

RESEARCH PAPER



Predicting functional long non-coding RNAs validated by low throughput experiments

Bailing Zhou^{a,b}, Yuedong Yang^{ib a,c,d}, Jian Zhan^d, Xianghua Dou^{a,b}, Jihua Wang^{a,b}, and Yaoqi Zhou^{ib a,d}

^aShandong Provincial Key Laboratory of Biophysics, Institute of Biophysics, Dezhou University, Dezhou, China; ^bCollege of Physics and Electronic Information, Dezhou University, Dezhou, China; ^cSchool of Data and Computer Science, Sun Yat-sen University, Guangzhou, China; ^dInstitute for Glycomics and School of Information and Communication Technology, Griffith University, Gold Coast, QLD, Australia

ABSTRACT

High-throughput techniques have uncovered hundreds and thousands of long non-coding RNAs (lncRNAs). Among them, only a tiny fraction has experimentally validated functions (EVlncRNAs) by low-throughput methods. What fraction of lncRNAs from high-throughput experiments (HTlncRNAs) is truly functional is an active subject of debate. Here, we developed the first method to distinguish EVlncRNAs from HTlncRNAs and mRNAs by using Support Vector Machines and found that EVlncRNAs can be well separated from HTlncRNAs and mRNAs with 0.6 for Matthews correlation coefficient, 64% for sensitivity, and 81% for precision for the independent human test set. The most useful features for classification are related to sequence conservations at RNA (for separating from HTlncRNAs) and protein (for separating from mRNA) levels. The method is found to be robust as the human-RNA-trained model is applicable to independent mouse RNAs with similar accuracy and to a lesser extent to plant RNAs. The method can recover newly discovered EVlncRNAs with high sensitivity. Its application to randomly selected 2000 human HTlncRNAs indicates that the majority of HTlncRNAs is probably non-functional but a large portion (nearly 30%) are likely functional. In other words, there is an ample number of lncRNAs whose specific biological roles are yet to be discovered. The method developed here is expected to speed up and reduce the cost of the discovery by prioritizing potentially functional lncRNAs prior to experimental validation. EVlncRNA-pred is available as a web server at <http://biophy.dzu.edu.cn/lncnapred/index.html>. All datasets used in this study can be obtained from the same website.

ARTICLE HISTORY

Received 8 November 2018
Revised 17 June 2019
Accepted 10 July 2019

KEYWORDS

Long non-coding RNAs; low throughput experiments; prediction; functional lncRNAs





Introduction


Advances in high-throughput sequencing and microarray technologies showed that most of the human genome transcribe into RNAs despite they were not coded for proteins [1–3]. Among these non-coding RNAs (ncRNAs), long transcripts (>200 nucleotides) with unknown functions were found prevalent with low expression, highly tissue-specific, and lack of strong cross-species conservation [4–6]. Some long ncRNAs (lncRNAs) have been confirmed as functional and disease-relevant using traditional low throughput techniques such as qRT-PCR, knockdown, Western blot, Northern blot, and luciferase reporter assays [7–11]. So far, more than 1000 lncRNAs in >70 species were experimentally validated and collected in a number of databases (lncRNADisease, lncRInter, lncRNAdb, PLNlncRBase) [12–15]. These databases were integrated into the comprehensive EVlncRNAs database [16], which collected all known EVlncRNAs up to May 2016 from 77 species.

The experimentally validated lncRNAs (EVlncRNAs), however, are only a tiny fraction of all transcribed ones. What percentage of transcribed lncRNAs is functional remains a subject of active debate [17]. It is known that some lncRNAs can be expressed due to a lack of fidelity in

transcription initiation by RNA polymerase II [18]. Hon et al [19]. found that 69% of 27,919 FANTOM CAT lncRNAs overlap with trait-associated single nucleotide polymorphisms. However, the overlap could be due to their genomic positions, rather than intrinsic functions coded in sequences [20]. Nevertheless, 31% of lncRNAs remain unaccounted for. Liu et al [21], on the other hand, found that only 3% (499/16,401) lncRNA loci are essential for robust cell growth, based on a large-scale knockdown using a CRISPR interference technique. More importantly, analysis of mutational loads suggests that ‘the functional fraction within the human genome cannot exceed 25% and is probably considerably lower’ [22]. Thus, a significant portion of transcribed lncRNAs is possibly non-functional.

Existence of non-functional but transcribed lncRNAs calls for computational methods to prioritize potentially functional lncRNAs prior to expensive and laborious experimental validations. Current computational tools on identifying lncRNAs have been focused on distinguishing expressed lncRNAs from coding RNAs [23], a challenging problem as some lncRNAs were coded for short peptides while others such as H19, Xist, Mirg, and Gtl2 have predicted coding regions of longer than 100 amino acids [24,25]. One approach for lncRNA

CONTACT Yaoqi Zhou  yaoqi.zhou@griffith.edu.au  Institute for Glycomics and School of Information and Communication Technology, Griffith University, Gold Coast, QLD 4222, Australia; Jihua Wang  jhw25336@126.com  Shandong Provincial Key Laboratory of Biophysics, Institute of Biophysics, Dezhou University, Dezhou 253023, China

 Supplemental data for this article can be accessed [here](#)

identification is cross-species comparison. Examples are CRITICA [26], PhyloCSF [27], CSTMiner [28] and RNAcode [29]. Because the majority of lncRNAs are not conserved across different species [30], many methods estimated coding potentials of a sequence by using a wide variety of features and machine-learning techniques. Examples are CPAT by logistic regression model [31], lncRNA-ID [32], lncRNAPred [33], FEELnc [34] and COME [35] based on random forest and PORTRAIT [36], CNCI [37], CPC [38], PLEK [39] and lncRScan-SVM [40] based on support vector machines. Using high-throughput experimental data has proven useful for further improving the accuracy of separating lncRNAs from mRNAs [35,41–43].

In this study, we address the question of whether or not lncRNAs experimentally validated (EVlncRNAs) by low throughput techniques are distinguishable from those lncRNAs obtained from high throughput experiments (HTlncRNAs). By using support vector machines and employing sequence-derived and HT experimental features in combination or separately, we showed that experimentally validated lncRNAs are identifiable from HTlncRNAs and mRNAs with reasonable accuracy. Moreover, a method trained and tested from human datasets is applicable to mouse RNAs with similar accuracy and to plant RNAs with somewhat lower accuracy. This indicates the robustness of the method developed for locating functional lncRNAs. The online server of EVlncRNA-pred and the datasets are freely available at <http://biophy.dzu.edu.cn/lncrnaped/index.html>.

