D308–D314 *Nucleic Acids Research, 2018, Vol. 46, Database issue* *Published online 11 November 2017*
doi: 10.1093/nar/gkx1107

NONCODEV5: a comprehensive annotation database for long non-coding RNAs

ShuangSang Fang[1,2,†], LiLi Zhang[2,3,†], JinCheng Guo[1,4], YiWei Niu[2,3], Yang Wu[1], Hui Li[1], LianHe Zhao[1,2], XiYuan Li[1], XueYi Teng[2,3], XianHui Sun[2,3], Liang Sun[1], Michael Q. Zhang[5], RunSheng Chen[3,*] and Yi Zhao[1,6,*]

[1]Key Laboratory of Intelligent Information Processing, Advanced Computer Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, [2]University of Chinese Academy of Sciences, Beijing 100049, China, [3]CAS Key Laboratory of RNA Biology, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China, [4]Department of Biochemistry and Molecular Biology, Shantou University Medical College, Shantou 515041, China, [5]School of Medicine, MOE Key Laboratory of Bioinformatics and Bioinformatics Division, Center for Synthetic and System Biology, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China and [6]Chinese Academy of Sciences, LuoYang Branch of Institute of Computing Technology, Luoyang, China

Received September 14, 2017; Revised October 14, 2017; Editorial Decision October 22, 2017; Accepted October 25, 2017

ABSTRACT

NONCODE (http://www.bioinfo.org/noncode/) is a systematic database that is dedicated to presenting the most complete collection and annotation of non-coding RNAs (ncRNAs), especially long non-coding RNAs (lncRNAs). Since NONCODE 2016 was released two years ago, the amount of novel identified ncRNAs has been enlarged by the reduced cost of next-generation sequencing, which has produced an explosion of newly identified data. The third-generation sequencing revolution has also offered longer and more accurate annotations. Moreover, accumulating evidence confirmed by biological experiments has provided more comprehensive knowledge of lncRNA functions. The ncRNA data set was expanded by collecting newly identified ncRNAs from literature published over the past two years and integration of the latest versions of RefSeq and Ensembl. Additionally, pig was included in the database for the first time, bringing the total number of species to 17. The number of lncRNAs in NONCODEv5 increased from 527 336 to 548 640. NONCODEv5 also introduced three important new features: (i) human lncRNA–disease relationships and single nucleotide polymorphism-lncRNA–disease relationships were constructed; (ii) human exosome lncRNA expression profiles were displayed; (iii) the RNA secondary structures of NONCODE human transcripts were predicted. NONCODEv5 is also accessible through http://www.noncode.org/.

INTRODUCTION

Whole transcriptome studies have revealed that protein-coding regions account for only 2% of the genome (1–3), although most of the human genome can be transcribed. The vast majority of transcribed sequences do not encode proteins and are called non-coding RNAs (ncRNAs). Long non-coding RNAs (lncRNAs) are ncRNAs that are >200 nt in length (4,5). Accumulating evidence has shown that lncRNAs play key roles in various biological processes, such as the circuitry controlling pluripotency and differentiation, imprinting control, immune responses, disease aetiology (6–8) and chromosome dynamics (9). The biological functions and mechanisms of lncRNAs have not been fully mined to date due to a systematic lack of comprehensive collation and summary. Consequently, we updated the NONCODE database to version 5.0 to remain up-to-date with the latest lncRNA discoveries. The data sources of NONCODEv5 include literature published since the last NONCODE update and the latest versions of several public databases (Ensembl (10), RefSeq (11), lncRNAdb (12) and LNCipedia (13)). The number of transcripts in NONCODEv5 reached 548 640 after removal of false and redundant lncRNAs.

Of the articles retrieved from PubMed concerning lncRNAs, we found that the vast majority studied lncRNA functions, especially the relationships between lncRNAs and diseases. The last version of NONCODE collected relationships between lncRNAs and various diseases using literature mining, differential lncRNA analysis utilizing public

[*]To whom correspondence should be addressed. Tel: +86 10 6260 0822; Fax: +86 10 6260 1356; Email: biozy@ict.ac.cn
Correspondence may also be addressed to RunSheng Chen. Tel: +86 10 6488 8543; Fax: +86 10 6487 7837; Email: rschen@ibp.ac.cn
[†]These authors contributed equally to the paper as first authors.

RNA-seq data and microarray data and mutation analyses from public genome-wide association study (GWAS) data. In NONCODEv5, relationships between lncRNAs and diseases were obtained from four lncRNA disease databases, and the relationships between lncRNAs and SNPs were obtained from LincSNP 2.0, which integrated eight database resources in the update. A total of 32 226 disease-related and 724 579 SNP-related records on human genes were included in the NONCODEv5 database.

Apart from the lncRNA–disease information discovered by researchers, the majority of associations remain unknown and need to be developed. Several novel research fields have provided valuable lncRNA–disease association research directions. A recent study showed that exosomes derived from tumour and normal cells greatly contributed to tumourigenesis, apoptosis, and chemotherapeutic resistance (14). To provide statistical support for lncRNA–disease association studies, NONCODEv5 collected six human exosome datasets from GEO (15), including six tumour cell lines and four tissues, and depicted the expression profiles of genes and transcripts included in NONCODEv5.

