Briefings in Bioinformatics, 22(3), 2021, 1–6

doi: 10.1093/bib/bbaa200
Problem Solving Protocol

NAMS webserver: coding potential assessment and functional annotation of plant transcripts

Kun Sun, Huating Wang and Hao Sun

Corresponding authors: Kun Sun, Shenzhen Bay Laboratory, Shenzhen 518132, China. Tel.: +86-0755-2641-9310; Fax: +86-755-8696-7710. E-mail: sunkun@szbl.ac.cn; Hao Sun, Department of Chemical Pathology, The Chinese University of Hong Kong, Hong Kong SAR 999077, China. Tel.: +852-3763-6048; Fax: +852-2636-5090. E-mail: haosun@cuhk.edu.hk

Abstract

Recent advances in transcriptomics have uncovered lots of novel transcripts in plants. To annotate such transcripts, dissecting their coding potential is a critical step. Computational approaches have been proven fruitful in this task; however, most current tools are designed/optimized for mammals and only a few of them have been tested on a limited number of plant species. In this work, we present NAMS webserver, which contains a novel coding potential classifier, NAMS, specifically optimized for plants. We have evaluated the performance of NAMS using a comprehensive dataset containing more than 3 million transcripts from various plant species, where NAMS demonstrates high accuracy and remarkable performance improvements over state-of-the-art software. Moreover, our webserver also furnishes functional annotations, aiming to provide users informative clues to the functions of their transcripts. Considering that most plant species are poorly characterized, our NAMS webserver could serve as a valuable resource to facilitate the transcriptomic studies. The webserver with testing dataset is freely available at http://sunlab.cpy.cuhk.edu.hk/NAMS/.

Key words: transcriptomics; RNA sequencing; long non-coding RNA; conservation; secondary structure

Introduction

Recent advance in transcriptomics (e.g. whole transcriptome sequencing technique) have uncovered large amounts of novel transcripts, a high proportion of which are long non-coding RNAs (lncRNAs), which are larger than 200 nucleotides and lack coding potential [1–4]. Studies have shown that lncRNAs are functional and many of them serve as important regulators. For instance, we previously had discovered various lncRNAs (e.g. Linc-Yy1) and demonstrated their interactions with transcription factors during myogenesis [5–10]. LncRNAs are also reported in plants and involved in various pathways, e.g. they regulate the splicing of the coding mRNAs and play roles in response to stress [11–17]. These studies demonstrate that most of the novel transcripts could be functional thus deserve further investigations.

Besides the inspiring progress in the past years, however, it is still complex and challenging to infer the functions of newly identified transcripts [18–22]. One important step toward functional annotation is to differentiate lncRNAs from mRNAs, as coding potential directly affects the design of downstream experiments [23]. Various computational algorithms have been

Kun Sun is an Assistant Professor in Shenzhen Bay Laboratory and holds a PhD degree in Bioinformatics and Genomics. His main research interests include long non-coding RNA identification and functional predictions in various species, data mining and visualizations.
Huating Wang is a Professor at the Department of Orthopaedics and Traumatology, The Chinese University of Hong Kong and holds a PhD degree in Molecular Biology. Her research interests mainly focus on the functional roles of non-coding RNAs in regulating gene expression in skeletal muscle stem cells and muscle regeneration.
Hao Sun is a Professor in Bioinformatics and Statistics at the Department of Chemical Pathology, The Chinese University of Hong Kong and holds a MSc degree in Computer Science as well as a PhD degree in Environmental Chemistry. His main interests include understanding the fundamental aspects of transcriptional regulation of both coding and non-coding genes using integrative approaches.
Submitted: 7 June 2020; **Received (in revised form):** 23 July 2020

