EPiC Series in Computing

Volume 60, 2019, Pages 190–199

Proceedings of 11th International Conference
on Bioinformatics and Computational Biology

Prediction Of Novel Pirna Rat Clusters Based On Mouse Pirna Clusters Using Downstream and Upstream Analysis

Tamer Aldwairi[1,2], Federico Hoffmann[1] and Andy D. Perkins[1]
[1]Mississippi State University
Starkville, MS 39762, USA
[2]Ursinus College
601 E. Main Street
Collegeville, PA 19426, USA
taa70email@gmail.com, fgh19@msstate.edu, perkins@cse.msstate.edu

Abstract

PiRNAs are a particular type of small non-coding RNA. They are distinct from miRNA in size as well as other characteristics, such as the lack of sequence conservation and increased complexity when compared to their miRNA counterparts. PiRNA is considered the largest class of sRNA that is expressed especially in the animal cells. piRNAs are derived from long single-stranded RNAs, which are transcribed from genomic clusters, in contrast to other small silencing RNAs. It has been speculated that one locus could generate more than one piRNA. PiRNA corresponding to repetitive elements is fewer in mammals than in other species like Drosophila and Danio rerio, which signifies that piRNA might have possessed or gained some additional functionality in mammals. While the functionality of piRNAs may not be fully understood, they are believed to be involved in gene silencing. In this paper, we will examine a novel approach to identify potential piRNA clusters based on genes downstream and upstream location and order.

1 Introduction

The inauguration of next-generation sequencing technologies has allowed us to better understand mechanisms such as gene regulation and especially gene silencing and its effects on the production of proteins. Proteins can be mediators and possess components for gene silencing. One of those proteins possessing components for gene silencing complexes is the Argonaute (Ago) protein, which is usually

O. Eulenstein, H. Al-Mubaid and Q. Ding (eds.), BiCOB 2019 (EPiC Series in Computing, vol. 60),
pp. 190–199

involved in RNA silencing [2]. The Ago proteins can be divided into two main subclades: the Ago subclade and the Piwi subclade. The biological function and mechanisms of the Ago subclade are much better known than the Piwi subclade since the former have been researched and investigated more in the past years. The Piwi subfamily has been identified in different organisms based upon its primary sequence, and it was found that most of those proteins are distinct to germ cells which suggest germline conserved functionality [11].

A small RNA that came to be known as Piwi-RNA (piRNA) was identified to interact with the Piwi subfamily [11]. piRNA is distinguished from other small RNAs by its sequence length and it's 5' uridine (U) residue biases. piRNAs are found within clusters, especially in mammals. Cluster positions are conserved, but the sequences within them are not [12]. Identifying piRNA in different species especially nonmodel organisms is considered a difficult problem due to the lack of conservation in secondary structure and sequence homology [7]. Added difficulty comes from the complexity and variation of transposable elements across eukaryotic genomes.

Transposable element variety could be attributed to many reasons and can be defined at many different levels of granularity. For example, the variation in the percentage of transposon-derived DNA in a genome and the element diversity within those eukaryotic genomes varies a lot. The transposon load ranges from a few percents in certain organisms to more than 90% percent in others. While the number of classes ranges from a few in certain species to more than 150 classes in others [3].

Finding and predicting piRNA clusters is an important field in bioinformatics due to the functionality of piRNA in germline development and because full functionality of piRNA are not yet determined. In our research, we used a computational approach to find potential rat piRNAs clusters based on the similarity between genes surrounding the mouse piRNA clusters and those within the rat genome. Even though there are previous studies that have addressed the issue of predicting piRNA. These studies [9, 17, 18, 19] differ in the way they address the problem; some of them address the prediction of piRNA transcripts [9] instead of predicting piRNA clusters; while others assume a uniform distribution of non-piRNA regions within the genome [18] and others refute that possibility [19]. Our study makes use of data sets generated by two of those studies [17, 18] to prove the validity of our prediction approach.

