D210–D215 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 1 December 2011*
doi:10.1093/nar/gkr1175

NONCODE v3.0: integrative annotation of long noncoding RNAs

Dechao Bu[1,2], Kuntao Yu[1,2], Silong Sun[1], Chaoyong Xie[1,2], Geir Skogerbø[3], Ruoyu Miao[1,4], Hui Xiao[1], Qi Liao[1], Haitao Luo[1], Guoguang Zhao[1,2], Haitao Zhao[4], Zhiyong Liu[1], Changning Liu[1], Runsheng Chen[3,*] and Yi Zhao[1,*]

[1]Bioinformatics Research Group, Key Laboratory of Intelligent Information Processing, Advanced Computer Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, PR China, [2]Graduate School of the Chinese Academy of Sciences, Beijing, PR China, [3]Bioinformatics Laboratory and National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing, PR China and [4]Department of Liver Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, CAMS & PUMC, Beijing 100730, China

Received September 15, 2011; Revised and Accepted November 13, 2011

ABSTRACT

Facilitated by the rapid progress of high-throughput sequencing technology, a large number of long noncoding RNAs (lncRNAs) have been identified in mammalian transcriptomes over the past few years. LncRNAs have been shown to play key roles in various biological processes such as imprinting control, circuitry controlling pluripotency and differentiation, immune responses and chromosome dynamics. Notably, a growing number of lncRNAs have been implicated in disease etiology. With the increasing number of published lncRNA studies, the experimental data on lncRNAs (e.g. expression profiles, molecular features and biological functions) have accumulated rapidly. In order to enable a systematic compilation and integration of this information, we have updated the NONCODE database (http://www.noncode.org) to version 3.0 to include the first integrated collection of expression and functional lncRNA data obtained from re-annotated microarray studies in a single database. NONCODE has a user-friendly interface with a variety of search or browse options, a local Genome Browser for visualization and a BLAST server for sequence-alignment search. In addition, NONCODE provides a platform for the ongoing collation of ncRNAs reported in the literature. All data in NONCODE are open to users, and can be downloaded through the website or obtained through the SOAP API and DAS services.

INTRODUCTION

Long noncoding RNAs (lncRNAs) were first characterized as mRNA-like noncoding RNAs in that they undergo splicing and have features such as a poly(A) signal/tail (1), while an arbitrary criterion of 'transcripts longer than 200 nucleotides' has later been added to its 'definition' (2,3). With the development of experimental technology, especially the high-throughput sequencing methods, and further advancement of computational prediction algorithm, an increasing number of lncRNAs is being identified in mammals. For example, thousands of conserved large intervening (or intergenic) noncoding RNAs (lincRNAs) were discovered in human and mouse by using chromatin signature analysis (4–6). Computational methods including ORF-Predictor and BLASTP pipeline (7) identified 5446 lncRNAs in the human genome, 1859 lncRNAs were found throughout the human genome by high-throughput sequencing across a prostate cancer cohort (8) and a reference catalog of 8195 lincRNAs were founded using RNA-seq data collected from ~4 billion RNA-seq reads across 24 human tissues and cell types (9). The functional properties of the lncRNAs are also rapidly being revealed. lncRNAs have already been shown to play key roles in imprinting control, circuitries controlling pluripotency and differentiation, immune responses, chromosome dynamics and human diseases (2,10). In parallel, the lncRNA studies

*To whom correspondence should be addressed. Tel: +86 10 62601010; Fax: +86 10 62601356; Email: biozy@ict.ac.cn
Correspondence may also be addressed to Runsheng Chen. Tel: +86 10 6488 8543; Fax: +86 10 6487 7837; Email: crs@sun5.ibp.ac.cn

