*Published online 4 October 2018* *Nucleic Acids Research, 2019, Vol. 47, Database issue* ***D63–D68***
*doi: 10.1093/nar/gky904*

# CRISPRlnc: a manually curated database of validated sgRNAs for lncRNAs

**Wen Chen [1,2,†], Guoqiang Zhang[1,†], Jing Li[1,†], Xuan Zhang[1], Shulan Huang[2], Shuanglin Xiang[2], Xiang Hu[2,*] and Changning Liu[1,*]**

[1]CAS Key Laboratory of Tropical Plant Resources and Sustainable Use, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Kunming 650223, China and [2]State Key Laboratory of Developmental Biology of Freshwater Fish, School of Life Sciences, Hunan Normal University, Changsha 410081, China

Received August 01, 2018; Revised September 18, 2018; Editorial Decision September 22, 2018; Accepted September 26, 2018

## ABSTRACT

**The CRISPR/Cas9 system, as a revolutionary genome editing tool for all areas of molecular biology, provides new opportunities for research on lncRNA's function. However, designing a CRISPR/Cas9 single guide RNA (sgRNA) for lncRNA is not easy with an unwarrantable effectiveness. Thus, it is worthy of collecting validated sgRNAs, to assist in efficiently choosing sgRNA with an expected activity. CRISPRlnc (http://www.crisprlnc.org or http://crisprlnc.xtbg.ac.cn) is a manually curated database of validated CRISPR/Cas9 sgRNAs for lncRNAs from all species. After manually reviewing more than 200 published literature, the current version of CRISPRlnc contains 305 lncRNAs and 2102 validated sgRNAs across eight species, including mammalian, insect and plant. We handled the ID, position in the genome, sequence and functional description of these lncRNAs, as well as the sequence, protoacceptor-motif (PAM), CRISPR type and validity of their paired sgRNAs. In CRISPRlnc, we provided the tools for browsing, searching and downloading data, as well as online BLAST service and genome browse server. As the first database against the validated sgRNAs of lncRNAs, CRISPRlnc will provide a new and powerful platform to promote CRISPR/Cas9 applications for future functional studies of lncRNAs.**

## INTRODUCTION

Long non-coding RNAs (lncRNAs), which were once regarded as junk sequences, are defined as non-protein-coding transcripts longer than 200 nt (1–3). Instead of being translated into proteins, lncRNAs exert their functions in cellular processes (4–6), organismal development (7,8) and diseases (9,10) directly in the shape of RNA. With an enormous range of applications of next generation sequencing, a large number of lncRNAs have been found in human (11,12), mouse (13,14), zebrafish (7,15,16), etc. At the same time, a large number of lncRNA databases have been created, and committed to large-scale collection and annotation of lncRNAs for various species. For example, NONCODE contained more than 350 000 lncRNA genes across 17 species (17). lncRNAdb manually collected and annotated about 300 functional lncRNAs that have been studied to date (18). However, due to the extremely complicated and inconstant mechanisms when compared with protein-coding genes, most of lncRNAs have not been well studied, despite that a host of researches have extensively demonstrated the significance and diversity of lncRNAs in regulatory functions.

CRISPR/Cas9 system, as a revolutionary gene editing tool for all the areas of molecular biology, can induce site-specific DNA cleavage by an RNA-guided DNA-endonuclease. CRISPR/Cas9 system can be used in the unmodified form of 'molecular scissors' (wtCas9) to edit parts of the genome by removing, adding or altering sections of the DNA sequence or to create a knockout genotype (CRISPRko) (19,20). After inactivating the nuclease domain to create a dead Cas9 (dCas9) (21), accessional function elements could conduce to branching out the applications of CRISPR/Cas9 system, such as transcriptional activation (CRISPRa) (22), transcriptional interference (CRISPRi) (23,24), gene editing (CRISPRedit) (25) and so on.

