D962–D968 *Nucleic Acids Research, 2021, Vol. 49, Database issue* *Published online 12 October 2020*
doi: 10.1093/nar/gkaa850

LncExpDB: an expression database of human long non-coding RNAs

Zhao Li[1,2,3,4,†], Lin Liu[1,2,3,4,†], Shuai Jiang[1,2,3], Qianpeng Li[1,2,3,4], Changrui Feng[1,2,3,4], Qiang Du[1,2,3,4], Dong Zou[1,2,3], Jingfa Xiao[1,2,3,4], Zhang Zhang[1,2,3,4,*] and Lina Ma[1,2,3,*]

[1]China National Center for Bioinformation, Beijing 100101, China, [2]National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, [3]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China and [4]University of Chinese Academy of Sciences, Beijing 100101, China

Received August 15, 2020; Revised September 12, 2020; Editorial Decision September 20, 2020; Accepted September 22, 2020

ABSTRACT

Expression profiles of long non-coding RNAs (lncRNAs) across diverse biological conditions provide significant insights into their biological functions, interacting targets as well as transcriptional reliability. However, there lacks a comprehensive resource that systematically characterizes the expression landscape of human lncRNAs by integrating their expression profiles across a wide range of biological conditions. Here, we present LncExpDB (https://bigd.big.ac.cn/lncexpdb), an expression database of human lncRNAs that is devoted to providing comprehensive expression profiles of lncRNA genes, exploring their expression features and capacities, identifying featured genes with potentially important functions, and building interactions with protein-coding genes across various biological contexts/conditions. Based on comprehensive integration and stringent curation, LncExpDB currently houses expression profiles of 101 293 high-quality human lncRNA genes derived from 1977 samples of 337 biological conditions across nine biological contexts. Consequently, LncExpDB estimates lncRNA genes' expression reliability and capacities, identifies 25 191 featured genes, and further obtains 28 443 865 lncRNA-mRNA interactions. Moreover, user-friendly web interfaces enable interactive visualization of expression profiles across various conditions and easy exploration of featured lncRNAs and their interacting partners in specific contexts. Collectively, LncExpDB features comprehensive integration and curation of lncRNA expression profiles and thus will serve as a fundamental resource for functional studies on human lncRNAs.

INTRODUCTION

Accumulating evidences have shown that long non-coding RNAs (lncRNAs) can act in *cis* or *trans* to perform diverse functions including regulating gene transcription and RNA splicing, modulating the activity or abundance of RNAs and proteins, and organizing nuclear domains (1). They are extensively implicated in cell-fate programming/reprogramming (2), differentiation (3), development (4) and especially in human diseases (5–7). While hundreds of thousands of human lncRNAs have been identified primarily attributable to the rapid development of high-throughput sequencing technology in recent years, only a small subset of them have been well characterized (8–10).

Nowadays, RNA-seq data are generated at exponential rates and scales, enabling the identification of lncRNAs and investigation of their expression profiles and accordingly providing the most direct evidence for identifying their potential functions across various biological conditions (11). Compared with mRNAs, particularly, lncRNAs feature higher tissue-specificity (12) and wider subcellular localizations (13), indicating their specific biological roles in gene expression regulation. Accordingly, multiple databases have been developed to integrate lncRNA expression profiles from different aspects. However, these databases have several limitations. First, they focus on specific biological contexts, e.g. normal and cancer tissue/cell (e.g. LncBook, RefLnc) (9,12,14–18), organ development (e.g. lncRNAtor) (17,19), subcellular localization (e.g. LncATLAS) (20) or exosome (e.g. NONCODE, exoRBase) (18,21). To our knowledge, lncRNA expression is associated with at least nine biological contexts of general interests, including normal tissue/cell line, cancer tissue/cell line, subcellular

*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, the first two authors should be regarded as Joint First Authors.

localization, exosome, cell differentiation, preimplantation embryo, organ development, circadian rhythm and virus infection. Second, existing databases have different numbers of lncRNAs with different annotation/curation criteria adopted (e.g. 140k lncRNA genes in LncBook (9), 16k in Expression Atlas (22); see a review in (11)), leading to the inconsistency of functional lncRNAs as well as their annotations. Third, none of them fully characterizes expression features of lncRNAs, including expression level, specificity, breadth, capacity, and co-expression network. Considering massive RNA-seq datasets across diverse conditions publicly available, it is desirable to have a dedicated expression database that comprehensively integrates expression profiles and systematically characterizes expression features of human lncRNAs.