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

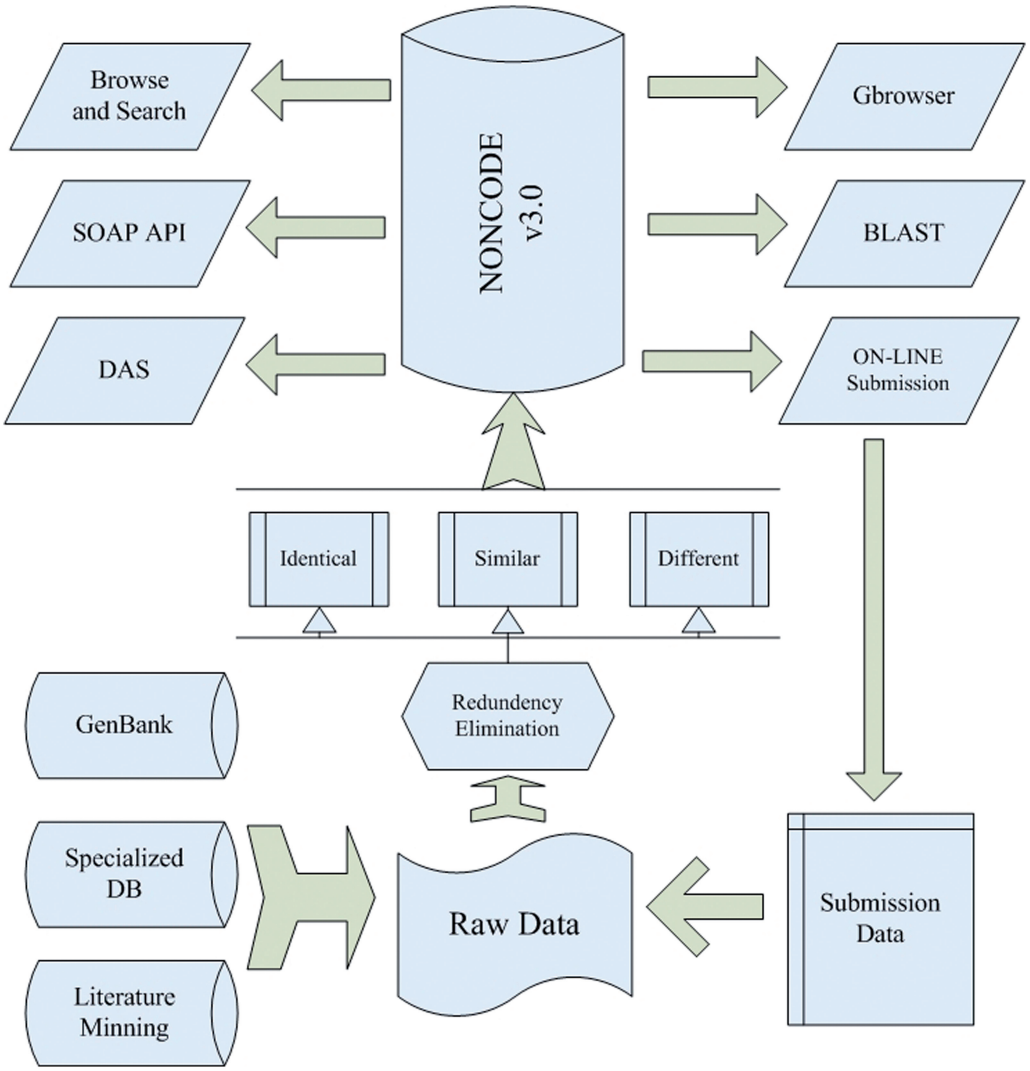

Figure 1. Overview of the NONCODE v3.0 Database. Raw data were mainly obtained from three types of sources: GenBank, specialized databases and literature. Sequences from different sources first go through redundancy elimination and are then included in the database. The ncRNAs in NONCODE can be accessed and analyzed by various tools and services, including a variety of search or browse options, a local Genome Browser for visualization, and a BLAST server for sequence-alignment search. ncRNA sequences can be download directly from the website or accessed through the SOAP API or DAS servers. In addition, an on-line submission system is provided for continuous collection of new ncRNAs.

have led to an accumulating amount of experimental data, such as expression profiles (11) and information on lncRNA functions in a variety of biological processes (3,12).

