Send Orders for Reprints to reprints@benthamscience.net

RESEARCH ARTICLE

Computational Identification of piRNAs Using Features Based on RNA Sequence, Structure, Thermodynamic and Physicochemical Properties

Isha Monga[1,*] and Indranil Banerjee[1,*]

[1]*Cellular Virology Laboratory, Department of Biological Sciences, Indian Institute of Science Education and Research, Mohali (IISER Mohali) Sector 81, S.A.S. Nagar, Mohali-140306, India*

ARTICLE HISTORY

Received: September 06, 2019
Revised: November 08, 2019
Accepted: November 22, 2019

DOI:
10.2174/1389202920666191129112705

Abstract: ***Rationale*:** PIWI-interacting RNAs (piRNAs) are a recently-discovered class of small non-coding RNAs (ncRNAs) with a length of 21-35 nucleotides. They play a role in gene expression regulation, transposon silencing, and viral infection inhibition. Once considered as "dark matter" of ncRNAs, piRNAs emerged as important players in multiple cellular functions in different organisms. However, our knowledge of piRNAs is still very limited as many piRNAs have not been yet identified due to lack of robust computational predictive tools.

***Methods*:** To identify novel piRNAs, we developed *piRNAPred*, an integrated framework for piRNA prediction employing hybrid features like *k*-mer nucleotide composition, secondary structure, thermodynamic and physicochemical properties. A non-redundant dataset (D^{3349} or $D^{1684p+1665n}$) comprising 1684 experimentally verified piRNAs and 1665 non-piRNA sequences was obtained from *piRBase* and *NONCODE,* respectively. These sequences were subjected to the computation of various sequence-structure based features in binary format and trained using different machine learning techniques, of which support vector machine (SVM) performed the best.

***Results*:** During the ten-fold cross-validation approach (10-CV), *piRNAPred* achieved an overall accuracy of 98.60% with Mathews correlation coefficient (MCC) of *0.97* and receiver operating characteristic (ROC) of 0.99. Furthermore, we achieved a dimensionality reduction of feature space using an attribute selected classifier.

***Conclusion*:** We obtained the highest performance in accurately predicting piRNAs as compared to the current state-of-the-art piRNA predictors. In conclusion, *piRNAPred* would be helpful to expand the piRNA repertoire, and provide new insights on piRNA functions.

Keywords: piRNA, classification, algorithm, prediction, non-coding RNA, physicochemical.

1. INTRODUCTION

Since its discovery, RNA interference (RNAi) and its effector non-coding RNAs (ncRNAs) namely, microRNAs (miRNAs), small interfering RNAs (siRNAs), and PIWI-interacting RNA (piRNAs) have revolutionized our understanding of the mechanisms regulating gene expression [1, 2]. The common key event in all the RNAi regulatory mechanisms is the binding of small ncRNAs to Argonaute (AGO), forming a ribonucleoprotein complex termed as RNA-induced silencing complex (RISC) [3]. The AGO family of proteins is phylogenetically divided into two sub-families: somatic AGO [4], and germ line specific PIWI (P-element induced wimpy testis) clade [2, 5]. Both miRNAs and siRNAs act through the AGO family and constitute the RISC complexes known as the miRISC and siRISC, whereas the small ncRNAs interacting with PIWI are known as the piRNAs [6].

piRNA is a recently discovered class of ncRNAs, which are in the length range of ~24-32 nucleotides [1, 7, 8]. Initially, piRNAs were described as repeat-associated siRNAs (rasiRNAs) because of their origin from the repetitive elements such as transposable sequences of the genome [9]. However, later it was identified that they acted *via* PIWI-protein [10]. In addition to having a role in the suppression of genomic transposons, various roles of piRNAs have been recently reported like regulation of 3' UTR of protein-coding genes *via* RNAi [11], transgenerational epigenetic inheritance to convey a memory of past transposon activity [12], and RNA-induced epigenetic silencing [13]. Furthermore, piRNA sequences are comparatively diverse than any other class of cellular ncRNAs and they constitute the most prevalent class of ncRNAs [7].

The overall mechanism of piRNA biogenesis is substantially different than that of the other ncRNAs. It includes siRNA, miRNA and long ncRNAs (lncRNAs) [13]. piRNAs

*Address correspondence to these authors at the Cellular Virology Laboratory, Department of Biological Sciences, Indian Institute of Science Education and Research, Mohali (IISER Mohali) Sector 81, S.A.S. Nagar, Mohali-140306, India; Tel: +91-7044936698;
E-mails: indranil@iisermohali.ac.in; mongaisha4@gmail.com

1389-2029/19 $58.00+.00 ©2019 Bentham Science Publishers

are generated from long, single-stranded RNAs, which are transcribed from the genomic loci termed as piRNA clusters [14-19]. Earlier studies suggested multiple mechanisms of piRNA biogenesis in different cell types, tissues and organisms [10]. However, recent reports proposed a single and unified model that explained the mechanism of piRNA biogenesis across a range of evolutionarily diverse organisms [20-23]. These studies indicated that piRNA biogenesis could be divided into two parts: 1) the piRNA-dependent amplification loop, known as the 'Ping-Pong cycle', and 2) the piRNA-independent generation of phased trailing piRNAs. The former pathway begins with a maternally inherited 1U-biased piRNA, known as the initiator piRNA [24]. The PIWI-bound initiator piRNA cleaves the complementary single-stranded long transcript sequence into a pre-pre-piRNA with a terminal 5' monophosphate [23]. The PIWI-bound pre-pre-piRNA undergoes subsequent RNA cleavage from its 5' end to produce responder pre-piRNA. The responder pre-piRNA further undergoes 3' end-processing to generate mature responder piRNA with 10A-bias. Subsequently, the mature responder piRNA enters the piRNA biogenesis cycle, acting as an initiator piRNA, and in turn, produces a new responder piRNA with 1U-bias, which is identical to the original initiator piRNA. The pathway operates as an amplification loop and hence, it is termed the 'Ping-Pong' cycle [7]. Thus, this arm of the unified model is dependent on a maternal initiator piRNA and also on the availability of long single-stranded piRNA precursor sequences. In contrast, phased piRNA generation is a piRNA-independent process, which operates through interaction with the mitochondrial endonuclease phospholipase D family member 6 (PDL6), an endonuclease present on the outer mitochondrial membrane. PDL6 cleaves the remaining 3' end of the pre-pre-piRNA in a repeated manner, producing an array of tail-to-head pre-piRNAs, which are known as phased trailing piRNAs [21]. Thus, the current model proposed that the initiator piRNAs produce responder piRNAs and trailing piRNAs, which were known as the secondary and primary piRNAs, respectively, in the old model [22]. Hence, owing to the extensive role of piRNAs in regulating various biological processes, the conserved features of their biogenesis, and their sequence diversity, genome-wide identification of novel piRNAs would be of great importance to help understand piRNA-guided gene regulatory mechanisms across different cell types and organisms.

In piRNA identification, the general alignment-centered algorithms such as basic local alignment search tool (BLAST) [25] and Multiple Expectation Maximization for Motif Elicitation (MEME) [26] are well known for their robustness, but not for accuracy due to divergence of piRNAs across different species [27]. Recently, next-generation sequencing (NGS)-based methods have emerged as a powerful platform to identify piRNAs in a high-throughput mode [28]. However, in addition to piRNA, sequencing-generated data may also harbor reads from several other small ncRNAs such as miRNAs, endogenous siRNAs (endo siRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), smaller fragments of mRNAs and long ncRNAs in the piRNA length range. Therefore, the length range alone cannot be a characteristic parameter to screen for piRNAs in small RNA-based sequencing platforms. Hence, more accurate and robust algorithms are required to distinguish the real piRNA sequences from the pseudo piRNAs (non-piRNAs) of similar lengths by employing various discriminative features.