To this end, we developed LncExpDB (https://bigd.big.ac.cn/lncexpdb), an expression database of human lncRNAs. Based on manual curation and standardized analysis, LncExpDB features comprehensive integration of high-quality expression profiles of human lncRNAs across diverse biological contexts and conditions. It estimates a wide range of expression features for each lncRNA gene, characterizes potentially functional lncRNAs and identifies lncRNA–mRNA interactions by co-expression networks (Figure 1). Moreover, LncExpDB is equipped with user-friendly web interfaces, providing functionalities for data query, browsing, visualization as well as easy access.

MATERIALS AND METHODS

LncRNA gene integration and curation

LncRNA transcripts were integrated from LncBook v1.2 (9), RefLnc (12), GENCODE v33 (23), CHESS v2.2 (24), FANTOM-CAT (25) and BIGTranscriptome (26), and curated with the following three steps: (i) To obtain a high-confidence lncRNA reference, redundant/questionable/incomplete transcripts were removed. Specifically, GffCompare (27) was used to compare different lncRNA entries and identify redundancy, mapping error, possible pre-mRNA fragment and polymerase run-on, which are allocated as '=', 's', 'e' and 'p', respectively. Single-exon transcripts that are part of multi-exon transcripts and located in their exon regions and transcripts with very short exons (<15 nt) at the 5′ and 3′ ends, were considered as incomplete transcripts. (ii) We further identified lncRNAs based on their sequence length and coding potential. Transcripts shorter than 200 nt were excluded. Four algorithms, namely, LGC (28), CPC2 (29), CPAT (30) and PLEK (31), were used for coding potential estimation; transcripts identified as lncRNAs by at least three algorithms were retained. iii) To provide a reliable list of human lncRNAs, lncRNAs that have no strand information were removed. Consequently, a total of 331 244 lncRNA transcripts were obtained. It is noted that LncBook integrated lncRNAs from GENCODE v27, NONCODE v5.0, LNCipedia v4.1, MiTranscriptome beta and HGNC (9), and about 10 databases thus have contributed to the lncRNA integration of LncExpDB, indicating a comprehensive coverage of human lncRNA resources.

Following the strategy used for the assignment of lncRNA transcripts (32,33), transcripts sharing exonic sequences in the same strand are linked together and considered as the same gene, and we assigned lncRNA transcripts into gene loci using GffCompare (27). Thus, a high-confidence list of 101 293 lncRNA genes were obtained. Accordingly, curated annotations of lncRNA genes as well as protein-coding genes were merged and derived from the above databases and GENCODE v33, respectively.

RNA-seq data collection

A total of 24 RNA-seq datasets across 1977 samples were collected from GEO (https://www.ncbi.nlm.nih.gov/gds/), SRA (https://www.ncbi.nlm.nih.gov/sra/) and ArrayExpress (https://www.ebi.ac.uk/arrayexpress/), covering 337 biological conditions of nine important biological contexts, including normal tissue/cell line, organ development, preimplantation embryo, cell differentiation, subcellular localization, exosome, cancer cell line, virus infection and circadian rhythm. The detailed information for each dataset is summarized in Supplementary Table S1.

Read mapping, quantification and normalization

A standardized RNA-seq analysis pipeline was employed for all samples. Specifically, Trimmomatic (version 0.39) (34) was used to filter out low-quality reads and remove the adapters. STAR (version 2.7.1a) (35) was used to map the sequencing reads to the human reference genome (version hg38/GRCh38 from UCSC). The uniquely mapped reads were counted using featureCounts (version 2.0.0) (36) with a strand-specific parameter '-s'. Kallisto (version 0.46.1) (37) was used to quantify transcript-level abundances of all samples. Considering the robustness of the TMM (weighted trimmed mean of M-values) method in read counts normalization (38), TMM in the package edgeR (39) was used to normalize read counts for samples within the same project.

