NONCODEv4: exploring the world of long non-coding RNA genes

Chaoyong Xie1,2, Jiao Yuan2,3, Hui Li1,2, Ming Li1,2, Guoguang Zhao1, Dechao Bu1,2, Weimin Zhu4, Wei Wu2,3, Runsheng Chen2,* and Yi Zhao1,*

1Bioinformatics Research Group, Advanced Computing Research Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, 2University of Chinese Academy of Sciences, Beijing 100049, China, 3Laboratory of Noncoding RNA, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China and 4Taicang Institute of Life Sciences Information, Suzhou 215400, China

Received September 15, 2013; Revised November 4, 2013; Accepted November 5, 2013

ABSTRACT

NONCODE (http://www.bioinfo.org/noncode/) is an integrated knowledge database dedicated to non-coding RNAs (excluding tRNAs and rRNAs). Non-coding RNAs (ncRNAs) have been implicated in diseases and identified to play important roles in various biological processes. Since NONCODE version 3.0 was released 2 years ago, discovery of novel ncRNAs has been promoted by high-throughput RNA sequencing (RNA-Seq). In this update of NONCODE, we expand the ncRNA data set by collection of newly identified ncRNAs from literature published in the last 2 years and integration of the latest version of RefSeq and Ensembl. Particularly, the number of long non-coding RNA (lncRNA) has increased sharply from 73,327 to 210,831. Owing to similar alternative splicing pattern to mRNAs, the concept of lncRNA genes was put forward to help systematic understanding of lncRNAs. The 56,018 and 46,475 lncRNA genes were generated from 95,135 and 67,628 lncRNAs for human and mouse, respectively. Additionally, we present expression profile of lncRNA genes by graphs based on public RNA-seq data for human and mouse, as well as predict functions of these lncRNA genes. The improvements brought to the database also include an incorporation of an ID conversion tool from RefSeq or Ensembl ID to NONCODE ID and a service of lncRNA identification. NONCODE is also accessible through http://www.noncode.org/.

INTRODUCTION

Non-coding RNAs (ncRNAs) constitute a significant fraction of the transcriptome (1). Widespread application of high-throughput RNA sequencing (RNA-seq), with the aid of computational methods, has revealed increasing number of ncRNAs identified from various organisms (2). Especially, long non-coding RNAs (lncRNAs), which are considered to be >200 nt in length and are often multiexonic (3), have been identified to play critical roles in various processes including embryonic development (4), dosage compensation (5) and immune response (6). Owing to their functional significance, databases that integrate comprehensive information about lncRNAs can be helpful for understanding biological processes. However, existing lncRNA resources such as lncRNAdb (7), ncRNAdb (8) and Rfam (9) fail to cover most of the newly identified lncRNAs in recent studies. Consequently, we updated the NONCODE database to version 4.0, to keep up-to-date with the latest discovery of lncRNAs. The number of lncRNA entries in NONCODE version 4.0 has increased to 210,831.

Despite a lack of protein-coding ability, lncRNAs are similar to mRNAs in many ways (10). LncRNAs are involved in alternative splicing patterns that resemble mRNAs (11). Moreover, the majority of lncRNAs are spliced with similar exon/intron lengths to protein-coding genes (11,12). Considering the increasing number of lncRNA transcripts, proposing lncRNA gene structures is now necessary to gain a systematic understanding of lncRNAs. However, existing resources merely describe gene structures of lncRNAs. Following the classical definition of ‘gene’ for protein-coding RNAs (13), NONCODE version 4.0 unites genomic sequences encoding a coherent...
set of overlapping long non-coding transcripts into an lncRNA gene. Because many lncRNAs reside within or overlap protein-coding loci (11), we then classified lncRNAs genes into four categories according to their genomic location in relation to protein-coding genes: antisense, intergenic, sense exonic and sense non-exonic, respectively.

