Analysis of variation in piRNA expression in the male germline of three *Mus* species.

Adrià Mitjavila Ventura
MSc in Bioinformatics and Biostatistics
Area 2 - Comparative and Regulatory Genomics

Elisabetta Mereu
David Arranz Merino

December 24, 2021

Title:	Analysis of the variation in piRNA expression in the male germline of three *Mus* species.
Author:	Adrià Mitjavila Ventura
Tutor:	Elisabetta Mereu
PRA:	David Merino Arranz
Date:	03/01/2022
Studies	MSC in Bioinformatics and Biostatistics
Area	Comparative and regulatory genomics
Language	English
Keywords	piRNA, transposons, IAP

Abstract

Piwi-interacting RNAs (piRNAs) are small non-coding RNAs expressed in the germline of most animals whose main function is to silence transposable elements (TEs) through base-pair complementarity. In the current work, we studied the piRNA expression and its variation in the male germ line of three closely related *Mus* species, providing the first small RNA datasets in two of these species. In addition, we evaluated factors that could influence piRNA expression and its diversity between species. First, we confirmed that our sequencing data was enriched in piRNAs. We also developed an approach to minimize length differences between orthologous regions that allowed performing multi-species differential expression analyses. In summary, we found that the piRNA-producing loci (piRNA clusters) and its expression have great differences between species. On the other hand, the most conserved piRNA clusters across species were those with higher expression of piRNAs. Finally, although we could not find significant associations between TEs and piRNA expression, we provide some examples consistent with a model where TE insertions alter production of piRNAs. Our results suggest that presence of transposon insertions may be the origin of a subset of new piRNA clusters. However, there must be additional factors explaining the piRNA diversity between species. Thus far, this has been the first piRNA study comparing closely related species within the mammalian clade and it is a first step towards unravelling the mechanisms by which piRNA-producing genes evolve.

Abstract - Català

Els ARN associats a Piwi (*piRNAs*) son ARN petits no codificants que s'expressen a la línia germinal de la major part d'animals i la seva funció principal és silenciar els transposons per mitjà de complementarietat de bases. En aquest treball hem estudiat l'expressió de *piRNAs* i la seva variació en la línia germinal de mascles de tres espècies del gènere *Mus*, incloent les primers dades d'ARN petits de testicles en dues d'aquestes. Hem avaluat factors que poguessin influir en l'expressió de *piRNAs* i la seva diversitat entre espècies. Primerament, hem confirmat que les dades de seqüenciació estaven enriquides en *piRNAs* i hem desenvolupat un mètode per minimitzar les diferències de longitud entre ortòlegs i poder realitzar anàlisis d'expressió diferencial amb diverses espècies. Resumidament, hem vist que els loci productors de *piRNA* (*piRNA clusters*) i la seva expressió tenen grans diferències entre espècies. D'altra banda, els *piRNA clusters* que produeixen més *piRNAs* són els més conservats entre espècies. Finalment, si bé no hem trobat associacions significatives globals entre la presència de transposons i l'expressió dels *piRNA clusters*, hem proporcionat exemples consistents amb un model on els transposons alteren la producció de *piRNAs*. En resum, els resultats suggereixen que els transposons poden explicar l'origen d'alguns *piRNA clusters*, encara que hi ha d'haver senyals addicionals que expliquin la variació dels *piRNAs* entre espècies. Fins al moment, aquest ha sigut el primer estudi sobre *piRNAs* comparant espècies de mamífers estretament relacionades i és un pas per desxifrar els mecanismes d'evolució dels gens productors de *piRNAs*.

Contents

List of tables

List of supplementary tables

List of figures

List of supplementary figures

1. Introduction

1.1. Context and motivation behind the work

Piwi-interacting RNAs (piRNAs) are small non-coding RNAs which are expressed in the germline of most animals. Their main function is to guide PIWI proteins to silence transposable elements (TEs) -at transcriptional and post-transcriptional level- through base-pair complementarity, something that is required for the normal progression of mammalian spermatogenesis. Indeed, in male mammals, all mutants lacking proteins for piRNA biogenesis are infertile. Furthermore, it has been shown that they also control expression of some genes (*Özata et al., 2019*).

Unlike other small RNAs like microRNAs (miRNAs) and small interfering RNAs (siRNAs), piRNAs derive -with some exceptions (*Ruby et al., 2006*)- from long single-stranded transcripts (piRNA precursors) that are transcribed from loci called piRNA clusters.

In the adult mammalian male germline, piRNA biogenesis occurs due to two different, but interacting, pathways: phased (primary) pathway and ping-pong (secondary) pathway. Precursor transcripts are usually processed and exported from the nucleus to be sliced by a piRNA-guided PIWI protein leaving a 5' monophosphorylated pre-pre-piRNA. This pre-pre-piRNA enters the phased biogenesis pathway -which might be partially guided by ribosomes (*Sun et al., 2020*)- to be consecutively sliced by PIWI-directed endonucleolytic cleavage generating pre-piRNAs that are trimmed 3'-to-5' to mature into piRNAs. The first of these piRNAs (responder piRNA) undergoes the ping-pong cycle to reinitiate the whole process and cleave the precursor transcript (*Gainetdinov et al., 2018*; *Özata et al., 2019*). During this whole process, proteins other than PIWIs have essential roles whose presence is mandatory for piRNA biogenesis and the whole spermatogenesis (e.g. the helicase Mov10l1 or the endonuclease Pld6, etc.) (*supp. figure 1*).

In postnatal mammalian male germ cells, three classes of piRNAs exist considering the stage of the spermatogenesis they start to accumulate: pre-pachytene, hybrid and pachytene piRNAs. Pre-pachytene piRNAs are produced from TE-rich intragenic regions during pre-pachytene stages of the spermatogenesis and are the most abundant piRNAs in fetal testes and isolated spermatogonia (*Özata et al., 2019*), whereas pachytene piRNAs arise mainly from TE-depleted (<20% of TEs) intergenic transcripts during the pachytene stage of the prophase I of meiosis, generally map to unique regions in the genome (*Li et al., 2013*) and are the most abundant piRNAs in the adult testes. On the contrary, hybrid piRNAs share characteristics with both pre-pachytene and pachytene piRNAs (*Li et al., 2013*).

In terms of function, pre-pachytene piRNAs are directed to repress TEs (*Ernst et al., 2017*), whereas pachytene piRNAs are much less understood: they are linked to massive mRNA elimination during spermatogenesis (*Gou et al., 2014*) and they are essential for male fertility (*Choi et al., 2021; Wu et al., 2020*). They may also target some TEs and, like hybrid piRNAs they might associate with ribosomes to fine-tune protein synthesis (*Sun et al., 2020; Sun et al., 2021*). However, the deletion of a subset

of pachytene piRNA clusters, in spite of being among the largest ones, does not produce any phenotype on male fertility (*Wu et al., 2020*), suggesting that their biological function is extensively redundant, yet to be explored or inexistent.

Moreover, despite the high conservation of PIWI pathway proteins, piRNAs evolve very fast, and they may have great variation across different species, even intra-species, either due to differences in the cluster sequences (which may force the transcripts to undergo the PIWI pathway) and/or differences in the expression level (*Chirn et al., 2015; Ozata et al., 2020*). One of the reasons for this rapid evolution could be because many piRNA clusters have -and target- TEs (*Assis et al., 2009*). TEs are known to be subject to several evolutionary forces and, in fact, there are large amounts of species-specific TE insertions and even within the same species TE variation is huge (*Lilue et al., 2018; Nellaker et al., 2012*).

TEs -sometimes referred to as mobile elements (ME) or transposons- are fragments of the genome that can be integrated to other locations, either by “cut-and-paste” or “copy-and-paste” approaches, thus, they are a great source of mutations and evolution. TEs come in many ways and from many origins (*Bourque et al., 2018*). The two main TE groups are DNA transposons and retrotransposons. Among others, the latter comprise short interspersed nuclear elements (SINE), long interspersed nuclear elements (LINE) and long terminal repeats (LTR), which include endogenous retroviruses (ERVs), LTRs that result from past retroviral infections and integrations in the germline.

Activity of TEs is more frequent in germ cells due to the existence of an epigenetic reprogramming that implies global DNA demethylation and other epigenetic changes to prime germ cells for totipotency (*Surani et al., 2010*). This transposon derepression jeopardizes the whole spermatogenesis in mammalian male germ line; hence, TE-silencing is crucial to ensure reproduction.

TE-silencing in mammalian male germline is carried out through transcriptional and post-transcriptional silencing mechanisms during the germ cell development and is promoted by PIWI proteins: MIWI2, a nuclear protein expressed in the embryonic germ cells; MIWI, a cytoplasmic protein expressed in postnatal germ cells from pachytene to round spermatid stages of spermatogenesis; and MILI, a cytoplasmic protein expressed in embryonic and postnatal germ cells (*Carmell et al., 2007; Ernst et al., 2017; Kuramochi-Miyagawa et al., 2001*).

In embryonic germ cells, piRNAs derived from TE-rich regions direct MILI to silence TEs through endonucleolytic cleavage (slicing) of TE transcripts -especially LINE1 and intracisternal A-particle (IAP), an endogenous retrovirus type K (ERV)-. This cleavage generates a piRNA that guides MIWI2 to the nucleus in order to bind the nascent TE transcript and recruit DNA methylation -and other repressing- machinery to transcriptionally silence the TE (*Carmell et al., Dev Cell, 2007; Kojima-Kita, Cell Rep, 2016; Pezic et al., Genes Dev, 2016*). In the case of LINE1, but not for IAP, this silencing is boosted due to piRNA amplification dependent on MILI slicing activity (*De Fazio et al., 2011*). Moreover, MILI can be linked to DNA methylation in a wider range of TEs (e.g. LINEs and LTRs) in a MIWI2-independent manner (*Manakov et al., 2015*).

On the other hand, in postnatal germ cells MIWI and MILI are guided from piRNAs that arise from discrete piRNA clusters to silence TEs in a posttranscriptional manner. During pre-pachytene stages of spermatogenesis MILI binds pre-pachytene TE-rich piRNAs and is directed to TE transcripts that escaped transcriptional gene silencing (i.e. DNA methylation promoted by MIWI2) to slice them (*Di Giacomo et al., 2013; Ernst et al., 2017*). Although pachytene piRNA generally have surprisingly low repeat content (*Girad et al., 2006*), some of them bind MIWI to also slice TEs that escaped transcriptional repression (*Ernst et al., 2017; Reuter et al., 2011*).

The relationship between piRNAs and TEs is strong and has been proven in several ways. In addition to the fact that the depletion of piRNA-related machinery leads to the derepression of TEs, it has been shown that, in chickens, the infection with avian leukosis virus (AVL), an endogenous retrovirus, induces the production of piRNAs from pre-existing piRNA clusters as a defence against the infection (*Sun et al., 2017*), hinting that TE activity may induce piRNA production.

Despite being non-coding, approximately half of the piRNAs in mice are produced from protein-coding transcripts that are selected to enter the PIWI pathway to produce piRNAs (*Li et al., 2013*), and these vary across different strains and species. One mechanism responsible for originating new clusters is duplication of a pre-existing one (*Assis & Kondrashov, 2009*). Nevertheless, what drives a protein-coding or intergenic transcript to enter the piRNA pathway in a species-specific -or strain-specific- way needs further investigation.

Some studies have shown that genomic context of a piRNA cluster has little importance in the selection of a locus to be transcribed into a piRNA precursor, since transgenic piRNA clusters inserted in ectopic locations still produced piRNAs (*Muerdter et al., 2012; Goh et al., 2015*). Others have shown that, although some orthologous genes produce piRNAs in one species but not in others, they are still transcribed to mRNA (*Chirn et al., 2015*). Considering this, if there is a signal that triggers a locus to be selected to enter the piRNA biogenesis pathway, it must be located within the sequence, and it likely acts at post-transcriptional level.

Some of the approaches used to unravel the triggers for piRNA biogenesis have been focussed on the comparison of piRNA clusters across distant-related species and also within species (*Assis & Kondrashov, 2009; Chirn et al., 2015; Ozata et al., 2020*), finding that piRNA producing loci are highly divergent but there are sets of conserved clusters that are putatively implied in conserved functions for the physiology of several species.

Nevertheless, no studies in mammals have been focussed on the comparison of closely related species, such as *Mus musculus*, *Mus caroli* and *Mus pahari*. More specifically, *Mus caroli* and *Mus pahari* diverged approximately 3 million years (MY) and 6 MY ago, respectively, from the *Mus musculus* lineage (*Thybert et al., 2018*). *Homo sapiens* (human) and *Rattus norvegicus* (rat) are the main organisms compared to *Mus musculus* regarding piRNAs and their lineages diverged from *Mus musculus* lineage approximately 91 and 15 MY ago, respectively (*Murphy et al., 2007*), comparisons between the mentioned *Mus* species may provide new insights on many aspects of piRNA biology.

Unraveling the qualitative (i.e. sequence) and quantitative (i.e. expression level) differences of the piRNA-producing loci across these three closely related species may help us understand the mechanisms of piRNA production itself, but also potentially the impact that repetitive sequences and TEs have on germline gene expression.

To achieve these goals, we analyzed small RNAs sequenced from adult testes of *Mus musculus* C57BL6 strain (house mouse), *Mus caroli* (ryukyu mouse) and *Mus pahari* (shrew mouse) and compared the repertoire, conservation and expression of piRNAs in these three closely related mammalian species.

1.2. Objectives

The general objectives of this project, which can be broken down into more specific ones, are:

1. To assess the quality of generated small RNA-seq data from the three species and compare their quality. This can be separated into:
    a. To count the number of raw, adaptor-trimmed, quality-filtered and aligned reads.
    b. To assess the proportion of reads that fit into the piRNA specific characteristics (e.g. size, U as first nucleotide...).
    c. To count the genomic features (i.e. exons, introns, UTRs, intergenic regions) where the reads map into.
2. To define a method for differential expression of orthologous genes and regions across closely related species. This can be broken down into:
    a. To find a method to determine orthologous and conserved regions across different species.
    b. To define a differential expression method that accounts for differences in gene length.
3. To test whether transposable element insertions and deletions have contributed to the divergence of piRNA production in mice. For this, we will aim:
    a. To identify regions that have additions and/or deletions of repeats and/or transposable elements in one species compared to the others.
    b. To test whether these species-specific transposable elements are associated with a change in the expression of the piRNA loci containing them.

1.3. Methods

This study was initially focussed in the differential expression analysis of piRNA clusters of six small RNA samples coming from adult testis of three *Mus* species (i.e. *Mus musculus* -the common mouse-, *Mus caroli* and *Mus pahari*) which share a common ancestor approximately 6 million years ago (*Thybert et al., 2018*).

Globally, the project could be divided in four main parts: (1) raw data pre-processing, (2) quality checks and exploratory analyses, (3) definition of piRNA clusters and orthologous regions in each species, and (4) expression analyses to look for differences in piRNA production and potential traits that explain these differences across species.

The first was the most straightforward part of the project. Briefly, it consists of using `Cutadapt` (*Martin, 2011*) to trim sequencing adaptors (`TGGAATTCTCGGGTGCCAAGG`) from the raw sequencing reads and filtering them by quality using `fastq_quality_filter` from the `FastX` toolkit (*http://hannonlab.cshl.edu/fastx_toolkit/*). Filtered reads were aligned to corresponding genome assemblies (i.e. GRCm39, CAROLI_EIJ_v1.1 and PAHARI_EIJ_v1.1) downloaded from ENSEMBL release 104 (*Howe et al., 2020*) using `Bowtie` (*Langmead et al., 2009*), and the resulting alignments were sorted, converted to BAM and indexed using `Samtools` (*Li et al., 2009*).

Secondly, we checked whether our small RNA-seq data fulfilled the characteristics of piRNA biogenesis. Using custom bash and R scripts, we calculated the first and tenth nucleotide compositions for each read as well as the read length distribution. We also used Perl programs (*Rosenkranz et al., 2015*) to compute the 5'-to-5' distance between reads on opposite strands -as a measure for the ping-pong effect- and the 3'-to-5' distance on consecutive reads -as a measure for the phased biogenesis-.

To define piRNA clusters, we used an existing annotation previously defined in *Mus musculus* by *Yu et al.* (2021). We also used `proTRAC` (*Rosenkranz and Zischler, 2012*) to predict *de novo* piRNA clusters based on our data. Obtaining the orthologous regions of these piRNA clusters in the other species, we considered two possible approaches. The first one, `liftOver`, was straightforward and allowed to convert most regions, but it did not account for length differences, something that could bias further expression analyses. On the other hand, ENSEMBL Compara Perl API was also used to obtain the coordinates and aligned sequences of the input regions from the Murinae multiple alignment (i.e. piRNA clusters). Using custom Bash and R scripts allowed us to retrieve the "conserved blocks" -blocks in the multiple alignment that have a match/mismatch in all the aligned organisms-, minimizing the differences in gene length and reducing potential biases in future analyses. Although ENSEMBL Compara Perl API returned fewer regions than `liftOver`, we decided to use the former one.

Lastly, estimation of the expression was performed using `featureCounts` with the aligned reads -BAM files- and each conserved blocks of the piRNA clusters -GTF files- as inputs. These estimates -raw counts- were imported into R and processed with DESeq2 to perform the differential expression analysis and detect significant differences in piRNA production across species. `DESeq2` -as most differential expression analysis tools- normalizes by library size and library composition (*Love et al., 2014*), but it does not account for differences in gene length and that is why the conserved blocks were used. In addition, TEs from `repeatMasker` annotation were downloaded from the UCSC Table Browser and split by transposon classes and subclasses. Each type of TE was intersected with the coordinates of the piRNA cluster in order to check the differences in piRNA cluster expression regarding the species-specific presence or absence of transposons and to test the potential associations.

A full description of the methods as well as the tools used can be found in *Section 6.3 Methods* and *supp. tables S5 and S6*.

1.4. Work plan

We planned our work in an iterative way that allowed us to perform the analyses with the previously defined clusters and, afterwards, the clusters predicted *de novo*. Pre-processing and quality controls were performed at the beginning and mostly in parallel. Also, the writing of this thesis and previous tasks was planned to be carried out in parallel with the different analysis, while we were obtaining and discussing the results.

A full description of the planning, including dates, tasks, milestones and Gantt charts can be found in *Section 6.4. Work plan.*

1.5. Summary of the results

Our method to obtain orthologous regions and conserved blocks from ENSEMBL Compara Perl API minimizes length differences, thus reducing potential biases in the results.

Although small RNA-seq data were not obtained from Piwi immunoprecipitates, the performed quality controls allowed us to assume that we were working with piRNAs.

Analyses of *de novo* clusters and their intersecting genes revealed great differences in piRNA production across closely related species. This is further supported by principal component analysis.

Highly expressed piRNA clusters tend to be more conserved -in terms of expression- in other species, as shown by the higher correlation in pachytene and top expressed clusters (Q1).

Significant links between presence of TEs and differential piRNA production could not be established among all clusters, but we showed several examples in which TEs could trigger the origin of piRNA production in a species-specific way.

1.6. Description of the chapters

In the following chapters, we describe the analyses that were performed, the obtained results and their discussion:

- **Data preparation**: description of the samples used for the analyses and the preparation of the data, including some analyses to decide which approach was followed to obtain the piRNA cluster orthologs in each species.
- **Quality and exploratory analyses**: some quality controls, including the composition of the first and tenth nucleotides of each reads, read length distribution and distinct measures for ping-pong amplification and phased biogenesis. Quick exploratory analyses on the Integrative Genomics Viewer (IGV) (*Robinson et al., 2011*) were also performed.
- **piRNA clusters from *Yu et al.* (*2021*)**: analyses using piRNA cluster defined by *Yu et al.* (*2021*), including:

    - Orthology: descriptive analyses of piRNA clusters and their conversion to other species.
    - Conservation of expression (I): analyses performed to show the link between the expression level in the original species and the conservation of the expression in the new species.
    - Expression and transposable elements (I): analyses that test potential associations between transposable elements and differential piRNA production across species.
- ***De novo* piRNA clusters**: analyses using the *de novo* piRNA clusters predicted with `proTRAC`, including:
    - Prediction of *de novo* clusters: brief description of how *de novo* clusters were predicted and some exploratory analysis comparing them with clusters from *Yu et al.* (*2021*).
    - Obstacles in *de novo* cluster prediction: report of some issues faced with the prediction of *de novo* clusters.
    - Genic and intergenic *de novo* piRNA clusters: first approximation to define genic and intergenic clusters to analyze cluster conservation and species-specific differences in the predicted clusters.
    - Conservation of expression (II): analyses performed to show the link between the expression level in the original species and the conservation of the expression in the new species.
    - Expression and transposable elements (II): analyses that test potential associations between transposable elements and differential piRNA production across species.
- **Discussion**: conclusions retrieved from the results.

2. Project design and execution

2.1. Data preparation

In the present study we studied piRNA production from 215 piRNA clusters defined by *Yu et al.* (2021) (hereafter, *Yuetal* clusters) as well as *de novo* clusters predicted by us, and the differences across closely related species. We analyzed small RNA (sRNA)-seq data from 6 samples belonging to three *Mus* species, i.e. *Mus musculus* (mouse), *Mus caroli* (ryukyu mouse) and *Mus pahari* (shrew mouse), in order to discover differential piRNA production across-species and the specific traits causing these variations. The data was processed following the workflow shown in *supp. figure 2*.

Since *Yuetal* clusters were defined in *Mus musculus*, we tested two methods to obtain the orthologous regions in the different species: (1) `liftOver` to convert the coordinates and (2) the ENSEMBL Compara Perl API (*supp. figure 3A*) to retrieve the conserved blocks (i.e. blocks in the sequence where all the assemblies from a multiple alignment show a match/mismatch and not a gap (*supp. figure 3B*)) from the Murinae multiple alignment. Basically, those piRNA clusters that could be mapped to species other than the original were considered orthologous -or conserved in terms of sequence-.

Each approach resulted in several advantages and drawbacks. On one hand, `liftOver` returned a high number of orthologous piRNA clusters (*figure 1A*), both in the original and new species, in addition to the easy control of duplications. Nevertheless, the gene length variation across the different species was great (*figure 1B*) and its correction for further differential expression (DE) analyses was difficult. On the other hand, the ENSEMBL Compara Perl API returned fewer orthologous piRNA clusters (*figure 1C*), including some duplications. However, as only the conserved blocks were used to estimate the gene expression, differences in gene length across the species could be ignored (*figure 1D*). Therefore, analyses comparing distinct species were based on the conserved blocks from the ENSEMBL Compara Murinae multiple alignment.

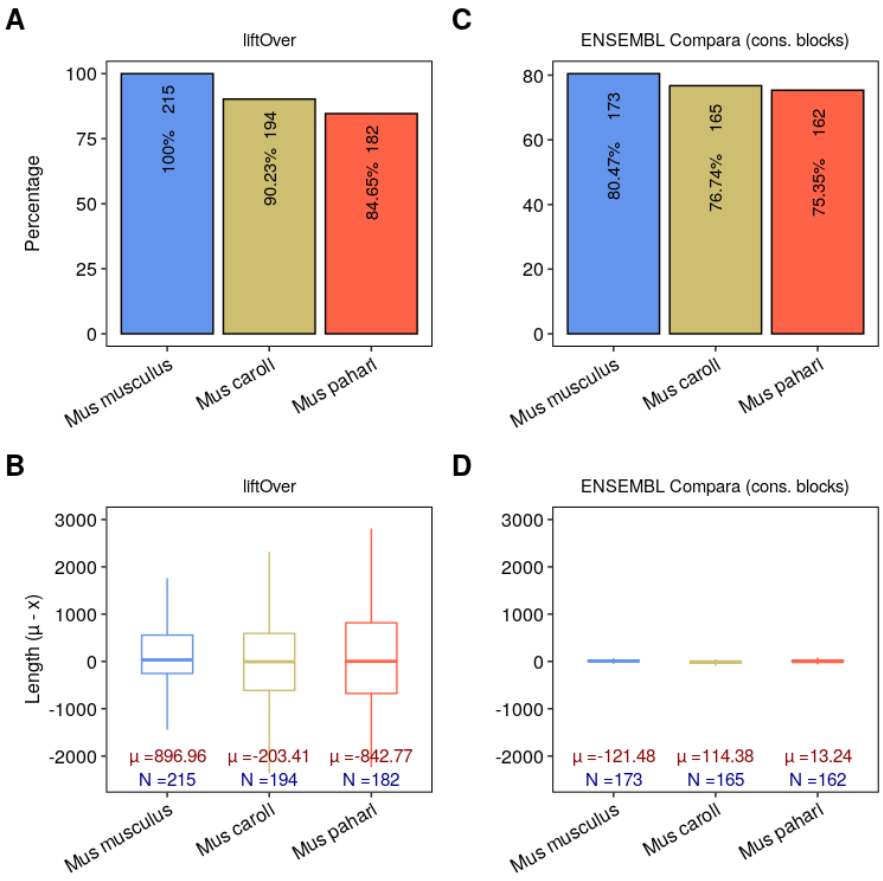


Figure 1. *Yuetal* clusters defined in *Mus musculus* and converted to *Mus caroli* and *Mus pahari*. (A and B) *Yuetal* clusters and orthologs obtained with liftOver: (A) percentatge and number of clusters and orthologs in each species; (B) difference between the mean length of a cluster and the length in each species. (C and D) *Yuetal* clusters and orthologs obtained with ENSEMBL Compara (conserved blocks): (C) percentatge and number of clusters and orthologs in each species; (D) difference between the mean length of a cluster and the length in each species.