Gene expression analysis

To obtain the normalized expression levels, TPM (transcripts per million), FPKM (fragments per kilobase of transcript per million mapped reads) and CPM (counts per million) were calculated (publicly accessible at https://bigd.big.ac.cn/lncexpdb/downloads). Expression values of both lncRNA genes and mRNA genes were averaged among biological replicates. All expressed genes (averaged expression values $\geq$ 1 TPM) were further ranked under a specific condition (time point/stage/tissue/cell/component/process); genes whose expression values are greater than the upper quantile are considered as 'H' (high expression level), those less than the lower quantile as 'L' (low expression level), and the remaining as 'M' (medium expression level). High-capacity lncRNAs (HCL) are defined as lncRNA genes that have the potential to be highly expressed and thus should have at least one H among all conditions, low-capacity lncRNAs (LCL) are those that have L in all conditions, and the remaining are medium-capacity lncRNAs (MCL).

Furthermore, featured genes were identified by using maSigPro (40) with significant time-course expression pat-

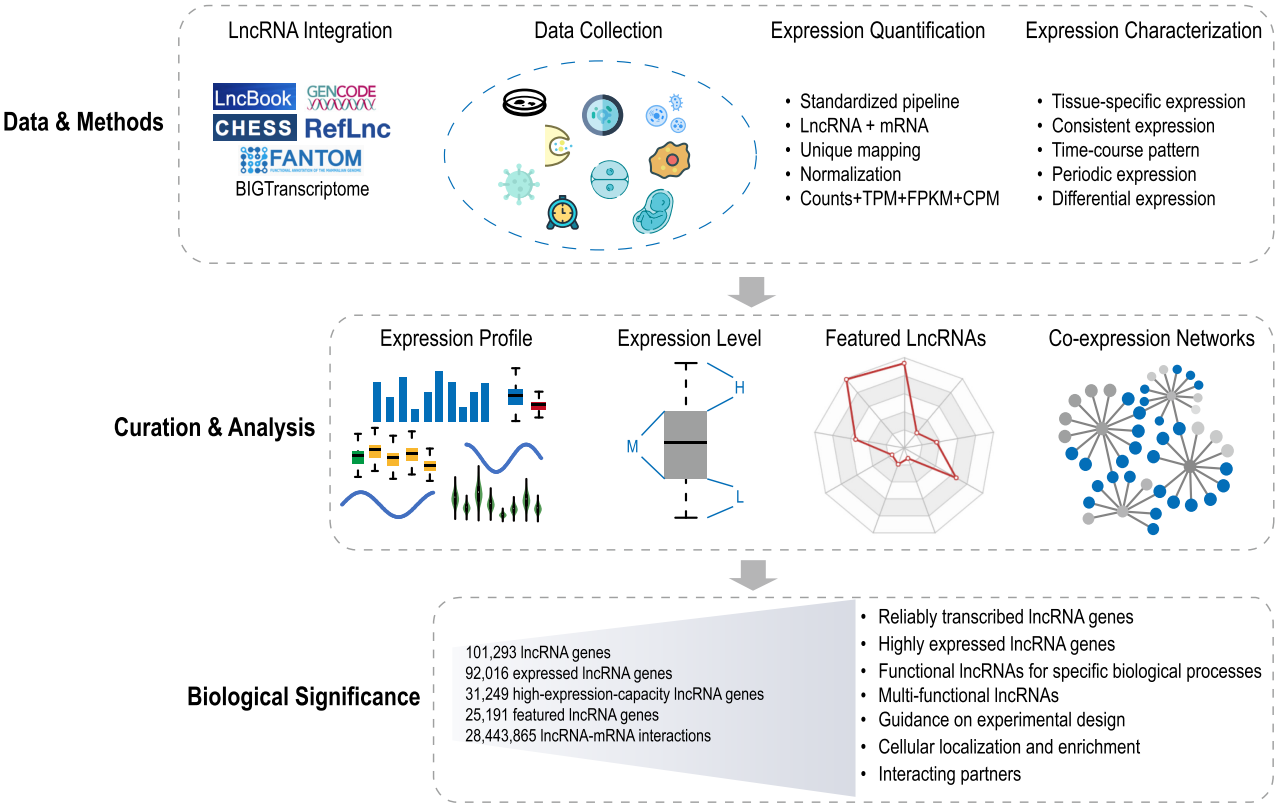

Figure 1. Data curation and analysis workflow of LncExpDB. LncExpDB curates and analyzes expression profiles of human lncRNAs across diverse biological contexts/conditions based on comprehensive integration, accurate quantification and specialized analysis, to systematically characterize expression signatures and identify featured lncRNA genes.