Results

Model performance for the full-feature model

Using positive samples collected in the EVlncRNAs dataset [16] and negative samples from lncRNAs and mRNAs from GENCODE [44] (see Methods), we built the training set from human RNAs and independent test sets from human, mouse and plant RNAs. Table 1 shows the results of the human 10-fold cross-validation and independent test by the support vector machines (SVM) model with 33 features (the full-feature model). The corresponding receiver operating characteristic (ROC) curves are shown in Figure 1. The results indicate that the model performs better on the test set (Matthews correlation coefficient, MCC, at 0.60 compared to 0.51, the area under the ROC curve, AUC, at 0.88 compared to 0.84). This is likely due to a slightly larger training set (799 positive samples) than in the cross-validation (719 positive samples as the remaining 80 of the 799 samples (10%) was

used as one-fold for 10-fold cross validation). When the model is further applied to mouse RNAs, there is a performance drop to the performance level similar to ten-fold cross-validation. It should be noted that the lower precision for the mouse test set is due to higher sensitivity as the threshold was set by the 10-fold cross-validation. If one adjusts the threshold to the sensitivity of 0.64, one would obtain a precision of 0.74. Nevertheless, the low standard deviation in ten-fold cross-validation performance over 100 randomly selected ten folds and the consistently high, cross-species performance (AUC > 0.84) in independent tests confirm the overall quality and robustness of the model developed.

Model performance for the sequence-only model

The above model employed some high-throughput experimental results including expression abundance and histone modification. However, these experimental data are not always available. Thus, we also built a model that requires the input of a sequence only. Table 1 and Figure 1 also present the results from the sequence-only model. The model performance is much more similar among the ten-fold cross-validation and two independent tests (human and mouse test sets) with MCC = 0.47, 0.51, 0.48 and AUC = 0.84, 0.85, and 0.85, respectively. This overall performance is slightly worse than the case when experimental data were employed, confirming the usefulness of expression abundance and histone modification in EVlncRNA prediction. On the other hand, if these experimental results are not available, the sequence-only model yields adequate accuracy in separating EVlncRNAs from HTlncRNAs and mRNAs as shown in Figure 1 and Table 1.

Model performance for the plant test set

The ability of the human-RNA trained model to predict mouse lncRNA indicates inherently similar characteristics of functional lncRNAs in human and mouse. It is of interest to know if plant EVlncRNAs can also be detected in a similar accuracy. In human and mouse, we used phastCons [45] scores provided by the UCSC [46] to represent the DNA sequence conservation. However, UCSC does not have phastCons scores of *Arabidopsis thaliana*. Thus, we re-trained all models without using DNA conservation scores.

Table 2 and Figure 2 compare the performance of the new full-feature and sequence-only models without DNA

Table 1. Performance of full-feature and sequence-only SVM models trained on human datasets.

		MCC	AUC	Accuracy	Sensitivity	Specificity	Precision
Full-feature model							
Human	CV ^a	0.513 ± 0.006	0.841 ± 0.006	0.791 ± 0.003	0.599 ± 0.011	0.887 ± 0.005	0.728 ± 0.005
	Test ^b	0.603	0.879	0.829	0.641	0.923	0.806
Mouse	Test ^c	0.512	0.859	0.765	0.777	0.759	0.617
Sequence-only model							
Human	CV ^a	0.471	0.841	0.774	0.551	0.887	0.709
	Test ^b	0.514	0.852	0.792	0.590	0.893	0.734
Mouse	Test ^c	0.481	0.846	0.745	0.777	0.729	0.589

^a10-fold cross-validation on the full training set. The mean and standard deviation are obtained from 100 random divisions of 10 folds in the training set. ^bTest results on the independent test set of the human RNAs. ^cTest results on the independent test set of the mouse RNAs.

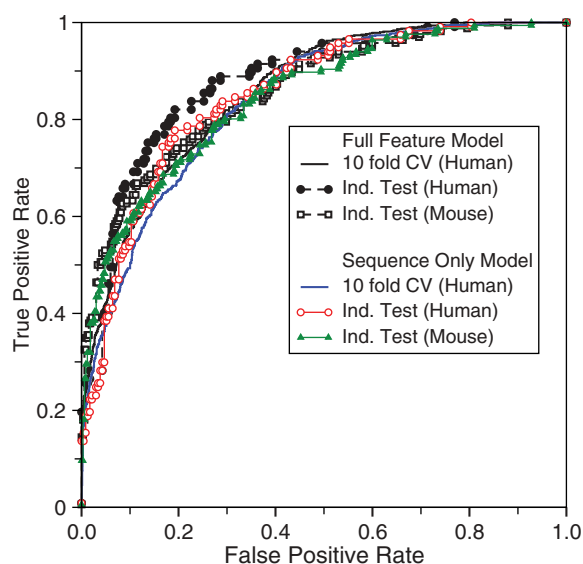


Figure 1. Receiver operating characteristic curves by full-feature and sequence-only models trained on human RNAs.

Table 2. Performance of full-feature and sequence-only SVM models (except DNA conservation scores) trained on human datasets.

	MCC	AUC	Accuracy	Sensitivity	Specificity	Precision
Full-feature model except DNA conservation						
Human-CV ^a	0.482	0.845	0.781	0.514	0.915	0.753
Human-Test ^b	0.560	0.869	0.812	0.581	0.927	0.800
Mouse-Test ^c	0.518	0.844	0.779	0.723	0.807	0.652
Plant-Test ^d	0.383	0.801	0.744	0.417	0.908	0.694
Sequence-only model except DNA conservation						
Human-CV ^a	0.448	0.833	0.763	0.562	0.864	0.675
Human-Test ^b	0.521	0.849	0.792	0.632	0.872	0.712
Mouse-Test ^c	0.414	0.818	0.719	0.711	0.723	0.562
Plant-Test ^d	0.221	0.725	0.689	0.283	0.892	0.567

^a10-fold cross-validation on the full training set. ^bTest results on the independent test set of the human RNAs. ^cTest results on the independent test set of the mouse RNAs. ^dTest results on the independent test set of the plant RNAs.

