DATABASE
The Journal of Biological Databases and Curation

Database, 2019, 1–6
doi: 10.1093/database/baz011
Database tool

Database tool

LIVE: a manually curated encyclopedia of experimentally validated interactions of lncRNAs

Gaole An[1,†], Jiaqi Sun[2,†], Chao Ren[1,†], Zhangyi Ouyang[1], Lingyun Zhu[2], Xiaochen Bo[1,*], Shaoliang Peng[2,3,*] and Wenjie Shu[1,*]

[1]Department of Biotechnology, Beijing Institute of Radiation Medicine, Beijing 100850, China, [2]Department of Biology and Chemistry, College of Liberal Arts and Sciences, National University of Defense Technology, Changsha, 410073 Hunan, China and [3]College of Computer Science and Electronic Engineering and National Supercomputing Centre in Changsha, Hunan University, Changsha 410082, China

*Corresponding author: Tel: +86 10 6693 2211; Fax: +86 10 6821 0077; Email: shuwj@bmi.ac.cn
Correspondence may also be addressed to Xiaochen Bo. Email: boxc@bmi.ac.cn and Shaoliang Peng.
Email: pengshaoliang@nudt.edu.cn or pengshaoliang@gmail.com

†These authors contributed equally to this work as joint first authors.

Citation details: An,G., Sun,J., Ren,C. *et al.* LIVE: a manually curated encyclopedia of experimentally validated interactions of lncRNAs. *Database* (2019) Vol. 2019: article ID baz011; doi:10.1093/database/baz011

Received 24 October 2018; Revised 15 January 2019; Accepted 15 January 2019

Abstract

Advances in studies of long noncoding RNAs (lncRNAs) have provided data regarding the regulatory roles of lncRNAs, which perform functional roles through interactions with other functional elements. To track the underlying relationships among lncRNAs, various databases have been developed as repositories for lncRNA data. However, the ability to comprehensively explore the diverse interactions between lncRNAs and other functional elements is limited. To this end, we developed LIVE (LncRNA Interaction Validated Encyclopaedia), an interactive resource to integrate the diverse interactions of functional elements with lncRNAs. LIVE is a manually curated database of experimentally validated interactions of lncRNAs with genes, proteins and other various functional elements. By mining publications, we constructed LIVE with the following three interaction networks: a binding interaction network, a regulation network and a disease network; then, we combined them to form a comprehensive lncRNA interaction network. The current release of LIVE contains the validated interactions of 572 lncRNAs in humans and mice with 103 proteins, 209 genes, 56 transcription factors and 194 diseases. LIVE provides an interactive interface with charts and figures to aid users in searching and browsing interactions with lncRNAs. LIVE will greatly facilitate further investigation into the regulatory roles of lncRNAs and is freely available.

Database URL: https://live.bioinfotech.org

Introduction

LncRNAs play important functional roles in regulating biological and cellular processes, including proliferation, differentiation and development, through interacting with various genes, proteins and functional elements (1, 2). With the rapid expansion of research on lncRNAs, a number of databases have been developed to satisfy the demand for exploration of the regulatory roles of lncRNAs.

Unlike the early databases developed to deposit candidate lncRNAs from both experimental validation and computational prediction studies, recent databases have focused on the specific functional roles of lncRNAs. LincSNP contains single nucleotide polymorphisms (SNPs) identified in human lncRNAs and transcription factor binding sites (3). LncReg collects the validated regulatory relationships of lncRNAs (4). Lnc2Meth was developed specifically to store relationships between human lncRNAs and DNA methylation (5). LncChrom includes validated lncRNA-chromatin interactions (6). However, these databases face accuracy problems due to the high false positive rate of data from computational predictions and high-throughput experiments. Thus, there is still an urgent demand for an experimentally validated lncRNA database that has been manually curated. EVlncRNAs manually curated publications to include experimentally validated lncRNA interactions (7), but the potential interaction networks beneath publications are not fully revealed.

Here we present LIVE (LncRNA Interaction Validated Encyclopaedia), a manually curated database of experimentally validated interactions between functional elements and lncRNAs from published literature available on PubMed prior to July 1, 2018. LIVE contains the validated interactions of 572 lncRNAs with 103 proteins, 209 genes, 56 transcription factors and 194 diseases in humans and mice. These validated interactions are classified into three networks, namely, a binding interaction network, a regulation network and a disease association network. By assembling these three networks, we generated a complete picture of the lncRNA interaction network with diverse types of functional regulatory elements and interactions. Moreover, an interactive interface and analysis kits are provided to help users mine the potential regulatory roles of lncRNAs in interaction networks.