Recent studies have shown that lncRNA secondary structures also serve as regulatory factors in biological processes and influence practically every step of the RNA life cycle, including RNA transcription, splicing, cellular localization, translation, and turnover (16). The RNA secondary structure is important for RNA functions and regulation, and there is growing interest in determining the RNA structures of many transcripts (17). To meet the requirements, we predicted the RNA secondary structures of the human transcripts collected in NONCODEv5 for the first time.

The updated NONCODEv5 is committed to building a one-stop knowledge gateway for lncRNAs in the areas of data collection, expression profile calculation in different tissues, lncRNA gene-disease relationship construction and RNA structure prediction.

DATA COLLECTION AND PROCESSING

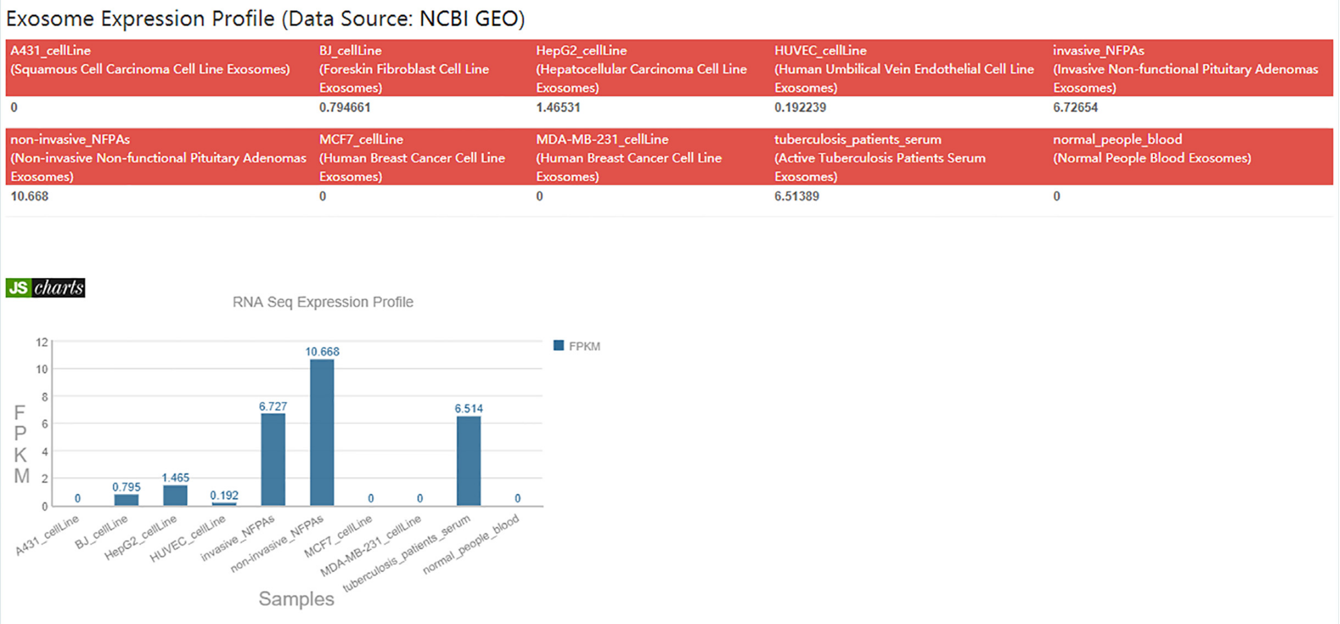
The data sources for NONCODEv5 include the previous versions of NONCODE (18–20), public literature and lncRNA databases. To obtain lncRNA information from published articles, keywords including ‘ncrna’, ‘noncoding’, ‘non-coding’, ‘no code’, ‘non-code’, ‘lncrna’ and ‘lincrna’ were searched on NCBI PubMed to identify studies published between 1 July 2015 and 22 June 2017. A total of 10 454 lncRNA-related articles were retrieved and excavated using artificial information. We retrieved the newly identified lncRNAs and their annotations from the supplementary material or websites of these articles. After screening papers manually, we focused on 70 research articles which identified novel ncRNAs in species included in the NONCODE species list. For articles which did not provide lists of ncRNAs, we asked the authors for detail information. Additionally, we collated the newest data from Ensembl, RefSeq, lncRNAdb, and LNCipedia and the old versions of the NONCODE data. In addition to the four databases, we also integrated The Arabidopsis Information Resource (TAIR) (21) and FlyBase (22) databases into NONCODEv5**.** All of the collected data were processed through a standard pipeline which is concordant with the previous versions of NONCODE for each species.

