*Nucleic Acids Research, 2018* **1**
*doi: 10.1093/nar/gky960*



# LncBook: a curated knowledgebase of human long non-coding RNAs

**Lina Ma[1,2,*,†], Jiabao Cao[1,2,3,†], Lin Liu[1,2,3,†], Qiang Du[4], Zhao Li[1,2,3], Dong Zou[1,2], Vladimir B. Bajic[5] and Zhang Zhang[1,2,3,*]**

[1]BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, [2]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China, [3]University of Chinese Academy of Sciences, Beijing 100049, China, [4]Anhui University of Technology, Maanshan 243032, China and [5]King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering Division (CEMSE), Thuwal 23955-6900, Kingdom of Saudi Arabia

Received August 15, 2018; Revised September 27, 2018; Editorial Decision October 04, 2018; Accepted October 10, 2018

## ABSTRACT

**Long non-coding RNAs (lncRNAs) have significant functions in a wide range of important biological processes. Although the number of known human lncRNAs has dramatically increased, they are poorly annotated, posing great challenges for better understanding their functional significance and elucidating their complex functioning molecular mechanisms. Here, we present LncBook (http://bigd.big.ac.cn/lncbook), a curated knowledgebase of human lncRNAs that features a comprehensive collection of human lncRNAs and systematic curation of lncRNAs by multi-omics data integration, functional annotation and disease association. In the present version, LncBook houses a large number of 270 044 lncRNAs and includes 1867 featured lncRNAs with 3762 lncRNA–function associations. It also integrates an abundance of multi-omics data from expression, methylation, genome variation and lncRNA–miRNA interaction. Also, LncBook incorporates 3772 experimentally validated lncRNA-disease associations and further identifies a total of 97 998 lncRNAs that are putatively disease-associated. Collectively, LncBook is dedicated to the integration and curation of human lncRNAs as well as their associated data and thus bears great promise to serve as a valuable knowledgebase for worldwide research communities.**

## INTRODUCTION

Long non-coding RNAs (lncRNA) have a variety of functions in many important biological processes (1–5) and are closely associated with various diseases (6–9). In recent years, rapid advances of next-generation sequencing technologies have triggered an explosion of newly discovered lncRNAs (especially in human) (9–12), primarily due to their highly tissue/cell-specific (5,13–15) and lineage/species-specific (14,16,17) nature. Accordingly, multiple databases have been constructed to archive lncRNA sequences and annotations (10,12,14,18,19), collect experimentally validated lncRNAs (2,3,9,20), curate lncRNAs-disease associations (6,7,9), and annotate miRNA–lncRNA interactions (21–23). Despite this, lncRNAs are still poorly annotated (24), posing great challenges for better understanding their functional significance and dissecting their complex functioning molecular mechanisms.

To harness collective intelligence for gathering and annotating human lncRNAs, we constructed LncRNAWiki (9) in 2015, a wiki-based platform for community curation of human lncRNAs. LncRNAWiki has frequently been updated by adding more experimentally validated lncRNAs, incorporating small peptides encoded in lncRNAs and associating lncRNAs with diseases (25,26). However, LncRNAWiki, built based on MediaWiki, has significant limitations on managing structured data and providing customized functionalities; functional annotations and sequence data are stored as unstructured text in MediaWiki, which makes it difficult to retrieve and show data items of interest. It would be desirable to organize large-scale annotations in a structured manner and to provide customized web functionalities with more friendly interfaces. More importantly, it is highly beneficial to integrate multi-omics data with the aim to significantly enrich and improve lncRNAs' annotations to support function inference.

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

Here we develop an expert-curation-based resource, LncBook (http://bigd.big.ac.cn/lncbook), as a complement to community-curation-based LncRNAWiki. LncBook features a comprehensive collection of human lncRNAs and systematic curation of lncRNAs by multi-omics data integration, functional annotation and disease association (Table 1). It houses a larger number of human lncRNAs that are not only derived from existing databases but also novel RNA assemblies based on RNA-seq data analysis. It includes community-contributed annotations from LncRNAWiki and expert-curated annotations curated from published literature, respectively. Particularly, it integrates a variety of multi-omics data, including expression, methylation, variation, and interaction, conducts functional annotation and incorporates a collection of lncRNA-disease associations. Equally important, LncBook organizes all relevant data in a structured manner, facilitating data browse/search with more enhanced efficiency and provides several useful tools for online analysis.

