GENOME DUPLICATION EVENTS AND SMALL RNA EXPRESSION AND FUNCTION

A thesis submitted to the University of Manchester for the degree of MPhil in the Faculty of Biology, Medicine and Health

2021

LUZ ROSALINA TINCOPA MARCA

SCHOOL OF BIOLOGICAL SCIENCES

Blank page

TABLE OF CONTENTS

Final Word Count: 22809

LIST OF FIGURES

LIST OF TABLES

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

A note on the alternative thesis format

This MPhil thesis is formatted in the alternative style, such that abstract, introduction, methods, results, and discussion are presented as stand-alone manuscripts intended for publication in peer-reviewed journals.

Manuscripts are preceded by a broad introduction that covers an overview of gene duplication, miRNAs, piRNAs. The manuscripts are followed by a general discussion and references.

DEDICATION

To science.

ACKNOWLEDGEMENTS

I am very thankful to my sponsor, the Peruvian Government, the National Council for Science, Technology and Technological Innovation – CONCYTEC. They covered all my expenses during my research at The University of Manchester.

To my family who supported me during my research.

To my supervisor, Sam Griffiths-Jones, for accepting me in his group and for his guidance.

To my co-supervisor, Matthew Ronshaugen, for his guidance and suggestions to improve my thesis analysis.

To Jamie Soul to guide me in my thesis analysis and from whom I learned most of the R programming, to Robert Maidstone, who help me in the statistical analysis and some analysis for my first manuscript, Michael Nelson and Craig Lawless from whom I asked for some support in computational issues and Linux. A Kunal Chopra for his suggestions in my thesis writing.

To Daniel Leite and Alistair McGregor for provided me information of the small RNA seq data, the accession number of the RNA seq data, the protein annotation and genome from *P. tepidariorum*, and for all additional information that they provided me in my thesis analysis.

To Dave Gerrad, Ana Kozomara and Thomas Bleazard, Mark Reardon, Edith, Steve, Zhengxue, and other members of the lab, who helped me in my thesis analysis and with whom I shared good moments in the office.

To David Robertson who gave me feedback from my first manuscript, and it was also good friend to talk about science, to Jean-marc Schwartz, from whom I learned to teach in a classroom and who also was a good friend who invited me to his lab social activities. To Simon hubbard from whom I also learned how to teach in a classroom, and I worked in his group for a short period.

I am grateful with Louise Hyde, who gave me support in the corrections and submission process.

I am also grateful with people who support me in my thesis analysis, and I forgot to mention them.

To my international and latinoamerican friends who also help me in my thesis analysis and were my social support in special to Samaneh, Zhisong, Moises, Stefano, Kunal, Ruben, Natalia, Javi, Raymundo, Namrita, Shalaw, Bihn, Felipe, Diana, Pablo, Andres, Javo, Gabriela,

Andy, Paulo, Oscar, Mariela, Giulia, Ana, Alex, Blanca, Ghader, Hirra, Johanna, Leonard, Nathalia, Igor, Zhabiz, and other people that I forgot their names.

.

Introduction

1. Introduction

1.1. Gene duplication

Genome duplications are an important evolutionary process, which can drive genomic diversity and create novel genes in eukaryotic genomes (Li, Musso, and Zhang 2008). Duplicated genes can arise at whole genome duplication (WGD) events and by small-scale duplication (SSD) (Figure 1A and Figure1B). Unequal cross over, retroposition, or segmental duplication are processes that result in SSD (Figure 2) (Zhang 2003). It has been proposed that vertebrate genomes have undergone two rounds of whole genome duplications. However, an alternative hypothesis is that there have been one whole genome duplications followed by extensive tandem or segmental duplication. Work in *Arabidopsis thaliana* showed that WGD had a strong impact on the gene set, but also highlight the importance of genes from small-scale duplications involved in metabolism, stress, and survival. WGD genes were enriched in transcription factors and transduction genes (Carlos and Ramirez-parra 2015).

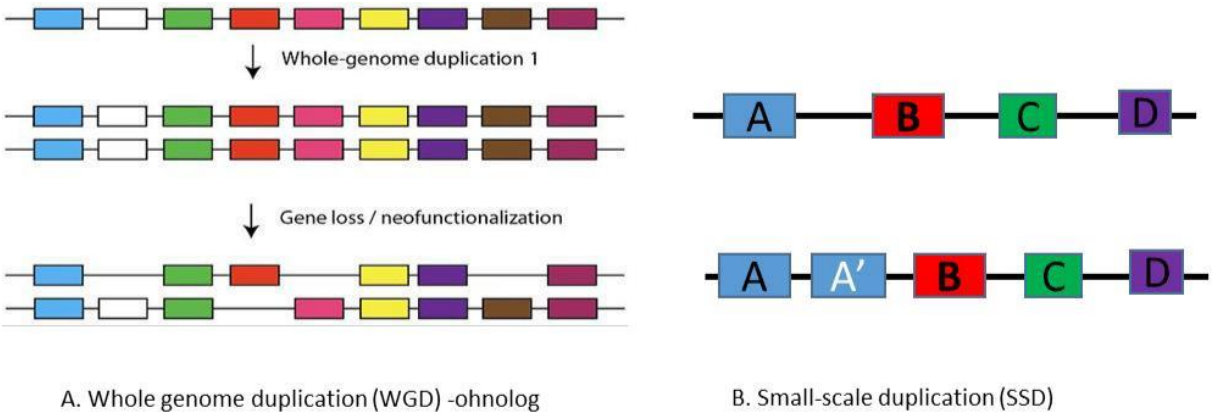

Figure 1.1. Duplicated genes. A. Whole genome duplicated genes and B. small-scale duplicated genes (modified from Venema 2011).

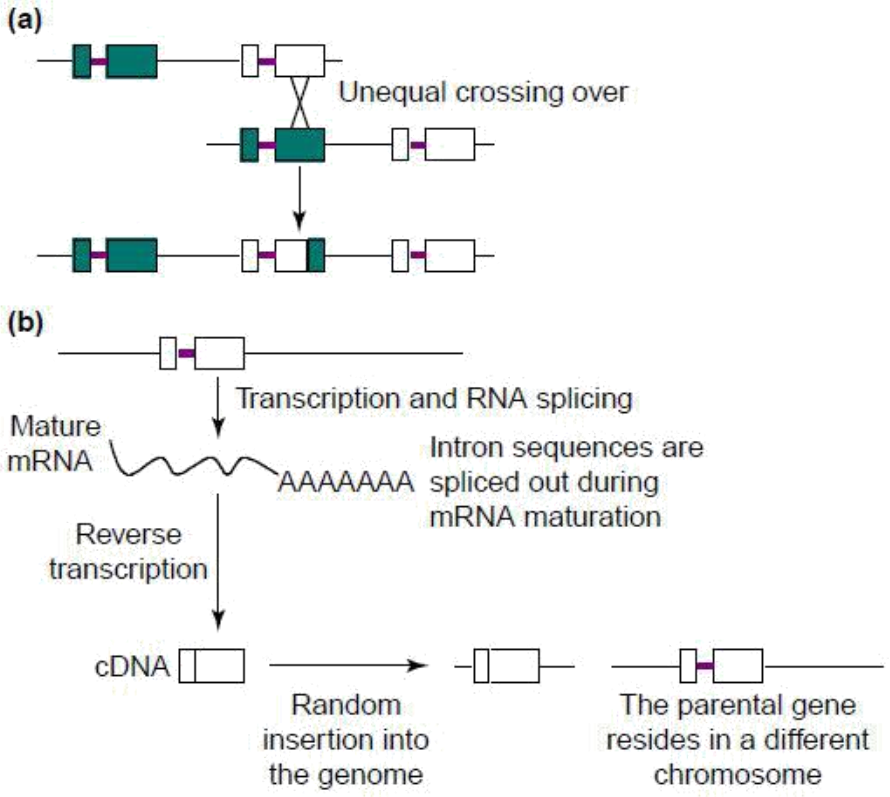

Figure 1.2. Two common modes of small-scale duplication (a) Unequal crossing over, which results in a recombination event in which the two recombining sites lie at non-identical locations in the two parental DNA molecules. (b) Retroposition, which occurs when a message RNA (mRNA) is retrotranscribed to complementary DNA (cDNA) and then inserted into the genome (modified version of Zhang 2003).

1.2. Gene dosage

Gene dosage is referred to the number of copies of specific gene in a cell. Changing the dosage of a gene by duplication can cause an imbalance in the stoichiometry of proteins in the cell. For example, changing the relative amounts of proteins in a complex might have adverse effects on the function of that complex (Birchler and Veitia 2010). It is proposed that whole genome duplication events do not perturb this equilibrium. However, duplication of fragments of DNA that contain genes encoding members of protein complexes will be more likely to have a dosage effect. Gene dosage can cause deleterious effects. For example, Charcot-Marie-Tooth disease type 1A (CMT1A) is a disease caused by a segmental duplication in the chromosome 17 p12-p11.2. Duplication of the PMP22 gene results in over-expression of the protein product, which can affect the fitness of the organism (King et al. 1998).

1.3. Fate of duplicated genes

A single gene duplication can end in different fates. One of the most common cases is when one copy loss its function and the other maintain it (nonfunctionalization; see Figure 3a). There are rare cases when one copy retains the original function and the other acquire a new function (neofunctionalization; see Figure 3b). In other cases, after a duplication event, the two genes acquire mutations that give them complementary function (subfunctionalization; see Figure 3b). Finally, there are cases when two copies are retained maintaining their function, so the organism may acquire genetic robustness against mutation (Figure 3c).

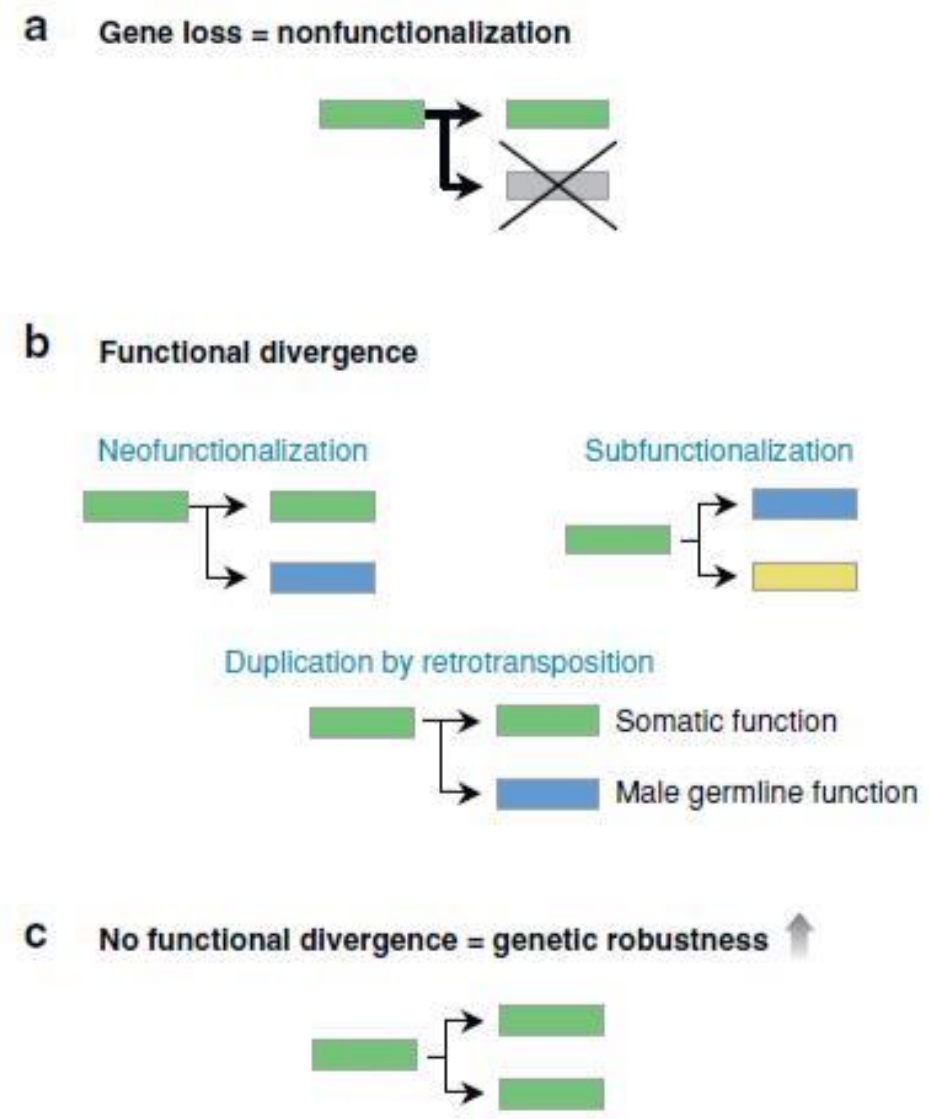

Figure 1.3. Evolutionary fate of single gene duplication (a-c). when one of copy loss function (a, nonfunctionalization); In rare instances, the functional duplicate gene copy and the ancestral gene diverge in function (b, neofunctionalization)

Retrotransposition is a special case of neofunctionalization, when the two copies evolve and acquire new function (b, subfunctionalization) and when two copies are retained without functional changes give as a results robustness (Conrad and Antonarakis 2007).

1.4. MicroRNAs

MicroRNAs (miRNAs) are small RNAs from 21-25 nucleotides long. Since the discovery of the first miRNAs, lin-4 that represses lin-14 mRNA and let-7 that represses lin-41 mRNA, in *C. elegans*, miRNAs have been found to be widespread and numerous in all animal and plants genomes, and in some viruses and single-celled organisms.

1.5. MicroRNA Biogenesis

In animals, three main proteins are involved in the biogenesis of miRNAs: Drosha, Dicer and AGO. In plants, there is no Drosha; Dicer replaces its function. MiRNAs are transcribed by RNA polymerase II. The Drosha-DGCR8 complex makes a sequential cleavage of the primary precursor (pri-miRNA) to convert it into 70-nucleotide precursor hairpin (pre-miRNA) in the nucleus. The pre-miRNA is then exported by exportin-5 to the cytoplasm, there Dicer-TRBP in mammals processes the pre-miRNA to produce a duplex of two 20-nt sequences, called the mature miRNA/miRNA* duplex. One strand of this duplex will be bound by the RNA induced silencing complex (RISC), which will target complementary mRNAs for translational repression, deadenylation or degradation. The choice of which strand of the mature miRNA duplex goes to form of RISC-complex is thought to depends on the instability of duplex at the 5´ end of the strand (Krol, Loedige, and Filipowicz 2010) (Ameres and Zamore 2013). Around half of animal miRNAs are processed from the introns of protein-coding genes, and around a third are found clustered with other miRNAs in the genome (Figure 1.4).

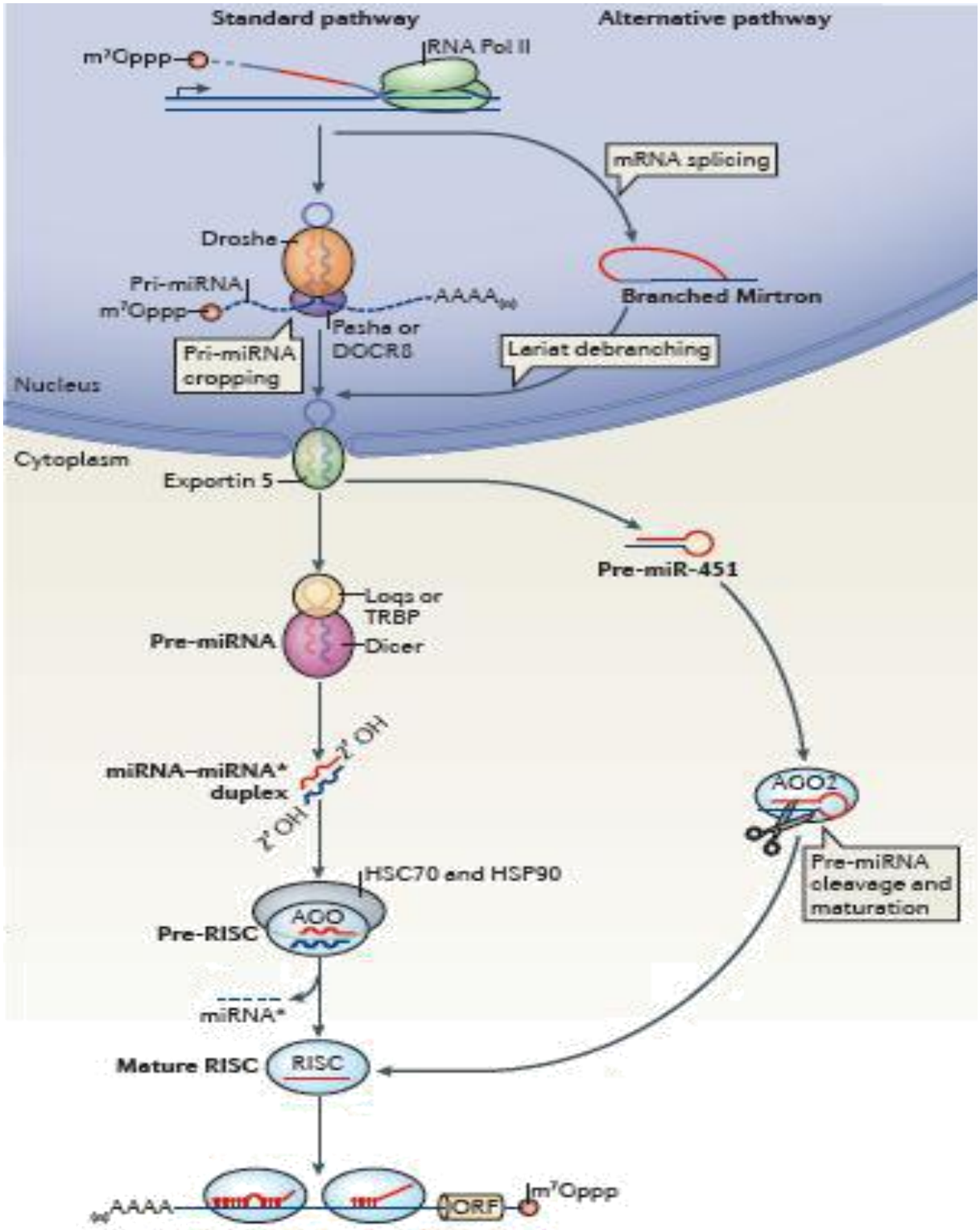

Figure 1.4. MiRNA biogenesis. In the standard miRNA (miRNA) biogenesis pathway, primary miRNA (pri-miRNA) transcripts are processed by Drosha in the nucleus and by Dicer in the cytoplasm. Dicer together with its dsRNA-binding partner TRBP (transactivation-response RNA-binding protein: in mammals) to liberate miRNA-miRNA* duplex. Supported by HSC70-HSP90 chaperone machinery, this duplex is loaded into an Argonaute (AGO) protein as a dsRNA. Subsequent maturation steps expel the miRNA*, producing a mature RNA-induced silencing complex (RISC) (modified from Ameres & Zamore 2013)

1.6. MicroRNA function

MiRNAs regulate gene expression at the post-transcriptional level: the mechanisms proposed by which miRNA regulate expression are translation inhibition, mRNA degradation. In mammals, each miRNA is predicted to have hundreds of targets (Ameres and Zamore 2013). The complete set of miRNAs in a genome regulate the majority of proteins, and miRNAs are therefore involved in almost all cellular processes (Friedman et al. 2009). To prove the function of individual miRNA is difficult due to the fact that loss of function experiments often does not show strong phenotypic effects. There are exceptions to this, for example the loss of function of miR-17~92 and miR-96 cause development problems in humans (de Pontual et al. 2011) (Mencía et al. 2009). Most miRNA function experiments have been done using over-expression or antisense molecules to take down the function of miRNA. Many of these experiments have highlighted the role of miRNAs in processes such as cell proliferation, development and disease (Wang et al. 2008).

1.7. MicroRNA target recognition

In animals, the methods that are used to predict targets of miRNAs use algorithms to find partially complementary between 5' region of miRNA with 3'UTR regions of mRNAs (B. P. Lewis, Burge, and Bartel 2005). There are certain criteria to recognize target of miRNAs: there should be a match between the 5' end of miRNA (called the miRNA "seed" region, and generally defined as nucleotides 2-8 of the miRNA) and the 3'UTR of the mRNA. There are different types of miRNA target sites: canonical sites, marginal sites, atypical sites and 3' compensatory sites (Figure 5). The canonical sites have 7–8-nt matches between the seed region of the miRNA and the target, marginal sites match 6-nt, and atypical sites match seed bases 2-7. For 3´-compensatory sites, Watson–Crick pairing centered on miRNA nucleotides 13–16 can compensate for a seed mismatch and thereby create a functional site.

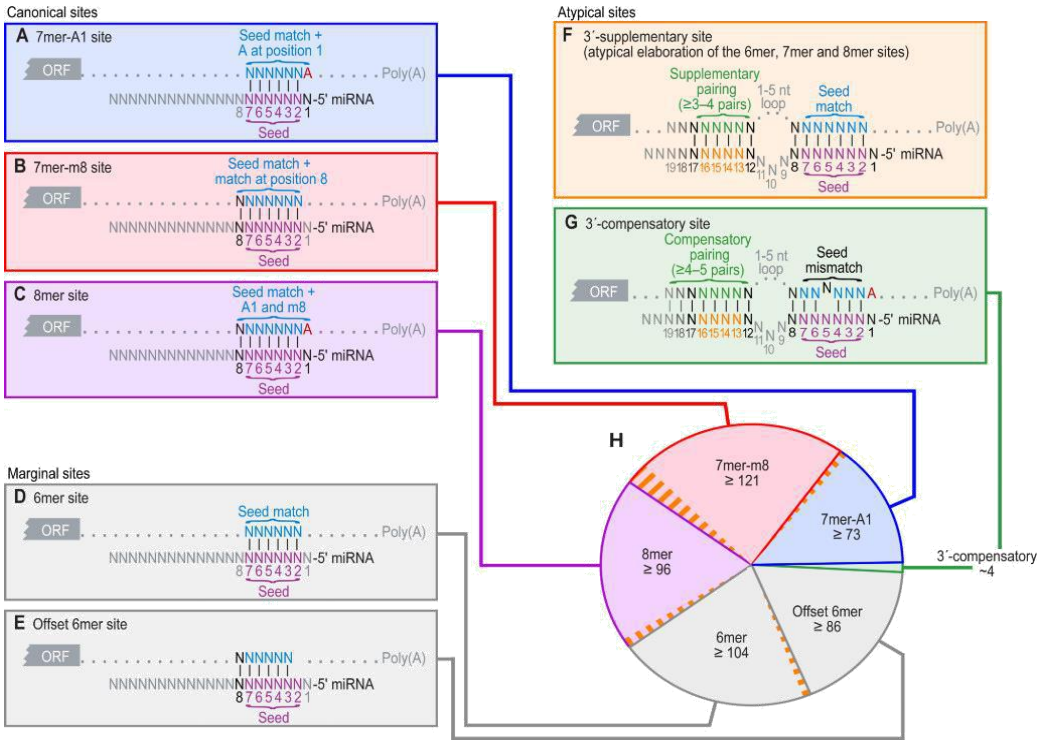

Figure 1.5. Types of miRNA target sites (A–G). The canonical, 7–8-nt seed-matched sites. Vertical dashes indicate contiguous Watson–Crick pairing. (D–E) Marginal, 6-nt sites matching the seed region. (F–G) Sites with productive 3′ pairing. For 3′-supplementary sites (F), Watson–Crick pairing usually centering miRNA nucleotides, 13–16 (orange) supplements a 6–8-nt site (A–E). At least 3–4 well-positioned contiguous pairs are typically required for increased efficacy, which explains why 3′-supplementary sites are atypical. For 3′-compensatory sites (G), Watson–Crick pairing usually centering on miRNA nucleotides 13–16 (orange) can compensate for a seed mismatch and thereby create a functional site. (H) Number of preferentially conserved mammalian sites matching a typical highly conserved miRNA (Friedman et al., 2008) (modified from Bartel 2009).

1.8. Development in *Parasteatoda tepidariorum*

P. tepidariorum is easy to maintain in the laboratory, has a short life cycle, a good number of offspring are produced per cocoon. From an experimental perspective it is amenable to techniques such as RNAi and *in situ* hybridisation. These features have helped to make *P. tepidariorum* an important model organism for developmental biology, evolution and genetics studies.

P. tepidariorum undergoes the short germ band mode of embryogenesis (Hilbrant, Damen, and McGregor 2012) (Figure 6). *P. tepidariorum* has 14 embryonic developmental stages 1: early cleavage (0-10h), stage 2: blastoderm (11-15h), stage 3: germ disc formation (16-27h), stage 4: primary thickening (28-30h), stage 5: cumulus migration (31-35h, up to 40 h), stage 6: dorsal field (41-50h), stage 7: germ band (51-55h), stage 8: prosoma limb band (56-75h), stage 9: limb differentiation (76-85h), stage 10: brain differentiation (86-96h), stage 11:

inversion (96-105h), stage 12: retraction (106-115h), stage 13: dorsal closure (116-140h), stage 14: ventral closure (141-185h). According to Pechmann 2016 (Pechmann 2016), the zygotic genome activation occurs late stage 2 and early stage 3. In general, in metazoa, in the embryogenesis process the mother provides not only the genetic information but also the cytoplasm in which are all the components necessary for the first stages of life (Vasudevan, Seli, and Steitz 2006) (Lee, Bonneau, and Giraldez 2014).

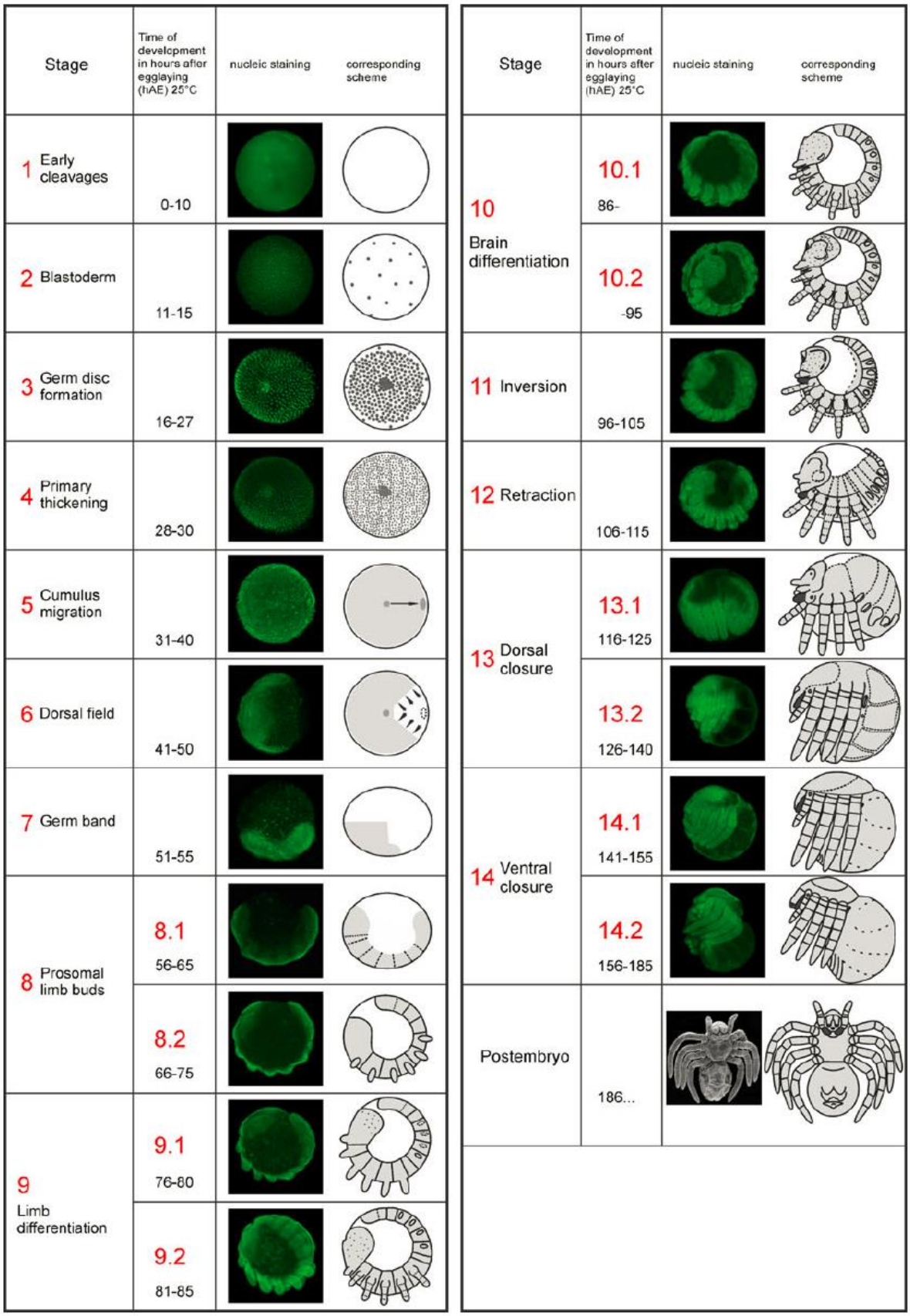

Figure 1.6. Stages of embryo and post embryo of *P. tepidariorum* **[**taken from Mittman 2012**]**

1.9. MicroRNA expression in embryogenesis in *P. tepidariorum*

Embryogenesis is defined by the processes that occur from fertilisation, morphogenesis changes, cell differentiation, organs formation, until the whole organism is ready. In this process, a series of events need to be synchronised and regulated. This regulation involves not only master transcriptional regulators such as transcription factors but also microRNAs.

It was proposed that microRNA have two main roles: a small fraction of microRNAs have a hierarchical function at early stage, and the majority of microRNAs play a role in cell differentiation in a late stage (Alberti and Cochella 2017). At early-stage microRNA are regulating genes in maternal clearance, cell proliferation, apoptosis, cell signalling. It was identified an early and late microRNA expression in *C.elegans* and *D. melanogaster*. They propose that the main function of microRNAs is in the late embryogenesis stages related to cell differentiation (Avital and Franc 2017). In *C. elegans* and *D. melanogaster* we observe bimodal expression.

Studies have shown that microRNAs have a powerful effect of regulation when acting in cluster or through multiple members of a family. For example, in zebrafish, it is known that microRNAs play a role in maternal clearance of mRNA, via over 90 members of the mir-430 family (Giraldez 2005).

1.10. Transposable elements

Transposable elements (TEs) are called also selfish elements, transposon or mobile elements. TEs are DNA sequences in the genome that have the ability to replicate and translocate. There is contraposition of views about their function in the genome. Some scientific works support that TEs confer an advantage in the genome while other results suggest they have deleterious effects (Rebollo, Romanish, and Mager 2012).

