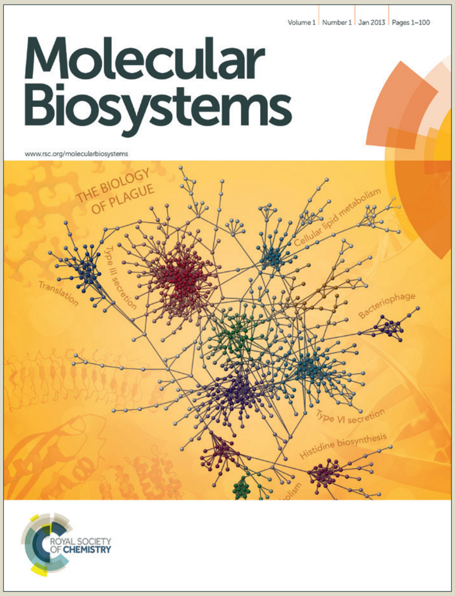

Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Molecular Biosystems
www.rsc.org/molecularbiosystems
THE BIOLOGY OF PLAGUE
Cellular lipid metabolism
Type III secretion
Translation
Bacteriophage
Type VI secretion
Histidine biosynthesis
ROYAL SOCIETY OF CHEMISTRY

piRNA identification based on motif discovery†‡

Xiuqin Liu,*[a,b] Jun Ding[a] and Fuzhou Gong*[b,c]

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX

DOI: 10.1039/b000000x

[a] School of Mathematics and Physics, University of Science and Technology Beijing, Beijing 100083, China;
[b] National Center for Mathematics and Interdisciplinary Sciences,Chinese Academy of Sciences, Beijing 100190, China;
[c] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.
*E-mail: mathlxq@163.com, fzgong@amt.ac.cn

† Electronic Supplementary Information (ESI) available: See DOI: 10.1039/b000000x/

‡ Author contributions: XQL conceived and developed the prediction method, conceived, designed and implemented the experiments, analyzed the result, and wrote the paper. JD developed the webserver. FZG analyzed the result and provided guidance and supervision.

Piwi-interacting RNA (piRNA) is a class of small non-coding RNAs about 24 to 32 nucleotides long associated with PIWI proteins, which are involved in germline development, transposon silencing, and epigenetic regulation. Identifying piRNA loci on the genome is very useful for further studies in the biogenesis and function of piRNAs. To accomplish this, we applied the computational biology tool Teiresias to identify motifs of variable length appearing frequently in mouse piRNA and non-piRNA sequences, respectively, and then proposed an algorithm for **pi**RNA **i**dentification **b**ased **o**n **m**otif **d**iscovery, termed "Pibomd" by using these sequence motifs as features in the Support Vector Machine (SVM) algorithm, a sensitivity of 91.48% and a specificity of 89.76% on a mouse test dataset could be achieved, much better results than those reported in previously published algorithms. We also trained an unbalanced SVM classifier (named as "Asym-Pibomd") that got a higher specificity (96.2%) and a lower Sensitivity (72.68%) than the Pibomd. The predicted ACC is less than that of the Pibomd, in spite of this, the predicted ACC (84.44%) of Asym-Pibomd is about ten percent more than that of k-mer method. Further analysis of the motif positions on the piRNA sequences showed that the piRNA sequences may contain information on the 5' and/or 3' end recognized by the piRNA processing apparatus of actual piRNA precursors. Furthermore, this prediction method can be found on a user-friendly web server found at http://app.aporc.org/Pibomd/.

Introduction

Piwi-interacting RNA (piRNA) is a specific class of RNAs having a length of about 24-32nt and interacting with a subset of Argonaute proteins, catalytic components of the RNA-induced silencing complex RISC (RNA-inducing silencing complex), related to Piwi[1-4]. Many studies have shown that piRNAs have a specific germ line function in repressing transposable elements. The repression is thought to involve heterochromatin formation and both transcriptional and post-transcriptional silencing[2, 5-9]. Mammalian piRNAs can be divided into pre-pachytene (26–28nt) piRNAs derived from retrotransposons and required for their silencing and pachytene (29–31nt) piRNAs whose function remains unknown[10]. piRNA biogenesis remains to be elucidated; however, two popularly accepted mechanisms have been reported. One is termed as the "Ping-Pong" Model[11, 12], based on the formation of a feed-forward amplification loop that "facilitate piRNA surveillance of transposon transcription in the germ line[10]." Another holds that piRNAs are derived from genomic regions depleted in transposons, generated by a pathway called ''primary processing'' present in somatic cells, and that they may have a role in the regulation of target mRNAs[13].