## MATERIALS AND METHODS

### Data collection

The human lncRNAs in LncBook were obtained not only from existing databases and published literature but also from novel RNA assemblies based on RNA-seq data analysis (Figure 1). Namely, we collected lncRNAs from several well-known lncRNA databases, including GENCODE v27 (14), NONCODE v5.0 (12), LNCipedia v4.1 (10) and Mi-Transcriptome beta (11). To obtain high-confidence lncRNAs, a set of strict criteria was adopted considering redundancy, background noise, mapping error, incomplete transcripts, length and coding potential. We used Cuffcompare (27) to compare different datasets to remove redundant, questionable or incomplete transcripts: (i) redundant transcripts were identified using Cuffcompare (27) with the comparison code ' = ' (which means complete match of intron chain), and then representative lncRNAs were selected according to their annotation quality; (ii) questionable transcripts were detected using Cuffcompare (27) with the comparison codes 'e', 'p' and 's'. Single-exon transcripts that are part of multi-exon transcripts and located in their exon regions, were regarded as incomplete lncRNAs. Also, transcripts with very short exons (<15 nt) at the 5′ and 3′ ends were considered as incomplete lncRNAs. In addition, transcripts shorter than 200 nt were excluded. Moreover, three algorithms, namely, LGC (an in-house tool publicly available at http://bigd.big.ac.cn/biocode/tools/BT000004), CPAT (28) and PLEK (29), were used for coding potential estimation, and we only retained transcripts that were identified as lncRNAs by all the three algorithms. As a consequence, an integrated, non-redundant and high-quality dataset containing 247 246 lncRNAs was obtained.

To identify novel lncRNAs, we downloaded 122 RNA-seq datasets from HPA (Human Protein Atlas) (30). FastaQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and Trimmomatic (31) were used for quality control. Hisat2 (32) was used for reads mapping, while StringTie (32) was used for assembling and merging transcripts. We compared the assembled transcripts with existing lncRNAs using Cuffcompare (27). As a result, a total of 21 815 novel lncRNAs were identified. At last, we integrated a large collection of lncRNAs from existing databases, novel RNA assemblies based on RNA-seq data analysis, and literature reported lncRNAs that were obtained from LncRNAWiki (9). After removing lncRNAs that were not traceable back to their genomic locations, we finally obtained a total of 270 044 non-redundant lncRNA transcripts, belonging to 140 362 gene loci. All these lncRNAs are publicly available at ftp://download.big.ac.cn/lncbook.

### Data integration and annotation

To profile expression levels for lncRNAs, two RNA-seq datasets were used: HPA (Human Protein Atlas, 32 normal human tissues are covered) (30) and GTEx (Genotype-Tissue Expression, 53 normal human tissues are covered) (33). We filtered out lncRNAs whose highest expression values are lower than 0.5 TPM/FPKM. Then, $\tau$-value (34) and cv (coefficient of variance) value were used to determine HK (housekeeping) lncRNAs ($\tau$-value $\leq 0.5$ and cv $\leq 0.5$) and TS (tissue-specific) lncRNAs ($\tau$-value $\geq 0.95$). To annotate methylation information of lncRNAs, bisulfite-seq data from TCGA (The Cancer Genome Atlas) (https://portal.gdc.cancer.gov) and ENCODE (The ENCyclopedia of DNA Elements) (https://www.encodeproject.org) were downloaded, covering nine cancers with both normal and cancer samples. We defined regions from −1500 bp relative to the transcription start sites as promoters, and calculated the methylation levels of both promoter and body regions of lncRNAs. In addition, we mapped the SNP sites in dbSNP (35) to the lncRNA loci, and annotated MAF (Minor Allele Frequency) values based on 1000 Genomes Project (36), pathogenic information in ClinVar (version 2017.9.05) (37) and COSMIC (version 85) (38) using ANNOVAR (39). TargetScan (40) and miRanda (41) were used to predict lncRNA–miRNA interactions and experimentally validated interactions were obtained from starBase v2.0 (21).

To provide high-quality annotations for experimentally validated lncRNAs, we systematically curated 1867 lncRNAs (that were sourced from LncRNAWiki (9)) with function annotations reported in 2501 publications and with controlled vocabularies describing their functioning mechanisms and biological processes they affect.