Recent studies employed multiple approaches to identify piRNAs with high accuracy in the genome of various organisms (Table **S1**). Betel *et al.* proposed an algorithm based on the position-specific scoring matrix (PSSM) of 10 nucleotides upstream and downstream of piRNA sequences identified from the pachytene stage of mouse spermatogenesis, and predicted mouse-derived conserved piRNAs with ~60-70% precision [29]. Nevertheless, developed on a homogeneous dataset and trained on static position-specific nucleotide usage, the method shows position-specific features for only one species and hence, has limitations in piRNA identification in other species. Zhang *et al.* developed an algorithm using *k*-mer nucleotide frequency to capture the sequence-based characteristics in piRNA sequences [27]. Since the algorithm adopted dynamic nucleotide frequency calculation over the static PSSMs, it was able to identify piRNAs in the organisms beyond the training set species. However, there was a need to include additional features to improve the accuracy of piRNA identification apart from the sequence-based features. Other studies integrated additional features than sequence alone like structure and thermodynamic energy, and developed different piRNA prediction algorithms. The algorithm *'PIANO'* was developed to predict piRNAs in *Drosophila melanogaster* using structural features and transposon interactions [30, 31], adopting local contiguous sequence-structure triplet-elements [27]. To predict mouse piRNAs using sequence motifs, '*Pibomd'* was developed [32]. The above tools were trained on homogeneous datasets generated from single species (Table **S1**). 'Accurate piRNA prediction' [33], 'GA-WE' (a genetic algorithm-based weighted ensemble method) [34] and '2L-piRNA' (a two-layer ensemble classifier) [35] utilized sequence, structure and *k*-mer spectrum profile-based features to achieve better performance. Although these tools demonstrated a high accuracy level in predicting piRNAs, there was a need to develop a more robust algorithm capable of analyzing heterogeneous datasets from multiple organisms with high accuracy.

In this study, a comprehensive and robust machine learning-based algorithm is developed to predict piRNA sequences relying on the hybrid features such as spectrum profile or *k*-mer features (*k*=1 to 5), secondary structure, thermodynamic energy and physicochemical properties of RNA dinucleotides extracted from the piRNA sequences of eight species: *Homo sapiens, Mus musculus, D. melanogaster, Caenorhabditis elegans, Danio* rerio*, Gallus gallus domesticus, Xenopus tropicalus, and Bombyx mori* (Table **S1**). During the 10-fold cross-validation (10-CV), our method reached an overall accuracy of >98% and a sensitivity of >98% in the above species. When compared with the existing algorithms, our hybrid predictive model *piRNAPred* demonstrated higher accuracy level. In conclusion, *piRNAPred* is the most updated piRNA identification method developed by integrating features of sequence, structure, thermodynamic energy and physicochemical properties extracted from a heterogeneous dataset across phylogenetically diverse organisms.

2. MATERIALS AND METHODS

2.1. Dataset Preparation

The experimentally verified positive piRNA and non-piRNA sequences were extracted from piRBase [28] and NONCODE [36], respectively. Since both the databases contain millions of sequences from numerous organisms, we removed redundant and overlapping sequences using a cluster database at high identity with tolerance (CD-Hit) software. Removal of the redundant sequences was achieved at different percentage identity level; we used 60% sequence identity. Sample sequences having 60% identical overlap among them were grouped into one cluster and represented as a single sequence. The rationale of clustering was to verify the redundancy/overlap of the sequences to reduce the number of sequences at different percentage sequence identity. Finally, a non-redundant benchmark dataset (D^{3349} or $D^{1684p+1665n}$) comprising 1684 experimentally verified piRNAs and 1665 non-piRNA sequences was obtained. These piRNA sequences were extracted from eight organisms (Table **S1**). Further, benchmark dataset was subjected to the computation of various sequence-structure based features in binary format, and trained using different machine learning techniques (MLTs). Fig. (**1**) entails the step-wise methodology adopted in the development of piRNAPred.

Apart from the above heterogeneous benchmark dataset representing sequences from evolutionarily distant species, we considered two homogeneous datasets in a tissue-specific manner. The first tissue-specific dataset was extracted from a single study, which involved the identification of piRNAs and non-piRNAs from the testes of C57BL/6P20$^{Miwi\ +/+}$ and C57BL/6P20$^{Miwi-/ADH}$ mice, respectively, using MIWI-Immunoprecipitation (IP) followed by sequencing [37]. There were 57,5786 piRNA and 55,1640 non-piRNA sequences in GSM822759 and GSM822762, respectively. These sequences were screened, processed and made non-redundant, which resulted in a total of 2000 positive and 1711 negative sequences. Similarly, the second tissue-specific dataset involved 1418 piRNA and 1418 pseudo-piRNA sequences form *M. musculus* [35]. These tissue-specific homogeneous datasets were subsequently subjected to features formulation and model development using 10-CV.

3. FEATURES UTILIZED

3.1. Spectrum Profile or *k*-mer Features

One of the important contributing features of predictive model development is the *k*-mer string [33, 35, 38]. In machine learning, a *k*-mer string is defined as particular *k*-mer tuple (1, 2, 3, 4, 5 or even higher) of nucleotide or amino acid sequences that can be used to identify some representative motifs within DNA or proteins. The overall idea of using different *k*-mer strings is the identification of differential nucleotide usage between the authentic piRNA and pseudo piRNA sequences. However, when the dataset sequences are of different length, spectrum profile or *k*-mer nucleotide composition (k-MNC) is used [34, 39]. *k*-MNC is defined as a total number of a particular nucleotide divided by the length of the sequence. *e.g.*, a piRNA sequence with five guanine residues and of length *n,* where *n=19* nucleotides, the percentage composition of guanine residues will be 26.31%. Hence, there are 4 mononucleotide composition-based features (A, T/U, G and C).

$$\text{Composition of } k = \frac{\text{Number of } k * 100}{\text{Number of all nucleotides}}$$

(*k*, any amino acid or nucleotide)

