Bioinformatics, 32(8), 2016, 1170–1177
doi: 10.1093/bioinformatics/btv729
Advance Access Publication Date: 16 December 2015
Original Paper

OXFORD

Gene expression

Computational identification of piRNA targets on mouse mRNAs

Jiao Yuan[1,2,†], Peng Zhang[1,†], Ya Cui[2], Jiajia Wang[1], Geir Skogerbø[2], Da-Wei Huang[1], Runsheng Chen[2,*] and Shunmin He[1,*]

[1]Key Laboratory of the Zoological Systematics and Evolution, Institute of Zoology and [2]CAS Key Laboratory of Rna Biology, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Ivo Hofacker

Received on 16 July 2015; revised on 20 November 2015; accepted on 9 December 2015

Abstract

Motivation: PIWI-interacting RNAs (piRNAs) are a class of small non-coding RNAs that are highly abundant in the germline. One important role of piRNAs is to defend genome integrity by guiding PIWI proteins to silence transposable elements (TEs), which have a high potential to cause deleterious effects on their host. The mechanism of piRNA-mediated post-transcriptional silencing was also observed to affect mRNAs, suggesting that piRNAs might play a broad role in gene expression regulation. However, there has been no systematic report with regard to how many protein-coding genes might be targeted and regulated by piRNAs.

Results: We trained a support vector machine classifier based on a combination of Miwi CLIP-Seq-derived features and position-derived features to predict the potential targets of piRNAs on mRNAs in the mouse. Reanalysis of a published microarray dataset suggested that the expression level of the 2587 protein-coding genes predicted as piRNA targets showed significant upregulation as a whole after abolishing the slicer activity of Miwi, supporting the conclusion that they are subject to piRNA-mediated regulation.

Availability and implementation: A web version of the method called pirnaPre as well as our results for browse is available at http://www.regulatoryrna.org/software/piRNA/piRNA_target_mRNA/index.php.

Contact: crs@sun5.ibp.ac.cn or heshunmin@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

PIWI-interacting RNAs (piRNAs), microRNAs (miRNAs) and short-interfering RNAs (siRNAs) are three major classes of small RNAs (Saxe and Lin, 2011). piRNAs differ from miRNAs and siRNAs in several aspects: (1) piRNAs show specific expression in germ cells (Houwing *et al.*, 2007); (2) most piRNAs have lengths in the range of 25–33 nt (Sai Lakshmi and Agrawal, 2008), greater than those of miRNAs and siRNAs, which are 20–23 nt in most cases, although *Caenorhabditis elegans* piRNAs with the unusual length of 21 nt constitute an exception to the general rule (Kim *et al.*, 2009; Weick and Miska, 2014); (3) piRNAs do not show any conservation in either sequence or secondary structure(Le Thomas *et al.*, 2014), except that there is a preference for a 5′ uridine (U) residue and an adenosine as the tenth nucleotide(Dannemann *et al.*, 2012); (4) piRNAs bind to Piwi-clade Argonauts, whereas miRNAs and siRNAs associate with the Ago-clade Argonauts (Khurana and Theurkauf, 2008); and (5) most piRNAs are derived from genomic piRNA clusters, which are discrete regions containing a large number of various types of transposable elements (TEs) (Theron *et al.*, 2014). Besides, piRNAs have distinguished biogenesis mechanisms which are independent of Dicer (Aravin *et al.*, 2007). The recently reported phased manner of piRNA biogenesis further supports that

piRNAs are more diverse than other classes of small RNAs (Han *et al.*, 2015; Mohn *et al.*, 2015; Siomi and Siomi, 2015).

The existence of piRNAs represents adaptive control mechanisms that protect the genomic architecture against TEs, which constitute a large fraction of the mammalian genome and are a constant threat to the host (Kelleher and Barbash, 2013; Reuter *et al.*, 2011; Zamudio and Bourc'his, 2010). PIWI proteins participate in the piRNA pathway(Le Thomas *et al.*, 2013; Ross *et al.*, 2014; Siomi *et al.*, 2011). The mutation of *Miwi* in the mouse results in male infertility and the upregulation of LINE1 retrotransposon transcripts (Reuter *et al.*, 2011). There have been similar observations made in flies when mutations in *piwi*, *aub* or *Ago3* occurred (Siomi *et al.*, 2011). The silencing machinery is similar to that of siRNAs and miRNAs (Ishizu *et al.*, 2012; Reuter *et al.*, 2011). A piRNA recognizes a transposon transcript target by complementarity, and the interacting PIWI protein then slices via its catalytic domain, cleaving the target at a position 10-nt downstream of the 5′ end of the piRNA, generating a 5'-monophosphate containing fragment (Ishizu *et al.*, 2012; Reuter *et al.*, 2011).

Recent studies suggest that piRNA-mediated cleavage acts not only on TEs (Weick and Miska, 2014) but also on mRNAs (Aravin *et al.*, 2001; Nishida *et al.*, 2007; Rouget *et al.*, 2010; Saito *et al.*, 2009; Zhang *et al.*, 2015). In silkworm, Kiuchi *et al.* (2014) even discovered that a single piRNA is responsible for primary sex determination by mediating cleavage of the Masc mRNA, demonstrating the functional importance of piRNA-mediated cleavage on mRNA. Our previous study revealed widespread Miwi/piRNA-mediated mRNA cleavage events in mouse testes and demonstrated the functional importance of the temporal cleavage of piRNA target mRNAs for spermiogenesis (Zhang *et al.*, 2015). Taking advantage of published global 5′ RACE (rapid amplification of 5′ complementary DNA ends) tags, we previously identified 169 piRNA-targeted mRNAs (Zhang *et al.*, 2015). The role of piRNAs in mediating mRNA cleavage is also evidenced in *C. elegans* (Lee *et al.*, 2012) and *D. melanogaster* (Brower-Toland*et al.*, 2007). However, there has been no systematic assessment with respect to how many protein-coding genes might be regulated by a piRNA-guided cleavage mechanism.