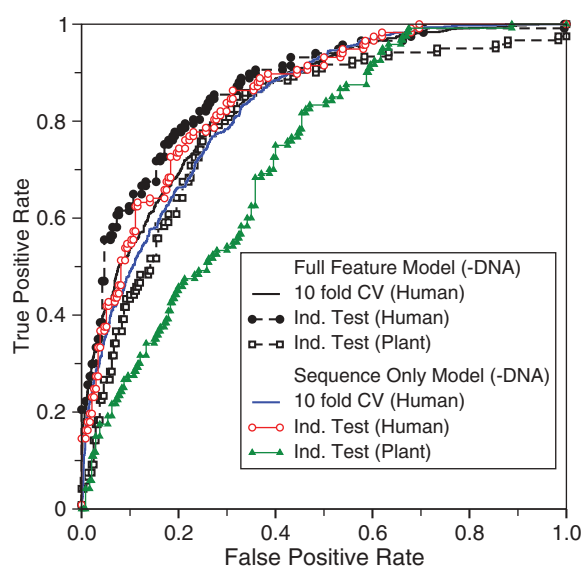


Figure 2. As in Figure 1 but for the model without DNA conservation features and tested by the plant RNAs.

conservation in 10-fold cross-validation and independent tests on the human, mouse, and plant RNAs by models trained on human RNAs. The results show that there is a large drop in performance when the human-RNA-trained model is applied to plant with MCC decreasing from 0.56 to 0.38 and AUC decreasing from 0.87 to 0.80 for the full feature model without DNA conservation. By comparison, the corresponding changes from human to mouse are 0.56 to 0.52 for MCC and 0.87 to 0.84 for AUC. This suggests that the difference between human and plant lncRNAs is larger than the difference between human and mouse lncRNAs. Nevertheless, the performance of applying the human-RNA trained model to plant remains high with AUC = 0.80 for the full-feature model and 0.73 for the sequence-only model (all without DNA conservations). This confirms the robustness of the model trained by human RNAs and the existence of basic common characteristics of EVlncRNAs from plant to human. This also suggests that plant-specific training when sufficient data is available may be necessary to maximize the classification capability.

The importance of individual features

To examine the classification power of each feature for separating EVlncRNAs from HTlncRNAs and mRNAs, we obtained the best performing single feature according to 10-fold cross-validation and compare them in Figure 3. For single-feature performance, the performance is measured by the difference (Δ AUC) between the AUC by the model using a single feature only and the AUC by random prediction (0.5) or between the AUC by the full feature model and the AUC by the model after removing a single feature. Figure 3 shows that according to feature removal, protein conservation has the highest Δ AUC value from the full feature model at 0.031, followed by RNA conservation (Δ AUC = 0.020). The best experimental feature is the H3K4me3 modification with the Δ AUC value at 0.008. These changes in Δ AUC are small. We can also measure the changes in precision at a fixed value of

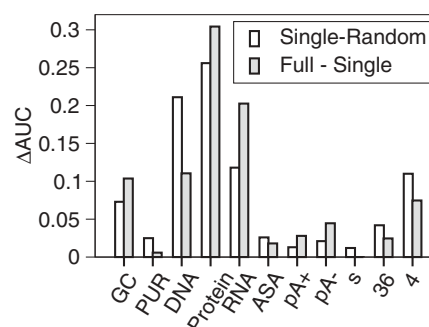


Figure 3. The difference in Area Under the ROC Curve (AUC) as a single feature. Here, GC denotes GC content; PUR: Purine content; DNA: DNA conservation; Protein: Protein conservation; RNA: RNA conservation; ASA: Accessible surface area; pA+: polyA+ RNA-seq; pA-: polyA- RNA-seq; s: small RNA-seq; 36: H3K36me3 modification; and 4: H3K4me3 modification. The difference is multiplied by 10 for removing a single feature only (filled bar), to facilitate comparison to the results of using a single feature (open bar).

sensitivity. The same trend is observed with two highest reductions of 10.3% and 8% in precision for removing protein and RNA conservation, respectively, and the largest reduction of 3.7% for removing the H3K4me3 modification in experimental features. This result is similar to the analysis of models based on a single feature only (Figure 3). All features contribute somewhat to lncRNA classifications except small RNA-seq experimental data.

The importance of protein conservation found in the above analysis is likely due to the presence of mRNAs in the negative samples. To further explore the features important for separating EVlncRNAs from HTlncRNAs only, we employed the training set (799 EVlncRNAs as the positive set and 799 HTlncRNAs without mRNAs as the negative set) and obtained the result of ten-fold cross-validation with the full feature model. Then, we removed the redundant single feature that led to the largest increase of the AUC value in each round until the AUC can no longer increase after feature removal. The final model eliminated three features (protein conservation, predicted RNA ASA, and purine content) that are not useful for distinguishing HTlncRNAs from EVlncRNAs. The most important remaining features according to the magnitude in AUC reduction after removal are RNA conservation ($\Delta\text{AUC} = 0.025$), followed by DNA conservation ($\Delta\text{AUC} = 0.007$) and the experimental feature of H3K36me3 modification ($\Delta\text{AUC} = 0.006$). The same trend is observed based on reduction in precision with a fixed sensitivity

Distinguishing from mRNAs

Our method was designed to separate EVlncRNAs from both HTlncRNAs and mRNAs as all of our negative sets in training and test sets contain 1:1 ratio of mRNA: HTlncRNAs. To further examine the capability of distinguishing from mRNAs, we built an additional test set by using the EVlncRNAs in our test set as the positive set and newly randomly selected mRNAs as an additional negative set. EVlncRNA-pred achieves a high AUC of 0.959, precision of 0.987, a specificity of 0.991, and the MCC value of 0.675. This higher performance in distinguishing from mRNA is consistent with our intuition that separating EVlncRNAs from mRNAs is easier than separating them from HTlncRNAs (see discussion).

