Bioinformatics, 34(22), 2018, 3825–3834
doi: 10.1093/bioinformatics/bty428
Advance Access Publication Date: 29 May 2018
Original Paper

Sequence analysis

LncADeep: an *ab initio* lncRNA identification and functional annotation tool based on deep learning

Cheng Yang[1,2], Longshu Yang[1], Man Zhou[1], Haoling Xie[1,3], Chengjiu Zhang[1], May D. Wang[2] and Huaiqiu Zhu[1,3,*]

[1]Department of Biomedical Engineering, College of Engineering, and Centre for Quantitative Biology, Peking University, Beijing 100871, China, [2]Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332, USA and [3]Peking University-Tsinghua University-National Institute of Biological Sciences (PTN) Joint PhD Program and College of Life Sciences, Peking University, Beijing 100871, China

*To whom correspondence should be addressed.
Associate Editor: Inanc Birol

Received on January 22, 2018; revised on April 20, 2018; editorial decision on May 18, 2018; accepted on May 23, 2018

Abstract

Motivation: To characterize long non-coding RNAs (lncRNAs), both identifying and functionally annotating them are essential to be addressed. Moreover, a comprehensive construction for lncRNA annotation is desired to facilitate the research in the field.

Results: We present LncADeep, a novel lncRNA identification and functional annotation tool. For lncRNA identification, LncADeep integrates intrinsic and homology features into a deep belief network and constructs models targeting both full- and partial-length transcripts. For functional annotation, LncADeep predicts a lncRNA's interacting proteins based on deep neural networks, using both sequence and structure information. Furthermore, LncADeep integrates KEGG and Reactome pathway enrichment analysis and functional module detection with the predicted interacting proteins, and provides the enriched pathways and functional modules as functional annotations for lncRNAs. Test results show that LncADeep outperforms state-of-the-art tools, both for lncRNA identification and lncRNA–protein interaction prediction, and then presents a functional interpretation. We expect that LncADeep can contribute to identifying and annotating novel lncRNAs.

Availability and implementation: LncADeep is freely available for academic use at http://cqb.pku.edu.cn/ZhuLab/lncadeep/ and https://github.com/cyang235/LncADeep/.

Contact: hqzhu@pku.edu.cn

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

1 Introduction

As the majority of non-coding RNAs, long non-coding RNAs (lncRNAs, length above 200 nt) (Fatica and Bozzoni, 2014) play important biological roles in dosage compensation, genomic imprinting, cell differentiation etc. (Fatica and Bozzoni, 2014; Guttman and Rinn, 2012), and have been implicated in human disease such as cancers (Gupta *et al.*, 2010). Although quite a number of lncRNAs have been characterized, the functions of most of the lncRNAs discovered currently remain unclear (Derrien *et al.*, 2012).

To comprehensively characterize newly discovered transcripts, two issues are essential in the field to be addressed: identifying lncRNAs and then inferring their functions. As for the former, lncRNA identification is still a challenge. First, lncRNAs and mRNAs share many similarities such as transcript length and

splicing structure (Ulitsky and Bartel, 2013; Guttman and Rinn, 2012), which complicate lncRNA identification. Facilitated by high-throughput sequencing technologies, RNA sequencing (RNA-seq) has become a prevalent method for studying lncRNAs. However, accurate full-length transcript assembly is impeded by the short reads from current RNA-seq techniques (Steijger *et al.*, 2013). For example, according to an assessment, the best-performing unguided assembly tools can identify only 21% of full-length mRNAs from human datasets, and over 60% of reconstructed transcripts are of partial-length (Steijger *et al.*, 2013). Thus, lncRNA identification is further complicated by partial-length mRNAs reconstructed from RNA-seq short reads, owing to the fact that partial-length mRNAs truncated at 5' and/or 3' end may lead to incomplete coding sequences (CDSs), which are prone to be misclassified as lncRNAs. In addition, in real data consisting of lncRNAs as well as both full- and partial-length mRNAs, the composition of full- and partial-length mRNAs actually varies case by case, which thus complicates the training of lncRNA identification algorithms.

Currently the lncRNA identification methods can be classified into two categories, reference-based and reference-free (or *ab initio*) methods. The reference-based methods, such as lncRScan-SVM (Sun *et al.*, 2015), COME (Hu *et al.*, 2016) and lncScore (Zhao *et al.*, 2016), require comprehensive reference genome annotation. For example, lncRScan-SVM relies on a reference genome with annotation to extract the features of exon length and count into a SVM. However these methods can suffer limitations for non-model organisms lacking whole genome sequence or gene annotation. As for the reference-free (*ab initio*) methods, such as CPC (Kong *et al.*, 2007), lncRNA-ID (Achawanantakun *et al.*, 2015), CPAT (Wang *et al.*, 2013), CPC2 (Kang *et al.*, 2017), CNCI (Sun *et al.*, 2013), PLEK (Li *et al.*, 2014), FEElnc (Wucher *et al.*, 2017), longdist (Schneider *et al.*, 2017) and lncRNA-MFDL (Fan and Zhang, 2015), several of them are noteworthy in lncRNA prediction with visible performances. However, most of them focus on only full-length transcripts. Although CNCI, lncScore and FEElnc realized the existence of partial-length mRNAs, they did not consider the various compositions of full- and partial-length mRNAs in real data and used only an arbitrary composition of full- and partial-length mRNAs for training, which could affect the performance of lncRNA identification. A brief introduction for these methods is provided in Supplementary Table S1.

After identifying lncRNAs from transcripts, one is certainly much more concerned to obtain their functional interpretation. However, comprehensively characterizing lncRNAs would be more of a challenge since the functions of lncRNAs are complicated. To exert biological functions, lncRNAs can interact with DNAs, RNAs and proteins (lncRNA–protein interactions), and protein is confirmed to be the first and principal partner of lncRNAs (Chu *et al.*, 2015; Guttman and Rinn, 2012; Yang *et al.*, 2015). Among the three types of interactions, the ones between lncRNAs and DNAs/RNAs are less studied at present, while lncRNA–protein interactions demonstrate crucial roles in the functioning of lncRNA, providing satisfactory details of how lncRNAs exert functions in various biological processes (Chu *et al.*, 2015).

Thus, identifying lncRNA–protein interactions is essential to understanding lncRNA functions and mechanisms in biological processes. Although several experimental approaches (McHugh *et al.*, 2014) are available for probing RNA–protein interactions, experimental approaches are expensive and time-consuming (Muppirala *et al.*, 2011). In contrast, computational approaches are more convenient and rapid, and can employ experimentally verified datasets to infer lncRNA–protein interactions. Therefore, it is essential to develop computational approaches for predicting lncRNA–protein interaction.

Several remarkable methods have been be addressed to lncRNA–protein interactions prediction, for example catRAPID (Bellucci *et al.*, 2011), GlobalScore (Cirillo *et al.*, 2016), RPISeq (Muppirala *et al.*, 2011), lncPro (Lu *et al.*, 2013), RPI-Pred (Suresh *et al.*, 2015), rpiCool (Akbaripour-Elahabad *et al.*, 2016) and IPMiner (Pan *et al.*, 2016). They integrated sequence or structure features of RNAs and proteins to predict interactions (Supplementary Table S2). However, the lncRNA–protein interactions prediction are still expected to be improved by more powerful methods. What is more, these tools were developed to end in predicting lncRNA–protein interactions and with interactions output only. To characterize the lncRNAs for their biological role, it is beneficial to answer their functional roles involved in biological pathways or functional modules. Up to now, there are no function annotation approaches to implement this step yet.