2 Background

2.1 PiRNA characteristics and their origins

PiRNA is small non-coding RNAs of length 26-31 nucleotides found in animal genomes that interact with proteins of the Piwi subfamily [1, 2]. The Piwi subfamily has been associated with the development of germ cells, self-renewal of stem cells and silencing of retrotransposons [5, 6]. Piwi proteins were initially named rasiRNAs because of their association with repeat elements and later named piRNAs based on their protein interaction patterns. piRNA is usually located in clusters within the genome. The functionality of the piRNA is not fully known but it appears to target transposons and is involved in defending the genome against transposable elements [3, 14, 15]. Such a connection has not been fully explored in mammals.

The full functionality and biogenesis of piRNA are not fully understood. This could be due mainly to the variation of piRNA sequences and the diversity of Piwi functionality [4]. It is not clear how

piRNA is generated but it is suggested that their biogenesis is different from both miRNA and siRNA due to the differences between the Ago and the Piwi proteins [2].

PiRNAs differ from other small RNAs in their length which is longer than miRNAs and siRNAs in mammals [8]. They are derived from one strand and show no clear secondary structural motifs [1]. piRNAs are known to lack primary sequence conservation.

PiRNAs are usually located in clusters within the genome. These clusters contain 10s to 1000s of piRNAs, and clusters sizes range from 1-100s of kb. It is also suggested that many piRNAs could be generated from one locus due to the fact that many piRNAs can be mapped to a small number of genomic loci. It is speculated that piRNA might have gained additional functionality in mammals as it has been observed that there are fewer piRNAs corresponding to repetitive elements in mammals compared to piRNAs in Drosophila and Danio rerio [8]. It has been argued through genetic analysis of flies and zebrafish that the Dicer enzyme plays a role in piRNA as it does in both miRNA and siRNA [14].

Certain studies indicate that piRNA is not generated from dsRNA precursors and that piRNA clusters, even though generated from both strands, could in some cases be mapped to one strand. Those reasons and others aided in the development of two main models to explain the origins of piRNA. The first model suggests that the generation of piRNA is done through processing of small RNA from long single-stranded transcripts. The second model suggests that the piRNA could be made as primary transcription products [14].

When piRNA recognizes a transposon, the mRNA is cleaved, which results in the destruction of the mRNA. As a result of the cleavage process, secondary piRNA derived from the mRNA is produced. This happens if the piRNA targets the sense part of the transposon. If it targets the antisense part of the transposon then the cleavage process results in the reproduction of the antisense piRNA, which can target transposons in a similar manner. This cycle of production and reproduction of the piRNAs is called the ping pong cycle [13].

PiRNAs from mammals and flies are usually derived from the post-transcriptional amplification, ping pong model [14].

It is important to note here that current methods for detecting piRNA do not necessarily detect lowly expressed piRNAs. That is why computational methods might provide an alternative approach for detecting piRNAs [7]

2.2 Transposable elements and their roles in piRNA

Transposable elements are genetic elements that have the potential of causing harm or damage to the host cell through its continuous movement within the genome (insertion, deletion, duplication). However, most transposable elements are in a non-active state meaning they are not duplicating or moving from one place to another in the genome. Even though active transposable elements might be potentially harmful, the genome has developed epigenetic mechanisms to suppress their activity [3, 16]. In a single cell, transposable elements can change their positions within genomes. Class I transposable elements, called retrotransposons can increase the copy number within the genome but class II, called DNA-transposons, do not. Generally speaking, transposable elements can have either a positive or a negative impact on the organization of an organism's genome and its progeny [3].

Retrotransposons movements do not necessitate a negative impact on the organism. The threat to the cell or the genome comes from the large variety of transposable element classes. Therefore the cell

should find a way to cope with such variety by controlling features that are common or shared among all the transposons. Organisms are responsible for transmitting a non-defective and working copy of their genome to their offspring. That is why these organisms have developed ways to adapt to the changing threats to protect the genomes within their germline cells. A key transposon regulator for this process is RNAi and the Piwi clade (Piwi, AGO3, Aubergine), which is part of the Argonaute protein, and essential for this regulation. While on the other hand the canonical Argonaute has been shown to be nonessential in the process [3].