STATISTICAL ANALYSIS OF NONCODE

NONCODEv5 contains 548 640 lncRNA transcripts from 17 species (human, mouse, cow, rat, chimpanzee, gorilla, orangutan, rhesus macaque, opossum, platypus, chicken, zebrafish, fruit fly, *Caenorhabditis elegans*, yeast, arabidopsis and pig). According to the definition of lncRNA gene (19), NONCODEv5 collected a total of 354 855 genes. A total of 96 308 and 87 774 genes were generated from 172 216 and 131 697 human and mouse transcripts, respectively. NONCODEv5 annotated the expression profiles from all human and mouse transcripts and genes, and some of these genes were annotated with predicted functions. Expression profiles in exosomes calculated for human species and the conservation of information between human and other species were provided. Moreover, the RNA secondary structures of the human transcripts were predicted in this version.

LncRNAs and diseases

Most genomic transcripts are noncoding, whereas only a small fraction of the transcripts encode proteins (2,23). Among these noncoding transcripts, lncRNAs are closely linked to diseases through several aspects, including the regulation of histone modifications, transcription, DNA methylation and chromatin remodeling and post-transcriptional regulation (24,25). Conformational changes of the RNA structure, expression levels and lncRNA-binding proteins can all contribute to dysfunction, including cancer and neurodegenerative disorders (26). For example, the lncRNA SNHG1 functions both in *cis* and *trans* mechanisms to contribute to tumour cell growth by regulating SLC3A2 and FUBP1 expression (27). The lncRNA MT1JP interacts with TIAR to regulate the p53 pathway, resulting in tumour suppression (28). Spinocerebellar ataxia type 8 (SCA8), which is a type of neurodegenerative disorder, has been shown to be related to expansion repeats of the lncRNA gene ATXN8OS (29). To provide a comprehensive resource for lncRNA–disease relationships, we collected information from four databases (LncRNADisease (30), Lnc2Cancer (31), MNDR (32) and LncRNAWiki (33)). LncRNADisease includes lncRNA–disease associations supported by biological experiments and predicted by computational methods. Lnc2Cancer is a manually curated database of experimentally supported lncRNAs associated with various human cancers. MNDR v2.0 integrates experimentally and computationally predicted diverse ncRNA–disease relationships from the literature and other database resource collections, such as NSDNA (34) and LincSNP (35). LncRNAWiki is a lncRNA knowledge base that contains lncRNA–disease association information. We summarized the collected lncRNA–disease relationships from these databases. However, lncRNA–disease associations predicted by computational methods were not included in NONCODEv5 because of uncertainty credibility, and only experimentally supported relationships were integrated in our final results. In summary, we obtained 32 226 records of lncRNA and disease-related information.



Exosome Expression Profile (Data Source: NCBI GEO)

A431_cellLine (Squamous Cell Carcinoma Cell Line Exosomes)	BJ_cellLine (Foreskin Fibroblast Cell Line Exosomes)	HepG2_cellLine (Hepatocellular Carcinoma Cell Line Exosomes)	HUVEC_cellLine (Human Umbilical Vein Endothelial Cell Line Exosomes)	invasive_NFPAs (Invasive Non-functional Pituitary Adenomas Exosomes)
0	0.794661	1.46531	0.192239	6.72654
non-invasive_NFPAs (Non-invasive Non-functional Pituitary Adenomas Exosomes)	**MCF7_cellLine (Human Breast Cancer Cell Line Exosomes)**	**MDA-MB-231_cellLine (Human Breast Cancer Cell Line Exosomes)**	**tuberculosis_patients_serum (Active Tuberculosis Patients Serum Exosomes)**	**normal_people_blood (Normal People Blood Exosomes)**
10.668	0	0	6.51389	0

Figure 1. NONHSAG000148.2 exosome expression profile.

LncRNAs and SNPs

Genome-wide association studies (GWASs) have revealed many of the genetic variants associated with diseases. At least one-third of genetic variants belong to noncoding regions (36,37). The most common type of genetic variant in the human genome is single nucleotide polymorphisms (SNPs). Furthermore, the density of SNPs in lncRNA regions is similar to the density in protein coding regions. Some lncRNA intervals even have a higher density of SNPs than the genome average (37). SNPs in lncRNAs may influence their partner mRNAs by disturbing splicing or structural stability (38,39). Thus, the lncRNA SNP–disease relationship needs to be well studied. Functions of this relationship have been revealed in multiple disease processes, including carcinogenesis (40). For example, a study found that the TT genotype of rs3787016 was related to the risk of breast cancer and the clinicopathological features of the tumour among premenopausal women (41). To integrate the SNP information with the NONCODEv5 human transcripts, we collected lncRNA–SNP associations from LincSNP 2.0 (35). LincSNP 2.0 is a database that links disease-associated SNPs to human large intergenic noncoding RNAs and has summarized 809 451 lncRNA SNPs from eight databases (dbGaP (42), GAD (43), GWAS Central (44), Johnson and O'Donnell (45), the NHGRI GWAS Catalog (46), PharmGKb (47), GWASdb (48) and GRASP (49)). Using information for SNP chromosome positions, we obtained lncRNA SNPs with Bedtools (50) Intersect. Thus, NONCODEv5 includes 724 579 SNPs that potentially exist in lncRNA intervals, which provide researchers with a convenient resource to acquire SNP–disease information for their lncRNA of interest.