To this end, a comprehensive construction of lncRNA annotation is expected to facilitate the research in the field. With the development of high-throughput sequencing technology, a large amount of transcripts will be sequenced and require functional interpretation. In this paper, we present LncADeep, which can not only identify lncRNAs, but also infer the functions of lncRNAs, while no tools can accomplish both yet. For lncRNA identification, LncADeep integrated sequence intrinsic features [e.g. entropy density profile (EDP)] and homology features (i.e. profile hidden Markov model-based alignment) into a deep belief network (DBN) of deep learning algorithm. Herein, we constructed the model to target both full- and partial-length transcripts. To our knowledge, LncADeep is the first tool considering the various compositions of full- and partial-length mRNAs in dataset. Results showed that LncADeep has outperformed state-of-the-art tools, as well as for cross-species lncRNA identification. For functional annotation, we first predicted a lncRNA's interacting proteins based on deep neural networks (DNNs) of deep learning algorithm, using both sequence and structure information. Results demonstrated that LncADeep achieved better performance compared with all the current methods. Furthermore, LncADeep integrated KEGG (Kanehisa *et al.*, 2010) and Reactome (Croft *et al.*, 2014) pathway enrichment analysis and functional module detection with the predicted interacting proteins, and provided the enriched pathways and functional modules as functional annotations for lncRNAs. As an *ab initio* lncRNA identification and functional annotation tool, we expect that LncADeep can contribute to identifying and annotating novel lncRNAs.

2 Materials and methods

2.1 Data description

Both RefSeq (Pruitt *et al.*, 2012) and GENCODE (Derrien *et al.*, 2012; Harrow *et al.*, 2012) provide comprehensive and well-annotated sequences for mRNAs and lncRNAs. In particular, the human mRNAs in RefSeq are of full-length. However about 36% of human mRNAs in GENCODE (Release 24) are of partial-length, which are mainly owning to the challenges of transcript assembling from high-throughput sequenced RNA-seq data (Harrow *et al.*, 2012). Herein, full-length mRNAs contains 5' untranslated region (UTR), CDS and 3' UTR, while partial-length mRNAs can miss 5' UTR or 3' UTR and the CDS can also be incomplete.

As aforementioned, we constructed two models for lncRNA identification, one targeting full-length transcripts and another targeting transcripts including both full- and partial-length ones.

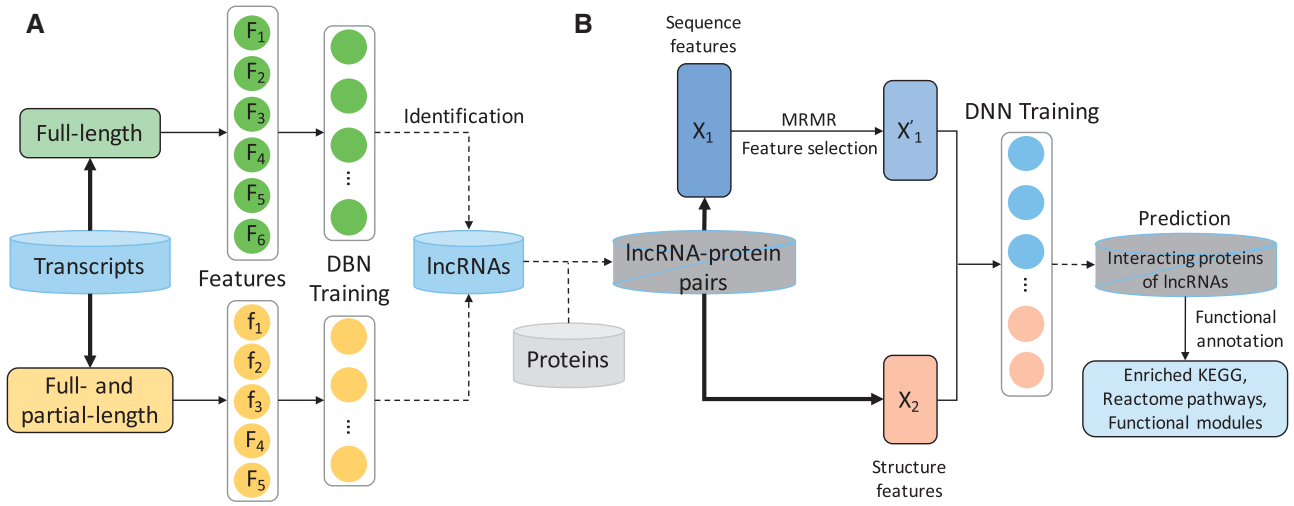

Fig. 1. The flowchart of LncADeep and using LncADeep for lncRNA identification and functional annotation. (**A**) We construct two models for lncRNA identification, one for full-length transcripts, and the other for transcripts including full- and partial-length. F_1–F_6 refer to ORF length and coverage, the EDP of ORF, mean hexamer score, Fickett nucleotide features, HMMER index, and UTR length and GC content, respectively. f_1–f_3 represent LCDS length and coverage, the EDP of LCDS and mean hexamer score, respectively. (**B**) We use sequence and structure features for the prediction of lncRNA–protein interaction. X_1, X'_1 and X_2 refer to sequence features, sequence features after feature selection, and structure features, respectively. The identified lncRNAs can be input for predicting lncRNA–protein interactions, then the interacting proteins can be used for inferring the functions of lncRNAs

Herein the full-length transcripts are composed of lncRNAs and full-length mRNAs, while the others consist of lncRNAs and full- and partial-length mRNAs. Full-length human mRNAs were collected from RefSeq Release 75, and mRNAs including full- and partial-length and lncRNAs were from GENCODE Release 24 (Supplementary Table S3). To assess the performance of cross-species lncRNA prediction, we collected mouse transcripts also from RefSeq and GENCODE database, as its experimentally verified mRNAs and lncRNAs are more abundant compared with those of other species (Supplementary Table S4).

To predict lncRNA–protein interactions for lncRNA function annotation, we downloaded data from the NPInter database (Yuan *et al.*, 2014), which collects experimentally verified lncRNA–protein interacting pairs. Since we focus on human lncRNA–protein interactions, we kept only the interacting pairs labeled with 'Homo sapiens' and 'ncRNA-protein binding', removing the interacting pairs whose ncRNA is shorter than 200 nt, and finally obtained 6204 lncRNA–protein interacting pairs (Supplementary Table S5). To construct the negative dataset for benchmarking, we referred to the method used in (Akbaripour-Elahabad *et al.*, 2016; Muppirala *et al.*, 2011), i.e. pair lncRNAs and proteins, exclude all known interactions, and randomly keep 6204 lncRNA–protein pairs as non-interacting ones. As for the release version of LncADeep, to make full use of the generated non-interacting lncRNA–protein pairs, we adopted EasyEnsemble (Liu *et al.*, 2009) for model training (Supplementary Section S1).

2.2 Methods for lncRNA identification

To identify lncRNA, we integrate several features of sequence content and homology as predictor variables into a DBN of deep learning architecture and construct two models shown in Figure 1A.