developed and proven reliable in this task, most of which utilize supervised machine learning methods (e.g. support vector machine) to train classification models based on various features extracted from a list of known coding and non-coding transcripts [4, 24–27]. However, since most of the plant species are poorly characterized, current tools are usually optimized for mammals. For example, previously we had developed iSeeRNA [4], which showed high accuracy for human and mouse species, but it is not suitable for plants, as it relies on cross-species conservation, while currently, the lack of high-quality reference genomes for most plant species makes it infeasible to deduce reliable conservation scores. A handful of software claim applicability in plants, such as CPC2 [24] and CNIT [25], the successors of CPC [26] and CNCI [27], respectively; however, the performance of these tools have only been tested on a limited number of species. In this work, we present NAMS (non-coding assessment of Magnoliophyta species) webserver, which provides online coding potential assessment specifically designed for plants. We have evaluated NAMS against state-of-the-art tools using a comprehensive dataset with more than 3 million transcripts from various plant species to demonstrate performance advantages of our classifier. Our webserver further provides functional annotations to help the users to find clues to the functions of their transcripts. NAMS webserver with testing dataset is freely available at http://sunlab.cpy.cuhk.edu.hk/NAMS/.

Methods

Figure 1 shows the schematic workflow of our NAMS webserver. The webserver takes the transcript sequences provided by the users as input, then it calculates the coding potential and performs functional annotations of the query transcripts. Currently, the webserver provides three classifiers: NAMS, a newly developed tool in this work, as well as CPC2 and CNIT, tools that have been demonstrated to be capable of processing plant transcripts. Supervised machine learning approaches are widely used in coding potential classifications; however, currently plant lncRNAs are poorly characterized and a high proportion of annotated ones originate from a limited number of model species (e.g. Arabidopsis [28]), which means supervised classifiers could be biased to the well-annotated species as they contribute the majority of transcripts in the training dataset. As a result, we have developed two modules for NAMS: the NAMS-SVM module is based on support vector machine (similar to iSeeRNA and CPC2) [29] and the NAMS-DT module is implemented using a Decision-Tree algorithm. In the current implementation of NAMS (Supplementary Figure S1), the key features are open reading frame (ORF) and homologue search (using blast software [30] against annotated protein sequences of Magnoliophyta species in the UniProt database [31]), which are both well-recognized features in this task [4, 26]. Intrinsic sequence features, e.g. Fickett TESTCODE score [32], have been proven informative and are heavily utilized in both CPC2 and CNIT; however, we find that TESTCODE scores show a systematic bias among various species. For example, TESTCODE scores for mRNAs are generally higher than lncRNAs in all species investigated in Supplementary Figure S2, but the scores for rice lncRNAs are comparable to Arabidopsis mRNAs while much higher than human mRNAs. We thus doubt that such sequence features may consist species-specific bias and do not use them in the current work. Briefly, in NAMS-SVM, we randomly selected 2000 well-annotated transcripts (1000 coding and 1000 non-coding) from Arabidopsis (similar to CPC2 and CNIT which train the model using Arabidopsis transcripts while find that the model works on most plant species) [28] using the ORF size and homologue search results (E-value, a statistical parameter accounting for the significance of the hit, and high-scoring segment pair score (HSP), a measurement of the biological relevance of the hit) as the features to train a classification model; while in NAMS-DT, for a given transcript, if its ORF is longer than 120 aa (amino acids), NAMS-DT checks the best hit during the homologue search: if the E-value is lower than 0.001 or the HSP score is higher than 200, the transcript will be classified as 'coding'; otherwise it will be reported as 'uncertain'; on the other hand, for a transcript whose ORF is shorter than 120 aa, if the E-value is larger than 0.001 or the HSP score is lower than 200, it will be classified as 'non-coding'; otherwise an 'uncertain' will be reported (Supplementary Figure S1). The thresholds used by NAMS-DT are carefully selected based on previous knowledge, tradeoff between sensitivity and specificity, as well as to minimize the proportion of 'uncertain' results (Supplementary Figure S3). Based on our evaluations, 'uncertain' results account for around 5–10% of the transcripts (notably, we find that CNIT also provides blank prediction results for some transcripts).