To detect piRNAs, a previous approach combined immunoprecipitation and deep sequencing in model and sequenced organisms[7]. However, this experimental model has some shortcomings. For example, this method may miss low-expressing or tissue-specific piRNAs. Thus, computational methods would be a good complementary technique to search for piRNAs. Such method could summarize the general properties from known piRNAs and then train them to predict novel ones[14].

Some computational methods have already been developed to detect piRNAs[14, 15]. Betel et al.[15] trained a support vector machine (SVM) classifier to discriminate between 5'RNA and all other uridine positions. Feature vectors were constructed by converting the 21-base sequences into 84-bit vectors such that each nucleotide position was converted to a 4-bit vector representing the RNA base. SVM training and classification was then performed, resulting in the identification of mouse piRNAs with a precision of 61–72%. However, their method could not efficiently detect those piRNAs derived from the 3'UTR of mRNA, which are not produced by forward loop amplification, as suggested in the Ping-Pong Model [16, 17]. To predict piRNAs, Zhang et al.[14]

implemented a k-mer algorithm which does not require a reference genome and gives much better performance. They used all the 1364 1-5nt strings and an improved Fisher algorithm by setting different cutoffs and elevating the precision rate to characterize the piRNA sequence in five model species: rat, mouse, human, fruit fly and nematode. The k-mer method performed much better with a precision of over 90% and a sensitivity of over 60%.

Previously published methods have used the sequence feature of fixed size. Liu et al. [18] carried out exhaustive searches for all possible sequence-structure motifs (ss-motifs) on potential RNA hairpin structures. A machine learning algorithm was implemented to successfully predict precursor miRNAs based on the identified string of variable-length sequence-structure motifs. This method was able to predict miRNA loci in the human genome with higher specificity and sensitivity than other previously published computational tools[18]. The present study was based on a similar idea, i.e., extracting sequence information of variable length, to better capture the real features of piRNA and other RNA sequences. In order to test this hypothesis, we proposed an SVM algorithm for **p**iRNA **i**dentification **b**ased on **m**otif **d**iscovery, termed "Pibomd", which was able to predict piRNAs in mouse with high specificity and sensitivity when only using 258 motifs appearing frequently.

Materials and Methods

Data

We downloaded the data from http://59.79.168.90/piRNA/download.php [14]. The positive piRNA data consisted of 173,090 sequences of 5 species (rat, mouse, human, fruit fly and nematode) from NONCODE version 2.0[19] and NCBI. The non-piRNA sequences were also derived from NONCODE version 2.0[19] and consisted of 193,321 sequences of various types, such as 'miRNA', 'piRNA', 'mlRNA', 'snoRNA', 'snRNA', 'tmRNA', 'SRP RNA', 'gRNA' , 'sbRNA', and 'snlRNA' [14]. The piRNA dataset consisted of 72,747 mouse piRNA sequences and 100,343 non-mouse piRNA sequences, with171,703 mouse sequences and 21,618 non-mouse sequences comprising the 193,321 sequences of the negative dataset.

Anastassios Vourekas et al.[20] used high-throughput sequencing after cross-linking and immunoprecipitation (HITS-CLIP) coupled with RNA-sequencing (RNA-seq) to characterize the genome-wide target RNA repertoire of Mili (Piwil2) and Miwi (Piwil1). Therefore, we downloaded Miwi and Mili (GEO accession: GSM684624 and GSM684620, respectively) mapped reads and then took out sequences in the range of 23–31nt for Mili and 25–33nt for Miwi mapped reads, respectively, according to their designation. After removing the duplicated sequences, 649052 Miwi-bound piRNA and 166988 Mili-bound piRNA sequences were left.

The training dataset was obtained by randomly selecting 5000 sequences from mouse piRNA and 5000 sequences from mouse non-piRNA sequences. Three types of test sets were used. The first type included 2,500 mouse piRNA sequences and 2,500 mouse non-piRNA sequences randomly selected from the above mouse piRNA and non-piRNA sets after removal of the sequences used for the training sets. The second type included 100343 non-mouse piRNA sequences from the above positive piRNA data. The third type included 649052 Miwi-bound piRNA and 166988 Mili-bound piRNA sequences.