After obtaining the conserved blocks from the piRNA clusters in *Mus musculus*, *Mus caroli* and *Mus pahari*, these were used to estimate the expression. Clusters with duplications or with great length differences were removed across species. Then, we used a Fligner-Killeen test to check the homoscedasticity -homogeneity of variance- of the data. As expected from RNA-seq, where variance grows with the mean expression, our data was not homoscedastic ($p_{\text{Fligner-Killen}}$ = 1.205e-3). Counts were normalized using `DESeq2` (mean of ratios normalization) (*Love et al., 2014*) or reads per kilobase per million mapped reads (RPKM) normalization and, when needed, variance-stabilizing transformation like `rlog()` was applied.

2.2. Quality and exploratory analyses

As a first check for the quality of the data, for each sample, we counted the sRNA reads in each processing step, including the reads mapping into piRNA clusters from different annotations, as shown in *supp. table1*.

Depending on which PIWI protein they bind, piRNAs tend to differ in length. Length distribution of piRNAs bound to MILI has a peak around 26-27 nucleotides, whereas for piRNAs bound to MIWI, the peak is around 30 nucleotides (*Gainetdinov et al., 2018*). After trimming adaptors and filtering by read quality, our data shows a unimodal

distribution of the length with a clear peak around 29-30 bp *(figure 2A)*, indicating that most piRNAs in our data bind to MIWI, as expected from pachytene piRNAs.

Furthermore, piRNAs present a bias towards uracil (U) as the first base (1U) or adenosine (A) as the tenth base (10A), depending on the piRNA biogenesis pathway. Our reads are clearly enriched for 1U (*figure 2B*), while 10A is more abundant than other bases, but its enrichment is not as clear as 1U (*figure 2C*). Since phased piRNA pathway is most common in whole testes, our data agrees with expected piRNA characteristics, although ping-pong amplification cannot be ruled out as the source of some piRNAs. In fact, all the samples have significant ping-pong amplification, observed as the 5'-to-5' distance between overlapping reads on opposite strands (*supp. table2; figure 2D*). However, the 3'-to-5' distance between consecutive reads - measure of phased piRNA biogenesis- shows a local -but not global- peak at 0 (*figure 2E*), indicating that our small RNA-seq data have more types of small RNAs other than phased piRNAs.

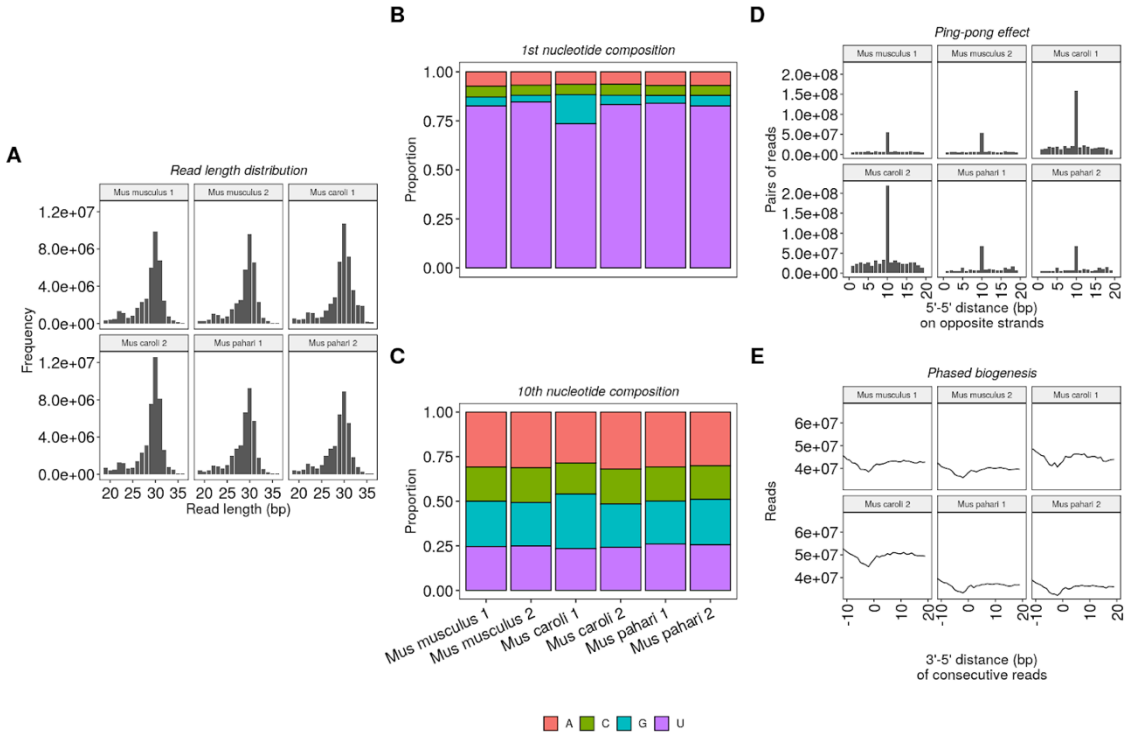


Figure 2. Quality controls of the small RNA-seq data: (A) read length distribution; (B) composition of first nucleotide (in proportion); (C) composition of the thenth nucleotide (in proportion); (D) 5'-to-5' distance of reads in oposite strands (measure of the ping-pong effect); (E) 3'-to-5' distance of reads on the same strand (measure of phased biogènesis).

Regarding the reads mapping to piRNA clusters, a quick exploration in the `Integrative Genomics Viewer (IGV)` (*Robinson et al., 2011*) shows that, for many *Yuetal* genic clusters, reads are mapped mainly to the last exon of the gene, indicating that most piRNAs are produced from 3'UTRs (*figure 3A*), agreeing with literature (*Gainetdinov et al., 2018*; *Li et al., 2013*). In intergenic piRNA clusters, reads tend to align across the whole region (*figure 3B*).

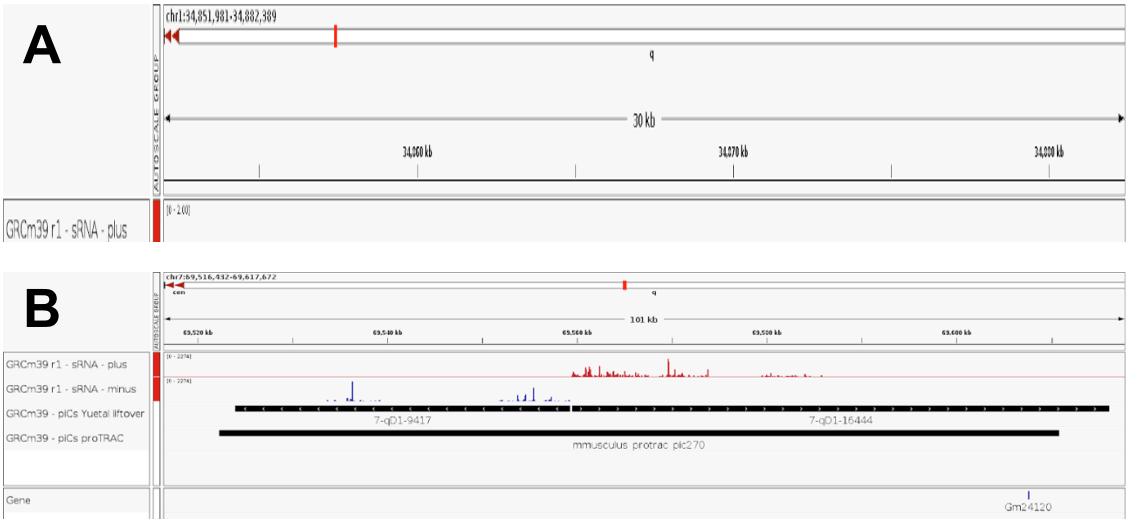


Figure 3. IGV snapshots showing small RNA-seq data and piRNA clusters from *Yu et al* in the GRCm39 assembly. (2015) and predicted *de novo* with proTRAC: (A) *Fam168b* gene, genic cluster *pi-Fam168b* and *de novo* cluster predicted in *Mus musculus* (*musculus1*); (B) intergenic clusters *7-qD1-9417* and *7-qD1-16444* (divergently transcribed from a bidirectional promoter) and bidirectional *de novo* cluster predicted in *Mus musculus* (cluster *musculus270*).

Finally, after estimating the small RNA expression in the *Yuetal* clusters, counts were imported into R to check the variability between the samples. Counts were normalized with `DESeq()` and variance-stabilized using `rlog()`. A principal components analysis (PCA) using the sRNA expression of *Yuetal* clusters showed great variation between the different species, but high similarity in the samples within the same species (*figure 4*), indicating that the species was a clear source of variation in piRNA production.

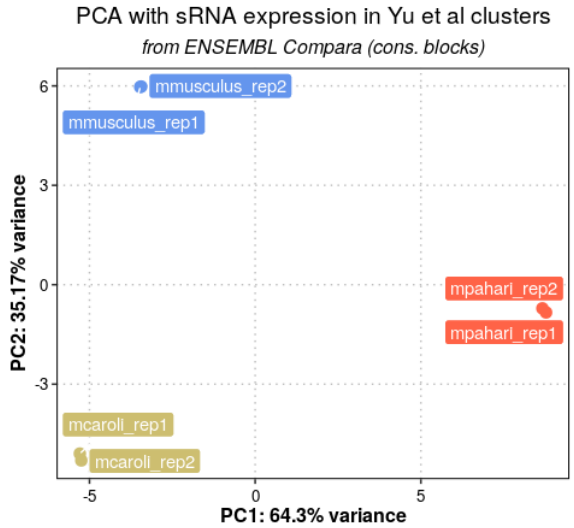


Figure 4. Principal component analysis with the small RNA expression in the clusters from *Yu et al. (2015)* and their orthologs obtained with ENSEMBL Compara (conserved blocks).

2.3. piRNA clusters from *Yu et al. (2021)*

2.3.1. Orthology

From the 215 original *Yuetal* clusters, only 173 regions were present in the Murinae multiple alignment (*figure 1C*), some of which were duplications that had to be removed from subsequent analyses. Considering orthologous clusters as those that could be mapped from *Mus musculus* (GRCm39) to other assemblies in the multiple alignment, *Mus caroli* had more orthology (165 regions) with *Mus musculus* than *Mus pahari* (162 regions) (*figures 1C* and *5A*). This is expected, since the former diverged from the *Mus musculus* lineage about 3 MY ago, whereas the latter diverged approximately 6 MY ago.

Since the pre-pachytene, hybrid and pachytene piRNA clusters have different sources and expression levels, it can be expected that they may also have different sequence conservation across several species. In our case, pre-pachytene and hybrid clusters had more orthology than pachytene clusters, since almost all of them were present in both *Mus caroli* and *Mus pahari*, while a lower proportion of pachytene piRNA clusters were maintained in the new species (*figure 5B*). Although this bias may not be relevant, it can be explained by the fact that most pachytene piRNA clusters lie within intergenic regions, which tend to be subjected to more mutations and quicker evolution.

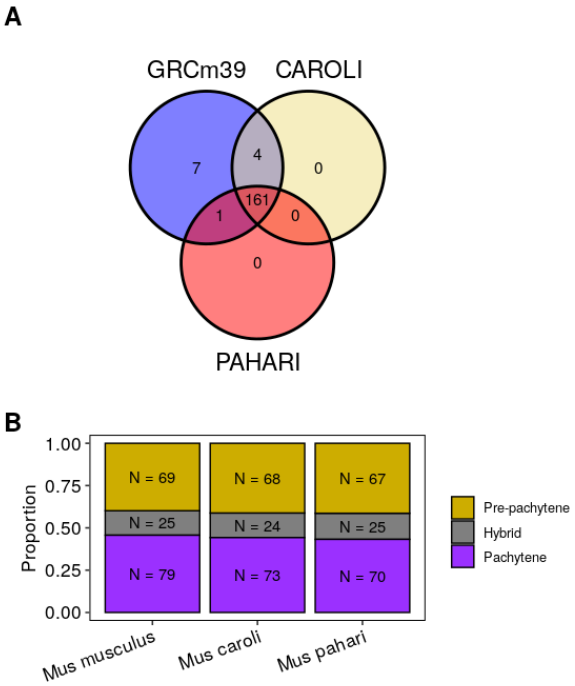


Figure 5. piRNA clusters from *Yu et al.*, (2021) and their orthologs obtained from ENSEMBL Compara (conserved blocks): (A) intersections of the clusters and their orthologs (i.e. how many could be mapped to each species); (B) proportion and number of each class of piRNA clusters.

2.3.2. Conservation of expression (I)

Only those clusters without duplications and with orthologs in the three species were retained, with 158 clusters remaining. In terms of expression -mean RPKM- *Yuetal* clusters were more expressed in *Mus musculus* than the other species, especially in the case of pre-pachytene clusters, which showed general significant differences, and hybrid clusters (*figure 6A*). Regarding pachytene clusters, which were the most expressed ones, there were no general differences between species (*figure 6A*). This raised two possible explanations: (1) expression of pachytene clusters is more conserved or (2) that higher expression level implies higher conservation of the expression across species.

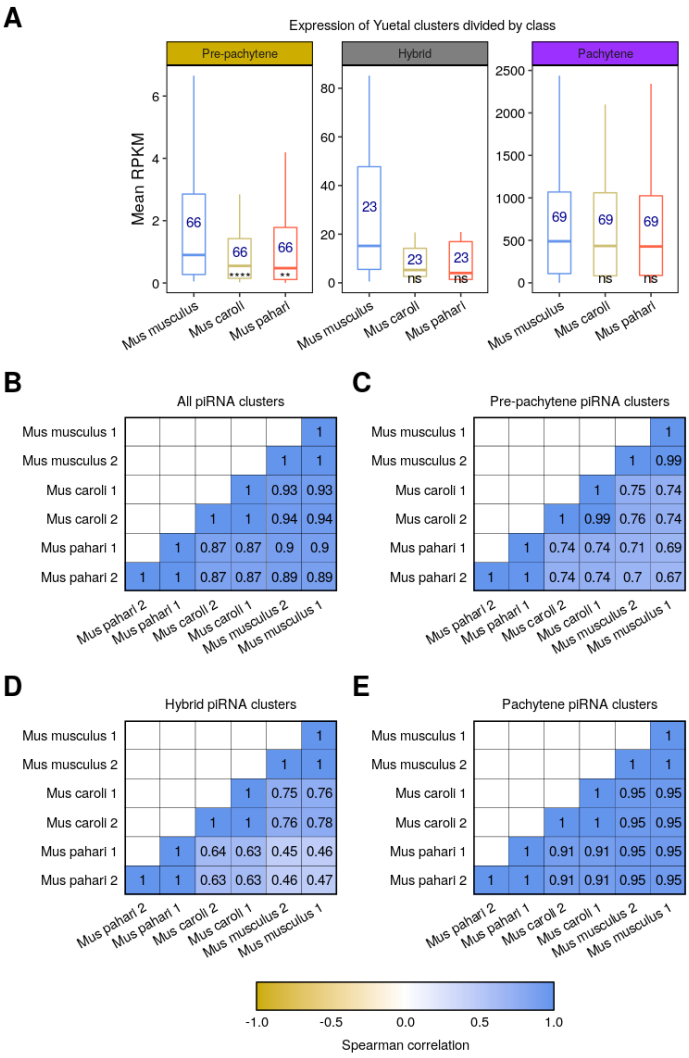


Figure 6. Expression and Spearman correlations of piRNA clusters from *Yu et al* (2021): (A) expression (RPKM) of clusters divided by class; (B) correlations between samples considering the DESeq2-normalized expression of clusters present in all the species; (C) correlations between samples considering DESeq2-normalized expression of pre-pachytene clusters present in all the species; (D) correlations between samples considering DESeq2-normalized expression of hybrid clusters present in all the species; (E) correlations between samples considering DESeq2-normalized expression of pachytene clusters present in all the species.

Spearman correlation showed that, although the expression in the *Yuetal* clusters correlates well across species (*figure 6B*), the piRNA production in pre-pachytene (*figure 6C*) and hybrid (*figure 6D*), but not pachytene (*figure 6E*), clusters had lower correlation between species, reinforcing the previous results. In addition, looking at cluster expression rather than the class showed that correlation between *Mus musculus* and *Mus caroli* or *Mus pahari* increased with the expression (*figure 7*).

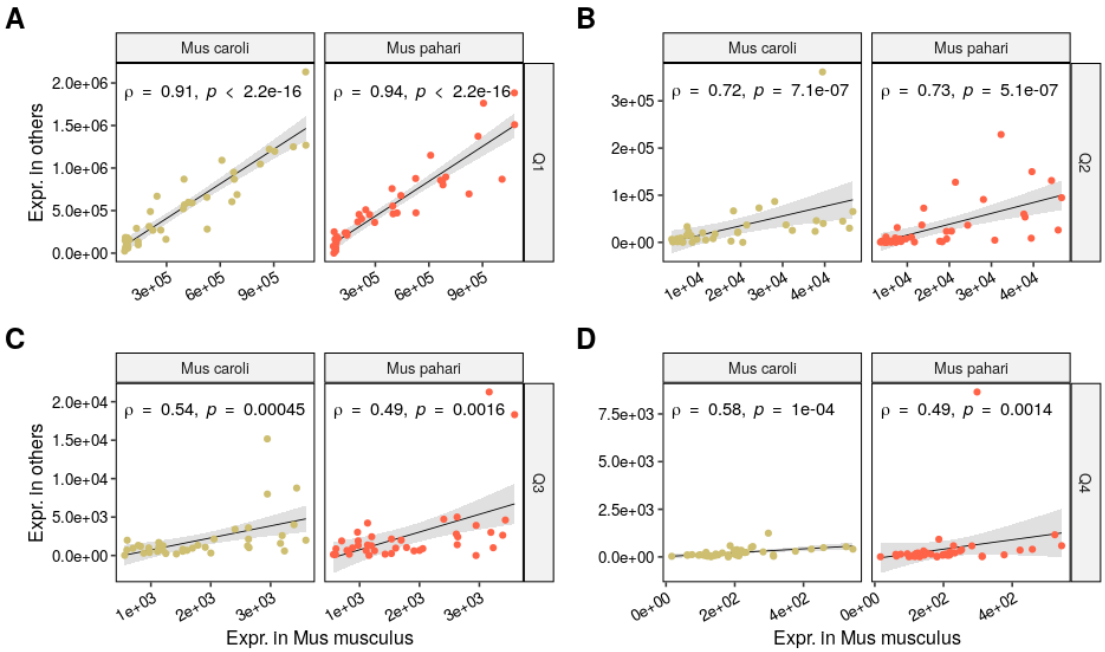


Figure 7. Scatter plots and Spearman correlations of the expression of clusters from *Yu et al.* (2021) and their orthologs divided by their expression (quartiles) in *Mus musculus*: (A) top expressed clusters (Q1); (B) clusters in the Q2; (C) clusters in the Q3; (D) clusters in the Q4.

2.3.3. Expression and transposable elements (I)

For differential expression (DE) analysis, we performed three contrasts comparing all our species in a pairwise manner. This is *Mus pahari vs Mus musculus* (hereafter, "PAHvsMUS"), *Mus caroli vs Mus musculus* (hereafter, "CARvsMUS") and *Mus pahari vs Mus caroli* (hereafter, "PAHvsCAR"). To call a piRNA cluster as differentially expressed (DEpiC), thresholds of $|log_2 foldchange| > 1$ and adjusted $p\text{-}value < 0.05$ were established.

As expected from the PCA, we observed several DEpiCs in all the contrasts (*figure 8A*). Curiously, all pairwise comparisons implying *Mus musculus* as reference had a tendency for downregulated over upregulated clusters, implying higher expression in this species. Although this is interesting, it is probably a bias arising from the fact that *Yuetal* clusters were defined in *Mus musculus*.

Among all the factors that may influence differential expression on a genomic feature, TEs -such as ERVs- make one of the most interesting possibilities, since they have been proven to be source of mutations and genetic polymorphisms (*Bourque et al., 2018*) and to influence gene expression and they show a strong link to piRNAs (*Cullen and Schorn, 2021*; *Sun et al., 2017*). Consequently, we wondered whether species-specific insertions of TEs could induce the differential piRNA production in some piRNA clusters across the species featuring this study. This hypothesis was strongly

suggested from a previous study from our lab (Tanya Vavoury, personal communication) that showed that *Noct* gene -also named cluster *pi-Ccrn4l*- had increased piRNA production in *Mus musculus* strains with an IAP insertion compared to strains that lacked the IAP in this gene (*supp. figure 4*). To check this, we first looked at all the species-specific insertions of two major retrotransposon classes (i.e. LINE and LTR), as well as more specific subclasses (i.e. LINE1, ERVs and IAP). From all the piRNA clusters, few had species-specific insertions of TEs (*supp. figure 5A*), being endogenous retroviruses (ERVs) from group K (ERVK) -in all species- and IAP -in *Mus musculus*- the most common ones. In addition, only a subset of these produced more piRNA in the species with the TE insertion (*supp. figure 5B*).

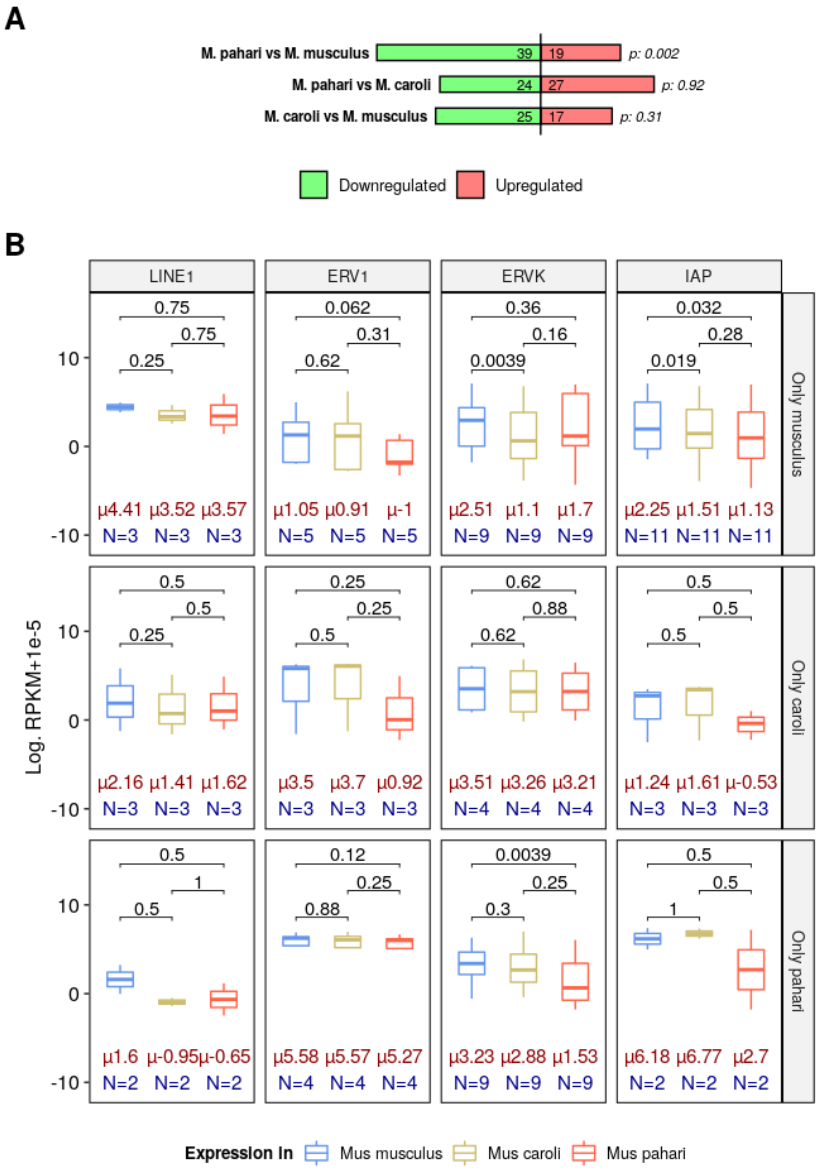


Figure 8. piRNA clusters from *Yu et al.* (2015). (A) Number of differentially expressed piRNA clusters (DEpiCs) in each contrast and `prop.test()` testing wether the proportion of upregulated or downregulated clusters is different from the expected 0.5. (B) Expression in log(RPKM+1e-5) of the clusters with species-specific insertions. Top strips indicate the inserted TE and right strips indicate which species the TE is inserted in. Text in blue indicates the number of observations in each boxplot and text in read shows the mean of the distribution. The nubmers on the boxplots are p-values from two-sided Wilcoxon signed-rank tests.