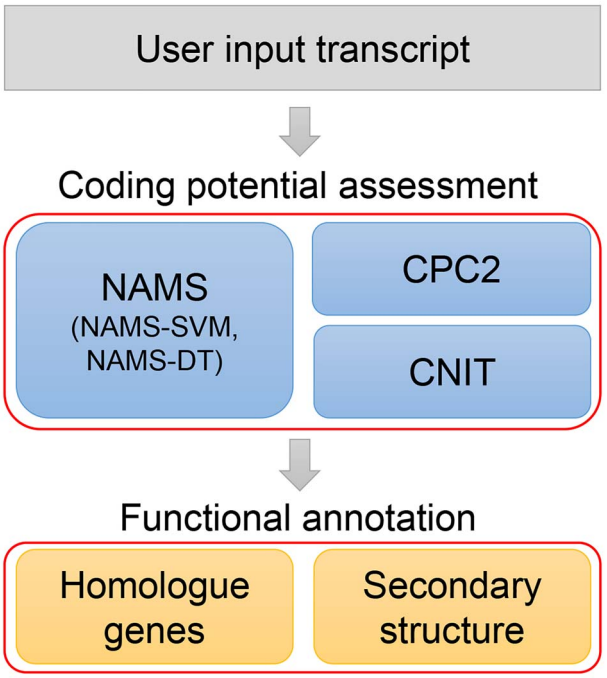

Figure 1. Schematic workflow of NAMS webserver. The webserver consists of two components: coding potential assessment and functional annotation of plant transcripts. The webserver takes the transcript sequences from the users as input, and predicts the coding potential using three classifiers: NAMS (which contains two modules: NAMS-DT and NAMS-SVM), CPC2 and CNIT. After that, the webserver further performs functional annotations of the query transcripts, including homologue search results and 2D secondary structure prediction.

To use the webserver, users are only required to submit their transcripts in FASTA format. The predicted results will be returned within a few seconds after submission. All the three classifiers are called to perform coding potential classification for the query transcripts. Based on our evaluations (see Results section), NAMS-DT shows the best performance, therefore, the webserver uses its prediction as to the primary classification result; for the transcripts that could not be classified by NAMS-DT (i.e. 'uncertain'), the webserver uses the prediction by CPC2

Table 1. Performance of NAMS on the well-annotated testing dataset

Species	Category	Number of transcripts	NAMS-DT accuracy (%)	NAMS-SVM accuracy (%)
Arabidopsis	mRNA	11 975	95.4	97.6
	lncRNA	1233	99.7	98.9
Maize	mRNA	286	95.0	98.3
	lncRNA	2534	99.6	98.9
Potato	mRNA	67	98.5	100.0
	lncRNA	978	98.9	95.0
Rice	mRNA	1506	95.7	96.8
	lncRNA	3726	85.4	83.4
Tomato	mRNA	130	90.6	94.6
	lncRNA	2932	99.8	99.4
Overall	mRNA	13 964	95.4	97.5
	lncRNA	11 403	95.1	93.6

as the final result. Besides coding potential assessment, the webserver also provides functional annotations to the user-supplied transcripts: for those predicted to be 'coding,' the potential homologue proteins with pairwise amino acid sequence alignment will be listed; otherwise, the 2D secondary structure (calculated by RNAfold software [33]) will be shown to the users.