3 Methods

3.1 Exploring candidate rat piRNAs

Recent comparisons between piRNA clusters among mice and rats suggest that expansions in the number of piRNA clusters are driven by increases in transposable element activity. To find out if the difference between piRNA clusters is due to genome annotation or are real biological differences between those species, we decided to choose two animal species for our piRNA analysis the mouse and rat. We chose to analyze two main databases known to contain piRNA data, piRNA Bank [10] and Johannes Gutenberg University of Mainz piNRA database (JGU database) [20]. We noticed in the latest release of piRNA bank that the number of piRNA clusters in the mouse (2710) are greater than 14 folds more than the one's in the rat (189). While in the JGU database the number of piRNAs is (171) in the mouse and (168) in the rat which shows close resemblance in the number of piRNAs found in both species. From the previous observations, we found out that the current methodologies of predicting piRNAs might not be correctly presenting all the piRNAs that are within the genome.

This observation led us to investigate the possibility that there are rat piRNA clusters that have not been found yet. We did that by mapping the 2 genes upstream and 2 genes downstream that surround the mouse clusters to their corresponding genes in the rat genome. Our approach shown in Figure 1 below can be summarized in five main steps:-

1. In the first step, we extract the mouse and rat genes from the source files of mouse and rat (in our case it was an Ensemble GTF files) and create a file that contains the names of the genes, the chromosome they are on and the starting and ending ranges of each gene. It is important to note here that we chose the largest range possible for the start and end of each gene. This means that if there are different alleles for the same gene we find the starting and ending positions that will guarantee that all the gene alleles are included in that range.

2. In the second step using the information that we gathered in the first step and the piRNA files for mouse and rat from both databases, we create a file that contains the closest two genes on the left and the right of the piRNA for both species and for both the databases we are using.

3. In the third step, we utilize the Bio-Mart Database to find the rat genes that mirror the mouse genes through mapping the gene names of the mouse with their associated gene names in the rat using the Ensembl Genes database.

4. In the fourth step, we compare each gene name and its location from the list of genes surrounding the piRNA in the mouse with the corresponding gene name and its location in the rat to create a link between the genes surrounding the mouse piRNA and the rat genes. If the rat gene had multiple corresponding gene names for the same gene name in mouse, then different rat gene names will be treated as separate entities.

5. In the fifth and last step, we find the genes in the rat that have a direct correspondence to the mouse genes in the same or the reverse order and if they contain piRNAs within them. Using the same approach we try to find candidate piRNA clusters in the rat. Figure 1: Below shows the steps used in our approach

