Universitat Oberta de Catalunya

uoc.edu

Analysis of variation in PIWI-interacting RNA (piRNA) expression in testes of different mouse strains.

Pío Sierra
Máster universitario en Bioinformática y bioestadística UOC-UB

Tanya Vavouri
Marc Maceira Duch

January 5, 2021

Project	Analysis of variation in PIWI-interacting RNA (piRNA) expression in testes of different mouse strains.
Author	Pío Sierra
Tutor	Tanya Vavouri
PRA	Marc Maceira Duch
Date	2021-01-05
Studies	Máster universitario en Bioinformática y bioestadística UOC-UB
Area	TFM-Estadística y Bioinformática
Language	English
Keywords	piRNAs, transposons, IAPs

Abstract

Piwi-interacting RNAs (piRNAs) are a class of small non-coding RNAs present in the germline of most animals which are responsible for silencing transposable elements (TEs) through base-pair complementarity. We performed an extensive study of the expression of piRNAs in the male germline of five inbred mouse strains. We tested for variation in piRNA expression between different mouse strains. Furthermore, we tested whether variation in transposon copies between mouse strains affects the expression of piRNAs. We analyzed piRNA expression from previously known piRNA clusters and also from piRNA clusters that we predicted *de novo*. The results, with the documented clusters as well as with the *de novo* ones show a significant correlation between piRNA differential expression and the different status of a very young retrotransposon, the murine intracisternal A-particle (IAP) in the strains being compared. We found that the presence of an IAP in the region being expressed, in just one of the strains, (±10kb) was highly correlated with differential expression of the cluster between both strains. To the best of our knowledge this is the most extensive study in which variation in piRNA clusters expression in testes between genetically diverse individuals has been reported in any mammalian species. This work suggests a mechanism for piRNA evolution and piRNA biogenesis through endogenous retrovirus insertions in genes and piRNA clusters.

Abstract - Español

Los ARN asociados a Piwi (piRNAs) son una clase de pequeños ARN no codificante presentes en la línea germinal de la mayoría de animales que son responsables de silenciar transposones (TEs) a través de complementariedad de bases. Hemos llevado a cabo un extensivo estudio de la expresión de piRNA en la línea germinal masculina de cinco líneas puras de ratones y hemos testeado la variación en la expresión de piRNA entre las distintas líneas. También hemos testeado si la variación entre las copias de transposones presentes en las distintas líneas de ratón afectaba a la expresión de piRNAs. Para ello hemos analizado la expresión de piRNA en agrupaciones (clusters) ya conocidos previamente y también en agrupaciones que hemos predicho *de novo*. Los resultados, tanto con los clusters ya documentados como con los predichos *de novo*, muestran una correlación significativa entre la expresión diferencial de piRNA y la diferente presencia en las dos líneas comparadas de un transposón muy joven (murine intracisternal A-particle, IAP). La presencia de un IAP en la región expresada (±10kb) en solo una de las dos líneas, está altamente correlacionada con una expresión diferencial del cluster correspondiente entre las dos líneas. Hasta donde hemos podido comprobar este es el estudio más extenso llevado a cabo en el que variación en la expresión de clusters de piRNAs ha sido documentada en cualquier especie de mamíferos. Este trabajo sugiere un mecanismo para la evolución y biogénesis de piRNAs a través de inserciones de retrovirus endógenos en genes y clusters de piRNA.

Contents

List of Figures

1 Introduction

1.1 Context and motivation behind the work

1.1.1 piRNAs

Piwi-interacting RNAs (piRNAs) are a class of small non-coding RNAs present in the germline of most animals. They are responsible for silencing transposable elements (TEs) through base-pair complementarity. piRNAs are 24-31 nucleotides long and their sequences present a bias to start with uracil (U) (aprox 75%). They were first identified as a novel class of small interfering RNA (siRNAs), and then associated with transposon silencing in most sexually reproducing animals. The biogenesis of piRNAs is different from the biogenesis of siRNAs and microRNAs (miRNAs) in that it is Dicer independent [Vagin et al., 2006],[Houwing et al., 2007]. piRNAs generally derive from long single-stranded precursor transcripts. [Vagin et al., 2006]. piRNA precursors are transcribed from genomic loci known as piRNA clusters. At least in the mouse genome, transcripts that produce piRNAs can be from protein-coding genes as well as from non-coding transcripts and from intergenic transposon insertions. piRNA sequences are very diverse and rarely conserved among species (reviewed in [Ozata et al., 2019]).

piRNAs bind PIWI proteins, a subfamily of the Argonaute family of proteins, which are expressed mainly in germ cells[Cenik and Zamore, 2011]. A unified model proposed for the biogenesis of piRNAs places PIWI-clade Argonautes at the center of it by performing an endonucleolytic cleavage that is the base for further stepwise slicing of the precursor transcripts (phased pre-piRNAs), and the pingpong effect, in which the production of two 10 bases complementary piRNA on opposite strands is amplified by the recursive action of two PIWI proteins, one acting on each strand [Gainetdinov et al., 2018]. Phased piRNAs create diversity, while the pingpong effect amplifies the production of piRNAs targeting TEs. It is essential that they can be expressed properly as the disruption of genes required to make pachytene piRNAs (piRNAs expressed mainly during the third stage of the prophase of meiosis) prevents production of mature sperm [Aravin et al., 2008].

1.1.2 Transposable elements

Transposable elements (TEs) are DNA sequences that have the ability to change their position within a genome either by a copy-paste or by a cut-paste mechanism [Bourque et al., 2018]. TEs are subdivided in several classes, including Long Terminal repeat (LTR) retrotransposons, which are classified in 21 different families [McCarthy and McDonald, 2004]. Of particular interest for this project is the LTR class of TEs and in particular the murine intracisternal A-particle (IAP) retrotransposon.

IAPs are endogenous retroviruses. The mouse genome contains around 10,678 IAPs [Elmer et al., 2020]. These elements contain 5' and 3' long terminal repeats (LTRs) with functional *gag*, *pro* and *pol* genes [Mietz et al., 1987].

They lack an extracellular phase due to the absence of a functional *env* gene [Mietz et al., 1987]. IAPs are very young transposons with evidence of recent activity in the mouse genome and are responsible for the most reported mutations in this genome [Gagnier et al., 2019]. IAPs can affect gene expression in multiple ways, at both the transcriptional and post-transcriptional level. For example, intronic IAPs can induce alternative RNA processing choices, including alternative splicing [Concepcion et al., 2015]. IAPs can also affect promoter activity, for example an IAP found in the promoter region of Stabilin2 (Stab2) likely interferes with normal expression in the liver sinusoidal endothelial cells [Maeda-Smithies et al., 2020]. An IAP has also been found to have an age-dependent effect on gene transcription, a non deleterious IAP in the first intron of the mouse gene Nocturnin was found to be expressed together with the gene and become active with aging [Barbot, 2002]. Some IAP TEs in mice exhibit inter-individual variability in cytosine methylation (VM IAPs), a phenomenon largely limited to IAPs after screening all TEs [Elmer et al., 2020].

The different IAP copies of the mouse genome vary according to their age and strain of origin. A study comparing IAPs in strains 129 and C57BL/6 found that half of the IAP elements obtained from embryonic stem cells derived from the first strain were absent from the second [Horie et al., 2007], suggesting many other IAPs might be present in the other strains. This was later confirmed in a comprehensive catalogue of 103,798 polymorphic TE variants generated using whole genome sequencing data from 13 classical laboratory and four wild derived mouse inbred strains [Nellåker et al., 2012]. Of the 13,666 IAPs found in total only 2,684 were present on the mouse reference (C57BL/6J) while 10,976 where not present on C57BL/6J but were present in some of the other 12 strains.

One relevant aspect of IAPs insertions is that the effect they have on genes can be reversed by a mutation in the nuclear RNA export factor 1 (Nxf1). One potential explanation of this effect is that it affects splicing by an unknown mechanism, although its activity in nuclear export cannot be ruled out either [Concepcion et al., 2009] [Concepcion et al., 2015]. A natural mutant variant of Nxf1 is found in the inbred mouse strain CAST/EiJ, one of the strains used in this work.

1.1.3 piRNAs and TEs

The TE silencing function of piRNAs seems to be conserved across all the species where piRNAs have been studied. The disruption of the piRNA pathway results in the activation of TEs in male mice [Kuramochi-Miyagawa et al., 2004], [Carmell et al., 2007]; male and female fruit flies [Wilson et al., 1996] and male and female zebrafish [Houwing et al., 2008]. Studies in fruit flies have revealed that a small number of genomic loci that produce piRNAs act as “archives” of past TE invasions and provide a memory of sequences that the piRNA pathway can detect and repress. Furthermore, in chickens new piRNAs have been detected in response to an infectious retroviral invasion. Also, there is some evidence that piRNAs associated with a TE erode away as the TE loses its activity [Sun et al., 2017]. This could be a hint that activation of piRNA clusters can be influenced by the activity of TEs.

1.1.4 piRNAs clusters

Mammalian piRNAs are expressed from a couple of hundred of genomic loci. Here we make use of a transcript-based set of piRNA clusters by [Li et al., 2013]. In their study three types of piRNA-producing clusters were identified, according to when their piRNAs first accumulate and how their expression changes during spermatogenesis: 84 pre-pachytene, 100 pachytene, and 30 hybrid loci. piRNAs start to be expressed in the developing germline of the male fetus and continue being expressed in the adult throughout the life of the animal. These pachytene piRNAs typically derive from intergenic regions [Li et al., 2013].

The distribution and extent of piRNA clusters is very relevant for an organism. Following a simple population genetics model the total size of the piRNA clusters of an organism must exceed 0.2% of a genome to repress TE invasions. Moreover, larger piRNA clusters accounting for up to 3% of the genome may be necessary when populations are small, transposition rates are high, and TE insertions are recessive [Kofler, 2020].

piRNAs have so far been analysed only from one mouse strain, the reference strain. The extent of piRNA expression variation between mouse strains is therefore unknown. Here we analyse the expression of piRNAs clusters in the male germline of five inbred mouse strains in 214 previously identified piRNA clusters and a total of 611 de novo predicted clusters. We found significant variation in expression between strains, providing evidence that genetic variation affects piRNA expression in mice. Since the function of piRNAs is tightly linked to transposable elements and transposable elements vary between strains, we tested whether variation in piRNA cluster expression correlates with variation in TEs. The results with both sets of piRNA clusters show a significant correlation between piRNA differential expression and the different status of an IAP in the strains being compared. We found that the presence of a transposon of that type in the region being expressed (±10kb) disrupted the regular expression of the cluster. To the best of our knowledge this is the first study in which a variation in piRNA expression between genetically diverse individuals has been reported in any mammalian species.

1.2 Objectives

The aim of this project is to test for variation in piRNA expression between mouse strains and test whether variable transposable elements are associated with this variation. Specifically, at the start of the project we set the following objectives.

1.2.1 General Objectives

1. Qualitative evaluation of the amount of variation in expression explained by each variable.

2. Identification of clusters over or under expressed due to genetic differences.

3. Perform an enrichment test, if we find over or under expressed clusters, to find if they overlap more polymorphic transposons than the rest of the piRNA clusters.

1.2.2 Specific Objectives

1.1 Quality check of all the data. Identify any possible outliers that should be excluded.

1.2 Prepare data and design matrix for the experiment.

1.3 Identify common patterns on each of the clusters with principal components analysis (PCA). Create visualizations for them.

2.1 Create a multiple regression experiment that can be used to identify the set of variables that provide the better explanation for the data.

2.2 Examine the raw data and use diagnostic plots to evaluate the results.

3.1 Perform an enrichment of the data to study with variation of transposons.

3.2 Document and create relevant data files for all the analyses performed, including the necessary scripts for reproducible research.

1.3 Methods

The initial plan was to perform a differential expression analysis of piRNA clusters on 49 small RNA samples from testis, plus eight mouse small RNA samples from isolated spermatogonia. The eight spermatogonia samples were from two inbred strains; four CAST/EiJ (CAST) and four C57BL/6 (BL6). Of the first 49 samples, 10 are from four different inbred strains (two C57BL/6 (BL6), three C3H/HeJ (C3H), two NOD/ShiLtJ (NOD) and three 129S1/SvImJ (129); and 39 of them from an outbred strain ICR (ICR) (Table 1). This last set consists of heterogeneous samples generated for different purposes and they belong to three-generation pedigrees in which some of the progenitor mice received a high fat diet which resulted in a slightly diabetic phenotype that was transmitted to some of their offspring. Therefore the 39 mice were phenotypically and genetically diverse. Furthermore, the ICR mice were sequenced in two batches. The first step of this TFM was to analyse all the samples, including the samples from ICR mice to understand how much of the variation in piRNA cluster expression between the ICR mice is explained by confounding variables such as batch, diet and phenotype.

There are several challenges associated with small RNA-seq data analysis related to the small size of piRNAs, in addition to the ones regular RNA-seq provides, as described in [Conesa et al., 2016]. We used the workflow of DEseq2 described on [Love et al., 2014] and [Love et al., 2020a], which allows us to use un-normalized counts and also internally corrects for library size. To prepare the data we first counted reads on piRNA clusters. To this end we used cutadapt [Martin, 2011] to remove the adapter(TGGAATTCTCGGGTGCCAAGGAACTCCAGTCAC) and `fastq_quality_filter` [Hannon, a] to remove the low quality reads (at least 90% of bases on read with quality score > 30). As a further quality check we used standard shell tools to extract counts on the first nucleotide composition for the reads (piRNAs are expected to start predominantly with uracile "U"). We used `bowtie` [Langmead, 2010] to map the reads to the primary assembly of GRCm38 and then we used `samtools` [Li et al., 2009] to sort and filter the bam files. Finally we used `featureCounts` [Liao et al., 2014], from the Subread package to create the counts for the clusters and an R script to merge them into a single file for the differential expression analysis. This same workflow was repeated for each batch of reads and each set of clusters used.

A full description of all the tools used on the different pipelines can be found in Appendix C.

1.4 Work plan

We prepared a work plan that included the need to iterate some of the activities and create specific tasks out of the general objectives for the deliverables that have to be produced. We included the deliverables as tasks, but we did not include specific times for them as they were being prepared in parallel with the other tasks. The full work plan including tasks, milestones, gantt charts and risks analysis can be found in Appendix D.

1.5 Summary of the results

The results of the project are the current document and a private github repository with all the scripts, results files and figures shared with the tutor (that will be made public once the results included in this report are sent for publication).

https://github.com/vavouri-lab/picVar-TFM

1.6 Description of chapters

In the rest of the chapters we describe the tasks performed and the findings of the project:

- Data preparation: Description of the different samples and other sources of information used.
- Exploratory analysis: Evaluation of the basic characteristics of the raw data and readiness for the rest of the tasks.
- Differential Expression Analysis (1): Results from the first run of Differential Expression Analysis using the gold-standard piRNA clusters from [Li et al., 2013].

- *De novo* prediction of piRNA clusters: Prediction of piRNA clusters from testis and spermatogonia smallRNA samples from five different inbred mouse strains.
- Differential Expression Analysis (2): Results from the second run of Differential Expression Analysis using the de novo clusters we generated previously.
- Discussion: Analysis of the results.

2 Project design and execution

2.1 Data preparation

In the present study we use the extensive annotation of mouse variable TEs from [Nellåker et al., 2012] to test if TE insertions and deletions have an effect on piRNA abundance at 214 piRNA clusters identified in [Li et al., 2013] and piRNA clusters predicted *de novo* by us. We analyzed data from 3 batches of small RNA data from a total of 57 mouse samples; 18 belonging to five different inbred strains, six C57BL/6 (BL6), four CAST/EiJ (CAST), three C3H/HeJ (C3H), two NOD/ShiLtJ (NOD) and three 129S1/SvImJ (129); and 39 outbred ICR (ICR) samples (Table 1). The data was processed following the workflow shown in Figure 1.

After the file for the total counts was prepared, we performed several quality checks on the cutadapt results for six of the samples. The results can be found in Appendix E. Additionally some plots were manually generated to check the nucleotide composition of the reads. All the results were consistent with data from small RNA-Seq enriched for piRNAs. We also wanted to confirm the special characteristics on our piRNAs reads according to the models of piRNA biogenesis [Gainetdinov et al., 2018]. As expected for piRNA enriched samples, we detected the documented bias for uracile at the first nucleotide [Stein et al., 2019] (Figure 2). We also found the pingpong signature (Figure 3) that is characteristic of one of the paths for piRNA generation. The pingpong signature is enriched for A at position 10. This comes as a result of the complementarity of the bases and the bias for U on the first base of the opposing piRNA. Finally we checked the resulting counts for missing values or outliers that could be a signal for problems with the data, but none were found.

We confirmed with a Fligner-Killeen test that the data was not homoscedastic, as expected for RNA_Seq data, where the variance grows with the mean [Love et al., 2020b]. This fact is handled later by DESeq2 directly, working with raw counts, but it means that in order to apply other types of analysis (principal component analysis (PCA), sample-distances) it is necessary to perform a variance transformation, for which we have selected the `rlog` transformation.