LncRNAs and Exosome expression profiles

Exosomes, which are produced by endocytosis, are small membrane vesicles that are secreted by most cells (51). Studies have shown that exosomes can be used for the early disease diagnosis of cancer and as potential drug targets. Exosomes can also change the target cell microenvironment and contribute to cancer metastasis. In a study of pancreatic cancer, the researchers found that the abundance of GPC1 (glypican-1)-positive exosomes in the sera of patients with early pancreatic cancer was significantly higher than the abundance in the normal population; this finding provides a significant basis for the early diagnosis of cancer (52). Another study of exosomes showed that metastatic tumour cells released exosome prior to departure. Exosome arrived at the metastatic organ and were taken in by the corresponding cells to change the state of these target cells and create a suitable condition for tumour growth (53). However, most previous studies have focused solely on exosomal mRNAs, miRNAs, proteins and lipids. A recent study showed that lncRNAs protected by exosomes were up-regulated and promoted cell proliferation and migration in non-small cell lung cancer (54). NONCODE has noted this field, which is full of potential, and has provided the first expression profiles of lncRNAs in exosome (Figure 1) based on high-throughput analyses of exosome RNA-sequencing data. Exosome-related RNA-seq data were collected from 6 RNA-sequencing datasets downloaded from the GEO database (Table 1). The exosome sources included six cell lines (A431, BJ, HepG2, HUVEC, MCF7 and MDA-MB-231) and four tissues [invasive non-functional pituitary adenomas (NFPAs), non-invasive NFPAs, tuberculosis patient serum and blood from normal individuals]. Then, the ribosome RNA sequencing reads were removed using sortmerna-2.1b (55). The read quality was filtered using Trimmomatic-0.36 (56) through a 4-base sliding window with an average quality threshold of 24, and reads with length less than 36 were dropped. The filtered reads were mapped to the genome with the split-aware aligner

Table 1. Exosome RNA-sequencing datasets

Exosome source	GSE ID	Description	Citation
Tuberculosis patients serum	GSE94907	Active tuberculosis patient serum exosomes	Lv L *et al.* 2017 Front Microbiol (59)
BJ cell line	GSE89926	Human untreated foreskin fibroblast exosomes	Prakash A *et al.* (unpublished)
HUVEC cell line	GSE89926	Human untreated endothelial cell exosomes	Prakash A *et al.* (unpublished)
Invasive NFPAs	GSE89779	Invasive non-functional pituitary adenomas exosomes	Ren Y *et al.* (unpublished)
Non-invasive NFPAs	GSE89779	Non-invasive non-functional pituitary adenoma exosomes	Ren Y *et al.* (unpublished)
MDA-MB-231 cell line	GSE58464	Breast cancer ultracentrifugation method extracted exosomes	Ghosh A *et al.* (unpublished)
MCF7 cell line	GSE58464	Breast cancer ultracentrifugation method extracted exosomes	Ghosh A *et al.* (unpublished)
A431 cell line	GSE76173	A431 squamous cell carcinoma cell line exosomes	Lefebvre FA *et al.* 2016 Sci Rep (60)
HepG2 cell line	GSE76173	HepG2 hepatocellular carcinoma cell line exosomes	Lefebvre FA *et al.* 2016 Sci Rep (60)
Normal people blood	GSE100206	Normal blood extracted exosomes	Li Y *et al.* 2015 Cell Res (61)

STAR (57) with '–outFilterScoreMinOverLread' 0.1 and '–outFilterMatchNminOverLread' 0.1. To avoid their influence, we removed duplicated reads using picard-tools-2.1.0 (http://broadInstitute.github.io/picard). Finally, we calculated the lncRNA expressions. We used featureCounts (58) to quantify read counts for each NONCODE gene or transcript with default parameters. For paired-end data, we set 'isPairedEnd = TRUE' and 'requireBothEndsMapped = TRUE' additionally. Then, we used DGEList and rpkm function from edgeR (59) to calculate the normalized expression levels in FPKM (fragments per kilobases per million mapped reads; counted on read pairs in case of paired-end data).