In this study, we trained a support vector machine (SVM) classifier to predict potential piRNA targets on protein-coding genes. Because base-pairing recognition by piRNAs as well as the physical association of Miwi are both critical for piRNA-guided cleavage in mouse male germ cells, the anti-Miwi crosslinking immunoprecipitation coupled with deep sequencing (CLIP-seq) data that we produced recently was used to obtain both piRNAs and target fragments. In particular, we introduced a combination of CLIP-Seq-derived features and position-derived features. In contrast to the previous identification of piRNA targets from 5′ RACE tags, which is restricted to the sequencing depth, we have undertaken a genome-wide scan across all sites of mRNA transcripts and computationally identified 3781 mRNAs of 2587 protein-coding genes as potential piRNA targets. This result was further validated by a microarray dataset that presented gene expression differences between catalytic mutant without Miwi slicer activity and the wild type. The dataset of piRNA targets in mRNAs is available for browsing, searching and download via http://www.regulatoryrna.org/software/piRNA/piRNA_target_mRNA/index.php.

2 Methods

2.1 Biologically relevant dataset

The potential piRNA target sites that we recently identified (Zhang *et al.*, 2015) were used as positive examples of our training dataset. We have previously reported 169 protein-coding genes with 193 target sites of piRNAs. Two protein-coding genes were removed because they have multiple genomic loci on different chromosomes. Consequently, the positive examples of our training dataset consist of 191 target sites in 167 protein-coding genes.

To construct reliable negative examples for the training dataset, we referred to the published microarray data performed on wild type and *Miwi* mutant mouse testis in two replicates (Reuter *et al.*, 2011). Considering that the expression level of genes that were not targeted by piRNAs should not be influenced when the slicer activity of Miwi is disrupted, we first extracted genes that showed little change in expression between wild type and the *Miwi* mutant (data from GEO: GSE32180). To equalize the amount of positive and negative examples, 191 were randomly selected from the resulting 3071 protein-coding genes that had an expression level not affected by the *Miwi* mutation. Each of these genes contributes one site by random selection, generating 191 sites as negative examples for the training dataset (See Supplementary Fig. S1 for details). Thus, the final size of the dataset was 382 sites from 358 specific protein-coding genes.

2.2 Antimiwi CLIP-seq processing

Anti-Miwi CLIP-seq was performed as described previously (Xue *et al.*, 2013) and deposited in the GEO database under the accession number GSE67683. The sequenced reads were first mapped to the mouse genome (mm9) using Bowtie (Langmead*et al.*, 2009). The reads that were successfully mapped were further classified into piRNAs and Miwi targets according to their length (Fig. 1). Reads with a length of 25–33 nt were classified as piRNAs, whereas longer reads (36 nt) were classified as Miwi targets. Reads with a shorter length were discarded because it was difficult to determine to which class they belonged. Of the ~48 million reads that were mapped to the mouse genome, a total of ~17 million reads corresponded to

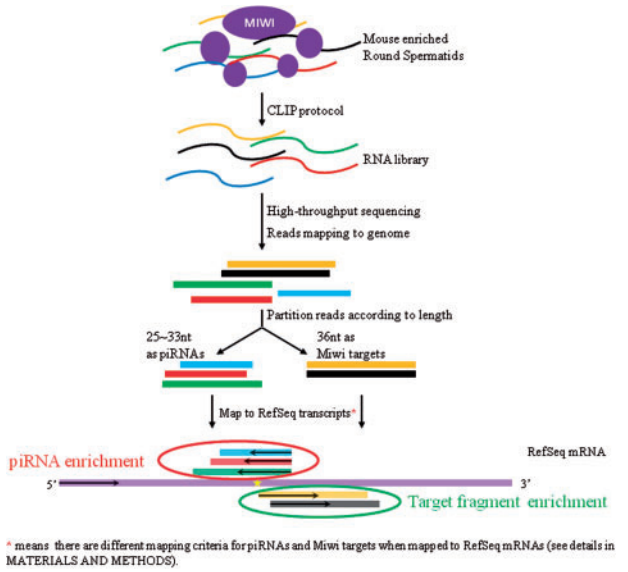

Fig. 1. Pipeline for extracting MIWI CLIP-Seq features (see Section 2). The RNA library was obtained following the CLIP protocol and subjected to high-throughput sequencing. The reads that were successfully mapped to the genome were then classified as piRNAs or Miwi targets according to their length. The classified reads were mapped to the mRNAs of RefSeq with Bowtie using different mapping criteria. The yellow star depicts a candidate piRNA cleavage site. In this presented example, the piRNA enrichment of the candidate site is counted as three, whereas the Miwi target enrichment is counted as two

piRNAs and ~10 million reads correspond to Miwi targets. Miwi targets with more than one genomic location were excluded from further consideration.

2.3 Support vector machine

We used an SVM (Boser, 1992) to build a classifier discriminating the target sites of piRNAs on mRNAs. SVM mapped the sample vectors into a high-dimensional feature space in which the samples may be separated by an optimal boundary through the kernel transformation. In our study, a radial basis function (RBF) kernel was used:

$$k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

where the parameter γ determines the similarity level of the features. In practice, a separating hyperplane may not exist when a problem is very noisy or complex. Thus slack variables $\xi_i \geq 0$ for all $i = 1 \ldots n$ are introduced to loosen the constraints as follows (Bennett and Mangasarian, 1992):

$$yi(\langle w, x_i \rangle + b) \geq 1 - \xi_i \quad \text{for all} \quad i = 1, \ldots, n.$$

A classifier that generalizes well is then obtained by adjusting both the classifier capacity $\|w\|$ and the sum of the slacks $\sum_i \xi_i$, which can be realized by minimizing the following objective function:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

subject to the constraints on ξi and (2), where the constant $C > 0$ determines the trade-off between margin maximization and training error minimization. Consequently, two parameters, γ and C should be determined in the classification model.

