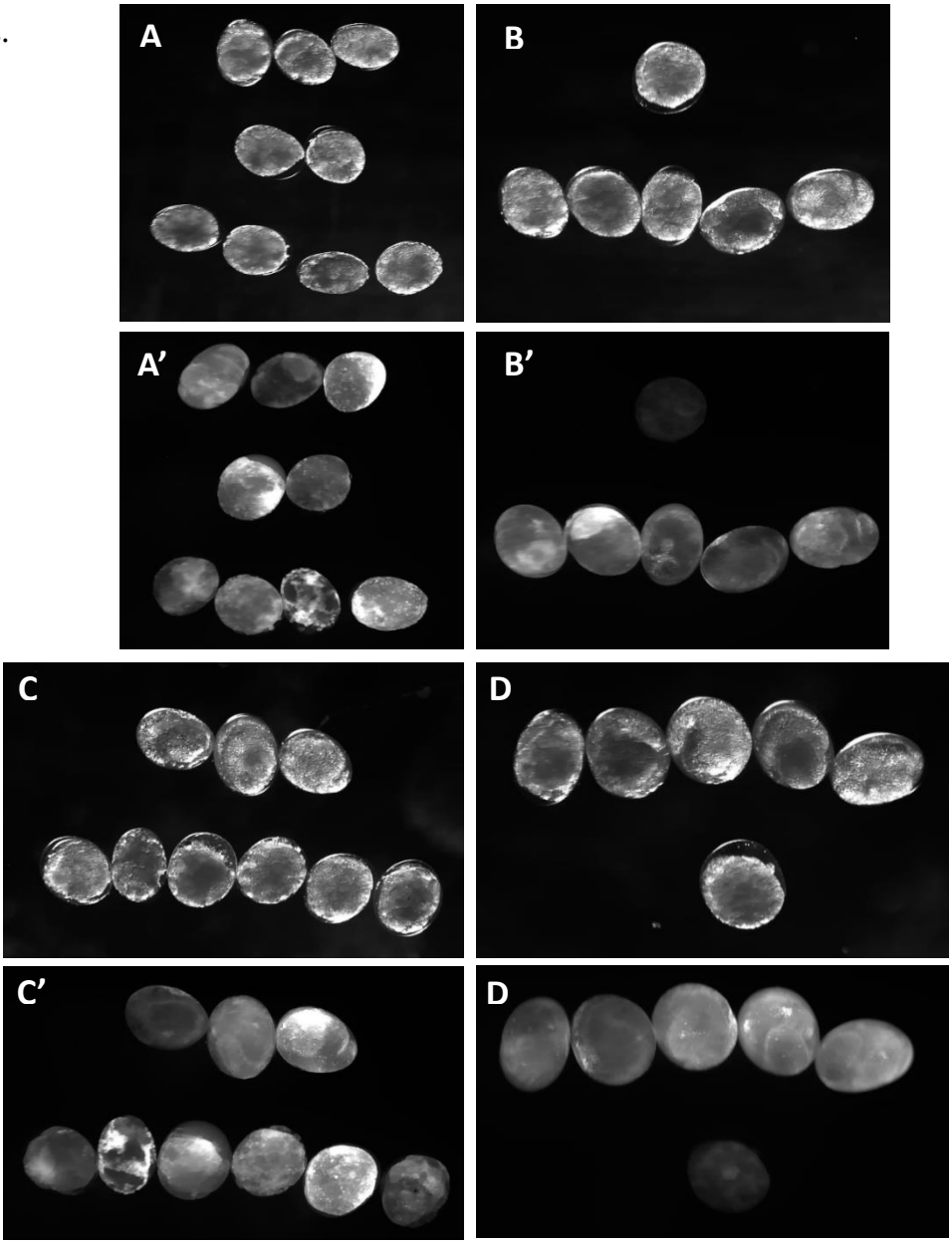
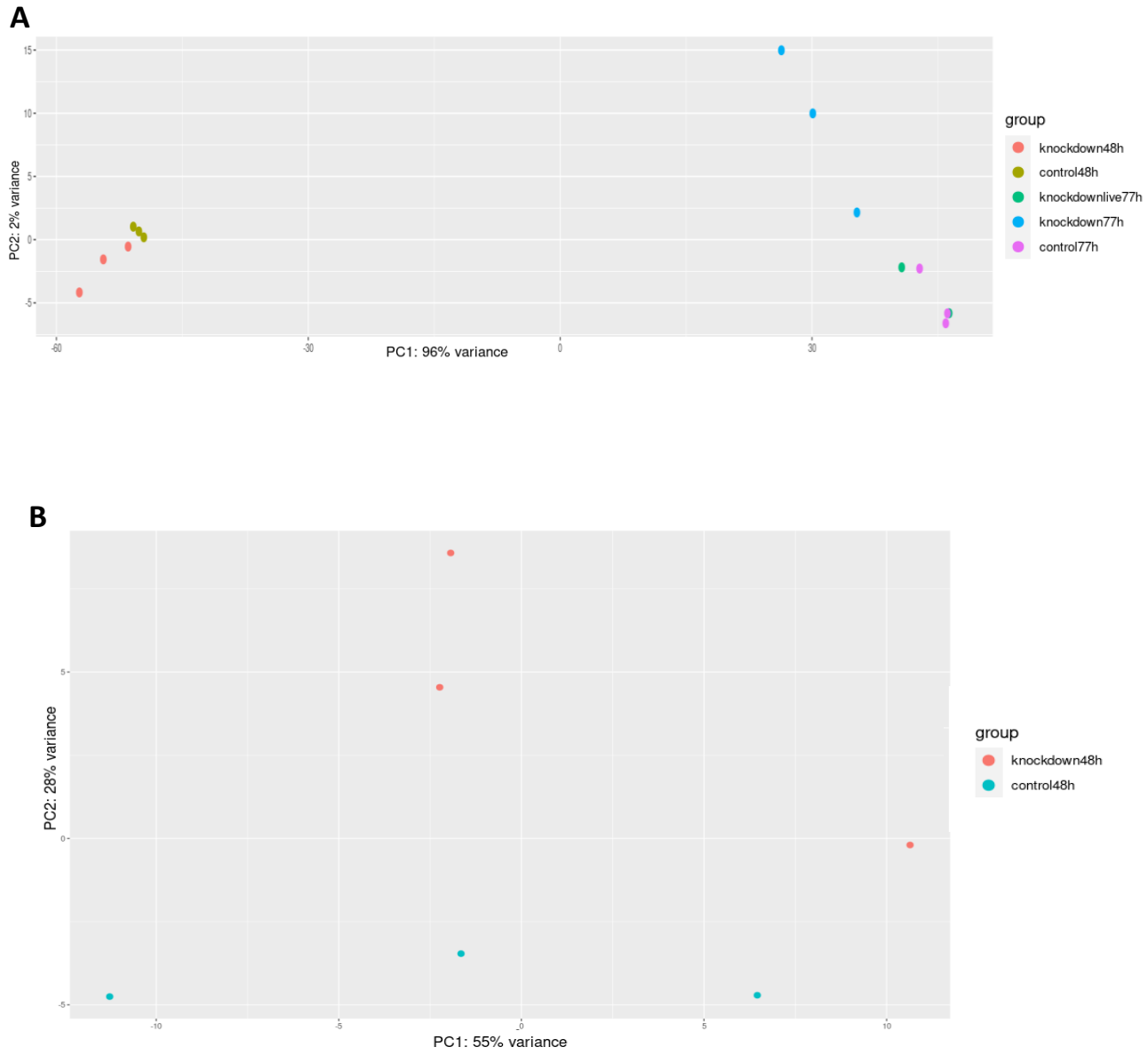


Figure 4.19. PCA plot showing clustering of different replicates of RNA-seq libraries. **(A)** Clustering showing all sequenced timepoints. **(B)** PCA showing clustering of RNA-seq libraries at the 48-hour time point only.



<u>Figure 4.20. Quality assessment of MBD2/3 RNAi RNA-seq data</u> (A) Bar plot showing the library sizes for each sample. All libraries exhibit similar sizes. **(B)** Boxplot of count distributions for each sample, indicating the median (horizontal line inside the box) and similar count distribution across all libraries with low median value. The results suggest high consistency between all RNA-seq libraries.

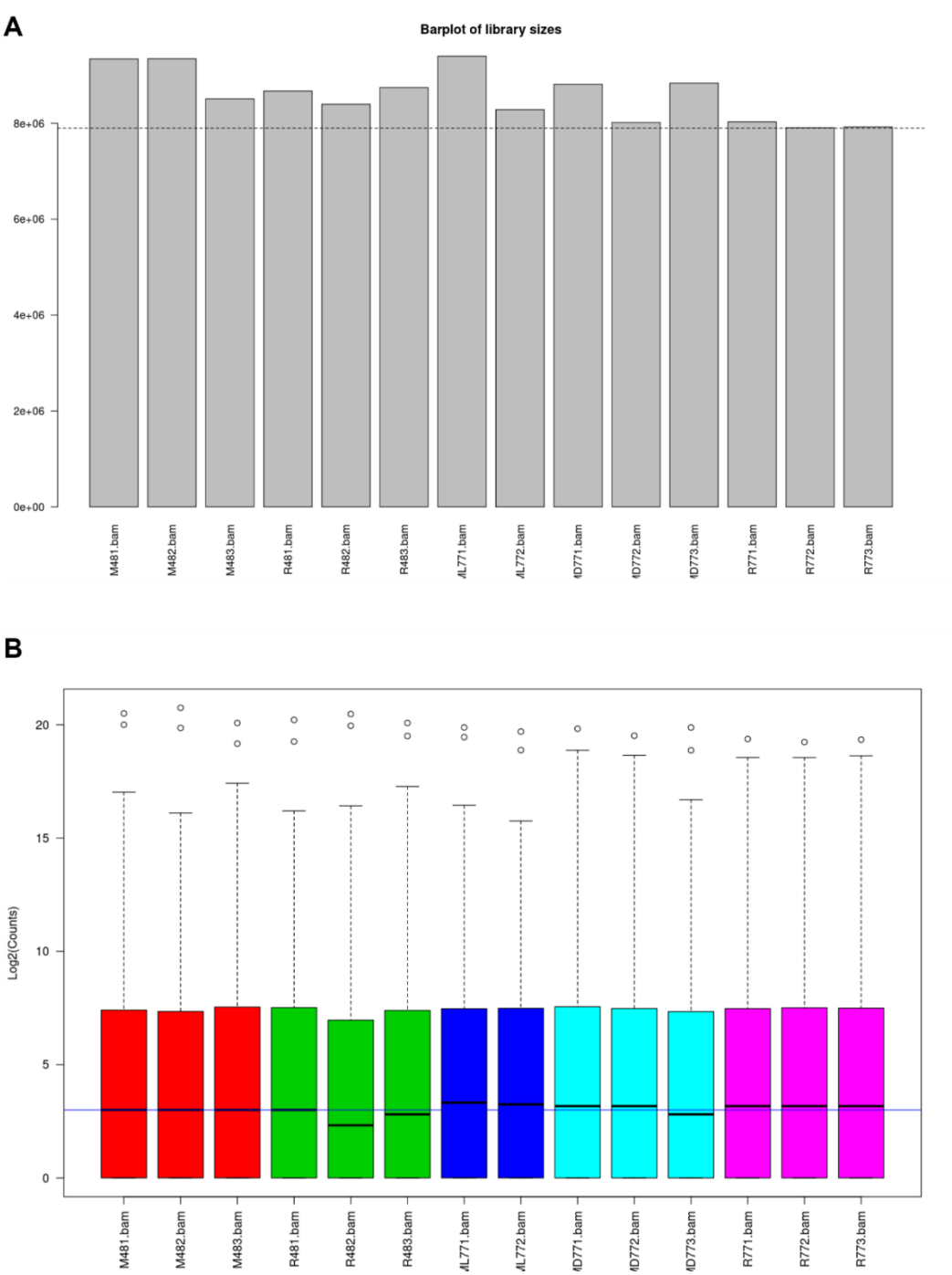


Figure 4.21. Volcano plot showing differentially expressed transcripts at 48-hour **(A)** and 77-hour **(B)** after MBD2/3 RNAi treatment. Differential expression analysis was performed using DESEq2 package with an adjusted *p*-value <0.05 and a fold change threshold of 1. The green dots indicate the upregulated genes, the black dots indicate the downregulated genes, and the grey dots indicate the genes with no significant differential expression.

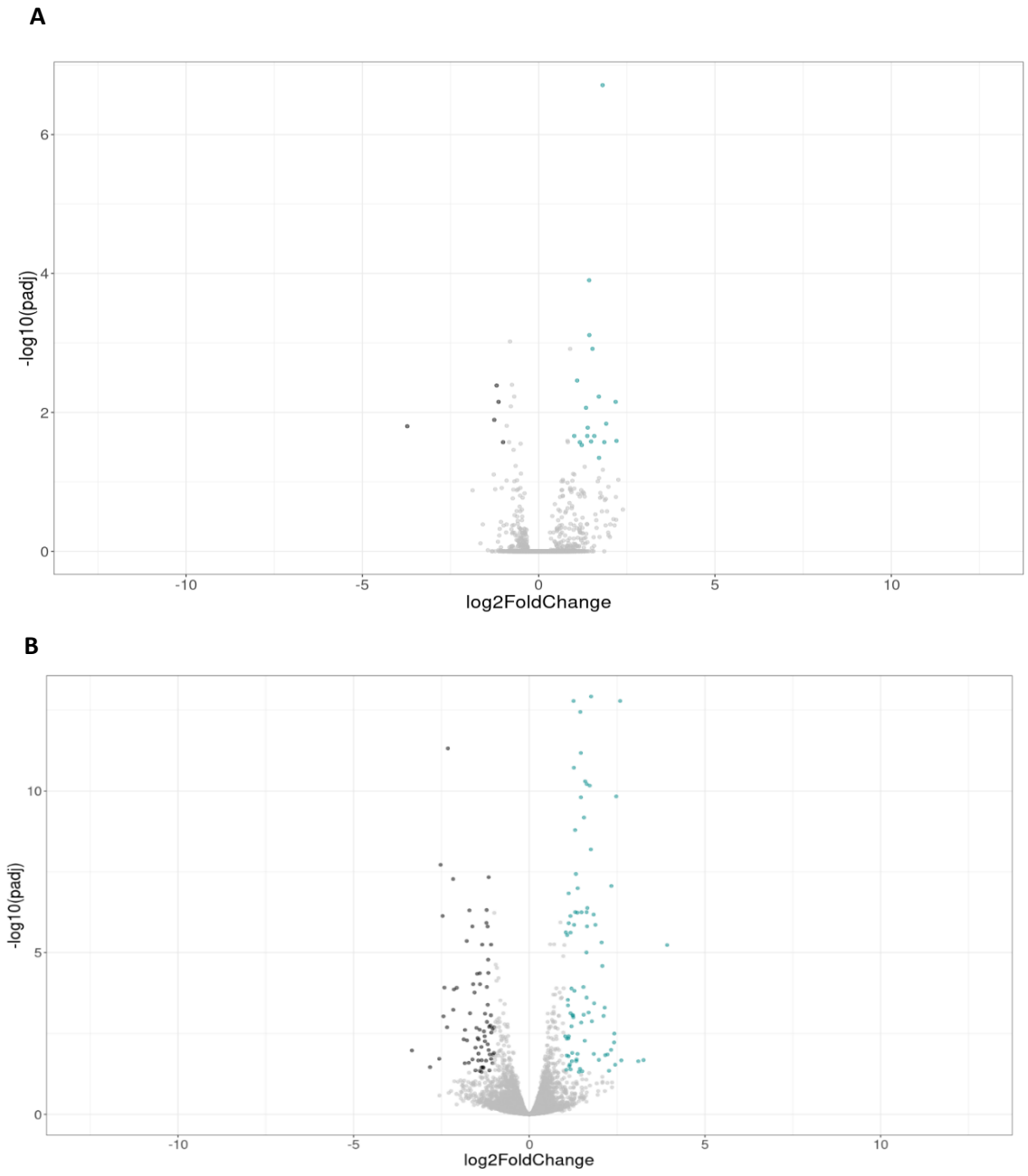


Figure 4.22. Dot plot showing the expression value of MBD2/3 in MBD2/3 RNAi and control RNAi embryos. at 48 hours **(A)** and 77 hours **(B)**. The y-axis represents the expression value of MBD2/3, and the x-axis represents the experimental group. The plots show significant decrease in the MBD2/3 expression in MBD2/3 RNAi group compared to the control group at both time points (>5 log-fold change at 48 hours and >4 log-fold change at 77 hours), indicating a successful knockdown of MBD2/3.

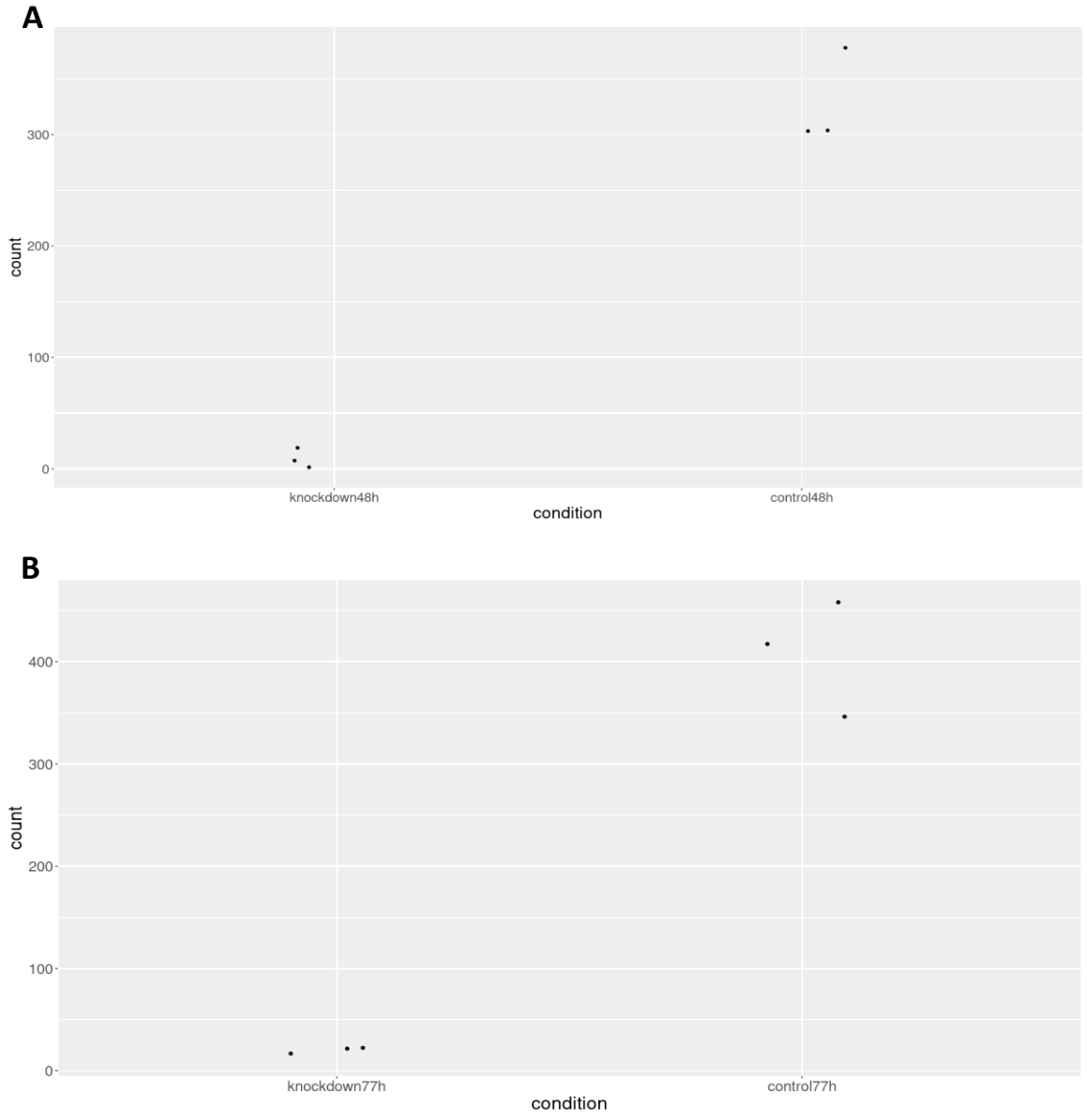


Table 4.8. Summary of genes excluded from DESeq analysis due to zero or low count.

	48h MBD2/3 RNAi timepoint	**77h MBD2/3 RNAi timepoint**
Total number of genes in the count matrix	22,756	22,756
Removed zero-count genes	4,871	5,234
Removed low-count genes	6,935	6,544
Total number of genes included in differential expression analysis	10,950	10,978
Total number of excluded genes	11,806	11,778

Table 4.9. Differentially Expressed Genes at 77 hours RNAi timepoint.

Upregulated genes				
Gene id	**padj**	**log2FoldChange**	**Top hit ortholog**	**GO term**
MSTRG.2478	6.14E-11	3.924252	--	--
MSTRG.3367	1.39E-06	3.252269	Glucose dehydrogenase	Oxidoreductase activity
MSTRG.5269	0.000997	3.1026	--	--
MSTRG.9050	5.65E-07	2.618494	Uncharacterized in *H. azteca*	--
MSTRG.10667	4.85E-06	2.586975	BZIP transcription factor	--
MSTRG.10668	0.005949	2.473838	bATF transcription factor	Heterochromatin assembly
MSTRG.10933	0.02925	2.443329	BED-type domain containing protein	DNA binding, Metal ion binding
MSTRG.15366	9.87E-06	2.419797	Sphingomyelin phosphodiesterase	Metal ion binding
MSTRG.15664	2.36E-06	2.414046	Acid phosphatase 2	Acid phosphatase activity
MSTRG.15691	1.21E-13	2.334179	--	--
MSTRG.17289	5.80E-07	2.327163	Coiled-coil domain containing protein	Transcription coactivator activity
MSTRG.17454	0.00071	2.266965	Pkd113	Carbohydrate binding
MSTRG.17624	0.001328	2.226634	Uncharacterized in *H. azteca*, (Contain RING-type domain)	Metal ion binding
MSTRG.19403	3.70E-08	2.160081	VWFC-domain containing protein	Transcription, DNA repair
MSTRG.19987	6.67E-10	2.144968	Uncharacterized in *H. azteca*	--
MSTRG.19988	6.74E-12	2.114545	--	--
MSTRG.20706	0.01583	2.0773	RNA-directed DNA polymerase, integrase H2C2-domain containing protein	DNA integration, nucleic acid binding
MSTRG.21544	0.000747	2.056999	--	--
MSTRG.21545	0.000152	1.977693	Translation initiation factor IF-2	Translation
MSTRG.22150	0.000865	1.881299	ERG transcription factor	Transcriptional regulation
MSTRG.24524	0.012596	1.846988	--	--
MSTRG.25685	4.17E-07	1.836818	RFX transcription factor	Transcription Factor
MSTRG.26610	0.000115	1.832079	--	--

MSTRG.27464	0.000369	1.779223	G-protein receptor FI-2-domain containing protein	Phototransduction
MSTRG.28071	0.04485	1.759029	--	--
MSTRG.28072	0.021403	1.751769	--	--
MSTRG.28074	6.46E-09	1.719441	--	--
MSTRG.28279	1.93E-11	1.685368	WAP-domain containing protein	Peptidase inhibitor activity
MSTRG.28281	1.62E-09	1.653269	Antimicrobial peptide type 2 llb	Peptidase inhibitor activity
MSTRG.28286	1.02E-07	1.644152	--	--
MSTRG.28292	5.59E-07	1.633084	--	--
MSTRG.29027	0.005287	1.631532	Uncharacterized in *H. azteca*	--
MSTRG.30514	5.60E-07	1.630664	PAX 6 Transcription factor	Transcriptional repressor
MSTRG.35459	0.001895	1.625602	RNA-directed DNA polymerase	DNA integration, nucleic acid binding
MSTRG.35860	0.022701	1.587778	--	--
MSTRG.36422	2.59E-05	1.578155	--	--
MSTRG.37621	0.020911	1.557033	--	--
MSTRG.39794	1.66E-13	1.547879	--	--
MSTRG.41288	0.004571	1.540775	Glutamine-fructose-6-phosphate-transamine-1	Glutamine metabolic process
MSTRG.41595	0.046472	1.501986	SH3-domain containing protein	
MSTRG.41700	0.000428	1.485691	Uncharacterized in *H. azteca*	--
MSTRG.42265	0.000833	1.478422	PRG transcription factor	Required for embryonic dorsal closure
MSTRG.43041	0.032299	1.470961	--	--
MSTRG.43720	0.013503	1.468697	Uncharacterized in *H. azteca*	--
MSTRG.44288	0.000129	1.453735	Saa1	Inflammatory response
MSTRG.45826	0.022701	1.442928	Salivary secreted protein	--
MSTRG.45828	0.022407	1.410864	Salivary secreted protein	--
MSTRG.45990	0.040673	1.378469	F-box leucine	Protein catabolic process
MSTRG.46116	0.003803	1.373089	Collagen alpha-1 (1v) chain	Extracellular matrix structural constituent
MSTRG.46154	0.048607	1.368067	F-box leucine repeat-rich protein	Protein catabolic process
MSTRG.46287	0.000288	1.345515	Uncharacterized in *H. azteca*, (Contain F-box domain)	--

MSTRG.49048	0.010105	1.335736	Cationic amino acid transporter	Transmembrane transporter activity
MSTRG.51581	8.67E-08	1.325638	Uncharacterized in *H. azteca*	--
MSTRG.52326	0.014693	1.306134	--	--
MSTRG.52337	0.013475	1.30343	Uncharacterized in *H. azteca,* (Contain RING-type domain)	
MSTRG.52352	0.000865	1.285264	Uncharacterized in *H. azteca,* (Contain RING-type domain)	
MSTRG.52366	1.49E-10	1.275246	--	--
MSTRG.52382	6.65E-07	1.268817	--	--
MSTRG.52386	0.000902	1.259825	--	--
MSTRG.52387	0.001449	1.252589	Uncharacterized in *H. azteca,* (Contain RING-type domain)	
MSTRG.52388	0.003168	1.241208	Uncharacterized in *H. azteca,* (Contain RING-type domain)	
MSTRG.52393	0.04174	1.235212	Uncharacterized transcription factor	--
MSTRG.57468	2.40E-06	1.219951	--	--
MSTRG.57565	0.039133	1.211444	--	--
MSTRG.59370	0.013991	1.200213	--	--
MSTRG.59716	0.000502	1.199981	--	--
MSTRG.60158	0.020594	1.176259	Serine/threonine-protein kinase	ATP binding, Metal ion binding, ribosome biogenesis
MSTRG.60159	0.000245	1.17439	Endo/exonuclease/phosphate-domain containing protein	Catalytic activity
MSTRG.60851	0.020027	1.173866	BcDNA.GH07269	
MSTRG.60853	2.88E-06	1.172477	--	--
MSTRG.60962	1.54E-06	1.152725	--	--
MSTRG.60963	1.66E-13	1.136987	--	--
MSTRG.60964	5.84E-06	1.119988	--	---
MSTRG.63084	1.23E-06	1.119673	Serine protease inhibitor Spn88Ea	Serine-type endopeptidase inhibitor activity
MSTRG.63227	0.028658	1.11859	Cuticular like protein	Catabolic process

MSTRG.64131	1.47E-07	1.11461	--	--
MSTRG.64140	0.004423	1.112156	Uncharacterized in *H. azteca*, (Contain RING-type domain)	
MSTRG.65394	7.38E-07	1.101447	Spn42Dd	Serine-type endopeptidase inhibitor activity
MSTRG.65655	0.00383	1.098336	Rho-GAP domain containing protein	Positive regulation of transcription by RNA polymerase II
MSTRG.65658	0.015089	1.079918	Rho-GAP domain containing protein	Positive regulation of transcription by RNA polymerase II
MSTRG.66556	3.64E-13	1.071328	C-type lectin domain family 2 member CTL9	Carbohydrate binding
MSTRG.66604	1.38E-06	1.067816	Nuclear receptor subfamily 2 group F member 1 (tII)	Transcription factor
MSTRG.66881	6.88E-11	1.053827	Uncharacterized in *H. azteca*	
MSTRG.68018	1.58E-10	1.040473	Ribosome binding protein (RRBP1)	Translation, RNA binding
MSTRG.68019	5.06E-11	1.020599	NFH protein	Maintenance of neural caliber
Downregulated genes				
Gene id	**padj**	**log2FoldChange**	**Top hit ortholog**	**GO term**
MSTRG.2899	0.014157	-1.00457	Cral/trio-domain containing protein	Lipid transfer activity
MSTRG.3330	7.38E-07	-1.00851	Ctsc	Protolysis involved in protein catabolic process
MSTRG.3478	0.043715	-1.03665	Protein associated with basic TF3 (BTF3)	Transcription factor
MSTRG.3748	0.004835	-1.04372	Frrs1	Metal ion binding
MSTRG.4939	0.001387	-1.04866	Acidic repeat-containing protein (ACRC)	Neuro peptide hormone activity
MSTRG.7060	0.021834	-1.08069	Indy	Transmembrane transporter activity
MSTRG.9004	0.00436	-1.08661	Cathepson (Ctsb)	Proteolysis
MSTRG.9031	1.55E-06	-1.0871	Uncharacterized in *H. azteca*	
MSTRG.9080	0.00041	-1.09157	OTOP1	Protein channel activity
MSTRG.10270	4.92E-07	-1.10688	DUF 885 domain-containing protein	

MSTRG.10272	0.025971	-1.1256	GDP-fucose protein O-fucosyl transferase (pofut 2)	Fucose metabolic process
MSTRG.10880	0.048607	-1.13696	Neurotrimin	Neurite outgrowth
MSTRG.11088	0.002414	-1.1403	Sphingosine-1-phosphatelyase	Fatty acid metabolic process
MSTRG.11265	0.002241	-1.15767	Amino Oxidase (Maoa)	Primary amine Oxidase activity
MSTRG.12182	0.021132	-1.15767	Dehydrogenase/reductase SDR family member	Oxidoreductase activity
MSTRG.12653	0.00479	-1.16724	RGM-domain family member B	BMP signaling pathway
MSTRG.14408	1.65E-05	-1.17207	Sphingomyelin phosphodiesterase	Metal ion binding
MSTRG.18554	0.021489	-1.18235	Heat-shock protein (HSpa19)	ATP binding
MSTRG.18970	0.000771	-1.18441	Glycine cleavage system D protein	Catabolic process
MSTRG.22749	0.000121	-1.18459	Acyl co enzyme A thioesterase 2 mitochondrial	Fatty acid metabolisom
MSTRG.25722	0.002019	-1.20525	rdx	Regulation of proteolysis, protein metabolic process
MSTRG.27961	0.020015	-1.20786	C2H2-type domain containing protein	Zinc-finger transcription factor
MSTRG.28774	4.34E-05	-1.2147	NPC2 Intracellular transporter	Cholesterol metabolic process transport
MSTRG.28852	0.000585	-1.22118	--	--
MSTRG.28856	0.00093	-1.23302	Dmel\CG3108 (LP17541p)	Proteolysis
MSTRG.28858	5.30E-08	-1.26016	Dmel\CG3108 (LP17541p)	Proteolysis
MSTRG.30121	0.005505	-1.26236	--	--
MSTRG.30150	0.00384	-1.27009	--	--
MSTRG.31102	0.002708	-1.2915	--	--
MSTRG.31936	0.00203	-1.31379	C-type lectin domain family 4 member E-like	Carbohydrate binding
MSTRG.32822	0.012597	-1.31467	Uncharacterized in *H. azteca,* (Contain zinc-finger domain)	
MSTRG.34514	0.005166	-1.34168	--	--

MSTRG.36978	0.045133	-1.3575	Integrin alpha 4	Integrin-mediated signaling pathway, cell adhesion
MSTRG.38366	4.37E-06	-1.35793	Lysosome membrane protein 2 (SCARB2)	Protein targeting to lysosome
MSTRG.41557	4.81E-07	-1.3608	NPC1	Cholesterol binding
MSTRG.41842	0.00017	-1.36314	Beta helix domain containing protein	
MSTRG.42111	0.043715	-1.40239	RNA directed DNA polymerase	Nucleic acid binding, DNA integration
MSTRG.43804	0.019134	-1.40801	CES4A	Carboxylic ester hydrolase activity
MSTRG.44007	0.008242	-1.40959	Fucosyltransferase	Protein glycosylation
MSTRG.44008	0.006808	-1.41388	--	--
MSTRG.44139	0.001803	-1.44341	Transmembrane protein 45 B (TMEM45A)	Transmembrane transporter activity
MSTRG.45164	0.013197	-1.44721	KYNU	Tryptophan catabolic process
MSTRG.45622	4.66E-08	-1.45294	MFSD8	Transmembrane transporter activity
MSTRG.46588	0.026203	-1.46932	ACSF2	Fatty acid metabolic process
MSTRG.47223	5.66E-06	-1.4828	CTRB1	Proteolysis
MSTRG.48751	0.010279	-1.5035	CASC3 exon junction complex subunit	RNA binding protein
MSTRG.49063	9.45E-05	-1.53327	Carboxypeptidase	Proteolysis, regulation of protein stability
MSTRG.50032	0.035195	-1.53344	Pkd1l3	Cellular response to acidic pH
MSTRG.51251	0.000122	-1.55798	Neprilysin (MME)	Proteolysis
MSTRG.51586	0.000115	-1.59824	MFS-domain containing protein	Transmembrane transporter activity
MSTRG.51939	0.010567	-1.61924	Shell matrix protein, SEA-domain containing protein	Nervous system development
MSTRG.52764	0.013991	-1.62148	Angiotensin converting enzyme (ACE)	Proteolysis, negative regulation of gene expression
MSTRG.53326	0.034588	-1.68743	Replication factor C subunit (RFC2)	ATP binding, DNA replication
MSTRG.53398	0.025352	-1.70628	ACADL	Leucine catabolic process
MSTRG.54233	4.52E-05	-1.72366	--	--
MSTRG.56994	0.036568	-1.77328	C2H2-type zinc-finger TF SP1	Regulation by transcription by RNA polymerase II
MSTRG.57962	1.21E-06	-1.78192	VMO1	Extracellular space

MSTRG.57963	4.24E-05	-1.83117	SLCO4A1	Transmembrane transporter activity
MSTRG.58963	9.47E-05	-1.83641	NKX2-2, Homeobox protein	DNA binding TF activity, RNA polymerase II specific
MSTRG.60362	0.019116	-1.86602	Trypsin-1	Proteolysis
MSTRG.60785	0.002948	-2.06722	Carboxylic ester hydrolase, transcription factor	Carboxylic ester hydrolase activity
MSTRG.61740	0.000747	-2.14585	Alkaline ceramidase (ACER2)	Ceramide metabolic process
MSTRG.61756	5.66E-06	-2.16524	Neurexin-4	Serine-type endopeptidase inhibitor activity
MSTRG.63909	0.002121	-2.16877	--	--
MSTRG.63914	0.000865	-2.31702	Chitin-binding type-2 domain containing protein	Carbohydrate metabolic process
MSTRG.63960	0.008671	-2.34027	TPR-Region-domain containing protein	
MSTRG.63984	4.87E-12	-2.41561	MBD2/3	
MSTRG.65285	0.034783	-2.44451	XPNPEP1	Proteolysis
MSTRG.65543	0.002453	-2.46815	Serine protease, PRSS!	Proteolysis
MSTRG.65885	1.93E-08	-2.52374	Chemotrypsin	Proteolysis
MSTRG.65887	0.000137	-2.5652	Chemotrypsin	Proteolysis
MSTRG.67176	1.54E-06	-2.82491	Sulfotransferase	Sulfotransferase activity
MSTRG.68021	0.001956	-3.34145	SLC34A2	Protein metabolic process

4.5 Discussion

In this chapter, we present functional work on DNA methylation-related genes and generate transcriptomic data in response to MBD2/3 knockdown in *Parhyale*. We found that DNA methylation machinery genes in *Parhyale*, DNMT1, DNMT3 and MBD2/3 are required for completion of normal embryogenesis. Therefore, *Parhyale* offers an exciting opportunity to functionally study the role of DNA methylation in developmentally well-defined arthropod species.

