D186–D191 Nucleic Acids Research, 2023, Vol. 51, Database issue *Published online 4 November 2022*
https://doi.org/10.1093/nar/gkac999

LncBook 2.0: integrating human long non-coding RNAs with multi-omics annotations

Zhao Li [1,2,3,†], Lin Liu [1,2,†], Changrui Feng [1,2,3,†], Yuxin Qin[1,2,3], Jingfa Xiao[1,2,3], Zhang Zhang [1,2,3,*] and Lina Ma [1,2,3,*]

[1]National Genomics Data Center & CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, [2]China National Center for Bioinformation, Beijing 100101, China and [3]University of Chinese Academy of Sciences, Beijing 100049, China

Received September 15, 2022; Revised October 11, 2022; Editorial Decision October 11, 2022; Accepted October 19, 2022

ABSTRACT

LncBook, a comprehensive resource of human long non-coding RNAs (lncRNAs), has been used in a wide range of lncRNA studies across various biological contexts. Here, we present LncBook 2.0 (https://ngdc.cncb.ac.cn/lncbook), with significant updates and enhancements as follows: (i) incorporation of 119 722 new transcripts, 9632 new genes, and gene structure update of 21 305 lncRNAs; (ii) characterization of conservation features of human lncRNA genes across 40 vertebrates; (iii) integration of lncRNA-encoded small proteins; (iv) enrichment of expression and DNA methylation profiles with more biological contexts and (v) identification of lncRNA–protein interactions and improved prediction of lncRNA-miRNA interactions. Collectively, LncBook 2.0 accommodates a high-quality collection of 95 243 lncRNA genes and 323 950 transcripts and incorporates their abundant annotations at different omics levels, thereby enabling users to decipher functional significance of lncRNAs in different biological contexts.

INTRODUCTION

LncBook, a curated resource of human long non-coding RNAs (lncRNAs), features comprehensive integration of human lncRNAs and systematic annotation with multi-omics data analysis (1). Since its inception in 2019, LncBook has been widely used in delineating the transcriptional landscape of human lncRNAs (2,3), uncovering lncRNAs' molecular signatures (4,5), and disentangling functional relevance of lncRNAs in human diseases (6–8). Over the past several years, considerable efforts have been devoted to identifying (9–13) and characterizing human lncRNAs at different omics levels across diverse biological contexts, e.g. disease, normal tissue/cell line, organ development, subcellular localization (14). Particularly, multiple lines of evidence have accumulated that sequence conservation is a fundamental indicator for lncRNA functional significance (15) and that lncRNA-encoded small proteins are involved in diverse functions and multiple diseases (16–18). Therefore, it is highly needed to integrate newly reported lncRNAs and characterize lncRNAs from multiple omics levels and in more biological contexts. Toward this end, here we perform comprehensive integration of lncRNAs as well as their curated annotations, including expression and DNA methylation profiles in multiple biological contexts, disease/trait-associated variants, lncRNA-miRNA interactions, lncRNA–protein interactions, evolutionary conservation features and small proteins. As a consequence, we provide an updated version of LncBook, which, in contrast to the previous version, has been significantly upgraded, expanded and enhanced (Table 1).

MATERIALS AND METHODS

LncRNA integration and curation

Based on the previous version, LncBook 2.0 integrated lncRNAs from five resources, including RefLnc (9), GENCODE v33 (10), CHESS v2.2 (11), FANTOM-CAT (12) and BIGTranscriptome (13). Transcripts with redundancy, background noise, and mapping error, as well as incomplete transcripts, short ones, and those that may encode proteins, were excluded (1). To improve the curation quality, in LncBook 2.0, we also removed lncRNA transcripts without strand information, and transcripts identified as miRNA precursors, small RNAs and pseudogenes according to the comparison results generated with GffCompare (19). In addition, four algorithms, viz., CPC2 (20), LGC (21), CPAT (22) and PLEK (23), were used for coding potential estimation, and transcripts identified as lncRNAs by at least three algorithms were retained. It is noted that the lncRNAs annotated by HGNC and GENCODE were

[*]To whom correspondence should be addressed. Tel: +86 10 8409 7845; Fax: +86 10 8409 7298; Email: malina@big.ac.cn
Correspondence may also be addressed to Zhang Zhang. Tel: +86 10 8409 7261; Fax: +86 10 8409 7298; Email: zhangzhang@big.ac.cn
[†]The authors wish it to be known that, in their opinion, these authors should be regarded as Joint First Authors.