2.4 SVM features

The SVM features that were used for classifier training were categorized into two groups: anti-Miwi CLIP-Seq-derived features and position-derived features. An analysis of anti-Miwi CLIP-Seq resulted in two features for a site to be predicted: the enrichment of 5′ ends of distinct piRNAs at position 10 nt downstream of the site and the enrichment of target fragments that were detected in the Miwi complex mapping to a region ranging from 150 nt upstream to 150 nt downstream of the site (Fig. 1). To generate the piRNA enrichment, the full-length mRNA sequences of the RefSeq database (Pruitt *et al.*, 2012) were scanned by Bowtie to search for sites with a possible piRNA interaction. Complementary base-pairing focused on the first 21 nucleotides from the 5′ end of piRNAs by an imperfect pattern. No more than three mismatches were allowed. At the same time, at least one perfect match for the first three nucleotides from the 5′ end of piRNAs was required. piRNA enrichment for a given site is calculated as the number of distinct piRNAs with 5′ ends that are located exactly 10 nt downstream of the target site based on base-pairing complementarity to the flanking sequence. piRNAs interacting with more than 100 mRNAs were eliminated because their targeting might result from the complementary base-pairing of simple sequence repeats that were difficult to distinguish from random hits. Additionally, the genomic locations of Miwi targets were compared to those of mRNAs by BEDTools (Quinlan, 2014). The enrichment of Miwi targets for a specific site was counted as the number of Miwi target reads overlapping the neighboring region of the site with 150 nt extended upstream and downstream. An empirical value of 155 was set as the upper limit for the enrichment of Miwi targets. Both of the anti-Miwi CLIP-Seq-derived features had quantitative values.

For position-derived features, we incorporated information from three aspects. First, the position of the candidate target site was compared to the architecture of the mRNA to which it belongs to determine whether it was located in the 5′UTR, CDS or 3′UTR. Second, 10 neighboring nucleotides upstream and downstream of the target site were extracted, generating a sequence of 20 nucleotides with the target site exactly in the middle. The sequence was then scanned to determine whether it consisted of short repeats (k-mer tandem repeat in which $k = 2$, 3 or 4). Third, the nucleotide selection (the use of A, U, C and G) at the 20 nucleotide positions represented the remaining features. All of the position-derived features had qualitative values (0 or 1). In total, we obtained a feature space of 86 dimensions, and all of the feature values were normalized to have real values with a mean of 0 and a standard deviation of 1.

2.5 Parameter optimization and classifier evaluation

Before applying a trained SVM classifier for large-scale prediction, the performance of our method should be evaluated. We first divided our dataset by random selection into two groups: 70% in the first group for SVM classifier training and the remaining 30% in the second group for independent validation. Data for validation should not participate in any of the steps of classifier training. Before training the SVM classifier on the first group, we should determine in advance two SVM parameters, γ and C, by 10-fold cross validation. The first group is further divided into 10 subgroups of equal size by random sampling. At each sampling, nine subgroups were used to train an SVM classifier given a specific pair of values for γ and C, while the remaining subgroup was used to test the performance of the trained classifier. The average prediction accuracy of the 10-fold cross validation was recorded to represent the performance of the specific pair of values for γ and C. After all of the combinations of values for γ and C were tested, the best pair of values was selected according to the best prediction accuracy. The SVM classifiers were trained on the whole of the first group with the selected parameters and applied to the second group. The sensitivity and specificity of prediction were calculated by setting different thresholds of the SVM score, thus generating an ROC curve that had an area that represented the performance of the SVM classifier. A larger area under the curve indicated better performance. To avoid possible bias in one categorization, the assessment described above was repeated 10 times and the ROC curve was plotted with the specificity and sensitivity averaged from the result of 10 repeated evaluations. Because a genome-wide prediction would be conducted across all of the sites of the full-length mRNA transcripts, evaluation at the gene level is more important. The evaluation at the gene level was different from previous evaluations in that a gene was considered a piRNA target as long as it contained at least one site that was predicted as 'positive' after all of its sites are scanned by the classifier.

2.6 Functional enrichment of piRNA target genes

A functional enrichment analysis was performed for predicted piRNA target genes with the GO Enrichment Analysis Software Toolkit (GOEAST) (Zheng and Wang, 2008). The resulting GO terms with a P-value<0.05 and log-odds ratio (LR) > 0.7 were considered statistically significant.

2.7 Public microarray data analysis

As described previously (Zhang *et al.*, 2015), the raw data (CEL file format) for samples of round spermatids (GSE32180) (Reuter *et al.*, 2011) and elongating spermatids (GSE59291) (Reuter*et al.*, 2011) from adult mouse testes were downloaded and load into R (Team,

2011). Signal intensities were normalized and log2 transformed by the affy package and rma package. For comparison of gene expression levels between the two different platforms, the expression levels were further normalized by the limma package.

3 Results

3.1 Performance of the SVM classifier

The performance of the SVM classifier was evaluated at two levels, the site level and the gene level, following the steps that were described in the Methods section. At the site level, a test applying a set of different threshold values to the SVM scores resulted in an area under the ROC curve of 87.25% (Fig. 2a), with a specificity of 89.48% and a sensitivity of 63.91% corresponding to a default threshold of zero. Since the trained SVM classifier would be applied to scan all of the sites across the full-length of the mRNA sequence of a candidate target gene to determine whether it is a target of piRNAs, the performance of the SVM classifier was reevaluated at the gene level. After all of the sites were scanned, an mRNA was determined as a potential target of piRNAs if it contained at least one site that was classified as 'positive'. The reevaluation gave an area under the ROC curve of 75.55% at the gene level (Fig. 2b), whereas the default threshold of zero resulted in a specificity of 82.07% and a sensitivity of 62.75%.

3.2 Identification of piRNA targets on mRNAs

A final SVM classifier was built based on the whole training dataset with the γ and C parameters optimized by a grid search upon 10-fold cross validation. The classifier was then applied to a total of 28 050 mRNAs from the mouse (the 358 genes used as training examples as well as genes with multiple genomic loci were excluded in advance) with a default threshold of zero. This generated a list of 3781 mRNAs from 2587 genes which contained 12 233 potential target sites which may be cleaved by piRNAs. On average, each target mRNA harbored more than three target sites. The distribution of number of predicted target sites harbored by each mRNA is shown in Figure 3a. It was also observed that piRNA target sites were enriched near the 3' ends of the mRNAs (Fig. 3b). Further analysis of flanking bases revealed a strong bias for U at the mRNA cleavage site and for A 10 nt downstream of the U residue (Fig. 3c).

Khurana and Theurkauf (2008) reported on genes that had mRNA expression levels that showed significant changes after truncation of the slicer activity of Miwi. These data were used to verify our predictions, considering that piRNA targeting genes should have elevated expression levels when the piRNA-guided cleavage mechanism was obstructed. The expression difference of the 2587 predicted piRNA target genes between the *Miwi* mutant and wild type has a positive trend compared with other genes (Fig. 4a), suggesting that the prediction result captures a substantial number of actual piRNA target genes.