The associations between lncRNA and disease were derived from LncRNADisease (6) and LncRNAWiki (9), which have been extracted from published literature with experimental evidence. Each lncRNA-disease association was annotated with disease name, MeSH Ontology (Medical Subject Headings 2018 name), dysfunction type, detailed description and publication. On the other hand, we predicted disease-associated lncRNAs according to the evidence from methylation, genome variation and lncRNA–miRNA interaction: (i) Methylation: In each cancer, lncRNA whose promoter region methylation level shows increase (decrease) in 80% cancer samples relative to normal samples is considered to be hypermethylated (hypomethylated). Thus, we considered one lncRNA to be cancer-associated if it is consistently hypermethylated or hypomethylated in at least eight cancers; (ii) Genome Varia-

**Table 1.** Key differences between LncRNAWiki 2015 and LncBook

| Data Item | LncRNAWiki 2015 | LncBook |
|---|---|---|
| LncRNA transcripts | 105 255 | **270 044** |
| Featured lncRNAs | 86 | **1867** |
| LncRNA–function associations | NA | **3762** |
| LncRNA-disease associations | NA | **3772** |
| Predicted disease-associated lncRNAs | NA | **97 998** |
| HK and TS lncRNAs | NA | **819 HK**<br>**49 115 TS** |
| Methylation profiles | NA | **Profiles in nine cancers** |
| Genome variations | NA | **92 725 757 SNPs** |
| LncRNA-miRNA interactions | NA | **129 690 817 predicted lncRNA–miRNA interactions** |
| Tools | BLAST | **BLAST, classification, coding potential prediction, ID conversion** |