Then, to test whether there was a link between TE insertions and piRNA production, we performed Fisher's exact tests comparing the presence of species-specific TEs and the number of DEpiCs. With a threshold of |*log2FoldChange*| > *1*, none of the tests gave significant results (*table 1*). Nevertheless, further analyses revealed significant changes ($P_{Wilcoxon}$ < 0.05) in the expression distribution of clusters with *Mus musculus*-specific ERVKs and IAPs, as well as clusters with *Mus pahari*-specific ERVs (*figure 8B*). Clusters with higher piRNA production in species with specific TEs -which include *pi-Ccrn4l*- are shown in *supp. figure 5B*.

Table 1. Fisher's tests associating species-specific TE insertions with the differentially expressed piRNA clusters from *Yu et al. (2021)* in different contrasts

Contrast	TE	TE only in	Fisher's test p-value
PAH vs MUS	LINE	*Mus musculus*	0.532
CAR vs MUS		*Mus musculus*	1
CAR vs MUS		*Mus caroli*	0.173
PAH vs CAR		*Mus caroli*	1
PAH vs MUS		*Mus pahari*	0.133
PAH vs CAR		*Mus pahari*	0.543
PAH vs MUS	LTR	*Mus musculus*	1
CAR vs MUS		*Mus musculus*	1
CAR vs MUS		*Mus caroli*	1
PAH vs CAR		*Mus caroli*	1
PAH vs MUS		*Mus pahari*	1
PAH vs CAR		*Mus pahari*	0.665
PAH vs MUS	LINE 1	*Mus musculus*	0.555
CAR vs MUS		*Mus musculus*	1
CAR vs MUS		*Mus caroli*	1
PAH vs CAR		*Mus caroli*	0.552
PAH vs MUS		*Mus pahari*	0.133
PAH vs CAR		*Mus pahari*	0.543
PAH vs MUS	ERV1	*Mus musculus*	1
CAR vs MUS		*Mus musculus*	1
CAR vs MUS		*Mus caroli*	0.566
PAH vs CAR		*Mus caroli*	0.244
PAH vs MUS		*Mus pahari*	0.297
PAH vs CAR		*Mus pahari*	0.306
PAH vs MUS	ERVK	*Mus musculus*	0.726
CAR vs MUS		*Mus musculus*	0.701
CAR vs MUS		*Mus caroli*	1
PAH vs CAR		*Mus caroli*	0.306
PAH vs MUS		*Mus pahari*	0.726
PAH vs CAR		*Mus pahari*	1
PAH vs MUS	IAP	*Mus musculus*	0.214
CAR vs MUS		*Mus musculus*	1
CAR vs MUS		*Mus caroli*	1
PAH vs CAR		*Mus caroli*	0.244
PAH vs MUS		*Mus pahari*	1
PAH vs CAR		*Mus pahari*	0.543

To further investigate a possible link between transposable elements and differential piRNA production, we manually inspected several of the piRNA clusters with species-specific transposons. From all the clusters with IAP present specifically in *Mus musculus*, only *pi-Ccrn4l* and *pi-Phf20* were significantly higher expressed in *Mus*

musculus compared to both, *Mus caroli* and *Mus pahari* (*figure 9*). Interestingly, these clusters had an intronic sense IAP insertion and the sequencing reads mapped across all the region (*supp. figure 6A*), whereas the rest of the IAP-containing non-differentially expressed clusters contained an antisense IAP and they were mainly intergenic or genic clusters with reads mapping to the 3'UTR (*supp. figure 6B*). In the case of *Mus caroli*-specific IAP insertions, all of them inserted antisense to a non-differentially expressed cluster.

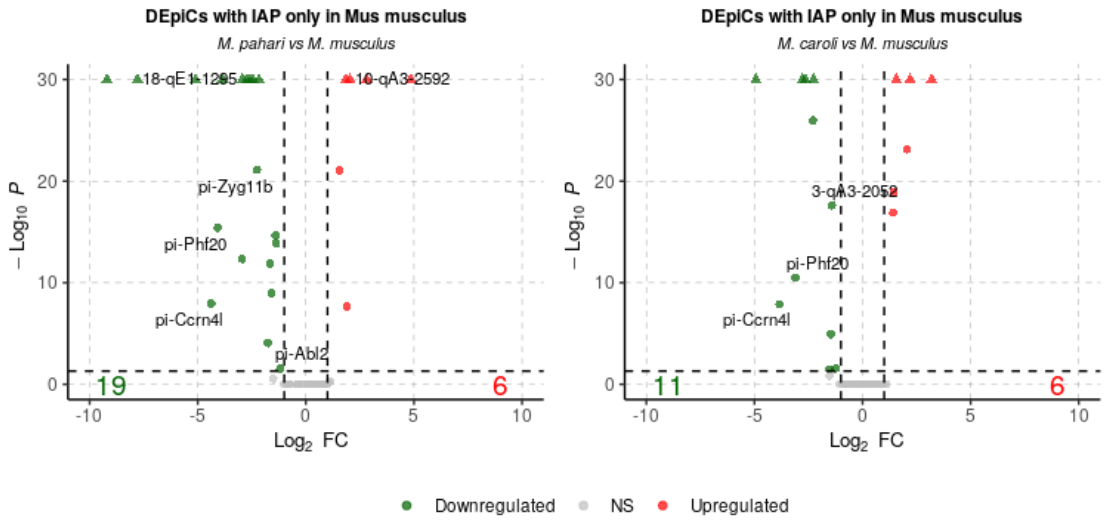


Figure 9. Volcano plots showing the differentially expressed clusters from *Yu et al. (2021)* in the contrasts "*Mus pahari* vs. *Mus musculus*" and *"Mus caroli* vs. *Mus musculus"*.

Altogether, these results suggest that, although TE insertions cannot be established as a common feature that triggers piRNA production in all piRNA clusters, their influence cannot be ruled out in the case of some clusters -like *pi-Ccrn4l*- and more variables may be taken into account. For example, traits like the stradedness of the insertion, whether it is exonic or intronic and whether it is in a gene or in an intergenic region could be considered.

2.4. *De novo* piRNA clusters

2.4.1. Prediction of *de novo* piRNA clusters

The *Yuetal* clusters were defined in *Mus musculus* using sRNA-seq, RNA-seq and ChIP-seq from postnatal testes in different stages of development. Since the present work is focussed on the comparison of *Mus musculus* with *Mus caroli* and *Mus pahari*, we decided to use `proTRAC` (*Rosenkranz and Zischler, 2012*) to predict *de novo* piRNA clusters (hereafter, *de novo* clusters or predicted clusters).

Starting from the adaptor-trimmed, quality-filtered reads, we followed the `proTRAC` workflow (*supp. figure 7*) to predict piRNA clusters in each sample. Only predicted clusters present in both replicates for each species were retained and further filtered to remove those overlapping TEs by 80%, remaining 330, 216 and 254 clusters for *Mus musculus*, *Mus caroli* and *Mus pahari*, respectively (*figure 10A*). Finally, clusters were sorted by chromosome and start position, and named using the species name and the corresponding number in the sorted list (e.g. *musculus1*, *caroli205*, *pahari53*).

These *de novo* clusters had similar length distribution between the different species (*figure 10B*), but less than one third overlapped with *Yuetal* clusters (*figures 10C, 10D and 10E*). Interestingly, from all the Yuetal clusters overlapping with de novo clusters, most (55) were common in the three species, while the rest were mainly present in *Mus musculus* (14) and in the intersection between *Mus musculus* and *Mus caroli* (15) or *Mus pahari* (12) (*figure 11A*). The *Yuetal* clusters intersecting with *de novo* clusters in all the species are mainly pachytene, while the others are mostly not pachytene (*figure 11B*). Moreover, the expression of Yuetal clusters overlapping clusters predicted in all three species was much higher than the ones overlapping with clusters predicted in some -not all- species or the ones not overlapping with any predicted cluster (*figure 11C*). Since proTRAC uses the reads to predict clusters, this agrees with our previous results linking the expression level and the expression conservation of piRNA clusters (*figures 6* and *7*).

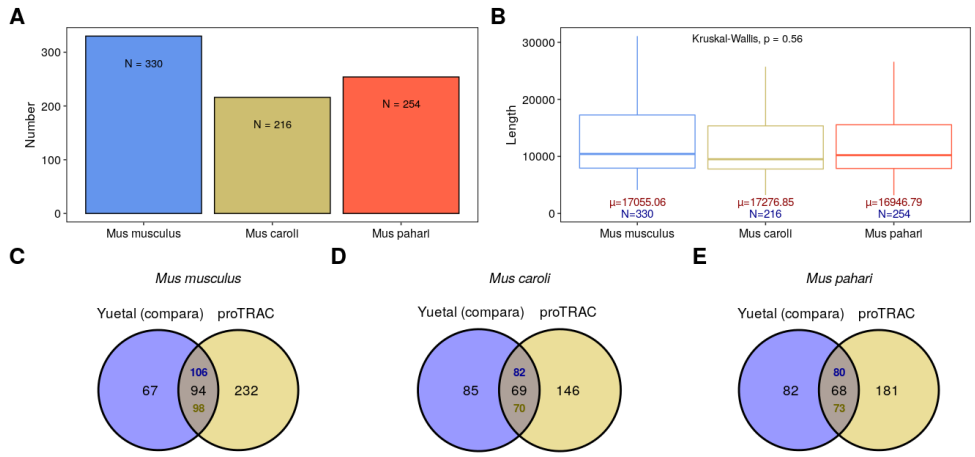


Figure 10. (A) Number of *de novo* clusters predicted in each species. (B) Length distribution of *de novo* clusters predicted in each species. Text in red indicate the mean of the distribution and text in blue indicates the number of observations in each boxplot. (C, D and E) Intersection of *de novo* clusters predicted in each species with clusters from *Yu et al.* (2021) obtained from ENSEMBL Compara: (C) *Mus musculus*; (D) *Mus caroli*; (E) *Mus pahari*.

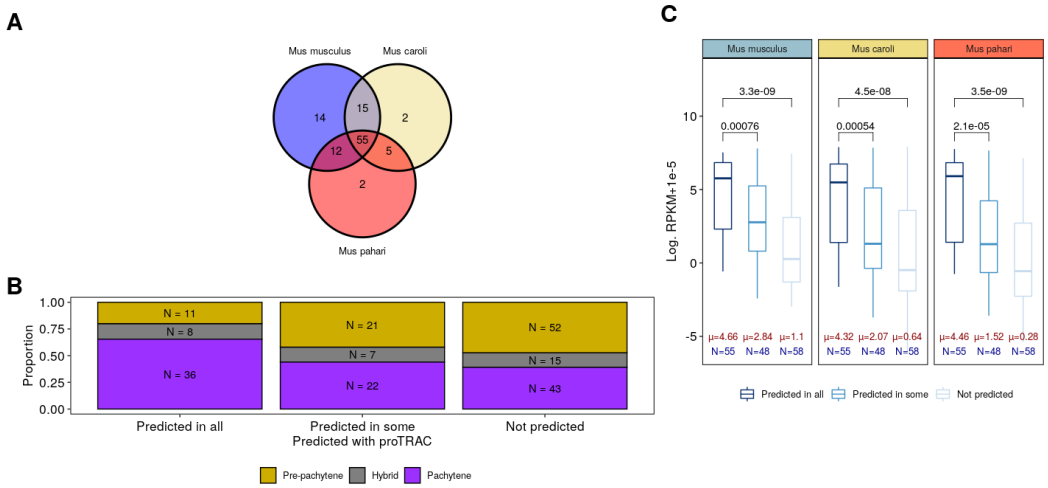


Figure 11. Clusters from *Yu et al. (2021)* intersecting with *de novo* clusters predicted in each species: (A) intersection between all the *Mus* species; (B) number and proportion of classes of *Yuetal* clusters intersecting with *de novo* clusters predicted in all, some or zero species; (C) Expression of *Yuetal* clusters intersecting with *de novo* clusters predicted in all, some or zero species. Expression is shown as the logarithm of RPKM+1e-5. We added 1e-5 as a pseudocount to avoid infinite values after the logarithmic transformation.

2.4.2. Obstacles in *de novo* prediction

Even though `proTRAC` is a useful tool for the study of piRNAs and piRNA clusters, we must note that it may have some problems. First, *de novo* clusters that are divergently transcribed from a bidirectional promoter are predicted as one bidirectional cluster (i.e. strand value is "." instead of "+" or "-") (*figure 3B*). This may generate complications in subsequent analysis where one needs to know the strand of each part of the cluster or where its promoter is. Secondly, although it has been shown that piRNAs arising from 3'UTRs come from full-length precursor transcripts (*Sun et al., 2021*), many of the resulting genic piRNA clusters comprised only the last part of the gene (i.e. 3'UTRs) (*figure 3A*). In a nutshell, if we need to look at putatively relevant sequence differences between species, but they are not present in the 3'UTR of the genes, finding them will be much harder.

2.4.3. Genic and intergenic *de novo* piRNA clusters

As an approximation to define *de novo* genic and intergenic clusters, we intersected our predicted clusters with protein coding genes present in the three species (i.e. orthologs). For each overlap, we required the same strandedness and a minimum overlap of 25%. Therefore, *de novo* clusters fulfilling these requirements were considered putative genic clusters, whereas the ones not meeting them were defined as putative intergenic clusters.

Approximately 50% of the *de novo* clusters overlapped protein-coding genes (i.e. putative genic clusters) whereas the rest lied in non-coding regions (i.e. putative intergenic clusters) (*figure 12A*). Among these genes, few were common across all - henceforth referred to as genic *Mus*-conserved piRNA clusters (MCpiCs)- or between two species, while most genes were species-specific (*figure 12B*).

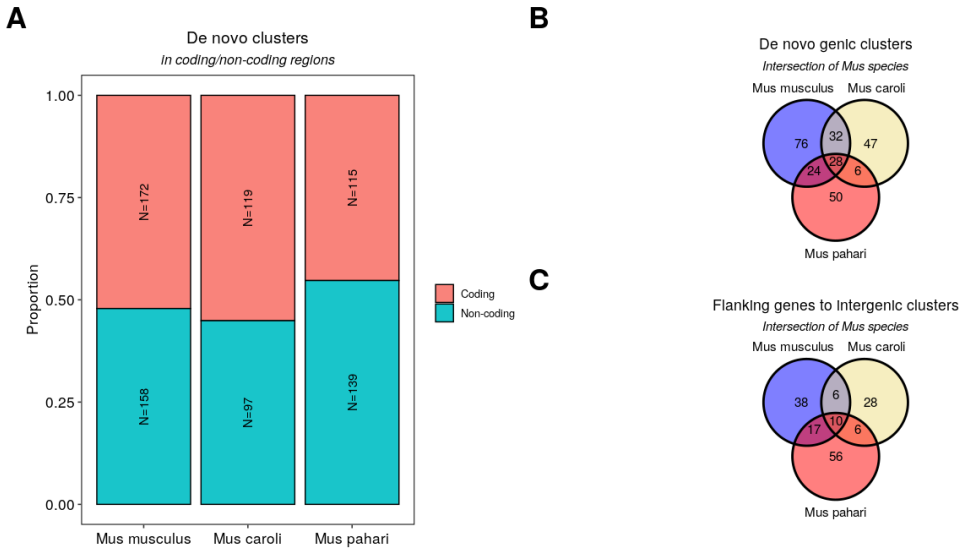


Figure 12. *De novo* clusters intersecting with protein-coding genes in each species: (A) number and proportion of clusters intersecting or not with protein-coding genes; (B) intersection of genes overlapping with *de novo* clusters in each species; (D) intersection of flanking genes to *de novo* intergenic clusters predicted in each species.

Regarding these genic *de novo* clusters (*figure 12B*), it is interesting to compare them with the 56 genic Eutherian-conserved piRNA clusters (hereafter, ECpiCs) defined by *Chirn et al.* (2015). Interestingly, only half of the 28 genic MCpiCs were present in the list of genic ECpiCs (*supp. figure 8A*), comprising genes like *Ago2* -involved in the miRNA-mediated gene silencing- or *Nr2c2* -potentially involved in gene regulation during spermatogenesis-. As expected, most of the species-specific genic clusters were not conserved within the Eutherian clade, although neither half of the genic MCpiCs were Eutherian-conserved (*supp. figure 8B*), including genes such as *Ago3* -involved in the miRNA-mediated gene silencing- or *Strbp* -involved in the spermatogenesis and sperm physiology-.

Then, for each *de novo* intergenic cluster, we searched the two flanking protein-coding genes. This is, the closest gene upstream and the closest gene downstream of the cluster. As with the genic clusters, only few pairs of flanking genes were common across all or between two species, while most were species-specific (*figure 12C*).

Although this approach might not exactly estimate the number of *de novo* piRNA clusters and their conservation, it is a first approximation that strongly suggests high differences in piRNA production across the three *Mus species*. Furthermore, although many *de novo* genic clusters are conserved also in Eutheria, most of them are not (*supp. figure 8*) indicating that, even those clusters conserved in our closely related species, are not conserved in other more distant species.

2.4.4. *De novo* genic clusters and transposable elements

As a result of the *de novo* cluster differences across different species, we asked ourselves whether insertions of TEs could select a gene to become a piRNA cluster in a species-specific manner. Using the coordinates of the genes overlapping with predicted clusters, from the 71 *Mus musculus*-specific genic clusters non-conserved in Eutheria (*supp. figure 8B*), only 4 had a specific IAP insertion (*Stxbp4*, *Zfp69*, *Abhd2*, *Ccdc15*). This number decreased to 2 in the case of the 46 *Mus caroli*-specific genic clusters (*Gm28051*, *Bicdl1*) and to 1 for the 49 *Mus pahari*-specific genic clusters (*Hjurp*). Looking at ERVKs, there were 4, 2 and 1 species-specific insertions for *Mus musculus*, *Mus caroli* and *Mus pahari*, respectively. This low number of species-specific IAP insertions, as well as ERVKs, suggest that new transposons are not a general trait by which a gene can be selected to enter the piRNA biogenesis pathway, but they can still be the trigger for some of these genes. Interestingly, all species-specific IAPs had the same characteristics mentioned in the *section 2.3.3*, if the insertion was antisense to the gene, the gene produced piRNAs mostly from the 3'UTR *(supp. figure 9A*), otherwise, the piRNAs were produced across all the gene, including introns (*supp. figure 9B*).

Concisely, these results suggest that IAP insertions are not responsible for all the species-specific clusters. However, their placement and their strandedness regarding the genes might be important because all tested genes with an intronic sense IAP insertion produces piRNAs all across the gene, whereas each studied gene with an antisense insertion generated piRNAs from the 3'UTR.

2.4.5. Conservation of expression (II)

In *Yuetal* clusters, the higher expression of a cluster in *Mus musculus* also implied greater expression of its orthologous regions in *Mus caroli* and *Mus pahari* (*figures 6 and 7*), implying that expression level and conservation of the expression are correlated.

To test this in the *de novo* predicted clusters, we used ENSEMBL Compara Perl API to retrieve their orthologous regions and the conserved blocks. This is, we obtained the orthologs of the *de novo* clusters predicted in *Mus musculus*, *Mus caroli* and *Mus pahari* and estimated the piRNA production in those clusters and their orthologs. As before, only those clusters without duplications, present in the three species and without great changes in gene length (i.e. 10% of the mean length across the species) were retained for the expression analyses.

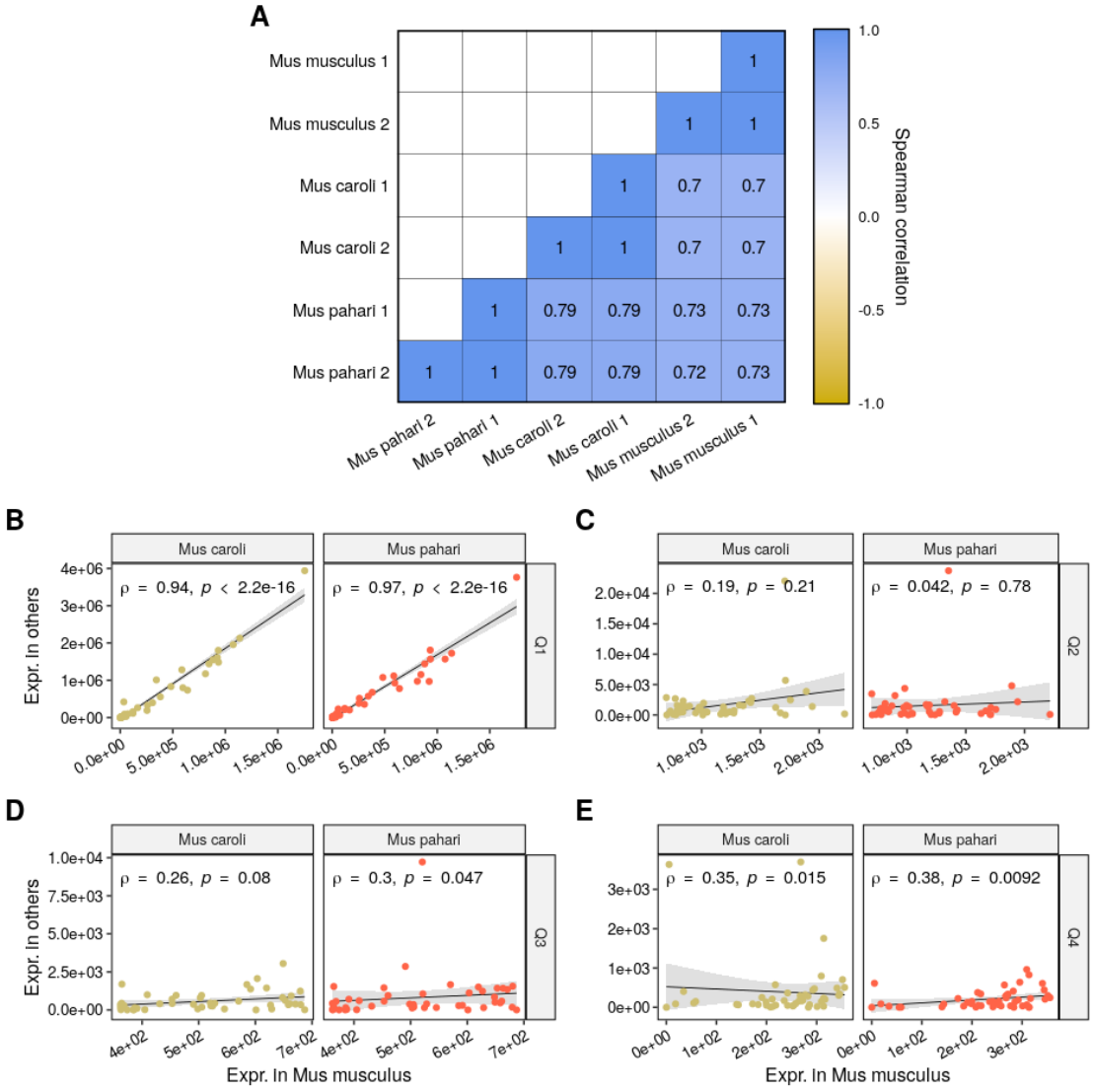


Figure 13. Spearman correlations of the *de novo* clusters predicted in *Mus caroli*: (A) correlation plot with all clusters present in all samples; (B) scatter plots and correlations the top 25% (Q1) expressed clusters in *Mus caroli*; (C) scatter plots and correlations of the clusters in the Q2; (D) scatter plots and correlations of the clusters in the Q3; (E) scatter plots and correlations of the clusters in the Q4. Expression is shown in DESeq2-normalized counts.

We computed the Spearman correlation between the expression of the *de novo* clusters predicted in one species (e.g. *Mus musculus*) and the expression of the orthologs in the other species (e.g. *Mus caroli* and *Mus pahari*). For the clusters predicted in *Mus musculus*, although the overall correlation (*figure 13A*) was not as high as in *Yuetal* clusters (*figure 6B*), the top expressed *de novo* clusters -first quartile (Q1)- were highly correlated across all species (*figure 13B*), while clusters in the second (Q2), third (Q3) and fourth (Q4) quartiles had much smaller correlations (*figures 13C, 13D* and *13E*). Similar results were obtained for the clusters predicted in *Mus caroli* (*supp. figure 10*) and in *Mus pahari* (*supp. figure 11*).

Together, alongside the analyses done with the *Yuetal* clusters, these results indicate that expression level strongly correlates with conservation.