2.2.1 Model for full-length transcripts

For the model targeting full-length transcripts, we use features including open reading frame (ORF) length and coverage, the EDP of ORF (Liu *et al.*, 2013; Zhu *et al.*, 2007), Mean hexamer score, UTR coverage and guanine-cytosine (GC) content, Fickett nucleotide feature and HMMER index. The description of these features can be referred to in Supplementary Section S2.

2.2.2 Model for transcripts including full- and partial-length

Currently RNA-seq data tend to produce transcripts mixed with a significant amount of partial-length mRNAs, which may be easily misclassified into lncRNAs and challenge the performance of lncRNA identification. For the model targeting transcripts including both full- and partial-length, we introduce a feature of the longest CDS (LCDS) to describe partial-length mRNAs.

Considering the existence of partial-length mRNA, we first define the ORF-based CDS representing the longest ORFs (LORFs) with their 3' ends missed. Then, given a transcript, we use a dynamic programming method, maximum subarray sum (MSS) (Bentley, 1984), to find a hexamer-based CDS with the maximum hexamer score (Supplementary Section S3). The total hexamer score $\lambda(S)$ for a given hexamer sequence $S = h_1 h_2 \ldots h_m$ is

$$\lambda(S) = \sum_{i=1}^{m} \log \frac{F_c(h_i)}{F_{nc}(h_i)} \quad (1)$$

where $F_c(h_i)$ and $F_{nc}(h_i)$ $(i = 1, 2, \ldots, 4096)$ represent in-frame coding and non-coding hexamer frequency, respectively. Finally, we choose the longer one as the LCDS from the ORF-based and the hexamer-based CDS. With the LCDS, we then calculate its length and coverage as features, where the coverage is the ratio of the length of LCDS to the transcript length.

In addition, we use other features, including mean hexamer score, the EDP of the LCDS, Fickett nucleotide feature and HMMER index. Among these features, mean hexamer score and the EDP are calculated on the LCDS, while Fickett nucleotide feature and HMMER index are calculated the same as in the model for full-length transcripts. However we did not use the UTR feature, since partial-length transcripts might lack 5' or 3'ends.

2.3 Methods for predicting lncRNA–protein interaction and inferring lncRNA functions

To characterize lncRNA–protein interacting pairs, we use both sequence and structure features shown in Figure 1B, which have been demonstrated useful in (Akbaripour-Elahabad *et al.*, 2016;

Muppirala *et al.*, 2011; Suresh *et al.*, 2015). For sequence features, each lncRNA is first encoded with a 256-dimensional vector corresponding to its EDP of *4-mer*s (Supplementary Section S4). Meanwhile we encode a lncRNA with the features used in lncRNA identification, including Fickett nucleotide feature and the features of LCDS (i.e. the EDP of the LCDS, LCDS length and coverage and mean hexamer score), which consist of a 47-dimensional vector (Supplementary Table S6).

Protein sequence is encoded with a 343-dimensional vector corresponding to its EDP of *3-mer*s in the 7-letter alphabet representation of the protein sequence (Supplementary Section S4). In total, each lncRNA–protein pair is represented by a 646 ($= 4^4 + 7^3 + 47$) dimensional sequence feature vector. However, too many features can incur overfitting when the size of training data is relatively small, since the number of experimentally verified lncRNA–protein interacting pairs is limited (herein only 6204 interacting pairs are available). To alleviate overfitting, we conduct feature selection using the minimal-redundancy-maximal-relevance (mRMR) criterion to select those most characterizing features (Peng *et al.*, 2005). Finally, we obtain 110 features among the 646 candidate features (Supplementary Table S7).

For the structure features, we include secondary structure, hydrogen-bonding and Van der Waals propensities for both lncRNA and protein (Lu *et al.*, 2013). To predict the secondary structure of a lncRNA, we use RNAfold (Lorenz *et al.*, 2011) (Supplementary Sections S4 and S5). For the other structure features, we follow the process described in lncPro (Lu *et al.*, 2013) (Supplementary Section S4). Finally the structure features of a lncRNA–protein pair are encoded using a 80-dimensional vector, and the sequence and structure features of lncRNA–protein pairs are further combined for predicting lncRNA–protein interaction.

To infer the functions of a lncRNA from its predicted interacting proteins, LncADeep integrates KEGG, Reactome pathway enrichment analysis and functional module detection on the interacting proteins. We downloaded reviewed human protein sequences from Uniprot database (UniProtConsortium et al., 2011), and obtained 20 121 protein sequences after filtering (Supplementary Section S6). Then, LncADeep predicts the interacting proteins of lncRNAs from the 20 121 proteins and conducts the function annotations with the predicted interacting proteins. For pathway enrichment analysis, LncADeep uses Fisher's exact test for the significance test, Benjamini–Hochberg (BH) method for the multiple testing correction and keeps enriched pathways whose adjusted *P*-value is <0.05. Proteins usually function as modules (Enright *et al.*, 2002), and interpreting the functional modules derived from the interacting proteins of lncRNAs can offer some helpful information for the functions of lncRNAs. To detect functional modules in the interacting proteins of a lncRNA, LncADeep uses Markov clustering (MCL) (Enright *et al.*, 2002) by integrating protein–protein interaction information provided by HIPPIE database (Alanis-Lobato *et al.*, 2016), which collects experimentally verified interactions from various reliable sources.

2.4 Deep learning architectures for lncRNA identification and lncRNA–protein interaction prediction

The extracted features from each transcript can be concatenated into a vector and then fed to a deep learning method for training and testing. Deep learning methods are good at discovering intricate hidden structures in high-dimensional data, which is particular helpful for classification problems (Min *et al.*, 2016). Many deep learning architectures have been available, such as DBN, DNN, convolutional neural network (CNN) and recurrent neural network (RNN) (Hinton *et al.*, 2006; LeCun *et al.*, 2015; Min *et al.*, 2016). Various architectures offer alternative approaches to further improve the classification performance apart from novel models.

In this paper, we implement a DBN, built as a stack of restricted Boltzmann machines (RBMs), to identify lncRNAs. Stacking a number of the RBMs learned layer by layer from bottom-up gives rise to a DBN. DBN can make use of the dataset for pre-training and obtain a good initialization point for the neural network, which helps to prevent overfitting and capture complex hidden information of the observed variables (LeCun *et al.*, 2015). After initialization, an output layer can be added and the whole neural network can be fine-tuned with backpropagation. DBN has shown successfulness and impressive performance in the field of bioinformatics, such as TransSpeciesDeepLearning (Chen *et al.*, 2015) and DeepQA (Cao *et al.*, 2016). We construct the DBN with the setting as suggested in (Hinton *et al.*, 2006): for the first two layers, we use Gaussian (visible)–Bernoulli (hidden) RBM; while for the other layers, we utilize Binary-Binary RBM. The architecture of the DBN is described in Supplementary Table S6.

Based on the sequence and structure features for lncRNA–protein pairs, we then construct the deep learning architecture with a DNN for predicting lncRNA–protein interaction. Similarly, the features for each lncRNA–protein pair are concatenated into a feature vector and then input to the DNN for classification (Fig. 1B). Inspired by the deep stacking network proposed in (Deng *et al.*, 2012), the DNN is built as described in Supplementary Figure S1 and Table S8. The rationale is that the prediction results from the previous step can offer additional helpful generalized information. In our implementation, we added dropout layers to prevent overfitting (Srivastava *et al.*, 2014) and used backpropagation to fine-tune the network.