Results

Accuracy of NAMS on well-annotated datasets

We first evaluated the performance of our NAMS classifier using well-annotated plant datasets. For NAMS-SVM, during training, it shows a 10-fold cross-validation accuracy of 99.1%; receiver operating characteristic (ROC) analysis reveals an area under the ROC curve (AUC) value larger than 0.999 (Supplementary Figure S4); in addition, the accuracies on classifying mRNAs (i.e. sensitivity) and non-coding RNAs (i.e. specificity) are 98.8% and 99.6%, respectively, demonstrating very high performance of NAMS-SVM on the training dataset. We then evaluated the performance of NAMS-SVM and NAMS-DT on a testing dataset obtained from CPC2 project. Notably, only RNAs longer than 200 bp are kept in the analysis, resulting in 13 964 mRNAs and 11 403 lncRNAs from five plant species [Arabidopsis (which transcripts have no overlap with the training dataset), Maize, Potato, Rice, Tomato]. Both NAMS-SVM and NAMS-DT modules show high performance on this dataset (Table 1). For example, NAMS-DT shows an overall accuracy of 95.4 and 95.1% in classifying mRNAs and lncRNAs, respectively; the median accuracies on mRNAs and lncRNAs across the five species are 95.4 and 99.6%, respectively, which performance is similar to CPC2 on this dataset. For NAMS-SVM, ROC analysis further reveals AUC values range from 0.973 to 0.999 (median: 0.998) among the five species (Supplementary Figure S4). These results demonstrate that NAMS is highly reliable in coding potential prediction of common plant species.

Comprehensive evaluation of NAMS and current tools

To comprehensively evaluate the performance of NAMS and current tools, annotated plant transcripts from Phytozome (v12) [34] and GreeNC (v1.12) [35] databases are collected. As a result, this testing dataset contains more than 3.2 million transcripts: the Phytozome database presents ~3 million mRNAs from 68 species and the GreeNC database presents ~200 thousand lncRNAs from 45 species. In addition, we also applied CPC2 and CNIT, two state-of-the-art classifiers, on this comprehensive dataset for performance comparisons. The summary of the evaluation results is shown in Table 2, and the detailed information could be found in Supplementary Table S1. Considering that this dataset consists of various species, we calculated both the overall accuracy (using all transcripts) as well as quantiles (including median) of accuracies across the species. Briefly, on the Phytozome transcripts, the performance is similar among NAMS, CPC2 and CNIT (Table 2). On the GreeNC transcripts, however, NAMS-DT module shows significantly higher accuracies than the others: the overall accuracies are 93.6, 68.9, 83.9 and 55.0%, respectively, for NAMS-DT, NAMS-SVM, CPC2 and CNIT. In addition, the median accuracy of NAMS-DT is also remarkably higher (96.3%) compared to NAMS-SVM (81.8%), CPC2 (80.8%) and CNIT (53.9%) among the 45 species in the GreeNC dataset. Furthermore, NAMS shows similar F-measure scores [36] (both NAMS-DT and NAMS-SVM are 0.929) while higher Matthews correlation coefficient (MCC) values [37] (NAMS-DT: 0.495, NAMS-SVM: 0.384) compared to CPC2 (F-measure score: 0.916; MCC value: 0.423) and CNIT (F-measure score: 0.943; MCC value: 0.359). ROC analysis further shows that NAMS-SVM performs slightly lower than CPC2 while higher than CNIT on this comprehensive dataset (Supplementary Figure S5). The performance of CNIT is not as excellent as that demonstrated in their original work. We think that the inconsistency may due to that CNIT is previously tested using the transcripts in Ensembl database [38], where the transcript sequences might be more complete than that in GreeNC database, as many species in GreeNC database have not been well-characterized. The results on this comprehensive testing dataset thus demonstrate the high accuracy and improved performance of NAMS, especially the NAMS-DT module in the classification of lncRNAs.

NAMS webserver and example outputs

To facilitate the usage of NAMS and provide an easy-to-access resource for the plant research society, we have set up a webserver freely available to all users. The users are only required to submit the sequences of their transcripts in FASTA format for prediction, and the result will be shown within a few seconds. Notably, the webserver will keep the result for each online query for 1 month, which allows the users to bookmark it for future reviews or share it with their colleagues. Two well-annotated transcripts are provided as examples to illustrate the usage and output of NAMS webserver in Figure 2, and more examples are provided in Supplementary Figure S6. The first one is a coding transcript, NM_101677.1 (*Arabidopsis thaliana* putative FKBP-type peptidyl-prolyl cis-trans isomerase 5). As