SampleID	Strain	Tissue	MouseID	AncestralDiet	MetabolicState	Batch
sample01	ICR	testis	20.3	CC	1	1
sample02	ICR	testis	25.4	CC	1	1
sample03	ICR	testis	20.3.1	CC	0	1
sample04	ICR	testis	25.4.1	CC	NA	1
sample05	ICR	testis	20.3.2.2	CC	0	1
sample06	ICR	testis	25.4.3.6	CC	0	1
sample07	ICR	testis	21.2	CC	5	1
sample08	ICR	testis	23.3	CO	5	1
sample09	ICR	testis	22.2.1	CO	4	1
sample10	ICR	testis	21.3.2.5	CO	4	1
sample11	ICR	testis	23.3.1.1	CO	1	1
sample12	ICR	testis	22.2	CO	5	1
sample13	ICR	testis	21.3.4	CO	1	1
sample14	ICR	testis	23.3.3	CO	3	1
sample15	ICR	testis	22.2.3.2	CO	4	1
sample16	ICR	testis	5.5	CC	0	2
sample17	ICR	testis	5.5.1	CC	0	2
sample18	ICR	testis	5.5.2	CC	0	2
sample19	ICR	testis	5.1	CC	1	2
sample20	ICR	testis	5.4.3	CC	NA	2
sample21	ICR	testis	25.4.4	CC	NA	2
sample22	ICR	testis	5.2	CC	0	2
sample23	ICR	testis	20.4	CC	2	2
sample24	ICR	testis	20.3.2.1	CC	0	2
sample25	ICR	testis	20.3.5.5	CC	1	2
sample26	ICR	testis	5.4	CC	1	2
sample27	ICR	testis	5.3	CC	0	2
sample28	ICR	testis	21.1	CO	2	2
sample29	ICR	testis	21.3	CO	NA	2
sample30	ICR	testis	23.2	CO	4	2
sample31	ICR	testis	23.4	CO	4	2
sample32	ICR	testis	24.1	CO	4	2
sample33	ICR	testis	24.4	CO	4	2
sample34	ICR	testis	21.3.3	CO	1	2
sample35	ICR	testis	22.2.4	CO	3	2
sample36	ICR	testis	24.1.1	CO	NA	2
sample37	ICR	testis	25.4.3.2	CO	0	2
sample38	ICR	testis	22.2.3.5	CO	0	2
sample39	ICR	testis	24.1.5.3	CO	1	2
sample40	BL6	spermatogonia	1	CC	0	3
sample41	BL6	spermatogonia	2	CC	0	3
sample42	BL6	spermatogonia	3	CC	0	3
sample43	BL6	spermatogonia	4	CC	0	3
sample44	CAST	spermatogonia	1	CC	0	3
sample45	CAST	spermatogonia	2	CC	0	3
sample46	CAST	spermatogonia	3	CC	0	3
sample47	CAST	spermatogonia	4	CC	0	3
sample48	129	testis	1	CC	0	3
sample49	129	testis	2	CC	0	3
sample50	129	testis	3	CC	0	3
sample51	C3H	testis	1	CC	0	3
sample52	C3H	testis	2	CC	0	3
sample53	C3H	testis	3	CC	0	3
sample54	NOD	testis	1	CC	0	3
sample55	NOD	testis	2	CC	5	3
sample56	BL6	testis	6.3.1	CC	0	3
sample57	BL6	testis	6.3.2	CC	0	3

Table 1: Full set of samples used during the project. The most relevant variables are Strain and Tissue. AncestralDiet refers to the type of diet received by the ancestor (CC: control, CO: overfed) MetabolicState is the phenotypic trait of diabetic level observed on the sample (0: normal, 5: acute)

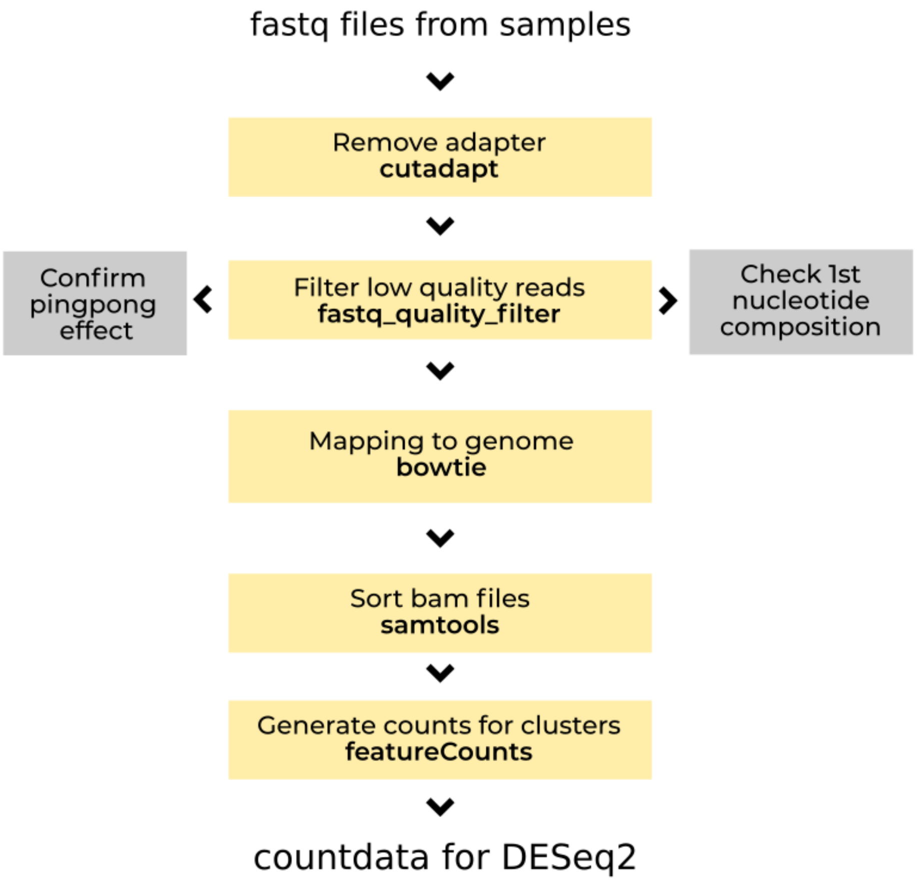


Figure 1: Workflow to prepare the counts for DESeq2 from the sequencing data.

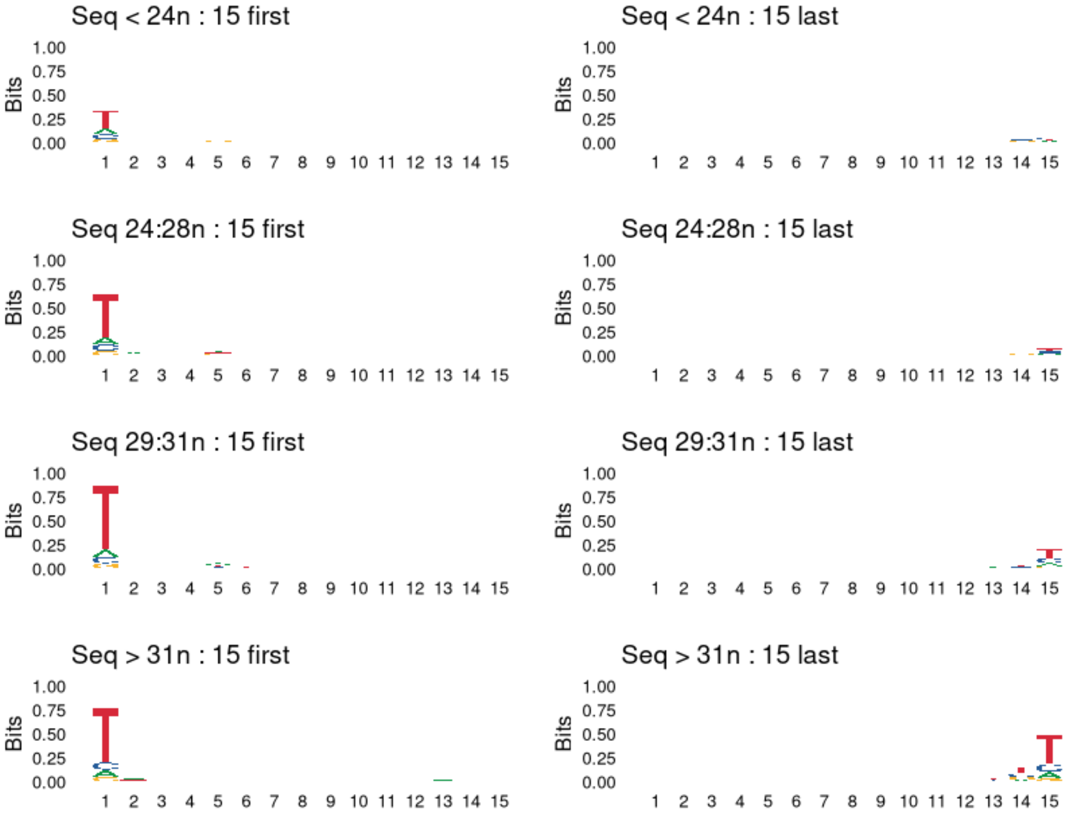


Figure 2: Composition of the first 15 nucleotides and the last 15 nucleotides of piRNAs from one sample by length of the sequence. The rest of the samples looked similar.

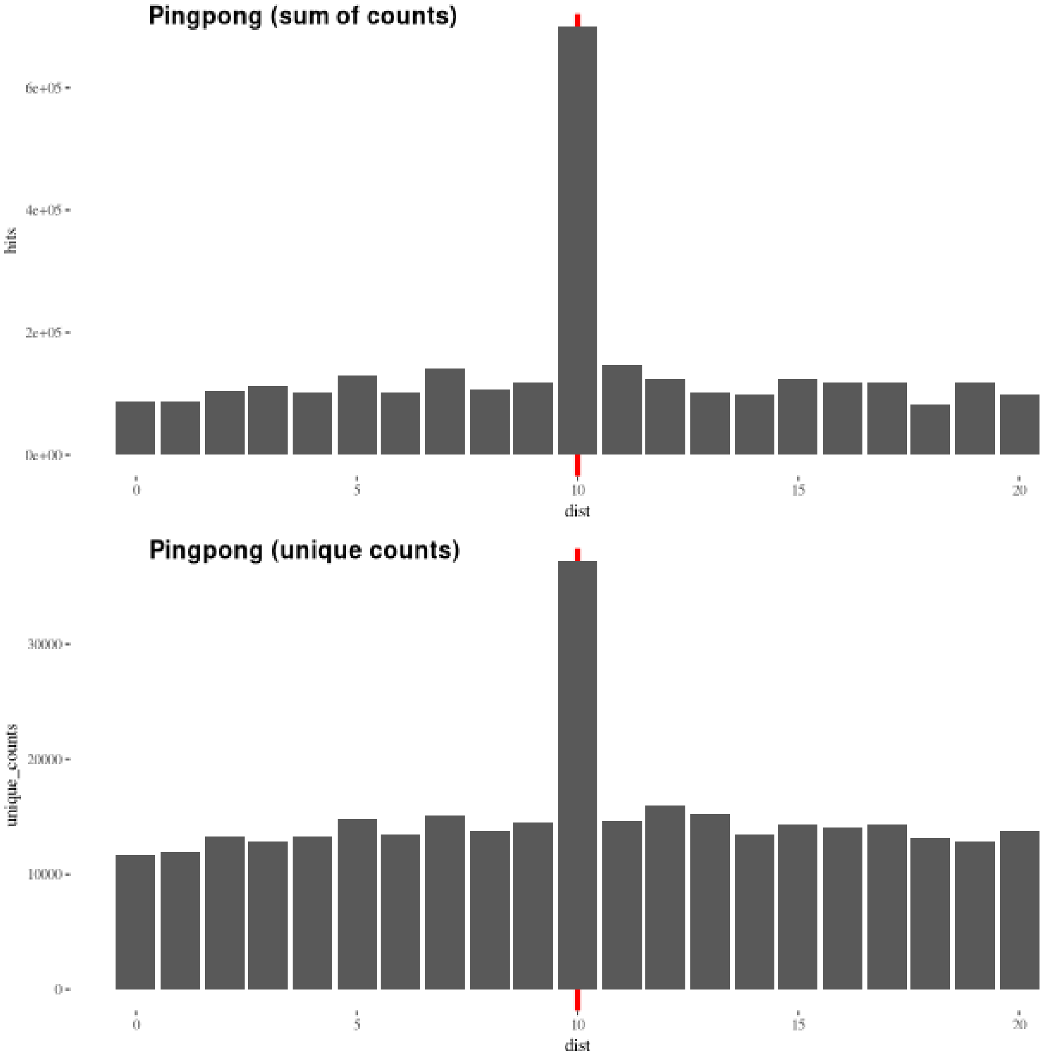


Figure 3: Distance between the 5' end of overlaping piRNAs on opposite strands for sample 57. For clusters with >20k reads a random sample of 20K reads was used.

2.2 Exploratory analysis

To analyse the variation in piRNA expression from known piRNA clusters among all the 57 samples we followed the standard DESeq2 [Love et al., 2014] workflow as described in [Love et al., 2020a]. We first checked for the relevance of different potential confounding factors on differential expression. We removed the samples with missing values for some of the factors (five had missing Metabolic State and/or Ancestral diet), and applied the `rlog` transformation to check the potential relevance of each factor using PCA. The effect of the tissue (testis/spermatogonia) had the largest weight by far (Additional figures 13 and 14). This clustering of the data was expected since spermatogonia RNA samples contain predominantly 'prepachytene' type of piRNAs while whole testes contain predominantly 'pachytene' type of piRNAs [Li et al., 2013]. After that we proceeded analysing the data from spermatogonia and testes separately. The results from PCA and sample distances for testes revealed that strain was the factor that explained most of the remaining variance (Figure 4 and Additional figure 15), with samples from inbred mouse strains (BL6, 129, C3H and NOD) clustered tightly together. Inbred mouse strains are nearly genetically identical

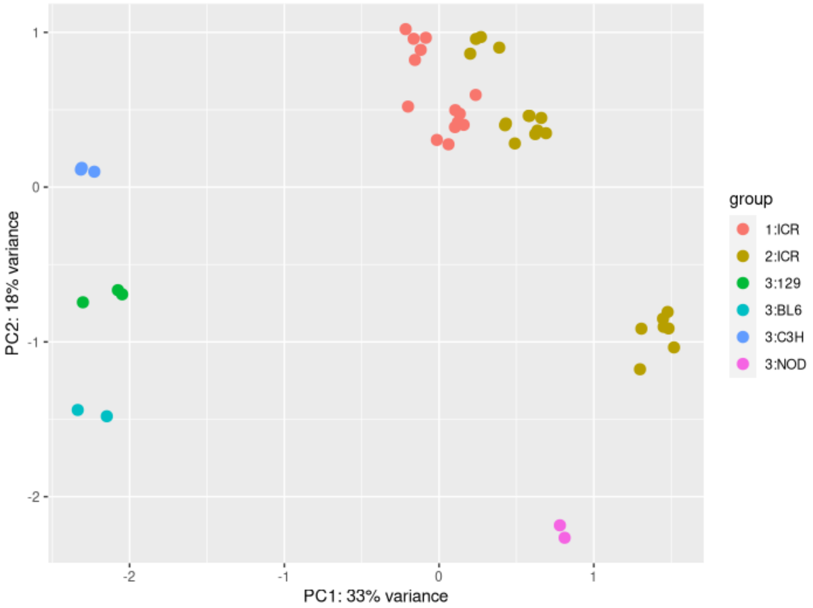


Figure 4: Principal components analysis of small RNAs mapping to known piRNA clusters in testis samples, coloured by batch and strain.

which suggests that genetic differences between mouse strains drive diverged expression of at least some piRNA clusters. Samples from the outbred mouse strain (ICR) were less tightly clustered. Mice of the ICR strain are genetically diverse. Altogether the data from PCA thus suggest that genetic differences between mice underlie differences in piRNA cluster expression.

2.3 Differential Expression Analysis (1)

Taking into account results of the exploratory analysis we chose to go ahead first with the testis samples testing for differential expression using just one factor at a time. From PCA, it seemed clear that strain was the main factor explaining the variance of the data, so we proceeded with the differential expression analysis on the samples from the five inbred strains (eight samples from spermatogonia from two inbred strains and ten samples from testes from four inbred strains). We first annotated the piRNA clusters with information on overlapping genes from Ensembl using the biomaRt R package [Durinck et al., 2009], repeats from the RepeatMasker [Smit, Hubley, R & Green, P., 2013] annotation from the UCSC Genome browser [Kent et al., 2002], refined repeats in mice from [Elmer et al., 2020] and finally with variable TEs from [Nellåker et al., 2012], which included information about the status of each TE on each inbred mouse strain (present, absent), which was the most important piece of information for our analysis.

Focusing just on testis samples of inbred strains allowed us to perform six contrasts. We selected from them the significant results as those that had ($log_2foldchange > 2$ and $p_{adj} < 0.05$). We selected this high fold change cut off because we were interested in large - not moderate - changes in expression between strains. Volcano plots on Figure 5 show the results for each contrast. One of the contrasts, C3H vs 129 did not produce any significant results with our thresholds. On the others we found two clusters showing particularly high differential expression ($log_2foldchange > 4$) which overlap the genes *Noct* and *Mrs2*, and two others which a lower differential expression just enough to pass the cut overlapping genes *Gm41109* and *Zbtb37*. A heatmap of the significantly differentially expressed piRNA clusters, shows also a dendrogram of the clustering for different piRNA clusters and strains according to the results (Additional figure 16).

We generally observed high correlation in piRNA cluster expression between all contrasts of inbred mouse strains. We also used the classification of clusters into prepachytene and pachytene piRNA clusters as defined by [Li et al., 2013] to highlight the clusters and we saw a large correlation between the level of expression and the type of cluster as defined by them according to the phase in which they found the cluster was more expressed (Figure 6).

From the pairwise comparisons on the testis samples for the four inbred strains we obtained four significantly differentially expressed piRNA clusters. Out of these four, the differential expression of the piRNA cluster overlapping *Noct* result was compatible with the hypothesis that the presence of a large intronic IAP in C57BL/6 and NOD/ShiLtJ could have a differential effect on

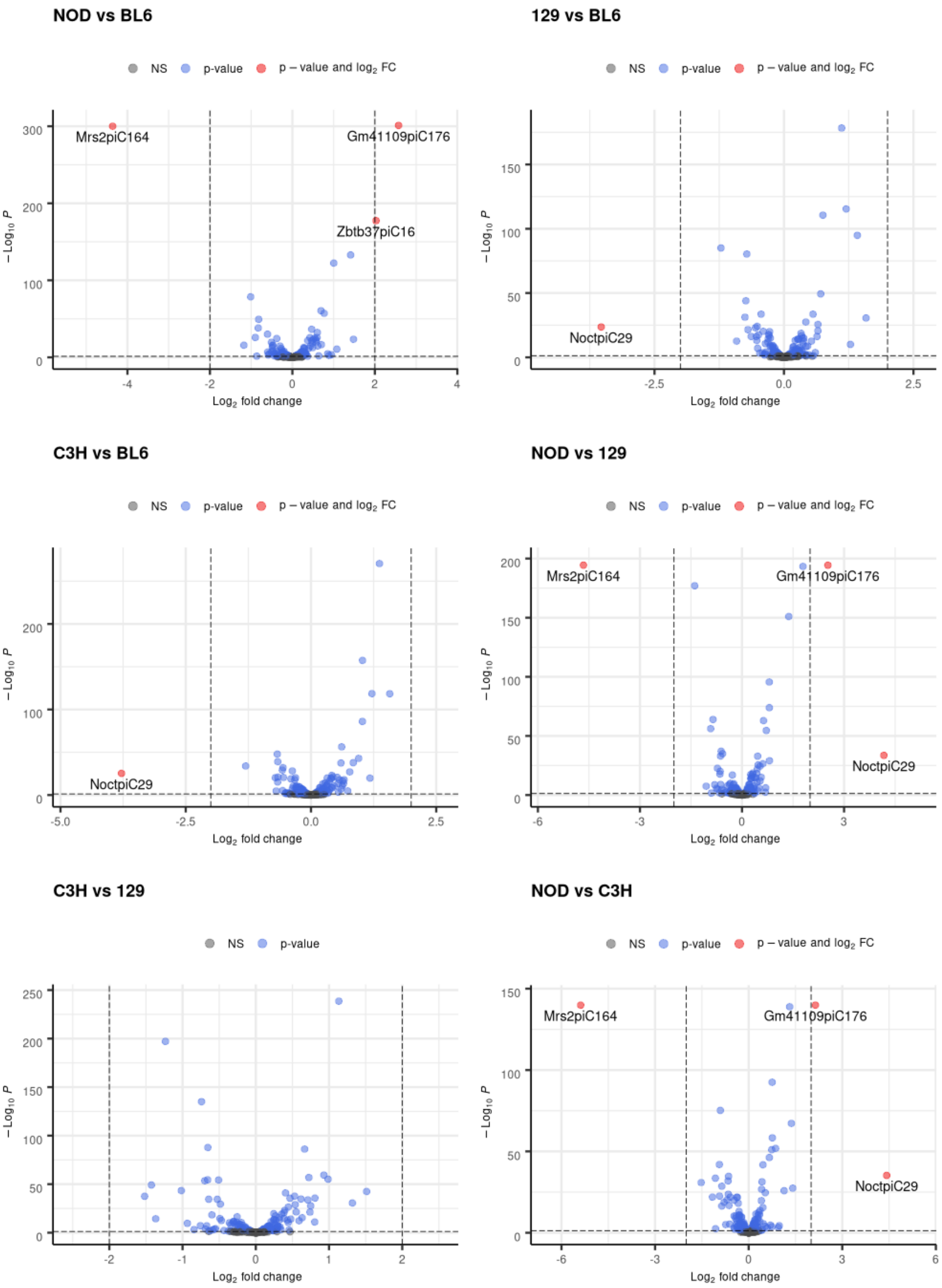


Figure 5: Volcano plots for the 6 contrast by strain performed on inbred samples.