The emergence of a large amount of RNA-seq data not only facilitates identification and characterization of lncRNAs but also provides clues to understanding expression patterns (14) and potential functions of lncRNAs (15). For human and mouse, NONCODE version 4.0 presents expression patterns across various tissues, as well as predicted functions of lncRNAs inferred from public RNA-seq data. Other improvements of NONCODE version 4.0 include iLncRNA, an online lncRNA identification pipeline based on user supplied data, and an ncRNA ID conversion tool allowing query of accessions from various RNA databases. An overview of updates in NONCODE version 4.0 is shown in Figure 1.

NONCODE has already proven to be an important resource in the realm of ncRNA databases, and is therefore incorporated into other ncRNA databases such as fRNAAdb — a large collection of ncRNAs (16), GeneCards — the comprehensive human gene compendium (17) and DIANA-LncBase — a database for miRNAs targets on lncRNAs (18). We believe that the recent improvements in NONCODE version 4.0 will significantly contribute to the enhancement of these and potentially other ncRNA databases.

DATA COLLECTION, REDUNDANCY ELIMINATION AND FILTRATION

Based on former versions of NONCODE (19–21), new data sets from literatures and other specialized databases were collected. For literature mining, we first retrieved literature published since May 1, 2011, from PubMed, using the key words ‘ncrna’, ‘noncoding’, ‘non-coding’, ‘no code’, ‘non-code’, ‘lncrna’ or ‘lincrna’, and found 4572 relevant articles. Then sequences, genome locations and other relevant information concerning transcripts from manually selected reports on new ncRNAs were retrieved. Next, the latest releases of Ensembl (22) and RefSeq (23) were integrated to supplement our manual curation efforts. In total, 118 148, 141 194 and 35 445 transcripts were retrieved from literature, Ensembl and RefSeq, respectively.

A process of redundancy elimination was then performed on the ncRNAs collected from literature and specialized databases mentioned earlier in text, together with existing data in NONCODE version 3.0. Cuffcompare program in Cufflinks suite (24) was used to map the whole ncRNA data set back to annotations of itself. Transcripts completely matching each other, annotated with class code ‘=’ by Cuffcompare, were considered to be redundant and grouped in a single record.

Each transcript we collected was considered non-coding in the resource it came from. However, the same transcript might be assigned mutually exclusive annotations in different resources due to respective standards of distinguishing non-coding from protein-coding RNAs.

Therefore, we used two screening criteria for all transcripts in NONCODE version 4.0 to ensure no inclusion of protein-coding transcripts. First, all transcripts kept by the redundancy elimination step were compared with a reference set containing known protein-coding RNAs from Ensembl and RefSeq by Cuffcompare. Those completely matching protein-coding transcripts, annotated with class code ‘=’, were discarded. Second, the coding potential of each transcript was evaluated by our CNCI program (25). Transcripts classified into coding sequences by CNCI were discarded. The left transcripts were kept with high confidence to be non-coding and entered into NONCODE. Finally, 595 854 ncRNAs were finally recorded. Data expansion of NONCODE mainly resulted from new collection of lncRNAs of human and mouse. In all of the 210 831 lncRNA transcripts of the final catalog of NONCODE, 95 135 and 67 628 come from human and mouse, respectively, whereas the other 48 068 come from other organisms.

DEFINITION AND CATEGORIZATION OF LncRNA GENES

LncRNAs are similar to mRNAs in regards to alternative splicing (26), thus we united genomic sequences encoding a coherent set of overlapping lncRNAs into an lncRNA gene, following the classical definition of ‘gene’ for protein-coding RNAs. Different transcripts that intersected any exons of one other and resided on the same DNA strand were considered to belong to the same gene and clustered into a single gene record. In this way, 56 018 and 46 475 lncRNA genes were generated from 95 135 and 67 628 lncRNAs for human and mouse, respectively. Both lncRNA transcripts and genes were designated systematically in NONCODE. LncRNA transcripts from a same organism were numbered subsequently, starting with ‘NON’ followed by a symbol representing the organism. For example, ‘NONMMUT000020’ denotes a transcript from mouse (the beginning ‘NON’ stands for ‘noncoding’; the following ‘MMU’ stands for ‘Mus musculus’; the next letter ‘T’ stands for ‘transcript’). Likewise, lncRNA genes were named sequentially with the middle letter ‘T’ replaced by ‘G’ representing ‘gene’.