terns (R-square ≥ 0.7 and P-value < 0.05) across different time points/stages. By using the index τ (41), time point/stage-specific genes and tissue/cell-specific genes were defined as those with $\tau \geq 0.9$. Consistently expressed genes were defined as those with $\tau \leq 0.35$. Additionally, specifically expressed genes or consistently expressed genes were strictly screened out with maximum expression values greater than 10 TPM. Differentially expressed genes ($|\log_2$ fold change$| \geq 1$ and adjusted P-value < 0.05) for virus infection and exosomes were identified with the help of DeSeq2 (42). Organelle-enriched genes were defined as specifically highly expressed genes for each subcellular organelle/compartment compared with the whole cell ($\log_2$ fold change ≥ 1 and adjusted P-value (FDR) < 0.05), with more stringent criteria than the study (FDR < 0.05) from which our data was derived (43). Circadian genes were identified (meta2d, BH.Q < 0.05) using the software MetaCycle (44).

LncRNA-mRNA interaction prediction

LncRNA-mRNA interactions were predicted by co-expression networks, which were identified using the Pearson correlation coefficient (P-value < 0.01 and $|r| \geq 0.5$). Nearly all datasets were used for co-expression network construction, except the dataset of 'COVID patients vs. normal control', due to extremely smaller sampling size ($n = 4$).

Implementation

LncExpDB was built with String Boot (https://spring.io/projects/spring-boot) as backend web framework and MySQL (https://www.mysql.com/) as database engine. Web interfaces were developed by JSP (Java Script Pages) and AJAX (Asynchronous JavaScript and XML). Bootstrap (https://getbootstrap.com/) was adopted as a front-end framework, providing a series of templates for designing web pages with consistent interface components. Also, data visualization was rendered by Highcharts (https://www.highcharts.com.cn/), Echarts (http://echarts.apache.org/zh/index.html), Plotly.js (https://plotly.com/javascript/), DataTables (https://datatables.net/) and UCSC Genome Browser (45).

DATABASE CONTENTS AND FEATURES

LncExpDB presents a comprehensive and high-quality collection of 101 293 human lncRNA genes (corresponding to 331 244 transcripts). It houses abundant expression profiles of these lncRNAs across 337 biological conditions that fall into nine important biological contexts, involving normal tissue/cell line, cancer cell line, subcellular localization, exosome, cell differentiation, preimplantation embryo, organ development, circadian rhythm and virus infection. Moreover, LncExpDB identifies 25 191 featured lncRNA genes and characterizes 28 443 865 co-expression interac-

tions between 24 508 lncRNA genes and 17 345 mRNA genes. The reference list of 101 293 lncRNA genes, lncRNA genes/transcripts' expression values as well as their expression analyzed results are all publicly available at https://bigd.big.ac.cn/lncexpdb/downloads.

LncRNA expression profiles

Based on comprehensive expression profiles across multiple biological contexts, LncExpDB features value-added curation and analysis to provide reliably transcribed lncRNA genes. Consequently, we find that 92 016 lncRNA genes (90.8%) are supported with reliable transcriptional evidence (threshold of expression value is 1 TPM), distributing unevenly across the nine biological contexts (Figure 2A). Among the reliably transcribed genes, the majority (82.6%) are expressed in at least two biological contexts, and 3318 lncRNAs (3.6%) are expressed in all the nine contexts (Figure 2B).

Moreover, considering massive expression profiles across abundant conditions in LncExpDB, we define high-capacity lncRNAs (HCL) as those that have the potential to be highly expressed (see Materials and Methods) in at least one biological condition. Therefore, HCL are those whose expression values are greater than the upper quantile in at least one condition. Likewise, low-capacity lncRNAs (LCL) are those whose expression values are always less than lower quantile in all biological conditions, and the remaining are medium-capacity lncRNAs (MCL). The expression levels for inferring expression capacities of the 101 293 lncRNAs across 337 biological conditions are available at https://bigd.big.ac.cn/lncexpdb/downloads. Our analyzed results show that, among the 92 016 expressed lncRNA genes across the nine biological contexts, 34% are HCL, 55.7% are MCL, and 10.3% are LCL and that 74.8% of HCL are specifically highly expressed under only one biological context (Figure 2C). However, the number of HCL does not change so much (336 genes are absent) without considering the context of cancer cell line, which show abnormal transcriptional regulation for a large number of genes. Although lncRNAs are often believed to be lowly expressed (12), these results indicate that one third of them have the capacity to be highly expressed under certain conditions and play important functions accordingly.