Materials and methods

To provide a comprehensive collection of experimentally validated interactions with lncRNAs, we manually curated the available publications to construct LIVE (Figure.1). First, we used 'LncRNA', 'Long noncoding RNA', and 'Long ncRNA' along with their plural forms as keywords to search the PubMed database up to July 1, 2018. The species mentioned in publications were limited to human and mouse. Second, we developed a word segment system to pre-process the abstracts and extract the keywords, including species, experiment type, disease and lncRNA. The word segment system is derived from Python module 'jieba.analyse.extract_tags' (https://pypi.org/project/jieba/), which is based on TF-IDF (term frequency—inverse document frequency) algorithm (8). TF-IDF algorithm is a statistical method used to evaluate the importance of a word to one of the documents in a document set. The importance of a word increases proportionally with the number of times it appears in a document, but decreases inversely with its frequency in the corpus. This word segment system helps extracting keywords such as disease, tissues and molecules from abstracts. Moreover, we standardized the keyword lists by comparing them with the existing databases (MeSH, HGNC, MGI and MalaCards) (9, 10). The keyword list was used to assist in describing the features of each publication by building a keyword index, which improved the efficiency and accuracy of the following manual curation. Third, we curated the abstracts to determine whether the described lncRNAs were associated with the keywords describing regulation or interaction. Fourth, we carefully curated the main text to extract the interactions and ensured that these interactions were experimentally validated. Finally, we labeled these interactions with the corresponding abstract keywords to aid the following categorization of interactions. In total, in the construction

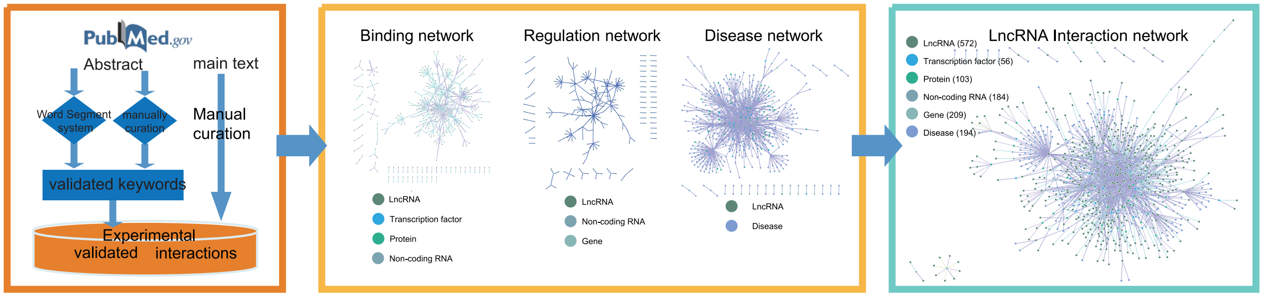
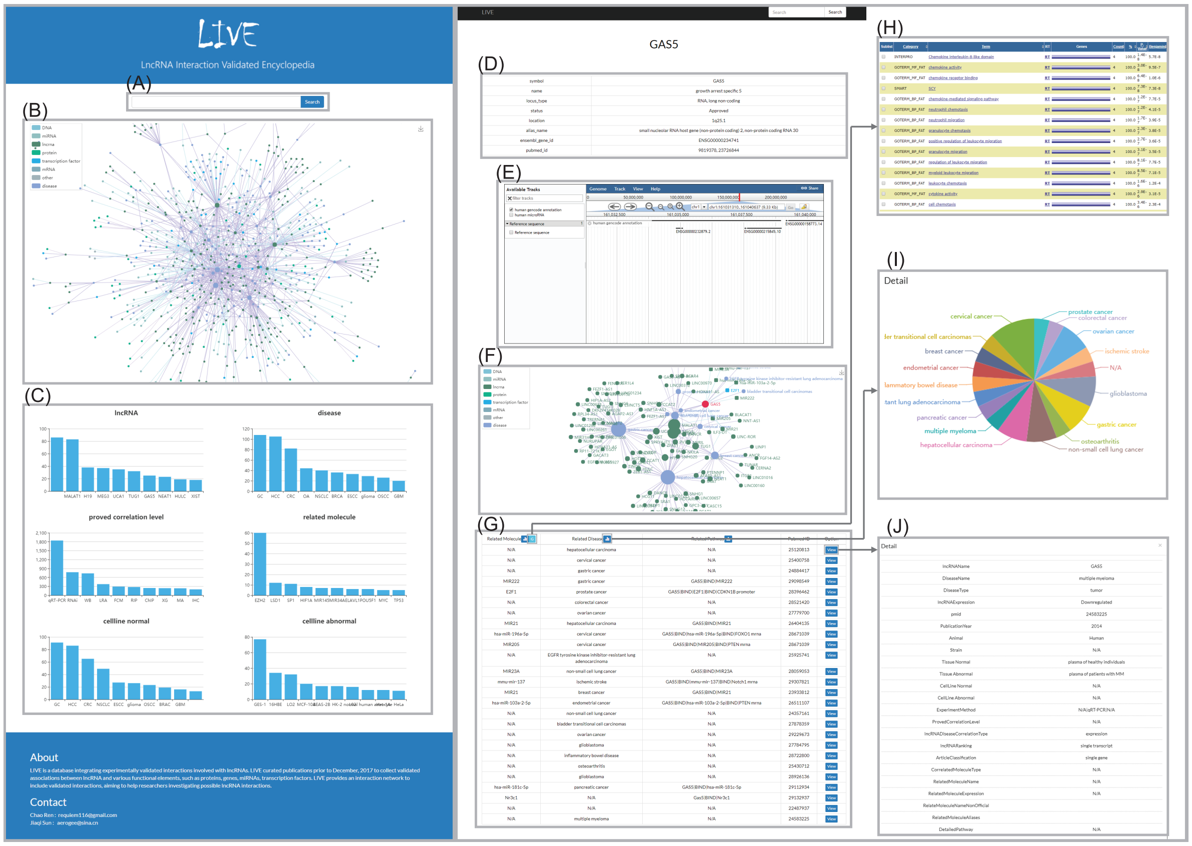