Comparison with other methods

To the best of our knowledge, the method reported here is the first technique for separating EVlncRNAs from HTlncRNAs and mRNAs. Existing techniques for lncRNA prediction are dedicated to separate HTlncRNAs from mRNAs. We do not expect that these previous methods could be useful for identifying EVlncRNAs from HTlncRNAs and mRNAs. To confirm this, Figure 4 reported the applications of the logistic regression model CPAT [31], the random forest model COME [35], and support vector machines models CNCI [37] and PLEK [39] to the human test set. Indeed, CNCI, CPAT, and PLEK methods are close to random predictions at low false positive rates whereas COME is unable to make any positive prediction until false positive rates are greater than 0.2. Overall prediction

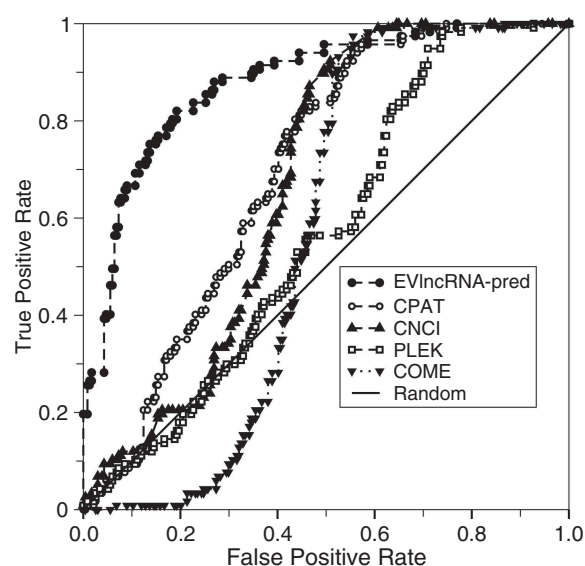


Figure 4. Receiver operating characteristic curves on the human test set by EVlncRNA-pred and several methods that were trained for separating expressed lncRNAs from mRNAs only.

of COME, CNCI, CPAT, and PLEK is better than random with AUCs ranging from 0.567, 0.672, 0.699, and 0.569, respectively. This is because mRNA belongs to the negative set whereas EVlncRNAs belongs to the positive set in training COME, CNCI, CPAT, and PLEK. We would like to emphasize that the comparison made in Figure 4 is not to illustrate the improvement of our method over previous techniques but to highlight the difference in the prediction goals.

Case studies

Tumour-specific EVlncRNAs

Tumour-specific EVlncRNAs is a large group in all known EVlncRNAs. In the EVlncRNA database, there were 446 and 72 tumour-specific lncRNAs in our training and test sets, respectively. The sensitivity of EVlncRNA-pred (the fraction of predicted EVlncRNAs in known EVlncRNAs) is 54% for these tumour-specific lncRNAs in the training set and 57% in the test set, which are close to a sensitivity of 60% in ten-fold cross-validation (Table 1), suggesting overall consistency of the method performance for specific types of EVlncRNAs.

CRISPRi-identified functional lncRNAs

Recently Liu et al. [21] developed a CRISPR interference technique for large-scale screening of lncRNA loci required for robust cell growth. Strictly speaking, the resulting 499 lncRNA loci discovered would require further validation by low-throughput experiments. However, it is of interest to examine the performance of EVlncRNA-pred for these newly discovered putatively functional lncRNAs. Among these 499 lncRNA loci, we located 194 lncRNAs with the gene structure information in the general transfer format, 59 lncRNAs of which are in the positive training set (known EVlncRNAs). Applying EVlncRNA-pred to the remaining 135 lncRNAs yields a sensitivity of 42%. It should be noted that the above putatively functional lncRNA loci were filtered

with a statistical confidence score (called ‘screen score’) >7 . This screen score was based on the effect size and the p-values of genes in each screening experiment, relative to the negative control gene distribution. If we increase this threshold from 7, 15 to 25, the sensitivity of our method will improve from 42%, 49%, to 67%. Concurrent increase of our method sensitivity and the experimental confidence score confirms the ability of EVlncRNAs to locate truly functional lncRNAs.

Newly discovered EVlncRNAs

We conducted a literature search for newly discovered EVlncRNAs because true positives from the EVlncRNA database are based on the literature prior to May 2016. We found that 24 new functional human lncRNAs are indeed classified as EVlncRNAs by EVlncRNA-pred. As shown in Supplementary Table S1, these new lncRNAs were experimentally validated by low-throughput techniques such as qRT-PCR, western blot, and knockdown. Their functions range from microRNA and protein binding to expression regulation although not all new lncRNAs have a clearly identified molecular-level function. For example, Li et al. [47] found that a lncRNA, SNHG20, has a significantly higher expression in Colorectal Cancer (CRC) tissues than in corresponding normal tissues from 107 CRC patients. SNHG20 regulated cell growth through modulation of a series of cell cycle-associated genes. Similarly, Lu et al. [48] found that a higher expression level of a lncRNA, SOX21-AS1, positively correlated with the tumour size and the advanced stage of tumor-node-metastasis (TNM), and the inhibition of SOX21-AS1 induced p57 expression. SNHG20 and SOX21-AS1 are classified as EVlncRNAs by EVlncRNA-pred.

Discussion

We have developed a method termed EVlncRNA-pred for selecting potentially functional lncRNAs from expressing lncRNAs found in high-throughput sequencing. Two different versions of the method were developed: one requires sequence information only whereas the other needs high-throughput experimental data in expression and histone modification. The results show that both versions can provide reasonably accurate separation of EVlncRNAs from HTlncRNAs and mRNAs, whereas the experimental data can provide an additional improvement from 0.47 to 0.51 for the Matthews correlation coefficient in ten-fold cross-validation. The method trained by human RNAs is robust as it performs equally well in mouse RNA classification and to a lesser extent in plant RNA classification.