Featured lncRNA genes

LncExpDB characterizes featured lncRNA genes that are specifically expressed in a certain cell line/tissue, 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. These featured genes are most likely to exert important functions under specific conditions (see details in Materials and Methods). Based on massive RNA-seq data, a total of 25 191 featured lncRNAs are identified, including 7922 in organ development, 7498 in normal tissue/cell line, 5292 in subcellular localization, 4343 in preimplantation embryo, 2907 in cancer cell line, 1740 in circadian rhythm, 1538 in exosome, 1232 in cell differentiation and 985 in virus infection (Figure 2D).

LncExpDB provides easy access to all identified featured lncRNAs, thereby helpful for users to perform in-depth investigations on lncRNAs that are specifically expressed in some conditions or involved in specific biological contexts. Importantly, most featured lncRNA genes are present in one biological context, while most protein-coding genes are featured in two or three biological contexts (Figure 2E). Nevertheless, 6362 lncRNA genes are found to be involved in at least two biological contexts. For example, three contexts, viz., organ development, normal tissue/cell line, and cancer cell line, share 191 featured lncRNA genes (Figure 2F), presumably providing important insights into their functional roles in regulating the development of normal tissues and cancers.

LncRNAs present a wider range of subcellular localizations (localized in nuclear, or cytoplasm, or both) compared with mRNAs (generally exported to the cytoplasm), and subcellular localizations, especially the enrichment/specific localization, provides valuable insights into cellular roles of lncRNAs (13). Based on data curation and analysis, we identify lncRNAs that show specific localizations and find that majority of these lncRNAs are enriched in nuclear (including nucleus, nuclear lamina and nucleolus) (Figure 2G). Among them, two well-characterized lncRNAs, *MALAT1* and *NEAT1*, are significantly enriched in five subcellular compartments. As circadian-associated lncRNA genes show large overlap with the enriched genes of cellular compartments (Figure 2D), we further investigated the distribution of the overlapped genes. We found the overlapped lncRNA genes are specifically enriched in nuclear lamina and nucleus, whereas the overlapped mRNA genes exhibit a wider distribution (Figure 2G), indicating different roles of lncRNAs and mRNAs in regulation of circadian rhythms.

LncRNA–mRNA interactions

To facilitate in-depth investigations on molecular mechanisms of featured lncRNAs, LncExpDB predicts lncRNA–mRNA interactions by co-expression networks. Totally, LncExpDB houses 28 443 865 predicted lncRNA–mRNA interactions; the majority of these interactions (96.4%) are present in one biological context, and there are 12 interactions found in five contexts (Figure 2H).

Interestingly, *TUG1* (https://bigd.big.ac.cn/lncexpdb/gene?geneid=HSALNG0134930), one of the extensively studied lncRNAs, is found to be highly interacted in five contexts (organ development, cell differentiation, subcellular localization, exosome and circadian rhythm) and is co-expressed with seven mRNAs in all the five contexts (Figure 2I). Also, *TUG1* is frequently co-expressed with 114 mRNAs under at least four contexts (Figure 2I) (Supplementary Table S2), and these mRNAs are significantly enriched in processes of transcription regulation and cell cycle (data not shown). Additionally, among the 114 mRNAs, *DDX17* and *IFT27* are located on the same chromosome (chr22), 750 and 578 kb distant from the locus of *TUG1*, respectively.

Data organization and access

The central entities of LncExpDB are lncRNA genes, and each lncRNA gene has a corresponding page, consisting