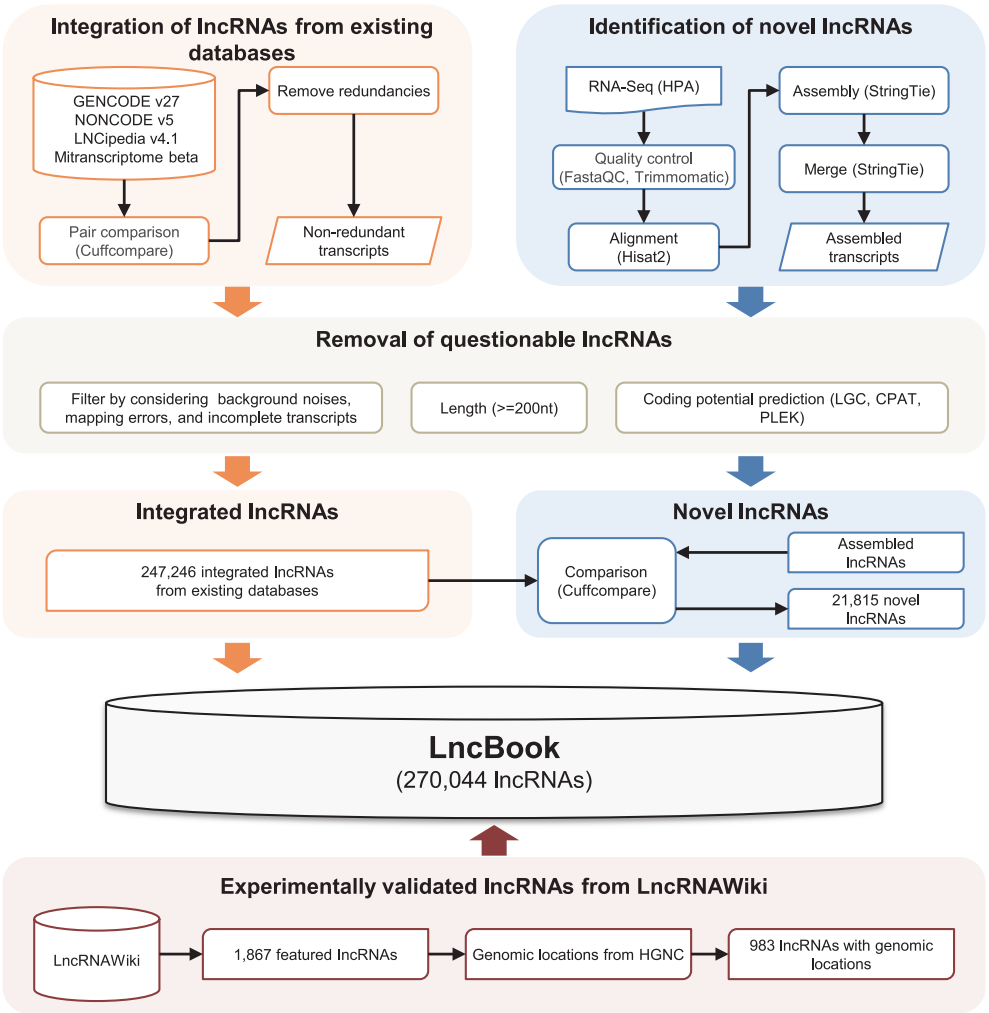


**Figure 1.** Workflow diagram of lncRNA curation and integration. The human lncRNAs in LncBook are compiled from (i) existing databases, (ii) novel RNA assemblies based on RNA-seq data analysis and (iii) published literature.

tion: any lncRNA overlapping disease-associated SNPs of COSMIC (OCCURRENCE $\geq 3$) or ClinVar in its genomic location we considered to be disease-associated; (iii) Interaction: any lncRNA interacting with at least 11 disease-associated miRNAs (associated with at least five diseases according to the Human microRNA Disease Database HMDD (42)) we considered to be disease-associated.

## Implementation

We developed LncBook using String Boot as back-end web framework and MySQL (http://www.mysql.org) as database engine. Web interfaces were developed by JSP (Java Server Pages) and AJAX (Asynchronous JavaScript and XML). Bootstrap (https://getbootstrap.com) was adopted as a front-end framework, which pro-

vides a series of templates for designing web pages with consistent interface components. Also, data visualization was powered by Highcharts (a charting library written in pure JavaScript), offering an easy way of adding interactive charts to any web site or application.

## DATABASE CONTENTS AND FEATURES

Compared with the existing lncRNA databases, LncBook features a comprehensive collection of human lncRNAs and systematic curation of lncRNAs' annotation by multi-omics data integration, function annotation and disease association (Table 1). In the current version, LncBook houses a total of 270,044 lncRNA transcripts, contains 1867 experimentally validated lncRNAs manually curated based on published literature, and annotates all lncRNAs by integration of large-scale multi-omics data including tissue expression profiles, cancer-associated methylation levels, genome variations and lncRNA–miRNA interactions. These 1867 featured lncRNAs that have been documented in published literature are systematically curated and annotated with functioning mechanisms and biological processes, resulting in 3762 lncRNA–function associations. LncBook also includes a total of 3772 experimentally validated lncRNA-disease associations and identifies 97 998 lncRNAs that are putatively associated with diseases. Also, a series of useful tools, such as coding potential prediction, sequence search, etc., are deployed in LncBook.

### Comprehensive collection of human lncRNAs

LncBook accommodates a comprehensive collection of 270 044 human lncRNAs (see details in 'Materials and Methods' section), including 247 246 lncRNAs obtained from existing databases, 1867 from LncRNAWiki and 21 815 novel lncRNAs identified based on RNA-seq data analysis, which together belong to 140 362 gene loci. LncBook manages human lncRNAs based on transcripts, where a unique accession number prefixed with HSALNT is assigned to each lncRNA transcript entity. Likewise, the lncRNA gene has an accession number prefixed with HSALNG. In LncBook, each transcript corresponds to a specific web page containing basic information (symbol, genomic context, length, exon number, GC content, classification, sequence, longest ORF length, coding potential), multi-omics data (expression, methylation, genome variation, lncRNA–miRNA interaction), function annotations and disease associations (Figure 2).

