Small Non-Coding RNA in Early Fly Development: Plasticity, Interactions and Improved Bioinformatic Tools

Lovisa Örkenby
Small non-coding RNA in early fly development: plasticity, interactions and improved bioinformatic tools

Lovisa Örkenby

Department of Biomedical and Clinical Sciences (BKV), Department of Cell- and Neurobiology
Faculty of Medicine and Health Sciences
Linköpings universitet, SE-581 83 Linköping, Sweden

Linköping 2023
This work is licensed under a Creative Commons Attribution 4.0 International License.
https://creativecommons.org/licenses/by/4.0

© Lovisa Örkenby, 2023

Published papers (Paper I and II) are open access.
Printed in Sweden by LiU-tryck, 2023

https://doi.org/10.3384/9789180753395
ISSN 0345-0082

During the course of the research underlying this thesis, Lovisa Örkenby was enrolled in Forum Scientium, a multidisciplinary doctoral program at Linköping University, Sweden

Cover image: Illustrative sncRNAs and an early Drosophila embryo (captured using Leica DMi8 by Alessandro Gozzo)
ABSTRACT

At fertilization, the male and female pronuclei undergo a transformation from germline to pluripotency as they fuse, marking the beginning of Drosophila embryogenesis. As the parental contributions decrease, the zygote takes control of its genome in a process called the maternal-to-zygotic transition (MZT). Several small non-coding RNAs (sncRNAs), a very large and diverse group of RNAs, have regulatory roles during this transition. This includes for example microRNAs (miRNAs) and piwi-interacting RNAs (piRNAs). Regulation by miRNAs mainly occurs through mediating maternal mRNA degradation, while piRNAs operate by repressing transposable elements (TEs) and regulating the nanos-induced embryonic body axis determination.

In this thesis, the complex and dynamic field of early Drosophila embryogenesis and sncRNAs are put in relation to the included papers. In Paper I, I explored the most stress-sensitive embryonic period and found that stress before the midblastula transition retains maternal miRNAs. These miRNAs impact zygotic gene activation by modulating the boundary factor Elba1, leading to compromised transcription control. Paper III examines the piRNA population during MZT. I find differences of unique piRNA sequences in embryos of different ages but not in target preferences, potentially highlighting the importance of constant repression of certain TEs. Paper II addresses specific difficulties with sncRNA seq data analysis and presents a bioinformatic framework to improve these analyses using a sequence-based strategy.

This thesis highlights the intricate interplay of sncRNAs in the critical period of early Drosophila embryogenesis and offers insights into their regulatory roles.

Keywords: Drosophila, Maternal-to-zygotic transition, sncRNA, miRNA, piRNA, embryogenesis, Zygotic gene activation
Små icke-kodande RNA (småRNA) är en grupp av RNA molekyler som inte används till att producera protein. Istället har de som funktion att på olika sätt skydda cellen exempelvis från angrepp av hoppande gener, så kallade transposoner, eller till att förhindra att fel gener uttrycks vid fel tillfällen.


Denna avhandling handlar om hur småRNA uttrycks och fungerar under denna tidsperiod i bananflugan. Vi har i två projekt mätt hur olika klasser av småRNA förändras från precis efter att ägget lagts av honan till dess att embryot själv börjar göra egna gentransskript.

Vi upptäckte att flera olika typer av småRNA förändrades mycket, varav en klass, microRNA, var den som ökade mest ju äldre embryot blev. Vi fann även att de maternellt laddade microRNAs och de som var gjorda av embryot skilde sig åt. Då vi utsatte embryot för en akut stress, fann vi att stress tidigt under MZT gör att de maternella microRNAs inte bryts ner, vilket normalt ska ske. Dessa kan då reglera proteinnivån av ett protein som är involverat i kontrollen av embryots genuttryck, vilket leder till en ökning av dessa gener.

Vi upptäckte även att de unika piwi-interagerande RNAs (piRNAs) skilde sig mellan tidig och sen MZT. Då piRNAs framför allt är involverade i tystningen av transposoner, undersökte vi om tidiga och sena piRNAs hade olika mål för reglering (targets). Jag fann att det inte var någon större skillnad i det piRNA som mamman laddade i ägget och det som embryot själv sedan började producera. Detta kan tyda på att tystningen av vissa transposoner är viktigare än andra och att mamman och embryot arbetar för att uppnå samma mål.
I det tredje arbetet som är inkluderat i denna avhandling diskuterar vi olika svårigheter vid sekvenseringsanalys av småRNA och presenterar ett nytt analysverktyg som behandlar flera av dessa bekymmer. Detta ramverk har även fördelen att det går att använda på en relativt enkel laptop, vilket gör att fler forskare har möjlighet att analysera denna typ av data och samtidigt hålla hög standard på analysen.
ACKNOWLEDGMENTS

First, I would like to thank my supervisor Anita Öst for taking me under her wing and believing in me. You have patiently taught me so much and helped me develop to become a more and more independent researcher.

I would also especially want to thank Rashmi Ramesh, my co-supervisor for all your support and advice. And also for all those very nice non-science related discussions, lunches, and fikas!

To Daniel Nätt, who was my co-supervisor. You had the absolute greatest amount of patience I have ever experienced when helping me take the first shaky programming steps. You will always be remembered.

I also want to express my great gratitude to everyone else in the Öst lab, for having succeeded in creating a simultaneously delightful and productive working environment. I have looked forward to meeting you every workday, even if the meetings are sometimes a bit too early for my taste. To Unn Kugelberg, who always knows how to solve EVERYTHING, lab-related or not. To Signe Skog, who is one of the sharpest minds and kindest people I’ve met. To Anna Asratian, my best desk buddy, whose sense of humor always makes me laugh. To Alessandro Gozzo, my “little brother in science” who has a never-ending enthusiasm that is highly contagious. I also want to give a shout-out to all previous colleagues and students who have worked in the lab. I’m so happy to have had such wonderful colleagues.

Jag vill även tacka mamma och pappa (Ingrid och Håkan Örkenby) och KalleP för att ni alltid stöttat mig, men även hjälppt mig sätta tillbaka fötterna på jorden när jag tappat perspektiv på vad som är viktigt i livet. Er villkorslösa kärlek är otrolig.

Jag vill även ge världens största tack till min stora kärlek Robin Kämpe, som inte bara hjälpt mig förbättra stora stycken av denna avhandling utan även har varit en klippa både hemma och i livet.
There are many more whom I would like to thank for making my time as a Ph.D. student fun and engaging.

Not least all the people at the cell biology department, floor 12, who always bring me to a good mode. I am so happy to have shared lunches and AWs with you, and I am convinced that this workplace is special because of the extremely lucky composition of people working here.

Also, all the amazing and smart fellow Ph.D. students and postdocs I have met through Forum Scientium. Meeting you has helped me cope with many issues relating to this education and letting me know I’m not alone.

And last but not least, to all my friends who have been there during this process. Especially to my “Corona family” Maike Bensberg, Markus Petersson, and Daniel Viberg, who helped me through the dark times. I am forever grateful.

I will always carry this time in my heart.
LIST OF PAPERS

This thesis is built upon two published papers (Paper I and II) and one manuscript (Paper III), listed below and referred to using Roman numerals in the text.


III. **Örkenby L** & Öst A, piRNAs during the maternal to zygotic transition in Drosophila melanogaster (manuscript)

* = authors contributed equally
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3' UTR</td>
<td>3' untranslated region</td>
</tr>
<tr>
<td>Ago1</td>
<td>Argonaute 1</td>
</tr>
<tr>
<td>Ago2</td>
<td>Argonaute 2</td>
</tr>
<tr>
<td>Ago3</td>
<td>Argonaute 3</td>
</tr>
<tr>
<td>ATAC</td>
<td>Assay for Transposase-Accessible Chromatin</td>
</tr>
<tr>
<td>Aub</td>
<td>Aubergine</td>
</tr>
<tr>
<td>BRAT</td>
<td>Brain Tumor</td>
</tr>
<tr>
<td>CCR4-NOT</td>
<td>Carbon Catabolite Repression – Negative on TATA-less</td>
</tr>
<tr>
<td>ENA</td>
<td>European Read Archive</td>
</tr>
<tr>
<td>FISH</td>
<td>Fluorescent In Situ Hybridization</td>
</tr>
<tr>
<td>H3K9me3</td>
<td>trimethylation on lysine 9 on histone 3</td>
</tr>
<tr>
<td>HP1</td>
<td>heterochromatin protein 1</td>
</tr>
<tr>
<td>Insv</td>
<td>Insensitive</td>
</tr>
<tr>
<td>LINE</td>
<td>Long Interspersed Elements</td>
</tr>
<tr>
<td>IncRNA</td>
<td>long non-coding RNA</td>
</tr>
<tr>
<td>LTR</td>
<td>Long Terminal Repeat</td>
</tr>
<tr>
<td>MBT</td>
<td>Mid-blastula transition</td>
</tr>
<tr>
<td>miRISC</td>
<td>microRNA-induced silencing complex</td>
</tr>
<tr>
<td>miRNA</td>
<td>microRNA</td>
</tr>
<tr>
<td>MZT</td>
<td>maternal-to-zygotic transition</td>
</tr>
<tr>
<td>nc</td>
<td>nuclear cycle</td>
</tr>
<tr>
<td>nt</td>
<td>nucleotide</td>
</tr>
<tr>
<td>PCA</td>
<td>principal component analysis</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>PEV</td>
<td>position-effect variegation</td>
</tr>
<tr>
<td>piRNA</td>
<td>piwi-interacting RNA</td>
</tr>
<tr>
<td>PNG</td>
<td>Pan Gu</td>
</tr>
<tr>
<td>PUM</td>
<td>Pumelo</td>
</tr>
<tr>
<td>qPCR</td>
<td>quantitative polymerase chain reactions</td>
</tr>
<tr>
<td>RISC</td>
<td>RNA-induced silencing complex</td>
</tr>
<tr>
<td>RNAi</td>
<td>RNA interference</td>
</tr>
<tr>
<td>rRF</td>
<td>rRNA fragments</td>
</tr>
<tr>
<td>rRNA</td>
<td>ribosomal RNA</td>
</tr>
<tr>
<td>RT-PCR</td>
<td>reverse transcription polymerase chain reactions</td>
</tr>
<tr>
<td>siRNA</td>
<td>small interfering RNA</td>
</tr>
<tr>
<td>sncRNA</td>
<td>small non-coding RNA</td>
</tr>
<tr>
<td>snoRNA</td>
<td>small nucleolar RNA</td>
</tr>
<tr>
<td>snRNA</td>
<td>small nuclear RNA</td>
</tr>
<tr>
<td>SRA</td>
<td>Sequence Read Archive</td>
</tr>
<tr>
<td>TAD</td>
<td>topologically associated domain</td>
</tr>
<tr>
<td>TE</td>
<td>transposable element</td>
</tr>
<tr>
<td>tRF</td>
<td>tRNA-derived fragment</td>
</tr>
<tr>
<td>wm4h</td>
<td>white mottled or In(1)wm4h</td>
</tr>
<tr>
<td>ZGA</td>
<td>Zygotic gene activation</td>
</tr>
<tr>
<td>Zuc</td>
<td>Zucchini</td>
</tr>
</tbody>
</table>
# Table of Contents