Motif extraction

The computational biology tool Teiresias [21] derived from IBM was used to discover the frequently occurring motifs. With the help of Teiresias, we were able to search for motifs of variable length in the two groups of training sequences, respectively. The options used here were ''Exact discovery'', ''Seq Version'' and "Only nucleic acid characters". The parameters used were L=3, W=5, and K=2,000, implying that any motif of length W = 5 positions containing at least L = 3 defined nucleotides and occurring in at least K=2,000 different sequences will be retained (see [16] for more details).

SVM for training and prediction

Support Vector Machine (SVM) was used to classify piRNAs versus non-piRNAs using the 258 features as input. The famous SVM software package LIBSVM was downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/[22]. First, the 258 features (the 10000 sequences in the training set) were rescaled linearly by the SVM-scale program to the interval [-1, 1] (svm-scale -l -1 -u 1mousetrain.txt>mousetrain.scale) to 1) prevent attributes in higher numeric ranges from dominating those in smaller numeric ranges and 2) avoid other numerical difficulties during the calculation. The test set features (5000 sequences in the test set) were also rescaled using the same scaling factors. The classifier model was trained with ‘svm-train’ $-C\ 2^C\ g\ \gamma$ and the default RBF kernel ($K(x,y)=e^{-\gamma\|x-y\|^2}$). As both the penalty parameter C and the RBF kernel parameter γ are critical for the SVM performance[23], they were optimally selected by an exhaustive grid-search strategy, which was carried out using ‘grid.py’, as described by Ng and Mishra[24]. We performed a 5-fold cross validation. In order to avoid over fitting the generalization, we employed the combined hyper parameters (C,γ), thus maximizing the 5-fold leave-one-out cross-validation accuracy rate as the default setting for training the classifier. The optimal values of the parameters (C,γ) are 32768 and 8, respectively. Finally, classification was conducted on the test set and independent evaluation datasets with ‘svm-predict’ (see [16] for more details).

For the purpose of genome screening, a highly specific detector is desirable. In order to meet that kind of requirement, a biased SVM classifier was also trained. The penalty for class “-1”was set 30 times over that of class“1” in order to improve the predicted specificity as much as possible to reduce the incidence of false-positives. Here we used the “-wi ” options in the libsvm software package to train an unbalanced classifier (svm-train –c 32768 –g 8 -w-1 30 w1 1), and we named it as Asym-Pibomd. Classification was conducted on the test set also.

Evaluating the importance of motifs in prediction performance

F-score is a simple technique which measures the discrimination of two sets of real numbers[25]. Here, we used F-score to evaluate the importance of motifs in prediction performance. Given a training vector $x_k, k = 1, 2, \cdots, n,$ if the number of positive and negative instances are n_+ and n_-, respectively, then the F-score of the i*th* motif is defined as

$$F(i) = \frac{(\overline{x}_i^{(+)} - \overline{x}_i)^2 + (\overline{x}_i^{(-)} - \overline{x}_i)^2}{\frac{1}{n_+ + 1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)} - \overline{x}_i^{(+)}) + \frac{1}{n_- + 1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)} - \overline{x}_i^{(-)})}$$

Where $\overline{x}_i$ $\overline{x}_i^{(+)}$ $\overline{x}_i^{(-)}$ are the average of the i*th* motif of the whole, positive and negative datasets, respectively; $x_{k,i}^{(+)}$ is the i*th* motif of the k*th* positive instance, and $x_{k,i}^{(-)}$ is the i*th* motif of the k*th* negative instance. As the F-score increases, the likelihood that this motif will become more discriminating also increases.

Results and Discussion

The variable-length sequence motif

The functionality of an RNA molecule is primarily determined by its nucleotide sequence. Accordingly, its molecular information can be represented by a string of letters (e.g., UCACAUAACUCCGGCCCUAC), denoting, in this case, the chemical composition of an RNA molecule that is 20 nucleotides long. We only extracted sequence information for piRNAs since they may lack conserved secondary structure motifs[26]. From these piRNA sequences, we extracted frequently occurring motifs of variable length (see Materials and Methods) which were subsequently used to distinguish piRNAs from non-piRNA sequences.