In order to compile this information and establish a comprehensive and systematic database, and thereby facilitating further exploration of the molecular mechanisms of lncRNAs, we have updated the NONCODE database to version 3.0 (NONCODE v3.0). As a first, NONCODE v3.0 now also includes expressional and functional lncRNA data (13,14) obtained from re-annotated microarray studies. At the same time, other classes of ncRNAs have also been updated. The number of ncRNA entries has been more than doubled since NONCODE v2.0, increasing from 206 226 to 411 552. Other improvements include upgraded BLAST and UCSC Genome Browser functions, and incorporation of SOAP API and DAS services, which will simplify queries, visualization and access to the large amounts of data in NONCODE v3.0. To simplify a continuous update of the data, an online submission system for new ncRNAs has also been provided. An overview of NONCODE v3.0 is shown in Figure 1. In conclusion, the aim of the NONCODE database is to provide a user-friendly web interface to browse, search, retrieve and update information on ncRNAs, and to facilitate further research of ncRNAs, gene networks and functional genomics. The NONCODE v3.0 database is freely accessible at http://www.noncode.org.

DATA COLLECTION

NONCODE v3.0 has, as far as possible, collected all published ncRNAs that have been experimentally verified or

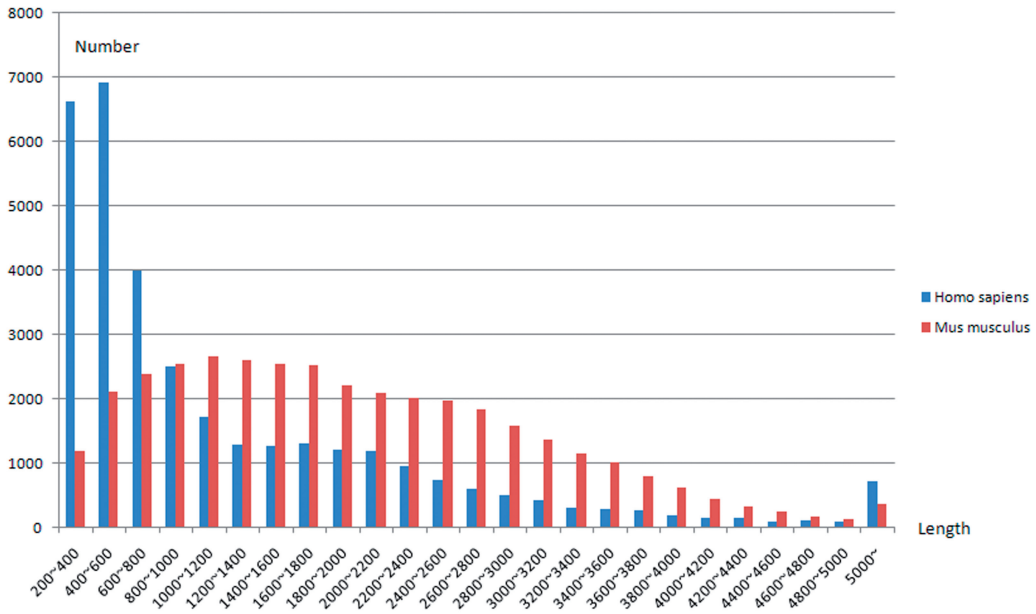

Figure 2. Length distribution of human and mouse lncRNAs.