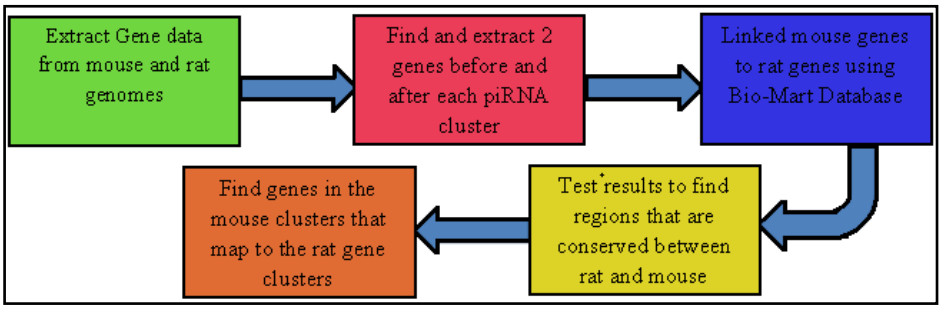

Figure 1: The steps used to find the candidate rat piRNA clusters

3.2 Filtering and extracting the Data

The filtering process goes as follows we start by filtering the piRNA clusters that not contain an exact 4 gene matches within the rat genome. We then find the indices of the rat piRNA clusters that have a one to one match with the mouse piRNA clusters. This means that for the mouse piRNA there is a corresponding, existing rat piRNA with the genes surrounding it has the same order and locations as the mouse piRNA. These rat piRNAs are removed when we are predicting new piRNAs since they are already identified as piRNA by other databases and methods. We then check if the rat genes surrounding our potential piRNAs clusters are on the same chromosome and if that the second gene ends before the start of the third gene (this is done mainly to find out if there is a space between the genes for a piRNA cluster to exist between them).

We then extract the potential rat piRNAs from the rat genome using the locations (chromosome, start and end positions) calculated through the one to one matching with the mouse piRNA clusters. Large Sequence ranges above a certain threshold were neglected due to the belief that such potential clusters would be too large and are not feasible. The threshold was set as the mouse maximum cluster size + 2 STD. The output is a text file that contains all the rat piRNA clusters that are below the threshold.

4 Results and Discussion

We analyzed the results from the approach we used and we concluded that the mouse genes surrounding the mouse piRNA clusters map to the genes in the rat clusters according to a number of patterns. We classified those patterns into five main cases, the normal case, the reverse case, the mixed case, the missing link, and the random case. Our classification takes into account the order and the locations at which the genes surrounding the rat cluster appear in comparison with the mouse piRNA clusters. It is important to note here that some of the rat genes that are the same as the mouse genes might have different names that is why the Bio-Mart database is used.

The Figures (3.2 to 3.6) below show these different cases. In the figures, we used the same colors to specify similar genes and we specified numbers on their right to indicate the order of appearance around the gene.

In the normal case, the genes surrounding the piRNA cluster in mouse have an exact match regarding their locations and the order of the genes in mouse and their corresponding rat genes. The normal case happens when we have a one to one correspondence between the genes in both mouse and rat, and the order in which they appear upstream or downstream is the same.

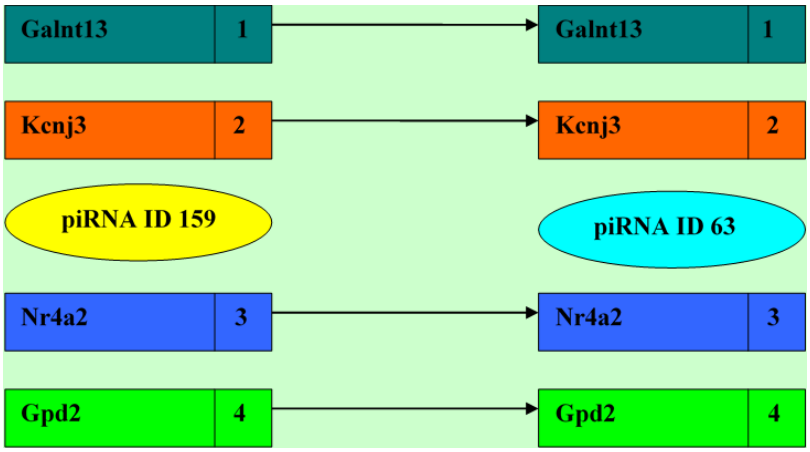

Figure 2: Normal case one to one correspondence

In the reverse case, we noticed that the 2 genes that were directly before the mouse piRNA cluster came directly after the rat piRNA cluster and vice versa. This could be due to a real change of order or could be due to the sequence read from the opposite direction, the reverse strand.

