D170–D172 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 13 November 2007*
doi:10.1093/nar/gkm1011

NONCODE v2.0: decoding the non-coding

Shunmin He[1,4], Changning Liu[2], Geir Skogerbø[1], Haitao Zhao[3], Jie Wang[1,4], Tao Liu[1], Baoyan Bai[1], Yi Zhao[2] and Runsheng Chen[1,2,*]

[1]Bioinformatics Laboratory and National Laboratory of Biomacromolecules, Institute of Biophysics, [2]Bioinformatics Research Group, Center for Advanced Computing Technology Research, Institute of Computing Technology, [3]Department of Liver Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, CAMS & PUMC and [4]Graduate 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 206 226, 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.

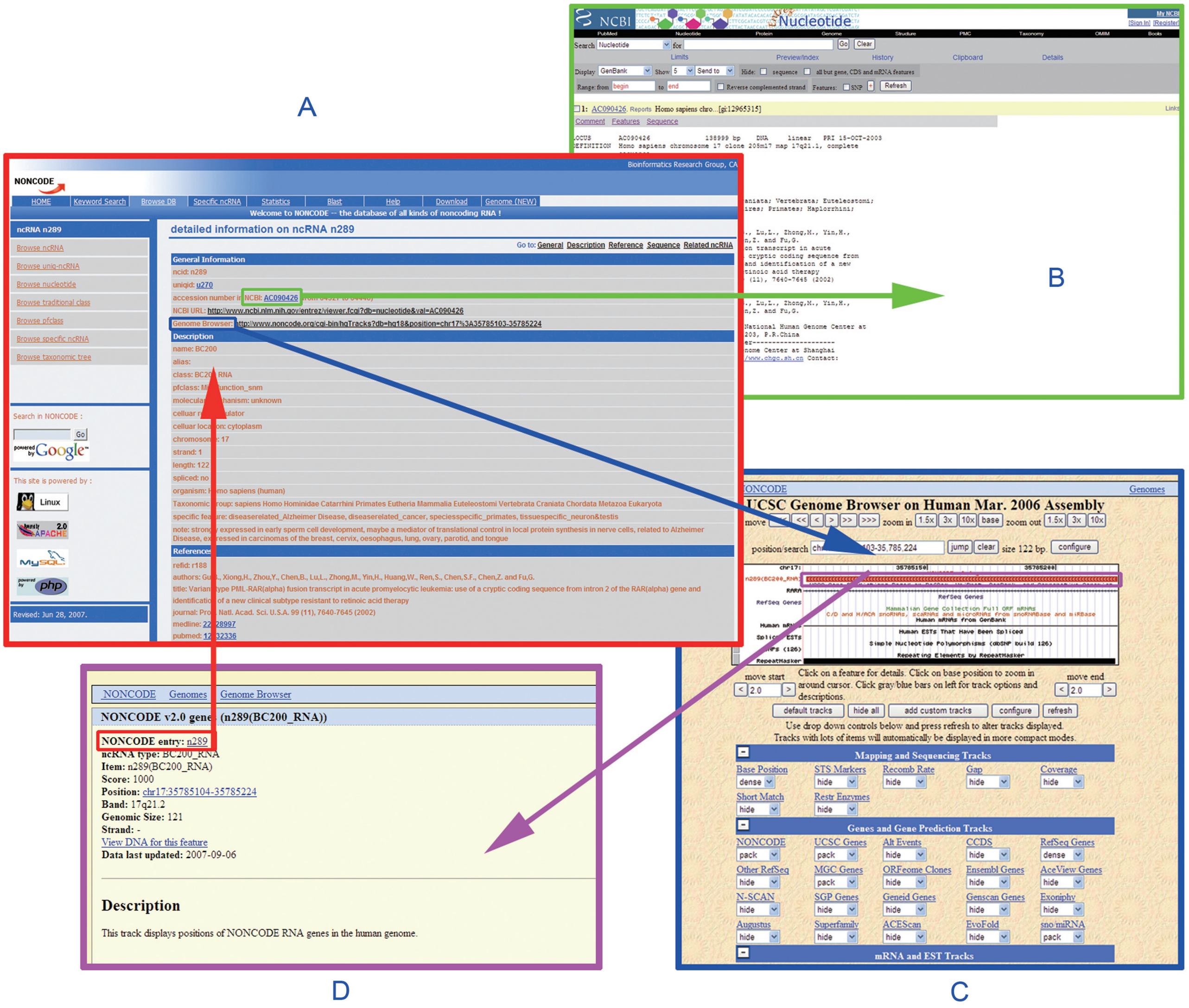
and the genomic mapping information of the sequences rely on data provided in the original GenBank records, the FANTOM3 Database (2), the UCSC Genome Browser Database (7), or directly from the reference literature.