In the first part of the chapter, we conducted CRISPR/Cas9 knockout experiments on three genes: DNMT1, DNMT3 and MBD2/3. Our initial approach involved replicating the *Distalless* knockout performed by Kao et al.in 2016, using the same gRNA. The purpose of this replication was to establish a positive control experiment. The reported phenotype of the *Distalless* knockout was animals with truncated limbs, and we successfully reproduced this outcome in our experiments.

Furthermore, we achieved successful knockouts of the genes we targeted, although obtaining DNA from mutants with lethal phenotypes in the early stages presented a challenge. Despite this obstacle, we were able to detect indels in the DNMT1 knockout embryos.

In the previous chapter, we presented our RNA-seq analysis, which revealed that all three genes were provided maternally. This finding explains why the embryos were able to develop until the third day of development before they died. Since we are targeting the zygotic genome, the maternal transcript would not be affected by our knockout experiment. Any effects resulting from the loss of the gene would only become visible after the complete degradation of the maternal transcript. Therefore, we decided to switch to a knockdown experiment, which targets both the maternal and zygotic copies of the gene. However, it's important to note that if the genes were deposited as proteins, RNAi would not have an impact on them. In this experiment, we focused on the MBD2/3 gene due to the limited availability of studies on its function in the existing literature.

In our knockdown experiment, we used RNAi to knockdown MBD2/3 and observed developmental arrest before the formation of the dorsal organ and midgut anlagen at stage 11 of embryogenesis, resulting in a lethal phenotype. RNA-seq data from MBD2/3 RNAi embryos showed that the knockdown of MBD2/3 led to alterations in the expression of transcription factors that are involved in the regulating gene transcription. To determine if there is any association between MBD2/3 and DNA methylation, we combined these results with those from an EM-seq experiment conducted in our lab to examine the methylation status of differentially expressed genes (DEG). We found that most of the DEG have methylated cytosine in their gene-body or promoter region at low levels (0.1 – 19%). However, we did not observe any correlation between methylation status and the direction of change in gene expression. Both upregulated and downregulated genes had low methylation levels, and there was no significant difference in methylation levels between them.

In vertebrates, MBD proteins selectively bind methylated DNA and recruit histone-modifying enzymes, forming the NuRD complex which bridges three major epigenetic mechanisms for gene regulation: histone deacetylation, nucleosome remodeling, and recognition of methylated DNA (Leighton and Williams Jr. 2019). In invertebrates lacking DNA methylation, such as Drosophila and planarians, MBD2/3 is conserved and functions independently of DNA methylation. However, in invertebrates with a DNA methylation system, MBD2/3 may function in a methylation-dependent manner. For example, the sponge species *Ephydatia muelleri*, that possesses a complete DNA methylation system, and MBD2/3 was shown to recognize and bind methylated DNA, thus functioning in a methylation dependent manner (Cramer et al., 2017).

The findings presented in this chapter provide evidence that MBD2/3 may have methylation-dependent functions. The majority of differentially expressed genes following MBD RNAi had low levels of methylation, while a few had no methylated cytosine at all. In vertebrates, MBD2, which can recognize methylated DNA, has been found to bind sites lacking 5-methylcytosine and 5-hydroxymethylcytoine (Reynolds et al., 2012), suggesting that MBD proteins may act in both DNA methylation-dependent and

independent manner. This possibility is also supported by the detection of unmethylated genes in the DEG list in *Parhyale*.

Gene-body methylation of coding genes in invertebrates is often divided into two classes: highly methylated genes and lowly methylated genes. Some studies hypothesize that the methylation class serves as a regulatory signal (Dixon and Matz, 2022; Dixon et al., 2018). Thus, highly methylated, and lowly methylated genes exhibit differences in their transcription status, and both classes undergo group-level changes in methylation and transcription in response to environmental changes. The correlation between DNA methylation and gene expression in *Parhayle* that was presented in the previous chapter revealed a positive correlation between highly expressed genes and highly methylated genes. However, the analysis of MBD2/3 RNAi embryos showed that most of the differentially expressed genes after knockdown of MBD2/3 have low gbM, suggesting that MBD2/3 might be recruiting the NuRD complex to lowly methylated class genes. However, this hypothesis and the association between MBD2/3 and the NuRD complex in *Parhyale* requires further investigations to be tested.

In summary, this work establishes the importance of DNA methylation mediator genes for *Parhyale's* embryogenesis and lays the foundation for further studies on the role of DNA methylation in invertebrate regenerative models, such as *Parhyale*.

Chapter V

Thesis discussion and future directions

Contents

5.1 Identifying transcriptional and translational dynamics during *Parhyale's* MZT & ZGA

In this thesis, we have generated comprehensive transcriptomic data for the early stages of *Parhyale's* embryogenesis, specifically during the maternal-to-zygotic transition (MZT) and zygotic-genome activation (ZGA) periods. This dataset will be valuable for future research focused on understanding the molecular events associated with MZT and ZGA. By integrating polyA+ and total RNA sequencing datasets, we have established a timeline of events during MZT, identified maternally provided genes, and determined the minor and major timepoints for zygotic genome activation. Furthermore, we have examined the expression patterns of genes in each category. Our dataset contains a wealth of important developmental regulators that can be prioritized for future research avenues. The accuracy of our analysis was enhanced by utilizing an extensively annotated and improved genome assembly, which will serve as a valuable resource for future bioinformatic analyses in *Parhyale*, enabling more informative research outcomes.

We successfully identified maternal mRNAs and assessed their expression patterns during the sequenced developmental timepoints. As transcription is silenced until ZGA, the embryo heavily relies on post-transcriptional regulation of maternally loaded transcripts. A crucial regulator of mRNA translation efficiency and stability is the length of the poly-A tail (Eckmann et al., 2011; Weill et al., 2012), which is controlled by two opposing processes: polyadenylation and deadenylation.

Through our RNA sequencing data, we gained insights into the dynamic landscape of maternal mRNAs during the MZT, and these changes were found to be consistent with alterations in poly-A tail length. In the early stages of development, maternal mRNAs were enriched in polyadenylated transcripts, indicating the translation of important mRNAs essential for initiating embryogenesis. However, this polyadenylation was followed by a general trend of deadenylation, suggesting a decay of maternal mRNAs as development progressed.

Notably, we observed a transcriptome-wide increase in polyadenylation coinciding with the major wave of ZGA. This coupling between the expression of maternal mRNAs and poly-A tail length started to diminish after

gastrulation, where transcriptional processes began to dominate. These observations align with the dynamics of poly-A tail length during MZT observed in other organisms such as mice, pigs, and fruit flies (Liu et al., 2021; Eichhorn et al., 2016).

Our findings have provided compelling evidence for the regulation of maternal mRNAs through poly-A tail length during early embryogenesis. In future studies, we can apply specific sequencing techniques to profile the poly-A tail length of maternal transcripts, allowing us to gain deeper understanding of how poly-A tail length influences the translation and clearance of maternal mRNAs.

One such technique is poly-A tail length profiling by sequencing (PAL-seq). PAL-seq enables the measurement of the poly-A length of individual genes, regardless of their size, providing a comprehensive view of poly-A tail dynamics. This approach has been successfully applied to various species, including mouse, zebrafish, and frog embryos (Subtenly et al., 2014).

Our analysis of *Parhyale's* early embryonic development revealed a two-wave pattern of ZGA, which is consistent with observations in numerous other species (Tadros & Lipshitz, 2009). The first wave of ZGA initiates as early as 32-cell stage, with a gradual increase in the number of activated genes. The second wave occurs around the start of blastodisc formation.

Furthermore, we discovered that early zygotic transcription in *Parhyale* is characterized by the presence of short, truncated transcripts with limited evolutionary conservation and few, if any introns. This phenomenon is also observed in other species (Heyn et al., 2014; Kwasnieski et al., 2019) and could be attributed to the limited time available for transcription due to the rapid cell-cycles that occur during the early embryonic development (Shermeon and O'Farrell, 1991; Tadros and Lipshitz, 2009). However, further investigations are needed to determine if the short genes in *Parhyale* are directly influenced by the duration of the cell-cycle during the early wave of ZGA.

The observation that some of the earliest transcribed genes either lack homologs or possess orthologs solely in *H. azteca*, a closely related amphipod, suggests that these genes might be classified as "orphan genes" or lineage

specific genes (LSGs), respectively. LSGs are genes without detectable homologs outside a specific monophyletic group (Cai et al., 2006). In contrast, "orphan genes" are unique to a single species (Domazet-Loso and Tautz, 2003). Such genes constitute a significant portion of the sequenced genomes; for instance, 23% of *C. elegans* genes are orphans (Zhou et al., 2015), and 6% of honeybee genes are insect-exclusive (Johnson and Tsutsui, 2011). It is commonly believed that these genes are novel or recent evolutionary additions. Novel genes can emerge through reorganization of pre-existing genes or by originating *de novo* (Carvunis et al., 2012; Long et al., 2003). The former includes several processes: the classical model of gene duplication; exon shuffling, where two or more exons from various genes combine ectopically; retro-position, where genes are duplicated to new genomic locations via reverse transcription of parental genes; lateral gene transfer, which involves the direct transfer of genes between organisms; and gene fusion/fission, where two neighboring genes are fused into a single gene (Long et al., 2003). *De novo* gene origination, referring to the evolutionary birth of new genes, has been observed across various lineages (Tautz and Domazet-Loso, 2011; Kaessmann H. 2010; Khalturin et al., 2009). This emergence often begins with the transcription of non-genic sequences, which later acquire open reading frames (ORFs). These ORF-bearing transcripts then access the translation machinery, producing functional proteins (Carvunis et al., 2012). Such *de novo* genes can play crucial roles, often being associated with lineage-specific adaptations and diversification (McLysaght and Guerzoni, 2015). Notably, these newly formed genes typically start off as short and initially lack introns (Yang and Huang, 2011).

In 2014, a study delved into the relationship between gene length, expression level, and genomic novelty. The research specifically explored the connection between gene length and gene duplication, as indicated by gene family size (Grishkevich and Yanai, 2014). They discovered that genes with multiple copies are generally shorter and exhibit lower expression levels than single-copy genes or those within small gene families. Shorter genes have higher survival rates following duplication, while longer genes often result in only partial duplication (Grishkevich and Yanai, 2014). The study further posited an evolutionary trend: as genes evolve, they tend to increase in length. This elongation can be attributed to the insertion of transposable elements, which

extends introns and overall gene size. Consequently, longer genes often possess more splicing variants due to their increased length and face challenges in successful duplication (Grishkevich and Yanai, 2014).

The accurate identification of novel genes presents substantial challenges. Central to this issue is the pivotal assumption that the identification of novel genes rests on the absence of detectable homologs in sister lineage or other organisms during sequence similarity searches (McLysaght and Guerzoni, 2015; Weisman et al., 2020). However, one study suggests that this lack of detectable homologs might merely result from computational similarity search algorithms failing to recognize out-of-lineage homologs. This is termed "homology detection failure" (Weisman et al., 2020). The study's authors argue that such detection failure stems from the reduced statistical significance of similarity due to homolog divergence over evolutionary time, which might drop below the set significance threshold (Weisman et al., 2020). Their model indicates that a significant number of genes classified as lineage-specific could actually be attributed to homology detection failure. Yet, genuine novelty shouldn't be dismissed, especially for short, rapidly evolving genes (Weisman et al., 2020).

Building on this, a subsequent study by the same researchers incorporated the annotation method into consideration. The identification of lineage-specific genes doesn't just rely on genome sequences but crucially on accurate genome annotation (Weisman et al., 2022). The authors suggested that "annotation heterogeneity" – the variability that arises from using different annotation methods – can lead to misidentification of lineage-specific genes. To validate this, they conducted analyses on four species clades with multiple public gene annotations, comparing lineage-specific gene numbers derived from varied annotations versus a single annotation method across the clade (Weisman et al., 2022). Their results were revealing: using heterogenous annotations, as opposed to a uniform one, consistently led to an uptick in the count of lineage-specific genes. The increase spanned from a few tens to several thousands, representing up to a 15-fold surge, thereby exaggerating the lineage-specific gene count (Weisman et al., 2022).

Interestingly, the most substantial surge was observed when one annotation method was uniformly applied to a clade, while a different one was used for outgroup species. In contrast, the smallest spike was when a mix of annotation methods was used across both groups (Weisman et al., 2022). Additionally, the study underscored

that other determinants, like the specific annotation method used and sequence attributes (e.g., length, expression level, and GC content), also influenced the effect size (Weisman et al., 2022).

We also used data we generated to identify genes transcribed at the onset of ZGA. To achieve this, we employed two different strategies that have been previously used to study ZGA genes (Graf et al., 2014). The first approach relied on the absence of expression in the maternal genes set as a proxy for detection of ZGA genes. The second strategy depended on the detection of intron signals as a marker of *de novo* transcription.

Our analysis revealed that zygotic gene activation in *Parhyale* began as early as stage 5 of development. The number of activated genes peaked during a minor and a major wave of ZGA. However, the activation was occurring to an increasing degree between the two waves rather than being discrete, which is consistent with observations from other species as well (Schulz and Harrison, 2019).

Notably, we observed an enrichment of genes encoding proteins with reverse-transcriptase (RT) domains among the earliest activated genes in *Parhyale*. This suggests a potential role for retroelements during the early stages of embryonic development in *Parhyale*. The activation of RT family genes during ZGA onset has also been observed in mouse embryos (Sciamanna et al., 2011). Investigating the significance of these genes in *Parhyale* could provide valuable insights into their role in early embryonic development.

Our analysis revealed the conservation of the pioneer zinc-finger transcription factor (TF) *Zelda*, which is present in the genome of many insect and crustacean species (Ribeiro et al., 2017). It has been shown to play an essential role during MZT in Drosophila and *Tribolium castaneum* (Liang et al., 2008; Gao et al., 2022). In *Parhyale*, the ortholog of *Zelda* was found to be loaded maternally and among the earliest activated genes (by 13-14 hours stage), suggesting a conserved role of this TF during MZT that could be further investigated. Additionally, we identified many other developmental TFs during the minor and major waves of ZGA. The major wave showed a higher proportion of evolutionary conserved TFs, while the minor wave exhibited more likely species-specific genes that have newly evolved. The dataset we generated represents a rich source for

further enhancing our understanding of *Parhyale hawaiensis* and early embryonic development in this arthropod model.

5.2 Role of DNA methylation during *Parhyale's* MZT & ZGA

DNA methylation exhibits dynamic patterns during ZGA in various species (Andersen et al., 2012; Liu et al., 2018; Deng et al., 2020). In *Parhyale*, we found all the genes involved in DNA methylation establishment and erasure were maternally loaded. These genes displayed their highest expression levels at the beginning of embryogenesis, except for TET2, the demethylation gene, which was upregulated again towards the end of embryogenesis. Our investigations targeting DNMTs and MBD2/3 demonstrated their essential role in normal embryogenesis, indicating the potential involvement of DNA methylation during *Parhyale's* MZT. To explore this further, a colleague in our lab generated a dataset profiling DNA methylation at three timepoints before and after ZGA, revealing dynamic changes in the methylation landscape. Following ZGA, there was a reduction in genome-wide methylation levels, suggesting a role for DNA methylation in *Parhyale's* MZT. Integrating this dataset with our transcriptomic data, we observed positive correlation between gene-body methylation and gene expression. Interestingly, non-methylated genes were predominantly lowly expressed or not expressed at all during the examined timepoints. These findings strongly suggest that gene-body DNA methylation plays a regulatory role in gene expression and early embryonic development in *Parhyale*. Consequently, *Parhyale* presents a valuable arthropod model for conducting future studies on DNA methylation in the context of early embryogenesis or regeneration.

Indeed, while it is known that epigenetic processes, including DNA methylation, play a role in regulating the MZT, their specific functions during this critical transition remain largely unknown. Further exploration is needed in the literature to unravel the mechanisms underlying DNA methylation during the MZT process. Multiple studies have emphasized the coordination and importance of DNA methylation and histone modifications in zygotic genome activation (Deng et al., 2020; Guo et al., 2014; Dahl et al., 2016). For instance, investigations conducted in human embryonic stem cells utilizing chromatin immunoprecipitation followed by

sequencing (ChIP-seq) revealed a negative correlation between the repressive histone mark H3K27me3 and DNA methylation. Regions with high H3K27me3 levels exhibited lower DNA methylation, whereas regions lacking H3K27me3 peaks showed higher DNA methylation levels (Guo et al., 2014). In future studies on *Parhyale,* it would be valuable to explore the potential interplay between DNA methylation and histone modifications during early embryogenesis, possibly through the application of techniques like ChIP-seq at various timepoints spanning ZGA. Uncovering the intricate relationship between these epigenetic marks during early embryogenesis has the potential to enhance our understanding of their regulatory role.

5.3 Establishing functional genetic tools to study DNA methylation function

To assess the significance of DNA methylation in *Parhyale* embryogenesis, we utilized CRISPR/Cas9 mutagenesis to knock out genes involved in DNA methylation machinery. The knockout of DNMT1, DNMT3, and MBD2/3 resulted in a lethal phenotype, with embryos consistently found dead by day 4 of development. These knockout embryos failed to reach stage 14, which is characterized by the emergence of the ovoid shaped midgut anlagen. Analysis of the RNA sequencing dataset revealed that all three genes were maternally provided, indicating that the delay in lethality was likely due to the loss of expression of zygotic transcripts. Maternally loaded mRNAs appeared to mask the effects of losing DNA methylation mediator genes during the initial stages of embryogenesis. To overcome this issue, we switched to RNAi knockdown experiments, aiming to extend the impact to maternal transcripts as well. We specifically targeted the MBD2/3 gene for knockdown, and the lethal phenotype was confirmed in this experiment. Detailed imaging analysis demonstrated that knockdown embryos were arrested at stage 11 of development, with some embryos failing to reach even this stage. Collectively, the functional experiments performed in this thesis confirm the essential role of DNMTs and MBD2/3 in normal embryonic development in *Parhyale,* highlighting the importance of DNA methylation during embryogenesis.

However, it would be ideal to conduct further experiments specifically targeting maternal transcripts of these genes. One approach could involve injecting adult females in the ovary with RNAi to induce knockdown,

followed by allowing them to mate and produce embryos deficient for DNA methylation genes. Similar experiments have been conducted in the cockroach *Blatella germanica*, where maternal DNMT1 and DNMT3 were targeted by RNAi (Ventós-Alfonso et al., 2020). In line with our findings in *Parhyale* from the RNA-seq dataset, DNMT3 expression levels were significantly lower than those of DNMT1. Successful knockdown was achieved for maternal DNMT1, while DNMT3 RNAi did not produce the desired effect, despite detecting expression levels in *B. germanica*. The authors of the study proposed that RNAi depletion may be challenging for genes with very low expression, such as DNMT3. It is worth noting that the difference in survival rates between knockout embryos and control embryos was smallest in the case of DNMT3, although still significant. DNMT1 knockout embryos exhibited a greater distinction in survival rates, and mutations in knockout embryos were confirmed through Sanger sequencing of mutant DNA. Additionally, DNMT1 activity in *Parhyale* was targeted using 5-azacytidine drug by our colleague Wei Wei, known to inhibit DNMT1 levels and subsequently reduce DNA methylation levels. The resulting phenotype aligned with that of DNMT1 knockout embryos, as the embryos died by the third day of development (~72 hours post fertilization). Thes observations from *Parhyale*, along with the provided example, suggest that DNMT1, the maintenance methyltransferase, may be more crucial for these species than the *de novo* methyltransferase DNMT3. Some *in vitro* studies have demonstrated that DNMT1 can also act as a *de novo* methyltransferase. Furthermore, DNMT3 has been lost altogether in several insect orders, including Odonata, Orthoptera, and Diptera (Bewick et al., 2017; Lewis et al., 2020). Further investigations are required to elucidate the role of DNMT3 in *Parhyale*. One approach would involve collecting RNA and DNA from DNMT3 knockout embryos and assessing changes in DNA methylation levels and gene expression in response to the loss of DNMT3.

We also attempted to use CRISPR/Cas9 knock-in to tag our genes of interest and track their expression throughout the life cycle of *Parhyale*. We employed a strategy similar to the one used in Kao et al. (2016) to tag the *Distalless* gene, which is involved in limb patterning. This strategy relies on the non-homologous end joining (NHEJ) repair mechanism in a homology-independent manner. Constructs containing the coding sequence joined to a fluorescent protein were generated for DNMT3 and MBD2/3, but unfortunately, we were

unable to achieve correct tagging of the endogenous genes. Although the tagging strategy was successful in Kao et al. (2016), they reported a low efficiency of 6.6% with a survival rate of 21.4%. This low efficiency could be a contributing factor to our unsuccessful tagging attempts. Additionally, the concentration of the tagging plasmid used in Kao et al. (2016) was 10 ng/µl, which we also followed. The negative result could be attributed to the low concentration of the donor template. In future experiments, it may be worthwhile to substantially increase the concentration and determine if this leads to successful knock-in. Furthermore, since this method relies on using the same guide RNAs (gRNAs) employed in the knockout experiment, it is possible that some gRNAs are more efficient than others. Testing more than five gRNAs per gene in the knockout experiment would be necessary, although this was challenging due to time limitations. However, this could be explored in future studies and combined with the construct we have generated. Lastly, it is worth mentioning that *Parhyale* has successfully produced genetic markers and transgenesis lines in previous studies (Pavlopoulos & Averfo, 2005; Rehm et al., 2009; Kontarakis et al., 2011; Kontarakis & Pavlopoulos, 2014). Therefore, alternative tagging strategies could be considered in future research to track the expression of DNA methylation genes.

Moreover, in situ hybridization is a successful strategy that has been applied in *Parhyale* and can be utilized to track the expression of our genes of interest throughout embryogenesis. Previous studies have employed in situ hybridization on embryos of different stages, spanning the entire embryonic development, to track the expression of Hox genes (Serano et al., 2016). We can adapt this approach by designing probes specific to DNA methylation genes and performing in situ hybridization on various embryonic stages as well as regenerating tissues of the animal.

5.4 Identifying the NuRD regulatory network in *Parhyale* and how it works with gene-body methylation

MBD2/3, as part of the DNA methylation machinery, was targeted in this thesis to investigate its role in normal embryonic development. Our knockout and knockdown experiments confirmed that MBD2/3 is essential for embryogenesis, as loss of MBD2/3 resulted in early embryonic lethality, with embryos arrested at stage 11 or earlier. In vertebrates, MBD2 and MBD3, the methyl-binding domain proteins, are part of the Nucleosome Remodeling and Deacetylase (NuRD) complex. They arose from the ancestral MBD2/3 duplication observed in invertebrates (Leighton and Williams, 2019). While MBD2 can recognize and bind methylated DNA, MBD3 lacks this ability. However, both proteins associate with the NuRD complex in mutually exclusive manner. Initially, the NuRD complex was characterized as transcriptional repressor, modulating chromatin accessibility of target genes to RNA polymerase II and transcription factors (Xue et al., 1998; Wade et al., 1999; Shao et al., 2020). However, it is now recognized that the NuRD complex participates in multiple mechanisms, including transcriptional activation, DNA repair and replication, and protein modification. For example, in the context of cancer, the NuRD complex has been shown to have roles in both promoting and suppressing tumorigenesis (Denslow & Wade, 2007; Lai & Wade, 2011).

Interestingly, MBD2/3 is present in several invertebrate species that have lost DNA methylation, such as *C. elegans*, Drosophila, and planarians (Jaber-Hijazi et al., 2013; Marhold et al., 2004; Gutierrez et al., 2007). Studies in Drosophila have demonstrated that MBD2/3 interacts with the NuRD complex during embryonic development, suggesting a conserved function independent of DNA methylation.

In the case of *Parhyale*, where DNA methylation is preserved, it is predominantly found in the gene-bodies of protein-coding genes, positively correlating with gene expression. Additionally, promoter methylation was observed for a subset of genes. This suggests that *Parhyale's* MBD2/3 may participate in transcriptional repression or activation as part of the NuRD complex.

In chapter 4 of this thesis, we collected RNA from MBD2/3 knockdown embryos at two timepoints post-injection in 1-cell embryos of *Parhyale* to understand the transcriptomic response to MBD2/3 loss and its correlation with DNA methylation. The analysis revealed an enrichment of transcription factors known for their developmental significance in other species among the upregulated genes at both examined timepoints. The number of upregulated genes consistently exceeded the number of downregulated genes, with no significant difference in their methylation status. Most of the differentially expressed genes exhibited low levels of gene-body methylation, and some of the upregulated genes displayed an intermediate level of promoter methylation, including an ortholog to C2H2 TF. Only limited number of Differentially expressed (DE) genes had no methylated cytosine at all. These results indicate the involvement of MBD2/3 in gene regulation, with the number of upregulated genes during 48 hours of development being 3.8 times higher than the number of downregulated genes. However, at 77 hours, the number of genes in both categories was similar.

The absence of genes with intermediate or high methylation level in the DE dataset suggests that MBD2/3 may have an independent function apart from DNA methylation. However, it is worth noting that most DE genes exhibited methylated cytosine, which could be sufficient for MBD2/3 to bind to the relevant methylated genes and recruit regulatory mechanism.

Parhyale's MBD2/3 contains the conserved C-terminus domain of a methyl-CpG binding protein, enabling it to recognize and bind methylated DNA. This is in contrast to the Drosophila and *C. elegans* orthologs, which have lost this ability. Considering the presence of methylated cytosine at the DE genes in our datasets, albeit at low levels, we propose that the function of MBD2/3 could be DNA methylation dependent.

In summary, the work presented in this thesis demonstrates a role for MBD2/3 during the embryonic development of *Parhyale*. However, it remains to be determined whether MBD2/3 functions as part of *Parhyale's* NuRD complex. If it does, it raises the question of whether MBD2/3 coordinates NuRD activity through binding to methylated DNA or independently of DNA methylation. Further functional and transcriptomic experiments will be instrumental in answering these interesting questions.

5.5 Using *Parhyale* as a model to study DNA methylation in the context of regeneration

In the introduction chapter, we discussed the remarkable regenerative ability of *Parhyale* to regrow their appendages following injury. Within eight days, the newly regenerated limb becomes encapsulated within the exoskeleton of the previously amputated limb, and molting facilitates the restoration of the regenerated limb's structure and function (Alwes et al. 2016, Konstantinides and Averof 2014, Grillo et al. 2016). The involvement of similar events between embryonic development and regeneration suggests that there may be shared molecular mechanisms underlying the generation of these structures, such as appendages. Several studies have supported this theory, demonstrating the re-usage of embryonic genes during regeneration in other organisms like amphibians and zebrafish, including *Oct4, Sox2, and Nanog* (Muneoka & Bryant, 1982; Nacu et al., 2011; Christen et al., 2010; Carlson et al., 2001).

However, it is important to acknowledge the significant differences between embryonic development and the process of regeneration. Unlike development, which has a consistent starting point in a stable environment, regeneration begins with an injury of unpredictable extent and timing. Additionally, the regeneration process occurs in a variable environment influenced by factors such as nutrition, season, microbe exposure, and hormonal cycles, which are not constant.

Moreover, the size difference between developing and regenerating organs can be significantly greater, differing by order of magnitude. These differences lead to the conclusion that regeneration is not simply a recapitulation of embryonic development, despite the observed similarities in some events. A recent study in *Parhyale* by Sinigaglia et al. (2022) explored this comparison further. The authors compared the transcriptional profiles of regenerating legs and developing embryos during the leg developmental stage of *Parhyale* to assess the extent of similarity between leg regeneration and embryonic development. While they identified an overlapping set of genes and regulatory interactions between the two datasets, the deployment of these genes does not follow the same temporal order. This discrepancy was observed not only in processes unique to

regeneration, such as wound healing and metabolic reprogramming but also in shared events like cell proliferation, differentiation, and patterning.

Therefore, it would be interesting to elucidate the dynamics of DNA methylation during the process of regeneration in *Parhyale* and determine if it is involved in the regulatory network that coordinates regeneration. Previous studies have investigated DNA methylation dynamics during regeneration in certain vertebrate models (Powell et al. 2013; Aguilar and Gardiner, 2015; Górnikiewicz et al. 2013; Barker and Tsai 2017). For instance, in zebrafish, the DNA methylation landscapes were compared during the transition of differentiated cells, specifically Muller Glia (MG), to their progenitors (MGPC) during retina regeneration (Powell et al. 2013). Differentially methylated bases (DMBs) were identified between the two cell types, but the analysis of DMBs located at the promoter region showed little correlation between methylation and gene expression of the relevant gene (Powell et al. 2013). In axolotl, DNMT activity was inhibited during regeneration (Aguilar and Gardiner, 2015). Downregulation of DNMT3a induced the expression of *Sp9*, a transcription factor required for de-differentiation processes in axolotl limb regeneration and the formation of an early blastema (Endo et al. 2004). However, regeneration regressed after developing to the medium bud stage and failed to complete (Aguilar and Gardiner, 2015).

In the future, we could perform genome-wide profiling of DNA methylation landscapes during *Parhyale* limb regeneration. This would also involve conducting comprehensive analysis of DNA methylation profiles for candidate genes at different cell types and different time points of regeneration. Additionally, we could sequence DNA methylation profiles of progenitors to identify hypo- or hyper-methylated genes by comparing them with differentiated cells, such as satellite-like progenitors versus muscle cells.

To trace cell lineages during regeneration, we can employ the technique of injecting lineage tracers into the blastomeres of eight-cell stage embryos in *Parhyale* (Gerberding et al. 2002, Alwes et al. 2011, Nast and Extavour, 2014). Cell lineage analysis has shown that each of the 8-cells gives rise to an exclusive germline, either ectoderm, mesoderm, or endoderm, which can enable the labeling of different progenitors during regeneration. Fluorescence-activated cell sorting (FACS) can then be utilized to isolate progenitors from an

amputated limb in the formed blastema and proliferating cells. Subsequently, libraries of cell-type-specific or time-point-specific DNA methylation at target loci can be analyzed and compared to identify changes in DNA methylation patterns.

Furthermore, a relatively simple and fast approach to investigate the contribution of DNA methylation to blastema formation and subsequent regeneration events is by chemically inhibiting DNA methylation during regeneration. This can be achieved by treating the injured limb with the same drug used for embryos (5-aza-2′-deoxycytidine) (Christman 2002). Inhibiting DNA methylation may not affect the formation of the blastema, particularly if demethylation is required at the beginning of regeneration. However, the absence of DNA methylation is expected to impede later differentiation steps.

Chapter VI

Materials and Methods

Contents

6.1 *Parhyale hawaiensis* culture

The wild-type *Parhyale* used in this project was obtained from the 'German' line. The animals were reared in plastic or glass aquarium tanks filled with artificial saltwater at a salinity of 30 parts per thousands (PPT) and crushed coral. The culture is maintained in a temperature-controlled (CT) room at 24-26 degree Celsius (°C), depending on the season. Pumps and air-stones were employed to provide aeration to the aquariums. Water is changed on a weekly basis, and the animals are fed fish and carrots once a week.

6.2 Embryos collection

Embryos are available throughout the year. Mating pairs are collected the day before, and the following morning, gravid females are gathered and placed in a solution of clove oil (Sigma) diluted at a ratio of 1:10,000 in artificial seawater for anaesthesia. Once the females are completely paralyzed, embryos are easily extracted from the ventral brood pouch using forceps and a sharp glass pipette. The pipette's tip is rounded by melting it in a Bunsen burner, ensuring a safe handling of the embryos. Subsequently, the embryos are transferred into a fresh petri dish containing filtered artificial seawater (FASW). The FASW is supplemented with the antibiotic Penicillin-Streptomycin (10,000 µg/ml, ThermoFisher Scientific) at a ratio of 1:100 and the antifungal Amphotericin B (250 µg/ml, Gibco, Merck Life) at a ratio of 1:200.