Unlike protein-coding genes, many lncRNAs are confined to nucleus, and some exert their molecular functions in a transcript-independent mode, meaning that the transcribing event of lncRNA in itself could affect target genes. Therefore, there are probably a lot of restrictions to use RNAi method for performing loss-of-function studies of lncRNAs (26). Contrastively, CRISPR/Cas9 has a huge advantage in lncRNA researches, resulting from its *in-cis* reg-

*To whom correspondence should be addressed. Tel: +86 691 8713009; Fax: +86 691 8713061; Email: liuchangning@xtbg.ac.cn
Correspondence may also be addressed to Xiang Hu. Tel: +86 731 88872095; Fax: +86 731 88872095; Email: huxiang@hunnu.edu.cn
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

ulative function in cell nucleus. Hence, CRISPR/Cas9 provides new opportunities for deeply researching lncRNA's functions, and is receiving increasing attention in the field of lncRNA studies.

The first step of CRISPR/Cas9 gene editing is to design a single guide RNA (sgRNA) to target your gene of interest. However, because sgRNAs vary widely in their activity and action models, designing a sgRNA is not easy due to an unwarrantable effectiveness. Thus, it is worthy of collecting validated sgRNA sequences, to assist in efficiently choosing sgRNA with an expected activity. For example, Varshney *et al*. had constructed CRISPRz to collect validated sgRNAs for zebrafish coding-genes (27). However, CRISPR/Cas9 applications for lncRNAs are much different from coding-genes, as indicated by many known works (26,28,29). For instance, it is not necessary for lncRNA to maintain an intact open reading frame for functioning. Besides, lncRNA as well as their surrounding coding/noncoding neighbors had complicated genomic architecture, like sense/antisense, intergenic/intragenic etc. Therefore, a validated sgRNA database specifically for lncRNAs is profoundly valuable for the relevant academic community.

In this study, we constructed CRISPRlnc (http://www.crisprlnc.org or http://crisprlnc.xtbg.ac.cn)—a manually curated database of validated sgRNAs for lncRNAs. After manually reviewing more than 200 published literature, the current version of CRISPRlnc contains 305 lncRNAs and 2102 validated sgRNAs across eight species, including mammalian, insect and plant. We handled the ID, position in the genome, sequence and functional description of these lncRNAs, as well as the sequence, protoacceptor-motif (PAM), CRISPR type and validity of their corresponding sgRNAs. In CRISPRlnc, we also provided the tools for browsing, searching and downloading all of the data covered, as well as online BLAST service and genome browse server. As the first database against the validated sgRNAs of lncRNAs, CRISPRlnc will give an efficient assistance when employing CRISPR/Cas9 system as tools for lncRNA genome editing. Moreover, the high-quality integrated information of validated sgRNAs for lncRNAs could be used as a gold standard dataset to guide the design of sgRNAs themselves.

## DATA COLLECTION AND DATABASE CONTENT

As shown in Figure 1, to exhaustively collect the validated data associated with CRISPR/Cas9 applications into lncRNA functional interrogation, first, we searched PubMed (30) to retrieve all relevant literature upon a list of keywords, such as 'lncRNA', 'long noncoding RNA' and 'lincRNA'. Next, the preliminary screening results were further traversed in title or abstract via our own in-house python script to extract those containing 'CRISPR', 'Cas9', 'Knockout' or 'Knockdown'. As a result, more than 200 articles are reserved. To ensure the high quality of our data, we manually checked and read over these papers. For each collected lncRNA, its ID, genomic position, nucleotide sequence and functional description are filed, along with the resources of its corresponding sgRNAs (including targeting sequence, PAM, CRISPR type and validity). Given the differential designation for the same gene between literatures, we normalized the gene ID and transcript ID of these lncRNAs in line with NCBI and Ensembl. The position of sgRNAs and lncRNAs in genome was obtained by aligning the consensus sequence to the reference genome using BLAT (31). Besides, in consideration of the positions of sgRNA that often link its functional effect—for example, the sgRNAs for CRISPRa or CRISPRi usually located near the transcription start sites of lncRNA (32,33)—and the expressional specificity of lncRNA in different tissues as well as the varying activity of CRISPR across different cell lines (32), we also integrated the information of 'the location of sgRNA relative to lncRNA' and 'the effective cell line of sgRNA' in our database.