2.4.6. piRNA expression and transposable elements (II)

Using the *de novo* clusters and their orthologs, we performed DE analyses. As in *Section 2.3.3*, to call a cluster as differentially expressed (DEpiC), thresholds of *adjusted p-value < 0.05* and *|log2FoldChange| > 1* had to be satisfied. Like *Yuetal* clusters (*figure 8A*), *de novo* clusters showed high numbers of DEpiCs (*figure 14*), with clear trends to higher expression in the species where the clusters were predicted. This may cause difficulties in the identification of differential traits like transposon insertions that may induce variations in piRNA production in different species. Moved by our hypothesis that linked IAPs to expression of piRNA clusters, we looked at species specific TE insertions in *de novo* clusters, extending their limits by 10kb upstream and downstream.

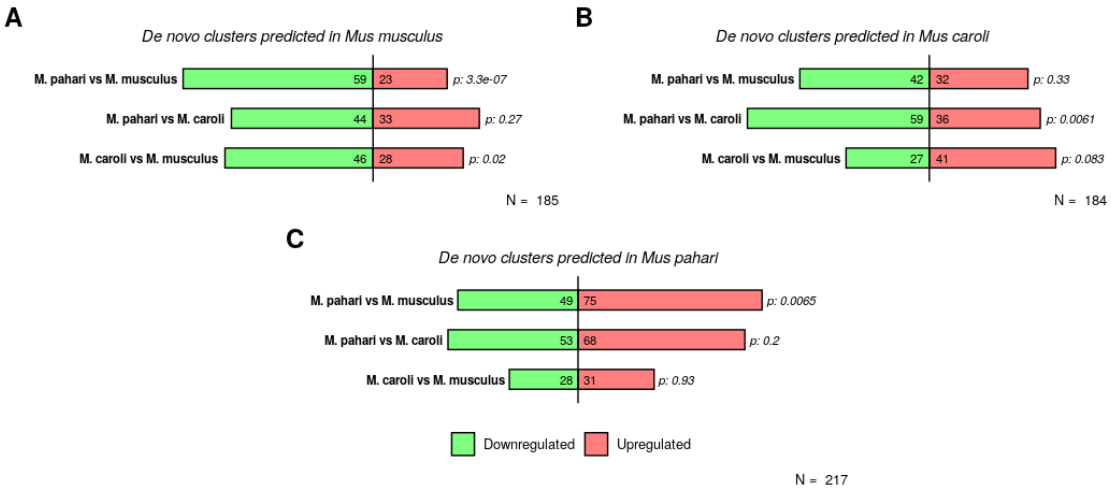


Figure 14. Number of differentially expressed piRNA clusters in each contrast and each *de novo* cluster annotation: (A) *de novo* clusters predicted in *Mus musculus*; (B) *de novo* clusters predicted in *Mus caroli*; (C) *de novo* clusters predicted in *Mus pahari*. `prop.test()` was done to test whether the proportion of upregulated or downregulated clusters was different from the expected proportion 0.5.

In the *de novo* clusters predicted in *Mus musculus*, ERVK, but also LINE1 and IAP, had many species-specific insertions (*figure 15A*), although fewer clusters were higher expressed in the species with the transposon insertion (*figure 15B*). The *de novo* clusters predicted in *Mus caroli* (*supp. figure 12*) and in *Mus pahari* (*supp. figure 13*) also had a high number of species-specific TEs, with *Mus musculus* bearing most of the ERVK and IAP insertions. *Mus musculus* tended to have more species-specific

ERVKs and IAPs than *Mus caroli* and *Mus pahari* likely because, globally, *Mus musculus* has many more ERVKs and IAPs than *Mus caroli* (~5-fold) and *Mus pahari* (~10-fold) (not shown).

Table 2. Fisher's test associating species-specific TE insertions with the differentially expressed piRNA clusters in different contrasts. piRNA clusters are *de novo* clusters predicted in *Mus musculus*.

Contrast	TE	Only in	Fisher'sP
PAH vs *MUS*	LINE	*Mus musculus*	1
CAR vs *MUS*		*Mus musculus*	1
CAR vs *MUS*		*Mus caroli*	1
PAH vs *CAR*		*Mus caroli*	0.53
PAH vs *MUS*		*Mus pahari*	1
PAH vs *CAR*		*Mus pahari*	0.311
PAH vs *MUS*	LTR	*Mus musculus*	1
CAR vs *MUS*		*Mus musculus*	0.552
CAR vs *MUS*		*Mus caroli*	1
PAH vs *CAR*		*Mus caroli*	1
PAH vs *MUS*		*Mus pahari*	0.4
PAH vs *CAR*		*Mus pahari*	1
PAH vs *MUS*	LINE 1	*Mus musculus*	1
CAR vs *MUS*		*Mus musculus*	1
CAR vs *MUS*		*Mus caroli*	0.346
PAH vs *CAR*		*Mus caroli*	0.314
PAH vs *MUS*		*Mus pahari*	1
PAH vs *CAR*		*Mus pahari*	**0.058***
PAH vs *MUS*	ERV1	*Mus musculus*	0.565
CAR vs *MUS*		*Mus musculus*	**0.04****
CAR vs *MUS*		*Mus caroli*	1
PAH vs CAR		*Mus caroli*	1
PAH vs MUS		*Mus pahari*	**0.038****
PAH vs CAR		*Mus pahari*	0.38
PAH vs MUS	ERVK	*Mus musculus*	0.471
CAR vs MUS		*Mus musculus*	1
CAR vs MUS		*Mus caroli*	0.418
PAH vs CAR		*Mus caroli*	0.38
PAH vs MUS		*Mus pahari*	0.147
PAH vs CAR		*Mus pahari*	0.261
PAH vs MUS	IAP	*Mus musculus*	1
CAR vs MUS		*Mus musculus*	0.335
CAR vs MUS		*Mus caroli*	0.166
PAH vs CAR		*Mus caroli*	1
PAH vs MUS		*Mus pahari*	0.517
PAH vs CAR		*Mus pahari*	1

To check whether TE insertions in a *de novo* cluster (±10kb) could increase the chance of differential expression, we performed Fisher's test with all the combinations of contrasts and species-specific transposons. As resulted with *Yuetal* clusters, few tests gave a significant result ($P_{Fisher} < 0.05$). In the *de novo* clusters predicted in *Mus musculus*, *Mus musculus*- and *Mus pahari*-specific ERV1 insertions were significantly associated with differential expression in "CAR vs MUS" ($p_{Fisher} = 0.04$) and "PAH vs

MUS" (p_{Fisher} = 0.038), respectively (*table 2*). Instead, Fisher's tests in the clusters predicted in *Mus caroli* (*supp. table3*) did not give any significant result. Finally, in the clusters predicted in *Mus pahari*, LINE1 insertions in *Mus pahari* increased the chance of differential expression in "PAH vs MUS" (p_{Fisher} = 0.028) (*supp. table4*). These results point out that, in some cases, TE insertions may be related to differential piRNA production.

Additional analyses focussing on the expression revealed significant changes ($p_{Wilcoxon}$ < 0.05) in the expression of clusters with species-specific transposons. For *de novo* clusters predicted in *Mus musculus*, clusters with *Mus musculus*-specific ERVK and IAP insertions were significantly higher expressed than *Mus caroli* (ERVK and IAP) and *Mus pahari* (only ERVK) (*supp. figure 14A*). In the case of *de novo* clusters predicted in *Mus caroli*, a significant increase in expression was observed when IAP inserted in *Mus caroli*. However, clusters with *Mus pahari*-specific ERVK insertions were also more expressed in *Mus caroli* (*supp. figure 15A*), suggesting that there may be a bias regarding the species of origin. Finally, all the significant expression changes detected in the clusters predicted in *Mus pahari* were always in favour of this species, regardless of in which species the TE was inserted (*supp. figure 16A*).

Finally, we manually inspected several of the *de novo* clusters bearing species-specific insertions of TEs. To restrict the analysis, we looked for those with higher expression in the species with the transposon (*figures 15B*, *S12B and S13B*), especially in those with great changes in expression. For example, *musculus152* -an intergenic cluster carrying a *Mus musculus*-specific IAP insertion- is highly expressed in *Mus musculus* but its expression is zero in *Mus caroli* and *Mus pahari*. Moreover, most reads mapping in *musculus152* mapped downstream of the IAP insertion (*supp. figure 17A*). Also, *caroli71* -an intergenic cluster with an *Mus caroli*-specific antisense IAP- had much higher expression in *Mus caroli* compared to *Mus musculus* and *Mus pahari*, and most reads mapped downstream of the IAP insertion (*supp. figure 17B*). Although this is not a constitutive proof that establishes a definitive link between piRNA production and IAP insertions, it serves as an example of piRNA clusters potentially linked to IAP insertions. In this case, the piRNA clusters were intergenic, the insertion antisense and the piRNAs started to be generated downstream of the IAP.

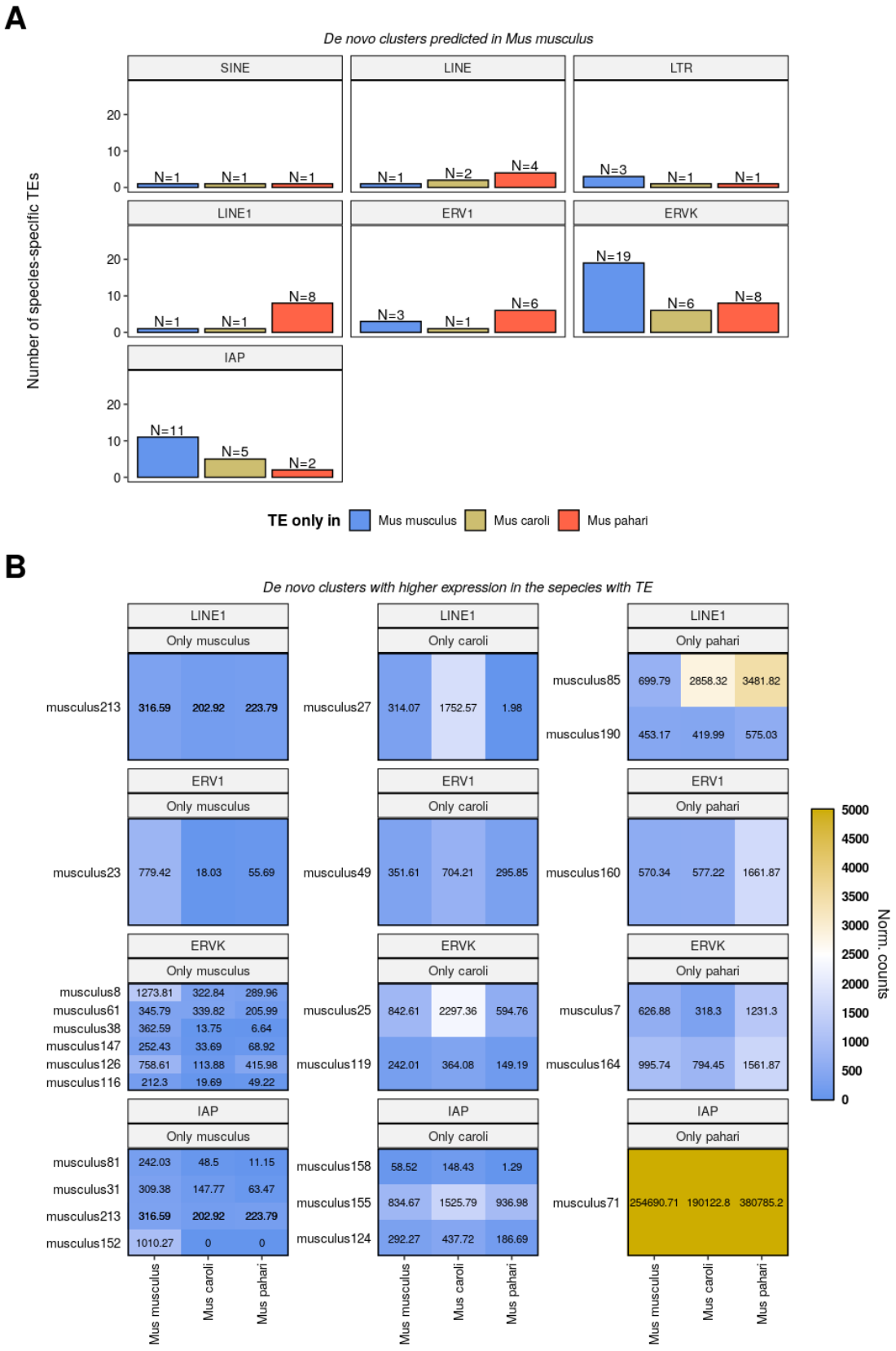


Figure 15. Clusters with species-specific transposon insertions (*de novo* clusters predicted in *Mus musculus*): (A) number of clusters with species-specific insertions of different TEs; (B) expression heatmaps of the clusters with higher expression in the species with the TE. Expression is shown in DESeq2-normalized counts.

2.5. Discussion

piRNAs are the main mechanism of TE silencing in the germline of most animals and control the expression of some genes. Here, we have performed an extensive study of the small RNA expression in the male adult testis of three closely related species, providing the first small RNA datasets from testis of *Mus caroli* and *Mus pahari*. Our quality controls show that small RNAs sequenced of testis from the three species are enriched in piRNAs, yet other small RNAs are present in our data. The first exploratory analysis shows that, even though the small RNA expression across the species is completely different, piRNA biogenesis mechanisms are similar, since proportions of 1U and 10A are comparable across all samples. Read length is around 30 nucleotides and there is a significant ping-pong signature in all samples. All this suggests that most piRNAs bind to MIWI, which is the most expressed in the pachytene stage (*Ernst et al., 2017*), and that both primary and secondary pathways are taking place in adult testes of the three species.

Despite all the similarities in the mechanisms of piRNA biogenesis -which are evolutionary conserved-, many clusters defined by *Yu et al.* (2019) in *Mus musculus* do not have orthology in *Mus caroli* or *Mus pahari*. Still, most *Yuetal* clusters have orthologous regions in the three studied species and approximately a third of these were differentially expressed in terms of piRNA production, showing a clear bias towards higher expression in *Mus musculus.* Although this bias may appear because *Yuetal* clusters were defined in *Mus musculus*, the fact that pachytene clusters were not significantly higher expressed in *Mus musculus* suggests that they may have greater conservation -in terms of expression- across species. On the other hand, *de novo* clusters predicted with `proTRAC` were extremely useful to study the piRNA production in each species and the differences between them. Since our *de novo* clusters were based only in data from testis from one developmental stage -adulthood- they greatly differed from *Yuetal* clusters -which were defined using data from several developmental stages-.

Nevertheless, most *Yuetal* clusters intersected *de novo* clusters, with pachytene clusters being enriched among those intersections. In addition, *Yuetal* clusters overlapping with clusters predicted in all species were significantly more expressed than other clusters, suggesting that either pachytene clusters or high expressed clusters -or both- tend to be conserved. In congruence with that, the expression of the top 25% expressed clusters (Q1) highly correlated with the expression with their orthologous regions, indicating that expression level in a species is associated with conserved expression in other species. Indeed, other studies focussing on distantly related Eutherian species (*Chirn et al., 2015*) have already shown that the most Eutherian-conserved piRNA clusters tend to yield more piRNAs than others. Considering this, it should be expected that greatest fold-changes appear in the less expressed clusters.

As an approach to define *de novo* clusters as genic or intergenic, we intersected them with protein-coding genes. Assuming that our approach correctly estimates genic and intergenic clusters, approximately half of the *de novo* clusters were genic, and from these, few were conserved across all species. Instead, most were species-specific, revealing great differences in piRNA production across the *Mus* genus that must be

further explored to find differential traits that can trigger the genes to enter the piRNA biogenesis pathway in a species but not in others. As a consequence of these great differences, making statistical associations between species-specific traits and differential piRNA production may be challenging, especially if the traits are linked with the expression of a small subset of piRNA clusters.

To the best of our knowledge, many factors could trigger the piRNA production from one locus. The current model suggests that these factors should be held within the sequence of the piRNA cluster rather than the genomic context, since piRNA clusters introduced in ectopic locations are still expressed (*Muerdter et al., 2012; Goh et al., 2015*). For example, pachytene clusters -mainly intergenic- require the binding of *A-Myb* transcription factor to their promoter, although it must not be the only characteristic that differentiates pachytene clusters from other loci, since *A-Myb* binds the Piwi protein genes, among others (*Li et al., 2013*). Moreover, expression of the precursor transcript alone is not sufficient for piRNA production, since distinct species have shown to express a gene which is processed into piRNA in one species, but it is only transcribed into the other (*Chirn et al., 2015*). Therefore, if there is a trigger that selects a transcript to be processed into piRNA it is likely within the cluster itself and it should exert its influence at post-transcriptional level.

Among all these factors, we decided to study species-specific insertions of transposons, more specifically retrotransposons such as LINEs and LTRs, and some of their respective subclasses. The link between TEs and differential piRNA biogenesis is strongly suggested by several studies. For instance, in chicken, infection with the AVL -an ERV- induced the production of piRNAs from pre-existing loci (*Sun et al., 2017*). Moreover, in previous studies from our lab, IAP -an ERV- activated piRNA production from the *Noct* gene in *Mus musculus* strains bearing this insertion, whereas other strains were not expressing piRNA from *Noct* and others (Tanya Vavouri, personal communication).

In spite of the strong association obtained in our previous work, we could not find significant (*Fisher's test, $p < 0.05$*) links between insertion of transposons and differential expression of piRNA clusters. Although this does not completely rule out TEs as potential elements that originate new piRNA clusters, they do not appear to be a general trigger for piRNA biogenesis. Nevertheless, *Noct* and *Phf20* -which were triggered for piRNA production upon the insertion of IAP in our previous study- are differentially expressed in our data and share differential traits with other piRNA clusters that also have a specific IAP insertion. In these two genes, IAP was inserted in *Mus musculus* sense to the gene and in an intronic region. Also, their expression in *Mus caroli* and *Mus pahari* was near to zero, something that should be expected if the mentioned insertion was the trigger to originate the piRNA cluster. Unfortunately, *Noct* and *Phf20* were not predicted in the *de novo* clusters, probably because they are pre-pachytene clusters and that our data coming from testis is enriched in cells on the pachytene stage. Nevertheless, some *de novo* intergenic clusters showed differential expression and antisense IAP insertions. Also, piRNAs mapped downstream of the IAP.

This hints that a possible mechanism in which a new IAP insertion could induce the production of piRNAs from a locus is by serving as target for initiator piRNA to trigger

the phased piRNA biogenesis. *Noct* and *Phf20* have reads mapped upstream of the IAP insertion and hence, phased biogenesis starting in the IAP should not be the main mechanism explaining differential piRNA expression for these genes. However, the de novo intergenic clusters we have used as examples are consistent with this model of IAP being target for the initiator of phased biogenesis. Therefore, further analyses regarding the effect of IAP -and other- transposons in genic and intergenic clusters -or in different classes- should be done to confirm this. Furthermore, sense and genomic location for the insertion should also be considered. Alternatively, IAP could interact with the splicing machinery (*Concepcion et al., 2009*), something that might explain why some piRNAs map in the introns of genes if the precursor transcript tends to be a full-processed transcript, or it could interact with the transporting machinery that places the precursor to the mitochondrial outer membrane.

Finally, although we have not tested it, an insertion of species-specific transposon could cause a cluster to disappear, either because there is a deleterious mutation that impedes the transcription of the gene or because it somehow blocks its entrance to piRNA biogenesis pathway.

In conclusion, mammalian piRNAs are essential for the defence against transposons as well as for other biological functions in germline. Nevertheless, their conserved functions confront the fact that piRNA clusters present a lot of differences across different species. Little is known about how piRNA clusters originate and evolve, or what drives differential piRNA production in different organisms. Nonetheless, although there are cases suggesting that insertions of endogenous retrovirus are contribution to the evolution of piRNAs and piRNA clusters, there must be additional, important sequence signals that drive evolution of piRNAs that we are currently missing.

3. Conclusions

This section is focussed on the conclusions related to the project and its realization, as well as the approaches we followed. For the discussion on the analyses and the corresponding conclusions, see *Section 2.6. Discussion*.

3.1. Take home messages

- piRNAs are difficult to study due to their repetitive sequence and small size.
- Differential expression analyses with multiple species are challenging due to length differences and non-conserved regions in the features to be compared.
- The ENSEMBL Compara Perl API is an invaluable tool to retrieve the multiple sequence alignment of genomic regions in different organisms.
- Retrieving the conserved blocks of the multiple alignment from ENSEMBL Compara mitigates the problems with gene length and non-conserved regions.

3.2. Goal achievement

All the initial goals were achieved. We checked the quality of our data, retrieved orthologous regions with minimal gene differences for differential expression analysis and tested a potential dependence of piRNA clusters and transposon insertions.

3.3. Planning and methodology

We followed the initial planning in a strict way, leaving some room for non-planned analyses and problems we could face. Regarding the methodology, the most challenging issue we faced in this project lies in the nature of comparing multiple species: differential expression analysis assumes that there are no differences in the length of the features being compared, but this is not true for distinct species. We considered the use of `liftOver` or the conserved blocks from ENSEMBL Compara to obtain orthologous regions of the annotated piRNA clusters. Although the latter method allowed to convert fewer regions, it minimized the length differences, and we chose to use it. About the rest of the analyses and tools, we used R and Bash custom scripts and functions that did not cause any major problem.

3.4. Lines of work to be explored

Some of the lines of work that have not been possible to carry out, but they are worth exploring are related to:

- Study whether the sense, relative position (i.e. intronic/exonic), distance (i.e. overlapping, upstream, downstream), number and the size of a TE insertion, specifically IAP, can influence the evolution of piRNAs.
- Check whether the class (i.e. pre-pachytene, hybrid or pachytene), promoter directionality or genomic context (i.e. genic/intergenic).
- Try other tools for piRNA and piRNA cluster identification, such as `PILFER` (*Ray et al., 2018*) and others. This may include benchmarking different tools.

4. Glossary

- **Pachytene**: third stage of the prophase I of the meiosis.
- **TE**: transposable element, transposon.
- **LINE**: long interspersed nuclear element.
- **LTR**: long terminal repeat.
- **ERV**: endogenous retrovirus.
- **IAP**: intracisternal A-particle retrotransposon.
- **Piwi**: p-element induced wimpy testis.
- **piRNA**: Piwi-interacting RNAs.
- **Precursor transcript**: transcript that is going to be processed into piRNAs.
- **DE**: differential expression.
- **DEpiC**: differentially expressed piRNA cluster.
- **ECpiC**: Eutherian-conserved piRNA cluster (*Chirn et al.*, *2015*).
- **MCpiC**: *Mus*-conserved piRNA cluster, as an analogous term to ECpiC.
- ***Yuetal* clusters**: piRNA clusters defined by *Yu et al.,* (2021).