6.3 Genome assembly and annotation improvement

6.3.1 Genomic DNA extraction

Genomic DNA was extracted from adult male and female *Parhyale* (German-line) using a phenol-chloroform-based extraction protocol. Two adult males or females were transferred to a 1.5 ml Eppendorf tube, and the artificial seawater (ASW) was aspirated from the tube. Next, 100 µl of lysis buffer (composed of 100 µM Tris-HCl pH 8.0, 100 µM NaCl, 50 µM EDTA, 0.5% SDS, and MiliQ H_2O) was added to the samples. The samples were flash-frozen in dry ice and kept at -80 °C overnight. After thawing, an additional 200 µl of the lysis buffer was added, and the samples were homogenized using a pestle for approximately 2 minutes. The pestle was then washed with an additional 200 µl of lysis buffer, resulting in a total volume of 500 µl of lysis buffer in the samples. Subsequently, the samples were digested with 20 mg/ml proteinase

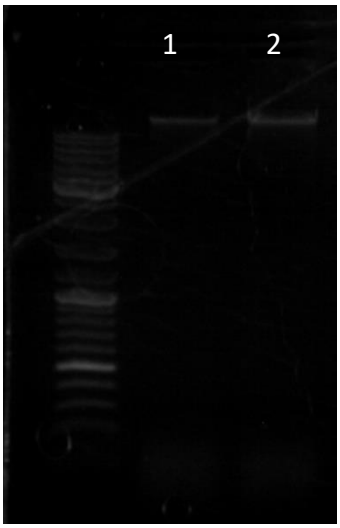
K (Thermo Scientific™, or NEB) and 20 µg/ml (2 µl) RNase A (Invitrogen or NEB). The samples were incubated in a shaking heat-block at 37 ºC and 500 rpm for 1 hour, followed by incubation at 65 ºC and 500 rpm for an additional 30 minutes in the same shaking heat-block. Following digestion, two washes with phenol (Sigma Aldrich, ≥ 99%, pH 8.0) and two washes with chloroform (Sigma Aldrich, ≥ 99.5) were performed. Each wash involved adding 1 volume of the respective solvent to the samples, gently inverting for 5 minutes by hand on ice, and centrifuging (13,000 rpm, 4 ºC) between the washes. After each centrifugation step, the upper aqueous layer was carefully transferred to a fresh tube for the next step. Next, the samples were washed once with 2 volumes of 100% ice-cold ethanol (Fisher Scientific) without any precipitation step, followed by three washes with 500 µl of 70% cold ethanol. After each wash, the samples were centrifuged for 20 minutes at 4 ºC, and the supernatant was carefully removed. Finally, the DNA pellet was air-dried at room temperature and resuspended in nuclease-free water (Invitrogen). The samples were then stored at -20 ºC. The quality of the extracted DNA was assessed using gel electrophoresis. Contamination indication was evaluated by checking that the 260/280 ratio fell within the range of 1.7 – 2.0 and the 260/230 ratio fell within the range of 2.0 – 2.2. The DNA concentration was measured using the Qubit system.

6.3.2 PacBio sequencing

High-quality genomic DNA (Figure 6.1) was sent to Novogene company for library construction, quality control, and PacBio Single-Molecule Real-Time (SMRT) sequencing. Since the PacBio data output can be highly influenced by the species, and this was the first time PacBio sequencing was performed on the *Parhyale* genome, we generated a single PacBio DNA library using the PacBio Sequel II HiFi/CCS library preparation method. The average length of the HiFi reads obtained was 17,480 bp (30.642 Gb of data), which aligns with the expected output of CCS HiFi data for animal DNA.

Figure 6.1. Genomic DNA samples used for PacBio sequencing. Panel (A) shows a gel electrophoresis image of the high-quality DNA samples sent for PacBio sequencing. Panel (B) provides measurements of the genomic DNA samples in A, including the total amount of DNA, concentration as measured by Qubit, and the 260/280 and 260/230 ratios obtained from Nanodrop analysis.

A



B

Sample No.	Volume(µl)	Amount(µg)	Concentration Qubit(ng/µl)	260/280	260/230
1	50	75000	1.5 mg/ml	2	2.34
2	50	47435	948.7	1.95	2.28

6.3.2 Scaffolding and gap filling

The existing reference genome assembly (*Phaw_5.0*) underwent re-scaffolding using the SAMBA tool with the following parameters: number of threads (-t) set to 30, scaffolding data type (-d) set to asm (corresponding to PacBio HiFi reads), minimum long-read alignment (-m) set to 5,000 bp, and maximum overhang (-o) set to 1 kb (Zimin and Salzberg, 2022). Subsequently, gap filling was performed using the TGS-GapCloser tool with the following parameters: minmap_arg set to '-x asm20', threads set to 20, and tgstype set to pb (Xu et al., 2020), utilizing the generated PacBio HiFi sequencing data.

The quality of the updated assembly (*Phaw_5.1*) was assessed using various approaches. Gaps within each scaffold were identified and quantified using the scaffoldgap2bed.py script (available at

https://github.com/lexnederbragt/sequencetools). Additionally, filtered paired-end Illumina reads were remapped to compare and evaluate the coverage and mapping rate. Assembly statistics were obtained using the assembly-stats tool (available at https://github.com/sanger-pathogens/assembly-stats), as presented in chapter 2.

To illustrate the differences in assembly gaps before and after PacBio integration, the IGV software (Robinson et al., 2011) was utilized.

6.3.4 Genome annotation

A reference-guided transcriptome assembly was conducted using all available Poly-A RNA-seq libraries. Initially, the previous *Parhyale* transcriptome (available at https://research.janelia.org/pavlopoulos/) was mapped to the *Phaw_5.1* assembly using the GMAP tool with the GTF option (Wu and Watanabe, 2005), resulting in the generation of a *.gtf* file. Subsequently, transcripts and gene models for each RNA-seq library were assembled using Stringtie2 (Kovaka et al., 2019) while incorporating the reference *.gtf* file. The Stringtie2 *–merge* parameter was employed to integrate all the generated annotation files into an initial set of *Parhyale* transcripts (n=131,116).

To reduce redundancy and obtain a set of non-redundant potential genes or loci with transcriptional activity, the number of transcripts was reduced through the application of CD-HIT, resulting in a final set of 74,420 transcripts. Predicted protein sequence and candidate coding regions were generated using *Transdecoder* (v5.5.0) (available at https://github.com/TransDecoder/TransDecoder) by selecting the longest predicted open reading frame (ORF) per transcripts (n=83,306). The list of longest peptides was further filtered to retain only peptides with hits for known protein domains, utilizing the *–retained_pfam_hits* and *–retain_blastp_hits* option.

Interproscan5(Jones et al., 2014) was employed to search for Pfam domains (Finn et al., 2014) within the longest peptides. Additionally, *Blastp_hits* (Altshul et al., 1990) were obtained by conducting a blast search against the UniProtKB/Swiss-Prot database, which is a manually annotated, non-redundant protein sequence dataset (UniProt Consortium, 2018). To reduce redundancy in the protein sequences (identity <

0.99), CD-HIT (Li and Godzik, 2006) was applied, resulting in a final set of 22,756 representative coding sequences.

The coding gene models were annotated using Interproscan5 to identify putative protein families based on both the assigned Gene Ontology (GO) annotation and the Pfam database. A similarity search was performed using the *blastp* function with an e-value ≤ 1e-06 cut-off between the *Parhyale* proteome and reference proteome datasets from 28 different metazoan species spanning various taxonomical groups. The *Orhtofinder* tool (Emms and Kelly, 2019) was utilized to identify potential orthologs shared between *Parhyale* and other model or non-model species.

6.3.5 Sequence and phylogenetic analysis of specific genes

The identity of each DNA methylation machinery gene in the *Pahw_5.1* assembly was verified using the *blastp* function. This involved comparing the *Parhyale* proteome with reference proteome datasets from various organisms, including *H. sapiens, M. musculus, D. pulex,* and *D. rario*. To confirm the exon and intron boundaries for each gene, alignment between the genomic DNA and mRNA was performed using ApE software (Davis and Jorgensen, 2022). Protein sequences for all other species were obtained from the NCBI database. Protein sequences for the DNA methylation machinery genes in *Parhyale* were generated using the mRNA sequences annotated in Kao et al. (2016) for DNMT2, MBD4, and TET2, or the mRNA sequences from clones of DNMT1, DNMT3, and MBD2/3. The ExPASy tool (Gasteiger et al., 2003) was utilized to define the protein sequences. Finally, sequences alignment and maximum-likelihood trees were constructed using MEGA7 (Kumar et al., 2016).

6.4 Cloning

Total RNA was extracted from pools of embryos at different developmental stages: early (stage 1 to stage 5), middle (stage 8 to stage 14), and late (stage 18 to stage 20). Additionally, RNA was extracted from multiple adult animals. The extraction process followed the TRIzol (Invitrogen) based protocol. Subsequently, cDNA was synthesized using the QuantiTect Reverse Transcription Kit (Qiagen).

To amplify target sequences, primers were designed based on *Parhyale's* transcriptome dataset, which covers the full-length coding sequences of the genes studied in this thesis (see primer list in Table 6.1). Polymerase Chain Reaction (PCR) was performed using *Parhyale* cDNA. The resulting amplicons were then ligated and cloned into the pGEM T-Easy Vector (Promega).

Plasmids containing the coding sequences were purified using either the Wizard plus SV Minipreps DNA Purification System (Promega) or the QIAprep Spin Miniprep Kit (Qiagen). All the plasmids were sent to the Source Bioscience sequencing company for Sanger sequencing.

Table 6.1. Primers used to amplify coding sequence of DNMT1, DNMT3, and MBD2/3.

Primer name	**Sequence (5'⟶ 3')**
DNMT1-1-For	GCTCGGAGTGGATATGGCTT
DNMT1-1-Rev	AGCCAGTACTACAGATGAAGCA
DNMT1-2-For	CGTTTGTTGAACTCAGCGCA
DNMT1-2-Rev	CTGGCAATGGTGTGGATGTG
DNMT1-3-For	CCATGCCCAGTTTGTGTGTG
DNMT1-3-Rev	TTATGAGCACGAGCATGCCA
DNMT1-4-For	ACCTGCGTCCTTTAGCTTCC
DNMT1-4-Rev	CTGCTGCCAACGCTTACAAG
DNMT1-5-For	TCTGGCTGTGGCGGTTTG
DNMT1-5-Rev	CGAGGAGTGTGTGCGTGTG
DNMT1-6-For	CCGATGGATCATGGACCAGG
DNMT1-6-Rev	GTCGGCCTTTGTGAGCTTTG
DNMT3-For	TGTGGCTCTTCGAGAACGTC
DNMT3-Rev	TTTAAGGATGCCGAAGCCGT
MBD2/3-For	TCGAACCAACTGTAGCTGCC
MBD2/3-Rev	TTGCTAAGGAGGCCTTCTGC

6.5 Transcriptome analysis of embryonic stages to identify MZT & ZGA

6.5.1 Sample preparation and library construction

For the analysis of early embryonic stages to investigate MZT (maternal-to-zygotic transition) and ZGA (zygotic genome activation), the embryos were collected as described previously. Subsequently, the embryos were washed in FASW/antibiotic/antifungal mix and staged using a bright field dissection microscope (GXM XTL3T101 Zoom Stereo Microscope).

To collect embryos at the 1-cell to 8-cell stages, the embryos were directly stored in TRIzol (Invitrogen) and flash-frozen in dry ice after dissection. For collecting later stages, synchronous early embryos were incubated at 28.33 ºC until they reached the desired stage. The Extavour lab staging table (https://www.extavourlab.com/wp-content/uploads/2017/10/Parhyale-staging-table.pdf), which follows the staging method described in Brown et al. (2005), was used. The slightly higher incubation temperature was chosen to facilitate reaching the desired stage in a shorter incubation time. Once the embryos reached the relevant stage, they were inspected using a Leica MZ16F fluorescence stereo microscope to remove any embryos that were either dead or clearly lagging behind the required stage.

Total RNA was extracted from a minimum of 30 mg of embryos per replicate using the TRIzol-based RNA extraction protocol. For each time-point, two or three replicates were generated. Each sample was used to generate mRNA libraries with poly-A enrichment and total RNA libraries with ribosomal RNA removal. The library preparation and sequencing processes were performed by Novogene (Novogene UK company Limited). Illumina sequencing platform was used for library construction, and paired end reads of 150 bp were generated.

6.5.2 Pre-analysis data processing

FastQC v0.11.9 (Andrews, 2010) was utilized to assess the quality of the raw transcriptomic libraries. Data filtering, adapter sequence trimming, and removal of low-quality reads and PCR duplicates were performed using TRIMMOMATIC with the following parameters: SLIDINGWINDOW:4:15, LEADING:3, TRAILING:3, MINLEN:36 (Bolger et al., 2014). The reads were then aligned to the *Parhyale Phaw_5.1* assembly using HISAT2 with default parameters (Kim et al., 2019).

To summarize the aligned reads for exon annotation and obtain gene-level exon counts, as well as for gene annotation to obtain gene-level gene-body counts, *FeatureCounts* (Liao et al., 2014) was employed. Gene-level intron counts were obtained by subtracting exon counts were from gene-body counts. Subsequently, the raw read counts were normalized as Transcripts Per Million (TPM) separately for exon and introns.

The median of intronic counts for each gene was calculated, and abnormal intron reads were filtered out based on the criteria that intronic regions with a count/median (intron-count) fold change (FC) ≥ 1.25 in at least 5 libraries. All reads overlapping these abnormally covered regions were discarded.

6.5.3 K-means clustering analysis

Both clustering analyses performed in the third chapter, namely the maternal transcripts polyA+/total RNA ratio clustering and exon-polyA clustering, were conducted using the K-means clustering algorithm (Kmeans function in R) (Forgy, E.W. 1965; Hartigan and Wong, 1979). The number of clusters was determined based on the sum of squared error (SSE).

For the exon-polyA clustering, a filtering criterion was applied to determine the genes included in the cluster. This purpose of this criterion was to ensure that lowly expressed genes were not excluded. Specifically, the criterion required the the sum of TPM values for each gene to be 3 or higher over a window of three consecutive stages. The clusters were visualized using a heatmap and a line-based plot. Additionally, the additional libraries (intron-polyA, exon-total, and intron-total) were reordered based on the clusters defined using the exon-polyA library. To ensure comparability, the genes within each cluster were scaled using Z-scores, bringing them to the same range, and the averaged Z-score at each timepoint in each library was plotted.

6.5.4 Detection of zygotic genome activation by first expression

To identify the first zygotic transcripts based on time-point-specific expression, we utilized the methodology described by Graf et al. (2014). We calculated the number of exonic reads for each gene and defined genes as "first expressed" based on specific criteria. A gene was considered "first expressed" if it had fewer than five reads during the 0-9 hours stage, indicating transcriptional silence, and at least 20 reads in one of the subsequent analyzed timepoints.

In addition to read count thresholds, we required that the transcript abundance in a given timepoint be significantly upregulated compared to previous stages. This was determined using a statistical approach:

the adjusted p-value had to be less than 0.05, and the fold change (FC) based on DESeq analysis had to be greater than 1.

The identification of the first expressed genes was performed separately using both the polyA+ and total RNA datasets. The reference time-point for the analysis was set as the 0-9 hours stage.

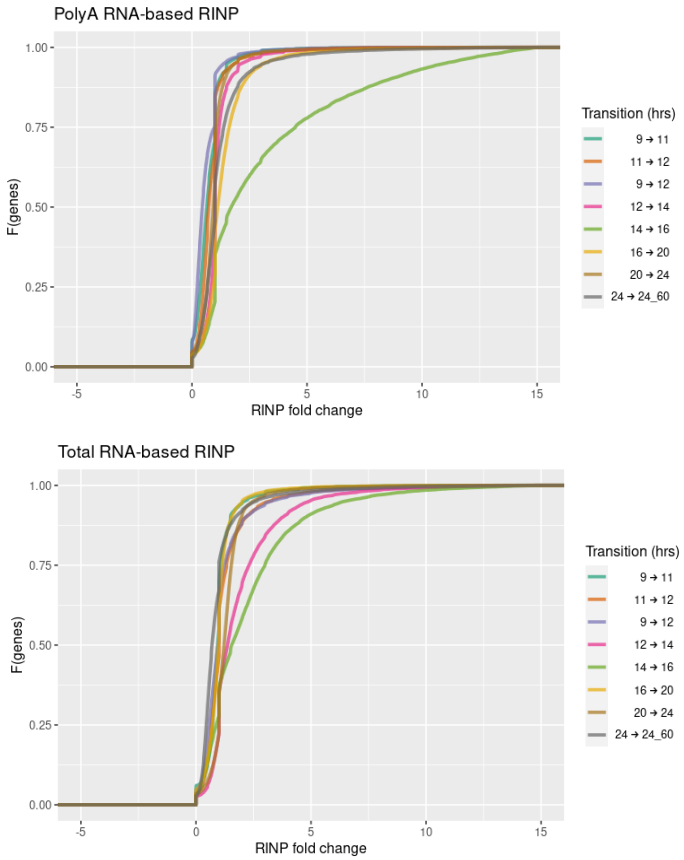
6.5.5 Detection of zygotic genome activation by intronic reads

The initial estimation of intron-specific gene coverage using the polyA+ or total RNA dataset was performed using the Bedtools coverage tool (Quinlan and Hall, 2010). The estimated counts, covered base pairs, and uncovered base pairs were aggregated for each gene. These values were then utilized to calculate the RINP (Ratio of Intronic read count to Not covered intronic Positions) score, which provides a measure for the coverage of all intronic sequences within a transcript. The RINP score was computed by summing up all mapped intronic reads for each gene and dividing it by the positions where no coverage was found.

The gene-level RINP score was averaged across replicates of each developmental timepoint that were screened. We established the 75th percentile of the RINP score obtained for the 0-9 hours stage 0.002603) (Chapter 3, Figure 3.12 A) as the threshold for intronic expression. This threshold was used to distinguish between background noise and actual intronic expression.

A gene was considered newly activated if the fold change in intronic expression between subsequent stages was equal to or greater than 2.5. This threshold was chosen based on the cumulative distribution of RINP score fold-change values. Figure 6.2 displays the cumulative distribution of RINP score fold-change between consecutive stages in polyA+ and total RNA libraries. The plot reveals that most of the transition points exhibit a uniform distribution of fold-change in their RINP scores. However, there are notable deviations observed during the transitions between 14-16 hours in the polyA+ library, as well as between 12-14 hours and 14-16 hours in the total RNA library. These transition points display a distinct pattern in the RINP score fold-change distribution, which is further supported by a notable increase in intron coverage observed at the 16-hour stages in polyA+ library and at both the 13-14-hour and16-hour stages in the total RNA library (refer to chapter 3, Figure 3.12 A).

Figure 6.2. Cumulative distribution plots of RINP score fold-change between consequetive sages. using **(A)** polyA+ and **(B)** total RNA libraries. Each line represents distribution of RINP score fold-change during the transition between two consequtive stages, as indicated in the graph key. The x-axis represents different values of RINP score fold-change, and the y-axis represents the fraction of genes.



6.5.6 Gene-body coverage analysis

To estimate the coverage of exon- and intron-based full-length transcripts at a local level, we divided each feature into 30 bp windows. We then used *FeatureCounts* to summarize the number of reads per window for exonic and intronic regions. The read counts for each window were transformed to the log2 scale, using an offset of 1 to avoid undefined values. To normalize the coverage, we divided the log2 local coverage by the maximum log-coverage, resulting in relative log-coverage values for each gene in each sample.

To visualize the coverage patterns across gene bodies, we divided all relative log-coverage values per gene into 20 sections. For genes where the length of the vector containing the windows-based relative log-coverage could not be evenly divided into 20 sections, we added the median value of the relative log-coverage to the middle of the vector. This adjustment ensured that the vector could be divided into 20 sections. Finally, we calculated the average relative log-coverage per section across genes to display a general coverage trend along the gene-body.

All statistical analyses and plots were conducted using R (https://www.R-project.org/).

The analysis of enzymatic methyl-seq (EM-seq) data was carried out by our colleague, Wei Wei. The codes for PacBio HiFi reads, RNA-seq, and EM-seq data processing and analysis have been uploaded to GitHub: (https://github.com/Mjaraespejo/Parhyale_MZT_ZGA).

6.6 CRISPR/Cas9 experiment

6.6.1 gRNA design and synthesis

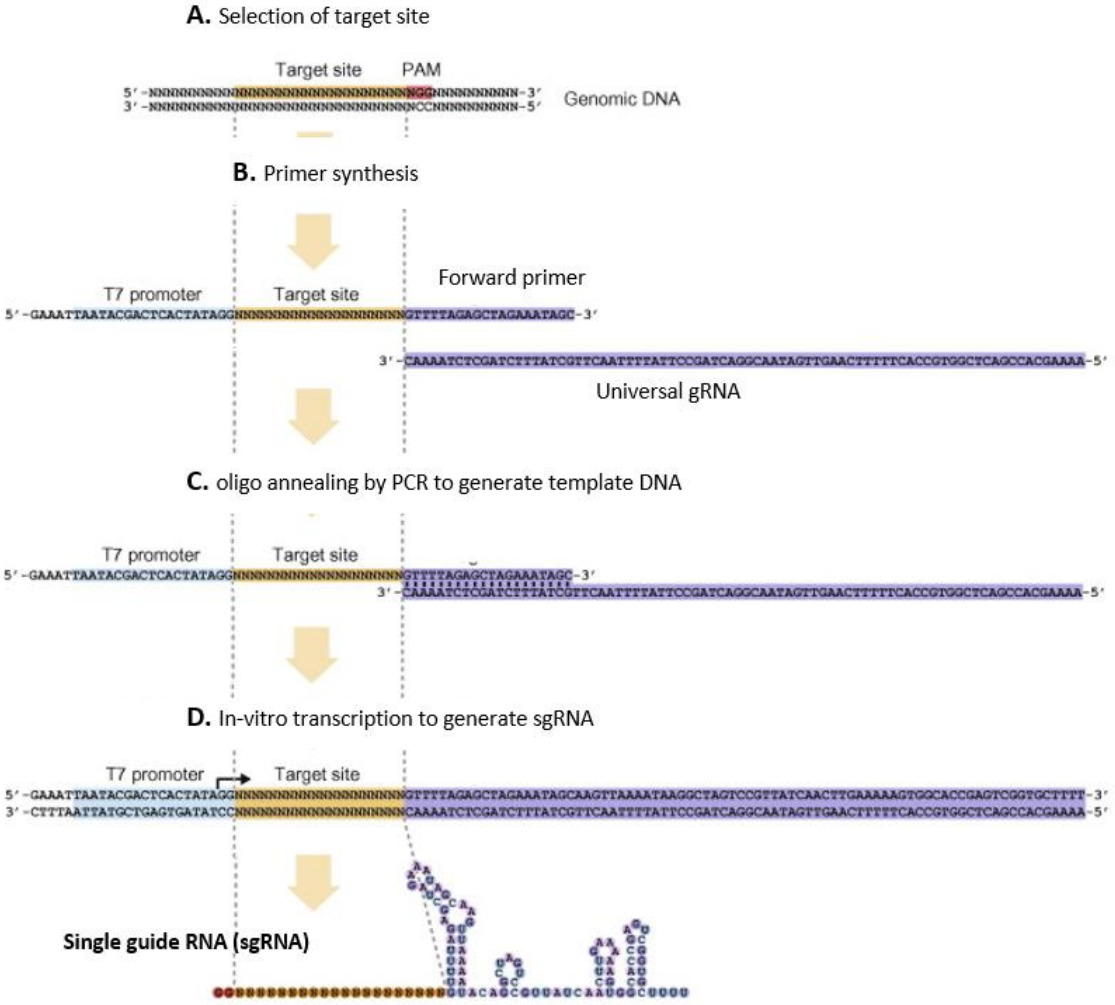
Single guide RNAs (sgRNAs) were designed using the ZiFit Targeter webtool (Sander et al., 2007; Sander et al., 2010). To synthesize the sgRNAs, primers were designed by incorporating the designed target sites into a tracrRNA sequence. The oligo annealing of the DNA template for the sgRNA was generated by PCR using Phusion High-Fidelity DNA polymerase (2U/µl) (Thermo Scientific™), with a specific forward primer for each sgRNA and a universal reverse primer for all guides (sequences listed in Table 6.2 below). The resulting amplicons were verified by gel electrophoresis and purified using either Wizard plus SV Minipreps DNA Purification System (Promega) or the QIAprep Spin Miniprep Kit (Qiagen).

The purified DNA template was then used for in-vitro transcription to generate sgRNA, following the protocol outlined in Bassett et al. (2014), using reagents from the T7 MEGAshortscript kit (Ambion). The quality of the synthesized sgRNA was checked on a 2% agarose gel, and their concentration was measured using the Qubit RNA Broad Range (BR) Assay kit (Invireogen™).

Table 6.2. Primers for sgRNA Synthesis and corresponding target sequence.

Primer name	Primer Sequence (5'⟶ 3')	gRNA name	gRNA target Sequence (including PAM sequence) (5' ⟶ 3')
Universal sgRNA	AAAAGCACCGACTCGGTGCCACTTTTTC AAGTTGATAACGGACTAGCCTTATTTTA ACTTGCTATTTCTAGCTCTAAAAC	_____	_____
DNMT1-1-For	GAAATTAATACGACTCACTATAG GGGATGGTGATATAACTGCAAGT TTTAGAGCTAGAAATAGC	DNMT1-1	GGATGGTGATATA ACTGCAAAGG
DNMT1-2-For	GAAATTAATACGACTCACTATAG GTCATATGACTCATCTGTGCCGTT TTAGAGCTAGAAATAGC	DNMT1-2	CCCTCATATGACTC ATCTGTGCC
DNMT3-1-For	GAAATTAATACGACTCACTATAG GGAGATGTACAAACGAGCCATGT TTTAGAGCTAGAAATAGC	DNMT3-1	CCGGAGATGTACA AACGAGCCAT
DNMT3-2-For	GAAATTAATACGACTCACTATAG GGCGGAGAAAGCGCCTATTCTGT TTTAGAGCTAGAAATAGC	DNMT3-2	GCGGAGAAAGCGC CTATTCTGGG
MBD2/3-1-For	GAAATTAATACGACTCACTATAG GAAAAACTGCCTCATAACACCGT TTTAGAGCTAGAAATAGC	MBD2/3-1	AAAAACTGCCTCAT AACACCTGG
MBD2/3-2-For	GAAATTAATACGACTCACTATAG GCAACTCATCCGCTTCCTGTTTTA GAGCTAGAAATAGC	MBD2/3-2	CAACTCATCCGCTT CCTCGG
MBD2/3-22-For	GAAATTAATACGACTCACTATAG GTCTTCGTTTGATTTCCGCACGTT TTAGAGCTAGAAATAGC	MBD2/3-22	TCTTCGTTTGATTT CCGCACTGG

Figure 6.3. Workflow of guide RNA (gRNA)production, adapted from Bassett et al., 2014. The figure illustrates the step-by-step process of producing gRNA. **(A)** Target site selection, the orange region indicates the selected target site in the genome where the gRNA will bind. The target site sequence is followed by the PAM sequence highlighted in red. **(B)** Primer design, the forward primer is designed to incorporate the target site sequence, represented by the orange region. The forward primer also includes the T7 promoter and overlaps with the universal sgRNA. **(C)** Oligo annealing performed through PCR to create template DNA. The annealing process combines the forward primer sequence, including the target site and the universal sgRNA. **(D)** In-vitro transcription, the template DNA undergoes in-vitro transcription using T7 polymerase, resulting in the production of mature gRNA.



6.6.2 In-vitro digestion assay

sgRNA cutting efficiency was tested using an in-vitro cleavage assay. We developed a protocol using plasmid DNA (pGEM T-Easy vector backbone) that contains the target site as a template. Five different reaction setups were prepared using NEBuffer r3.1, a restriction enzyme that cuts the plasmid once (Dra III or NcoI), the gRNA to be tested, the Cas9 protein, and the plasmid in various combinations.

The first reaction served as a negative control, where Cas9, gRNA, and the restriction enzyme were excluded. In the second reaction, only Cas9 was excluded, allowing the plasmid to be linearized by the restriction enzyme alone. The third reaction contained all the components, enabling the plasmid to be first linearized by the restriction enzyme and then cleaved by the Cas9/gRNA complex. The fourth reaction excluded the restriction enzyme to evaluate if the Cas9/gRNA complex could linearize the plasmid. The final reaction served as a negative control, excluding Cas9 while keeping the gRNA only, with the expectation that the plasmid would remain circular.

All reactions were assembled at room temperature and incubated at 37 ^{o}C for one hour. Subsequently, they were analysed by running them on a 1% agarose gel to assess the cutting efficiency for each gRNA and Cas9 batch.

6.6.3 CRISPR/Cas9 knockout injection

Microinjection was performed using the Eppendorf microinjection system, which consists of the FemtoJet 4i Eppendorf electronic microinjector, TransferMan 4r, Eppendorf Femtotip II injection capillaries, and Zeiss Axiovert S 100 TV inverted microscope. The microinjection process involved placing 1 or 2 cell embryos on a glass slide containing a 2% piece of stepped agarose.

The Cas9 protein was provided in recombinant protein form (PNA Bio Inc) and was resuspended in nuclease-free water at a starting concentration of 1 µg/µl. The Cas9 protein (500 ng/µl) was mixed with the gRNA (250 ng/µl) at a ratio of 2:1, resulting in an injection mixture. This mixture was incubated at 37 ^{o}C for 5 -7 minutes and then transferred to ice. Phenol red (5X Sigma-Aldrich) was added to a final

concentration of 0.05% to aid in visualizing the injection mixture. In some injections, Rhodamine B (Sigma-Aldrich) was also added to confirm successful injection of the mixture into the embryos.

After injection, the embryos were kept in FASW (filtered artificial seawater) supplemented with the mentioned antibiotic and antifungal agents. They were then incubated at 26 ºC and checked daily for scoring and imaging. The hatchlings resulting from the injected embryos were raised in aerated FASW to observe their behaviour and check for any post-hatching phenotypes.

6.6.4 CRISPR/Cas9 knock-in construct design and synthesis

For knock-in experiments, we utilized a construct based on the on generated in the study by Kao et al. (2016), with modifications to the coding sequence and the 3'UTR sequence to match the relevant sequences of MBD2/3 or DNMT3. The tagging plasmids were designed for CRISPR/Cas9 knock-in using a Non-Homologous End Joining (NHEJ) strategy.

Each construct contained either the full or partial coding sequence, which was fused in-frame with the T2A self-cleavage peptide, H2B *Parhyale* histone sequence, and mRuby2 fluorescent protein. These elements were inserted into the pGEM T-Easy vector backbone (Promega). Restriction enzyme digestion was employed to cut and paste to the desired sequence into the construct. Primers were designed to amplify the coding sequence and the 3'UTR sequence, incorporating adapters for ligation to the plasmid. PCR was performed to amplify the primers, followed by gel purification using the QIAquick Gel Extraction Kit (Qiagen). The amplified fragments were then digested using suitable restriction enzymes (NEB) to facilitate ligation with the digested plasmid.

After ligation, the plasmid constructs were transformed into DH5α competent cells (Thermo Fisher Scientific) and incubated overnight at 37 ºC. The transformed cells were subsequently isolated using the QIAprep Spin Miniprep Kit (Qiagen) and sent for Sanger sequencing to verify the correctness of the ligation before further processing. Additional details of the construct sequence can be found in Appendix N and O. For the injection step, the same procedure described earlier for knockout injection was followed, with the addition of the plasmid construct at a final concentration of 10 ng/µl.

6.7 RNAi experiment

6.7.1 Generation of double-stranded RNA (dsRNA)

The majority of the MBD2/3 coding sequence (CDS) was cloned into the pPR-T4P vector, and the cloned sequence was amplified by PCR using M13 primers. The purified PCR amplicon served as the template for an in-vitro transcription reaction with T7 RNA polymerase (NEB) at 37 ºC for 3 hours. To prevent RNA degradation, RNase-OUT (Invitrogen) was included in the reaction. After transcription, the samples were treated with DNase (Invitrogen) to remove the PCR template. Precipitation was performed using chilled 100% ethanol and sodium-acetate (PH 5.2) (PanReac AppliChem).

The resulting pellet was incubated at 75 ºC for 5 minutes and then gradually cooled down by incubating at incrementally decreasing temperatures in 20-second intervals, starting from 74 ºC. The samples were cooled to 37 ºC and maintained at this temperature for approximately 32 minutes. Finally, the pellet was incubated at 25 ºC for 15 minutes. The synthesized dsRNA was analysed on a 1% agarose gel, and the concentration was determined from the gel image using ImageJ software. The dsRNA was stored at -20 ºC in single injection aliquots.

The primers used to generate fragments for dsRNA production were MBD2/3i-F and MBD2/3i-R for MBD2/3 RNAi. For the negative control RNAi injection, the primers used were Ruby-F and Ruby-R. In the case of GFP RNAi, a plasmid containing the GFP sequence cloned into the pPR-T4P vector was available in the laboratory.

Table 6.3. List of primers to generate dsRNAs used in this project.

Primer name	**Sequence (5'⟶ 3')**
MBD2/3i-F	CATTACCATCCCGGCTATGAATGCTCCGCACTT
MBD2/3i-R	CCAATTCTACCCGGGTGGCGGTGTCAGTTATTT
Ruby-F	CATTACCATCCCGCCACCAATTCAAATGCACAG
Ruby-R	CCAATTCTACCCGGAACTTGGCAACTGCGTGT

6.7.2 RNA interference (RNAi)

CleanCap EGFP mRNA (Tebu Bio) was used in all RNAi injections. For the positive control experiments, dsRNA against GFP was used to knockdown injected EGFP mRNA expression, while dsRNA against MBD2/3 at a matched concentration was used as a control. Scoring was based on examining RNAi injections under a Leica MZ16F Fluorescence Stereo Microscope, and the percentage of fluorescent embryos was recorded for each injection batch.