3 Results

To evaluate the prediction performance of LncADeep and other existing tools, we use two independent quantities, Sn (sensitivity), Sp (specificity) and an average measure Hm (the harmonic mean of sensitivity and specificity), they are defined as follows: $Sn = TP/(TP + FN)$, $Sp = TP/(TP + FP)$, $Hm = (2 \times Sn \times Sp)/(Sn + Sp)$, where TP, TN, FP and FN represent true positive, true negative, false positive and false negative, respectively. Herein, Sn measures the ratio of actual positives which are correctly identified, Sp measures the ratio of true positives in all predicted positives and Hm is a composite measure used as an aggregated performance score for the evaluation of algorithms (Liu *et al.*, 2013).

3.1 Performance of lncRNA identification

As a comprehensive annotation tool for lncRNA, LncADeep manifests above all in accurately identifying lncRNAs from newly sequenced transcripts. In this subsection, we present the lncRNA identification accuracies of our LncADeep, as well as the performance comparison with current representative tools providing executable programs, including the reference-based tools as COME (Hu *et al.*, 2016), lncScore (Zhao *et al.*, 2016) and lncRScan-SVM (Sun *et al.*, 2015), and the *ab initio* tools as CPC (Kong *et al.*, 2007), CPC2 (Kang *et al.*, 2017), CPAT (Wang *et al.*, 2013), CNCI (Sun *et al.*, 2013), PLEK (Li *et al.*, 2014), longdist (Schneider *et al.*, 2017), lncRNA-MFDL (Fan and Zhang, 2015) and FEElnc (Wucher *et al.*, 2017). To benchmark these tools, it is fair to retrain all

Table 1. Comparison of performances for lncRNA identification by LncADeep and other tools with 10-fold cross validation on human transcripts

Category	Tool	Full-length transcripts[a]			Full- and partial-length transcripts[b]					
					Full- and partial-length[c]			100% partial-length		
		Sn (%)	Sp (%)	Hm (%)	Sn (%)	Sp (%)	Hm (%)	Sn (%)	Sp (%)	Hm (%)
Reference-based	COME	94.3±0.3	96.0±0.3	95.1±0.2	86.9±0.6	**97.2±0.3**	91.8±0.3	86.9±0.6	**96.1±0.4**	91.2±0.3
	lncScore	93.7±0.5	94.3±0.6	94.0±0.3	93.7±0.6	89.7±0.6	91.6±0.4	93.7±0.6	83.5±0.5	88.3±0.4
	lncRScan-SVM	96.5±0.2	93.1±0.4	94.8±0.3	81.8±0.8	94.9±0.4	87.9±0.4	81.8±0.8	91.6±0.3	86.4±0.5
Reference-free	CPC	**98.9±0.1**	84.8±0.6	91.3±0.3	**98.9±0.1**	69.0±0.4	81.3±0.3	**98.9±0.1**	57.5±0.3	72.7±0.3
	CNCI	97.3±0.2	88.2±0.5	92.6±0.2	97.3±0.2	79.3±0.7	87.4±0.5	97.3±0.2	70.1±0.6	81.5±0.4
	CPAT	95.4±0.4	93.6±0.5	94.5±0.3	89.8±0.7	88.0±0.4	88.9±0.5	89.8±0.7	80.6±0.6	85.0±0.6
	CPC2	94.4±0.5	92.4±0.5	93.4±0.4	94.4±0.5	70.1±0.4	80.4±0.4	94.4±0.5	52.5±0.4	67.5±0.4
	FEElnc	96.7±0.3	95.5±0.4	96.1±0.3	92.5±0.4	91.1±0.7	91.8±0.4	92.5±0.4	86.1±0.7	89.2±0.5
	PLEK	98.1±0.2	95.5±0.2	96.8±0.2	98.1±0.2	70.2±0.3	81.8±0.2	98.1±0.2	55.1±0.3	70.5±0.2
	longdist	98.6±0.2	44.5±0.2	61.3±0.2	98.6±0.2	51.4±0.2	67.6±0.2	98.6±0.2	55.1±0.2	70.7±0.2
	lncRNA-MFDL	93.9±0.5	94.8±0.5	94.3±0.4	93.9±0.4	80.8±0.4	86.9±0.3	93.9±0.4	69.4±0.6	79.8±0.5
	LncADeep	98.1±0.2	**97.2±0.3**	**97.7±0.2**	93.8±0.5	94.5±0.4	**94.2±0.3**	93.8±0.5	90.3±0.6	**92.0±0.5**

Note: Bold values indicate the highest value of each metric.

[a]The dataset is from Refseq (full-length mRNAs) and GENCODE (lncRNAs). LncADeep was trained for the model targeting full-length transcripts.

[b]The dataset is from GENCODE (lncRNAs and full- and partial-length mRNAs). LncADeep was trained for the model targeting transcripts including full- and partial-length.

[c]The composition of mRNAs in test set is 65% full-length and 35% partial-length, which matches the composition of full- and partial-length mRNAs in GENCODE dataset. The test sets can be downloaded from the homepage of LncADeep.

programs on the same training set and test against the same test set. Among the above tools, COME, CPAT, lncRScan-SVM, longdist, PLEK and FEElnc provided model-training options, while longdist and PLEK's retraining is very time-consuming as mentioned in users' manual so we did not retrain these programs (Supplementary Table S9). For the tools without available retraining option, namely CPC, CPC2, CNCI, lncRNA-MFDL and lncScore, we used their pre-built models.

We first compared the performance of lncRNA identification on full-length transcripts. Herein we used 10-fold cross validation to construct training and test sets, as illustrated in Supplementary Figure S2. The data is randomly divided into 10 subsets. Of the 10 subsets, one is retained as the test set, and the other nine ones are used as the training set. This process is repeated 10 times so that each subset is used as validation set. Then the averaged results from 10-folds can produce an estimation for the performance of a method. With 10-fold cross validation, LncADeep achieves an average sensitivity of 98.1% and specificity of 97.2%, meanwhile an average harmonic mean of 97.7%. Compared with other tools on the same test sets, LncADeep has the highest harmonic mean, including the highest specificity, and rather a high sensitivity only slightly lower than that of CPC (while CPC's specificity as 84.8% is much lower than LncADeep's as 97.2%, and leading to the harmonic mean of CPC as 91.3% evidently lower than LncADeep's 97.7%). In total, our method has the highest accuracy among all lncRNA identification tools, and certainly outperforms the existing tools. The comparison of identification results by LncADeep and other tools are shown in Table 1a. Also the receiver operating characteristic (ROC) curves of LncADeep and other tools may be referred to Supplementary Figure S3. Among the tools compared with in this study, it should be pointed out that the tools, CNCI, CPC, CPC2, PLEK, longdist, lncScore and lncRNA-MFDL, have not been retrained in the process of 10-fold cross validation. So their identification performances might be over-estimated since their training sets have overlaps with the test sets. In particular, PLEK's pre-built model also used mRNAs from Refseq and lncRNAs from GENCODE for training, which overlapped with the datasets we used, and this is why PLEK achieved relatively high performance, yet lower than that of LncADeep.

