Published online 11 November 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D145–D148***
doi:10.1093/nar/gkl837

fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences

Taishin Kin[1,*], Kouichirou Yamada[3], Goro Terai[2], Hiroaki Okida[2], Yasuhiko Yoshinari[4], Yukiteru Ono[3], Aya Kojima[2], Yuki Kimura[3], Takashi Komori[2] and Kiyoshi Asai[1,5]

[1]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Aomi 2-42, Koto-ku, Tokyo 135-0064, Japan, [2]Intec Web and Genome Informatics, 1-3-3 Shinsuna, Koto-ku, Tokyo 136-0075, Japan, [3]Information and Mathematical Science Laboratory, 1-5-21, Oh-tsuka, Bunkyo-ku, Tokyo 112-0012, Japan, [4]Mitsubishi Research Institute, 2-3-6, O-temachi, Chiyoda-ku, Tokyo 100-8140, Japan and [5]Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, 5-1-5 Kashiwa-no-ha, Chiba 277-8583, Japan

Received August 15, 2006; Revised September 20, 2006; Accepted October 6, 2006

ABSTRACT

There are abundance of transcripts that code for no particular protein and that remain functionally uncharacterized. Some of these transcripts may have novel functions while others might be junk transcripts. Unfortunately, the experimental validation of such transcripts to find functional non-coding RNA candidates is very costly. Therefore, our primary interest is to computationally mine candidate functional transcripts from a pool of uncharacterized transcripts. We introduce fRNAdb: a novel database service that hosts a large collection of non-coding transcripts including annotated/non-annotated sequences from the H-inv database, NONCODE and RNAdb. A set of computational analyses have been performed on the included sequences. These analyses include RNA secondary structure motif discovery, EST support evaluation, *cis*-regulatory element search, protein homology search, etc. fRNAdb provides an efficient interface to help users filter out particular transcripts under their own criteria to sort out functional RNA candidates. fRNAdb is available at http://www.ncrna.org/

INTRODUCTION

fRNAdb is a database that helps in annotating non-coding transcripts acquired from publicly available databases. H-inv: human full-length non-coding cDNAs (1); NONCODE: experimentally validated non-coding transcripts (2); and RNAdb: non-coding transcripts curated from the literature, human chromosome 7 project, and RIKEN antisense pipeline and other putative non-coding RNAs (3). Details are shown in Table 1. Each transcript is analyzed for various features such as maximum ORF length, the number of protein homologs, the average conservation score, transcription regulatory element motifs, existence of CpG islands and so on (listed in Table 2) that help in filtering out promising non-coding candidates. Transcripts can be filtered with fRNAdb's main listing interface in many different ways (see Figure 1). This main listing interface is linked to our custom UCSC Genome Browser (4) for functional RNAs equipped with our RNA-specific original custom tracks that are specific to screening of functional RNA. Users can inspect a transcript of interest from a genomic view with rich genomic information surrounding the mapped transcript. The information includes the UCSC original tracks such as known genes, genome conservation and Affymetrix transcriptome tracks (5), and our original tracks such as conserved potential secondary structure, existence of known RNA secondary structure motifs and significant RNA secondary structure *Z*-score regions (for details see Table 3).

fRNAdb

fRNAdb provides two types of interfaces. The first page presents a list of all transcripts rendered as a table with

Table 1. Data sources of fRNAdb

Source	Num. seq. (mapped)
H-inv 2.0 (non-protein coding transcripts)	5489 (5217)
NONCODE	5339 (576)
RNAdb	2865 (1306)
RNAdb (literature curation)	1446 (524)
RNAdb (human chromosome 7 project)	306 (299)
RNAdb (RIKEN antisense pipeline)	1113 (486)
Total	13 693 (7102)

*To whom correspondence should be addressed. Tel: +81 3 355 8059; Fax: +81 3 355 8081; Email: kin-taishin@aist.go.jp