TE sequences are classified as retrotransposon and DNA transposon, based on the mechanism of transposition. The retrotransposon class can be divided into two subgroups according to the mechanism of chromosomal integration - long terminal repeat (LTR) retrotransposon and non-long terminal repeat (non-LTR). Non-LTR retrotransposon can be further sub-divided into short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs) (Bourque et al. 2018).

The number of TEs in arthropod genomes is highly variable and there is no correlation between the genome size and the number of TEs identified in each species. For example, for

Acanthoscurria geniculata with a genome size of 7.2 Gb, 2024 TEs have been identified; *Mesobuthus martensii* with a genome size of 925 Mb, there are 1400 known TEs; *Apis cerana* with a genome size of 228 Mb, has 87 TEs; and *Apis mellifera* whose genome size is 250 Mb, has 143 TEs (Wu and Lu 2019).

1.11. PIWI- interacting RNAs (piRNAs)

PiRNAs are single stranded RNA from 24-32 nt long, specific to animals. PiRNAs are processed from a single-stranded precursor, distinct from the hairpin precursor of microRNAs and the double-stranded siRNA precursor. piRNAs are immensely diverse molecules (Ozata et al. 2019), the function of which is silencing transposon expression in germ line at the transcriptional level and post-transcriptional level. In the last years, researchers have found that piRNAs also regulate the expression of genes and defend against viral infection (Ozata et al. 2019). Another feature is that piRNAs are expressed in hundreds of thousands in contrast to other small RNAs (Han and Zamore 2014a).

PiRNAs interact with piwi protein family members, which are a sub-family clade from argonaute proteins: Aubergine (aub), ago3, piwi. Aub and ago3 and are components of the ping pong cycle that silence transposons while piwi is located in the nucleus, and silence at the transcriptional level through histone modification using the Panoramix and Asterix as machinery (Parhad and Theurkauf 2019).

In *Drosophila*, piRNA are transcribed from piRNA clusters that are generally localised in pericentromeric regions, telomeric sequences and euchromatin (Yin and Lin 2007). In mice, piRNAs are transcribed from transposable elements sequences, 3'UTR from mRNA, lncRNA genes (Figure 7). PiRNAs are expressed in clusters, including from TE sequences, 3'UTR from mRNA or lncRNA (Han and Zamore 2014a). The best-studied piRNAs are derived from TE sequences.

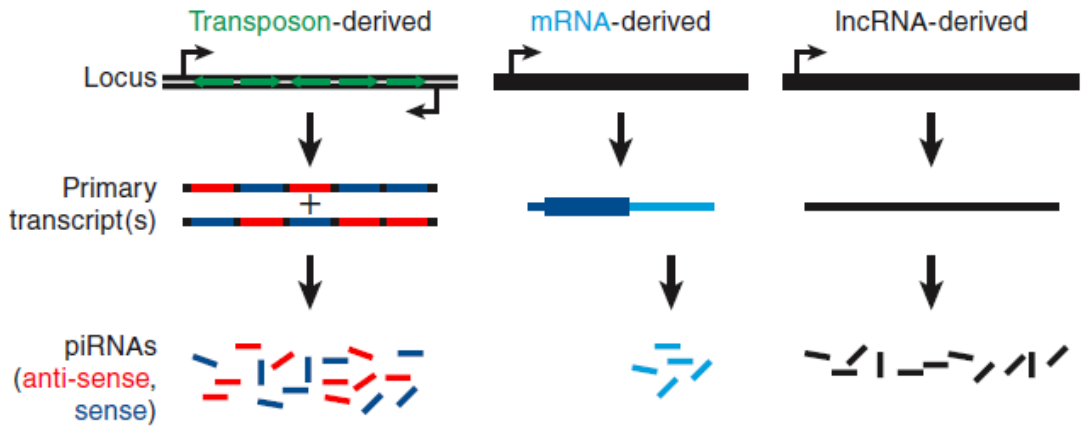

Figure 1.7. Types of piRNA in flies and mice [taken from Han&Zamore 2014]

1.12. Primary piRNA

Most of what we know about piRNA biogenesis and function comes from studies in *Drosophila*. The primary piRNAs are transcribed from a piRNA cluster, which is a region that has invasion of transposon sequences (Figure 8) (Parhad and Theurkauf 2019). In Drosophila piRNAs are produced from 3 different sources: first the germ line nurse cells produce piRNAs in both strands; the follicle produces piRNAs that are transcribed in one strand and their transcripts are very similar to the mRNA. In the fly ovaries, piRNAs are produced to target TEs. There is formation of the complex Yb –piwi-piRNA, to this complex zuc protein bind and cleavage the precursors then the nibbler protein cut to the right size of piRNAs and the modification in 2'-o is made by Hen1 protein. There is successive cleavage when piRNAs bind to the 5'end and this piRNA produced then bind to the piwi protein and are silence TE in the nucleus (Parhad and Theurkauf 2019).

1.13. Secondary piRNAs are linked to the ping pong cycle

Secondary piRNAs are the piRNAs that are generated as products of the so-called ping-pong cycle. The cycle starts when aub binds to the antisense piRNA and together as complex align with the sense transposon element transcript. The aub protein then cleaves the transposon transcript. This fragment binds to ago3 and this complex align to the antisense piRNA cluster. Ago3 cleaves the targeted TE sequences generating again a fragment that bind to aub protein to enter a new cycle (Figure 9). In this way, the product of ago3-piRNA complex can be used as a substrate for the production of primary piRNA. The cleavage site of the ago3 and aub are in the nucleotide 10^{th} counting from the 5'end. This cycle can be originated by maternally deposited piRNAs. DEAD box helicase, tudor domain protein Qin and vasa protein are involved in this cycle (Parhad and Theurkauf 2019). This ping-pong mechanism allows production of piRNAs at high levels to silence TE fast and efficiently.

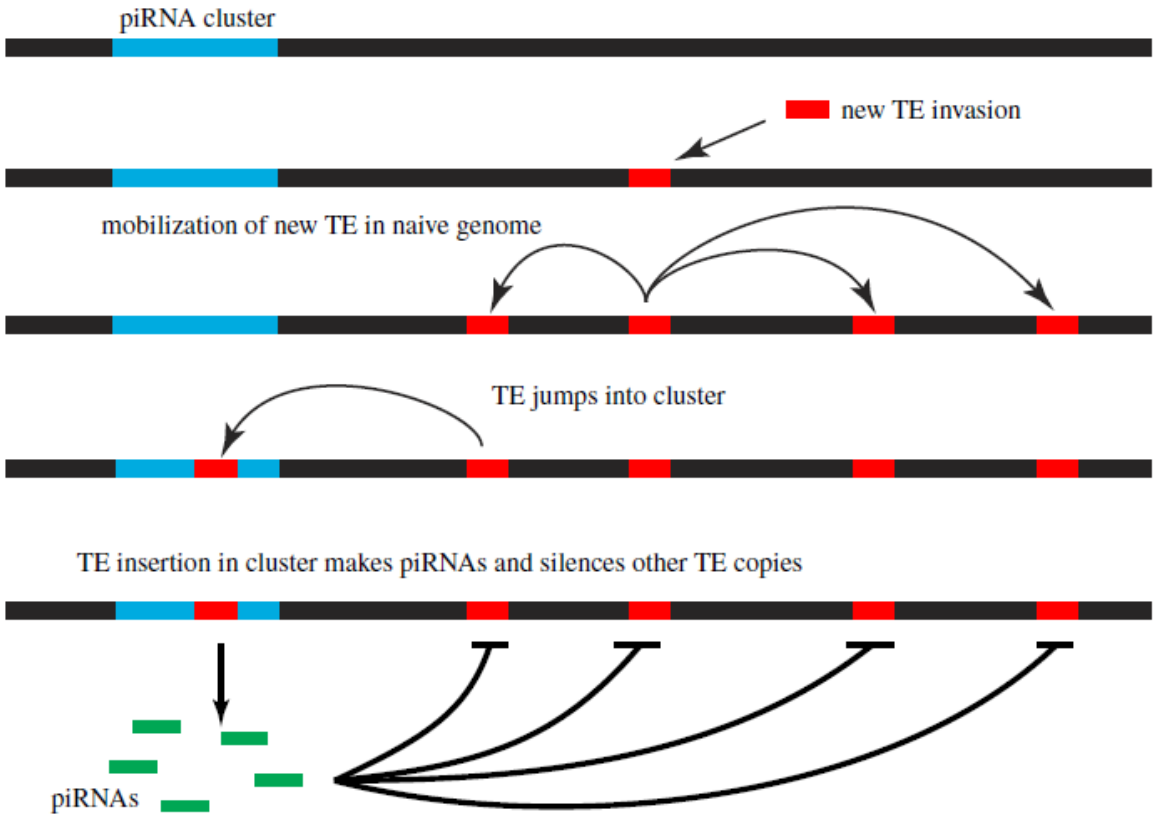

Figure 1.8. Transposable element (TE) jumps in to piRNA cluster. piRNAs are generated against the new TE [taken from Parhad and Theurkauf 2019]

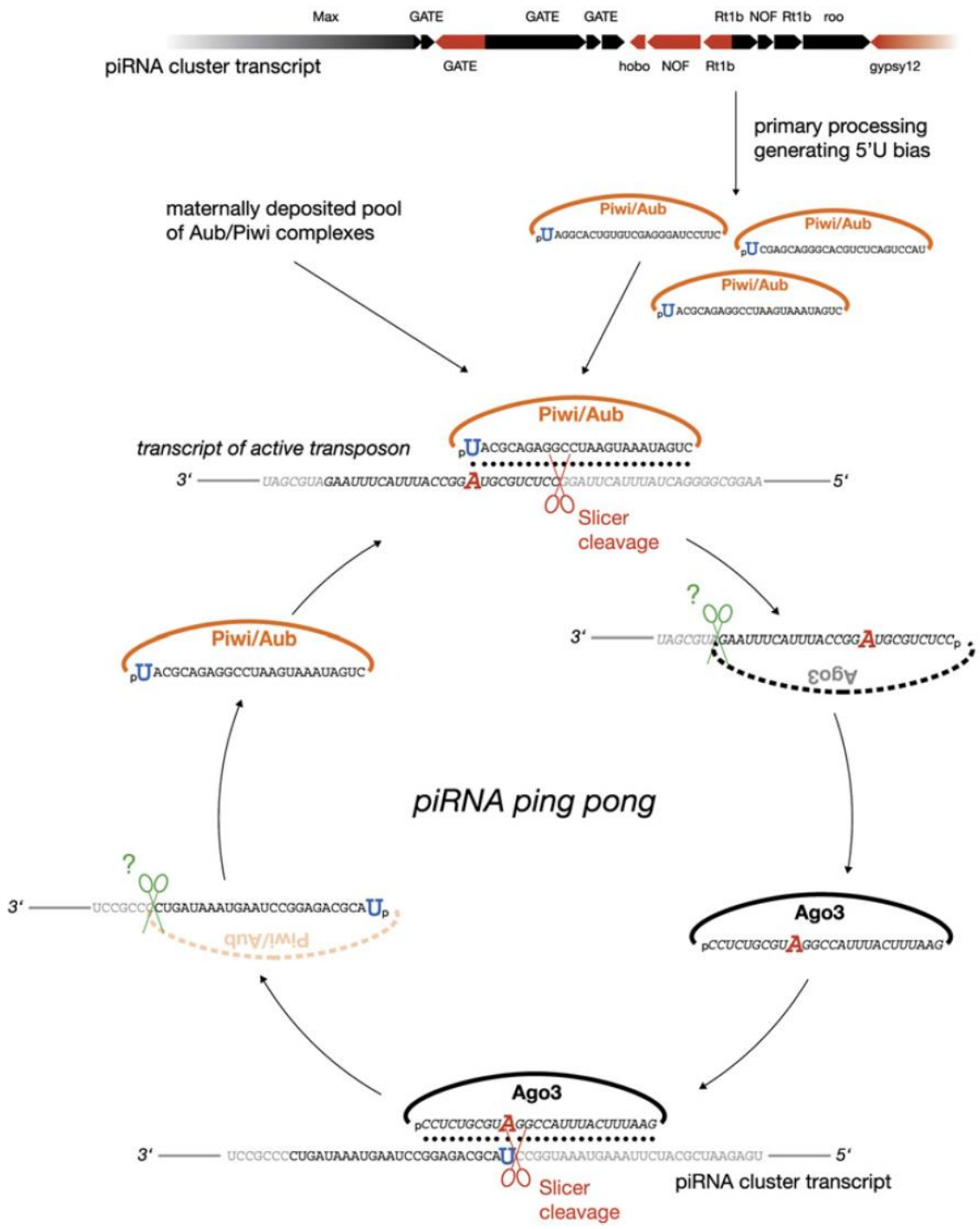

Figure 1.9. Ping-pong cycle [taken from Brennecke et al., 2007].

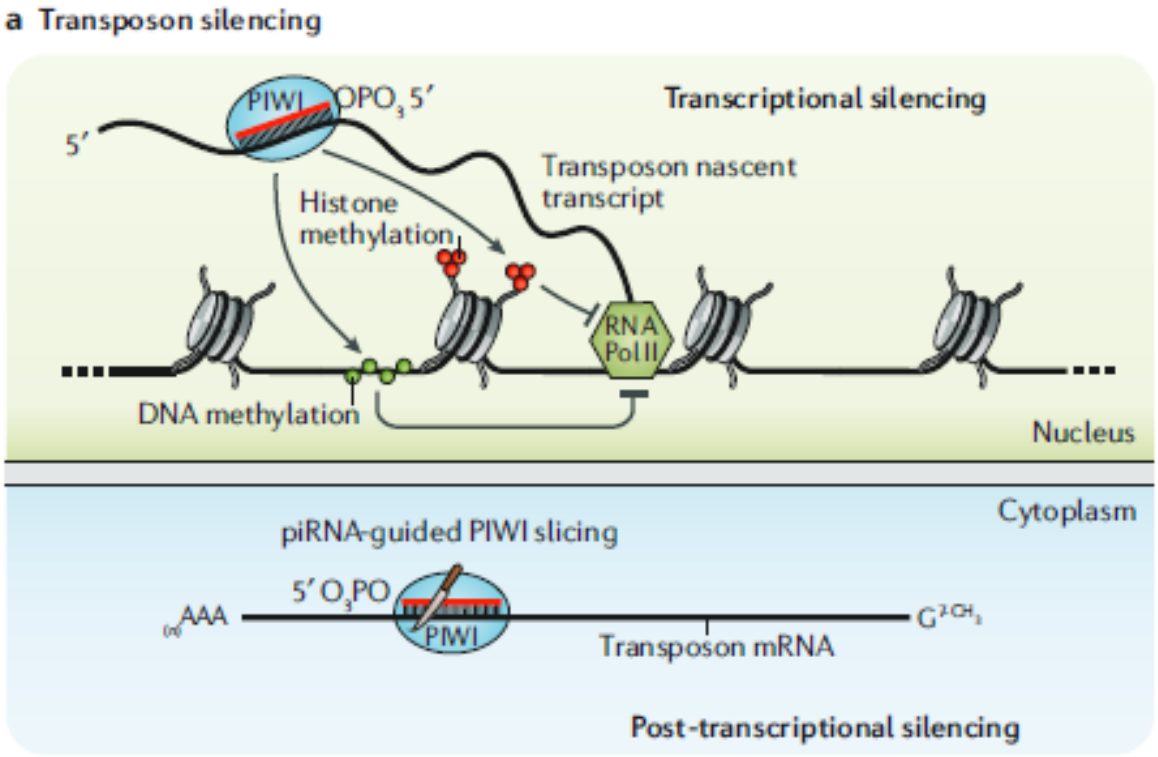

Figure 1.10. Transposon silencing at transcriptional and post-transcriptional by piRNAs[taken from Ozata et al., 2019]

Do miRNAs preferentially regulate ohnologs genes in vertebrates?

2. Do miRNAs preferentially regulate ohnologs genes in vertebrates?

2.1. Abstract

MiRNAs are small non-coding RNAs approximately 20-22 nucleotides long, the function of which is to post-transcriptionally regulate the expression of most genes in animals and plants. Gene duplication is an important evolutionary process that can drive diversity and novelty. However, gene duplication can generate a gene dosage imbalance in the cell. Here, we investigate the possible roles of miRNAs in buffering gene dosage post duplication. We propose that small-scale duplications (SSDs) can generate a stoichiometric imbalance in gene products. However, miRNAs may play a role in buffering expression of such genes limiting the imbalance. Alternatively, genes duplicated during whole genome duplication (WGD) processes may not create such an imbalance, and therefore will show evidence of being less targeted by miRNAs. To address this scenario, we determine the properties of targeted WGD and SSD genes. We predicted miRNA target sites in 3'UTRs of SSD, WGD and single copy genes in human, mouse, rat, pig, dog and chicken genomes. Contrary to our hypothesis, we found that in human, mouse, rat, miRNAs preferentially regulate WGD genes. In addition, we found that miRNAs preferentially regulated haplo-insufficient genes.

2.2. Introduction

MicroRNAs are short non-coding RNAs ranging between 21-25 nt in length. They have the function to regulate the gene expression through repression of translation, for example, via deadenylation of mRNA in the cytoplasm. MicroRNAs are present throughout the animal kingdom.

Duplication of genes provides a raw source of material to the genome. The proposition of the whole genome duplication in vertebrates is controversial and hotly debated. It has been proposed that the complexity of vertebrates is due to the whole genome duplications event. In line with this, some propose two rounds of whole genome duplication (WGD), others support one whole genome duplication and successive small scale duplication, and others still propose that there occurred only segmental duplications. In relation to time of duplication, WGD genes are old genes, SSD genes can be old, intermediate or recently duplicated genes (Singh, Arora, and Isambert 2015).

In human, ohnologos that do not pass SSD events, and also suffer from copy number variation (CNV), are more likely to linked to human disease (Makino and Mclysaght 2010).

In humans, there is a classification of genes according to the threshold of molecules produced for a gene expression. Haplosufficient genes are so called because a single copy is enough to obtain the normal phenotype. Logically then, haploinsufficient genes are those needed in the diploid state to have a normal phenotype. In humans, these genes are linked to dominant disease.

The essential genes are vital for survival, with their absence leading to lethality. In human, it was found that there is an overlap between essential genes and ohnologs. Contrary to what happens in *C. elegans* genes that arise from a whole genome duplication in mouse and humans are not redundant, but rather, essential genes (Makino, Hokamp, and Mclysaght 2008).

Changing the dosage of a gene by duplication can cause an imbalance in the stoichiometry of proteins in the cell. For example, changing the relative amounts of proteins in a complex might have adverse effects on the function of that complex (Birchler and Veitia, 2010). It is proposed that whole genome duplication (WGD) events do not perturb this equilibrium. However small-scale duplications (SSD) that contain genes encoding members of protein complexes will be more likely to have a dosage effect. We predict that miRNAs may preferentially target and regulate small-scale duplicate genes rather than whole genome duplication genes. Previous study investigated how microRNAs regulate preferentially the duplicated genes as one group (WGD and SSD) rather than single copy genes in human and

mouse (Li, Musso, and Zhang 2008). However, there is no study on microRNA preference for a type of duplicated gene. In this chapter, we explore whether microRNAs regulate preferentially WGD genes rather than SSD genes in human, mouse, rat, pig, dog and chicken.

2.3. Methods

2.3.1. Properties of targeted WGD and SSD gene pairs and gene families

2.3.1.1. Identification of gene pairs and gene families in WGD and SSDr

The pair and gene family list for WGD genes in human were obtained from the ohnolog database (http://ohnologs.curie.fr/). The ohnolog database provided 3 classifications of WGD genes based on strict (WGDs), intermediate (WGDi) and relaxed (WGDr) criteria (Singh, Arora, and Isambert 2015). We subtracted the WGD genes from the paralogous gene list to define the list of small-scale duplicated genes (SSD). Then we get SSDs, SSDi and SSDr, respectively, we considered the SSDr as the most suitable category for our analysis because only in this case we get a more stringent gene list for SSD. With this SSDr gene list, we selected its corresponding pair in the paralog gene list from Ensembl database v88 (Aken et al. 2017). For SSDr gene families, we used the list of SSDr pairs of genes that we generated previously, and then expanded to generate families of genes, which could be connected by small-scale duplication events. These families of genes formed a disjoint set structure, since a gene being a member of two families must imply that those families are connected into a single set. We used a standard disjoint-set data structure to reflect a graph with genes as nodes and pairwise duplication events as edges. We followed the union-find algorithm to join nodes to form maximal families of connected genes. Briefly, the data structure used a dictionary to match a representative node to its set of connected genes and held a pointer for each node to the representative node for its family. The union-find algorithm then considered each pairwise edge in turn, and either merged two families if the two connected nodes belonged to different groups, or added a node to a family if it had not been encountered before, we use a python script to get the SSDr gene family list

2.3.1.2. Conservation ratio (CR)

We downloaded the high confidence human mature miRNAs sequences from miRBase version 21 (http://www.mirbase.org/). We use the SeedVicious program to predict miRNAs

targets sites in the longest 3'UTR that recognises a seed sequence (Bartel 2009). We created a matrix between miRNAs and targets genes of those miRNAs, this information later was used to calculate the conservation ratio. We considered when miRNAs target that gene as 1 and when not as 0 scores. We do not consider how many target sites has each miRNAs in one 3'UTR

The conservation ratio was calculated as the sum of the total miRNAs that target n genes, excluding the miRNAs that only target one gene; this result was then divided by multiplication of n number of genes and number of unique miRNAs shared by n genes. We obtained the conservation ratio value using a script in R. This script was applied for the analysis of gene pairs and gene families in WGD and SSDr.

CR= sum miRNAs target least 2 genes–sumMiRNAs target only one gene/number of genes*sum unique miRNAs

2.3.1.3. Percent identity (PI)

In order to get the percent identity for pairwise alignment and multiple sequence alignment of 3UTRs, we used the Clustalw alignment program (Thompson, Gibson, and Higgins 2003) with default parameters for pairwise (gapextension=0.1, ktup=1, windowlength=5, topdiag=5, pairgap=3) and for multiple sequence alignment (gapextension=0.20, gapdistance=5, noendgaps=no, numiter=1, clustering=nj. For both we used gapopen=10 and DNA weightmatrix=clustalw).

The percent identity was calculated as the number of identical positions divided by the length of complete alignment. For more than 2 sequences, we used the multiple sequences alignment generated by Clustalw and then extracted all the possible pair sequences alignments, corrected the complete alignment for each pair and then calculated the PI for each of them, finally we averaged those values. We obtained the percent identity pairwise and multiple sequence alignment using an R script.

2.3.1.4. Spearman rank correlation

In order to see the degree of correlation between the percent identity and the conservation ratio values in WGD and SSD genes, we performed the Spearman rank correlation for pair and gene family in WGD and SSDr genes.

2.3.1.5. Gene description and Gene enrichment analysis

We performed a gene ontology (GO) enrichment analysis for SSDr, we selected the peak CR=1 of the density pairwise conservation ratio plot. We use DAVID method for the analysis (D. W. Huang, Sherman, and Lempicki 2009). The aim of this analysis was to investigate if the gene pairs in SSD peak, at CR=1, were recently duplicated genes. Enrichment analysis can provide information about their function, localisation, or biological processes. To get the annotation information for genes description, we retrieve information from ensembl v88 for human.

2.3.2. miRNAs regulation of WGD, SSD and single copy in human, mouse, rat, pig, dog, and chicken.

2.3.2.1. Identification of WGD, SSD and single copy genes.

Human, mouse, rat, pig, dog, and chicken genes were categorised as WGD, small-scale duplicates (SSD), and single copy genes following the steps shown in (Figure 6). Briefly, we retrieved all gene IDs from Ensembl version 88 using Biomart. We then used Ensembl (v88) and Biomart to retrieve lists of all paralogous gene IDs. We subtracted duplicated genes from all gene IDs to define the list of single copy genes. Genes classified as WGDs were downloaded from the ohnologs database (Singh, Arora, and Isambert 2015) (http://ohnologs.curie.fr/). We subtracted the WGD genes from the paralogous gene list to define the list of small-scale duplicated genes. The WGD database provided 3 classifications of WGD genes based on strict, intermediate and relaxed criteria. The same analysis was carried out independently on all 3 datasets.

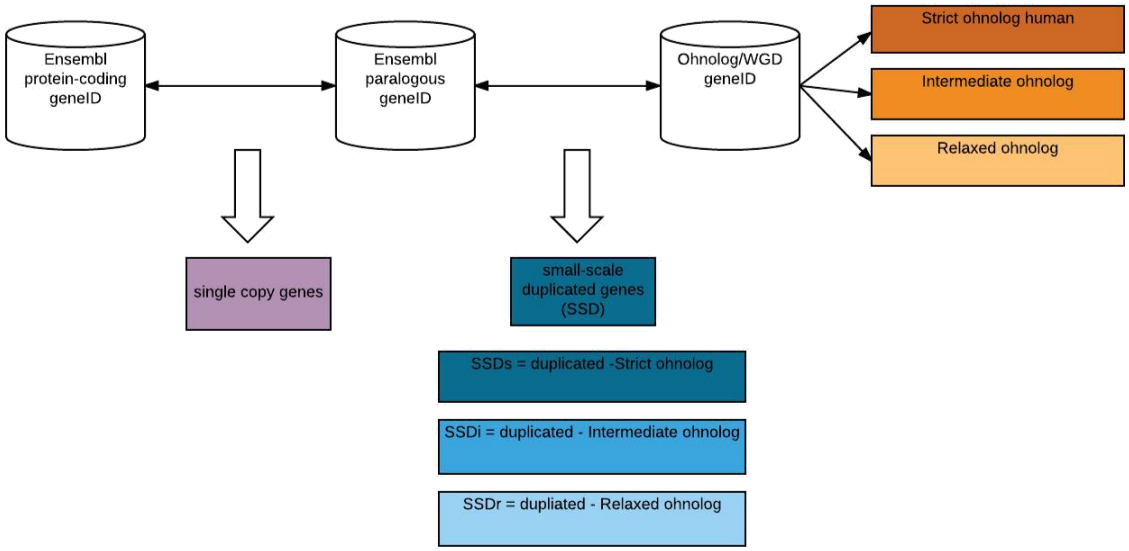

Figure 2.1. The steps involved in identifying WGDs, small-scale duplicate and single copy genes in human, mouse, rat, pig, dog, and chicken.

2.3.2.2. MiRNA target prediction

We obtained 3'UTR sequences of all human, mouse, rat, pig, dog and chicken transcripts from Ensembl version 88 using Biomart, then we chose the longest transcript per gene. For human, mouse and chicken we use the mature high confidence miRNAs and for dog, pig and rat the mature miRNAs, miRNAs sequences were downloaded from the miRBase version 21 (http://www.mirbase.org/). We used the SeedVicious perl script to predict miRNA target sites (Marco 2018) based on the seed concept as described previously (Bartel 2009). Since longer UTRs are expected to have more miRNA binding sites, we normalised the number of predicted binding sites by the longest length of the UTR, creating a numerical variable representing the miRNA binding site density per gene. The densities per gene values were compared between single copy, SSD and WGD genes sets obtained using the methodology in section 2.3.2.2. To test that very long or very short 3'UTRs were not biasing the results, we removed 3'UTR sequences with less than 50 bp and longer than 10 kb.

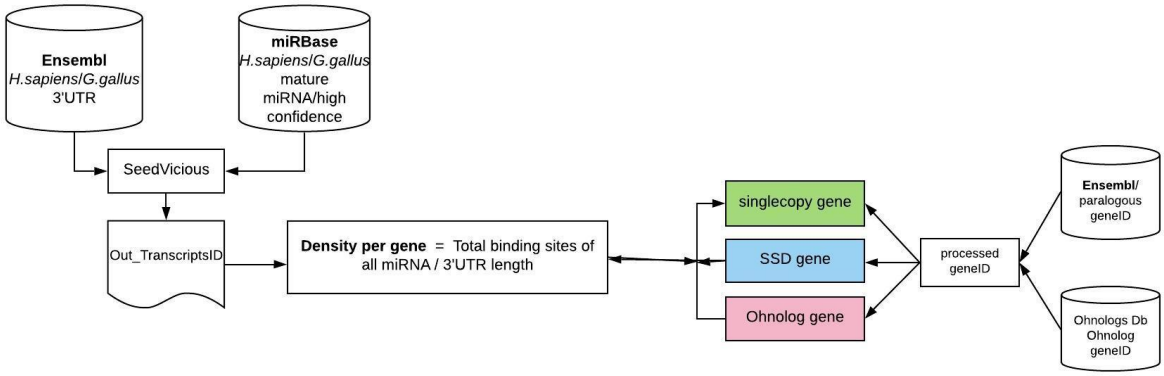

Figure 2.2. Flowchart describing the steps to compare miRNAs binding site density per gene values among small-scale duplicate, WGD and single copy genes.

2.3.2.3. Statistical analysis

T-test was used to compare the miRNA binding site density per gene between each pair of WGD, SSD and single copy genes in human, mouse, rat, pig, dog and chicken independently. We performed independent analyses using the gene sets derived from the strict WGD and strict SSD classifications. We also compared the strict WGD set with the SSD derived from the relaxed WGD classification, the latter representing the most conservative definition of SSD gene duplicates.

2.3.3. Fast and slow evolving WGD genes

We used gene pairs of WGD genes, from the strict criteria classification. The classifications in slow and fast evolved genes were made by Mark Reardon (personal communication) based on the distance to root, the root of each gene is not a gene but a duplication or speciation node that is available in the ensemble compare. We compared the slow and fast WGD genes using the binding site density per gene. We use t-test to test whether one group was preferentially targeted rather than the other by miRNAs.

2.3.4. Do miRNAs regulate haploinsufficient rather than haplosufficient genes in human?

The list of genes was downloaded from DECIPHER database (https://decipher.sanger.ac.uk/). A subset of the table was done in which we have the Ensembl geneID, gene name and the HI values. The values are from 0 to 100%. We consider as haploinsufficient genes above 80% as there were enrichment for HI in the highest 20%

mentioned in the article (N. Huang et al. 2010) and to determine HS genes, we consider the lowest 20%, we considered that this will be more accurate according with the enrichment criteria. With these lists we compare the binding site density per gene for human obtained in the section (2.3.2.2). We performed a t-test to see whether there were preferential regulations of haploinsufficient or haplosufficient genes by miRNAs.

2.3.5. Physical and non-physical gene interaction in human

We obtained the gene set list from BIOGRID database. This is a database that contains information about genes from genetic and physical interaction from different species. We chose the human subset. We decided to consider the co-fractionation and co-purification method as physical list of genes, we subtracted the physical from the total protein coding gene list obtained from ensembl version 88 to define the non-physical list of genes. Then later, we compare the binding site density per gene obtained for human (section 2.3.2.2) for these two lists. We performed a t-test to observe a differential regulation between these two groups. We used R package for the visualization of the density values in physical and non-physical gene list.