We then report the identification performances of LncADeep applied on the transcripts including partial-length mRNAs, which certainly has more application significance for processing real data. Our analysis has demonstrated that the ratio of partial-length mRNAs in the training set affects the lncRNA identification performance on the same test set (Supplementary Fig. S5). So it is more reasonable for the algorithm to train a classifier for a given ratio of partial-length mRNAs within the training set as well as introduce the mathematical model of partial-length transcripts as mentioned in section Materials and methods. To this end, we designed a majority voting strategy as follows. The LncADeep model was first trained on training sets with various ratios of partial-length mRNAs. Herein we trained 21 classifiers, and each with the percentage of full-length mRNAs from 0 to 100% while partial-length mRNAs from 100 to 0% both with a step of 5%. Supplementary Figure S4 displays the details for the construction of training sets for the 21 classifiers. When using these 21 classifiers for lncRNA identification, each classifier gives an output, then the decision will be voted by 21 outputs. Our results showed that majority voting strategy overall outperformed a single classifier trained on a specific ratio of partial-length mRNAs (Supplementary Fig. S5).

We thus compared the 10-fold cross-validation performances of LncADeep and other tools on transcripts including both full- and partial-length. The results were listed in Table 1b, and the test sets in this case were constructed with details in Supplementary Figure S4. It is of reference value to first examine the performance on the test sets with 100% partial-length transcripts. Among all the tools, LncADeep achieves the highest harmonic mean of 92.0%. Although the CPC program has the highest sensitivity of 98.9%, it presents a much lower specificity of 57.5% compared to LncADeep's 90.3%. Similarly, COME achieves the highest specificity of 96.1%, but it does a lower sensitivity of 86.9% compared to LncADeep's 93.8%. Despite a comparable performance herein,

COME actually relies on experimental information and reference genomes, whereas LncADeep is an *ab initio* method which does not require these information and still outperforms COME. Furthermore, we constructed various test sets by randomly combining full- and partial-length mRNAs (Supplementary Fig. S4) and compared the performance of LncADeep and other tools on them. Table 1c illustrates the performance on the test set with a composition 65% full-length and 35% partial-length mRNAs, which matches that of GENCODE dataset, where LncADeep achieves the highest harmonic mean of 94.2% with a noticeable advantage than others. Similarly, CPC reaches the highest sensitivity of 98.9%, but presents a much lower specificity of 69.0% compared to LncADeep's 94.5%. COME has the highest specificity of 97.2%, but a lower sensitivity of 86.9% compared to LncADeep's 93.8%. Figure 2 illustrates five tools, LncADeep, COME, CPAT, FEElnc and lncScore with the best performances, while Supplementary Figure S6 displays the comparison for all tools. Besides, Supplementary Figures S7 and S8 illustrate the ROC curves, where LncADeep achieves the highest AUC, conforming with the above results. It is clear that LncADeep outperforms all other tools consistently no matter how the ratio of partial-length mRNAs varies.

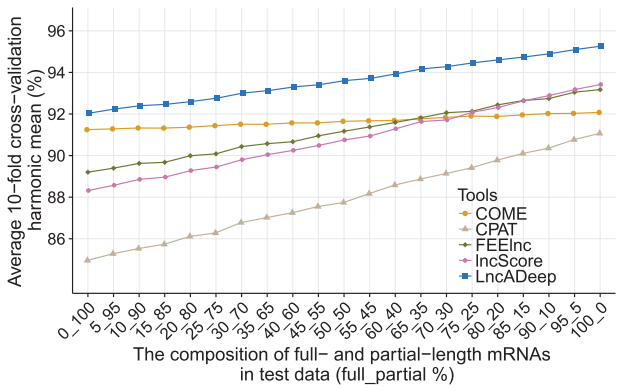

Fig. 2. The performance of lncRNA identification on human transcripts with various compositions of full- and partial-length mRNAs

In summary, with a series of cross-validation tests both for full-length transcripts and partial-length transcripts, we showed that the identification performance of LncADeep is higher than all other tools, even with the over-estimation for several of them trained by data overlapping with test sets.

3.2 Performance of cross-species lncRNA identification

In view of genetic conservation and diversity, it is worth expecting that the lncRNA predictors can be applied well in cross-species identification. To this end, we go to present the performance of cross-species lncRNA identification by LncADeep and other tools, with their algorithms trained on human data. Herein we chose the mouse for cross-species identification, because it is evolutionarily close to human, and moreover there are relatively abundant experimentally verified lncRNAs and mRNAs for mouse. In other words, we used human data as the training set and mouse data as the test set to evaluate the performance of cross-species identification. Since lncRScan-SVM cannot be applied for cross-species identification, we did not include it. We first performed the test against the data for full-length transcripts, the results are shown in Table 2a. Similarly, the ROC curves of LncADeep and the other tools are presented in Supplementary Figure S9. It is clear that LncADeep keeps the highest accuracy (with harmonic mean of 96.7%) of cross-species identification on full-length transcripts (Table 2a). The harmonic mean of COME, CPC and PLEK dropped to <90%. It is also notable that the specificity and harmonic mean of COME dropped to less than 60%, which are much lower than its performance on human data, owing to its highly dependence on the experimental information and reference genome, which might not be conservative across species.

In addition, we conducted cross-species identification for mouse transcripts including full- and partial-length (Table 2b), and LncADeep still achieved the best performance with the highest harmonic mean of 91.2% on the test sets with 100% partial-length transcripts. To comprehensively measure the performance of cross-species lncRNA identification on transcripts including full- and partial-length, we constructed various test sets by randomly combining full- and partial-length mouse mRNAs, where the percentage of full-length mRNAs ranges from 0 to 100% with a step of 5%

Table 2. Comparison of performances for cross-species lncRNA identification by LncADeep and other tools with on mouse transcripts

Category	Tool	Full-length transcripts[a]			Full- and partial-length transcripts[b]					
					Full- and partial-length[c]			100% partial-length		
		Sn (%)	Sp (%)	Hm (%)	Sn (%)	Sp (%)	Hm (%)	Sn (%)	Sp (%)	Hm (%)
Reference-based	COME	98.4	41.1	58.0	98.4	51.8	67.9	98.4	46.5	63.1
	lncScore	93.5	92.4	92.9	93.5	90.9	92.1	93.5	83.6	88.2
Reference-free	CPC	**98.6**	73.3	84.1	**98.6**	70.0	81.9	**98.6**	56.0	71.5
	CNCI	96.2	85.5	90.6	96.2	81.1	88.0	96.2	69.3	80.6
	CPAT	94.9	92.0	93.4	89.9	90.4	90.2	89.9	82.0	85.8
	CPC2	93.9	89.7	91.8	93.9	73.2	82.3	93.9	52.3	67.2
	FEElnc	94.8	92.7	93.7	90.7	91.1	90.9	90.7	84.1	87.3
	PLEK	90.2	75.5	82.2	90.2	66.3	76.4	90.2	50.5	64.8
	longdist	97.6	33.5	49.9	97.6	47.9	64.2	97.6	54.8	70.2
	lncRNA-MFDL	95.9	91.7	93.8	95.9	89.4	92.6	95.9	77.6	85.8
	LncADeep	97.0	**96.3**	**96.7**	95.1	**93.3**	**94.2**	95.1	**87.7**	**91.2**

Note: Bold values indicate the highest value of each metric.