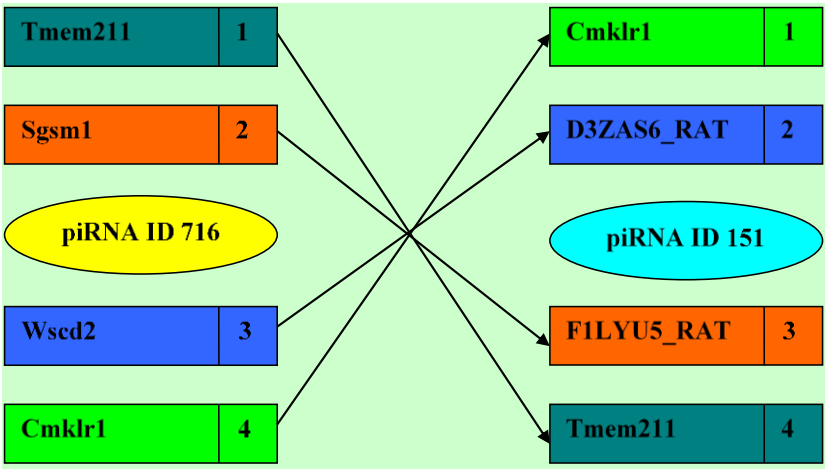

Figure 3: Reverse Case Genes match opposite sides of the piRNA

In the special or mixed case, the genes surrounding the rat candidate clusters have a mix of the normal and the reverse cases together

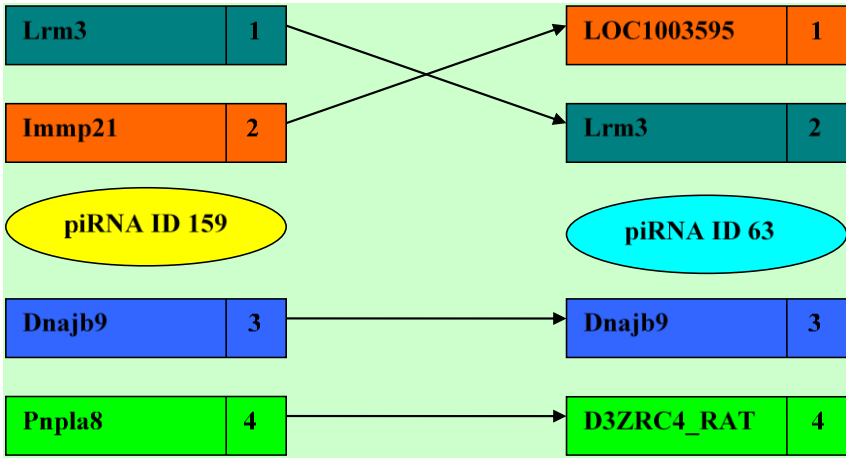

Figure 4: Mixed case which is a mix of normal and opposite matches

In the missing link case, one of the genes is missing a link meaning that two genes in mouse point to one gene in rat and so the other gene has a missing link. This case could have other possibilities too like the other way around. In the figure below, we can see that both the third and the fourth gene in mouse point to the fourth gene in the rat.

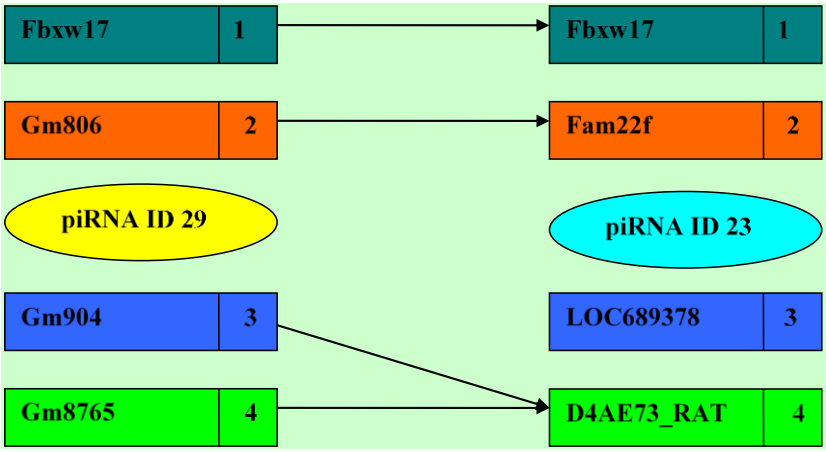

Figure 5: Missing link case were two genes in mouse point to one gene in rat

In the Random case, we found that the genes surrounding the mouse have no specific order in the rat piRNA cluster. The genes that are surrounding the mouse piRNA for different piRNA IDs could be mapped to different rat potential piRNAs with different genes order, numbers and locations.