2.3.6. Essential genes

We downloaded Ensembl gene IDs for essential and non-essential genes from the Online GEne Essentiality database (Chen et al. 2012) (http://ogeedb.embl.de/#overview:). We used data from the large genomic wide analysis, where genes were classified as essential, when reducing the expression of these genes using RNAi, caused the inhibition in growth in five cell lines. We assessed the differential regulation of essential and non-essential genes considering miRNAs binding site density per gene value

2.4. Results

2.4.1. Properties WGD and SSDr

2.4.1.1. Identification of pair and gene family in WGD and SSD

The identification of the WGD gene pair and WGD gene family was developed in a previous study (Singh, Arora, and Isambert 2015). We considered the lists from previous work and also develop a match evaluation against the list of available 3'UTRs in human. We obtained

2,519 WGD gene pairs and 1,279 WGD gene families. In addition, we obtained 39,029 SSDr gene pairs and 2,202 SSDr gene families for the analysis.

2.4.1.2. Conservation ratio

We obtained 544 high confidence mature human miRNA sequences from miRBase version 21 (http://www.mirbase.org/). We predicted miRNA target sites in 20,920 longest 3'UTRs transcripts that recognised a seed sequence (Bartel 2009). We calculated the conservation ratio as the sum total miRNAs target all genes excluding the genes that are only targeted once; this result was divided by multiplication of number of genes and number of unique miRNAs. Our aim was to determine how WGD and SSD genes have evolved in terms of shared target sites by miRNAs in gene pairs and gene families. Then later these results allow us to compare between WGD and SSD gene pairs and gene families (Figure 8).

2.4.1.3. SSDr genes present wide and higher conservation ratio than WGD genes

WGD and SSD are duplicated genes that differ not only how they have arisen, enriched for certain classes of genes, process to be preserved (Davis and Petrov 2005). Our initial hypothesis was to determine whether miRNAs regulate SSD preferentially rather than WGD to conserve the stoichiometry balance in the cell. We decided first to investigate whether the two copies of WGD or SSD are targeted differentially by miRNAs, then later compare between groups. The number of pairs analysed of WGD and SSDr are 2, 519 and 35, 989 respectively and the number of WGD and SSDr gene families are 1, 279 and 2, 202.

We can observe that conservation ratio (CR) values of WGD and SSD present a bimodal distribution. We can observe that CR values of WGD pairs are from 0.0 to 0.5 approximately while in SSDr gene pair, we also can observe that the majority of values are from 0.0 to 0.5 but in addition, there are values that go from 0.5 to 0.9 and ending in a small peak arisen that reach at CR=1 (Figure 8).

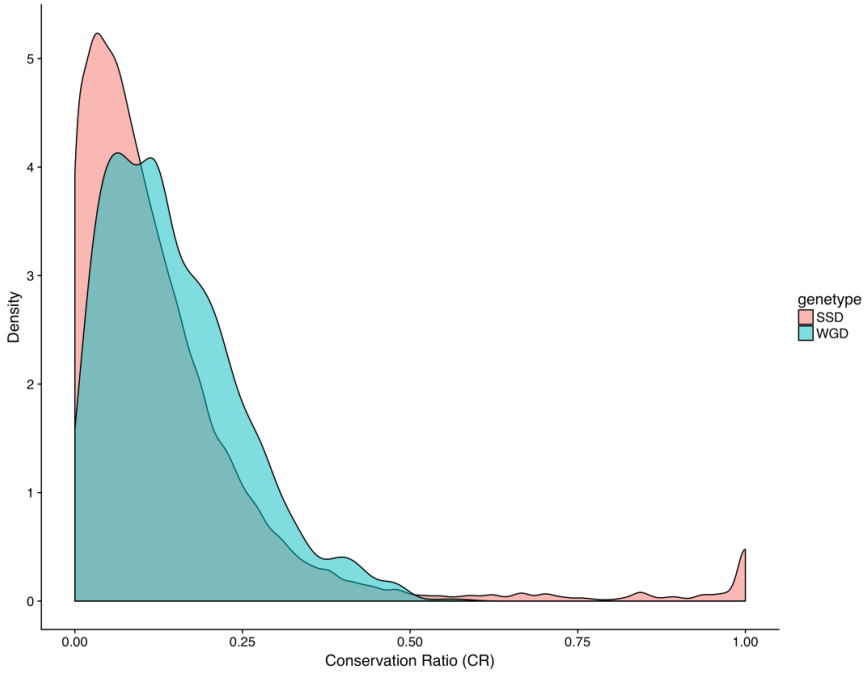

Figure 2.3. Density distribution of Pairwise Conservation Ratio of the conservation of miRNA target sites in the 3'UTRs calculated for WGD and SSD gene pairs in human.

We also observe that conservation ratio (CR) values of WGD gene families are from 0.0 to 0.5 approximately and in SSDr gene families, we observe a bimodal distribution. We further observe that the majority of values are from 0.0 to 0.5, but in addition there are values that go from 0.5 to 0.9 and ending in a small peak arisen that reach at CR=1 (Figure 9). Gene pairs and family represent the same pattern of distribution values.

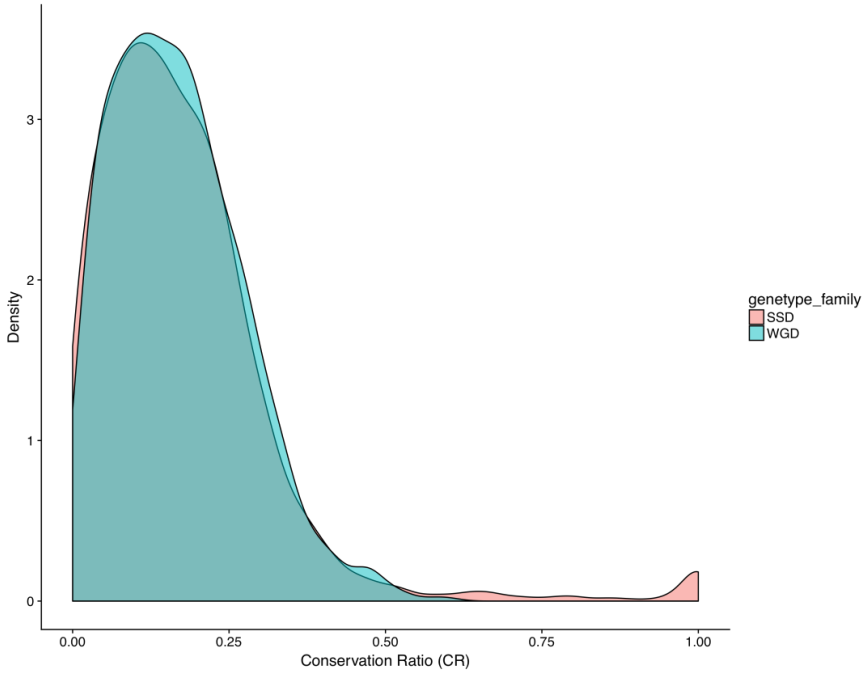

Figure 2.4. Density distribution of Family Conservation Ratio of the conservation of miRNAs target sites in the 3'UTRs calculated for WGD and SSDr gene family in human

2.4.1.4. SSDr gene pair presented a wide and higher percent identity than WGD gene pair.

We decided to calculate the percent identity for each gene pair and gene family in WGD and SSD, as a method to validate the conservation ratio measurement. Since the target recognition by miRNAs is dependent of the sequence identity, we expect a correlation between those parameters. We obtained the percent identity using Clustalw using the default parameters for gene pairs and gene family. When we compare the WGD and SSDr gene pairs and gene family (Figure 10, Figure 11), we can observe similarities in the distribution, WGD values go from 0 to 50% approximately of percent identity and as well SSDr from 0 to 50% approximately, however the SSDr also presented bimodal distribution with values from 50 to 100% percent identity, having a small peak of 100% percent identity.

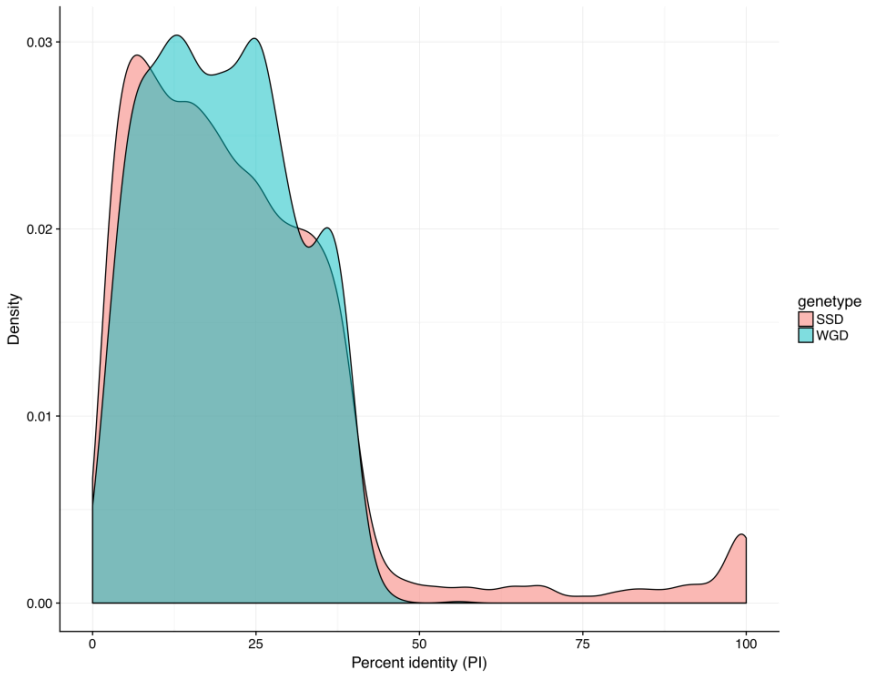

Figure 2.5. Density distribution of Percent identity in the 3'UTRs calculated for WGD and SSDr gene pair in human

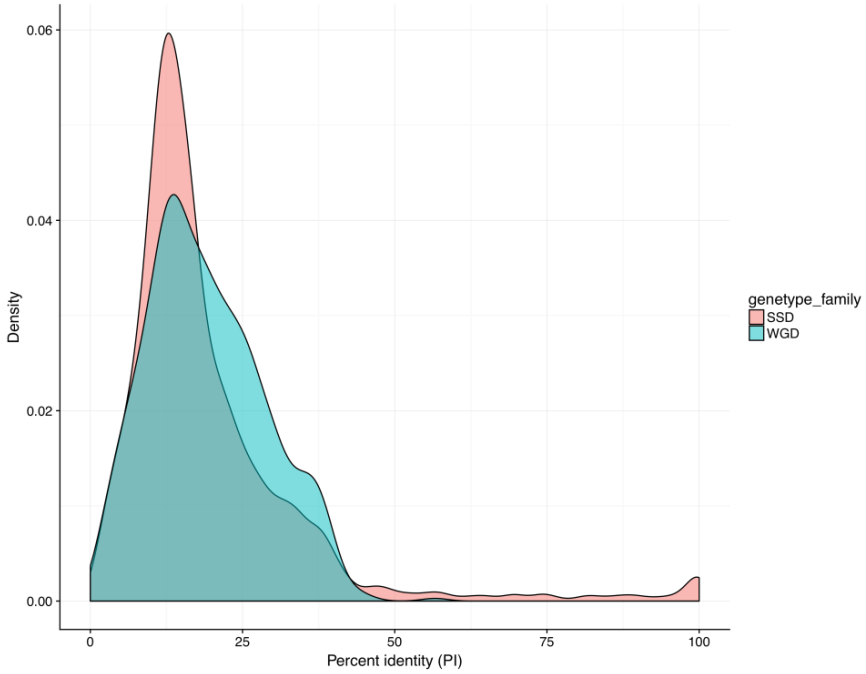

Figure 2.6. Density distribution of Percent identity in the 3'UTRs calculated for WGD and SSDr gene family in human

2.4.1.5. SSDr gene pairs show different degree of miRNA regulation than WGD gene pairs

We wanted to observe how miRNAs targeted gene pair and gene families of the WGD and SSD. In order to get that, we plotted the CR and PI for WGD and SSD for gene pairs (Figure 12, 13, 16) and gene families (Figure 14, 15 and 17). In order to understand the comparison, it is necessary to explain that CR=0 means that those pairs of sequences do not share any miRNAs, and when CR=1 it means that those pairs of sequences share all miRNAs between them.

For WGD gene pairs (Figure 12), the CR values go from 0 to 0.6 and the PI from 0 to 40%. The majority of values are located within CR=0 to 0.4 and within PI=0 to 40. When the value of CR=0 and PI is from 0 to 4%, we found a common pattern, the low percent identity and CR values is because those alignments are between small and large sequences, for example sequences of 160bp and 1,100bp.

For SSDr gene pairs (Figure 13), the CR values go from 0 to 1.0 and PI from 0 to 100%. The majority of gene pairs (80%) are concentrated within CR= 0 to 0.4 and within PI=0 to 40. We also can observe gene pairs distributed CR=0.4 to 1.0 and PI =40 to 100. Only in this type of duplication, SSDr, we can observe that at least 50 pair of genes has CR 1.0 and PI=100%. SSDr gene pair also present 2612 pairs with CR=0 value. When we look for CR=0 and PI=0%,

we can observe that those genes has again the same pattern found in WGD gene pair, the length difference in sequence are very big, for example some of them has the following length 57bp with 6,872bp; 87bp with 4,249bp; 71bp with 2,479bp; 53bp with 5,120bp; 64bp with 2,006bp. For all, the case the small sequence matches 100% within the large sequences.

For SSDr with CR=0 and PI=25%, we can observe that those genes have relatively small length and for example in some cases: 157bp with 228 bp; 73bp with 126bp; 24bp with 386bp; 163bp with 86bp; 81bp with 168bp. For SSDr with CR=0, PI=75, we found 2 genes with small sequences forming small alignments, for example, we found an alignment of 100bp, length sequences 81bp and 163bp; 91bp and 99bp,

For SSDr with CR=1 and PI=100%, we found 459 gene pairs with CR=1 and PI=69-100%, the majority values are 100%. For example, we have for 69% we have an alignment of 128bp and 132b with 93 bp length, for CR=1 and 98% we have an alignment of 170 bp and 166 bp with 167 as sequence length. For CR=1 and PI=100%, we found alignment with the following length 238 bp, 1,882bp, 99 bp, 689bp, 471bp, 240bp, 475bp, 409bp, 1,882bp, 910bp, 240bp and in many cases of this alignments we observe that one gene was common in each pair of genes, so hence the alignments were of the same length. We got a list 918 genes and using this list we retrieved 234 gene descriptions from Ensembl database and gene enrichment analysis

2.4.1.6. SSDr gene families show different degree of miRNA regulation than WGD gene families

The number of gene families analysed of WGD and SSDr are 1,271 and 2,202, respectively. When we compare the plots (Figure 16 and 17), SSDr show a different degree of regulation by miRNAs. For SSDr gene families, with CR=0, PI=0 to 5%, we can observe again the same pattern as we found in WGD gene pairs and SSDr gene pairs: a small sequence aligned with a large sequence. When parameters like CR=0 and PI=close to 40% were used, we found that sequences and their alignments were relative small, ranging from 120bp to 360pb length.

When parameters like CR=0.6 and PI=27% were used, we found 4 genes whose sequences and their alignments were relatively large, such us 5353bp, 5680bp, 6055bp, 9331bp length. When parameters like CR=0.5 and PI=26% were used, we found that sequences and their alignments were relatively large, such as 9110bp, 9609bp, 5219bp length. When parameters like CR=0.5 and PI=35% were used, we also found that sequences and their alignments were relatively large, such 6040bp, 5755bp, 4826bp length. When parameters like

CR=0.4 and PI=30% were used, we found that sequences and their alignments, such as 3773bp, 5520bp, 5839bp length.

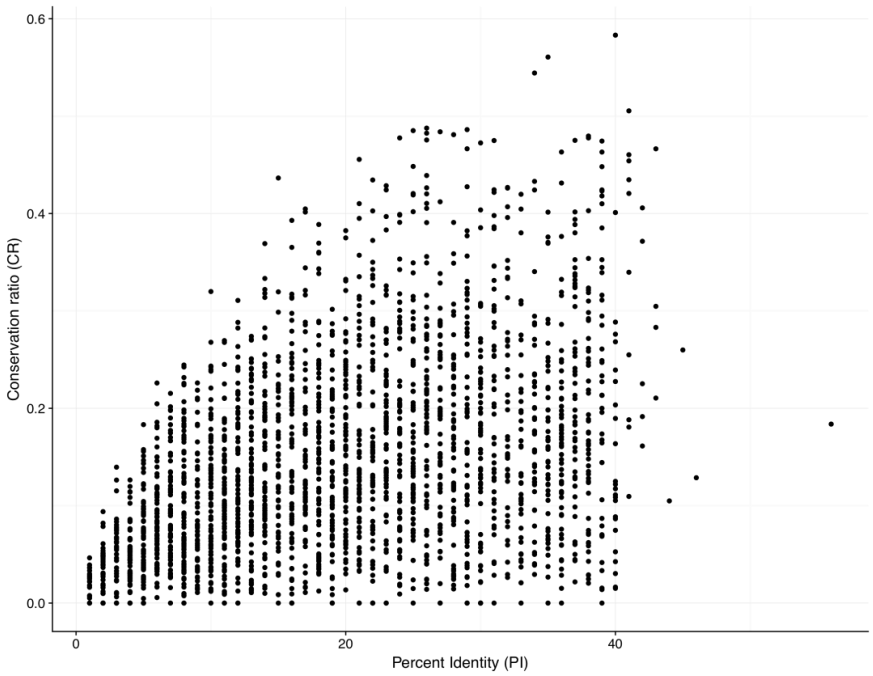

Figure 2.7. Correlation between percent identity (PI) and conservation ratio (CR) in WGD 3'UTRs in human.

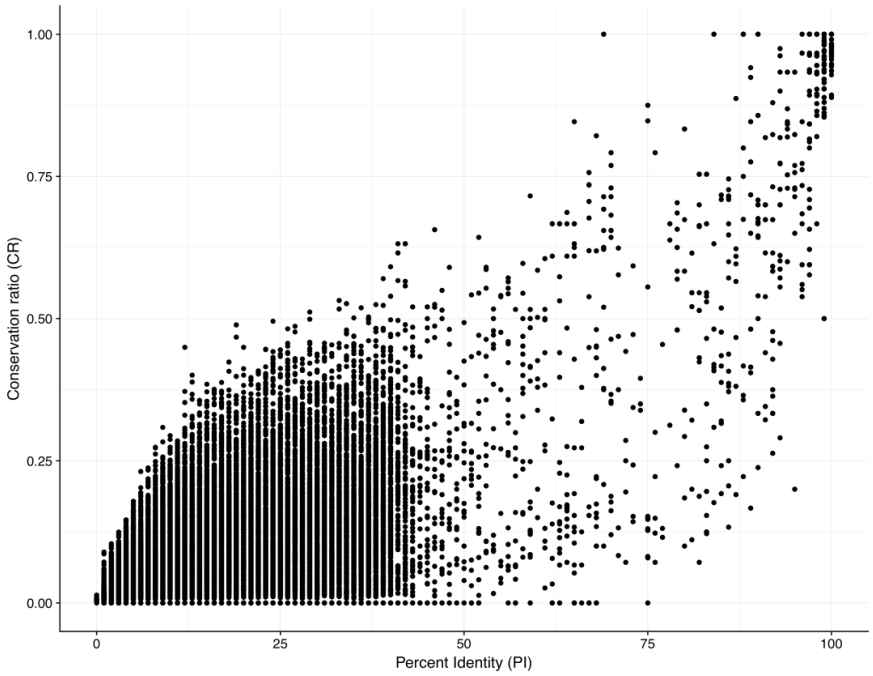

Figure 2.8. Correlation between percent identity (PI) and conservation ratio (CR) in SSDr 3'UTRs in human.

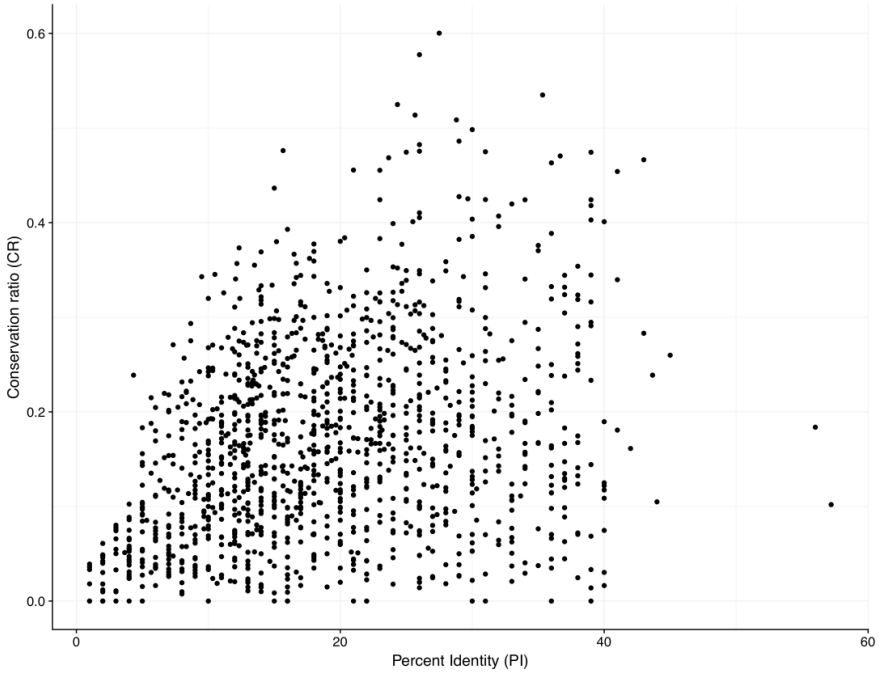

Figure 2.9. Correlation between percent identity (PI) and conservation ratio (CR) in WGD family 3'UTRs in human.

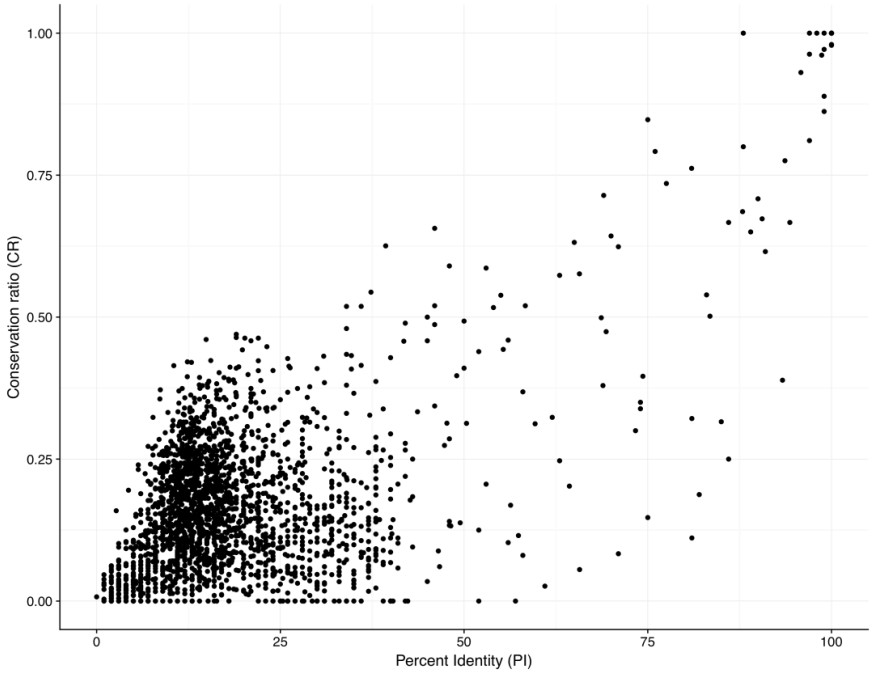

Figure 2.10. Correlation between percent identity (PI) and conservation ratio (CR) in SSDr family 3'UTRs in human.

We can observe that the highest CR value of WGD gene pairs is close to 0.6 while the highest CR value of SSDr instead reaches 1.0. The plots also show that they have a great number of pair of genes located from CR=0.0 to CR=0.4 between WGD and SSDr (Figure 16).

Figure 2.11. Correlation between percent identity (PI) and conservation ratio (CR) in WGD and SSD 3'UTRs in human.

Figure 2.12. Correlation between percent identity (PI) and conservation ratio (CR) in WGD and SSDr family 3'UTRs in human.

Table 1.1. Spearman rank correlation between SSD and WGD families and pairs

CR&PI	Rho value	Spearman rank correlation (p-value)
WGD	0.4142041	2.2e-16
SSDr	0.4704847	2.2e-16
WGD family	0.3109981	2.2e-16
SSDr family	0.242922	2.2e-16

2.4.1.7. Gene description and gene enrichment analysis

We found a small peak in the density plot of the SSD genes; this peak was arisen at CR=1 value. In order to know what kind of gene pair was in that peak, we decide to get information about its gene description and enrichment analysis using the DAVID method. The most abundant genes was described as cancer/testis antigen family 45 member (18), protocadherin gamma subfamily A (16), G antigen (10), defensin beta (8), proline rich (7), olfactory receptor family (6), RNA binding motif protein, Y-linked, family (6), TBC1 domain family member (6), family with sequence similarity (6), histone (6), testis specific protein, Y-linked (5), POTE ankyrin domain family (5), nuclear pore complex interacting protein family (5), TP53 target (4), SPATA31 subfamily A (4), PRAME family member (4), variable charge, Y-linked (2), chromosome X open reading frame (4), SSX family member (4), tripartite motif containing 49 (3), basic charge, Y-linked, 2 (3), RANBP2-like and GRIP domain (3), tripartite motif (3), sperm acrosome associated (2), defensin alpha (2), LIM zinc finger domain (2), heat shock transcription factor family (2), potassium voltage-gated channel (2), protease, serine(2), U2 small nuclear RNA auxiliary factor 1 (2). The enrichment analysis using DAVID method gave us that those genes were enriched with genes related to the spermatogenesis, gonadal mesoderm development, nucleosome assembly GO term (D. W. Huang, Sherman, and Lempicki 2009)(Figure 16, 18)

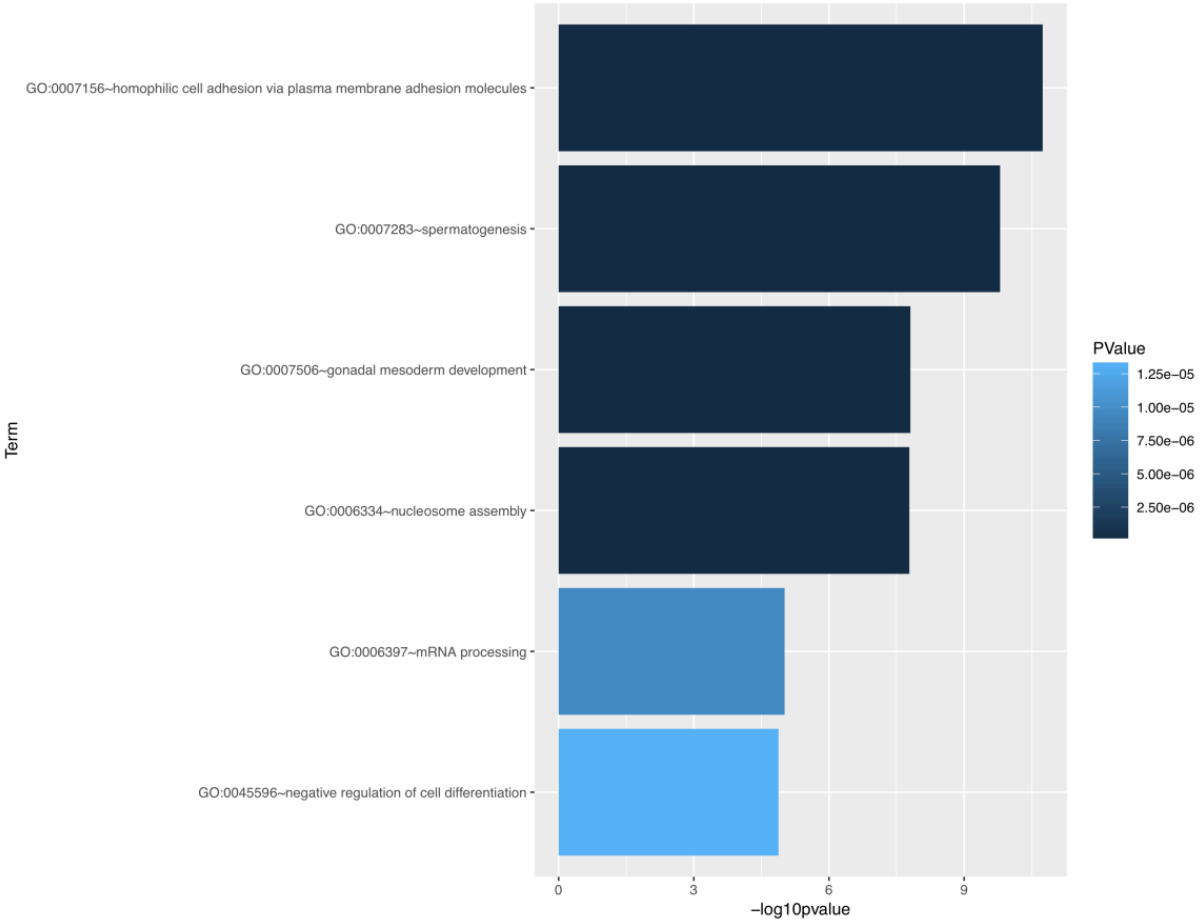

Figure 2.13. Gene enrichment analysis GO terms for the peak CR=1 in SSDr gene pairs

2.4.2. MiRNAs regulation in six vertebrates

2.4.2.1. Identification of WGD, SSD and single copy genes in human, mouse, rat, pig, dog and chicken.

The identification of WGD genes has been approached in previous studies using sequence similarity and synteny (Singh, Arora, and Isambert 2015) (Makino and McLysaght 2010). However, the identification of SSD genes has been done only in an indirect way (Makino and McLysaght 2010). In our study, we consider a previous WGD genes classification (Singh, Arora, and Isambert 2015) that categorises WGD with strict, intermediate and relaxed stringency. We used all 3 datasets, together with the Ensembl classification of paralogous genes to obtain lists of SSD, single copy and the final WGD genes for our analysis, using the following steps. We downloaded all 22, 285 protein-coding gene that have annotated 3'UTRs in human (Ensembl version 88) using Biomart. We also used Biomart to obtain a list of 13,583 paralogous genes, which were considered as the "duplicated gene IDs". We subtracted

duplicated genes from 22,285 geneIDs to give a list of 7,214 single copy genes. In order to get the WGD genes that have annotated 3'UTR sequences, we intersected the "duplicated gene IDs" list with gene IDs from 3,544 strict WGD, 5,504 intermediate WGD and 7,831 relaxed WGD genes classifications from the WGD database and we got 3,415; 5,289; 7,440 final WGD gene, respectively. We subtracted the WGD genes from the "duplicated genes" and to give a list of 10,168 strict small-scale duplication (SSDs), 8,294 intermediate small-scale duplication (SSDi) and 6,143 relaxed small-scale duplication (SSDr). For mouse, rat, pig, dog and chicken the number of 3'UTRs sequences obtained are in table 2.