Moreover, we found a significant decrease in the expression level of 2587 predicted piRNA target genes in elongating spermatids (GSE59291) compared with those in round spermatids (GSE32180) based on global expression profiling analysis (Fig. 4b). This result might be explained by the hypothesis that piRNAs ensure germ cell development by suppressing specific genes through a cleavage mechanism.

The predicted piRNA target genes were then subjected to pathway enrichment analysis (Fig. 5). Nearly one-quarter of the target genes were associated with metal ion binding. In addition, a significant proportion of the total predicted target genes were enriched in the regulation of nitrogen compound metabolic process and biosynthetic process. Although piRNAs had been considered to have specific expression in germ cells, this result suggests that piRNAs might have largely unexplored functions.

3.3 Contribution of each feature

To investigate which features were most important in piRNA target prediction, we ranked all features by data mining. The OneR classifier and Ranker methods of the Weka software were used to evaluate the features, and the 20 most important features are shown in Table 1. Miwi CLIP-Seq-derived features occupied the first two positions in the rank. Of them, piRNA enrichment was the most important. mRNA fragment enrichment, as inferred from Miwi CLIP-Seq, was the second most important, further indicating that Miwi is necessary in the piRNA-guided cleavage mechanism. The third ranked feature concerns about genomic annotation, with 3'UTR being preferred. Next are features regarding the nucleotide composition. Of all the positions along the sequence of 20 nucleotides, the preference of uracil (U) at the first position and adenine

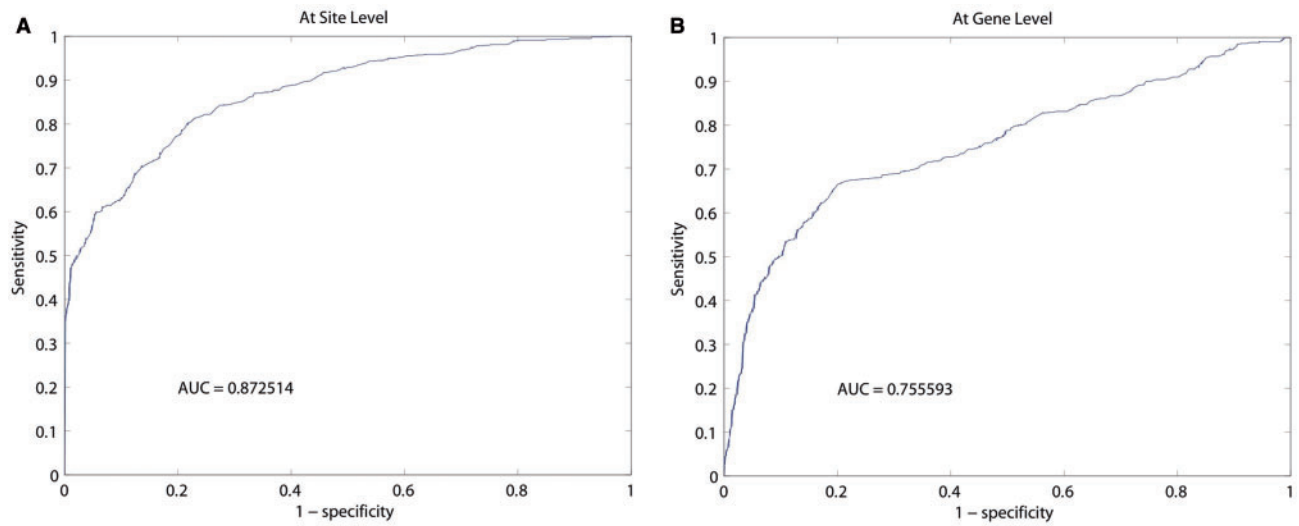

Fig. 2. Performance of the SVM classifier as evaluated with 10-fold cross-validation at the site level (**A**) and the gene level (**B**)