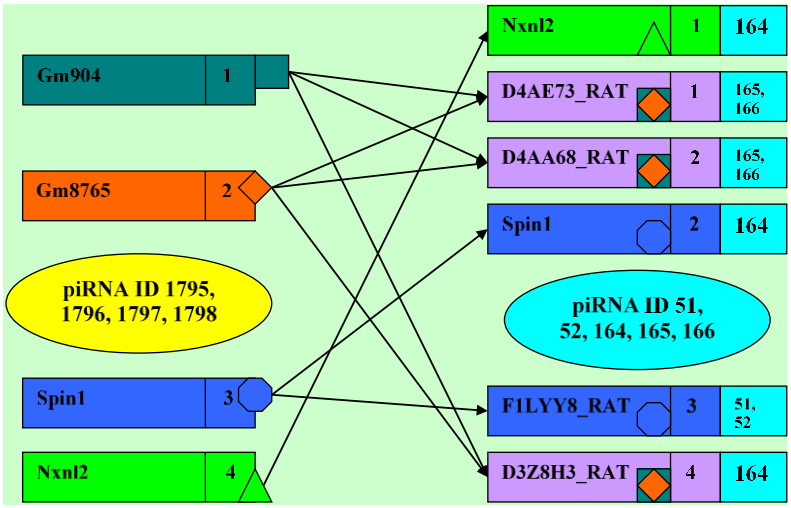

Figure 6: Random case where there are no specific patterns for the matches

Analyzing the piRNA data for the mouse and the rat data extracted from the JGU Database, we observed that if we take the normal case and the reverse case (since the reverse case could be the genes just being on the opposite strand based) we found that we are able to identify 18 piRNA matches (17 were unique matches) out of the 168 piRNAs in rat. A match here means that we are able to identify within the region between the genes in rat, a piRNA that was identified by the JGU database or by the piRNA bank database in the other case. This accounts for approximately more than 10 % of the possible piRNAs that were found in the rat. In addition, if we look at the piRNA data extracted from piRNA bank we found we are able to identify 15 piRNA matches (11 were unique matches) out of the 189 piRNA clusters in the rat. This accounts for approximately 8% of the possible piRNAs that were found in the rat. We also found that including other cases like the missing link case or the special case will increase the percentage of piRNA matches found in rat depending on the number of cases that we add. These results clearly show that it is possible to predict certain piRNAs through mapping the genes upstream and downstream surrounding the piRNA.

5 Conclusions

The upstream and downstream genes for each cluster of the mouse piRNA was mapped to the corresponding rat genes. The mapping of these genes resulted in a number of relationships. These relations range from a one-to-one direct mapping relation to a complicated complex random mapping relation among the genes. Using the direct mapping relation we identified a number of genomic regions in rat for further exploration. Those candidate regions can be divided into two groups. The first group is regions that match an already known piRNA cluster that was identified by other databases. While the second group has no known piRNA cluster, but co-linearity of homologous genes indicate possible conservation of the piRNA cluster.

Competing Interests

The authors declare that they have no competing interest.

Authors' Contributions

TA, AP, and FH developed the concept. TA, AP wrote the manuscript.

Acknowledgments

This work was supported by the National Science Foundation under award EPS-0903787.

References

[1] “Molecular Biology Select,” Cell, vol. 126, no. 2, pp. 223–225, Jul. 2006.

[2] A. G. Seto, R. E. Kingston, and N. C. Lau, “The Coming of Age for Piwi Proteins,” Molecular Cell, vol. 26, no. 5, pp. 603–609, Aug. 2007.