Based on the original CRISPR/Cas9 system, researchers had developed a variety of technological innovation to satisfy differentiated intervention on gene function. Therefore, in this work, we collected data upon different CRISPR/Cas9-based systems, including CRISPR knockout (CRISPRko), CRISPR activation (CRISPRa), CRISPR interference (CRISPRi) and CRISPR edit (CRISPRedit). To assist in efficiently choosing sgRNA with an expected activity, we further divided our data into three types according to the sgRNA validity as stated in the corresponding literature: (i) recommended high-activity sgRNAs, referring to those that had been either clearly demonstrated as powerful executors in one study or, multi-employed by several studies; (ii) experimentally validated sgRNAs, which has some low-throughput experiments to verify effectiveness; (iii) sgRNAs with partial experimental validation are those designed by experts with high-throughput experimental validation—for example, using a high-throughput genomic deletion strategy, 51 lncRNAs were found to be capable of regulating human cancer cell growth when knocked out via a CRISPR system which constructs thousands of targeted sgRNAs in a lentiviral paired-guide RNA library (34).

The current version of CRISPRlnc contains 305 lncRNAs and 2102 validated sgRNAs across 8 species, including mammalian, insect and plant. The statistics of CRISPRlnc is summarized in Figure 2. Figure 2A shows the sgRNA distribution in various species. The collected sgRNA data are mainly from human (*Homo sapiens*, 1777 in total), mouse (*Mus musculus*, 79 in total) and fly (*Drosophila melanogaster*, 233 in total). Figure 2B shows the type of these data, most of which are CRISPR interference (CRISPRi, about 52.1%); while the minimum type is CRISPRedit (about 0.2%). According to the Top10 lncRNAs possessing the designed sgRNAs (Figure 2C, regardless of type III sgRNAs supported by high-throughput experiments), CRISPR/Cas9 applications for lncRNAs are currently focused on some potentially important lncRNAs in human, such as NEAT1 (Nuclear Enriched Abundant Transcript 1), PVT1 (Plasmacytoma variant translocation 1) and MALAT1 (Metastasis Associated Lung Adenocarcinoma Transcript 1). We also calculated the localization tendency of sgRNAs on lncRNAs, including up/downstream region of gene and gene body near 5′/3′ end (Figure 2D). Our results revealed that the collected sgRNAs were more likely to locate at gene body near 5′ end (61.4% in 5′ end versus 12.6% in 3′ end), as well as more likely at gene upstream as opposed to downstream (20.0% versus 5.9%).