In this work, we have randomly chosen 799 HTlncRNA and mRNA sequences to match in number to the largest training set we currently have for EVlncRNAs. The equal number was chosen to maximize learning [49]. To confirm the randomness for the choice of 799 HTlncRNAs and mRNAs, we have randomly selected 9 additional sets of 799 HTlncRNAs and mRNAs. The results of 10 fold-cross validations for the 10 sets are 0.475 ± 0.015 for MCC and 0.842 ± 0.01 for AUC. These small standard deviations indicate unbiased choices of negative sets. Furthermore, to confirm the usefulness of setting the ratio to 1, we systematically

expanded the training set by increasing the ratio from 1 to 1.5, 2, 3, and 4. We found that there is a reduction of AUC values from 0.879 to 0.866, 0.808, 0.790, and 0.799, respectively, for the human independent test set as the ratio increases. We also built another test set with a ratio of 1:4:4 for EVlncRNA: HTlncRNA: mRNA. We observed a similar reduction of AUC values for this larger test set (from 0.864 to 0.855, 0.799, 0.785, and 0.792, respectively). Thus, the model trained by the data with the ratio of 1 has the best performance not only for the test set with the ratio of 1 but also for the ratio of 1:4:4.

One revealing fact is that the most useful features for classification are related to conservations at protein levels followed by RNA levels. It turns out that protein conservation is the most important for separating from mRNAs whereas RNA conservation is the most important for separating from HTlncRNAs. This result provides additional confidence for the method developed. Although sequence conservation signal for lncRNA is in general weak [50], it remains one essential feature for functional lncRNAs [51–54]. The weak conservation signal for lncRNA, compared to the stronger conservation signal for proteins, makes the separation of EVlncRNA from HTlncRNA more challenging than the separation of HTlncRNA from mRNA. The result reported here indicates that the conservation signal can be picked up by a machine learning technique to highlight the intrinsic difference between those experimentally validated lncRNAs and those somehow expressed in high-throughput sequencing.

Here, we assumed from the outset that all lncRNAs reported in GENCODE are negative samples after excluding experimentally validated ones. This assumption was made despite the training set may contain a significant number of false negatives, which are truly functional lncRNAs yet to be validated by low throughput experiments. If the majority of the presumed negatives were false negatives, one would not be able to develop a method to separate positive from negatives during training. The fact that a highly robust method can be made indicates that false negatives are not dominant and there is a population in HTlncRNAs separable from known EVlncRNAs. The existence of such a population in HTlncRNAs that is distinct from known EVlncRNAs itself is interesting, as transcriptional noise could be a source for some of the lncRNAs found by high-throughput experiments [18].

Using a negative set containing some false negatives is a common practice in machine learning because negatives are always more difficult to prove. For example, in studying pathogenic genetic variations, genetic variants found in 1000 genome projects on healthy individuals [55] are considered as neutral (non-disease causing) [56]. However, this assumption may not be correct for some late-onset disorders, in particular. It was shown that removing potential false negatives (the genetic variants with low minor allele frequency and potentially pathogenic [57]) reduces the performance of the method trained. This suggests that having more data is more important than reducing potential false negatives in the training set [56].

To further examine the effect of potential errors in negatives, we randomly added 5% or 10% errors to nine-folds in the training set by assigning HTlncRNAs to EVlncRNAs and

EVlncRNAs to HTlncRNAs and testing the method for the remaining fold. This was repeated 10 times (ten-fold cross-validation). We also randomly selected 5% or 10% errors 10 separate times to obtain an average effect. Introducing 5% and 10% errors lead to the average MCC values changed only slightly from 0.513 to 0.496 and 0.480, respectively. The small changes due to assignment errors indicate that our method is robust against potential assignment errors in the training set. However, one has to be cautious that not all positive predictions are functional lncRNAs as the fraction of correct predictions in positive predictions is at 81% for the human test set (i.e. 19% are incorrect). Moreover, the coverage of functional lncRNAs (sensitivity) is at 64% due to the small training set. That is, predictions may miss many functional RNAs, tissue-specific lncRNAs, in particular. Nevertheless, the method should be already useful for prioritizing potentially functional lncRNAs for further experimental validation. In the meantime, we hope to further improve sensitivity and precision in the near future when a much larger dataset is available for deep learning.

To estimate the fraction of potentially functional EVlncRNAs in HTlncRNAs, we randomly selected 2000 human lncRNAs in NONCODE database [58]. None is found to overlap with known EVlncRNAs. Among them, 566 lncRNAs were classified as functional lncRNAs that can be validated by low-throughput experiments. This would place the fraction at 28.3% (566/2000) in expressed lncRNAs potentially with biological roles. In other words, the majority of expressing lncRNAs are likely non-functional. However, the fraction of functional ones is nearly 30% and the majority of their biological roles is yet to be investigated.

It is of interest to know the computational requirement of EVlncRNA-pred. Because our method makes a prediction according to 100-base blocks, the CPU time is linearly dependent on the length of a lncRNA. It takes about 198 seconds for a 1000-nucleotide RNA (ENST00000511331.1), 288 seconds for a 4000-nucleotide RNA (ENST00000592187.1), and 2115 seconds for a 15,145-nucleotide RNA (ENST00000608023.1) on an Intel Xeon E5 2.3GHz machine. Thus, EVlncRNA-pred will be computationally efficient for large-scale screening of functional lncRNAs.

Conclusions

In summary, we have developed the first bioinformatics tool to identify potentially functional lncRNAs from numerous lncRNAs and mRNAs found in high-throughput experiments. The classification performance of the method EVlncRNA-pred is reasonably high with AUC >0.84 for independent tests on human and mouse RNAs and >0.8 for plant RNAs. This indicates that the model built here is already useful for prioritizing functional lncRNAs for validation by low throughput experiments. The method along with the training and test datasets is freely available for experimental biologists at <http://biophy.dzu.edu.cn/lncrnared/index.html>.

Materials and methods

Training and test datasets for human lncRNAs

Most previous methods for separating lncRNAs from mRNAs were trained by using lncRNAs from GENCODE [44] as the positive dataset. These lncRNAs were obtained from the ENCODE project [59] by using a variety of high-throughput techniques and annotated by a combination of computational analysis, sequence comparison, and manual annotation. Here we treated them as the negative dataset after excluding all known experimentally validated, functional lncRNAs from a recently curated database EVlncRNAs [16] (the positive dataset). Because EVlncRNAs is far from a complete dataset for functional lncRNAs, our negative dataset likely contains some false negatives. As we discussed in the discussion section, this should not prevent us addressing the question if current experimentally validated lncRNAs (denoted as EVlncRNAs for convenience) are separable from lncRNAs from high-throughput (HT) experiments (denoted as HTlncRNAs for convenience).