Table 1. Data statistics of two LncBook versions

Data item		Version 2.0	Version 1.0
LncRNA integration and curation	Transcript	323 950	270 044
	Gene	95 243	140 356
	Quality control	Remove pseudogenes, small RNAs, miRNA precursors, and transcripts without strand information	-
	Reference file	LncRNA; LncRNA & other genes	LncRNA
Multi-omics annotation	Evolutionary conservation	Conservation features across 40 vertebrates	-
	Small protein	34 012 small proteins	-
	Genome variation	959 138 disease/trait-associated variants	92 725 757 SNPs
	DNA methylation profile	14 cancers and 2 neurodevelopmental disorders	9 cancers
	Expression profile	9 biological contexts	1 biological context
	LncRNA–protein interaction	772 745 lncRNA–protein interactions	-
	LncRNA–miRNA interaction	Predicted with TargetScan, miRanda and RNAhybrid	Predicted with TargetScan and miRanda

retained regardless of the coding potential. Following the strategies used by GENCODE (24) and NONCODE (25), we assigned lncRNA transcripts overlapped in their exonic regions in the same strand into the same gene. To meet different demands of data analysis, lncRNA gene annotation file and the integrated annotation file with both lncRNA genes and other genes (derived from GENCODE) were provided.

Data integration and analysis

To perform sequence conservation analysis, genome references, gene annotation files and paired alignment chain files for human and 40 vertebrates were downloaded from UCSC Genome Browser Gateway (https://hgdownload.soe.ucsc.edu) (26). We identified lncRNA homologous sequences/genes by considering alignment length and comparing with introns' alignments in different species. As lncRNAs are in general lack of high sequence conservation, 20% transcript coverage has been adopted to identify homologous lncRNAs (27), so that we collected alignments that are at least 50 nt in length and with >20% transcript coverage. Meanwhile, to reduce the impacts of evolutionary distance, the homologous sequences/genes were determined for each species if the alignment performance (measured by alignment length and identity) of lncRNAs exceeds the introns' Q50 threshold, which represents the intermediate level of intron alignments. LncRNA gene age was defined as the earliest occurrence time of its homologous sequence across 40 species, which, from latest to earliest, are 'Homo' (human specific), 'Hominini', 'Homininae', 'Hominidae', 'Hominoidea', 'Catarrhini', 'Simiiformes', 'Haplorrhini', 'Primates', 'Euarchontoglires', 'Boreoeutheria', 'Eutheria', 'Theria', 'Mammalia', 'Amniota', 'Tetrapoda' and 'Euteleostomi'.

High-confidence variants and associations were collected from COSMIC (28), ClinVar (29) and GWAS Catalog (30). For COSMIC (28), the variants labeled as 'Confirmed somatic mutation' were retained. Suggested by COSMIC (https://cancer.sanger.ac.uk/cosmic/analyses), the variants whose FATHMM-MKL scores > 0.7 were defined as disease-related (pathogenic) variants. Variants with definite labels such as 'benign' or 'pathogenic' in ClinVar (29) were collected. For GWAS Catalog (30), we collected the associations with P-value < 5×10^{-8}, which has been widely used to determine association between a common genetic variant and a trait of interest (31,32). Finally, names of diseases and traits were unified according to Human Phenotype Ontology (33) and Experimental Factor Ontology (34), respectively. All the variants were allocated to lncRNAs by BEDTools (35).

To characterize the DNA methylation profiles of lncRNAs across human diseases, 16 publicly accessible bisulfite-seq datasets were collected from TCGA (The Cancer Genome Atlas) (36) and GEO (Gene Expression Omnibus) (37), covering 14 cancers and 2 neuro-developmental disorders. Here, promoter region was defined as -1500 bps relative to the transcription start site and averaged methylation level of all CpGs in promoter or body region was calculated. Differentially methylated lncRNA genes were identified by considering the significance of fold change, P-value, maximum and minimum methylation levels.

Expression profiles of human lncRNAs were obtained from LncExpDB (38), covering 337 biological conditions that can be further classified into nine biological contexts, namely, normal tissue/cell line, organ development, preimplantation embryo, cell differentiation, subcellular localization, exosome, cancer cell line, virus infection and circadian rhythm. To determine gene expression capacity, genes whose expression values are greater than the upper quantile of whole transcriptome (includes both lncRNA genes and protein-coding genes) in at least one biological condition are considered as high expression capacity, those less than the lower quantile as low capacity, and the remaining as medium capacity. Featured lncRNA genes are those that are specifically expressed in a certain cell line/tissue, consistently expressed across different cell lines/tissues, differentially expressed in the context of cancer or virus infection, enriched in a specific organelle, dynamically expressed during cell differentiation or embryo/organ development, or periodically expressed with circadian rhythm. In addition,