The injections were performed using the same settings and equipment as described earlier for the CRISPR injections. During the injection, embryos were injected with a mix of dsRNA and Dextran (Alexa Flour 546; 10,000 MW; Invitrogen) for visualization. In some injections, EGFP mRNA was also added to enhance visualization. The injected embryos were then incubated at 26 ^{o}C and observed daily until hatching. Embryos were also stained with Hoechst dye (bisbenzimide H 33342, Sigma-Aldrich) at a final concentration of 5 μg/μl. The staining was performed overnight after injection, allowing the dye to permeate the embryos and label the nuclei.

6.7.3 Imaging and phenotypic scoring

Bright field and fluorescent images were acquired using a Leica MZ16F Fluorescence Stereo Microscope. The captured images were processed using Fiji software (Schindelin et al., 2012), and representative images were selected for inclusion in the figures. Embryos were monitored for normal development stages based on the staging criteria outlined in the Browne et al., 2005 paper. The first sign of lethality was observed at stage 11, where knockdown embryos lacked the characteristic aggregation of cells that would form the midgut anlagen and/or the dorsal organ, as described in the paper. By stage 14, it was evident that knockdown animals were decreased and exhibited delayed development compared to the normal embryos.

6.7.4 Sample preparation and library construction for MBD2/3 knockdown RNA-seq analysis

For collecting samples from MBD2/3 knockdown embryos and the negative control embryos, the injection procedure described earlier was followed. For the 48-hour timepoint, embryos from both conditions were injected at the 1-cell stage and then incubated at 26 ºC until reaching 48 hours of development. At this point, the embryos were collected in TRIzol (Invitrogen) and flash-frozen using dry ice.

For the 77-hour timepoint, 1-cell embryos were injected in the same manner and incubated until reaching 77-hour of development. During the incubation period, careful and regular observations were conducted to identify and remove any dead embryos under a microscope, allowing for the detection of non-viable embryos, which were subsequently removed. Furthermore, the water in which the embryos were incubated was periodically changed to maintain a clean and healthy environment for their development. At 77 hours, embryos injected with dsRNA against MBD2/3 were examined under a Leica MZ16F Fluorescence Stereo microscope to differentiate between embryos that were developing normally and those that were affected by the knockdown. Normally developing embryos were excluded from further analysis, while the group of embryos showing developmental arrest between stage 11 and prior to stage 14 were divided into two parts. Around70% of this group was stored in TRIzol and flash-frozen using dry ice, while the remaining 30% were kept in incubation to score the percentage of lethality in that batch. Batches with a survival rate of 50% or below proceeded to RNA extraction, while batches with a survival rate higher than 50% were excluded from the RNA extraction procedure.

For the control group injected with dsRNA against Ruby with matching concentration, around 70% of the batch was stored in TRIzol for later RNA extraction, and the remaining 30% continued development and exhibited normal development with a survival rate of 78-90%

Total RNA was extracted from the collected samples using a TRIzol-based RNA extraction protocol. For the 48-hour timepoint, three replicates were used for both the knockdown and control group. For the 77-

hour timepoint, three replicates were used for the control group and the knockdown group, while only two replicates were used from the knockdown embryos that escaped RNAi and developed normally.

The extracted RNA samples were shipped to Novogene for mRNA (Poly-A enrichment) library generation. Libraries were constructed using Illumina sequencing platform, and paired end reads of 150 bp were generated.

6.7.5 Statistical analysis

The paired end RNA-seq reads were subjected to trimming using TRIMMOMATIC, specifically applying the following parameters: SLIDINGWINDOW:4:15, LEADING:3, TRAILING:3, MINLEN:36. This process aimed to remove adapter sequences and low-quality reads. Subsequently, the trimmed reads were aligned to the *Parhyale Phaw_5.1* assembly using HISAT2 with default parameters.

To quantify mRNA expression, the *FeatureCounts* function from the Rsubread package was employed, with the 'exon' feature utilized. For the analysis of differential expression (DF), the DESeq2 R package (Love et al., 2014) was employed, following the standard analysis pipeline. After storing the read counts in a DESeq2 object, separate subsets were created for each timepoint (RNAi versus control), and DF analysis was performed using the DESeq function. A *p-value* threshold of 0.05 and a log2 fold change threshold of 1 were applied to identify genes that were differentially expressed.

Bibliography

Bibliography

Aanes, H., Østrup, O., Andersen, I. S., Moen, L. F., Mathavan, S., Collas, P., & Alestrom, P. (2013). Differential transcript isoform usage pre- and post-zygotic genome activation in zebrafish. *BMC Genomics*, *14*(1), 331.

Aanes, H., Winata, C. L., Lin, C. H., Chen, J. P., Srinivasan, K. G., Lee, S. G. P., Lim, A. Y. M., Hajan, H. S., Collas, P., Bourque, G., Gong, Z., Korzh, V., Aleström, P., & Mathavan, S. (2011). Zebrafish mRNA sequencing deciphers novelties in transcriptome dynamics during maternal to zygotic transition. *Genome Research*, *21*(8), 1328–1338.

Abe, K.-I., Yamamoto, R., Franke, V., Cao, M., Suzuki, Y., Suzuki, M. G., Vlahovicek, K., Svoboda, P., Schultz, R. M., & Aoki, F. (2015). The first murine zygotic transcription is promiscuous and uncoupled from splicing and 3' processing. *The EMBO Journal*, *34*(11), 1523–1537.

Aguilar, C., & Gardiner, D. M. (2015). DNA Methylation Dynamics Regulate the Formation of a Regenerative Wound Epithelium during Axolotl Limb Regeneration. *PloS One*, *10*(8), e0134791.

Albalat, R., Martí-Solans, J., & Cañestro, C. (2012). DNA methylation in amphioxus: from ancestral functions to new roles in vertebrates. *Briefings in Functional Genomics*, *11*(2), 142–155.

Alberstat, E. J., Chung, K., Sun, D. A., Ray, S., & Patel, N. H. (2022). Combinatorial interactions of Hox genes establish appendage diversity of the amphipod crustacean Parhyale hawaiensis. *BioRxiv*.

Ali-Murthy, Z., Lott, S. E., Eisen, M. B., & Kornberg, T. B. (2013). An essential role for zygotic expression in the pre-cellular Drosophila embryo. *PLoS Genetics*, *9*(4), e1003428.

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.

Alwes, F., Enjolras, C., & Averof, M. (2016). Live imaging reveals the progenitors and cell dynamics of limb regeneration. *ELife*, *5*, e19766.

Alwes, F., Hinchen, B., & Extavour, C. G. (2011). Patterns of cell lineage, movement, and migration from germ layer specification to gastrulation in the amphipod crustacean Parhyale hawaiensis. *Developmental Biology*, *359*(1), 110–123.

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 30.

Ameur, A., Zaghlool, A., Halvardson, J., Wetterbom, A., Gyllensten, U., Cavelier, L., & Feuk, L. (2011). Total RNA sequencing reveals nascent transcription and widespread co-

transcriptional splicing in the human brain. *Nature Structural & Molecular Biology*, *18*(12), 1435–1440.

Amodeo, A. A., Jukam, D., Straight, A. F., & Skotheim, J. M. (2015). Histone titration against the genome sets the DNA-to-cytoplasm threshold for the *Xenopus* midblastula transition. *Proceedings of the National Academy of Sciences*, *112*(10), E1086–E1095.

Andersen, I. S., Reiner, A. H., Aanes, H., Aleström, P., & Collas, P. (2012). Developmental features of DNA methylation during activation of the embryonic zebrafish genome. *Genome Biology*, *13*(7), R65.

Aoki, F., Worrad, D. M., & Schultz, R. M. (1997). Regulation of transcriptional activity during the first and second cell cycles in the preimplantation mouse embryo. *Developmental Biology*, *181*(2), 296–307.

Arkhipova, I. R. (2005). Mobile genetic elements and sexual reproduction. *Cytogenetic and Genome Research*, *110*(1–4), 372–382.

Arnold, A., Rahman, M. M., Lee, M. C., Muehlhaeusser, S., Katic, I., Gaidatzis, D., Hess, D., Scheckel, C., Wright, J. E., Stetak, A., Boag, P. R., & Ciosk, R. (2014). Functional characterization of C. elegans Y-box-binding proteins reveals tissue-specific functions and a critical role in the formation of polysomes. *Nucleic Acids Research*, *42*(21), 13353–13369.

Aronson, B. E., Rabello Aronson, S., Berkhout, R. P., Chavoushi, S. F., He, A., Pu, W. T., Verzi, M. P., & Krasinski, S. D. (2014). GATA4 represses an ileal program of gene expression in the proximal small intestine by inhibiting the acetylation of histone H3, lysine 27. *Biochimica et Biophysica Acta*, *1839*(11), 1273–1282.

Artieri, C. G., & Fraser, H. B. (2014). Evolution at two levels of gene expression in yeast. *Genome Research*, *24*(3), 411–421.

Atallah, J., & Lott, S. E. (2018). Evolution of maternal and zygotic mRNA complements the early Drosophila embryo. *PLOS Genetics*, *14*(12), 1–29.

Atlasi, Y., & Stunnenberg, H. G. (2017). The interplay of epigenetic marks during stem cell differentiation and development. *Nature Reviews Genetics*, *18*(11), 643–658.

Ball, M. P., Li, J. B., Gao, Y., Lee, J.-H., LeProust, E. M., Park, I.-H., Xie, B., Daley, G. Q., & Church, G. M. (2009). Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature Biotechnology*, *27*(4), 361–368.

Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, *27*(2 Pt 1), 299–308.

Barker, S. J., & Tsai, L.-H. (2017). MethyLock: DNA Demethylation Is the Epigenetic Key to Axon Regeneration. *Neuron*, *94*(2), 221–223.

Barnard, J. L. (Jerry L., & Karaman, G. S. (1991). In *The Families and genera of marine gammaridean Amphipoda (except marine gammaroids)*. Australian Museum.

Barnard, J. L. (1965). "Marine Amphipoda of atolls in Micronesia." *Proceedings of the United States National Museum.*, *117*((3516)), 459–552.

Barra, V., Schillaci, T., Lentini, L., Costa, G., & Di Leonardo, A. (2012). Bypass of cell cycle arrest induced by transient DNMT1 post-transcriptional silencing triggers aneuploidy in human cells. *Cell Division*, *7*(1), 2.

Bashir, A., Klammer, A., Robins, W. P., Chin, C.-S., Webster, D., Paxinos, E., Hsu, D., Ashby, M., Wang, S., Peluso, P., Sebra, R., Sorenson, J., Bullard, J., Yen, J., Valdovino, M., Mollova, E., Luong, K., Lin, S., LaMay, B., … Schadt, E. E. (2012). A hybrid approach for the automated finishing of bacterial genomes. *Nature Biotechnology*, *30*(7), 701–707.

Bashirullah, A., Cooperstock, R. L., & Lipshitz, H. D. (2001). Spatial and temporal control of RNA stability. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(13), 7025–7028.

Bassett, A., & Liu, J.-L. (2014). CRISPR/Cas9 mediated genome engineering in Drosophila. *Methods (San Diego, Calif.)*, *69*(2), 128–136.

Baugh, L. R., Hill, A. A., Slonim, D. K., Brown, E. L., & Hunter, C. P. (2003). Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. *Development*, *130*(5), 889–900.

Bazzini, A. A., Lee, M. T., & Giraldez, A. J. (2012). Ribosome Profiling Shows That miR-430 Reduces Translation Before Causing mRNA Decay in Zebrafish. *Science*, *336*(6078), 233–237.

Benoit, P., Papin, C., Kwak, J. E., Wickens, M., & Simonelig, M. (2008). PAP- and GLD-2-type poly(A) polymerases are required sequentially in cytoplasmic polyadenylation and oogenesis in Drosophila. *Development*, *135*(11), 1969–1979.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., … Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), 53–59.

Berleth, T., Burri, M., Thoma, G., Bopp, D., Richstein, S., Frigerio, G., Noll, M., & Nüsslein-Volhard, C. (1988). The role of localization of bicoid RNA in organizing the anterior pattern of the Drosophila embryo. *The EMBO Journal*, *7*(6), 1749–1756.

Bestor, T. H. (1999). Sex brings transposons and genomes into conflict. *Genetica*, *107*(1), 289–295.

Bewick, A. J., Sanchez, Z., Mckinney, E. C., Moore, A. J., Moore, P. J., & Schmitz, R. J. (2019). Dnmt1 is essential for egg production and embryo viability in the large milkweed bug, Oncopeltus fasciatus. *Epigenetics & Chromatin*, *12*(1), 6.

Bewick, A. J., Vogel, K. J., Moore, A. J., & Schmitz, R. J. (2017). Evolution of DNA Methylation across Insects. *Molecular Biology and Evolution*, *34*(3), 654–665.

Bilandžija, H., Laslo, M., Porter, M. L., & Fong, D. W. (2017). Melanization in response to wounding is ancestral in arthropods and conserved in albino cave species. *Scientific Reports*, *7*(1), 17148.

Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research*, *8*(7), 1499–1504.

Bird, A. P., & Wolffe, A. P. (1999). Methylation-induced repression--belts, braces, and chromatin. *Cell*, *99*(5), 451–454.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & Development*, *16*(1), 6–21.

Blythe, S. A., & Wieschaus, E. F. (2016). Establishment and maintenance of heritable chromatin structure during early Drosophila embryogenesis. *ELife*, *5*, e20148.

Boag, P. R., Nakamura, A., & Blackwell, T. K. (2005). A conserved RNA-protein complex component involved in physiological germline apoptosis regulation in C. elegans. *Development*, *132*(22), 4975–4986.

Boetzer, M., & Pirovano, W. (2014). SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics*, *15*(1), 211.

Bogan, S. N., Johnson, K. M., & Hofmann, G. E. (2020). Changes in Genome-Wide Methylation and Gene Expression in Response to Future pCO2 Extremes in the Antarctic Pteropod Limacina helicina antarctica. *Frontiers in Marine Science*, *6*.

Bouniol, C., Nguyen, E., & Debey, P. (1995). Endogenous transcription occurs at the 1-cell stage in the mouse embryo. *Experimental Cell Research*, *218*(1), 57–62.

Bouvet, P., & Wolffe, A. P. (1994). A role for transcription and FRGY2 in masking maternal mRNA within Xenopus oocytes. *Cell*, *77*(6), 931–941.

Braude, P., Bolton, V., & Moore, S. (1988). Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. *Nature*, *332*(6163), 459–461.

Braude, P., Pelham, H., Flach, G., & Lobatto, R. (1979). Post-transcriptional control in the early mouse embryo. *Nature*, *282*(5734), 102–105.

Brinkman, A. B., Gu, H., Bartels, S. J. J., Zhang, Y., Matarese, F., Simmer, F., Marks, H., Bock, C., Gnirke, A., Meissner, A., & Stunnenberg, H. G. (2012). Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Research*, *22*(6), 1128–1138.

Briscoe, J., Sussel, L., Serup, P., Hartigan-O'Connor, D., Jessell, T. M., Rubenstein, J. L., & Ericson, J. (1999). Homeobox gene Nkx2.2 and specification of neuronal identity by graded Sonic hedgehog signalling. *Nature*, *398*(6728), 622–627.

Browne, W. E., Price, A. L., Gerberding, M., & Patel, N. H. (2005). Stages of embryonic development in the amphipod crustacean, Parhyale hawaiensis. *Genesis*, *42*(3), 124–149.

Bruce, H. S., Jerz, G., Kelly, S. R., McCarthy, J., Pomeranz, A., Senevirathne, G., &, & Patel, N. H. (2021). Hybridization chain reaction (HCR) in situ protocol. *Protocols.Io.*

Bruce, H. S., & Patel, N. H. (2020). Knockout of crustacean leg patterning genes suggests that insect wings and body walls evolved from ancient leg segments. *Nature Ecology & Evolution*, *4*(12), 1703–1712.

Burgess, D. J. (2018). Genomics: Next regeneration sequencing for reference genomes. *Nature Reviews. Genetics*, *19*(3), 125.

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, *31*(12), 1119–1125.

Bushati, N., Stark, A., Brennecke, J., & Cohen, S. M. (2008). Temporal Reciprocity of miRNAs and Their Targets during the Maternal-to-Zygotic Transition in Drosophila. *Current Biology*, *18*(7), 501–506.

Cai, J. J., Woo, P. C. Y., Lau, S. K. P., Smith, D. K., & Yuen, K.-Y. (2006). Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota. *Journal of Molecular Evolution*, *63*(1), 1–11.

Calvo, L., Birgaoanu, M., Pettini, T., Ronshaugen, M., & Griffiths-Jones, S. (2022). The embryonic transcriptome of Parhyale hawaiensis reveals different dynamics of microRNAs and mRNAs during the maternal-zygotic transition. *Scientific Reports*, *12*(1), 174.

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., & Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, *18*(1), 188–196.

Carlson, M. R. J., Komine, Y., Bryant, S. V, & Gardiner, D. M. (2001). Expression of Hoxb13 and Hoxc10 in Developing and Regenerating Axolotl Limbs and Tails. *Developmental Biology*, *229*(2), 396–406.

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., Brar, G. A., Weissman, J. S., Regev, A., Thierry-Mieg, N., Cusick, M. E., & Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, *487*(7407), 370–374.

Cavallini, B., Huet, J., Plassat, J.-L., Sentenac, A., Egly, J.-M., & Chambon, P. (1988). A yeast activity can substitute for the HeLa cell TATA box factor. *Nature*, *334*(6177), 77–80.

Chaw, R. C., & Patel, N. H. (2012). Independent migration of cell populations in the early gastrulation of the amphipod crustacean Parhyale hawaiensis. *Developmental Biology*, *371*(1), 94–109.

Chen, L., Dumelie, J. G., Li, X., Cheng, M. H. K., Yang, Z., Laver, J. D., Siddiqui, N. U., Westwood, J. T., Morris, Q., Lipshitz, H. D., & Smibert, C. A. (2014). Global regulation of mRNA translation and stability in the early Drosophila embryo by the Smaug RNA-binding protein. *Genome Biology*, *15*(1), R4.

Chen, S. S., Fitzgerald, W., Zimmerberg, J., Kleinman, H. K., & Margolis, L. (2007). Cell-cell and cell-extracellular matrix interactions regulate embryonic stem cell differentiation. *Stem Cells (Dayton, Ohio)*, *25*(3), 553–561.

Chen, T., Ueda, Y., Dodge, J. E., Wang, Z., & Li, E. (2003). Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Molecular and Cellular Biology*, *23*(16), 5594–5605.

Christen, B., Robles, V., Raya, M., Paramonov, I., & Belmonte, J. C. I. (2010). Regeneration and reprogramming compared. *BMC Biology*, *8*(1), 5.

Christman, J. K. (2002). 5-Azacytidine and 5-aza-2'-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy. *Oncogene*, *21*(35), 5483–5495.

Ciechanover, A. (2005). Proteolysis: from the lysosome to ubiquitin and the proteasome. In *Nature reviews. Molecular cell biology* (Vol. 6, Issue 1, pp. 79–87).

Clark-Hachtel, C. M., & Tomoyasu, Y. (2020). Two sets of candidate crustacean wing homologues and their implication for the origin of insect wings. *Nature Ecology & Evolution*, *4*(12), 1694–1702.

Collart, C., Owens, N. D. L., Bhaw-Rosun, L., Cooper, B., De Domenico, E., Patrushev, I., Sesay, A. K., Smith, J. N., Smith, J. C., & Gilchrist, M. J. (2014). High-resolution analysis of gene activity during the Xenopus mid-blastula transition. *Development*, *141*(9), 1927–1939.

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P. D., Wu, X., Jiang, W., Marraffini, L. A., & Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science (New York, N.Y.)*, *339*(6121), 819–823.

Coombe, L., Li, J. X., Lo, T., Wong, J., Nikolic, V., Warren, R. L., & Birol, I. (2021). LongStitch: high-quality genome assembly correction and scaffolding using long reads. *BMC Bioinformatics*, *22*(1), 534.

Cotton, A. M., Price, E. M., Jones, M. J., Balaton, B. P., Kobor, M. S., & Brown, C. J. (2015). Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Human Molecular Genetics*, *24*(6), 1528–1539.

Coudert, E., Gehant, S., de Castro, E., Pozzato, M., Baratin, D., Neto, T., Sigrist, C. J. A., Redaschi, N., Bridge, A., & Consortium, T. U. (2022). Annotation of biologically relevant ligands in UniProtKB using ChEBI. *Bioinformatics*, *39*(1).

Cramer, J. M., Pohlmann, D., Gomez, F., Mark, L., Kornegay, B., Hall, C., Siraliev-Perez, E., Walavalkar, N. M., Sperlazza, M. J., Bilinovich, S., Prokop, J. W., Hill, A. L., & Williams, D. C. J. (2017). Methylation specific targeting of a chromatin remodeling complex from sponges to humans. *Scientific Reports*, *7*, 40674.

Crickard, J. B., Lee, J., Lee, T.-H., & Reese, J. C. (2017). The elongation factor Spt4/5 regulates RNA polymerase II transcription through the nucleosome. *Nucleic Acids Research*, *45*(11), 6362–6374.

Crosby, I. M., Gandolfi, F., & Moor, R. M. (1988). Control of protein synthesis during early cleavage of sheep embryos. *Journal of Reproduction and Fertility*, *82*(2), 769–775.

Cui, J., Sartain, C. V, Pleiss, J. A., & Wolfner, M. F. (2013). Cytoplasmic polyadenylation is a major mRNA regulator during oogenesis and egg activation in Drosophila. *Developmental Biology*, *383*(1), 121–131.

Cunningham, C. B., Ji, L., Wiberg, R. A. W., Shelton, J., McKinney, E. C., Parker, D. J., Meagher, R. B., Benowitz, K. M., Roy-Zokan, E. M., Ritchie, M. G., Brown, S. J., Schmitz, R. J., & Moore, A. J. (2015). The Genome and Methylome of a Beetle with Complex Social Behavior, Nicrophorus vespilloides (Coleoptera: Silphidae). *Genome Biology and Evolution*, *7*(12), 3383–3396.

Dahl, J. A., Jung, I., Aanes, H., Greggains, G. D., Manaf, A., Lerdrup, M., Li, G., Kuan, S., Li, B., Lee, A. Y., Preissl, S., Jermstad, I., Haugen, M. H., Suganthan, R., Bjørås, M., Hansen, K., Dalen, K. T., Fedorcsak, P., Ren, B., & Klungland, A. (2016). Broad histone H3K4me3 domains in mouse oocytes modulate maternal-to-zygotic transition. *Nature*, *537*(7621), 548–552.

Dalle Nogare, D. E., Pauerstein, P. T., & Lane, M. E. (2009). G2 acquisition by transcription-independent mechanism at the zebrafish midblastula transition. *Developmental Biology*, *326*(1), 131–142.

Dana, J. D. (1853). Crustacea. Part II. In *In: United States Exploring Expedition. During the years 1838, 1839, 1840, 1841, 1842. Under the command of Charles Wilkes. U. S. N. C. Sherman. Philadelphia* (Vol. 14, pp. 689–1618).

Dattani, A., Sridhar, D., & Aziz Aboobaker, A. (2019). Planarian flatworms as a new model system for understanding the epigenetic regulation of stem cell pluripotency and differentiation. *Seminars in Cell & Developmental Biology*, *87*, 79–94.

Davis, D. L. (1985). Culture and storage of pig embryos. *Journal of Reproduction and Fertility. Supplement*, *33*, 115–124.

Davis, M. W., & Jorgensen, E. M. (2022). ApE, A Plasmid Editor: A Freely Available DNA Manipulation and Visualization Program. *Frontiers in Bioinformatics*, *2*.

de Mendoza, A., Lister, R., & Bogdanovic, O. (2019). Evolution of DNA Methylome Diversity in Eukaryotes. *Journal of Molecular Biology*. (a)

de Mendoza, A., Hatleberg, W. L., Pang, K., Leininger, S., Bogdanovic, O., Pflueger, J., Buckberry, S., Technau, U., Hejnol, A., Adamska, M., Degnan, B. M., Degnan, S. M., & Lister, R. (2019). Convergent evolution of a vertebrate-like methylome in a marine sponge. *Nature Ecology & Evolution*, *3*(10), 1464–1473. (b)

De Renzis, S., Elemento, O., Tavazoie, S., & Wieschaus, E. F. (2007). Unmasking Activation of the Zygotic Genome Using Chromosomal Deletions in the Drosophila Embryo. *PLOS Biology*, *5*(5), 1–16.

Deng, M., Zhang, G., Cai, Y., Liu, Z., Zhang, Y., Meng, F., Wang, F., & Wan, Y. (2020). DNA methylation dynamics during zygotic genome activation in goat. *Theriogenology*, *156*, 144–154.

Denslow, S. A., & Wade, P. A. (2007). The human Mi-2/NuRD complex and gene regulation. *Oncogene*, *26*(37), 5433–5438.

Diamante, L., & Martello, G. (2022). Metabolic regulation in pluripotent stem cells. *Current Opinion in Genetics & Development*, *75*, 101923.

Dixon, G. B., Bay, L. K., & Matz, M. V. (2014). Bimodal signatures of germline methylation are linked with gene expression plasticity in the coral Acropora millepora. *BMC Genomics*, *15*(1), 1109.

Dixon, G., Liao, Y., Bay, L. K., & Matz, M. V. (2018). Role of gene body methylation in acclimatization and adaptation in a basal metazoan. *Proceedings of the National Academy of Sciences*, *115*(52), 13342–13346.

Dixon, G., & Matz, M. (2022). Changes in gene body methylation do not correlate with changes in gene expression in Anthozoa or Hexapoda. *BMC Genomics*, *23*(1), 234.

Domazet-Loso, T., & Tautz, D. (2003). An evolutionary analysis of orphan genes in Drosophila. *Genome Research*, *13*(10), 2213–2219.

Du, Q., Luu, P.-L., Stirzaker, C., & Clark, S. J. (2015). Methyl-CpG-binding domain proteins: readers of the epigenome. *Epigenomics*, *7*(6), 1051–1073.

Duan, J., Rieder, L., Colonnetta, M. M., Huang, A., Mckenney, M., Watters, S., Deshpande, G., Jordan, W., Fawzi, N., & Larschan, E. (2021). CLAMP and Zelda function together to promote Drosophila zygotic genome activation. *ELife*, *10*, e69937.

Eckersley-Maslin, M. A., Alda-Catalinas, C., & Reik, W. (2018). Dynamics of the epigenetic landscape during the maternal-to-zygotic transition. *Nature Reviews Molecular Cell Biology*, *19*(7), 436–450.

Eckmann, C. R., Rammelt, C., & Wahle, E. (2011). Control of poly(A) tail length. *Wiley Interdisciplinary Reviews. RNA*, *2*(3), 348–361.

Edgar, B. A., & Datar, S. A. (1996). Zygotic degradation of two maternal Cdc25 mRNAs terminates Drosophila's early cell cycle program. *Genes & Development*, *10*(15), 1966–1977.

Edgar, B. A., & Schubiger, G. (1986). Parameters controlling transcriptional activation during early drosophila development. *Cell*, *44*(6), 871–877.

Edwards, J. R., Yarychkivska, O., Boulard, M., & Bestor, T. H. (2017). DNA methylation and DNA methyltransferases. *Epigenetics & Chromatin*, *10*, 23.

Eichhorn, S. W., Subtelny, A. O., Kronja, I., Kwasnieski, J. C., Orr-Weaver, T. L., & Bartel, D. P. (2016). mRNA poly(A)-tail changes specified by deadenylation broadly reshape translation in Drosophila oocytes and early embryos. *ELife*, *5*.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., … Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science (New York, N.Y.)*, *323*(5910), 133–138.

Elango, N., & Yi, S. V. (2008). DNA methylation and structural and functional bimodality of vertebrate promoters. *Molecular Biology and Evolution*, *25*(8), 1602–1608.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238.

Endo, T., Bryant, S. V, & Gardiner, D. M. (2004). A stepwise model system for limb regeneration. *Developmental Biology*, *270*(1), 135–145.

English, A. C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D. M., Reid, J. G., Worley, K. C., & Gibbs, R. A. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PloS One*, *7*(11), e47768.

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., & Marks, D. S. (2003). MicroRNA targets Drosophila. *Genome Biology*, *5*(1), R1.

Estella, C., Rieckhof, G., Calleja, M., & Morata, G. (2003). The role of buttonhead and Sp1 in the development of the ventral imaginal discs of Drosophila. *Development (Cambridge, England)*, *130*(24), 5929–5941.

Extavour, C. G. (2005). The fate of isolated blastomeres with respect to germ cell formation in the amphipod crustacean Parhyale hawaiensis. *Developmental Biology*, *277*(2), 387–402.

Fatemi, M., Hermann, A., Gowher, H., & Jeltsch, A. (2002). Dnmt3a and Dnmt1 functionally cooperate during de novo methylation of DNA. *European Journal of Biochemistry*, *269*(20), 4981–4984.

Feng, Q., & Zhang, Y. (2001). The MeCP1 complex represses transcription through preferential binding, remodeling, and deacetylating methylated nucleosomes. *Genes & Development*, *15*(7), 827–832.

Feng, S., Cokus, S. J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M. G., Hetzel, J., Jain, J., Strauss, S. H., Halpern, M. E., Ukomadu, C., Sadler, K. C., Pradhan, S., Pellegrini, M., & Jacobsen, S. E. (2010). Conservation and divergence of methylation patterning in plants and animals. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(19), 8689–8694.

Ferg, M., Sanges, R., Gehrig, J., Kiss, J., Bauer, M., Lovas, A., Szabo, M., Yang, L., Straehle, U., Pankratz, M. J., Olasz, F., Stupka, E., & Müller, F. (2007). The TATA-binding protein regulates maternal mRNA degradation and differential zygotic transcription in zebrafish. *The EMBO Journal*, *26*(17), 3945–3956.

Ferrier, D. E., & Akam, M. (1996). Organization of the Hox gene cluster in the grasshopper, Schistocerca gregaria. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(23), 13024–13029.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., & Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research*, *42*(Database issue), D222-30.

Firmino, J., Carballo, C., Armesto, P., Campinho, M. A., Power, D. M., & Manchado, M. (2017). Phylogeny, expression patterns and regulation of DNA Methyltransferases in early development of the flatfish, Solea senegalensis. *BMC Developmental Biology*, *17*(1), 11.

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., Korlach, J., & Turner, S. W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods, 7*(6), 461–465.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics, 21*, 768–769.

Frei, R. E., Schultz, G. A., & Church, R. B. (1989). Qualitative and quantitative changes in protein synthesis occur at the 8-16-cell stage of embryogenesis in the cow. *Journal of Reproduction and Fertility*, *86*(2), 637–641.

Gao, L., Wu, K., Liu, Z., Yao, X., Yuan, S., Tao, W., Yi, L., Yu, G., Hou, Z., Fan, D., Tian, Y., Liu, J., Chen, Z.-J., & Liu, J. (2018). Chromatin Accessibility Landscape in Human Early Embryos and Its Association with Evolution. *Cell*, *173*(1), 248-259.e15.

Gao, S., Shauang, X., Gao, T., Ruixue, L., Xinyi, Z., Zhang, Y., & Zhang, K. (2022). Transcriptome Analysis Reveals the Role of Zelda in the Regulation of Embryonic and Wing Development of Tribolium Castaneum. *Available at SSRN*.

Gargioli, C., & Slack, J. M. W. (2004). Cell lineage tracing during Xenopus tail regeneration. *Development (Cambridge, England)*, *131*(11), 2669–2679.

Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D., & Bairoch, A. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Research*, *31*(13), 3784–3788.

Gatzmann, F., Falckenhayn, C., Gutekunst, J., Hanna, K., Raddatz, G., Carneiro, V. C., & Lyko, F. (2018). The methylome of the marbled crayfish links gene body methylation to stable expression of poorly accessible genes. *Epigenetics & Chromatin*, *11*(1), 57.

Gavis, E. R., & Lehmann, R. (1992). Localization of nanos RNA controls embryonic polarity. *Cell*, *71*(2), 301–313.

Gerberding, M., Browne, W. E., & Patel, N. H. (2002). Cell lineage analysis of the amphipod crustacean Parhyale hawaiensis reveals an early restriction of cell fates. *Development (Cambridge, England)*, *129*(24), 5789–5801.

Gil, O. D., Zanazzi, G., Struyk, A. F., & Salzer, J. L. (1998). Neurotrimin mediates bifunctional effects on neurite outgrowth via homophilic and heterophilic interactions. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *18*(22), 9312–9325.

Gildor, T., & Ben-Tabou de-Leon, S. (2015). Comparative Study of Regulatory Circuits in Two Sea Urchin Species Reveals Tight Control of Timing and High Conservation of Expression Dynamics. *PLOS Genetics*, *11*(7), 1–19.

Giraldez, A. J., Mishima, Y., Rihel, J., Grocock, R. J., Dongen, S. Van, Inoue, K., Enright, A. J., & Schier, A. F. (2006). Zebrafish MiR-430 Promotes Deadenylation and Clearance of Maternal mRNAs. *Science*, *312*(5770), 75–79.

Giribet, G., & Edgecombe, G. D. (2019). The Phylogeny and Evolutionary History of Arthropods. *Current Biology*, *29*(12), R592–R602.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews. Genetics*, *17*(6), 333–351.

Górnikiewicz, B., Ronowicz, A., Podolak, J., Madanecki, P., Stanisławska-Sachadyn, A., & Sachadyn, P. (2013). Epigenetic basis of regeneration: analysis of genomic DNA methylation profiles in the MRL/MpJ mouse. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, *20*(6), 605–621.

Graf, A., Krebs, S., Zakhartchenko, V., Schwalb, B., Blum, H., & Wolf, E. (2014). Fine mapping of genome activation in bovine embryos by RNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(11), 4139–4144.