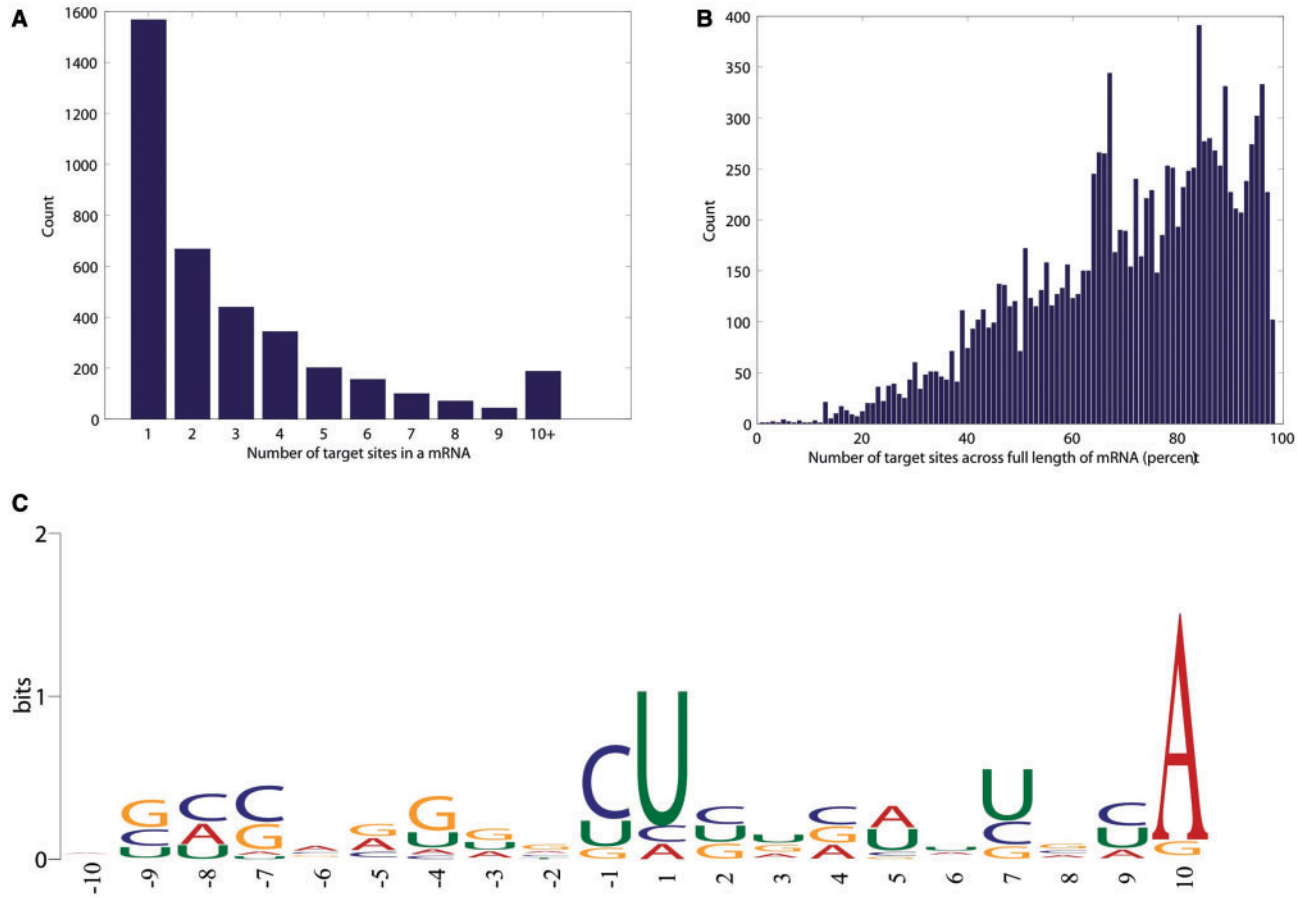

Fig. 3. Predicted target sites of piRNAs on mRNAs. (**A**) The distribution of the number of target sites that harbored by an mRNA. The majority of mRNAs that were predicted to be targeted by piRNAs contain only one target site. (**B**) Enrichment of target sites across the full-length mRNA sequences. The length of the mRNA sequences is scaled to be 100. (**C**) The motif that was generated from the flanking sequences of the predicted piRNA target sites. The logo was generated by the MEME suite using sequences neighboring 12 233 target sites with 20 nucleotides. The number marked below represents the distance to the target site, '-' representing upstream

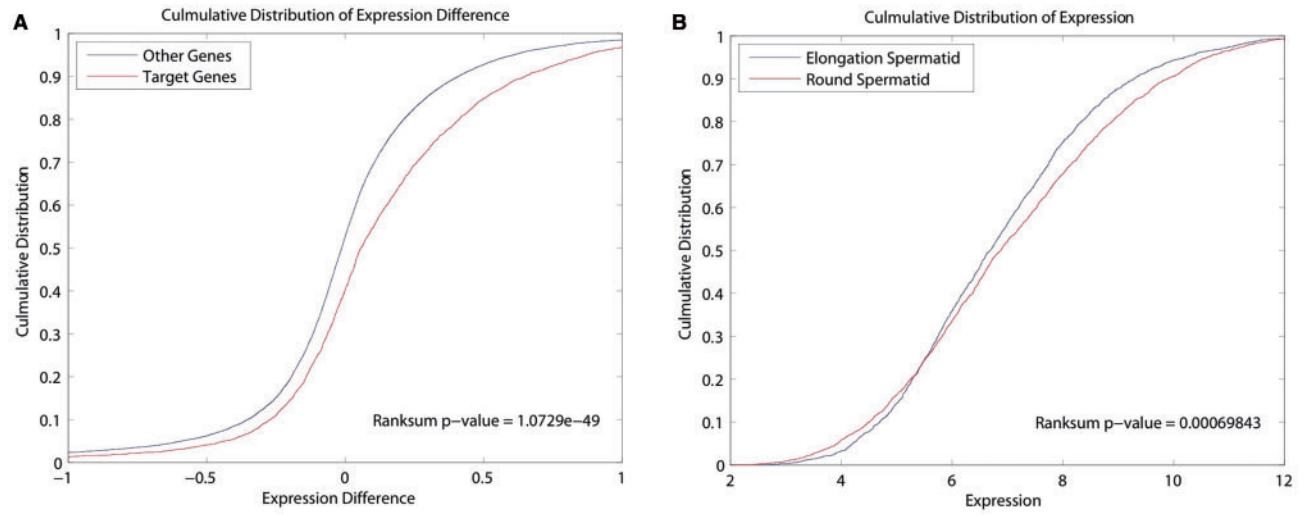

Fig. 4. Validation of the predicted piRNA target genes by microarray data. (**A**) Accumulative distribution of expression difference of the predicted target genes (red line) and other genes (blue line) between the catalytic mutant (*Miwi*$^{-/ADH}$) and control (*Miwi*$^{+/-}$) samples (GSE32180). The expression of individual genes is log2 transformed and duplicate averaged. The gene expression difference is presented as the expression in the catalytic mutants minus that in the controls. The *P* value was calculated by the Wilcoxon rank sum test. (**B**) Accumulative distribution of the expression of predicted target genes in round spermatids (GSE32180; red line) and elongating spermatids (GSE59291; blue line). The expression of individual genes is log2 transformed and duplicate averaged