Abstract ............................................................................................................................ iii

Populärvetenskaplig sammanfattning ........................................................................ iv

Acknowledgments .......................................................................................................... vi

List of Papers ................................................................................................................ viii

Abbreviations ............................................................................................................... ix

1 Introduction .............................................................................................................. 13

   1.1 Aim of thesis ....................................................................................................... 14

   1.2 Ethical considerations ....................................................................................... 14

2 The Different Small Non-Coding RNA Classes ................................................. 15

   2.1 microRNA ......................................................................................................... 15

   2.2 Piwi-interacting RNA ....................................................................................... 17

   2.3 Other sncRNA Classes ..................................................................................... 19

      2.3.1 tRNA-derived fragments and tRNA halves ............................................. 20

      2.3.2 Small Nuclear RNA .................................................................................. 21

      2.3.3 Small nucleolar RNA .............................................................................. 21

      2.3.4 rRNA fragments ..................................................................................... 22

3 Creation of the Embryonic Environment ............................................................... 22

   3.1 Short about the egg ............................................................................................ 22

   3.2 Short about the sperm ....................................................................................... 23

   3.3 Fertilization ....................................................................................................... 23

   3.4 The parental sncRNA contributions ................................................................ 24

4 From parental to zygotic control .......................................................................... 26
4.1 The Maternal-to-Zygotic Transition – the Battle Begins 26
4.2 Zygotic gene activation – the Victorious Zygote 28
5 Epigenetics and higher order chromatin structures in the early embryo 29
5.1 sncRNAs and their role in regulating chromatin structure 31
6 methods to study sncRNA 32
6.1 Laboratory methods 32
6.2 Bioinformatic analysis 35
7 Methodology used in this thesis 37
7.1 Why Drosophila? 37
7.2 Sampling and choice of intervention 38
7.3 Pros and cons of using Position-effect variegation for studying the chromatin environment 39
7.4 Validating the connection between miRNAs and Elba1 40
7.5 Bioinformatic analysis of the sncRNA seq data 40
8 Summary of papers and results 42
Paper I - Stress-sensitive dynamics of miRNAs and Elba1 in Drosophila embryogenesis 42
Paper II - Seqpac: a framework for sRNA-seq analysis in R using sequence-based counts 44
Paper III - piRNAs during the maternal to zygotic transition in Drosophila melanogaster 45
9 Concluding Remarks and Future Perspectives 48
10 References 50
1 INTRODUCTION

At fertilization, the male and female pronuclei must transform their specialized germline configuration to pluripotency. This brief moment is the very foundation of embryogenesis and the essence of a new beginning. Shortly after, when the parental contributions and epigenetic signatures are almost completely erased, the zygote needs to take control of its genome, establish a new chromatin architecture, and start transcription (Hug et al., 2017; Yuan & O’Farrell, 2016; Stadler et al., 2017). The progression into zygotic independence is precisely orchestrated during the maternal-to-zygotic transition (MZT) to avoid deviating from the normal development path or lethality. This is first regulated by parentally loaded transcripts and proteins and then increasingly by the zygotic transcription of genes and regulatory small non-coding RNAs (sncRNAs), such as micro RNAs (miRNAs) and piwi-interacting RNAs (piRNAs) (reviewed in (Vastenhouw et al., 2019; Rojas-Ríos & Simonelig, 2018)).

SncRNAs are a large and diverse group of short non-coding RNAs that differ in origin, biogenesis, and function. miRNAs, for example, are transcribed into long pre-miRNAs, processed by Dicer-1, and loaded into Argonaute 1. The miRNA-RNA-induced silencing complex (miR-ISC) recognizes target mRNA by a complementary seed, which initiates cleavage or translational stalling ((Agarwal et al., 2018) reviewed in (Bartel, 2004)). miRNAs play a crucial role in the degradation of maternal mRNA during the MZT, a mechanism that is conserved in several species ((Bushati et al., 2008; Giraldez Antonio J. et al., 2006) reviewed in (Tadros & Lipshitz, 2009; Vastenhouw et al., 2019)). piRNAs are usually transcribed from piRNA clusters consisting of remnants of transposable elements (TEs) (reviewed in (Huang et al., 2017)). In the embryo, they are one of the key silencers of active TEs, which are highly abundant in the genomes of almost all life forms, but also responsible for the regulation of a growing number of known mRNAs (reviewed in (Rojas-Ríos & Simonelig, 2018)).

With the exception of miRNAs and piRNAs, other classes of sncRNAs also play important roles during embryogenesis and reproduction. For
example, cleavage products from tRNA, (5’ halves) regulate histone biogenesis and are involved in transmitting information from farther to offspring (Sharma et al, 2016; Chen et al, 2016). During *Drosophila* development, sncRNAs have various functions that help create specific environments within the embryo by diminishing the parental loadings, securing differentiation (e.g. at the forming pole cells), and cell type-specific transcription (Vastenhouw et al, 2019; Rojas-Ríos & Simonelig, 2018; Bushati et al, 2008). It is therefore valuable to address their expression and functionality throughout this important period of life.

### 1.1 AIM OF THESIS

The regulatory network governing the beginning of life, such as MZT, is surprisingly conserved between species. Since the first discoveries of sncRNAs not only their *modus operandi*, but also their regulation and impact on development and disease, has proven to be far more complex than anticipated. Here, I have focused on compiling the functional significance of sncRNA during early *Drosophila* embryogenesis and put this into the context of what I have learned and published. In addition, as sequencing has become cheaper, the amount of sequenced sncRNA data, created in-house or publicly available on storages such as from the Sequence Read Archive (SRA), has increased. This calls for simpler more accessible analysis tools that keep quality- as well as sequence information. I have therefore addressed and suggested further perspectives on this issue.

### 1.2 ETHICAL CONSIDERATIONS

All data generated for the papers included in this thesis have been solely generated using *Drosophila melanogaster* as the model organism. Due to this, the ethical guidelines and principles applied in animal research, such as those governing the use of vertebrate animals, are not applicable. For Paper II, all downloaded and used sequence data from SRA or European Read Archive (ENA) is publicly available and does not require authorized access.
2 THE DIFFERENT SMALL NON-CODING RNA CLASSES

When small RNAs were first identified during the 1960s, they were initially believed to be byproducts from the biogenesis or degradation products of longer RNAs. It was not until the late 90s - early 2000s when the laboratory methods were sophisticated enough to properly study these RNAs that it was understood that these short RNA fragments were of different and distinct classes having real biological relevance (reviewed (Mattick & Amaral, 2023)). Small non-coding RNAs (sncRNAs) are short RNAs, < 200 nucleotides (nt) but usually 20-36 nt long, that are not translated into proteins. Instead, sncRNAs usually have regulatory roles on targets due to different degrees of sequence complementary, which can lead to mRNA decay, translational inhibition, regulation of transposable elements (TEs), or histone biogenesis. Similar to mRNAs, sncRNAs have to go through processing steps in order to fulfill their functional roles, although the biogenesis differs widely from class to class. Many sncRNAs are dependent on some or several Argonaute proteins, which are the catalytic components of the RNA-induced silencing complex (RISC). There are many different Argonaute proteins and they are found both in the cytoplasm as well as inside the nucleus. The following text gives a brief introduction to a few of the most well-characterized sncRNAs.

2.1 MICRORNA

MicroRNAs (miRNAs) are found in all species from complex multicellular animals such as mammals, to plants and unicellular species such as algae. Although miRNAs can vary a lot between species, several examples of conserved sequences or functions have been reported over long evolutionary distances (reviewed in (Mattick & Amaral, 2023)). The mature miRNA is approximately 23-25 nt long and can regulate mRNA levels when loaded into the Argonaute 1 (Ago1) protein.

MiRNAs are produced from specific clusters of different individual miRNAs, introns, or other non-coding RNA transcripts (reviewed in (Quesnelle et al, 2023)). They are transcribed by RNA polymerase II, which generates a long pri-miRNA precursor consisting of a central hairpin flanked by 1 kb long sequences. Still inside the nucleus, the pri-miRNA precursor is recognized and cleaved by the microprocessor
complex consisting of the endonuclease Drosha with co-factors. This complex binds double-stranded RNA and cuts the pri-miRNA precursor at the base of the hairpin creating a 60 bp long pre-miRNA. The pre-miRNA is then exported to the cytosol and recognized and cleaved by another endonuclease called Dicer-1 that binds and cuts the pre-miRNA at the loop generating a double-stranded RNA duplex. Chaperone proteins then help load the double-stranded RNA into Argonaute 1 (Ago1) where one of the strands, called the passenger strand, is removed. Interestingly, both RNA strands could work as functional miRNAs. It is believed that the orientation of how the double-stranded RNA enters the complex is the main determining factor of which strand will be used (reviewed in (Quesnelle et al, 2023)) (Fig. 1).