Motif extraction

An SVM-based classifier was developed to test the efficacy of the motifs in classifying piRNA sequences from non-piRNA sequences. The Teiresias tool was used to discover frequently occurring motifs in 5000 mouse non-piRNA sequences and 5000 mouse piRNA sequences, and 254 and 73 motifs were extracted, respectively. After the elimination of redundancy motifs, 258 motifs remained (Table S1). We then developed a program to compute the frequency of each motif in the training sequences: a 258*10,000 feature matrix and a 258*5000 feature matrix which could be used to construct a classifier and test that it had been generated.

Motifs evaluation

F-score was adopted to analyze the discriminatory power of the motifs (see Materials and Methods for details). Based on its ability to distinguish piRNA sequences from negative sequences, the motifs were subsequently ranked in descending order according to their F-score (Table S1). The fifteen motifs showing greatest discriminatory power are listed in Table 1. We employed 258 motifs (10000 mouse sequences in the training data) to train an SVM classifier (Fig.1). Pibomd reached a prediction accuracy of over 90%, with a sensitivity of 91.48% and a specificity of over 89.76% (Table 2). We also drew an ROC curve for Pibomd (figure not shown). The area under the curve (AUC) was 0.953, which may further imply the potential of sequence motifs in piRNA prediction.

The same training and testing datasets were used to train an unbalanced SVM classifier, termed, as noted above, Asym-Pibomd, which achieves a higher specificity of 96.2% and a lower sensitivity of 72.68% than the balanced SVM classifier Pibomd. However, the predicted ACC is 84.44%, which is less than that of Pibomd (Table 3).

Pibomd identifies novel and non-mouse piRNAs

It is well known that piRNA is not conserved among the different species. Therefore, we asked if the sequence motifs derived from mouse would be representative of piRNAs of most organisms. To address this question, we applied Pibomd to 100343 non-mouse piRNA sequences and found that it could distinguish 86% of them. When it was applied to the new version of the Miwi and Mili IP dataset, including 649052 Miwi-bound piRNAs and 166988 Mili-bound piRNAs, Pibomd identified 82.55% (535797/649052) and 76% (126910/166988), respectively(Table 4).

Comparison with k-mer scheme

The results obtained for Pibomd are comparable to those achieved by previously published computational methods of piRNA prediction, in particular, the k-mer scheme (Table 3). The k-mer schema used all of the 1364 1nt-5nt strings and an improved Fisher Discriminant algorithm to identify piRNA sequences in five model species; the algorithm reached high prediction precision and low sensitivity in those species. In order to make a more effective comparison, we trained a k-mer scheme based on the training dataset used in the present work and tested it on the test set, achieving a very high prediction specificity of 97.16% and very low sensitivity of 48.32% in those species, with accuracy of 72.74%. Afterwards, we trained a k-mer scheme on the training dataset used in Zhang et al.[14] (120000 piRNAs and 120000 non-piRNAs) and tested it on the dataset used in the present work. Again, we achieved a prediction specificity of 98.4% and very low sensitivity 52.04%, with accuracy of 75.22%. Obviously, the k-mer scheme consistently demonstrates high specificity at the cost of sensitivity, and the predicted accuracy

(~75%) is much less than that of Pibomd (90.62%).

As previously noted, we also trained an unbalanced SVM classifier, Asym-Pibomd, at the expense of the predicted ACC (see Material and Methods for details). It achieved higher specificity (96.2%) and lower sensitivity (72.68%) than Pibomd. The predicted ACC is much smaller than that of Pibomd. In spite of this, the predicted ACC (84.44%) of Asym-Pibomd is about ten percent more than that of the k-mer method.

Betel[15] trained an SVM classifier to only discriminate between 5'piRNA and other uridine positions with a precision of 61–72%. However, this method cannot efficiently detect those piRNAs derived from the 3'UTR of mRNA, which are not produced by the 'Ping-Pong Model'.