LncRNAs and RNA secondary structure

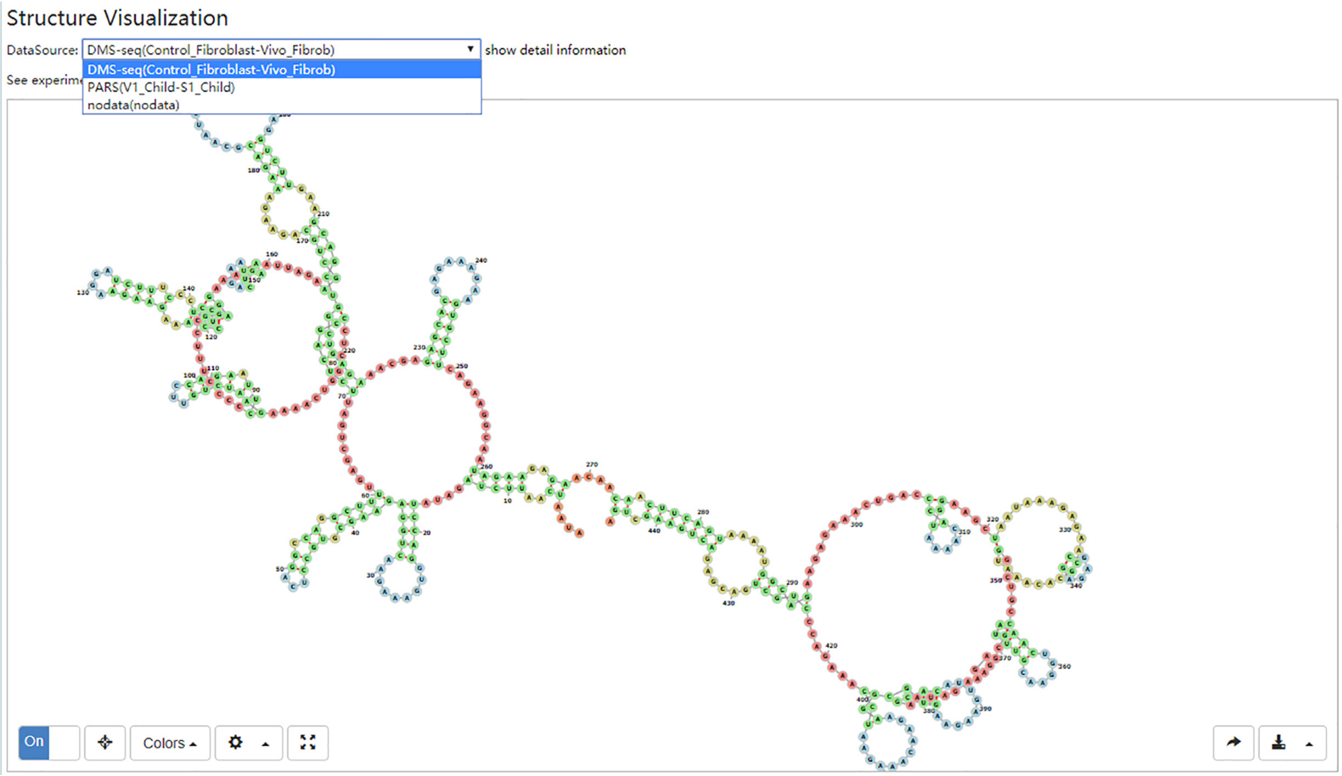
In parallel to the genetic code for protein synthesis, a secondary layer of information is embedded in all RNA transcripts in the form of the RNA structure. The RNA structure influences practically every step in the gene expression program (62).

Several high-throughput technologies have been developed recently to probe RNA secondary structures at the transcriptome level in human. These technologies are based on enzyme cleavage or chemical modification of nucleotides with specific structural states (e.g. loop regions or double-stranded regions), which can be detected by high-throughput sequencing via the stops they cause during reverse transcription (RT stops) (63). For example, parallel analysis of the RNA structure (PARS) utilizes RNase V1 and nuclease S1 simultaneously to probe RNA structures (63). DMS-seq or Structure-seq uses the small molecule dimethyl sulfate (DMS) to modify adenines and cytosines in single-stranded states both *in vivo* and *in vitro* (64,65).

We selected representative human structure probing data (Table 2) reprocessed by the RNAex web server (63) and predicted the RNA secondary structures of the human transcripts in NONCODEv5 restrained by these data using RME (66), which is a software program that can incorporate multiple types of experimental probing data for NONCODE human transcripts. To remain in agreement with RNAex, we transformed the genomic positions of the NONCODE transcript to hg19 using liftover (http://genome.ucsc.edu/cgi-bin/hgLiftOver) and used the RME parameter trained and provided by RNAex for each dataset. This approach only predicts minority RNA secondary structures through dataset correction with the limitation of sequencing data coverage. The V1 Child-S1 Child data pair only covered 4756 transcripts, whereas the data pair Control Fibroblast-Vivo Fibroblast covered 5,795 transcripts of a total number of 172 216 transcripts. Moreover, 3059 transcripts were covered by both V1 Child-S1 Child and Control Fibroblast-Vivo Fibroblast. We also used RME without any dataset restraint. The RNA secondary structure of each transcript can be accessed through the NONCODE website; this option is supported by the RNA secondary structure visualization tool forna (67), which allows RNA secondary structures to be displayed directly in the browser (Figure 2). NONCODEv5 can provide the corresponding RNA secondary structures predicted in different datasets.

DISCUSSION

NONCODEv5 contains a total of 548 640 transcripts. Compared with NONCODE 2016, the number of human transcripts increased from 167 150 to 172 216, and the number of mouse transcripts increased from 130 558 to 131 697. The increase in transcripts is not much larger than the last update. NONCODE 2016 collected a large quantity of new transcripts due to the rapid increase in second generation sequencing data. Although the identification of novel lncRNAs was much slower in recent years, many researchers changed their study focus from novel lncRNA detection to lncRNA function and annotation. For example, some researchers concentrated on the relationships between lncRNAs and diseases, how lncRNA secondary structure information influenced the expression of lncRNAs and how the expression of lncRNAs in exosomes contributed to disease diagnosis. NONCODEv5 made the corresponding adjustments to these changes in the collection, analysis, and storage methods for the above information. However, annotation data are still scarce for lncRNAs. For example, in consideration of limited resources, six human-related exosome datasets were collected, but only two exosome datasets for mice were found; therefore, we only calculated exosome expression for human, and data from other species need to



Structure Visualization

DataSource: DMS-seq(Control_Fibroblast-Vivo_Fibrob)

Figure 2. NONHSAT001763.2 RNA secondary structure prediction.