5. Bibliography

Assis, R., & Kondrashov, A. S. (2009). Rapid repetitive element-mediated expansion of piRNA clusters in mammalian evolution. Proceedings of the National Academy of Sciences, 106(17), 7079–7082. https://doi.org/10.1073/pnas.0900523106

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements. Genome Biology, 19(1). https://doi.org/10.1186/s13059-018-1577-z

Carmell, M. A., Girard, A., van de Kant, H. J. G., Bourc'his, D., Bestor, T. H., de Rooij, D. G., & Hannon, G. J. (2007). MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. Developmental Cell, 12(4), 503–514. https://doi.org/10.1016/j.devcel.2007.03.001

Chirn, G., Rahman, R., Sytnikova, Y. A., Matts, J. A., Zeng, M., Gerlach, D., Yu, M., Berger, B., Naramura, M., Kile, B. T., & Lau, N. C. (2015). Conserved piRNA Expression from a Distinct Set of piRNA Cluster Loci in Eutherian Mammals. PLOS Genetics, 11(11), e1005652. https://doi.org/10.1371/journal.pgen.1005652

Choi, H., Wang, Z., & Dean, J. (2021). Sperm acrosome overgrowth and infertility in mice lacking chromosome 18 pachytene piRNA. PLOS Genetics, 17(4), e1009485. https://doi.org/10.1371/journal.pgen.1009485

Concepcion, D., Flores-García, L., & Hamilton, B. A. (2009). Multipotent genetic suppression of retrotransposon-induced mutations by nxf1 through fine-tuning of alternative splicing. PLoS Genetics, 5(5), e1000484. https://doi.org/10.1371/journal.pgen.1000484

Cullen, H., & Schorn, A. J. (2020). Endogenous retroviruses walk a fine line between priming and silencing. Viruses, 12(8), 792. https://doi.org/10.3390/v12080792

De Fazio, S., Bartonicek, N., Di Giacomo, M., Abreu-Goodger, C., Sankar, A., Funaya, C., Antony, C., Moreira, P. N., Enright, A. J., & O'Carroll, D. (2011). The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements. Nature, 480(7376), 259–263. https://doi.org/10.1038/nature10547

Di Giacomo, M., Comazzetto, S., Saini, H., De Fazio, S., Carrieri, C., Morgan, M., Vasiliauskaite, L., Benes, V., Enright, A. J., & O'Carroll, D. (2013). Multiple Epigenetic Mechanisms and the piRNA Pathway Enforce LINE1 Silencing during Adult Spermatogenesis. Molecular Cell, 50(4), 601–608. https://doi.org/10.1016/j.molcel.2013.04.026

Ding, D., Liu, J., Dong, K., Midic, U., Hess, R. A., Xie, H., Demireva, E. Y., & Chen, C. (2017). PNLDC1 is essential for piRNA 3′ end trimming and transposon silencing

during spermatogenesis in mice. Nature Communications, 8(1). https://doi.org/10.1038/s41467-017-00854-4

Ernst, C., Odom, D. T., & Kutter, C. (2017). The emergence of piRNAs against transposon invasion to preserve mammalian genome integrity. Nature Communications, 8(1). https://doi.org/10.1038/s41467-017-01049-7

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Girard, A., Sachidanandam, R., Hannon, G. J., & Carmell, M. A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature, 442(7099), 199–202. https://doi.org/10.1038/nature04917

Goh, W. S. S., Falciatori, I., Tam, O. H., Burgess, R., Meikar, O., Kotaja, N., Hammell, M., & Hannon, G. J. (2015). piRNA-directed cleavage of meiotic transcripts regulates spermatogenesis. Genes & Development, 29(10), 1032–1044. https://doi.org/10.1101/gad.260455.115

Gou, L.-T., Dai, P., Yang, J.-H., Xue, Y., Hu, Y.-P., Zhou, Y., Kang, J.-Y., Wang, X., Li, H., Hua, M.-M., Zhao, S., Hu, S.-D., Wu, L.-G., Shi, H.-J., Li, Y., Fu, X.-D., Qu, L.-H., Wang, E.-D., & Liu, M.-F. (2014). Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. Cell Research, 24(6), 680–700. https://doi.org/10.1038/cr.2014.41

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., … Flicek, P. (2020). Ensembl 2021. Nucleic Acids Research, 49(D1), D884–D891. https://doi.org/10.1093/nar/gkaa942

Karolchik, D. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Research, 32(90001), 493D – 496. https://doi.org/10.1093/nar/gkh103

Kojima-Kita, K., Kuramochi-Miyagawa, S., Nagamori, I., Ogonuki, N., Ogura, A., Hasuwa, H., Akazawa, T., Inoue, N., & Nakano, T. (2016). MIWI2 as an effector of DNA methylation and gene silencing in embryonic male germ cells. Cell Reports, 16(11), 2819–2828. https://doi.org/10.1016/j.celrep.2016.08.027

Kuff, E. L., Fewell, J. E., Lueders, K. K., DiPaolo, J. A., Amsbaugh, S. C., & Popescu, N. C. (1986). Chromosome distribution of intracisternal A-particle sequences in the Syrian hamster and mouse. Chromosoma, 93(3), 213–219. https://doi.org/10.1007/bf00292740

Kuramochi-Miyagawa, S., Kimura, T., Yomogida, K., Kuroiwa, A., Tadokoro, Y., Fujita, Y., Sato, M., Matsuda, Y., & Nakano, T. (2001). Two mouse piwi-related genes: Miwi and mili. Mechanisms of Development, 108(1–2), 121–133. https://doi.org/10.1016/s0925-4773(01)00499-3

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., Carey, V. (2013). Software for Computing and Annotating Genomic Ranges. PLoS Computational Biology, 9. doi: 10.1371/journal.pcbi.1003118

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, X. Z., Roy, C. K., Dong, X., Bolcun-Filas, E., Wang, J., Han, B. W., Xu, J., Moore, M. J., Schimenti, J. C., Weng, Z., & Zamore, P. D. (2013). An Ancient Transcription Factor Initiates the Burst of piRNA Production during Early Meiosis in Mouse Testes. Molecular Cell, 50(1), 67–81. https://doi.org/10.1016/j.molcel.2013.02.016

Liao, Y., Smyth, G. K., & Shi, W. (2013). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Lilue, J., Doran, A. G., Fiddes, I. T., Abrudan, M., Armstrong, J., Bennett, R., Chow, W., Collins, J., Collins, S., Czechanski, A., Danecek, P., Diekhans, M., Dolle, D.-D., Dunn, M., Durbin, R., Earl, D., Ferguson-Smith, A., Flicek, P., Flint, J., … Keane, T. M. (2018). Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. Nature Genetics, 50(11), 1574–1583. https://doi.org/10.1038/s41588-018-0223-8

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12). https://doi.org/10.1186/s13059-014-0550-8

Mager, D. L., & Stoye, J. P. (2015). Mammalian endogenous retroviruses. Microbiology Spectrum, 3(1). https://doi.org/10.1128/microbiolspec.mdna3-0009-2014

Manakov, S. A., Pezic, D., Marinov, G. K., Pastor, W. A., Sachidanandam, R., & Aravin, A. A. (2015). MIWI2 and MILI Have Differential Effects on piRNA Biogenesis and DNA Methylation. Cell Reports, 12(8), 1234–1243. https://doi.org/10.1016/j.celrep.2015.07.036

Marini, F., & Binder, H. (2019). pcaExplorer: An R/Bioconductor package for interacting with RNA-seq principal components. BMC Bioinformatics, 20(1). https://doi.org/10.1186/s12859-019-2879-1

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal, 17(1), 10. https://doi.org/10.14806/ej.17.1.200

Muerdter, F., Olovnikov, I., Molaro, A., Rozhkov, N. V., Czech, B., Gordon, A., Hannon, G. J., & Aravin, A. A. (2011). Production of artificial piRNAs in flies and mice. RNA, 18(1), 42–52. https://doi.org/10.1261/rna.029769.111

Murphy, W. J., Pringle, T. H., Crider, T. A., Springer, M. S., & Miller, W. (2007). Using genomic data to unravel the root of the placental mammal phylogeny. Genome Research, 17(4), 413–421. https://doi.org/10.1101/gr.5918807

Nellåker, C., Keane, T. M., Yalcin, B., Wong, K., Agam, A., Belgard, T. G., Flint, J., Adams, D. J., Frankel, W. N., & Ponting, C. P. (2012). The genomic landscape shaped by selection on transposable elements across 18 mouse strains. Genome Biology, 13(6), R45. https://doi.org/10.1186/gb-2012-13-6-r45

Ono, M., Kitasato, H., Ohishi, H., & Motobayashi-Nakajima, Y. (1984). Molecular cloning and long terminal repeat sequences of intracisternal A-particle genes in Mus caroli. Journal of Virology, 50(2), 352–358. https://doi.org/10.1128/jvi.50.2.352-358.1984

Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D., & Zamore, P. D. (2018). PIWI-interacting RNAs: Small RNAs with big functions. Nature Reviews Genetics, 20(2), 89–108. https://doi.org/10.1038/s41576-018-0073-3

Özata, D. M., Yu, T., Mou, H., Gainetdinov, I., Colpan, C., Cecchini, K., Kaymaz, Y., Wu, P.-H., Fan, K., Kucukural, A., Weng, Z., & Zamore, P. D. (2019). Evolutionarily conserved pachytene piRNA loci are highly divergent among modern humans. Nature Ecology & Evolution, 4(1), 156–168. https://doi.org/10.1038/s41559-019-1065-1

Pezic, D., Manakov, S. A., Sachidanandam, R., & Aravin, A. A. (2014). piRNA pathway targets active LINE1 elements to establish the repressive H3K9me3 mark in germ cells. Genes & Development, 28(13), 1410–1428. https://doi.org/10.1101/gad.240895.114

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., & Manke, T. (2014). deepTools: A flexible platform for exploring deep-sequencing data. Nucleic Acids Research, 42(W1), W187–W191. https://doi.org/10.1093/nar/gku365

Ray, R., & Pandey, P. (2018). piRNA analysis framework from small RNA-Seq data by a novel cluster prediction tool - PILFER. Genomics, 110(6), 355–365. https://doi.org/10.1016/j.ygeno.2017.12.005

Reuter, M., Berninger, P., Chuma, S., Shah, H., Hosokawa, M., Funaya, C., Antony, C., Sachidanandam, R., & Pillai, R. S. (2011). Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. Nature, 480(7376), 264–267. https://doi.org/10.1038/nature10672

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. Nature Biotechnology, 29(1), 24–26. https://doi.org/10.1038/nbt.1754

Rosenkranz, D., Han, C.-T., Roovers, E. F., Zischler, H., & Ketting, R. F. (2015). Piwi proteins and piRNAs in mammalian oocytes and early embryos: From sample to sequence. Genomics Data, 5, 309–313. https://doi.org/10.1016/j.gdata.2015.06.026

Rosenkranz, D., & Zischler, H. (2012). proTRAC - A software for probabilistic piRNA cluster detection, visualization and analysis. BMC Bioinformatics, 13(1). https://doi.org/10.1186/1471-2105-13-5

Ruby, J. G., Jan, C., Player, C., Axtell, M. J., Lee, W., Nusbaum, C., Ge, H., & Bartel, D. P. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. Cell, 127(6), 1193–1207. https://doi.org/10.1016/j.cell.2006.10.040

Sun, Y. H., Wang, R. H., Du, K., Zhu, J., Zheng, J., Xie, L. H., Pereira, A. A., Zhang, C., Ricci, E. P., & Li, X. Z. (2021). Coupled protein synthesis and ribosome-guided piRNA processing on mRNAs. Nature Communications, 12(1). https://doi.org/10.1038/s41467-021-26233-8

Sun, Y. H., Xie, L. H., Zhuo, X., Chen, Q., Ghoneim, D., Zhang, B., Jagne, J., Yang, C., & Li, X. Z. (2017). Domestic chickens activate a piRNA defense against avian leukosis virus. ELife, 6. https://doi.org/10.7554/elife.24695

Sun, Y. H., Zhu, J., Xie, L. H., Li, Z., Meduri, R., Zhu, X., Song, C., Chen, C., Ricci, E. P., Weng, Z., & Li, X. Z. (2020). Ribosomes guide pachytene piRNA formation on long intergenic piRNA precursors. Nature Cell Biology, 22(2), 200–212. https://doi.org/10.1038/s41556-019-0457-4

Surani, M. A., & Hajkova, P. (2010). Epigenetic reprogramming of mouse germ cells toward totipotency. Cold Spring Harbor Symposia on Quantitative Biology, 75(0), 211–218. https://doi.org/10.1101/sqb.2010.75.010

Thybert, D., Roller, M., Navarro, F. C. P., Fiddes, I., Streeter, I., Feig, C., Martin-Galvez, D., Kolmogorov, M., Janoušek, V., Akanni, W., Aken, B., Aldridge, S., Chakrapani, V., Chow, W., Clarke, L., Cummins, C., Doran, A., Dunn, M., Goodstadt, L., … Flicek, P. (2018). Repeat associated mechanisms of genome evolution and function revealed by the Mus caroli and Mus pahari genomes. Genome Research, 28(4), 448–459. https://doi.org/10.1101/gr.234096.117

Wang, L., Wang, S., & Li, W. (2012). RSeQC: Quality control of RNA-seq experiments. Bioinformatics, 28(16), 2184–2185. https://doi.org/10.1093/bioinformatics/bts356

Wu, P.-H., Fu, Y., Cecchini, K., Özata, D. M., Arif, A., Yu, T., Colpan, C., Gainetdinov, I., Weng, Z., & Zamore, P. D. (2020). The evolutionarily conserved piRNA-producing locus pi6 is required for male mouse fertility. Nature Genetics, 52(7), 728–739. https://doi.org/10.1038/s41588-020-0657-7

Yang, Z., Chen, K.-M., Pandey, R. R., Homolka, D., Reuter, M., Janeiro, B. K. R., Sachidanandam, R., Fauvarque, M.-O., McCarthy, A. A., & Pillai, R. S. (2016). PIWI Slicing and EXD1 Drive Biogenesis of Nuclear piRNAs from Cytosolic Targets of the Mouse piRNA Pathway. Molecular Cell, 61(1), 138–152. https://doi.org/10.1016/j.molcel.2015.11.009

Yu, T., Fan, K., Özata, D. M., Zhang, G., Fu, Y., Theurkauf, W. E., Zamore, P. D., & Weng, Z. (2021). Long first exons and epigenetic marks distinguish conserved

pachytene piRNA clusters from other mammalian genes. Nature Communications, 12(1). https://doi.org/10.1038/s41467-020-20345-3

6. Suppelementary material

6.1. Supplementary tables

Supp. Table 1. Biological and sequencing information, number of reads and expression counts in different annotations for each sample. Asterisk (*) indicates that the corresponding annotation was not initially defined in these species.

Sample	Age	Tissue	Assembly	Sequencing adapter
Mus musculus replicate 1	14 weeks	Whole testes	GRCm39	TGGAATTCTCGGGTGCCAAGG
Mus musculus replicate 2	14 weeks	Whole testes	GRCm39	TGGAATTCTCGGGTGCCAAGG
Mus caroli replicate 1	9-10 weeks	Whole testes	CAROLI_EIJ_v1.1	TGGAATTCTCGGGTGCCAAGG
Mus caroli replicate 2	9-10 weeks	Whole testes	CAROLI_EIJ_v1.1	TGGAATTCTCGGGTGCCAAGG
Mus pahari replicate 1	8 weeks	Whole testes	PAHARI_EIJ_v1.1	TGGAATTCTCGGGTGCCAAGG
Mus pahari replicate 2	8 weeks	Whole testes	PAHARI_EIJ_v1.1	TGGAATTCTCGGGTGCCAAGG

Sample	Raw reads	Trimmed reads	Filtered reads	Aligned reads
Mus musculus replicate 1	49312103	44532620	38214919	36770278
Mus musculus replicate 2	44144051	41401353	35578711	34347199
Mus caroli replicate 1	53885386	50810911	43600732	40425463
Mus caroli replicate 2	54677733	51699764	44523496	41942804
Mus pahari replicate 1	45442424	42632343	36611022	32348873
Mus pahari replicate 2	46260817	42860036	36561199	32357333

Sample	Counts in *Yuetal* clusters	Counts in *de novo* clusters (*MUS*)	Counts in *de novo* clusters (*CAR*)	Counts in *de novo* clusters (*PAH*)
Mus musculus replicate 1	20752590	21485748	22045335*	22298927*
Mus musculus replicate 2	20047122	20797988	21310649*	21553681*
Mus caroli replicate 1	20033499*	19995856*	21244080	21532090*
Mus caroli replicate 2	23662354*	23594772*	24971737	25422103*
Mus pahari replicate 1	18700335*	18228778*	19200200*	20624883
Mus pahari replicate 2	18227929*	17758049*	18718261*	20099482

Supp. Table 2. Ping-pong Z-scores obtained with `TBr2_pingpong.pl` for each replicate. A Z-score > 1.6449 is equivalent to a p-value < 0.05, and a Z-score > 2.3264 to a p-value < 0.01. Alternative hypothesis is the presence of piRNA production by ping-pong effect.

Sample	Ping-pong Z-score
Mus musculus replicate 1	48.8170
Mus musculus replicate 1	52.9612
Mus caroli replicate 1	44.3188
Mus caroli replicate 1	37.2372
Mus pahari replicate 1	19.1476
Mus pahari replicate 1	19.3694

Supp. Table 3. Fisher's test associating species-specific TE insertions with the differentially expressed piRNA clusters in different contrasts. piRNA clusters are *de novo* clusters predicted in *Mus caroli*.

Contrast	TE	TE only in	Fisher's p-value
PAH vs *MUS*	LINE	*Mus musculus*	1
CAR vs *MUS*		*Mus musculus*	0.342
CAR vs *MUS*		*Mus caroli*	1
PAH vs *CAR*		*Mus caroli*	1
PAH vs *MUS*		*Mus pahari*	1
PAH vs *CAR*		*Mus pahari*	0.498
PAH vs *MUS*	LTR	*Mus musculus*	0.353
CAR vs *MUS*		*Mus musculus*	1
CAR vs *MUS*		*Mus caroli*	1
PAH vs *CAR*		*Mus caroli*	1
PAH vs *MUS*		*Mus pahari*	1
PAH vs *CAR*		*Mus pahari*	1
PAH vs *MUS*	LINE 1	*Mus musculus*	0.553
CAR vs *MUS*		*Mus musculus*	0.27
CAR vs *MUS*		*Mus caroli*	0.301
PAH vs *CAR*		*Mus caroli*	0.622
PAH vs *MUS*		*Mus pahari*	0.667
PAH vs *CAR*		*Mus pahari*	1
PAH vs *MUS*	ERV1	*Mus musculus*	0.283
CAR vs *MUS*		*Mus musculus*	0.495
CAR vs *MUS*		*Mus caroli*	0.547
PAH vs *CAR*		*Mus caroli*	1
PAH vs *MUS*		*Mus pahari*	0.187
PAH vs *CAR*		*Mus pahari*	0.432
PAH vs *MUS*	ERVK	*Mus musculus*	0.16
CAR vs *MUS*		*Mus musculus*	**0.092***
CAR vs *MUS*		*Mus caroli*	1
PAH vs *CAR*		*Mus caroli*	0.611
PAH vs *MUS*		*Mus pahari*	0.283
PAH vs *CAR*		*Mus pahari*	0.741
PAH vs *MUS*	IAP	*Mus musculus*	1
CAR vs *MUS*		*Mus musculus*	0.168
CAR vs *MUS*		*Mus caroli*	0.449
PAH vs *CAR*		*Mus caroli*	1
PAH vs *MUS*		*Mus pahari*	1
PAH vs *CAR*		*Mus pahari*	0.109

Supp. Table 4. Fisher's test associating species-specific TE insertions with the differentially expressed piRNA clusters in different contrasts. piRNA clusters are *de novo* clusters predicted in *Mus pahari*.

Contrast	TE	TE only in	Fisher's p-value
PAH vs *MUS*	LINE	*Mus musculus*	1
CAR vs *MUS*		*Mus musculus*	1
CAR vs *MUS*		*Mus caroli*	1
PAH vs *CAR*		*Mus caroli*	0.605
PAH vs *MUS*		*Mus pahari*	**0.028****
PAH vs *CAR*		*Mus pahari*	0.372
PAH vs *MUS*	LTR	*Mus musculus*	0.493
CAR vs *MUS*		*Mus musculus*	1
CAR vs *MUS*		*Mus caroli*	1
PAH vs *CAR*		*Mus caroli*	1
PAH vs *MUS*		*Mus pahari*	1
PAH vs *CAR*		*Mus pahari*	1
PAH vs *MUS*	LINE 1	*Mus musculus*	1
CAR vs *MUS*		*Mus musculus*	1
CAR vs *MUS*		*Mus caroli*	1
PAH vs *CAR*		*Mus caroli*	0.724
PAH vs *MUS*		*Mus pahari*	**0.028****
PAH vs *CAR*		*Mus pahari*	0.67
PAH vs *MUS*	ERV1	*Mus musculus*	0.365
CAR vs *MUS*		*Mus musculus*	0.581
CAR vs *MUS*		*Mus caroli*	1
PAH vs *CAR*		*Mus caroli*	0.348
PAH vs *MUS*		*Mus pahari*	1
PAH vs *CAR*		*Mus pahari*	0.104
PAH vs *MUS*	ERVK	*Mus musculus*	0.497
CAR vs *MUS*		*Mus musculus*	0.402
CAR vs *MUS*		*Mus caroli*	0.287
PAH vs *CAR*		*Mus caroli*	0.193
PAH vs *MUS*		*Mus pahari*	0.683
PAH vs *CAR*		*Mus pahari*	0.426
PAH vs *MUS*	IAP	*Mus musculus*	1
CAR vs *MUS*		*Mus musculus*	1
CAR vs *MUS*		*Mus caroli*	0.581
PAH vs *CAR*		*Mus caroli*	0.623
PAH vs *MUS*		*Mus pahari*	0.369
PAH vs *CAR*		*Mus pahari*	0.193

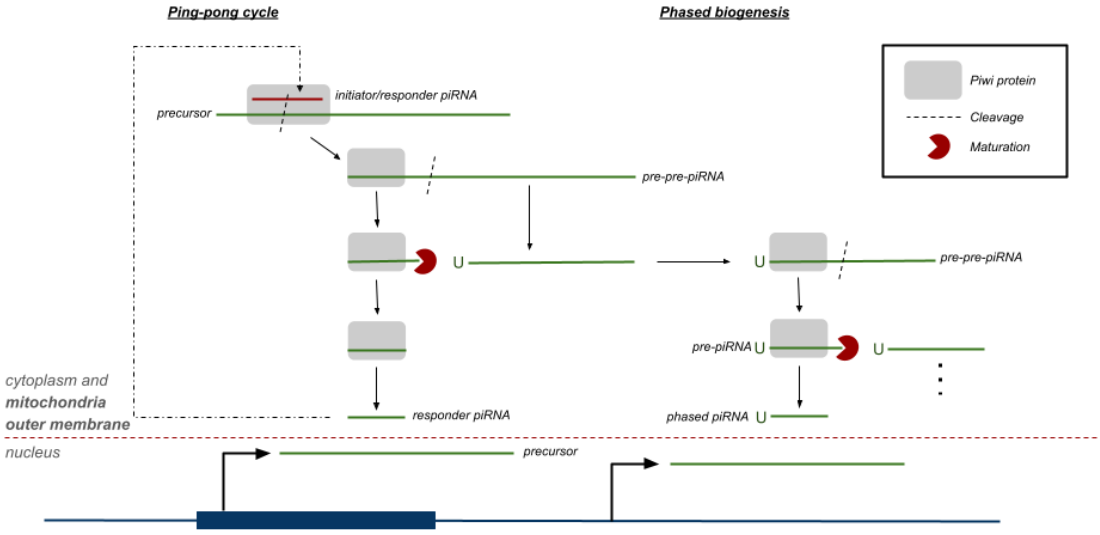
Supp. Table 5. Versions and parameters for the software used in some of the processes and analyses performed in this report.

Process	Software	Version	Parameters	Reference
Trim adaptor from reads	`Cutadapt`	3.4	`-a <adaptor_sequence> -O 9 -j 0 -m 19 -M 36 --trimmed-only`	*Martin, 2011*
Quality filtering of reads	`fastq_quality_filter`	0.0.14	`-q 30 -p 90 -Q 33`	http://hannonlab.cshl.edu/fastx_toolkit/
Read mapping to reference genome	`Bowtie`	1.2	`-v 1 -M 1 --best --strata -q -p 5 --seed 666 --time -S`	*Langmead et al., 2009*
SAM operations	`Samtools`	1.10 (`htslib` 1.10)	*depends on the command*	*Li et al., 2009*
Count reads in genomic features	`featureCounts`		`-Q 1 -T 4 -R BAM -F GTF -O --minOverlap 18 -s 0 -a <annot.gtf> -t <feature> -g <attribute>`	*Liao et al., 2013*
BAM to BigWig	`bamCoverage`	3.5.1	`--outFileFormat bigwig --effectiveGenomeSize <effectiveGenomeSize> --normalizeUsing CPM --samFlagExclude/--samFlagInclude 16`	*Ramírez et al., 2014*
BED operations	`Bedtools`	2.29.2	*depends on the command*	*Quinlan et al., 2010*
Quality controls on the FASTQ files	`FastQC`	0.11.9	`-o <outdir> -f fastq -t 2 --noextract`	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Multi-sample quality report	`MultiQC`	1.6	`-o <outdir> --file-list <list_of_inputs.txt> -f -v -n "multiqc_mus_sp"`	*Ewels et al., 2016*
Distribution of reads in genomic locations (i.e. exons, introns, intergenic…)	`RSeQC - read_distribution.py`	4.0.0	`-i <bam_file> -r <annot.bed>`	*Wang et al., 2012*
Collapse redundant reads (for *proTRAC*)	`TBr2_collapse.pl`	2.1	*default*	*Rosenkranz et al., 2015*
Filter low complexity reads (for *proTRAC*)	`TBr2_duster.pl`	2.1	*default*	*Rosenkranz et al., 2015*
Mapp reads to reference genome (for *proTRAC*)	`sRNAmapper.pl`	1.0.4	`-alignments best`	*Rosenkranz et al., 2015*
Weight multimapping reads (for *proTRAC*)	`reallocate.pl`	1.1	`10000 1000 b 0`	
Predict piRNA clusters	`proTRAC.pl`	2.4.3	`-geneset <annot.gtf> -repeatmasker <repeatmasker.out> -pdens 0.01 -pimin 21 -pmax 35 -pisize 0.75 -1Tor10A 0.75`	*Rosenkranz et al., 2012*
Compute ping-pong signature.	`Tbr2_pingpong.pl`	2.1	*default*	*Rosenkranz et al., 2015*

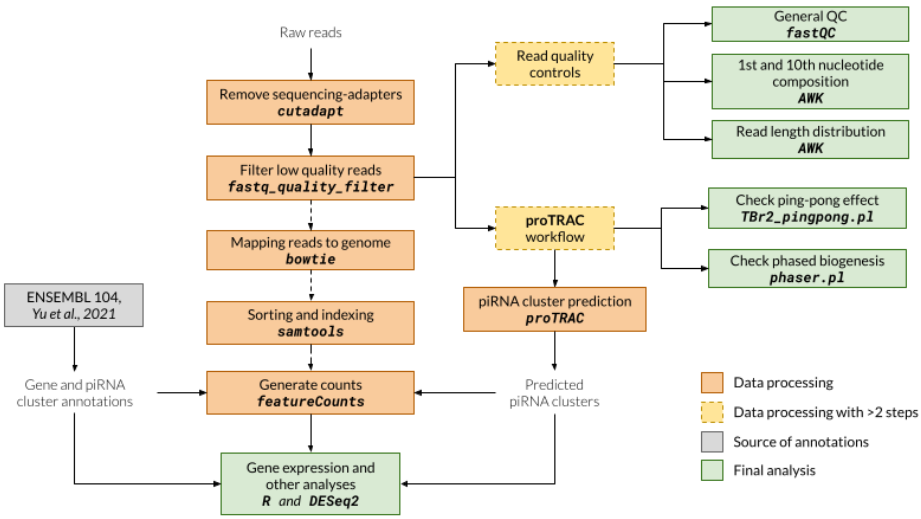
Supp. Table 6. Versions and sources/references for the R packages used in this study.

