NONCODE 2016: an informative and valuable data source of long non-coding RNAs

Yi Zhao1,†, Hui Li2,3,†, Shuangsong Fang2,3, Yue Kang3,4, Wei wu3,4, Yajing Hao3,4, Ziyang Li2, Dechao Bu2, Ninghui Sun2, Michael Q. Zhang1,* and Runsheng Chen4,*

1School of Medicine, MOE Key Laboratory of Bioinformatics and Bioinformatics Division, Center for Synthetic and System Biology, TNLST/Department of Automation, Tsinghua University, Beijing 100084, China, 2Key Laboratory of Intelligent Information Processing, Advanced Computer Research Center, State Key Laboratory of Computer Architecture, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, 3University of Chinese Academy of Sciences, Beijing 100049, China and 4CAS Key Laboratory of RNA Biology, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

Received September 15, 2015; Revised October 29, 2015; Accepted October 31, 2015

ABSTRACT

NONCODE (http://www.bioinfo.org/noncode/) is an interactive database that aims to present the most complete collection and annotation of non-coding RNAs, especially long non-coding RNAs (IncRNAs). The recently reduced cost of RNA sequencing has produced an explosion of newly identified data. Revolutionary third-generation sequencing methods have also contributed to more accurate annotations. Accumulative experimental data also provides more comprehensive knowledge of IncRNA functions. In this update, NONCODE has added six new species, bringing the total to 16 species altogether. The IncRNAs in NONCODE have increased from 210 831 to 527,336. For human and mouse, the IncRNA numbers are 167,150 and 130,558, respectively. NONCODE 2016 has also introduced three important new features: (i) conservation annotation; (ii) the relationships between lncRNAs and diseases; and (iii) an interface to choose high-quality datasets through predicted scores, literature support and long-read sequencing method support. NONCODE is also accessible through http://www.noncode.org.

INTRODUCTION

Recent whole transcriptome studies have revealed that about three quarters of the human genome is capable of being transcribed, while protein-coding regions account for just 2% of the genome (1–3). Therefore, the vast majority of transcribed sequences do not encode proteins, and are called non-coding RNA (ncRNA). Accumulating evidence shows that non-coding RNAs play key roles in various biological processes, such as imprinting control, the circuitry controlling pluripotency and differentiation, immune responses, and chromosome dynamics (4). ncRNAs are as important as protein-coding genes to cellular functions (5,6). Notably, a growing number of long ncRNAs (lncRNAs), which are considered to be >200 nt in length and are often multieexonic (7), have been implicated in disease etiology (8–10). It is therefore of great importance to collect ncRNA information and store this information in a one-stop knowledge gateway for ncRNAs, the NONCODE database.

The development of high-throughput sequencing methodologies has reduced the cost of RNA sequencing, and as a result there has been an explosive rise in the number of newly identified IncRNAs. For example, in 2015, Chinnaiyan et al. established a consensus set of 384,066 predicted transcripts from 7256 RNA-seq libraries, which were designated as the MiTranscriptome assembly (11). Since then the revolutionary advancement of sequencing methods, such as single-molecule long-read techniques, leads us closer to the real lncRNA transcriptome. Given sufficient material, amplification-free sequencing of full-length cDNA molecules provides a more direct view of RNA molecules (12). NONCODE has collected data from literature published since the last update and includes the latest versions of several public databases (Ensembl (13), RefSeq(14), lncRNAdb (15) and GENCODE (16)). After the removal of false and redundant IncRNAs, NONCODE contains a total of 527,336 transcripts.

In addition to the identification of new lncRNAs, data on the genetics and biochemical properties of lncRNAs has accumulated rapidly. Of the papers retrieved from Pubmed for IncRNAs, we found that the vast majority studied IncRNA
function, especially the relationship between lncRNAs and disease (8–10). In large-scale searches for single-base differences between diseased and healthy individuals, about 40% of the disease-related differences show up in genomic regions outside of protein-coding genes. This implicates non-coding regions as vital for genetic risk factors of disease (2).