**Figure 1. miRNA biogenesis and mRNA silencing.** Pri-miRNA precursors are transcribed from miRNA genes creating a hairpin loop flanked by 1 kb bases. These are cleaved by the microprocessor and the pre-miRNA is then exported to the cytosol where Dicer-1 cleaves the loop, leaving a double-stranded miRNA duplex. This is loaded into Ago1 and the passenger strand is discarded. The miRISC can then recognize target mRNA leading to their degradation or translational repression.
miRNAs primarily target mRNAs, but they are also found inside the nucleus where they can target pre-mRNAs, enhancer RNAs, long non-coding RNAs (lncRNAs), and other nuclear RNAs. Their mode of silencing depends on where the miRNA-induced silencing complex (miRISC) is located (Cosacak et al., 2018) as well as in which context it operates. In general, miRNAs recognize regions on their target 3’ untranslated regions (3’UTRs) which are approximately 7 bp long seed regions that are complementary to the nucleotides 2-8 on the miRNAs (Bartel, 2004; Agarwal et al., 2018). Greater complementarity gives stronger target repression (Agarwal et al., 2018).

The binding of miRISC to a target initiates the recruitment of other protein complexes and target repression can occur either by translational silencing or through target degradation. This occurs through e.g. recruited poly(A)-binding proteins or deadenylases which interfere with the translation initiation factors or destabilize and degrade the transcript (reviewed in (Quesnelle et al., 2023)). Although miRNAs are primarily known for their regulatory roles on mRNAs, miRNAs are themselves regulated to avoid excessive target elimination. This can be achieved by for example poly-adenylation or other posttranscriptional modifications, regulation of intracytoplasmic localization, cleavage within the RISC, or through target-directed miRNA decay ((Cosacak et al, 2018) reviewed in (Quesnelle et al, 2023)).

2.2 Piwi-interacting RNA

Piwi-interacting RNAs (piRNAs) are a very diverse group of sncRNA whose function was identified in Drosophila testis, where non-functional piwi biogenesis gives de-repression of transposable elements (TEs) and the gene Stellate leading to infertility (Aravin et al, 2001). piRNAs are mostly found in the germlines as well as during early embryogenesis where they help guarding the DNA integrity. They interact with a specific subgroup of Argonaute proteins called the piwi-proteins, which, in Drosophila, consists of Piwi, Aubergine (Aub), and Argonaute 3 (Ago3), where the Piwi-protein name is the abbreviation of P-element-induced wimpy testicles. piRNAs are typically 20-30 nt long, but the size can vary. One piRNA characteristic is a 2’O-methylation (methylaion on the ribose 2’-OH group of the RNA) on the 3’, which is
recognized by the PAZ domain of the piwi-proteins. The piRNA sequences are poorly conserved between species, and even lineages, although their regulatory mechanisms are conserved.

piRNAs are usually transcribed from defined piRNA clusters, where many consist of remnants of TEs (Huang et al, 2017)). TEs are genomic parasitic elements that are highly abundant in all life forms. They can transpose themselves to different places inside the genome, potentially disturbing genes, and the genomic integrity. As new TEs infect, these can by time be converted to piRNA clusters giving a defense against future transpositions. Except for piRNA clusters or transposons, piRNAs can also be derived from intergenic transcripts, pseudogenes, and 3’UTRs of mRNAs (Robine et al, 2009). The piRNA precursor exits the nucleus and is associated with one of the piwi-proteins and thereafter processed, independently of Dicer, in one of two mechanisms: either the phasing mechanism or the ping-pong amplification mechanism, or a combination of the two (Fig 2).

When processed through the phasing mechanism, the piRNA precursor undergoes cleavage by the enzyme Zucchini (Zuc), which is active on the outside of the mitochondrial membrane. Initially, this occurs by Aub-Zuc interaction and is then followed by Piwi-Zuc guided cleavage giving several new adjacent piRNAs. These piRNAs typically have the characteristic 1st nucleotide 5’ U-bias (Han et al, 2015). The ping-pong mechanism is instead active in the nuage, which is a membrane-less organelle located near the nucleus. Here, an interplay of pre-existing piRNAs, piRNA precursors, and active TEs contributes to both the silencing of TEs and amplification of piRNAs.

This is done by the cleavage of piRNA precursors loaded into Aub, which relies on the complementarity between an Ago3-loaded piRNA and the piRNA precursor and results in the generation of an Aub-loaded piRNA. This newly formed piRNA recognizes andcleaves an active TE, producing a new piRNA loaded into Ago3. The interplay of Aub and Ago3 leaves a 10 nt complementary overlap and an A-bias at the 10th nucleotide (reviewed in (Huang et al, 2017)). The target silencing and the biogenesis of piRNAs are therefore in many aspects highly intertwined, although piRNAs additionally silence targets transcriptionally. Piwi-piRNA complex, which can be nuclear, can recruit epigenetic
factors, such as the heterochromatin protein 1 (HP1a) or Rhino (a *Drosophila* HP1 homolog) to initiate trimethylation on lysine 9 on histone 3 (H3K9me3) and transcriptional repression (Brower-Toland *et al.*, 2007; Wang & Elgin, 2011; Le Thomas *et al.*, 2013) reviewed in (Tóth *et al.*, 2016).

Although piRNAs are primarily involved in the repression of TEs, they have been demonstrated to have a wider repertoire of regulation, such as the regulation of an expanding number of mRNAs. This regulatory role is especially interesting during the maternal-to-zygotic transition (see Chapter 4).

![Diagram](image)

**Figure 2. piRNA biogenesis and silencing of transposable elements.** piRNA precursors are transcribed from uni- (or dual) stranded piRNA clusters and thereafter processed through Zucchini (Zuc)-dependent phasing or the ping-pong mechanism. In the ping-pong mechanism, active transposable elements (TEs) are cleaved and used as piRNAs for the production and amplification of other piRNAs.

### 2.3 Other sncRNA Classes
In addition to miRNAs and piRNAs, there are many other types of sncRNAs. Several, but not all of them are to some extent discussed or examined in this thesis. I have therefore included short summaries of, for this thesis relevant, sncRNAs and excluded others. For excellent
overviews of functions and biogenesis of e.g. small interfering RNAs (siRNAs), or the longer lncRNAs, please see (Ghildiyal & Zamore, 2009; Li et al, 2019).

2.3.1 tRNA-derived fragments and tRNA halves
Initially thought to be degradation products of tRNAs, tRNA-derived fragments (tRFs) and tRNA halves have been shown to have several specific regulatory functions (Lee et al, 2009). As the name implies, tRFs and tRNA halves are processed from the cloverleaf-like (in secondary structure) mature tRNAs (reviewed in (Kumar et al, 2016; Shen et al, 2018)). The tRNA halves are generated by different endoribonucleases (e.g. angiogenin) that cleave in the anticodon loop and give both 3’ and 5’ tRNA halves. They are sometimes referred to as stress-induced tRNA fragments as several tRNA halves have been connected to and increased by stress (reviewed in (Hou et al, 2022)). In Paper I (Örkenby et al, 2023), we found that also early embryonic stress increases one of these 5’ tRNA halves, although the biological relevance of this increase is not further investigated within the context.

There are also shorter fragments produced from either the 3’- or the 5’ end and cleavage in the T or D loop. There are also internally derived fragments (tRFi), that cover the area from the D to the T loop, and fragments derived from the 3’ trailer (tRF-1) (reviewed in (Su et al, 2020b)). tRFs and tRNA halves have been connected to a growing number of important regulatory functions such as TE repression, translational control, gene silencing, chromatin organization, transgenerational inheritance, and cell proliferation. In addition, they have also been shown to regulate the biogenesis of other sncRNA classes such as snRNA and snoRNAs (see below) as well as the biogenesis of histones (reviewed in (Kumar et al, 2016; Su et al, 2020b)). Their role in regulating early embryogenesis is only speculative, but their dynamic expression and stress sensitivity have also been detected in the maternal-fetal placental interphase in mice and are suggested to have a role in the communication between the mother and fetus (Su et al, 2020a). In addition, tRFs have been implicated in neurodevelopment and neurodegenerative diseases (reviewed in (Hou et al, 2022)). Their function in paternal inheritance, histone biogenesis, sncRNA production, and TE silencing (Boskovic et al, 2020; Sharma et al, 2016; Chen et al, 2016) could
also point to a potential epigenetic regulatory role for tRFs and tRNA halves in early embryogenesis.

2.3.2 Small Nuclear RNA
Small nuclear RNAs (snRNAs) are Uridine-rich RNAs that are located in the nucleus. They were one of the first sncRNAs to be discovered in the 1970s, although their function at that point was not clear (reviewed in (Mattick & Amaral, 2023)). They are a critical component of the spliceosome where they guide the complex to pre-mRNA introns. This is done by the recognition and base-pairing to sequence elements inside the introns (reviewed in (Wilkinson et al, 2020)). In addition to their involvement in splicing, the snRNAs U1 and U4 are involved in transcriptional initiation, transcription structure, and chromatin architecture. The snRNA U2 is involved in RNA pol II pausing (Caizzi et al, 2021).

2.3.3 Small nucleolar RNA
Small nucleolar RNAs (snoRNAs) are mainly located in the nucleolus, which is a membrane-less compartment inside the nuclei in which ribosomal RNA (rRNA) biogenesis occurs. They are primarily produced from transcript introns and the mature snoRNAs are approximately 60-300 nt in length (reviewed in (Bratkovič & Rogelj, 2014)). There are two major classes of snoRNAs, the C/D-box and the H/ACA-box snoRNAs, which are distinguished by distinctive motifs and secondary structures. While both classes share the overall function of guiding enzymatic complexes to facilitate RNA modifications, the two classes have different means of accomplishing this. The C/D-box snoRNA guides enzymes to perform 2’-O-methylation (methylation on the RNA ribose 2’-OH group), whilst the H/ACA box snoRNA guides pseudouridylation, converting uridine to pseudouridine (reviewed in (Bratkovič & Rogelj, 2014)).