Table 1.2 Number of geneIDs obtained in each step sequences available for human, mouse, rat, pig, dog and chicken

Category	Human	Mouse	Rat	Pig	Dog	Chicken
Protein Coding	22285	22232	22250	26712	19856	18346
Paralogous	13583	15058	10980	8020	6298	5796
Singlecopy	7214	5514	4218	2736	2443	2954
SOhnolog	3415	3430	2621	1830	1711	928
IOhnolog	5289	5315	4073	2770	2622	1318
Rohnolog	7440	7528	5782	3910	3664	1829
SSDs	10168	11628	8359	6190	4587	4868
SSDi	8294	9743	6907	5250	3676	4478
SSDr	6143	7530	5198	4110	2634	3967

2.4.2.2. WGD genes are preferentially targeted by miRNAs in human, mouse, rat, and pig.

A previous study showed that duplicated genes are preferentially targeted by miRNAs in humans (Li, Musso, and Zhang 2008). We wanted to know whether WGDs and SSD genes are differentially targeted by miRNAs. As discussed in the introduction, duplicated genes can create a stoichiometry imbalance, and we might expect that SSD create more imbalance than WGD, although this proposition is controversial. In order to answer this question, we followed these steps: downloaded mature high confidence miRNA from miRBase version 21 for human, mouse and chicken and for dog, pig and rat; we used the mature miRNAs and obtained 3'UTR sequences from Ensembl version 88 using Biomart for all six vertebrates. We then used the SeedVicious perl script to predict miRNA binding sites in the 3'UTR genes. Using an R script, we calculated the miRNA binding site "density per gene". Later, we used t-test to compare the distributions of "density per gene" values between pairs of WGD, SSD and single copy genes.

Since we have used 3 different WGD genes classifications - strict, intermediate and relaxed - we tested the following two combinations: strict WGD versus the relaxed SSD genes (experiment 1) and strict WGD versus the strict SSD (experiment 2) because we wanted to see if we can get the same results using the strict classification of SSD (experiment1) and when we

include other genes that are more dubious the classification (experiment 2). The t-test results are shown in table 3.

The results of t-test in experiments showed that the mean miRNA binding sites per gene, so we made a pair comparison because we wanted to know whether WGD, SSD or single copy genes are preferentially regulated by microRNAs. We performed the two-side t-test for the experiment 1 and 2. The results show that WGD is more highly regulated than SSDs and SSDr in human, mouse, rat.

In order to have a graphical observation of our density values, we plot the distribution of the "density per gene" value for SSD, WGDs, and single copy gene for human, mouse, rat, pig, dog and chicken using the R package (Figure 18). In addition, we show scatter plot of 3'UTR length versus number of binding sites and the length differences in the WGD, SSD, and single copy genes. We have plotted the results for all experiments, but we only show the plot of experiment 1 as similar density curves were found for experiment 2 (Figure 19).

Table 2.3. Results of t-test for number of miRNAs binding sites density in individual genes compared between WGD, SSD and single copy genes.

Human	Pair-comparison	T-test (pvalue)	mean of x mean of y
	Sohnolog> SSDs	0.04191	0.0627380 0.06243239
	Sohnolog>singlecopy	0.1657	0.06273805 0.06254754
	SSDs>singlecopy	0.7467	0.06243239 0.06254754
	singlecopy>SSDs	0.2533	0.06254754 0.06243239
	Sohnolog> SSDr	0.005705	0.06273805 0.06222667
	Sohnolog>singlecopy	0.1657	0.06273805 0.06254754
	SSDr>singlecopy	0.9464	0.06222667 0.06254754
	singlecopy>SSDr	0.05358	0.06254754 0.06222667
Mouse	Sohnolog> SSDs	0.0004039	0.07444790 0.07366342
	Sohnolog>singlecopy	0.09863	0.0744479 0.0740985
	SSDs>singlecopy	0.972	0.07366342 0.07409850
	singlecopy> SSDs	0.02805	0.07409850 0.07366342
	Sohnolog> SSDr	3.66E-07	0.07444790 0.07317024
	Sohnolog>singlecopy	0.09863	0.0744479 0.0740985
	SSDr>singlecopy	0.9999	0.07317024 0.07409850
	singlecopy> SSDr	0.0001158	0.07409850 0.07317024
Rat	Sohnolog> SSDs	0.04094	0.08154433 0.08092832
	Sohnolog> singlecopy	0.02958	0.08154433 0.08078004
	SSDs>singlecopy	0.3243	0.08092832 0.0878004
	singlecopy>SSDs	0.6757	0.08078004 0.08092832
	Sohnolog> SSDr	0.004572	0.08154433 0.08053843
	Sohnolog> singlecopy	0.02958	0.08154433 0.08078004
	SSDr>singlecopy	0.749	0.08053847 0.08078004
	singlecopy>SSDr	0.251	0.08078004 0.08053847
Pig	Sohnolog>SSDs	0.2672	0.04516970 0.04499588
	Sohnolog> singlecopy	0.0011792	0.04516970 0.04420566
	SSDs> singlecopy	0.001921	0.04499588 0.04420566
	singlecopy>SSDs	0.9981	0.04420566 0.04499588
	Sohnolog>SSDr	0.07233	0.04516970 0.04472559
	Sohnolog> singlecopy	0.001792	0.04516970 0.04420566
	SSDr> singlecopy	0.04076	0.04472559 0.04420566
	singlecopy>SSDr	0.9592	0.04420566 0.04472559
Dog	Sohnolog>SSDs	0.925	0.05047239 0.05087088
	Sohnolog>singlecopy	0.146	0.05047239 0.05014620
	SSDs> singlecopy	0.00261	0.05087088 0.05014620
	singlecopy>SSDs	0.9974	0.05014620 0.05087088
	Sohnolog>SSDr	0.9561	0.05047239 0.05101123
	Sohnolog>singlecopy	0.146	0.05047239 0.05014620
	SSDr> singlecopy	0.002002	0.05101123 0.05014620
	singlecopy>SSDr	0.998	0.05014620 0.05101123
Chicken	Sohnolog>SSDs	0.407	0.01883922 0.01878611
	Sohnolog>singlecopy	0.7785	0.01883922 0.01902567
	SSDs>singlecopy	0.9199	0.01878611 0.01902567
	singlecopy>SSDs	0.08006	0.01902567 0.01878611
	Sohnolog>SSDr	0.3046	0.01883922 0.01872081
	Sohnolog>singlecopy	0.7785	0.01883922 0.01902567
	SSDr>singlecopy	0.9564	0.01872081 0.01902567
	singlecopy> SSDr	0.04356	0.01902567 0.01872081

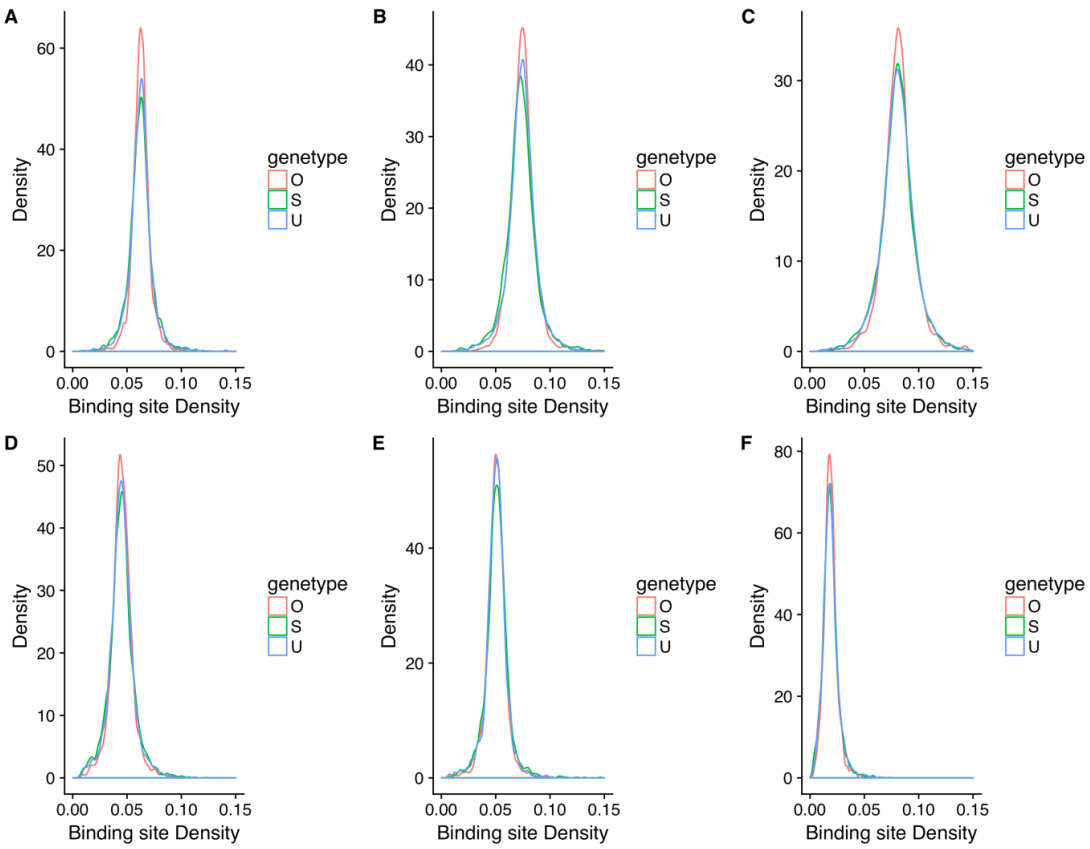

Figure 2.14. The distribution of miRNAs binding sites density for strict WGD (red), SSDr (green) and single copy (blue) for experiment 1. (A) human, (B) mouse, (C) rat, pig, dog and chicken. O, Whole genome duplicated gene; S, small-scale duplicated gene

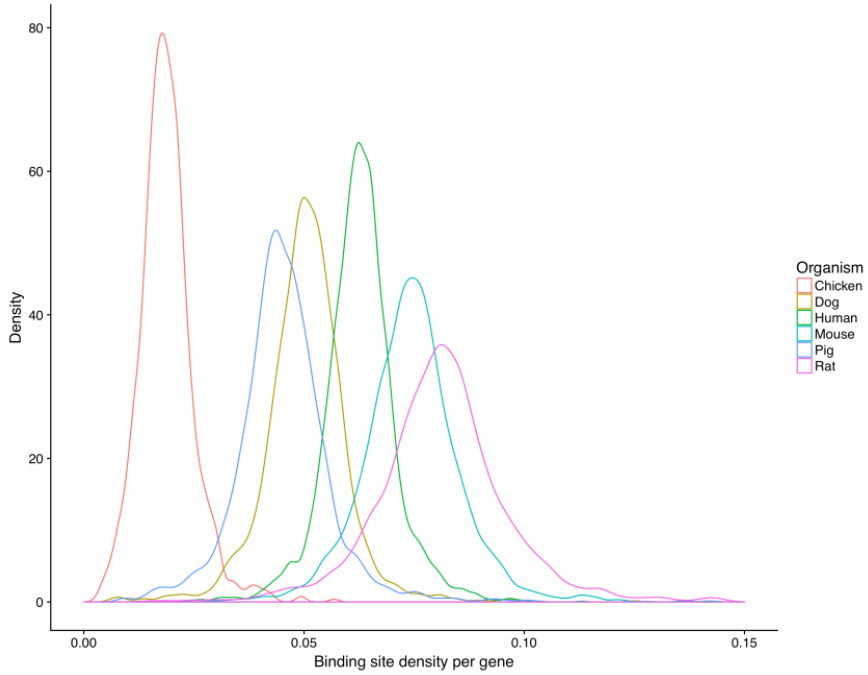

Figure 2.15. Comparative distribution of miRNAs binding sites density for WGD genes in human (green), mouse (dark green), rat (pink), pig (blue), dog (yellow) and chicken (orange).

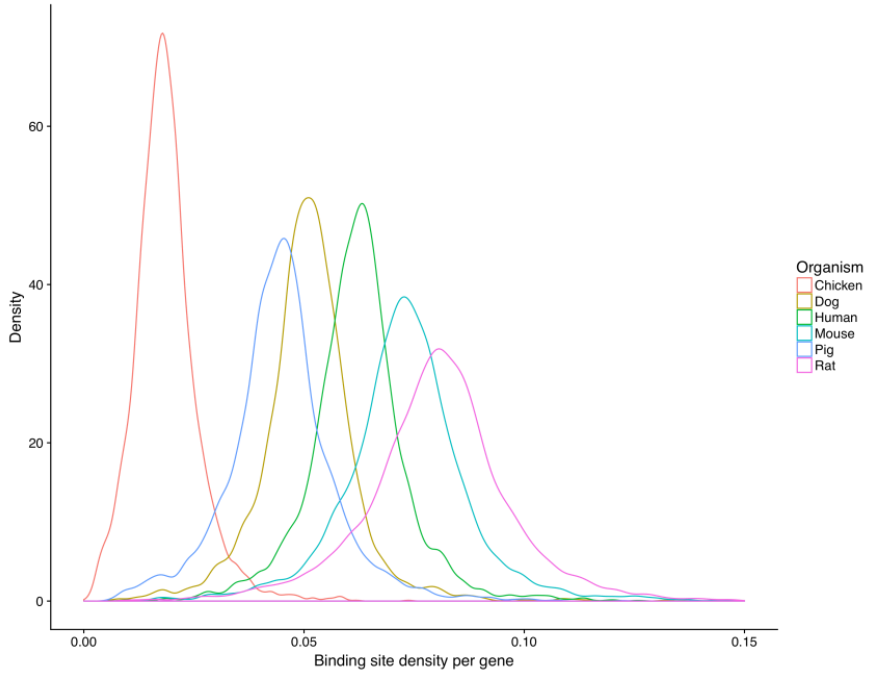

Figure 2.16. Comparative distribution of miRNAs binding sites density for SSD genes in human (green), mouse (dark green), rat (pink), pig (blue), dog (yellow) and chicken (orange).

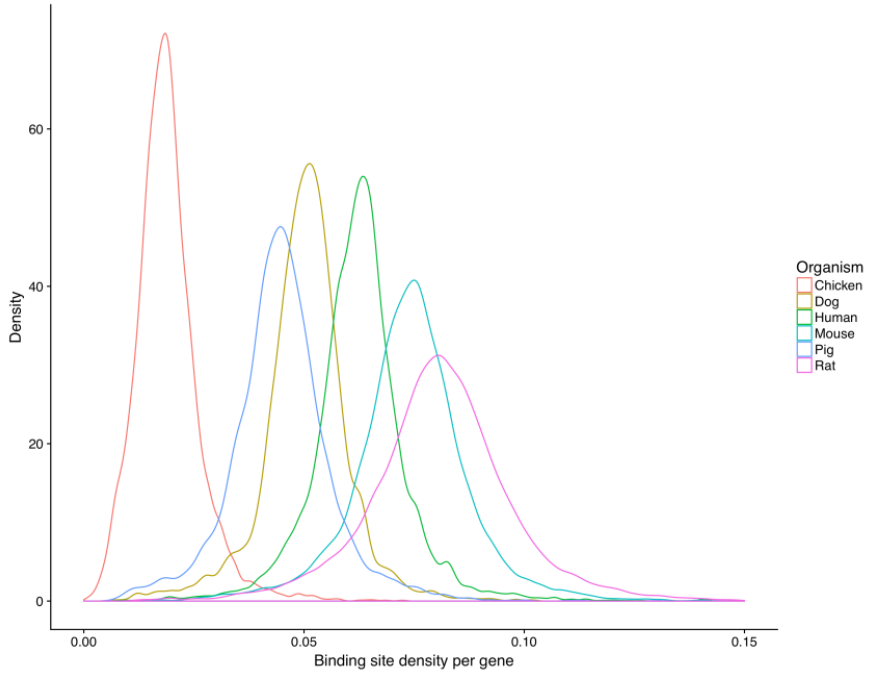

Figure 2.17. Comparative distribution of miRNAs binding sites density for single copy genes in human (green), mouse (dark green), rat (pink), pig (blue), dog (yellow) and chicken (orange).

2.4.3. Fast and slow evolving WGD genes are equally regulated by miRNAs

We found that there is no preferential regulation by miRNAs for slow and fast evolve WGD genes using t-test, being the mean of gene A and mean of gene B (Which is the fast and slow) 0.06281088 and 0.06252448 respectively (p-value = 0.1956). (Figure 23).

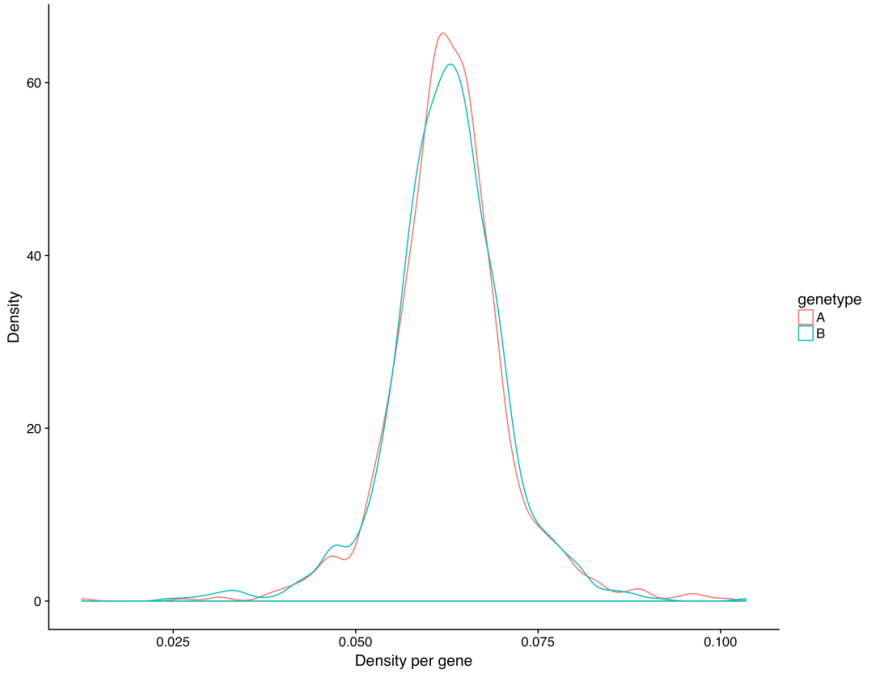

Figure 2.18. Distribution of miRNAs binding site density per gene in slow and fast WGD genes

2.4.4. Haplo-insufficient genes are preferentially regulated by miRNAs in human

The total number of genes used for the analysis were 3863 haploinsufficient (HI), 3893 haplosufficient (HS), and 11,928 genes that were excluded from the analysis (E). The result of the t-test gave the results that the HI is preferentially regulated than HS with a p-value of 0.004394 and with a mean of HI and HS of 0.06299407 and 0.06233690 respectively (Figure 24 and 25)

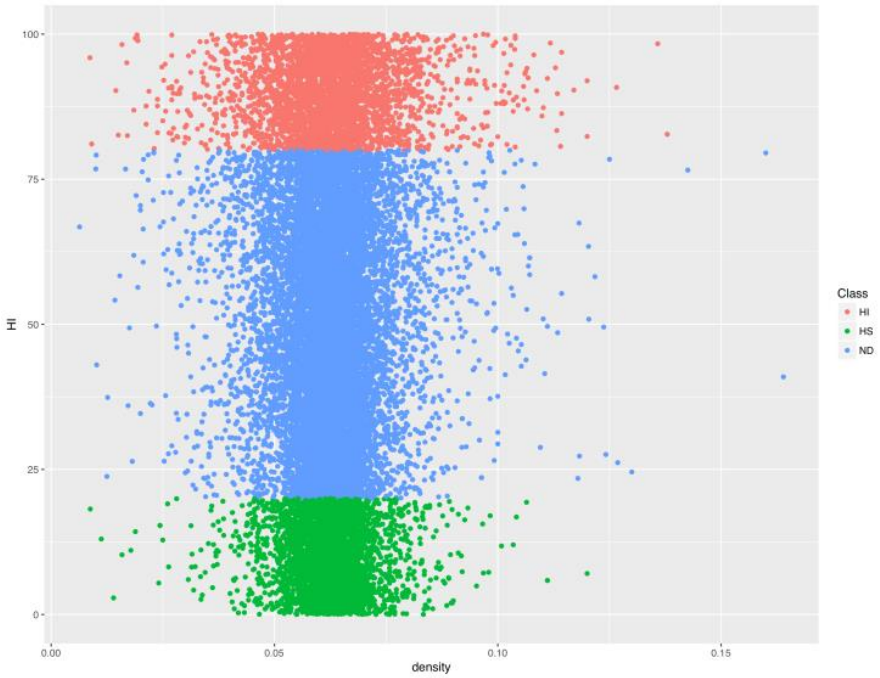

Figure 2.19. Distribution of binding site per gene in haploinsufficient (HI) and haplosufficient (HS) genes in human

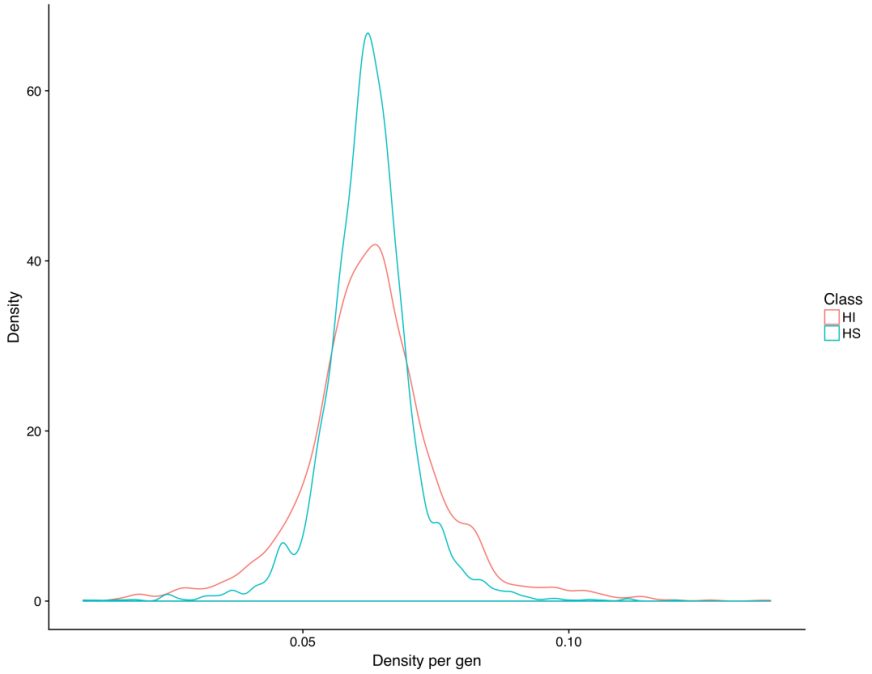

Figure 2.20. Distribution of miRNA binding site density per gene in haploinsufficient (HI) and haplosufficient (HS) genes in human.

2.4.5. Physical and non-physical interaction genes have no preferential regulation by miRNAs

We obtained 3,710 ensembl geneIDs for the list of physical genes and 17,648 as non-physical genes. The t-test gave the results that neither physical and non-physical genes were

regulated preferentially, with a p-value of 0.7685 and with a mean of physical and non-physical of 0.06216061 and 0.06231448 respectively (Figure 25 and 26)

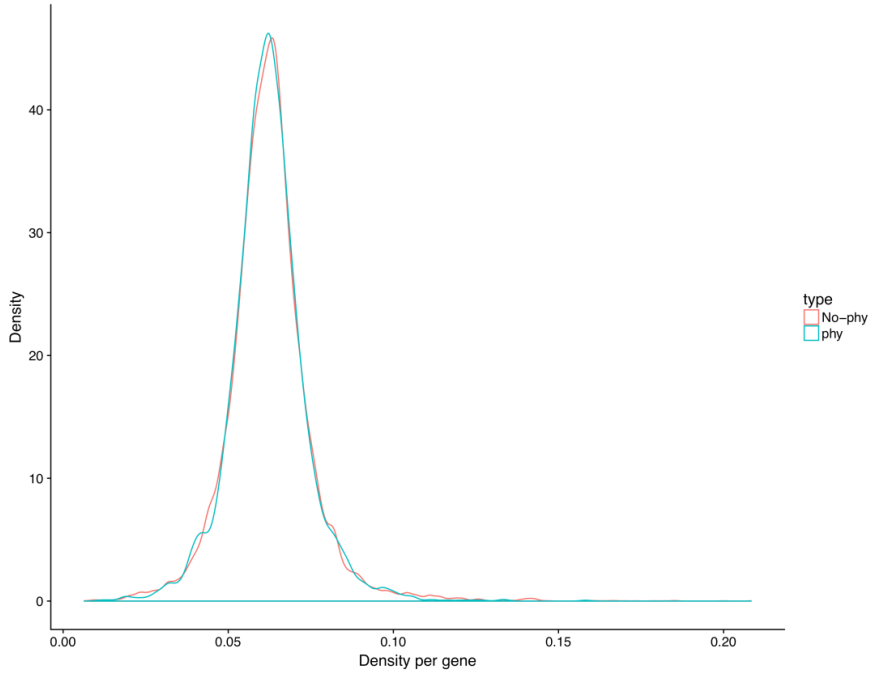

Figure 2.21. Distribution of Density per gene Values in physical and non-physical interaction genes

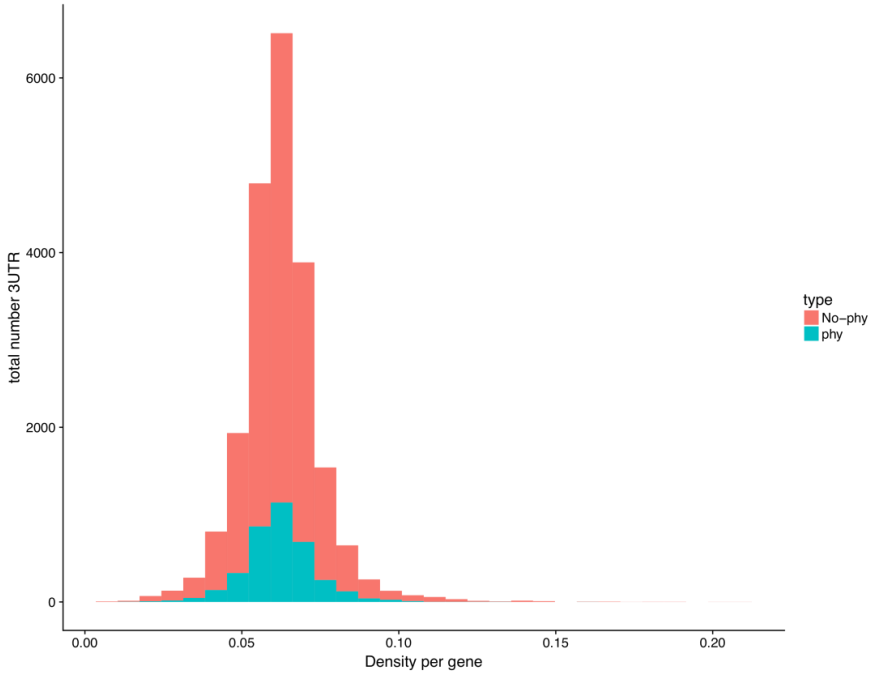

Figure 2.22. Histogram of Density per gene Values in physical and non-physical interaction genes

2.4.6. Non-essential genes are primarily targeted by miRNAs

There are many ways to evaluate the importance of a gene in a genome; one of them is "essentiality". Essential genes are those that cause lethality and sterility on deletion. A previous study in mouse showed that WGDs are more likely than SSD genes to be essential (Makino, Hokamp, and Mclysaght 2008). In human, they found a correlation between WGD genes and their essentiality, based on their dosage balance. Our interest was to find a connection in the regulation by miRNAs in human essential genes and WGDs. To accomplish this, we downloaded 20,684 Ensembl gene IDs from the Online GEne Essentiality database and cross-sectioned with the 21,713 gene IDs of the density per gene values used in the previous analysis above. We matched 19,383 Ensembl gene IDs, of which 1,520 were essential genes and 17,863 non-essential genes. We performed Wilcoxon two-sided test, which showed that the mean miRNA target site density non-essential and essential genes were 0.391 and 0.385 respectively, with a p-value 0.00121. This result tells us that non-essential genes are more regulated by miRNAs than essential genes (Figure 27).

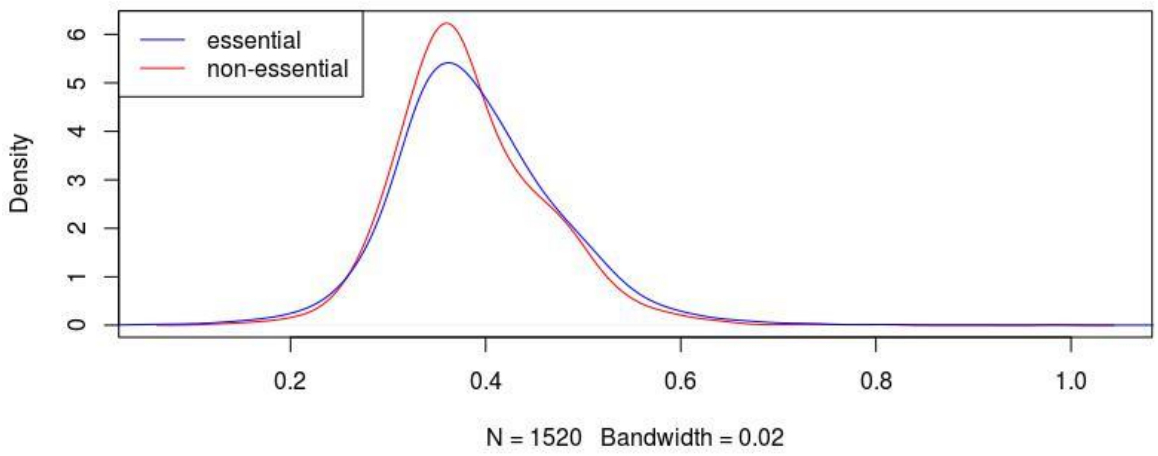

Figure 2.23. Distribution of miRNA target site density per gene in human essential and non-essential genes. N is equal to number of essential genes

2.5. Discussion

2.5.1. Properties of WGD an SSD

Genome duplication is an important process in Evolutionary Biology because it is a source of new genes in the organisms. We know that SSD event has happened in different time

points during evolution in human being, however the WGD in humans only happened in two rounds times, so we expect to find SSD genes as ancient as WGD genes and SSD that was recently duplicated. The analysis of how miRNAs targeted to WGD gene pairs or SSD gene pairs, can allow us to quantify how many conserved target sites is shared within the gene pairs and family members in WGD and SSD genes. As shown in Figure 8, for WGD gene pairs, the CR values range from 0 to 0.5 approximately. We also can observe that the SSD has the same pattern of distribution for CR from 0.0 to 0.5, but in addition the SSD has gene pairs with values from 0.5 to 0.90 ending in a small peak at CR=1.0. This density plots show that there are ancient SSD and WGD genes that are within value CR 0 to 0.5 and there are SSD that has a more recent event duplication process, pair of genes that goes CR from 0.5 to 1.0. CR. CR =0 represent miRNAs that share few target sites in the sequences, CR=1 miRNAs that shared almost all target sites in the sequences.