We first created a positive human test set from EVlncRNAs that were not contained in GENCODE V19 so that we can create a set of newly discovered, experimentally validated lncRNAs. This test set was obtained by using CD-HIT [60] to remove redundant sequences with more than 80% sequence similarity with HTlncRNAs in GENCODE V19 and among themselves. We have chosen 80% sequence identity cutoff because statistics suggest a significant reduction in secondary structure similarity for RNA sequences with <80% sequence identity [61]. Moreover, it is the lowest sequence identity cutoff allowed by the program CD-HIT [60]. This cutoff was also employed previously for establishing non-redundant RNA sequences [62,63]. A total of 117 human EVlncRNAs were obtained as an independent positive test set. The remaining human lncRNAs from EVlncRNAs were used to generate the positive training set after removing redundant sequences by CD-HIT from the independent test set and among themselves. This leads to a training set of 799 human EVlncRNAs. The negative sets for HTlncRNAs (799 training and 117 independent test HTlncRNAs) were randomly selected from GENCODE V19 while ensuring <80% sequence identity among themselves and from the positive sets.

In addition to the HTlncRNA set as the negative set, we also included mRNAs from GENCODE V19 as the negative set. These mRNAs were randomly selected with <80% sequence similarity between each other and from selected HTlncRNAs and EVlncRNAs. The number of mRNAs is set to the same as the size of the two positive sets. Thus, the final human training dataset contains 799 EVlncRNAs (positive), 799 HTlncRNAs (negative) and 799 mRNAs (negative). The final human independent test set contains 177 EVlncRNAs (positives), 177 HTlncRNAs (negatives) and 177 mRNAs (negatives). Using both HTlncRNAs and mRNAs in the negative sets is to ensure that our method can separate EVlncRNAs from either HTlncRNAs or mRNAs.

Here, we have set the ratio of EVlncRNA:HTlncRNA:mRNA to 1:1:1 in training and test sets. The purpose is to

maximize learning by undersampling the negative samples [49]. To examine the effect of the ratio, we also built the training set for EVlncRNA:HTlncRNA:mRNA at 799:1200:1200 (1:1.5:1.5), 799:1600:1600 (1:2:2), 799:2400:2400 (1:3:3), and 799:3200:3200 (1:4:4), respectively. These additional mRNAs and HTlncRNAs were randomly selected with <80% sequence similarity between each other and from previously selected mRNAs, HTlncRNAs and EVlncRNAs. Similarly, we built an additional independent test set EVlncRNA:HTlncRNA:mRNA at the ratio of 117:468:468 (1:4:4) with the same positive samples in the independent test set but expanded the sample sizes of HTlncRNAs and mRNAs. A ratio of 1:4 between EVlncRNA and HTlncRNA is close to the real-world situation as we shall see.

Independent test sets from mouse and plant lncRNAs

To further test the robustness of the methods developed, we established independent test sets by using the mouse and plant lncRNAs. Similar to human datasets, the positive sets for plant and mouse were obtained from the EVlncRNAs database [16]. There were 166 mouse EVlncRNAs after removing redundant sequences with more than 80% sequence similarity to the human set (both positive and negative sets) and among themselves. We randomly selected 166 mouse HTlncRNAs from GENCODE V19 as the negative set after removing the redundant sequences from the mouse positive set, the human set and among themselves. In addition, we have randomly selected 166 mouse mRNA set from GENCODE V19 with the same sequence similarity cutoff to remove redundancy. The final mouse test set contains 166 EVlncRNAs (positives), 166 HTlncRNAs (negatives) and 166 mRNAs (negatives).

We used the EVlncRNAs of *Arabidopsis thaliana* in the EVlncRNAs database [16] to construct the positive set for plant. After removing redundant sequences with more than 80% sequence similarity with the human set and among themselves, 120 *Arabidopsis thaliana* EVlncRNAs were obtained as the plant positive set. The HTlncRNAs and mRNAs of *Arabidopsis thaliana* were from the Ensembl Plants database [64]. Similar to the mouse negative set, an equal number of HTlncRNAs and mRNAs of *Arabidopsis thaliana* were randomly selected after removing the redundant sequences from the plant positive set, the human set, and among themselves. The final plant test set contains 120 EVlncRNAs (positives), 120 HTlncRNAs (negatives) and 120 mRNAs (negatives).

Input features

All features were block-averaged similar to previous studies [35,37,39,65]. Each block has 100 nucleotides, centered at 50, 100, 150, and etc. until the entire sequence is covered by blocks. For a given feature, the value or average value of each block was calculated to represent the block. The final feature values are average, maximum and variance of values of the blocks that covered the entire sequence. Following features are calculated.

Features based on sequences

Features based on sequences are employed to develop the sequence-only model.

GC content. The percentage of G and C in a sequence block was calculated.

Purine (PUR) content. The percentage of purines (G and A) in a sequence block was calculated.

DNA conservation score. The phastCons [45] scores provided by the UCSC [46] represented the DNA sequence conservation of human and mouse. The phastCons scores for human DNAs are from phastCons100way, and the scores for mouse DNAs are from phastCons60way. However, no similar scores are available for plant DNAs. We have simply set these values to zero when applying to the plant set (also see below).

Protein conservation score. The protein conservation score was calculated by BLASTx that searches a given nucleotide sequence against the protein sequence in the UniProt database [66].

RNA conservation score. Infernal ('INFERence of RNA Alignment') [67] was employed for searching Rfam databases [68] for RNA structure and sequence similarities. It is an implementation of a special case of profile stochastic context-free grammars called covariance models (CMs). A CM is like a sequence profile, but it scores a combination of sequence conservation and RNA secondary-structure conservation.

Predicted solvent accessible surface area (ASA) of RNA. RNA ASA values were predicted by RNAsnap [62].

Features based on high-throughput experimental results

The above sequence-based features together with the features based on high-throughput experimental results (see below) are utilised to develop the full-feature model.

Expression abundance. The reads per kilobase per million (RPKM) for each sequence were calculated from polyA+, polyA- and small RNA-seq data. The maximum scores for five cell lines (GM12878, K562, H1-hESC, HeLa-S3, and HepG2) were assigned to human sequences. For the mouse and *Arabidopsis thaliana*, the RNA-seq data of various tissues were used. These feature values were obtained from COME [35].