Although snoRNAs are mainly recognized for their involvement in modifying rRNAs and snRNAs, separate miRNA-like functions have also been demonstrated when snoRNAs are further processed into snoRNA-derived RNAs (Taft et al, 2009). Some of these RNAs are even classified as miRNAs or as piRNAs, and snoRNAs can therefore be viewed as sources for these more characterized sncRNA classes (reviewed in (Wajahat et al, 2021)).
2.3.4 rRNA fragments

Although sometimes neglected as degradation products of rRNA, rRNA fragments (rRFs) have been shown to have conserved cleavage sites and are found in most animals, plants, and fungi (reviewed in (Lambert et al, 2019)). In Drosophila, rRFs were shown to derive from mature rRNAs both from the nuclear and the mitochondrial genome and associate with either the Argonaute proteins Ago1 or Ago2 (Guan & Grigoriev, 2020). This is indicative that rRFs can have similar functions as siRNA (Ago2 loaded) or as miRNAs (Ago1 loaded). rRNAs have in addition been shown to template miRNAs (Chak et al, 2015). Interestingly, potential rRF seed regions were enriched for targeting genes involved in the development, including Brain Tumor (BRAT) (see Chapter 4) (Guan & Grigoriev, 2020).

3 CREATION OF THE EMBRYONIC ENVIRONMENT

3.1 SHORT ABOUT THE EGG

The Drosophila oocyte develops within an egg chamber together with 15 supportive nurse cells and follicle cells that surround the whole complex. Nutrients, mRNAs, and components needed for transcription and translation are loaded into the egg from the nurse cells during oogenesis, as the oocyte initially is meiotically stalled and transcriptionally inactive. The oogenesis occurs in 14 distinct stages, and it is not until the last three that the egg matures. As the oocyte is initially transcriptionally silent, the maturation process is started by translation and other post-transcriptional events (reviewed in (Avilés-Pagán & Orr-Weaver, 2018)).

In the immature oocyte, the transcripts have very short poly(A)-tails, protecting them from degradation, but also inhibiting translation. During maturation and egg activation, the translation of selected mRNAs is regulated by either poly-adenylation or poly(A)-tail shortening (Eichhorn et al 2016, Subtelny et al 2014). Interestingly, the egg is activated before fertilization, through mechanical forces, Ca\(^{2+}\) influx, and rehydration as the egg progresses into the oviduct. Proteins involved with embryonic patterning, division, and zygotic gene activation increase during egg activation (reviewed in (Avilés-Pagán & Orr-Weaver, 2018)).
3.2 SHORT ABOUT THE SPERM

In *Drosophila*, the whole spermatogenesis is completed in approximately five days. Both the germline stem cells and the somatic stem cells are localized at the most apical tip of the testis. As the germline stem cells go through mitosis, they and the somatic cells (now called cyst cells) form a cyst, where the two cyst cells pairs and encases the gonialblast. The gonialblast differentiates and continues to go through additional rounds of mitosis followed by meiosis, all within the cyst until they have reached 64 round spermatids (reviewed in *Demarco et al*, 2014).

Spermiogenesis starts immediately after the completion of meiosis in the round spermatids. The mitochondria fuse, creating two larger structures called the nebenkern, which localize close to the nucleus. As the spermatids start to elongate, the cyst is polarized and the sperm heads are placed on one side and the growing tails on the other side. The nucleus starts getting a less round, more elongated morphogenesis (reviewed in *(Fabian & Brill, 2012)*). Almost all histones are replaced by protamines or protamine-like proteins, leading to highly condensed DNA (reviewed in *(Loppin et al, 2015)*). As chromatin condenses, the transcription is strongly reduced. Although it was thought to be no transcription at this point, this is now being questioned as transcription has been detected also in later spermatids *(Vedelek et al, 2018; Witt et al, 2019)*.

The elongated spermatids are connected through ring canals as well as other bridges. As spermatogenesis proceeds, an individualization complex forms and travels along each spermatid, separating them and removing excess organelles. The mature sperm, which is approximately 1.8-2 mm long, is stored in the seminal vesicles in waiting for mating (reviewed in *(Fabian & Brill, 2012)*).

3.3 FERTILIZATION

The male *Drosophila* only transfers a few thousand sperms at mating, which is little compared to most other animals (reviewed in *(Loppin et al, 2015)*). The female does however store sperm efficiently in the spermatheca or the seminal receptacle (reviewed in *(Avilés-Pagán & Orr-Weaver, 2018; Loppin et al, 2015)*) and can release single sperms that
enter the egg through the micropyle, a specialized opening located at the anterior side of the egg. The whole sperm enters and coils up at this side of the egg. Following the entry, the sperm protamines are replaced by histones with the help of maternally provided factors (reviewed in (Lop- pin et al, 2015)). In addition, upon sperm entry, the female genetic content can continue meiosis, creating four female meiotic products. The sperm then creates a sperm aster that catches one of the nearest female nuclei which becomes the female pronucleus, and the rest becomes polar bodies. DNA replication occurs synchronously in the male and female pronucleus and the first zygotic division can complete.

3.4 THE PARENTAL sncRNA CONTRIBUTIONS

Both maternal and paternal load of sncRNAs has been shown crucial for embryo development in several species ((Rödel et al, 2013; Kugler et al, 2013), reviewed in (Skvortsova et al, 2018)). In Papers I and II in this thesis, we demonstrate that in the Drosophila stage 1 embryo, which is equivalent to the very first embryonic nuclear cycles, there is a large proportion of rRFs (50%) followed by 37% piRNAs and 10% miRNAs of the total sncRNA pool (Fig 3). These sncRNAs are presumed to be maternally (or paternally) provided. During Drosophila oogenesis, piRNAs scan the environment of potential targets inside the nurse cell nuages. These are then amplified by the ping-pong mechanism and loaded from nurse cells into the oocyte through ring canals (reviewed in (Rojas-Ríos & Simonelig, 2018)).

Maternally provided piRNAs, as well as the piwi-proteins (which are also maternally provided), have regulatory roles during embryogenesis including transcriptional silencing of TEs, germline development, and even regulating the nanos transcript thereby ultimately the body axis formation (Fabry et al, 2021; Rouget et al, 2010; Barckmann et al, 2015) (see Chapter 4). Interestingly, the maternal miR-2 family also has a regulatory role in the anterior – posterior-axis formation by cooperating with Bicoid and thereby regulating the Bicoid-induced repression (Rödel et al, 2013). In addition, certain maternal miRNAs potentially stabilize germ cell formation in Drosophila, as loss of the maternal mir-
9c and mir-969 leads to a reduced number of embryonic germ cells (Kugler et al., 2013).

Figure 3. Proportions of sncRNAs in embryos at different stages during MZT. The image is adapted from Fig 1 in Paper III.

In Paper 1, we found that maternal miRNAs are retained after exposure to heat shock stress at the very beginning of embryogenesis. We further found indications that dysregulation of these maternal miRNAs can lead to abnormal miRNA-induced regulation of early zygotic transcripts (Örkenby et al., 2023). The maternal miRNAs are usually marked and degraded during the maternal-to-zygotic transition similar to other maternal mRNAs (see Chapter 3). This has been shown to happen through polyadenylation by the poly(A) polymerase Wispy (Lee et al., 2014).

The paternal load of sncRNAs has a significant impact on the offspring. However, if the loading happens into the sperm or the seminal fluid, as well as how it is transmitted to the egg and thereby the next generation is still not clear. The sperm optimizes its structure and content for efficient fertilization. In several species, rapid changes in sperm sncRNAs following exposures to stressors or diets have been detected (Ramesh et al., 2023; Rodgers et al., 2015; Chen et al., 2016). Some studies in mice have even shown that sperm sncRNAs, especially tRFs, tRNA halves, and miRNAs, affect the offspring (Sharma et al., 2016; Chen et al., 2016; Rodgers et al., 2015). The relatively transcriptionally silent spermatid and sperm speak against de novo transcription of epigenetic factors or sncRNAs from these cells, at least from the nuclear genome. Instead, either the response occurs in less differentiated germ cells and is kept throughout spermiogenesis, the loading comes from the surrounding environment e.g. from cyst cells or seminal fluid, processing of sncRNA precursors occurs affecting the sncRNA content, or it is transcription
from the mitochondrial genome that could be potential responders to stress or other environmental changes.

Worth noting is that not all histones are replaced by protamines during spermiogenesis, and the non-replaced histones also keep some of their modifications. This occurs at specific loci such as some miRNA clusters and Hox genes (reviewed in (Skvortsova et al, 2018)), potentially speaking in favor of active transcription of sncRNA in late spermatids. In our lab, we have sequenced Drosophila mature sperm from seminal vesicles and found that in Drosophila sperm, the main sncRNA classes found (when excluding rRFs) were mitochondrial tRFs (47 %) followed by miRNAs (21.7 %) and piRNAs (15.9 %) (Ramesh et al, 2023). In this study, we show that miRNA and mitochondrial sncRNA respond acutely to short-term intervention with dietary sugar. Whether these changes originate in the sperm itself, or if they are transported to sperm via extracellular vesicles is not clear yet.

4 FROM PARENTAL TO ZYGOTIC CONTROL

4.1 THE MATERNAL-TO-ZYGOTIC TRANSITION — THE BATTLE BEGINS.
In animals with sexual reproduction, mRNAs, proteins and most other parentally (primarily maternally) loaded factors and molecules must be degraded to facilitate zygotic transcription. This process is known as the maternal-to-zygotic transition (MZT) and is largely conserved between species although the timing of the events may differ and the exact molecular key players are not always homologs (reviewed in (Hamm & Harrison, 2018)). In Drosophila, the MZT is well studied and shares several similarities with a broad variety of species such as humans, mice, X.laevis (African clawed frog), and zebrafish.

In Drosophila, the decay of maternal mRNAs is dependent on several different pathways where the first is prepared by the mother herself. The maternally provided Smaug mRNA, which has been translationally repressed by Pumelo (PUM), Brain Tumor (BRAT), and potentially other factors during Drosophila oogenesis is released upon egg activation
Smaug can therefore attract the deadenylase complex “Carbon Catabolite Repression – Negative on TATA-less” (CCR4-NOT) and target a large number of maternal transcripts (Tadros et al., 2007). In addition to Smaug, other maternally provided factors such as BRAT and PUM (Laver et al., 2015) as well as Pan Gu (PNG) are also essential for the maternally directed mRNA decay pathway. The maternal factors are together responsible for the deadenylation and clearance of about half of the maternally loaded mRNAs during MZT (reviewed in (Hamm & Harrison, 2018)).