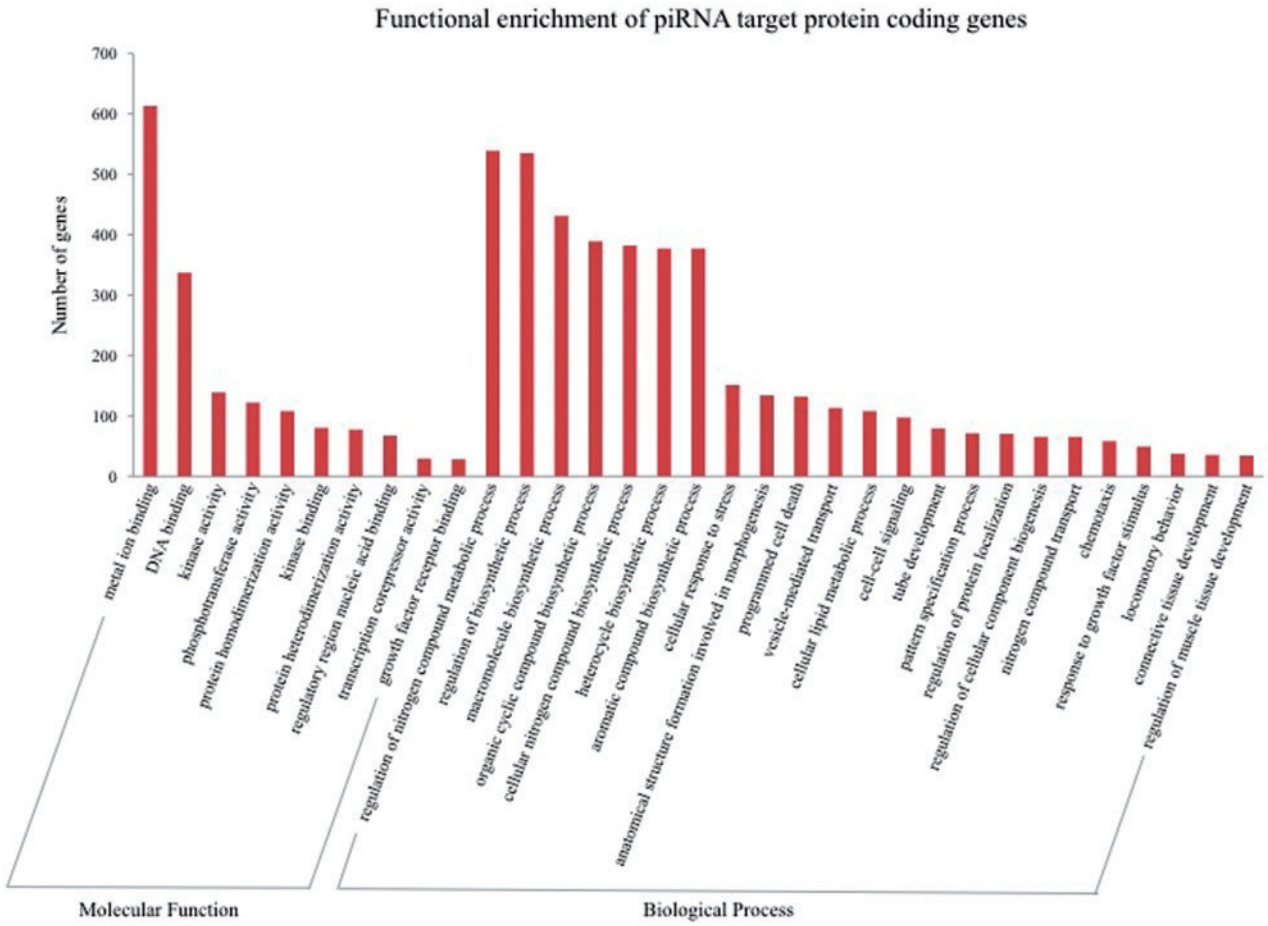

Fig. 5. Functional enrichment result for the predicted target genes by GOEAST. The resulted GO terms of level 4 with a $P < 0.05$ and a log-odds ratio > 0.7 were considered statistically significant and are depicted

Table 1 The top 20 contributing features

Rank	Rank score	Feature
1	82.722	Enrichment of piRNAs
2	71.204	Enrichment of MIWI targets
3	70.681	3′ UTR
4	67.801	U on 1st position
5	67.539	A on 10th position
6	66.492	CDS
7	61.78	U on 6th position
8	60.995	A on 5th position
9	60.471	G on 1st position
10	60.209	U on -10th position
11	60.209	G on 9th position
12	59.948	U on 9th position
13	58.9	G on -1st position
14	58.639	U on -4th position
15	58.639	C on 5th position
16	58.639	G on 5th position
17	58.115	C on -1st position
18	58.115	G on 6th position
19	58.115	U on 7th position
20	57.592	C on 10th position

(A) at the tenth position downstream of the target site are the most outstanding, showing complementarity with the preferred usage of U and A at the first and tenth positions from the 5' end of piRNAs.

3.4 Characteristics of mRNA-targeting piRNAs

Of all of the piRNAs that were identified by the anti-Miwi CLIP-Seq experiment, 16 657 piRNA reads representing 3385 different piRNA species contributed to cleavage of the predicted targets on mRNAs. The length profile of these piRNA species showed a stronger peak at 30 nt than that of the total piRNA species from the anti-Miwi CLIP-Seq (Fig. 6a). An unexpected large percentage of these mRNA-targeting piRNA species (~88.54%) were derived from repetitive genomic sequences, whereas the corresponding percentage of repeat-derived piRNAs in the total piRNA population from the anti-Miwi CLIP-Seq data was less than 30% (Fig. 6b). Among the different repeat classes, the SINE family constituted the highest proportion (~66.44%), with Alu-like elements representing largest amount within the SINEs (~42.48%). A *de novo* motif analysis was performed by the MEME Suite (Bailey *et al.*, 2009) and discovered a significant consensus sequence residing in these mRNA-targeting piRNAs (Fig. 6c). More than 700 mRNA-targeting piRNA species harbored the identified motif starting from the first nucleotide at the 5′ end. The bias towards U at the 5′ end and A at the tenth position downstream indicated the ping-pong interaction. Nontheless, this motif only represented ~23% of all of the mRNA-targeting piRNA species, indicating that this motif alone would be inadequate to distinguish mRNA-regulating piRNAs.

4 Discussion

Using the piRNA targets that we previously identified (Zhang *et al.*, 2015) as positive examples of our training dataset, we made