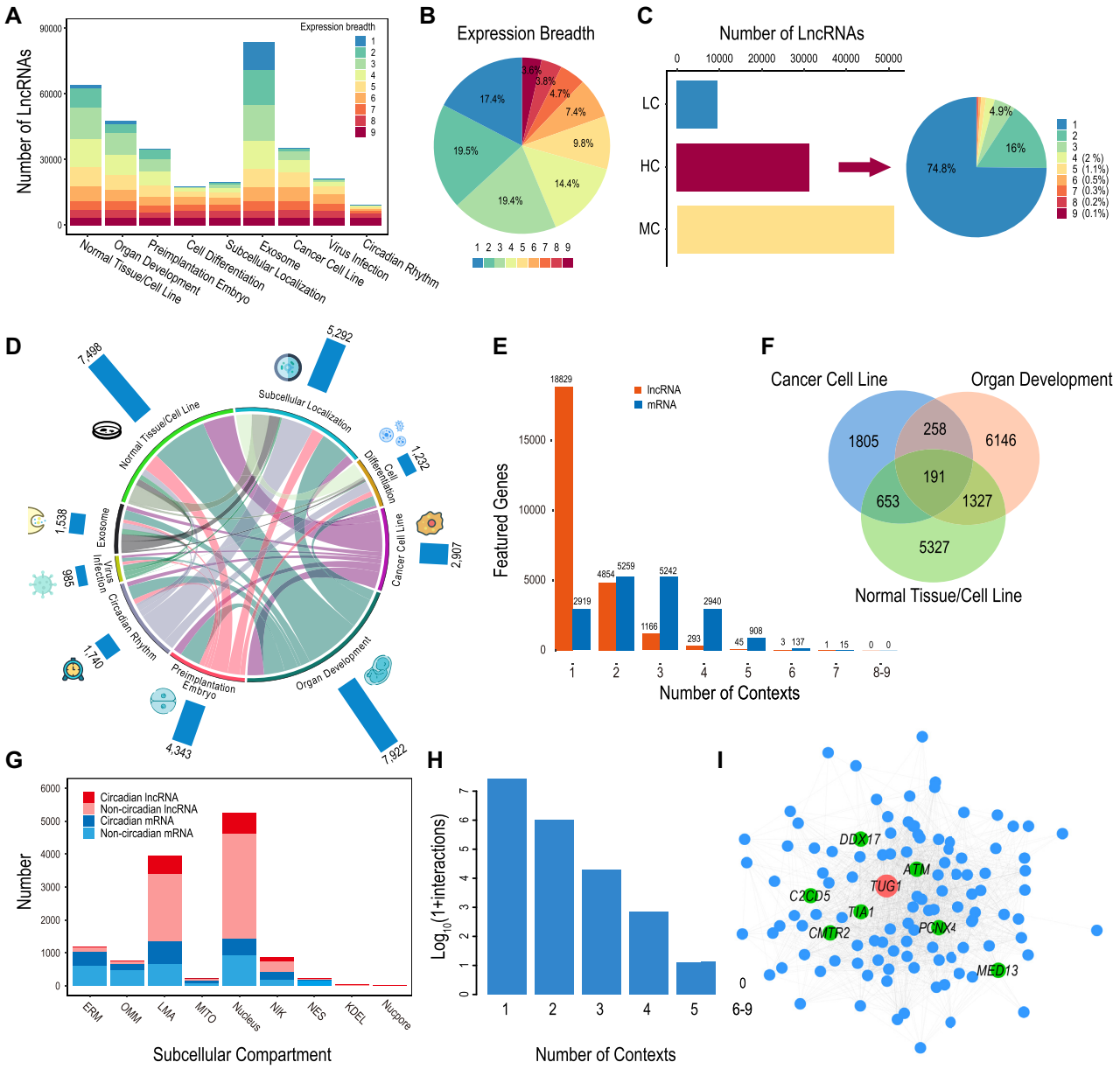

Figure 2. Expression features of lncRNA genes. (**A**) Number of expressed lncRNA genes across nine biological contexts; (**B**) Distribution and expression breadth of lncRNA genes; (**C**) Expression capacity of lncRNA genes (HC = high-capacity, LC = low-capacity, MC = medium-capacity); (**D**) Distribution of featured lncRNA genes across nine biological contexts and the correlations between contexts; (**E**) Distribution of featured lncRNA and mRNA genes; (**F**) Venn diagram of featured lncRNA genes in three biological contexts, namely, organ development, normal tissue/cell line and cancer cell line; (**G**) Comparison of subcellular compartment enrichment between lncRNAs and mRNAs, and distribution of circadian genes that are enriched in specific subcellular compartments (ERM: ER membrane, OMM: outer mito. Membrane, LMA: nuclear lamina, MITO: mitochondrial matrix, nucleus: NLS, NIK: nucleolus, NES: cytosol, KDEL: ER lumen, NucPore: nuclear pore); (**H**) Distribution of lncRNA–mRNA interactions across the nine biological contexts; (**I**) Interaction network between *TUG1* and its co-expressed mRNAs in at least four biological contexts. Green dots represent the mRNAs that are co-expressed with *TUG1* in five biological contexts.

of two main parts, viz., basic information (e.g. gene symbol, genome context, length, exon number, classification and corresponding transcripts information), and expression profiles. For each lncRNA, LncExpDB profiles its gene expression landscape across all collected conditions and visualizes its expression profiles in an interactive manner. It organizes all relevant data in a structured manner to facilitate gene-, dataset- and context-based data browse/search. It enables visualization of various expression profiles of a specific lncRNA in one page, facilitates exploration of featured genes and their related co-expression networks, and provides useful functionalities for capturing expression landscape across different biological conditions.