Interestingly, maternally loaded piRNAs have also been shown to direct the degradation of the nanos mRNA in the soma by the direction of Smaug and CCR4-NOT (Rouget et al., 2010; Barckmann et al., 2015). The nanos mRNA is spread throughout the embryo and is translationally repressed and degraded at the anterior (head region) during the first 2-3 hours of embryogenesis. This later creates an anterior-posterior gradient of the translated nanos and is crucial for the head and thorax segmentation (Bergsten & Gavis, 1999). The nanos 3’UTR was shown to have potential binding sites for piRNAs originating from the 412 and roo transposable elements (Rouget et al., 2010; Barckmann et al., 2015). Interestingly, opposite to the Aub-piRNA RISC function in the soma, the same piRNA complex was shown to stabilize the nanos mRNA with the help of Wispy in the germplasm, further supporting the segregation (Dufourt et al., 2017).

Towards the end of the MZT, the zygotic contributions to the maternal mRNA decay become more prominent. In several species, zygotic miRNAs are important regulators of maternal mRNA decay. As described in Paper I, we and others have found that there is a great increase of miRNAs during the MZT, which does not compare to any other group of snRNAs (Örkenby et al., 2023; Luo et al., 2016; Bushati et al., 2008). Although the miRNAs, or their targets, are not conserved between species, the overall mechanisms have a similar modus operandi (reviewed in (Vastenhouw et al., 2019)). These miRNAs are miR-430 in zebrafish, miR-427 in X.laevis (African clawed frog), and miRNAs from the miR-309 cluster in Drosophila (Giraldez Antonio J. et al, 2006; Lund et al, 2009; Bushati et al, 2008). In Drosophila, the miR-309 cluster consists of 8 miRNA genes, encoding 6 different miRNAs, having 5 different seed regions (as mir-3 and mir-309 have identical seeds), and is present
throughout the whole embryo except for in the pole cells (Bushati et al., 2008). These miRNAs are responsible for the recognition and degradation of approximately 14% of all maternal transcripts during the later phases of MZT (Bushati et al., 2008). The miR-309 cluster is dependent on both Smaug and the maternal transcription factor Zelda (Luo et al., 2016; Fu et al., 2014).

### 4.2 ZYGOTIC GENE ACTIVATION — THE VICTORIOUS ZYGOTE

Several factors prevent the zygotic genome from being transcribed directly after fertilization. One straightforward example is that several of the maternally loaded proteins act as transcriptional repressors (reviewed in (Vastenhouw et al., 2019). Contradictory, several of the maternally provided transcripts also encode general and specific transcription factors needed for zygotic transcription. The de novo transcription is therefore dependent on the timing of the expression of different factors provided by the mother.

The initial very short nuclear cycle (nc) lengths are also a limiting factor for the full onset of zygotic transcription (reviewed in (Vastenhouw et al., 2019)). Following fertilization, there are 13 rapid syncytial nuclear divisions, lacking gap phases, creating approximately 6000 nuclei that share the same cytoplasm. These nuclei are initially divided inside the embryo but they start migrating to the periphery of the embryo at nc 10 at the syncytial blastoderm stage. It is not until cycles 11-13 that the tempo is progressively slowed down, the cycles lengthened and reach approximately 21 min at nc 13 and increase to 1 h at nc 14 (reviewed in (Yuan et al., 2016; Avilés-Pagán & Orr-Weaver, 2018)). In Drosophila, the mid-blastula transition (MBT), which is when the nuclear cycles are no longer synchronous, occurs at this stage and so do cellularization and the full zygotic gene activation.

Despite the limited time, there is still some zygotic transcription during the first rapid mitotic cycles, and RNA polymerase II has been detected at approximately 100 genes at this time. (Chen et al., 2013). Already in nc 8, although only estimated to be 12 seconds, there is a theoretical transcriptional window of approximately 200 bp (Strong et al., 2020). This window increases with each cycle. In concordance with the time constraint, the pre-MBT transcripts are short, lack introns, and show
signs of being aborted (Chen et al., 2013; Kwasnieski et al., 2019; De Renzis et al., 2007), with the potential explanation simply being that mitosis terminates transcription. This minor wave of zygotic transcription at nc 8-12 is followed by a major wave of zygotic transcription of approximately one-third of all genes at nc 14.

Early zygotic transcription is dependent on master-regulatory transcription factors such as the maternally provided Zelda (Liang et al., 2008). Except for ensuring the transcription of the miR-309 cluster (discussed above), Zelda facilitates the transcription of several other zygotic genes. Zelda does this by recognizing specific genomic motifs, forming nuclear hubs, and attracting other transcriptional regulators, such as transcription factors (Liang et al., 2008; Harrison et al., 2011). In addition, Zelda is also suggested to be essential for chromatin accessibility and the first chromatin 3D structure (Brennan et al., 2023; Hug et al., 2017), which is essential for all transcription.

ZGA is also initially limited by the high concentration of maternally loaded histones. These are believed to either cause replication stress affecting the cell cycle, or even to compete with the binding of e.g. Zelda (Chari et al., 2019). As more mitotic events occur, the concentration of unbound histones decreases, and nucleosomes are arranged with specific histone variants that also regulate the transcription. One example is the gradual accumulation of the histone variant H2A.Z in the nucleus before ZGA, which is enriched at several of the genes where Zelda is not bound. This histone variant is needed for the transcription of these genes (Ibarra-Morales et al., 2021). Another example is the embryonic linker histone H1 variant dBigH1 which has the opposite and repressive effect on zygotic transcription (Pérez-Montero et al., 2013) (see Chapter 5).

5  EPIGENETICS AND HIGHER ORDER CHROMATIN STRUCTURES IN THE EARLY EMBRYO

Chromatin has several layers of regulation determining DNA packing density and accessibility. These regulate the cell-to-cell differences in gene expression and are necessary to progress the development. Many
of these regulatory layers are well demonstrated in the early *Drosophila* embryo as compartmentalization and chromatin environments form and mature.

As for ZGA (see Chapter 4), histone variants play a role in chromatin accessibility. Upon fertilization, the paternal histones are replaced by maternally provided histones and soon after, a similar course of events happens within the female pronucleus. One of these replacements is the special linker histone H1, called dBigH1, which can regulate transcription (see Chapter 4) and chromatin compaction (Pérez-Montero *et al.*, 2013). dBigH1 is specific for the developmental time before the MBT. During MBT, a somatic type of H1 replaces the dBigH1 throughout the whole embryo except for in the developing germ cells, which are still transcriptionally silent ((Pérez-Montero *et al.*, 2013) reviewed in (Hamm & Harrison, 2018)).

Modifications on the histone tails are one more layer of important factors in chromatin regulation. There is a plethora of different types of modifications and many different proteins involved in catalyzing or removing them (reviewed in (Mattick & Amaral, 2023)). In general, acetylation on lysines or arginines is usually present in more actively transcribed regions as they neutralize the DNA-histone charge making the DNA more accessible, whilst trimethylation on lysines typically marks the less transcriptionally active facultative or constitutive heterochromatin. Facultative heterochromatin differs from constitutive by being differently organized in different cell types, while constitutive typically marks the chromosome centers (at centromeres) and ends (telomeres) in all cells. The early zygotic chromatin has overall little histone modifications. The earliest modifications are acetylations that can be detected around nc 8, thereafter, more histone acetylation marks are deposited from cycles 8 to 13 (Li *et al.*, 2014). At cellularization and ZGA, methylations and other histone modifications are heavily enriched ((Li *et al.*, 2014; Yuan & O’Farrell, 2016) reviewed in (Hamm & Harrison, 2018)). One of the methyltransferases, Eggless (or SetDB1), has been shown to initiate the formation of heterochromatin in *Drosophila* embryos (Seller *et al.*, 2019). Accumulation of Eggless attracts heterochromatin protein 1a (HP1a) that binds to tri-methylations of the lysine at histone 3 (H3K9me3), which in turn attracts other methyltransferases such as Su(var)3-9, which further maintains the chromatin state.
Higher-order chromatin compartmentalization, such as the formation of topologically associated domains (TADs) is also becoming more prominent at MBT and cellularization, although there are some detected before (Hug et al, 2017). The compartmentation of chromatin gives that e.g. genes or enhancers that lie far from each other on the DNA can be nearby in the three-dimensional (3D) space. In mammals, the TAD boundaries are delimited by regions in which the insulating binding protein CTCF and cohesion bind (reviewed in (Dehingia et al, 2022)). These control chromatin loop formation linking e.g. enhancers and promoters.

In Drosophila, however, there are several other insulating binding factors in addition to CTCF, which have comparably less effect on compartmentalization but might primarily have barrier functions either activating or inhibiting the transcriptional environment. For example, one complex of early embryonic-specific insulating binding factors, the Elba-complex and Insensitive (Insv), was shown to affect zygotic gene activation (Ubeurschär et al, 2019). In paper I, we find a similar increase in embryonic transcription after a stress-induced decrease of Elba1, which is one of the proteins in the Elba complex (Örkenby et al, 2023).

5.1 sncRNAs and their role in regulating chromatin structure
How chromatin-modifying proteins find their target loci on the DNA is not yet fully understood. However, this process is believed to some extent to be RNA-guided. There are several examples from the piRNA pathway in which the piwi-piRNA complex attracts epigenetic factors to TEs to ensure their transcriptional repression in the form of H3K9me3 repressive marks (reviewed in (Tóth et al, 2016)). The piwi-piRNA complex has been shown to interact both with HP1a (Brower-Toland et al, 2007; Wang & Elgin, 2011) which is one of the main factors needed for heterochromatin formation, and with the HP1 variant Rhino (Klattenhoff et al, 2009), which is involved with the creation of piRNA clusters. Interestingly, the piwi-piRNA complex has also been shown to interact with the histone H1, and loss of piwi gives less H1 leading to de-repression of TEs (Iwasaki et al, 2016). In addition, one recent study showed that maternal piRNAs are needed to properly direct the heterochromatin at TE insertion sites during early embryogenesis, proving an
epigenetic role for maternally inherited piRNAs in TE silencing (Fabry et al., 2021).

In mice, there have been several reports of paternal tRFs’ impact on offspring following paternal diet intervention (Chen et al., 2016; Sharma et al., 2016). A later study showed that tRFs play a key role in the production of other sncRNAs, such as snoRNA and certain snRNAs, thereby regulating the levels of histones and transcriptional repression of the TE MERVL (Boskovic et al., 2020).