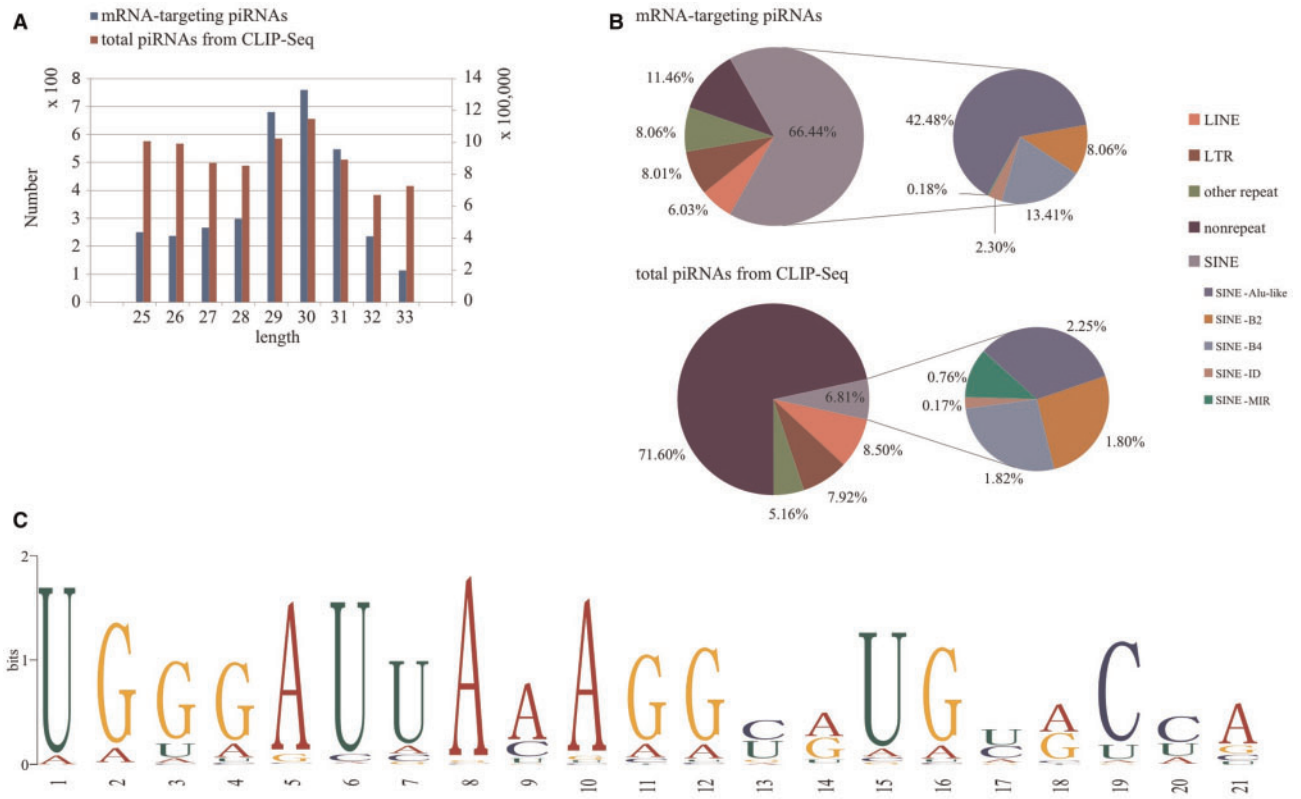

Fig. 6. Characteristics of mRNA-targeting piRNAs. (**A**) Length profiles of mRNA-targeting piRNA species and total piRNA species from CLIP-Seq. (**B**) Diagram depicting the genomic distribution of the mRNA-targeting piRNA species and the total piRNA species from CLIP-Seq relative to repetitive sequences. (**C**) Enriched motif that was generated from mRNA-targeting piRNA species by MEME

genome-wide de novo predictions and provided a much larger list of mRNAs that might be targeted and cleaved by piRNAs. In this study, we took advantage of CLIP-Seq data, which are specific to fragments that are bound by the Miwi protein, whereas the previous study was based on 5′ RACE sequencing data, which contains not only fragments that are produced by piRNA-mediated cleavage but also other biological processes. Moreover, the previous study was restricted to the coverage of 5′ RACE sequencing data on mRNAs, whereas the depth of the CLIP-Seq data that were used in this study makes it possible to systematic explore piRNA targets on mRNAs. In addition, the incorporation of position-derived features enabled the discovery of piRNA target sites that might be missed by the absence of complementary piRNAs from sequenced reads. Our method presents robustness as repeated random selection of negative examples for training the model did not resulted in much difference with regarding to the predicted list of piRNA target genes. It could also be applied to other tissue types which have both training examples and anti-Miwi CLIP-Seq data.

In addition to the repression of mRNAs by cleavage in this study, another repression mechanism is mediated by piRNAs that induces mRNA deadenylation, resulting in mRNA degradation. For example, piRNAs target the nanos 3′ UTR with imperfect complementarity, thus facilitating nanos mRNA deadenylation and normal anteroposterior embryonic patterning in *Drosophila* (Rouget *et al.*, 2010). Recently, Gou *et al.*(2014) reported that more than 7000 mRNAs were associated with Miwi, of which approximately 60% experienced destabilization by Miwi in the elongating spermatids of mouse. Together with mRNAs that are involved in piRNA-mediated cleavage, a catalog of nearly ten thousand protein-coding genes are potentially mediated by piRNAs, suggesting a prosperous role of piRNAs in gene expression regulation. Conclusively, piRNAs have mechanisms that act to repress protein-coding gene expression similar to miRNAs, either causing the cleavage of the target or the degradation of the target. Both the miRNA pathway and the piRNA pathway require recognition based on base-pairing and the participation of proteins from the Argonaute protein family. However, because piRNAs are restricted to expression in germ cells, piRNA-mediated regulation might take place primarily during spermatogenesis.

5 Conclusion

It has been well accepted for a long period of time that the emergence of piRNAs plays a role in genome defense by silencing TEs; however, many recent studies have shown that piRNA-guided silencing mechanisms are also involved in the regulation of non-transposon transcripts. The first example of a piRNA targeting mRNA was described in the *Drosophila* embryo (Rouget *et al.*, 2010). A similar phenomenon was also observed in *C. elegans*, mouse and silkworm (Kiuchi *et al.*, 2014), suggesting that piRNAs could regulate gene expression via a mechanism similar to that of miRNAs. Although there have been many published programs aimed at predicting miRNA targets as well as public databases curating experimentally verified or computational predicted targets of miRNAs, there have been no studies assessing how many protein-coding genes might be regulated by piRNAs. In our study, piRNA targets on mRNAs were systematically identified at the genome-wide level using an SVM classifier. The classifier was trained based on the incorporation of Miwi CLIP-Seq-derived features and position-derived features and achieved an area of the ROC curve of

87.25% at the site level and 75.55% at the gene level. Of all of the features that were used, Miwi CLIP-Seq-derived features play a dominate role in the piRNA enrichment of the 5′ end 10 nt downstream of the target site occupying the top ranked position. Of the position-derived features, the genomic annotation is highly ranked, with 3′UTR as a preferred choice, and nucleotide usage on the first and tenth positions downstream of the target site is the next highest ranked. Genome-wide prediction based on the trained classifier resulted in 3781 mRNAs of 2587 protein-coding genes as piRNA targets. These genes show significant upregulation as a whole after the slicer activity of Miwi was abolished, suggesting that the predicted list of piRNA targets is reliable. This work might provide valuable clues for biologists who are interested in the function of piRNAs in protein-coding genes. More details of the prediction results in this study and the online version of the method called pirnaPre are available at http://www.regulatoryrna.org/software/piRNA/piRNA_target_mRNA/index.php.