Considering that the genomic context of lncRNAs may provide suggestions about their functional role, we subsequently classified lncRNA genes into the following biotypes according to their location with respect to protein-coding genes: (i) antisense, which have transcripts that intersect protein-coding genes on the opposite strand; (ii) intergenic, which are a subset of non-coding RNA loci located between protein-coding genes; (iii) sense exonic, which have transcripts that intersect protein-coding exons on the same strand; and (iv) sense non-exonic, which overlap with protein-coding genes in respect to transcription boundaries but not overlap in respect to processed exons.

Take mouse as an example, applying this categorization automatically to the lncRNA data set of NONCODE
version 4.0 results in the following distribution: antisense (6653), intergenic (19 067), sense exonic (12 111) and sense non-exonic (9312). See Figure 2 for further details on the lncRNA genes.

ncRNA ANNOTATION

One significant characteristic of NONCODE is its comprehensive annotation information. Each transcript in NONCODE is annotated with the following information: (i) basic description, including the ncRNA name, alias, sequence, length, genomic location, coding potential assessment by CPC (27) and CNCI, organisms and references; (ii) biological information, concerning its function, cellular role, cellular location and process function class (PfClass); and (iii) expression indication, including independent sources of multi-tissue expression profiles and potential function predicted based on a coding–non-coding co-expression network (28,29), especially for lncRNAs.

In this update, lncRNA genes are also annotated with two important features, as follows:

Presentation of lncRNA gene expression

We made full use of public RNA-seq data of human and mouse to provide indication of lncRNA functions. Human BodyMap 2.0 data (ENA archive: ERP000546) from human across 16 tissues and another RNA-seq data set from mouse across six different tissues (ENA archive: ERP000591) were downloaded. Cufflinks assembled transfrags from these raw RNA-seq for human and mouse, respectively. Cuffcompare then compared assembled transcripts with a reference annotation set composed of lncRNA genes from NONCODE version 4.0. At the same time, Cuffdiff calculated the FPKM of each reference lncRNA gene, representing expression level of it. The expression profile of each lncRNA gene across various tissue types is presented as a bar graph.

Assignment of lncRNA gene function

Functional predictions may guide and assist future investigations of lncRNAs. We applied lnc-GFP (30), a bi-colored network-based global function predictor, to the same RNA-seq data mentioned earlier in text to predict probable functions for lncRNA genes. A total of 20 100 lncRNA genes in NONCODE version 4.0 have been annotated with potential functions with a suitable parameter setting.

SERVICE UPDATE

The NONCODE database is based on MySQL and the Web site is powered by an Apache server. NONCODE has a user-friendly interface with a number of convenient browse and search options. Several useful services are available for users to access the NONCODE data, including BLAST, UCSC Genome Browser, SOAP API, DAS and an online submission system. UCSC Genome Browser has been updated in the new NONCODE version, whereas all other services are new additions. Furthermore, two online services have been added, which are the ID conversion tool and iLncRNA, the lncRNA identification pipeline.
different databases. In such situations, an ID conversion tool is necessary to facilitate more efficient user queries. For example, a transcript variant of the lncRNA gene termed HOTAIR (31) was assigned identifier ‘NR_003716’ in RefSeq. The ID conversion tool of NONCODE version 4.0 would recognize and convert it to ‘NONHSAT028508’. So far, NONCODE version 4.0 supports ID mapping between NONCODE identifiers and accessions from RefSeq and Ensembl.