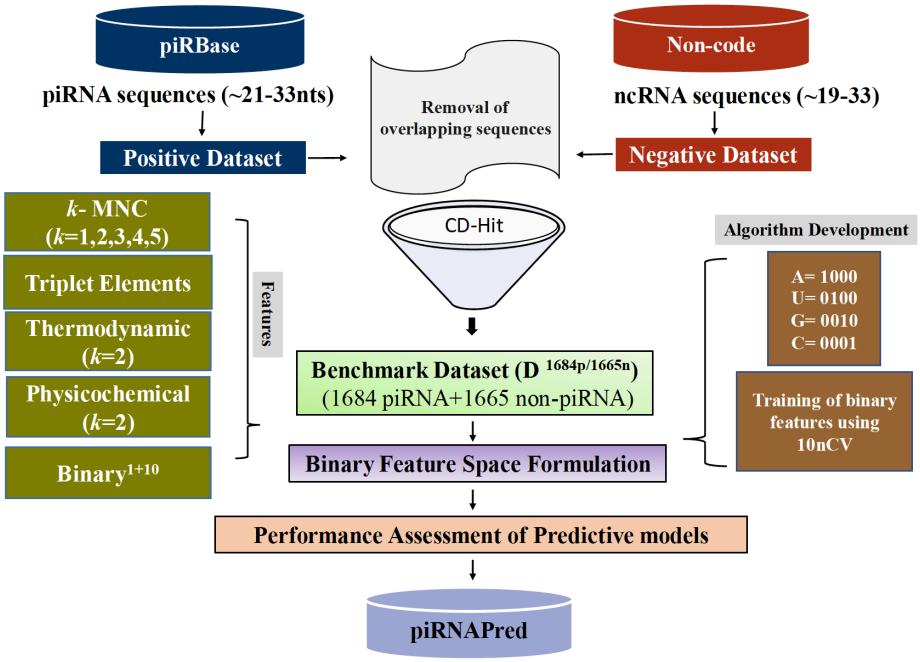

Fig. (1). Schematic illustration of the overall workflow adopted to develop piRNAPred: Left and right arm demonstrates the processing of piRNA and non-piRNA sequence from piRBase and NONCODE, respectively to generate a dataset ($D^{1684p+1665n}$) followed by their downstream conversion into sequence, structure, thermodynamic, physiochemical and BINARY^{1+10} feature space and predictive model development. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Similarly, there will be 16-dinucleotide composition-based features (AA, AU, AC, AG, UA, UU, UC, UG, CA, CU, CC, CG, GA, GU, GC, GG). From *k*=1 to 5, they are designated as mono-, di-, tri-, tetra- and penta-nucleotide compositions. Hybrid predictive models (*k*=1, 2/1, 2, 3 *etc*.) involve higher-order combinations like hybrid of mono- and di-nucleotide composition is (*md)*, mono-, di- and tri-nucleotide is *mdt,* mono-, di-, tri- and tetra-nucleotide *is mdtt* and mono-, di-, tri-, tetra- and penta-nucleotide is *mdttp.* Thus, *k*-MNC can be utilized to identify the motifs or regions of interest, which can distinguish the positive sequences from the negative ones in terms of differential nucleotide usage. Here, we employed a hybrid model of *k*=1-5, resulting into a total of 1364 strings of feature space.

3.2. Sequence-Structure Triplet Elements (SSTE)

A set of 32 features of *local contiguous structure-sequence triplet element (SSTE)* was proposed by Xue *et al.*, for distinguishing the hairpins of real pre-miRNAs and pseudo pre-miRNAs [31]. These SSTEs can also be incorporated for extracting real piRNAs from pseudo piRNA sequences. In a secondary structure, there are two states for every nucleotide, paired or unpaired, which are designated by brackets: either "(" or ")" or dots "." Correspondingly, standard secondary structure notation in RNAfold was developed in the Vienna RNA package where, a paired nucleotide near the 5'-side "(", can form pair with ")" nucleotide at 3' end [40]. To convert them into a binary format, the paired state was designated as "1", while the unpaired state as "0". Further, an *SSTE* feature was calculated, which consisted of the secondary structure state of three contiguous nucleotides with the identity of the nucleotide in the middle [*e.g.* A(((]. For example, there will be 8 possible SSTEs of "adenine" as ["A(((", "A((.", "A(..", "A(.(", "A.((", "A.(.", "A..(" and "A..."]. Therefore, there will be 32 (4 × 8) possible structure-sequence combinations for the nucleotides A, U, G, and C (Table **S2**) [31].

3.3. Thermodynamic Energy of Contiguous Dinucleotides

It was reported that thermodynamic properties of miRNA and siRNA sequences contribute to determining strand bias, molecule's function, and longevity [41]. The thermodynamic energy feature space in the study corresponds to Gibbs free energy of continuous two nucleotides ina siRNA/miRNA sequence [41]. These features were incorporated to develop computational models for siRNA prediction and demonstrated considerable improvement [42, 43]. Therefore, to check the potential of thermodynamic energy of contiguous dinucleotides in piRNA identification from false positive sequences, we calculated 16 thermodynamic energy-related patterns, and develop predictive models.

3.4. RNA Physicochemical Properties

There are six physicochemical properties of RNA dinucleotides namely rise, roll, shift, tilt, slide, and twist (Table **S3**). Their values have been used for the prediction of ncRNAs such as miRNA and piRNA [33]. The dinucleotide combination of an individual physicochemical property leads to the generation of 16 features, and these six RNA physicochemical properties collectively create a feature space of 96 strings.

3.5. Binary Profile of 1U and 10A Bias of piRNAs

We have included the information about the relative preference of nucleotides at 1st and 10th position of the primary piRNA and secondary piRNA in the form of a new predictive feature termed as the binary profile of 1+10 position ($BINARY^{1+10}$). This feature reflects the relative occupancy of a particular nucleotide (A/U/G/C) at a given position of the piRNA sequence. Thereafter, this position-specific occupancy was converted into SVM readable binary format resulting in a total of 16 features (4 for 1st and 4 for 10th position, respectively). Thus, it provided information about the relative preference of a nucleotide at each position of piRNA (Table **1**).

All the above features were combined in a hybrid model *i.e. k*-MNC with *k*=1-5, leading to 1364 vectors, 32 SSTE, 16 thermodynamic energy-, 96 physicochemical property-related features and 16 features for $BINARY^{1+10}$ for 1U and 10A bias of piRNA making up to a total of 1516 features space (1364+32+16+96+16). As computational algorithms are only trained on binary notation (0 or 1), these features were further needed to be transformed into binary symbols to implement the feature vectors in different MLTs. In the present study, binary symbolisation for standard 4 nucleotides was in a string of 4 characters: A, U, G and C, and are denoted by 1000, 0101, 0010, and 0001, respectively [38]. Further, the classifier was trained on 1 for "ON" and 0 for "OFF" signals for a particular feature. In this way, we could provide sequence, structure or any other biological features into binary format to train the classifier. The main goal of computing composition is to change the varied length of a sequence to a fixed length vector. This step was crucial while performing the classification of sequences by MLTs.

4. ALGORITHM DEVELOPMENT

We used SVM^{light} software package (http://svmlight.joachims.org/) to train the 1516 features of $D^{1684p+1665n}$ and to develop the predictive model in 10-CV mode. SVM^{light} is an implementation of a supervised MLT called support vector machines (SVM), which was proposed by Vapnik [44]. It works on the principle of pattern-recognition in the given dataset during training, and predicts the test-set based on these patterns. We used radial basis function (RBF) kernel having kernel width parameter (gamma, *g*) and regularization parameter (c) to build the SVM-based models in the range of 0.000001 to 5 and 0.00001 to 2000, respectively. Apart from SVM, we utilized other MLTs (Random Forest (RF)), Bagging, classification *via* regression, J48 pruned tree, naive Bayes and Instance Based Learner (Ibk)) using *Weka* package (Table **2**).

5. FEATURE SELECTION TO ACHIEVE DIMENSIONALITY REDUCTION

An optimal classifier depends upon features used for training. Therefore, to screen out the maximum relevant subspace out of 1516 features, we employed attribute selected classifier in the *Waikato Environment for Knowledge Analysis (Weka)* package [45], and retrieved reduced feature space and provide 198 maximum relevance features (Table **S4**). Further, we checked the performance of these 198 features in different MLTs (Table **S5**).

6. STATISTICAL ASSESSMENT OF ALGORITHM

The performance of predictive models on the test set was assessed with the help of the statistical equation Mathews correlation coefficient (MCC). It is generally applied to calculate correlation between the actual and predicted values along with other statistical measures namely, percentage sensitivity (Sn), specificity (Sp), accuracy (Ac), true positive (TP), true negative (TN), false positive (FP), and false negative (FN) [35, 46]. These equations are provided below:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$Senstivity(Sn) = [TP/(TP + FN)] * 100$$

$$Specificity(Sp) = [TN/(TN + FP)] * 100$$

$$Accuracy(Ac) = [TP + TN/(TP + FP + TN + FN)] * 100$$

7. RESULTS

7.1. Performance of Individual Features During 10-CV

7.1.1. Spectrum Profile or k-mer Features

The k-mer models *i.e.* mono-nucleotide (mono-NT), di-NT, tri-NT, tetra-NT and penta-NT achieved a MCC of 0.35, 0.37, 0.41, 0.51, and 0.56, respectively. We also assessed the performance of hybrid *k*-mer features *e.g.* a combination of mono- and di-nucleotide termed as md hybrid model. The mono-di (MD), mono-tri (MT), di-tri (DT), mono-di-tri (MDT), mono-di-tri-tetra (MDTT) and mono-di-tri-tetra-penta (MDTTP) achieved a minimum MCC of 0.38 to a maximum MCC of 0.44. Among all the *k*-mer models, MDTTP performed the best, and hence, it was selected to combine with structural and thermodynamic features to create a hybrid model (Table **1**).