In Figure 10, we show the percent identity of each pair of sequences in WGD and SSD gene pairs are between 0% to 40% approximately, however SSD present pair of genes that not only share 40% of percent identity but also from 40% to 90% and some pairs create a small peak of 100 percent identity. It is important to consider that WGD genes are 450m years old and there are SSD genes that are as ancient as WGD genes. The low percent identity found in WGD and SSD pair and family genes (Figure 10,11) could be explained by: 1) Splicing processes that create a different variant of the 3'UTR; hence we can get different sequences for 3'UTRs in duplicated genes, so this misalignment can make, we compare sequences very different, 2) miss annotation of the 3'UTR, despite that human, is the best annotated vertebrate organism, it still has a significant fraction of miss annotations which would affect the alignments of small sequences against large sequences.

Previous analysis has been made for human WGD and SSD genes (Acharya and Ghosh 2016). These have shown that WGD genes show less functional similarity than SSD genes, WGD genes are found in more sub cellular localization than SSD, WGD gene have different gene expression pattern than SSD genes. WGD genes show more adaptation to a new function compared to SSD, WGD genes are more evolutionary conserved compared to SSD genes, WGD genes are associated with more functions than SSD genes using the GO biological process and Pfam domain analysis. Furthermore, WGD genes are associated with more essential genes than SSD genes and WGD genes are disease associated than SSD genes (Acharya and Ghosh 2016), our study only corroborate that WGD genes share less microRNA within gene pairs or gene families compared to SSD recently duplicated genes

WGD and SSD genes have to undergo for different process to be preserved, WGD pass through genomic rearrangements while, SSD genes have to be fixed in the population (Inoue et al. 2015). For SSD, small number of genes survives this process because they are exposed to natural selection, and only the duplication of genes that confer an advantage to the organism will be preserved (Lynch 2001). Evidence for how only a few SSD genes are preserved is show in Figure 8 and 9. When the CR values in SSD gene pairs ranges from 0.5 to 1.0 the thinner line in the density plot show that there are low numbers of gene pairs that has those values until a small peak is reached at CR=1.0, that line extension and small peak show in the SSD gene pairs and gene families, represent sequences that are still share high CR value and. In the Figure 8 and 9, what is not clear is that why SSD genes have a great number of gene pair and gen families located in this region. If SSD duplicated genes are more exposed to natural selection, why the majority of their gene in SSD pairs has the same pattern distribution of WGD gene pairs? The answer could be that those ancient genes at some points are exposure to relaxation process.

We expected to find a correlation between the conservation ratio and percent identity of the gene pair and gene family analysis in WGD and SSD genes. Conservation ratio depends on miRNAs target sites, which is sequences based, so we expect taking any pair of sequence we will find a degree of correlation. In general, there are some common patterns in the four correlation plots for WGD and SSD genes (Figure 12, 13, 14, 15, 16 and 17). When we have values CR=0 and PI from 0-5% the alignments are between small and large sequences. This feature probably makes that they do not share miRNAs in common as they only share small nucleotide sequence in common. When we take values CR=0.4 and PI from 0-40%, in this region of the plots, we have long sequences; when we have CR=0 and PI is close to 100%, the alignments are small and finally only in SSD pair of genes we can see when the CR=1 and PI=100, we verified that those pair and gen family sequences has 100% we found alignment with different length sequences.

In SSDr gene pairs and SSDr gene families, we observed in our plots a small peak at CR=1 (Figure 13 and 15). When we compare the CR values and the PI they got the maximum value 1.0 and PI 100, we propose that this result can be explained by two alternatives: 1) they are recently duplicated SSDr and they have not diverged so much since the duplication event, hence they share all the same miRNAs that targeting them or 2) this is the product of gene annotation error, since even for human it is common to have the same sequences named as two different genes. In order to investigate this, we decided to do an enrichment analysis on this peak, the results gave us spermatogenesis, gonadal mesoderm development, nucleosome

assembly GO terms. Additionally, we also retrieved information about gene description of those genes from that peak, we found 18 out of 918 genes cancer/testis antigen family 45 member, 16 out of 918 protocadherine gamma subfamily, G antigen, defensina beta, proline rich and olfactory receptor family. The olfactory, testis antigen family and G antigen are mentioned as recently duplicated genes (Guschanski, Warnefors, and Kaessmann 2017). However, mapping those Ensembl gene IDs in the human genome to make sure that they are not allelic variant of the same gen annotated as different genes.

The CR analysis in the WGD duplication can show as how miRNAs are shared in gene pair and gene families in WGD and SSD. Results have shown in SSD gen pair and gene families has not change in sequence nucleotides of the 3'UTR and they have shared all miRNAs between pair and member of the families. To look if for those genes with that characteristic, if they are highly repressed by miRNAs, a number of binding sites or target sites determination is necessary. Guschanski et al., 2017 show that a recently duplicated (SSD) gene shows a low expression compared to an ancient duplicated gene (WGD).

We propose it is necessary an additional comparison between the conservation ratio and age of duplication event of WGD and SSD.

2.5.2. WGDs are preferentially targeted by miRNAs in human, mouse, rat, pig

Regarding our hypothesis of the stiochometric imbalance, we expected that the WGD genes cannot cause a stiochometric protein problem in the cell because the duplicated all the proteins, however the SSD event can create an imbalance as the duplication create a disparity in the protein interaction in complex proteins, this. Our results give the WGD duplication is more highly regulated by miRNAs compared with SSD genes in human, mouse, rat and pig.

We propose that the SSD is not highly regulated as we propose in our hypothesis, because the process by which the SSD genes have to be preserved is more stringent and give less chance to be preserved, they need to be fixed in the populations and in other organism very few of them are preserved (Davis and Petrov 2005) (Lynch 2001). So, there is not enough time that miRNAs can apply a role in decreasing the expression of the SSD because they disappear from the population before to be fixed, the probability that such genes be fixed in the population is 1/4N (Lynch 2001).

The fact some single copies are more regulated that the SDS duplication can be explained because those genes where before duplicated genes that comes from a whole genome duplication

2.5.3. Haploinsufficient genes are preferentially regulated by miRNAs in human

This analysis could be improved if we link haplosufficient and haploinsufficient genes to WGD and SSD, if we grouped them in haploinsufficient-whole genome duplicated genes (HI-WGD), haploinsuficient-small scale duplicated genes (HI-SSD), haplosuficient-whole genome duplicated genes (HS-WGD) and haplosuficient-small scale duplicated genes (HS-SSD), the first two analysis to identify what kind of haploinsuficient gene are preferentially regulated by miRNA, and the last two analysis to know if when they are separated can be a differential in this group and we could not detect for slight difference. The fact that HI genes are preferentially regulated it has sense considering the evolutionary point of view. One of the features of HI genes is that they are longer genes, not necessarily longer 3UTRs but it is something that could be checked. Another feature is that these genes are expressed in liver in the early embryo stage.

2.5.4. Physical versus non-physical gene list

We did not get any preferentially regulation by miRNAs for any of these groups. We consider that the classification of the groups of physical or non –physical interaction for the analysis was not accurate. Maybe a better way to do this analysis could be, select specific protein that we consider as complex proteins such the ribosomes.

2.5.5. Non-essential genes regulation by miRNA

It was found that WGDs were enriched with essential genes compared with SSD genes (Acharya and Ghosh 2016).To our knowledge, no similar study has been reported. Further work is required to test the impact of the relationship between essential genes and WGDs on the miRNA target density differences, as well at 3'UTR length distributions.

Blank page

microRNA and protein expression in *Parasteatoda tepidariorum* embryogenesis

3. MicroRNA and protein expression in *Parasteatoda tepidariorum* embryogenesis

3.1. Abstract

MicroRNAs are short non-coding RNA ranging in size from 21 to 24 nucleotides. They regulate the expression of protein coding genes at the post-transcriptional level. MicroRNAs are expressed throughout embryogenesis in animals and are suggested to have important roles in development. Most developmental studies of microRNA expression have been performed in Drosophila melanogaster, which has a long germ mode of development. However, the ancestral mode of development in arthropods is short germ. *P. tepidariorum* is a short germ model arachnid, with emerging use for developmental studies. We obtained a microRNA profile from 10 time points spanning the first 95 hours of *P. tepidariorum* embryogenesis. We identified eight cluster of microRNAs across the ten developmental time points. The main cluster were highly expressed at early and late stage of embryogenesis. We identify clusters that are highly expressed in the early cleavage and blastoderm formation, cluster with high microRNA expression at later embryo, and a small cluster of microRNAs whose expression changes significantly during the maternal zygotic transition. There are clusters of microRNAs that we suggest may be important for the middle stage development, and for limb and brain differentiation.

3.2. Introduction

MicroRNAs are short non-coding RNAs, ranging in size from 21 to 24 nucleotides. Their functions are to regulate gene expression at the post-transcriptional level, both in animals and plants. There are at least two mechanisms of regulation: degradation of the mRNA and the blocking the translation into proteins. In animals, microRNAs generally bind target sites located in the 3'UTR of the mRNA.

In the last years, the function of many microRNAs have been studied, and it is possible to distinguish microRNA that act in cell-specific type and others microRNA that act in a more global manner. There are two kinds of groups of microRNAs, expressed at early stages and later stages of embryogenesis. MicroRNAs that are highly expressed at early stages of embryogenesis regulate the maternal mRNA clearance, cell communication, control proliferation, apoptosis and microRNA highly expressed in later stages are related to cell differentiation (Alberti and Cochella 2017).

P. tepidariorum is easy to maintain in the laboratory, has a short life cycle, a good number of offspring are produced per cocoon. From an experimental perspective it is amenable to techniques such as RNAi and in situ hybridization (Hilbrant, Damen, and McGregor 2012). These features have helped to make *P. tepidariorum* an important model organism for developmental biology, evolution and genetics studies. *P. tepidariorum* is known to develop via the short germ developmental mechanism, which is ancestral in arthropods. *P. tepidariorum* has a genome and transcriptome sequences published.

The *P. tepidariorum* genome size is of 1443.9 Mb with 1,642 scaffolds obtained (Schwager et al. 2017). 148 microRNAs were identified that represent 66 families (Leite et al. 2016). This study was made by taking a pool of ten stages. This is the first work in studying the microRNA expression profile of *P. tepidariorium* embryogenesis of these ten stages.

3.3. Methods

3.3.1. Small RNA sequences analysis

The RNA-seq data were obtained from NCBI with accession number PRJEB13119. Samples were collected from midpoints of the times for each stage define in Mittman(Mittmann and Wolff 2012). The time developmental points were 5 h, 13 h, 21.5 h, 29 h, 35.5 h, 45.5 h, 53 h, 65.5 h, 80.5 h, 90.5 h. The adapters were removed from reads using the Cutadapt tool (https://cutadapt.readthedocs.io/en/stable/). Reads shorter than 17-nucleotides were then discarded. tRNAs were predicted in the genome using the tRNAscan-SE (Lowe and Eddy 1996), and then reads mapping to those predicted tRNAs were removed. Remaining reads were then mapped against the *P. tepidariorum* dovetail genome assembly using bowtie2 (Langmead et al. 2009) with the following parameters (--very-sensitive). The reads mapped to the previously annotated microRNAs were filtered using htseq-count with the following parameter (htseq-count -f bam -r pos -s yes) and reads were normalized by counts per million (CPM) using edgeR R package (Robinson, McCarthy, and Smyth 2010). A heatmap of microRNA expression was produced using pheatmap in R package.

3.3.2. mRNA sequences analysis

The experiment of time developmental embryogenesis was made in a laboratory in Japan (Iwasaki-Yokozawa, Akiyama-Oda, and Oda 2018). The mRNA sequencing files were obtained from European Nucleotide Archive (ENA) with PRJNA448775 accession number (https://www.ebi.ac.uk/ena/browser/view/PRJNA448775). Trim Galore v.0.6.4 was used to remove the adaptor and the sequences less than 15 bp long running by default parameters (https://github.com/FelixKrueger/TrimGalore). Cleaned reads were mapped to *P. tepidariorum* genome dovetail version using HISAT2 v.2.2.1 (Kim et al. 2019) with default parameters. FeatureCounts v 2.0.3 (Liao, Smyth, and Shi 2014) was run by default for count the mapped reads, the cDNA library was strand specific, considering the -t gene. The protein annotation was provided by the McGregor lab (http://mcgregor-evo-devo-lab.net/McGregor_lab/home.html). DESeq2 v.1.30.1 (Love, Huber, and Anders 2014) were used for the normalization and an adaptation of the Turner's script were used for the plots (https://gist.github.com/stephenturner/f60c1934405c127f09a6)

3.3. Results

3.4.1. Expression of microRNAs in *P. tepidariorum*.

148 microRNA was characterized that belong to the 66 families in *P. tepidariorum* embryo (Leite et al. 2016). *P. tepidariorum* has fourteen developmental stages that spanning 185 hours. We analysed how microRNAs are expressed across the ten first developmental stages that consist of early cleavage, blastoderm, germ disc formation, primary thickening, cumulus migration, dorsal field, germ band, prosomal limb buds, limb differentiation, and brain differentiation.

We mapped the cleaned reads to microRNA sequences; these microRNAs sequences were previously annotated (Leite et al. 2016). We count the reads using htseq-count and the data were normalized by counts per million (CPM), heatmap and line plot was produced. The heatmap (Figure 3.1) and the line plot (Figure 3.2) have very similar overview.

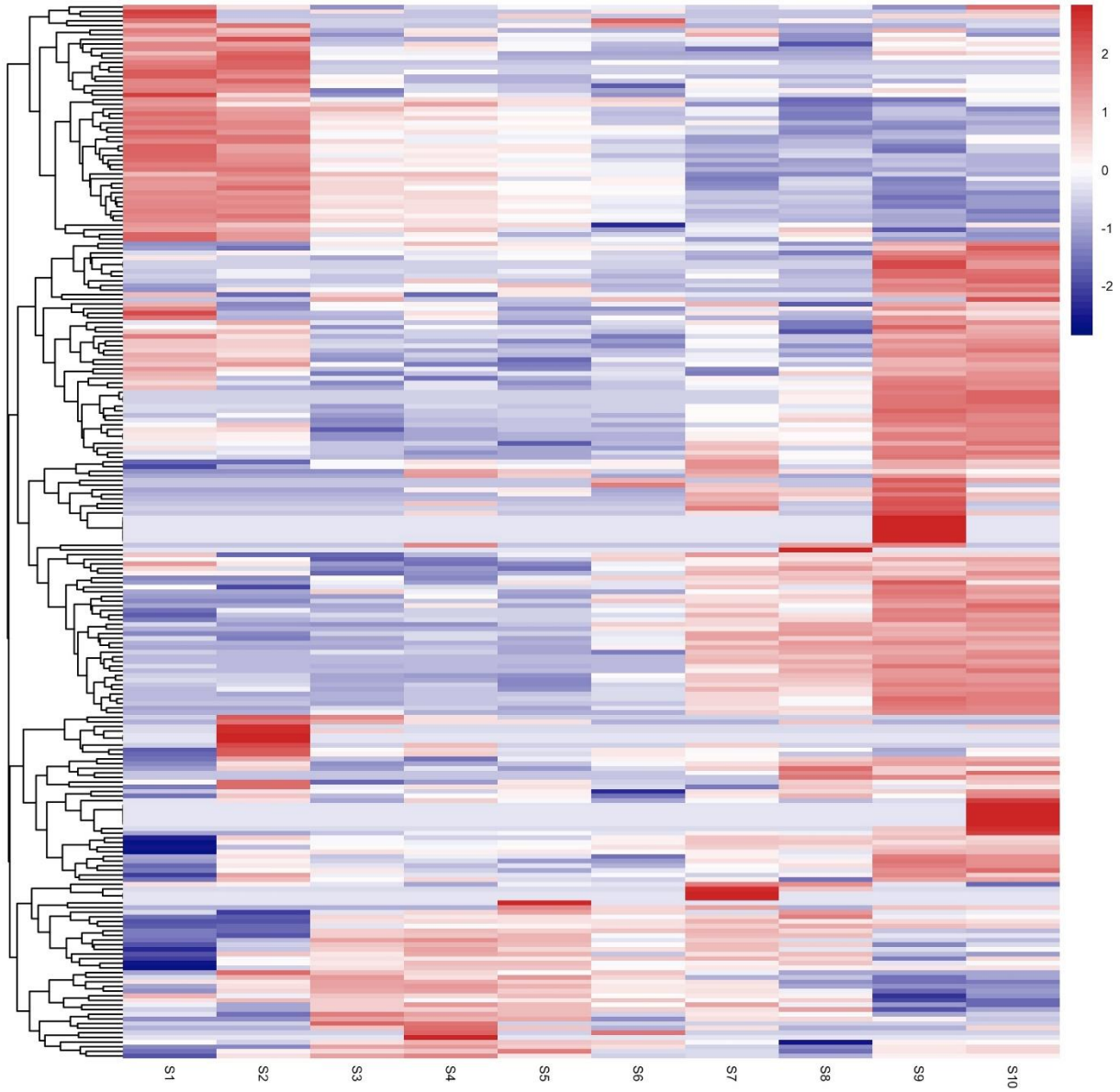

Figure 3.1. Heatmap showing microRNA expression across ten (S1-S10) development time points in *P. tepidariorum* embryogenesis. The red colour represents the upregulated genes and the blue colour the downregulated genes. Each row represents the expression of a microRNA across ten stages (columns).

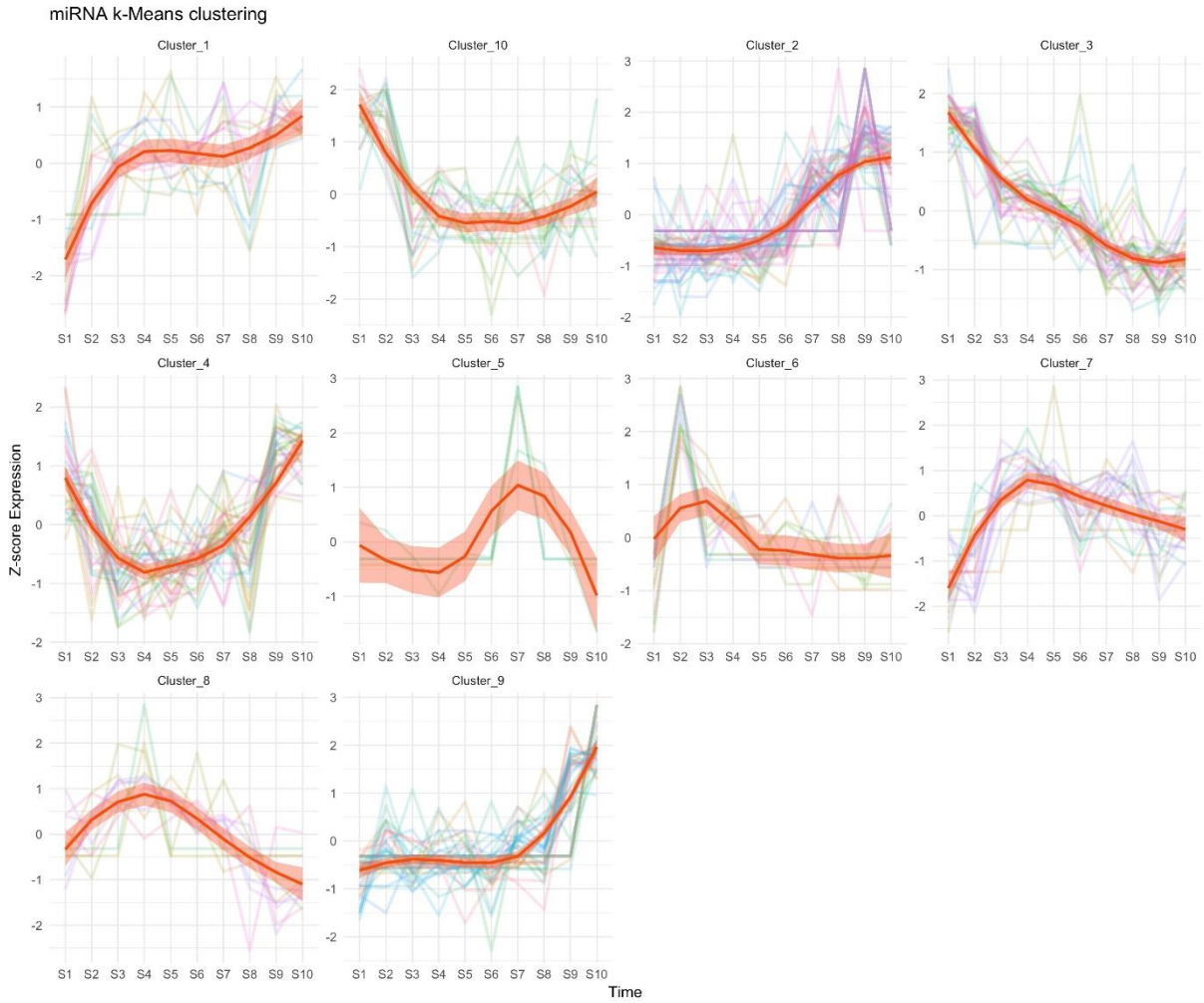

Figure 3.2. Line plot of clusters of co-expression microRNAs across differential ten first stages (S1-S10) of *P. tepidariorum* embryogenesis.

The ten first stages from *P. tepidariorum* are early cleavage (S1-5h), blastoderm (S2-13h), germ disc formations (S3-21.5h), primary thickening (S4-29h), cumulus migration (S5-35.5h), dorsal field (S6-45.5h), germ band (S7-53h), prosomal limb buds (S8-65.5h), limb differentiation (S9- 80.5h), brain differentiation (S10-90.5h).

Figure 3.1 shows the expression of the ten stages of *P. tepidariorum*. We can observe that in the stage 1 and 2, there are a highly expressed microRNA, which correspond to the cluster_3 and cluster_10 (figure 3.2), the microRNAs are important in the early cleavage and blastoderm formation. Later those microRNAs are downregulated and remain low express until the end of embryogenesis. In S2, we can observe another small cluster that is highly expressed; this may be the zygote microRNAs expression. On S1 and S2 stage top, we can observe a cluster of microRNAs that are highly expressed at early stage and low expressed in the later stages (figure 3.1). This is exactly the cluster_10 in the line plot (very high in S1 and S2 stage

and decrease in the others). At the germ disc formation (S3), primary thickening (S4), cumulus migration (S5), microRNAs in clusters 1, 7 and 8 appear to be upregulated, remaining at high levels until S5 but then decreasing, the microRNAs might be important for middle stage of development. At the stage of S9 and S10 which are limb differentiation and brain differentiation, we observed two main clusters that are highly expressed – clusters 2 and 9. Clusters 5 and 6 contain microRNAs with highly variable expression patterns, and we are unable to draw obvious conclusions about those microRNAs.

Heatmap and line plot are similar analysis. The difference is the k-means clustering, since the heatmap is unsupervised meaning that we do not have to ask how many clusters you want while the line plot is supervised so we tell the function the number of clusters you need (10 because the number of samples), and this will cluster the microRNA based on their gene expression in ten defined clusters. In the line plot we can clearly see clusters of microRNAs that follow a specific trajectory across time.

3.4.2. Protein expression in *P. tepidariorum* embryogenesis

To study the protein expression of *P.tepidariorum* the reads were removed and sequences longer than 15 bp nucleotides long were mantained. The reads were mapped against the *P. tepidariorum* genome, and count was done with featureCounts, the normalization and differential expression were estimated with DESeq2.

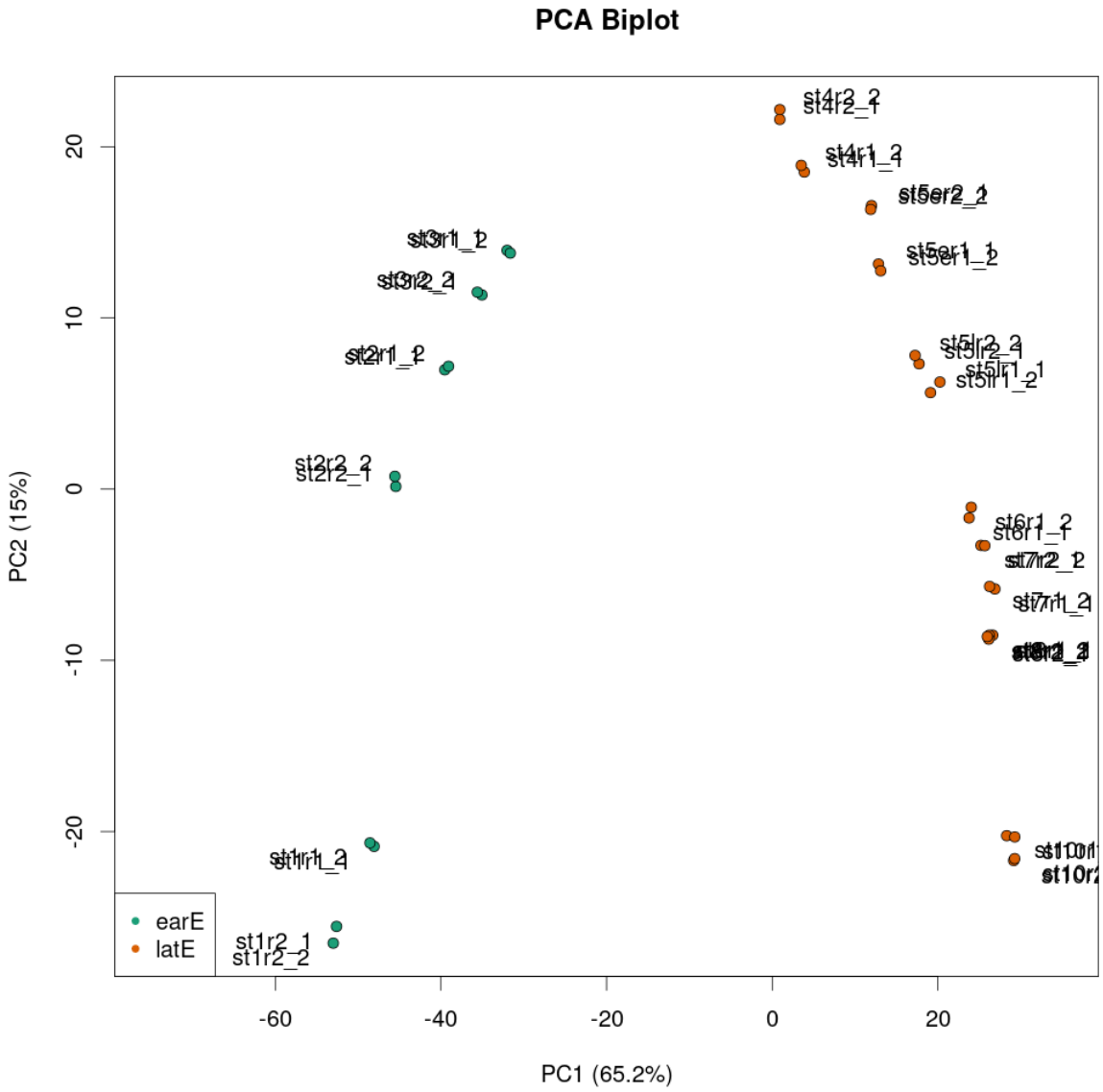

Figure 3.3. Principal component analysis of the ten developmental time points in *P.tepidariorum* embryogenesis. Protein expression in stages: st1, st2, st3, st4, st5e, st5l, st6, st7, st8, st10.

The PCA analysis from protein gene expression in the ten developmental stages (st1, st2, st3, st4, st5e, st5l, sta6, st7, st8, st10) show that the PC1 represent the 65% of the variance and PC2 15%, together represent 80% of the total variance. There are two main groups labelled as early (earE) and late (latE) embryogenesis stages, the earE group contain st1, st2, st3 and late group contain st4, st5e, st5l, st6, st7, st8, st10. There is a big shift between earE and latE groups. In earE group shows that there are two subgroups st1 forming a group and a little far

from st2 and st3. Inside latE group there are three subgroups st4, st5e, st5l, another from st6, st7, and st8 and the last group with st10. It also important to mention that the biological replicates are similar, and this is a good sign of the experiment and sequencing

The earE and latE group in figure 3.4 represent the similarity in expression as we are representing developmental time points. The subgroups can explain the similar expression of close time point. The variation of the gene expression inside of subgroups are seems to be gradually as they are cluster together and maybe the big shift can be explained by big expression changes in the embryo as the maternal zygotic transition and brain formation in the spider. The difference of st1 gene expression it could be due to in this stage, the molecules of the mother are predominant in this stage, including proteins.