[a]Full-length mouse dataset is composed of 12 529 lncRNAs (from GENCODE) and 29 739 mRNAs (from RefSeq).

[b]Full- and partial-length mouse dataset is composed of 12 529 lncRNAs (from GENCODE) and 56 744 mRNAs (39 079 full-length and 17 665 partial-length, from GENCODE).

[c]The composition of mRNAs in test set is 70% full-length and 30% partial-length, which matches the composition of full- and partial-length mRNAs in GENCODE dataset. The test sets can be downloaded from the homepage of LncADeep.

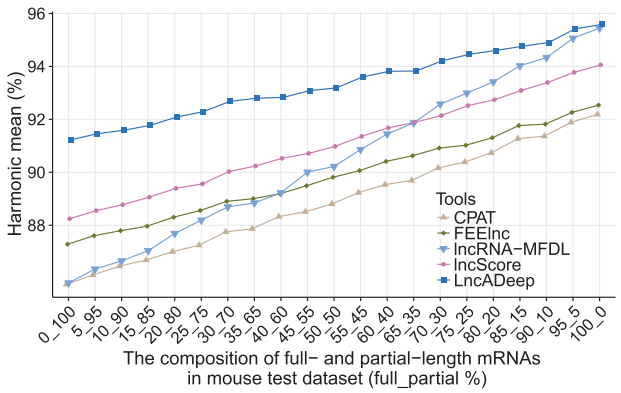

Fig. 3. The performance of cross-species lncRNA identification on mouse transcripts with various compositions of full- and partial-length mRNAs

(Supplementary Fig. S10). Figure 3 displays the five tools with the best performance on cross-species lncRNA identification (Supplementary Fig. S11 plotted all benchmarked tools). Similar to Figure 2, LncADeep consistently outperformed other tools for cross-species lncRNA identification on mouse data. The ROC curves (Supplementary Figs S12 and S13) are also consistent with the above results. To summarize, the results showed that LncADeep also evidently outperforms state-of-the-art tools for cross-species lncRNA identification on mouse transcripts including full- and partial-length.

3.3 Performance of lncRNA–protein interaction prediction and lncRNA functional annotation

For LncADeep as a comprehensive annotation pipeline, it is essential, and last but not least, to provide functional interpretation for lncRNAs identified in the preceding step. This is usually proceeded with predicting which protein is involved into a lncRNA's interaction, and then inferring the lncRNA's functional role. Herein, we first report the performance of lncRNA–protein interaction prediction by LncADeep. To benchmark the performance, we compared LncADeep with several state-of-the-art tools for predicting RNA–protein interactions, i.e. lncPro (Lu *et al.*, 2013), RPISeq (Muppirala *et al.*, 2011), RPI-pred (Suresh *et al.*, 2015), rpiCool (Akbaripour-Elahabad *et al.*, 2016) and IPMiner (Pan *et al.*, 2016). Among these tools, we could retrain the IPMiner tool since it provides retraining option. But for the others, we could only use the tools (lncPro and rpiCool) directly or online servers (RPISeq and RPI-pred) as provided. Considering the unavailability of stand-alone tools of catRAPID (Bellucci *et al.*, 2011) and GlobalScore (Cirillo *et al.*, 2016), we tried to submit lncRNA–protein pairs online and download the results, however, it was time-consuming and could not be conducted in large scale. Therefore, we did not include catRAPID and GlobalScore for benchmarking.

As shown in Table 3, with 5-fold cross validation, LncADeep demonstrates an average sensitivity of 97.0%, slightly lower than the highest sensitivity of 99.1% by RPISeq with RF mode, and specificity of 85.4% on average, slightly lower than the highest specificity of 85.6% by IPMiner. However, for the total performance of harmonic mean, LncADeep reaches up to 90.8% on average, evidently outperforming RPISeq with RF mode (66.5%), IPMiner (87.6%) and all other tools (Table 3). Note that the specificities and the harmonic means of lncPro, RPISeq and RPI-pred were much lower than that of rpiCool and IPMiner, suggesting that they tend to predict much more lncRNA–protein pairs as interacting ones. Nevertheless, LncADeep achieves the best performance on the prediction of lncRNA–protein interaction when compared with both rpiCool and IPMiner.

Table 3. Comparison of performances for predicting lncRNA–protein interaction by LncADeep and other tools with 5-fold cross validation

Tools	Sn (%)	Sp (%)	Hm (%)
lncPro	80.3±0.9	52.2±0.6	63.2±0.4
RPISeq (RF)	**99.1±0.2**	50.1±0.1	66.5±0.1
RPISeq (SVM)	93.5±0.7	50.2±0.2	65.3±0.4
RPI-pred	88.0±0.3	49.8±0.6	63.6±0.5
rpiCool	92.0±0.8	83.3±0.8	87.5±0.6
IPMiner	89.8±1.1	**85.6±0.7**	87.6±0.6
LncADeep	97.0±0.5	85.4±0.8	**90.8±0.4**

Note: RPISeq has two prediction modes, SVM and random forests (RF). Bold values indicate the highest value of each metric.

As described in section Materials and methods, LncADeep was designed to integrate KEGG and Reactome pathway enrichment analysis and functional module detection to infer all lncRNAs' functional information based on their predicted interacting proteins. It should be emphasized that LncADeep is the first tool known to automatically predict lncRNA–protein interactions as well as provide significative functional annotations for lncRNAs. So, as an *ab intio* comprehensive annotation tool (Supplementary Fig. S14), with sequenced transcripts input, LncADeep provides not only identified lncRNAs with the best performance, but also predicted lncRNA–protein interactions more accurately, and moreover, the lncRNAs' functional interpretation. Herein we try to discuss LncADeep's performance of functional inference, although there are not any tools known to implement this job yet. We used LncADeep to annotate the functions for all 27 384 lncRNAs collected in GENCODE Release 24. The functional annotations including KEGG and Reactome pathways and functional modules are provided as Supplementary files which can be downloaded online (http://cqb.pku.edu.cn/ZhuLab/lncadeep/). Here, we focused on KEGG and Reactome pathways, which are more explicit for annotating functions compared to functional modules (Supplementary Section S6). Consequently, LncADeep annotated 702 675 associations with 140 KEGG pathways and 1 839272 associations with 422 Reactome pathways for the 27 384 lncRNAs, with an average of 25 KEGG and 67 Reactome pathways associated with each lncRNA, conforming the complexity of lncRNA functions (Chu *et al.*, 2015; Guttman and Rinn, 2012; Yang *et al.*, 2015). Since IPMiner ranked next to LncADeep on lncRNA–protein interaction prediction, as a comparison, we used IPMiner (Pan *et al.*, 2016) (its release version) to predict interacting proteins for lncRNAs (IPMiner is very time-consuming and we cannot used it for all the 27 384 lncRNAs, and thus we randomly sampled 100 lncRNAs) and then conducted KEGG and Reactome pathway enrichment analysis with the predicted interacting proteins. For these 100 randomly sampled lncRNAs, LncADeep annotated 2555 associations with 78 KEGG pathways and 6697 associations with 216 Reactome pathways, with an average of 25 KEGG and 67 Reactome pathways associated with each lncRNA. In contrast, using IPMiner-predicted interacting proteins, we only obtained 473 associations with 37 KEGG pathways and 2840 associations with 84 Reactome pathways, with an average of only 5 KEGG and 28 Reactome pathways associated with each lncRNA, which were much smaller than that of LncADeep, at least indicating that LncADeep can give more detailed functional annotations for lncRNAs. Therefore, as LncADeep outperformed IPMiner on predicting lncRNA–protein interactions, LncADeep is expected to provide evidently better functional annotations.