R package	Version	Description	Reference
DESeq2	1.32.0	Differential gene expression analysis based on the negative binomial distribution	*Love et al., 2014*
pcaExplorer	2.18.0	Visualizaton of RNA-seq data based on principal component analysis	*Marini and Binder, 2019*
plyr	1.8.6	Tools for splitting, applying and combining data	*https://github.com/hadley/plyr*
dplyr	1.0.7	Grammar of data manipulation.	*https://dplyr.tidyverse.org/*
tidyr	1.1.3	Functions that help you tidy the data	*https://tidyr.tidyverse.org/*
magrittr	2.0.1	Operators and functions to make the code more readable	*https://magrittr.tidyverse.org/*
purrr	0.3.4	Enhancers R's functional programming. Help dealing with lists.	*https://purrr.tidyverse.org/*
stringr	1.4.0.9000	Work with strings as easy as possible	*https://stringr.tidyverse.org/*
tibble	3.1.6	Modern re-imaging of a data frame.	*https://tribble.tidyverse.org/*
janitor	2.1.0.9000	Simple tools for data cleaning in R	*https://github.com/sfirke/janitor*
bedtoolsr	2.30.0.1	R package wrapping bedtools	*https://github.com/PhanstielLab/bedtoolsr*
plyranges	1.13.1	A fluent interface for manipulating GenomicRanges	*https://sa-lee.github.io/plyranges/*
GenomicRanges	1.46.1	Representation and manipulation of genomic intervals	*Lawrence et al., 2013*
ggplot2	3.3.5	A system for declaratively creating graphics	*https://ggplot2.tidyverse.org*
ggpubr	0.4.0	'ggplot2'-based publication ready plots	*https://github.com/kassambara/ggpubr*
ggh4x	0.2.1.9000	'ggplot2' extension with options for facets, etc	*https://github.com/teunbrand/ggh4x*
ggrepel	0.9.1.9999	Repel overlapping text labels away from each other	*https://github.com/slowkow/ggrepel*
cowplot	1.1.1	Streamlined Plot Theme and Plot Annotations for 'ggplot2'	*https://github.com/wilkelab/cowplot*
patchwork	1.1.0.9000	Combine 'ggplots' easily.	*https://patchwork.data-imaginist.com/*
plotmics	5.1.0	Visualize omics and sequencing data in R	*https://github.com/amitjavilaventura/plotmics*

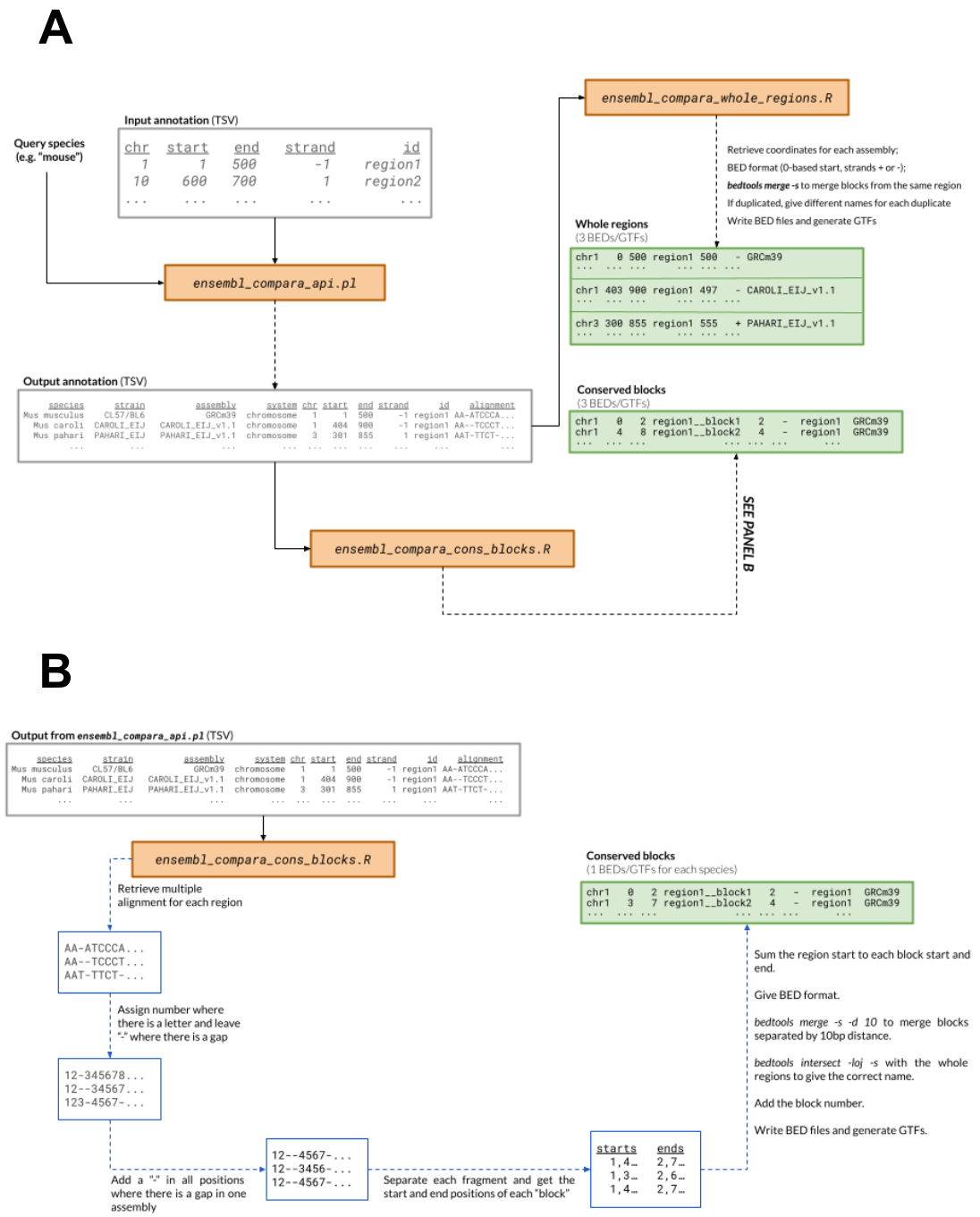
6.2. Supplementary figures



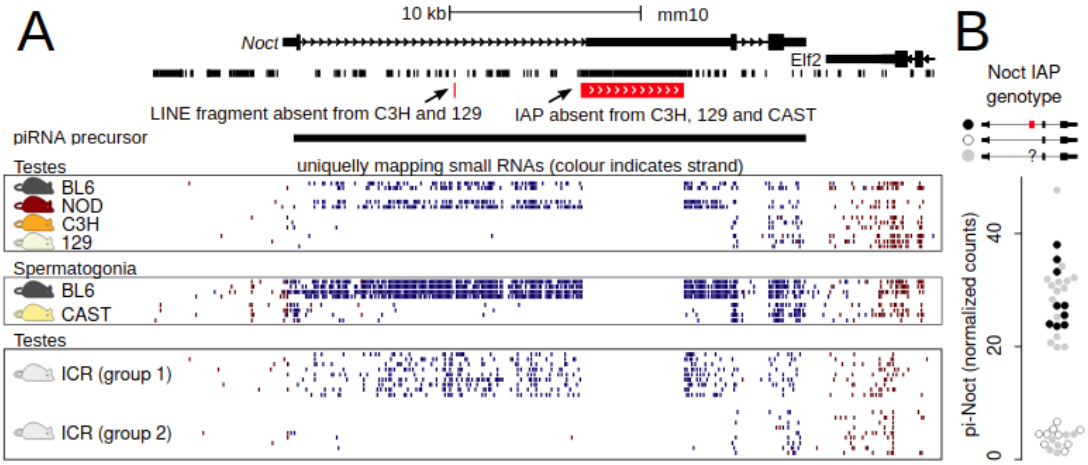
Supp. Figure 1. piRNA biogenesis in most animals. First, a piRNA cluster is transcribed and the precursor transcript is transported to the cytoplasm and mitocondria outer membrane. Then, a Piwi protein guided by an initiator piRNA slices the precursor transcript, initiating the phased piRNA pathway. Some of these piRNAs (responder piRNAs) may act as initiators to restart the whole process, producing a cycle of amplification called ping-pong pathway.



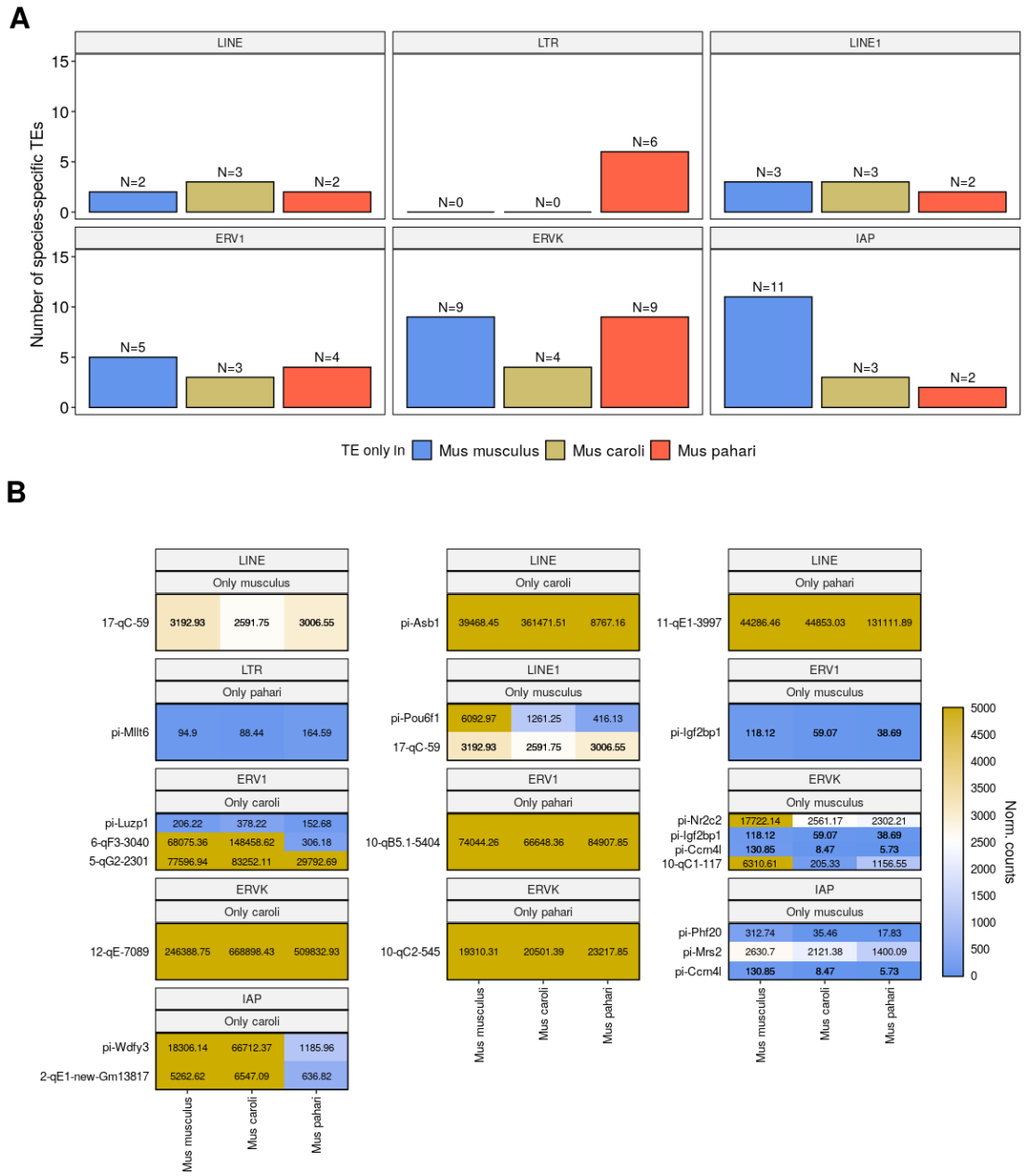
Supp. Figure 2. General workflow followed in this project. Gray rectangle is the source of gene and piRNA cluster annotations, orange rectangles represent processing steps, yellow oranges include more than one processing step and green rectangles are endpoint analyses.



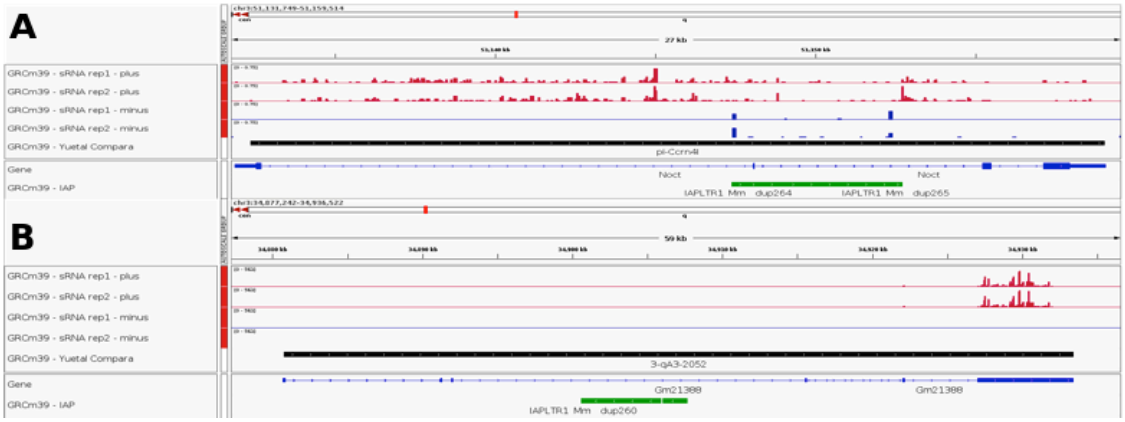
Supp. Figure 3. Workflow followed to obtain the orthologous regions from ENSEMBL Compara Perl API (A) and retrieve the coserved blocks (B).



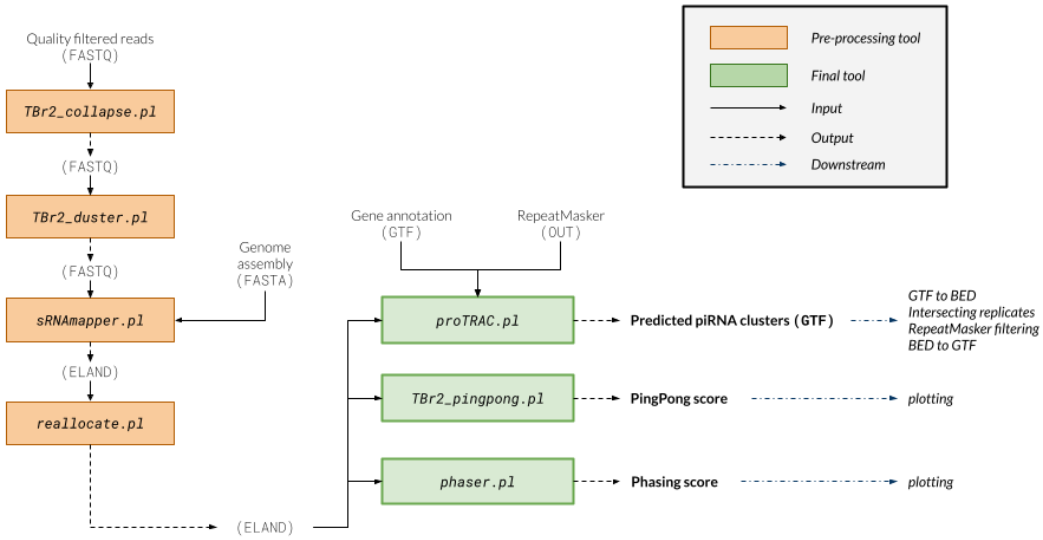
Supp. Figure 4. *Noct* gene in several strains of *Mus musculus*. (A) browser track with the coordinates, presence of transposable elements and small RNA reads in the *Noct* gene. (B) Normalized expression of *pi-Noct/pi-Ccrn4l* -piRNA cluster arising from *Noct*- regarding the presence or absence of IAP. ICR mice are from an outbred strain. ICR mice from group 1 have an IAP insertion, while group 2 mice do not have it. Source: Tanya Vavouri personal communication (not published).



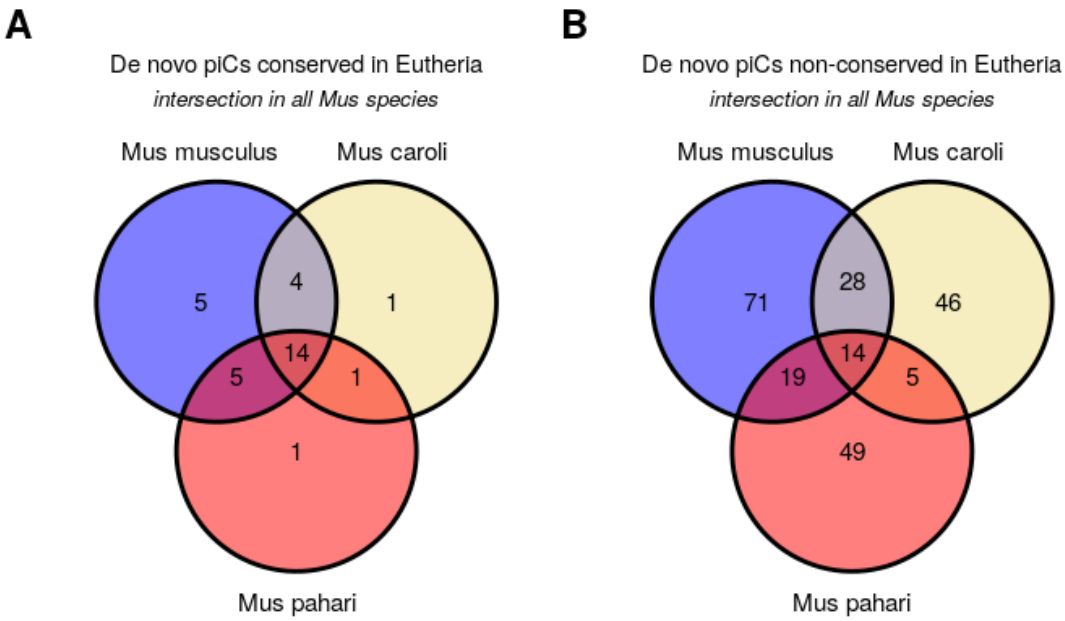
Supp. Figure 5. Clusters with species-specific transposon insertions (*Yuetal* clusters): (A) number of clusters with species-specific insertions of different TEs; (B) expression heatmaps with the DESeq2-normalized counts of the *Yuetal* clusters with higher expression in the species with the TE.



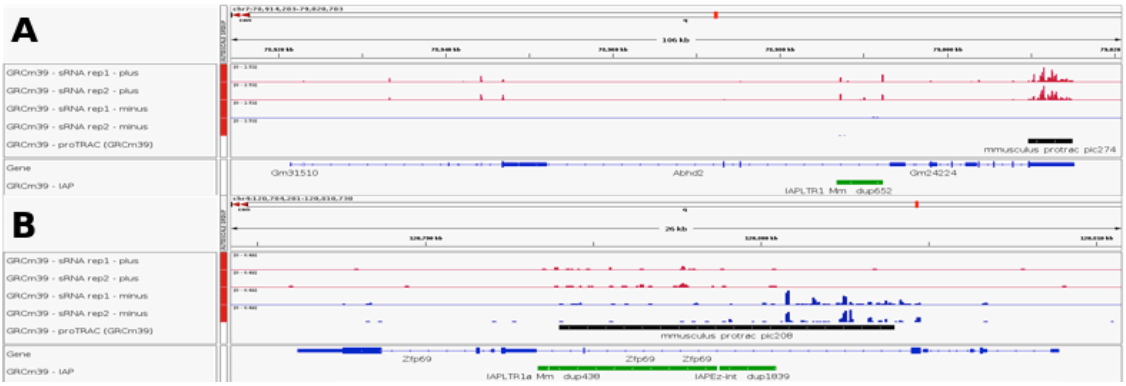
Supp. Figure 6. IGV snapshots of *Yuetal* clusters with IAP only in *Mus musculus*: (A) *Noct* gene / *pi-Ccrn4l* cluster with a sense IAP insertion and reads mapping all along the cluster; (B) intergenic cluster *3-qA3-2052* with an antisense IAP insertion and reads mapping only in the 3'UTR.