Grafi, G., Zemach, A., & Pitto, L. (2007). Methyl-CpG-binding domain (MBD) proteins in plants. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, *1769*(5), 287–294.

Grillo, M., Konstantinides, N., & Averof, M. (2016). Old questions, new models: unraveling complex organ regeneration with new experimental approaches. *Current Opinion in Genetics & Development*, *40*, 23–31.

Grishkevich, V., & Yanai, I. (2014). Gene length and expression level shape genomic novelties. *Genome Research*, *24*(9), 1497–1503.

Guo, H., Zhu, P., Yan, L., Li, R., Hu, B., Lian, Y., Yan, J., Ren, X., Lin, S., Li, J., Jin, X., Shi, X., Liu, P., Wang, X., Wang, W., Wei, Y., Li, X., Guo, F., Wu, X., … Qiao, J. (2014). The DNA methylation landscape of human early embryos. *Nature*, *511*(7511), 606–610.

Guo, Y., Zhao, S., Sheng, Q., Guo, M., Lehmann, B., Pietenpol, J., Samuels, D. C., & Shyr, Y. (2015). RNAseq by Total RNA Library Identifies Additional RNAs Compared to Poly(A) RNA Library. *BioMed Research International*, *2015*, 862130.

Gupta, T., & Extavour, C. G. (2013). Identification of a putative germ plasm in the amphipod Parhyale hawaiensis. *EvoDevo*, *4*(1), 34.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England)*, *29*(8), 1072–1075.

Gutierrez, A., & Sommer, R. J. (2004). Evolution of dnmt-2 and mbd-2-like genes in the free-living nematodes Pristionchus pacificus, Caenorhabditis elegans and Caenorhabditis briggsae. *Nucleic Acids Research*, *32*(21), 6388–6396.

Gutierrez, A., & Sommer, R. J. (2007). Functional diversification of the nematode mbd2/3 gene between Pristionchus pacificus and Caenorhabditis elegans. *BMC Genetics*, *8*, 57.

Haberle, V., Li, N., Hadzhiev, Y., Plessy, C., Previti, C., Nepal, C., Gehrig, J., Dong, X., Akalin, A., Suzuki, A. M., van IJcken, W. F. J., Armant, O., Ferg, M., Strähle, U., Carninci, P., Müller, F., & Lenhard, B. (2014). Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature*, *507*(7492), 381–385.

Hackett, J. A., & Surani, M. A. (2013). DNA methylation dynamics during the mammalian life cycle. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1609), 20110328.

Hadzhiev, Y., Qureshi, H. K., Wheatley, L., Cooper, L., Jasiulewicz, A., Van Nguyen, H., Wragg, J. W., Poovathumkadavil, D., Conic, S., Bajan, S., Sik, A., Hutvàgner, G., Tora, L., Gambus, A., Fossey, J. S., & Müller, F. (2019). A cell cycle-coordinated Polymerase II transcription compartment encompasses gene expression before global genome activation. *Nature Communications*, *10*(1), 691.

Hai, T. W., Liu, F., Coukos, W. J., & Green, M. R. (1989). Transcription factor ATF cDNA clones: an extensive family of leucine zipper proteins able to selectively form DNA-binding heterodimers. *Genes & Development*, *3*(12B), 2083–2090.

Hainer, S. J., McCannell, K. N., Yu, J., Ee, L.-S., Zhu, L. J., Rando, O. J., & Fazzio, T. G. (2016). DNA methylation directs genomic localization of Mbd2 and Mbd3 in embryonic stem cells. *ELife*, *5*.

Hamatani, T., Carter, M. G., Sharov, A. A., & Ko, M. S. H. (2004). Dynamics of Global Gene Expression Changes during Mouse Preimplantation Development. *Developmental Cell*, *6*(1), 117–131.

Hampsey, M. (1998). Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiology and Molecular Biology Reviews : MMBR*, *62*(2), 465–503.

Hannibal, R. L., Price, A. L., & Patel, N. H. (2012). The functional relationship between ectodermal and mesodermal segmentation in the crustacean, Parhyale hawaiensis. *Developmental Biology*, *361*(2), 427–438. (a)

Hannibal, R. L., Price, A. L., Parchem, R. J., & Patel, N. H. (2012). Analysis of snail genes in the crustacean Parhyale hawaiensis: insight into snail gene family evolution. *Development Genes and Evolution*, *222*(3), 139–151. (b)

Harris, C. J., Scheibe, M., Wongpalee, S. P., Liu, W., Cornett, E. M., Vaughan, R. M., Li, X., Chen, W., Xue, Y., Zhong, Z., Yen, L., Barshop, W. D., Rayatpisheh, S., Gallego-Bartolome, J., Groth, M., Wang, Z., Wohlschlegel, J. A., Du, J., Rothbart, S. B., … Jacobsen, S. E. (2018). A DNA methylation reader complex that enhances gene transcription. *Science (New York, N.Y.)*, *362*(6419), 1182–1186.

Harrison, M. M., & Hamm, D. C. (n.d.). *Regulatory principles governing the maternal-to-zygotic transition: insights from Drosophila melanogaster*.

Harrison, M. M., Li, X.-Y., Kaplan, T., Botchan, M. R., & Eisen, M. B. (2011). Zelda Binding in the Early Drosophila melanogaster Embryo Marks Regions Subsequently Activated at the Maternal-to-Zygotic Transition. *PLOS Genetics*, *7*(10), 1–13.

Hartigan, J. A., & Wong, M. A. (1979). A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *28*(1), 100–108.

Harvey, S. A., Sealy, I., Kettleborough, R., Fenyes, F., White, R., Stemple, D., & Smith, J. C. (2013). Identification of the zebrafish maternal and paternal transcriptomes. *Development*, *140*(13), 2703–2710.

He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., Sun, Y., Li, X., Dai, Q., Song, C.-X., Zhang, K., He, C., & Xu, G.-L. (2011). Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science*, *333*(6047), 1303–1307.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8.

Héberlé, É., & Bardet, A. F. (2019). Sensitivity of transcription factors to DNA methylation. *Essays in Biochemistry*, *63*(6), 727–741.

Hellman, A., & Chess, A. (2007). Gene body-specific methylation on the active X chromosome. *Science (New York, N.Y.)*, *315*(5815), 1141–1143.

Hendrich, B., & Bird, A. (1998). Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Molecular and Cellular Biology*, *18*(11), 6538–6547.

Hendrich, B., Guy, J., Ramsahoye, B., Wilson, V. A., & Bird, A. (2001). Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development. *Genes & Development*, *15*(6), 710–723.

Hendrich, B., Abbott, C., McQueen, H., Chambers, D., Cross, S., & Bird, A. (1999). Genomic structure and chromosomal mapping of the murine and human Mbd1, Mbd2, Mbd3, and Mbd4 genes. *Mammalian Genome*, *10*(9), 906–912.

Hendrich, B., & Tweedie, S. (2003). The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends in Genetics: TIG*, *19*(5), 269–277.

Hendrickson, P. G., Doráis, J. A., Grow, E. J., Whiddon, J. L., Lim, J.-W., Wike, C. L., Weaver, B. D., Pflueger, C., Emery, B. R., Wilcox, A. L., Nix, D. A., Peterson, C. M., Tapscott, S. J., Carrell, D. T., & Cairns, B. R. (2017). Conserved roles of mouse DUX and human DUX4

inactivating cleavage-stage genes and MERVL/HERVL retrotransposons. *Nature Genetics*, *49*(6), 925–934.

Hernández de Madrid, B., & Casanova, J. (2018). GATA factor genes in the Drosophila midgut embryo. *PloS One*, *13*(3), e0193612.

Heyn, P., Kircher, M., Dahl, A., Kelso, J., Tomancak, P., Kalinka, A. T., & Neugebauer, K. M. (2014). The earliest transcribed zygotic genes are short, newly evolved, and different across species. *Cell Reports*, *6*(2), 285–292.

Holliday, R. (2006). Dual Inheritance. In W. Doerfler & P. Böhm (Eds.), *DNA Methylation: Basic Mechanisms* (pp. 243–256). Springer Berlin Heidelberg.

Holt, C. E., & Bullock, S. L. (2009). Subcellular mRNA Localization in Animal Cells and Why It Matters. *Science*, *326*(5957), 1212–1216.

Hon, G. C., Rajagopal, N., Shen, Y., McCleary, D. F., Yue, F., Dang, M. D., & Ren, B. (2013). Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nature Genetics*, *45*(10), 1198–1206.

Howell, C. Y., Bestor, T. H., Ding, F., Latham, K. E., Mertineit, C., Trasler, J. M., & Chaillet, J. R. (2001). Genomic imprinting disrupted by a maternal effect mutation in the Dnmt1 gene. *Cell*, *104*(6), 829–838.

Huff, J. T., & Zilberman, D. (2014). Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. *Cell*, *156*(6), 1286–1297.

Iliakis, G., Wang, H., Perrault, A. R., Boecker, W., Rosidi, B., Windhofer, F., Wu, W., Guan, J., Terzoudi, G., & Pantelias, G. (2004). Mechanisms of DNA double strand break repair and chromosome aberration formation. *Cytogenetic and Genome Research*, *104*(1–4), 14–20.

Ito, S., Shen, L., Dai, Q., Wu, S. C., Collins, L. B., Swenberg, J. A., He, C., & Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science (New York, N.Y.)*, *333*(6047), 1300–1303.

Iwano, H., Nakamura, M., & Tajima, S. (2004). Xenopus MBD3 plays a crucial role in an early stage of development. *Developmental Biology*, *268*(2), 416–428.

Jaber-Hijazi, F., Lo, P. J. K. P., Mihaylova, Y., Foster, J. M., Benner, J. S., Tejada Romero, B., Chen, C., Malla, S., Solana, J., Ruzov, A., & Aziz Aboobaker, A. (2013). Planarian MBD2/3 is required for adult stem cell pluripotency independently of DNA methylation. *Developmental Biology*, *384*(1), 141–153.

Jackson, M., Krassowska, A., Gilbert, N., Chevassut, T., Forrester, L., Ansell, J., & Ramsahoye, B. (2004). Severe global DNA hypomethylation blocks differentiation and induces histone

hyperacetylation in embryonic stem cells. *Molecular and Cellular Biology*, *24*(20), 8862–8871.

Jaenisch, R., & Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, *33*(3), 245–254.

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 239.

Jeltsch, A. (2006). Molecular enzymology of mammalian DNA methyltransferases. *Current Topics in Microbiology and Immunology*, *301*, 203–225.

Jeltsch, A. (2010). Phylogeny of Methylomes. *Science*, *328*(5980), 837–838.

Jeltsch, A., Ehrenhofer-Murray, A., Jurkowski, T. P., Lyko, F., Reuter, G., Ankri, S., Nellen, W., Schaefer, M., & Helm, M. (2017). Mechanism and biological role of Dnmt2 in Nucleic Acid Methylation. *RNA Biology*, *14*(9), 1108–1123.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)*, *337*(6096), 816–821.

Jjingo, D., Conley, A. B., Yi, S. V, Lunyak, V. V, & Jordan, I. K. (2012). On the presence and role of human gene-body DNA methylation. *Oncotarget*, *3*(4), 462–474.

Johnson, B. R., & Tsutsui, N. D. (2011). Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. *BMC Genomics*, *12*, 164.

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews. Genetics*, *13*(7), 484–492.

Joseph, S. R., Pálfy, M., Hilbert, L., Kumar, M., Karschau, J., Zaburdaev, V., Shevchenko, A., & Vastenhouw, N. L. (2017). Competition between histone and transcription factor binding regulates the onset of transcription in zebrafish embryos. *ELife*, *6*, e23326.

Jukam, D., Shariati, S. A. M., & Skotheim, J. M. (2017). Zygotic Genome Activation in Vertebrates. *Developmental Cell*, *42*(4), 316–332.

Jurkowska, R. Z., Jurkowski, T. P., & Jeltsch, A. (2011). Structure and function of mammalian DNA methyltransferases. *Chembiochem : A European Journal of Chemical Biology*, *12*(2), 206–222.

Jurkowski, T. P., & Jeltsch, A. (2011). Burning off DNA Methylation: New Evidence for Oxygen-Dependent DNA Demethylation. *ChemBioChem*, *12*(17), 2543–2545.

Kaczynski, J., Cook, T., & Urrutia, R. (2003). Sp1- and Krüppel-like transcription factors. *Genome Biology*, *4*(2), 206.

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. *Genome Research*, *20*(10), 1313–1326.

Kaji, K., Nichols, J., & Hendrich, B. (2007). Mbd3, a component of the NuRD co-repressor complex, is required for development of pluripotent cells. *Development (Cambridge, England)*, *134*(6), 1123–1132.

Kanno, M., Chalut, C., & Egly, J.-M. (1992). Genomic structure of the putative BTF3 transcription factor. *Gene*, *117*(2), 219–228.

Kao, D., Lai, A. G., Stamataki, E., Rosic, S., Konstantinides, N., Jarvis, E., Di Donfrancesco, A., Pouchkina-Stancheva, N., Sémon, M., Grillo, M., Bruce, H., Kumar, S., Siwanowicz, I., Le, A., Lemire, A., Eisen, M. B., Extavour, C., Browne, W. E., Wolff, C., … Aboobaker, A. (2016). The genome of the crustacean Parhyale hawaiensis, a model for animal development, regeneration, immunity and lignocellulose digestion. *ELife*, *5*.

Kao, D., Lai, A. G., Stamataki, E., Rosic, S., Konstantinides, N., Jarvis, E., Di Donfrancesco, A., Pouchkina-Stancheva, N., Sémon, M., Grillo, M., Bruce, H., Kumar, S., Siwanowicz, I., Le, A., Lemire, A., Eisen, M. B., Extavour, C., Browne, W. E., Wolff, C., … Aboobaker, A. (2016). The genome of the crustacean *Parhyale hawaiensis*, a model for animal development, regeneration, immunity and lignocellulose digestion. *ELife*, *5*, e20062.

Kao, D., Lai, A. G., Stamataki, E., Rosic, S., Konstantinides, N., Jarvis, E., Donfrancesco, A. Di, Pouchkina-Stancheva, N., Sémon, M., Grillo, M., Bruce, H., Kumar, S., Siwanowicz, I., Le, A., Lemire, A., Eisen, M. B., Extavour, C., Browne, W. E., Wolff, C., … Aboobaker, A. (2016). The genome of the crustacean Parhyale hawaiensis, a model for animal development, regeneration, immunity and lignocellulose digestion. *ELife*, *5*.

Katsuyama, T., & Paro, R. (2011). Epigenetic reprogramming during tissue regeneration. *FEBS Letters*, *585*(11), 1617–1624.

Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R., & Bosch, T. C. G. (2009). More than just orphans: are taxonomically-restricted genes important in evolution? *Trends in Genetics : TIG*, *25*(9), 404–413.

Khan, P., Linkhart, B., & Simon, H.-G. (2002). Different regulation of T-box genes Tbx4 and Tbx5 during limb development and limb regeneration. *Developmental Biology*, *250*(2), 383–392.

Kiebler, M. A., Hemraj, I., Verkade, P., Köhrmann, M., Fortes, P., Marión, R. M., Ortín, J., & Dotti, C. G. (1999). The mammalian staufen protein localizes to the somatodendritic domain of cultured hippocampal neurons: implications for its involvement in mRNA transport. *The*

Journal of Neuroscience: The Official Journal of the Society for Neuroscience, *19*(1), 288–297.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360.

Konstantinides, N., & Averof, M. (2014). A common cellular basis for muscle regeneration in arthropods and vertebrates. *Science (New York, N.Y.)*, *343*(6172), 788–791.

Kontarakis, Z., & Pavlopoulos, A. (2014). Transgenesis in non-model organisms: the case of Parhyale. *Methods in Molecular Biology (Clifton, N.J.)*, *1196*, 145–181.

Kontarakis, Z., Pavlopoulos, A., Kiupakis, A., Konstantinides, N., Douris, V., & Averof, M. (2011). A versatile strategy for gene trapping and trap conversion in emerging model organisms. *Development*, *138*(12), 2625–2630.

Kosugi, S., Hirakawa, H., & Tabata, S. (2015). GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. *Bioinformatics (Oxford, England)*, *31*(23), 3733–3741.

Kovaka, S., Zimin, A. V, Pertea, G. M., Razaghi, R., Salzberg, S. L., & Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, *20*(1), 278.

Kucharski, R., Maleszka, J., Foret, S., & Maleszka, R. (2008). Nutritional control of reproductive status in honeybees via DNA methylation. *Science (New York, N.Y.)*, *319*(5871), 1827–1830.

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, *33*(7), 1870–1874.

Kwasnieski, J. C., Orr-Weaver, T. L., & Bartel, D. P. (2019). Early genome activation in Drosophila is extensive with an initial tendency for aborted transcripts and retained introns. *Genome Research*, *29*(7), 1188–1197.

Lai, A. Y., & Wade, P. A. (2011). Cancer biology and NuRD: a multifaceted chromatin remodelling complex. *Nature Reviews. Cancer*, *11*(8), 588–596.

Lamond, A. I. (1993). The spliceosome. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *15*(9), 595–603.

Latham, K. E., Garrels, J. I., Chang, C., & Solter, D. (1991). Quantitative analysis of protein synthesis in mouse embryos. I. Extensive reprogramming at the one- and two-cell stages. *Development (Cambridge, England)*, *112*(4), 921–932.

Laver, J. D., Li, X., Ray, D., Cook, K. B., Hahn, N. A., Nabeel-Shah, S., Kekis, M., Luo, H., Marsolais, A. J., Fung, K. Y. Y., Hughes, T. R., Westwood, J. T., Sidhu, S. S., Morris, Q.,

Lipshitz, H. D., & Smibert, C. A. (2015). Brain tumor is a sequence-specific RNA-binding protein that directs maternal mRNA clearance during the Drosophila maternal-to-zygotic transition. *Genome Biology*, *16*(1), 94. (a)

Laver, J. D., Marsolais, A. J., Smibert, C. A., & Lipshitz, H. D. (2015). Chapter Two - Regulation and Function of Maternal Gene Products During the Maternal-to-Zygotic Transition in Drosophila. In H. D. B. T.-C. T. in D. B. Lipshitz (Ed.), *The Maternal-to-Zygotic Transition* (Vol. 113, pp. 43–84). Academic Press. (b)

Law, J. A., & Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews. Genetics*, *11*(3), 204–220.

Le Guezennec, X., Vermeulen, M., Brinkman, A. B., Hoeijmakers, W. A. M., Cohen, A., Lasonder, E., & Stunnenberg, H. G. (2006). MBD2/NuRD and MBD3/NuRD, two distinct complexes with different biochemical and functional properties. *Molecular and Cellular Biology*, *26*(3), 843–851.

Lécuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T. R., Tomancak, P., & Krause, H. M. (2007). Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*, *131*(1), 174–187.

Lee, H., Gurtowski, J., Yoo, S., Marcus, S., McCombie, W. R., & Schatz, M. (2014). Error correction and assembly complexity of single molecule sequencing reads. *BioRxiv*.

Lee, M. T., Bonneau, A. R., & Giraldez, A. J. (2014). Zygotic genome activation during the maternal-to-zygotic transition. *Annual Review of Cell and Developmental Biology*, *30*, 581–613.

Lee, M. T., Bonneau, A. R., Takacs, C. M., Bazzini, A. A., DiVito, K. R., Fleming, E. S., & Giraldez, A. J. (2013). Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition. *Nature*, *503*(7476), 360–364.

Lee, M., Choi, K.-H., Oh, J.-N., Kim, S.-H., Lee, D.-K., Choe, G. C., Jeong, J., & Lee, C.-K. (2021). SOX2 plays a crucial role in cell proliferation and lineage segregation during porcine pre-implantation embryo development. *Cell Proliferation*, *54*(8), e13097.

Lee, T., Zhai, J., & Meyers, B. C. (2010). Conservation and divergence in eukaryotic DNA methylation. *Proceedings of the National Academy of Sciences*, *107*(20), 9027–9028.

LeGendre, J. B., Campbell, Z. T., Kroll-Conner, P., Anderson, P., Kimble, J., & Wickens, M. (2013). RNA targets and specificity of Staufen, a double-stranded RNA-binding protein in Caenorhabditis elegans. *The Journal of Biological Chemistry*, *288*(4), 2532–2545.

Leighton, G., & Williams, D. C. J. (2019). The Methyl-CpG-Binding Domain 2 and 3 Proteins and Formation of the Nucleosome Remodeling and Deacetylase Complex. *Journal of Molecular Biology*.

Leung, B. (2021). *Evolution and development of spinal cord stem cells and cell type diversity*. University of Oxford.

Lewis, S. H., Ross, L., Bain, S. A., Pahita, E., Smith, S. A., Cordaux, R., Miska, E. A., Lenhard, B., Jiggins, F. M., & Sarkies, P. (2020). Widespread conservation and lineage-specific diversification of genome-wide DNA methylation patterns across arthropods. *PLOS Genetics*, *16*(6), 1–24.

Li, C., Fan, Y., Li, G., Xu, X., Duan, J., Li, R., Kang, X., Ma, X., Chen, X., Ke, Y., Yan, J., Lian, Y., Liu, P., Zhao, Y., Zhao, H., Chen, Y., Yu, Y., & Liu, J. (2018). DNA methylation reprogramming of functional elements during mammalian embryonic development. *Cell Discovery*, *4*(1), 41.

Li, E., Bestor, T. H., & Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell*, *69*(6), 915–926.

Li, H., Janssens, J., De Waegeneer, M., Kolluru, S. S., Davie, K., Gardeux, V., Saelens, W., David, F. P. A., Brbić, M., Spanier, K., Leskovec, J., McLaughlin, C. N., Xie, Q., Jones, R. C., Brueckner, K., Shim, J., Tattikota, S. G., Schnorrer, F., Rust, K., … Zinzen, R. P. (2022). Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. *Science (New York, N.Y.)*, *375*(6584), eabk2432.

Li, L., Lu, X., & Dean, J. (2013). The Maternal to Zygotic Transition in Mammals. *Molecular Aspects of Medicine, 34*(5), 919-938.

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, *22*(13), 1658–1659.

Li, Y., Mei, N.-H., Cheng, G.-P., Yang, J., & Zhou, L.-Q. (2021). Inhibition of DRP1 Impedes Zygotic Genome Activation and Preimplantation Development in Mice. *Frontiers in Cell and Developmental Biology*, *9*, 788512.

Liang, H.-L., Nien, C.-Y., Liu, H.-Y., Metzstein, M. M., Kirov, N., & Rushlow, C. (2008). The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. *Nature*, *456*(7220), 400–403.

Liao, J., Karnik, R., Gu, H., Ziller, M. J., Clement, K., Tsankov, A. M., Akopian, V., Gifford, C. A., Donaghey, J., Galonska, C., Pop, R., Reyon, D., Tsai, S. Q., Mallard, W., Joung, J. K., Rinn, J. L., Gnirke, A., & Meissner, A. (2015). Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. *Nature Genetics*, *47*(5), 469–478.

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, *30*(7), 923–930.

Lindeman, D. (1991). Natural history of the terrestrial amphipod Cerrorchestia hyloraina Lindeman (Crustacea: Amphipoda; Talitridae) in a Costa Rican cloud forest. *Journal of Natural History*, *25*(3), 623–638.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., & Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, *462*(7271), 315–322.

Liu, G., Wang, W., Hu, S., Wang, X., & Zhang, Y. (2018). Inherited DNA methylation primes the establishment of accessible chromatin during genome activation. *Genome Research*, *28*(7), 998–1007.

Liu, H., Wei, Z., Dominguez, A., Li, Y., Wang, X., & Qi, L. S. (2015). CRISPR-ERA: a comprehensive design tool for CRISPR-mediated gene editing, repression and activation. *Bioinformatics (Oxford, England)*, *31*(22), 3676–3678.

Liu, Y., Wu, K., Shao, F., Nie, H., Zhang, J., Li, C., Hou, Z., Wang, J., Zhou, B., Zhao, H., & Lu, F. (2021). Dynamics of poly(A) tail length and non-A residues during the human oocyte-to-embryo transition. *BioRxiv*.

Liubicich, D. M., Serano, J. M., Pavlopoulos, A., Kontarakis, Z., Protas, M. E., Kwan, E., Chatterjee, S., Tran, K. D., Averof, M., & Patel, N. H. (2009). Knockdown of Parhyale Ultrabithorax recapitulates evolutionary changes in crustacean appendage morphology. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(33), 13892–13896.

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, *21*(10), 597–614.

Long, M., Betrán, E., Thornton, K., & Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nature Reviews. Genetics*, *4*(11), 865–875.

Lott, S. E., Villalta, J. E., Schroth, G. P., Luo, S., Tonkin, L. A., & Eisen, M. B. (2011). Noncanonical Compensation of Zygotic X Transcription in Early Drosophila melanogaster Development Revealed through Single-Embryo RNA-Seq. *PLOS Biology*, *9*(2), 1–13.

Lou, S., Lee, H.-M., Qin, H., Li, J.-W., Gao, Z., Liu, X., Chan, L. L., Kl Lam, V., So, W.-Y., Wang, Y., Lok, S., Wang, J., Ma, R. C., Tsui, S. K.-W., Chan, J. C., Chan, T.-F., & Yip, K. Y. (2014). Whole-genome bisulfite sequencing of multiple individuals reveals complementary roles of promoter and gene body methylation in transcriptional regulation. *Genome Biology*, *15*(7), 408.

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), 550.

Ludikhuize, M. C., & Rodríguez Colman, M. J. (2021). Metabolic Regulation of Stem Cells and Differentiation: A Forkhead Box O Transcription Factor Perspective. *Antioxidants & Redox Signaling*, *34*(13), 1004–1024.

Lund, E., Liu, M., Hartley, R. S., Sheets, M. D., & Dahlberg, J. E. (2009). Deadenylation of maternal mRNAs mediated by miR-427 in Xenopus laevis embryos. *RNA (New York, N.Y.)*, *15*(12), 2351–2363.

Lunyak, V. V, & Rosenfeld, M. G. (2008). Epigenetic regulation of stem cell fate. *Human Molecular Genetics*, *17*(R1), R28-36.

Luo, J., Lyu, M., Chen, R., Zhang, X., Luo, H., & Yan, C. (2019). SLR: a scaffolding algorithm based on long reads and contig classification. *BMC Bioinformatics*, *20*(1), 539.

Lyko, F., Foret, S., Kucharski, R., Wolf, S., Falckenhayn, C., & Maleszka, R. (2010). The honeybee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biology*, *8*(11), e1000506.

Mao, Z., Bozzella, M., Seluanov, A., & Gorbunova, V. (2008). DNA repair by nonhomologous end joining and homologous recombination during cell cycle in human cells. *Cell Cycle (Georgetown, Tex.)*, *7*(18), 2902–2906.

Marhold, J., Brehm, A., & Kramer, K. (2004). The Drosophila methyl-DNA binding protein MBD2/3 interacts with the NuRD complex via p55 and MI-2. *BMC Molecular Biology*, *5*(1), 20. (a)

Marhold, J., Kramer, K., Kremmer, E., & Lyko, F. (2004). The Drosophila MBD2/3 protein mediates interactions between the MI-2 chromatin complex and CpT/A-methylated DNA. *Development*, *131*(24), 6033–6039. (b)

Martin, A., Serano, J. M., Jarvis, E., Bruce, H. S., Wang, J., Ray, S., Barker, C. A., O'Connell, L. C., & Patel, N. H. (2016). CRISPR/Cas9 Mutagenesis Reveals Versatile Roles of Hox Genes in Crustacean Limb Specification and Evolution. *Current Biology: CB*, *26*(1), 14–26.

Materna, S. C., Nam, J., & Davidson, E. H. (2010). High accuracy, high-resolution prevalence measurement for the majority of locally expressed regulatory genes in early sea urchin development. *Gene Expression Patterns*, *10*(4), 177–184.

Mathavan, S., Lee, S. G. P., Mak, A., Miller, L. D., Murthy, K. R. K., Govindarajan, K. R., Tong, Y., Wu, Y. L., Lam, S. H., Yang, H., Ruan, Y., Korzh, V., Gong, Z., Liu, E. T., & Lufkin, T. (2005). Transcriptome Analysis of Zebrafish Embryogenesis Using Microarrays. *PLOS Genetics*, *1*(2), null.

Matsumoto, K., Meric, F., & Wolffe, A. P. (1996). Translational repression dependent on the interaction of the Xenopus Y-box protein FRGY2 with mRNA. Role of the cold shock

domain, tail domain, and selective RNA sequence recognition. *The Journal of Biological Chemistry*, *271*(37), 22706–22712.

Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., D'Souza, C., Fouse, S. D., Johnson, B. E., Hong, C., Nielsen, C., Zhao, Y., Turecki, G., Delaney, A., Varhol, R., Thiessen, N., Shchors, K., Heine, V. M., Rowitch, D. H., Xing, X., Fiore, C., … Costello, J. F. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, *466*(7303), 253–257.

McGinnis, S., & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, *32*(Web Server issue), W20-5.

McLysaght, A., & Guerzoni, D. (2015). New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *370*(1678), 20140332.

Meissner, A. (2010). Epigenetic modifications in pluripotent and differentiated cells. *Nature Biotechnology*, *28*(10), 1079–1088.

Menafra, R., & Stunnenberg, H. G. (2014). MBD2 and MBD3: elusive functions and mechanisms. *Frontiers in Genetics*, *5*, 428.

Mendelowitz, L., & Pop, M. (2014). Computational methods for optical mapping. *GigaScience*, *3*(1), 33.

Mendizabal, I., Shi, L., Keller, T. E., Konopka, G., Preuss, T. M., Hsieh, T.-F., Hu, E., Zhang, Z., Su, B., & Yi, S. V. (2016). Comparative Methylome Analyses Identify Epigenetic Regulatory Loci of Human Brain Evolution. *Molecular Biology and Evolution*, *33*(11), 2947–2959.

Mishima, Y., & Tomari, Y. (2016). Codon Usage and 3' UTR Length Determine Maternal mRNA Stability in Zebrafish. *Molecular Cell*, *61*(6), 874–885.

Morrison, J. I., Lööf, S., He, P., & Simon, A. (2006). Salamander limb regeneration involves the activation of a multipotent skeletal muscle satellite cell population. *Journal of Cell Biology*, *172*(3), 433–440.

Muneoka, K., & Bryant, S. V. (1982). Evidence that patterning mechanisms in developing and regenerating limbs are the same. *Nature*, *298*(5872), 369–371.

Myers, A. A. (1985). Shallow-water, coral reef and mangrove Amphipoda (Gammaridea) of Fiji. *Records of The Australian Museum, Supplement*, *5*, 1–143.

Nacu, E., & Tanaka, E. M. (2011). Limb regeneration: a new development? *Annual Review of Cell and Developmental Biology*, *27*, 409–440.

Nagamori, I., Kobayashi, H., Shiromoto, Y., Nishimura, T., Kuramochi-Miyagawa, S., Kono, T., & Nakano, T. (2015). Comprehensive DNA Methylation Analysis of Retrotransposons in Male Germ Cells. *Cell Reports*, *12*(10), 1541–1547.

Nast, A. R., & Extavour, C. G. (2014). Ablation of a single cell from eight-cell embryos of the amphipod crustacean Parhyale hawaiensis. *Journal of Visualized Experiments: JoVE*, *85*.

Nestorov, P., Battke, F., Levesque, M. P., & Gerberding, M. (2013). The Maternal Transcriptome of the Crustacean Parhyale hawaiensis Is Inherited Asymmetrically to Invariant Cell Lineages of the Ectoderm and Mesoderm. *PLOS ONE*, *8*(2), 1–14.

Ng, H. H., Zhang, Y., Hendrich, B., Johnson, C. A., Turner, B. M., Erdjument-Bromage, H., Tempst, P., Reinberg, D., & Bird, A. (1999). MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nature Genetics*, *23*(1), 58–61.

Ni, P., Zhong, Z., Xu, J., Huang, N., Zhang, J., Nie, F., Zhao, H., Zou, Y., Huang, Y., Li, J., Xiao, C.-L., Luo, F., & Wang, J. (2023). DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. *BioRxiv*.

Noort, R. J., Christopher, G. A., & Esseltine, J. L. (2021). Pannexin 1 Influences Lineage Specification of Human iPSCs. *Frontiers in Cell and Developmental Biology*, *9*.

Nüsslein-Volhard, C., Frohnhöfer, H. G., & Lehmann, R. (1987). Determination of anteroposterior polarity in Drosophila. *Science (New York, N.Y.)*, *238*(4834), 1675–1681.

O. Pourquié. (2009). *Hox genes*. Elsevier Academic Press.

Okano, M., Bell, D. W., Haber, D. A., & Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, *99*(3), 247–257.

Okitsu, C. Y., & Hsieh, C.-L. (2007). DNA methylation dictates histone H3K4 methylation. *Molecular and Cellular Biology*, *27*(7), 2746–2757.

Ou, S., Liu, J., Chougule, K. M., Fungtammasan, A., Seetharam, A. S., Stein, J. C., Llaca, V., Manchanda, N., Gilbert, A. M., Wei, S., Chin, C.-S., Hufnagel, D. E., Pedersen, S., Snodgrass, S. J., Fengler, K., Woodhouse, M., Walenz, B. P., Koren, S., Phillippy, A. M., … Ware, D. (2020). Effect of sequence depth and length in long-read assembly of the maize inbred NC358. *Nature Communications*, *11*(1), 2288.

Owens, N. D. L., Blitz, I. L., Lane, M. A., Patrushev, I., Overton, J. D., Gilchrist, M. J., Cho, K. W. Y., & Khokha, M. K. (2016). Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in Development. *Cell Reports*, *14*(3), 632–647.

Ozhan-Kizil, G., Havemann, J., & Gerberding, M. (2009). Germ cells in the crustacean Parhyale hawaiensis depend on Vasa protein for their maintenance but not for their formation. *Developmental Biology*, *327*(1), 230–239.

Pace, R. M., Grbić, M., & Nagy, L. M. (2016). Composition and genomic organization of arthropod Hox clusters. *EvoDevo*, *7*, 11.