Table 2. The structure-probing data used in NONCODE

Data type	Sample	Condition	Raw data	Citation
DMS-seq	Control Fibroblast Vivo Fibroblast	in vitro in vivo	GSE45803	Rouskin *et al.*, 2014 Nature (64)
PARS	V1 Child S1 Child	in vitro in vitro	GSE50676	Wan *et al.*, 2014 Nature (62)

be collected in the future. Furthermore, the datasets used to predict RNA secondary structures only covered a small portion of the human reference genome, and thus the RNA secondary structures predicted with the datasets were retained. We will continue to follow up with the latest released datasets to enrich the annotation of lncRNAs.

As a comprehensive database of non-coding RNAs, NONCODE devoted itself to collect lncRNAs thoroughly from literatures and other databases, and annotate these lncRNAs exhaustively. NONCODE is one of the expert databases of RNAcentral, which is a public resource that offers integrated access to a comprehensive and up-to-date set of non-coding RNA sequences. Currently, NONCODE covered the sequence, structure, expression, function, conservation, disease relevance and many other aspects of lncRNAs. Compared to other lncRNA databases, such as GENCODE/Ensembl/Refseq, NONCODE collected more lncRNA transcripts, and provided unique annotations of lncRNAs, such as RNA secondary structure, expression of exosome, association between lncRNA and disease. When researchers were concerned with the whole repository of lncRNA isoforms, or interested in the above-mentioned unique annotations, it is a good choice to use NONCODE. And when they paid more attentions to representative lncRNA isoforms, NONCODE also supports to retrieve lncRNAs subsets to meet the quality demands of researchers since NONCODE 2016 (18). Several subsets were provided according to the source of the data, such as from literature, Refseq, Ensembl and other databases supported. And a lot of quality controls were implemented on these subsets, including exon number, transcript length, and CNCI score for these transcripts.

ACKNOWLEDGEMENTS

We thank Zhi John Lu for help with the RNA secondary structure predictions of human transcripts in NONCODEv5.

FUNDING

National Natural Science Foundation of China [31371320, 31401119, 31501066]; National Science Foundation for Young Scholars [31701141, 31701149]. Funding for open access charge: Our institution will pay the publication fee.

Conflict of interest statement. None declared.