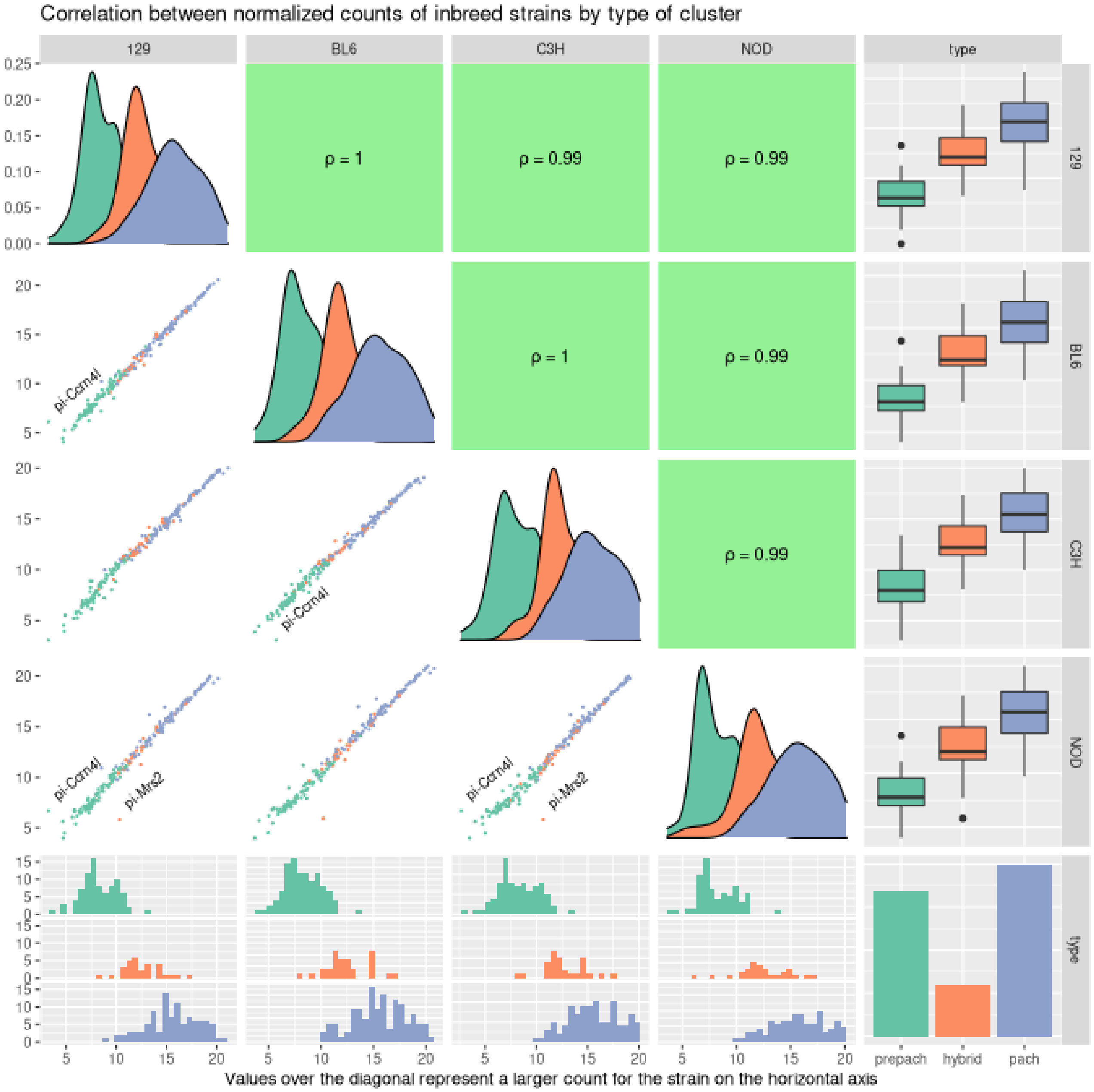


Figure 6: Correlation between normalized counts of inbred strains by type of cluster.

the expression of the cluster (Figure 7). Another cluster overlapping the gene *Mrs2*, shows a similar pattern (Additional figure 17), but the IAP present in the RepeatMaster annotation [Smit, Hubley, R & Green, P., 2013] is missing from the list curated in [Nellåker et al., 2012] that was the base for our analysis (Figure 8). After inspection of the region using genome browsers, we found that in NOD/ShiLtJ there is a large deletion (structural variant) annotated by The Mouse Genomes Project [Adams et al., 2015] that overlaps an annotated IAP so we came to the conclusion that indeed there is an IAP on three out of the four strains. More details on the position of the IAPs in these two genes can be seen in Figure 9 and Additional figure 20.

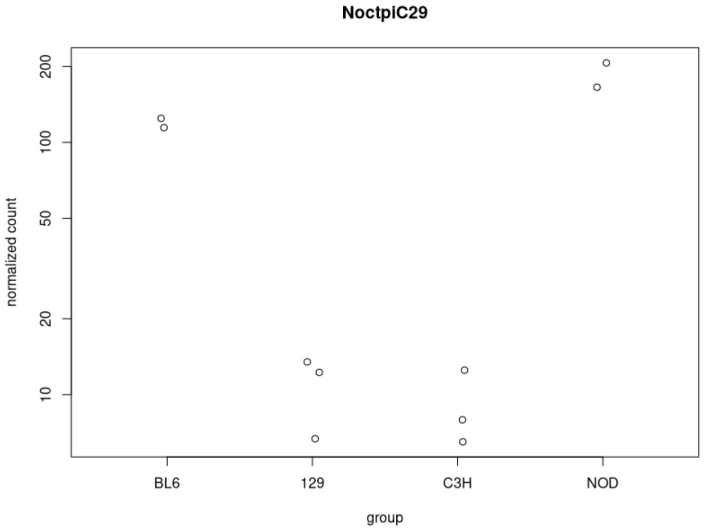
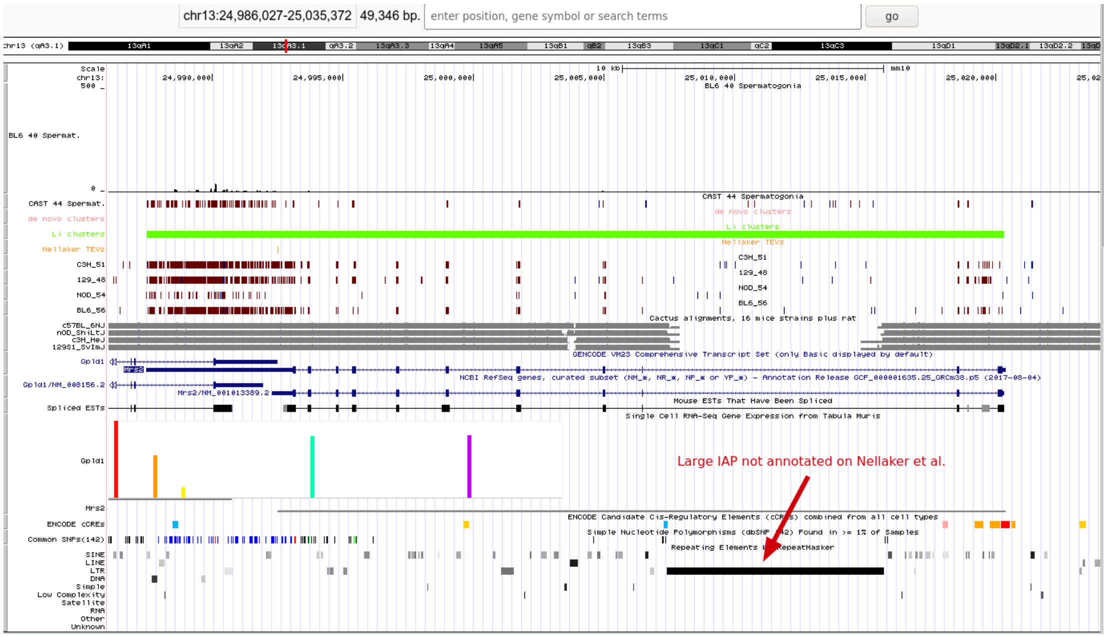


Figure 7: Normalized counts for Li cluster 29 (*Noct*).

To test our hypothesis that an IAP insertion affects the production of piRNAs from a cluster we checked the relationship between presence of variable TEs and differential piRNA cluster expression in each pair of strains. Of the 214 piRNA clusters examined, only seven have at least one IAP variant between the strains. That is, very few of the piRNA clusters overlap an IAP that is known to vary between the four inbred strains. We used the Fisher's exact test to test the hypothesis that the differential expression of a piRNA cluster depends on whether the cluster contains an IAP that is missing on one of the two strains. Although the number of differentially expressed clusters and variable IAPs was very small the null hypothesis could be rejected in all cases (Additional table 3). It should be noted that for the contingency tables we just used the information on IAPs from [Nellåker et al., 2012]. Had we included the *Mrs2* IAP that we found to be variable by visual inspection to the list of annotated variable IAPs,


chr13:24,986,027-25,035,372
49,346 bp.
enter position, gene symbol or search terms
go
chr13 (qA3.1)
13qA1
13qA2
13qA3.1
qA3.2
13qA3.3
13qA4
13qA5
13qB1
qB2
13qB3
13qC1
qC2
13qC3
13qD1
13qD2.1
13qD2.2
13qD
Scale
chr13:
10 kb
mm10
24,990,000
24,995,000
25,000,000
25,005,000
25,010,000
25,015,000
25,020,000
25,02
500 _
0 _
BL6 40 Spermatogonia
BL6 40 Spermat.
CAST 44 Spermatogonia
CAST 44 Spermat.
de novo clusters
Li clusters
Nellaker TEVs
C3H_51
129_48
NOD_54
BL6_56
Cactus alignments, 16 mice strains plus rat
c57BL_6NJ
nOD_ShiLtJ
c3H_HeJ
129S1_SvImJ
GENCODE VM23 Comprehensive Transcript Set (only Basic displayed by default)
Gpld1
Mrs2
NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release GCF_000001635.25_GRCm38.p5 (2017-08-04)
Gpld1/NM_008156.2
Mrs2/NM_001013389.2
Mouse ESTs That Have Been Spliced
Spliced ESTs
Single Cell RNA-Seq Gene Expression from Tabula Muris
Gpld1
Mrs2
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
ENCODE cCREs
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

Large IAP not annotated on Nellaker et al.

Figure 8: *Mrs2* gene on the UCSC browser with the intronic IAP missing on the annotations by [Nellåker et al., 2012].

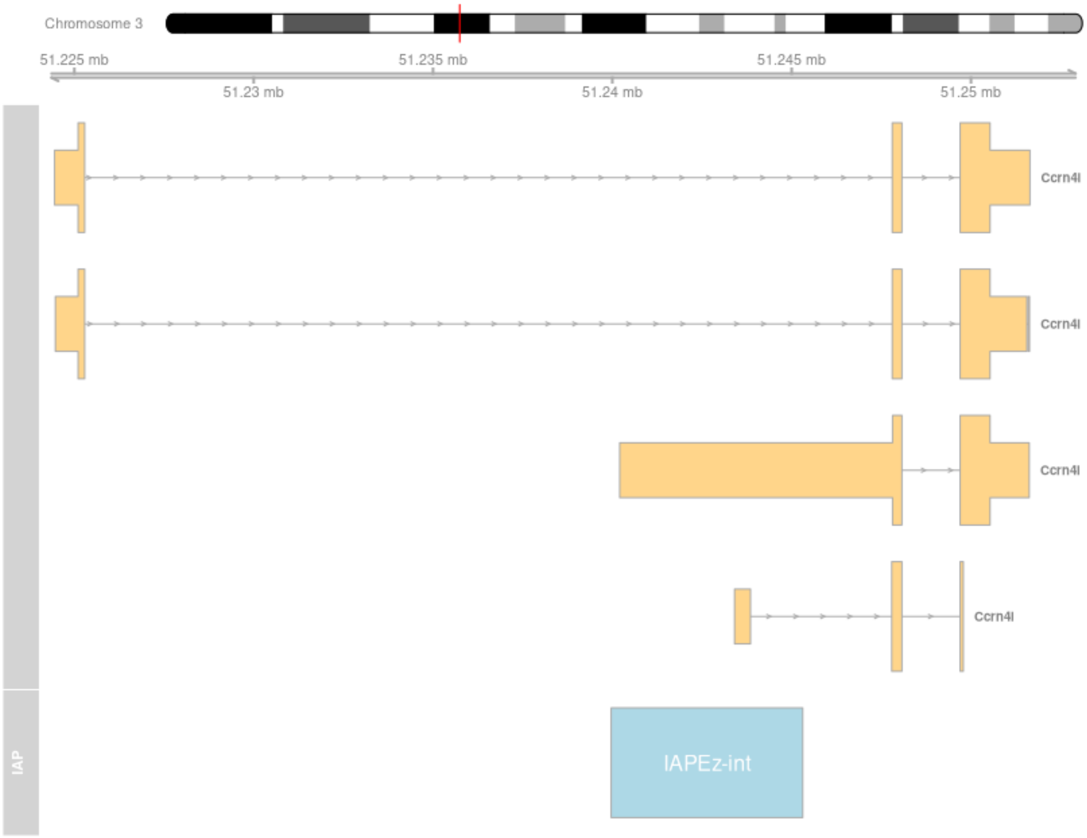


Figure 9: Transcripts of *Noct* (*Ccrn4l*) and relative position of the variable IAP associated with it.

the result would have been even more significant. It remains to be understood why there is no differential piRNA cluster expression for all of the seven clusters which have variable IAPs.

We repeated the process with the samples from the C57BL/6 and CAST/EiJ inbred mouse strains. We found eight significant differentially expressed clusters ($log_2 foldchange > 2$ and $p_{adj} < 0.05$) (Additional figure 21). At least six of them have associated IAPs or other types of ERVs. The higher number of significantly down-regulated piRNA clusters in CAST/EiJ could be explained by the suppressing effect of the CAST Nxf1 mutant variant. [Concepcion et al., 2015].

Finally, we wanted to confirm if there was any effect on the outbred mice that could be attributed to the metabolic state of the animals (note that a subset of the animals were glucose intolerant). Our reasoning was to test if these samples could also be used to test whether genetic variation likely underlies piRNA cluster expression variation, since different ICR mice contain different genetic variants. The metabolic state of the animals was previously assessed by a collaborator (JC Jimenez-Chillaron, personal communication) and coded in a scale from 0 (normal) to 5 (most extreme metabolic phenotype). We analysed the samples with metabolic status as the only factor, that we converted into a binary factor ($MetabolicState > 4 \rightarrow TRUE$) to check if there was any differential expression related to the metabolic status. The more significant results were just moderately significantly expressed ($log_2 foldchange < 2$) and matched two of the clusters for which we had previously found a significant differential expression related to the presence of an IAP on their loci (*Noct*, *Mrs2*) and one cluster which showed differential expression in the spermatogonia test and also is know to have a variable IAP present (*Phf20*) (Results can be found on Appendix E: Additional files). These results suggest that genetic differences between animals of the ICR strain likely explain the observed variation. In the case for *Noct*, a variable IAP in the first intron perfectly explains this variation (based on genotyping PCR results, Vavouri personal communication). We have also seen on the inbreed strains that differential expression of *Mrs2* and *Phf20* correlates with a variable IAP on them too, so we suspect this could be that case also here. However, as other genetic variants of animals of this strain are not known, we cannot be sure about the type of genetic variation responsible for piRNA cluster expression variation in mice of this strain".

In summary, these results (Table 2 and Additional table 4) provide support to the hypothesis that TE (especially IAP) variation between strains is linked to changes in piRNA cluster expression in the mouse male germline.

CAST vs BL6	piRNA cluster was	IAP missing in one strain	
	Differentialy expressed	FALSE	TRUE
	FALSE	198	6
	TRUE	8	2

$p_{val} = 0.047$

Table 2: Contingency table and Fisher test p-value for the spermatogonia samples comparison on whether the cluster was differentialy expressed and whether there is and IAPs present on just one of the strains according to [Nellåker et al., 2012] using the 214 clusters by [Li et al., 2013].

2.4 De novo prediction of piRNA clusters

The piRNA clusters from [Li et al., 2013] were generated from data belonging only to the reference (C57BL/6J) strain for Mus musculus. We wanted to test whether we were missing clusters of piRNAs present on any of the other four strains but not in the reference. To generate the clusters we decided to use proTRAC [Rosenkranz and Zischler, 2012]. With proTRAC we predicted piRNA clusters starting with raw reads from two samples of each of the combinations of strain and tissues (testes and spermatogonia). Then we took only those clusters which were present in both of the samples for each strain, to make sure that the predicted clusters were consistent at the strain level. We then merged the predicted clusters from all the strains. Finally we removed the clusters for which a large TE from [Nellåker et al., 2012] was overlaping at least 80% of the cluster, as they could be an artefact of multimapping reads (Additional figure 22). TEs are repeated along the genome and for that reason any match of a small RNA in them is actually very hard to map to a precise TE. From the list of variable TEs by [Nellåker et al., 2012] those which are not present on the reference strain can not be mapped to it, and hence were not used to filter overlapping clusters. In the end we predicted a total of 611 clusters, which overlaped 149 of the previous clusters by [Li et al., 2013] (69,6%). In particular the overlap was of 63 out of 100 pachytene clusters (63%), 24 out of 30 hybrid clusters (80%) and 62 out of 84 prepachytene clusters (73,8%). (Figure 10, Additional figures 24 and 23).

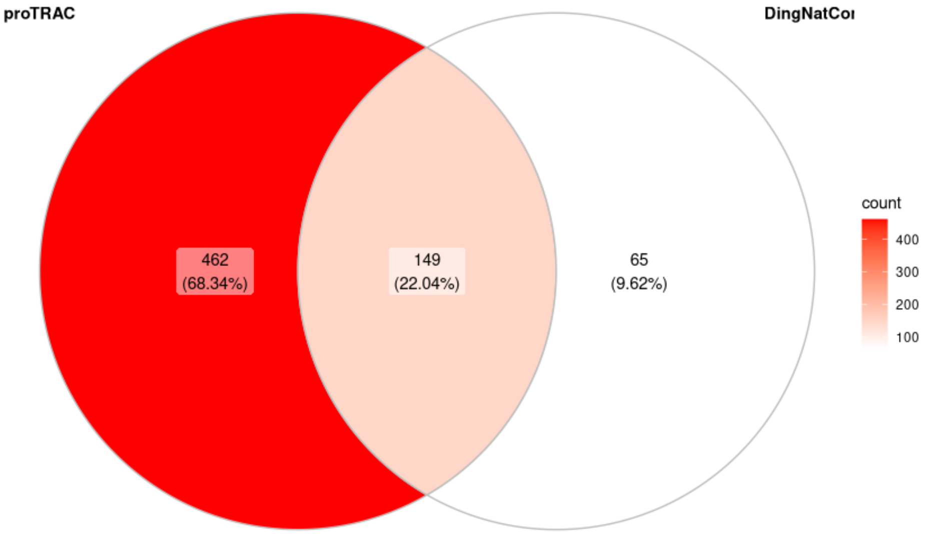


Figure 10: Common clusters to [Li et al., 2013] and our de novo clusters.

2.5 Differential Expression Analysis (2)

Using these 611 clusters we counted small RNA reads mapping to them and repeated the differential expression analysis that we previously did with the [Li et al., 2013] clusters. As the clusters were generated using data from the two types of tissues (testes and spermatogonia), we found that a small number of the clusters had really low counts, so we removed all clusters with no samples with $counts > 9$.