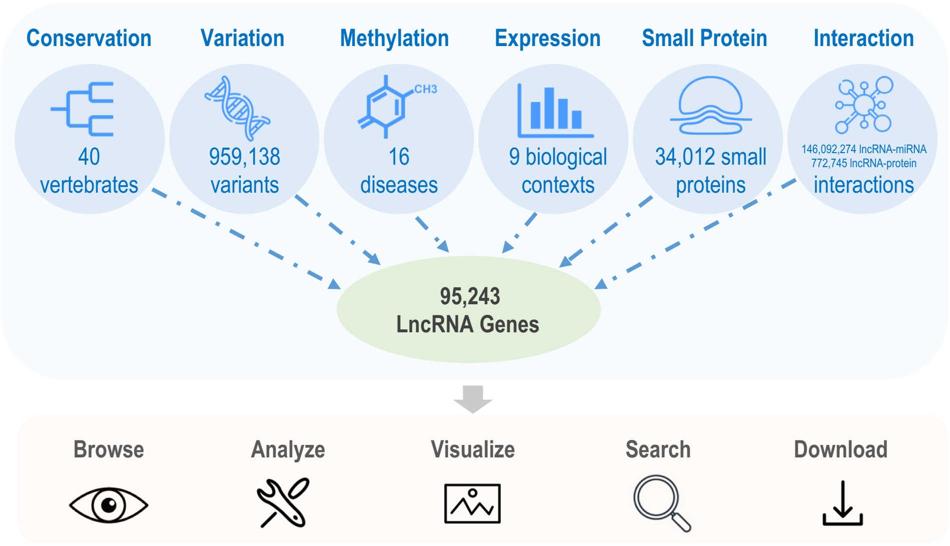


Figure 1. Data contents and organization in LncBook 2.0. A comprehensive, high-quality collection of human lncRNAs is annotated at different omics levels and organized with user-friendly web interfaces for searching, browsing, visualizing, analyzing and downloading.

subcellular localization and tissue/normal cell/cancer cell specificity of lncRNA genes were characterized based on the expression profiles, and related information was listed in ‘Gene Summary’.

Small proteins supported by Ribo-seq or mass spectrometry evidence were integrated from SmProt (39). Small proteins were mapped onto the lncRNAs with BEDtools (35) and those entirely and uniquely falling within lncRNA transcripts were retained.

LncRNA–protein interactions were identified based on the collection of 848 077 RBP (RNA Binding Protein) binding sites of 150 RBPs in HepG2 and K562 cell lines from ENCODE (40). We mapped the RBP binding sites onto the lncRNAs with BEDtools (35), and RBP binding sites entirely and uniquely falling within lncRNA transcripts were retained.

Three tools, including miRanda (41), TargetScan (42), and RNAhybrid (43), were used to predict more lncRNA-miRNA interactions. Interactions supported by all the three tools as well as those by any two tools were listed in the ‘Interaction’ section. Additionally, interactions predicted by only one tool were provided in ‘Downloads’.

Implementation

LncBook 2.0 was implemented based on Spring Boot (https://spring.io/projects/spring-boot), MySQL (https://www.mysql.com) and Apache Tomcat Server (https://tomcat.apache.org). Web interfaces were developed by HTML5, CSS3, AJAX (Asynchronous JavaScript and XML), JQuery (https://jquery.com), Bootstrap (https://getbootstrap.com) and Semantic UI (https://semantic-ui.com). In addition, data visualization was powered by HighCharts (https://www.highcharts.com.cn), ECharts (https://echarts.apache.org), Plotly.js (https://plotly.com), and DataTables (https://datatables.net). Web tools were set up by HTML widgets, NCBI BLAST+ (44) and in-house python scripts.