In order to enable a systematic compilation and integration of this information, we added the relationships between lncRNAs and diseases to the annotations of NONCODE. The sources for these annotations were derived from literature mining, differential lncRNA analysis utilizing public RNA-seq data and microarray data and mutation analysis from public genome-wide association study (GWAS) data.

Along with the ever increasing number of lncRNAs and the amount and functional study data, genome-wide conservation information is required for biologists to study the mechanisms of lncRNA actions. In order to explore the conservation information of lncRNAs, NONCODE collected six new mammalian species (chimpanzee, gorilla, orangutan, rhesus macaque, opossum and platypus) (17). Conservation annotation is available on the information page of each NONCODE lncRNA gene. Users can browse the conserved counterparts of any human lncRNA gene in other species through a phylogenetic tree layout. This conservation information should greatly increase the convenience of studying lncRNA functions.

DATA COLLECTION AND PROCESSING

Similar to the former iterations of NONCODE (18–21), the source of NONCODE 2016 includes the previous versions of NONCODE, the collated literature and other public databases. We searched PubMed using the key words ‘ncrna’, ‘noncoding’, ‘non-coding’, ‘no code’, ‘non-code’, ‘lncrna’ and ‘lincrna’, and found 6532 new articles since 1 June 2013 (the last collection date for NONCODE). We retrieved the newly identified lncRNAs and their annotations from the supplementary material or web site of these articles. Together with the newest data from Ensembl, RefSeq, lncRNAdb, GENCODE and the old versions of NONCODE data, literature data were processed through a standard pipeline for each species. The pipeline included the following six steps:

(i) Format normalization. All of the input data were processed into bed or gtf formats based on one assembly version, for example, hg38 for human and mm10 for mouse.

(ii) Combination. All of the normalized data files were combined together using the Cuffcompare program in the Cufflinks suite (22). After eliminating redundancy, every new transcript ID and the accompanying resources were extracted.

(iii) Filtering protein-coding RNA. We filtered out protein-coding RNA using two methods. Firstly, the RNA was compared with the coding RNA in RefSeq and Ensembl, and the ‘=’ and ‘c’ transcripts were excluded. Secondly, the RNA was filtered through the Coding-Non-Coding Index (CNCI) (23) program and only the RNAs considered non-coding by CNCI were kept.

(iv) Information retrieval. We assigned each transcript a name according to the criterion of NONCODE v4 and extracted basic information such as location (24), exons, length, assembly sequence, source, etc.

(v) Advanced annotation. Advanced annotations included expression profiles, predicted functions, conservation, disease information, etc. Human expression profiles were collected from 16 tissues of the Human BodyMap 2.0 data (ENA archive: ERP000546) and eight cell lines (GEO accession no. GSE30554), while mouse data was collected from six different tissues (ENA archive: ERP000591). Functions for the lncRNA genes were predicted by lnc-GFP (25), a coding-non-coding co-expression network (26,27) based global function predictor.

(vi) Web presence. The new NONCODE has provided completely new web pages. More annotation information has been added and a more user-friendly interface has been introduced.

STATISTICS OF NONCODE

NONCODE contains 527,336 lncRNA transcripts from 16 species (human, mouse, cow, rat, chimpanzee, gorilla, orangutan, rhesus macaque, opossum, platypus, chicken, zebrafish, fruitfly, Caenorhabditis elegans, yeast and Arabidopsis). According to the definition of lncRNA genes (18), NONCODE collected 337,880 genes altogether. A total of 101,700 and 86,935 genes were generated from 167,150 and 130,558 lncRNAs from human and mouse (shown in Table 1), respectively. Following the nomenclature of NONCODE v4 (18), both lncRNA transcripts and genes were designated systematically: NON+ three characters (representing a species) +T (transcript) or G (gene) + six sequential numbers. NONCODE has annotated expression profiles from all the human and mouse transcripts and genes, and a large number of these genes were annotated with predicted functions.