MiRNAs are well-known for their post-transcriptional gene silencing activity but are unlikely to have a direct impact on chromatin structure. However, it is not impossible for miRNAs to have indirect roles impacting the chromatin structure by e.g. regulating epigenetic factors, such as histone deacetylases or methyltransferases. In Paper I, we found that stress-sensitive miRNAs decrease the Elba1 levels at cellularization, which opens new, but indirect, regulatory roles for this sncRNA in chromatin 3D organization. We believe that in the case of Elba1, the regulatory effects on chromatin organization are modest. Full knockout of any of the genes from the Elba complex does not affect viability, which is affected in CTCF knockout models (Gambetta & Furlong, 2018). However, we still detect effects that are strong enough to affect position-effect variegation (PEV) in the white mottled fly (ln(1)wm4h) (see Chapter 7) indicating that the disruption of this protein interferes with the first constitutive heterochromatin (Örkenby et al., 2023).

6 METHODS TO STUDY SNCRNA

6.1 LABORATORY METHODS

The ability to study sncRNAs has advanced a lot since the development and reduced cost of high-throughput sequencing. At the beginning of the 2000s, Sanger sequencing of the human genome cost between 0.5 and 1 billion dollars (reviewed in (Reuter et al., 2015). Both the sequencing methods as well as the cost have evolved a lot since then and today, sequencing is a relatively common method to study the genome, chromatin environment, transcription, and much more.
There are several sequencing methods available, but Illumina's NextSeq is one of the most common next-generation sequencing platforms used. To study the RNA environment using this system, the extracted RNAs are first prepared by ligating specific adaptor sequences. These are later recognized by the flow cell inside the sequencer. The RNA is reversely transcribed to the more stable cDNA and amplified. For studies involving more than one sample, sample-specific indexes are ligated to the samples and, in the case of sncRNA studies, a size-selection step is performed and RNAs of the desired sizes are isolated. The libraries are then loaded into a flow cell that contains complementary sequences to one of the adapters on the sample DNA, allowing them to bind.

During the sequencing process, the prepared libraries are amplified, creating approximately 1000 copies of each unique sequence, and complementary template strands are created. From these, second strands are synthesized by incorporating different fluorescently labeled nucleotides. One nucleotide is added per cycle, and images are captured simultaneously for all clusters of sequences, with each base having its unique fluorescent signal. This cyclic process is repeated for the desired number of cycles, adding one more base to the sequence length. When sequencing sncRNAs, usually the whole sequence will be detected within 70 cycles (including adaptors and indexes). The high-resolution fluorescent images are used to interpret and convert the unique nucleotide bases and their order (see Fig 4). This raw data of sequence information (fastq) can then be used in downstream bioinformatic analyses.

It is important to be aware that modifications on the sncRNA molecules can affect which RNAs will be captured using different library preparation methods (Kugelberg et al, 2021). For example, when studying tRFs or piRNAs, modifications to the fragment can introduce biases in library preparation and sequencing results (Su et al, 2020b).
Figure 4. Schematic illustration of fluorescent nucleotide incorporation during a sequencing run. Fluorescent nucleotides are incorporated into complementary bases on the template strand. The fluorophores are removed between each cycle, and new ones can be incorporated. As one nucleotide is incorporated per cycle, the unique fluorescent signals can be captured per cycle and flow cell cluster. When the sequencing run is over, the signals are interpreted as bases and a fastq file containing sequences and quality scores can be retained.

In addition to sequencing-based techniques, several other methods are used to study sncRNAs, serving different purposes in sncRNA research. Reverse transcription polymerase chain reactions (RT-PCR) and quantitative PCR (qPCR) are popular methods to determine the presence or quantity of specific sncRNAs or genes in a sample. RT-PCR involves reverse transcription that converts RNA to complementary cDNA and amplification using primers designed for a specific gene or e.g. miRNA. Another popular method to detect and quantify specific sncRNAs is Northern blotting, in which RNA is separated by size through electrophoresis, transferred to a membrane, and hybridized with a specific labeled probe for the sncRNA of interest.

In Situ Hybridization and Fluorescent In Situ Hybridization (FISH) can be used to study the localization of sncRNAs within a tissue. Also this method relies on the hybridization of specific RNA probes (fluorescently labeled in FISH) and imaging.
For studying the functions and effects of sncRNAs, over- or underexpression of specific sncRNAs can be used. Techniques such as RNA interference (RNAi) or the newer CRISPR-Cas9 are common to silence or reduce the expression of specific targets. The RNAi system uses sncRNAs and RISCs to silence complementary target RNAs (see Chapter 2), leading to their knockdown. The newer CRISPR-Cas9 on the other hand, can introduce mutations or delete specific genomic sequences, effectively knocking out target sncRNA. miRNA-sponges is an additional method to reduce the number of active miRNAs (Ebert et al, 2007). Here, the miRNA sponge acts as a “mock target” that is introduced through a gene construct. The miRNA sponge contains several miRNA binding sites and can therefore efficiently block miRNA function. To study the effect of overexpression, extra copies of sncRNA can be introduced into the genome. This is often achieved through the CRISPR-Cas9 system.

To study sncRNA-protein interactions, several methods are based on crosslinking the RNA and the protein of interest. The specific protein is immunoprecipitated and isolated, and the RNA is then released. These methods can be followed by, for example, sequencing or qPCR.

A luciferase assay can be used to experimentally validate sncRNA-target interactions, such as miRNA-target silencing (Clément et al, 2015). In this method, cells are transfected with a reporter construct containing the 3'UTR of the target gene downstream of the luciferase gene as well as the miRNA of interest. If the miRNA indeed binds to the target 3' UTR, it will affect the expression of the luciferase reporter gene, which can be measured using a substrate and light emission quantification.

6.2 BIOINFORMATIC ANALYSIS

As previously discussed, sequencing is an increasingly popular method to study sncRNAs. This has led to a constantly increasing amount of big datasets, where one sncRNA fastq file can be between 1-5 Gb if generated using single-end sequencing for 70 cycles. If the cycle number is increased, the file sizes can be significantly larger. This exponential increase in data demands efficient methods for data analysis and interpretation.
The initial workflow for analyzing total RNA, mRNA, and sncRNA sequencing shares the same basic principles. The pre-processing of the fastq reads includes removing the adaptor sequences that were introduced during the RNA library preparation and quality filtering. Quality filtering removes low-quality reads that have a high number of bases with low quality scores from the sequencing. In many workflows, the pre-processed reads are subsequently aligned to reference genomes. In this process, overlapping RNAs that align to the same feature are grouped and counted under that specific feature. This information is usually then summarized into count tables and used for downstream analyses. The initial analytical steps usually require access to robust Linux-based systems due to the computational demands, storage requirements, and code compatibility. Alternatively, these steps can be run remotely through access to a research computing center or a computer cluster, such as Uppmax.

Analyzing sncRNAs presents a number of challenges when compared to mRNA analysis. Firstly, counting features are effective for mRNA and in many cases also for miRNAs, but it becomes less suitable for highly diverse sncRNA classes such as piRNAs. Many sncRNAs come from repetitive sequences that are distributed across several genomic locations, making it impossible to determine their origin. In addition, the study of sncRNAs can be difficult as several classes are templates for other classes, e.g. both snorRNAs and rRNAs template and hosts miRNAs and piRNAs (see Chapter 2). Because of this, several sncRNA reads can align to different species within the same sncRNA class, or even to different sncRNA classes. It is therefore beneficial to save multiple alignments for each sncRNA read.

One common approach to address the issue with multiple mapping is hierarchical annotation, where the user specifies the order of reference databases for which the fastq files should be matched. Once a hit is found within a reference, the sequence is annotated accordingly and no longer compared against other references. This approach is useful when investigating a single, well-characterized group of sncRNA, such as miRNAs. Another approach for handling multiple mappings is to consider all reference hits and divide the count number by the number of hits, assigning a divided count score to each hit. An additional way to handle,
or at least bring to notice, multiple mappings is to count sequences rather than features before the alignment. This approach significantly reduces the data size compared to raw data, especially if used together with a noise filter removing single-occurring sequences (noise) that can be a substantial portion of the data. Annotation to reference genomes can then be performed much quicker, and all alignments can be retained alongside the read sequences. In Paper II, we implemented this later method (Skog et al, 2023).

Downstream analysis of sncRNA is much dependent on the research question. However, it generally includes several of the listed essential tasks:

- Differential expression analysis, where the expression of sncRNAs is compared across different conditions or samples.
- Annotation and classifications, where the sncRNA are characterized depending on features of sncRNA class or genomic origin.
- Clustering analysis, where similar expression patterns are visualized to reveal trends or patterns.
- Target prediction analysis, where the binding compatibility (e.g. in seed regions) and binding energy are taken into account. This is often followed by a target enrichment analysis, such as gene ontology analysis.

Although computational prediction models have advanced, additional experimental validation is beneficial to provide robust evidence of sncRNA-target interactions and regulation.

## 7 Methodology Used in This Thesis

### 7.1 Why Drosophila?

*Drosophila* has been a popular model organism for biological studies for more than a century and there is therefore a plethora of well-characterized genetic tools available for studies in this model. Studies within *Drosophila* have historically led to several important findings such as basic concepts of chromosomes, mutations, and genes (not least the Hox genes) laying the foundation for several of the most basic biological
concepts used today (reviewed in (Letsou & Bohmann, 2005)). Studying early development, the *Drosophila* embryo is very beneficial to work with as it is relatively easy to get a large sample number, and the embryogenesis is both rapid and well characterized. In addition, the development occurs outside the mother, which facilitates staging in living embryos.

Although several fundamental principles are relatively similar between species, there are a great number of exceptions. *Drosophila* (as well as *C.elegans* and yeast) has for example very little DNA cytosine methylation, which is one of the key regulatory principles of chromatin in mammals (reviewed in (Greenberg & Bourc’his, 2019)). Findings from studies using any model should therefore always be carefully tested and evaluated before drawing general conclusions involving other species.