Figure. 1. Data flow of LIVE. LIVE curated publications involved with lncRNA interactions from human and mouse to construct database.



About

LIVE is a database integrating experimentally validated interactions involved with lncRNAs. LIVE curated publications prior to December, 2017 to collect validated associations between lncRNA and various functional elements, such as proteins, genes, miRNAs, transcription factors. LIVE provides an interaction network to include validated interactions, aiming to help researchers investigating possible lncRNA interactions.

Contact

Chao Ren : requiem116@gmail.com
Jiaqi Sun : aerogee@sina.cn

symbol	GAS5
name	growth arrest specific 5
locus_type	RNA, long non-coding
status	Approved
location	1q25.1
alias_name	small nucleolar RNA host gene (non-protein coding) 2, non-protein coding RNA 30
ensembl_gene_id	ENSG00000234741
pubmed_id	9819378, 23726844

Related Molecule	Related Disease	Related Pathway	PubmedID	Option
N/A	hepatocellular carcinoma	N/A	25120813	View
N/A	cervical cancer	N/A	25400758	View
N/A	gastric cancer	N/A	24884417	View
MIR222	gastric cancer	GAS5|BIND|MIR222	29098549	View
E2F1	prostate cancer	GAS5|BIND|E2F1|BIND|CDKN1B promoter	28396462	View
N/A	colorectal cancer	N/A	28521420	View
N/A	ovarian cancer	N/A	27779700	View
MIR21	hepatocellular carcinoma	GAS5|BIND|MIR21	26404135	View
hsa-miR-196a-5p	cervical cancer	GAS5|BIND|hsa-miR-196a-5p|BIND|FOXO1 mrna	28671039	View
MIR205	cervical cancer	GAS5|BIND|MIR205|BIND|PTEN mrna	28671039	View
N/A	EGFR tyrosine kinase inhibitor-resistant lung adenocarcinoma	N/A	25925741	View
MIR23A	non-small cell lung cancer	GAS5|BIND|MIR23A	28059053	View
mmu-mir-137	ischemic stroke	GAS5|BIND|mmu-mir-137|BIND|Notch1 mrna	29307821	View
MIR21	breast cancer	GAS5|BIND|MIR21	23933812	View
hsa-miR-103a-2-5p	endometrial cancer	GAS5|BIND|hsa-miR-103a-2-5p|BIND|PTEN mrna	26511107	View
N/A	non-small cell lung cancer	N/A	24357161	View
N/A	bladder transitional cell carcinomas	N/A	27878359	View
N/A	ovarian cancer	N/A	29229673	View
N/A	glioblastoma	N/A	27784795	View
N/A	inflammatory bowel disease	N/A	28722800	View
N/A	osteoarthritis	N/A	25430712	View
N/A	glioblastoma	N/A	28926136	View
hsa-miR-181c-5p	pancreatic cancer	GAS5|BIND|hsa-miR-181c-5p	29112934	View
Nr3c1	N/A	Gas5|BIND|Nr3c1	29132937	View
N/A	N/A	N/A	22487937	View
N/A	multiple myeloma	N/A	24583225	View