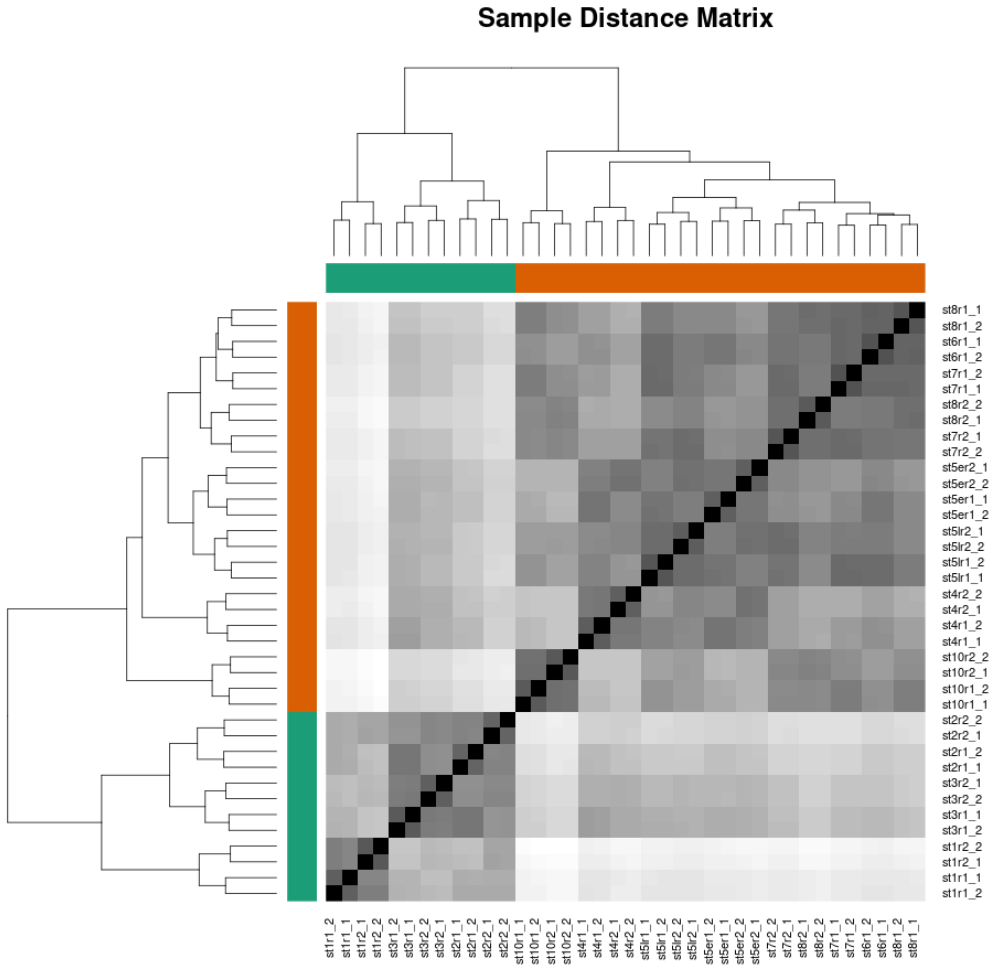

Figure 3.4. Hierarchical clustering of the ten developmental time points of embryogenesis from *P.tepidariorum* for proteins expression

In the figure 3.5 show pairwise comparison. We can observe clearly in the figure 3.5 similar expression in the earE stages and also similar expression as they are cluster together in latE stages

Table 3.1. Top differential regulated protein genes expression of ten developmental time point of embryogenesis in *P. tepidariorum*

Gene	log2FoldChange	padj
aug3.g9562.t1	-4.410764995	6.19E-124
aug3.g3173.t1	-6.581983416	2.34E-122
aug3.g9725.t1	-8.762622148	5.57E-105
aug3.g14183.t1	-5.936502633	3.98E-98
aug3.g9196.t1	-3.993824843	3.69E-97
aug3.g17477.t1	3.609664696	6.63E-97
aug3.g27589.t1	-5.638909088	5.18E-90
aug3.g23929.t2	-5.348508582	1.71E-88
aug3.g129.t1	-7.957270562	2.22E-85
aug3.g13384.t1	2.856035094	1.47E-84
aug3.g8887.t1	2.235413341	6.19E-79
aug3.g15668.t1	-8.350953698	1.16E-78
aug3.g10890.t1	-6.267703793	3.83E-78
aug3.g2321.t1	-3.511866096	5.21E-78
aug3.g10586.t1	-8.621004828	9.28E-75
aug3.g9728.t1	-8.919321609	1.07E-74
aug3.g14054.t1	3.389339344	5.20E-74
aug3.g18398.t1	-7.700255512	2.25E-73
aug3.g13141.t1	-8.333982368	4.44E-73
aug3.g18792.t1	-7.212051246	4.44E-73

We identified the top twenty regulated protein genes based on p-value adjustment from the ten developmental time points of embryogenesis in *P. tepidariorum* (Table 3.1)

3.4. Discussion

Eight microRNAs clusters were identified using k-means clustering (Figure 3.2), some of these clusters are highly expressed, only at early, only at late, or at early and late embryogenesis stages. The early and late embryo expression is a pattern common to *D. melanogaster* and *C.elegans*, it was proposed that this feature is general to animal development (Avital and Franc 2017). We have not investigated yet what are those microRNAs expressed at early stages and late stages, but in *C. elegans* and *D. melanogaster*, they identify that microRNA expressed at later stages of embryogenesis are conserved and the ones of early stage as young genes.

A recent review proposed that this early and late expression pattern can explain the main characteristics and roles that microRNAs play in the animal developmental. Alberti., 2017 proposed that the early expressed microRNAs have the function in the maternal clearance, cell proliferation, apoptosis, and cell signalling. Later expressed microRNAs are often involved in cell differentiation. For example, the mir-309 cluster in Drosophila is expressed early after zygotic genome activation. When it is knocked down, the cells accumulate maternally deposited mRNAs. However, Avital et al., 2017 proposed that that the main roles of microRNAs are at late stages of embryogenesis. They proposed a model in which microRNA do not regulate early embryogenesis but rather later stage for the differentiation of cell types.

Avital et al., 2017 found that microRNA expressed late were conserved and function as fine-tuning to regulate their targets, in contrast microRNAs that were expressed at early stages are young and specific to the genus and work as repressor. They proposed a bimodal role in the function of microRNA: repressor and tuning, they also proposed that the double function of microRNAs as repressor at early stage and a fine-tuning at late stage could have its origin on the type of ago protein that bind to microRNA.

We obtained a profile expression of microRNAs in the first ten stage of embryogenesis of *P. tepidariorum*. However, we still need to investigate what are those genes that are highly expressed in the different stages and make a correlation between the highly microRNA expression with the downregulation for the target genes.

Blank page

Pingpong cycle is present in *Parasteatoda tepidariorum* embryo

4. Pingpong cycle is present in *Parasteatoda tepidariorum* embryo

4.1. Abstract

Piwi-associated RNAs (piRNAs) are single strand RNA molecules ranging between ~24-32 nucleotides, specific to animals. piRNAs are inmensely diverse molecules and are expressed in hundreds of thousands. The first studies of piRNAs come from *Drosophila*, in which they act to silence the expression of transposable elements (TEs). It is thought that this mechanism is vital for the protection of genome integrity. *Parastetoda tepidariorum* is a common house spider that has been emerging as a model organism owing to its short life cycle and ease of manipulation in the lab. *P. tepidariorum* and *D. melanogaster* shared a common ancestor ~ 550 mya. *P. tepidarioum* has short germband segmentation. We annotate 103,123 TEs sequences in *P. tepidariorum* genome using RepeatMasker, which represent 0.96% of the genome. We identify that the total piRNAs exhibit uracil bias at the first nucleotide position and the small proportion of piRNA exhibit the adenine at the tenth nucleotide position. We identified abundant piRNAs. The ping-pong signature was present in piRNA transcribed from TE and protein sequences.

4.2. Introduction

Transposable elements (TE) are repetitive DNA sequences that have the ability of movement and copying from one region to another within the genome (McGurk and Barbash 2018). TEs were first discovered in maize by Barbara McClintock (McClintock 1950). TE sequences are classified as retrotransposon and DNA transposon, based on the mechanism of transposition. The retrotransposon class can be divided into two subgroups according to the mechanism of chromosomal integration - long terminal repeat (LTR) retrotransposon and non-long terminal repeat retrotransposon (non-LTR). Non-LTR retrotransposon can be further sub-divided into short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs) (Bourque et al. 2018). Significant proportions of large animal genomes are made up of TEs. The commonly used TE and repeat finding tool, RepeatMasker, masks 56% of the human genome as interspersed repeats and low complexity regions (http://www.repeatmasker.org/). ~20% of the *Drosophila melanogaster* genome and 20% of the *Arabidopsis thaliana* genome are similarly masked (Adrion et al. 2017) (Legrand et al. 2019). Within arthropods, the proportion of the genome made up of TEs varies significantly, from less than 6% in Belgica Antarctica (*Antartic midge)* to more than 58% in the malaria mosquito (*Anophales gambiae*) (Petersen et al. 2019).

Insertions of transposons in different genomic regions can disrupt gene function. Therefore, a mechanism to protect the genome integrity is vital, in particular in the germline because any changes will be passed down to the progeny (Siomi et al. 2011a). piRNAs act to silence TE expression at the transcriptional and post-transcriptional level. piRNAs are single-stranded, non-coding RNAs, ranging in size from 24 to 32 nt. These short non-coding RNAs are associated with PIWI proteins, including Argonaut proteins, piwi, Aubergine (aub) and AGO3. piRNAs are sequence complex – they do not appear to have any enriched sequence motifs, only a bias for uridine at the first position. piRNAs are derived from TE sequences, from 3'UTR of mRNA and intergenic long non-coding RNAs (Han and Zamore 2014b). There are thousands to hundreds of thousands of different piRNAs in each animal genome (Siomi et al. 2011b) (Siomi et al. 2011a).

Most studies have shown that piRNAs play a role in silencing transposons. However, in the last few years other functions of piRNAs have been discovered. For example, in the silk moth (*Bombyx mori*) specific piRNAs cause the sex determination, in *C. elegans*(Kiuchi et al. 2014), a piRNA iniciate transgenerational memory (Sarkar, Volff, and Vaury 2017). Hence the importance to study piRNA in other organism, to discover new functions of piRNA.

P. tepidariorum, an arachnid with a whole genome duplication event (Schwager et al. 2017), is an emerging model organism, used for its short life cycle and easy maintenance in the laboratory. *P.tepidariorum* and *D.melanogaster* shared a common ancestor ~ 550mya (Misof et al. 2014). *P. tepidarioum* has a short germ band segmentation mode of development, which is an ancestral mode in arthropods, different to the most well know dipteran model organism, *D. melanogaster* that has long germ band (Paese et al. 2018). The genome assembly of *P. tepidariorum* consists of 1443.9 Mb (Schwager et al. 2017), and small RNA transcriptome has been sequenced (Leite et al. 2016). It was identified that piwi is expressed in early embryogenesis in *P. tepidariorium*(Schwager, Meng, and Extavour 2015). TEs have been annotated, but in an older version of the *P. tepidariorum* genome and both somatic and germline piRNA were identified in *P. tepidariorum* (S. H. Lewis et al. 2018).

In this chapter, we describe an annotation of transposable elements in the *P. tepidariorum* genome dovetail version and present the first investigation of the expression of piRNAs and transposon in *P. tepidariorum* embryogenesis, covering early cleavage until brain formation. We also report for the first time the ping-pong signature in piRNAs transcribed from TE and protein sequences in the *P. tepidariorum* embryogenesis.

4.3. Methods

4.3.1. Annotation of transposable elements in the *P. tepidariorum* genome

The *P. tepidariorum* genome sequence dovetail version was provided by the McGregor lab (http://mcgregor-evo-devo-lab.net/McGregor_lab/home.html). The genome used is an updated version from the assembly 2.0 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000365465.2), consisting of 16,542 scaffolds.

TEs were annotated using three approaches. For the first annotation, *P. tepidariorum* transposable elements were annotated using the TEAnnotator pipeline (https://github.com/SamuelHLewis/TEAnnotator) (S. H. Lewis et al. 2018). The TEAnnotator pipeline produces two annotations and merges them into a single non-redundant one. The first annotation uses RepeatMasker (http://www.repeatmasker.org) to identify homologous sequences of transposable elements from the metazoan library (https://www.dfam.org/home). The second annotation is produced using RepeatModeler (http://www.repeatmasker.org/RepeatModeler/). RepeatModeler uses the *P.tepidariorum*

genome as input, identifies repeat elements and builds consensus models of predicted novel TE sequences. These potential novel TE sequences are then used as a library of transposable elements to run RepeatMasker against the genome. Then the two annotations in gff are merged, and sequences less than 100 nucleotides long were filtered out. For the second method, *P. tepidariorum* TEs were annotated using RepeatMasker and the metazoan library of transposable elements from Dfam_consensus 3.0 (https://dfam.org/home). The third approach used RepeatMasker and the transposable elements library from *Acanthoscurria geniculata*, one of the best annotated arachnids and has the TEs sequences available, the sequences were download from ArTEdb (http://artedb.net/).

4.3.2. Small RNA sequence analysis

P. tepidariorum small RNA deep sequencing data were retrieved from the Sequence Read Archive, accession PRJEB13119 (https://www.ncbi.nlm.nih.gov/bioproject/PRJEB13119). The ten available developmental time points correspond to the first ten stages of the *P. tepidariorum* embryogenesis: 5 h, 13 h, 21.5 h, 29 h, 35.5 h, 45.5 h, 53 h, 65.5 h, 80.5 h, 90.5 h. The RNA for each developmental point was extracted from multiple embryos contained in a single cocoon. Sequencing was performed by Illumina HiSeq 2000. 3' adaptor sequences were removed, and 24-32 nucleotides were selected using Cutadapt program (https://cutadapt.readthedocs.io/en/stable/), we followed a standard pipeline from a Siomi lab, in our piRNA size selection. The reads were collapsed using fastx-toolkit from Hannon Lab (http://hannonlab.cshl.edu/fastx_toolkit/).

We retrieved 3783 Arachnida ribosomal RNA sequences from (https://www.arb-silva.de/browser/) and 37 Araneae ribosomal RNA sequences from NCBI (https://www.ncbi.nlm.nih.gov/nuccore/?term=ribosomal+RNA+araneae). Reads mapping to tRNAs and rRNAs, using Bowtie 1.1.1 allowing three mismatches, were removed from all small RNA datasets. tRNAs were annotated using tRNAscan-SE-2.0 (http://lowelab.ucsc.edu/tRNAscan-SE/).

The cleaned reads then were mapped using Bowtie 1.1.1 (Langmead et al. 2009) to TEs sequences of *P. tepidariorum*. The ssviz R package was used to calculate counts and size distribution of mapped reads and analyse ping-pong signature Low D (2021), in order to use the ssviz package, this needs the usage of fastx_collapser option of FASTX-toolkit.

Ping-pong analysis was tested in piRNAs that mapped TE sequences and piRNAs that mapped in protein sequences. It is known that in flies and mice piRNAs are transcribed from TE sequences, from 3`UTR of mRNA and lncRNA genes (Han and Zamore 2014a). Currently, we only have the annotation of TE sequences and protein sequences in *P. tepidariorum*, so we made the ping-pong analysis for these sequences. The protein annotation of *P. tepidariorum* was provided by McGregor Lab (http://mcgregor-evo-devo-lab.net/McGregor_lab/home.html)

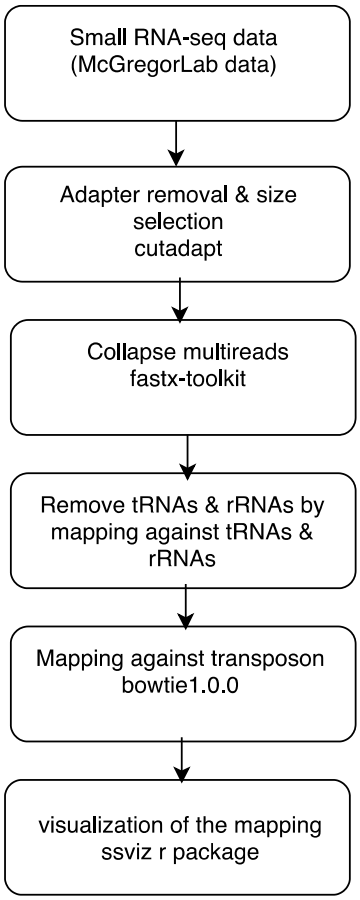

Figure 4.1. Workflow of the piRNA analysis

4.3.3. piRNA expression in *P. tepidariorum* embryo

P. tepidariorum small RNAseq files were download from European Nucleotide Archive (ENA) with PRJEB13119 accession number (https://www.ebi.ac.uk/ena/browser/view/PRJEB13119?show=reads). The 3' adapter were removed and reads from 24-32 nucleotides were selected using Cutadapt program (https://cutadapt.readthedocs.io/en/stable/). The ribosomal RNA was removed from reads, the rRNA sequences were obtained from the ribosomal RNA database (https://www.arb-silva.de/). The cleaned reads were mapped with bowtie against the *P. tepidariorum* dovetail genome

version. The mapping was using Bowtie 1.1.1 (Langmead et al. 2009), allowing one mismatch and to choose unique mapping reads, multi-mapped reads were excluded from the analysis. We use featureCounts v2.0.3 to count the reads and use the transposon annotation. DESeq2 v.1.30.1 (Love, Huber, and Anders 2014) was used for the normalization and an adaptation of the Turner`s script for the plots (https://gist.github.com/stephenturner/f60c1934405c127f09a6)

We first run the DESeq2 with condition time course and after evaluating the PCA plots we observed in PCA plots results by, -M, -Mf. -MfO. We observed that the RNA seq st6r2_1, st6r2_2 was an outlier and was located very far from its replicate st6r1, we exclude it from the analysis,

4.3.4. mRNA sequences analysis

The experiment was made for a research group in Japan. Embryos were obtained from two different pairs of parents independently (two biological replicates experiment) from ten developmental stage points, the time was measured after egg laying: stage1 16 nucleus, 9hr; stage2 after 256 nucleus, 15hr; stage3 forming germ disc, 20-21 hr; stage4 the germ disc with the white spot at the centre, 25-26 hr; stage5 early the cumulus appeared at the centre of the germ disc, 31-32 hr; stage5 late the cumulus moved to the rim of the germ disc, 36-37 hr; stage6 the cumulus disappeared, 42-43 hr; stage7 segmentation, 50 hr; stage8 the germ band is formed, 60 hr; stage10 the limb segments appeared 78 hr (https://www.e-celldev.jp/pt_spider2/search_stexpress_ew.php).

The embryos were maintained at 25˚C, it was taken between 10-100 embryos for each developmental time point. RNA extraction and fragmentation were made by Ambion and New England Biolabs kits respectively. NEBNext Ultra Directional RNA Library Prep Kit for Illumina and NEBNext Multiplex Oligos for Illumina were used for the library preparation, the libraries were strand specific. Illumina MiSeq was used for the sequencing. The experiment had two biological replicates for each time point (10) and two MiSeq run (Iwasaki-Yokozawa, Akiyama-Oda, and Oda 2018).

P. tepidariorum mRNA seq sequencing data were retrieved from European Nucleotide Archive (ENA) web page with PRJNA448775 accession number (https://www.ebi.ac.uk/ena/browser/view/PRJNA448775?show=reads).

The adaptors and sequences shorter than 15bbp were removed with Trim Galore v. 0.6.4 (https://github.com/FelixKrueger/TrimGalore) running by default parameters.

The cleaned reads were mapped to *P. tepidariorum* genome dovetail version with HISAT2 v. 2.2.1 (Kim et al. 2019) with the default parameters. The mapped reads then were counted using featureCounts v2.0.3 (Liao, Smyth, and Shi 2014) considering strand specific, considering ID instead of gene_id, and CDS instead of exon in bam file (-s 1 -g ID -t CDS). Besides the default run presented here, the program was also run with multimapping, multimapping-fraction, and multimapping-fraction-overlapping parameters activated (data do not show).

The transposon annotation was generated in a previous analysis. The count read matrix data generated by featureCounts were used as input for DESeq2 v.1.30.1(Love, Huber, and Anders 2014) and adaptation of Turner`s script was used for the plots (https://gist.github.com/stephenturner/f60c1934405c127f09a6)

We first run the DESeq2 with condition time course and after evaluating the PCA plots we observed in PCA plots results by, -M, -Mf. -MfO. We observed that the RNA seq st6r2_1, st6r2_2 was an outlier and was located very far from its replicate st6r1, we exclude this biological replicate.

4.4. Result

4.4.1. Annotation of transposable elements in *P. tepidariorum*

The genome sequences from *P. tepidariroum* have been published, and microRNAs have annotated (Schwager et al. 2017), (Leite et al. 2016). To study the most abundant piRNAs, we decided to annotate the transposable elements in *P. tepidariorum* as the majority and best studied piRNAs come from TE sequences (Han and Zamore 2014a). There is a previous annotation of TE in *P. tepidariorum* but with an older version of the genome assembly (S. H. Lewis et al. 2018).

103,123 TEs sequences were annotated in the *P. tepidariorum* genome using RepeatMasker using the metazoan TEs library from Dfam 3.0. (Table 4.1). These TEs sequences represent 0.96% of the genome. 67,857 DNA transposons were the most abundant TE type representing 0.64% of the total TE identified, 29,393 hobo-activator and 20,339 Tc1-IS630-pogo transposon families. 22,142 retroelements were annotated and 10,228 LINEs and 8,778 LTR were the most abundant families 3,310 Gypsy/DIRS1 (Table 4.1).

118,377 TE were annotated using RepeatMasker with *Acanthoscurria geniculata* TE sequences as library, representing 1.35% of the *P. tepidariorum* genome. We obtained 192,127 TEs using TEAnnotator pipeline, this number is very similar obtained from Lewis results; 203,585 TEs that use the same pipeline.

Our results show that using annotation from RepeatMasker with transposable elements from metazoan and the *A.geniculata* libraries, we obtained a similar number of TEs 103,123 and 118,377 respectively. 1857 TE sequences were obtained using RepeatModeler for *A. geniculata* (7.2 Gb genome size) (Wu and Lu 2019).

Table 4.1 Main families of transposable elements annotated using RepeatMasker for *P.tepidariorum* usig metazoa Dfam as TE library

Transposable elements	Number of elements	Length occupied (bp)	% of genome
Retroelements	22142	3176856 bp	0.22%
SINEs	3136	190037 bp	0.01%
Penelope	1561	209762 bp	0.01%
LINEs:	10228	1517744 bp	0.11%
L2/CR1/Rex	408	28080 bp	0.00%
R1/LOA/Jockey	1710	234393 bp	0.02%
R2/R4/NeSL	9	637 bp	0.00%
RTE/Bov-B	3028	673875 bp	0.05%
L1/CIN4	1443	104695 bp	0.01%
LTR elements:	8778	1469075 bp	0.10%
BEL/Pao	2059	403684 bp	0.03%
Ty1/Copia	180	48957 bp	0.00%
Gypsy/DIRS1	3310	761331 bp	0.05%
Retroviral	2710	217500 bp	0.02%
DNA transposons	67857	9196538 bp	0.64%
hobo-Activator	29393	3031225 bp	0.21%
Tc1-IS630-Pogo	20339	4606084 bp	0.32%
PiggyBac	1631	186169 bp	0.01%
Tourist/Harbinger	1631	86502 bp	0.01%
Mirage, P-element, Transib	293	20026 bp	0.00%
Rolling-circles	1523	177023 bp	0.01%

Unclassified:	6983	1318582 bp	0.09%
Total interspersed repeats:		13691976 bp	0.95%

4.4.2. Identification and abundant piRNAs in *Parasteatoda tepidariorum* embryos

Our ten embryonic samples from *P. tepidariorum* represent the first 95 hours of development: early cleavage (5 h), blastoderm (13 h), germ disc formation (21.5 h), primary thickening (29 h), cumulus migration (35.5 h), dorsal field (45.5 h), germ band (53 h), prosomal limb buds (65.5 h), limb differentiation (80.5 h), brain differentiation (90.5 h).

In order to identify the piRNAs from small RNA seq, we followed an established pipeline that select piRNAs size from 24-32 nt long. The reads were collapsed using fastx-toolkit. Then the cleaned reads were mapped using Bowtie 1.1.0 to the TEs sequences from *P. tepidariorum* with the parameter -v1, which allows one mismatch in the alignment and default for the other parameters. We obtained reads that uniquely mapped and multi-mapped reads. These reads mapped were considered to the ping-pong analysis using ssviz package in R.

In order to confirm the presence of piRNAs in our small RNA libraries, we searched for certain features in the sequences. First, piRNAs that interact piwi, aub, ago3 proteins have a bias for uracil at the first nucleotide position and adenine bias at the tenth nucleotide position (Figure 4.2). Second, piRNAs range in size from 24 to 32 nucleotides (Figure 4.3). Third, piRNAs that are derived piRNA cluster are mainly antisense to the TE sequence (Tóth et al. 2016) (Figure 4.4).

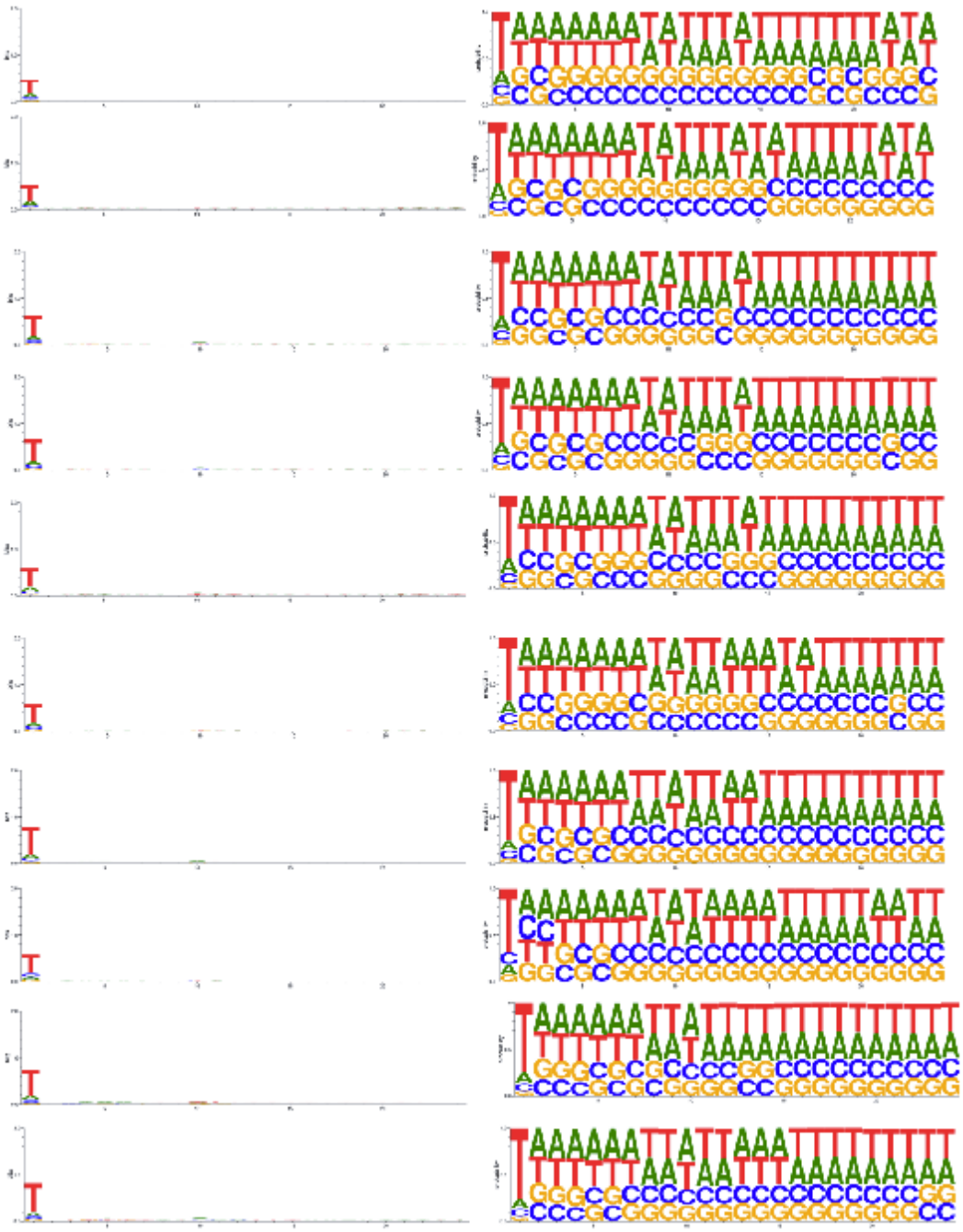

Figure 4.2. Nucleotide sequence bias. The left column shows the bias in bits and the right column the bias in probability, from top to the bottom the ten developmental points: 5 h, 13 h, 21.5 h, 29 h, 35.5 h, 45.5 h, 53 h, 65.5 h, 80.5 h, 90.5 h.

We detected enrichment for uracil at the first position in all ten stages Stage 9 (80.5 h) and stage 10 (90.5h) reported the highest uracil bias. In addition, the bias for adenine at the tenth position nucleotide also was identified in few cases (Figure 4.2).

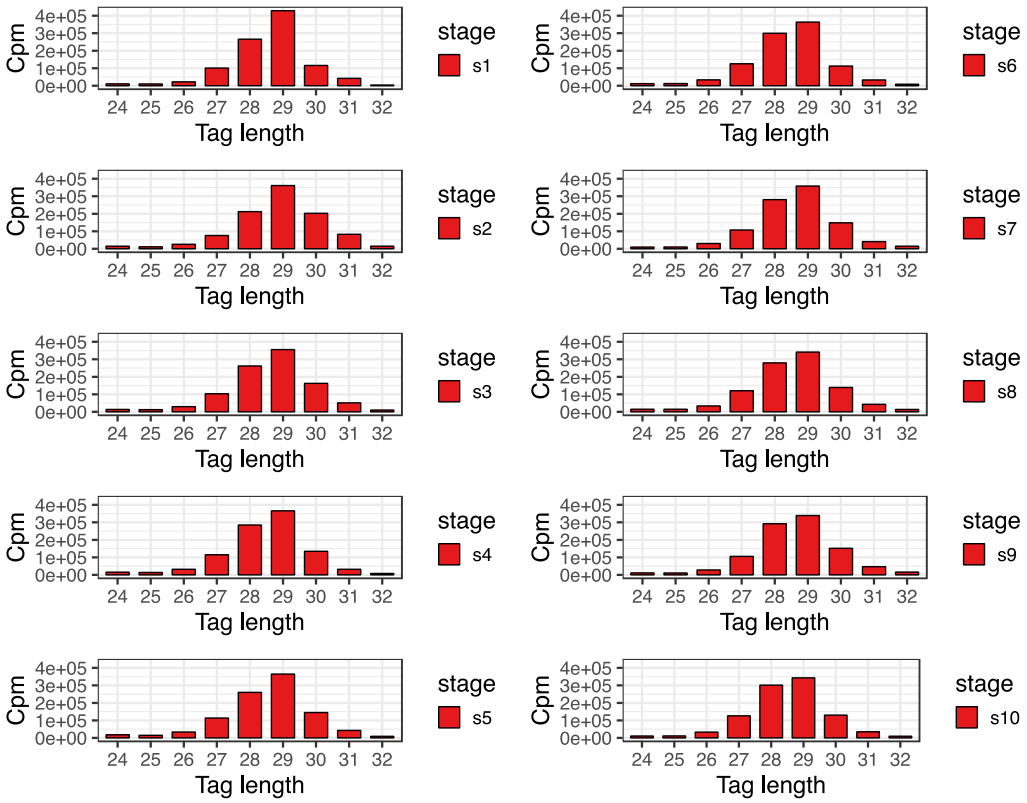

Figure 4.3. Size distribution of mapped reads to transposable elements sequences. Stage 1 (5 h), stage 2(13 h), stage 3 (21.5 h), stage 4 (29 h), stage 5 (35.5 h), stage 6 (45.5 h), stage 7 (53 h), stage 8 (65.5 h), stage 9 (80.5 h), stage 10 (90.5 h) from *P. tepidariorum.*

Another feature is that piRNAs are single RNA strand of 24-32nt. We mapped the cleaned reads to TE sequences of *P. tepidariorum* and then calculated the size distribution from the bam files using ssviz package in R. In figure 4.3 we find that the most abundant reads that map to TEs sequences are between 26 and 31 nt long.