Pibomd identifies mouse CLIP piRNAs and chromosome cluster piRNAs more efficiently

piRBase is a manually curated resource of piRNAs. It focuses on the functional analyses of piRNAs, as well as piRNA annotation. piRBase collected piRNAs obtained from such methods as small RNA sequencing, protein IP or chromatography, and protein CLIP. The protein CLIP method is a high-quality means of identifying piRNAs. Therefore, we downloaded the clip piRNAs from http://www.regulatoryrna.org/database/piRNA/. Altogether, we found 2092708 Clip piRNA sites in piRBase. The k-mer method could identify 696740 (696740/2092708=33.3%; see Fig.2A), and Asym-Pibomd could identify 966093(966093/2092708=46.16%; see Fig.2A). The specificity of Asym-Pibomd is only two percent lower than that of the k-mer method, but the sensitivity of k-mer is much less (see Table 3) than that of Asym-Pibomd. These results show that Asym-Pibomd has the capacity to identify more piRNA sites, even though false-positives are slightly increased over those in the k-mer method. Pibomd could identify 1614483(1614483/2092708=77.15%; see Fig.2A). Furthermore, most of the piRNA sites identified by the k-mer method were also recognized by Asym-Pibomd

(561830/696740=80.64%) and Pibomd (669709/696740=96.12%).

Most piRNAs are generated from genomic piRNA clusters. A piRNA cluster from 44329185 to 44352221 is found on theNo.1 chromosome. Using piRBase, 399 piRNAs were identified on the No.1 chromosome from 44329185 to 44352221. The k-mer method was able to identify 85 (85/399=21.3%; see Fig.2B) piRNAs, and Asym-Pibomd could identify 131(131/399=33%; see Fig.2B) piRNAs. Furthermore, Pibomd could identify 222(222/399=55.64%; see Fig.2B) piRNAs. Similarly, most of the piRNA sites identified by the k-mer method were also recognized by Asym-Pibomd (561830/696740=80.64%) and Pibomd (669709/696740=96.12%).

Motif positions on the piRNA sequence

In order to determine if the distribution of the individual motif on the piRNA sequences is different from that of the non-piRNAs, we counted the number of motifs appearing on the piRNA and non-piRNA sequences separately and analyzed the distribution. For simplicity, we only analyzed the distribution of the motifs on the piRNAs having a length of 29nt and 30nt, which accounted for about fifty percent of all piRNAs(Fig.3). We plotted the distribution of the motifs along the piRNA sequences. The motifs could be divided into two categories: those headed by a nucleotide “T” and those headed by the remaining three nucleotides (Fig.4A). Considering the entire sequence, the first category of motifs would typically appear on the 5’ end of the piRNA sequences (Fig.4 A). This observation is in accordance with nucleotide preference at the 3’end position of piRNAs, which is enriched for 3’uridine. The second category of motifs seemed to appear with significantly less frequency on the 5’and 3’ ends of the piRNA sequences (Fig.4 B). On the other hand, it has been shown that the motifs were uniformly distributed on the

Table 1. Discriminatory power of the top 15 motifs. The discriminatory power of the motifs that distinguish piRNA sequences from negative RNAs is calculated using the F-score, and the 15 motifs with greatest discriminatory power are listed here (“N” denotes any nucleotide).

No.	1	2	3	4	5
Motifs	ANAA	CNTG	CNTA	CTNT	CNTC
F-score	0.044701	0.042748	0.042698	0.041646	0.041185
No.	6	7	8	9	10
Motifs	CAC	CNTNT	GNCA	ATA	ANTT
F-score	0.040529	0.039217	0.035025	0.028029	0.025241
No.	11	12	13	14	15
Motifs	TGNNT	GNAC	CTT	ANNCT	AGNG
F-score	0.02471	0.023792	0.023612	0.023527	0.023306

Table 2. Definitions of precision and sensitivity of prediction.

	Predicted positive	Predicted negative
Actual positives	True positives (TP) =2287	False negatives (FN)=213
Actual negatives	False positives (FP)=256	True negatives(TN)=2244

Sensitivity (sn)
Specificity(sp)
Accuracy(ACC)
	Sn =2287/2500=91.48%
	Sp =2244/2500=89.76%
	ACC=4531/5000=90.62%	

Table 3.Comparison with k-mer scheme

	Sp(%)	Sn(%)	ACC(%)
k-mer	98.4	52.04	75.22
Pibomd	89.76	91.48	90.62
Asym-Pibomd	96.2	72.68	84.44
Betel's method	/	/	72

Table 4. Pibomd prediction accuracy for independent evaluation datasets.