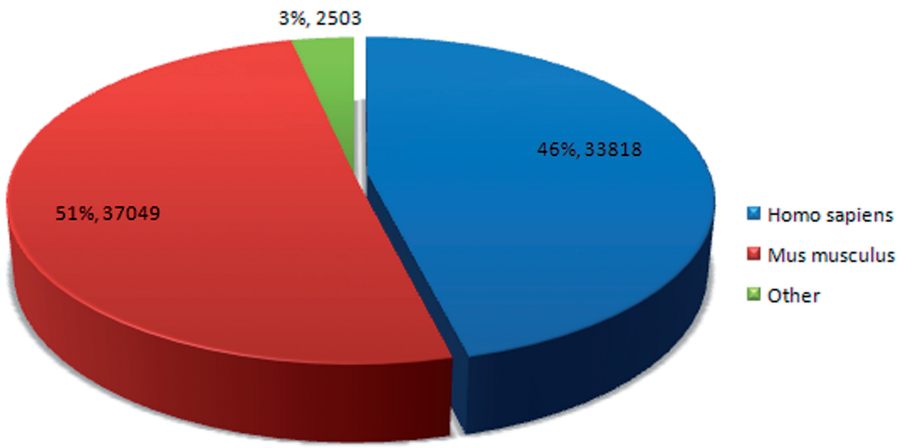

Figure 3. Distribution of the lncRNAs on organisms.

identified by computational methods. It presently contains 411 552 public sequences distributed on 134 ncRNA classes and 26 cellular processes. The 411 552 ncRNA entries in the database are collected from 1239 different organisms, and 73 370 of the entries represent lncRNAs, covering nearly all published human and mouse lncRNAs (see Figures 2 and 3 for further details on the lncRNAs). In other words, NONCODE is one of the most comprehensive and systematic ncRNA databases. The new data included in NONCODE v3.0 have mainly been obtained from the following three types of sources.

GenBank

We extracted ncRNAs from GenBank with the keywords 'ncRNA', 'miRNA', 'piRNA', 'snRNA', 'snoRNA', 'snmRNA', 'tmRNA', 'SRP RNA' or 'gRNA' by using the pipeline built in NONCODE v1.0/v2.0 (15,16). Altogether 8954 unique new ncRNAs which had not been collected by other databases were obtained and labeled as 'from GenBank'.

Specialized databases

The latest releases of a number of well-known databases were used as the sources for NONCODE: RNAdb v2.0 (17), fRNAdb v3.0 (18), H-InvDB v7.5 (19), FANTOM3 (20), lncRNAdb (21), miRBase v17.0 (22), RefSeq (23), UCSC (24) and Ensembl (25). The ncRNAs were first extracted from these databases, then passed through the redundancy elimination operation (below), and finally entered into NONCODE.

Literature sources

Several new types of ncRNAs have recently been reported in the literature [e.g. lincRNAs (4–6,9) and eRNAs (26)], but have not yet been included by any of the specialized databases. We therefore constructed a pipeline for mining the literature of recently published ncRNAs. We first used EFetch to retrieve literature published since 1 January 2009, from PubMed, employing the key words 'ncRNA', 'noncoding', 'non-coding', 'noncode' or 'non-code', and got 2605 relevant articles. After manually selecting reports on new ncRNAs, we retrieved sequences, genome locations and other relevant information concerning these transcripts. This resulted in 20 252 new ncRNAs being entered into the database.

REDUNDANCY ELIMINATION

The data sources mentioned above necessarily contain a varying extent of overlapping data, and a step to eliminate redundancies across different sources is necessary. To decide whether two primary entries might represent the same ncRNA, we took into account their accession numbers, organism information and sequence similarity. The latter was measured by the identity, e-value and overlap-ratio as returned by Blast alignments. The overlap-ratio for each entry was calculated as the proportion of the length of matched sequence compared with the whole length of sequence, and the overlap-ratio of both entries was taken into consideration. According to the above information, two entries derived from the same organism fell into one of three categories: (i) Identical entries. These are entries with identical sequences and genomic locations, and with non-conflicting accession numbers ('non-conflicting accession numbers' refer to cases when both entries have the same accession number or at least one entry does not have an accession number). (ii) Similar entries. These are entries with similar sequence information (overlap-ratio>0.8, identity>0.8, e-value<1e-10). (iii) Different ncRNAs. These are entries that do not fall into categories (i) and (ii). The 'identical entries' were finally integrated as one record in NONCODE, whereas the 'similar ncRNAs' were all retained, but assigned with the same 'uniqID' to indicate their relationship. After the redundancy elimination step, a total of 411 552 ncRNAs were finally recorded in NONCODE v3.0.