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). In the



Welcome to NONCODE -- the database of all kinds of noncoding RNA !

detailed information on ncRNA n289

General Information

ncid: n289
uniqid: u270
accession number in NCBI: AC090426 from 64327 to 64448)
NCBI URL: http://www.ncbi.nlm.nih.gov/entrez/viewer.fcgi?db=nucleotide&val=AC090426
Genome Browser: http://www.noncode.org/cgi-bin/hgTracks?db=hg18&position=chr17%3A35785103-35785224

Description

name: BC200
alias:
class: BC200 RNA
pfclass: Miscfunction_snm
molecular mechanism: unknown
celluar regulator
celluar location: cytoplasm
chromosome: 17
strand: 1
length: 122
spliced: no
organism: Homo sapiens (human)
Taxonomic group: sapiens Homo Hominidae Catarrhini Primates Eutheria Mammalia Euteleostomi Vertebrata Craniata Chordata Metazoa Eukaryota
specific feature: diseaserelated_Alzheimer Disease, diseaserelated_cancer, speciesspecific_primates, tissuespecific_neuron&testis
note: strongly expressed in early sperm cell development, maybe a mediator of translational control in local protein synthesis in nerve cells, related to Alzheimer Disease, expressed in carcinomas of the breast, cervix, oesophagus, lung, ovary, parotid, and tongue

References

refid: r188
authors: Gu,B., Xiong,H., Zhou,Y., Chen,B., Lu,L., Zhong,M., Yin,H., Huang,W., Ren,S., Chen,S.F., Chen,Z. and Fu,G.
title: Variant-type PML-RAR(alpha) fusion transcript in acute promyelocytic leukemia: use of a cryptic coding sequence from intron 2 of the RAR(alpha) gene and identification of a new clinical subtype resistant to retinoic acid therapy
journal: Proc. Natl. Acad. Sci. U.S.A. 99 (11), 7640-7645 (2002)
medline: 22028997
pubmed: 12032336

Revised: Jun 28, 2007.

Search: Nucleotide
Display: GenBank Show: 5
Range: from begin to end

1: AC090426. Reports Homo sapiens chro...[gi:12965315]

```
LOCUS       AC090426             138999 bp    DNA     linear   PRI 15-OCT-2003
DEFINITION  Homo sapiens chromosome 17 clone 205m17 map 17q21.1, complete
aniata; Vertebrata; Euteleostomi;
ires; Primates; Haplorrhini;
., Lu,L., Zhong,M., Yin,H.,
n,Z. and Fu,G.
on transcript in acute
 cryptic coding sequence from
and identification of a new
tinoic acid therapy
 (11), 7640-7645 (2002)
., Lu,L., Zhong,M., Yin,H.,
n,Z. and Fu,G.
National Human Genome Center at
203, P.R.China
er---------------------
nome Center at Shanghai
/www.chgc.sh.cn Contact:
```

UCSC Genome Browser on Human Mar. 2006 Assembly

position/search: ch…103-35,785,224 size 122 bp.

RefSeq Genes
Mammalian Gene Collection Full ORF mRNAs
C/D and H/ACA snoRNAs, scaRNAs and microRNAs from snoRNABase and miRBase
Human mRNAs from GenBank
Human ESTs That Have Been Spliced
Simple Nucleotide Polymorphisms (dbSNP build 126)
Repeating Elements by RepeatMasker

move start: 2.0 move end: 2.0

Click on a feature for details. Click on base position to zoom in around cursor. Click gray/blue bars on left for track options and descriptions.

Use drop down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes.

