NONCODE v2.0: decoding the non-coding

Shunmin He1,4, Changning Liu2, Geir Skogerbo1, Haitao Zhao3, Jie Wang1,4, Tao Liu1, Baoyan Bai1, Yi Zhao2 and Runsheng Chen1,2,*

1Bioinformatics Laboratory and National Laboratory of Biomacromolecules, Institute of Biophysics, 2Bioinformatics Research Group, Center for Advanced Computing Technology Research, Institute of Computing Technology, 3Department of Liver Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, CAMS & PUMC and 4Graduate School of the Chinese Academy of Sciences, Beijing, China

Received September 15, 2007; Revised October 23, 2007; Accepted October 24, 2007

ABSTRACT
The NONCODE database is an integrated knowledge database designed for the analysis of non-coding RNAs (ncRNAs). Since NONCODE was first released 3 years ago, the number of known ncRNAs has grown rapidly, and there is growing recognition that ncRNAs play important regulatory roles in most organisms. In the updated version of NONCODE (NONCODE v2.0), the number of collected ncRNAs has reached 206226, including a wide range of microRNAs, Piwi-interacting RNAs and mRNA-like ncRNAs. The improvements brought to the database include not only new and updated ncRNA data sets, but also an incorporation of BLAST alignment search service and access through our custom UCSC Genome Browser. NONCODE can be found under http://www.noncode.org or http://noncode.bioinfo.org.cn.

INTRODUCTION
The considerable number of non-coding RNAs (ncRNAs) that has been detected in the past few years was largely unexpected (1–3). Although the functions of the many recently identified ncRNAs remain mostly unknown, increasing evidence stands in support of the notion that ncRNAs represent a diverse and important functional output of most genomes (4). NONCODE is an integrated knowledge database dedicated to ncRNAs. All ncRNAs in NONCODE were filtered automatically from GenBank (5) and the literature, and were then later manually curated. With the exception of rRNAs and tRNAs, all classes of reported ncRNAs are included. The aim of the database is to provide a platform that will facilitate both bioinformatic as well as experimental research. In addition to containing sequence data, NONCODE provides a user-friendly interface, a visualization platform and a convenient search option, allowing efficient recovery of sequences, regulatory elements in the flanking sequences, related publications and other information.

DATA COLLECTION AND ANNOTATION
Data collection and annotation for NONCODE v2.0 was carried out in a similar fashion as for version 1.0 and can be briefly described as follows: GenBank entries constituted the major source of NONCODE. We searched PubMed (6) with a list of ncRNA keywords, such as ‘ncRNA’, ‘snoRNA’, ‘snRNA’, ‘tmRNA’, ‘SRP RNA’, ‘gRNA’, etc., and thereafter consulted the literature matched with them and extracted more ncRNA keywords. The downloaded GenBank files (gbfiles) were then filtered using these keywords, and the filtered entries were subsequently confirmed by manual curation. For all obtained ncRNA records, basic information related to sequence, name, alias, length, ncRNA class, organism, references and accession number in GenBank were extracted and entered into the NONCODE database. Each ncRNA sequence was checked for redundancies using Perl scripts, and each cluster of redundant sequences was given a non-redundant NONCODE accession number (UniqID, i.e. unique ncRNA i.d.). In addition to the ‘traditional’ ncRNA classification system, NONCODE v1.0 introduced the alternative ‘process function class (PfClass)’ system based on the biological processes or functions in which an ncRNA is involved, and one or more of the 26 PfClasses were also assigned to all ncRNAs in NONCODE v2.0. Moreover, a subset of ncRNAs has been divided into nine additional categories according to whether they are gender- or tissue-specific or associated with tumors and diseases, etc. Where possible, NONCODE also provides additional annotations, such as information on function, cellular role, cellular location, chromosomal localization and splicing. The annotations

*To whom correspondence should be addressed. Tel: +86 10 6488 8543; Fax: +86 10 6487 7837; Email: crs@sun5.ibp.ac.cn
Correspondence may also be addressed to Yi Zhao. Tel: +86 10 6260 0564; Fax: +86 10 6260 0602; Email: biozy@ict.ac.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
and the genomic mapping information of the sequences rely on data provided in the original GenBank records, the FANTOM3 Database (2), the UCSC Genome Browser Database (7), or directly from the reference literature.

DATABASE CONTENT AND CLASSIFICATION
The purpose of the database is to serve the research community by organizing information concerning all types of ncRNAs (except tRNAs and rRNAs) from all groups of organisms. As of August 2007, the NONCODE database includes over 206,226 non-redundant sequences from 861 organisms. The significant growth in the amount of data, compared with the 5339 non-redundant sequences in the previous edition published in 2005, is primarily due to systematic identification of mRNA-like ncRNA transcripts (2) and the discovery of Piwi-interacting RNAs (piRNAs) through large-scale cDNA sequencing (1,3,8). Other novel ncRNAs, such as stem-bulge RNAs (sbRNAs) (9), snRNA-like RNAs (snlRNAs) (9) and a number of unclassified ncRNA transcripts were mainly obtained from our laboratory and other published literature (10–12). According to the traditional classification system, NONCODE v2.0 contains three novel classes of ncRNAs, the sbRNAs, the snlRNAs and the piRNAs, whereas the number of PfClasses is the same as in NONCODE v1.0 (i.e. 26), with sbRNAs and snlRNAs corresponding to the ‘Miscfunction_snm’ and piRNAs to ‘RNA-processing_cleavage’ PfClass.

DATABASE ACCESS
All sequences can be directly downloaded from the webpage. Sequences can be searched using accession numbers found in GenBank, name, traditional class, PfClass, organism and UniqID in NONCODE. In addition to access to NONCODE database records, search results are also linked to full GenBank entries (Figure 1).
Saccharomyces cerevisiae, Caenorhabditis elegans and Homo sapiens. NcRNA loci of these species may be viewed through the NONCODE track in the Genome Browser. Other common tracks concerning basic information on these species, such as mRNA genes, ESTs and so on, have also been retrieved from the UCSC Genome Browser Database. For the above three species, ncRNA entries in the NONCODE database can be directly linked to the Genome Browser; similarly, NONCODE ncRNA annotations may be accessed through the Genome Browser (Figure 1). The database can be accessed through the following URL: http://www.noncode.org/ or http://noncode.bioinfo.org.cn.

FUTURE DIRECTIONS

As new ncRNAs are being progressively discovered, we will continue to update the NONCODE database. Submissions of new ncRNAs are invited, and should be sent to noncode@ict.ac.cn. Within the coming year, we will continue to add Genome Browser services for other model organisms, such as mouse and fly. Given the increasing amount of ncRNA data and the emergence of ncRNA prediction software [e.g. QRNA (13), RNAz (14)], we will attempt to establish a service for ncRNA prediction based on the mentioned softwares and the information in the NONCODE database.

ACKNOWLEDGEMENTS

Sequence data were downloaded from NCBI GenBank (ftp://ftp.ncbi.nih.gov/genbank). The authors thank Lisa Caviglia for careful corrections. This work was supported by the National Key Basic Research & Development Program (973), under the Grant Nos. 2002CB713805 and 2003CB715907, the National Sciences Foundation of China, under Grant Nos. 30630040, 30570393 and 30600729, the Data Sharing Network of China Essential Medicine Science, under Grant No. 2005DKA32402. Funding to pay the Open Access publication charges for this article was provided by the National Sciences Foundation of China, under Grant No.30570393.

Conflict of interest statement. None declared.

REFERENCES