7.1.2. SSTE Based Features

SSTE based models achieved excellent performance with a percentage accuracy of 95.61, sensitivity and specificity of 93.53 and 97.72, respectively.

7.1.3. Thermodynamic Energy- based Features

Thermodynamic energy-based predictive models scored percentage sensitivity, specificity, accuracy of 76.13, 81.38, and 78.74, respectively, with an MCC of 0.58.

7.1.4. RNA Physicochemical Property-based Features

The 96 RNA physicochemical property-based features exhibited percentage sensitivity, specificity, and accuracy of 77.26, 76.70, and 76.98 respectively, with an MCC of 0.54.

7.1.5. Binary Profile of 1U and 10A Bias of piRNA

The 16 features reflecting the position-specific relative occupancy of a particular nucleotide exhibited percentage sensitivity, specificity, and accuracy of 69.24, 44.85, and 57.12, respectively with an MCC of 0.15 (Table **1**).

7.2. Performance of Hybrid Features During 10-CV

For assessing the performance of hybrid models with MDTTP and other features during 10-CV, we employed SSTE (*A*), thermodynamic energy-based features (*B*), RNA physicochemical property-based features (*C*) and BINARY^{1+10} (D). Out of all predictive feature combinations, MDTTP+A+B+C *i.e.* a hybrid of MDTTP, SSTE, thermodynamic energy and RNA physicochemical properties achieved the best result as it exhibited 98.60% accuracy, 98.57% sensitivity, 98.62% specificity, and 0.97 MCC followed by MDTTP+ABC+BINARY^{1+10}, which performed almost equally well: 98.24% accuracy, 99.05% sensitivity, 97.42% specificity and 0.96 MCC. We also plotted the performance of the three best predictive models in the graphic analysis using the receiver operating characteristic (ROC) plot (Fig. **2**).

Additionally, to reflect the performance of a predictive model trained on homogeneous dataset from a single tissue, we developed two models trained on the piRNA and non-piRNA sequences in a single tissue-specific manner from mouse testes and ovary. The dataset was formulated into 1516 features, accessed using 10-CV resulting in models MDTTP+ABC+Binary^{1+10} (*testes)* and MDTTP+ABC+Binary^{1+10} (*ovary*) for testes and ovary dataset, respectively. Further, we performed external validation by checking the performance of the testes model to predict the ovary dataset and *vice-versa* (Table **S6**).

7.3. Performance of MDTTP+A+B+C Trained on Other MLTs

To check the performance of other MLTs on the 1508 feature space, we performed 10-CV. However, the SVM performed the best of all the tested MLTs, followed by random forest, bagging, and classification-*via*-regression (Table **2**). The best predictive model was termed as "*piRNAPred*", and has been provided to users for the prediction of piRNAs (https://github.com/IshaMonga/piRNAPred).

7.4. Cross-validation

We adopted the 10-CV method for validating our classifier. During 10-CV, the complete dataset was randomly divided into 10 sets, of which the model was trained on 9 sets (the training set) and 1 set was kept aside for testing (the test set) (Fig. **S1**). This process was recursively repeated ten-times, and the performance of ten steps was averaged to provide the final assessment of the predictive model [38, 42].

7.5. Comparison with Other State of the Art Predictors

Since the existing piRNA prediction algorithms were developed on dissimilar features employing diverse MLTs, therefore they exhibited different sensitivity and specificity levels: piRNA- 72.47% and 95.53% [27], PIANO- 95.89% and 94.61% [30], *Pibomd*- 91.48% and 89.76% [32], *accurate piRNA prediction*- 83.10% and 82.10% [33], *GA-WE* - 90.6% and 78.3% [34] and *2L-piRNA*- 88.3% and 83.9% [35]. On the other hand, *piRNAPred* achieved 98.57% sensitivity, 98.62% specificity, 98.60% accuracy and 0.97 MCC (Table **3**).

Table 1. Performance of different predictive models using SVM during 10-fold cross-validation.

S. No.	Predictive Model	Features	No. of Features	Thres	TP	FP	TN	FN	Sn (%)	Sp (%)	Acc (%)	Mcc	g	c	ROC
1	MONO	*k*-MNC based features	4	0	1426	867	798	258	84.7	47.93	66.41	0.35	1	5	0.75
2	DI		16	0.1	1153	531	1134	531	68.5	68.11	68.29	0.37	5.00E-05	50	0.73
3	TRI		64	0.1	1189	501	1164	495	70.6	69.91	70.26	0.41	5.00E-05	500	0.76
4	TETRA		256	0	1126	282	1383	558	66.9	83.06	74.92	0.51	0.01	1	0.81
5	PENTA		1024	0	1363	412	1253	321	80.9	75.26	78.11	0.56	0.0005	50	0.85
6	MD	Hybrid *k*-MNC based features	20	0	1212	574	1091	472	72	65.53	68.77	0.38	1.00E-05	300	0.74
7	MT		68	0.1	1165	484	1181	519	69.2	70.93	70.05	0.40	5.00E-05	100	0.75
8	DT		80	0	1228	557	1108	456	72.9	66.55	69.75	0.40	0.0001	50	0.76
9	MDT		84	0.1	1172	499	1166	512	69.6	70.03	69.81	0.40	1.00E-05	1000	0.75
10	MDTT		340	0	1233	554	1111	451	73.2	66.73	69.99	0.40	0.0001	10	0.76
11	MDTTP		1364	-0.1	1308	572	1093	376	77.7	65.65	71.69	0.44	5.00E-05	300	0.79
12	A	SSTE based features	32	0	1575	38	1627	109	93.5	97.72	95.61	0.91	0.0001	10	0.70
13	B	Thermo-dynamic energies of RNA dinucleo-tides	16	0.1	1282	310	1355	402	76.1	81.38	78.74	0.58	0.05	250	0.86
14	C	Physico-chemical properties of RNA dinucleo-tides	96	0	1301	388	1277	383	77.3	76.7	76.98	0.54	1.00E-05	5	0.85
15	BINARY^{1+10}	position specific NT usage of 1st and 10th posi-tion (Bina-ry)	16	0	817	642	522	363	69.24	44.85	57.12	0.15	1.00E-005	100	0.35
16	AB	Hybrids of SSTE, Thermo-dynamic energies of RNA dinucleo-tides and Physico-chemical properties of RNA dinucleo-tides (ABC) with MDTTP	48	-0.3	1653	27	1638	31	98.2	98.38	98.27	0.97	0.0005	500	1.00
17	AC		128	0	1468	230	1435	216	87.2	86.19	86.68	0.73	1.00E-05	1000	0.93
18	BC		112	0	1301	388	1277	383	77.3	76.7	76.98	0.54	1.00E-05	5	0.85
19	ABC		144	-0.1	1502	275	1390	182	89.2	83.48	86.35	0.73	1.00E-05	500	0.93
20	MDTTP + A		1396	0	1592	99	1566	92	94.5	94.05	94.3	0.89	5.00E-05	100	0.98
21	MDTTP + B		1380	0	1436	333	1332	248	85.3	80	82.65	0.65	5.00E-05	200	0.90
22	MDTTP + C		1460	0	1369	127	1538	315	81.3	92.37	86.8	0.74	0.01	5	0.89

(Table 1) contd….