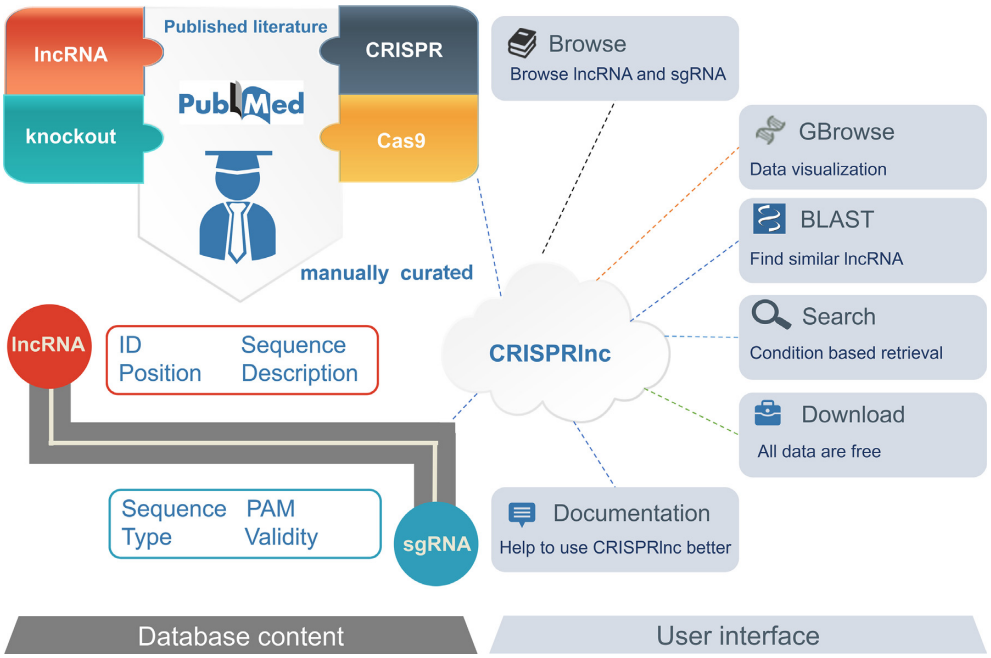


**Figure 1.** Overview of the CRISPRlnc database.

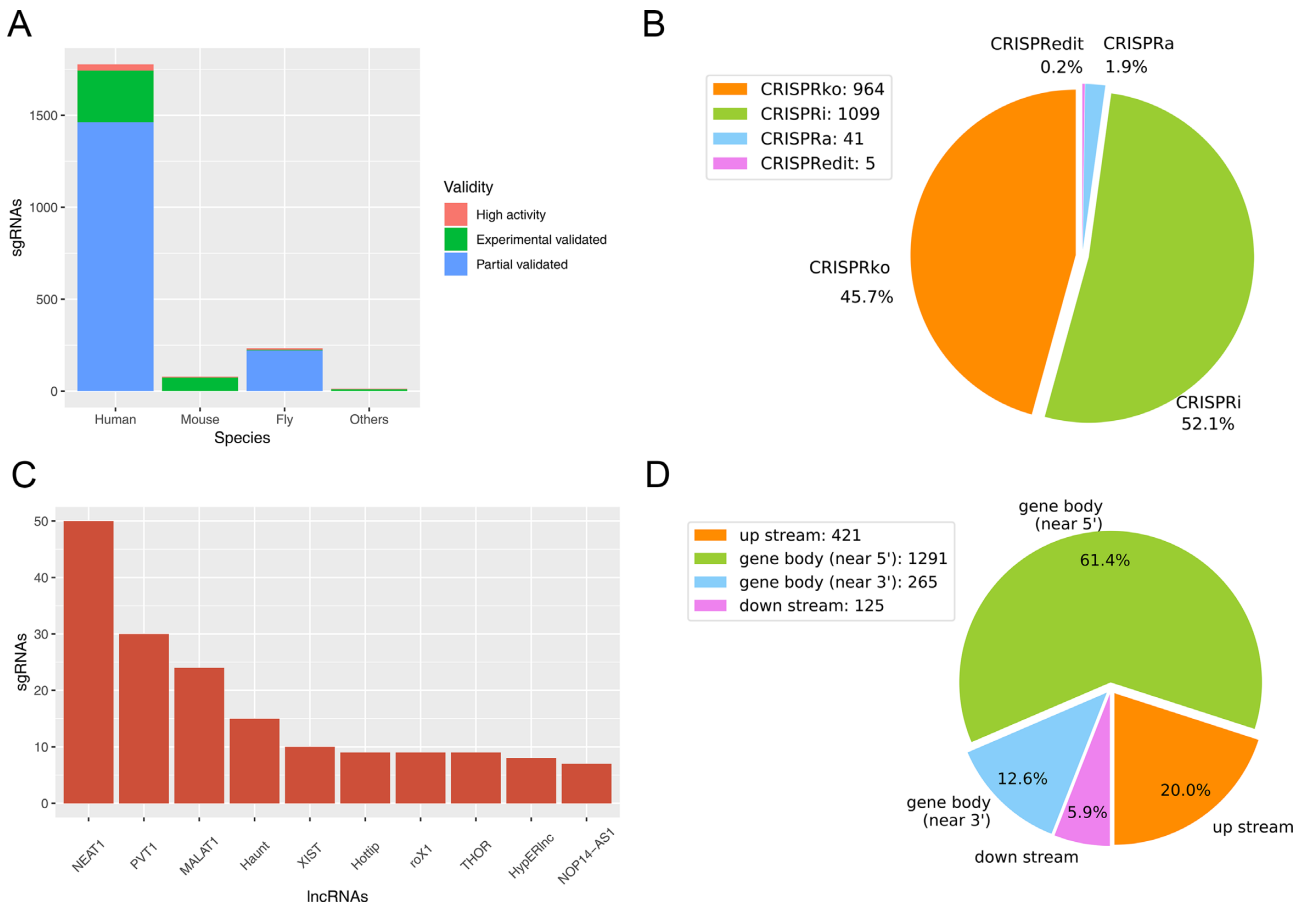


**Figure 2.** Statistics of the CRISPRlnc database. (**A**) The distribution of sgRNAs in various species. (**B**) The functional types of sgRNA. (**C**) The Top10 lncRNAs that possess known designed sgRNAs, without regard to those high-throughput experimental data. (**D**) The position distribution of sgRNAs on lncRNAs.