For lack of a gold standard dataset for lncRNA functions, it is difficult to quantitatively evaluate the performance of inferred functions of lncRNAs. Therefore, to demonstrate that LncADeep can provide helpful suggestions on lncRNA functioning, we took four well-studied lncRNAs as examples through comparing their inferred functions (by LncADeep and IPMiner) with the reported functions from literatures (please see Supplementary Section S6). Examples showed that LncADeep can give informative functional annotations which well conform the known functions of lncRNAs and evidently outperform IPMiner. For lncRNAs whose functions remain unclear, we have demonstrated that our LncADeep makes a distinguishing effort to turn the situation to infer their functions and provide helpful hints for biologists.

4 Discussion

We have shown overall high performance of LncADeep on a suite of automated annotations for lncRNAs. In addition, we have benchmarked the run time of LncADeep and other tools under the same computing environment (Supplementary Table S9). For lncRNA identification, the time cost of LncADeep was competitive among state-of-the-art tools (Supplementary Section S7). For predicting lncRNA–protein interaction, LncADeep was the fastest among the tools which can predict interactions in large scale (Supplementary Section S8). Consequently, it is expected to be of significance for the research community to conduct RNA-seq data curation and to understand lncRNA-associated meaning. Clearly, the mathematical modeling for lncRNA-related features and the optimization schemes by deep learning algorithm facilitated the success in this study. We expect that LncADeep can not only accurately identify lncRNAs but also offer informative functional annotations for lncRNAs.

For the features used in the model targeting full-length transcripts, we used the EDP of ORF, whose dimension accounts for over two-thirds of the total feature dimension. The success of LncADeep on lncRNA identification again supported the hypothesis that protein-coding and non-coding ORFs have different distributions in the EDP phase space, which can be caused by various selection pressures during evolution (Liu *et al.*, 2013; Zhu *et al.*, 2007). Although most lncRNA identification tools can reach an accuracy over 90% on full-length transcripts, LncADeep has achieved the highest one with 97.7%. Actually, a tiny improvement of the accuracy is not a trivial matter: since there are a large number of lncRNAs, 1% improvement on the accuracy indicates hundreds of correctly identified lncRNAs. With the development of sequencing technology, the third-generation sequencing technologies are emerging and being prevalent. The length of reads sequenced by the third-generation sequencing technologies suffices for sequencing full-length transcripts, where LncADeep achieves high accuracy on lncRNA identification.

In this study, we developed our model addressing the issue of partial-length transcripts, which is a key point that has been overlooked for a long term by most of the current tools. To distinguish lncRNAs from mRNAs, one challenge is to avoid misclassifying partial-length mRNAs into lncRNAs. So we proposed the LCDS-based model for a partial-length mRNA containing partial CDS, which targets on transcripts including partial-length. Supplementary Section S9 showed that LCDS had comparable performance with LORF on detecting CDS on full-length mRNAs, moreover outperformed LORF on partial-length mRNAs. The reason is, LCDS incorporates the hexamer-based CDS, which will not be affected even if the start and stop codons of mRNAs are missed. We also evaluated the performance of various features in lncRNA identification. Results showed that using only LCDS-related features, the harmonic mean of lncRNA identification on partial-length transcripts reached up to 89.9% (Supplementary Section S10), while that of most tools was lower than 89.5% (Table 1b), demonstrating the discriminative power of LCDS. Actually in real data, the composition of full- and partial-length mRNAs is unknown, then it is not appropriate to train a specific classifier for any given real dataset. Therefore, we proposed to use majority voting, and results have shown the effectiveness of our strategy (Figs 2 and 3). Moreover, even without majority voting, LncADeep still showed better performance than other tools (Supplementary Fig. S15). Besides, Ji *et al.* (2015) suggested that many annotated lncRNAs could have coding potential and thus it is important to avoid false positives for lncRNA identification. The LncADeep tool has outperformed other *ab intio* tools with the highest specificity and harmonic mean (Tables 1 and 2), showing its advantages and underlining the importance of identification accuracy of lncRNAs.

In addition to lncRNA identification, LncADeep predicts lncRNA–protein interactions, outperforming state-of-the-art tools and showing several advantages. For instance, LncADeep uses mRMR for feature selection and keeps only the most discriminative features, which helps to reduce overfitting. Besides, LncADeep integrates both sequence and structure features, which is more robust than using only one kind of feature (Supplementary Section S11). Furthermore, LncADeep is a user-friendly stand-alone tool which can predict interactions in large scale. In contrast, RPISeq, RPI-pred and rpiCool cannot be used for large scale interaction prediction (Supplementary Section S8). Finally, as a lncRNA functional annotation tool, LncADeep offers annotations automatically for lncRNAs, and case studies have shown that LncADeep can provide helpful functional annotations.

The performance of LncADeep has demonstrated the effectiveness of deep learning methods, which are capable of learning sophisticated hidden structures in data. In LncADeep, we used DBN for lncRNA identification and DNN for predicting lncRNA–protein interaction. The performance of LncADeep indicates that DBN and DNN are powerful deep learning architectures. Compared with the shallow machine learning architectures (such as SVM and logistic regression), deep learning methods can gradually integrate simple features into complex features, which are better at discovering intricate hidden structures in high-dimensional data and achieve more suitable hierarchical representations (Min *et al.*, 2016). However, LncADeep can also suffer some drawbacks on predicting lncRNA–protein interactions. In fact, the training dataset is relatively small, which covers only a small amount of interacting lncRNA–protein pairs. To make full use of the capability of deep learning, more training data is required to reveal the hidden interacting mechanisms between lncRNA and protein. In addition, generating non-interacting lncRNA–protein pairs by pairing lncRNAs and proteins and excluding known interacting pairs can be of risk, as some lncRNA–protein pairs can be potential unverified interacting pairs, which might bias the trained model. However, it is difficult to verify non-interacting lncRNA-proteins, because to some extent, the interaction between lncRNA and protein can only be verified rather than excluded. Moreover, the function annotation of lncRNAs might also be biased since it is dependent on the predicted interacting proteins of lncRNAs. Nevertheless, inferring the functions of lncRNAs through its interacting proteins provides an alternative approach to investigate lncRNAs and can help to offer biological insights. In future work, we plan to collect more experimentally verified interacting lncRNA–protein pairs and tackle the problem related to

non-interacting pairs. Since LncADeep is readily adapted, we expect to give more informative functional annotations for lncRNAs with additional training datasets.