LncRNAs and Diseases

Definitive evidence has proven that transcription of the non-coding genome has produced functional RNAs (1). In particular, lncRNAs have been implicated in biological, developmental, and pathological processes, and acted through mechanisms such as chromatin reprogramming, cis regulation at enhancers, and post-transcriptional regulation of mRNA processing (28). lncRNAs are therefore considered to be important regulators of tissue physiology and disease processes including cancer (11). Although we have collected functional interactions between ncRNAs and biomolecules in NPInter (29–31), we think it is also necessary to include disease information into NONCODE.

The data retrieval pipeline is listed in Figure 1. Recent published papers have been explored, and the proven associations between NONCODE transcripts and diseases has been integrated into the latest version. The NONCODE assembly also assessed the overlaps of transcripts with the unique disease-associated Single-Nucleotide Polymorphisms (SNPs) from a catalog of GWASs (32) and the SNP database (dbSNP) (33). There were also a lot of rela-
Table 1. Transcript and gene statistics for NONCODE

| Species            | Number of lncRNA transcripts | Number of lncRNA genes |
|--------------------|------------------------------|------------------------|
| human              | Homo sapiens                 | 167,150                | 101,700                |
| mouse              | Mus musculus                 | 130,558                | 86,935                 |
| cow                | Bos taurus                   | 23,599                 | 18,189                 |
| rat                | Rattus rattus                | 29,070                 | 25,114                 |
| chimpanzee         | Pan troglodytes              | 18,604                 | 13,224                 |
| gorilla            | Gorilla gorilla              | 20,785                 | 17,140                 |
| orangutan          | Pongo pygmaeus               | 15,601                 | 13,432                 |
| rhesus macaque     | Macaca mulatta               | 9325                   | 6125                   |
| opossum            | Monodelphis domestica        | 21,014                 | 14,135                 |
| platypus           | Ornithorhynchus anatinus     | 11,518                 | 9394                   |
| chicken            | Gallus gallus                | 13,085                 | 9688                   |
| zebrafish          | Danio rerio                  | 5000                   | 3635                   |
| fruitfly           | Drosophila melanogaster      | 54,818                 | 13,890                 |
| C. elegans         | Caenorhabditis elegans       | 3269                   | 2746                   |
| yeast              | Saccharomyces cerevisiae     | 60                     | 56                     |
| Arabidopsis        | Arabidopsis thalana          | 3853                   | 2477                   |
| Total              |                              | 527,336                | 337,880                |

Figure 1. Disease related data acquisition pipeline

LncRNA CONSERVATION

Compared to protein-coding genes and small RNAs (e.g. miRNAs and snoRNAs), several reports have suggested that lncRNAs are modestly conserved (11). Most lncRNAs are less conserved in sequence (34), but there are still many lncRNAs that are conserved in their genomic loci, exonic sequences and promoter regions (35). These are preserved across multiple species, attesting to their important functional potentials (36).

Benefiting from next-generation sequencing technologies, ncRNAs are now more easily identified via transcript sequencing. NONCODE has added six new species, mainly from multi-species RNA-Seq data (37,38). An evolutionary tree from 12 commonly studied species (human, mouse, cow, rat, chicken, zebrafish, chimpanzee, gorilla, orangutan, rhesus macaque, opossum and platypus) was constructed using methods introduced in phyloNONCODE (39). Each human lncRNA gene counterpart from the other listed species can be retrieved through browsing the evolutionary tree (shown in Figure 2). The counterpart of each lncRNA was computed using the UCSC LiftOver tool (40). In brief, LiftOver utilized BLASTZ (41), an independent implementation of the Gapped BLAST algorithm specifically designed for aligning two long genomic sequences, as a core algorithm to detect homologous regions in other genomes. After mapping to the second species, the counterpart region was intersected with the second species transcript. Users can browse the transcript...
Figure 2. Conservation annotation for NONHSAG200087.

QUALITY OF LncRNAs

Short-read sequencing technology allows high-throughput identification of lncRNAs. However, a proper method of in silico transcript reconstruction is an ongoing challenge. According to an assessment by the Paul Bertone group, <40% of known transcripts were well assembled from Homo sapiens RNA-seq data. The complexity of higher eukaryotic genomes imposes severe limitations on transcript recall and splice product discrimination that are likely to remain limiting factors for the analysis of current-generation RNA-seq data (42). Furthermore, multiple amplification steps during library preparation complicate the quantification of expression levels.