Table 2. Performance of NAMS and current tools on the comprehensive testing dataset

	Phytozome (mRNA)				GreeNC (lncRNA)			
	NAMS-DT	NAMS-SVM	CPC2	CNIT	NAMS-DT	NAMS-SVM	CPC2	CNIT
Overall	87.2	88.5	85.4	92.1	93.6	68.9	83.9	55.0
Quantile 1	85.7	84.1	86.3	90.7	93.3	52.8	77.1	48.0
Median	88.1	88.4	88.9	93.0	96.3	81.8	80.8	54.1
Quantile 3	90.1	92.5	90.9	94.6	97.8	91.7	87.5	59.5

Notes: All values are shown in percentages; Phytozome database contains ~3 million mRNAs from 68 species, and GreeNC databases contains ~200 thousand lncRNAs from 45 species. The prediction accuracies are calculated for each species and summarized in this table.

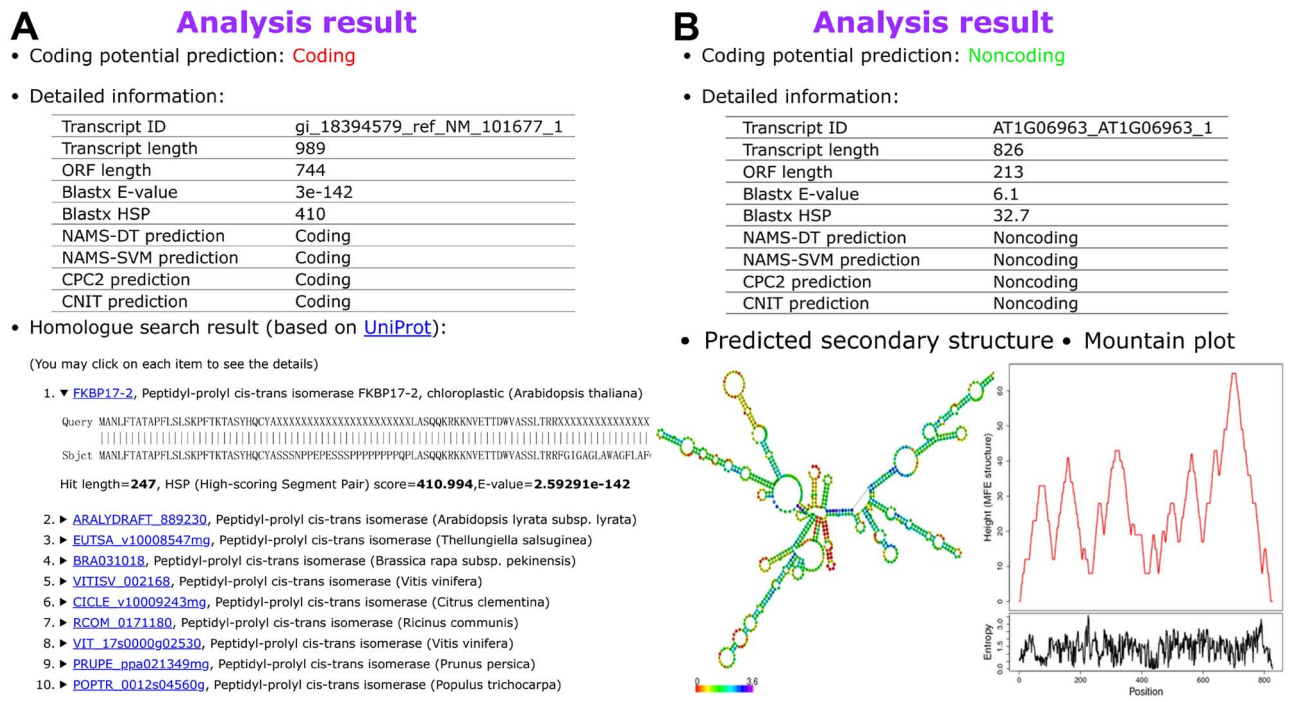

Figure 2. Snapshots of the NAMS webserver. Example outputs of an annotated (A) coding and (B) non-coding transcript.