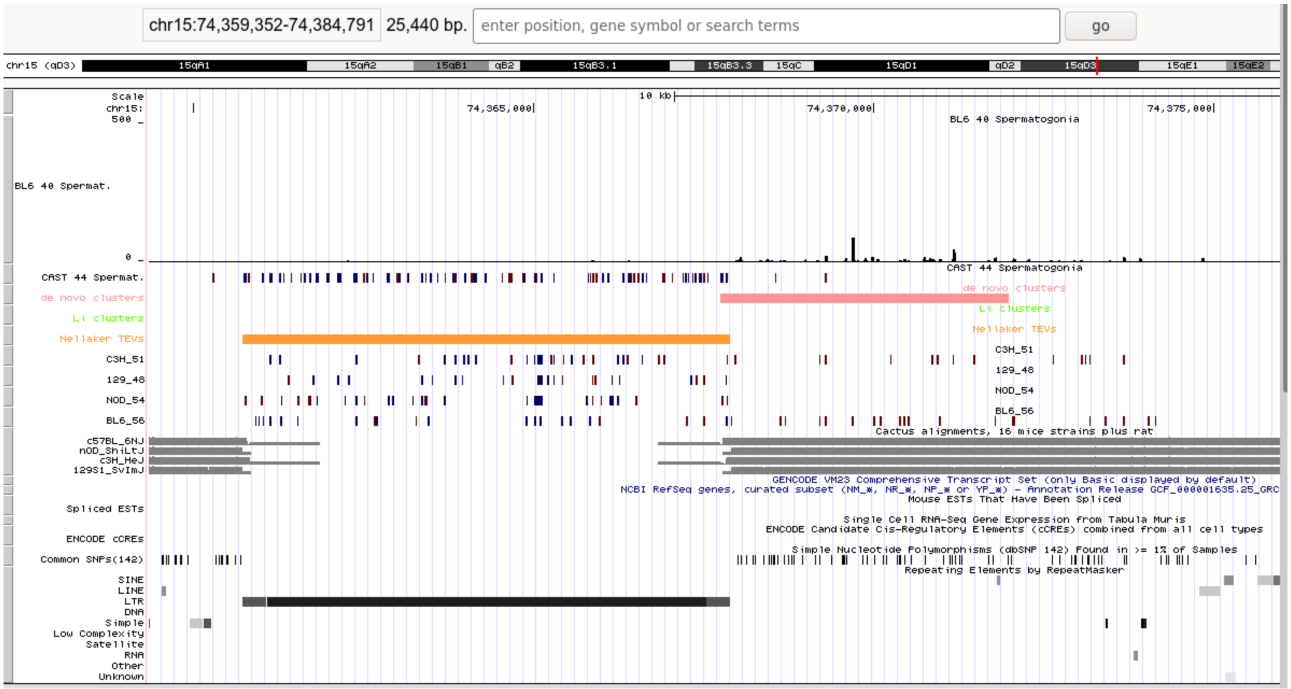

chr15:74,359,352-74,384,791
25,440 bp.
enter position, gene symbol or search terms
go
chr15 (qD3)
Scale
chr15:
10 kb
74,365,000
74,370,000
74,375,000
BL6 40 Spermatogonia
BL6 40 Spermat.
CAST 44 Spermat.
CAST 44 Spermatogonia
de novo clusters
Li clusters
Nellaker TEVs
C3H_51
129_48
NOD_54
BL6_56
Cactus alignments, 16 mice strains plus rat
c57BL_6NJ
nOD_ShiLtJ
c3H_HeJ
129S1_SvImJ
GENCODE VM23 Comprehensive Transcript Set (only Basic displayed by default)
NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release GCF_000001635.25_GRC
Mouse ESTs That Have Been Spliced
Spliced ESTs
Single Cell RNA-Seq Gene Expression from Tabula Muris
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
ENCODE cCREs
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

Figure 11: Browser view of an intergenic de novo cluster PTc181 (pink), with a large IAP (orange) downstream of the cluster (Cluster is expressed on the - strand).

We proceeded with the same type of cluster annotation as with the previous differential expression analysis. The results are shown as volcano plots and a heatmap (Additional figures 25 and 26). In total, 49 clusters had significant differential expression between at least two of the strains ($log_2 foldchange > 2$ and $p_{adj} < 0.05$). In addition to confirming all the previous results where there was a cluster overlap, we were able to identify three new clusters where the presence or absence of an IAP had an effect on piRNA expression (PTc181, PTc567, PTc463) (Figure 11, Figure 12). PTc463 in particular is affected by the presence of an IAP just detected in NOD, and it is an example of a potential new piRNA cluster in a strain that could not have been identified on the reference or using data about repeats from the reference strain only.

Again we used Fisher's exact tests to check if the presence of a variable IAP is associated with differentially expressed piRNA clusters (Additional table 4).

We identified a total of 16 clusters with a variable IAP between at least two strains, but for which no significant differential expression was found. At least

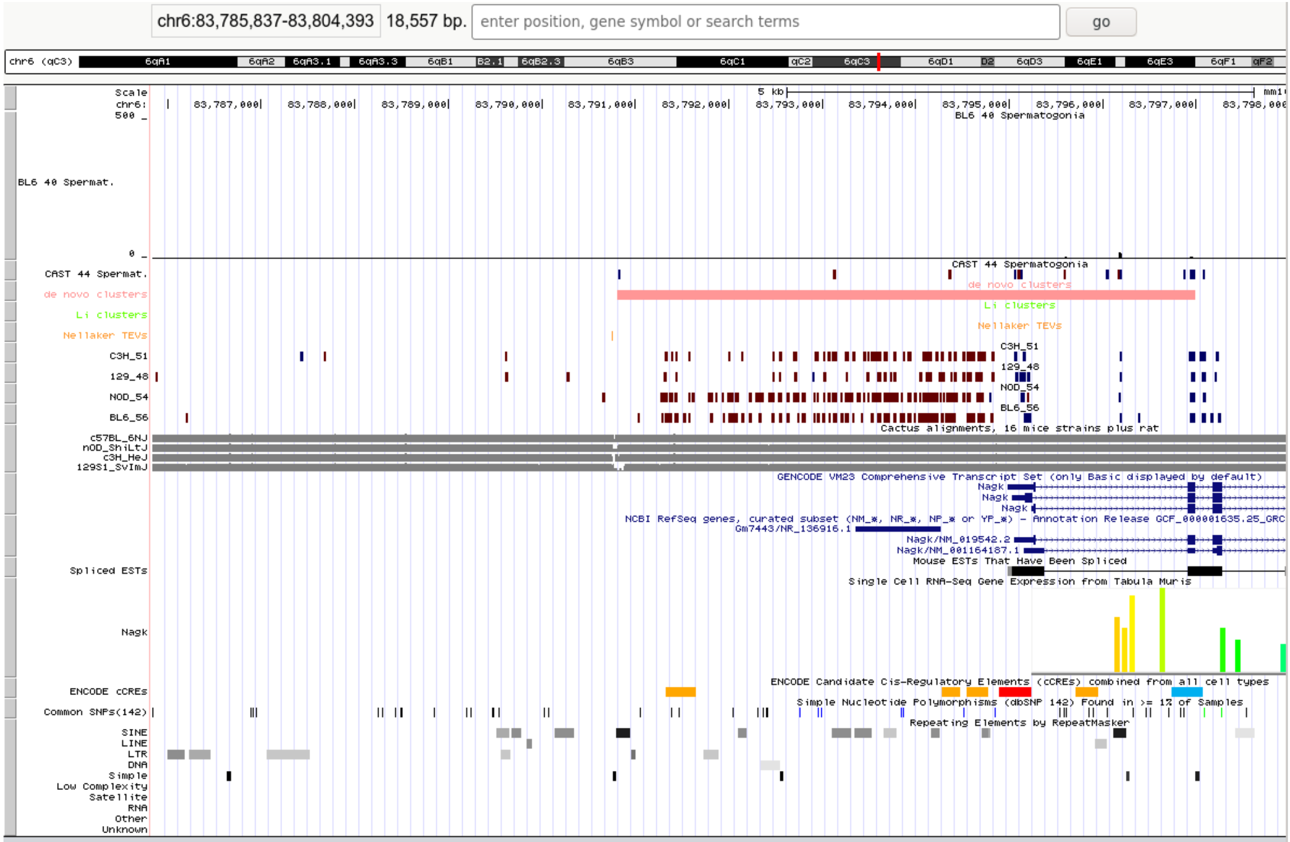

chr6:83,785,837-83,804,393
18,557 bp.
enter position, gene symbol or search terms
go
chr6 (qC3)
Scale
chr6:
5 kb
mm10
83,787,000
83,788,000
83,789,000
83,790,000
83,791,000
83,792,000
83,793,000
83,794,000
83,795,000
83,796,000
83,797,000
83,798,000
BL6 40 Spermatogonia
500 _
BL6 40 Spermat.
0 _
CAST 44 Spermatogonia
CAST 44 Spermat.
de novo clusters
Li clusters
Nellaker TEVs
C3H_51
129_48
NOD_54
BL6_56
Cactus alignments, 16 mice strains plus rat
C57BL_6NJ
NOD_ShiLtJ
C3H_HeJ
129S1_SvImJ
GENCODE VM23 Comprehensive Transcript Set (only Basic displayed by default)
Nagk
NCBI RefSeq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release GCF_000001635.25_GRC
Gm7443/NR_136916.1
Nagk/NM_019542.2
Nagk/NM_001164187.1
Mouse ESTs That Have Been Spliced
Spliced ESTs
Single Cell RNA-Seq Gene Expression from Tabula Muris
Nagk
ENCODE Candidate Cis-Regulatory Elements (cCREs) combined from all cell types
ENCODE cCREs
Simple Nucleotide Polymorphisms (dbSNP 142) Found in >= 1% of Samples
Common SNPs(142)
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

Figure 12: Browser view of an intergenic de novo cluster PTc463 (pink), in which the reference genome does not include any IAP, but for which [Nellåker et al., 2012] has identified a large IAP only present on the NOD strain just downstream of the cluster (Cluster is expressed on the - strand).

two of these clusters include the presence of another, complete IAP in them. If we assume that just one IAP is required to have an effect on piRNA expression from a transcribed region, then simply having one IAP would be sufficient to explain it and variation at other IAPs would not have any effect. In at least one case there was also a complete ERVK between the cluster and the IAP. We also found that six of them belonged to transcripts that were barely expressed at all according to the data from [Gainetdinov et al., 2018]. On the rest we could not identify any reason which could prevent them from showing a significant differential expression, so further analysis needs to be done to understand the additional conditions which trigger the effect.

2.6 Discussion

piRNAs present in the germline of most animals are responsible for silencing TEs through base-pair complementarity. We have performed an extensive study of the expression of piRNAs in the male germline of five inbred mouse strains. Our sequenced small RNA data do not come from PIWI immunoprecipitated RNA samples, we used small RNAs sequenced from whole testes and spermatogonia. Even though by definition only small RNAs immunoprecipitated with PIWI proteins can be called piRNAs, testes and spermatogonia are highly enriched for piRNAs, and the quality checks showed that the small RNAs we analyse have the characteristic length and first nucleotide composition of piRNAs. This allowed us to work on the assumption that, indeed, we were dealing with piRNAs. This was again confirmed by the level of mapping of the reads that we show to the clusters, up to 80% for the testis samples. For spermatogonia, the percentage of reads mapping to piRNA clusters is significantly less (30%). It is unclear at this point what small RNAs the remaining of the reads correspond to and would require further investigation. Our analysis showed that there are differences in the expression of some piRNA clusters and these differences can be attributed to the genetic background (i.e. strain). We analyzed piRNA expression from 214 previously known piRNA clusters from [Li et al., 2013] and also from 611 piRNA clusters that we predicted *de novo*. The results, with the documented clusters as well as with the *de novo* ones show a significant correlation between piRNA differential expression and the different status of a very young retrotransposon, the murine intracisternal A-particle (IAP) in the strains being compared. We found that the presence of a transposon of that type in the region being expressed (±10kb) increases the chance that the cluster is differentially expressed in the strains where the IAP is present or absent ($p_{val} < 0.5$ for all Fisher's exact tests). This result would be compatible with previous observations that the signals that define piRNA generative loci must lie within the clusters themselves rather than being implicit in their genomic position [Muerdter et al., 2012]. We confirmed clusters where this effect had already been documented, specifically for the cluster overlapping *Noct* (Casas and Vavouri, personal communication) and we also confirmed our theory with five new piRNA clusters (two from the previously published piRNA clusters and three from our de novo predicted clusters) in which the same effect correlated to the presence/absence of an IAP was confirmed. This work suggests that at least part of the sequence of IAP transposons somehow interact with the mechanism of piRNA biogenesis in the mouse male germline. A potential explanation would be an interaction between IAPs and the splicing machinery [Concepcion et al., 2009] or the export of RNA from the nucleus [Pippadpally and Venkatesh, 2020], either of which would likely affect piRNA production. The work presented in this TFM is the most extensive analysis of piRNA cluster expression variation in testes of genetically diverse individuals in any mammalian species. It remains an open question why the effect is sometimes to enhance the expression of piRNAs and other times to reduce it, so the mechanism by which the IAP interacts with the piRNAs biogenesis needs to be clarified.

piRNAs are important guardians of animal genomes against active transposons and they achieve these through epigenetic and post-transcriptional silencing of active transposons ([Siddiqi and Matushansky, 2012], [Landry et al., 2013], [Calcagno et al., 2019] and reviewed in [Liu et al., 2018]). Although a lot is known about piRNA biogenesis, there remain gaps in our knowledge. Furthermore, how piRNA clusters evolve is poorly understood. The relationship we just presented between IAPs and piRNAs expression suggests that young endogenous retrovirus insertions may trigger piRNA cluster evolution.

3 Conclusions

Here we list the conclusions related to the realization of project and the tools and methods used. For conclusions on the analysis performed see previous section "Discussion".

3.1 Main take home messages

- piRNAs are challenging to analyse due to their small size and repetitive sequence.
- The DESeq2 workflow is a comprehensive tool for differential expression analysis of small RNA-Seq data.
- The UCSC Genome Browser and Ensembl are invaluable tools to put in context genome-related information and to aggregate genome annotations in a meaningful way.
- Good organization of the documentation from the beginning is paramount to being able to access it correctly and to keeping control over it.
- Scheduled sessions with the tutor or other investigators is the best way to help you challenge continuously your ideas and keep you on the edge.
- The R language provides all kinds of libraries needed for this kind of work, including mature ones for very specialized graphics.

3.2 Initial scope achievement

All the targets from the initial scope were achieved early on, which allowed us to extend the scope (prediction de novo clusters described in section 2.4) and extend the reach and the depth of the analysis.

3.3 Planning and methodology

We consider the planning was correct. It was important to leave room for changes and to understand that this kind of work is almost always of an iterative nature, in which several iterations to debug the processes are needed. The main tool used, DESeq2, is a mature tool and it was easy to prepare the data for it and to use it to perform multiple contrasts at once. As for the rest, we used a combination of shell and R tools, depending on the task at hand. We found no issues with any of them.

3.4 Lines of work to be explored

- Identify if the presence / absence of one or some of the genes of an IAP (*gag, pro, pol, env*) is more correlated with the differential expression.
- The analysis of the data also showed a small bias to have uracil on the 3' end. As this has not been documented before further work would be needed to see if this was an artifact or if there is any effect related to piRNA biogenesis which could account for it, similar to the larger bias for uracil on the 5' end.
- Understand better the relationship between the actual location of the IAP related to the cluster and its effect (distance, sense, relative position, strand, size of the transcript).
- Try to identify new piRNA clusters with the use of more advanced tools like IpiRld [Boucheham et al., 2017], a Multiple Kernel Learning approach based on several piRNA features.
- We used information on repeats from several sources, but one we could not use was directly from Repbase due to the licence [Jurka et al., 2005], which we expect to have deeper information about the age of the different TEs in the genome.
- All the analysis performed has been based on mappings of the reads to the reference genome. More precise results, but harder to compare, could be achieved using the full assembly for each of the strains.

4 Glossary

- **IAP:** murine intracisternal A-particle retrotransposon.
- **Pachytene:** third stage of the prophase of meiosis.
- **PCA:** principal component analysis.
- **piRNA:** PIWI-interacting RNA. A class of small non-coding RNA expressed in animal cells.
- **rlog:** regularized-logarithm transformation.
- **Spermatogonia:** undifferenciated male germ cell.
- **TE:** transposable element, transposon.
- **TEV:** transposable element variant.

5 Bibliography

References

Adams, D. J., Doran, A. G., Lilue, J. and Keane, T. M. (2015). The Mouse Genomes Project: a repository of inbred laboratory mouse strain genomes. Mammalian Genome *26*, 403–412.

Aravin, A. A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Fejes Toth, K., Bestor, T. and Hannon, G. J. (2008). A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Molecular cell *31*, 785–799.

Barbot, W. (2002). Epigenetic regulation of an IAP retrotransposon in the aging mouse: progressive demethylation and de-silencing of the element by its repetitive induction. Nucleic Acids Research *30*, 2365–2373.

Boucheham, A., Sommard, V., Zehraoui, F., Boualem, A., Batouche, M., Bendahmane, A., Israeli, D. and Tahi, F. (2017). IpiRId: Integrative approach for piRNA prediction using genomic and epigenomic data. PLOS ONE *12*, e0179787.

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. and Feschotte, C. (2018). Ten things you should know about transposable elements. Genome Biology *19*, 199.

Calcagno, D. Q., Mota, E. R. d. S., Moreira, F. C., de Sousa, S. B. M., Burbano, R. R. and Assumpção, P. P. (2019). Role of PIWI-Interacting RNA (piRNA) as Epigenetic Regulation. In Handbook of Nutrition, Diet, and Epigenetics, (Patel, V. B. and Preedy, V. R., eds), pp. 187–209. Springer International Publishing Cham.

Carmell, M. A., Girard, A., Kant, H. J. G. v. d., Bourc'his, D., Bestor, T. H., Rooij, D. G. d. and Hannon, G. J. (2007). MIWI2 Is Essential for Spermatogenesis and Repression of Transposons in the Mouse Male Germline. Developmental Cell *12*, 503–514.

Cenik, E. S. and Zamore, P. D. (2011). Argonaute proteins. Current Biology *21*, R446–R449.

Concepcion, D., Flores-García, L. and Hamilton, B. A. (2009). Multipotent Genetic Suppression of Retrotransposon-Induced Mutations by Nxf1 through Fine-Tuning of Alternative Splicing. PLoS Genetics *5*.

Concepcion, D., Ross, K. D., Hutt, K. R., Yeo, G. W. and Hamilton, B. A. (2015). Nxf1 Natural Variant E610G Is a Semi-dominant Suppressor of IAP-Induced RNA Processing Defects. PLoS Genetics *11*.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X. and Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. Genome Biology *17*.

Durinck, S., Spellman, P. T., Birney, E. and Huber, W. (2009). Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor package biomaRt. Nature protocols *4*, 1184–1191.

Elmer, J. L., Hay, A. D., Kessler, N. J., Bertozzi, T. M., Ainscough, E. and Ferguson-Smith, A. C. (2020). Genomic properties of variably methylated retrotransposons in mouse. preprint Genetics.

Gagnier, L., Belancio, V. P. and Mager, D. L. (2019). Mouse germ line mutations due to retrotransposon insertions. Mobile DNA *10*.

Gainetdinov, I., Colpan, C., Arif, A., Cecchini, K. and Zamore, P. D. (2018). A Single Mechanism of Biogenesis, Initiated and Directed by PIWI Proteins, Explains piRNA Production in Most Animals. Molecular cell *71*, 775–790.e5.

Hannon FASTX-Toolkit.

Horie, K., Saito, E.-S., Keng, V. W., Ikeda, R., Ishihara, H. and Takeda, J. (2007). Retrotransposons influence the mouse transcriptome: implication for the divergence of genetic traits. Genetics *176*, 815–827.

Houwing, S., Berezikov, E. and Ketting, R. F. (2008). Zili is required for germ cell differentiation and meiosis in zebrafish. The EMBO Journal *27*, 2702–2711.