To some extent, third-generation sequencing technologies reduced the noise. This provides a more comprehensive assessment of the true complexity of the transcriptome. Given sufficient material, amplification-free and fragmentation-free sequencing of full-length cDNA molecules provides a more direct view of RNA molecules (12). NONCODE contacted the authors of the third-generation single-molecule long-read survey of the human transcriptome paper (12). After our analysis, the single-molecule lncRNA transcripts were included into NONCODE.

To meet the quality demands of researchers, NONCODE provides a subset searching interface. Users can choose the subset which is considered high quality. The quality controls include the source of the data, literature support, other database support and long-read sequencing method support. The controls also include selection of exon numbers, the lengths of the transcripts and prediction tools support. The web interface will return the subset according to the conditions users chose and allow users to download the data.

DISCUSSION

NONCODE 2016 contains 527,336 lncRNAs from 16 different species, this compares favorably with other lncRNA databases. For example, LNCipedia (human only) contains 111 685 transcripts (43), lncRNAtor (human, mouse, fly, zebrafish, worm and yeast) contains 34 605 transcripts (44), while the lncRNAWiki (human only) contains 105 255 transcripts (45). As mentioned above, technical limitations imposed by short-read sequencing lead to a number of computational challenges in transcript reconstruction and quantification. For many transcripts, automated methods failed to identify all of the constituent exons, and in cases in which all exons were reported, the protocols tested often failed to assemble the exons into complete isoforms (42). Considering this point, NONCODE filtered out some datasets. For example, although we have obtained all the data from MiTranscriptome (11), which contains 384 066 human lncRNAs from 7256 RNA-seq libraries, the detection of precise RefSeq splicing patterns from MiTranscriptome was only 31%, and the fraction of annotated genes within the entire MiTranscriptome was only 46%. Although it is reasonable to assume that unannotated transcription is unique to specific lineages, the low RefSeq detection rate was unusual. We therefore made a decision that NONCODE would not include MiTranscriptome data in the current version. In the future, we will attempt to make clear the real reason(s). Perhaps a more comprehensive construction tool is required to answer this question.

ACKNOWLEDGEMENT

We wish to thank Dr Greg Vatcher, Dr Jianjun Luo and Adistan Wahed for carefully reading our manuscript.
FUNDING

National High Technology Research and Development Program (‘863’ Program) of China [2014AA021103, 2014AA021502]; Training Program of the Major Research Plan of the National Natural Science Foundation of China [91229120]; National Natural Science Foundation of China [31371320, 31401119]; Chinese Academy of Science Strategic Project of Leading Science and Technology [XDA01020402]. Funding for open access charge: National High Technology Research and Development Program (‘863’ Program) of China [2014AA021103].

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. Pennisi,E. (2010) Shining a light on the genome’s ‘dark matter’. Science, 330, 1614–1614.

3. Ponting,C.P., Oliver,P.L. and Reik,W. (2009) Evolution and functions of long non coding RNAs. Cell, 136, 629–643.

4. Guttman,M., Domínguez,J., Carey,B.W., Garber,M., Grenier,J.K., Munson,G., Young,A.B., Ach,R., Bruhn,L. et al. (2007) lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature, 477, U295–U260.

5. Rinn,J.L., Kertesz,M., Wang,J.K., Squazzo,S.L., Xu,X., Brugmann,S.A., Goodnough,L.H., Helms,J.A., Farnham,P.J., Segal,E. et al. (2011) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell, 149, 1311–1323.

6. Merkin,J., Russell,C., Chen,P. and Burge,C.B. (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science, 338, 1593–1599.

7. Wang,K.C. and Chang,H.Y. (2011) Molecular mechanisms of long noncoding RNAs reveals global properties and specific subclasses. Genes Dev., 25, 1915–1927.