[3] C. D. Malone and G. J. Hannon, “Molecular Evolution of piRNA and Transposon Control Pathways in Drosophila,” Cold Spring Harb Symp Quant Biol, vol. 74, pp. 225–234, 2009.

[4] G. Wang and V. Reinke, “A C. elegans Piwi, PRG-1, regulates 21U-RNAs during spermatogenesis,” Curr. Biol., vol. 18, no. 12, pp. 861–867, Jun. 2008.

[5] D. N. Cox, A. Chao, and H. Lin, “piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells,” Development, vol. 127, no. 3, pp. 503–514, Feb. 2000.

[6] A. I. Kalmykova, M. S. Klenov, and V. A. Gvozdev, “Argonaute protein PIWI controls mobilization of retrotransposons in the Drosophila male germline,” Nucleic Acids Res., vol. 33, no. 6, pp. 2052–2059, 2005.

[7] Y. Zhang, X. Wang, and L. Kang, “A k-mer scheme to predict piRNAs and characterize locust piRNAs,” Bioinformatics, vol. 27, no. 6, pp. 771–776, Mar. 2011.

[8] K. A. O’Donnell and J. D. Boeke, “Mighty Piwis Defend the Germline against Genome Intruders,” Cell, vol. 129, no. 1, pp. 37–44, Jun. 2007.

[9] Y. Zhang, X. Wang, and L. Kang, “A k-mer scheme to predict piRNAs and characterize locust piRNAs,” Bioinformatics, vol. 27, no. 6, pp. 771–776, Mar. 2011.

[10] “piRNA Database.” [Online]. Available: http://pirnabank.ibab.ac.in/. [Accessed: 28-Oct-2018].

[11] L. Peters and G. Meister, “Argonaute Proteins: Mediators of RNA Silencing,” Molecular Cell, vol. 26, no. 5, pp. 611–623, Aug. 2007.

[12] C. D. Malone and G. J. Hannon, “Small RNAs as Guardians of the Genome,” Cell, vol. 136, no. 4, pp. 656–668, Feb. 2009.

[13] A. A. Aravin, R. Sachidanandam, D. Bourc'his, C. Schaefer, D. Pezic, K. F. Toth, T. Bestor, and G. J. Hannon, "A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice," Mol. Cell, vol. 31, no. 6, pp. 785–799, Sep. 2008.

[14] A. A. Aravin, G. J. Hannon, and J. Brennecke, "The Piwi-piRNA Pathway Provides an Adaptive Defense in the Transposon Arms Race," Science, vol. 318, no. 5851, pp. 761–764, Nov. 2007.

[15] V. V. Vagin, A. Sigova, C. Li, H. Seitz, V. Gvozdev, and P. D. Zamore, "A distinct small RNA pathway silences selfish genetic elements in the germline," Science, vol. 313, no. 5785, pp. 320–324, Jul. 2006.

[16] R. K. Slotkin and R. Martienssen, "Transposable elements and the epigenetic regulation of the genome," Nat Rev Genet, vol. 8, no. 4, pp. 272–285, Apr. 2007.

[17] S. S. Lakshmi and S. Agrawal, "piRNABank: a web resource on classified and clustered Piwi-interacting RNAs," Nucl. Acids Res., vol. 36, no. suppl 1, pp. D173–D177, Jan. 2008.

[18] D. Rosenkranz and H. Zischler, "proTRAC - a software for probabilistic piRNA cluster detection, visualization and analysis," BMC Bioinformatics, vol. 13, no. 1, p. 5, Jan. 2012.

[19] I. Jung, J. C. Park, and S. Kim, "piClust: A density based piRNA clustering algorithm," Computational Biology and Chemistry, vol. 50, pp. 60–67, Jun. 2014.

[20] D. Rosenkranz and H. Zischler, "proTRAC - a software for probabilistic piRNA cluster detection, visualization and analysis," BMC Bioinformatics, vol. 13, no. 1, p. 5, 2012.