To sum up, we developed a novel lncRNA identification and functional annotation tool, LncADeep, based on deep learning algorithms. With the reconstructed transcripts input, LncADeep can identify lncRNAs, predict lncRNA–protein interactions and provide functional annotations (including enriched KEGG, Reactome pathways and functional modules) for lncRNAs. LncADeep outperformed state-of-the-art lncRNA identification tools on both full-length transcripts and transcripts including partial-length. This advantage endows the program more application significance for processing real data for RNA-seq community. Furthermore, LncADeep outperformed state-of-the-art tools on predicting lncRNA–protein interactions. With the predicted lncRNA–protein interactions, LncADeep provides rich functional annotations, conforming with the known functions, for lncRNAs. To our best knowledge, LncADeep is the first tool that can identify lncRNAs, predict lncRNA–protein interactions, and moreover provide functional annotations for lncRNAs. Currently, there are still large amounts of lncRNAs to be identified, while the functions of most lncRNAs remain unclear. We expect that the LncADeep tool can thus contribute to identifying and annotating novel lncRNAs, and providing helpful functional information for investigating the associations among lncRNAs, gene regulation and diseases, and then facilitate the large-scale automatic genome annotation.

Acknowledgements

We would like to thank Dr. Po-yen Wu and especially Dr. Chanchala Kaddi of Georgia Institute of Technology for their helpful advices to the writing improvement.

Funding

This work was supported by the National Key Research and Development Program of China (2017YFC1200205), the National Natural Science Foundation of China (31671366 and 91231119) and the Special Research Project of 'Clinical Medicine + X' by Peking University.

Conflict of Interest: none declared.

References

Achawanantakun,R. *et al*. (2015) LncRNA-id: long non-coding RNA identification using balanced random forests. *Bioinformatics*, **31**, 3897–3905.

Akbaripour-Elahabad,M. *et al*. (2016) rpiCOOL: a tool for in silico RNA–protein interaction detection using random forest. *J. Theor. Biol*., **402**, 1–8.

Alanis-Lobato,G. *et al*. (2016) HIPPIE v2. 0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Res*., **45**, D408–D414.

Bellucci,M. *et al*. (2011) Predicting protein associations with long noncoding RNAs. *Nat. Methods*, **8**, 444–445.

Bentley,J. (1984) Programming pearls: algorithm design techniques. *Commun. ACM*, **27**, 865–873.

Cao,R. *et al*. (2016) DeepQA: improving the estimation of single protein model quality with deep belief networks. *BMC Bioinformatics*, **17**, 495.

Chen,L. *et al*. (2015) Trans-species learning of cellular signaling systems with bimodal deep belief networks. *Bioinformatics*, **31**, 3008–3015.

Chu,C. *et al*. (2015) Technologies to probe functions and mechanisms of long noncoding RNAs. *Nat. Struct. Mol. Biol*., **22**, 29–35.

Cirillo,D. *et al*. (2016) Quantitative predictions of protein interactions with long noncoding RNAs. *Nat. Methods*, **14**, 5–6.

Croft,D. *et al*. (2014) The reactome pathway knowledgebase. *Nucleic Acids Res*., **42**, D472–D477.

Deng,L. *et al*. (2012) Scalable stacking and learning for building deep architectures. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, 2012*. IEEE.

Derrien,T. *et al*. (2012) The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*., **22**, 1775–1789.

Enright,A.J. *et al*. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*., **30**, 1575–1584.

Fan,X.N. and Zhang,S.W. (2015) lncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning. *Mol. Biosyst*., **11**, 892–897.

Fatica,A. and Bozzoni,I. (2014) Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet*., **15**, 7–21.

Gupta,R.A. *et al*. (2010) Long non-coding RNA hotair reprograms chromatin state to promote cancer metastasis. *Nature*, **464**, 1071–1076.

Guttman,M. and Rinn,J.L. (2012) Modular regulatory principles of large non-coding RNAs. *Nature*, **482**, 339–346.

Harrow,J. *et al*. (2012) GENCODE: the reference human genome annotation for the encode project. *Genome Res*., **22**, 1760–1774.

Hinton,G.E. *et al*. (2006) A fast learning algorithm for deep belief nets. *Neural Comput*., **18**, 1527–1554.

Hu,L. *et al*. (2016) COME: a robust coding potential calculation tool for lncRNA identification and characterization based on multiple features. *Nucleic Acids Res*., **45**, e2.

Ji,Z. *et al*. (2015) Many lncRNAs, 5' UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife*, **4**, e08890.

Kanehisa,M. *et al*. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res*., **38**, D355–D360.

Kang,Y.-J. *et al*. (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res*., **45**, W12–W16.

Kong,L. *et al*. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*., **35**, W345–W349.

LeCun,Y. *et al*. (2015) Deep learning. *Nature*, **521**, 436–444.

Li,A. *et al*. (2014) PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics*, **15**, 311.

Liu,X.-Y. *et al*. (2009) Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man. Cybern. B Cybern*., **39**, 539–550.

Liu,Y. *et al*. (2013) Gene prediction in metagenomic fragments based on the svm algorithm. *BMC Bioinformatics*, **14**, S12.

Lorenz,R. *et al*. (2011) ViennaRNA package 2.0. *Algorithms Mol. Biol*., **6**, 26.

Lu,Q. *et al*. (2013) Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genomics*, **14**, 651.

McHugh,C.A. *et al*. (2014) Methods for comprehensive experimental identification of RNA-protein interactions. *Genome Biol*., **15**, 203.

Min,S. *et al*. (2016) Deep learning in bioinformatics. *Brief. Bioinform*., **18**, 851–869.

Muppirala,U.K. *et al*. (2011) Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics*, **12**, 489.

Pan,X. *et al*. (2016) IPMiner: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. *BMC Genomics*, **17**, 1.

Peng,H. *et al*. (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell*., **27**, 1226–1238.

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

Schneider,H.W. *et al*. (2017) A support vector machine based method to distinguish long non-coding RNAs from protein coding transcripts. *BMC Genomics*, **18**, 804.

Srivastava,N. *et al*. (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res*., **15**, 1929–1958.

Steijger,T. *et al*. (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods*, **10**, 1177–1184.

Sun,L. *et al*. (2013) Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res*., **41**, e166.

Sun,L. *et al*. (2015) lncRScan-SVM: a tool for predicting long non-coding RNAs using support vector machine. *PLoS One*, **10**, e0139654.

Suresh,V. *et al*. (2015) RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. *Nucleic Acids Res*., **43**, 1370–1379.

Ulitsky,I. and Bartel,D.P. (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell*, **154**, 26–46.

UniProtConsortium. *et al*. (2011) Reorganizing the protein space at the universal protein resource (uniprot). *Nucleic Acids Res*., **40**, D71–D75.

Wang,L. *et al*. (2013) CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res*., **41**, e74.

Wucher,V. *et al*. (2017) FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res*., **45**, e57.

Yang,Y. *et al*. (2015) Unveiling the hidden function of long non-coding RNA by identifying its major partner-protein. *Cell Biosci*., **5**, 1.

Yuan,J. *et al*. (2014) NPInter v2.0: an updated database of ncRNA interactions. *Nucleic Acids Res*., **42**, D104–D108.

Zhao,J. *et al*. (2016) lncScore: alignment-free identification of long noncoding RNA from assembled novel transcripts. *Sci. Rep*, **6**, 34838.

Zhu,H. *et al*. (2007) MED: a new non-supervised gene prediction algorithm for bacterial and archaeal genomes. *BMC Bioinformatics*, **8**, 97.