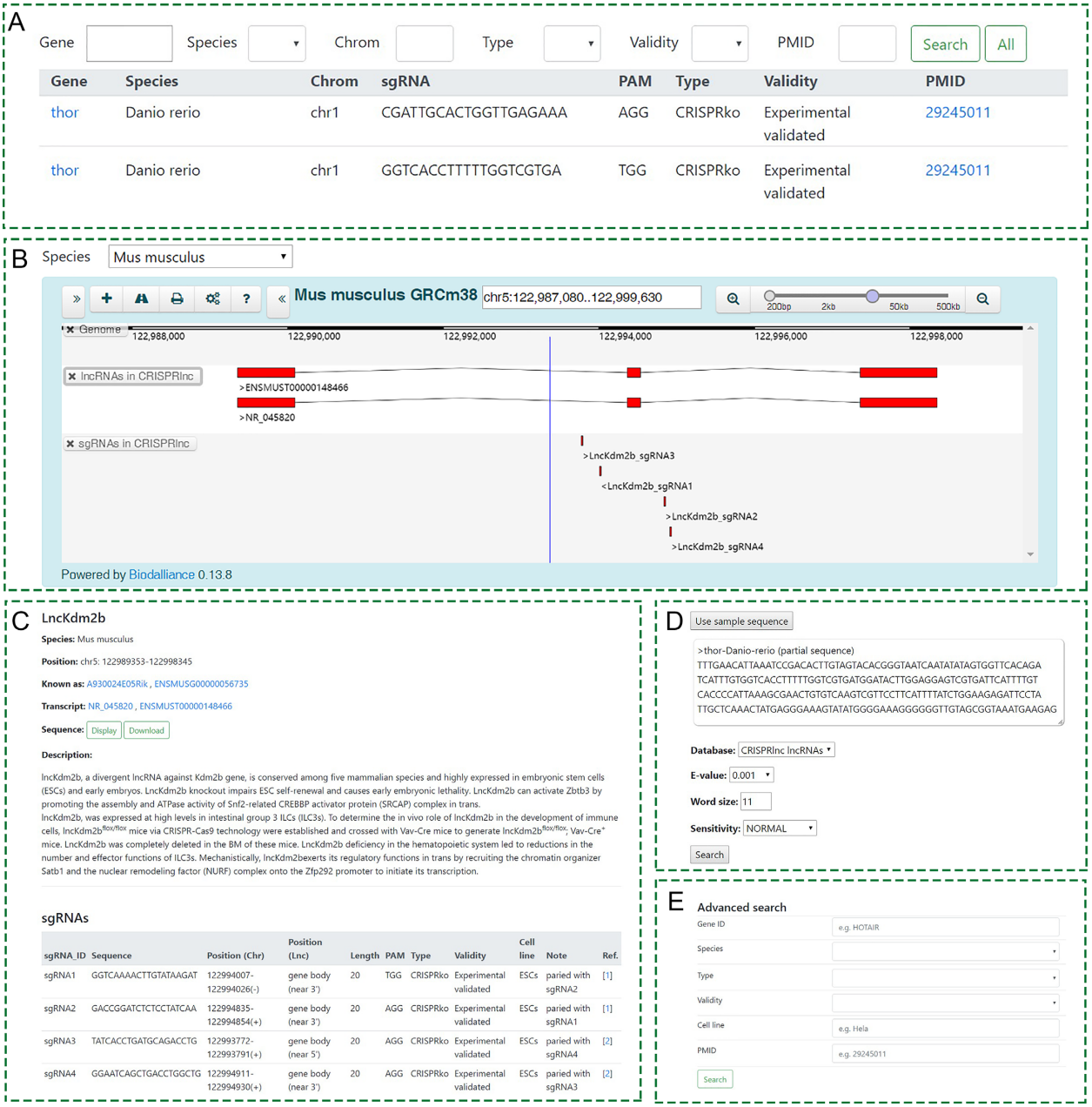


**Figure 3.** User interface of the CRISPRlnc database. (**A**) The 'Browse' page. (**B**) 'GBrowse', the genome browser. (**C**) The 'Details' page. (**D**) 'Online BLAST' search. (**E**) The 'Advanced Search' page.

## DATABASE ORGANIZATION AND WEB INTERFACE

### Database structure and implementation

The CRISPRlnc database is running on an Ubuntu Linux server (version 18.04), while Nginx (version 1.14.0) as the web server and SQLite (version 3.22.0) as the database server. The website is developed using Python based on Django web-framework (version 1.11). The web-frontend is developed using the Bootstrap framework (version 4.1.1). The GBrowser function is developed using Biodalliance (version 0.13.8) (35). The BLAST function is developed using django-blastplus (version 0.4.0), which is a simple Django app to conduct web-based homology search with blast+ (https://pypi.python.org/pypi/django-blastplus/).

## USER INTERFACE

For biologists to better access the information of CRISPRlnc, we established a user-friendly website. As shown in Figure 1, in CRISPRlnc, we provided the tools for browsing, searching and downloading all data, as well as online BLAST service and genome browse server. Each section contains sufficient help materials to ensure an easy use without any prerequisite knowledge or experience.

In 'Browse', user can retrieve all lncRNAs and sgRNAs in CRISPRlnc (Figure 3A). User can find the basic annotations on the 'Browse' page, such as gene ID, sgRNA sequence, PAM motif, CRISPR type and validity. The detailed information page for a specific gene can be accessed by clicking on the gene ID. The linked page is mainly containing lncRNA's information (Gene ID, transcript ID, genome location, sequence and function description), and a sgRNA table which contains all validated sgRNA information related to this lncRNA (Figure 3C). In the sgRNA table, each sgRNA is marked with a source of literature, so you can backtrack to the original documents.