ncRNA ANNOTATION

One significant characteristic of NONCODE is its comprehensive annotation information. Each sequence in NONCODE is annotated with (i) basic information including the ncRNA name, alias, sequence, length, organisms, references etc. and (ii) additional information concerns its function, cellular role, cellular location and process function class (PfClass). In this update, four important attributes, two of which are dedicated to lncRNAs, have been added as follows:

Coding potential assessment

Since not all published ncRNAs have undergone detailed experimental analysis, we calculated a coding potential calculator score [CPC score (27)] and Coding Non-Coding Index (CNCI (software in-house)) for each ncRNA to evaluate its coding potential. This will enable the user to quickly identify transcripts whose coding potential may need further scrutiny.

Mapping information

For most ncRNAs, we have collected its mapping information from its original source. The remaining ncRNAs have been mapped to its reference genome using BLAT, the top one hit with >99% match to the reference genome being retained as its 'locus'. The mapped locations can be viewed in the UCSC Genome Browser.

Expression profiles

Three independent sources of multi-tissue expression profiles have been included to facilitate the functional study of 27 408 lncRNAs. These are the FANTOM customer-designed microarray data which contain the expression profile of 10 874 mouse lncRNAs across 20 tissues (28), the re-annotated expression profiles of Affymetrix arrays (13), which contain expression profiles of 343 human lncRNAs across 65 tissues, and 4075 mouse lncRNAs across 22 tissues, and 13 565 human lincRNAs expression profiles from RNA-seq data across 22 tissues and cell lines (9). As more re-annotated or other lncRNA microarray data, along with RNA-seq data, are made available, these will be integrated within NONCODE.

Potential functions

Functional predictions may guide and assist future investigations of lncRNAs. A total of 1635 lncRNAs have been annotated with potential functions that have been predicted based on a Coding-Noncoding co-expression network (13,14). The estimated 'quality' of each functional prediction is indicated by a *P*-value.

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. BLAST and UCSC Genome Browser have been upgraded in the new NONCODE version, while other three services are new additions.

Browse and search

Two browse options, 'Browse by expression profile' and 'Browse by functional prediction', have been added to the new NONCODE version. These ensure rapid access to the expression profiles or to information on potential functions of the ncRNAs. Search by GO term functional

keywords is also supported. All browse and search results can be exported instantly from the query page. Besides, searching results can be filtered by species (human or mouse) and transcript length (more or less than 200 nt). These options will render browsing and searching more convenient for the user.

SOAP API service

Simple Object Access Protocol (SOAP) is a protocol specification used for exchanging structured information between client and server computers in the implementation of Web Services. NONCODE now provides a SOAP API that can be easily accessed for custom query. Users could get their query results by writing short codes that calls six SOAP query functions, including ncRNADetails(), QueryByRNA(), QueryByClass(), QueryByReference(), QueryByNucleotide() and QueryByLength(). The SOAP API service in NONCODE can be accessed via the following URL: http://www.noncode.org/soapApi.html. No installation of any program or package is needed to use the functions.

DAS service

Distributed Annotation System (DAS) allows sequence annotations to be decentralized among multiple third-party annotators and integrated on an as-needed basis on the client side (29), which facilitates integration and collation of ncRNA annotations from multiple servers. The DAS service is now available in NONCODE. It provides access to all annotation data for current assemblies featured in NONCODE, and can be visited via: http://www.noncode.org/das.html. Several examples have been illustrated to guide construction of DAS queries and fetch NONCODE tracks through the DAS server.

Online submission of new ncRNAs

In order to maintain an up-to-date and comprehensive resource, we encourage users to submit their own data to NONCODE. The submission page offers three different submission options: (i) if the data have already been submitted to NCBI, the user can just submit the NCBI accession number to us; otherwise, (ii) if the data are small, the user can paste them into a text box in FASTA format; or (iii) if the amount of data is large, the user can upload a FASTA format file to our server. In order to ensure data quality, we recommend users provide their names and email addresses. Email is especially necessary for quick and convenient communication.