Test set	Number of piRNAs	Accuracy (%)
Non-mouse piRNAs	100343	86%
Miwi-bound piRNAs	649052	82.55%
Mili-bound piRNAs	166988	76%

whole non-piRNA sequences. Furthermore, the distribution of the "TGA" motif between the piRNA and non-piRNA sequences was significantly different by the independent two-sample K-S test ($p<0.05$). This indicates that the piRNA sequences may contain information on the 5' and/or 3' end, as recognized by the piRNA processing apparatus of actual piRNA precursors.

Web Server Guide

For the convenience of the other researchers, a web server for implementing the algorithm and the software code is freely available at http://app.aporc.org/Pibomd/. Here are the step-by-step instructions.

Step1. Open the web server at http://app.aporc.org/Pibomd/, and you will see the dialogue window on your computer screen (Fig.5). Click on the "Read Me" button, and you will see a brief introduction describing the Pibomd predictor.

Step2. Either paste or upload the query sequences in FASTA format in the text box at the center of your dialogue window (Fig.5). A sample FASTA format can be seen when you click on the "Example" button on the top of the input frame.

Step3. Chose the model type. Users can choose the model type according to their requirements. Pibomd method can achieve high predicted ACC. Asym-Pibomd gets relatively lower predicted ACC than pibomd with higher specificity and lower sensitivity.

Step4. Click on the "Submit" button to get the predicted result.

Step5. Click on the "Data" button to download the data to train and test the Pibomd predictor.

Step6. Click on the "Citation" button to find documents related to the Pibomd algorithm.

Conclusions

Pibomd was developed to predict piRNAs by using only variable-length motifs that frequently appear in RNA sequences as features, using the SVM algorithm. At the outset, this work was only intended to improve piRNA prediction software based solely on sequence information. However, when using the Teiresias tool to identify sequence motifs of variable length in piRNA and non-piRNA sequences, it was discovered that we achieved more accurate piRNA identification than that of other previously published software and that we obtained a very informative catalogue of nucleotide information in the motifs that distinguish piRNAs from other kinds of RNA sequences in the genome. Further analysis of the motif positions on the piRNA sequences led to important clues suggesting that the sequence information of RNA sequences is utilized by the piRNA processing apparatus.

It is unfortunate that we cannot get the result of a genome screen due to the computational complexity and computational ability. The present version of Pibomd could not search for the mouse genome yet. We plan to fulfil it in the following work.

Acknowledgements

This work was supported by the National Key Basic Research & Development Program (973) under Grant No. 2011CB808000, the National Natural Science Foundation of China under Grant No. 11101028 and No. 11301024,the Foundation of Key Laboratory of Random Complex Structures and Data Science, Academy of Mathematics &Systems Science, Chinese Academy of Sciences (No. 2008DP173182), and by NCMIS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Notes and references

1. A. Aravin, D. Gaidatzis, S. Pfeffer, M. Lagos-Quintana, P. Landgraf, N. Iovino, P. Morris, M. J. Brownstein, S. Kuramochi-Miyagawa, T. Nakano, M. Chien, J. J. Russo, J. Ju, R. Sheridan, C. Sander, M. Zavolan and T. Tuschl, *Nature*, 2006, 442, 203-207.
2. J. Brennecke, A. A. Aravin, A. Stark, M. Dus, M. Kellis, R. Sachidanandam and G. J. Hannon, *Cell*, 2007, 128, 1089-1103.
3. A. Girard, R. Sachidanandam, G. J. Hannon and M. A. Carmell, *Nature*, 2006, 442, 199-202.
4. S. T. Grivna, E. Beyret, Z. Wang and H. Lin, *Genes Dev*, 2006, 20, 1709-1714.
5. K. Saito, K. M. Nishida, T. Mori, Y. Kawamura, K. Miyoshi, T. Nagami, H. Siomi and M. C. Siomi, *Genes Dev*, 2006, 20, 2214-2222.
6. V. V. Vagin, A. Sigova, C. Li, H. Seitz, V. Gvozdev and P. D. Zamore, *Science*, 2006, 313, 320-324.
7. H. Yin and H. Lin, *Nature*, 2007, 450, 304-308.
8. A. A. Aravin, R. Sachidanandam, D. Bourc'his, C. Schaefer, D. Pezic, K. F. Toth, T. Bestor and G. J. Hannon, *Mol Cell*, 2008, 31, 785-799.
9. A. K. Lim, L. Tao and T. Kai, *J Cell Biol*, 2009, 186, 333-342.
10. M. Ghildiyal and P. D. Zamore, *Nature reviews. Genetics*, 2009, 10, 94-108.
11. A. A. Aravin, G. J. Hannon and J. Brennecke, *Science*, 2007, 318, 761-764.
12. C. D. Malone and G. J. Hannon, *Cell*, 2009, 136, 656-668.
13. E. J. Lee, S. Banerjee, H. Zhou, A. Jammalamadaka, M. Arcila, B. S. Manjunath and K. S. Kosik, *RNA*, 2011, 17, 1090-1099.