S. No.	Predictive Model	Features	No. of Features	Thres	TP	FP	TN	FN	Sn (%)	Sp (%)	Acc (%)	Mcc	g	c	ROC
23	MDTTP +AB	Hybrid of *k*-MNC based features+ SSTE, thermody-namic, RNA physico-chemical properties and posi-tion spe-cific NT usage of 1st and 10th position (*Binary*)	1412	0	1608	71	1594	76	95.5	95.74	95.61	0.91	5.00E-05	100	0.99
24	MDTTP +BC		1476	0.1	1065	9	1656	619	63.2	99.46	81.25	0.67	0.01	5	0.89
25	MDTTP +AC		1492	0	1488	227	1438	196	88.4	86.37	87.37	0.75	1.00E-05	50	1.00
26	MDTTP +ABC		1508	0.1	1660	23	1642	24	98.57	98.62	98.60	0.97	1.00E-05	50	0.99
27	MDTTP+ ABC+ BINARY^{1+10}		1516	0	1668	43	1622	16	99.05	97.42	98.24	0.96	1.00E-005	50	0.99

Abbreviations: *Acc*, Accuracy; *diNT*, dinucleotide; *c*, Regularization parameter; *FP*, False Positive; *FN*, False Negative; *g*, Gamma (a kernel density parameter); *k-MF*, k-Mer Features OR *k-MNC*, k-Mer Nucleotide Composition; *MMP*, Mismatch Profile; *MCC*, Mathews Correlation Coefficient, *pseDNC*, pseudo Dinucleotide Composition; *PCPseDNC*, Parallel Correlation Pseudo Dinucleotide Composition; *PSSM*, Position-Specific Scoring Matrix; *SSTE*, Structure-Sequence Triplet Elements; *SSP*, Subsequence Profile; *Sn*, Sensitivity; *Sp*, Specificity; *SSTE*, Structure-Sequence Triplet Elements (A); *B*, Thermo-dynamic energies of contiguous dinucleotides; *C*, RNA physicochemical properties of adjoining dinucleotides (diNTs); *AB*, hybrid of SSTE and Thermo-dynamic energies of contiguous diNTs; Thres, threshold; TN, True Negative; TP, True Positive, *AC*, hybrid of SSTE and RNA physicochemical property of adjoining diNTs; ABC, hybrid of SSTE, thermo-dynamic energy and RNA physicochemical property of contiguous dinucleotides; ROC, Receiver Operating Characteristic.

Table 2. Performance of 1508 features on different machine learning techniques during 10-fold cross-validation.

S. No.	Machine Learning Technique	TP	FN	TN	FP	Sn (%)	Sp (%)	Acc (%)	MCC
1	SVM (*piRNAPred*)	1660	24	1642	23	98.57	98.62	98.60	0.97
2	Random Forest	1616	68	1647	18	95.96	98.92	97.43	0.95
3	Bagging	1611	73	1628	37	95.67	97.78	96.72	0.93
4	Classification *via* Regression	1586	98	1576	89	94.18	94.65	94.42	0.89
5	J48 Pruned tree	1553	131	1515	150	92.22	90.99	91.61	0.83
6	Naive Bayes	920	764	1362	303	54.63	81.8	68.14	0.38
7	IbK	1302	382	610	1055	77.32	36.64	57.09	0.15

Abbreviations: *Acc*, Accuracy; *FP*, False Positive; *FN*, False Negative; *MCC*, Mathews Correlation Coefficient, *Sn*, Sensitivity; *Sp*, Specificity; SVM, Support Vector Machines; TN, True Negative; TP, True Positive.

8. DISCUSSION

Cell growth, homeostasis and phenotypic expression of cellular functions have underlying complex regulations operating both at the gene expression and epigenetic levels [8, 10]. Gene expression regulation is majorly governed by small ncRNA-based surveillance and silencing in the biological systems [7, 47]. Among different players of RNAi, piRNAs are the most recently discovered and highly diverse class of ncRNAs [2]. They were first reported to govern the gonadal cell development by regulating the expression of transposons in the mouse germ cells [48-50], and loss of their expression leads to sterility [32-34]. Recent findings also suggest their roles in post-transcriptional gene regulation [11, 51], transgenerational epigenetic inheritance [12], regulation of mRNA decay [52], and localization in germplasm part of the embryo during embryogenesis [53]. Although many piRNAs have been recently discovered in some species, their identification has largely remained elusive in most of the organisms due to a lack of robust tools capable of identifying piRNA across phylogenetically diverse organisms with high accuracy.

The identification of piRNAs is majorly carried out by the NGS-based methods in which, many sequences from other ncRNAs fall within the range of piRNA sequence-length. Hence, to differentiate the piRNAs from the false-positive sequences, a more accurate and robust algorithm was warranted [33]. Recently, various methods were published which employed different features and used piRNAs

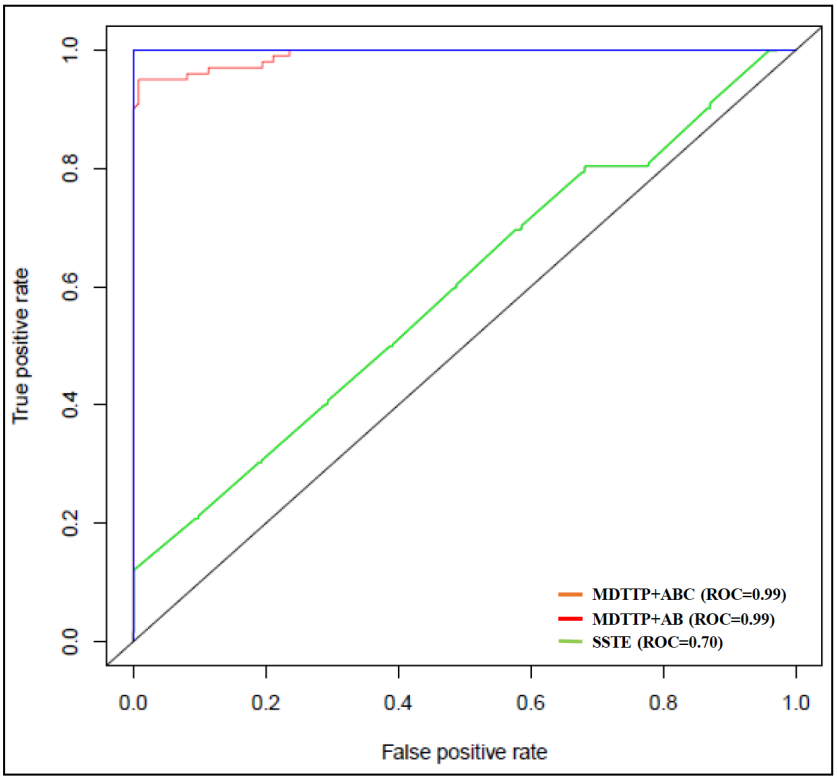

Fig. (2). ROC curve demonstrating the performance of three best predictive models: ROC curve demonstrating the performance of the top three predictive models MDTTP+A+B+C (ROC: 0.99), MDTTP+A+B (ROC: 0.99) and SSTE/ Triplet Elements (ROC: 0.70) respectively [x-axis and y-axis represents true positive rate (*tpr*) and false positive rate (*fpr*)]. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

Table 3. Comparison of piRNAPred with current state-of-the-art piRNA prediction methods.