Pálfy, M., Joseph, S. R., & Vastenhouw, N. L. (2017). The timing of zygotic genome activation. *Current Opinion in Genetics & Development*, *43*, 53–60.

Papatsenko, D., & Levine, M. (2011). The Drosophila gap gene network is composed of two parallel toggle switches. *PloS One*, *6*(7), e21145.

Parchem, R. J., Poulin, F., Stuart, A. B., Amemiya, C. T., & Patel, N. H. (2010). BAC library for the amphipod crustacean, Parhyale hawaiensis. *Genomics*, *95*(5), 261–267.

Paris, M., Wolff, C., Patel, N. H., & Averof, M. (2022). The crustacean model Parhyale hawaiensis. *Current Topics in Developmental Biology*, *147*, 199–230.

Passmore, L. A., & Coller, J. (2022). Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression. *Nature Reviews. Molecular Cell Biology*, *23*(2), 93–106.

Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., Fan, L., Sandelin, A., Rinn, J. L., Regev, A., & Schier, A. F. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Research*, *22*(3), 577–591.

Paulsen, M., & Ferguson-Smith, A. C. (2001). DNA methylation in genomic imprinting, development, and disease. *The Journal of Pathology*, *195*(1), 97–110.

Pavlopoulos, A., & Averof, M. (2005). Establishing genetic transformation for comparative developmental studies in the crustacean Parhyale hawaiensis. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(22), 7888–7893.

Pavlopoulos, A., Kontarakis, Z., Liubicich, D. M., Serano, J. M., Akam, M., Patel, N. H., & Averof, M. (2009). Probing the evolution of appendage specialization by Hox gene misexpression in an emerging model crustacean. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(33), 13897–13902.

Pavlopoulos, A., & Wolff, C. (2020). Crustacean limb morphogenesis during normal development and regeneration. *Developmental Biology and Larval Ecology: The Natural History of the Crustacea, Volume 7*, 45.

Pavlopoulos, A., & Wolff, C. (2020). 46C2Crustacean Limb Morphogenesis during Normal Development and Regeneration. In *Developmental Biology and Larval Ecology: The Natural History of the Crustacea, Volume 7*. Oxford University Press.

Peat, J. R., Dean, W., Clark, S. J., Krueger, F., Smallwood, S. A., Ficz, G., Kim, J. K., Marioni, J. C., Hore, T. A., & Reik, W. (2014). Genome-wide bisulfite sequencing in zygotes identifies demethylation targets and maps the contribution of TET3 oxidation. *Cell Reports*, *9*(6), 1990–2000.

Piro, V. C., Faoro, H., Weiss, V. A., Steffens, M. B. R., Pedrosa, F. O., Souza, E. M., & Raittz, R. T. (2014). FGAP: an automated gap closing tool. *BMC Research Notes*, *7*, 371.

Planques, A., Kerner, P., Ferry, L., Grunau, C., Gazave, E., & Vervoort, M. (2021). DNA methylation atlas and machinery in the developing and regenerating annelid Platynereis dumerilii. *BMC Biology*, *19*(1), 148.

Ponger, L., & Li, W.-H. (2005). Evolutionary Diversification of DNA Methyltransferases in Eukaryotic Genomes. *Molecular Biology and Evolution*, *22*(4), 1119–1128.

Poovachiranon, S., Boto, K., & Duke, N. (1986). Food preference studies and ingestion rate measurements of the mangrove amphipod Parhyale hawaiensis (Dana). *Journal of Experimental Marine Biology and Ecology*, *98*(1), 129–140.

Poovachiranon~, S., Boto, K., & Duke, N. (1986). FOOD PREFERENCE STUDIES AND INGESTION RATE MEASUREMENTS OF THE MANGROVE AMPHIPOD PARHYALE HA WAZENSZS (Dana)'. In *Mar. Biol. Ecol* (Vol. 98, pp. 129–140).

Powell, C., Grant, A. R., Cornblath, E., & Goldman, D. (2013). Analysis of DNA methylation reveals a partial reprogramming of the Müller glia genome during retina regeneration. *Proceedings of the National Academy of Sciences*, *110*(49), 19814–19819.

Powers, T. P., Hogan, J., Ke, Z., Dymbrowski, K., Wang, X., Collins, F. H., & Kaufman, T. C. (2000). Characterization of the Hox cluster from the mosquito Anopheles gambiae (Diptera: Culicidae). *Evolution & Development*, *2*(6), 311–325.

Price, A. L., Modrell, M. S., Hannibal, R. L., & Patel, N. H. (2010). Mesoderm and ectoderm lineages in the crustacean Parhyale hawaiensis display intra-germ layer compensation. *Developmental Biology*, *341*(1), 256–266.

Price, A. L., & Patel, N. H. (2008). Investigating divergent mechanisms of mesoderm development in arthropods: the expression of Ph-twist and Ph-mef2 in Parhyale hawaiensis. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *310B* (1), 24–40.

Putnam, N. H., O'Connell, B. L., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R., Troll, C. J., Fields, A., Hartley, P. D., Sugnet, C. W., Haussler, D., Rokhsar, D. S., & Green, R. E. (2016). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Research*, *26*(3), 342–350.

Qin, M., Wu, S., Li, A., Zhao, F., Feng, H., Ding, L., & Ruan, J. (2019). LRScaf: improving draft genomes using long noisy reads. *BMC Genomics*, *20*(1), 955.

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

Rai, K., Jafri, I. F., Chidester, S., James, S. R., Karpf, A. R., Cairns, B. R., & Jones, D. A. (2010). Dnmt3 and G9a cooperate for tissue-specific development in zebrafish. *The Journal of Biological Chemistry*, *285*(6), 4110–4121.

Rai, K., Nadauld, L. D., Chidester, S., Manos, E. J., James, S. R., Karpf, A. R., Cairns, B. R., & Jones, D. A. (2006). Zebra fish Dnmt1 and Suv39h1 regulate organ-specific terminal differentiation during development. *Molecular and Cellular Biology*, *26*(19), 7077–7085.

Ram, P. T., & Schultz, R. M. (1993). Reporter gene expression in G2 of the 1-cell mouse embryo. *Developmental Biology*, *156*(2), 552–556.

Ramos, A. P., Gustafsson, O., Labert, N., Salecker, I., Nilsson, D.-E., & Averof, M. (2019). Analysis of the genetically tractable crustacean Parhyale hawaiensis reveals the organisation of a sensory system for low-resolution vision. *BMC Biology*, *17*(1), 67.

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., & Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nature Protocols*, *8*(11), 2281–2308.

Rauluseviciute, I., Drabløs, F., & Rye, M. B. (2020). DNA hypermethylation associated with upregulated gene expression in prostate cancer demonstrates the diversity of epigenetic regulation. *BMC Medical Genomics*, *13*(1), 6.

Rehm, E. J., Hannibal, R. L., Chaw, R. C., Vargas-Vila, M. A., & Patel, N. H. (2009). The crustacean Parhyale hawaiensis: a new model for arthropod development. *Cold Spring Harbor Protocols*, *2009*(1), pdb.emo114.

Rehm, E. J., Hannibal, R. L., Chaw, R. C., Vargas-Vila, M. A., & Patel, N. H. (2009). In situ hybridization of labeled RNA probes to fixed Parhyale hawaiensis embryos. *Cold Spring Harbor Protocols*, *2009*(1), pdb.prot5130. (a)

Rehm, E. J., Hannibal, R. L., Chaw, R. C., Vargas-Vila, M. A., & Patel, N. H. (2009). Fixation and dissection of Parhyale hawaiensis embryos. *Cold Spring Harbor Protocols*, *2009*(1), pdb.prot5127. (b)

Rehm, E. J., Hannibal, R. L., Chaw, R. C., Vargas-Vila, M. A., & Patel, N. H. (2009). Injection of Parhyale hawaiensis blastomeres with fluorescently labeled tracers. *Cold Spring Harbor Protocols*, *2009*(1), pdb.prot5128. (c)

Reynolds, S., Wilson, C., Austin, J., & Hooper, L. (2012). Effects of psychotherapy for anxiety in children and adolescents: a meta-analytic review. *Clinical Psychology Review*, *32*(4), 251–262.

Ribeiro, L., Tobias-Santos, V., Santos, D., Antunes, F., Feltran, G., de Souza Menezes, J., Aravind, L., Venancio, T. M., & da Fonseca, R. (2017). Evolution and multiple roles of the Pancrustacea specific transcription factor zelda in insects. *PLOS Genetics*, *13*(7), 1–25.

Riviere, G., Wu, G.-C., Fellous, A., Goux, D., Sourdaine, P., & Favrel, P. (2013). DNA methylation is crucial for the early development in the Oyster C. gigas. *Marine Biotechnology (New York, N.Y.)*, *15*(6), 739–753.

Roberts, R. J., Carneiro, M. O., & Schatz, M. C. (2013). The advantages of SMRT sequencing. In *Genome biology* (Vol. 14, Issue 7, p. 405).

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. In *Nature biotechnology* (Vol. 29, Issue 1, pp. 24–26).

Roder, K., Hung, M. S., Lee, T. L., Lin, T. Y., Xiao, H., Isobe, K. I., Juang, J. L., & Shen, C. J. (2000). Transcriptional repression by Drosophila methyl-CpG-binding proteins. *Molecular and Cellular Biology*, *20*(19), 7401–7409.

Roloff, T. C., Ropers, H. H., & Nuber, U. A. (2003). Comparative study of methyl-CpG-binding domain proteins. *BMC Genomics*, *4*(1), 1.

Rošić, S., Amouroux, R., Requena, C. E., Gomes, A., Emperle, M., Beltran, T., Rane, J. K., Linnett, S., Selkirk, M. E., Schiffer, P. H., Bancroft, A. J., Grencis, R. K., Jeltsch, A., Hajkova, P., & Sarkies, P. (2018). Evolutionary analysis indicates that DNA alkylation damage is a byproduct of cytosine DNA methyltransferase activity. *Nature Genetics*, *50*(3), 452–459.

Rothbart, S. B., Krajewski, K., Nady, N., Tempel, W., Xue, S., Badeaux, A. I., Barsyte-Lovejoy, D., Martinez, J. Y., Bedford, M. T., Fuchs, S. M., Arrowsmith, C. H., & Strahl, B. D. (2012). Association of UHRF1 with methylated H3K9 directs the maintenance of DNA methylation. *Nature Structural & Molecular Biology*, *19*(11), 1155–1160.

Rothe, M., Pehl, M., Taubert, H., & Jäckle, H. (1992). Loss of gene function through rapid mitotic cycles in the Drosophila embryo. *Nature*, *359*(6391), 156–159.

Saito, M., & Ishikawa, F. (2002). The mCpG-binding domain of human MBD3 does not bind to mCpG but interacts with NuRD/Mi2 components HDAC1 and MTA2. *The Journal of Biological Chemistry*, *277*(38), 35434–35439.

Sallés, F. J., Lieberfarb, M. E., Wreden, C., Gergen, J. P., & Strickland, S. (1994). Coordinate Initiation of *Drosophila* Development by Regulated Polyadenylation of Maternal Messenger RNAs. *Science*, *266*(5193), 1996–1999.

Sander, J. D., Maeder, M. L., Reyon, D., Voytas, D. F., Joung, J. K., & Dobbs, D. (2010). ZiFiT (Zinc Finger Targeter): an updated zinc finger engineering tool. *Nucleic Acids Research*, *38*(Web Server issue), W462-8.

Sander, J. D., Zaback, P., Joung, J. K., Voytas, D. F., & Dobbs, D. (2007). Zinc Finger Targeter (ZiFiT): an engineered zinc finger/target site design tool. *Nucleic Acids Research*, *35*(suppl_2), W599–W605.

Sandler, J. E., Irizarry, J., Stepanik, V., Dunipace, L., Amrhein, H., & Stathopoulos, A. (2018). A Developmental Program Truncates Long Transcripts to Temporally Regulate Cell Signaling. *Developmental Cell*, *47*(6), 773-784.e6.

Sandler, J. E., & Stathopoulos, A. (2016). Quantitative Single-Embryo Profile of Drosophila Genome Activation and the Dorsal–Ventral Patterning Network. *Genetics*, *202*(4), 1575–1584.

Santos, F., Hendrich, B., Reik, W., & Dean, W. (2002). Dynamic reprogramming of DNA methylation in the early mouse embryo. *Developmental Biology*, *241*(1), 172–182.

Sarda, S., Zeng, J., Hunt, B. G., & Yi, S. V. (2012). The Evolution of Invertebrate Gene Body Methylation. *Molecular Biology and Evolution*, *29*(8), 1907–1916.

Sarkies, P. (2022). Encyclopaedia of eukaryotic DNA methylation: from patterns to mechanisms and functions. *Biochemical Society Transactions*, *50*(3), 1179–1190.

Satoh, A., Gardiner, D. M., Bryant, S. V, & Endo, T. (2007). Nerve-induced ectopic limb blastemas in the axolotl are equivalent to amputation-induced blastemas. *Developmental Biology*, *312*(1), 231–244.

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., & Cardona, A. (2012). Fiji: an open-source platform for biological-image analysis. *Nature Methods*, *9*(7), 676–682.

Schulz, K. N., & Harrison, M. M. (2019). Mechanisms regulating zygotic genome activation. *Nature Reviews. Genetics*, *20*(4), 221–234.

Schulz, N. K. E., Wagner, C. I., Ebeling, J., Raddatz, G., Diddens-de Buhr, M. F., Lyko, F., & Kurtz, J. (2018). Dnmt1 has an essential function despite the absence of CpG DNA methylation in the red flour beetle Tribolium castaneum. *Scientific Reports*, *8*(1), 16462.

Schüpbach, T., & Wieschaus, E. (1986). Maternal-effect mutations altering the anterior-posterior pattern of the Drosophila embryo. *Roux's Archives of Developmental Biology*, *195*(5), 302–317.

Sciamanna, I., Vitullo, P., Curatolo, A., & Spadafora, C. (2011). A reverse transcriptase-dependent mechanism is essential for murine preimplantation development. *Genes*, *2*(2), 360–373.

Semotok, J. L., Cooperstock, R. L., Pinder, B. D., Vari, H. K., Lipshitz, H. D., & Smibert, C. A. (2005). Smaug Recruits the CCR4/POP2/NOT Deadenylase Complex to Trigger Maternal Transcript Localization in the Early Drosophila Embryo. *Current Biology*, *15*(4), 284–294.

Sen, G. L., Reuter, J. A., Webster, D. E., Zhu, L., & Khavari, P. A. (2010). DNMT1 maintains progenitor function in self-renewing somatic tissue. *Nature*, *463*(7280), 563–567.

Serano, J. M., Martin, A., Liubicich, D. M., Jarvis, E., Bruce, H. S., La, K., Browne, W. E., Grimwood, J., & Patel, N. H. (2016). Comprehensive analysis of Hox gene expression in the amphipod crustacean Parhyale hawaiensis. *Developmental Biology*, *409*(1), 297–309.

Seydoux, G., & Fire, A. (1994). Soma-germline asymmetry in the distributions of embryonic RNAs in Caenorhabditis elegans. *Development (Cambridge, England)*, *120*(10), 2823–2834.

Seydoux, G., Mello, C. C., Pettitt, J., Wood, W. B., Priess, J. R., & Fire, A. (1996). Repression of gene expression in the embryonic germ lineage of C. elegans. *Nature*, *382*(6593), 713–716.

Sha, Q.-Q., Zhang, J., & Fan, H.-Y. (2019). 579-590 National Key Research and Developmental Program of China (2017YFC1001500, 2016YFC1000600), National Natural Science Foundation of China (31671558, 31890781), and The Key Research and Development Program of Zhejiang Province. *Biology of Reproduction*, *101*(3).

Sha, Q.-Q., Zhang, J., & Fan, H.-Y. (2019). A story of birth and death: mRNA translation and clearance at the onset of maternal-to-zygotic transition in mammals†. *Biology of Reproduction*, *101*(3), 579–590.

Shao, S., Cao, H., Wang, Z., Zhou, D., Wu, C., Wang, S., Xia, D., & Zhang, D. (2020). CHD4/NuRD complex regulates complement gene expression and correlates with CD8 T cell infiltration in human hepatocellular carcinoma. *Clinical Epigenetics*, *12*(1), 31.

Shatkin, A. J., & Manley, J. L. (2000). The ends of the affair: Capping and polyadenylation. *Nature Structural Biology*, *7*(10), 838–842.

Shen, L., Inoue, A., He, J., Liu, Y., Lu, F., & Zhang, Y. (2014). Tet3 and DNA replication mediate demethylation of both the maternal and paternal genomes in mouse zygotes. *Cell Stem Cell*, *15*(4), 459–471.

Shermoen, A. W., & O'Farrell, P. H. (1991). Progression of the cell cycle through mitosis leads to abortion of nascent transcripts. *Cell*, *67*(2), 303–310.

Shindo, Y., & Amodeo, A. A. (2019). Dynamics of Free and Chromatin-Bound Histone H3 during Early Embryogenesis. *Current Biology*, *29*(2), 359-366.

Shoemaker, C. R. (1956). “Observations on the Amphipod Genus Parhyale.” *Proceedings of the United States National Museum*, *106*((3372)), 345–358.

Sinigaglia, C., Almazán, A., Lebel, M., Sémon, M., Gillet, B., Hughes, S., Edsinger, E., Averof, M., & Paris, M. (2022). Distinct gene expression dynamics in developing and regenerating crustacean limbs. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(27), e2119297119.

Smith, Z. D., Chan, M. M., Mikkelsen, T. S., Gu, H., Gnirke, A., Regev, A., & Meissner, A. (2012). A unique regulatory phase of DNA methylation in the early mammalian embryo. *Nature*, *484*(7394), 339–344.

Srinageshwar, B., Maiti, P., Dunbar, G. L., & Rossignol, J. (2016). Role of Epigenetics in Stem Cell Proliferation and Differentiation: Implications for Treating Neurodegenerative Diseases. *International Journal of Molecular Sciences*, *17*(2).

St Johnston, D., Beuchle, D., & Nüsslein-Volhard, C. (1991). Staufen, a gene required to localize maternal RNAs in the Drosophila egg. *Cell*, *66*(1), 51–63.

Stamataki, E., & Pavlopoulos, A. (2016). Non-insect crustacean models in developmental genetics including an encomium to Parhyale hawaiensis. *Current Opinion in Genetics & Development*, *39*, 149–156.

Stancheva, I., Hensey, C., & Meehan, R. R. (2001). Loss of the maintenance methyltransferase, xDnmt1, induces apoptosis in Xenopus embryos. *The EMBO Journal*, *20*(8), 1963–1973.

Su, J., Shao, X., Liu, H., Liu, S., Wu, Q., & Zhang, Y. (2012). Genome-wide dynamic changes of DNA methylation of repetitive elements in human embryonic stem cells and fetal fibroblasts. *Genomics*, *99*(1), 10–17.

Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H., & Bartel, D. P. (2014). Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature*, *508*(7494), 66–71.

Sun, D. A., Bredeson, J. V, Bruce, H. S., & Patel, N. H. (2022). Identification and classification of cis-regulatory elements in the amphipod crustacean Parhyale hawaiensis. *Development (Cambridge, England)*, *149*(11).

Sun, D. A., & Patel, N. H. (2019). The amphipod crustacean Parhyale hawaiensis: An emerging comparative model of arthropod development, evolution, and regeneration. *Developmental Biology*, *8*(5), 1759-7684.

Sun, J., Yan, L., Shen, W., & Meng, A. (2018). Maternal Ybx1 safeguards zebrafish oocyte maturation and maternal-to-zygotic transition by repressing global translation. *Development*, *145*(19).

Susor, A., Jansova, D., Cerna, R., Danylevska, A., Anger, M., Toralova, T., Malik, R., Supolikova, J., Cook, M. S., Oh, J. S., & Kubelka, M. (2015). Temporal and spatial regulation of translation in the mammalian oocyte via the mTOR-eIF4F pathway. *Nature Communications*, *6*(6078).

Suter, B., Romberg, L. M., & Steward, R. (1989). Bicaudal-D, a Drosophila gene involved in developmental asymmetry: localized transcript accumulation in ovaries and sequence similarity to myosin heavy chain tail domains. *Genes & Development*, *3*(12A), 1957–1968.

Suzuki, H., Kanchiku, T., Imajo, Y., Yoshida, Y., Nishida, N., Gondo, T., Yoshii, S., & Taguchi, T. (2015). Artificial collagen-filament scaffold promotes axon regeneration and long tract reconstruction in a rat model of spinal cord transection. *Medical Molecular Morphology*, *48*(4), 214–224.

Suzuki, M. M., & Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, *9*(6), 465–476.

Suzuki, M. M., Kerr, A. R. W., De Sousa, D., & Bird, A. (2007). CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Research*, *17*(5), 625–631. (b)

Suzuki, S., Ono, R., Narita, T., Pask, A. J., Shaw, G., Wang, C., Kohda, T., Alsop, A. E., Marshall Graves, J. A., Kohara, Y., Ishino, F., Renfree, M. B., & Kaneko-Ishino, T. (2007). Retrotransposon silencing by DNA methylation can drive mammalian genomic imprinting. *PLoS Genetics*, *3*(4), e55. (b)

Svilar, D., Goellner, E. M., Almeida, K. H., & Sobol, R. W. (2011). Base excision repair and lesion-dependent subpathways for repair of oxidative DNA damage. *Antioxidants & Redox Signaling*, *14*(12), 2491–2507.

Tadros, W., Goldman, A. L., Babak, T., Menzies, F., Vardy, L., Orr-Weaver, T., Hughes, T. R., Westwood, J. T., Smibert, C. A., & Lipshitz, H. D. (2007). SMAUG Is a Major Regulator of Maternal mRNA Destabilization in Drosophila and Its Translation Is Activated by the PAN GU Kinase. *Developmental Cell*, *12*(1), 143–155.

Tadros, W., & Lipshitz, H. D. (2009). The maternal-to-zygotic transition: a play in two acts. *Development (Cambridge, England)*, *136*(18), 3033–3042.

Tan, M. H., Au, K. F., Yablonovitch, A. L., Wills, A. E., Chuang, J., Baker, J. C., Wong, W. H., & Li, J. B. (2013). RNA sequencing reveals a diverse and dynamic repertoire of the Xenopus tropicalis transcriptome over development. *Genome Research*, *23*(1), 201–216.

Tanaka, H. V., Ng, N. C. Y., Yang Yu, Z., Casco-Robles, M. M., Maruo, F., Tsonis, P. A., & Chiba, C. (2016). A developmentally regulated switch from stem cells to dedifferentiation for limb muscle regeneration in newts. *Nature Communications*, *7*(1), 11069.

Tautz, D., & Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews. Genetics*, *12*(10), 692–702.

Thomsen, S., Anders, S., Janga, S. C., Huber, W., & Alonso, C. R. (2010). Genome-wide analysis of mRNA decay patterns during early Drosophiladevelopment. *Genome Biology*, *11*(9), R93.

Thomson, J. P., Skene, P. J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A. R. W., Deaton, A., Andrews, R., James, K. D., Turner, D. J., Illingworth, R., & Bird, A. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature*, *464*(7291), 1082–1086.

Thorp, J. H., & Rogers, D. C. (2011). Chapter 17 - Aquatic Sow Bugs, Scuds, and Opossum Shrimp: Subphylum Crustacea, Class Malacostraca, Superorder Peracarida. In J. H. Thorp & D. C. Rogers (Eds.), *Field Guide to Freshwater Invertebrates of North America* (pp. 147–156). Academic Press.

Thorp, J., & Rogers, D. C. (2011). *Introduction to Freshwater Invertebrates in the Phylum Arthropoda* (pp. 109–119).

Tomazou, E. M., & Meissner, A. (2010). Epigenetic regulation of pluripotency. *Advances in Experimental Medicine and Biology*, *695*, 26–40.

Tost, J. (2009). DNA methylation: an introduction to biology and the disease-associated changes of a promising biomarker. *Methods in Molecular Biology (Clifton, N.J.)*, *507*, 3–20.

Tryselius, Y., & Hultmark, D. (1997). Cysteine proteinase 1 (CP1), a cathepsin L-like enzyme expressed in the Drosophila melanogaster haemocyte cell line mbn-2. *Insect Molecular Biology*, *6*(2), 173–181.

Tse, O. Y. O., Jiang, P., Cheng, S. H., Peng, W., Shang, H., Wong, J., Chan, S. L., Poon, L. C. Y., Leung, T. Y., Chan, K. C. A., Chiu, R. W. K., & Lo, Y. M. D. (2021). Genome-wide detection of cytosine methylation by single molecule real-time sequencing. *Proceedings of the National Academy of Sciences*, *118*(5), e2019768118.

Tserevelakis, G. J., Velentza, S., Liaskas, I., Archontidis, T., Pavlopoulos, A., & Zacharakis, G. (2022). Imaging Parhyale hawaiensis embryogenesis with frequency domain photoacoustic microscopy: A novel tool in developmental biology. In *Journal of biophotonics* (Vol. 15, Issue 12, p. e202200202).

Tsumura, A., Hayakawa, T., Kumaki, Y., Takebayashi, S., Sakaue, M., Matsuoka, C., Shimotohno, K., Ishikawa, F., Li, E., Ueda, H. R., Nakayama, J., & Okano, M. (2006). Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. *Genes to Cells: Devoted to Molecular & Cellular Mechanisms*, *11*(7), 805–814.

Tweedie, S., Ng, H. H., Barlow, A. L., Turner, B. M., Hendrich, B., & Bird, A. (1999). Vestiges of a DNA methylation system in Drosophila melanogaster? In *Nature genetics* (Vol. 23, Issue 4, pp. 389–390).

Tyurin-Kuzmin, P. A., Molchanov, A. Y., Chechekhin, V. I., Ivanova, A. M., & Kulebyakin, K. Y. (2020). Metabolic Regulation of Mammalian Stem Cell Differentiation. *Biochemistry. Biokhimiia*, *85*(3), 264–278.

UniProt Consortium, T. (2018). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, *46*(5), 2699.

Vargas-Vila, M. A., Hannibal, R. L., Parchem, R. J., Liu, P. Z., & Patel, N. H. (2010). A prominent requirement for single-minded and the ventral midline in patterning the dorsoventral axis of the crustacean Parhyale hawaiensis. *Development (Cambridge, England)*, *137*(20), 3469–3476.

Vastenhouw, N. L., Cao, W. X., & Lipshitz, H. D. (2019). The maternal-to-zygotic transition revisited. *Development (Cambridge, England)*, *146*(11).

Ventós-Alfonso, A., Ylla, G., Montañes, J.-C., & Belles, X. (2020). DNMT1 Promotes Genome Methylation and Early Embryo Development in Cockroaches. *IScience*, *23*(12), 101778.

Waddington, C. H. (2014). *The Strategy of the Genes*. Taylor & Francis.

Wade, P. A. (2001). Methyl CpG-binding proteins and transcriptional repression. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *23*(12), 1131–1137.

Wade, P. A., Gegonne, A., Jones, P. L., Ballestar, E., Aubry, F., & Wolffe, A. P. (1999). Mi-2 complex couples DNA methylation to chromatin remodelling and histone deacetylation. *Nature Genetics*, *23*(1), 62–66.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One*, *9*(11), e112963.

Wang, L., Zhang, J., Duan, J., Gao, X., Zhu, W., Lu, X., Yang, L., Zhang, J., Li, G., Ci, W., Li, W., Zhou, Q., Aluru, N., Tang, F., He, C., Huang, X., & Liu, J. (2014). Programming and inheritance of parental DNA methylomes in mammals. *Cell*, *157*(4), 979–991.

Wang, M., Ly, M., Lugowski, A., Laver, J. D., Lipshitz, H. D., Smibert, C. A., & Rissland, O. S. (2017). ME31B globally represses maternal mRNAs by two distinct mechanisms during the \textit{Drosophila} maternal-to-zygotic transition. *ELife*, *6*, e27891.

Wang, Q. T., Piotrowska, K., Ciemerych, M. A., Milenkovic, L., Scott, M. P., Davis, R. W., & Zernicka-Goetz, M. (2004). A Genome-Wide Study of Gene Activity Reveals Developmental

Signaling Pathways in the Preimplantation Mouse Embryo. *Developmental Cell*, *6*(1), 133–144.

Wang, X., Fang, X., Yang, P., Jiang, X., Jiang, F., Zhao, D., Li, B., Cui, F., Wei, J., Ma, C., Wang, Y., He, J., Luo, Y., Wang, Z., Guo, X., Guo, W., Wang, X., Zhang, Y., Yang, M., … Kang, L. (2014). The locust genome provides insight into swarm formation and long-distance flight. *Nature Communications*, *5*, 2957.

Wang, X., Xiong, X., Cao, W., Zhang, C., Werren, J. H., & Wang, X. (2019). Genome Assembly of the A-Group Wolbachia in Nasonia oneida Using Linked-Reads Technology. *Genome Biology and Evolution*, *11*(10), 3008–3013.

Wang, Y., Jorda, M., Jones, P. L., Maleszka, R., Ling, X., Robertson, H. M., Mizzen, C. A., Peinado, M. A., & Robinson, G. E. (2006). Functional CpG methylation system in a social insect. *Science (New York, N.Y.)*, *314*(5799), 645–647.

Wang, Y.-H., Tsai, D.-Y., Ko, Y.-A., Yang, T.-T., Lin, I.-Y., Hung, K.-H., & Lin, K.-I. (2019). Blimp-1 Contributes to the Development and Function of Regulatory B Cells. *Frontiers in Immunology*, *10*, 1909.

Warren RL. (2016). RAILS and Cobbler: scaffolding and automated finishing of draft genomes using long DNA sequences. *J Open Source Softw*, *1*(7), 116.

Warren, R. L., Yang, C., Vandervalk, B. P., Behsaz, B., Lagman, A., Jones, S. J. M., & Birol, I. (2015). LINKS: Scalable, alignment-free scaffolding of draft genomes with long reads. *GigaScience*, *4*, 35.

Waters, T. R., & Swann, P. F. (1998). Kinetics of the action of thymine DNA glycosylase. *The Journal of Biological Chemistry*, *273*(32), 20007–20014.

Watson, M., & Warr, A. (2019). Errors in long-read assemblies can critically affect protein prediction. *Nature Biotechnology*, *37*(2), 124–126.

Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Pääbo, S., Rebhan, M., & Schübeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics*, *39*(4), 457–466.

Wei, Z., Angerer, R. C., & Angerer, L. M. (2006). A database of mRNA expression patterns for the sea urchin embryo. *Developmental Biology*, *300*(1), 476–484.

Weill, L., Belloc, E., Bava, F.-A., & Méndez, R. (2012). Translational control by changes in poly(A) tail length: recycling mRNAs. *Nature Structural & Molecular Biology*, *19*(6), 577–585.

Weisman, C. M., Murray, A. W., & Eddy, S. R. (2020). Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biology*, *18*(11), e3000862.

Weisman, C. M., Murray, A. W., & Eddy, S. R. (2022). Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Current Biology : CB*, *32*(12), 2632-2639.

White, M. D., Bissiere, S., Alvarez, Y. D., & Plachta, N. (2016). Mouse Embryo Compaction. *Current Topics in Developmental Biology*, *120*, 235–258.

Winata, C. L., Łapiński, M., Pryszcz, L., Vaz, C., bin Ismail, M. H., Nama, S., Hajan, H. S., Lee, S. G. P., Korzh, V., Sampath, P., Tanavde, V., & Mathavan, S. (2018). Cytoplasmic polyadenylation-mediated translational control of maternal mRNAs directs maternal-to-zygotic transition. *Development*, *145*(1).

Wolf, T., Droste, J., Gren, T., Ortseifen, V., Schneiker-Bekel, S., Zemke, T., Pühler, A., & Kalinowski, J. (2017). The MalR type regulator AcrC is a transcriptional repressor of acarbose biosynthetic genes in Actinoplanes sp. SE50/110. *BMC Genomics*, *18*(1), 562.

Wolff, C., Tinevez, J.-Y., Pietzsch, T., Stamataki, E., Harich, B., Guignard, L., Preibisch, S., Shorte, S., Keller, P. J., Tomancak, P., & Pavlopoulos, A. (2018). Multi-view light-sheet imaging and tracking with the MaMuT software reveals the cell lineage of a direct developing arthropod limb. *ELife*, *7*, e34410.

Workman, J. L., & Kingston, R. E. (1998). ALTERATION OF NUCLEOSOME STRUCTURE AS A MECHANISM OF TRANSCRIPTIONAL REGULATION. *Annual Review of Biochemistry*, *67*(1), 545–579.

Wu, H., Coskun, V., Tao, J., Xie, W., Ge, W., Yoshikawa, K., Li, E., Zhang, Y., & Sun, Y. E. (2010). Dnmt3a-dependent nonpromoter DNA methylation facilitates transcription of neurogenic genes. *Science (New York, N.Y.)*, *329*(5990), 444–448.

Wu, T. D., & Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, *21*(9), 1859–1875.

Xiang, H., Zhu, J., Chen, Q., Dai, F., Li, X., Li, M., Zhang, H., Zhang, G., Li, D., Dong, Y., Zhao, L., Lin, Y., Cheng, D., Yu, J., Sun, J., Zhou, X., Ma, K., He, Y., Zhao, Y., … Wang, J. (2010). Single base–resolution methylome of the silkworm reveals a sparse epigenomic map. *Nature Biotechnology*, *28*(5), 516–520.

Xu, G.-C., Xu, T.-J., Zhu, R., Zhang, Y., Li, S.-Q., Wang, H.-W., & Li, J.-T. (2019). LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *GigaScience*, *8*(1).

Xu, L., Mao, A., Liu, H., Gui, B., Choy, K. W., Huang, H., Yu, Q., Zhang, X., Chen, M., Lin, N., Chen, L., Han, J., Wang, Y., Zhang, M., Li, X., He, D., Lin, Y., Zhang, J., Cram, D. S., & Cao, H. (2020). Long-Molecule Sequencing: A New Approach for Identification of Clinically Significant DNA Variants in α-Thalassemia and β-Thalassemia Carriers. *The Journal of Molecular Diagnostics: JMD*, *22*(8), 1087–1095.