In CRISPRlnc, we also provided some user-friendly web services. In 'GBrowse', you can view lncRNAs, sgRNAs and other genomic annotation according to their locations in genome (Figure 3B). The sgRNAs from some genome-scale CRISPR/Cas9 screens (32–34) were placed in a GBrowse track. For a lncRNA with a given genomic location, we recommend you use 'GBrowse' to find out whether it has validated sgRNAs. In addition, you can find a lncRNA or its homologs through BLAST sequence similarity search in CRISPRlnc (Figure 3D). Maybe the lncRNA in one species you are interested in does not have validated sgRNAs, but if you are lucky, you may find its homolog has. Moreover, you can find the lncRNAs or sgRNAs you are interested in using 'Advanced Search', by providing/selecting a list of keywords, such as gene ID, species, CRISPR type, validity, cell line and PMID (Figure 3E).

## DISCUSSION AND FUTURE DEVELOPMENT

Until now, tools with various characteristics for *ab initio* designing sgRNA are crowded, whereas systematic collection of validated sgRNAs is rare. Thus, we developed CRISPRlnc, a database that provides a comprehensive list of validated CRISPR/Cas9 sgRNAs for lncRNAs from all species. CRISPRlnc will provide not only an efficient assistance for genome editing of lncRNAs, but also a gold standard dataset that is crucial for subsequent development of sgRNA design tools of lncRNAs. We believed that CRISPRlnc will thrive in the fields related to lncRNA and CRISPR/Cas9 studies.

In the future, to respond to the rapid growth of researches in lncRNA and CRISPR/Cas9, we will continue to manually curate newly validated sgRNAs for lncRNAs and update the database every 3 months. In addition, we will continue to integrate more online tools and data sources to make CRISPRlnc more resourceful and more usable. As the first database specifically targeting to the validated sgRNAs of lncRNAs, we sincerely hope to get the support and advice from the academic community to help us improve it. And it is greatly appreciated for submitting data to us.