Houwing, S., Kamminga, L. M., Berezikov, E., Cronembold, D., Girard, A., Elst, H. v. d., Filippov, D. V., Blaser, H., Raz, E., Moens, C. B., Plasterk, R. H. A., Hannon, G. J., Draper, B. W. and Ketting, R. F. (2007). A Role for Piwi and piRNAs in Germ Cell Maintenance and Transposon Silencing in Zebrafish. Cell *129*, 69–82.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research *110*, 462–467.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, a. D. (2002). The Human Genome Browser at UCSC. Genome Research *12*, 996–1006.

Kofler, R. (2020). piRNA Clusters Need a Minimum Size to Control Transposable Element Invasions. Genome Biology and Evolution *12*, 736–749.

Kuramochi-Miyagawa, S., Kimura, T., Ijiri, T. W., Isobe, T., Asada, N., Fujita, Y., Ikawa, M., Iwai, N., Okabe, M., Deng, W., Lin, H., Matsuda, Y. and Nakano, T. (2004). Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. Development *131*, 839–849.

Landry, C. D., Kandel, E. R. and Rajasethupathy, P. (2013). New mechanisms in memory storage: piRNAs and epigenetics. Trends in Neurosciences *36*, 535–542.

Langmead, B. (2010). Aligning Short Sequencing Reads with Bowtie. Current Protocols in Bioinformatics *32*.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology *10*, R25.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

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. and Zamore, P. D. (2013). An Ancient Transcription Factor Initiates the Burst of piRNA Production During Early Meiosis in Mouse Testes. Molecular cell *50*, 67–81.

Liao, Y., Smyth, G. K. and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930.

Liu, J., Zhang, S. and Cheng, B. (2018). Epigenetic roles of PIWI-interacting RNAs (piRNAs) in cancer metastasis (Review). Oncology Reports *40*, 2423–2434.

Love, M., Ahlmann-Eltze, C., Anders, S. and Huber, W. (2020a). DESeq2: Differential gene expression analysis based on the negative binomial distribution.

Love, M. I., Anders, S. and Huber, W. (2020b). Analyzing RNA-seq data with DESeq2.

Love, M. I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology *15*, 550.

Maeda-Smithies, N., Hiller, S., Dong, S., Kim, H.-S., Bennett, B. J. and Kayashima, Y. (2020). Ectopic expression of the Stabilin2 gene triggered by an intracisternal A particle (IAP) element in DBA/2J strain of mice. Mammalian Genome *31*, 2–16.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12.

McCarthy, E. M. and McDonald, J. F. (2004). Long terminal repeat retrotransposons of Mus musculus. Genome Biology *5*, R14.

Mietz, J. A., Grossman, Z., Lueders, K. K. and Kuff, E. L. (1987). Nucleotide sequence of a complete mouse intracisternal A-particle genome: relationship to known aspects of particle assembly and function. Journal of Virology *61*, 3020.

Muerdter, F., Olovnikov, I., Molaro, A., Rozhkov, N. V., Czech, B., Gordon, A., Hannon, G. J. and Aravin, A. A. (2012). Production of artificial piRNAs in flies and mice. RNA *18*, 42–52.

Nellåker, C., Keane, T. M., Yalcin, B., Wong, K., Agam, A., Belgard, T. G., Flint, J., Adams, D. J., Frankel, W. N. and Ponting, C. P. (2012). The genomic landscape shaped by selection on transposable elements across 18 mouse strains. Genome Biology *13*, R45.

Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D. and Zamore, P. D. (2019). PIWI-interacting RNAs: small RNAs with big functions. Nature Reviews Genetics *20*, 89–108.

Pippadpally, S. and Venkatesh, T. (2020). Deciphering piRNA biogenesis through cytoplasmic granules, mitochondria and exosomes. Archives of Biochemistry and Biophysics *695*, 108597.

Rosenkranz, D. and Zischler, H. (2012). proTRAC - a software for probabilistic piRNA cluster detection, visualization and analysis. BMC Bioinformatics *13*, 5.

Siddiqi, S. and Matushansky, I. (2012). Piwis and piwi-interacting RNAs in the epigenetics of cancer. Journal of Cellular Biochemistry *113*, 373–380.

Smit, Hubley, R & Green, P. (2013). RepeatMasker.

Stein, C. B., Genzor, P., Mitra, S., Elchert, A. R., Ipsaro, J. J., Benner, L., Sobti, S., Su, Y., Hammell, M., Joshua-Tor, L. and Haase, A. D. (2019). Decoding the 5' nucleotide bias of PIWI-interacting RNAs. Nature Communications *10*.

Sun, Y. H., Xie, L. H., Zhuo, X., Chen, Q., Ghoneim, D., Zhang, B., Jagne, J., Yang, C. and Li, X. Z. (2017). Domestic chickens activate a piRNA defense against avian leukosis virus. eLife *6*, e24695.

Vagin, V. V., Sigova, A., Li, C., Seitz, H., Gvozdev, V. and Zamore, P. D. (2006). A Distinct Small RNA Pathway Silences Selfish Genetic Elements in the Germline. Science *313*, 320–324.

Wilson, J. E., Connell, J. E. and Macdonald, P. M. (1996). aubergine enhances oskar translation in the Drosophila ovary. Development *122*, 1631–1639.

A Appendix: Tables

129 vs BL6	piRNA cluster was	IAP missing in one strain	
	Differentialy expressed	FALSE	TRUE
	FALSE	210	0
	TRUE	3	1

$p_{val} = 0.019$

C3H vs BL6	piRNA cluster was	IAP missing in one strain	
	Differentialy expressed	FALSE	TRUE
	FALSE	211	0
	TRUE	2	1

$p_{val} = 0.014$

NOD vs BL6	piRNA cluster was	IAP missing in one strain	
	Differentialy expressed	FALSE	TRUE
	FALSE	209	2
	TRUE	2	1

$p_{val} = 0.042$

C3H vs 129	piRNA cluster was	IAP missing in one strain	
	Differentialy expressed	FALSE	TRUE
	FALSE	211	0
	TRUE	3	0

NA

NOD vs 129	piRNA cluster was	IAP missing in one strain	
	Differentialy expressed	FALSE	TRUE
	FALSE	208	1
	TRUE	3	2

$p_{val} = 0.001$

NOD vs C3H	piRNA cluster was	IAP missing in one strain	
	Differentialy expressed	FALSE	TRUE
	FALSE	209	1
	TRUE	2	2

$p_{val} = 0.001$

Table 3: Contingency tables and Fisher's exact tests p-values for the inbred testis samples comparison on whether the cluster was differentialy expressed and whether there is and IAPs present on just one of the strains according to [Nellåker et al., 2012] using the 214 clusters by [Li et al., 2013].

129 vs BL6	piRNA cluster was	IAP missing in one strain	
	Differentialy expressed	FALSE	TRUE
	FALSE	567	20
	TRUE	8	2

$p_{val} = 0.049$

C3H vs BL6	piRNA cluster was	IAP missing in one strain	
	Differentialy expressed	FALSE	TRUE
	FALSE	575	12
	TRUE	8	2

$p_{val} = 0.020$

NOD vs BL6	piRNA cluster was	IAP missing in one strain	
	Differentialy expressed	FALSE	TRUE
	FALSE	571	17
	TRUE	7	2

$p_{val} = 0.030$

C3H vs 129	piRNA cluster was	IAP missing in one strain	
	Differentialy expressed	FALSE	TRUE
	FALSE	570	15
	TRUE	10	2

$p_{val} = 0.043$

NOD vs 129	piRNA cluster was	IAP missing in one strain	
	Differentialy expressed	FALSE	TRUE
	FALSE	565	23
	TRUE	6	3

$p_{val} = 0.005$

NOD vs C3H	piRNA cluster was	IAP missing in one strain	
	Differentialy expressed	FALSE	TRUE
	FALSE	559	23
	TRUE	11	4

$p_{val} = 0.003$

Table 4: Contingency tables and Fisher's exact tests p-values for the inbred testis samples comparison on whether the cluster was differentialy expressed and whether there is and IAPs present on just one of the strains according to [Nellåker et al., 2012] using the 611 de novo clusters.

B Appendix: Figures

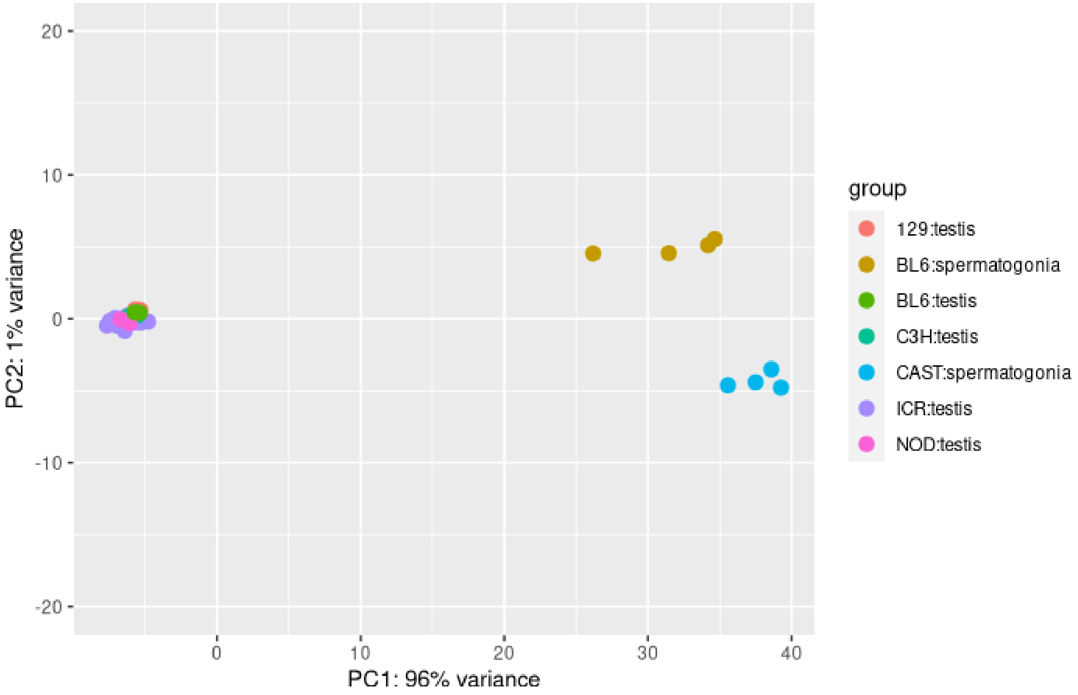


Figure 13: Principal components analysis of small RNAs mapping to known piRNA clusters in all samples, coloured by batch and strain.

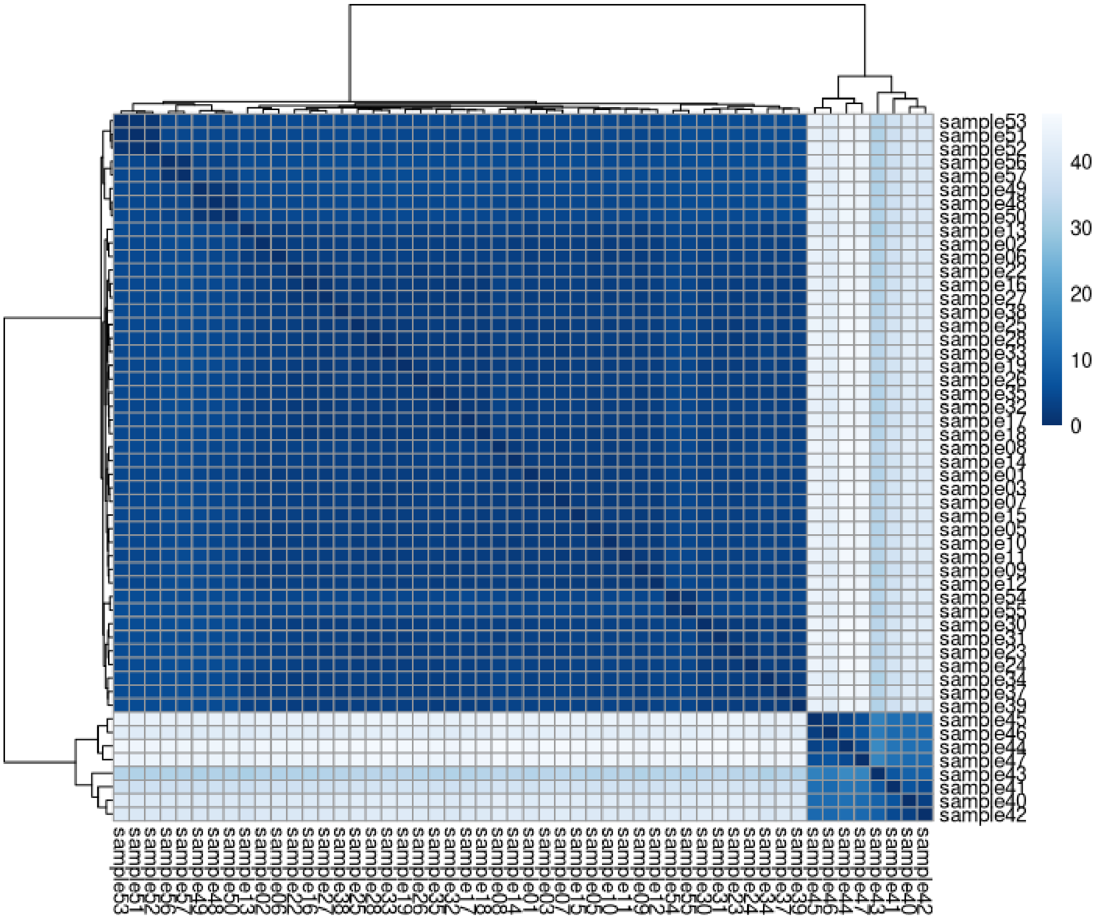


Figure 14: Sample distances of small RNAs mapping to known piRNA clusters with all samples, including spermatogonia (bottom 8).

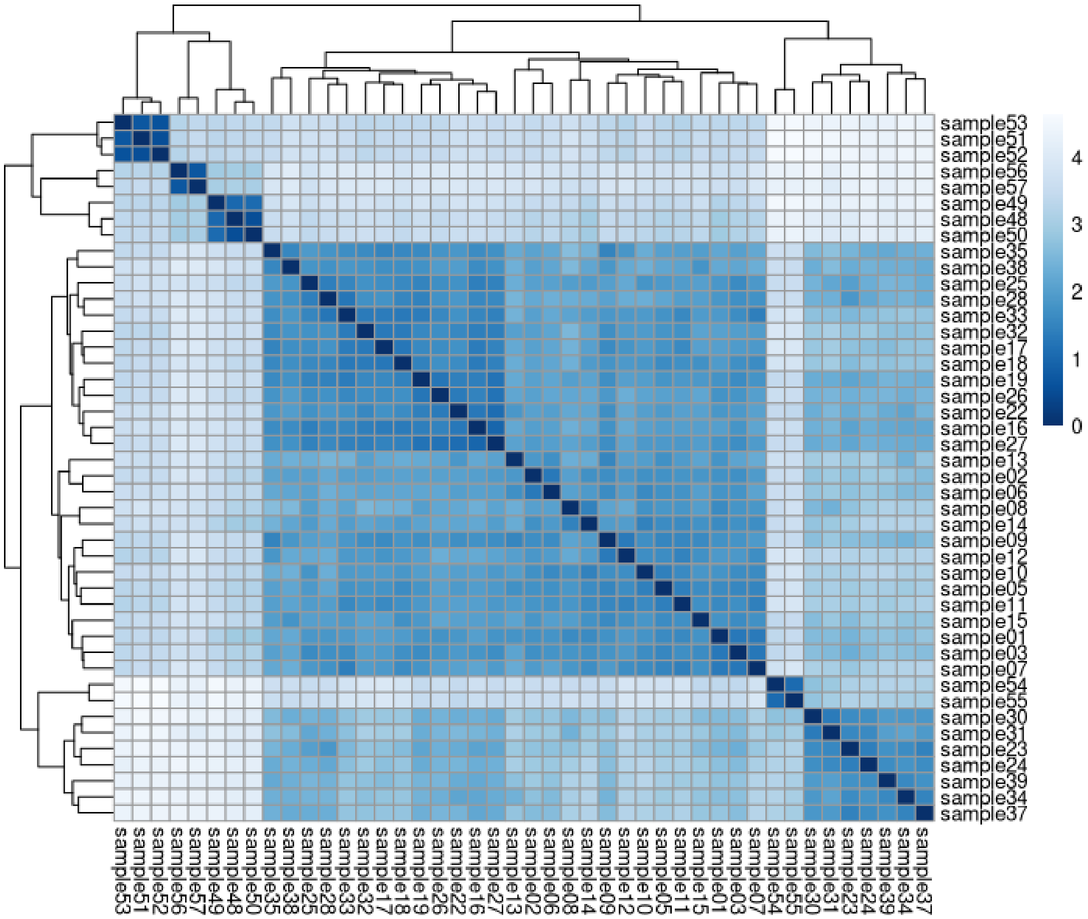


Figure 15: Sample distances of small RNAs mapping to known piRNA clusters for testis samples

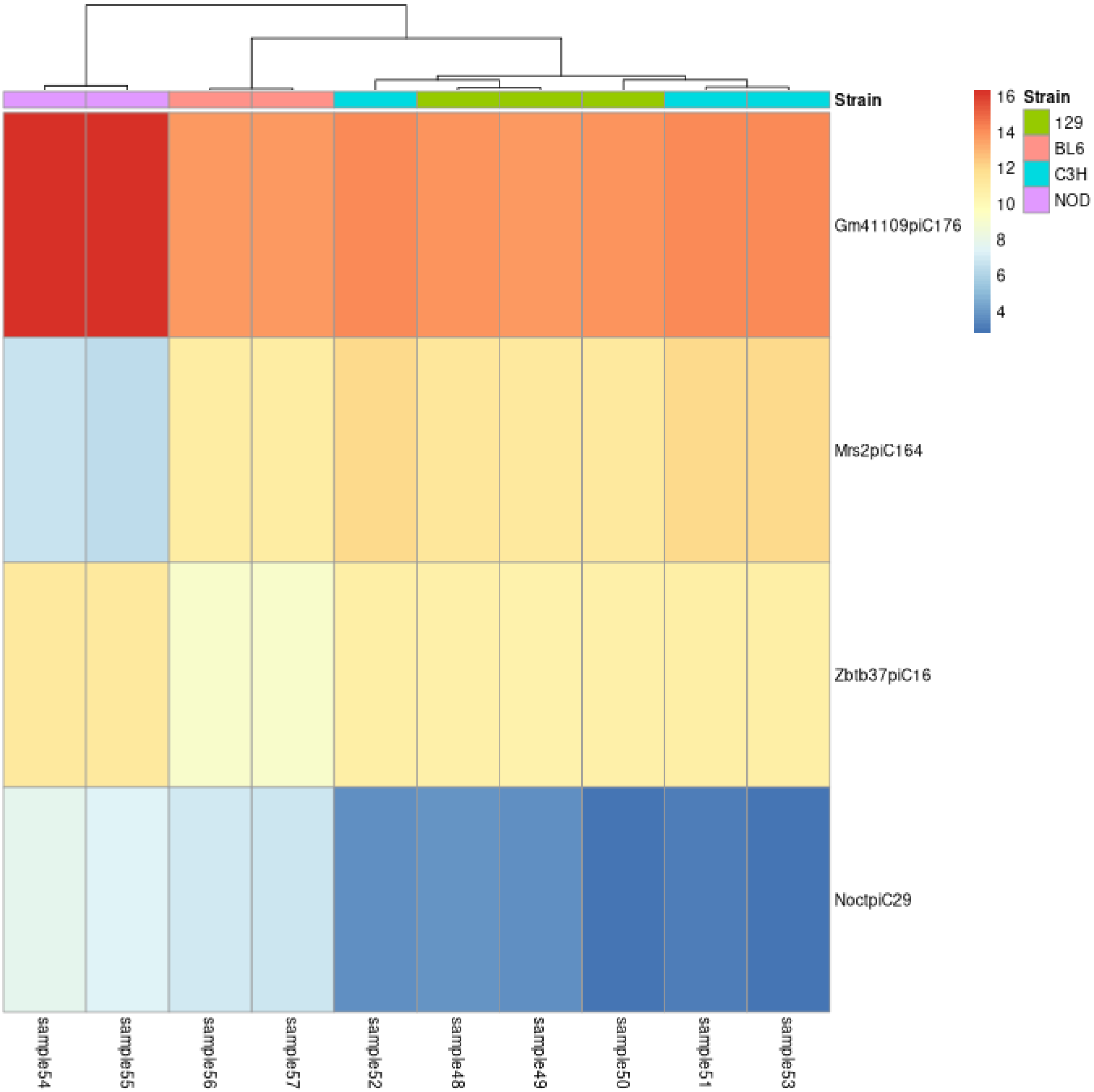


Figure 16: Heatmap of $log_2 foldchange$ for the top differential expressed genes by strain and type.