14. Y. Zhang, X. Wang and L. Kang, *Bioinformatics*, 2011, 27, 771-776.
15. D. Betel, R. Sheridan, D. S. Marks and C. Sander, *PLoS Comput Biol*, 2007, 3, e222.
16. P. P. Das, M. P. Bagijn, L. D. Goldstein, J. R. Woolford, N. J. Lehrbach, A. Sapetschnig, H. R. Buhecha, M. J. Gilchrist, K. L. Howe, R. Stark, N. Matthews, E. Berezikov, R. F. Ketting, S. Tavare and E. A. Miska, *Mol Cell*, 2008, 31, 79-90.
17. N. Robine, N. C. Lau, S. Balla, Z. Jin, K. Okamura, S. Kuramochi-Miyagawa, M. D. Blower and E. C. Lai, *Curr Biol*, 2009, 19, 2066-2076.
18. X. Liu, S. He, G. Skogerbo, F. Gong and R. Chen, *PLoS One*, 2012, 7, e32797.
19. C. Liu, B. Bai, G. Skogerbo, L. Cai, W. Deng, Y. Zhang, D. Bu, Y. Zhao and R. Chen, *Nucleic Acids Res*, 2005, 33, D112-115.
20. A. Vourekas, Q. Zheng, P. Alexiou, M. Maragkakis, Y. Kirino, B. D. Gregory and Z. Mourelatos, *Nat Struct Mol Biol*, 2012, 19, 773-781.
21. I. Rigoutsos and A. Floratos, *Bioinformatics*, 1998, 14, 55-67.
22. Chih-Chung Chang and Chih-Jen Lin, *ACM Transactions on Intelligent Systems and Technology,*, 2011, 2, 1-27.
23. K. e. a. Duan, *Neurocomputing*, 2003, 51, 41-59.
24. K. L. Ng and S. K. Mishra, *Bioinformatics*, 2007, 23, 1321-1330.
25. Y. W. Chen, & Lin, C. J. , 2005.
26. M. e. a. Kandhavelu, *Journal of Bioinformatics and Sequence Analysis*, 2009, 31-40.

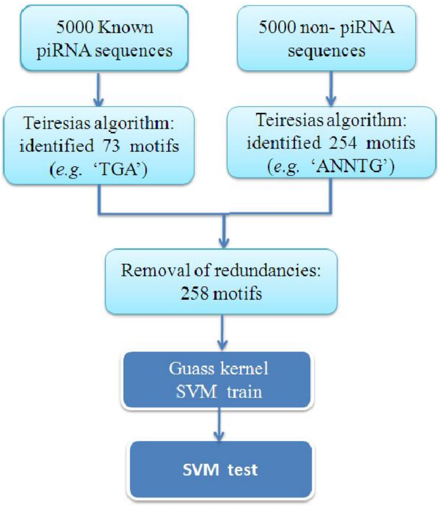

5000 Known piRNA sequences
5000 non- piRNA sequences
Teiresias algorithm: identified 73 motifs (e.g. 'TGA')
Teiresias algorithm: identified 254 motifs (e.g. 'ANNTG')
Removal of redundancies: 258 motifs
Guass kernel SVM train
SVM test

Fig.1 Pibomd pipeline.