### Multi-omics data integration

LncBook integrates a variety of multi-omics data, enriching lncRNAs with abundant annotations in expression, methylation, genome variation and interaction with miRNAs (see details in 'Materials and Methods' section). For any given lncRNA, LncBook profiles its expression levels across all collected tissues and visualizes its expression profile in a bar chart, greatly facilitating users to explore functional significance. Based on these expression profiles across different tissues, LncBook further identifies a total of 819 HK lncRNAs, which are consistently expressed in almost all tissues. Similarly, it also obtains 49 115 TS lncRNAs, which are expressed specifically in one or few tissues. All HK and TS lncRNAs are publicly available at http://bigd.big.ac.cn/lncbook/expression. Also, for each lncRNA, LncBook provides methylation levels of both promoter and body regions in normal and cancer samples across nine cancers, which are summarized in a table and visualized in a dot plot. According to the methylation analysis results, only 583 lncRNAs are always hypomethylated in cancers, contrasting to 27 723 lncRNAs that are always hypermethylated. LncBook also collects 92 725 757 SNPs from dbSNP (35) residing in 197 799 lncRNA transcripts. Among all these SNPs, there are 7571 pathogenic SNPs from ClinVar (37) overlapping 2280 lncRNA transcripts and 79 012 pathogenic SNPs from COSMIC (38) (OCCURRENCE $\geq$ 3) overlapping 26 008 lncRNA transcripts. Also, LncBook includes 145 lncRNA–miRNA interactions supported by experimental evidence from starBase (21), as well as 129 690 817 interactions predicted by TargetScan (40) and miRanda (41).

### Function annotation

Although a large number of lncRNAs have been identified in human, only a small fraction of them have experimental evidence with supported publications. According to the current collection of LncBook, there are only 1867 out of all 270 044 lncRNAs that have been documented experimental validation. Based on manual curation of 2632 publications, LncBook provides comprehensive function annotations for these 1867 featured lncRNAs; 1653 lncRNAs have function annotation, while 1502 lncRNAs are linked to different diseases, leading to 3762 lncRNA–function associations. Specifically, each lncRNA–function association in LncBook is described using controlled vocabularies in light of functioning mechanism and biological process in which they are involved. Regarding functioning mechanism, LncBook adopts six controlled terms with each having different number of associations: transcriptional regulation (397 associations), ceRNA (182 associations), splicing regulation (19 associations), translational control (17 associations), protein localization (4 associations) and RNAi (3 associations). For biological process, LncBook adopts two terms, namely, pathogenic process and developmental process; function annotation of featured lncRNAs shows that most of them are involved in cancer and other diseases (3598 associations), compared to developmental process (53 associations).

### LncRNA-disease association

Considering that most of the functionally studied lncRNAs are closely associated with human diseases, LncBook integrates 3772 lncRNA-disease associations, derived not only from LncRNADisease (6) and LncRNAWiki (9) but also curated based on 2337 publications. LncBook describes each association with disease name, dysfunction type, detailed description, MeSH disease ontology and publication. According to the current information contained in LncBook, all lncRNAs in LncBook are associated with 462 diseases and 28 MeSH disease terms. Among all the terms, 'Neoplasms' has the largest number of associations (2888 associations), followed by the term 'Digestive System Dis-


**Figure 2.** Screenshots of web pages for a lncRNA transcript. For example, *HOTAIR* is extensively annotated with an abundance of multi-omics data including (**A**) expression, (**B**) methylation, (**C**) variation, (**D**) interaction and the systematic curations of (**E**) function and (**F**) disease based on published literature.

eases' that has 888 associations. LncBook-contained information also reveals that among all the disease-associated lncRNAs, *HOTAIR*, *MALAT1*, *H19*, *MEG3*, *CDKN2B-AS1*, *PVT1*, *NEAT1* and *GAS5* are extensively studied and each of them is associated with at least 30 different diseases.

Additionally, based on an abundance of methylation, genome variation and lncRNA–miRNA interaction, LncBook predicts a total of 97 998 lncRNAs that are potentially associated with diseases (see details in 'Materials and Methods' section). Briefly speaking, one lncRNA is putatively believed to be disease-associated only if any evidence for that can be obtained from methylation, genome variation and/or lncRNA–miRNA interaction. For a specific lncRNA under investigation, supporting evidence can be that, for example, its methylation change relates to disease, it overlaps pathogenic variations, or it frequently interacts with disease-associated miRNAs. As a consequence, LncBook contains a collection of 97 998 disease-associated lncRNAs, where 607 are supported by three sources of evidence, namely, methylation, genome variation and lncRNA–miRNA interaction, 13 257 are supported by two of them, and 84 134 are supported by only one of them. All these disease-associated lncRNAs can be found at http://bigd.big.ac.cn/lncbook/disease.