Acknowledgements

We are grateful to Xiuqin Liu for thoughtful discussions and valuable comments on the manuscript. We also thank Hanchen Huang for discussions about SVM performance evaluation.

Funding

Ministry of Science and Technology of China (2014AA021103, 2012AA020402 and 2011CB504605); Chinese Academy of Science Strategic Project of Leading Science and Technology (XDA01020402); HPC Platform, Scientific Information Center, Institute of Zoology, CAS.

Conflict of Interest: none declared.

References

Aravin,A.A. *et al.* (2007) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science*, **318**, 761–764.

Aravin,A.A. *et al.* (2001) Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. *Curr. Biol.*, **11**, 1017–1027.

Bailey,T.L. *et al.* (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.

Boser,E. *et al.* (1992) A training algorithm for optimal margin classifiers. In: *Proceedings of the fifth annual workshop on Computational learning theory*, 144–152.

Brower-Toland,B. *et al.* (2007) Drosophila PIWI associates with chromatin and interacts directly with HP1a. *Genes Dev.*, **21**, 2300–2311.

Dannemann,M. *et al.* (2012) Transcription factors are targeted by differentially expressed miRNAs in primates. *Genome Biol. Evol.*, **4**, 552–564.

Gou,L.T. *et al.* (2014) Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. *Cell Res.*, **24**, 680–700.

Han,B.W. *et al.* (2015) Noncoding RNA. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. *Science*, **348**, 817–821.

Houwing,S. *et al.* (2007) A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell*, **129**, 69–82.

Ishizu,H. *et al.* (2012) Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines. *Genes Dev.*, **26**, 2361–2373.

Kelleher,E.S. and Barbash,D.A. (2013) Analysis of piRNA-mediated silencing of active TEs in Drosophila melanogaster suggests limits on the evolution of host genome defense. *Mol. Biol. Evol.*, **30**, 1816–1829.

Khurana,J.S. and Theurkauf,W.E. (2008) *piRNA function in germline development*. In: *StemBook*. Cambridge.

Kim,V.N. *et al.* (2009) Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.*, **10**, 126–139.

Kiuchi,T. *et al.* (2014) A single female-specific piRNA is the primary determiner of sex in the silkworm. *Nature*, **509**, 633–636.

Bennett,K.P. and Mangasarian,O.L. (1992) Robust linear programming discrimination of two linearly inseparable sets. *Optim. Methods Softw.*, **1**, 23–34.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Le Thomas,A. *et al.* (2013) Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes Dev.*, **27**, 390–399.

Le Thomas,A. *et al.* (2014) To be or not to be a piRNA: genomic origin and processing of piRNAs. *Genome Biol.*, **15**, 204.

Lee,H.C. *et al.* (2012) *C. elegans* piRNAs mediate the genome-wide surveillance of germline transcripts. *Cell*, **150**, 78–87.

Mohn,F. *et al.* (2015) Noncoding RNA. piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. *Science*, **348**, 812–817.

Nishida,K.M. *et al.* (2007) Gene silencing mechanisms mediated by Aubergine piRNA complexes in Drosophila male gonad. *RNA*, **13**, 1911–1922.

Pruitt,K.D. *et al.* (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.

Quinlan,A.R. (2014) BEDTools: The Swiss-Army tool for genome feature analysis. *Curr. Protoc. Bioinformatics*, **47**, 11 12 11–11 12 34.

Reuter,M. *et al.* (2011) Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing,. *Nature*, **480**, 264–267.

Ross,R.J. *et al.* (2014) PIWI proteins and PIWI-interacting RNAs in the soma. *Nature*, **505**, 353–359.

Rouget,C. *et al.* (2010) Maternal mRNA deadenylation and decay by the piRNA pathway in the early Drosophila embryo. *Nature*, **467**, 1128–1132.

Sai Lakshmi,S. and Agrawal,S. (2008) piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res.*, **36**, D173–D177.

Saito,K. *et al.* (2009) A regulatory circuit for piwi by the large Maf gene traffic jam in Drosophila. *Nature*, **461**, 1296–1299.

Saxe,J.P. and Lin,H. (2011) Small noncoding RNAs in the germline. *Cold Spring Harb. Perspect. Biol.*, **3**, a002717

Siomi,H. and Siomi,M.C. (2015) RNA. Phased piRNAs tackle transposons. *Science*, **348**, 756–757.

Siomi,M.C. *et al.* (2011) PIWI-interacting small RNAs: the vanguard of genome defence. *Nat. Rev. Mol. Cell Biol.*, **12**, 246–258.

Team,R.D.C. (2011) *R: A language and environment for statistical computing*. The R Foundation for Statistical Computing.

Theron,E. *et al.* (2014) Distinct features of the piRNA pathway in somatic and germ cells: from piRNA cluster transcription to piRNA processing and amplification. *Mob. DNA*, **5**, 28.

Weick,E.M. and Miska,E.A. (2014) piRNAs: from biogenesis to function. *Development*, **141**, 3458–3471.

Xue,Y. *et al.* (2013) Direct conversion of fibroblasts to neurons by reprogramming PTB-regulated microRNA circuits. *Cell*, **152**, 82–96.

Zamudio,N. and Bourc'his,D. (2010) Transposable elements in the mammalian germline: a comfortable niche or a deadly trap? *Heredity (Edinb)*, **105**, 92–104.

Zhang,P. *et al.* (2015) MIWI and piRNA-mediated cleavage of messenger RNAs in mouse testes. *Cell Res.*, **25**, 193–207

Zheng,Q. and Wang,X.J. (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res.*, **36**, W358–W363.