Table 2. List of attributes

S. no.	Description	Number of trascripts	Min/max
1	Length of the sequence (nt)	13 693	15/107 797
2	Number of exons	7166	0/60
3	Number of overlapping ESTs	4184	0/6490
4	Number of mapped positions	7158	0/892
5	GC-content (%)	13 693	4/87
6	Maximum length of potential ORF (amino acids)	12 655	0/1664
7	Percentage of bases that is covered with repeat elements	6460	0/100
8	Repeat elements reside proximal upstream/downstream	2219	
9	Known gene that is a potential sense/antisense of this transcript (exon overlapping required)	936	
10	Number of protein homologs (GenBank NR)	5811	0/250
11	Known gene that includes this transcript within its intron	951	
12	Known gene region that overlaps with the mapping extent of this transcript (strand not considered)	4245	
13	Known gene that overlaps with this transcript within its intron in different strand	965	
14	Known gene where this transcript is possibly a part of its 3′-UTR	757	
15	Known gene where this transcript is possibly a part of its 5′-UTR	77	
16	Known gene within upstream 5 kb	1011	
17	Known gene within downstream 5 kb	402	
18	Average conservation score over the mapped exonic region	6184	0/93
19	Maximum conservation score over the mapped exonic region	5741	0/98
20	Maximum conservation score within 500 base upstream from the mapped 5′ terminal	6878	0/255
21	Overlapping UCSC ultra conserved region	24	0/4
22	Number of canonical splice signals in this transcript	751	0/30
23	Number of poly(A) signals in this transcript	8081	0/199
24	Number of CpG island	1353	0/4
25	Associated transposon free region	1137	
26	Number of RFAM known RNA motifs in this transcript	5511	0/12
27	Number of RNAz predictive RNA motifs in this transcript	1185	0/24
28	Number of EvoFold predictive RNA motifs in this transcript	888	0/7
29	Maximum *Z*-score of RNA secondary structure over this transcript. Scores lower than −6 are significant. Higher scores are considered insignificant. Stored scores= raw score × −10	252	0.0/121.0
30	Number of cell lines responding to Affy probes in exon regions of this transcript (Affymetrix Transcriptome Phase 2 Tiling Array Analyses)	1593	0/11

The number of applicable transcripts and the range of the attributes are shown.

35 columns including ones for the attributes described in Table 2 (Figure 1B). The tabular control panel is placed above the table, which presents five tabs labeled 'Basic', 'DB/ID', 'Expert', 'Sort' and 'Column' (Figure 1A). The Basic tab contains the basic filters: a collection of frequently used filters that provide simple and quick selection of transcripts that match common criteria of functional non-coding RNAs. For example, checking 'Mapped' to select only genome-mapped transcripts, 'Well conserved at best (Max > 50%)' for transcripts that have maximum conservation score >50% among 17 vertebrates (4) in their exonic regions, 'EST-supported' for reliable expression evidence, 'Tiny ORF (<40 aa)' enriching for non-coding transcripts, 'Low Repeat Coverage (<30%)' for no repeat element contamination, 'No protein homolog' for another condition which enriches non-coding transcripts, 'No overlapping known gene' is for removing the possibility of being part of a protein-coding gene transcript. After checking the boxes, the 'refresh' button runs filtering action and presents results. Our example conditions yield nine hits including one H-inv non-protein coding cDNA and eight RNAdb literature-curated miRNAs. In other words, these criteria match real functional RNAs and also indicate that one non-coding transcript shares the same properties. Clicking on the ID of this transcript produces a detailed view of this transcript shown in Figure 2. This feature visualizer shows graphical representation of a variety of sequence elements found in the transcript including *cis*-regulatory elements, repeat elements, EST mapping regions and six frame stop codon positions. There are many different ways to filter these non-coding transcripts and there are many more potential candidates hidden in this dataset. More details of the basic filters are provided on the website.

The rest of the tabs offer additional functionality to further improve usability. The DB/ID tab contains DB selection and ID selection boxes. The DB selection box allows you to limit the target databases from currently available databases: H-inv, NONCODE and RNAdb. The ID selection box lets you choose target transcripts that match given string patterns. For example, specifying 'FR000001' (fRNAdb ID) in this box limits the target transcript FR000001 alone. The wild-card '%' is allowed for pattern matching. Specifying 'LIT%' lets you limit the search to targets whose original IDs start with 'LIT'. The string pattern is matched against ID, Acc. and Original columns. The Expert tab provides an interface to specify multiple conditions that let you perform more complex filtering than the basic filters. Please refer to the website for more details about the expert filters. The Sort tab has a sorting interface that lets you sort the table with multiple sorting keys. The Column tab allows you to limit visible columns of the main listing table. Since the 35-column table is too wide for ordinary browsers to display on a single screen, you can narrow the width of the table with this interface for better visibility.