S. No.	Tool	Description of Features	Species from which piRNAs Taken	Sn (%)	Sp (%)	ACC (%)	MCC	References
1	piRNA	k-MF (k=5, 1364)	*Homo sapiens, Mus musculus, Drosophila melanogaster, Caenorhabditis elegans*	72.47	95.5	NA	NA	(Zhang *et al.*, 2011)
2	PIANO	SSTE	*D. melanogaster*	95.89	94.6	95.27	NA	(Wang *et al.*, 2014)
3	Pibomd	SSP	*M. musculus*	91.48	89.8	90.62	NA	(Liu *et al.*, 2014)
4	Accurate piRNA prediction	k-MF , MMP, SSP, PSSM, pseDNC, SSTE	*H. sapiens, M. musculus , D. melanogaster*	83.10	82.10	82.6	0.651	(Luo *et al.*, 2016)
5	GA-WE	k-MF, PCPseDNC, PSSM		90.6	78.3	84.4	0.694	(Li *et al.*, 2016)
6	2L-piRNA	pseDNC, C	*M. musculus*	88.3	83.9	86.1	0.723	(Liu *et al.*, 2017)
7	piRNApred	k-MNC, SSTE, B and C	*H. sapiens, M. musculus , D. melanogaster , C. elegans Danio rerio, Gallus gallus domesticus , Xenopus tropicalus, Bombyx mori*	98.57	98.6	98.6	0.97	Algorithm proposed in the current study

Abbreviations: *Acc*, Accuracy; *k-MF*, k-Mer Features OR *k-MNC*, k-Mer Nucleotide Composition OR SP, Spectrum Profile; *MMP*, Mismatch Profile; *MCC*, Mathews Correlation Coefficient, *pseDNC*, pseudo Dinucleotide Composition; *PCPseDNC*, Parallel Correlation Pseudo Dinucleotide Composition; *PSSM*, Position-Specific Scoring Matrix; *SSTE*, Structure-Sequence Triplet Elements; *SSP*, Subsequence Profile; *Sn*, Sensitivity; *Sp*, Specificity; *SSTE*, Structure-Sequence Triplet Elements (A); *B*, Thermo-dynamic energies of contiguous dinucleotides; *C*, RNA physicochemical properties of adjoining dinucleotides (diNTs).

from different species. Integration of the *k*-mer spectrum profile in the algorithm was found to better distinguish the real piRNAs than the pseudo ones, and these *k*-mer nucleotide strings escalated to hybrid models by combining different *k*-mer nucleotide combinations [33]. Zhang *et al.*, proposed a *k*-mer approach to identify piRNAs from five organisms *i.e.* rat, mouse, human, fruit fly and nematode [27]. Further, all the ncRNAs have a unique secondary structure, which can be computed by SSTE [31]. This approach was utilized to classify piRNAs in *D. melanogaster* [30]. However, this tool had a homogeneous training dataset, and this limited the discovery of novel piRNAs in the other non-related organisms [54]. 'Pibomd' extended the use of sequence-based motifs in the prediction of *M. musculus* piRNAs [32]. However, a comprehensive dataset was needed to capture all the possible sequence-based motifs and predict piRNAs from various evolutionary diverse organisms. Other algorithms like GA-WE [34] and 2L-piRNA [35] used pseudo dinucleotide composition (pseKNC), PSSM, mismatch profiles, *etc*., and achieved high accuracy of the classifier. However, the existing piRNA prediction algorithms did not combine various other features like secondary structure, thermodynamic energy and physicochemical properties of RNA, and did not utilize piRNA sequences from phylogenetically diverse organisms, which we considered crucial for the prediction of piRNAs with high accuracy.

In the present study, we collected 1684 positive piRNA sequences from eight different species (Table **S1**). The length range of 1684 piRNAs was 24-33 nucleotides, and they were statistically distributed in 21 followed by 24, 29 and 31 nucleotides. To compare piRNAs and non-piRNAs, we investigated the composition-specific properties in detail and calculated the composition up to 5^{th} order of nucleotides that resulted in 1364 vector space. The accuracy of the classifier increased from k=1 to 5, suggesting that a combination of higher-order nucleotide usage (5-mer in our case) would better distinguish piRNAs from the non-piRNA sequences (Table **1**). Apart from training the classifier on the differential nucleotide usage by incorporating *k*-MNC, we also trained our classifier on the relative position-specific occupancy of nucleotides at 1^{st} and 10^{th} position in the primary and secondary piRNAs, respectively, which resulted into a feature space of 16. However, their accuracy (57.12%) suggested that these features alone were not enough to distinguish the piRNAs from the non-piRNA sequences. Further, the high accuracy of the triplet element features suggested that sequence and secondary structure together could be a potent and discriminative property to segregate piRNAs from the pseudo sequences of similar length. However, there was moderate accuracy in the classifier trained on the properties based on Gibbs free energy and physicochemical values of dinucleotides. Importantly, the hybrid model *piRNAPred* which incorporated the *k*-mer spectrum profile, SSTE, thermodynamic properties of continuous dinucleotides and physicochemical features, demonstrated the highest accuracy among all the models with an MCC of 0.97.

CONCLUSION AND FUTURE IMPLICATIONS

In conclusion, the *piRNAPred* demonstrated the highest accuracy in predicting piRNAs as compared to the existing piRNA prediction tools. We hope it would be helpful in expanding our current understanding of piRNA biology by predicting novel piRNAs in different organisms. In the future, we will update more piRNAs from the other species, and develop a user-friendly interface. The scripts, predictive models, datasets and other supplementary material are provided on https://github.com/IshaMonga/piRNAPred.

LIST OF ABBREVIATIONS

Acc	=	Accuracy
diNT	=	Dinucleotide
c	=	Regularization parameter
g	=	Gamma (a kernel density parameter)
k-MF	=	k-Mer Features OR k-MNC, k-mer nucleotide composition
MMP	=	Mismatch Profile
MCC	=	Mathews Correlation Coefficient
pseDNC	=	Pseudo Dinucleotide Composition
PCPseDNC	=	Parallel Correlation Pseudo Dinucleotide Composition
PSSM	=	Position-Specific Scoring Matrix
SSTE	=	Structure-Sequence Triplet Elements
SSP	=	Subsequence Profile
Sn	=	Sensitivity
Sp	=	Specificity

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable.

HUMAN AND ANIMAL RIGHTS

No animals/humans were used for studies that are the basis of this research.

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

All materials including computational workflow, predictive models, scripts and datasets utilized in this work are provided on GitHub repository at: https://github.com/Isha Monga/piRNAPred.

FUNDING

We acknowledge the Indian Institute of Science Education and Research, Mohali (IISER Mohali) for providing the financial support for this work.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

[1] Carmell, M.A.; Xuan, Z.; Zhang, M.Q.; Hannon, G.J. The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes Dev.,* **2002**, *16*(21), 2733-2742. [http://dx.doi.org/10.1101/gad.1026102] [PMID: 12414724]

[2] Thomson, T.; Lin, H. The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. *Annu. Rev. Cell Dev. Biol.,* **2009**, *25*, 355-376. [http://dx.doi.org/10.1146/annurev.cellbio.24.110707.175327] [PMID: 19575643]

[3] Kawamata, T.; Tomari, Y. Making RISC. *Trends Biochem. Sci.,* **2010**, *35*(7), 368-376. [http://dx.doi.org/10.1016/j.tibs.2010.03.009] [PMID: 20395147]

[4] Joshua-Tor, L. The Argonautes. *Cold Spring Harb. Symp. Quant. Biol.,* **2006**, *71*, 67-72. [http://dx.doi.org/10.1101/sqb.2006.71.048] [PMID: 17381282]

[5] Cox, D.N.; Chao, A.; Baker, J.; Chang, L.; Qiao, D.; Lin, H. A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. *Genes Dev.,* **1998**, *12*(23), 3715-3727. [http://dx.doi.org/10.1101/gad.12.23.3715] [PMID: 9851978]

[6] Meister, G.; Landthaler, M.; Patkaniowska, A.; Dorsett, Y.; Teng, G.; Tuschl, T. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol. Cell,* **2004**, *15*(2), 185-197. [http://dx.doi.org/10.1016/j.molcel.2004.07.007] [PMID: 15260970]

[7] Czech, B.; Hannon, G.J. One loop to rule them all: the ping-pong cycle and piRNA-guided silencing. *Trends Biochem. Sci.,* **2016**, *41*(4), 324-337. [http://dx.doi.org/10.1016/j.tibs.2015.12.008] [PMID: 26810602]

[8] Czech, B.; Munafò, M.; Ciabrelli, F.; Eastwood, E.L.; Fabry, M.H.; Kneuss, E.; Hannon, G.J. piRNA-guided genome defense: from biogenesis to silencing. *Annu. Rev. Genet.,* **2018**, *52*, 131-157. [http://dx.doi.org/10.1146/annurev-genet-120417-031441] [PMID: 30476449]