### 7.2 Sampling and Choice of Intervention

For Paper I in this thesis, we have sampled *Drosophila* embryos of the earliest embryonal stages (stages 1-5, covering egg laying to ZGA) and heat-shocked and control stage 5 embryos. The snRNA seq datasets of embryos of stages 1-5 created for Paper I were used for Paper III and for the test data set included in the Seqpac R-package (Paper II).

To facilitate the staging, the egg collections were performed in very narrow intervals of 30 minutes. In addition, the collected eggs were kept in an incubator holding a constant temperature of 22°C until dechorionation. The staging of embryos was performed using a bright field microscope and the criterias for embryo staging that are described in (Bownes, 1975). Counting the number of nuclei would have given a more exact staging, but would require either a fly strain expressing e.g. a fluorescent H2A.Z (Lott et al, 2011), fixation of embryos, or a much longer time for inspection. We reasoned that since we sampled for RNA extraction, fixation was not an alternative, and keeping the time as short as possible was in our prime interest. In addition, in the case of the embryos exposed to stressors and their matched controls, we could validate the staging transcriptionally after RNA seq. For details on sample collection and transcriptional validation, see Paper I (Örkenby et al, 2023).
In Paper I, we used heat shock stress as a means to explore stress-sensitive embryonic periods. This stressor was chosen as we needed a temporary stressor that could be precisely distributed and could be applied during all stages of development. A similarly performed heat shock intervention in young embryos was reported not to affect the lethality (Seong et al., 2011). However, we experienced difficulties with early embryonic lethality using a similar protocol. We therefore decreased the intervention from one hour to 30 minutes. We did, however, detect that 30 minutes of heat stress is enough to affect position-effect variegation in the white mottled strain if exposed before MBT (Örkenby et al., 2023).

7.3 Pros and cons of using Position-effect variegation for studying the chromatin environment

In position-effect variegation (PEV), a reporter gene lies within close contact with the heterochromatic barriers, making the expression of the reporter gene sensitive to the spreading of heterochromatin (reviewed in (Elgin & Reuter, 2013)). In Paper I, we used the \textit{ln(I)wm4h} (Muller, 1930) that uses the \textit{white} gene, responsible for the wild-type red eye pigmentation, on chromosome X as the reporter gene (Örkenby et al., 2023). The advantage of this reporter is easy tracking of heterochromatin spreading as this is visible on variegating eye pigmentation. It is therefore possible to measure the effect on a large number of flies, which is an advantage when studying the effect of stressors where the individual variation can be large.

On the downside, only one genomic locus of heterochromatin spreading is measured per PEV strain. To get a fuller view of the chromatin landscape, other methods such as Assay for Transposase-Accessible Chromatin (ATAC) sequencing, where the global genomic accessibility is captured, should be used (Grandi et al., 2022). However, this method is much more costly and should preferably not be used for screening of e.g. stress-sensitive developmental periods, but rather for evaluating the effect of a defined intervention or knockdown/knockout model.
7.4 Validating the Connection between miRNAs and Elba1

To study the effect of a specific gene or protein on a system, or to interrupt a specific pathway, knockdown or even knockout models of a specific factor can be used in several model systems, including Drosophila. Completely or partly removing one protein usually gives greater measurable effects than studying the effect on the system after an environmental exposure or intervention.

In Paper I, we used a mutant strain for Dicer-1 as one way to validate that the reduced level of Elba1 following stress is dependent on a functional miRNA biogenesis pathway. We analyzed the Elba1 RNA levels using qPCR in heterozygous eggs and found that this was correct. We further performed an RNA immunoprecipitation for Ago1, to validate that the Elba1 transcript indeed interacts with the miRISC. We also included an in silico miRNA analysis based on TargetScan Fly’s script (Agarwal et al., 2018) where we found one candidate miRNA having a 6-mer complementary seed sequence at the Elba1 3’UTR and several miRNAs having 7-mer seed sequences in the coding sequence of Elba1.

7.5 Bioinformatic Analysis of the sncRNA Seq Data

During the work with Paper I, the Seqpac package (Paper II) was still under development. The pre-processing of the sncRNA seq data was therefore performed using adaptor trimming from Cutadapt (Martin, 2011), quality filtering using FastQC (Andrews, 2015) as well as the SPORTS.1.0 pipeline (Shi et al., 2018) for counting sequences and aligning to reference genomes using a hierarchical strategy. However, some premature structures and functions later integrated within the Seqpac package were used for downstream analyses such as the PAC-list object (see Paper II) and premature versions of “pheno targets”. “Pheno targets” use information from the experimental metadata (Pheno), for subsiding e.g. normalized counts based on a defined feature, such as intervention, for analysis.

The finished, and extended, target object feature is greatly useful e.g. for subsiding sncRNA in the dataset based on annotation features, such as all reads aligning to miRNAs, piRNAs, or both. This was used both in Paper II to demonstrate how Seqpac can be implemented, and also in Paper III. In Paper III, the feature was first used to exclude reads that
did not align to the *Drosophila* reference genome, and later to subside the reads aligning to piRNAs, and had secondary alignments to TEs or gene features. In addition, the feature was used together with the Seqpac integrated DEseq2’s build-in Wald test after negative binominal fitting (Love *et al*, 2014) for differential expression analysis of piRNAs between different embryonal stages. Similarly, the feature was used for principal component analysis.

From pre-processing to advanced piRNA analysis, the Seqpac package (Paper II) was used for almost all analyses in Paper III, with some minor exceptions. Several of the functions developed for tRF analysis were successfully used for analyzing potential piRNA targets such as TE classes, subgroups, and target coverage. For future piRNA analyses, it would be beneficial to integrate functions for piRNA ping-pong signatures and potentially also for detecting piRNA phasing processing.
8 SUMMARY OF PAPERS AND RESULTS

Paper I - “Stress-sensitive dynamics of miRNAs and Elba1 in Drosophila embryogenesis” examines the dynamics of the small non-coding RNA (sncRNA) environment, focusing on miRNA, at the very beginning of Drosophila embryogenesis, and how this environment is influenced by stress and the subsequent implications it has.

Paper II - “Seqpac: a framework for sRNA-seq analysis in R using sequence-based counts” describes current difficulties with bioinformatic analysis specific to sncRNA and presents a new R-based tool (Seqpac) that deals with several of these issues. Using the Seqpac framework, the user can analyze sncRNA from fastq (raw data) until publication-ready figures on a standard non-Linux laptop without losing the sequence integrity.

Paper III (manuscript) - “piRNAs during the maternal-to-zygotic transition in Drosophila melanogaster” examines the piwi-interacting RNA (piRNA) population during the maternal to zygotic transition in Drosophila embryogenesis. Using Seqpac (Paper II) we were able to analyze piRNA complementarity to transposable elements as well as to maternally loaded mRNA.

Paper I - STRESS-SENSITIVE DYNAMICS OF miRNAs AND ELBA1 IN DROSOPHILA EMBRYOGENESIS

Relating to the developmental-origins-of-health-and-disease hypothesis, major public health issues such as obesity, heart disease, and mental health problems may have their etiology in sub-optimal conditions during sensitive periods of development (Suzuki, 2018). In Drosophila one such stressor could be heat shock (Seong et al, 2011). In this study, we investigated the impact of environmental stress on the expression of sncRNA during the initial stages of Drosophila embryogenesis, with the primary objective of identifying a vulnerable developmental period in which stress could potentially induce persistent memories. We additionally aimed to examine the simultaneous expression of genes and sncRNA within individual Drosophila embryos to uncover crucial interactions between sncRNA and genes during this sensitive window.
To do so, we used heat shock treatments at various time points during embryogenesis to identify the most sensitive period in which stress induces lasting effects on the chromatin state. To achieve this, we used the white-mottled (wm4h) position-effect-variegation fly strain, which possesses an eye color marker that enables us to track the spreading of pericentric heterochromatin on chromosome X. Our findings indicate that stress experienced before the midblastula transition (MBT) has the greatest impact on inducing enduring effects.

Using this insight, we characterized the small non-coding RNA (sncRNA) environment using sncRNA sequencing during the earliest stages of embryogenesis, including the initial nuclear divisions until the cellularization and formation of mature heterochromatin. We observed variations in multiple sncRNA biotypes across different developmental stages, particularly within the miRNA family.

To examine the effects of heat shock during the identified pre-MBT period, we conducted sequencing of both total RNA and sncRNA from cellularized embryos. Intriguingly, our analysis revealed the upregulation of 184 distinct miRNA sequences following heat shock. These miRNAs primarily consisted of retained maternal miRNAs that would typically undergo degradation during the maternal-to-zygotic transition (MZT) under normal conditions.

By comparing the increased miRNAs to downregulated total RNA, we observed an inverse correlation with a set of early zygotic genes that undergo transcription before zygotic gene activation. One gene in particular, Elba1, which plays a role in boundary formation, drew our attention. Using a Dicer-1 knockdown fly strain, we found that functional miRNA biogenesis is necessary for the downregulation of Elba1 following heat shock during the sensitive period. Furthermore, RNA immunoprecipitation of the Argonaute1 protein provided additional confirmation of the interaction between miRNA-RISC and Elba1.

As Elba1, along with its counterparts Elba2 and Elba3, forms an insulating binding complex, we investigated the relationship between upregulated genes and these factors. Our analysis revealed a strong inverse correlation between these factors and a cluster of genes, indicating a leaky genome at these specific loci due to the loss of Elba1. To validate this, we used CUT&RUN and observed a reduction in Elba1 binding near the
transcription start sites of these genes following heat shock. To further confirm the role of Elba1, we crossed knockout strains of the Elba family members with the \( w^{mth} \) strain and observed that the loss of these factors replicated the effects seen after heat shock on the PEV reporter.

Based on these findings, we propose a model wherein stress during the sensitive pre-MBT period leads to the retention of maternal miRNAs that target Elba1, ultimately resulting in a leaky genome.

**Paper II - Seqpac: A framework for sRNA-seq analysis in R using sequence-based counts**

In recent years, RNA sequencing has become increasingly affordable and available, making sequencing a common method for studying global RNA expression. The bioinformatic pipelines for handling this data type have evolved at the same pace and have become more accessible for researchers within the biomedical field. However, the bioinformatic processing of sncRNA data still presents various challenges and hurdles that necessitate alternative approaches for effective handling.