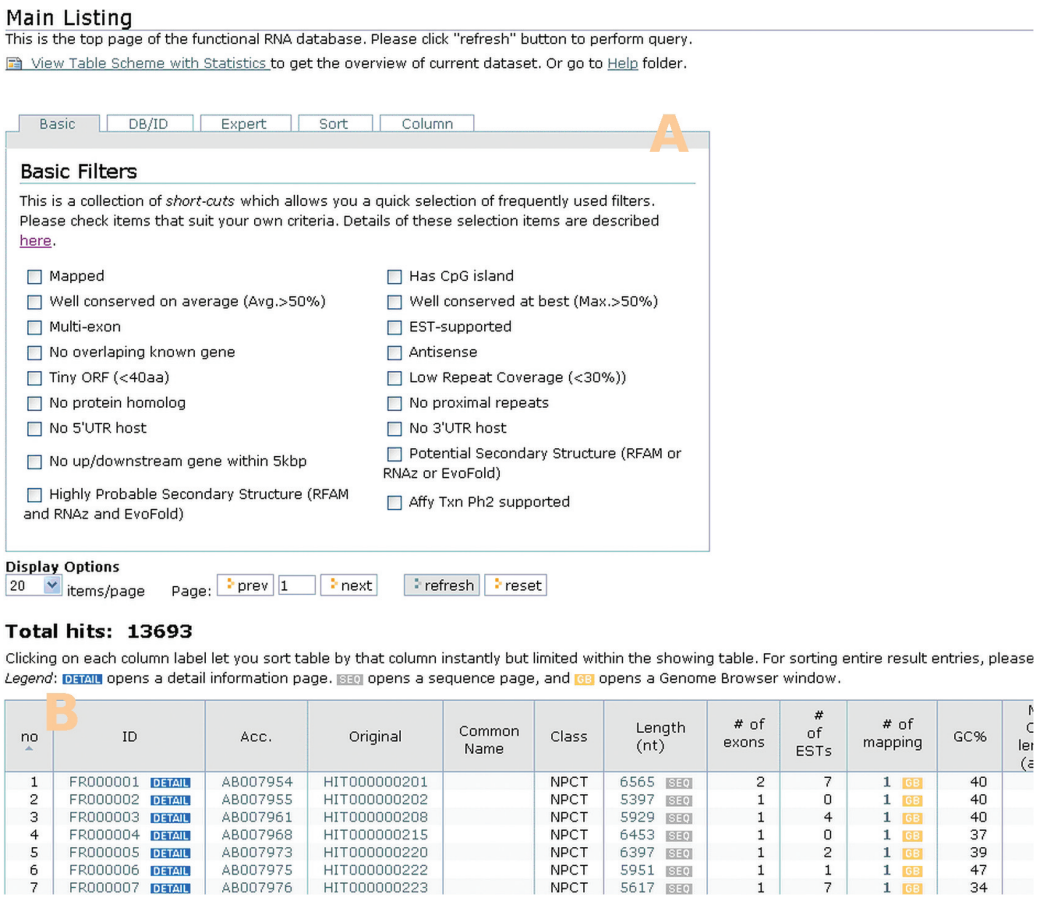

Figure 1. The first page shows a set of selection interfaces (**A**) and the listing table of 13 693 transcripts (**B**).