REFERENCES

1. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
2. Ponting,C.P., Oliver,P.L. and Reik,W. (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.
3. Pennisi,E. (2010) Shining a light on the genome's 'dark matter'. *Science*, **330**, 1614.
4. Fu,M., Zou,C., Pan,L., Liang,W., Qian,H., Xu,W., Jiang,P. and Zhang,X. (2016) Long noncoding RNAs in digestive system cancers: Functional roles, molecular mechanisms, and clinical implications (Review). *Oncol. Rep.*, **36**, 1207–1218.
5. Li,T., Mo,X., Fu,L., Xiao,B. and Guo,J. (2016) Molecular mechanisms of long noncoding RNAs on gastric cancer. *Oncotarget*, **7**, 8601–8612.
6. Gupta,R.A., Shah,N., Wang,K.C., Kim,J., Horlings,H.M., Wong,D.J., Tsai,M.C., Hung,T., Argani,P., Rinn,J.L. *et al.* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, **464**, 1071–1076.
7. Tsai,M.C., Manor,O., Wan,Y., Mosammaparast,N., Wang,J.K., Lan,F., Shi,Y., Segal,E. and Chang,H.Y. (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, **329**, 689–693.
8. Weakley,S.M., Wang,H., Yao,Q. and Chen,C. (2011) Expression and function of a large non-coding RNA gene XIST in human cancer. *World J. Surg.*, **35**, 1751–1756.
9. Guttman,M., Donaghey,J., Carey,B.W., Garber,M., Grenier,J.K., Munson,G., Young,G., Lucas,A.B., Ach,R., Bruhn,L. *et al.* (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*, **477**, 295–300.
10. Yates,A., Akanni,W., Amode,M.R., Barrell,D., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P., Fitzgerald,S., Gil,L. *et al.* (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.
11. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
12. 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.
13. 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**, 4363–4364.
14. Tickner,J.A., Urquhart,A.J., Stephenson,S.A., Richard,D.J. and O'Byrne,K.J. (2014) Functions and therapeutic roles of exosomes in cancer. *Front. Oncol.*, **4**, 127.
15. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
16. Wan,Y., Kertesz,M., Spitale,R.C., Segal,E. and Chang,H.Y. (2011) Understanding the transcriptome through RNA structure. *Nat. Rev. Genet.*, **12**, 641–655.
17. Wan,Y., Qu,K., Ouyang,Z. and Chang,H.Y. (2013) Genome-wide mapping of RNA structure using nuclease digestion and high-throughput sequencing. *Nat. Protoc.*, **8**, 849–869.
18. Zhao,Y., Li,H., Fang,S., Kang,Y., Wu,W., Hao,Y., Li,Z., Bu,D., Sun,N., Zhang,M.Q. *et al.* (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.*, **44**, D203–D208.
19. 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.
20. Bu,D., Yu,K., Sun,S., Xie,C., Skogerbo,G., Miao,R., Xiao,H., Liao,Q., Luo,H., Zhao,G. *et al.* (2012) NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Res.*, **40**, D210–D215.
21. Berardini,T.Z., Reiser,L., Li,D., Mezheritsky,Y., Muller,R., Strait,E. and Huala,E. (2015) The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis*, **53**, 474–485.
22. Gramates,L.S., Marygold,S.J., Santos,G.D., Urbano,J.M., Antonazzo,G., Matthews,B.B., Rey,A.J., Tabone,C.J., Crosby,M.A., Emmert,D.B. *et al.* (2017) FlyBase at 25: looking to the future. *Nucleic Acids Res.*, **45**, D663–D671.
23. Kapranov,P., Cheng,J., Dike,S., Nix,D.A., Duttagupta,R., Willingham,A.T., Stadler,P.F., Hertel,J., Hackermuller,J., Hofacker,I.L. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, **316**, 1484–1488.
24. Hao,Y., Zhang,L., Niu,Y., Cai,T., Luo,J., He,S., Zhang,B., Zhang,D., Qin,Y., Yang,F. *et al.* (2017) SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Brief. Bioinform.*, bbx005.
25. Schmitz,S.U., Grote,P. and Herrmann,B.G. (2016) Mechanisms of long noncoding RNA function in development and disease. *Cell. Mol. Life Sci.*, **73**, 2491–2509.
26. Wapinski,O. and Chang,H.Y. (2011) Long noncoding RNAs and human disease. *Trends Cell Biol.*, **21**, 354–361.
27. Sun,Y., Wei,G., Luo,H., Wu,W., Skogerbo,G., Luo,J. and Chen,R. (2017) The long noncoding RNA SNHG1 promotes tumor growth through regulating transcription of both local and distal genes. *Oncogene*, doi:10.1038/onc.2017.286.
28. Liu,L., Yue,H., Liu,Q., Yuan,J., Li,J., Wei,G., Chen,X., Lu,Y., Guo,M., Luo,J. *et al.* (2016) LncRNA MT1JP functions as a tumor suppressor by interacting with TIAR to modulate the p53 pathway. *Oncotarget*, **7**, 15787–15800.
29. Qureshi,I.A. and Mehler,M.F. (2012) Emerging roles of non-coding RNAs in brain evolution, development, plasticity and disease. *Nat. Rev. Neurosci.*, **13**, 528–541.
30. 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.
31. 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.
32. Wang,Y., Chen,L., Chen,B., Li,X., Kang,J., Fan,K., Hu,Y., Xu,J., Yi,L., Yang,J. *et al.* (2013) Mammalian ncRNA–disease repository: a global view of ncRNA-mediated disease network. *Cell Death Dis*, **4**, e765.
33. 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.
34. Wang,J., Cao,Y., Zhang,H., Wang,T., Tian,Q., Lu,X., Lu,X., Kong,X., Liu,Z., Wang,N. *et al.* (2017) NSDNA: a manually curated database of experimentally supported ncRNAs associated with nervous system diseases. *Nucleic Acids Res.*, **45**, D902–D907.
35. Ning,S., Yue,M., Wang,P., Liu,Y., Zhi,H., Zhang,Y., Zhang,J., Gao,Y., Guo,M., Zhou,D. *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.
36. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9362–9367.
37. Jin,G., Sun,J., Isaacs,S.D., Wiley,K.E., Kim,S.T., Chu,L.W., Zhang,Z., Zhao,H., Zheng,S.L., Isaacs,W.B. *et al.* (2011) Human polymorphisms at long non-coding RNAs (lncRNAs) and association with prostate cancer risk. *Carcinogenesis*, **32**, 1655–1659.
38. Chung,S., Nakagawa,H., Uemura,M., Piao,L., Ashikawa,K., Hosono,N., Takata,R., Akamatsu,S., Kawaguchi,T., Morizono,T. *et al.* (2011) Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. *Cancer Sci.*, **102**, 245–252.
39. Burd,C.E., Jeck,W.R., Liu,Y., Sanoff,H.K., Wang,Z. and Sharpless,N.E. (2010) Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. *PLoS Genet.*, **6**, e1001233.
40. Li,L., Sun,R., Liang,Y., Pan,X., Li,Z., Bai,P., Zeng,X., Zhang,D., Zhang,L. and Gao,L. (2013) Association between polymorphisms in

long non-coding RNA PRNCR1 in 8q24 and risk of colorectal cancer. *J. Exp. Clin. Cancer Res.*, **32**, 104.