DISCUSSION AND FUTURE DIRECTIONS

LncExpDB is dedicated to the integration and curation of human lncRNAs, identification and characterization of featured genes with potentially important functions, and construction of co-expression interactions with protein-coding genes across various biological contexts/conditions. Based on comprehensive integration, stringent curation, and systematic analysis of massive RNA-seq datasets, LncExpDB has great potential in deepening our understanding of lncRNAs' functions and thus serves as a valuable resource for the global research community. Future directions are to include more datasets (such as TCGA (46) and GTEx (47)) and accordingly to increase more diversity of biological contexts/conditions such as immune reaction, drug therapy, and also single-cell resolution. Meanwhile, in-depth expression profile analysis will be conducted by association with other omics data. Given the complexity and diversity of lncRNA–mRNA interactions, we are considering to characterize the lncRNA-involved interactions with protein, RNA, DNA and compound through analyzing other high-throughput sequencing data (e.g. CLIP-seq, SPLASH and PARIS) and literature curation. In addition, more user-friendly interfaces and tools will be developed and enhanced in aid of functional annotation, correlation analysis and interactive visualization of various expression profiles.

DATA AVAILABILITY

LncExpDB is an expression database of human long non-coding RNAs (https://bigd.big.ac.cn/lncexpdb).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Qiheng Qian and Yang Zhang for their valuable comments and discussions in this work.

FUNDING

Strategic Priority Research Program of the Chinese Academy of Sciences [XDA19050302, XDB38030400]; National Key Research & Development Program of China [2017YFC0907502, 2015AA020108]; Youth Innovation Promotion Association of Chinese Academy of Sciences [2019104]; National Natural Science Foundation of China [31871328]; 13th Five-year Informatization Plan of Chinese Academy of Sciences [XXH13505-05]; K. C. Wong Education Foundation (to Z.Z.); International Partnership Program of the Chinese Academy of Sciences [153F11KYSB20160008]. Funding for open access charge: Strategic Priority Research Program of the Chinese Academy of Sciences [XDB38030400].
Conflicts of interest statement. None declared.