Table 3. Functional RNA-specific tracks

Track	Description
RNAz folds (15)	Secondary structure annotation of RNAz
ENOR (16)	ENOR (expressed non-coding region) [lifted from mm5]
Erdmann (6)	Erdmann non-coding RNAs
NONCODE (2)	Mapping information of NONCODE RNAs
RNAdb (3)	Mapping information of RNAdb RNAs
RNA Clusters	Small RNA genes often reside close to each other forming clusters. This track represents computationally identified RNA clusters in human genome
Rfam seed folds	Genomic search results with INFERNAL and covariance models generated from RFAM seeds
Rfam full	BLAT mapping results for RFAM full sequence dataset
antisense ChenJ NAR2004 (17)	Sense–antisense pairs among UCSC known genes
tRNAscan-SE (18)	tRNA genes predicted by tRNAscan-SE
Ultra conserved elements (19)	100% conserved elements (⩾200 bp) in human, rat and mouse
Ultra conserved elements 17 way	100% conserved elements in 17 vertebrates (longer than 50 bp)
Transposon free region (20)	Regions longer than 5 kb or 10 kb containing no LINEs, SINEs and LTRs
Human accelerated region (14)	HAR non-coding gene candidates predicted by (14)
Z-score	Regions with *Z*-score lower (lower is better) than −6 (actual track score = *Z*-score × −10)

UCSC GENOME BROWSER FOR FUNCTIONAL RNAs

We mirrored the UCSC Genome Browser and added our custom tracks specific to functional RNAs and miRNAs as shown in Tables 3 and 4. Most of the tracks have their own sources and reference papers. Our original tracks are RNA clusters, Rfam seed folds, tRNAscan-SE, Ultra Conserved Elements 17way and *Z*-score (details are shown in Table 3). Besides, we mapped RNA sequences from public functional RNA sequence databases including Erdmann (6), NONCODE, RNAdb and Rfam. The UCSC Genome Browser has several tracks for miRNA genes and targets but we added more tracks including miRBase (7) known miRNA genes, miRNAMap (8) and Berezikov's predicted miRNA genes (9), TarBase (10) known miRNA targets, and predicted miRNA targets from RNAhybrid (11), PicTar 4 species and 5 species (12), miRBase targets and T-ScanS miRNA targets

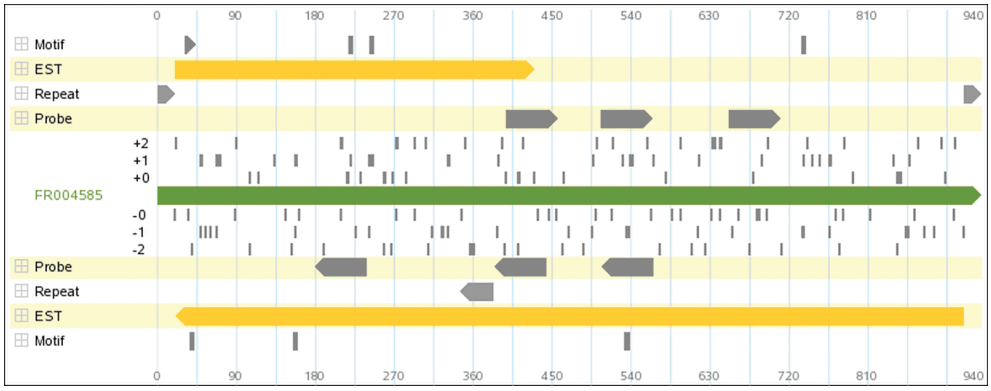

Figure 2. mRNA view of a transcript. Regulatory elements, EST positions, splice positions, repeat elements, six frame stop codons are visualized along with the full span of a cDNA.

Table 4. miRNA-specific tracks

Track	Description
Known miRNAs	miRBase known miRNAs
Predicted miRNAs	miRNAMap and Berezikov's predicted miRNAs
Known targets	TarBase experimentally verified miRNA target sites
Predicted targets	RNAhybrid, PicTar, miRBase and T-ScanS-predicted miRNA target sites

(13). Our custom tracks can be downloaded by using Table browser which can be accessed via 'Table' menu of the UCSC Genome Browser.

In the near future, fRNAdb will include more transcripts from other sequence databases or non-coding gene prediction results. For example, Human Accelerated Region (14) is currently included as our custom track of the Genome Browser. Sequences of these non-coding gene candidates will be included in fRNAdb. We will also add more attributes to fRNAdb. Especially attributes representing expression patterns of the transcripts or protein genes related to the transcripts.