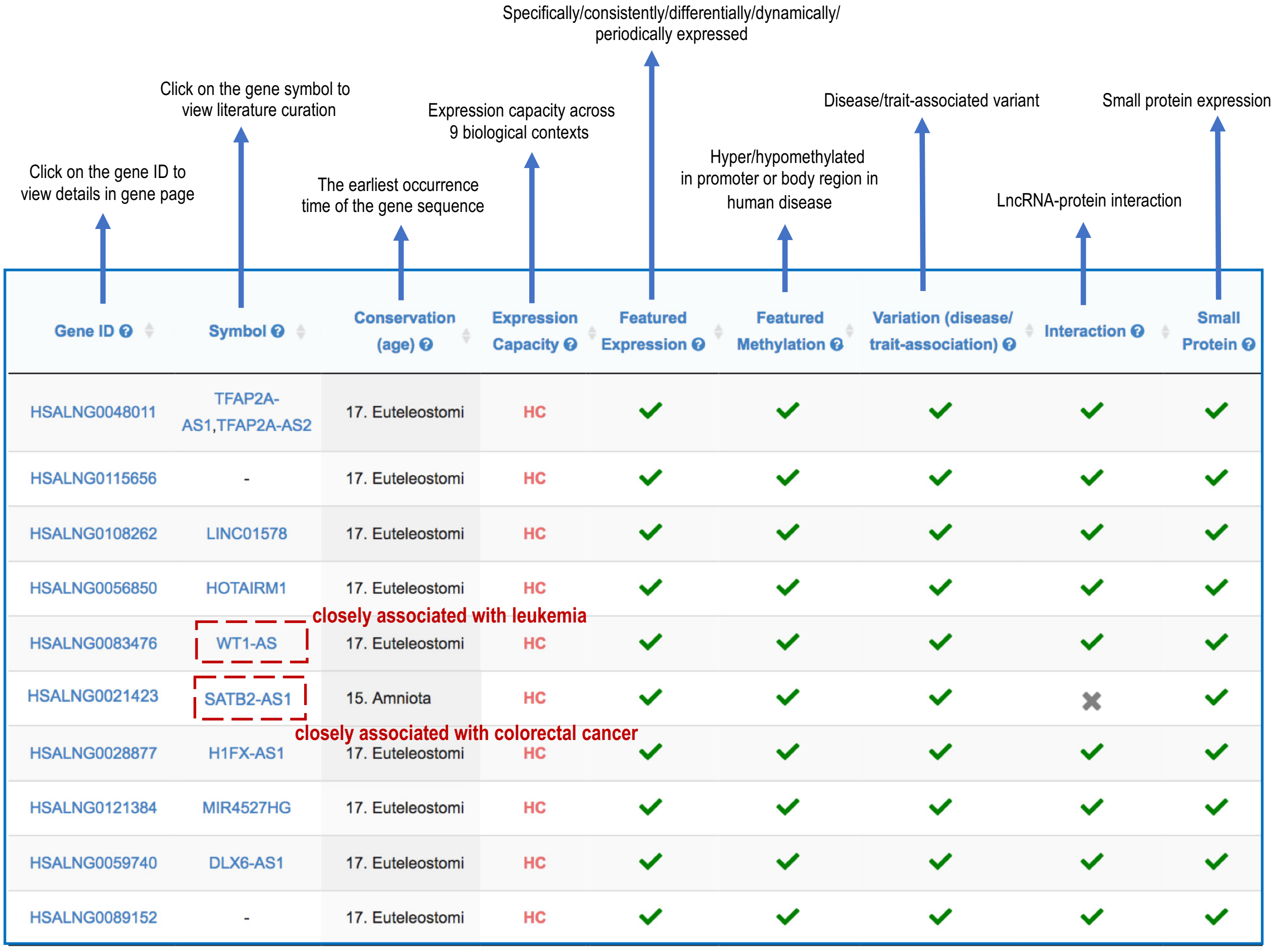
IMPROVED CONTENT AND NEW FEATURES

Expanded lncRNA list and enriched multi-omics annotations

LncBook is devoted to providing a comprehensive and high-quality collection of human lncRNAs as well as their annotations based on multi-omics analysis and value-added curation. Compared with the previous version, LncBook 2.0 is significantly improved in the quality of human lncRNA genes, and the comprehensiveness of their multi-omics annotations (Table 1).

LncBook 2.0 features a full list of human lncRNAs by comprehensively integrating lncRNAs from different resources and curating the lncRNAs with more strict criteria (see details in Materials and Methods). As a result, it incorporates 119 722 new transcripts and 9632 new genes, updates the structure of 21 305 genes and provides a high-quality collection of 323 950 lncRNA transcripts and 95 243 genes, compared with 270 044 transcripts and 140 356 genes of the previous version. As transcripts overlapped in the exonic regions in the same strand are assigned as the same gene, LncBook 2.0 presents a decreased gene count in spite of the increased number of transcripts. Based on this list, it incorporates more annotations by including new omics profiles and covering more biological contexts.

We characterize conservation features of human lncRNA genes across 40 vertebrates, identify 139 306 homologous genes for 22 347 human lncRNA genes, and integrate 34 012 lncRNA-encoded small proteins for 5743 lncRNA genes. Expression profiles have been enriched with more biological contexts, which are increased from 1 (normal tissue/cell line) to 9 (normal tissue/cell line, organ development, preimplantation embryo, cell differentiation, subcellular localization, exosome, cancer cell line, virus infection, circadian rhythm). Moreover, disease types for DNA methylation profiles have been increased from 9 cancers to 14 cancers and 2 neurodevelopmental disorders have been included as well. As a result, LncBook 2.0 contains a total of 24 157 featured lncRNA genes for expression (specifical