shown in Figure 1A, the first part of the result page is the coding potential prediction, where the query transcript is predicted as 'coding' (highlighted in red). The second part is the detailed information for classification including the ORF size and key statistics (i.e. E-value and HSP score) in homologue search. For query transcripts predicted as coding, the webserver shows the top 10 hits during the homologue search. Users could click the arrows to view the detailed information, which contains a pairwise alignment and statistics when measuring the similarity of the two sequences (Figure 2A). In this case, it is very likely that the query transcript belongs to a gene that is homologue to known peptidyl-polyl *cis-trans* isomerases in various plant species. The other example is AT1G06963.1, which is a lncRNA in *Arabidopsis thaliana* [28] and predicted to be 'non-coding' by NAMS (highlighted in green; Figure 2B). For such a non-coding transcript, the webserver provides the 2D secondary structure along with a mountain plot (Figure 2B). This information (e.g. the harpins) could help the users find out the potential functional domains for downstream mechanism studies. Considering that most plant species are poorly annotated, we believe that our annotations could assist the users to find valuable clues to the functions of their transcripts.

Discussion

The development of high-throughput whole transcriptome sequencing technology has revealed millions of novel transcripts in the past few years. To identify the coding potential of these transcripts, various computational predictors have been developed and demonstrated high accuracy. However, most of the current classifiers are built for mammals and very few of them are able to work well for plants. In this work, we present NAMS webserver, which contains a coding potential predictor specifically designed and optimized for plant species, as well as functional annotations aiming to provide the users' clues to the potential functions of their transcripts. NAMS had been comprehensively evaluated using three datasets collected from various sources and comprised of more than 3 million transcripts (Table 2). As a result, our evaluations have demonstrated the unsatisfactory performance of current tools as well as the advantage of NAMS. On other hand, no computational software is perfect even though it shows overall high accuracy; therefore, the strategy used in our webserver that combines various prediction tools could provide more reliable results to the users. For example, Supplementary Figure S6A shows

a well-characterized mRNA that codes the EPFL1 (EPIDERMAL PATTERNING FACTOR-like protein 1) protein, while it is predicted to be non-coding by CPC2; Supplementary Figure S6B shows an annotated lncRNA which is predicted to be coding by CNIT. The lncRNA in Supplementary Figure S6B has a 'reviewed' status in RefSeq gene annotation [39] but has not been fully characterized yet; our functional annotation reveals multiple stem-loop structures, suggesting the high probability of a functional lncRNA. Hence, our webserver could provide both reliable coding potential assessment, as well as valuable functional annotations for the users.

In coding potential assessment field, supervised machine learning approaches have been widely adapted and demonstrated high performance in various studies (mostly in mammals) [4, 24–27]. However, in this work, we show that currently such methods may not be optimal for plants. In fact, in our NAMS algorithm, the NAMS-SVM and NAMS-DT modules utilize the same features while employ different classification algorithms. The performance of these two modules is both very high and comparable on the testing dataset consisting of well-annotated species (Table 1), while drastically different on the comprehensive dataset which contains transcripts from various species that are poorly characterized. The unsatisfactory performance of CPC2 and CNIT on the comprehensive dataset further suggests that at the current stage, supervised machine learning approaches may not be the most preferred methods for plants. We think that this could be attributed to the circumstance that current studies on plants mainly focus on a limited number of model species (e.g. Arabidopsis); therefore biases the classification models; sequence features (e.g. Fickett TESTCODE) are also commonly used in various classifiers while our analysis reveals species-specific bias in such features; therefore, selection of features for machine learning approaches may also affects their performances. Our NAMS-DT algorithm is thus valuable under such scenario. In the future, we think that with the exploration of more plant species, supervised methods should be much improved toward the most reliable tools to facilitate transcriptomics studies in plants. In the meantime, considering the performance advantage of NAMS-DT, we think that that it is worthwhile to integrate NAMS into transcriptome data processing pipelines [23, 40] for more sensitive *de novo* lncRNA identifications in following studies.