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	INTERPRO	Chemokine interleukin-8-like domain	RT		4	100.0	1.4E-8	5.7E-8
□	GOTERM_MF_FAT	chemokine activity	RT		4	100.0	3.0E-8	9.5E-7
□	GOTERM_MF_FAT	chemokine receptor binding	RT		4	100.0	6.4E-8	1.0E-6
□	SMART	SCY	RT		4	100.0	7.3E-8	7.3E-8
□	GOTERM_BP_FAT	chemokine-mediated signaling pathway	RT		4	100.0	1.2E-7	7.7E-5
□	GOTERM_BP_FAT	neutrophil chemotaxis	RT		4	100.0	1.2E-7	4.1E-5
□	GOTERM_BP_FAT	neutrophil migration	RT		4	100.0	1.7E-7	3.9E-5
□	GOTERM_BP_FAT	granulocyte chemotaxis	RT		4	100.0	2.3E-7	3.8E-5
□	GOTERM_BP_FAT	positive regulation of leukocyte migration	RT		4	100.0	2.7E-7	3.6E-5
□	GOTERM_BP_FAT	granulocyte migration	RT		4	100.0	3.1E-7	3.5E-5
□	GOTERM_BP_FAT	regulation of leukocyte migration	RT		4	100.0	8.1E-7	7.7E-5
□	GOTERM_BP_FAT	myeloid leukocyte migration	RT		4	100.0	8.5E-7	7.1E-5
□	GOTERM_BP_FAT	leukocyte chemotaxis	RT		4	100.0	1.6E-6	1.2E-4
□	GOTERM_MF_FAT	cytokine activity	RT		4	100.0	2.9E-6	3.1E-5
□	GOTERM_BP_FAT	cell chemotaxis	RT		4	100.0	3.4E-6	2.3E-4

lncRNAName	GAS5
DiseaseName	multiple myeloma
DiseaseType	tumor
lncRNAExpression	Downregulated
pmid	24583225
PublicationYear	2014
Animal	Human
Strain	N/A
Tissue Normal	plasma of healthy individuals
Tissue Abnormal	plasma of patients with MM
CellLine Normal	N/A
CellLine Abnormal	N/A
ExperimentMethod	N/A|qRT-PCR|N/A
ProvedCorrelationLevel	N/A
lncRNADiseaseCorrelationType	expression
lncRNARanking	single transcript
ArticleClassification	single gene
CorrelatedMoleculeType	N/A
RelatedMoleculeName	N/A
RelatedMoleculeExpression	N/A
RelateMoleculeNameNonOfficial	
RelatedMoleculeAliases	
DetailedPathway	N/A

Figure. 2. Interface of LIVE. (A–C) Main page. Search box (A), integrated network of all interactions (B) and metadata of database (C). (D–J) Detail page. Basic information of lncRNA (D), genomic browser (E), interaction network (F), brief publication information (G), GO analysis results (H), pie chart (I) and detailed information (J).

of the LIVE database, over 10 000 publications were cross-curated, and each publication was curated by at least two specialists.

Curated interactions are categorized and integrated into three different interaction networks: binding interaction network, regulation network and disease association network. Publications providing experimental evidence of lncRNA binding with proteins, microRNA and transcription factors by methods such as RIP or CoIP are used to construct binding interaction network. Regulation associations based on expression experiment such as qRT-PCR or knockdown experiments are used to construct regulation network. The disease association network is constructed based on the publications in which lncRNAs are described to play validated functional roles in diseases. These three networks are assembled into the final lncRNA interaction network by merging the shared terms in their respective networks.