CONCLUSION

Compared with the previous versions of NONCODE, version 3.0 is a step towards a more integrated knowledge database, particularly with respect to lncRNAs. The total number of ncRNAs and the number of lncRNAs, functional annotations and on-line services have all been expanded (shown in Table 1). Beyond mere sequence information, NONCODE V3.0 also integrates various kinds of informative content, such as genome context, process function, coding potential score, re-annotated expression data, potential functions etc. NONCODE is designed to enable integration with other resources, including the UCSC Genome Browser, GenBank and other databases, and NONCODE thus provides a location from which researchers can obtain a wide range of information regarding their genes of interest. Although currently the expression data and annotations of predicted functions are only integrated with a small portion of the lncRNA entries in NONCODE, we expect this to increase as more data are published.

Table 1. Comparison of NONCODE v1.0, v2.0 and v3.0

Version	Total ncRNA number	lncRNA number	Functional annotation, etc.	Services
1.0	5339	1557	PfClass	Browse, Search, Download
2.0	206 226	35 805	PfClass	Browse, Search, Download, Blast, Genome Browser
3.0	411 552	73 370	PfClass Expression profile, predicted functions	Browse, Search, Download, Blast, Genome Browser, Soap API, DAS, On-line Submission

The decreasing cost and improved depth of the RNA-sequencing technology have already enabled numerous transcriptome studies in a variety of species. As a result of this, it is expected that huge numbers of lncRNAs will be rapidly identified and characterized in the near future. NONOCDE will continue to keep track of and promptly collect these data into the database. Also, the central role of lncRNAs in the molecular etiology of complex diseases, such as cancer, will make them a persistent research hotspot. Therefore, we expect that NONCODE will stay as an informative and valuable data source on the biological roles of lncRNAs for the scientific community.

FUNDING

National Natural Science Foundation of China (No. 31071137, 31000586, 30970623); National Key Basic Research and Development Program (973) under (Grant Nos.0997011001); Knowledge Innovation Program of the Chinese Academy of Sciences (KSCX2-EW-R-01,KSCX2-EW-R-0102); The Natural Science Foundation of Jiangsu province (BK2008231); Sci-tech Innovation Team of Jiangsu University (2008-018-02). Funding for open access charge: Knowledge Innovation Program of the Chinese Academy of Sciences (KSCX2-EW-R-01).

Conflict of interest statement. None declared.

REFERENCES

1. Erdmann,V.A., Szymanski,M., Hochberg,A., de Groot,N. and Barciszewski,J. (1999) Collection of mRNA-like non-coding RNAs. *Nucleic Acids Res.*, **27**, 192–195.