8. Gupta,R.A., Shah,N., Wang,K.C., Yeh,Y., Zhou,Y., Lai,C., Wang,K., Ng,A.S. et al. (2009) Predicting functions of the non-coding transcriptome by comparative genomics. Nature, 460, 1313–1317.

9. Merkin,J., Russell,C., Chen,P. and Burge,C.B. (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science, 338, 1593–1599.

10. Xie,C.Y., Yuan,J., Li,H., Li,M., Zhao,G.G., Bu,D.C., Zhu,W.M., Wu,W., Chen,R.S. and Zhao,Y. (2014) NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res., 42, D98–D103.

11. Quek,X.C., Thomson,D.W., Maag,J.L., Bartonieck,N., Signal,B., Clark,M.B., Gloss,B.S. and Dinger,M.E. (2015) lncRNAdb v2.0: expanding the reference database for functional long non-coding RNAs. Nucleic Acids Res., 43, D161–D173.

12. Merkin,J., Russell,C., Chen,P. and Burge,C.B. (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science, 338, 1593–1599.

13. Merkin,J., Russell,C., Chen,P. and Burge,C.B. (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science, 338, 1593–1599.

14. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Stransky,N., Bader,J.S., see,B. et al. (2012) Ensembl 2012: an update on the evolution mechanisms of long non-coding RNAs act in the circuitry controlling pluripotency and differentiation. Nature, 477, U295–U260.

15. Wang,K.C. and Chang,H.Y. (2011) Long non-coding RNA as modular scaffold of histone modification complexes. Science, 329, 689–693.

16. Weakley,S.M., Wang,H., Yao,Q.Z. and Chen,C.Y. (2011) Expression and function of a large non-coding RNA gene XIST in human cancer. World J. Surg., 35, 1751–1756.

17. Merkin,J., Russell,C., Chen,P. and Burge,C.B. (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science, 338, 1593–1599.

18. Putzi,J., Balasubramanian,A. and Saha,B. (2013) A new step towards a comprehensive understanding of the human transcriptome. Nucleic Acids Res., 41, D124–D126.

19. Merkin,J., Russell,C., Chen,P. and Burge,C.B. (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science, 338, 1593–1599.

20. Yoshida,T., Nishizaka,M., Kato,Y., Oshida,M., Morita,R. et al. (2013) Long non-coding RNA as master regulator of cancer metastasis. Nat. Rev. Mol. Cell Biol., 14, 43–57.

21. Merkin,J., Russell,C., Chen,P. and Burge,C.B. (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science, 338, 1593–1599.
2011 The evolution of gene expression levels in mammalian organs. Nature, 478, 343–348.

38. Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, I.C., Grutzner, F. and Kaessmann, H. (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature, 505, 635–640.

39. Bu, D., Luo, H., Jiao, F., Fang, S., Tan, C., Liu, Z. and Zhao, Y. (2015) Evolutionary annotation of conserved long non-coding RNAs in major mammalian species. Sci. China Life Sci., 58, 787–798.

40. Dreszer, T.R., Karolchik, D., Zweig, A.S., Hinrichs, A.S., Raney, B.J., Kuhn, R.M., Meyer, L.R., Wong, M., Sloan, C.A., Rosenbloom, K.R. et al. (2012) The UCSC Genome Browser database: extensions and updates 2011. Nucleic Acids Res., 40, D918–D923.

41. Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D. and Miller, W. (2003) Human-mouse alignments with BLASTZ. Genome Res., 13, 103–107.

42. Steijger, T., Abril, J.F., Engstrom, P.G., Kokocinski, F., Hubbard, T.J., Guigo, R., Harrow, J., Bertone, P. and Consortium, R. (2013) Assessment of transcript reconstruction methods for RNA-seq. Nat. Methods, 10, 1177–1184.

43. 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 (vol 43, pg D174, 2015). Nucleic Acids Res. 43, 4363–4364.

44. Park, C., Yu, N., Choi, I., Kim, W. and Lee, S. (2014) IncRNAtor: a comprehensive resource for functional investigation of long non-coding RNAs. Bioinformatics, 30, 2480–2485.

45. 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.