In summary, considering that most plant species are poorly annotated, through the integration of coding potential assessment and functional annotations, we believe that our easy-to-use and multifunctional NAMS webserver could serve as a valuable resource to the plant research society.

Key Points

- Comprehensive evaluation using more than 3 million plant transcripts from various species reveals unsatisfactory performance of current coding potential tools.
- We have developed a new tool, NAMS, specifically designed for plants and outperforms state-of-the-art tools in coding potential classification.
- We have set up a user-friendly webserver with coding potential assessment and functional annotations for plant transcripts freely available to all users.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgement

We thank Mr. Qiang Sun and Mr. Xing Zhao, both from the Department of Orthopaedics and Traumatology, The Chinese University of Hong Kong, for technical assistance in deploying the webserver. We also thank the National Supercomputing Center in Shenzhen for providing computing support. This work was supported by Shenzhen Bay Laboratory and Hong Kong Research Grants Council of the Hong Kong Special Administrative Region (Grant No. 14116918, 14120619, 14102315, 14100415, 14133016, 14106117, 14100018), Focused Innovations Scheme B (Grant No.1907307), National Natural Science Foundation of China (Grant No. 31871304), NSFC-RGC Joint Research Program (Grant No. N_CUHK413/18), and CUHK direct grant 2018.015.

Author contributions

Conceive of the study: K.S. and H.S.; implement the software: K.S.; analyze data and write the manuscript: K.S., H.W. and H.S.

Data Availability

No new data were generated or analysed in support of this research.

References

1. Fatica A, Bozzoni I. Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet* 2014;**15**:7–21.
2. Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. *Nat Rev Genet* 2009;**10**:155–9.
3. St Laurent G, Wahlestedt C, Kapranov P. The landscape of long noncoding RNA classification. *Trends Genet* 2015;**31**:239–51.
4. Sun K, Chen X, Jiang P, *et al*. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC Genomics* 2013;**14**:S7.
5. Zhou L, Sun K, Zhao Y, *et al*. Linc-YY1 promotes myogenic differentiation and muscle regeneration through an interaction with the transcription factor YY1. *Nat Commun* 2015;**6**:10026.
6. Lu L, Sun K, Chen X, *et al*. Genome-wide survey by ChIP-seq reveals YY1 regulation of lincRNAs in skeletal myogenesis. *EMBO J* 2013;**32**:2575–88.
7. Zhao Y, Zhou J, He L, *et al*. MyoD induced enhancer RNA interacts with hnRNPL to activate target gene transcription during myogenic differentiation. *Nat Commun* 2019;**10**:5787.
8. Sun K, Zhou L, Zhao Y, *et al*. Genome-wide RNA-seq and ChIP-seq reveal Linc-YY1 function in regulating YY1/PRC2 activity during skeletal myogenesis. *Genom Data* 2016;**7**:247–9.