LIVE was developed based on the Django framework, with JavaScript for the front-end and MySQL for the back-end. The Bootstrap framework was deployed to provide compatibility for mobiles, tablets and desktop computers. To improve the visualization tools, Jbrowse (11) and echarts3 were used to provide interactive charts and graphs.

Results

Interface and functions

User-friendly interfaces were developed to provide direct searching and browsing (Figure 2). The search box on the main page enables users to search any keywords, including lncRNAs, proteins, genes, microRNAs and diseases. LIVE supports fuzzy searches and autocompletes with all keywords to help users investigate the results of interest. The page will redirect to the corresponding detailed page if the term exactly matches one in the database or provide all candidate results if the term partially matches terms in the database. The detailed pages for the lncRNAs consist of four sections, namely, basic information, interactive network, genomic browser and interaction details. Detailed pages for other terms, such as genes and proteins, only provide the genomic browser and interaction details. For each entry, the interactive network connected with it (represented as a node) is visualized. The node size reflects the degree of each node, indicating its importance in the interaction

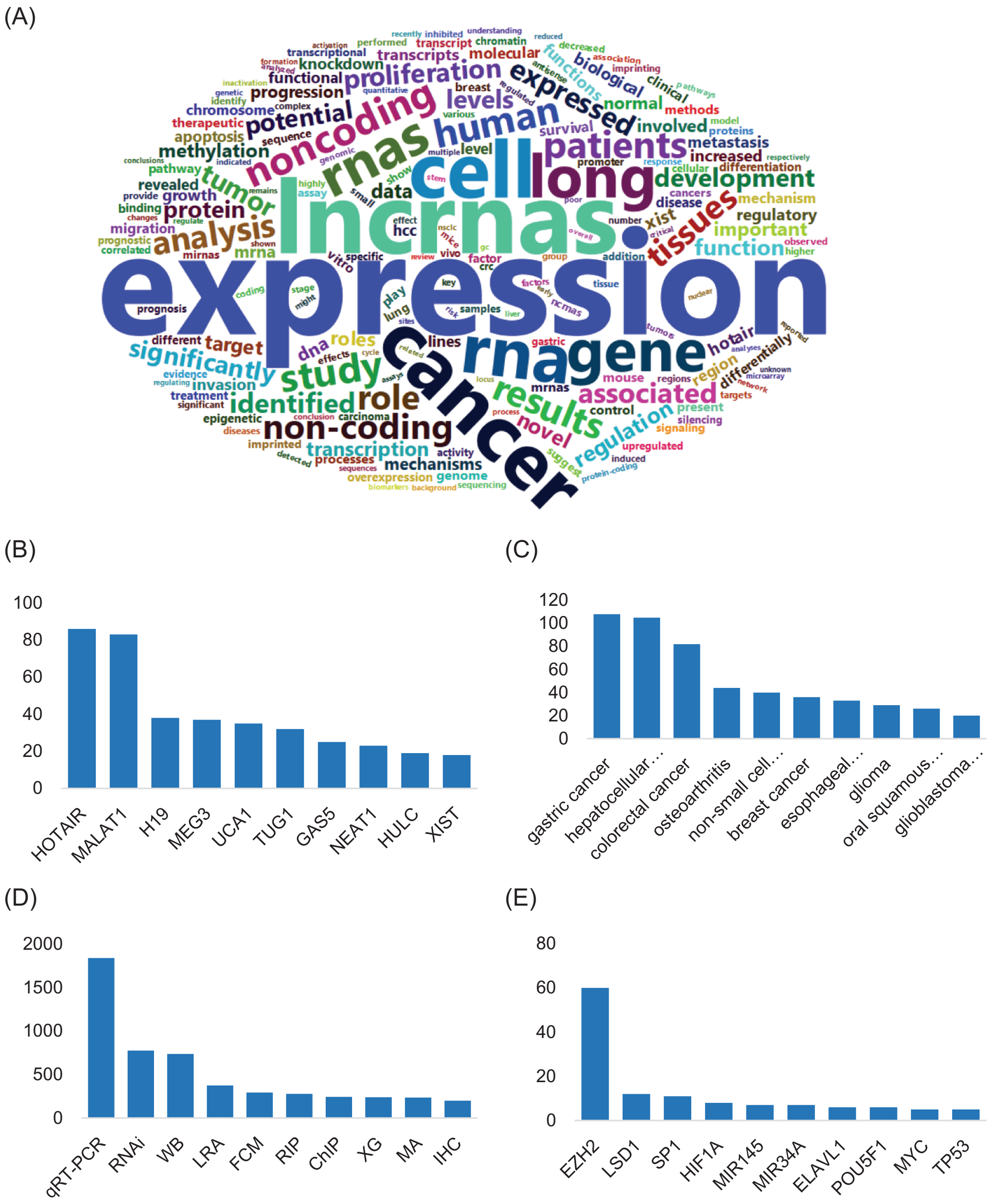


Figure. 3. Metadata of LIVE. (A) Word cloud graph of keywords in LIVE. (B) Top 10 lncRNAs involved with lncRNA interactions ranking by number of interactions, (C) number of top 10 diseases involved with lncRNA interactions ranking by number of validated interactions evolved, (D) number of top 10 experiment type used in all interactions ranking by number of publications and (E) number of top 10 functional elements used in all interactions ranking by number of validated interactions evolved.

network. Moreover, we expanded the interaction network to include the nodes that share nodes directly connected with the entry for which the user searched. The genomic browser section contains a genomic browser for each lncRNA entry to illustrate its nearby genomic elements. For entries in the interaction details section, external links to the standard databases are provided for available terms, such as GeneCard for genes, UniPort for proteins and MalaCard for diseases. A pie button and a GO button, in addition to column labels, provide analysis results in pop-out windows. More detailed information is provided by clicking the view button, such as sample and tissue information, experimental methods, detailed pathway descriptions and publication information.

Database metadata