Xu, X., Li, G., Li, C., Zhang, J., Wang, Q., Simmons, D. K., Chen, X., Wijesena, N., Zhu, W., Wang, Z., Wang, Z., Ju, B., Ci, W., Lu, X., Yu, D., Wang, Q.-F., Aluru, N., Oliveri, P., Zhang, Y. E., … Liu, J. (2019). Evolutionary transition between invertebrates and vertebrates via methylation reprogramming in embryogenesis. *National Science Review*, *6*(5), 993–1003.

Xue, Y., Wong, J., Moreno, G. T., Young, M. K., Côté, J., & Wang, W. (1998). NURD, a novel complex with both ATP-dependent chromatin-remodeling and histone deacetylase activities. *Molecular Cell*, *2*(6), 851–861.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., Huang, J., Li, M., Wu, X., Wen, L., Lao, K., Li, R., Qiao, J., & Tang, F. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structural & Molecular Biology*, *20*(9), 1131–1139.

Yang, X., Han, H., De Carvalho, D. D., Lay, F. D., Jones, P. A., & Liang, G. (2014). Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell*, *26*(4), 577–590.

Yang, Z., & Huang, J. (2011). De novo origin of new genes with introns in Plasmodium vivax. *FEBS Letters*, *585*(4), 641–644.

Yartseva, V., & Giraldez, A. J. (2015). Chapter Six - The Maternal-to-Zygotic Transition During Vertebrate Development: A Model for Reprogramming. In H. D. Lipshitz (Ed.), *The Maternal-to-Zygotic Transition* (Vol. 113, pp. 191–232). Academic Press.

Ye, C., Hill, C. M., Wu, S., Ruan, J., & Ma, Z. (Sam). (2016). DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. *Scientific Reports*, *6*(1), 31900.

Yoder, J. A., Walsh, C. P., & Bestor, T. H. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics: TIG*, *13*(8), 335–340.

Yuan, Y., Bayer, P. E., Batley, J., & Edwards, D. (2017). Improvements in Genomic Technologies: Application to Crop Genomics. *Trends in Biotechnology*, *35*(6), 547–558.

Zemach, A., McDaniel, I. E., Silva, P., & Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science (New York, N.Y.)*, *328*(5980), 916–919.

Zemach, A., & Zilberman, D. (2010). Evolution of eukaryotic DNA methylation and the pursuit of safer sex. *Current Biology: CB*, *20*(17), R780-5.

Zenk, F., Zhan, Y., Kos, P., Löser, E., Atinbayeva, N., Schächtle, M., Tiana, G., Giorgetti, L., & Iovino, N. (2021). HP1 drives de novo 3D genome reorganization in early Drosophila embryos. *Nature*, *593*(7858), 289–293.

Zeremski, M., Stricker, J. R., Fischer, D., Zusman, S. B., & Cohen, D. (2003). Histone deacetylase dHDAC4 is involved in segmentation of the Drosophila embryo and is regulated by gap and pair-rule genes. *Genesis*, *35*(1), 31–38.

Zhang, G., Fang, X., Guo, X., Li, L., Luo, R., Xu, F., Yang, P., Zhang, L., Wang, X., Qi, H., Xiong, Z., Que, H., Xie, Y., Holland, P. W. H., Paps, J., Zhu, Y., Wu, F., Chen, Y., Wang, J., … Wang, J. (2012). The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, *490*(7418), 49–54.

Zhang, J., Zhang, Y.-L., Zhao, L.-W., Guo, J.-X., Yu, J.-L., Ji, S.-Y., Cao, L.-R., Zhang, S.-Y., Shen, L., Ou, X.-H., & Fan, H.-Y. (2019). Mammalian nucleolar protein DCAF13 is essential for ovarian follicle maintenance and oocyte growth by mediating rRNA processing. *Cell Death & Differentiation*, *26*, 1251–1266.

Zhang, X., & Jacobs, D. (2022). A Broad Survey of Gene Body and Repeat Methylation in Cnidaria Reveals a Complex Evolutionary History. *Genome Biology and Evolution*, *14*(2).

Zhang, Y., Ng, H. H., Erdjument-Bromage, H., Tempst, P., Bird, A., & Reinberg, D. (1999). Analysis of the NuRD subunits reveals a histone deacetylase core complex and a connection with DNA methylation. *Genes & Development*, *13*(15), 1924–1935.

Zhang, Y., Yin, C., Zhang, T., Li, F., Yang, W., Kaminski, R., Fagan, P. R., Putatunda, R., Young, W.-B., Khalili, K., & Hu, W. (2015). CRISPR/gRNA-directed synergistic activation mediator (SAM) induces specific, persistent and robust reactivation of the HIV-1 latent reservoirs. *Scientific Reports*, *5*(1), 16277.

Zhou, K., Huang, B., Zou, M., Lu, D., He, S., & Wang, G. (2015). Genome-wide identification of lineage-specific genes within Caenorhabditis elegans. *Genomics*, *106*(4), 242–248.

Zimin, A. V, Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, *29*(21), 2669–2677.

Zimin, A. V, & Salzberg, S. L. (2022). The SAMBA tool uses long reads to improve the contiguity of genome assemblies. *PLOS Computational Biology*, *18*(2), 1–11.

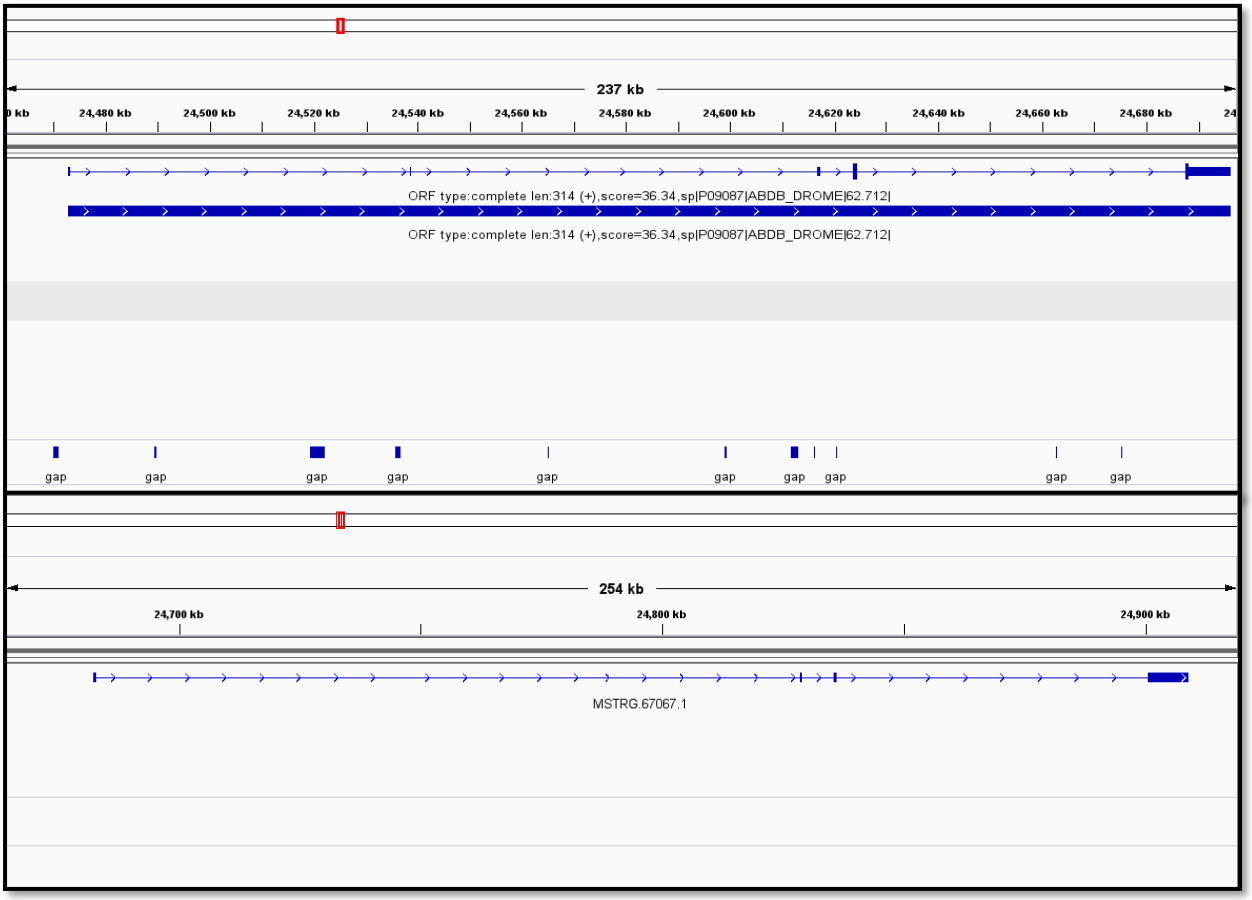
Zwier, M. V, Verhulst, E. C., Zwahlen, R. D., Beukeboom, L. W., & van de Zande, L. (2012). DNA methylation plays a crucial role during early Nasonia development. *Insect Molecular Biology*, *21*(1), 129–138.

Appendix

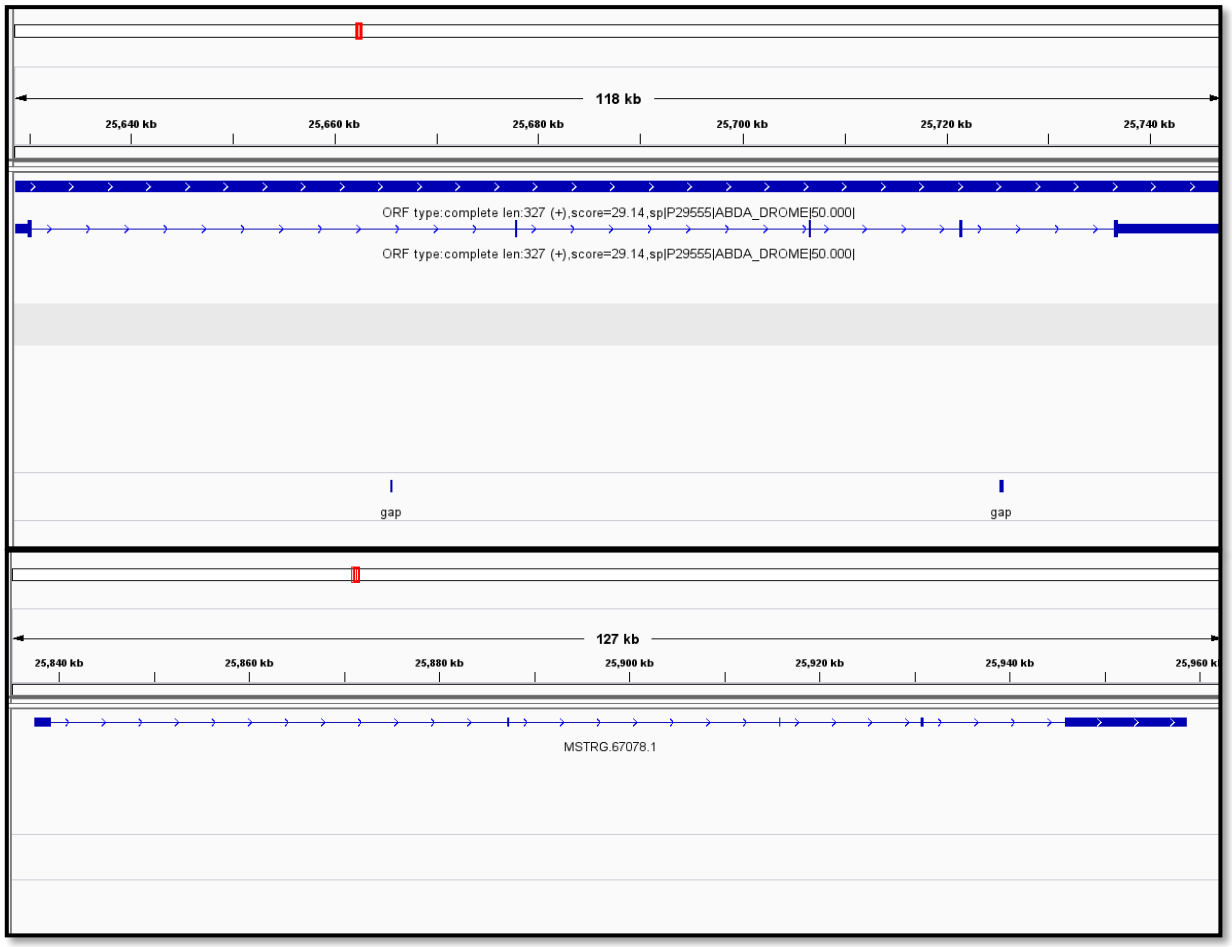
Appendix A. IGV screenshot of each Hox gene structure with indicated gaps in the gene.

The top panel shows *Phaw_5.0* version and bottom panel shows *Phaw_5.1*.

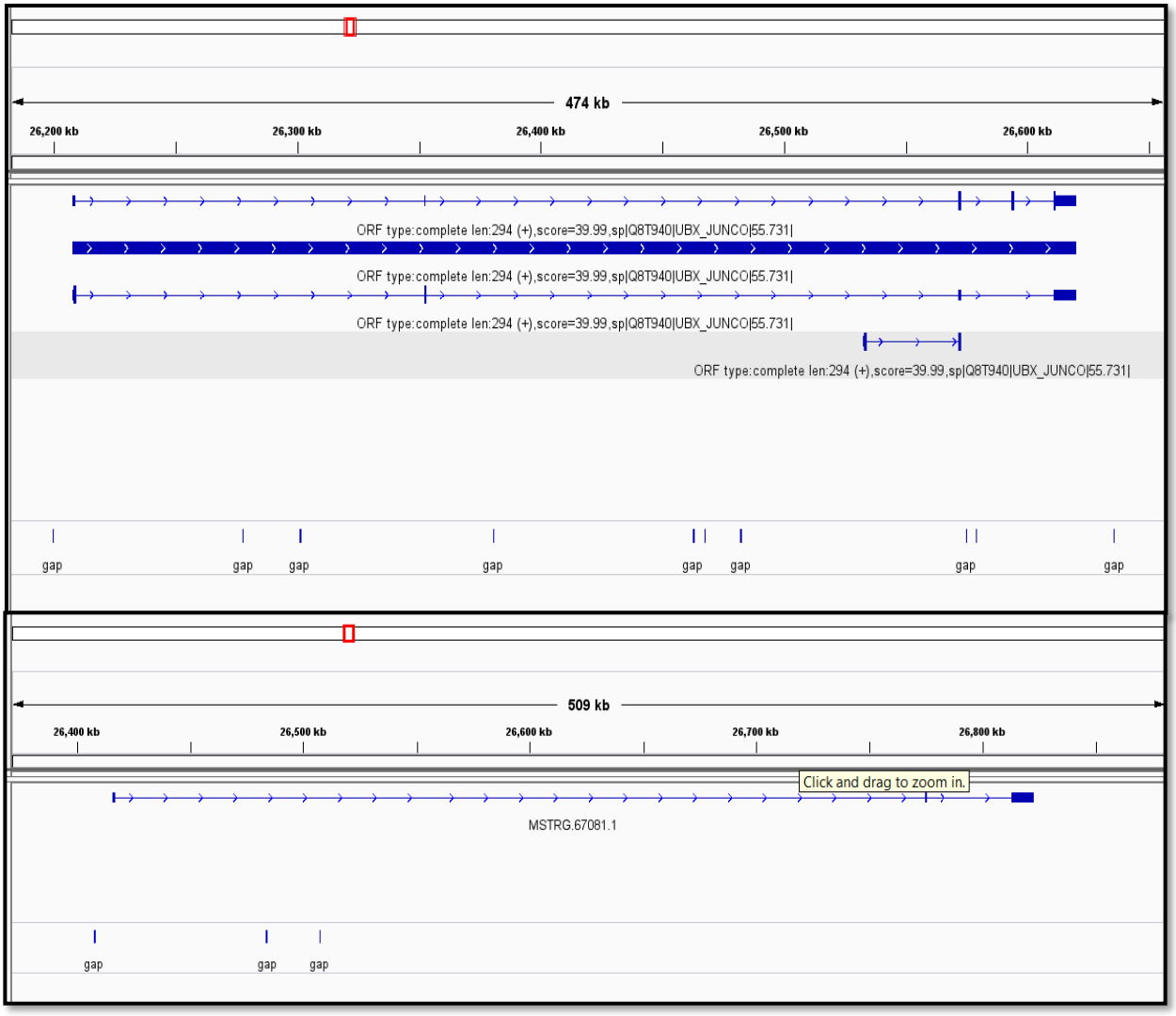
Abdominal-B



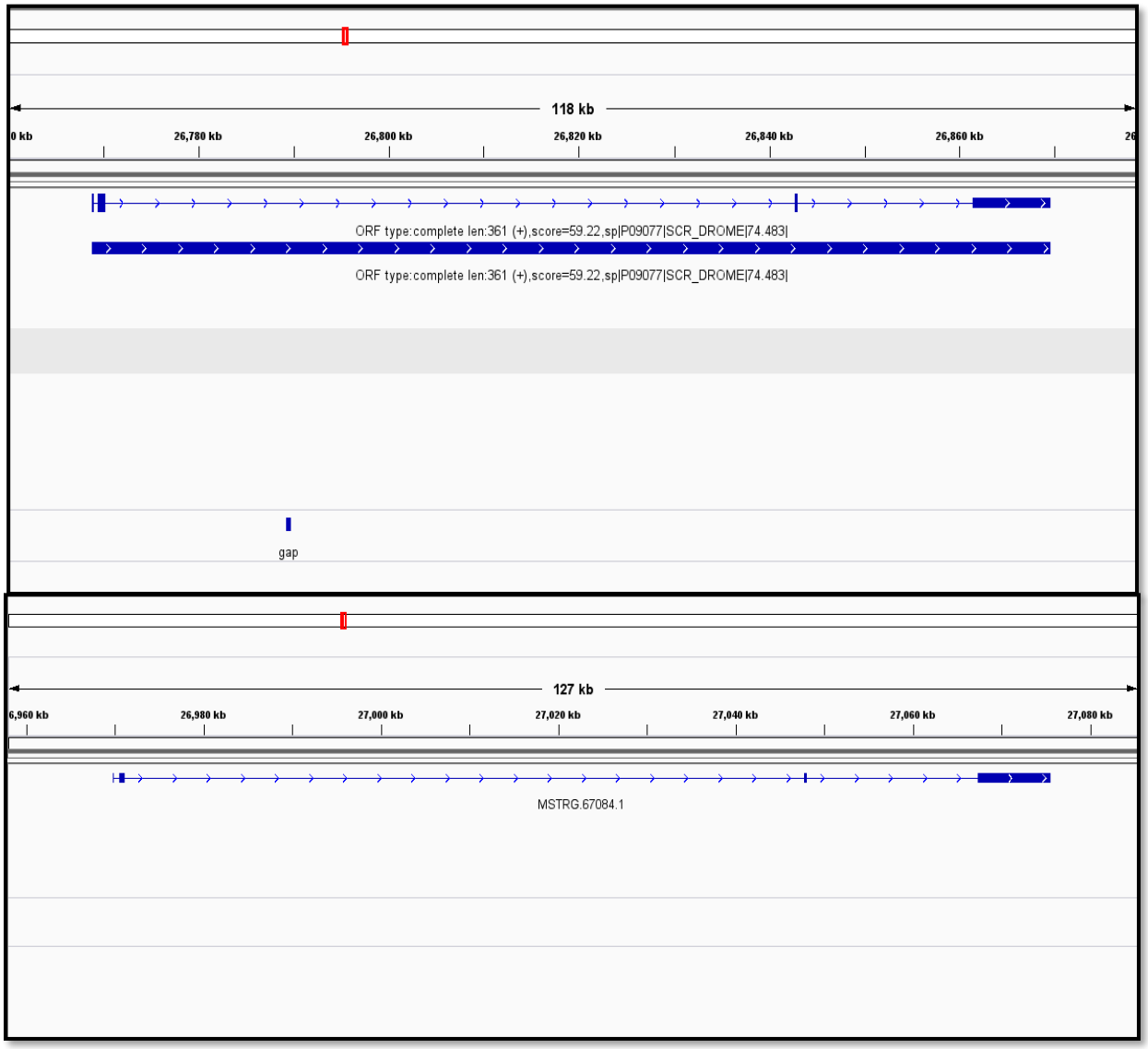
Abdominal-A


118 kb
25,640 kb
25,660 kb
25,680 kb
25,700 kb
25,720 kb
25,740 kb
ORF type:complete len:327 (+),score=29.14,sp|P29555|ABDA_DROME|50.000|
ORF type:complete len:327 (+),score=29.14,sp|P29555|ABDA_DROME|50.000|
gap
gap
127 kb
25,840 kb
25,860 kb
25,880 kb
25,900 kb
25,920 kb
25,940 kb
MSTRG.67078.1

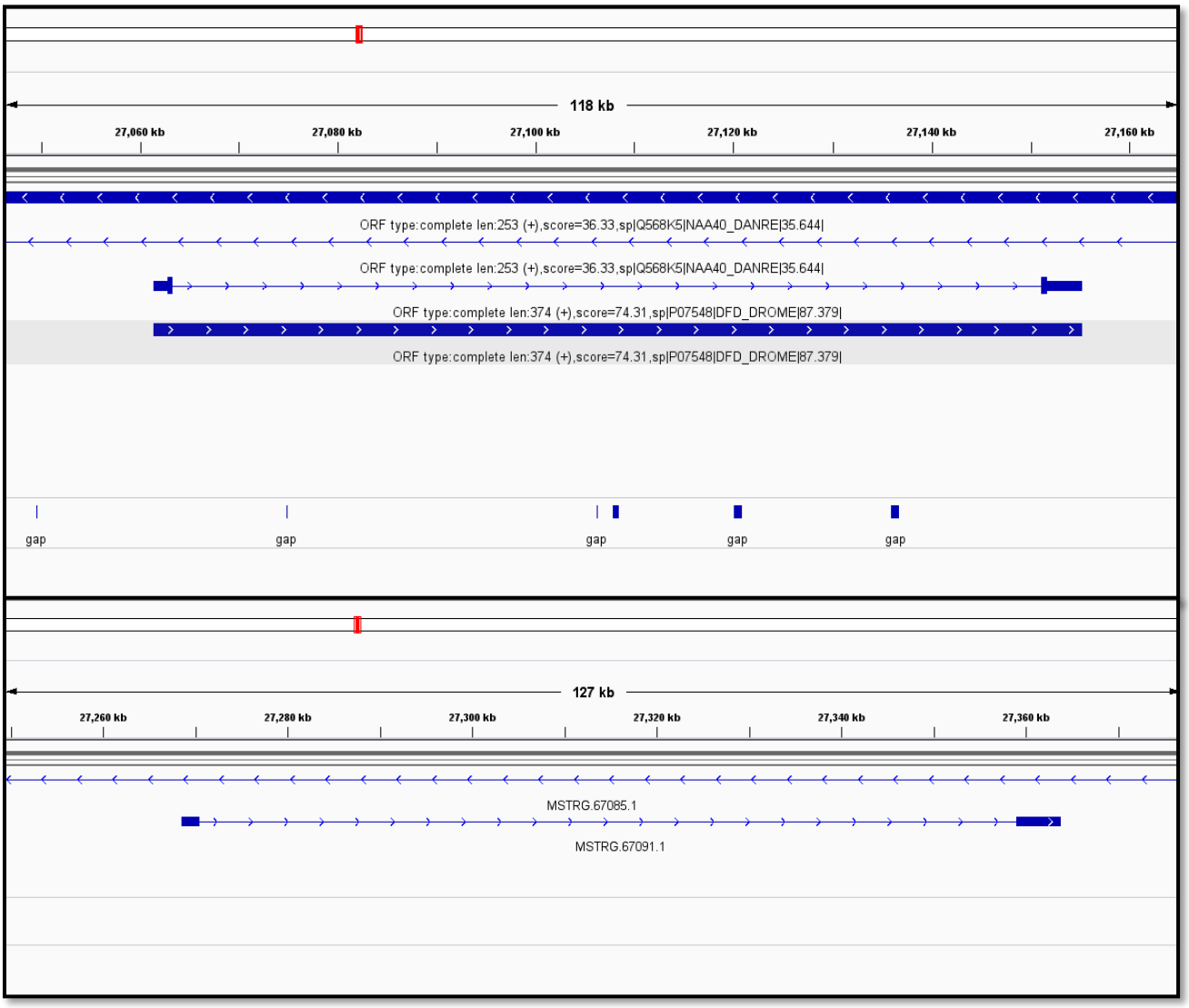
Antenapedia-Ultrabithorax



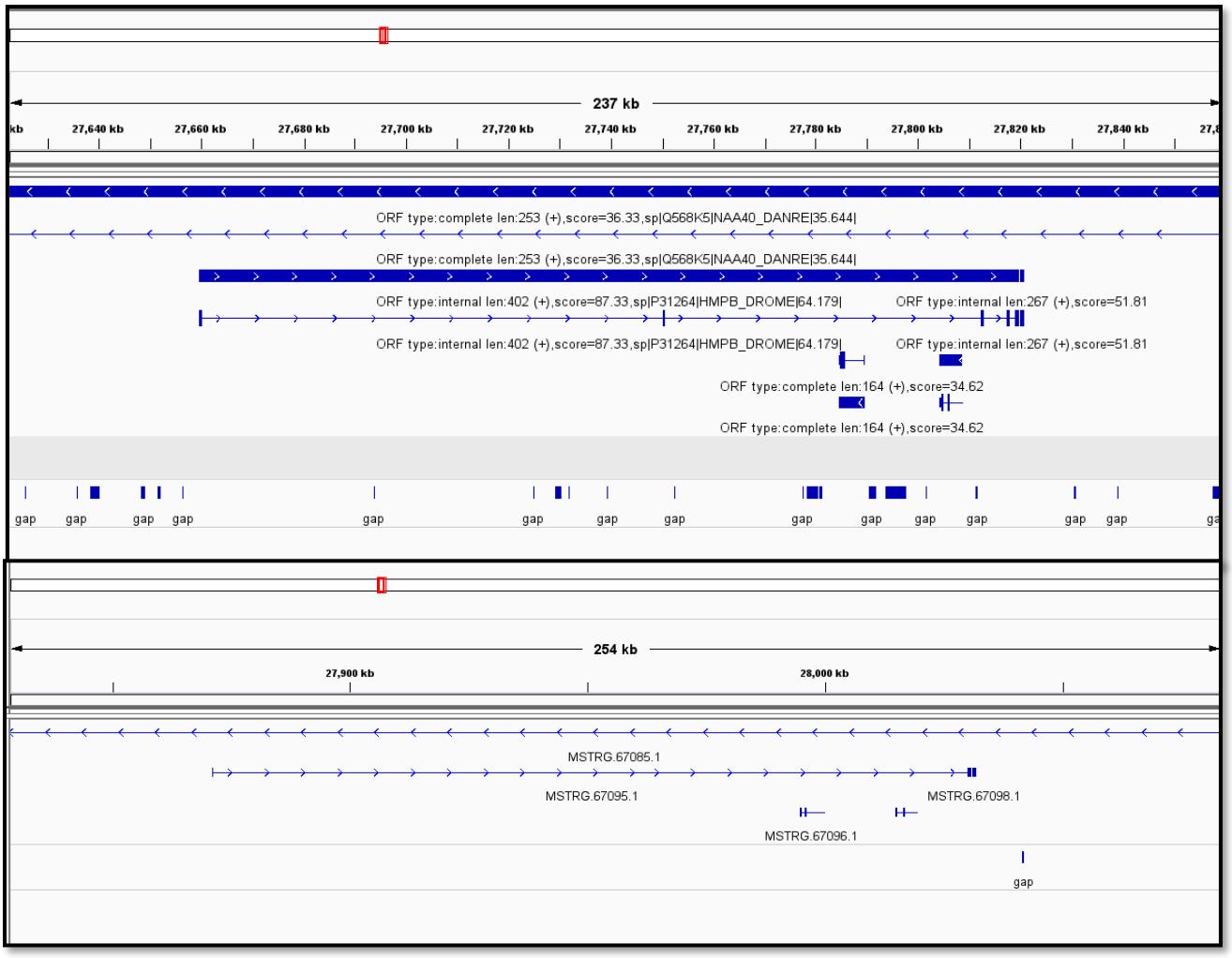
SCR


118 kb
0 kb
26,780 kb
26,800 kb
26,820 kb
26,840 kb
26,860 kb
ORF type:complete len:361 (+),score=59.22,sp|P09077|SCR_DROME|74.483|
ORF type:complete len:361 (+),score=59.22,sp|P09077|SCR_DROME|74.483|
gap
127 kb
6,960 kb
26,980 kb
27,000 kb
27,020 kb
27,040 kb
27,060 kb
27,080 kb
MSTRG.67084.1

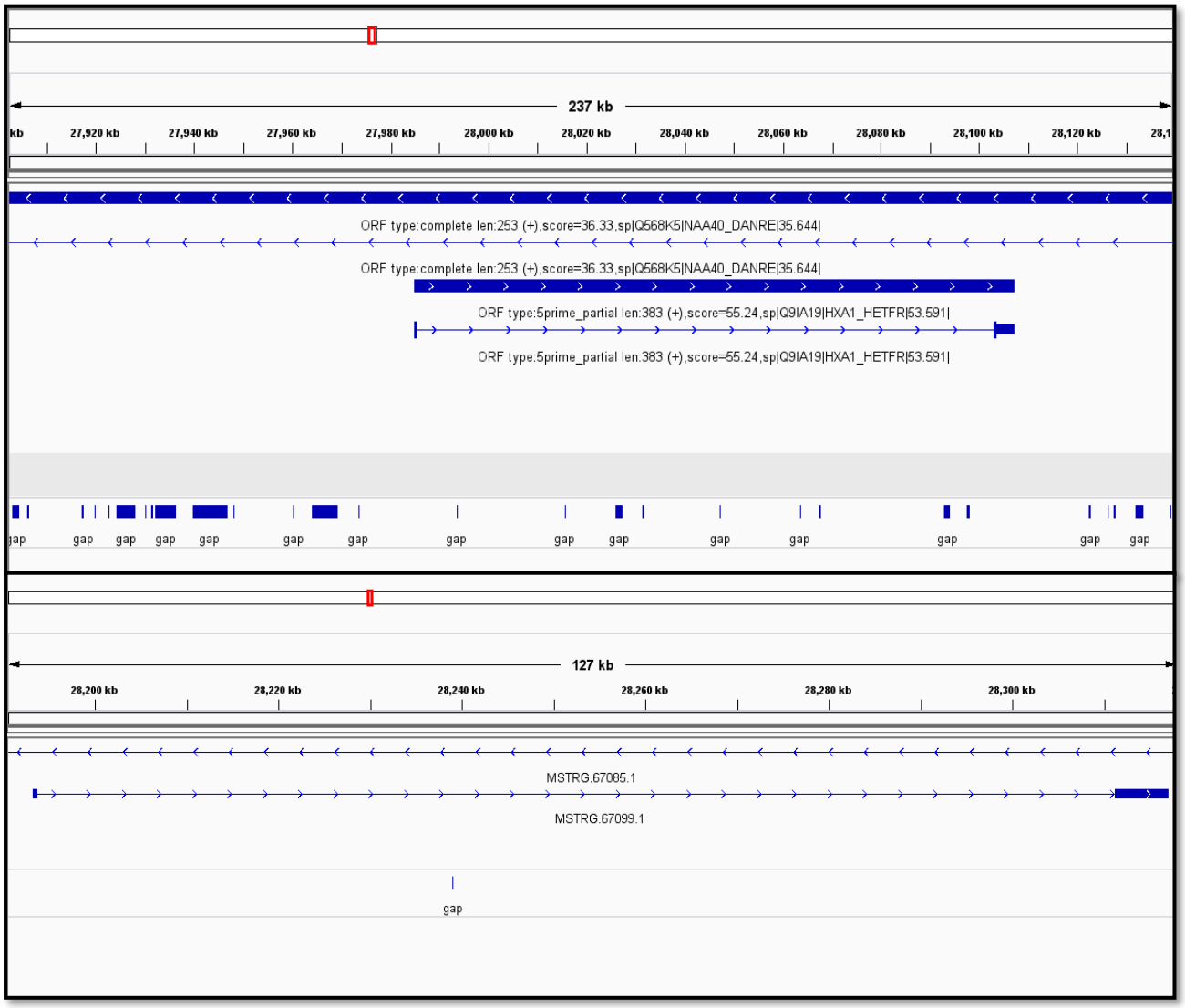
Deformed


118 kb
27,060 kb
27,080 kb
27,100 kb
27,120 kb
27,140 kb
27,160 kb
ORF type:complete len:253 (+),score=36.33,sp|Q568K5|NAA40_DANRE|35.644|
ORF type:complete len:253 (+),score=36.33,sp|Q568K5|NAA40_DANRE|35.644|
ORF type:complete len:374 (+),score=74.31,sp|P07548|DFD_DROME|87.379|
ORF type:complete len:374 (+),score=74.31,sp|P07548|DFD_DROME|87.379|
gap
gap
gap
gap
gap
127 kb
27,260 kb
27,280 kb
27,300 kb
27,320 kb
27,340 kb
27,360 kb
MSTRG.67085.1
MSTRG.67091.1

Proboscipedia



Labial.