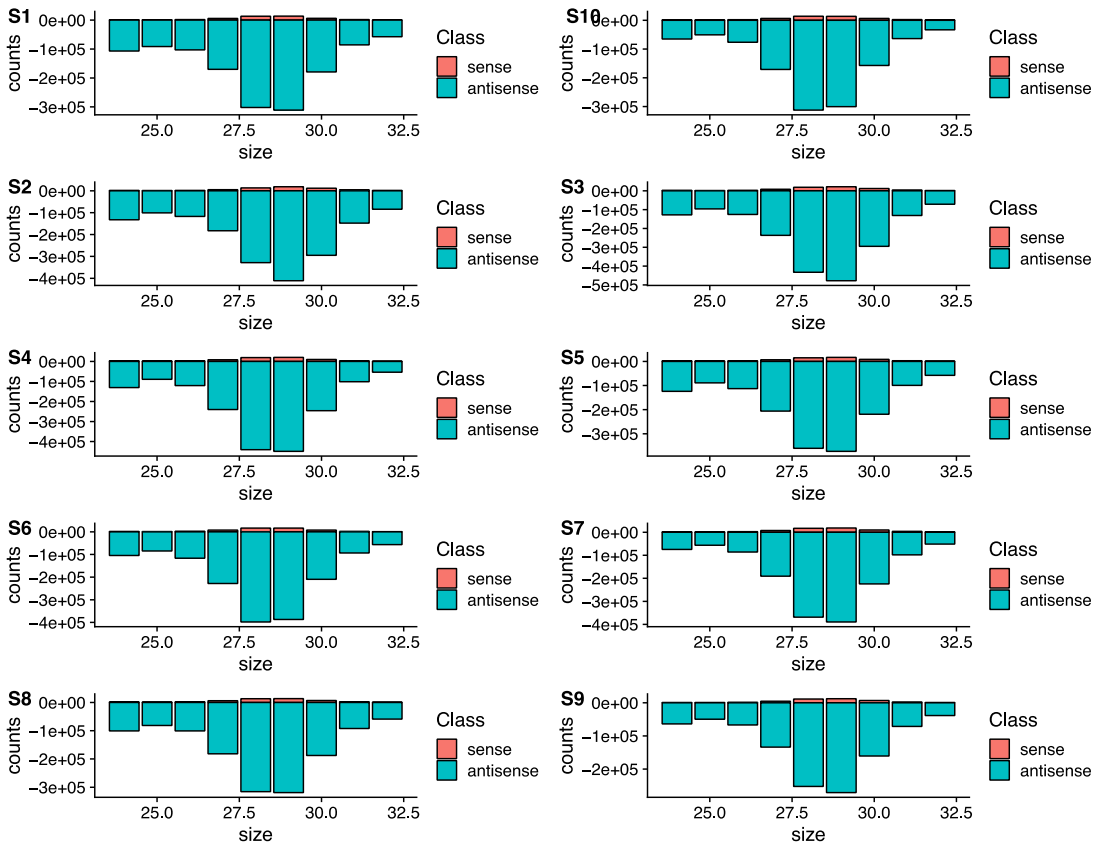

Figure 4.4. Sense and antisense reads mapped to the transposable elements of *P.tepidariorum*. Stage 1 (5 h), stage 2 (13 h), stage 3 (21.5 h), stage 4 (29 h), stage 5 (35.5 h), stage 6 (45.5 h), stage 7 (53 h), stage 8 (65.5 h), stage 9 (80.5 h), stage 10 (90.5 h) from *P. tepidariorum*.

To obtain Figure 4.4, we extracted sense and antisense reads mapped to the TE sequences using samtools (http://www.htslib.org/) and we plotted. We observed that the reads mapped to the transposable elements are in majority anti sense strand.

The pingpong cycle is a defense mechanism, post-transcriptional level at the cytoplasm by which first the piRNAs in complex with aub binds to transposon and cleavage it, then this new DNA fragment binds to ago3 to align to a piRNA cluster and cleavage this piRNA, creating another DNA fragment which enters again to the new cycle. This cycle generates the secondary piRNAs in an effective to silence transposon.

Our results show that the ping-pong cycle was present in all ten stages of *P. tepidariorum* embryogenesis for piRNAs transcribed from TEs (Figure 4.5). Since early cleavage until the brain formation. We observed that since first stage, early cleavage the piRNAs are being produced in the embryo actively, before of the zygote genome activation (end of stage2). We also identified at stage 5 (35.5hours) cumulus migration the highest production of piRNA followed by stage 9 (80.5h) limb differentiation.

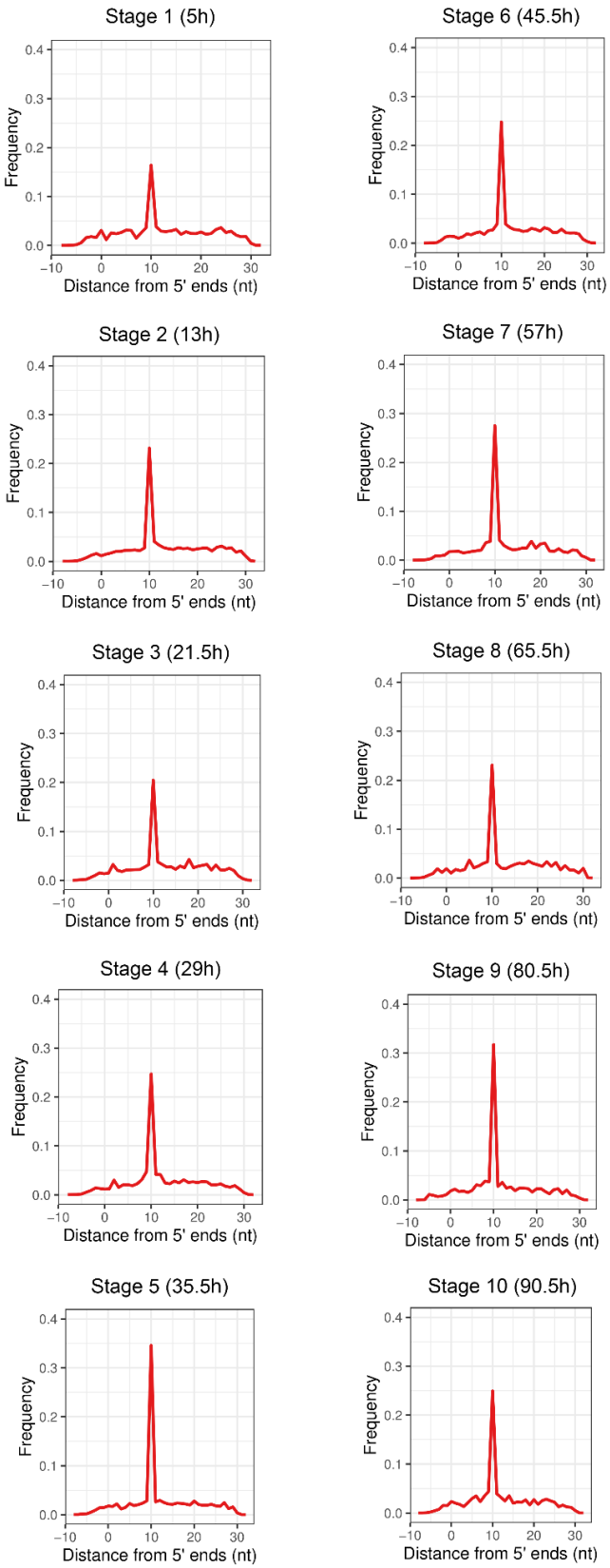

Figure 4.5. Ping-pong signature in the ten first stages of *P. tepidariorum* embryogenesis. piRNAs transcribed from TEs.

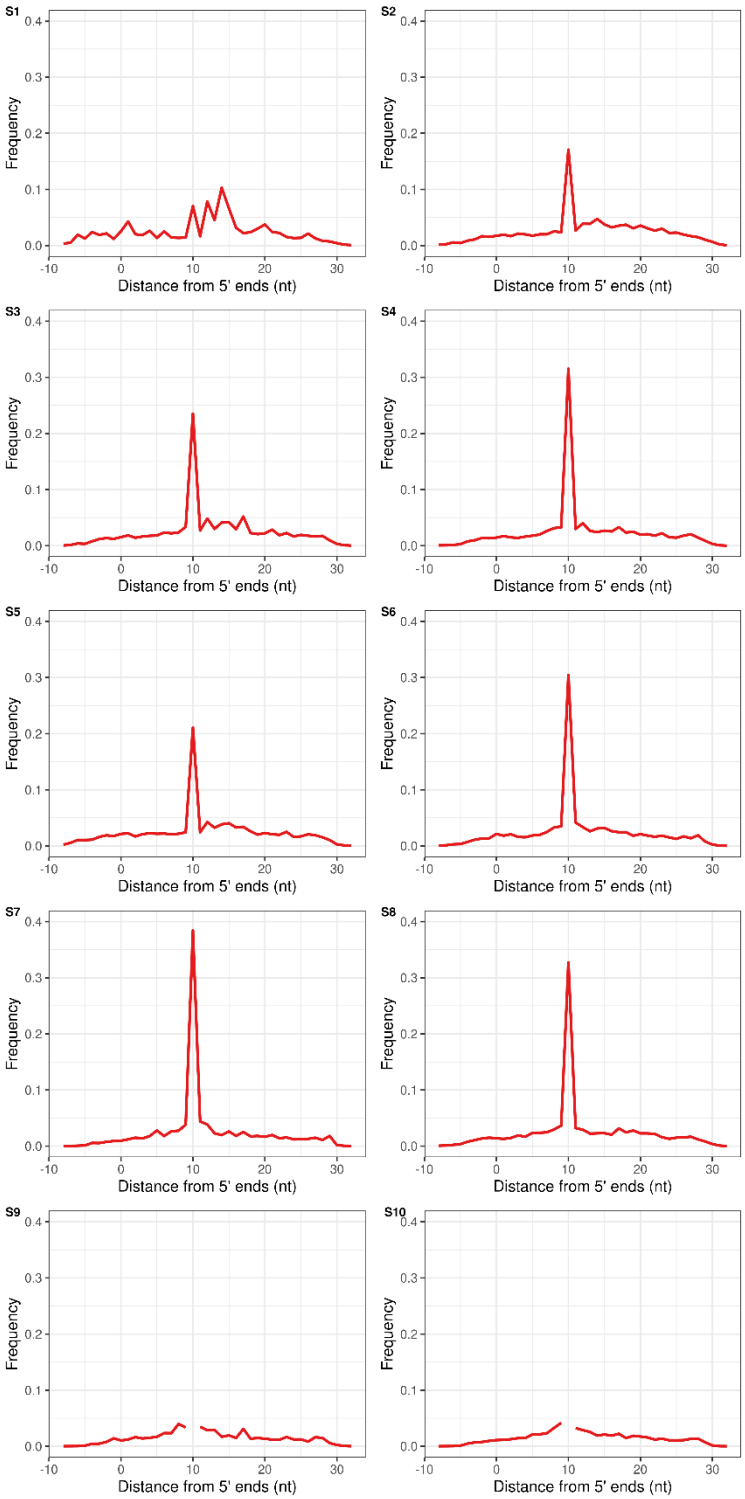

Figure 4.6. Ping-pong signature in the ten first stages of *P. tepidariorum* embryogenesis. piRNAs transcribed from protein sequences.

Ping-pong signature was present in S2, S3, S4, S5, S6, S7, and S8 stages of *P. tepidariorum* embryogenesis from piRNA transcribed from protein sequences (3`UTR) (Figure 4.6)

4.4.3. piRNA expression in *P. tepidariorum* embryo

To study the piRNA expression, we analysed small RNA sequencing from ten time point of embryogenesis of *P. tepidariorum*. We removed the adapter and the ribosomal RNA from the reads before the mapping. Our results show that close to 50% of the reads mapped were mapping in unique loci from small RNAs against the *P. tepidariorun* genome.

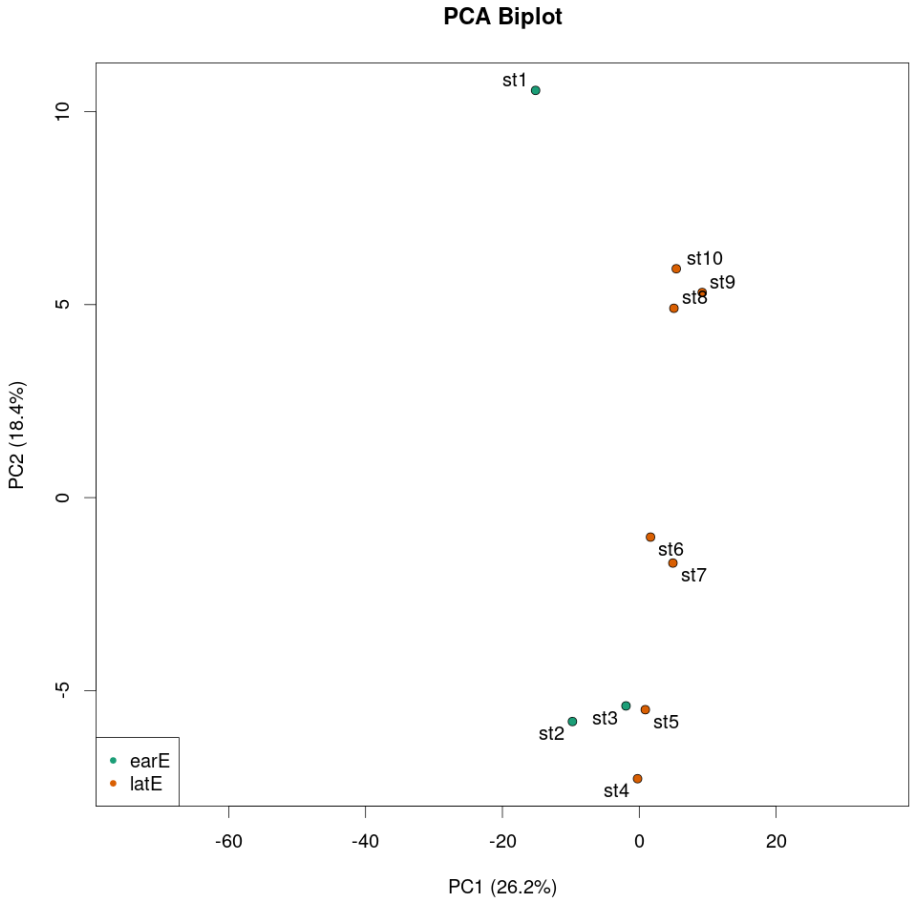

Figure 4.7. Principal component analysis of the ten developmental time points from *P.tepidariorum* embryogenesis from piRNAs.

PC1 represent 26% of variance and PC2 18%, together are 44% of the total variance. The ten stages do not show an early and late pattern expression. They form some small groups, such st2 and st3, another group of st4 and st5, another group st6 and st7 and the last group of st8, st9 and st10; st1 are not grouped. These results could be true, and the microRNAs expression can change during the stages. However, for this analysis additional replicates can add strength to the analysis.

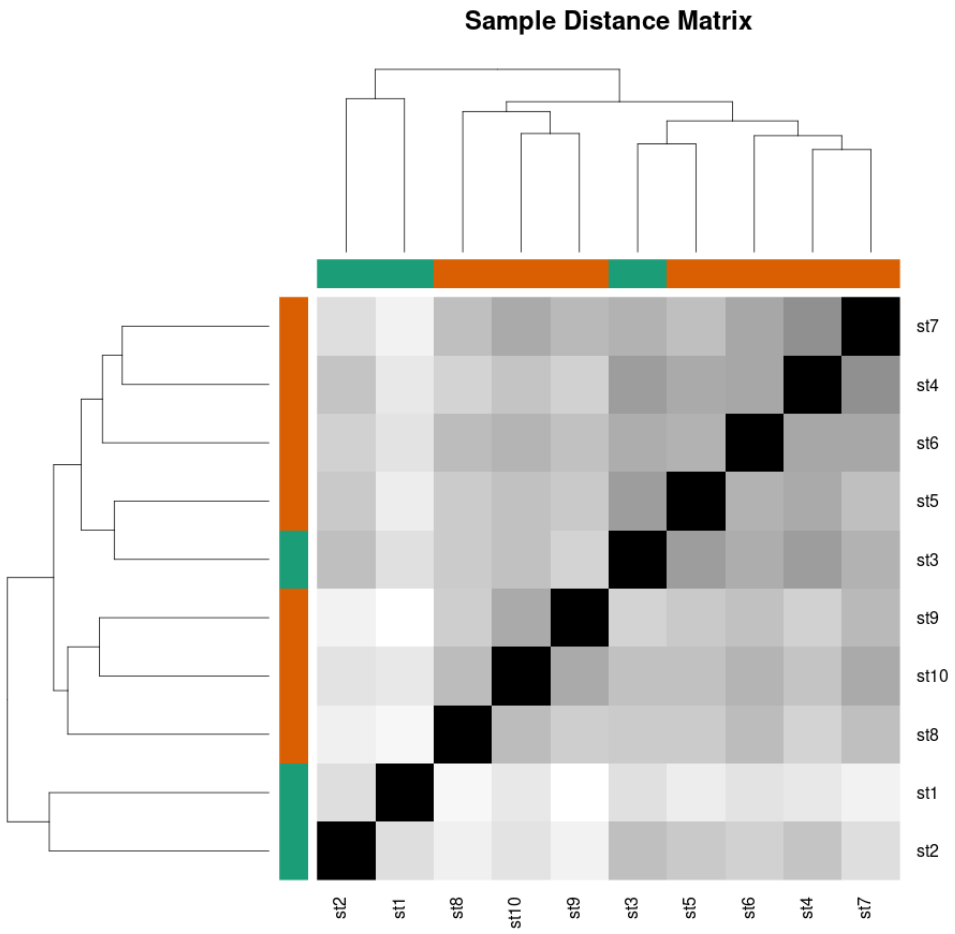

Figure 4.8. Hierarchical clustering of the ten developmental time points of embryogenesis from *P.tepidariorum*

Table 4.2. Top regulated genes from the piRNA expression of ten developmental time point of embryogenesis in *P. tepidariorum*

Gene	log2FoldChange	padj
Halyomorpha_halys-5_family-9382_21_1152	-4.023732993	0.000447

Our differential expression analysis can allow us to identify a piRNAs that are transcribed from the Halyomorpha_halys-5_family-9382 transposon (Table 4.2). As our approach to consider early and later piRNA expression pattern was wrong. There is still a time-series

strategy that can allows us to discover piRNAs upregulated or downregulated in the embryo of *P. tepidariorum*.

4.4.4. Transposon expression in *P tepidariorum* embryo

DESeq2 was used for the normalization and analysis of the differential expression in transposon from ten developmental time points of *P.tepidariorum.* The pattern of the gene expression from RNA samples was similar for all (-M, -Mf. -MfO) featureCounts parameters, we only observe that RNA-seq sample from sta6r2(st6r2_1, st6r2_2) was very far from the group, so we eliminated the st6r2 for the analysis of the gene expression.

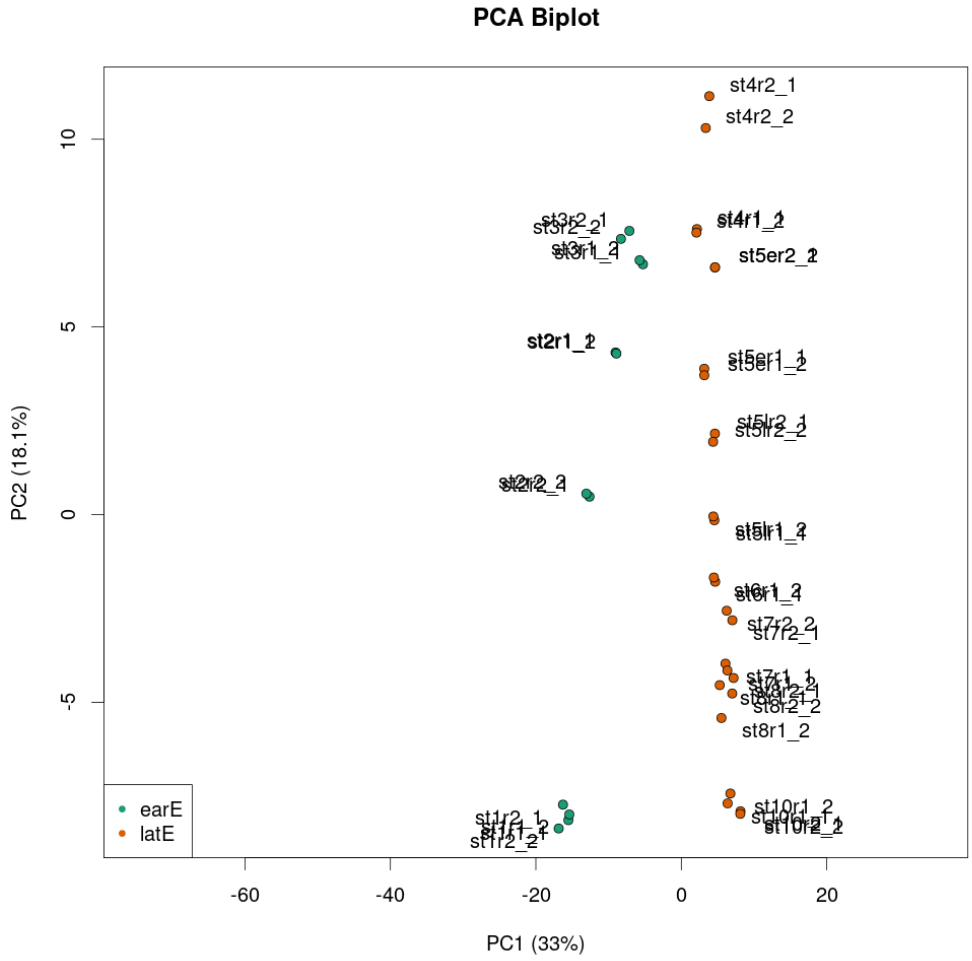

Figure 4.9. Principal component analysis of the ten developmental time points of embryogenesis from *P.tepidariorum*

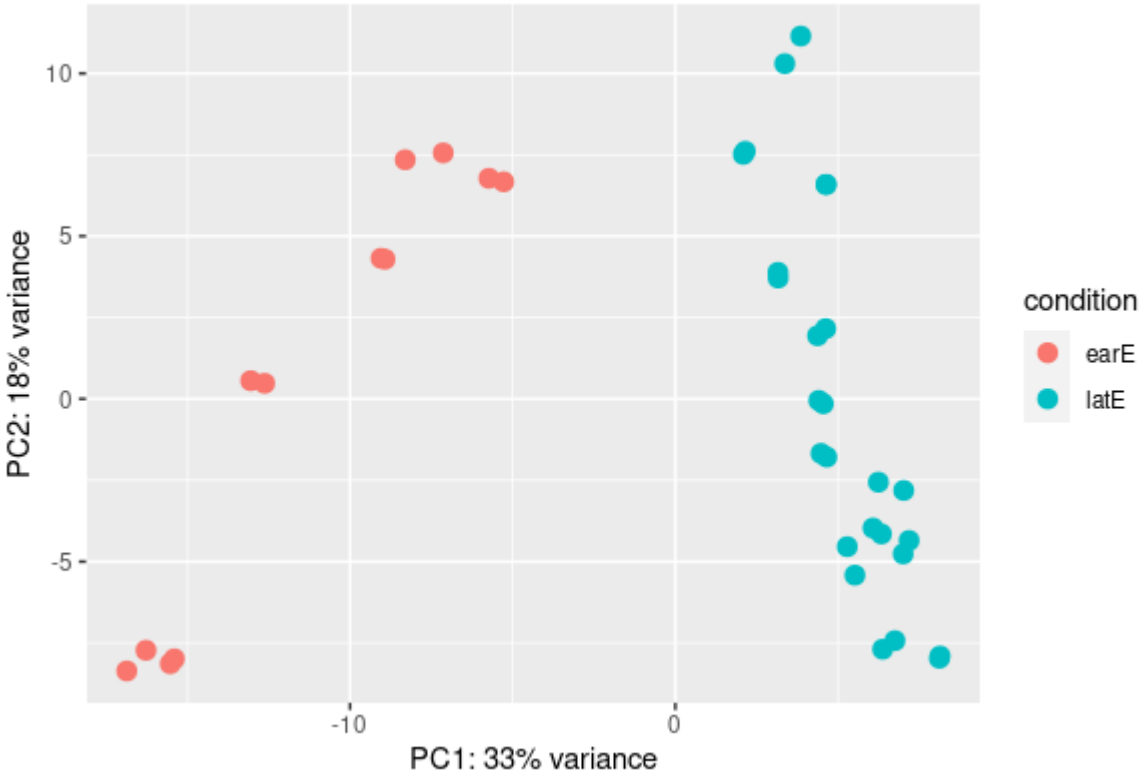

Figure 4.10. Principal component analysis of the ten developmental time points of embryogenesis from *P.tepidariorum*

The PC1 explain 33% of the variance and PC2 18% of the variance, together represent a total of 51% of the total variance. PCA result shows two main groups, we called early embryogenesis (earE) and late embryogenesis (latE). The earE group contain st1, st2, st3 and the latE group st4, st5e, st5l, st6, st7, st8, st10. In earE group, st1 is distant from st2 and st3, a slightly difference in gene expression. In latE, there is a slightly difference in gene expression among the st4, st5e, st5l, st6, st7, st8, and st10. In many cases de biological replicates are grouped together. Here also we observe a big shift from st3 and st4 as we observe in protein expression analysis section (Figure 4.9, Figure 4.10).

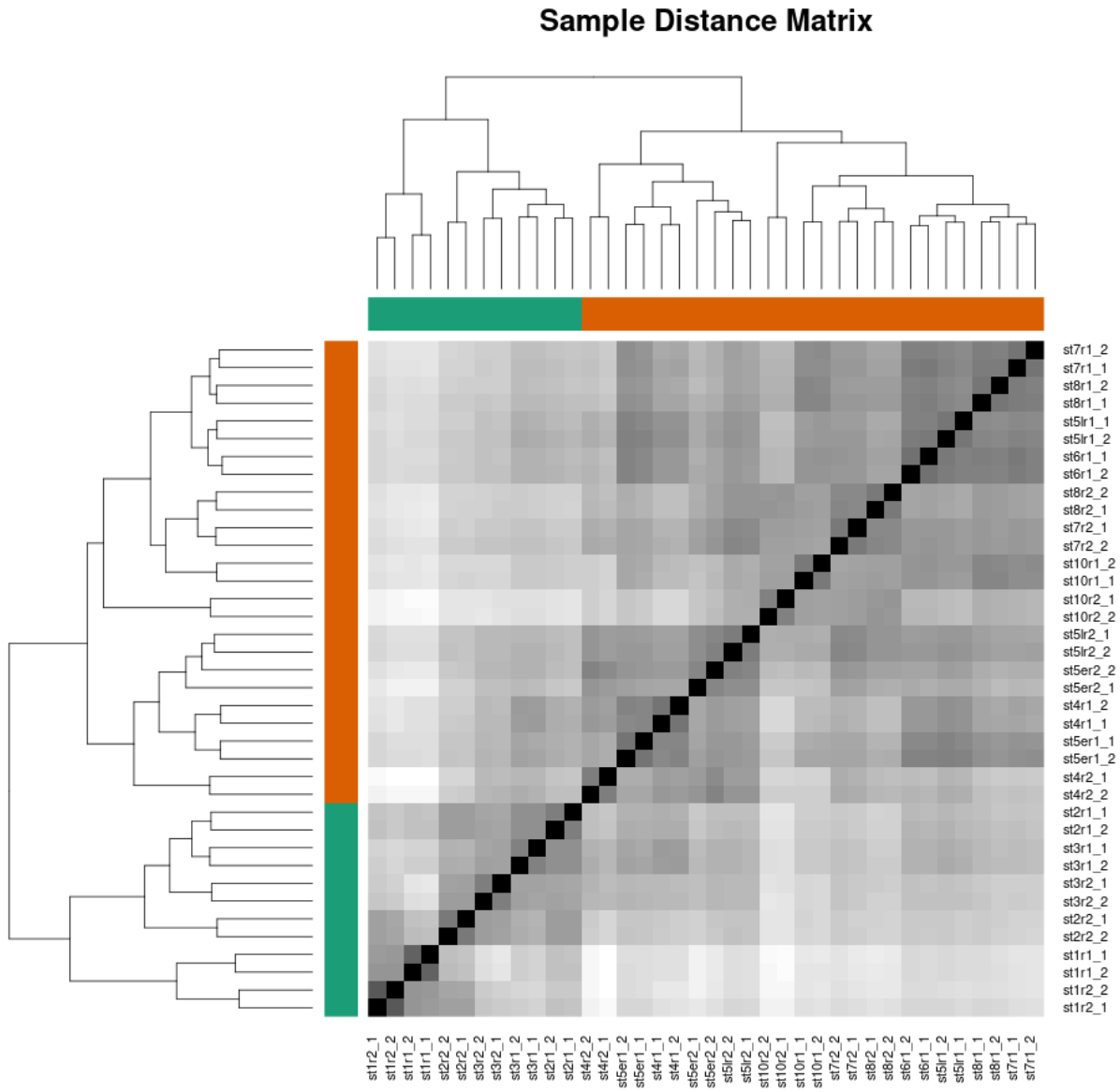

Figure 4.11. hierarchical clustering of the ten developmental time points from *P.tepidariorum for* transposon genes

We can observe in the hierarchical clustering, there is a clustering of the earE group as well the lateE groups (Figure 4.11).