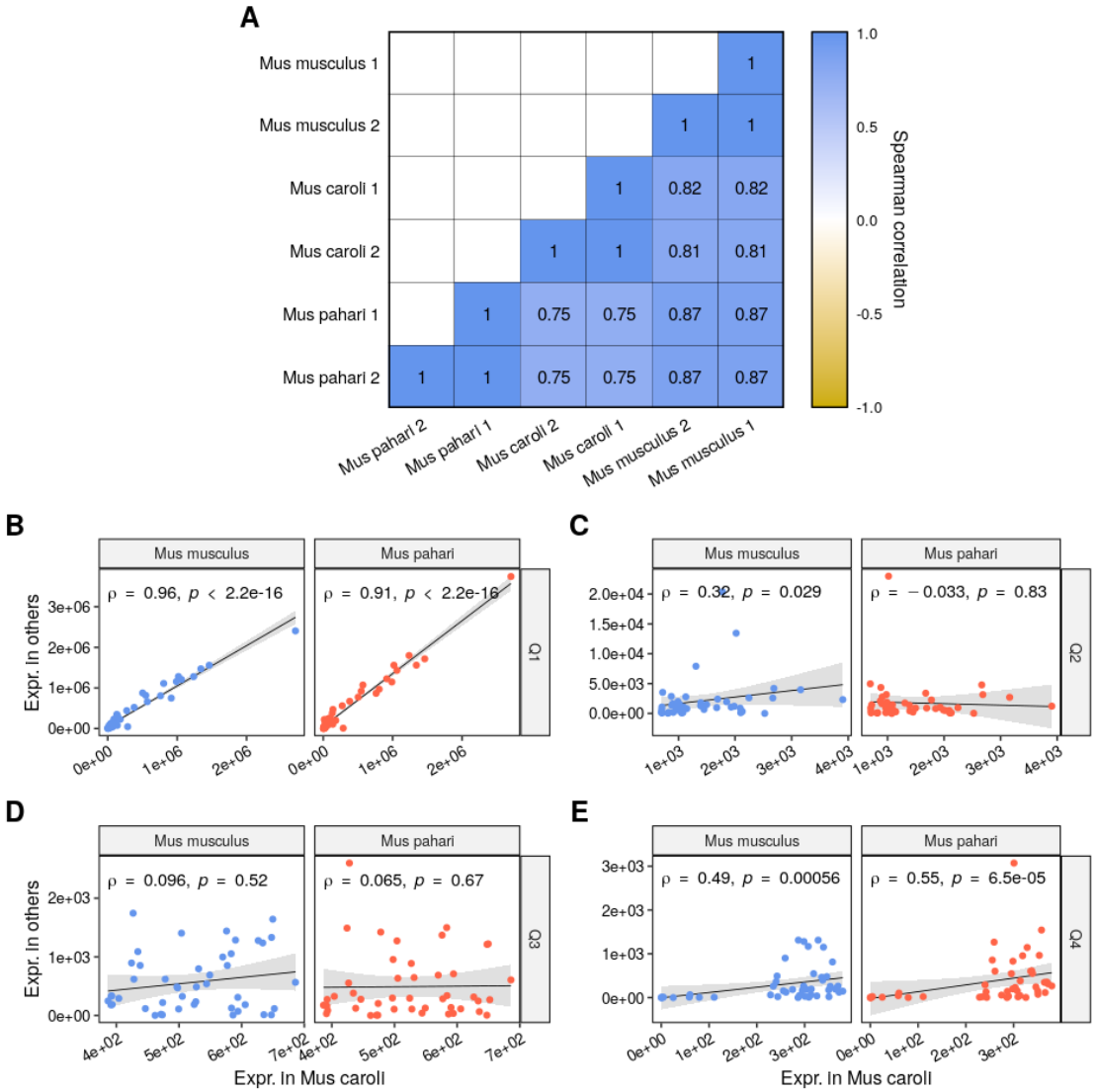
Supp. Figure 7. Workflow followed to predict the *de* novo clusters with proTRAC and compute the phasing score -3'-to-5' distance of consecutive reads in the same strand- and the Ping-pong score -5'-to-5' distance of overlapping reads on diferent strands-. All tools mentioned in the image are developed by *Rosenkranz et al. 2012* and *Rosenkranz et al., 2015.*



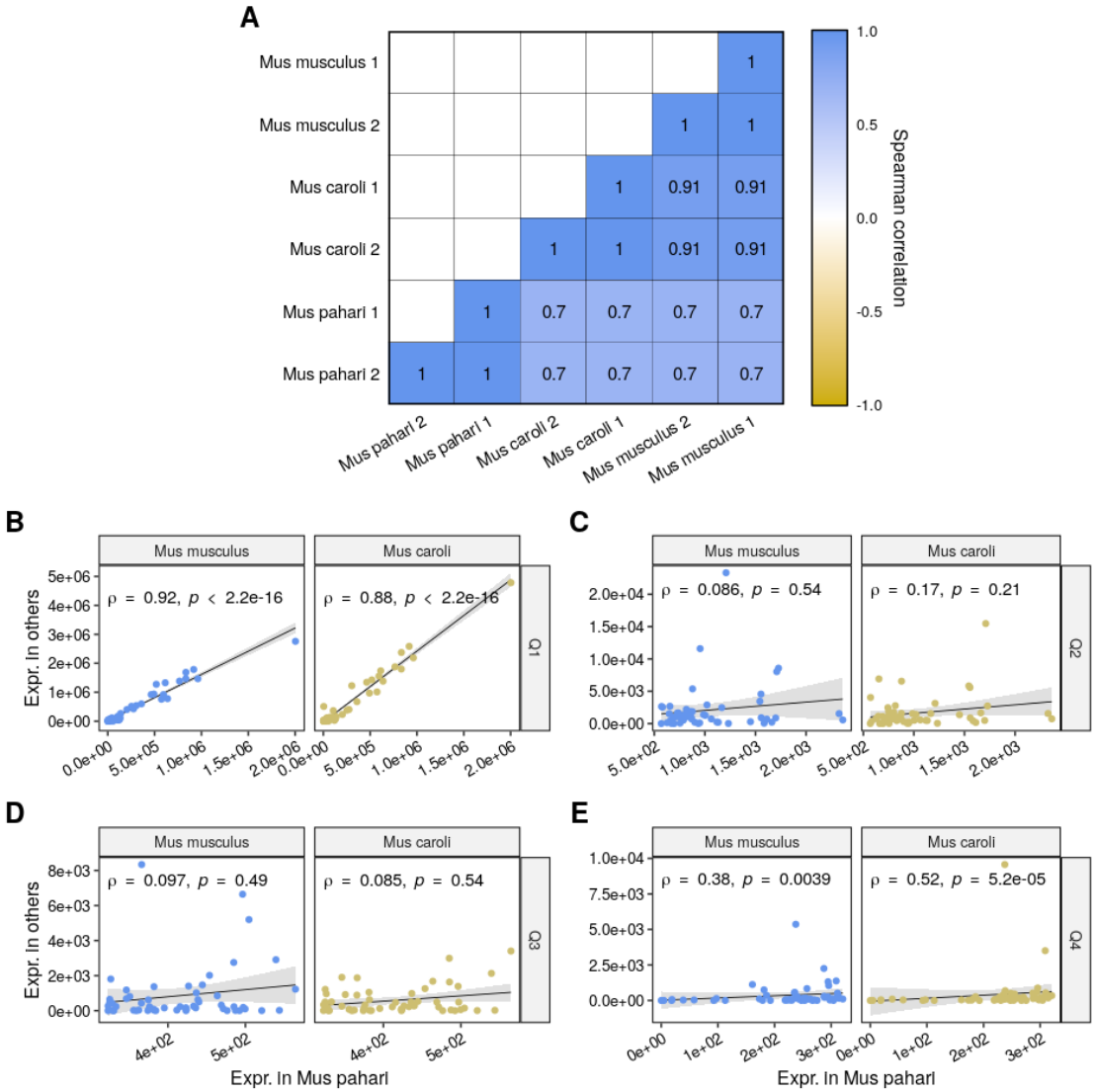
Supp. Figure 8. Intersection of the *novo* clusters predicted in all the studied *Mus* species: (A) clusters present in the Eutherian-conserved piRNA clusters (ECpiCs) obtained by *Chirn et al.* (2015); (B) *de novo* genic clusters not present in the ECpiCs.



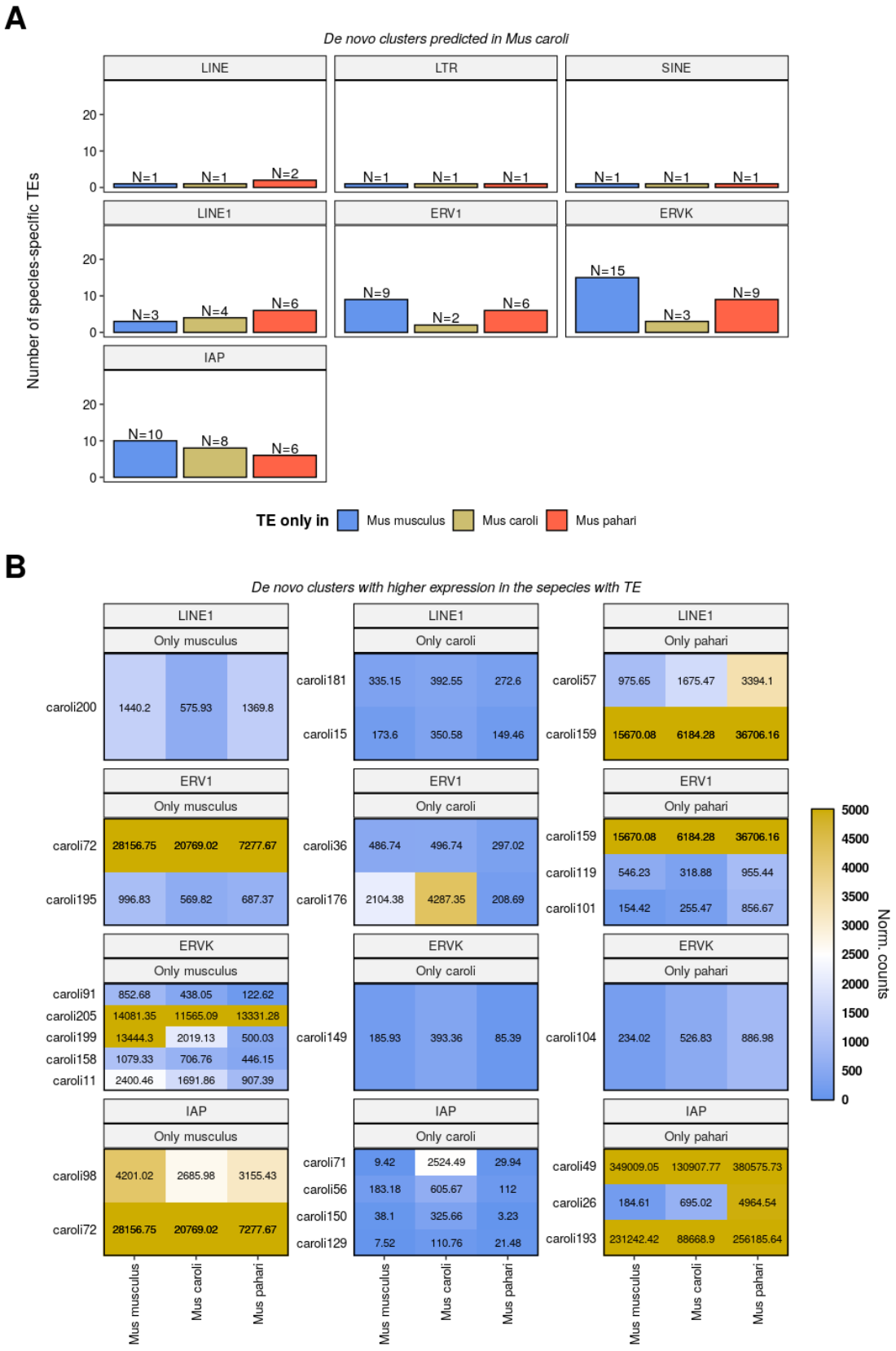
Supp. Figure 9. IGV snapshots of genes with IAP insertions and predicted *de novo* clusters: (A) *Abhd2* gene with an antisense IAP insertion and reads mainly in the 3'UTR; (B) *Zfp69* gene with a sense IAP insertion and reads mapping across a large region of the gene.



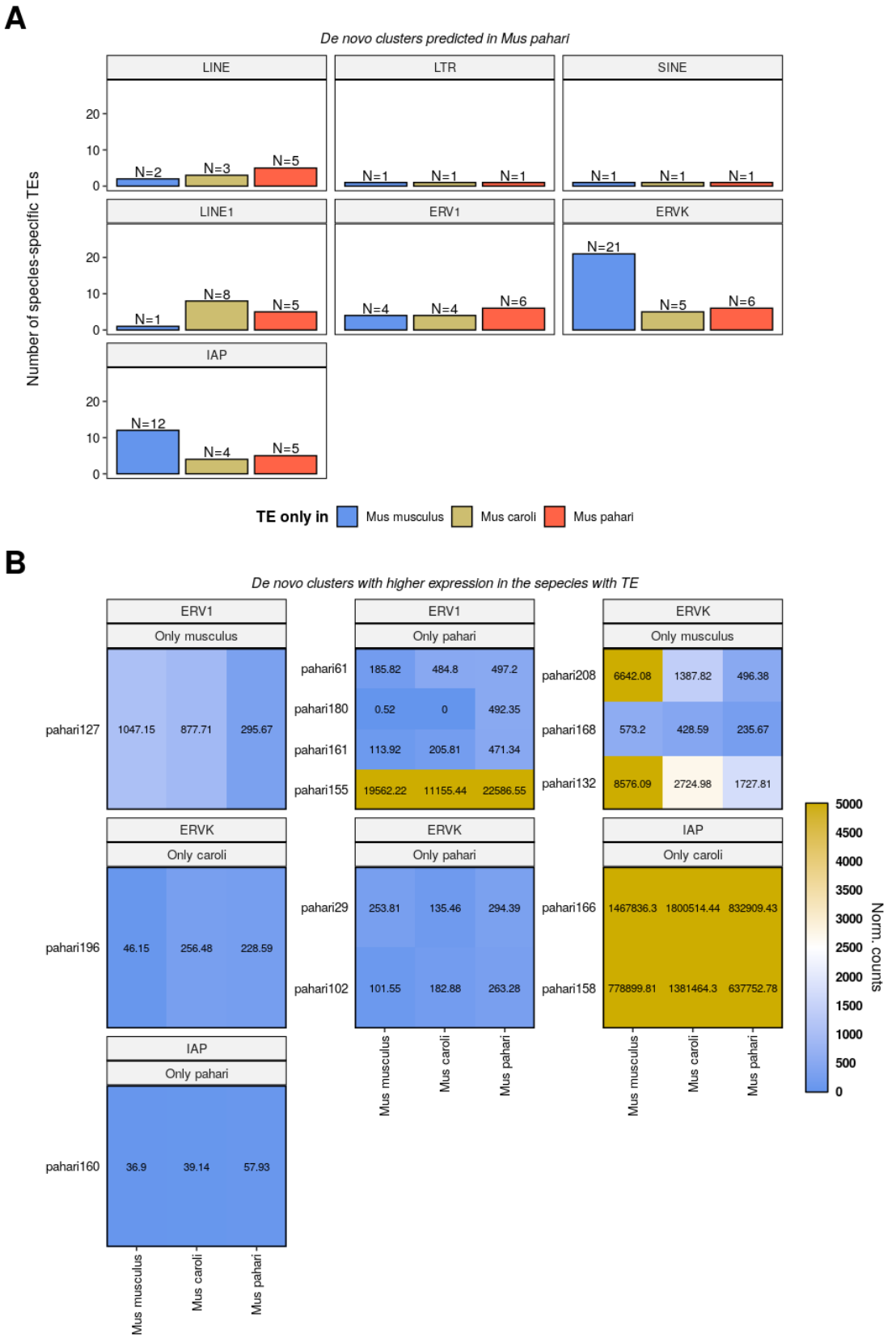
Supp. Figure 10. Spearman correlations of the *de novo* clusters predicted in *Mus caroli*: (A) correlation plot with all clusters and all samples; (B) scatter plots and correlations the top 25% (Q1) expressed clusters in *Mus caroli*; (C) scatter plots and correlations of the clusters in the Q2; (D) scatter plots and correlations of the clusters in the Q3; (E) scatter plots and correlations of the clusters in the Q4. Expression is shown in DESeq2-normalized counts.



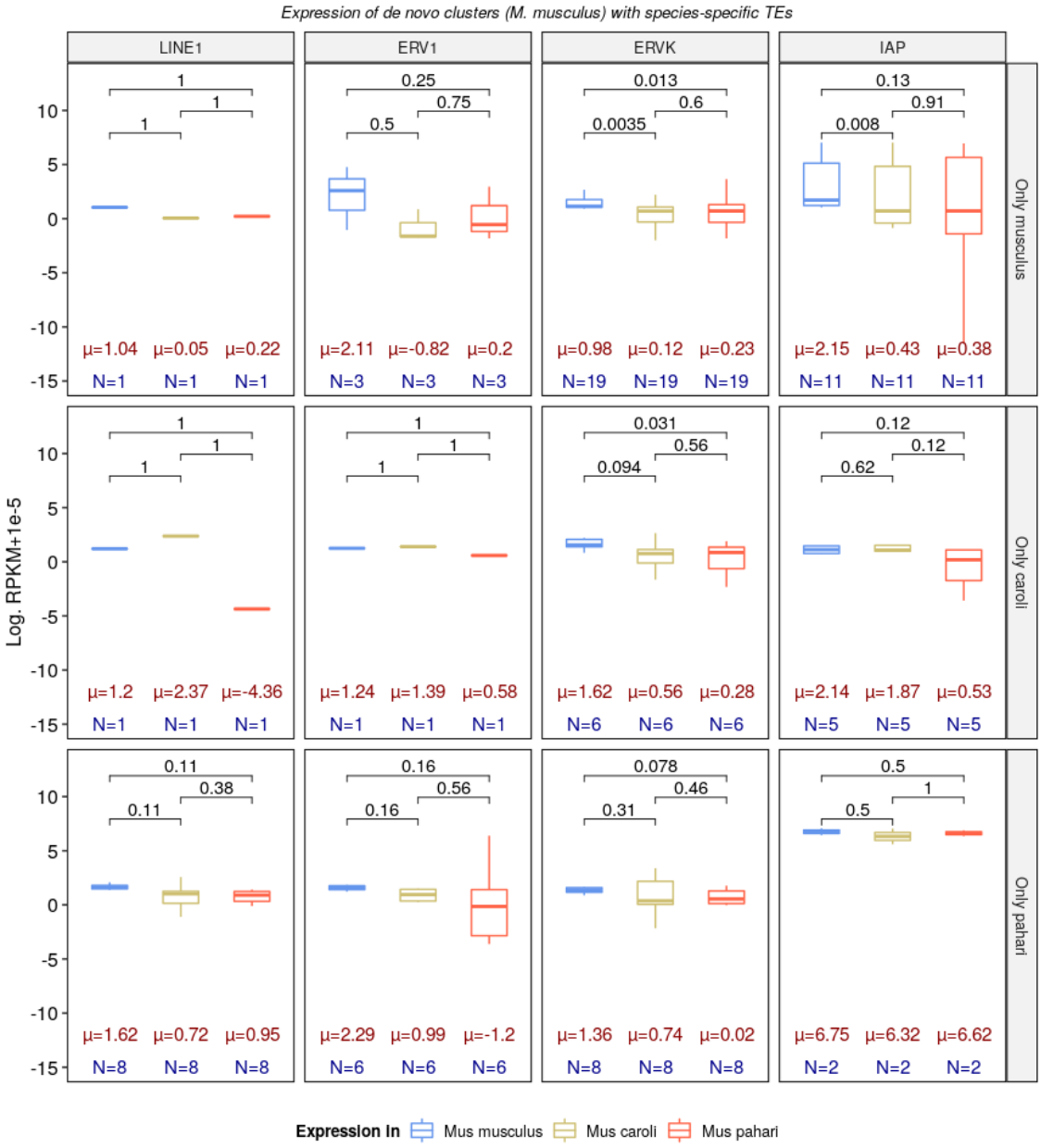
Supp. Figure 11. Spearman correlations of the *de novo* clusters predicted in *Mus pahari*: (A) correlation plot with all clusters and all samples; (B) scatter plots and correlations the top 25% (Q1) expressed clusters in *Mus pahari*; (C) scatter plots and correlations of the clusters in the Q2; (D) scatter plots and correlations of the clusters in the Q3; (E) scatter plots and correlations of the clusters in the Q4. Expression is shown in DESeq2-normalized counts.



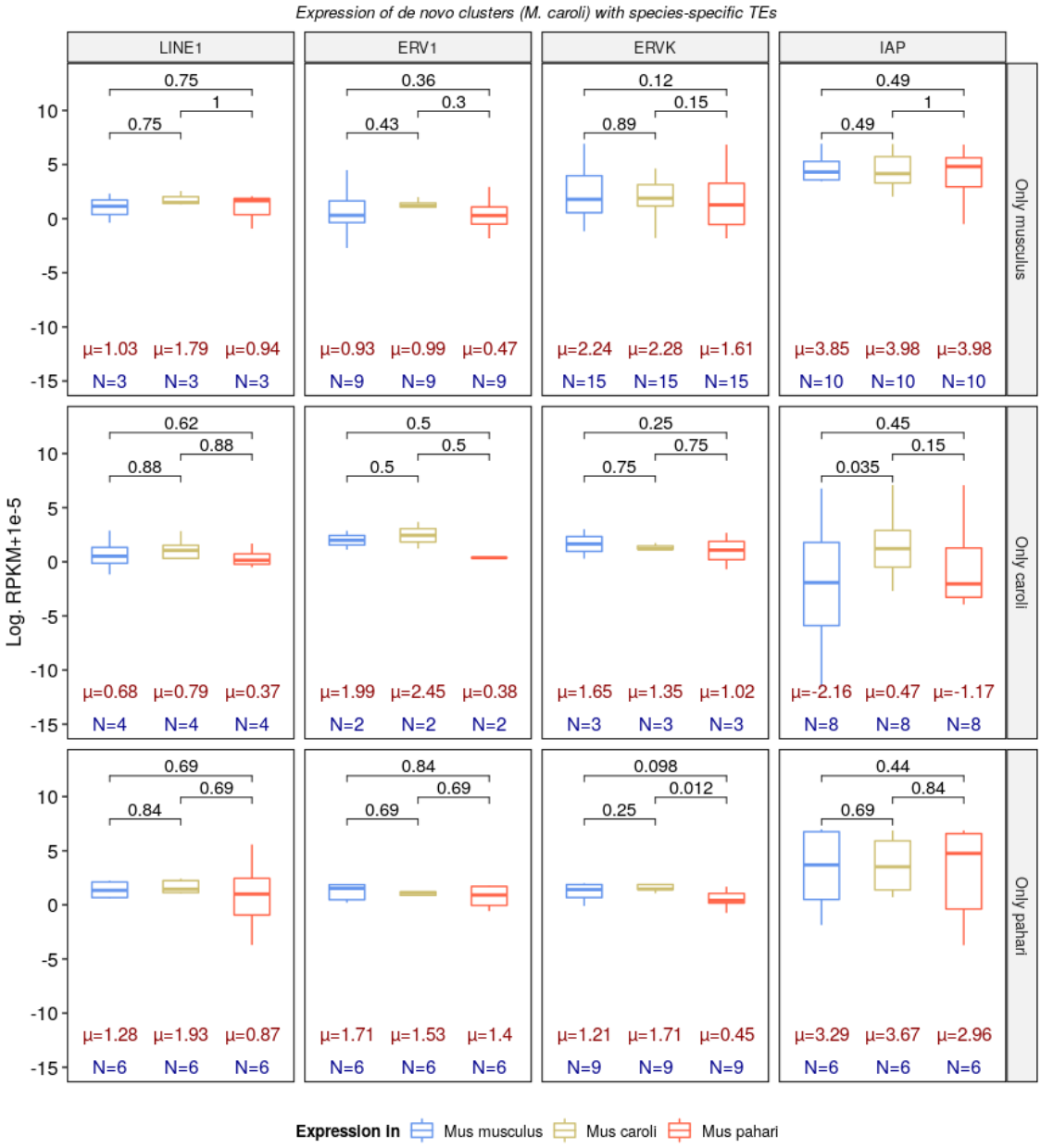
Supp. Figure 12. Clusters with species-specific transposon insertions (*de novo* clusters predicted in *Mus caroli*): (A) number of clusters with species-specific insertions of different TEs; (B) expression heatmaps of the clusters with higher expression in the species with the TE. Expression is shown in DESeq2-normalized counts.



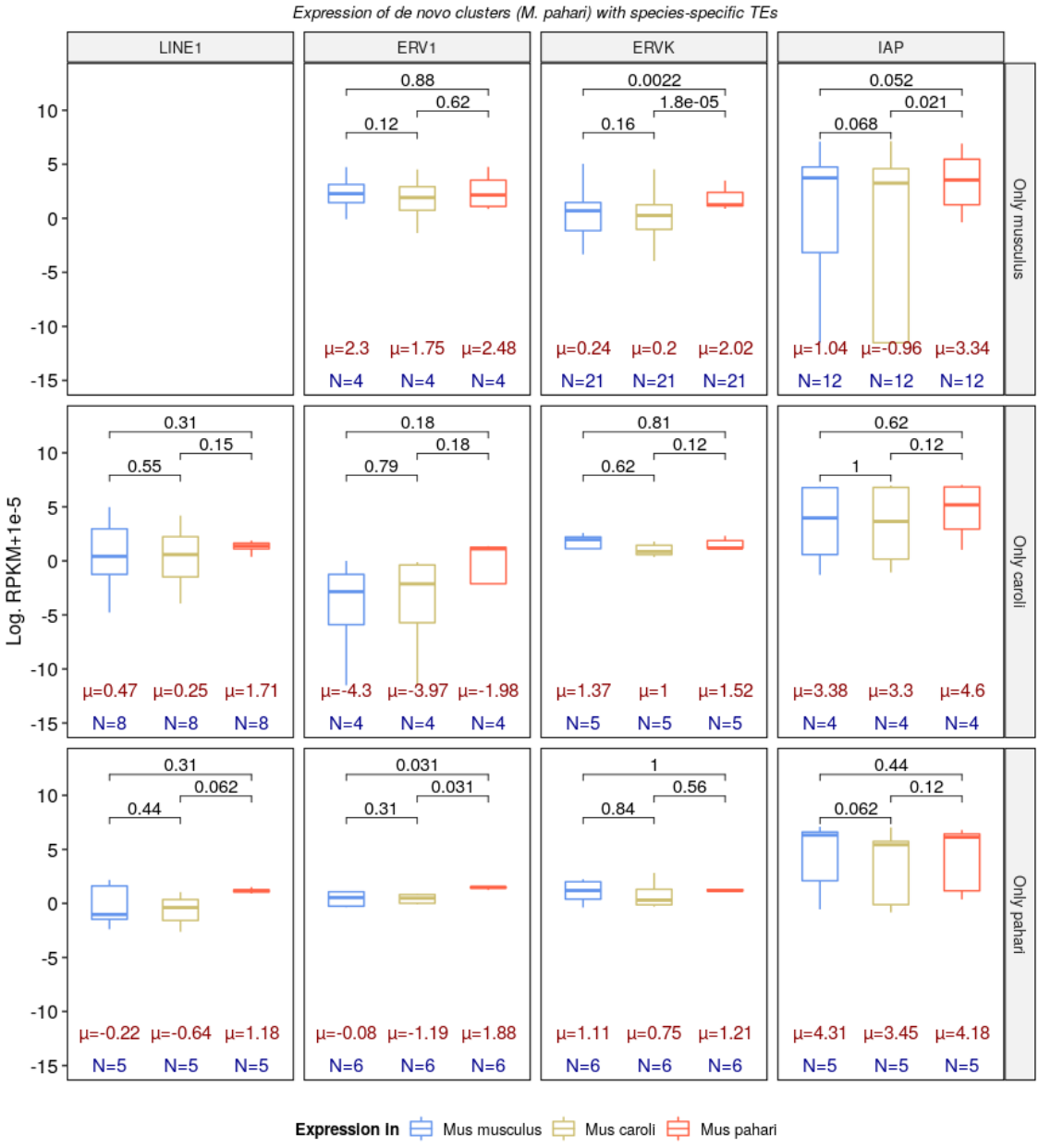
Supp. Figure 13. Clusters with species-specific transposon insertions (*de novo* clusters predicted in *Mus pahari*): (A) number of clusters with species-specific insertions of different TEs; (B) expression heatmaps of the clusters with higher expression in the species with the TE. Expression is shown in DESeq2-normalized counts.