2. Mercer,T.R., Dinger,M.E. and Mattick,J.S. (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, **10**, 155–159.
3. Nagano,T. and Fraser,P. (2011) No-nonsense functions for long noncoding RNAs. *Cell*, **145**, 178–181.
4. Mitchell Guttman,I.A., Garber,M., French,C., Lin,M.F., Feldser,D., Huarte,M., Zuk,O., Carey,B.W., Cassady,J.P. and Cabili,M.N. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.
5. Khalil,A.M., Guttman,M., Huarte,M., Garber,M., Raj,A., Rivea Morales,D., Thomas,K., Presser,A., Bernstein,B.E. and Van Oudenaarden,A. (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl Acad. Sci. USA*, **106**, 11667–11672.
6. Guttman,M., Garber,M., Levin,J.Z., Donaghey,J., Robinson,J., Adiconis,X., Fan,L., Koziol,M.J., Gnirke,A. and Nusbaum,C. (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, **28**, 503–510.
7. Jia,H., Osak,M., Bogu,G.K., Stanton,L.W., Johnson,R. and Lipovich,L. (2010) Genome-wide computational identification and manual annotation of human long noncoding RNA genes. *RNA*, **16**, 1478–1487.
8. Prensner,J.R., Iyer,M.K., Balbin,O.A., Dhanasekaran,S.M., Cao,Q., Brenner,J.C., Laxman,B., Asangani,I.A., Grasso,C.S. and Kominsky,H.D. (2011) Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat. Biotechnol.*, **29**, 742–749.
9. Cabili,M.N., Trapnell,C., Goff,L., Koziol,M., Tazon-Vega,B., Regev,A. and Rinn,J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.*, **25**, 1915–1927.
10. Taft,R.J., Pang,K.C., Mercer,T.R., Dinger,M. and Mattick,J.S. (2010) Non-coding RNAs: regulators of disease. *J. Pathol.*, **220**, 126–139.
11. Dinger,M.E., Pang,K.C., Mercer,T.R., Crowe,M.L., Grimmond,S.M. and Mattick,J.S. (2009) NRED: a database of long noncoding RNA expression. *Nucleic Acids Res.*, **37**, D122–D126.
12. Guttman,M., Donaghey,J., Carey,B.W., Garber,M., Grenier,J.K., Munson,G., Young,G., Lucas,A.B., Ach,R. and Bruhn,L. (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*, **477**, 295–300.
13. Liao,Q., Liu,C., Yuan,X., Kang,S., Miao,R., Xiao,H., Zhao,G., Luo,H., Bu,D. and Zhao,H. (2011) Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic Acids Res.*, **39**, 3864–3878.
14. Liao,Q., Xiao,H., Bu,D., Xie,C., Miao,R., Luo,H., Zhao,G., Yu,K., Zhao,H. and Skogerb,G. (2011) ncFANs: a web server for functional annotation of long non-coding RNAs. *Nucleic Acids Res.*, **39**, W118–W124.
15. Liu,C., Bai,B., Skogerb,G., Cai,L., Deng,W., Zhang,Y., Bu,D., Zhao,Y. and Chen,R. (2005) NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Res.*, **33**, D112–D115.
16. He,S., Liu,C., Skogerb,G., Zhao,H., Wang,J., Liu,T., Bai,B., Zhao,Y. and Chen,R. (2008) NONCODE v2. 0: decoding the non-coding. *Nucleic Acids Res.*, **36**, D170–D172.
17. Pang,K.C., Stephen,S., Dinger,M.E., Engström,P.G., Lenhard,B. and Mattick,J.S. (2006) RNAdb 2.0—an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res.*, **35**, D178–D182.
18. Mituyama,T., Yamada,K., Hattori,E., Okida,H., Ono,Y., Terai,G., Yoshizawa,A., Komori,T. and Asai,K. (2009) The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Res.*, **37**, D89–D92.
19. Yamasaki,C., Murakami,K., Takeda,J., Sato,Y., Noda,A., Sakate,R., Habara,T., Nakaoka,H., Todokoro,F. and Matsuya,A. (2010) H-InvDB in 2009: extended database and data mining resources for human genes and transcripts. *Nucleic Acids Res.*, **38**, D626–D632.
20. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B. and Wells,C. (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
21. Amaral,P.P., Clark,M.B., Gascoigne,D.K., Dinger,M.E. and Mattick,J.S. (2011) lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.*, **39**, D146–D151.
22. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
23. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.
24. Fujita,P.A., Rhead,B., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Cline,M.S., Goldman,M., Barber,G.P., Clawson,H. and Coelho,A. (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.
25. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S. and Fitzgerald,S. (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.
26. Kim,T.K., Hemberg,M., Gray,J.M., Costa,A.M., Bear,D.M., Wu,J., Harmin,D.A., Laptewicz,M., Barbara-Haley,K. and Kuersten,S. (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**, 182–187.
27. Kong,L., Zhang,Y., Ye,Z.Q., Liu,X.Q., Zhao,S.Q., Wei,L. and Gao,G. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, **35**, W345–W349.
28. Bono,H., Yagi,K., Kasukawa,T., Nikaido,I., Tominaga,N., Miki,R., Mizuno,Y., Tomaru,Y., Goto,H. and Nitanda,H. (2003) Systematic expression profiling of the mouse transcriptome using RIKEN cDNA microarrays. *Genome Res.*, **13**, 1318–1323.
29. Dowell,R., Jokerst,R., Day,A., Eddy,S. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.