Available publication abstracts were segmented to investigate the popular concerns in lncRNA studies. A large number (73%) of publications mentioned the expression patterns of lncRNAs and other genes, with ‘expression’ and ‘co-expression’ frequently noted. The regulatory functions

Table 1. Comparison with other lncRNA interaction database

Database	Data type					Interaction type			Manual curation	Metadata analysis	Interactive graphs
	Disease	TF	proteins	genes	ncRNA	Binding network	Regulation network	Disease network			
LncReg	No	No	Yes	Yes	No	No	Yes	Yes	Partially	No	No
LnChrom	Yes	Yes	Yes	No	No	Yes	No	Yes	Partially	No	No
Lnc2Meth	No	No	No	No	No	No	Yes	No	Totally	No	No
LincSNP	No	No	No	No	Yes	No	Yes	Yes	No	No	Yes
EVlncRNAs	Yes	Yes	Yes	Yes	Yes	No	No	No	Totally	No	No
LIVE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Totally	Yes	Yes

of lncRNAs are focused on derivatives of the term ‘regulate’ mentioned in more than half the publications. Interestingly, ‘methylation’ and ‘proliferation’ are hot keywords, as ∼35% of the analyzed publications involved those terms (Figure 3A).

LIVE also provides statistics regarding the involved functional elements and diseases in the lncRNA interaction network to facilitate research further exploring the functional roles of lncRNAs. Metadata analysis of publications illustrated preference and tendency in lncRNA studies (Figure 3B–E). A total of 83% of lncRNA-related studies are focused on tumors and cancer. The top 10 most studied lncRNAs are in 35% of the total studies. HOTAIR and gastric cancer are the most studied lncRNA and disease, respectively. The validated interactions curated from publications demonstrated that lncRNAs interact with functional elements and diseases in a very complex manner. HOTAIR, the most studied lncRNA, interacts with 26 various functional elements, such as EZH2 and PCBP1, and 40 diseases, including gastric cancer, breast cancer and colon cancer (12–16). Moreover, HOTAIR plays different roles in regulatory networks by acting in diverse ways, such as by binding with microRNAs or cis-regulating nearby genes. Additionally, 57 different lncRNAs play functional roles in gastric cancer by interacting with 42 different functional elements, including microRNAs and transcription factors. Furthermore, HOTAIR-related functional elements also interact with other lncRNAs in different tissues and diseases. Notably, 93% of lncRNAs are associated with more than one disease, and 98% are associated with more than one lncRNA. These metadata provided by our LIVE database will greatly facilitate the ability of researchers to clarify the complex interactions of lncRNAs in future studies.