REFERENCES

1. Kopp,F. and Mendell,J.T. (2018) Functional classification and experimental dissection of long noncoding RNAs. *Cell*, **172**, 393–407.
2. Flynn,R.A. and Chang,H.Y. (2014) Long noncoding RNAs in cell-fate programming and reprogramming. *Cell Stem Cell*, **14**, 752–761.
3. Lee,S., Seo,H.H., Lee,C.Y., Lee,J., Shin,S., Kim,S.W., Lim,S. and Hwang,K.C. (2017) Human long noncoding RNA regulation of stem cell potency and differentiation. *Stem Cells Int*, **2017**, 6374504.
4. Sarropoulos,I., Marin,R., Cardoso-Moreira,M. and Kaessmann,H. (2019) Developmental dynamics of lncRNAs across mammalian organs and species. *Nature*, **571**, 510–514.
5. Fernandes,J.C.R., Acuna,S.M., Aoki,J.I., Floeter-Winter,L.M. and Muxel,S.M. (2019) Long non-coding RNAs in the regulation of gene expression: physiology and disease. *Noncoding RNA*, **5**, 17.
6. Hu,X., Liao,S., Bai,H., Gupta,S., Zhou,Y., Zhou,J., Jiao,L., Wu,L., Wang,M., Chen,X. *et al.* (2020) Long noncoding RNA and predictive model to improve diagnosis of clinically diagnosed pulmonary tuberculosis. *J. Clin. Microbiol.*, **58**, e01973-19.
7. Bao,Z., Yang,Z., Huang,Z., Zhou,Y., Cui,Q. and Dong,D. (2019) LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res.*, **47**, D1034–D1037.
8. Volders,P.J., Anckaert,J., Verheggen,K., Nuytens,J., Martens,L., Mestdagh,P. and Vandesompele,J. (2019) LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Res.*, **47**, D135–D139.
9. 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.
10. Ma,L., Li,A., Zou,D., Xu,X., Xia,L., Yu,J., Bajic,V.B. and Zhang,Z. (2015) LncRNAWiki: harnessing community knowledge in collaborative curation of human long non-coding RNAs. *Nucleic Acids Res.*, **43**, D187–D192.
11. 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.
12. 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.
13. Carlevaro-Fita,J. and Johnson,R. (2019) Global positioning system: understanding long noncoding RNAs through subcellular localization. *Mol. Cell*, **73**, 869–883.
14. Li,J., Han,L., Roebuck,P., Diao,L., Liu,L., Yuan,Y., Weinstein,J.N. and Liang,H. (2015) TANRIC: an interactive open platform to explore the function of lncRNAs in cancer. *Cancer Res.*, **75**, 3728–3737.
15. Zheng,L.L., Li,J.H., Wu,J., Sun,W.J., Liu,S., Wang,Z.L., Zhou,H., Yang,J.H. and Qu,L.H. (2016) deepBase v2.0: identification, expression, evolution and function of small RNAs, LncRNAs and circular RNAs from deep-sequencing data. *Nucleic Acids Res.*, **44**, D196–D202.
16. 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.
17. Park,C., Yu,N., Choi,I., Kim,W. and Lee,S. (2014) lncRNAtor: a comprehensive resource for functional investigation of long non-coding RNAs. *Bioinformatics*, **30**, 2480–2485.
18. 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.
19. Cardoso-Moreira,M., Halbert,J., Valloton,D., Velten,B., Chen,C., Shao,Y., Liechti,A., Ascencao,K., Rummel,C., Ovchinnikova,S. *et al.* (2019) Gene expression across mammalian organ development. *Nature*, **571**, 505–509.
20. Mas-Ponte,D., Carlevaro-Fita,J., Palumbo,E., Hermoso Pulido,T., Guigo,R. and Johnson,R. (2017) LncATLAS database for subcellular localization of long noncoding RNAs. *RNA*, **23**, 1080–1087.
21. Li,S., Li,Y., Chen,B., Zhao,J., Yu,S., Tang,Y., Zheng,Q., Li,Y., Wang,P., He,X. *et al.* (2018) exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes. *Nucleic Acids Res.*, **46**, D106–D112.
22. Papatheodorou,I., Fonseca,N.A., Keays,M., Tang,Y.A., Barrera,E., Bazant,W., Burke,M., Fullgrabe,A., Fuentes,A.M., George,N. *et al.* (2018) Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.*, **46**, D246–D251.

23. 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.
24. 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.
25. 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.
26. You,B.H., Yoon,S.H. and Nam,J.W. (2017) High-confidence coding and noncoding transcriptome maps. *Genome Res.*, **27**, 1050–1062.
27. Pertea,G. and Pertea,M. (2020) GFF Utilities: GffRead and GffCompare. *F1000Research*, **9**, 304.
28. 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.
29. 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.
30. 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.
31. Li,A., Zhang,J. and Zhou,Z. (2014) PLEK: a tool forpredicting long non-coding RNAs and messenger RNAs based on an improved k-merscheme. *BMC Bioinformatics*, **15**, 311.
32. 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.
33. 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.
34. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
35. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
36. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
37. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.
38. Dillies,M.A., Rau,A., Aubert,J., Hennequet-Antier,C., Jeanmougin,M., Servant,N., Keime,C., Marot,G., Castel,D., Estelle,J. *et al.* (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.*, **14**, 671–683.
39. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
40. Nueda,M.J., Tarazona,S. and Conesa,A. (2014) Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics*, **30**, 2598–2602.
41. Itai,Y. (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, **21**, 650–659.
42. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
43. Fazal,F.M., Han,S., Parker,K.R., Kaewsapsak,P., Xu,J., Boettiger,A.N., Chang,H.Y. and Ting,A.Y. (2019) Atlas of subcellular RNA localization revealed by APEX-Seq. *Cell*, **178**, 473–490.
44. Wu,G., Anafi,R.C., Hughes,M.E., Kornacker,K. and Hogenesch,J.B. (2016) MetaCycle: an integrated R package to evaluate periodicity3 in large scale data. *Bioinformatics*, **32**, 3351–3353.
45. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res*, **12**, 996–1006.
46. Cancer Genome Atlas Research, N., Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The cancer genome atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
47. Consortium,G.T. (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.