Histone modification. The ChIP-seq data from H3K36me3 and H3K4me3 modification were used to calculate the signal over a sequence. The averaged input-normalized signals of five cell lines (GM12878, K562, H1-hESC, HeLa-S3, and HepG2) were used for human sequences. The ChIP-seq data of various tissues were used for sequences of mouse and *Arabidopsis thaliana*. These values were obtained from COME [35].

Ribosome profiling is another possible experimental feature. However, a previous study suggested its minor contribution to separation of HTlncRNA from mRNA [35]. We expect that it is less useful for separating EVlncRNA from

HTlncRNA as both are not translated into peptides or proteins. As a result, this feature was not employed in this study.

Support vector machines

We used SVM with the RBF kernel implemented in LIBSVM version 3.22 [69] to build our model. We optimized the parameters C and gamma using the grid search algorithm implemented in LIBSVM.

Cross-validation and independent test

We performed 10-fold cross-validation on the training set. In this cross-validation, the training set was randomly divided into ten folds, and each fold was tested in turn by using the remaining nine folds for training. To examine whether the results are consistent for different divisions of the dataset, we conducted 10-fold cross validations 100 times by randomly dividing the training set 100 times. We also used the whole training set to train the model and tested the model on independent test sets.

Performance evaluation criteria

The performance of our method was evaluated by Matthews correlation coefficient (MCC), receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), accuracy, sensitivity, specificity, and precision. The equations are as below.

$$\text{MCC} = \frac{(\text{TP} \times \text{TN} - \text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}} \quad (1)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (2)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \quad (4)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

where TP and TN represent the positive and negative samples that have been correctly predicted, respectively, FP and FN represent the positive and negative samples that have been falsely predicted, respectively. MCC is essentially a correlation coefficient between predicted and actual binary classifications with values between -1 to 1 with zero for random prediction. It is a balanced measure for unequal-sized positive and negative samples. Sensitivity is the fraction of predicted true EVlncRNAs in all true EVlncRNAs. Specificity is the fraction of predicted true negatives in all true negatives. Precision is the fraction of true EVlncRNAs in all predicted EVlncRNAs.

Data and software availability

EVlncRNA-pred is available as a web server at <http://biophy.dzu.edu.cn/lncrnaped/index.html>. All datasets used in this study can be obtained from the same website.

Acknowledgments

We thank Hucheng Tang for helping develop the web-based server.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Natural Science Foundation of China [61671107, 61271378, 61801081]; Taishan Scholars Program of Shandong province of China [Tshw201502045]; National Health and Medical Research Council of Australia [1121629 to Y.Z.]; Australia Research Council [DP 180102060 to Y.Z.]; and Talent Introduction Project of Dezhou University of China [320111 to B.Z.].

ORCID

Yuedong Yang  <http://orcid.org/0000-0002-6782-2813>
Yaoqi Zhou  <http://orcid.org/0000-0002-9958-5699>

References

- [1] Djebali S, Davis CA, Merkel A, et al. Landscape of transcription in human cells. *Nature*. 2012;489:101–108.
- [2] Bertone P, Stolc V, Royce TE, et al. Global identification of human transcribed sequences with genome tiling arrays. *Science*. 2004;306:2242–2246.
- [3] Johnson JM, Edwards S, Shoemaker D, et al. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet*. 2005;21:93–102.
- [4] Ponting CP, Oliver PL, Reik W. Evolution and functions of long noncoding RNAs. *Cell*. 2009;136:629–641.
- [5] Mercer TR, Dinger ME, Sunken SM, et al. Specific expression of long noncoding RNAs in the mouse brain. *Proc Natl Acad Sci U S A*. 2008;105:716–721.
- [6] Pauli A, Valen E, Lin MF, et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res*. 2012;22:577–591.
- [7] Brannan CI, Dees EC, Ingram RS, et al. The product of the H19 gene may function as an RNA. *Mol Cell Biol*. 1990;10:28–36.
- [8] Brockdorff N, Ashworth A, Kay GF, et al. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell*. 1992;71:515–526.
- [9] Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. *Nat Rev Genet*. 2009;10:155–159.
- [10] Kung JT, Colognori D, Lee JT. Long noncoding RNAs: past, present, and future. *Genetics*. 2013;193:651–669.
- [11] Gibb EA, Brown CJ, Lam WL. The functional role of long non-coding RNA in human carcinomas. *Mol Cancer*. 2011;10:38.
- [12] Quek XC, Thomson DW, Maag Jesper LV, et al. lncRNADB v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res*. 2015;43:D168–D173.
- [13] Xuan H, Zhang L, Liu X, et al. PLNlncRbase: A resource for experimentally identified lncRNAs in plants. *Gene*. 2015;573:328–332.