## DISCUSSION AND FUTURE DIRECTIONS

LncBook is dedicated to the integration and curation of human lncRNAs as well as their associated data. In harmony with LncRNAWiki that is a community-curated resource, LncBook serves as an expert-curated knowledgebase that integrates a comprehensive collection of human lncRNAs and contains multi-omics data, function annotations and disease associations. The current implementation of LncBook houses a large number of 270 044 lncRNAs and includes 1867 featured lncRNAs with 3762 lncRNA–function associations. It also integrates an abundance of multi-omics data from expression, methylation, genome variation and lncRNA–miRNA interaction. Also, LncBook includes 3772 experimentally validated lncRNA-disease associations and identifies 97 998 lncRNAs that are putatively disease-associated. However, of note, this does not mean that these disease-associated lncRNAs play causative roles in diseases (24). Taken together, LncBook is a curated knowledgebase of human lncRNAs and has the potential to serve as a valuable resource for worldwide research communities. Future developments of LncBook include regular integration of newly discovered lncRNAs, incorporation of high-quality annotations through literature curation and identification of differentially expressed lncR-

NAs in normal and disease samples. We also plan to integrate full-length lncRNAs from additional databases such as FANTOM CAT (43) and BIGTranscriptome (44). In addition, more user-friendly tools will be developed in aid of functional annotation and interactive visualization of various omics data. We also look forward to comments and suggestions from researchers worldwide, aiming to build LncBook into an encyclopedia of human lncRNAs.

## ACKNOWLEDGEMENTS

We thank Mengwei Li, Jian Sang, Yuansheng Zhang and Yanqing Wang for valuable comments and assistances in this work.

## FUNDING

Strategic Priority Research Program of the Chinese Academy of Sciences [XDA19050302, XDB13040500]; National Key Research and Development Program of China [2017YFC0907502, 2015AA020108]; The 13th Five-year Informatization Plan of Chinese Academy of Sciences [XXH13505-05]; International Partnership Program of the Chinese Academy of Sciences [153F11KYSB20160008]; National Natural Science Foundation of China [31200978]; King Abdullah University of Science and Technology (KAUST) [BAS/1/1606-01-01]. Funding for open access charge: Strategic Priority Research Program of the Chinese Academy of Sciences [XDB13040500].
*Conflict of interest statement.* None declared.