## FUNDING

National Natural Science Foundation of China [31471220, 91440113]; Xishuangbanna Tropical Botanical Garden Start-up Fund; 'Top Talents Program in Science and Technology' from Yunnan Province; Fund of State Key Laboratory of Developmental Biology of Freshwater Fish [2017KF003]; Scientific Research Fund of Hunan Provincial Education Department [15CY006]; Cooperative Innovation Center of Engineering and New Products for Developmental Biology of Hunan Province [20134486]. Funding for open access charge: National Natural Science Foundation of China [91440113].
*Conflict of interest statement.* None declared.

## REFERENCES

1. Wang,K.C. and Chang,H.Y. (2011) Molecular mechanisms of long noncoding RNAs. *Mol. Cell*, **43**, 904–914.
2. Iyer,M.K., Niknafs,Y.S., Malik,R., Singhal,U., Sahu,A., Hosono,Y., Barrette,T.R., Prensner,J.R., Evans,J.R., Zhao,S. *et al.* (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.*, **47**, 199–208.
3. Ulitsky,I. and Bartel,D.P. (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell*, **154**, 26–46.
4. Fatica,A. and Bozzoni,I. (2014) Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.*, **15**, 7–21.
5. Devaux,Y., Zangrando,J., Schroen,B., Creemers,E.E., Pedrazzini,T., Chang,C.P., Dorn,G.W. 2nd, Thum,T., Heymans,S. and Cardiolinc network (2015) Long noncoding RNAs in cardiac development and ageing. *Nat. Rev. Cardiol.*, **12**, 415–425.
6. Greco,C.M. and Condorelli,G. (2015) Epigenetic modifications and noncoding RNAs in cardiac hypertrophy and failure. *Nat. Rev. Cardiol.*, **12**, 488–497.
7. Ulitsky,I., Shkumatava,A., Jan,C.H., Sive,H. and Bartel,D.P. (2011) Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, **147**, 1537–1550.
8. Lin,N., Chang,K.Y., Li,Z., Gates,K., Rana,Z.A., Dang,J., Zhang,D., Han,T., Yang,C.S., Cunningham,T.J. *et al.* (2014) An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. *Mol. Cell*, **53**, 1005–1019.
9. Huarte,M. (2015) The emerging role of lncRNAs in cancer. *Nat. Med.*, **21**, 1253–1261.
10. Ling,H., Vincent,K., Pichler,M., Fodde,R., Berindan-Neagoe,I., Slack,F.J. and Calin,G.A. (2015) Junk DNA and the long non-coding RNA twist in cancer genetics. *Oncogene*, **34**, 5003–5011.
11. Derrien,T., Johnson,R., Bussotti,G., Tanzer,A., Djebali,S., Tilgner,H., Guernec,G., Martin,D., Merkel,A., Knowles,D.G. *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.
12. Iyer,M.K., Niknafs,Y.S., Malik,R., Singhal,U., Sahu,A., Hosono,Y., Barrette,T.R., Prensner,J.R., Evans,J.R., Zhao,S. *et al.* (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.*, **47**, 199–208.
13. Grote,P., Wittler,L., Hendrix,D., Koch,F., Wahrisch,S., Beisaw,A., Macura,K., Blass,G., Kellis,M., Werber,M. *et al.* (2013) The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Dev. Cell*, **24**, 206–214.
14. Sauvageau,M., Goff,L.A., Lodato,S., Bonev,B., Groff,A.F., Gerhardinger,C., Sanchez-Gomez,D.B., Hacisuleyman,E., Li,E., Spence,M. *et al.* (2013) Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife*, **2**, e01749.
15. Pauli,A., Valen,E., Lin,M.F., Garber,M., Vastenhouw,N.L., Levin,J.Z., Fan,L., Sandelin,A., Rinn,J.L., Regev,A. *et al.* (2012) Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.*, **22**, 577–591.
16. Kaushik,K., Leonard,V.E., Kv,S., Lalwani,M.K., Jalali,S., Patowary,A., Joshi,A., Scaria,V. and Sivasubbu,S. (2013) Dynamic expression of long non-coding RNAs (lncRNAs) in adult zebrafish. *PLoS One*, **8**, e83616.
17. Fang,S., Zhang,L., Guo,J., Niu,Y., Wu,Y., Li,H., Zhao,L., Li,X., Teng,X., Sun,X. *et al.* (2018) NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.*, **46**, D308–D314.
18. Quek,X.C., Thomson,D.W., Maag,J.L., Bartonicek,N., Signal,B., Clark,M.B., Gloss,B.S. and Dinger,M.E. (2015) lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.*, **43**, D168–D173.

19. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.
20. Mali,P., Yang,L., Esvelt,K.M., Aach,J., Guell,M., DiCarlo,J.E., Norville,J.E. and Church,G.M. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823–826.
21. Qi,L.S., Larson,M.H., Gilbert,L.A., Doudna,J.A., Weissman,J.S., Arkin,A.P. and Lim,W.A. (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, **152**, 1173–1183.
22. Konermann,S., Brigham,M.D., Trevino,A.E., Joung,J., Abudayyeh,O.O., Barcena,C., Hsu,P.D., Habib,N., Gootenberg,J.S., Nishimasu,H. *et al.* (2015) Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, **517**, 583–588.
23. Mali,P., Aach,J., Stranges,P.B., Esvelt,K.M., Moosburner,M., Kosuri,S., Yang,L. and Church,G.M. (2013) CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.*, **31**, 833–838.
24. Gilbert,L.A., Larson,M.H., Morsut,L., Liu,Z., Brar,G.A., Torres,S.E., Stern-Ginossar,N., Brandman,O., Whitehead,E.H., Doudna,J.A. *et al.* (2013) CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell*, **154**, 442–451.
25. Komor,A.C., Kim,Y.B., Packer,M.S., Zuris,J.A. and Liu,D.R. (2016) Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, **533**, 420–424.
26. Goyal,A., Myacheva,K., Gross,M., Klingenberg,M., Duran Arque,B. and Diederichs,S. (2017) Challenges of CRISPR/Cas9 applications for long non-coding RNA genes. *Nucleic Acids Res.*, **45**, e12.
27. Varshney,G.K., Zhang,S., Pei,W., Adomako-Ankomah,A., Fohtung,J., Schaffer,K., Carrington,B., Maskeri,A., Slevin,C., Wolfsberg,T. *et al.* (2016) CRISPRz: a database of zebrafish validated sgRNAs. *Nucleic Acids Res.*, **44**, D822–D826.
28. Ho,T.T., Zhou,N., Huang,J., Koirala,P., Xu,M., Fung,R., Wu,F. and Mo,Y.Y. (2015) Targeting non-coding RNAs with the CRISPR/Cas9 system in human cell lines. *Nucleic Acids Res.*, **43**, e17.
29. Ghosh,S., Tibbit,C. and Liu,J.L. (2016) Effective knockdown of Drosophila long non-coding RNAs by CRISPR interference. *Nucleic Acids Res.*, **44**, e84.
30. Coordinators,N.R. (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **46**, D8–D13.
31. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
32. Liu,S.J., Horlbeck,M.A., Cho,S.W., Birk,H.S., Malatesta,M., He,D., Attenello,F.J., Villalta,J.E., Cho,M.Y., Chen,Y. *et al.* (2017) CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science*, **355**, eaah7111.
33. Joung,J., Engreitz,J.M., Konermann,S., Abudayyeh,O.O., Verdine,V.K., Aguet,F., Gootenberg,J.S., Sanjana,N.E., Wright,J.B., Fulco,C.P. *et al.* (2017) Genome-scale activation screen identifies a lncRNA locus regulating a gene neighbourhood. *Nature*, **548**, 343–346.
34. Zhu,S., Li,W., Liu,J., Chen,C.H., Liao,Q., Xu,P., Xu,H., Xiao,T., Cao,Z., Peng,J. *et al.* (2016) Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. *Nat. Biotechnol.*, **34**, 1279–1286.
35. Down,T.A., Piipari,M. and Hubbard,T.J. (2011) Dalliance: interactive genome viewing on the web. *Bioinformatics*, **27**, 889–890.