41. Xu,T., Hu,X.X., Liu,X.X., Wang,H.J., Lin,K., Pan,Y.Q., Sun,H.L., Peng,H.X., Chen,X.X., Wang,S.K. *et al.* (2017) Association between SNPs in Long Non-coding RNAs and the Risk of Female Breast Cancer in a Chinese Population. *J. Cancer*, **8**, 1162–1169.
42. Mailman,M.D., Feolo,M., Jin,Y., Kimura,M., Tryka,K., Bagoutdinov,R., Hao,L., Kiang,A., Paschall,J., Phan,L. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.*, **39**, 1181–1186.
43. Becker,K.G., Barnes,K.C., Bright,T.J. and Wang,S.A. (2004) The genetic association database. *Nat. Genet.*, **36**, 431–432.
44. Beck,T., Hastings,R.K., Gollapudi,S., Free,R.C. and Brookes,A.J. (2014) GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur. J. Hum Genet.*, **22**, 949–952.
45. Johnson,A.D. and O'Donnell,C.J. (2009) An open access database of genome-wide association results. *BMC Med. Genet.*, **10**, 6.
46. Welter,D., MacArthur,J., Morales,J., Burdett,T., Hall,P., Junkins,H., Klemm,A., Flicek,P., Manolio,T., Hindorff,L. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.
47. Altman,R.B. (2007) PharmGKB: a logical home for knowledge relating genotype to drug response phenotype. *Nat. Genet.*, **39**, 426.
48. Li,M.J., Liu,Z., Wang,P., Wong,M.P., Nelson,M.R., Kocher,J.P., Yeager,M., Sham,P.C., Chanock,S.J., Xia,Z. *et al.* (2016) GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, **44**, D869–D876.
49. Leslie,R., O'Donnell,C.J. and Johnson,A.D. (2014) GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. *Bioinformatics*, **30**, i185–i194.
50. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
51. Thery,C., Zitvogel,L. and Amigorena,S. (2002) Exosomes: composition, biogenesis and function. *Nat. Rev. Immunol.*, **2**, 569–579.
52. Melo,S.A., Luecke,L.B., Kahlert,C., Fernandez,A.F., Gammon,S.T., Kaye,J., LeBleu,V.S., Mittendorf,E.A., Weitz,J., Rahbari,N. *et al.* (2015) Glypican-1 identifies cancer exosomes and detects early pancreatic cancer. *Nature*, **523**, 177–182.
53. Hoshino,A., Costa-Silva,B., Shen,T.L., Rodrigues,G., Hashimoto,A., Tesic Mark,M., Molina,H., Kohsaka,S., Di Giannatale,A., Ceder,S. *et al.* (2015) Tumour exosome integrins determine organotropic metastasis. *Nature*, **527**, 329–335.
54. Zhang,R., Xia,Y., Wang,Z., Zheng,J., Chen,Y., Li,X., Wang,Y. and Ming,H. (2017) Serum long non coding RNA MALAT-1 protected by exosomes is up-regulated and promotes cell proliferation and migration in non-small cell lung cancer. *Biochem. Biophys. Res. Commun.*, **490**, 406–414.
55. Kopylova,E., Noe,L. and Touzet,H. (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, **28**, 3211–3217.
56. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
57. 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.
58. 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.
59. 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.
60. Lefebvre,F.A., Benoit Bouvrette,L.P., Perras,L., Blanchet-Cohen,A., Garnier,D., Rak,J. and Lecuyer,E. (2016) Comparative transcriptomic analysis of human and Drosophila extracellular vesicles. *Sci. Rep.*, **6**, 27680.
61. Li,Y., Zheng,Q., Bao,C., Li,S., Guo,W., Zhao,J., Chen,D., Gu,J., He,X. and Huang,S. (2015) Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Cell Res.*, **25**, 981–984.
62. Wan,Y., Qu,K., Zhang,Q.C., Flynn,R.A., Manor,O., Ouyang,Z., Zhang,J., Spitale,R.C., Snyder,M.P., Segal,E. *et al.* (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, **505**, 706–709.
63. Wu,Y., Qu,R., Huang,Y., Shi,B., Liu,M., Li,Y. and Lu,Z.J. (2016) RNAex: an RNA secondary structure prediction server enhanced by high-throughput structure-probing data. *Nucleic Acids Res.*, **44**, W294–W301.
64. Rouskin,S., Zubradt,M., Washietl,S., Kellis,M. and Weissman,J.S. (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, **505**, 701–705.
65. Ding,Y., Tang,Y., Kwok,C.K., Zhang,Y., Bevilacqua,P.C. and Assmann,S.M. (2014) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, **505**, 696–700.
66. Wu,Y., Shi,B., Ding,X., Liu,T., Hu,X., Yip,K.Y., Yang,Z.R., Mathews,D.H. and Lu,Z.J. (2015) Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. *Nucleic Acids Res.*, **43**, 7247–7259.
67. Kerpedjiev,P., Hammer,S. and Hofacker,I.L. (2015) Forna (force-directed RNA): simple and effective online RNA secondary structure diagrams. *Bioinformatics*, **31**, 3377–3379.