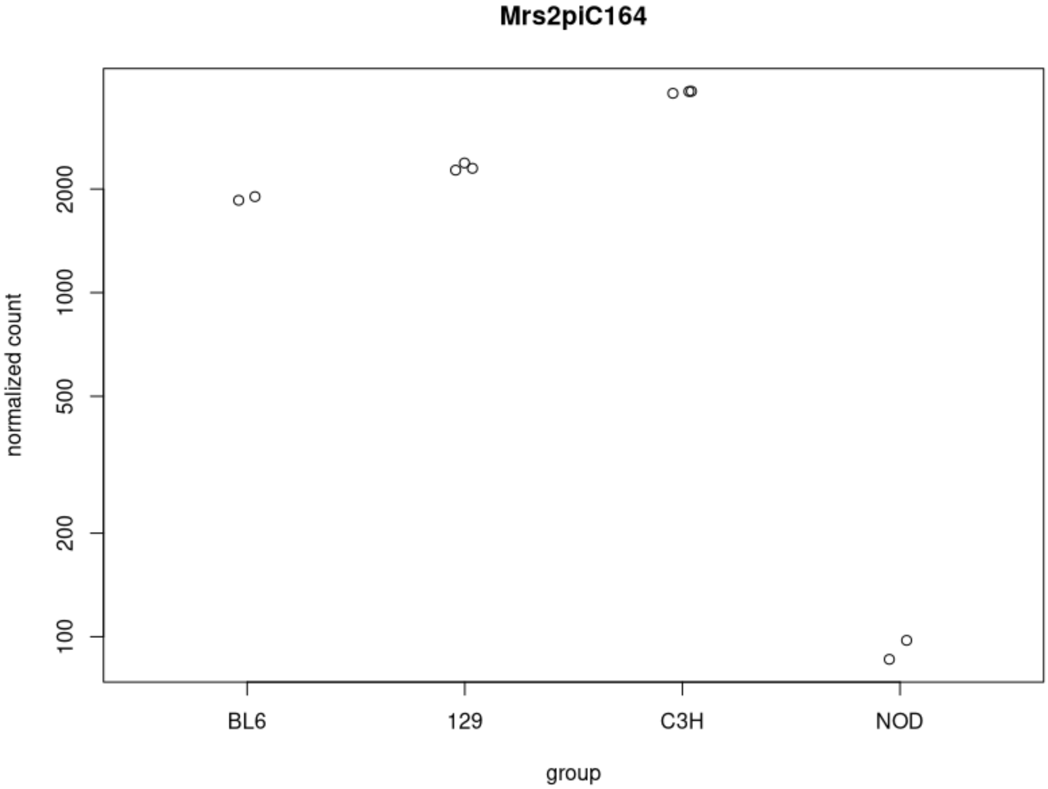


Figure 17: Normalized counts for Li cluster 164 (*Mrs2*).

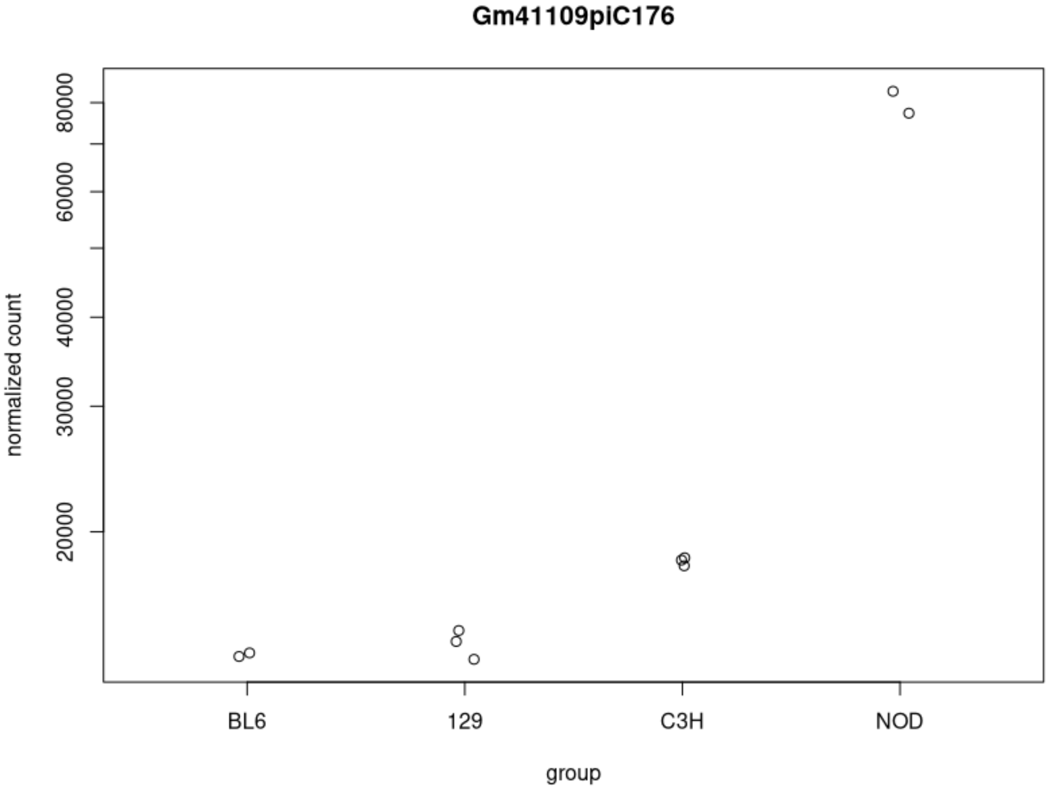


Figure 18: Normalized counts for Li cluster 176 (Gm41109).

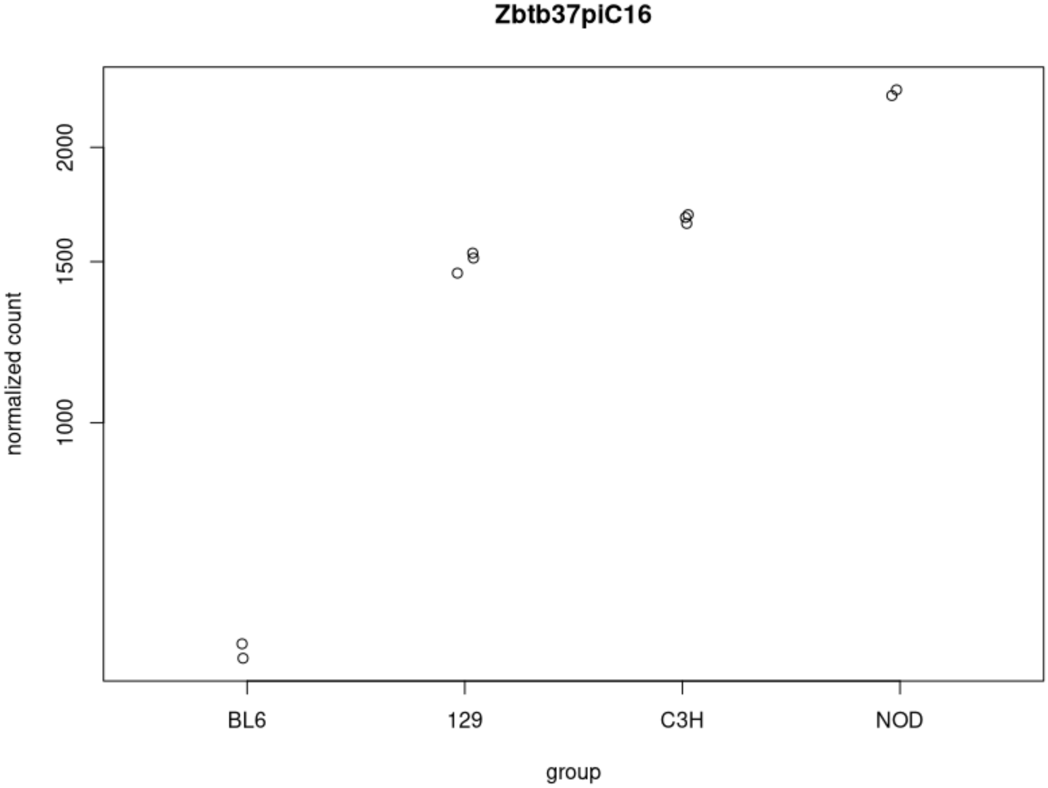


Figure 19: Normalized counts for Li cluster 29 (Zbtb37).

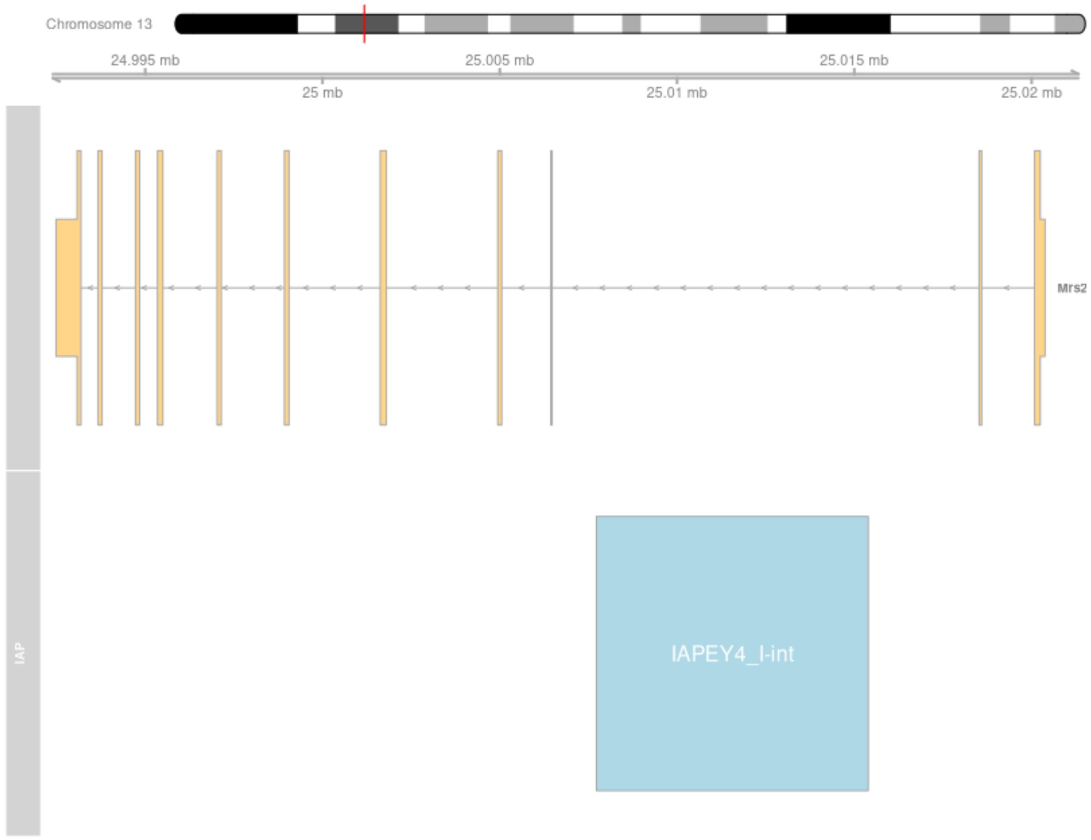


Figure 20: Transcripts of *Mrs2* and relative position of the variable IAP associated with it.

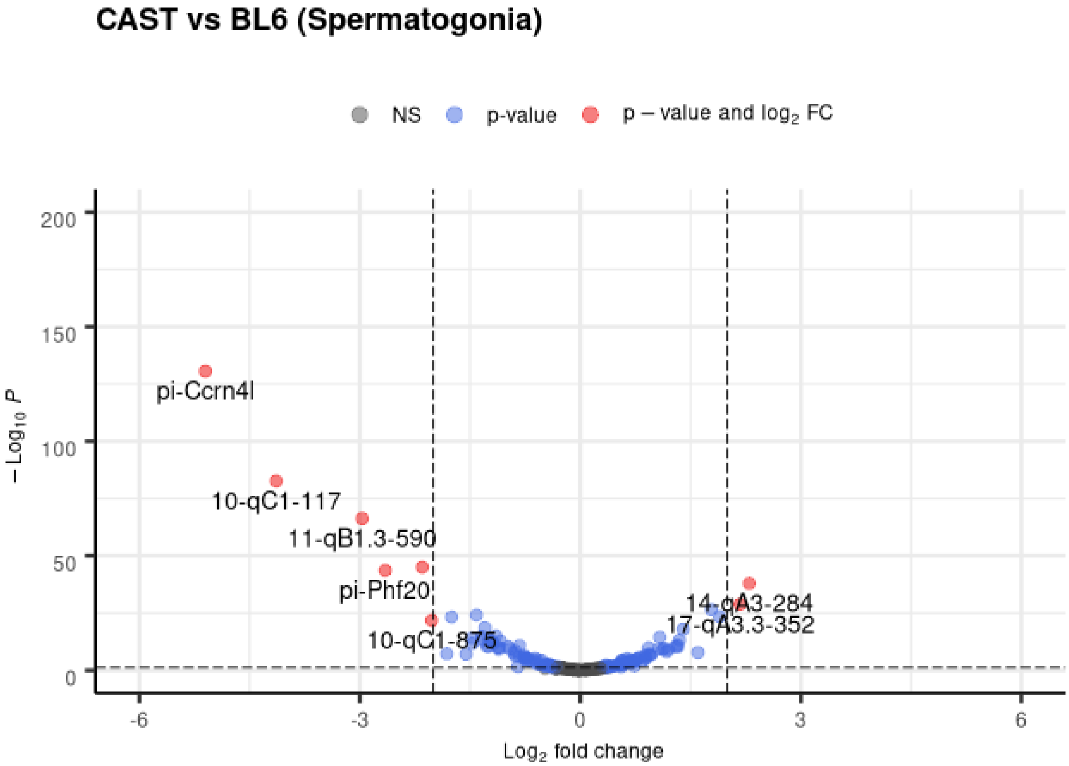


Figure 21: Volcano plot of differentially expressed genes of CAST vs BL6 (spermatogonia).

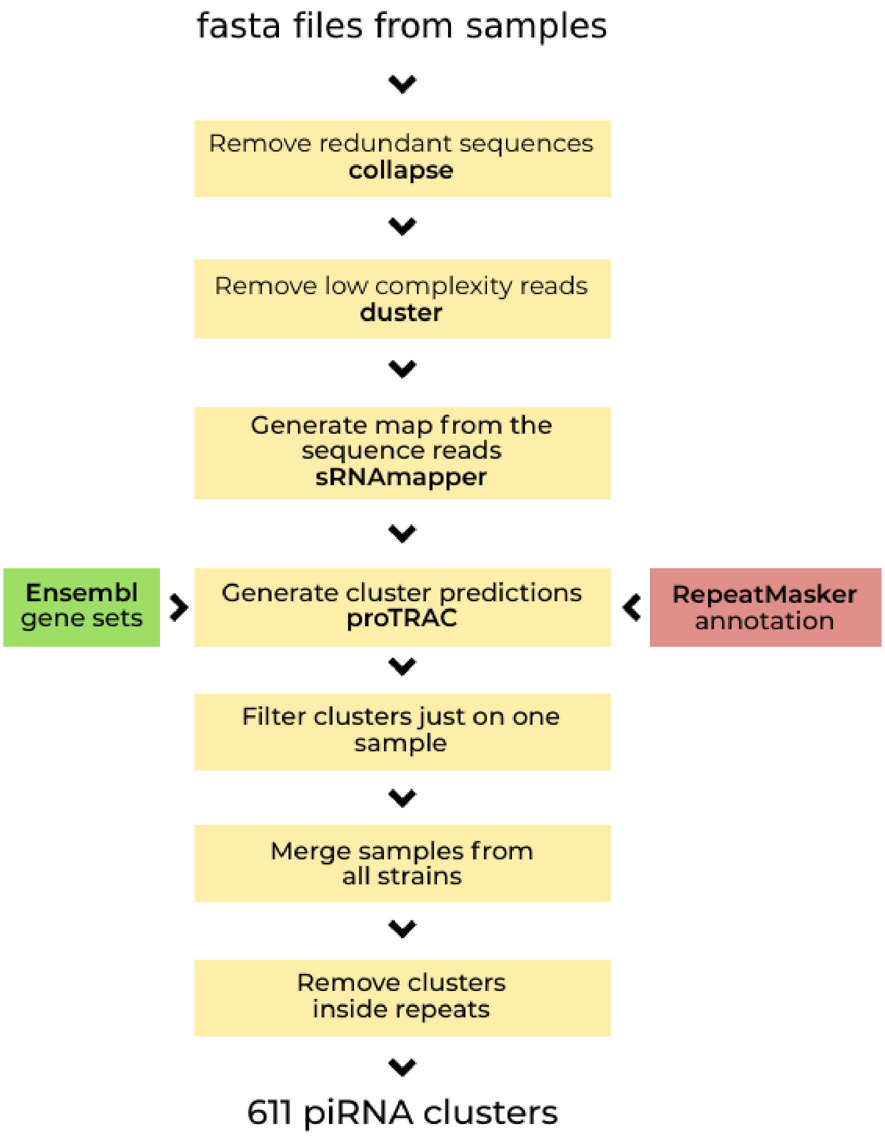

fasta files from samples
Remove redundant sequences
collapse
Remove low complexity reads
duster
Generate map from the sequence reads
sRNAmapper
Ensembl gene sets
Generate cluster predictions
proTRAC
RepeatMasker annotation
Filter clusters just on one sample
Merge samples from all strains
Remove clusters inside repeats
611 piRNA clusters

Figure 22: Workflow to create de novo piRNA clusters.