Click on the gene ID to view details in gene page

Click on the gene symbol to view literature curation

The earliest occurrence time of the gene sequence

Expression capacity across 9 biological contexts

Specifically/consistently/differentially/dynamically/periodically expressed

Hyper/hypomethylated in promoter or body region in human disease

Disease/trait-associated variant

LncRNA-protein interaction

Small protein expression

Gene ID	Symbol	Conservation (age)	Expression Capacity	Featured Expression	Featured Methylation	Variation (disease/trait-association)	Interaction	Small Protein
HSALNG0048011	TFAP2A-AS1,TFAP2A-AS2	17. Euteleostomi	HC	✓	✓	✓	✓	✓
HSALNG0115656	-	17. Euteleostomi	HC	✓	✓	✓	✓	✓
HSALNG0108262	LINC01578	17. Euteleostomi	HC	✓	✓	✓	✓	✓
HSALNG0056850	HOTAIRM1	17. Euteleostomi	HC	✓	✓	✓	✓	✓
HSALNG0083476	WT1-AS	17. Euteleostomi	HC	✓	✓	✓	✓	✓
HSALNG0021423	SATB2-AS1	15. Amniota	HC	✓	✓	✓	✗	✓
HSALNG0028877	H1FX-AS1	17. Euteleostomi	HC	✓	✓	✓	✓	✓
HSALNG0121384	MIR4527HG	17. Euteleostomi	HC	✓	✓	✓	✓	✓
HSALNG0059740	DLX6-AS1	17. Euteleostomi	HC	✓	✓	✓	✓	✓
HSALNG0089152	-	17. Euteleostomi	HC	✓	✓	✓	✓	✓

closely associated with leukemia

closely associated with colorectal cancer

Figure 2. Screenshot of 'Genes' page. Multi-omics features including expression capacity, featured expression pattern, featured methylation pattern, disease/trait-associated variation, lncRNA–protein interaction, and small protein expression, are summarized in a table, which enables customized filtration and sort. Functional evidences for *SATB2-AS1* and *WT1-AS* inferred from multi-omics association analysis and experimental evidences are described in the manuscript.

ly/consistently/differentially/dynamically/perio-dically expressed) and 19 543 featured lncRNA genes for DNA methylation (hyper/hypomethylated in promoter or body region). Also, we curate 959 138 disease/trait-associated variants associated with 50 165 lncRNA genes, identify 772 745 lncRNA–protein interactions for 2005 lncRNA genes, and predict 146 092 274 lncRNA-miRNA interactions for all lncRNA genes.

Database contents and organization

LncBook 2.0 is a gene-centric resource with user-friendly web interfaces for searching, browsing, visualizing, analyzing and downloading (Figure 1). A lncRNA gene corresponds to a web page, which is composed of nine sections, including gene summary, transcript information, coding potential, conservation, variation, methylation, expression, small protein and interaction. For any lncRNA gene, the annotations are summarized in various tables, sequence conservation status across 40 vertebrates is displayed in a phylogenetic tree, methylation levels of both promoter and body regions across 16 diseases are visualized in boxplots, and expression profiles across 337 biological conditions are represented in a bar chart. Meanwhile, LncBook 2.0 allows interactive visualization of literature curation results in LncRNAWiki (45), expression profiles across diverse biological contexts in LncExpDB (38) and relevant annotations in the integrated databases. In addition, LncBook 2.0 provides dedicated web pages for each omics resource with abundant descriptive terms to enable various customized comparisons. Moreover, curated multi-omics features of all lncRNAs are summarized in a tabular form in the 'Genes' page. Based on these annotations, LncBook 2.0 presents a series of useful statistics and analysis results in the 'Statistics' page, and deploys several useful tools for online analysis. Equally important, all the associated data are publicly available in the 'Downloads' page and all tables and figures could be freely downloadable in LncBook 2.0.

Functional lncRNA identification and exploration

As an alternative to experimental examination, bioinformatics association study serves as an efficient approach to investigate the putative function of lncRNAs with the analysis of multi-omics data across various biological contexts. Therefore, LncBook 2.0 is committed to providing the list of high-quality functional evidences from evolutionary conservation, genome variation, DNA methylation, gene expression, small protein and lncRNA-mediated interactions, which could be overviewed for all collected lncRNA genes (Figure 2).

Users can start from highly conserved lncRNA genes with more multi-omics associations, for example, by setting up the filters: gene age ≥14, high expression capacity, featured gene for both expression and methylation, possessing disease/trait-associated variants, and encoding small proteins. Consequently, a list of 100 lncRNA genes are obtained (Supplementary Table S1). According to multi-omics associations, we find that *SATB2-AS1* is suggested to be closely associated with colorectal cancer. It is highly expressed in colon and rectum, hypermethylated in colon adenocarcinoma, possesses large intestine carcinoma-associated variants, and its encoded small proteins (SPROHSA260428 and SPROHSA260429) are also detected in colorectal cancer samples. Consistently, annotations in LncRNAWiki show that *SATB2-AS1* has been reported to be involved in colorectal cancer (46,47). Another lncRNA, *WT1-AS*, is suggested to be closely associated with leukemia, as it is highly expressed and hyper-methylated in leukemia samples, and the encoded small proteins (SPROHSA65308, SPR0HSA264911 and SPROHSA326667) are also detected in leukemia samples. Consistently, *WT1-AS* has been experimentally validated to play an important role in leukemia (48,49). Also, LncBook 2.0 provides homologous genes for these lncRNAs, which would offer new insights into the exploration of the biological function in different species. Among the 100 lncRNA genes, notably, 50 are functionally uncharacterized up to now, which can be regarded as valuable candidates for experimental investigation and in-depth functional research. Of course, we believe that the more multi-omics associations do not necessarily represent the more important function, and users are encouraged to perform customized selection by different omics features of their own interest.