Table 4.3. Top differential regulated transposon genes of ten developmental time point of embryogenesis in *P. tepidariorum*

Gene	log2FoldChange	padj
TE-X-12_DR_637_988	-4.218243884	2.03E-131
EnSpm-6N1_DR_938_1080	-7.188555142	5.07E-59
TE-X-12_DR_94_993	-1.947423746	2.63E-44
TE-X-12_DR_346_642	-1.92925111	9.91E-36
MamTip1_162_267	2.962410608	4.77E-32
TE-X-12_DR_95_989	-1.858618559	8.77E-30
Halyomorpha_halys-5_family-3063_126_1211	1.930386332	4.15E-28
HSMAR1_216_1054	2.787576885	2.13E-27
Halyomorpha_halys-1_family-399_239_360	-1.829122827	1.43E-25
Halyomorpha_halys-5_family-257_394_695	-5.171110619	1.43E-25
fAlbLTRK12b_1338_1384	2.178868058	4.57E-25
Halyomorpha_halys-6_family-945_34_585	-1.771206618	1.68E-23
Halyomorpha_halys-6_family-1682_670_798	-2.212335554	3.74E-23
Halyomorpha_halys-6_family-1682_173_451	2.133549207	1.14E-21
Charlie29b_721_998	-4.890758968	2.15E-21
hAT-N131_DR_221_270	2.273491447	1.05E-20
RTE-2_Testu_1393_2221	2.902749354	5.97E-20
TE-X-12_DR_355_978	-1.550886564	1.18E-19
TE-X-12_DR_284_894	1.887275847	2.04E-19

We identified the top twenty regulated transposon genes based on p-value adjustment from the ten developmental time points of embryogenesis in *P. tepidariorum* (Table 4.3).

4.5. Discussion

4.5.1. TE annotation

Our annotation using RepeatMasker with the metazoan library from Dfam is similar to the one obtained using the *A. geniculata* TE library, with 103,123 and 118,377 respectively. Using TEAnnotator (RepeatMasker and RepeatModeler together) on the metazoa TE library together allowed us to further discover novel TE sequences 192,127 total TEs sequences. This annotation of our TEs sequences was similar to the TEannotator results from Lewis 203,585 TEs annotation with an older version of *P. tepidariorum* genome (S. H. Lewis et al. 2018), our similar result is a good indication of replicability of this method .

Overall, we find considerably lower proportions of the genome are TEs relative to other arthropods. Considering the annotation by RepeatMasker, we find fewer than 1% of the sequences are TEs. In contrast, the Belgica Antarctica (*Antartic midge)* has 6% of its genome TEs and the malaria mosquito (*Anopheles gambiae)* has over half of its genome made up of TEs (54%.) (Petersen et al. 2019).

It is well known that genome assembly is challenging when there are insertions into TE sequences and high copy number of TE, it could be the current genome coverage of the *P. tepidariorum* is lacking great percentage of the TE sequences, that could be the reason we do not identify them. This could be addressed using long-read sequencing to improve the genome coverage of *P. tepidariorum*.

Two of our annotations are based on number of sequences of the TEs in the database, Dfam3 and *A. geniculata* TE library. These each annotation depends on the number of TE sequences in the database to identify in *P. tepidariorum* genome. TEAnnotator pipeline annotation, furthermore, to use the homology identification that use RepeatMasker, use RepeatModeler which allows us to identify the novo TE sequences in the genome.

To identify additional TEs, we also can use all sequences from the Arthropod Transposable Elements Database ((http://artedb.net) in the annotation, we only have used the *A. geniculata* library.

Improvement in the identification of the TE sequences in eucaryotes was implemented in an Extensive de-novo TE Annotator (EDTA) pipeline for the TE annotation (Ou et al. 2019) , it is a pipeline that implement de-novo TE libraries then later it can be used for the TE annotation. In addition, the authors suggested to add manual curated TE libraries. They considered the following parameters: sensitivity, precision, accuracy and specificity false discovery rate, and f1 parameters to evaluate the annotation. They found RepeatModeler has the best annotation related to the identification the repeat sequences. The best results in LTR transposons identification were achieved using LTRharverst, LRT_retriever, and LRT_FINDER, the last one implemented by them to run in parallel to get faster results. RepeatModeler for the non-LRT sequences, TIR_learner showed a better result for the TIR transposons. They also identified when there are miss classification of the TE elements can cause a false discovery for the others TE elements, so it is very important to do a curation of the libraries. In summary, The EDTA showed a good result in plants and animals, with different size genome (Ou et al. 2019).

4.5.2. miRNA identification in the *P. tepidariorum* embryo

The mapping programs use in the small RNA seq analysis do not fully solve the problem of precision and sensitivity which are related to randomly choosing one best alignment and not considering multi-mapped reads respectively (Johnson et al. 2016).

As it was mentioned in the results, we used bowtie -v1, allowing one mismatch, the other parameters were run by default, by default k=1, k report the first alignment find per read, this mean for our mapping it was reported one hit per read. In summary, considering the parameters we use for the run bowtie, we got reads that map uniquely and reads that map in multi region/loci of the genome (in this case, TE sequences), bowtie in case of multi-mapping reads select exactly one hit arbitrarily. The disadvantage of choosing only -v 1 and k 1, and other parameters by default, it is that the hits chosen in the multi mapping reads not necessarily are the best alignments. In order to address if the mapping results considering uniquely and multi-mapping reads are similar to reads that only map uniquely, we decide to obtain the uniquely mapping reads and compare it, to draw a conclusion.

we run bowtie with the parameters to obtain the read that map uniquely (v1 -a –best -m1). It was obtained that for the first stage were 50% of the reads are unique and 50 % multimapped reads, for the rest of the nine stages were more than 50% of multi-mapped reads. It does not to change the conclusion that all of the are piRNAs that come from TE sequence.

4.5.3. Transposon expression in *P. tepidariorum* embryo

To consider only unique mapping reads or unique mapping reads and multimapping reads, both strategies are right since we look at the same loci and comparable data. For our transposon expression analysis, we decided to consider only unique mapping analysis, as it is more conservative and with obtention of almost 50 % of the unique mappers reads is a good (Deschamps-Francoeur, Simoneau, and Scott 2020).

An explanation of hight percentage of multi-mapped reads could be due to TE sequences are represented multiple times in the *P. tepidariorum* genome dovetail version.

4.6. Conclusion

We annotate TE sequences using three methods and we obtain similar results using RepeatMasker and a similar result that use Lewis et al., 2018 for the novo TEs annotation.

We found abundant piRNA in all the ten first stages of embryogenesis, cleavage until brain formation in *P. tepidariorum*. We identified abundant piRNAs, with the characteristic of the ping-pong mechanism of biogenesis, in each of the ten stages of *P. tepidariorum* embryogenesis.

General discussion

5. General discussion

5.1. Chapter 1: Do miRNAs preferentially regulate ohnologs genes in vertebrates?

Despite our hypothesis and previous proposition that microRNAs may regulate the imbalancein SSD events, we found that microRNA preferentially regulate ohnologs rather than SSD genes in mouse, rat and human.

We found that haploinsufficient genes are preferentially regulated by miRNAs in human. As future work we can get more information if we divide the data as WGD and SSD, if we groupthem in HI-WGD, HI-SSD, HS-WGD and HS-SSD.

We did not get any preferential regulation for essential genes using human cell culture however other studies found that WGDs were enriched with essential genes compared with SSD genes (Acharya and Ghosh 2016) (Makino, Hokamp, and Mclysaght 2008). To our knowledge, no similar study has been reported. Further work is required to test the impact of the relationship between essential genes and WGDs on the miRNA target density differences,as well at 3'UTR length distributions. Analyze WGD essential genes versus WGD non- essential genes.

The importance of the first chapter lies in that it is the first time we use microRNA to study the regulatory function in WGD and SSD genes as independent group. There is a previous work but this was made using duplicate genes as one sample, to investigate if microRNAs regulate preferentially duplicated genes rather than single copy genes in mouse and human (Li, Musso, and Zhang 2008)

5.2. Chapter 2: microRNA expression in *Parasteatoda tepidariorum*

MicroRNA profile results show an expression of microRNA at early and late stages in *P. tepidariorum*. Those highly expressed at the early stage correspond to the early cleavage and blastoderm, and the later stage corresponds to limb and brain differentiation. Showing the results in the line plot, we can see the presence of eight microRNA clusters. Some clusters arehighly expressed at early stages, other highly expressed at late stages, others still expressed atearly and late stage, and some others highly expressed at the

intermediate stages of embryogenesis. A bimodal pattern of expression is found in *C. elegans* and *D.melanogaster* embryos (Avital and Franc 2017). The bimodal model proposes that there are two main groups of microRNA expression at the embryogenesis level -1) few numbers of microRNAs with hierarchy that act at the beginning of the embryo and have core function, and 2) the majority of microRNAs expressed at later stages that regulate the cell differentiation. We observe that microRNAs have a spatio-temporal behaviour in *P. tepidariorum* embryo.

What we need to do next is to identify what are those microRNAs or families of microRNAs that are regulating the early cleavage, blastoderm, limb and brain formation. We have RNA- seq data available for protein coding of *P. tepidariorum*. We will process this data in order to answer what are the target genes of this microRNAs cluster in the embryogenesis and make a correlation between the high expression of microRNAs and the down regulation of their target genes at the embryogenesis in *P. tepidariorum*

5.3. Chapter 3: Pingpong cycle present in Parasteatoda tepidariorum embryo

The importance of this chapter is that is the first work to look into the presence of piRNA in the *P. tepidariorum* embryo. It is also important to mention that this study in different model organism is important within the context of piRNA, to help discover new functions of the piRNAs as many functions are unknown.

We identify the presence of piRNA in all ten stages of embryogenesis of *P. tepidariorum*. piRNAs are highly diverse in their sequences but show certain features such as the size 24- 32nt, the presence of the bias of uridine at the first position of the nucleotide and small proportion of piRNAs showing adenine in the position 10th, some of them are actively involved in the pingpong cycle.

We also obtained information that piRNAs and transposons were inherited maternally in our sample because we could see the pingpong cycle at the early cleavage state and this is only possible when there is an active response against transposons. We get information that the embryos since early stage are producing antisense piRNA to activate the pingpong cycle.

In the pingpong cycle we observe that in some stages, there is an increment of piRNA production, for example at stage 5 (35.5h) that correspond to the cumulus migration. Further work needs to be done in order to know the origin of the other piRNA cluster. We have worked with piRNA clusters originated from the transposon elements sequences. Further work will need to be done in order to obtain the profile expression of piRNA that are transcribed from 3'UTR of mRNA, and the long non-condign RNA

Further analysis to identify the proteins involved in the production of the primary andsecondary piRNA production, as we know *P. tepidariorum* has undergone a whole genome duplication event. *In situ* hybridisation will need to be undertaken to identify the piRNA-piwicomplex or piRNA-ago3 and piRNA-aub in the embryo. None of this work has been done in

P. tepidariorum yet and all information obtained will be valuable. *P. tepidariorum* has whole genome duplication. It has observed that duplication coding and non-coding sequences in the

P. tepidariorum (Schwager et al. 2017). What happened with the coding RNA (piwi, aub and ago3 proteins) of a *P. tepidariorum* after a whole genome duplication event?. For future workwill be explore whether PIWI proteins (piwi, aub and ago3) genes fate in a neofunctionalization or subfunctionalization. To approach what are the functions of these proteins after a WGD. First, we can identify these protein sequences in the small RNA-seq and the RNA-seq expression, there is a previous work testing the expression of piwi in *P. tepidariorum* embryo (Schwager, Meng, and Extavour 2015). Test the presence in vivo through in situ hybridization or test the lack of function using RNAi (Hilbrant, Damen, and McGregor 2012).

5.4. Integration of the chapters

In the first project, chapter 1, I answer whether microRNA regulate to a SSD gene rather thanWGD gene. This chapter allow me to use genomic data, tools and databases available for the analysis how microRNA regulate those genes, however as we know microRNA has spatio- temporal dynamic, to do transcriptome profiling of microRNA give us a precise information what is happening in the cell or tissue than a genomic analysis.

General References

Acharya, Debarun, and Tapash C Ghosh. 2016. “Global Analysis of Human Duplicated Genes Reveals the Relative Importance of Whole-Genome Duplicates Originated in the Early Vertebrate Evolution.” *BMC Genomics*, 1–14. https://doi.org/10.1186/s12864-016-2392-0.

Adrion, Jeffrey R., Michael J. Song, Daniel R. Schrider, Matthew W. Hahn, and Sarah Schaack. 2017. “Genome-Wide Estimates of Transposable Element Insertion and Deletion Rates in Drosophila Melanogaster.” *Genome Biology and Evolution* 9 (5): 1329–40. https://doi.org/10.1093/gbe/evx050.

Aken, Bronwen L., Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Friederike Bernsdorff, Jyothish Bhai, Konstantinos Billis, et al. 2017. “Ensembl 2017 Ia Gir On.” *Nucleic Acids Research* 45 (14): 1–8. https://doi.org/10.1093/nar/gkw1104.

Alberti, Chiara, and Luisa Cochella. 2017. “A Framework for Understanding the Roles of MiRNAs in Animal Development.” *Development (Cambridge)* 144 (14): 2548–59. https://doi.org/10.1242/dev.146613.

Ameres, Stefan L, and Phillip D Zamore. 2013. “Diversifying MicroRNA Sequence and Function.” *Nature Reviews. Molecular Cell Biology* 14 (8): 475–88. https://doi.org/10.1038/nrm3611.

Avital, Gal, and Gustavo S Franc. 2017. “Bimodal Evolutionary Developmental MiRNA Program in Animal Embryogenesis” 35 (3): 646–54. https://doi.org/10.1093/molbev/msx316.

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

Birchler, James A., and Reiner A. Veitia. 2010. “The Gene Balance Hypothesis: Implications for Gene Regulation, Quantitative Traits and Evolution.” *New Phytologist* 186 (1): 54–62. https://doi.org/10.1111/j.1469-8137.2009.03087.x.

Bourque, Guillaume, Kathleen H. Burns, Mary Gehring, Vera Gorbunova, Andrei Seluanov, Molly Hammell, Michaël Imbeault, et al. 2018. “Ten Things You Should Know about Transposable Elements.” *Genome Biology* 19 (1): 199. https://doi.org/10.1186/s13059-018-1577-z.

Carlos, Juan, and Elena Ramirez-parra. 2015. “Whole Genome Duplications in Plants : An Overview from Arabidopsis” 66 (22): 6991–7003. https://doi.org/10.1093/jxb/erv432.

Chen, Wei-hua, Pablo Minguez, Martin J Lercher, and Peer Bork. 2012. “OGEE : An Online Gene Essentiality Database” 40 (November 2011): 901–6. https://doi.org/10.1093/nar/gkr986.

Conrad, Bernard, and Stylianos E. Antonarakis. 2007. “Gene Duplication: A Drive for Phenotypic Diversity and Cause of Human Disease.” *Annual Review of Genomics and Human Genetics* 8 (1): 17–35. https://doi.org/10.1146/annurev.genom.8.021307.110233.

Davis, Jerel C, and Dmitri A Petrov. 2005. “Do Disparate Mechanisms of Duplication Add Similar Genes to the Genome?” *Trends in Genetics : TIG* 21 (10): 548–51. https://doi.org/10.1016/j.tig.2005.07.008.

Deschamps-Francoeur, Gabrielle, Joël Simoneau, and Michelle S. Scott. 2020. “Handling Multi-Mapped Reads in RNA-Seq.” *Computational and Structural Biotechnology Journal* 18 (January): 1569–76. https://doi.org/10.1016/J.CSBJ.2020.06.014.

Friedman, Robin C, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. 2009. “Most Mammalian MRNAs Are Conserved Targets of MicroRNAs.” *Genome Research* 19 (1): 92–105. https://doi.org/10.1101/gr.082701.108.

Giraldez, A. J. 2005. “MicroRNAs Regulate Brain Morphogenesis in Zebrafish.” *Science* 308 (5723): 833–38. https://doi.org/10.1126/science.1109020.

Guschanski, Katerina, Maria Warnefors, and Henrik Kaessmann. 2017. “The Evolution of Duplicate Gene Expression in Mammalian Organs.” *Genome Research* 27 (9): 1461–74. https://doi.org/10.1101/GR.215566.116/-/DC1.

Han, Bo W., and Phillip D. Zamore. 2014a. “PiRNAs.” *Current Biology* 24 (16): 730–33. https://doi.org/10.1016/j.cub.2014.07.037.

———. 2014b. “PiRNAs.” *Current Biology* 24 (16): 730–33. https://doi.org/10.1016/j.cub.2014.07.037.

Hilbrant, Maarten, Wim G.M. Damen, and Alistair P. McGregor. 2012. “Evolutionary Crossroads in Developmental Biology: The Spider Parasteatoda Tepidariorum.” *Development (Cambridge, England)* 139 (15): 2655–62. https://doi.org/10.1242/dev.078204.

Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2009. “Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources.” *Nature Protocols* 4 (1): 44–57. https://doi.org/10.1038/nprot.2008.211.

Huang, Ni, Insuk Lee, Edward M. Marcotte, and Matthew E. Hurles. 2010. “Characterising and Predicting Haploinsufficiency in the Human Genome.” *PLoS Genetics* 6 (10): 1–11. https://doi.org/10.1371/journal.pgen.1001154.

Inoue, Jun, Yukuto Sato, Robert Sinclair, Katsumi Tsukamoto, and Mutsumi Nishida. 2015. "Rapid Genome Reshaping by Multiple-Gene Loss after Whole-Genome Duplication in Teleost Fish Suggested by Mathematical Modeling." *Proceedings of the National Academy of Sciences of the United States of America* 112 (48): 14918–23. https://doi.org/10.1073/pnas.1507669112.

Iwasaki-Yokozawa, Sawa, Yasuko Akiyama-Oda, and Hiroki Oda. 2018. "Genome-Scale Embryonic Developmental Profile of Gene Expression in the Common House Spider Parasteatoda Tepidariorum." *Data in Brief* 19: 865–67. https://doi.org/10.1016/j.dib.2018.05.106.

Johnson, Nathan R., Jonathan M. Yeoh, Ceyda Coruh, and Michael J. Axtell. 2016. "Improved Placement of Multi-Mapping Small RNAs." *G3: Genes, Genomes, Genetics* 6 (7): 2103–11. https://doi.org/10.1534/g3.116.030452.

Kim, Daehwan, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. 2019. "Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype." *Nature Biotechnology* 37 (8): 907–15. https://doi.org/10.1038/s41587-019-0201-4.

King, P H, R Waldrop, J R Lupski, and L G Shaffer. 1998. "Charcot-Marie-Tooth Phenotype Produced by a Duplicated PMP22 Gene as Part of a 17p Trisomy-Translocation to the X Chromosome." *Clinical Genetics* 54 (5): 413–16. https://doi.org/10.1111/j.1399-0004.1998.tb03755.x.

Kiuchi, Takashi, Hikaru Koga, Munetaka Kawamoto, Keisuke Shoji, Hiroki Sakai, Yuji Arai, Genki Ishihara, et al. 2014. "A Single Female-Specific PiRNA Is the Primary Determiner of Sex in the Silkworm." *Nature* 509 (7502): 633–36. https://doi.org/10.1038/nature13315.

Krol, Jacek, Inga Loedige, and Witold Filipowicz. 2010. "The Widespread Regulation of MicroRNA Biogenesis, Function and Decay." *Nature Reviews Genetics* 11 (9): 597–610. https://doi.org/10.1038/nrg2843.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biology* 10 (3). https://doi.org/10.1186/gb-2009-10-3-r25.

Lee, Miler T., Ashley R. Bonneau, and Antonio J. Giraldez. 2014. "Zygotic Genome Activation During the Maternal-to-Zygotic Transition." *Annual Review of Cell and Developmental Biology* 30 (1): 581–613. https://doi.org/10.1146/annurev-cellbio-100913-013027.

Legrand, Sylvain, Thibault Caron, Florian Maumus, Sol Schvartzman, Leandro Quadrana, Eléonore Durand, Sophie Gallina, et al. 2019. “Differential Retention of Transposable Element-Derived Sequences in Outcrossing Arabidopsis Genomes.” *Mobile DNA* 10 (1): 1–17. https://doi.org/10.1186/s13100-019-0171-6.

Leite, Daniel J., Maria Ninova, Maarten Hilbrant, Saad Arif, Sam Griffiths-Jones, Matthew Ronshaugen, and Alistair P. McGregor. 2016. “Pervasive MicroRNA Duplication in Chelicerates: Insights from the Embryonic MicroRNA Repertoire of the Spider Parasteatoda Tepidariorum.” *Genome Biology and Evolution* 8 (7): 2133–44. https://doi.org/10.1093/gbe/evw143.

Lewis, Benjamin P, Christopher B Burge, and David P Bartel. 2005. “Conserved Seed Pairing, Often Flanked by Adenosines, Indicates That Thousands of Human Genes Are MicroRNA Targets.” *Cell* 120 (1): 15–20. https://doi.org/10.1016/j.cell.2004.12.035.

Lewis, Samuel H., Kaycee A. Quarles, Yujing Yang, Melanie Tanguy, Lise Frézal, Stephen A. Smith, Prashant P. Sharma, et al. 2018. “Pan-Arthropod Analysis Reveals Somatic PiRNAs as an Ancestral Defence against Transposable Elements.” *Nature Ecology and Evolution* 2 (1): 174–81. https://doi.org/10.1038/s41559-017-0403-4.

Li, Jingjing, Gabriel Musso, and Zhaolei Zhang. 2008. “Preferential Regulation of Duplicated Genes by MicroRNAs in Mammals.” *Genome Biology* 9 (8): 1–10. https://doi.org/10.1186/gb-2008-9-8-r132.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30 (7): 923–30. https://doi.org/10.1093/bioinformatics/btt656.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 1–21. https://doi.org/10.1186/s13059-014-0550-8.

Lowe, Todd M., and Sean R. Eddy. 1996. “TRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence.” *Nucleic Acids Research* 25 (5): 955–64. https://doi.org/10.1093/nar/25.5.0955.

Lynch, Michael. 2001. “The Molecular Natural History of the Human Genome.” *Trends in Ecology & Evolution* 16 (8): 420–22. https://doi.org/10.1016/S0169-5347(01)02242-X.

Makino, Takashi, Karsten Hokamp, and Aoife Mclysaght. 2008. “The Complex Relationship of Gene Duplication and Essentiality,” 152–55. https://doi.org/10.1016/j.tig.2009.02.001.

Makino, Takashi, and Aoife Mclysaght. 2010. “Ohnologs in the Human Genome Are Dosage Balanced and Frequently Associated with Disease” 107 (20). https://doi.org/10.1073/pnas.0914697107.

Makino, Takashi, and Aoife McLysaght. 2010. “Ohnologs in the Human Genome Are Dosage Balanced and Frequently Associated with Disease.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (20): 9270–74. https://doi.org/10.1073/pnas.0914697107.

Marco, Antonio. 2018. “SeedVicious: Analysis of MicroRNA Target and near-Target Sites.” *PLoS ONE* 13 (4): 1–9. https://doi.org/10.1371/journal.pone.0195532.

McClintock, B. 1950. “The Origin and Behavior of Mutable Loci in Maize.” *Proceedings of the National Academy of Sciences of the United States of America* 36 (6): 344–55. https://doi.org/10.1073/pnas.36.6.344.

McGurk, Michael P., and Daniel A. Barbash. 2018. “Double Insertion of Transposable Elements Provides a Substrate for the Evolution of Satellite DNA.” *Genome Research* 28 (5): 714–25. https://doi.org/10.1101/gr.231472.117.

Mencía, Ángeles, Silvia Modamio-Høybjør, Nick Redshaw, Matías Morín, Fernando Mayo-Merino, Leticia Olavarrieta, Luis A Aguirre, et al. 2009. “Mutations in the Seed Region of Human MiR-96 Are Responsible for Nonsyndromic Progressive Hearing Loss.” *Nature Genetics* 41 (5): 609–13. https://doi.org/10.1038/ng.355.

Misof, B., S. Liu, K. Meusemann, R. S Peters, and Et Al. 2014. “ゲノム系統学は昆虫の進化のタイミングとパターンを解決する Phylogenomics Resolves the Timing and Pattern of Insect Evolution.” *Science* 346 (6210): 763–67. https://doi.org/10.1017/CBO9781107415324.004.

Mittmann, Beate, and Carsten Wolff. 2012. “Embryonic Development and Staging of the Cobweb Spider Parasteatoda TepidariorumC. L. Koch, 1841 (Syn.: Achaearanea Tepidariorum; Araneomorphae; Theridiidae).” *Development Genes and Evolution* 222 (4): 189–216. https://doi.org/10.1007/s00427-012-0401-0.

Ou, Shujun, Weija Su, Yi Liao, Kapeel Chougule, Jireh R.A. Agda, Adam J. Hellinga, Carlos Santiago Blanco Lugo, et al. 2019. “Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline.” *Genome Biology* 20 (1): 1–18. https://doi.org/10.1186/s13059-019-1905-y.

Ozata, Deniz M., Ildar Gainetdinov, Ansgar Zoch, Dónal O'Carroll, and Phillip D. Zamore. 2019. "PIWI-Interacting RNAs: Small RNAs with Big Functions." *Nature Reviews Genetics* 20 (2): 89–108. https://doi.org/10.1038/s41576-018-0073-3.

Paese, Christian L.B., Anna Schoenauer, Daniel J. Leite, Steven Russell, and Alistair P. McGregor. 2018. "A SoxB Gene Acts as an Anterior Gap Gene and Regulates Posterior Segment Addition in a Spider." *ELife* 7: 1–18. https://doi.org/10.7554/eLife.37567.

Parhad, Swapnil S., and William E. Theurkauf. 2019. "Rapid Evolution and Conserved Function of the PiRNA Pathway." *Open Biology* 9 (1). https://doi.org/10.1098/rsob.18.0181.

Pechmann, Matthias. 2016. "Formation of the Germ-Disc in Spider Embryos by a Condensation-like Mechanism." *Frontiers in Zoology* 13 (1): 1–13. https://doi.org/10.1186/s12983-016-0166-9.

Petersen, Malte, David Armisén, Richard A. Gibbs, Lars Hering, Abderrahman Khila, Georg Mayer, Stephen Richards, Oliver Niehuis, and Bernhard Misof. 2019. "Diversity and Evolution of the Transposable Element Repertoire in Arthropods with Particular Reference to Insects." *BMC Evolutionary Biology* 19 (1): 1–15. https://doi.org/10.1186/s12862-018-1324-9.

Pontual, Loïc de, Evelyn Yao, Patrick Callier, Laurence Faivre, Valérie Drouin, Sandra Cariou, Arie van Haeringen, et al. 2011. "Germline Deletion of the MiR-17~92 Cluster Causes Skeletal and Growth Defects in Humans." *Nature Genetics* 43 (10): 1026–30. https://doi.org/10.1038/ng.915.

Rebollo, Rita, Mark T Romanish, and Dixie L Mager. 2012. "Transposable Elements : An Abundant and Natural Source of Regulatory Sequences for Host Genes." https://doi.org/10.1146/annurev-genet-110711-155621.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics (Oxford, England)* 26 (1): 139–40. https://doi.org/10.1093/BIOINFORMATICS/BTP616.

Sarkar, Arpita, Jean Nicolas Volff, and Chantal Vaury. 2017. "PiRNAs and Their Diverse Roles: A Transposable Element-Driven Tactic for Gene Regulation?" *FASEB Journal* 31 (2): 436–46. https://doi.org/10.1096/fj.201600637RR.

Schwager, Evelyn E, Yue Meng, and Cassandra G Extavour. 2015. "Vasa and Piwi Are Required for Mitotic Integrity in Early Embryogenesis in the Spider Parasteatoda Tepidariorum" 402: 276–90.

Schwager, Evelyn E., Prashant P. Sharma, Thomas Clarke, Daniel J. Leite, Torsten Wierschin, Matthias Pechmann, Yasuko Akiyama-Oda, et al. 2017. “The House Spider Genome Reveals an Ancient Whole-Genome Duplication during Arachnid Evolution.” *BMC Biology* 15 (1): 1–27. https://doi.org/10.1186/s12915-017-0399-x.

Singh, Param Priya, Jatin Arora, and Hervé Isambert. 2015. “Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes.” *PLoS Computational Biology* 11 (7). https://doi.org/10.1371/journal.pcbi.1004394.

Siomi, Mikiko C., Kaoru Sato, Dubravka Pezic, and Alexei A. Aravin. 2011a. “PIWI-Interacting Small RNAs: The Vanguard of Genome Defence.” *Nature Reviews Molecular Cell Biology* 12 (4): 246–58. https://doi.org/10.1038/nrm3089.

———. 2011b. “PIWI-Interacting Small RNAs: The Vanguard of Genome Defence.” *Nature Reviews Molecular Cell Biology* 12 (4): 246–58. https://doi.org/10.1038/nrm3089.

Thompson, Julie D., Toby. J. Gibson, and Des G. Higgins. 2003. “Multiple Sequence Alignment Using ClustalW and ClustalX.” *Current Protocols in Bioinformatics* 00 (1): 2.3.1-2.3.22. https://doi.org/10.1002/0471250953.BI0203S00.

Tóth, Katalin Fejes, Dubravka Pezic, Evelyn Stuwe, and Alexandre Webster. 2016. “The PiRNA Pathway Guards the Germline Genome Against Transposable Elements.” *Advances in Experimental Medicine and Biology* 886: 51. https://doi.org/10.1007/978-94-017-7417-8_4.

Vasudevan, Shobha, Emre Seli, and Joan A Steitz. 2006. “Metazoan Oocyte and Early Embryo Development Program : A Progression through Translation Regulatory Cascades,” 138–46. https://doi.org/10.1101/gad.1398906.Rao.

Wang, Yangming, Scott Baskerville, Archana Shenoy, Joshua E Babiarz, Lauren Baehner, and Robert Blelloch. 2008. “Embryonic Stem Cell-Specific MicroRNAs Regulate the G1-S Transition and Promote Rapid Proliferation.” *Nature Genetics* 40 (12): 1478–83. https://doi.org/10.1038/ng.250.

Wu, Changcheng, and Jian Lu. 2019. “Diversification of Transposable Elements In.” *Genes*.

Yin, Hang, and Haifan Lin. 2007. “An Epigenetic Activation Role of Piwi and a Piwi-Associated PiRNA in Drosophila Melanogaster” 450 (November): 3–8. https://doi.org/10.1038/nature06263.

Zhang, Jianzhi. 2003. “Evolution by Gene Duplication: An Update.” *Trends in Ecology & Evolution* 18 (6): 292–98. https://doi.org/10.1016/S0169-5347(03)00033-8.

References R Package

Low D (2021). ssviz: A small RNA-seq visualizer and analysis toolkit. R package version 1.28.0.