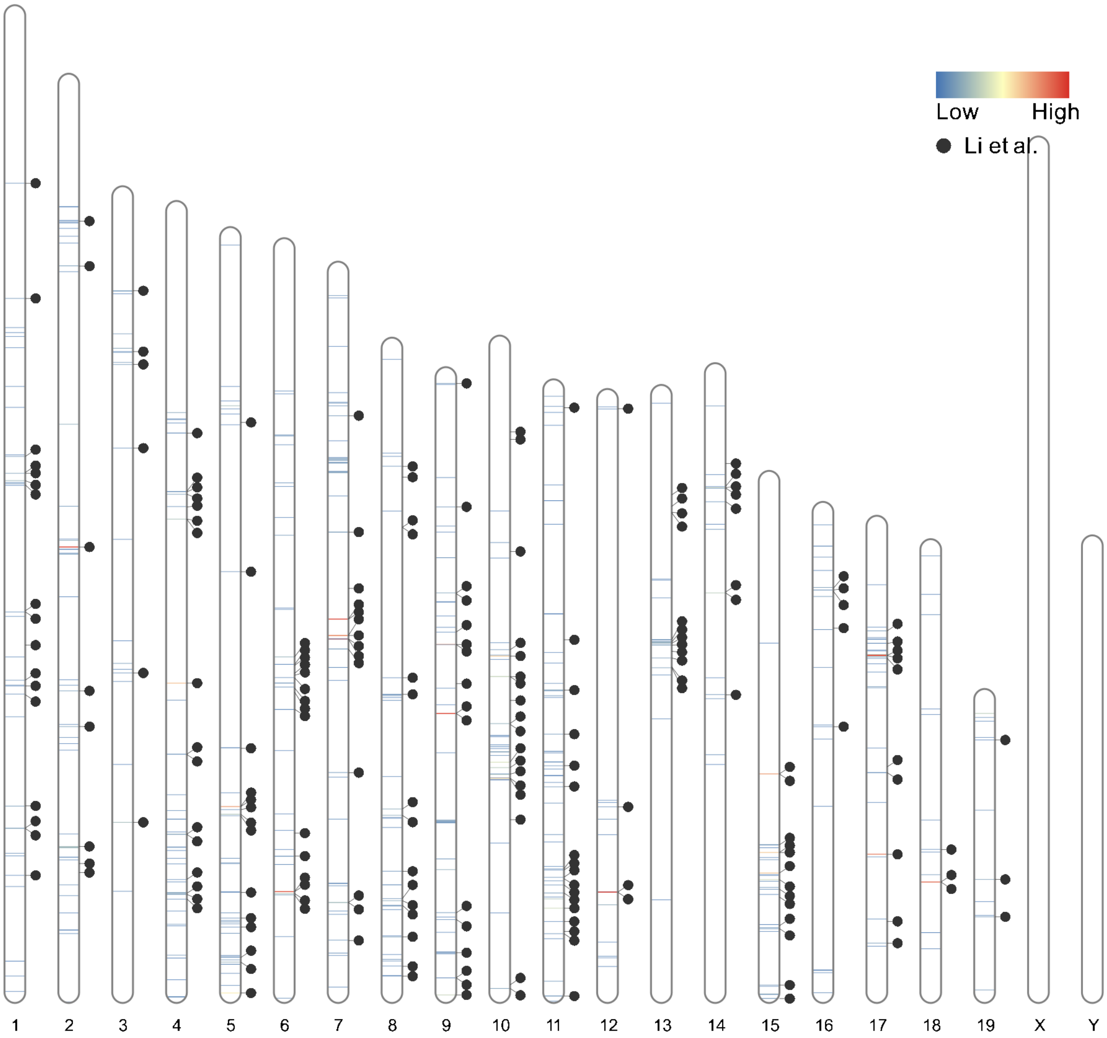


Figure 23: Karyogram with the clusters from [Li et al., 2013]. The scale signals the relative expression of the clusters.

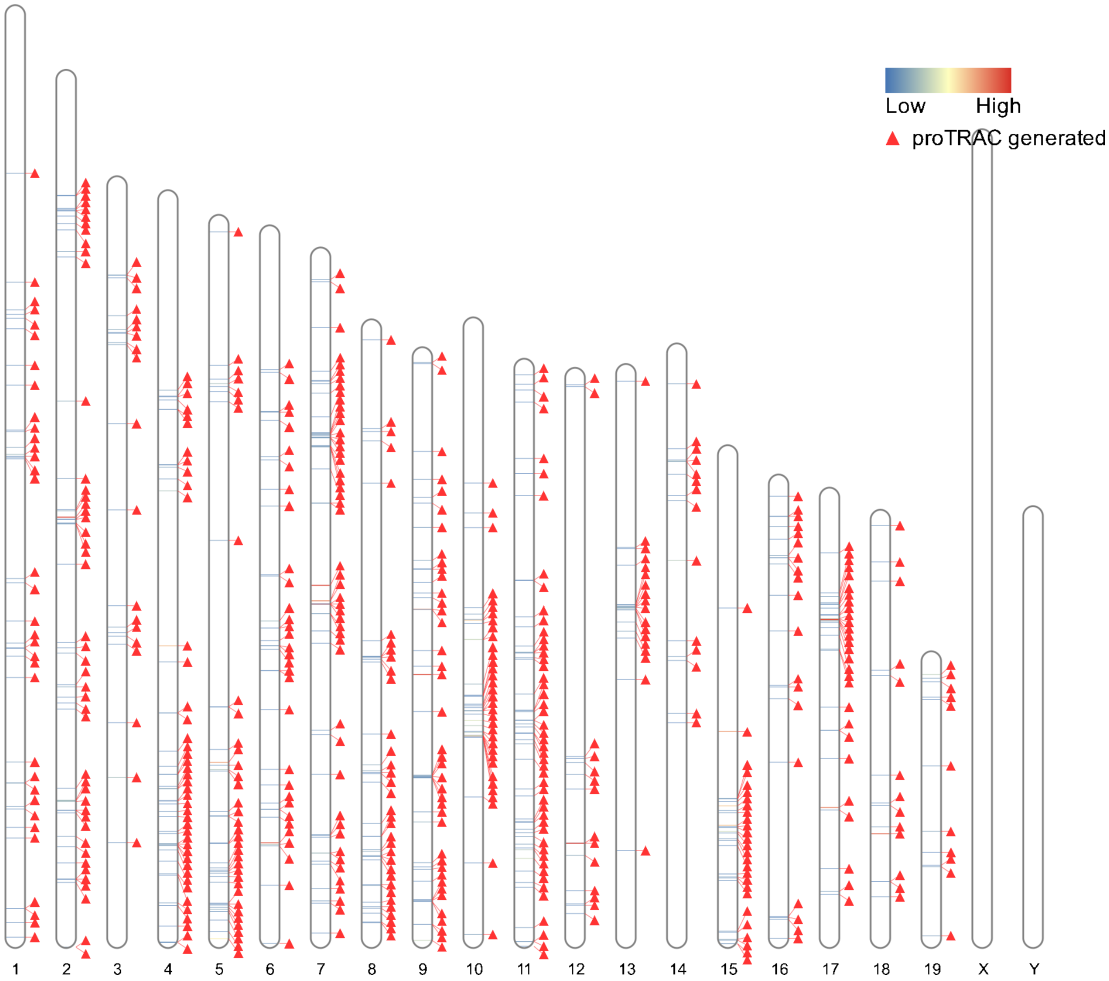


Figure 24: Karyogram with the de novo piRNA clusters. The scale signals the relative expression of the clusters.

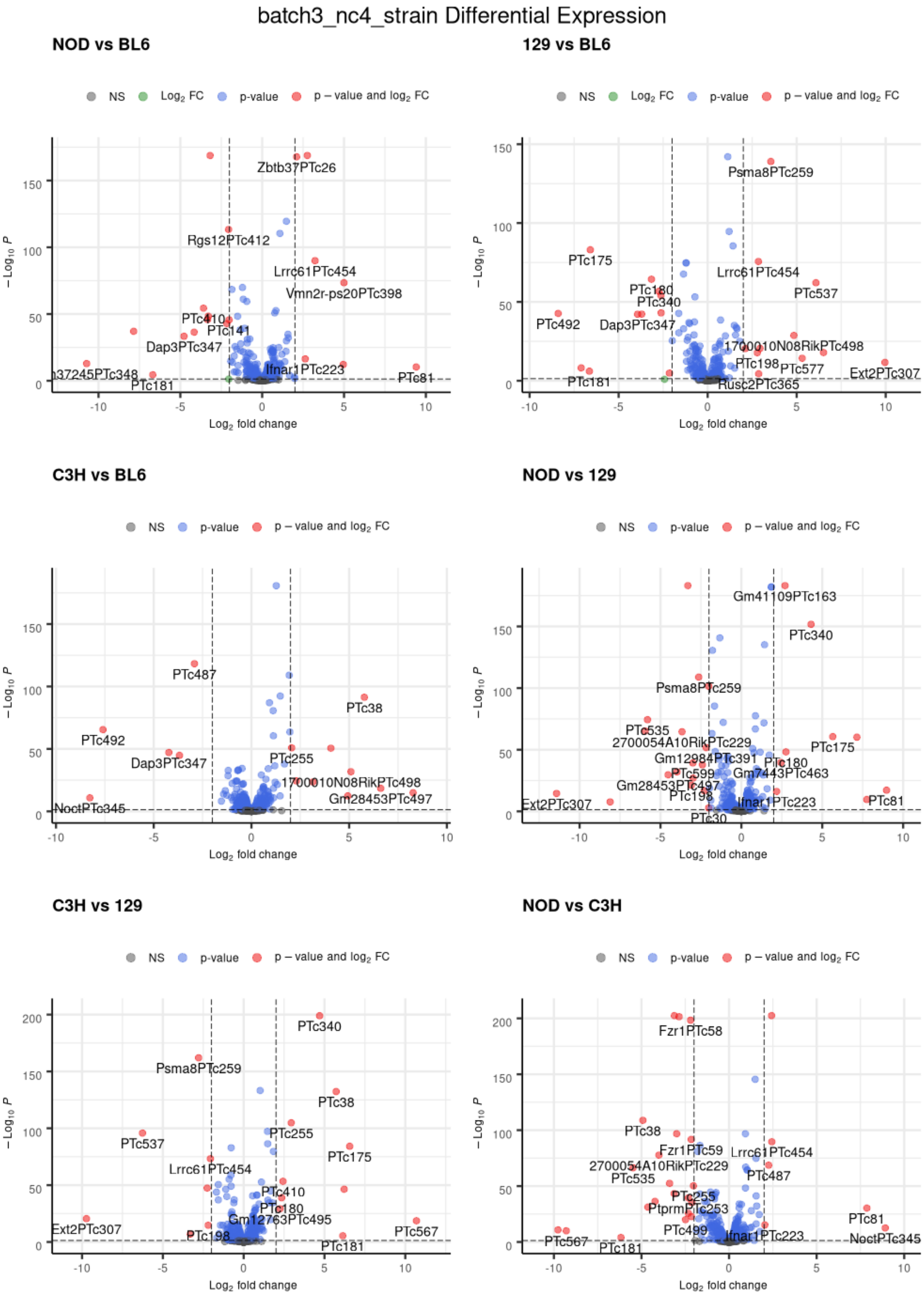


Figure 25: Volcano plots for the six contrast by strain performed on inbred samples using the de novo clusters.

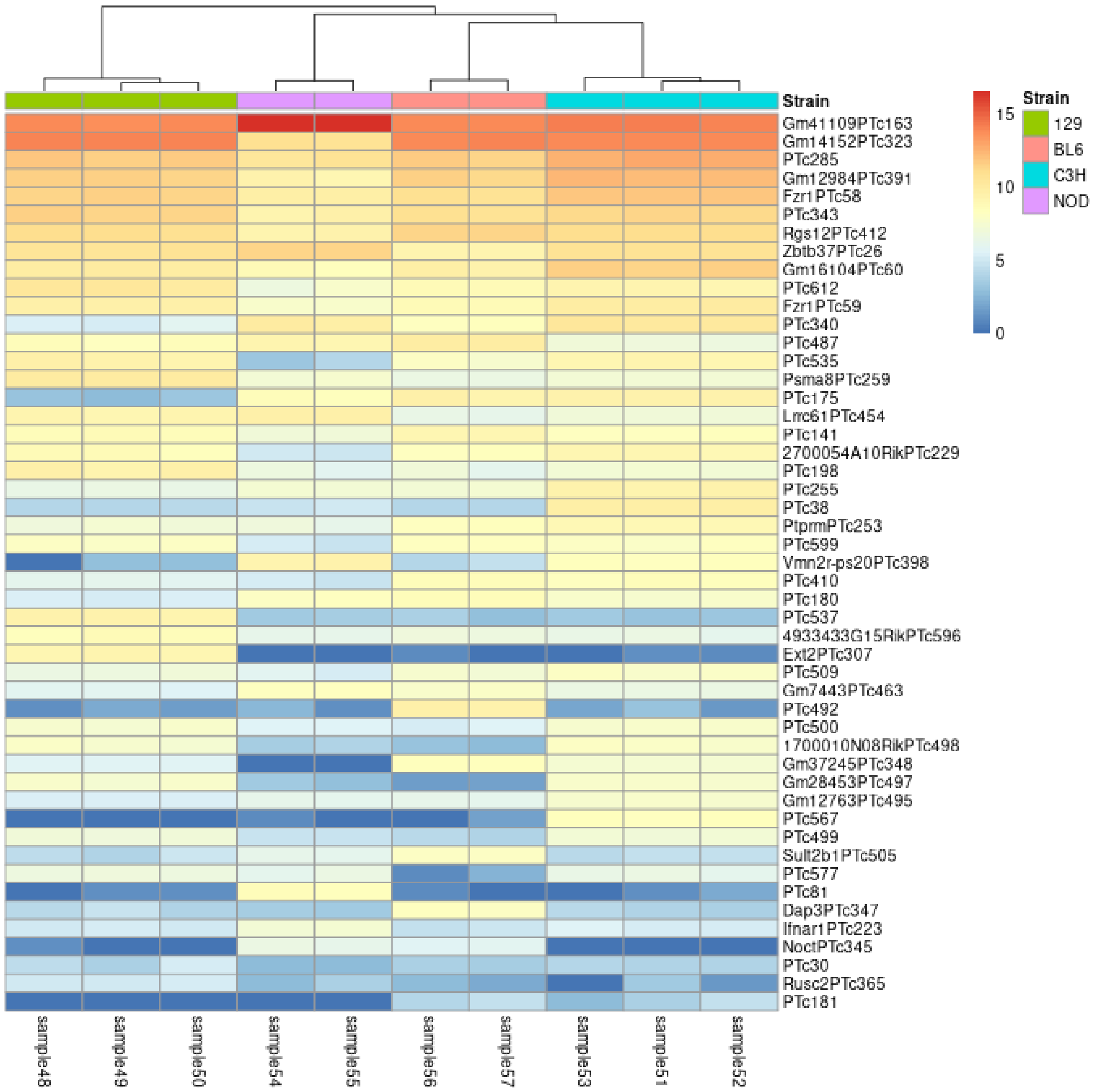


Figure 26: Heatmap of $log_2 foldchange$ for the top differential expressed genes by strain and type using the de novo clusters.

C Appendix: Tools

C.1 Preparation of the raw sequencing data:

- **Cutadapt 2.10:** Finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing reads. [Martin, 2011]
- **FASTX-Toolkit 0.0.14:** A collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing. [Hannon, a]
- **Bowtie 1.2.3:** It aligns short sequences to a reference genome. It first indexes the genome with a Burrows-Wheeler index.[Langmead et al., 2009]
- **SAMtools 1.11:** For reading, writing, editing, indexing and viewing SAM, BAM, CRAM format files. [Li et al., 2009]
- **featureCounts 2.0.1:** part of the Subread package, a general-purpose read aligner which can align both genomic DNA-seq and RNA-seq reads. featureCounts counts reads to genomic features. [Liao et al., 2014]

C.2 Creation of de novo clusters:

- **proTRAC:** a software for probabilistic piRNA cluster detection, visualization and analysis [Rosenkranz]
- **Intervals 0.15.2:** tools for Working with Points and Intervals.
- **ggVennDiagram 0.3:** functions to generate 2-4 sets Venn plots.
- **RIdeogram 0.2.2:** drawing SVG Graphics to Visualize and Map Genome-Wide Data on Idiograms

C.3 Differential Expression Analysis:

- **biomaRt 2.44.4:** nterface to BioMart databases
- **GGally 2.0.0:** ggplot2 extension that adds several functions to reduce the complexity of combining geoms with transformed data.
- **gridExtra 2.3:** functions to arrange multiple grid-based plots on a page.
- **EnhancedVolcano 1.6.0:** volcano plots with enhanced colouring and labeling.
- **ggrepel 0.9.0:** provides geoms for ggplot2 to repel overlapping text labels:
- **tidyr 1.1.2:** tools to create tidy data.

- **RColorBrewer 1.1-2:** ready-to-use color palettes for creating beautiful graphics.
- **pheatmap 1.0.12:** function to draw clustered heatmaps where one has better control over some graphical parameters such as cell size, etc.
- **ggplot2 3.3.2:** a system for declaratively creating graphics.
- **dplyr 1.0.2:** a grammar of data manipulation.
- **rtracklayer 1.48.0:** R interface to genome annotation files and the UCSC genome browser.
- **DESeq2 1.28.1:** Differential gene expression analysis based on the negative binomial distribution.
- **SummarizedExperiment 1.18.2:** SummarizedExperiment container.
- **DelayedArray 0.14.1:** A unified framework for working transparently with on-disk and in-memory array-like datasets.
- **matrixStats 0.57.0:** High-performing functions operating on rows and columns of matrices.
- **Biobase 2.48.0:** Base functions for Bioconductor.
- **GenomicRanges 1.40.0:** Representation and manipulation of genomic intervals.
- **GenomeInfoDb 1.24.2:** Utilities for manipulating chromosome names, including modifying them to follow a particular naming style.
- **IRanges 2.22.2:** Foundation of integer range manipulation in Bioconductor.
- **S4Vectors 0.26.1:** Foundation of vector-like and list-like containers in Bioconductor.
- **BiocGenerics 0.34.0:** S4 generic functions used in Bioconductor
- **ggbio 1.36.0:** Visualization tools for genomic data
- **ensembldb 2.12.1:** Utilities to create and use Ensembl-based annotation databases.
- **EnsDb.Mmusculus.v79 2.99.0:** Ensembl based annotation package for mus musculus.
- **AnnotationFilter 1.12.0:** Facilities for Filtering Bioconductor Annotation Resources.
- **AnnotationDbi 1.50.3:** Manipulation of SQLite-based annotations in Bioconductor.
- **Gviz 1.32.0:** Plotting data and annotation information along genomic coordinates.

D Appendix: Work Plan

We prepared a work plan that included the need to iterate some of the activities and created specific tasks out of the general objectives for the deliverables that have to be produced. We included the deliverables as tasks, but we did not include specific times for them as they would be performed in parallel with the other tasks.