DISCUSSION AND FUTURE DEVELOPMENTS

As an important resource of the National Genomics Data Center (50), LncBook, in close partnership with LncExpDB (38) and LncRNAWiki (45), serves as a fundamental resource to provide comprehensive and high-quality lncRNAs and their annotations. Considering the growing volume of human lncRNAs, we plan to develop an automatic pipeline and web server to ease lncRNA integration and curation, and classify these lncRNAs by building collaborations with field experts in RNAcentral (51). With more lncRNAs identified in different species, we plan to improve the conservation annotation by including the results generated from lncRNA sequence alignments. Moreover, to better decipher the characteristics of lncRNAs, we will continue to include new omics features, such as lncRNA–DNA/RNA interactions, histone modification regulation, lncRNA modifications/edits and structures, integrate more biological contexts, and perform comparisons between lncRNAs and other types of genes (e.g. protein-coding gene). With the incorporation of more datasets and annotations, we also plan to develop a robust metric to estimate the confidence level of lncRNA gene and accordingly provide a high-confidence list of functional lncRNAs.

DATA AVAILABILITY

LncBook 2.0 is freely available online at https://ngdc.cncb.ac.cn/lncbook.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Qianpeng Li, Qiheng Qian, Xiaonan Liu, Mochen Zhang and Chang Liu for their valuable comments and discussions in this work.

FUNDING

Strategic Priority Research Program of the Chinese Academy of Sciences [XDB38030400, XDA19050302]; Youth Innovation Promotion Association of Chinese Academy of Sciences [2019104]; National Natural Science Foundation of China [32030021, 31871328]. Funding for open access charge: Strategic Priority Research Program of the Chinese Academy of Sciences.
Conflict of interest statement. None declared.

REFERENCES

1. Ma,L., Cao,J., Liu,L., Du,Q., Li,Z., Zou,D., Bajic,V.B. and Zhang,Z. (2019) LncBook: a curated knowledgebase of human long non-coding RNAs. *Nucleic Acids Res.*, **47**, D128–D134.
2. Lin,Y., Pan,X. and Shen,H.-B. (2021) lncLocator 2.0: a cell-line-specific subcellular localization predictor for long non-coding RNAs with interpretable deep learning. *Bioinformatics*, **37**, 2308–2316.
3. Kraczkowska,W. and Jagodzinski,P.P. (2019) The long non-coding RNA landscape of atherosclerotic plaques. *Mol. Diagn. Ther.*, **23**, 735–749.
4. Cai,B., Cai,J., Yin,Z., Jiang,X., Yao,C., Ma,J., Xue,Z., Miao,P., Xiao,Q., Cheng,Y. *et al.* (2021) Long non-coding RNA expression profiles in neutrophils revealed potential biomarker for prediction of renal involvement in SLE patients. *Rheumatology (Oxford)*, **60**, 1734–1746.
5. Zhao,X., Tang,D., Chen,X., Chen,S. and Wang,C. (2021) Functional lncRNA-miRNA-mRNA networks in response to baicalein treatment in hepatocellular carcinoma. *Biomed. Res. Int.*, **2021**, 8844261.
6. Liu,X., Xu,Y., Wang,R., Liu,S., Wang,J., Luo,Y., Leung,K.S. and Cheng,L. (2021) A network-based algorithm for the identification of moonlighting noncoding RNAs and its application in sepsis. *Brief. Bioinform.*, **22**, 581–588.
7. Turjya,R.R., Khan,M.A. and Mir Md Khademul Islam,A.B. (2020) Perversely expressed long noncoding RNAs can alter host response and viral proliferation in SARS-CoV-2 infection. *Future Virol.*, **15**, 577–593.
8. Doke,M., McLaughlin,J.P., Cai,J.J., Pendyala,G., Kashanchi,F., Khan,M.A. and Samikkannu,T. (2022) HIV-1 tat and cocaine impact astrocytic energy reservoirs and epigenetic regulation by influencing the LINC01133-hsa-miR-4726-5p-NDUFA9 axis. *Mol. Ther. Nucleic Acids*, **29**, 243–258.
9. Jiang,S., Cheng,S.J., Ren,L.C., Wang,Q., Kang,Y.J., Ding,Y., Hou,M., Yang,X.X., Lin,Y., Liang,N. *et al.* (2019) An expanded landscape of human long noncoding RNA. *Nucleic Acids Res.*, **47**, 7842–7856.