[9] Aravin, A.A.; Lagos-Quintana, M.; Yalcin, A.; Zavolan, M.; Marks, D.; Snyder, B.; Gaasterland, T.; Meyer, J.; Tuschl, T. The small RNA profile during *Drosophila melanogaster* development. *Dev. Cell,* **2003**, *5*(2), 337-350. [http://dx.doi.org/10.1016/S1534-5807(03)00228-4] [PMID: 12919683]

[10] Siomi, M.C.; Sato, K.; Pezic, D.; Aravin, A.A. PIWI-interacting small RNAs: the vanguard of genome defence. *Nat. Rev. Mol. Cell Biol.,* **2011**, *12*(4), 246-258. [http://dx.doi.org/10.1038/nrm3089] [PMID: 21427766]

[11] Kotelnikov, R.N.; Klenov, M.S.; Rozovsky, Y.M.; Olenina, L.V.; Kibanov, M.V.; Gvozdev, V.A. Peculiarities of piRNA-mediated post-transcriptional silencing of Stellate repeats in testes of *Drosophila melanogaster*. *Nucleic Acids Res.,* **2009**, *37*(10), 3254-3263. [http://dx.doi.org/10.1093/nar/gkp167] [PMID: 19321499]

[12] Tiwari, B.; Kurtz, P.; Jones, A.E.; Wylie, A.; Amatruda, J.F.; Boggupalli, D.P.; Gonsalvez, G.B.; Abrams, J.M. Retrotransposons mimic germ plasm determinants to promote transgenerational inheritance. *Curr. Biol.,* **2017**, *27*(19), 3010-3016.e3. [http://dx.doi.org/10.1016/j.cub.2017.08.036] [PMID: 28966088]

[13] Ishizu, H.; Siomi, H.; Siomi, M.C. Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines. *Genes Dev.,* **2012**, *26*(21), 2361-2373. [http://dx.doi.org/10.1101/gad.203786.112] [PMID: 23124062]

[14] Brennecke, J.; Aravin, A.A.; Stark, A.; Dus, M.; Kellis, M.; Sachidanandam, R.; Hannon, G.J. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell,* **2007**, *128*(6), 1089-1103. [http://dx.doi.org/10.1016/j.cell.2007.01.043] [PMID: 17346786]

[15] Ding, D.; Liu, J.; Dong, K.; Midic, U.; Hess, R.A.; Xie, H.; Demireva, E.Y.; Chen, C. PNLDC1 is essential for piRNA 3′ end trimming and transposon silencing during spermatogenesis in mice. *Nat. Commun.,* **2017**, *8*(1), 819. [http://dx.doi.org/10.1038/s41467-017-00854-4] [PMID: 29018194]

[16] Ipsaro, J.J.; Haase, A.D.; Knott, S.R.; Joshua-Tor, L.; Hannon, G.J. The structural biochemistry of Zucchini implicates it as a nuclease in piRNA biogenesis. *Nature,* **2012**, *491*(7423), 279-283. [http://dx.doi.org/10.1038/nature11502] [PMID: 23064227]

[17] Kawaoka, S.; Izumi, N.; Katsuma, S.; Tomari, Y. 3′ end formation of PIWI-interacting RNAs *in vitro*. *Mol. Cell,* **2011**, *43*(6), 1015-1022. [http://dx.doi.org/10.1016/j.molcel.2011.07.029] [PMID: 21925389]

[18] Nishida, K.M.; Saito, K.; Mori, T.; Kawamura, Y.; Nagami-Okada, T.; Inagaki, S.; Siomi, H.; Siomi, M.C. Gene silencing mechanisms mediated by Aubergine piRNA complexes in Drosophila male gonad. *RNA,* **2007**, *13*(11), 1911-1922. [http://dx.doi.org/10.1261/rna.744307] [PMID: 17872506]

[19] Horwich, M.D.; Li, C.; Matranga, C.; Vagin, V.; Farley, G.; Wang, P.; Zamore, P.D. The Drosophila RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. *Curr. Biol.,* **2007**, *17*(14), 1265-1272. [http://dx.doi.org/10.1016/j.cub.2007.06.030] [PMID: 17604629]

[20] Gainetdinov, I.; Colpan, C.; Arif, A.; Cecchini, K.; Zamore, P.D. A single mechanism of biogenesis, initiated and directed by PIWI proteins, explains pirna production in most animals. *Mol. Cell,* **2018**, *71*(5), 775-790.e5. [http://dx.doi.org/10.1016/j.molcel.2018.08.007] [PMID: 30193099]

[21] Mohn, F.; Handler, D.; Brennecke, J. Noncoding RNA. piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. *Science,* **2015**, *348*(6236), 812-817. [http://dx.doi.org/10.1126/science.aaa1039] [PMID: 25977553]

[22] Homolka, D.; Pandey, R.R.; Goriaux, C.; Brasset, E.; Vaury, C.; Sachidanandam, R.; Fauvarque, M-O.; Pillai, R.S. PIWI slicing and rna elements in precursors instruct directional primary piRNA biogenesis. *Cell Rep.,* **2015**, *12*(3), 418-428. [http://dx.doi.org/10.1016/j.celrep.2015.06.030] [PMID: 26166577]

[23] Han, B.W.; Wang, W.; Li, C.; Weng, Z.; Zamore, P.D. Noncoding RNA. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. *Science,* **2015**, *348*(6236), 817-821. [http://dx.doi.org/10.1126/science.aaa1264] [PMID: 25977554]

[24] Ozata, D.M.; Gainetdinov, I.; Zoch, A.; O'Carroll, D.; Zamore, P.D. PIWI-interacting RNAs: small RNAs with big functions. *Nat. Rev. Genet.,* **2019**, *20*(2), 89-108. [http://dx.doi.org/10.1038/s41576-018-0073-3] [PMID: 30446728]

[25] Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.,* **1990**, *215*(3), 403-410. [http://dx.doi.org/10.1016/S0022-2836(05)80360-2] [PMID: 2231712]

[26] Bailey, T.L.; Boden, M.; Buske, F.A.; Frith, M.; Grant, C.E.; Clementi, L.; Ren, J.; Li, W.W.; Noble, W.S. MEME suite: tools for motif discovery and searching. *Nucleic Acids Res,* **2009**, *37*, W202. [http://dx.doi.org/https://doi.org/10.1093/nar/gkp335]

[27] Zhang, Y.; Wang, X.; Kang, L. A k-mer scheme to predict piRNAs and characterize locust piRNAs. *Bioinformatics,* **2011**, *27*(6), 771-776. [http://dx.doi.org/10.1093/bioinformatics/btr016] [PMID: 21224287]

[28] Wang, J.; Zhang, P.; Lu, Y.; Li, Y.; Zheng, Y.; Kan, Y.; Chen, R.; He, S. PiRBase: A comprehensive database of PiRNA sequences. *Nucleic Acids Res.,* **2018**. [http://dx.doi.org/10.1093/nar/gky1043] [PMID: 30371818]

[29] Betel, D.; Sheridan, R.; Marks, D.S.; Sander, C. Computational analysis of mouse piRNA sequence and biogenesis. *PLOS Comput. Biol.,* **2007**, *3*(11), e222. [http://dx.doi.org/10.1371/journal.pcbi.0030222] [PMID: 17997596]

[30] Wang, K.; Liang, C.; Liu, J.; Xiao, H.; Huang, S.; Xu, J.; Li, F. Prediction of piRNAs using transposon interaction and a support vector machine. *BMC Bioinformatics,* **2014**, *15*, 419. [http://dx.doi.org/10.1186/s12859-014-0419-6] [PMID: 25547961]

[31] Xue, C.; Li, F.; He, T.; Liu, G-P.; Li, Y.; Zhang, X. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics,* **2005**, *6*, 310. [http://dx.doi.org/10.1186/1471-2105-6-310] [PMID: 16381612]