Discussion

By mining and manually curating publications to extract experimentally validated interactions, we generated LIVE, which presents the comprehensive interactions of lncRNAs with genes, proteins, ncRNAs, transcription factors and diseases in humans and mice. Moreover, LIVE provides powerful search engines and extremely conveniently interactive visualization tools. Further metadata based on the validated interactions can help researchers investigate the possible interactions with lncRNAs and explore the functional roles of lncRNAs. Comparing with similar databases, our LIVE illustrate its specialty (Table 1).

LIVE will continue to be curated and updated in the coming years, and the next update will focus on providing more kits to help analyze the metadata of the included literature.

In summary, LIVE provides a highly interactive visual database of experimentally validated interactions with lncRNAs. LIVE will facilitate the ability of researchers to investigate the potential roles and mechanisms of lncRNAs in diverse biological processes and diseases.

Funding

Major Research Plan of the National Key R&D Program of China (2016YFC0901600); the National Natural Science Foundation of China (U1435222 and 61772543); and National Key R&D Program of China (2018YFC0910405).

Conflict of interest. None declared.

References

1. Kung,J.T., Colognori,D. and Lee,J.T. (2013) Long noncoding RNAs: past, present, and future. *Genetics*, **193**, 651–669.
2. Goff,L.A. and Rinn,J.L. (2015) Linking RNA biology to lncRNAs. *Genome Res.*, **25**, 1456–1465.
3. Ning,S., Yue,M., Wang,P. *et al.* (2017) LincSNP 2.0: an updated database for linking disease-associated SNPs to human long non-coding RNAs and their TFBSs. *Nucleic Acids Res.*, **45**, D74–D78.
4. Zhou,Z., Shen,Y., Khan,M.R. *et al.* (2015) LncReg: a reference resource for lncRNA-associated regulatory networks. *Database (Oxford)*, **2015**.

5. Zhi,H., Li,X., Wang,P. *et al.* (2018) Lnc2Meth: a manually curated database of regulatory relationships between long non-coding RNAs and DNA methylation associated with human disease. *Nucleic Acids Res.*, **46**, D133–D138.
6. Yu,F., Zhang,G., Shi,A. *et al.* (2018) LnChrom: a resource of experimentally validated lncRNA-chromatin interactions in human and mouse. *Database (Oxford)*, **2018**.
7. Zhou,B., Zhao,H., Yu,J. *et al.* (2018) EVLncRNAs: a manually curated database for long non-coding RNAs validated by low-throughput experiments. *Nucleic Acids Res.*, **46**, D100–D105.
8. Salton,G. and Yu,C.T. (1973) On the construction of effective vocabularies for information retrieval. *SIGIR Forum*, **9**, 48–60.
9. Yates,B., Braschi,B., Gray,K.A. *et al.* (2017) Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.*, **45**, D619–D625.
10. Smith,C.L., Blake,J.A., Kadin,J.A. *et al.* (2018) Mouse Genome Database (MGD)-2018: knowledgebase for the laboratory mouse. *Nucleic Acids Res.*, **46**, D836–D842.
11. Buels,R., Yao,E., Diesh,C.M. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.
12. Ren,Y., Jia,H.H., Xu,Y.Q. *et al.* (2018) Paracrine and epigenetic control of CAF-induced metastasis: the role of HOTAIR stimulated by TGF-ss1 secretion. *Mol. Cancer*, **17**, 5.
13. Liu,X.H., Sun,M., Nie,F.Q. *et al.* (2014) Lnc RNA HOTAIR functions as a competing endogenous RNA to regulate HER2 expression by sponging miR-331-3p in gastric cancer. *Mol. Cancer*, **13**, 92.
14. Wu,Z.H., Wang,X.L., Tang,H.M. *et al.* (2014) Long non-coding RNA HOTAIR is a powerful predictor of metastasis and poor prognosis and is associated with epithelial-mesenchymal transition in colon cancer. *Oncol. Rep.*, **32**, 395–402.
15. Li,P., Zhang,X., Wang,L. *et al.* (2017) lncRNA HOTAIR contributes to 5FU resistance through suppressing miR-218 and activating NF-kappaB/TS signaling in colorectal cancer. *Mol. Ther. Nucleic Acids*, **8**, 356–369.
16. Zhang,Z.Z., Shen,Z.Y., Shen,Y.Y. *et al.* (2015) HOTAIR long noncoding RNA promotes gastric cancer metastasis through suppression of poly r(C)-binding protein (PCBP) 1. *Mol. Cancer Ther.*, **14**, 1162–1170.