10. Frankish,A., Diekhans,M., Ferreira,A.M., Johnson,R., Jungreis,I., Loveland,J., Mudge,J.M., Sisu,C., Wright,J., Armstrong,J. *et al.* (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, **47**, D766–D773.
11. Pertea,M., Shumate,A., Pertea,G., Varabyou,A., Breitwieser,F.P., Chang,Y.C., Madugundu,A.K., Pandey,A. and Salzberg,S.L. (2018) CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.*, **19**, 208.
12. Hon,C.C., Ramilowski,J.A., Harshbarger,J., Bertin,N., Rackham,O.J., Gough,J., Denisenko,E., Schmeier,S., Poulsen,T.M., Severin,J. *et al.* (2017) An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature*, **543**, 199–204.
13. You,B.H., Yoon,S.H. and Nam,J.W. (2017) High-confidence coding and noncoding transcriptome maps. *Genome Res.*, **27**, 1050–1062.
14. Li,Q., Li,Z., Feng,C., Jiang,S., Zhang,Z. and Ma,L. (2020) Multi-omics annotation of human long non-coding RNAs. *Biochem. Soc. Trans.*, **48**, 1545–1556.
15. Diederichs,S. (2014) The four dimensions of noncoding RNA conservation. *Trends Genet.*, **30**, 121–123.
16. Zhang,C., Zhou,B., Gu,F., Liu,H., Wu,H., Yao,F., Zheng,H., Fu,H., Chong,W., Cai,S. *et al.* (2022) Micropeptide PACMP inhibition elicits synthetic lethal effects by decreasing CtIP and poly(ADP-ribosyl)ation. *Mol. Cell*, **82**, 1297–1312.
17. Li,X.L., Pongor,L., Tang,W., Das,S., Muys,B.R., Jones,M.F., Lazar,S.B., Dangelmaier,E.A., Hartford,C.C. and Grammatikakis,I. (2020) A small protein encoded by a putative lncRNA regulates apoptosis and tumorigenicity in human colorectal cancer cells. *Elife*, **9**, e53734.
18. Huang,J.Z., Chen,M., Chen Gao,X.C., Zhu,S., Huang,H., Hu,M., Zhu,H. and Yan,G.R. (2017) A peptide encoded by a putative lncRNA HOXB-AS3 suppresses colon cancer growth. *Mol. Cell*, **68**, 171–184.
19. Pertea,G. and Pertea,M. (2020) GFF utilities: gffread and gffcompare. *F1000Res*, **9**, 304.
20. Kang,Y.J., Yang,D.C., Kong,L., Hou,M., Meng,Y.Q., Wei,L. and Gao,G. (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.*, **45**, W12–W16.
21. Wang,G., Yin,H., Li,B., Yu,C., Wang,F., Xu,X., Cao,J., Bao,Y., Wang,L., Abbasi,A.A. *et al.* (2019) Characterization and identification of long non-coding RNAs based on feature relationship. *Bioinformatics*, **35**, 2949–2956.
22. Wang,L., Park,H.J., Dasari,S., Wang,S., Kocher,J.P. and Li,W. (2013) CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res.*, **41**, e74.
23. Li,A., Zhang,J. and Zhou,Z. (2014) PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinf.*, **15**, 311.
24. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.*, **22**, 1760–1774.
25. Xie,C., Yuan,J., Li,H., Li,M., Zhao,G., Bu,D., Zhu,W., Wu,W., Chen,R. and Zhao,Y. (2014) NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res.*, **42**, D98–D103.
26. Lee,B.T., Barber,G.P., Benet-Pages,A., Casper,J., Clawson,H., Diekhans,M., Fischer,C., Gonzalez,J.N., Hinrichs,A.S., Lee,C.M. *et al.* (2022) The UCSC genome browser database: 2022 update. *Nucleic Acids Res.*, **50**, D1115–D1122.
27. Guo,C.J., Ma,X.K., Xing,Y.H., Zheng,C.C., Xu,Y.F., Shan,L., Zhang,J., Wang,S., Wang,Y., Carmichael,G.G. *et al.* (2020) Distinct processing of lncRNAs contributes to Non-conserved functions in stem cells. *Cell*, **181**, 621–636.
28. Tate,J.G., Bamford,S., Jubb,H.C., Sondka,Z., Beare,D.M., Bindal,N., Boutselakis,H., Cole,C.G., Creatore,C., Dawson,E. *et al.* (2019) COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **47**, D941–D947.
29. Landrum,M.J., Lee,J.M., Benson,M., Brown,G.R., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Jang,W. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.
30. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E. *et al.* (2019) The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
31. Fadista,J., Manning,A.K., Florez,J.C. and Groop,L. (2016) The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur. J. Hum. Genet.*, **24**, 1202–1205.