The R environment offers a multitude of powerful statistical tools, and a user-friendly interface, and is already extensively used by biologists for data analysis. In line with this, we have developed a comprehensive framework, Seqpac, for studying sncRNA using the R language. This framework incorporates various innovative strategies for addressing challenges such as multiple mapping and complex group designs. Our primary objective in constructing this package was to provide a robust and user-friendly alternative for sncRNA seq analysis without lowering the quality of analysis or visual presentation.

After sequencing and de-multiplexing, the raw sequence data is stored as fastq files, which contain all bases for each read sequence along with quality scores for each base. In the Seqpac package, we have integrated functions for adaptor trimming, quality filtering, and counting unique sequences, taking fastq files as input. Since counting sequences rather than features results in large files, Seqpac’s “make_counts” function comes with a low-level user-defined filter in which it is possible to remove e.g. sequences that occur only once in a single sample, efficiently
reducing the file sizes without losing sequence depth. This feature enables efficient storage and analysis of sequence data while still retaining valuable information.

In seqpac, we have implemented annotation functions that preserve both mapped features as well as the read sequences, ensuring easy sequence accessibility. This allows for flexible annotations to additional reference genomes or RNA species. The Seqpac framework is centered around a unique set of functions that constructs and analyzes a PAC-object, which encompasses experimental metadata (Pheno), Annotation, and Counts tables. The subsequent analyses within seqpac are built upon this PAC object, allowing for streamlined and cohesive downstream analyses. To enhance the ease of use of the package, we have included pheno- and annotation target objects with easily changeable parameters in several functions. The user is therefore able to continuously track the analysis process and test new parameters using the extensive visualization capabilities provided. In addition to differential expression, nucleotide bias, size distributions, and biotype comparisons, Seqpac also offers detailed coverage analysis over user-defined reference files.

In the paper, we provide a few examples of applications. For example, by reanalyzing publically available sncRNA datasets using Seqpac, we found that Mycoplasma infection will influence the overall sncRNA read expression in the infected sample even after excluding reads associated with mycoplasma. Overall, our approach emphasizes the importance of preserving the integrity of the sequence in the data. We underscore this significance and provide a comprehensive framework, Seqpac, to facilitate such analyses and maintain data integrity.

**PAPER III - piRNAs DURING THE MATERNAL TO ZYGOTIC TRANSITION IN DROSOPHILA MELANOGASTER**

Approximately one-fifth of the *Drosophila* genome consists of transposable elements (TEs) (Biémont & Vieira, 2006). TEs are selfish genomic elements that can, driven by various mechanisms, copy and paste themselves into the genome. While they have played an important role in evolution, they are a constant threat to genome integrity as they have the
potential to transpose and disrupt gene coding sequences. The main defense against TE transposition is the piwi-interacting RNA (piRNA) silencing pathway (reviewed in Huang et al., 2017). This class of sncRNAs is particularly active in germ cells, gonads, and during early embryogenesis where genome stability is of extra importance.

Precursor piRNAs are typically transcribed from inactive TEs, but can, for example, also originate from genes. These precursors undergo processing through either a Zucchini-dependent pathway, generating several new piRNAs, or a pathway involving the two piwi-proteins Aubergine (Aub) and Argonaute 3 (Ago3). The latter pathway not only amplifies piRNAs but also silences active TEs through endonuclease activity. In this manuscript, we aimed to characterize the piRNA landscape in Drosophila during the maternal-to-zygotic transition (MTZ), a critical period when the embryo shifts from maternal control to independence.

To achieve this, we used a sncRNA dataset produced from embryos of stages 1-5, which covers the early nuclear divisions up to the zygotic gene activation. We reanalyzed this dataset using Seqpac (Skog et al., 2023) which is an R-based sncRNA framework well-suited for diverse sncRNA classes, even those that lack a structured nomenclature, such as piRNAs.

We first compared the overall sncRNA compositions at each developmental stage. Intriguingly, we found that piRNA peaks in stages 1 and 4, indicating two different piRNA bursts during this developmental period. The first burst is likely maternally provided, as stage 1 corresponds to the period when maternal RNA decay has had minimal impact. The second burst likely consists of either processed maternal piRNA precursors, or newly emerging zygotic piRNAs. Using principal component analysis (PCA) and Euclidean clustering of the unique piRNA reads, we separated these two stages from each other. We found that although the unique piRNAs had different expression profiles, their target alignments were remarkably similar. These targets primarily included TEs but also a considerable portion of genes.

Of the genic piRNAs, we observed a large proportion that was potentially targeting maternal mRNAs and that this portion was higher in
stage 4 compared to stage 1. This suggests a role for piRNAs in maternal mRNA decay, which is in line with previous findings of piRNA regulation of the body-axis determinator nanos (Barckmann et al, 2015; Dufourt et al, 2017).

Regarding the TE-targeting piRNAs, we found that piRNAs predominantly target TEs in the antisense orientation. Furthermore, we discovered that the top targeted TE superfamilies included Bel-Pao, Gypsy, Jockey, and R1, all of which are retrotransposons but from both the Long Terminal Repeat (LTR) and the Long Interspersed Elements (LINE) classes. Doc, a LINE retrotransposon, was significantly more targeted in stage 4 than in stage 1. Notably, the piRNAs aligning to this TE were almost exclusively in the antisense orientation. To our interest, the R1A1-element, which showed the highest enrichment in stage 4 (although not significant), had more piRNAs aligning in the sense than antisense orientation.

Based on these findings, we conclude that maternally loaded piRNAs undergo a replacement process during the MZT, resulting in newly generated piRNAs that differ in sequence from the maternally loaded. However, the alterations in these piRNAs predominantly affect their sequences, with only minor variations in their target preferences.
9 CONCLUDING REMARKS AND FUTURE PERSPECTIVES

The sncRNA field is constantly evolving and more knowledge about their regulation, pathways, and functions is unlocked every day. A search in Pubmed using the terms “miRNA” or “piRNA” generated 19,132 and 304 article hits respectively when only including those published last year (2022). It is therefore humbling to reflect on one’s own contributions to the field. In this thesis, I have summarized some of the known key functions of small non-coding RNAs’ impact on early *Drosophila* embryogenesis and put them in relation to the findings of our studies (Papers I, II, and III).

We found that the sncRNA expression is highly dynamic during the maternal-to-zygotic transition and that there is a clear distinction between the maternal and the zygotic contributions of miRNAs and piRNAs. We have found that maternal miRNAs can have abnormal functions if retained and present in periods where they should not be. This regulation involves targeting one architectural factor, Elba1, which is temporally expressed during a short embryonic period. Loss of this factor leads to increased zygotic transcription, potentially due to leakiness causing the spreading of active and loss of repressive chromatin marks.

We discovered that the main difference in piRNA populations before and after zygotic independence is mainly due to differences in sequences, and not target preferences or origins, which was our initial hypothesis. This nuanced replacement needs further investigation, as it might lead us towards a better understanding of TE repression during early embryogenesis. Finally, we have provided examples of why it is important to never give up the actual sequences in favor of features in sncRNA studies. In this work, we demonstrate that doing so can lead to hidden biases that are not transparent for either the readers or the researcher herself. We further present a potential solution, addressing several of the pitfalls of analyzing sncRNAs, packaged as a comprehensive R-tool for sncRNA seq analysis.

To continue and develop the work based on the findings presented in this thesis, I would suggest that further investigations could include the study of the initial *de novo* transcription of embryonic sncRNAs. Based on their relatively short length (precursors are also relatively short), it is likely that these are some of the first ones transcribed. In Paper I, we
found indications of this from the miR-309 cluster already in stage 2, corresponding to nc 2-8. As the upregulation of miRNAs is a known stress response (Leung & Sharp, 2010), this could be a clever way for the embryo to further control and protect itself from a non-beneficial environment. Providing proof of this would, however, include other methods such as PRO-seq or GRO-seq to detect de novo transcription.

Furthermore, other interesting experiments could include the analysis of chromatin 3D structures following stress in the sensitive pre-MBT period using e.g. HiC. This method, together with CUT&RUN of certain histone modifications (e.g. H3K9me3), could give additional pieces of information on how pre-MBT stress influences the ZGA through chromatin organization or barrier functions. In addition, tracking a candidate miRNA, e.g. mir-283, using the FISH technique and simultaneously stain for a stress granule marker would be of great interest to further examine the mechanisms for maternal miRNA retention.

To gain a better understanding of piRNA function in early embryos, matched sequencing datasets containing total RNA (including TEs) and sncRNAs from the same RNA extractions during MZT could be used to further investigate their interactions during this developmental window. In the future, we hope to also integrate a workflow for analyzing long RNA seq reads in Seqpac, facilitating these kinds of comparisons.

In addition to exciting new knowledge to the field, we have also provided publicly available sequencing datasets, which could continue to be of value to the field. For Paper I, we produced sncRNA datasets that, in fine resolution, cover the maternal-to-zygotic transition. These datasets are freely available for any researcher to use, e.g. analysis of tRFs, as they are publicly available on SRA. For Paper I, we also produced matched sncRNA and total RNA seq data prepared from the same RNA extractions, which could be of great value.

The framework for sncRNA seq analysis described in Paper II is also a contribution to the field as this R-package enables sncRNA analysis on an ordinary Windows laptop. More researchers can therefore, with no extra expenses, mine the jungle of available sncRNA datasets, potentially leading to exciting new discoveries.
10 References


Bergsten SE & Gavis ER (1999) Role for mRNA localization in translational activation but not spatial restriction of nanos RNA. *Development* 126: 659–669


Boskovic A, Bing XY, Kaymak E & Rando OJ (2020) Control of noncoding RNA production and histone levels by a 5’ tRNA fragment. *Genes Dev* 34: 118–131


Drosophila embryo is determined by transcription factor pioneering and enhancer activation. *Dev Cell*


Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal Vol 17 No 1 Gener Seq Data Anal*

Mattick J & Amaral P (2023) RNA, the Epicenter of Genetic Information: A New Understanding of Molecular Biology CRC Press

Muller HJ (1930) Types of visible variations induced by X-rays in Drosophila. *J Genet* 22: 299–334


Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

https://doi.org/10.3384/9789180753395