9. Sun K, Wang H, Sun H. mTFkb: a knowledgebase for fundamental annotation of mouse transcription factors. *Sci Rep* 2017;**7**:3022.
10. Wang L, Zhao Y, Bao X, *et al*. LncRNA Dum interacts with Dnmts to regulate Dppa2 expression during myogenic differentiation and muscle regeneration. *Cell Res* 2015;**25**:335–50.
11. Budak H, Kaya SB, Cagirici HB. Long non-coding RNA in plants in the era of reference sequences. *Front Plant Sci* 2020;**11**:276.
12. Ahmed W, Xia Y, Li R, *et al*. Non-coding RNAs: functional roles in the regulation of stress response in brassica crops. *Genomics* 2020;**112**:1419–24.
13. Hou J, Lu D, Mason AS, *et al*. Non-coding RNAs and transposable elements in plant genomes: emergence, regulatory mechanisms and roles in plant development and stress responses. *Planta* 2019;**250**:23–40.
14. Liu X, Hao L, Li D, *et al*. Long non-coding RNAs and their biological roles in plants. *Genomics Proteomics Bioinformatics* 2015;**13**:137–47.
15. Guan D, Yan B, Thieme C, *et al*. PlaMoM: a comprehensive database compiles plant mobile macromolecules. *Nucleic Acids Res* 2017;**45**:D1021–8.
16. Guan D, Shao J, Zhao Z, *et al*. PTHGRN: unraveling post-translational hierarchical gene regulatory networks using PPI, ChIP-seq and gene expression data. *Nucleic Acids Res* 2014;**42**:W130–6.
17. Zhang YC, Chen YQ. Long noncoding RNAs: new regulators in plant development. *Biochem Biophys Res Commun* 2013;**436**:111–4.
18. Cao H, Wahlestedt C, Kapranov P. Strategies to annotate and characterize long noncoding RNAs: advantages and pitfalls. *Trends Genet* 2018;**34**:704–21.
19. Zhou J, Zhang S, Wang H, *et al*. LncFunNet: an integrated computational framework for identification of functional long noncoding RNAs in mouse skeletal muscle cells. *Nucleic Acids Res* 2017;**45**:e108.
20. Axtell MJ, Meyers BC. Revisiting criteria for plant MicroRNA annotation in the era of big data. *Plant Cell* 2018;**30**:272–84.
21. Zhou B, Yang Y, Zhan J, *et al*. Predicting functional long noncoding RNAs validated by low throughput experiments. *RNA Biol* 2019;**16**:1555–64.
22. Chen X, Yan CC, Zhang X, *et al*. Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief Bioinform* 2017;**18**:558–76.
23. Sun K, Zhao Y, Wang H, *et al*. Sebnif: an integrated bioinformatics pipeline for the identification of novel large intergenic noncoding RNAs (lincRNAs)—application in human skeletal muscle cells. *PLoS One* 2014;**9**:e84500.
24. Kang YJ, Yang DC, Kong L, *et al*. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res* 2017;**45**:W12–6.
25. Guo JC, Fang SS, Wu Y, *et al*. CNIT: a fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition. *Nucleic Acids Res* 2019;**47**:W516–22.
26. Kong L, Zhang Y, Ye ZQ, *et al*. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res* 2007;**35**: W345–9.
27. 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.
28. Berardini TZ, Reiser L, Li D, *et al*. The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *Genesis* 2015;**53**:474–85.
29. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol* 2011;**2**:1–27.
30. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
31. Uniprot Consortium. UniProt: a hub for protein information. *Nucleic Acids Res* 2015;**43**:D204–12.
32. Fickett JW. Recognition of protein coding regions in DNA sequences. *Nucleic Acids Res* 1982;**10**:5303–18.
33. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, *et al*. ViennaRNA package 2.0. *Algorithms Mol Biol* 2011;**6**:26.
34. Goodstein DM, Shu S, Howson R, *et al*. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 2012;**40**:D1178–86.
35. Paytuvi Gallart A, *et al*. GREENC: a wiki-based database of plant lncRNAs. *Nucleic Acids Res* 2016;**44**:D1161–6.
36. Chinchor, N., MUC-4 evaluation metrics, in *Proceedings of the 4th Conference on Message Understanding. 1992, Association for Computational Linguistics*: McLean, Virginia. 22–9.
37. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta* 1975;**405**:442–51.
38. Kersey PJ, Allen JE, Allot A, *et al*. Ensembl genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic Acids Res* 2018;**46**:D802–8.
39. O'Leary NA, Wright MW, Brister JR, *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;**44**: D733–45.
40. Sun K. Ktrim: an extra-fast and accurate adapter- and quality-trimmer for sequencing data. *Bioinformatics* 2020; **36**:3561–2.