32. Jannot,A.S., Ehret,G. and Perneger,T. (2015) P < 5×10(-8) has emerged as a standard of statistical significance for genome-wide association studies. *J. Clin. Epidemiol.*, **68**, 460–465.
33. Kohler,S., Gargano,M., Matentzoglu,N., Carmody,L.C., Lewis-Smith,D., Vasilevsky,N.A., Danis,D., Balagura,G., Baynam,G., Brower,A.M. *et al.* (2021) The human phenotype ontology in 2021. *Nucleic Acids Res.*, **49**, D1207–D1217.
34. Malone,J., Holloway,E., Adamusiak,T., Kapushesky,M., Zheng,J., Kolesnikov,N., Zhukova,A., Brazma,A. and Parkinson,H. (2010) Modeling sample variables with an experimental factor ontology. *Bioinformatics*, **26**, 1112–1118.
35. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
36. Hutter,C. and Zenklusen,J.C. (2018) The cancer genome atlas: creating lasting value beyond its data. *Cell*, **173**, 283–285.
37. Clough,E. and Barrett,T. (2016) The gene expression omnibus database. *Methods Mol. Biol.*, **1418**, 93–110.
38. Li,Z., Liu,L., Jiang,S., Li,Q., Feng,C., Du,Q., Zou,D., Xiao,J., Zhang,Z. and Ma,L. (2021) LncExpDB: an expression database of human long non-coding RNAs. *Nucleic Acids Res.*, **49**, D962–D968.
39. Li,Y., Zhou,H., Chen,X., Zheng,Y., Kang,Q., Hao,D., Zhang,L., Song,T., Luo,H., Hao,Y. *et al.* (2021) SmProt: a reliable repository with comprehensive annotation of small proteins identified from ribosome profiling. *Genomics Proteomics Bioinformatics*, **19**, 602–610.
40. Davis,C.A., Hitz,B.C., Sloan,C.A., Chan,E.T., Davidson,J.M., Gabdank,I., Hilton,J.A., Jain,K., Baymuradov,U.K., Narayanan,A.K. *et al.* (2018) The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, **46**, D794–D801.
41. Betel,D., Koppal,A., Agius,P., Sander,C. and Leslie,C. (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.*, **11**, R90.
42. Nam,J.W., Rissland,O.S., Koppstein,D., Abreu-Goodger,C., Jan,C.H., Agarwal,V., Yildirim,M.A., Rodriguez,A. and Bartel,D.P. (2014) Global analyses of the effect of different cellular contexts on microRNA targeting. *Mol. Cell*, **53**, 1031–1043.
43. Kruger,J. and Rehmsmeier,M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res.*, **34**, W451–W454.
44. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinf.*, **10**, 421.
45. Liu,L., Li,Z., Liu,C., Zou,D., Li,Q., Feng,C., Jing,W., Luo,S., Zhang,Z. and Ma,L. (2022) LncRNAWiki 2.0: a knowledgebase of human long non-coding RNAs with enhanced curation model and database system. *Nucleic Acids Res.*, **50**, D190–D195.
46. Wang,Y.Q., Jiang,D.M., Hu,S.S., Zhao,L., Wang,L., Yang,M.H., Ai,M.L., Jiang,H.J., Han,Y., Ding,Y.Q. *et al.* (2019) SATB2-AS1 suppresses colorectal carcinoma aggressiveness by inhibiting SATB2-Dependent snail transcription and epithelial-mesenchymal transition. *Cancer Res.*, **79**, 3542–3556.
47. Xu,M., Xu,X., Pan,B., Chen,X., Lin,K., Zeng,K., Liu,X., Xu,T., Sun,L., Qin,J. *et al.* (2019) LncRNA SATB2-AS1 inhibits tumor metastasis and affects the tumor immune cell microenvironment in colorectal cancer by regulating SATB2. *Mol. Cancer*, **18**, 135.
48. Dallosso,A.R., Hancock,A.L., Malik,S., Salpekar,A., King-Underwood,L., Pritchard-Jones,K., Peters,J., Moorwood,K., Ward,A., Malik,K.T. *et al.* (2007) Alternately spliced WT1 antisense transcripts interact with WT1 sense RNA and show epigenetic and splicing defects in cancer. *RNA*, **13**, 2287–2299.
49. Wang,W., Lyu,C., Wang,F., Wang,C., Wu,F., Li,X. and Gan,S. (2021) Identification of potential signatures and their functions for acute lymphoblastic leukemia: a study based on the cancer genome atlas. *Front. Genet*, **12**, 656042.
50. CNCB-NGDC Members and Partners. (2022) Database resources of the national genomics data center, china national center for bioinformation in 2022. *Nucleic Acids Res.*, **50**, D27–D38.
51. RNAcentral Consortium (2021) RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.*, **49**, D212–D220.