iLncRNA

Owing to the development of next-generation sequencing technology, ncRNAs are now more easily and more accurately identified by sequencing transcriptomes (32). In this update, we provide iLncRNA an online pipeline for lncRNA identification based on assembled gtf files. As shown in Figure 3, transcript files in gff or gtf format either mapped and assembled from raw RNA-seq data or generated in other way are required as input for the identification pipeline. First, transcripts from input files <200 nt would be removed. The remaining transcripts were considered putative lncRNA transcripts, which were then subjected to the same filtration process comprises Cuffcompare and CNCI described earlier in text, to exclude potential protein-coding transcripts. Cuffcompare would be used once more to remove transcripts completely matching with pseudogenes from Ensembl. There would be an additional step to distinguish

![Figure 2](https://example.com/fig2.png)  
**Figure 2.** Details of lncRNA transcripts and genes. (A) Exon number distribution of human and mouse lncRNA transcripts. (B) Length distribution of human and mouse lncRNA transcripts. (C) Number of transcripts per gene for human and mouse. (D) Distribution of human and mouse lncRNA genes according to categorization.

![Figure 3](https://example.com/fig3.png)  
**Figure 3.** Pipeline for identification of lncRNAs for users. Refer to main text for details.
novel lncRNAs from known lncRNAs. Still by using Cuffcompare, lncRNAs not completely matching with collected lncRNAs from NONCODE would be defined as novel lncRNAs. Finally, predicted result would return to submitter of uploaded data after further validation. At the same time, this result would in turn be collected by NONCODE if authorized by the submitter.

**DISCUSSION**

The decreasing cost and improved capability of RNA-sequencing technology has lead to numerous transcriptome data from a variety of species. As a result of this, large numbers of ncRNAs are being rapidly identified and characterized (33). Due to this situation, we updated the NONCODE database to version 4.0, to keep track of newly identified ncRNAs. Of the newly collected data, lncRNAs constitute the majority. Particularly, lncRNAs were discovered to be involved in alternative splicing patterns that resemble mRNAs. Accumulating records of lncRNA transcripts makes it both possible and necessary to establish the concept of lncRNA genes. Consequently, NONCODE version 4.0 is a step toward a more integrated knowledge database with respect to definition and categorization of lncRNA genes.

**RNA-seq** is also increasingly being used for gene expression profiling. Through analysis of two sets of public RNA-seq data, NONCODE version 4.0 presents expression profiles of lncRNAs across different tissues from human and mouse, respectively, as bar graphs. Moreover, potential function of lncRNAs of human and mouse is inferred from the same data by lnc-GFP, a bi-colored network-based global function predictor.

Service improvements of NONCODE include not only updating existing tools such as BLAST and UCSC Genome Browser to the latest versions but also adding an ID conversion tool, enabling queries of accessions from different resources. To help users with their own RNA-seq data, we have also provided an online pipeline for lncRNA identification, which is named iLncRNA. User supplied files in gtf or gff format assembled from raw RNA-seq using TopHat (24) or other tools can be further analyzed by this pipeline. NONCODE version 4.0 in turn would consider collecting these data into the database. Moreover, with NPInter, the ncRNA interaction databases (34,35), cooperatting with our platform, NONCODE will stay as an informative and valuable data source for the study of lncRNAs.

**ACKNOWLEDGEMENT**

We thank Lu Yuyang for checking our annotation data. We also thank Andrew Plygawko for carefully reading our manuscript and Prof. Xiaohua Shen of Tsinghua University for help during the course of this work.

**FUNDING**

National High-tech Research and Development Projects 863 [2012AA020402, 2012AA022501], National Key Basic Research and Development Program 973 [2009CB825401], Training Program of the Major Research plan of the National Natural Science Foundation of China [91229120], National Natural Science Foundation of China [31371320]. Funding for open access charge: Training Program of the Major Research plan of the National Natural Science Foundation of China [91229120].

**Conflict of interest statement.** None declared.

**REFERENCES**