Supp. Figure 14. Expression of clusters (*de novo* clusters predicted in *Mus musculus*) with species-specific insertions of several TEs. Top strips indicate the inserted transposon. Right strips indicate the species that have the insertion of the transposon. Expression is represented in logarithm of the RPKM + 1e-5. We added 1e-5 as a pseudocount to avoid the infinite values after the logarithm transformation. Text in red represent the mean of the distribution. Text in blue show the number of observations in each boxplot. The number on the boxplot are p-values of two-sided Wilcoxon signed-rank tests.

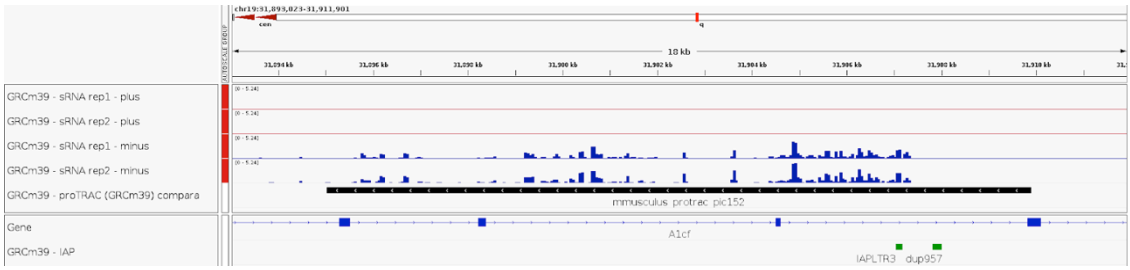


Supp. Figure 15. Expression of clusters (*de novo* clusters predicted in *Mus caroli*) with species-specific insertions of several TEs. Top strips indicate the inserted transposon. Right strips indicate the species that have the insertion of the transposon. Expression is represented in logarithm of the RPKM + 1e-5. We added 1e-5 as a pseudocount to avoid the infinite values after the logarithm transformation. Text in red represent the mean of the distribution. Text in blue show the number of observations in each boxplot. The number on the boxplot are p-values of two-sided Wilcoxon signed-rank tests.

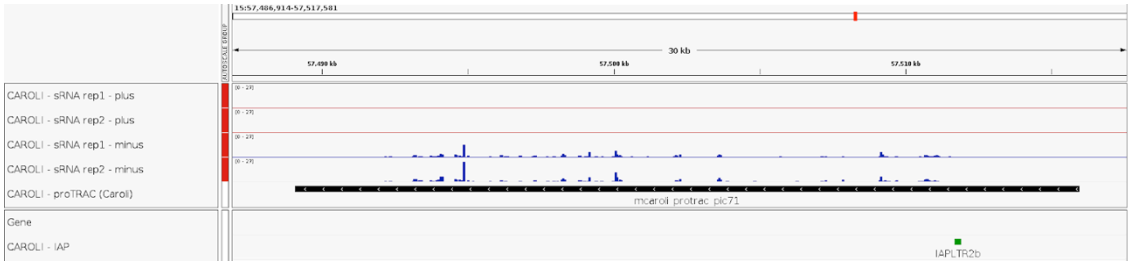


Supp. Figure 16. Expression of clusters (*de novo* clusters predicted in *Mus pahari*) with species-specific insertions of several TEs. Top strips indicate the inserted transposon. Right strips indicate the species that have the insertion of the transposon. Expression is represented in logarithm of the RPKM + 1e-5. We added 1e-5 as a pseudocount to avoid the infinite values after the logarithm transformation. Text in red represent the mean of the distribution. Text in blue show the number of observations in each boxplot. The number on the boxplot are p-values of two-sided Wilcoxon signed-rank tests.

A



B



Supp. Figure 17. IGV snapshots of *de novo* intergenic clusters predicted in *Mus musculus* (A) and *Mus caroli* (B). Antisense IAP insertions are shown in green. Most small RNA-seq reads map downstream of the IAP insertions.

6.3. Methods

Code and software

Supp. tables 5 and 6 show the versions, parameters and the references of the software used for some of the processes in our analyses.

Statistics and plotting

Statistical analyses and plotting were performed in R (v4.1.0) and RStudio (v1.4.1717), except those analyses and plots done by external software. Unless otherwise stated, all the performed *t tests* and *Wilcoxon tests* were two-sided.

Assemblies and annotations

Unless otherwise specified, the used reference genomes and genome annotations were downloaded from ENSEMBL release 104 (*Howe et al., 2021*), which corresponds with the following assemblies: GRCm39 (primary assembly) for the *Mus musculus* C57BL6 strain (common name: mouse or house mouse), CAROLI_EIJ_v1.1 (top level assembly) for *Mus caroli* (common name: Ryukyu mouse) and PAHARI_EIJ_v1.1 (top level assembly) for *Mus pahari* (common name: Shrew mouse). Each reference genome was indexed using `bowtie build` (*Langmead et al., 2009*).

In some cases (i.e. UCSC…), the assemblies for *Mus caroli* and *Mus pahari* can also be found under the identificators GCF_900094665.1 and GCF_900095145.1, respectively.

RepeatMasker annotations were downloaded from UCSC Table Browser (*Karolchick et al., 2004*). In the case of *Mus pahari* and *Mus caroli*, each type of repeat (i.e. SINE, LINE, LTR…) was separated in different files, so they were formatted and merged into one annotation using a custom bash script.

Small RNA-seq data processing

Total RNA was extracted from adult testis of *Mus musculus* (C57BL6 strain), *Mus caroli* and *Mus pahari*, with 2 biological replicates for each species. TruSeq small mRNA kit from Illumina was used for library preparation. Sequencing was done single-end with HiSeq-2500 from Illumina with ~42 million reads as target sequencing depth. Raw reads were provided in FASTQ files.

Sequencing adaptors were trimmed from raw reads using `Cutadapt` (*Martin, 2011*) and the resulting trimmed reads were further filtered based on quality using `fastq_quality_filter` from the `FastX` toolkit (*http://hannonlab.cshl.edu/fastx_toolkit/*).

Then, filtered reads were mapped to the corresponding reference genome using `bowtie` (*Langmead et al., 2009*) and, for multi-mapping reads, only one alignment was

provided randomly (*--seed 666*). The final output in SAM format was converted to BAM, sorted and indexed using `samtools` (*Li et al., 2009*).

These BAM files were split with `samtools` to separate reads from the minus and the plus strands. Then BigWig files were created from these split files with `Deeptools'` (*Ramírez et al., 2014*) `bamCoverage` function, using CPM as normalization method and an effective genome size calculated with `faCount` from UCSC Kent's tools.

Small RNA-seq quality controls

`FastQC` (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to perform a quality control of the reads in raw, adaptor-trimmed and quality-filtered FASTQ files.

The script `read_distribution.py` from `RSeQC` (*Wang et al., 2012*) was used to count how many reads mapped to different genomic regions (i.e. exons, introns, UTRs, intergenic…) with the BAM files as input.

Custom bash scripts were used to retrieve the composition of the 1st and 10th nucleotides, and the length of each read in the filtered FASTQ files and custom R scripts were used to generate the corresponding plots.

`TBr2_pinpong.pl` from the *NGS toolbox* (*Rosenkranz et al., 2015*) was used to retrieve the significance of the ping-pong pathway in our samples and, again, R scripts were used to generate the corresponding plots.

Finally, `MultiQC` (*Ewels et al., 2016*) was used to generate a multi-sample quality report with the data from `FastQC`, `read_distribution.py` bowtie and `featureCounts` (*Liao et al., 2013*).

To show read information for each sample, we used custom R code to create the data in *supp. table1*.

piRNA cluster annotation

The piRNA cluster annotation used in this project was first defined by *Li et al.* (*2013*) in *Mus musculus* (NCBI37/mm9) and recently refined by *Yu et al.* (2021) (GRCm38/mm10), from which our annotation was retrieved.

The piRNA precursor annotations from *Li et al.* (*2013*) and *Yu et al.* (2021) were defined by using data from diverse sequencing assays in several developmental stages of mouse testes. RNA-seq was to annotate the mouse testis transcriptome and transcripts with 100 rpkm of unique mapping piRNAs (obtained by small RNA-seq) were selected for manual annotation. The boundaries of the piRNA producing transcripts were further refined using CAGE-seq, RNA polymerase II ChIP-seq and H3K4me3 ChIP-seq for the 5' ends, and PAS-seq for the 3' ends (*Li et al., 2013*). *Yu et*

al. also used ChIP-seq data for A-Myb and Btbd18 transcription factors to improve the annotation (2015).

The original file included 467 transcripts from 215 piRNA clusters (464 and 214 in *Li et al.* (*2013*)), whose coordinates were collapsed using `bedtools merge -s` (*Quinlan et al., 2010*) to obtain the coordinates of the clusters. Then, since this annotation was provided in GRCm38/mm10 format, the coordinates were converted to GRCm39/mm39 using `liftOver` with default parameters.

piRNA clusters from *Yu et al. (2021)* were classified into 101 pachytene, 30 hybrid and 84 pre-pachytene clusters using the information provided in the original paper and in *Ding et al.* (2017). From the pachytene clusters, 30 were divergently transcribed in pairs from bidirectional promoters with a median distance of 127 bp (*Li et al., 2013*) between mates. This information was not provided in any of the annotations, so we used `bedtools closest -S -d` to find the pairs of closest clusters in different strands and then, pairs with distance between mates greater than 500 bp were removed. This returned 15 pairs (30 clusters) with a median distance of 123 bp between mates.

De novo piRNA cluster prediction

To do a prediction of piRNA clusters in each sample, we used the `proTRAC` pipeline, which includes read processing using the `NGS toolbox` (*Rosenkranz et al., 2015*) and the cluster prediction using `proTRAC` (*Rosenkranz et al., 2012*) itself (*supp. figure 7*).

First, `TBr2_collapse.pl` (*NGS toolbox*) removed redundant reads from the quality-filtered FASTQ files and added the read count information in the FASTQ header. `TBr2_duster.pl` (*NGS toolbox*) filtered the low-complexity reads from the collapsed FASTQs and the remaining reads were mapped to the reference genome using `sRNAmapper.pl` (*Rosenkranz et al., 2015*), returning all mapping reads in ELAND format. Alignments of multi-mapping reads were weighted depending on the transcription level of their regions using `reallocate.pl`. Finally, `proTRAC.pl` (*Rosenkranz et al., 2012*) was used to predict the clusters in each sample using the weighted ELAND file, a GTF genome annotation and a repeat masker annotation as inputs. Predicted clusters for each sample were outputted from `proTRAC` in GTF format.

BED files were generated from each GTF file using a custom bash script. Then, BED files for the two replicates from each species were intersected using `bedtools intersect` (*Quinlan et al., 2010*) returning the original coordinates of both replicates (`-wo`). The overlapping clusters with the same strandedness were then merged using `bedtools merge`. If a cluster had an 80% overlap with repeats from repeatMasker, it was removed.

De novo genic and intergenic clusters

To call predicted clusters as genic or intergenic, we used `bedtools intersect` to look at the protein-coding genes overlapping with piRNA clusters. Same strandedness (`-s`) and a minimum overlap of 25% were required (`-f 0.25`)

In the case of intergenic clusters, the flanking genes were obtained with `bedtools closest` using the flags `-iu` (i.e. ignore upstream) to get the downstream gene and `-id` (i.e. ignore downstream) to get the upstream gene.

Orthologous regions

To obtain the list of orthologous genes across *Mus musculus*, *Mus caroli* and *Mus pahari,* we used BioMart from ENSEMBL release 104 (*Howe et al., 2021*) to retrieve the *Mus musculus* gene names and id, and the gene ids of the orthologous genes in *Mus caroli* and *Mus pahari*. This process returned 24167 genes with orthologs present in the three species.

When other annotations were used (e.g. piRNA clusters), in order to be able to use them in all the species we searched for the orthologous regions. For this, two approaches were followed: (1) using `liftOver -minMatch 0.7` and (2) using ENSEMBL Compara Perl API (*Howe et al., 2021*) with the Murinae multiple alignment. From this last approach, we considered the whole region retrieved from the multiple alignment and the conserved blogs, which are those segments of the multiple alignment that have a match or mismatch -not gaps- for all the considered assemblies.

ENSEMBL Compara

We used ENSEMBL Compara Perl API to retrieve the multiple alignment of the Murinae species in the ENSEMBL release 104. Coordinates (1-based start) of desired regions and a query species (e.g. “Mouse”, “Mus musculus”, “Mus caroli”...) were given as input to the perl script, which returned the coordinates and aligned sequences for the input regions in all the assemblies in the alignment. Only those assemblies belonging to *Mus musculus* (CL57/BL6 strain), *Mus caroli* and *Mus pahari* were retained.

To retrieve the coordinates of the whole regions, custom bash and R scripts were used. Shortly, for each input region and each assembly, we took the coordinates (i.e. chromosome, start, end and strand) given by the ENSEMBL Compara Perl API. Since some regions were split in different alignment blocks, we used `bedtools merge` to merge the blocks belonging to the same input region. In case one of the input regions suffered a duplication, numbers were assigned to each duplication (e.g. “region__v1”) (*supp. figure 3A*).

Custom bash and R scripts were also used to retrieve the coordinates of the conserved blocks -segments of the sequence that had a match or mismatch (but not a gap) in all the assemblies-. Briefly, for each input region, we took the whole multiple alignment and assigned the position of each base, excluding the gaps, which were assigned a 0. If a position had a gap in one of the assemblies, that position was given a gap in all the assemblies, leaving blocks of equal length in all the assemblies (i.e. conserved blocks). Finally, the first and last position of each block were added to the start coordinate of the input region to get the coordinates of the conserved blocks. Finally, the coordinates of the conserved blocks were merged with `bedtools merge` and intersected with the

whole regions with `bedtools intersect` to give the correct id to each conserved block (*supp. figure 3B*).

Intersections with transposable elements

To study the effects of TE insertions in piRNA clusters, we retrieved LINEs, LTRs and SINEs from each repeatMasker annotation. Then, we intersected them with the piRNA cluster and each intersection was repeated three times in order to study the effects of the TE orientation: regardless of the strand, TE sense to the cluster and TE antisense to the cluster.

To do the intersection, in the case of *Yuetal* clusters and their orthologs (from ENSEMBL Compara) we intersected the piRNA cluster annotation with each TE annotation (e.g. LINE, LTR…) using `bedtools intersect`. Instead, in the case of the *de novo* clusters and their orthologs (from ENSEMBL Compara) we extended the search for TEs by 10kb to solve the fact that genic piRNA clusters are predicted mainly in the 3'UTR. To do so, we used `bedtools closest -d` to report the distance and retain only those features closer than 10kb.

Estimation of expression

To estimate the small RNA expression in the different features (i.e. genes, piRNA clusters), we used `featureCounts` (*Liao et al., 2013*) with `-Q 1` (minimum mapping quality of 1, which excludes multi-mapping reads) `-O` (count multi-overlapping reads) `-minOverlap 18` (minimum overlap between reads and features of 18 bp) and, unless otherwise stated, `-s 0` (count reads regardless the strand) and `-R BAM` (return the reads in BAM format when `-s 0`). Reads sense (`-s 1`) and antisense (`-s 1`) to the feature were also counted in separated analysis for which reads were not reported (`-R` was not specified).

Small RNA expression analyses

For small RNA expression analysis, raw counts from `featureCounts` were imported into R and normalized using reads per kilobase per million mapped reads (RPKM) or the `DESeq2` normalization method, depending on whether we were comparing different regions (e.g. pre-pachytene vs pachytene clusters) or the same regions across different species.

Differential expression (DE) analysis was done with `DESeq2` and to consider a feature as differentially expressed in a certain contrast the following conditions had to be satisfied: (1) Benjamini-Hochberg (*Benjamini et al., 1995*) -default p-value adjusting method in `DESeq2`- adjusted p-value lower than 0.05 and (2) a change in its expression greater than 2-fold change (i.e., |*log2(fold change)*| > 1).

Principal component analysis

Raw counts from `featureCounts` were imported into R, normalized and variance-stabilized with `DESeq()` and `rlog()` functions from `DESeq2` R package (*Love et al., 2014*). To perform the principal component analysis, these normalized variance-stabilized counts were used as input for the `pcaplot()` from the `pcaExplorer` R package (*Marini et al., 2019*), using the 500 top variable genes/regions or all of them if there were less than 500.

6.4. Work plan

We developed a work plan to achieve the defined objective on time, with a detailed list of tasks (see *Section 6.4.1. Tasks*) as well as a calendar to have a good control of the time (see *Section 6.4.2. Calendar*). Dates for the several proposed milestones were also established (see *Section 6.4.3. Milestones*) and we identified possible risks and

6.4.1. Tasks

The defined tasks and subtasks are listed below:

1. Assessment of the quality of the small RNA-seq data:
    1.1. Pre-processing:
        1.1.1. Trim raw reads to remove adaptors.
        1.1.2. Filter trimmed reads to remove low-quality reads.
        1.1.3. Map filtered reads to the reference genome.
    1.2. Quality control of the data:
        1.2.1. Perform FastQC analysis on the raw, trimmed and filtered reads.
        1.2.2. Perform read distribution analysis on the aligned reads.
        1.2.3. Run MultiQC to merge all previous reports in one.
        1.2.4. Retrieve length, first nucleotide and tenth nucleotide of each read and compute proportion of U in the first nucleotide and A in the tenth.
2. Analysis of gene expression:
    2.1. Find orthologous genes and piRNA clusters.
    2.2. Count reads mapping to and orthologous genes piRNA clusters.
    2.3. Perform exploratory analysis.
        2.3.1. Normalize raw counts.
        2.3.2. Perform PCA and plot results.
        2.3.3. Draw heatmaps and boxplots of gene expression.
        2.3.4. Compute distances between samples.
    2.4. Perform differential expression analysis with DESeq2.
        2.4.1. Perform pairwise contrasts.
        2.4.2. Draw diagnostic plots (e.g. volcano plots…).
        2.4.3. Iterate through tasks 2.1 to 2.4 with diverse piRNA cluster annotations.

3. Test the link between transposable elements and piRNA expression:

    3.1. Intersect piRNA cluster annotations from task 2 with the repeats and transposable elements (e.g. from repeatMasker).

    3.2. Search species-specific repeats/transposable elements in orthologous regions.

    3.3. Test whether this species-specific repeats/transposable elements affect gene expression.

6.4.2. Calendar

Supp. Figure 18. Gantt chart with the time schedule of each task. Done with the online tool https://clickup.com/features/gantt-chart-view.

6.4.3. Milestones

We have defined two types of milestones in the project: (1) the ones that are crucial to the progress of the project and (2) the PAC milestones, which are linked to the writing of the thesis and need to be delivered on a determined date.

The main milestones in the project have suggested dates to make the project flow better and are the following ones:

- Milestone 1: Obtain annotations for orthologous piRNA clusters and genes across the three *Mus* species (27/10/2021).
- Milestone 2: Results of the differential expression analyses (06/12/2021).
- Milestone 3: Code, documentation and figures (14/12/2021)

The PAC milestones are:

- PAC 0 - Definition of the project contents (23/09/2021).
- PAC 1 - Definition of the work plan (04/10/2021).

- PAC 2 - First report (08/11/2021).
- PAC 3 - Second report (09/12/2021).
- PAC 4 - Final report/memory (24/12/2021).
- PAC 5a - Presentation (03/01/2022).
- PAC 5b - Defense (21/01/2022).

6.4.4. Possible risks

The possible risks that we have identified are the following ones:

- Data loss:

  Mitigation plan: in order to mitigate or avoid the data loss, all the data, code and documentation will be duplicated and stored in the local computer and, at least, another location. Scripts, Rmarkdown and HTML reports, as well as small intermediate files and figures will be stored in a private Gitlab repository, creating a website for the visualization of the data. Raw data files and big intermediate files (e.g. BAM files) will be stored separately in backup disks accessible from the cloud, as well as hard disk drives.

- Time shortage:

  Mitigation plan: we have planned tasks and milestones with enough time to do them, as well as possible reruns of the experiments.

- Multifactorial analysis may not be able to account for all factors:

  Mitigation plan: our data was obtained in one laboratory, from the same tissue (testis) and in one sequencing run. Hence batch effect and “tissue” variables are not present in our experiment, leaving the “species” as the only variable to account for.

- Not finding orthologous regions across multiple species:

  There are many concerns when referring to orthologous regions across different species. It is likely that a gene/region in one species (e.g. *Mus musuclus*) has duplicates in other species, making the comparison more difficult. Also, even if a gene/region has one orthologous locus in another species, it is likely that the query gene in the first species and the orthologous locus have different gene lengths, making differential expression analysis harder to carry out. We have defined several ways to define orthologous regions across different species, each of them with a different mitigation plan.

  1. Retrieving lists of orthologous genes across all the three studied species from ENSEMBL BioMart. Here we confront two problems: (1) having duplicates of one gene and (2) we are not able to find intergenic orthologous regions. To mitigate having duplicates (1), the genes with more than one copy in any species are removed from the whole list. However the orthologs from intergenic regions cannot be found using this approach.
  2. Using liftOver from UCSC to convert the coordinates of the regions from one species (e.g. *Mus musculus*) to the other. By default, liftOver returns only one output region for each input region, so the problem with

duplicates is not present. However, the *chain files* (derived from the pairwise alignment between the two desired species) which liftOver uses to do the conversion may not be created. To mitigate this, we have already asked for the creation of these *chain files*. But the difference in gene length cannot be approached and we may not retrieve orthologs for all the desired regions.

3. To account for the gene length and also the problem of the duplicates, ENSEMBL Compara can be used to retrieve the multiple alignment of all the Murinae species, obtain the coordinates of the desired regions in all the species and then retrieve the coordinates of the conserved segments where there is no gap in any species. This will result in several segments per region, which will have the same or very similar length in all the species, resolving the problem of gene length.

6.5. Additional files

Some supplementary files used in the analyses will be available upon request. The list of these supplementary files is shown below:

Additional file 1: `yuetal_clusters.xlsx` – Clusters from *Yu et al.* (2021), their coordinates (original and orthologous regions from ENSEMBL Compara) and other information like the class and the directionality. The coordinates of the orthologs obtained with ENSEMBL Compara, the counts in the conserved blocks and the results for the differential expression analyses.

Additional file 2: `protrac_clusters_musculus.xlsx` – *de novo* clusters predicted in *Mus musculus,* their coordinates (original and orthologous regions from ENSEMBL Compara) and other information like the class and the directionality. The coordinates of the orthologs obtained with ENSEMBL Compara, the counts in the conserved blocks and the results for the differential expression analyses.

Additional file 3: `protrac_clusters_caroli.xlsx` – *de novo* clusters predicted in *Mus caroli,* their coordinates (original and orthologous regions from ENSEMBL Compara) and other information like the class and the directionality. The coordinates of the orthologs obtained with ENSEMBL Compara, the counts in the conserved blocks and the results for the differential expression analyses.

Additional file 4: `protrac_clusters_pahari.xlsx` – *de novo* clusters predicted in *Mus pahari,* their coordinates (original and orthologous regions from ENSEMBL Compara) and other information like the class and the directionality. The coordinates of the orthologs obtained ENSEMBL Compara, the counts in the conserved blocks and the results for the differential expression analyses.

Additional file 5: `ensembl_compara_perl_api.pl` – Perl script to access the ENSEMBL Compara database and retrieve the coordinates of the desired regions from the Murinae multiple alignment.

Additional file 6: `ensembl_compara_get_coords.sh` – Bash script that modifies the output of `ensembl_compara_perl_api.pl` and calls two R scripts to retrieve the whole regions and the conserved blocks from ENSEMBL Compara.

Additional file 7: `ensembl_compara_whole_regions.R` – R script to retrieve the whole regions using the output of `ensembl_compara_perl_api.pl`.

Additional file 8: `ensembl_compara_consblocks.R` – R script to retrieve the conserved blocks using the output of `ensembl_compara_perl_api.pl`.