Mapping and Sequencing Tracks

Base Position: dense	STS Markers: hide	Recomb Rate: hide	Gap: hide	Coverage: hide
Short Match: hide	Restr Enzymes: hide			

Genes and Gene Prediction Tracks

NONCODE: pack	UCSC Genes: pack	Alt Events: hide	CCDS: hide	RefSeq Genes: dense
Other RefSeq: hide	MGC Genes: pack	ORFeome Clones: hide	Ensembl Genes: hide	AceView Genes: hide
N-SCAN: hide	SGP Genes: hide	Geneid Genes: hide	Genscan Genes: hide	Exoniphy: hide
Augustus: hide	Superfamily: hide	ACEScan: hide	EvoFold: hide	sno/miRNA: pack

mRNA and EST Tracks

NONCODE v2.0 genes (n289(BC200_RNA))

NONCODE entry: n289
ncRNA type: BC200_RNA
Item: n289(BC200_RNA)
Score: 1000
Position: chr17:35785104-35785224
Band: 17q21.2
Genomic Size: 121
Strand: -
View DNA for this feature
Data last updated: 2007-09-06

Description

This track displays positions of NONCODE RNA genes in the human genome.

Figure 1. Links between the NONCODE ncRNA annotations, the Genome Browser and NCBI. (**A**) The NONCODE database window with ncRNA annotations. (**B**) The corresponding NCBI annotation. (**C**) The corresponding Genome Browser window. (**D**) The link from Genome Browser to NONCODE.

current version of the database, we also included the online BLAST service (NCBI wwwBLAST version 2.2.17) which allows sequence similarity searches against the entire NONCODE v2.0 database.

In this updated version of NONCODE, a UCSC Genome Browser for NONCODE was constructed for *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

1. Lau,N.C., Seto,A.G., Kim,J., Kuramochi-Miyagawa,S., Nakano,T., Bartel,D.P. and Kingston,R.E. (2006) Characterization of the piRNA complex from rat testes. *Science*, **313**, 363–367.
2. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B. *et al*. (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
3. Girard,A., Sachidanandam,R., Hannon,G.J. and Carmell,M.A. (2006) A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*, **442**, 199–202.
4. Mattick,J.S. and Makunin,I.V. (2006) Non-coding RNA. *Hum. Mol. Genet*., **15**, R17–R29.
5. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2007) GenBank. *Nucleic Acids Res*., **35**, D21–D25.
6. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al*. (2007) Database resources of the national center for biotechnology information. *Nucleic Acids Res*., **35**, D5–D12.
7. Kuhn,R.M., Karolchik,D., Zweig,A.S., Trumbower,H., Thomas,D.J., Thakkapallayil,A., Sugnet,C.W., Stanke,M., Smith,K.E. *et al*. (2007) The UCSC genome browser database: update 2007. *Nucleic Acids Res*., **35**, D668–D673.
8. Aravin,A., Gaidatzis,D., Pfeffer,S., Lagos-Quintana,M., Landgraf,P., Iovino,N., Morris,P., Brownstein,M.J., Kuramochi-Miyagawa,S. *et al*. (2006) A novel class of small RNAs bind to MILI protein in mouse testes. *Nature*, **442**, 203–207.
9. Deng,W., Zhu,X., Skogerbo,G., Zhao,Y., Fu,Z., Wang,Y., He,H., Cai,L., Sun,H. *et al*. (2006) Organization of the *Caenorhabditis elegans* small non-coding transcriptome: genomic features, biogenesis, and expression. *Genome. Res*., **16**, 20–29.
10. Huang,Z.P., Chen,C.J., Zhou,H., Li,B.B. and Qu,L.H. (2007) A combined computational and experimental analysis of two families of snoRNA genes from *Caenorhabditis elegans*, revealing the expression and evolution pattern of snoRNAs in nematodes. *Genomics*, **89**, 490–501.
11. Zemann,A., op de Bekke,A., Kiefmann,M., Brosius,J. and Schmitz,J. (2006) Evolution of small nucleolar RNAs in nematodes. *Nucleic Acids Res*., **34**, 2676–2685.
12. Xie,Z., Allen,E., Fahlgren,N., Calamar,A., Givan,S.A. and Carrington,J.C. (2005) Expression of Arabidopsis MIRNA genes. *Plant Physiol*., **138**, 2145–2154.
13. Rivas,E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8.
14. Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.