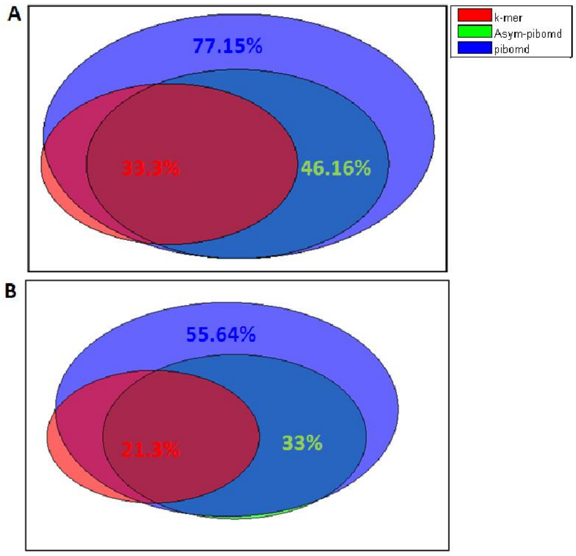


Fig.2 Pibomd identifies mouse CLIP piRNAs more efficiently than the k-mer method. Venn diagram showing the intersections between mouse piRNAs identified by the k-mer method[14] (red circle), Pibomd predictions (large blue circle), and Asym-Pibomd predictions (small green circle). (A) Clip piRNA sites in piRBase , (B) chromosome cluster piRNAs.

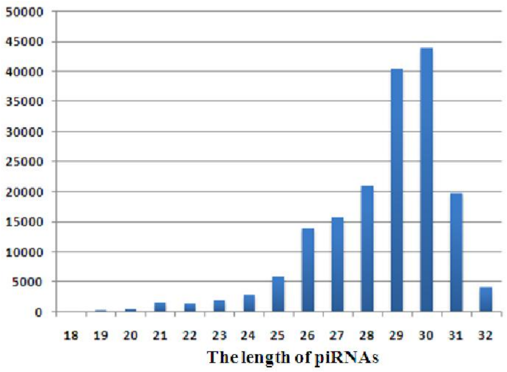


Fig.3 Frequency distribution of the length of piRNAs.

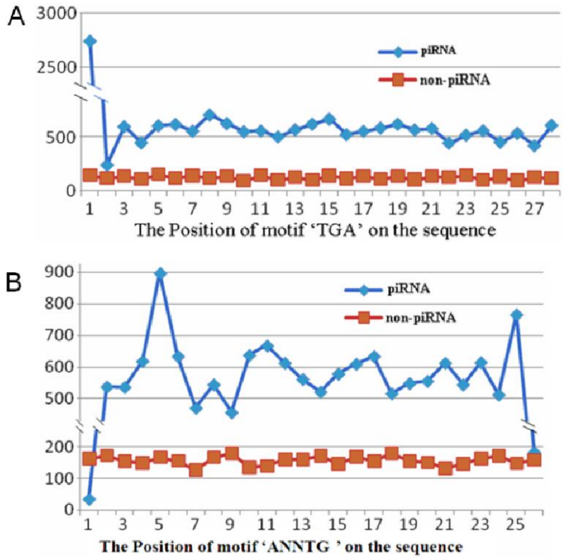
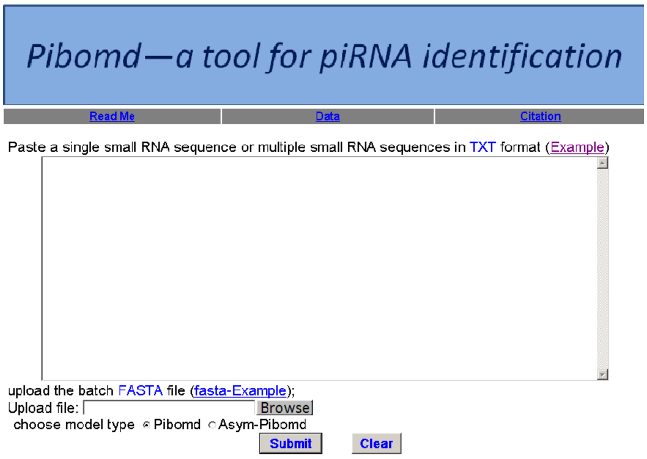


Fig.4 Position of motifs. (A) Motif 'TGA' usually appears on the 5' end of the piRNA sequences. (B) Motif 'ANNTG' appears with significantly less frequency on the 5'and 3' ends of piRNA sequences ("N" denotes any nucleotide).


Pibomd—a tool for piRNA identification

Paste a single small RNA sequence or multiple small RNA sequences in TXT format (Example)

upload the batch FASTA file (fasta-Example);

Upload file:

choose model type ● Pibomd ○ Asym-Pibomd

Fig.5 Screen shot showing the Pibomd web server. Its website address is at http://app.aporc.org/pibomd/