237 kb
27,920 kb
27,940 kb
27,960 kb
27,980 kb
28,000 kb
28,020 kb
28,040 kb
28,060 kb
28,080 kb
28,100 kb
28,120 kb
ORF type:complete len:253 (+),score=36.33,sp|Q568K5|NAA40_DANRE|35.644|
ORF type:complete len:253 (+),score=36.33,sp|Q568K5|NAA40_DANRE|35.644|
ORF type:5prime_partial len:383 (+),score=55.24,sp|Q9IA19|HXA1_HETFR|53.591|
ORF type:5prime_partial len:383 (+),score=55.24,sp|Q9IA19|HXA1_HETFR|53.591|
gap
127 kb
28,200 kb
28,220 kb
28,240 kb
28,260 kb
28,280 kb
28,300 kb
MSTRG.67085.1
MSTRG.67099.1
gap

Appendix B. Validation of *Odd-paired* inserted fragment after PacBio data integration to the assembly. Alignment between Opa from Phaw_5.0 and Phaw_5.2 assemblies and sanger sequencing output of amplified CDS using MAFT tool, the alignment shows identical matching between the region filled by PacBio data and the sequenced region.

```
Phaw_5.0        GGAGAGAAGCCATTCAAGTGCGAATATGAGGGCTGCGACAGAAGATTTGCAAATTCGTCA
PCR             GGAGAGAAGCCATTCAAGTGCGAATATGAGGGCTGCGACAGAAGATTCGCAAATTCGTCA
Phaw_5.1        GGAGAGAAGCCATTCAAGTGCGAATATGAGGGCTGCGACAGAAGATTTGCAAATTCGTCA
                ********************************************.************

Phaw_5.0        GACAGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
PCR             GACAGGAAAAAGCACTCGCATGTCCACACGTCAGACAAACCTTACAACTGTAAAGTGAGA
Phaw_5.1        GACAGGAAAAAGCACTCGCACGTCCACACGTCAGACAAACCTTACAACTGTAAAGTGAGA
                *****

Phaw_5.0        NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
PCR             GGCTGTGACAAATCCTACACCCATCCTTCATCCCTCAGAAAGCACATGAAAGTCCATGGC
Phaw_5.1        GGCTGTGACAAATCCTACACCCATCCTTCATCCCTCAGAAAGCACATGAAAGTCCATGGC


Phaw_5.0        NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN----------------------
PCR             AAGAGTCCACCTCCGTCAGGCAGCGGTTACGAAAGCGACGACTCCACCACCACGACGGGA
Phaw_5.1        AAGAGTCCACCTCCGTCAGGCAGCGGTTACGAAAGCGACGATTCCACCACCACGACGGGA


Phaw_5.0        ------------------------------------------------------------
PCR             ACCTCCGCCTCCAATTCCAATATTCCCAGNTCCACTACACAAGTTCCTGGCCAGGGACAA
Phaw_5.1        ACCTCCGCCTCCAATTCCAATATTCCCAGTTCCACTACACAAGTTCCTGGCCAGGGACAA


Phaw_5.0        ------------------------------------------------------------
PCR             ACCAGTATAAGTGGTAGCAGGGAATGTTCGTTAAATAGTACACCAATACCACCAACACCA
Phaw_5.1        ACCAGTATAAGTGGTAGCAGGGAATGTTCGTTAAATAGTACACCAATACCACCAACACCA


Phaw_5.0        ------------------------------------------------------------
PCR             ACCGCAGGTGTAACTTCGCACAATACCACTACTCCCCATAACGCCAATTTAAGTGAGTGG
Phaw_5.1        ACCGCAGGTGTAACTTCGCACAATACCACTACTCACCATAACGCCAATTTAAGTGAGTGG


Phaw_5.0        ----------------------------------NNNNNNNNNNNNNNNNNNNCCTA-
PCR             TATGTGTGCCAAAACGCGGCCGGNNTGCCTACACNNNCCCAGCAATGAACATAGTCCTAT
Phaw_5.1        TATGTGTGCCAGAGCGCGGCCGGTATGCCTACAC-CTCCCAGCAATGAACATAGTCCTA-
                                                                      ****

Phaw_5.0        TTCCTCCTATCCCTGGTACAATTTTGCCCGC-ATCTACGGTTTATTAGCAAA-TATTGTA
PCR             TTCCTCCTATCCCTGGTACAATTTTGCCCCCNANCCACGGTTNNNTAACAAANNATTGGA
Phaw_5.1        TTCCTCCTATCCCTGGTACAATTTTGCCCGC-ATCTACGGTTTATTAG------------
                ------------------------------------------------------------
```

Appendix C. Histone modifying genes in *Parhyale hawaiensis.*

	Histone	Site	Histone-modifying enzymes	Function	P. hawaiensis gene ID
Acetylation	H2A	Lys4	Esa1	transcriptional activation	-
		Lys5	Tip60	transcriptional activation	MSTRG.14997.3.p1
			p300/CBP	transcriptional activation	MSTRG.67809.1.p1
		Lys7	Hat1	unknown	MSTRG.51752.1.p1
	H2B	Lys5	p300, ATF2	transcriptional activation	MSTRG.67809.1.p1
		Lys11	Gcn5	transcriptional activation	MSTRG.51124.1.p1
		Lys12	p300/CBP	transcriptional activation	MSTRG.67809.1.p1
			ATF2	transcriptional activation	MSTRG.65654.1.p1
		Lys15	p300/CBP	transcriptional activation	MSTRG.67809.1.p1
			ATF2	transcriptional activation	MSTRG.65654.1.p1
		Lys16	Gcn5, Esa1	transcriptional activation	MSTRG.51124.1.p1
		Lys20	p300	transcriptional activation	MSTRG.67809.1.p1
	H3	Lys4	Esa1	transcriptional activation	
		Lys9	Gcn5	transcriptional activation	MSTRG.51124.1.p1
			SRC-1		MTRG.66040.1.p1
		Lys14	Gcn5	transcriptional activation	MSTRG.51124.1.p1
			PCAF	transcriptional activation	MSTRG.51124.1.p1
			Esa1, Tip60	transcriptional activation, DNA repair	MSTRG.14997.3.p1
			SRC-1	transcriptional activation	MTRG.66040.1.p1
			Elp3	transcriptional activation	
			TAF1	RNA polymerase II transcription	MSTRG.52048.1.p1
			Sas2	euchromatin	
			Sas3	transcriptional activation	
			p300	transcriptional activation	MSTRG.67809.1.p1

		Lys18	Gcn5	transcriptional activation, DNA repair	MSTRG.51124.1.p1
			p300/CBP	DNA replication, transcriptional activation	MSTRG.67809.1.p1
		Lys23	Gcn5	transcriptional activation, DNA repair	MSTRG.51124.1.p1
			Sas3	transcriptional activation	
			p300/CBP	transcriptional activation	MSTRG.67809.1.p1
		Lys27	Gcn5	transcriptional activation	MSTRG.51124.1.p1
		Lys36	Gcn5	transcriptional activation	MSTRG.51124.1.p1
		Lys56	Spt10	transcriptional activation	
	H4	Lys5	Hat1	histone deposition	MSTRG.51752.1.p1
			Esa1, Tip60	transcriptional activation, DNA repair	MSTRG.14997.3.p1
			ATF2	transcriptional activation	MSTRG.65654.1.p1
			p300	transcriptional activation	MSTRG.67809.1.p1
		Lys8	Gcn5, PCAF	transcriptional activation	MSTRG.51124.1.p1
			Esa1, Tip60	transcriptional activation, DNA repair	MSTRG.14997.3.p1
			ATF2	transcriptional activation	MSTRG.65654.1.p1
			Elp3	transcriptional activation	MSTRG.10685.1.p1
			p300	transcriptional activation	MSTRG.67809.1.p1
		Lys12	Hat1	histone deposition, telomeric silencing	MSTRG.67809.1.p1
			Esa1, Tip60	transcriptional activation, DNA repair	MSTRG.14997.3.p1
			p300	transcriptional activation	MSTRG.67809.1.p1
		Lys16	Gcn5	transcriptional activation	MSTRG.51124.1.p1
			MOF (Kat8)	transcriptional activation	MSTRG.14997..p1
			Esa1, Tip60	transcriptional activation, DNA repair	MSTRG.14997.3.p1

			ATF2	transcriptional activation	MSTRG.65654.1.p1
			Sas2	euchromatin	
		Lys91	Hat1	chromatin assembly	MSTRG.51752.1.p1
			Hat2	chromatin assembly	MSTRG.31604.1.p1

	Histone	**Site**	**Histone-modifying enzymes**	**Function**	**Transcriptome ID**
Methylation	H1	Lys26	Ezh2	transcriptional silencing	MSTRG.1221.12.p1
	H2A	Arg3	PRMT1	transcriptional activation, repression	MSTRG.64417.1.p1
		Arg3	PRMT5	transcriptional activation, repression	Not found in the new annotation
		Arg3	PRMT6	transcriptional activation, repression	MSTRG.13999.1.p1
		Arg3	PRMT7	transcriptional activation, repression	MSTRG.37862.1.p1
	H3	Arg2	PRMT5	transcriptional repression	Not found in the new annotation
		Arg2	PRMT6	transcriptional repression	MSTRG.13999.1.p1
		Arg8	PRMT5	transcriptional activation, repression	Not found in the new annotation
		Arg8	PRMT6	transcriptional activation, repression	MSTRG.13999.1.p1
		Arg17	CARM1	transcriptional activation	MSTRG.21339.3.p1
		Arg26	CARM1	transcriptional activation	MSTRG.21339.3.p1
		Arg42	CARM1	transcriptional activation	MSTRG.21339.3.p1
		Lys4	Set1	permissive euchromatin (di-Me)	MSTRG.39452.1.p1
			Set7/9	transcriptional activation (tri-Me)	
			MLL, ALL-1	transcriptional activation	MSTRG.25526.1.p1
			Ash1	transcriptional activation	MSTRG.51236.1.p1
		Lys9	Suv39h, Clr4	transcriptional silencing (tri-Me)	MSTRG.28229.2.p1

			G9a	transcriptional repression genomic imprinting	MSTRG.27940.1.p1
			SETDB1	transcriptional repression (tri-Me)	MSTRG.2381.1.p1
			Dim-5, Kryptonite	DNA methylation (tri-Me)	
			Ash1	transcriptional activation	MSTRG.51236.1.p1
		Lys27	Ezh2	transcriptional silencing, X inactivation (tri-Me)	MSTRG.1221.12.p1
			G9a	transcriptional silencing	MSTRG.27940.1.p1
		Lys36	Set2	transcriptional activation	MSTRG.3354.2.p1
		Lys79	Dot1	euchromatin, transcriptional activation, checkpoint response	MSTRG.9989.2.p1
	H4	Arg3	PRMT1/6	transcriptional activation	MSTRG.64417.1.p1
			PRMT5/7	transcriptional repression	MSTRG.13999.1.p1
		Lys20	PR-Set7	transcriptional silencing (mono-Me)	MSTRG.9450.1.p1
			Ash1	transcriptional activation	MSTRG.51236.1.p1
			Set9	checkpoint response	

	Histone	**Site**	**Histone-modifying enzymes**	**Function**	**Transcriptome ID**
Phosphorylation	H2A	Ser1	MSK1	transcriptional repression	MSTRG.16386.3.p1
		Ser129	ATR, Tet1	DNA repair	MSTRG.27129.1.p1
		Ser139 (H2A.X)	ATR	DNA repair	MSTRG.27129.1.p1
		Ser139 (H2A.X)	ATM	DNA repair	MSTRG.67741.1.p1
		Thr119	NHK1	mitosis	MSTRG.36871.1.p1
		Thr120	Bub1	mitosis, transcriptional repression	MSTRG.21165.2.p1
		Thr120	VprBP	mitosis, transcriptional repression	MSTRG.38235.1.p1
		Thr142	WSTF	apoptosis, DNA repair	MSTRG.30524.1.p1
	H2B	Ser10	Ste20	apoptosis	MSTRG.61038.1.p1

		Ser14	Mst1	apoptosis	MSTG.15482.2.p1
		Ser33	TAF1	DNA repair	MSTRG.52048.1.p1
		Ser36	AMPK	transcriptional activation	MSTRG.10245.1.p1
	H3	Ser10	Aurora-B kinase	mitosis, meiosis	MSTRG.9270.1.p1
			MSK1, MSK2	immediate-early gene activation	MSTRG.16386.3.p1
			IKK-alpha	transcriptional activation	MSTRG.2476.1.p1
			Snf1	transcriptional activation	MSTRG.10245.1.p1
		Ser28 (mammals)	Aurora-B kinase	mitosis	MSTRG.9270.1.p1
			MSK1, MSK2	immediate-early gene activation	MSTRG.16386.3.p1
		Thr3	Haspin/Gsg2	mitosis	MSTRG.20620.1.p1
		Thr6	PKCB1	unknown	MSTRG.16688.1.p1
		Thr11	Dlk/Zip	mitosis	
		Tyr41	JAK2	transcriptional activation	MSTRG.26406.1.p1
		Tyr45	PKCdelta	apoptosis	MSTRG.56742.3.p1
	H4	Ser1	CK2	DNA repair	MSTRG.63543.1.p1

	Histone	**Site**	**Histone-modifying enzymes**	**Function**	**Transcriptome ID**
Ubiquitylation	H2A	Lys119 (mammals)	Ring2	spermatogenesis	MSTRG.46712.1.p1
	H2B	Lys120	UbcH6	meiosis	MSTRG.33675.1.p1
		Lys123	Rad6	transcriptional activation, euchromatin	MSTRG.2092.1.p1

	Histone	**Site**	**Histone-modifying enzymes**	**Function**	**Transcriptome ID**
Sumoylation	H2A	Lys126	Ubc9	transcriptional repression	MSTRG.2910.1.p1
	H2B	Lys6 or Lys7	Ubc9	transcriptional repression	MSTRG.2910.1.p1
	H4	N-terminal tail	Ubc9	transcriptional repression	MSTRG.2910.1.p1

Appendix D. Nucleotide coding sequences of candidate genes in this study.

Please note that the full-length coding sequence of DNMT1 was confirmed through cloning. However, after updating the assembly, it was discovered that the original transcript had been incorrectly split into two transcripts due to a mis-assembly error.

- **(DNMT1) MSTRG.16902.1.P1, MSTRG.16904.1.P2**

ATGGCTTCTATTATTCCTCCTGATATTGTGGACAGGGTTAACCAAACATGGGAAGAATAT
GAGGATGGTGATATAACTGCAAAGGGTGCTGTCAAAAGATTGCGTTCTCTTCTTGGCCCT
CATATGACTCATCTGTGCCTCAAAAAACTTACATATTTGGACAATACTTTGGAGACCGGT
GGTATCAACGAGCGGCAGTTCTACGAGCAGGCCTTGGAAGTATTTGAGTCATTCAACAA
ACTCTCACATGAAGATTCTGAAAAACCCTCACACTCCTCAGTGCAGGCCGAGAATGACC
AGAACCTTGCGAATGGATCCTATGCGTTCAGTGTTTTACCTGCGCAGCATTCGACATCAG
TTGATGACGATGCTGAAAGCTCCAATGAATCTGTTGGTAAAATTAGTGTGAAATCAATTA
AATCCTTGCAAGCTTCAGATGCCCATGAGTGTGTTGCCCCTGGCAAGACAGATGTAGCTT
TTAGAGAAATTCACCATGTTTCAGCCTCATGTACTGATAAGGGTACTTGTGTTGCGGCTA
CTAAAAATGGGTCAAATCTGGAAGAAAATCTCATCCCTAATGCCTCCCTGAAGAAGCTT
CCTCTTGCCAAGAAAATAGGAGGTAAGCTAAAAGCCAGCAACAAGAGACAAAAATCCA
TCACTGAAATGTTCTCGCTCGTTTCCAAGAGTAATGGTGCAGCTTCAGCATCTCAGGGCA
AGGACTCTCTAAGTAGCACACTACCCGACACTGTTGAGGGCAGTAAATCTACTGGCGCT
CTCAATGTCAAAATGGAGATGCCTGGAAGATCTAACTCTTCTACTGTTCAGCATGTTCCT
GTGGCTAATAGAATGTCAGGCAATGATGTTGCTAGTAGTAATGGTGCTCCCATTAAGCAC
ATTACGAACCTAAATGTAAAACACGAAATCGAATCTGCGTTTGTTGAACTCAGCGCAGA
TCTCCAGCCGTCTGATAACATGAAAGCCAGTACTACAGATGAAGCATGTCCTGATCATAC
AACAAAACCTCCCGACGTGAAAGAGTGTCAAAGCTCGACAGGTGTTGTTAGTGCACAAA
TGGTGGAACACTCTGGGGGTACTGTTGCTGCTAGTGATAGGCATGCAAAAGACTCTAGA
GGCTGTGCTGAGGATGTAGATGCTTCAGGCTCTGCCACTTCCGATTGCTTCTCTTCAATGT
TTCATTCTGATGATAACACCAATGATGAATTTGAAGGTTTTCCTGCTGCTAAGAGACTGA
AGCTGGCGGAGACAGAATGGCGTGAAGAGTCCAGTGCTACGAAGAAAACTAAGGCGCC
AGTGGAGCCCAAGCCTCGGTGTCCCATATGCCGCCAACTGCTGGACTCGGAGAATCTCT
ATCACTATGAGGGACATCCACAGAATGCCGCGGAAGAGTTCATTGCTCTCACGGACAGT
CGGCTCTCTCTTGTTCTGAGTGATGACATTGACGAGCGCCCGCAGCACAAGCTTACCTCC
TTCACTGTGTATGACAAGCAGGGCCACATGTGTCCTTTCGATAGTGGCCTCATTGAGCGC
AATGTTCTGCTCTATACATCTGGTTACATCAAGCCTATATTTGCTGAGGACCCATCACCC
GAGGATGGTGTGCCTGTCATGGATGTGGGGCCCATCAACGAGTGGTGGATCTCTGGCTTT
GATGGAGGAGAGAAGGCACTGCTGGGCTTCTCAACGGGCTATGCTGAATATGTCCTCAT
GGAGCCCACGCCCGTCTATGAGCGATTTGTGAATGCTGTCATGGAAAAGATCTACTTGA
GCAAGCTGGTCATTGAATTCCTAAATGATGCTGAGGATGGGTCGTACGAGGATCTGTTAC
ACGTGCTGCAGACTGCTGTGCCTCCCCAAGGAGTTGTCCTCAGCGAAGATTCCTTACTCC
GCCATGCCCAGTTTGTGTGTGACCAAGTACACAATTTTGACCTCGCTGGCAATGGTGTGG
ATGTGCTCATCACTCACCCGTGCCTGCGGTCACTTGTTGATCTGGCTGGGGTAACTCTCG
GCAAGCGAGGTGCCATGAAAACAAGAGGCGTGAGAATGAAAACCGTCAAGAAGATGCC
CAAGTGGACTAAAGCAACAACAACACCTCTTGTTCGTCAGTGTTTCGATCAGTTCTTCGC
CGACCAAATGGAAATGCGCAACAGTGGCAAAAGCAATGATGATGACGACAAGGAAAAA
GGAGGACCCAGGCGCATGCGCTGTGGTGTCTGTGAGGCCTGCCTCAGGCAGGACTGTGG
AAAATGCACCTCGTGCGCTGATATGATCAAGTTTGGCGGCAGTGGTCGGAGCAAGCAGT
GTTGTAAAGAGCGACGGTGCCCCAATATGATGCTCGCTGATCATGAGGATGACGATGAA
GACGCAGCAATAAGCAGATTGGCTAATTTACGAGATGCTGGCTCTCGAACTAAGACGCA

TCACGTCAAAAAAGCTAACAACTTATTGAGCTGGGTGGGTGAGCCTGTCAGTCGTACCA
GCAAGCGCACTTATTACGAGACTGCCAAAGTTGATGATGATCTGGTCACCAGAGGGGAC
TGCGTTCAGATAGAGCCAGACCCTGGCACTGGAGAATTACCATACATCGCCCGTGTTGTG
TCGTTGTGGGAGGATGCCTCTGGTGAGAAGAGTCTGCATGCCGACTGGTACTGCAGAGG
TGCGGACACCATCTTAGGAGAGACGAGCGACCCCCAGGAGCTTTTCATTGTAGACAACT
GCGAGGATATTCCCTTAGCATCCATTATGAAGAAGGTGAGGGTTGTTCCTCACTTACCCC
CGCCTAACTGGTCTCTACTGGGAGGCATCTCTCATCCTGAGGACCTGCGTCCTTTAGCTT
CCGATGACACCCATAGCTTCTACTACCAGTTGGCTTATGAGCACGAGCATGCCAGGTTCC
AGCATATGAGTCCAGAGAAGAATGAGCTATTGCCCAACAGCACAGACTGCCCCTGCTGC
GAGCGACTACGGCAAGCAGAACTGAAGGAGCTGTGCATACTCTCTGAACCAACGTCCCA
GACTGAGTTTCAAAGCCTGCTGCATATGGATCAACGCTACGCTGTTGGGGACGCTGTTCT
TGTGGACCCTGCTGCTTTCAATTTCAAGGTGAAGCTACCAAGTGTGAGTGGGAAAAAAC
CTCAACTGGAGCAAGTAGATGAAGAACTGTATCCGGAGTATTATCGTGTGACGCAGAAA
ATTAAGGGCAGCAATACTGATACATCAGAACCATTCAGAGTTGCTCTCATCACTGGAATT
AGAGTTACAGTTACAGGCAAAGCCGAGGTAGGTAGCGGCAGTGCGCTGGAGCCTGCTGA
TGTGACTGCCGTGCTGCGCAAACTTTACCGCCCTGAGAACACGCACAGAGGTCCTCCTGC
TGCCTATCAGGCTCCTCTCAACCTCCTCTACTGGAGTGAAGAAGAGGCAACTGTGACTTT
CAGTTCCATTTATGGCAAGTGTACAGTCGTTTATGGTGAGAACGTTACACTGTCACAAAA
TGAATTCTTCCTACAAGGCCCGTACAGGTTCATTTTCACTCAAGCTTATGATGCCACCAC
GCAGCAGTTCACTGAACCTCCTAAGCACGCCATGCTATATGGCGCTCCCGGAAAGGGCA
AAGGGAAAGGAAAAGGTTCAAGCAACAGAAAGGAGAACTGTGCTAATAAAGGCAATGT
CCCCGAAGATTATCCCAGCATCTCAAGGCGGCTGCGTACCCTGGATGTCTTCTCTGGCTG
TGGCGGTTTGTCTGAAGGTTTTCATCAAGCTGGGCTGGCAGAAAGCTGCTGGGCAATAG
AGGTATTTGAACCTGCTGCCAACGCTTACAAGTTAAATAACCCCGATGCTACAGTCTTCA
CTGATGACTGTAACCTGCTCCTTCGGATGGCAATGGAGGATAATGACTGCAACGCCAAA
GGCCAAAAACTGCCCAAGAAGGGCGATGTAGAGCTGTTGTGTGGAGGTCCGCCGTGTCA
GGGCTTCAGTGGCATGAACCGCTTCAACTCGAGACAATATTCTCAATTTAAGAACTCCCT
CGTGGCATCTTACCTGTCGTACTGTGACTTCTATCGGCCCCGTTTTTTCCTACTGGAGAAC
GTACGCAACTTTGTGTCGTACAAGTGCGGCATGGTTTTGCAGCTGACACTTCGAGTGCTT
GTGCAAATGGGCTACCAGTGTACTTTTGGCATACTGCAGGCAGGCAGCTATGGCGTTGC
GCAGACGAGACGCAGAGCAATAGTGCTGGCTGCTGCCCCTGGTGAGGAGCTGCCGTTCT
ATCCTGAGCCCCTACACTCGTTTGCACCCCATGCCTGTACTCTCTCTGCTGCTGTGGGCGA
TGTTAAGTATAAGAGCAACTGCAGGTGGAGTGTGTGCGGGCCACTGCGCACCATCACCG
TCCGAGACACCCTCAGCGATCTGCAGCCCATCACTAGCGGCGGGGGCAGGGAGCAGGTG
GCCTACTCTACAGAGCCCGAGTCTCACCTACAGCAGCTGCTTCGAGGTGCCGGTGGCGA
CTCGTTACTGTTGGACCATCAGTGCAAACCCTTGTCTGATTTGGTGGTGGCTCGCATGCA
GCACATCCCTACTGCCCCAGGCAGCGACTGGCGCATGCTGCCAAATAAACAAGTCCGTC
TCCCCGATGGATCATGGACCAGGAAATTGGAGTACAATTACGACGACAAGCGCAATGGC
AAGTCAAGCGAAGGATACTTACGAGGAGTGTGTGCGTGTGCCACTGGGGAGGCGTGTGA
CCCGCTAGACAAGCAGCACCACACGCTTATTCCTTGGTGTCTGCCCCATACTGCCAATAG
GCACAACAACTGGGCAGGTCTTTATGGCAGACTCGAGTGGGACGGCTTCTTTTCTACTAC
AGTTACCAACCCTGAGCCCATGGGCAAACAGGGCCGTGTGCTACACCCAGAACAGCACC
GTGTGGTGAGTGTGCGCGAGTGTGCCCGCTCCCAGGGCTTCCCTGACAACTATCGCTTCT
ATGGATCCCTGATAGAGAAGCACAGGCAGGTCGGCAATGCTGTACCCCCGCCCATGGCA
CGGGCGATTGGTCTTGAAATTCGGAAAAGCATTGCGTCTGCTGAGAAAAATCAACAGGC
GTAA

- **(DNMT3) MSTRG.11921.1.P1**

ATGGTCGTCTGTTGCTGTGTGACAAGACCGACTGCTATATGGTGTACTGTACGGAGTGTG
TAGAGTTGCTGGTGGGTCGGGAGTTTATGGAGAAGACTGTGGAGAGCCGAGATACCTGG
ACCTGCTTCTTGTGCACCTCTTACTCCCCTGATACCCATGGCTTGCTGCAGCCCAGGCCTC
ATTGGATCTTCCAGCTGGAGCAAATTCGCCAGGATTGTCAATTCTTGCAGTTGCCGTCGC
TGCGTCACATGAGCCTCAATTGCACAAAAGAGACGAAGAAGCCACTGAGGGTACTGTCT
CTCTTTGACGGCATCTCAACAGGTCTGTATTGCCTCGACCGCTTAGGACTCGACGTGGAA
GTGTACTTCGCCTCTGAGATAAGCGACTCTGCTCTACTTGTGTCAGAAACTCACTTTCGG
GGTCGCGTGTCTAGACTGGGGGACGTCCGCAAGATCACGAAAGATCAGATCAGCAACAT
GGCACCCATTGATCTGCTCATTGGTGGATCCCCGTGCTCGGACCTAAGTGGTGTCAATCC
GAACCGCAAGGGGCTGTTTGATCCGTCTGGTTCTGGAGTTTTGTTCTTTGAATTTGTTCGT
CTGCTGAAGTGCGTAGCCAAAAGCAACGGGAACCATCCCCTCGTGTGGCTCTTCGAGAA
CGTCGCCTCCATGCCGGAGATGTACAAACGAGCCATCAGCAAGCATCTTCAGACGGATC
CCGTATGTGTTGATGCTCGTCTGTTCTCACCCATGCGGAGAAAGCGCCTATTCTGGGGTA
ATATTCCAGATATGGTAGATTTAACTTCATCTACGTCTGATGACAGCCTACCACGCCTTG
CGGACTACATTGAACCTGTCTTCGGCCGACGTGCAGTGGTGTCGCACGTGCCGTGTATCA
CCACGAGCAGCAGCACCACCACCACAGTGGACCAGGGACCAGGCTGGCCGGTGATAGT
GGGAGGCCGCGGAGACTCGTTGTGGATCACCGAAATTGAGAAGATATTTGGCTTCCCCT
CGCATTTCACTGACGCGGGGGACCTGCTGCCCCGCCAGCGGCAGCAGCTGCTCGGGAAG
GCTTGGAGTGTGCCGGTGCTGGTGCACTTGTTGACCCCTTTAAAGGAGATTTTCCGTACG
GTATCGTCGTAA

- **(MBD2/3) MSTRG.63984.1.P1**

ATGAGTGTGCAAATTCAGCGCCGTCGCTATGAATGCTCCGCACTTCCAAAAGGTTGGAAGAG
AGAAGAAGTTGTCCGTAAAAACTGCCTCATAACACCTGGAAGATTAGATGTATATTATTATA
GTCCATGTGGGAAAAGAATCCGCAACAAGCCTCAACTCATCCGCTTCCTCGGGGAATCTGTT
GATCTTTCTTCGTTTGATTTCCGCACTGGCAAGATCAACCCAATGCTTGTACGCAAGAAAAAA
TCTAAAGGAACTCTGTGTGACTACAGCCGAGGGCTTCGAAGTGACAGCTGTCTAGTGCCCCC
CATTCGACAGACCGCTTCTATATTCAAGCAACCTGTAACTGTTGTGAGGACGCAGGCCCGAG
GGGAAGTCAAGAATGATATTAAGCATGGTACTCAAGACAAGCCAAAACAGGTGTTCTGGGA
GAAAAGGTTGCAAGGTCTCGGCGCGACGTGCAATGTTAAGTGGCCTCTGATGGAAGACCTAC
AACTGCCCTCTGCACTCAAGCCTGTGGGGCCTCAGATGTCCAGTCACACTGTACTGCAGAGTC
TGGCCACAGCCCTGCACGTGGTGTCGGGCCCCATCACCGGCCAGACCGCCTCTGCGCAGGTC
CTCCAGTCCAATCCTGCTGTCTTCATCAATCCTGACCAGCCTCTGGTCGAGCAAGTGCACATT
CGAGACGCTGATATCTGCAGTCAAGAGGCGCGGGTGGACCAGATCCGGAACCGGCTAGAGC
GAGCCATTCGCCAAATAACTGACACCGCCACCCCTAGAAACTCCTGA

Appendix E. Knock-in Construct sequences

Note: construct sequence files are best viewed using ApE software found in https://jorgensen.biology.utah.edu/wayned/ape/

(DNMT3 Construct)

https://drive.google.com/file/d/1WvrkqfoSXNrcWZJPP6zhVvNd7jtug60g/view?usp=share_link

(MBD2/3 Construct)

https://drive.google.com/file/d/11x8J7ZTvWIzGcBX9xgVYsOe7TdI7a1Vt/view?usp=share_link

Appendix F. Summary of dataset generated in chapter III.

Library		Number of reads	Alignment rate (%)	Uniquely mapped	Multi mapped
Sequencing method	**Samples**				
PolyA+ RNA-seq	**0-9hr_rep1**	45,609,500	82.07	69.08	5.92
	0-9hr_rep2	35,309,025	82.13	69.01	6.13
	0-9hr_rep3	30,614,500	81.96	68.56	6.07
	11hr_rep1	25,060,574	81.13	68.36	5.70
	11hr_rep2	29,839,465	82.26	68.02	7.34
	12hr_rep1	19,116,044	79.70	63.58	6.34
	12hr_rep2	21,394,095	81.31	65.42	7.72
	12hr_rep3	20,216,379	81.37	64.93	7.50
	13-14hr_rep1	20,337,286	80.14	64.31	6.25
	13-14hr_rep2	21,650,467	80.20	64.55	6.24
	13-14hr_rep3	19,526,435	80.70	65.18	6.44
	16hr_rep1	22,276,171	84.60	65.94	11.75
	16hr_rep2	20,629,563	83.99	65.19	10.96
	20hr_rep1	19,129,596	85.57	67.49	10.60
	20hr_rep2	28,807,974	84.44	67.02	9.94
	24hr_rep2	23,092,648	85.22	68.20	9.53
	24hr_rep2	21,643,893	85.65	69.57	9.76
	24-60hr_rep1	34,034,387	84.76	69.60	8.10
	24-60hr_rep2	28,794,064	85.08	69.64	8.45
	24-60hr_rep3	21,543,419	86.04	68.02	7.34
Total RNA-seq	**0-9hr_rep1**	33,282,689	83.88	66.77	8.53
	0-9hr_rep2	37,821,313	84.09	66.76	8.66
	0-9hr_rep3	32,197,721	84.35	67.04	8.63
	11hr_rep1	32,334,730	83.73	65.68	9.13
	11hr_rep2	36,276,611	85.61	63.87	11.81
	12hr_rep1	32,679,654	82.72	65.08	8.54
	12hr_rep2	29,209,827	86.68	63.93	12.59
	12hr_rep3	32,315,621	83.85	64.72	9.39

	13-14hr_rep1	30,683,765	83.04	64.31	8.95
	13-14hr_rep2	31,369,797	82.83	64.08	9.37
	13-14hr_rep3	28,924,485	83.56	65.02	8.50
	16hr_rep1	36,286,541	86.77	56.89	20.84
	16hr_rep2	31,747,347	86.00	57.58	20.24
	20hr_rep1	27,492,624	86.75	57.18	21.03
	20hr_rep2	29,837,311	86.20	57.54	20.25
	24hr_rep1	35,063,272	85.72	57.53	18.96
	24hr_rep2	31,426,041	86.54	56.75	19.77
	24-60hr_rep1	30,257,286	86.97	63.31	15.20
	24-60hr_rep2	30,462,897	85.81	62.03	15.09
	24-60hr_rep3	30,964,324	87.56	61.93	17.30

Appendix G. The transcript list for each category identified in the analysis of *Parhyale hawaiensis* MZT & ZGA.

List of maternal transcripts identified in *Parhyale hawaiensis*.

https://drive.google.com/file/d/1bOpnpMUD_a1ScJHhUI_EC5lertJN20A-/view?usp=share_link

List of all minor wave genes.

https://drive.google.com/file/d/1QvnduPeMoH-Ckxchjqae5rqZReNBsPv0/view?usp=share_link

List of all major wave genes.

https://drive.google.com/file/d/1aHsOOLLlcKHlTK2cICR6A7AhGAjdQ9lw/view?usp=share_link

List of genes only provided maternally.

https://drive.google.com/file/d/1pniBCePCHZ8fSP4hmApAL7aLNTsPfBFo/view?usp=share_link

List of only Zygotic genes (not expressed maternally).

https://drive.google.com/file/d/11eRJXpg2hAZ7nxXbNTYL9x1XtWY1ZwUr/view?usp=share_link

List of Maternal-Zygotic genes.

https://drive.google.com/file/d/1GvMsz3Kf6t48M3Kj8ZNJO5AKKciKuEDF/view?usp=share_link