## REFERENCES

1. Ma,L., Bajic,V.B. and Zhang,Z. (2013) On the classification of long non-coding RNAs. *RNA Biol.*, **10**, 925–933.
2. 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.
3. Salhi,A., Essack,M., Alam,T., Bajic,V.P., Ma,L., Radovanovic,A., Marchand,B., Schmeier,S., Zhang,Z. and Bajic,V.B. (2017) DES-ncRNA: A knowledgebase for exploring information about human micro and long noncoding RNAs based on literature-mining. *RNA Biol.*, **14**, 963–971.
4. Dykes,I.M. and Emanueli,C. (2017) Transcriptional and Post-transcriptional gene regulation by long Non-coding RNA. *Genomics Proteomics Bioinformatics*, **15**, 177–186.
5. 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**, aah7111.
6. Chen,G., Wang,Z., Wang,D., Qiu,C., Liu,M., Chen,X., Zhang,Q., Yan,G. and Cui,Q. (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.*, **41**, D983–D986.
7. Ning,S., Zhang,J., Wang,P., Zhi,H., Wang,J., Liu,Y., Gao,Y., Guo,M., Yue,M., Wang,L. *et al.* (2016) Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res.*, **44**, D980–D985.
8. Chen,X., Yan,C.C., Zhang,X. and You,Z.H. (2017) Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.*, **18**, 558–576.
9. Ma,L.N., Li,A., Zou,D., Xu,X.J., 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.
10. Volders,P.J., Verheggen,K., Menschaert,G., Vandepoele,K., Martens,L., Vandesompele,J. and Mestdagh,P. (2015) An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res.*, **43**, D174–D180.
11. 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.
12. 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.
13. Cabili,M.N., Trapnell,C., Goff,L., Koziol,M., Tazon-Vega,B., Regev,A. and Rinn,J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.
14. 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.
15. Alam,T., Uludag,M., Essack,M., Salhi,A., Ashoor,H., Hanks,J.B., Kapfer,C., Mineta,K., Gojobori,T. and Bajic,V.B. (2017) FARNA: knowledgebase of inferred functions of non-coding RNA transcripts. *Nucleic Acids Res.*, **45**, 2838–2848.
16. Paralkar,V.R., Mishra,T., Luan,J., Yao,Y., Kossenkov,A.V., Anderson,S.M., Dunagin,M., Pimkin,M., Gore,M., Sun,D. *et al.* (2014) Lineage and species-specific long noncoding RNAs during erythro-megakaryocytic development. *Blood*, **123**, 1927–1937.
17. Hezroni,H., Koppstein,D., Schwartz,M.G., Avrutin,A., Bartel,D.P. and Ulitsky,I. (2015) Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.*, **11**, 1110–1122.
18. Szczesniak,M.W., Rosikiewicz,W. and Makalowska,I. (2016) CANTATAdb: a collection of plant long Non-Coding RNAs. *Plant. Cell Physiol.*, **57**, e8.
19. The RNAcentral Consortium (2017) RNAcentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Res.*, **45**, D128–D134.
20. Zhou,B., Zhao,H., Yu,J., Guo,C., Dou,X., Song,F., Hu,G., Cao,Z., Qu,Y., Yang,Y. *et al.* (2018) EVLncRNAs: a manually curated database for long non-coding RNAs validated by low-throughput experiments. *Nucleic Acids Res.*, **46**, D100–D105.
21. Li,J.H., Liu,S., Zhou,H., Qu,L.H. and Yang,J.H. (2014) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.*, **42**, D92–D97.
22. Paraskevopoulou,M.D., Vlachos,I.S., Karagkouni,D., Georgakilas,G., Kanellos,I., Vergoulis,T., Zagganas,K., Tsanakas,P., Floros,E., Dalamagas,T. *et al.* (2016) DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. *Nucleic Acids Res.*, **44**, D231–D238.
23. Miao,Y.R., Liu,W., Zhang,Q. and Guo,A.Y. (2018) lncRNASNP2: an updated database of functional SNPs and mutations in human and mouse lncRNAs. *Nucleic Acids Res.*, **46**, D276–D280.
24. Uszczynska-Ratajczak,B., Lagarde,J., Frankish,A., Guigo,R. and Johnson,R. (2018) Towards a complete map of the human long non-coding RNA transcriptome. *Nat. Rev. Genet.*, **19**, 535–548.
25. BIG Data Center Members. (2017) The BIG Data Center: from deposition to integration to translation. *Nucleic Acids Res.*, **45**, D18–D24.
26. BIG,Data Center Members. (2018) Database resources of the BIG data center in 2018. *Nucleic Acids Res.*, **46**, D14–D20.
27. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.
28. 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.
29. 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 Bioinformatics*, **15**, 311.

30. Uhlen,M., Fagerberg,L., Hallstrom,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,A., Kampf,C., Sjostedt,E., Asplund,A. *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.
31. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
32. Pertea,M., Kim,D., Pertea,G.M., Leek,J.T. and Salzberg,S.L. (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.*, **11**, 1650–1667.
33. The GTEx Consortium. (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.
34. Yanai,I., Benjamin,H., Shmoish,M., Chalifa-Caspi,V., Shklar,M., Ophir,R., Bar-Even,A., Horn-Saban,S., Safran,M., Domany,E. *et al.* (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, **21**, 650–659.
35. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
36. Abecasis,G.R., Altshuler,D., Auton,A., Brooks,L.D., Durbin,R.M., Gibbs,R.A., Hurles,M.E. and McVean,G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.
37. Landrum,M.J., Lee,J.M., Benson,M., Brown,G., Chao,C., Chitipiralla,S., Gu,B., Hart,J., Hoffman,D., Hoover,J. *et al.* (2016) ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.*, **44**, D862–D868.
38. Forbes,S.A., Beare,D., Boutselakis,H., Bamford,S., Bindal,N., Tate,J., Cole,C.G., Ward,S., Dawson,E., Ponting,L. *et al.* (2017) COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.*, **45**, D777–D783.
39. Wang,K., Li,M. and Hakonarson,H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.
40. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
41. Betel,D., Wilson,M., Gabow,A., Marks,D.S. and Sander,C. (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res.*, **36**, D149–D153.
42. Li,Y., Qiu,C., Tu,J., Geng,B., Yang,J., Jiang,T. and Cui,Q. (2014) HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.*, **42**, D1070–D1074.
43. 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.
44. You,B.H., Yoon,S.H. and Nam,J.W. (2017) High-confidence coding and noncoding transcriptome maps. *Genome Res.*, **27**, 1050–1062.