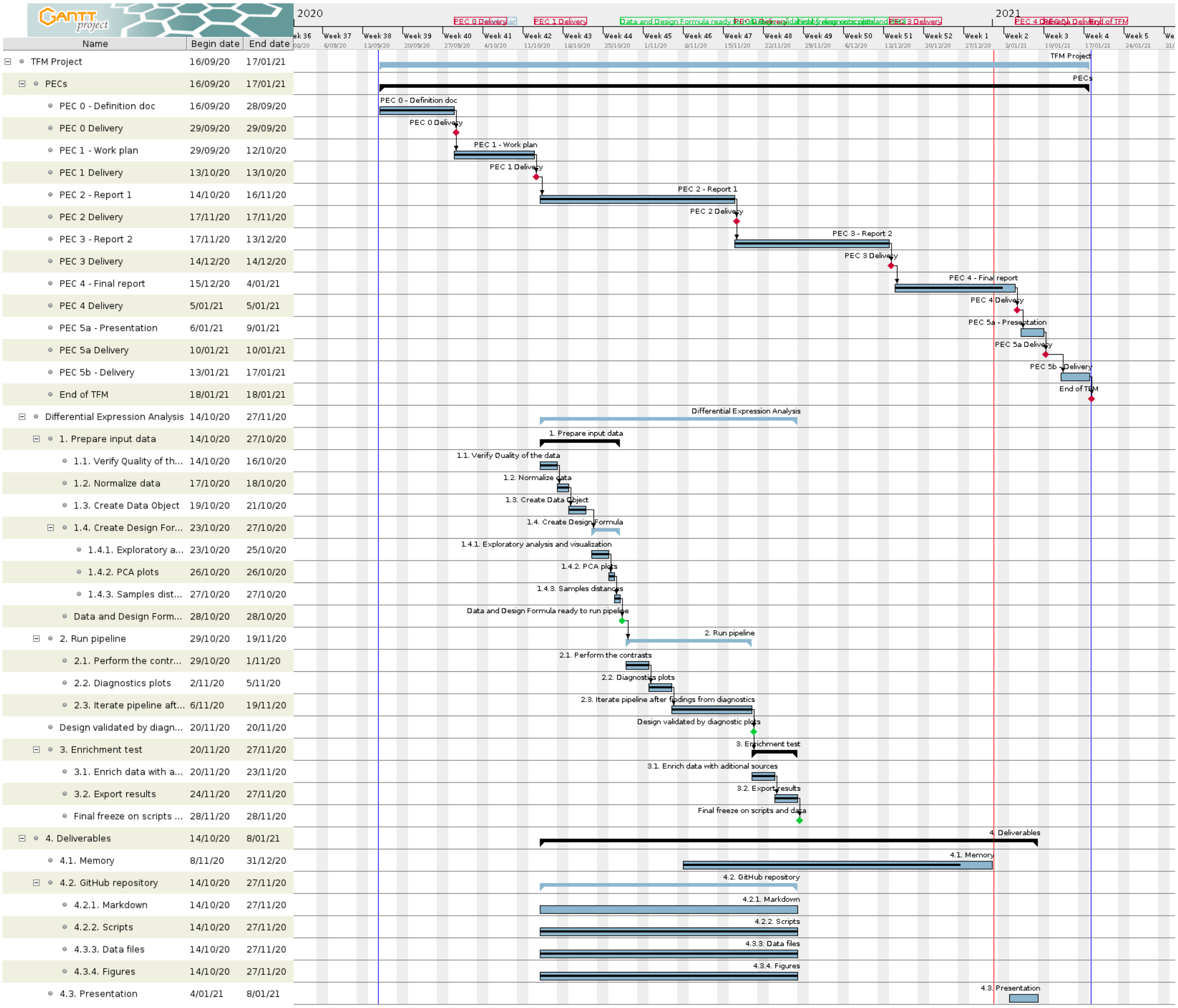
D.1 Tasks

Differential Expression Analysis

1. Prepare input data - 13 days
    - 1.1. Verify Quality of the data -3 days
    - 1.2. Normalize data - 2 days
    - 1.3. Create Data Object - 3 days
    - 1.4. Create Design Formula - 5 days
        - 1.4.1. Exploratory Analysis and Visualization - 3 days
        - 1.4.2. PCA plots - 1 days
        - 1.4.3. Samples distances - 1 days
2. Run pipeline - 22 days
    - 2.1. Perform contrasts - 4 days
    - 2.2. Diagnostics plots - 4 days
    - 2.3. Iterate pipeline after findings from diagnostics - 14 days
3. Enrichment tests - 8 days
    - 3.1. Enrich data with additional sources - 4 days
    - 3.2. Export results - 4 days
4. Deliverables
    - 4.1. Memory
    - 4.2. GitHub repository
        - 4.2.1. Markdown
        - 4.2.2. Scripts
        - 4.2.3. Data files
        - 4.2.4. Figures
    - 4.3. Presentation

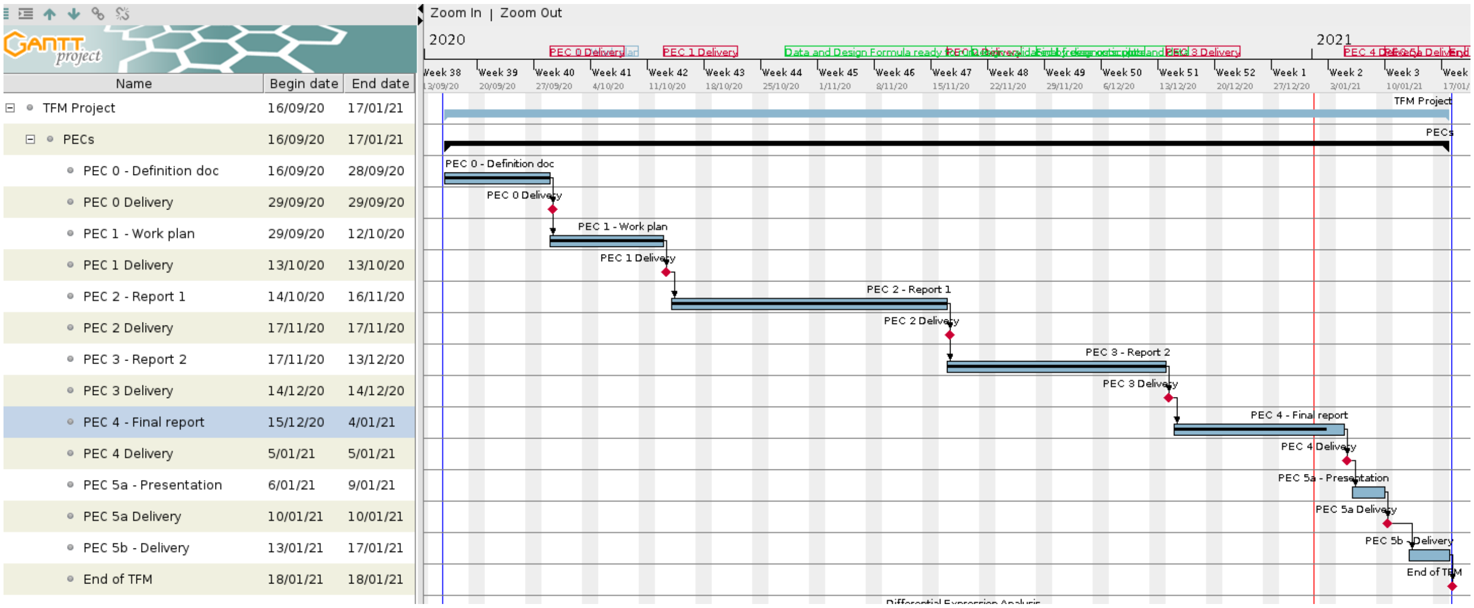
D.2 Calendar

In order to make the different tasks more visible we include not just a generic view of the calendar, but also details of it at a different time resolution.



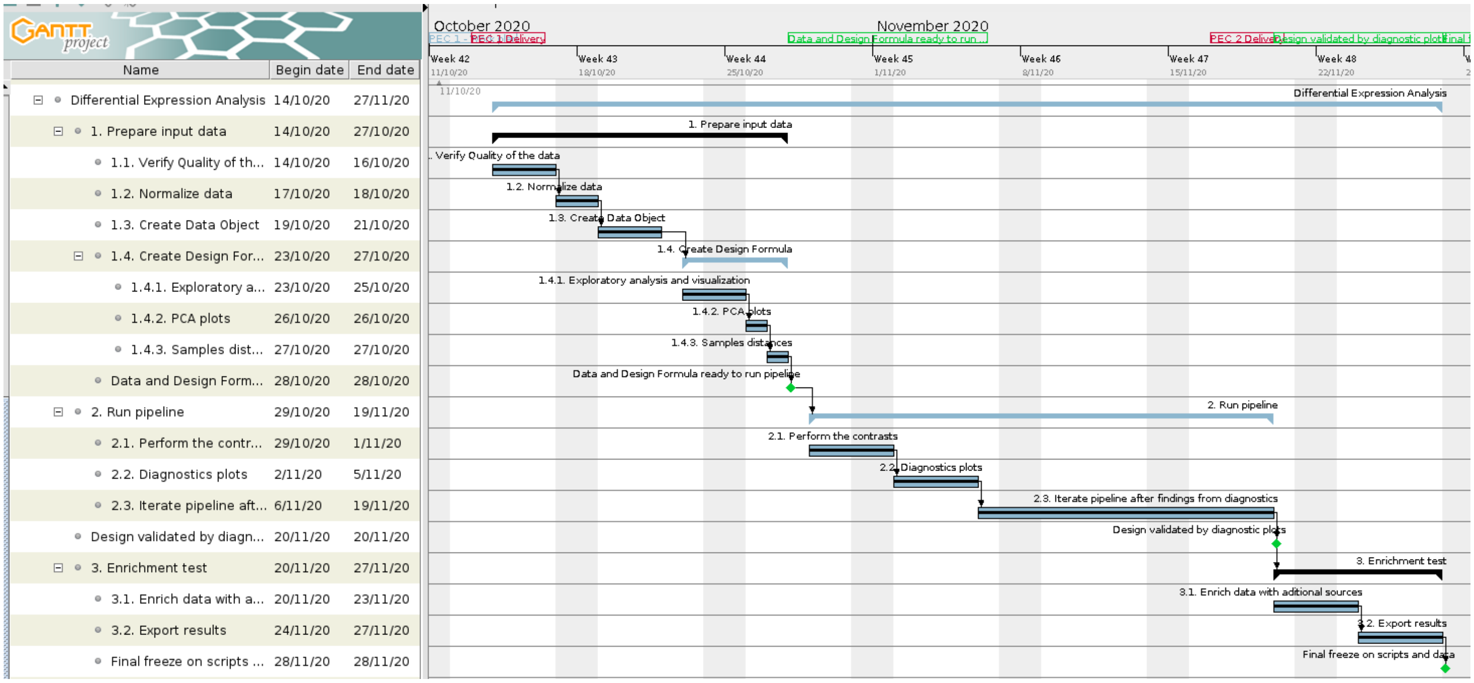
Name	Begin date	End date
TFM Project	16/09/20	17/01/21
PECs	16/09/20	17/01/21
PEC 0 - Definition doc	16/09/20	28/09/20
PEC 0 Delivery	29/09/20	29/09/20
PEC 1 - Work plan	29/09/20	12/10/20
PEC 1 Delivery	13/10/20	13/10/20
PEC 2 - Report 1	14/10/20	16/11/20
PEC 2 Delivery	17/11/20	17/11/20
PEC 3 - Report 2	17/11/20	13/12/20
PEC 3 Delivery	14/12/20	14/12/20
PEC 4 - Final report	15/12/20	4/01/21
PEC 4 Delivery	5/01/21	5/01/21
PEC 5a - Presentation	6/01/21	9/01/21
PEC 5a Delivery	10/01/21	10/01/21
PEC 5b - Delivery	13/01/21	17/01/21
End of TFM	18/01/21	18/01/21
Differential Expression Analysis	14/10/20	27/11/20
1. Prepare input data	14/10/20	27/10/20
1.1. Verify Quality of th...	14/10/20	16/10/20
1.2. Normalize data	17/10/20	18/10/20
1.3. Create Data Object	19/10/20	21/10/20
1.4. Create Design For...	23/10/20	27/10/20
1.4.1. Exploratory a...	23/10/20	25/10/20
1.4.2. PCA plots	26/10/20	26/10/20
1.4.3. Samples dist...	27/10/20	27/10/20
Data and Design Form...	28/10/20	28/10/20
2. Run pipeline	29/10/20	19/11/20
2.1. Perform the contr...	29/10/20	1/11/20
2.2. Diagnostics plots	2/11/20	5/11/20
2.3. Iterate pipeline aft...	6/11/20	19/11/20
Design validated by diagn...	20/11/20	20/11/20
3. Enrichment test	20/11/20	27/11/20
3.1. Enrich data with a...	20/11/20	23/11/20
3.2. Export results	24/11/20	27/11/20
Final freeze on scripts ...	28/11/20	28/11/20
4. Deliverables	14/10/20	8/01/21
4.1. Memory	8/11/20	31/12/20
4.2. GitHub repository	14/10/20	27/11/20
4.2.1. Markdown	14/10/20	27/11/20
4.2.2. Scripts	14/10/20	27/11/20
4.3.3. Data files	14/10/20	27/11/20
4.3.4. Figures	14/10/20	27/11/20
4.3. Presentation	4/01/21	8/01/21

Figure 27: Work plan: General view



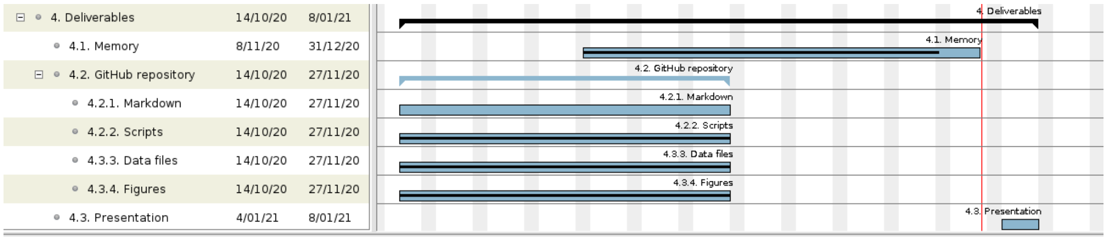
Name	Begin date	End date
TFM Project	16/09/20	17/01/21
PECs	16/09/20	17/01/21
PEC 0 - Definition doc	16/09/20	28/09/20
PEC 0 Delivery	29/09/20	29/09/20
PEC 1 - Work plan	29/09/20	12/10/20
PEC 1 Delivery	13/10/20	13/10/20
PEC 2 - Report 1	14/10/20	16/11/20
PEC 2 Delivery	17/11/20	17/11/20
PEC 3 - Report 2	17/11/20	13/12/20
PEC 3 Delivery	14/12/20	14/12/20
PEC 4 - Final report	15/12/20	4/01/21
PEC 4 Delivery	5/01/21	5/01/21
PEC 5a - Presentation	6/01/21	9/01/21
PEC 5a Delivery	10/01/21	10/01/21
PEC 5b - Delivery	13/01/21	17/01/21
End of TFM	18/01/21	18/01/21

Figure 28: Work plan: Focus on PECs timeline


October 2020
November 2020
PEC 1 -
PEC 1 Delivery
Data and Design Formula ready to run ...
PEC 2 Delivery
Design validated by diagnostic plots
Final f
Week 42
11/10/20
Week 43
18/10/20
Week 44
25/10/20
Week 45
1/11/20
Week 46
8/11/20
Week 47
15/11/20
Week 48
22/11/20
11/10/20
Differential Expression Analysis
1. Prepare input data
. Verify Quality of the data
1.2. Normalize data
1.3. Create Data Object
1.4. Create Design Formula
1.4.1. Exploratory analysis and visualization
1.4.2. PCA plots
1.4.3. Samples distances
Data and Design Formula ready to run pipeline
2. Run pipeline
2.1. Perform the contrasts
2.2. Diagnostics plots
2.3. Iterate pipeline after findings from diagnostics
Design validated by diagnostic plots
3. Enrichment test
3.1. Enrich data with aditional sources
3.2. Export results
Final freeze on scripts and data

Name	Begin date	End date
Differential Expression Analysis	14/10/20	27/11/20
1. Prepare input data	14/10/20	27/10/20
1.1. Verify Quality of th...	14/10/20	16/10/20
1.2. Normalize data	17/10/20	18/10/20
1.3. Create Data Object	19/10/20	21/10/20
1.4. Create Design For...	23/10/20	27/10/20
1.4.1. Exploratory a...	23/10/20	25/10/20
1.4.2. PCA plots	26/10/20	26/10/20
1.4.3. Samples dist...	27/10/20	27/10/20
Data and Design Form...	28/10/20	28/10/20
2. Run pipeline	29/10/20	19/11/20
2.1. Perform the contr...	29/10/20	1/11/20
2.2. Diagnostics plots	2/11/20	5/11/20
2.3. Iterate pipeline aft...	6/11/20	19/11/20
Design validated by diagn...	20/11/20	20/11/20
3. Enrichment test	20/11/20	27/11/20
3.1. Enrich data with a...	20/11/20	23/11/20
3.2. Export results	24/11/20	27/11/20
Final freeze on scripts ...	28/11/20	28/11/20

Figure 29: Work plan: Focus on the Differential Expression Analysis



4. Deliverables	14/10/20	8/01/21
4.1. Memory	8/11/20	31/12/20
4.2. GitHub repository	14/10/20	27/11/20
4.2.1. Markdown	14/10/20	27/11/20
4.2.2. Scripts	14/10/20	27/11/20
4.3.3. Data files	14/10/20	27/11/20
4.3.4. Figures	14/10/20	27/11/20
4.3. Presentation	4/01/21	8/01/21

Figure 30: Work plan: Focus on the deliverables

D.3 Milestones

We defined two types of milestones:

PEC related milestones (mark the deliverables we need to produce and deliver on time):

- 29/09/20: PEC 0 - Definition doc
- 13/10/20: PEC 1 - Work plan
- 17/11/20: PEC 2 - Report 1
- 14/12/20: PEC 3 - Report 2
- 05/01/21: PEC 4 - Final report
- 10/01/21: PEC 5a - Presentation
- 18/01/21: PEC 5b - Delivery

Differential Expression Analysis milestones (those actually related to the actual analysis):

- 28/10/20: Data and Design Formula ready to run pipeline
- 20/11/20: Design validated by diagnostic plots
- 28/11/20: Final freeze on scripts and data. All figures and exports should be prepared by this date.

D.4 Risks analysis

We identified the following risks to the realization of the project:

Time shortage

Mitigation Plan: the tasks and milestones have been planned to account for at least the need to do one rerun of the experiment. In addition several personal activities will be delayed to allow for extra time to work on the TFM.

Scope definition

Mitigation Plan: this is a tricky one, as it will be hard to change midway the scope of the project. To avoid any issues we have shrunk the initial scope to the bare minimum that still is meaningful with the expectation that hopefully we will be able to expand the initial scope if we have the time for it.

Resources shortage

Mitigation plan: The analysis requires certain IT resources that need to be in place and working correctly to be able to deliver the results. We have already checked that we count with the necessary resources and we even have a backup computer if that was necessary.

Data loss

Mitigation plan: all the data used, along with the scripts and documentation created, will be stored in at least two separate locations to prevent any loss in case of catastrophic failure.

Multi factor analysis does not allow to account for all factors

Mitigation plan: when some factors are a lineal combination of others it can be challenging to obtain meaningful data regarding them, also the effect of interactions has to be taken into account on the design matrix. We plan to start using this last method with DESeq2 to see if we can account for Batch and other variables. If it does not perform as expected we will consider moving the analysis to limma, which offers some alternatives to handle this.

Failure to find new significantly variable expressed clusters

Mitigation plan: Not exactly a risk other than in the sense that the scope of the project would be reduced, but still we would need to make the necessary confirmations to confirm the result, which in it self would also be interesting as it would confirm previous findings. Of course we would adjust both the fold change threshold and the α value to make sure that a really high over or under expression of cluster 29 (the previously identified differentially expressed cluster) does not hide other more subtle but still relevant differential expression on other clusters.

E Appendix: Supplementary files

This is a description of the supplementary files that eventually will be available.

Additional file 1:

`/data/private/allSampleIdentifiers_mouseSmallRNA_private_enriched.txt`
Identifiers of all samples, with information for Strain, Batch, Tissue, Phylogenetic tree, Ancestral diet, Metabolic status and presence of IAP Insertion on cluster pic29.

Additional file 2:

`/out/piRNACounts.txt`
Table with all the counts for each cluster, with 3 additional columns for chromosome, and start and end of the cluster on mm_10 assembly.

Additional file 3:

`/data/private/protrac_clusters_int_rellocate_with_sp_noTEVs.gtf`
611 de novo clusters predicted on this project.

Additional file 4:

`/out_DE/batch3_li_strain/results/`
DESeq2 annotated results for the six contrasts by strain using the Li et al. clusters.

Additional file 5:

`/out_DE/batch3_nc4_strain/results/`
DESeq2 annotated results for the six contrasts by strain using the *de novo* clusters.

Additional file 6:

`/figs/cluster_counts_by_IAP_insertion/`
Normalized counts for each cluster grouped by AIP insertion status.

Additional file 7:

`/figs/cluster_counts_not_normalized/`
Raw counts by cluster.

Additional file 8:

`/out_DE/batch1_li_metabolicstatus/results/`
DESeq2 annotated results for the contrasts by metabolic status (ICR) using the Li et al. clusters.