[32] Liu, X.; Ding, J.; Gong, F. piRNA identification based on motif discovery. *Mol. Biosyst.,* **2014**, *10*(12), 3075-3080. [http://dx.doi.org/10.1039/C4MB00447G] [PMID: 25230731]

[33] Luo, L.; Li, D.; Zhang, W.; Tu, S.; Zhu, X.; Tian, G. Accurate prediction of transposon-derived piRNAs by integrating various

sequential and physicochemical features. *PLoS One,* **2016**, *11*(4), e0153268.
[http://dx.doi.org/10.1371/journal.pone.0153268] [PMID: 27074043]

[34] Li, D.; Luo, L.; Zhang, W.; Liu, F.; Luo, F. A genetic algorithm-based weighted ensemble method for predicting transposon-derived piRNAs. *BMC Bioinformatics,* **2016**, *17*(1), 329.
[http://dx.doi.org/10.1186/s12859-016-1206-3] [PMID: 27578422]

[35] Liu, B.; Yang, F.; Chou, K-C. 2L-piRNA: A two-layer ensemble classifier for identifying piwi-interacting RNAs and their function. *Mol. Ther. Nucleic Acids,* **2017**, *7*, 267-277.
[http://dx.doi.org/10.1016/j.omtn.2017.04.008] [PMID: 28624202]

[36] Bu, D.; Yu, K.; Sun, S.; Xie, C.; Skogerbø, G.; Miao, R.; Xiao, H.; Liao, Q.; Luo, H.; Zhao, G.; Zhao, H.; Liu, Z.; Liu, C.; Chen, R.; Zhao, Y. NONCODE v3.0: integrative annotation of long non-coding RNAs. *Nucleic Acids Res.,* **2012**, *40*(Database issue), D210-D215.
[http://dx.doi.org/10.1093/nar/gkr1175] [PMID: 22135294]

[37] Reuter, M.; Berninger, P.; Chuma, S.; Shah, H.; Hosokawa, M.; Funaya, C.; Antony, C.; Sachidanandam, R.; Pillai, R.S. Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. *Nature,* **2011**, *480*(7376), 264-267.
[http://dx.doi.org/10.1038/nature10672] [PMID: 22121019]

[38] Monga, I.; Qureshi, A.; Thakur, N.; Gupta, A.K.; Kumar, M. ASPsiRNA: A resource of ASP-siRNAs having therapeutic potential for human genetic disorders and algorithm for prediction of their inhibitory efficacy. *G3 (Bethesda),* **2017**, *7*(9), 2931-2943.
[http://dx.doi.org/10.1534/g3.117.044024] [PMID: 28696921]

[39] Qureshi, A.; Thakur, N.; Monga, I.; Thakur, A.; Kumar, M. VIRmiRNA: a comprehensive resource for experimentally validated viral miRNAs and their targets. *Database (Oxford),* **2014**, *2014*, bau103-bau103.
[http://dx.doi.org/10.1093/database/bau103] [PMID: 25380780]

[40] Lorenz, R.; Bernhart, S.H.; Höner Zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P.F.; Hofacker, I.L. ViennaRNA Package 2.0. *Algorithms Mol. Biol.,* **2011**, *6*, 26.
[http://dx.doi.org/10.1186/1748-7188-6-26] [PMID: 22115189]

[41] Khvorova, A.; Reynolds, A.; Jayasena, S.D. Functional siRNAs and miRNAs exhibit strand bias. *Cell,* **2003**, *115*(2), 209-216.
[http://dx.doi.org/10.1016/S0092-8674(03)00801-8] [PMID: 14567918]

[42] Qureshi, A.; Thakur, N.; Kumar, M. VIRsiRNApred: a web server for predicting inhibition efficacy of siRNAs targeting human viruses. *J. Transl. Med.,* **2013**, *11*, 305.
[http://dx.doi.org/10.1186/1479-5876-11-305] [PMID: 24330765]

[43] Shabalina, S.A.; Spiridonov, A.N.; Ogurtsov, A.Y. Computational models with thermodynamic and composition features improve siRNA design. *BMC Bioinformatics,* **2006**, *7*, 65.
[http://dx.doi.org/10.1186/1471-2105-7-65] [PMID: 16472402]

[44] Vapnik, V.N. *The Nature of Statistical Learning Theory*; Springer-Verlag: Berlin, Heidelberg, **1995**.
[http://dx.doi.org/10.1007/978-1-4757-2440-0]

[45] Frank, E.; Hall, M.; Trigg, L.; Holmes, G.; Witten, I.H. Data mining in bioinformatics using Weka. *Bioinformatics,* **2004**, *20*(15), 2479-2481.
[http://dx.doi.org/10.1093/bioinformatics/bth261] [PMID: 15073010]

[46] Ahmed, F.; Raghava, G.P.S. Designing of highly effective complementary and mismatch siRNAs for silencing a gene. *PLoS One,* **2011**, *6*(8), e23443.
[http://dx.doi.org/10.1371/journal.pone.0023443] [PMID: 21853133]

[47] Kim, V.N.; Han, J.; Siomi, M.C. Biogenesis of small RNAs in animals. *Nat. Rev. Mol. Cell Biol.,* **2009**, *10*(2), 126-139.
[http://dx.doi.org/10.1038/nrm2632] [PMID: 19165215]

[48] Aravin, A.; Gaidatzis, D.; Pfeffer, S.; Lagos-Quintana, M.; Landgraf, P.; Iovino, N.; Morris, P.; Brownstein, M.J.; Kuramochi-Miyagawa, S.; Nakano, T.; Chien, M.; Russo, J.J.; Ju, J.; Sheridan, R.; Sander, C.; Zavolan, M.; Tuschl, T. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature,* **2006**, *442*(7099), 203-207.
[http://dx.doi.org/10.1038/nature04916] [PMID: 16751777]

[49] Girard, A.; Sachidanandam, R.; Hannon, G.J.; Carmell, M.A. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature,* **2006**, *442*(7099), 199-202.
[http://dx.doi.org/10.1038/nature04917] [PMID: 16751776]

[50] Grivna, S.T.; Beyret, E.; Wang, Z.; Lin, H. A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev.,* **2006**, *20*(13), 1709-1714.
[http://dx.doi.org/10.1101/gad.1434406] [PMID: 16766680]

[51] Aravin, A.A.; Klenov, M.S.; Vagin, V.V.; Bantignies, F.; Cavalli, G.; Gvozdev, V.A. Dissection of a natural RNA silencing process in the *Drosophila melanogaster* germ line. *Mol. Cell. Biol.,* **2004**, *24*(15), 6742-6750.
[http://dx.doi.org/10.1128/MCB.24.15.6742-6750.2004] [PMID: 15254241]

[52] Barckmann, B.; Pierson, S.; Dufourt, J.; Papin, C.; Armenise, C.; Port, F.; Grentzinger, T.; Chambeyron, S.; Baronian, G.; Desvignes, J-P.; Curk, T.; Simonelig, M. Aubergine iCLIP reveals piRNA-dependent decay of mRNAs involved in germ cell development in the early embryo. *Cell Rep.,* **2015**, *12*(7), 1205-1216.
[http://dx.doi.org/10.1016/j.celrep.2015.07.030] [PMID: 26257181]

[53] Vourekas, A.; Alexiou, P.; Vrettos, N.; Maragkakis, M.; Mourelatos, Z. Sequence-dependent but not sequence-specific piRNA adhesion traps mRNAs to the germ plasm. *Nature,* **2016**, *531*(7594), 390-394.
[http://dx.doi.org/10.1038/nature17150] [PMID: 26950602]

[54] Sai Lakshmi, S.; Agrawal, S. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res.,* **2008**, *36*(Database issue), D173-D177.
[http://dx.doi.org/10.1093/nar/gkm696] [PMID: 17881367]