- [14] Chen G, Wang Z, Wang D, et al. lncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* 2013;41:D983–D986.
- [15] Liu C-J, Gao C, Ma Z, et al. lncRInter: A database of experimentally validated long non-coding RNA interaction. *J Genet Genomics.* 2017;44:265–268.
- [16] Zhou B, Zhao H, Yu J, et al. eVLncRNAs: a manually curated database for long non-coding RNAs validated by low-throughput experiments. *Nucleic Acids Res.* 2018;46:D100–D105.
- [17] Palazzo AF, Lee ES. Non-coding RNA: what is functional and what is junk? *Front Genet.* 2015;6:2.
- [18] Struhl K. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol.* 2007;14:103–105.
- [19] Hon C-C, Ramilowski JA, Harshbarger J, et al. An atlas of human long non-coding RNAs with accurate 5' ends. *Nature.* 2017;543:199–204.
- [20] Toiber D, Leprévier G, Rotblat B. Long noncoding RNA: noncoding and not coded. *Cell Death Discov.* 2017;3:16104.
- [21] Liu SJ, Horlbeck MA, Cho SW, et al. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science.* 2017;355:eaah7111.
- [22] Graur D. An upper limit on the functional fraction of the human genome. *Genome Biol Evol.* 2017;9:1880–1885.
- [23] Jalali S, Kapoor S, Sivasdas A, et al. Computational approaches towards understanding human long non-coding RNA biology. *Bioinformatics.* 2015;31:2241–2251.
- [24] Dinger ME, Pang KC, Mercer TR, et al. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol.* 2008;4:e1000176.
- [25] Prasanth KV, Spector DL. Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes Dev.* 2007;21:11–42.
- [26] Badger JH, Olsen GJ. CRITICA: coding region identification tool invoking comparative analysis. *Mol Biol Evol.* 1999;16:512–524.
- [27] Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics.* 2011;27:i275–i282.
- [28] Castrignanò T, Canali A, Grillo G, et al. CSTminer: a web tool for the identification of coding and noncoding conserved sequence tags through cross-species genome comparison. *Nucleic Acids Res.* 2004;32:W624–W627.
- [29] Washietl S, Findeiß S, Müller SA, et al. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA.* 2011;17:578–594.
- [30] Necșulea A, Soumillon M, Warnefors M, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature.* 2014;505:635–640.
- [31] Wang L, Park HJ, Dasari S, et al. CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res.* 2013;41:e74.
- [32] Achawanantakun R, Chen J, Sun Y, et al. lncRNA-ID: long non-coding RNA identification using balanced random forests. *Bioinformatics.* 2015;31:3897–3905.
- [33] Pian C, Zhang G, Chen Z, et al. lncRNApred: classification of long non-coding RNAs and protein-coding transcripts by the ensemble algorithm with a new hybrid feature. *PLoS ONE.* 2016;11:e0154567.
- [34] Wucher V, Legeai F, Hédan B, et al. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.* 2017;45:e57.
- [35] Hu L, Xu Z, Hu B, et al. COME: a robust coding potential calculation tool for lncRNA identification and characterization based on multiple features. *Nucleic Acids Res.* 2017;45:e2.
- [36] Arrial RT, Togawa RC, Brigido M. Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: case study of the pathogenic fungus *Paracoccidioides brasiliensis*. *BMC Bioinformatics.* 2009;10:239.
- [37] Sun L, Luo H, Bu D, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* 2013;41:e166.
- [38] Kong L, Zhang Y, Ye Z-Q, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 2007;35:W345–W349.
- [39] Li A, Zhang J, Zhou Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics.* 2014;15:1.
- [40] Sun L, Liu H, Zhang L, et al. lncRScan-SVM: a tool for predicting long non-coding RNAs using support vector machine. *PLoS ONE.* 2015;10:e0139654.
- [41] Lu ZJ, Yip KY, Wang G, et al. Prediction and characterization of noncoding RNAs in *C. elegans* by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res.* 2011;21:276–285.
- [42] Lv J, Liu H, Huang Z, et al. Long non-coding RNA identification over mouse brain development by integrative modeling of chromatin and genomic features. *Nucleic Acids Res.* 2013;41:10044–10061.
- [43] Ramos Alexander D, Diaz A, Nellore A, et al. Integration of genome-wide approaches identifies lncRNAs of adult neural stem cells and their progeny in vivo. *Cell Stem Cell.* 2013;12:616–628.
- [44] Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 2012;22:1760–1774.
- [45] Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 2005;15:1034–1050.
- [46] Tyner C, Barber GP, Casper J, et al. The UCSC genome browser database: 2017 update. *Nucleic Acids Res.* 2017;45:D626–D634.
- [47] Li C, Zhou L, He J, et al. Increased long noncoding RNA SNHG20 predicts poor prognosis in colorectal cancer. *BMC Cancer.* 2016;16:655.
- [48] Lu X, Huang C, He X, et al. A novel long non-coding RNA, SOX21-AS1, indicates a poor prognosis and promotes lung adenocarcinoma proliferation. *Cell Physiol Biochem.* 2017;42:1857–1869.
- [49] Akbani R, Kwek S, Japkowicz N. Applying support vector machines to imbalanced datasets. In: Boulicaut J-F, Esposito F, Giannotti F, et al., editors. *Machine Learning: ECML 2004*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2004. p. 39–50.
- [50] Johnsson P, Lipovich L, Grander D, et al. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim Biophys Acta.* 2014;1840:1063–1071.
- [51] Guttman M, Amit I, Garber M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature.* 2009;458:223–227.
- [52] Brown CJ, Hendrich BD, Rupert JL, et al. The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell.* 1992;71:527–542.
- [53] Tang Z, Wu Y, Yang Y, et al. Comprehensive analysis of long non-coding RNAs highlights their spatio-temporal expression patterns and evolutionary conservation in *Sus scrofa*. *Sci Rep.* 2017;7:43166.
- [54] Diederichs S. The four dimensions of noncoding RNA conservation. *Trends Genet.* 2014;30:121–123.
- [55] Clarke L, Zheng-Bradley X, Smith R, et al. The 1000 genomes project: data management and community access. *Nat Methods.* 2012;9:459–462.
- [56] Zhao H, Yang Y, Lin H, et al. DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. *Genome Biol.* 2013;14:R23.
- [57] Tennessen JA, Bigham AW, O'Connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science.* 2012;337:64–69.
- [58] Fang S, Zhang L, Guo J, et al. NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.* 2018;46:D308–D314.
- [59] Consortium TEP. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489:57–74.

- [60] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*. 2006;22:1658–1659.
- [61] Capriotti E, Marti-Renom MA. Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. *BMC Bioinformatics*. 2010;11:322.
- [62] Yang Y, Li X, Zhao H, et al. Genome-scale characterization of RNA tertiary structures and their functional impact by RNA solvent accessibility prediction. *RNA*. 2017;23:14–22.
- [63] Guruge I, Taherzadeh G, Zhan J, et al. B-factor profile prediction for RNA flexibility using support vector machines. *J Comput Chem*. 2018;39:407–411.
- [64] Bolser DM, Staines DM, Perry E, et al. Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomic data. In: van Dijk ADJeditor. *Plant Genomics Databases: methods and Protocols*. New York: Springer New York; 2017. p. 1–31.
- [65] Lin MF, Carlson JW, Crosby MA, et al. Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. *Genome Res*. 2007;17:1823–1836.
- [66] UniProt Consortium T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2018;46:2699.
- [67] Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. 2013;29:2933–2935.
- [68] Kalvari I, Argasinska J, Quinones-Olvera N, et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res*. 2018;46:D335–D342.
- [69] Chang -C-C, Lin C-J. LIBSVM: a library for support vector machines. *Acm T Intel Syst Tec*. 2011;2:27.