ACKNOWLEDGEMENTS

This research is partially supported by the Functional RNA project funded by Ministry of Economy, Trade and Industry (METI). We thank Dr. Paul Horton for his kind help. Funding to pay the Open Access publication charges for this article was provided by National Institute of Advanced Industrial Science and Technology (AIST).

Conflict of interest statement. None declared.

REFERENCES

1. Imanishi,T., Itho,T., Suzuki,Y., O'Donovan,C., Fukuchi,S., Koyanagi,K.O., Barrero,R.A., Tamura,T., Yamaguchi-Kabata,Y., Tanino,M. *et al*. (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol*., **2**, 856–875.
2. Liu,C., Bai,B., Skogerbo,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.
3. Pang,K.C., Stephen,S., Engstrom,P.G., Tajul-Arifin,K., Chen,W., Wahlestedt,C., Lenhard,B., Hayashizaki,Y. and Mattick,J.S. (2005) RNAdb—a comprehensive mammalian noncoding RNA database. *Nucleic Acids Res*., **33**, D125–D130.
4. Hinrichs,A.S., Karolchik,D., Baertsch,R., Barber,G.P., Bejerano,G., Clawson,H., Diekhans,M., Furey,T.S., Harte,R.A., Hsu,F. *et al*. (2006) The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res*., **34**, D590–D598.
5. Cheng,J., Kapranov,P., Drenkow,J., Dike,S., Brubaker,S., Patel,S., Long,J., Stern,D., Tammana,H., Helt,G. *et al*. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, **308**, 1149–1154.
6. Szymanski,M., Erdmann,V.A. and Barciszewski,J. (2003) Noncoding regulatory RNAs database. *Nucleic Acids Res*., **31**, 429–431.
7. Griffiths-Jones,S. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*., **34**, D140–D144.
8. Hsu,P.W., Huang,H.D., Hsu,S.D., Lin,L.Z., Tsou,A.P., Tseng,C.P., Stadler,P.F., Washietl,S. and Hofacker,I.L. (2006) miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genome. *Nucleic Acids Res*., **34**, D135–D139.
9. Berezikov,E., Guryev,V., van de Belt,J., Wienholds,E., Plasterk,R.H. and Cuppen,E. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, **120**, 21–24.
10. Sethupathy,P., Corda,B. and Hatzigeorgiou,A.G. (2006) TarBase: a comprehensive database of experimentally supported animal microRNA targets. *RNA*, **12**, 192–197.
11. Kuger,J. and Rehmsmeier,M. (2006) RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res*., **34**, W451–W454.
12. Krek,A., Grun,D., Poy,M.N., Wolf,R., Rosenberg,L, Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C., Stoffel,M. *et al*. (2005) Combinatorial microRNA target predictions. *Nature Genet*., **37**, 495–500.
13. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
14. Pollard,K.S., Salama,S.R., Lambert,N., Lambot,M.A., Coppens,S., Pedersen,J.S., Katzman,S., King,B., Onodera,C., Siepel,A. *et al*. (2006) An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*, **443**, 167–172.
15. Washietl,S., Hofacker,I.L., Lukasser,M., Huttenhofer,A. and Stadler,P.F. (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol*., **23**, 1383–1390.
16. Furuno,M., Pang,K.C., Ninomiya,N., Fukuda,S., Frith,M.C., Bult,C., Kai,C., Kawai,J., Carninci,P., Hayashizaki,Y. *et al*. (2006) Clusters of internally primed transcripts reveal novel long noncoding RNAs. *PLoS Genet*., **2**, e37.
17. Chen,J., Sun,M., Kent,W.J., Huang,X., Xie,H., Wang,W., Zhou,G., Shi,R.Z. and Rowley,J.D. (2004) Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res*., **32**, 4812–4820.
18. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*., **25**, 955–964.
19. Bejerano,G., Pheasant,M., Makunin,I., Stephen,S., Kent,W.J., Mattick,J.S. and Haussler,D. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.
20. Simons,C., Pheasant,M., Makunin,I.V. and Mattick,J.S. (2005) Transposon-free regions in mammalian genome. *Genome Res*., **16**, 164–172.