Published online 23 October 2008 *Nucleic Acids Research, 2009, Vol. 37, Database issue* ***D89–D92***
doi:10.1093/nar/gkn805

The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs

Toutai Mituyama[1,*], Kouichirou Yamada[2], Emi Hattori[2], Hiroaki Okida[3], Yukiteru Ono[2], Goro Terai[3], Aya Yoshizawa[3], Takashi Komori[3] and Kiyoshi Asai[1,4]

[1]National Institute of Advanced Industrial Science and Technology (AIST), Computational Biology Research Center (CBRC), Tokyo 135-0064, [2]Information and Mathematical Science Laboratory, Inc., Department of Life Sciences, Tokyo 112-0012, [3]INTEC Systems Institute, Inc., Biobusiness Division, Tokyo 136-0075 and [4]Graduate School of Frontier Sciences, Department of Computational Biology, University of Tokyo, Chiba 277-8583, Japan

Received September 13, 2008; Revised October 9, 2008; Accepted October 10, 2008

ABSTRACT

We developed a pair of databases that support two important tasks: annotation of anonymous RNA transcripts and discovery of novel non-coding RNAs. The database combo is called the Functional RNA Database and consists of two databases: a rewrite of the original version of the Functional RNA Database (fRNAdb) and the latest version of the UCSC GenomeBrowser for Functional RNA. The former is a sequence database equipped with a powerful search function and hosts a large collection of known/predicted non-coding RNA sequences acquired from existing databases as well as novel/predicted sequences reported by researchers of the Functional RNA Project. The latter is a UCSC Genome Browser mirror with large additional custom tracks specifically associated with non-coding elements. It also includes several functional enhancements such as a presentation of a common secondary structure prediction at any given genomic window ⩽500 bp. Our GenomeBrowser supports user authentication and user-specific tracks. The current version of the fRNAdb is a complete rewrite of the former version, hosting a larger number of sequences and with a much friendlier interface. The current version of UCSC GenomeBrowser for Functional RNA features a larger number of tracks and richer features than the former version. The databases are available at http://www.ncrna.org/.

INTRODUCTION

Large-scale transcription analyses such as the H-invitational (1) and Fantom (2) projects reported a large number of transcripts that could not be associated with coding genes, and which were thus left unclassifiable. Several investigations revealed that these unclassifiable transcripts contain novel non-coding genes (3–5). The Functional RNA Database (fRNAdb) 1.0 (6) focused on acquiring and providing lines of evidence to infer non-coding-ness for these unclassifiable transcripts to help filter out candidates for non-coding genes. However, drastic changes in the situation surrounding non-coding RNA research spurred us to move on to the next phase of database development. A transcriptome analysis for natural RNA transcripts utilizing high-throughput sequencing is one of the most attractive topics among recent research activities. Due to the abundance of sequence data produced by deep sequencing, computational analysis plays an important role in the rapid sequence mapping and annotation of anonymous sequences. In particular, a sequence database is the most crucial part of computational analysis. Total RNAs extracted from a cell tend to have diverse compositions even though RNAs are extracted via immunoprecipitation of specific proteins (7–9). They contain tRNAs, rRNAs, coding mRNAs, varieties of transposons and non-coding RNAs including miRNAs and snoRNAs together with a fair amount of anonymous transcripts meeting no existing annotations although they can be mapped to a genome. Such transcripts may contain evidence of novel non-coding RNA genes. In order to adopt the large-scale sequence data from deep sequencing, we have completely redesigned and rebuilt fRNAdb. The major changes include increase of hosting sequences (from 13 693 to 509 795), sequence ontology (SO, http://song.sourceforge.net/) classification, keyword search function and Blast search service. The details given in the next section are new features for the current version.

fRNAdb

fRNAdb is a sequence database hosting a large collection of non-coding RNA sequence data from public

*To whom correspondence should be addressed. Tel: +81 3 3599 8059; Fax: +81 3 3599 8081; Email: mituyama-toutai@aist.go.jp non-coding databases: H-invDB rel. 5.0 (1), FANTOM3 (2), miRBase 10.0 (10), NONCODE v1.0 (11), Rfam v8.1 (12), RNAdb v2.0 (13) and snoRNA-LBME-db rel. 3 (14). Although these databases contain many identical sequences, fRNAdb consolidates them to a set of unique sequences. Therefore, one fRNAdb sequence can have multiple accessions and multiple source organisms.

A sequence can have one or more mapping loci in multiple genomes, gene association using mapping information, sequence similarity information between other registered sequences, and reference information. All sequences are mapped to multiple genomes (humans, mice, rats and fruit flies) in order to determine potential loci and potential homologs. The mapping loci can be viewed in our UCSC GenomeBrowser for Functional RNA for visual inspection with a number of tracks showing versatile genomic elements provided by the original UCSC Genome Browser and our additional tracks detailed in the next section.

fRNAdb allows users to search the sequences through keywords associated with them. Various kinds of information are associated with a sequence, as shown in Figure 1. The keywords are extracted from an identifier, description text, accession, SO, source organism, cross reference information, associated gene names, title/abstract/author text of reference papers, genome/chromosome/cytoband and sequence length. Common English words that may hinder efficient keyword search are eliminated from the index using the English dictionary of the open source spell checker *aspell* (http://aspell.net/).

Statistics of keywords associated with fRNAdb sequences can be browsed at the fRNAdb::Statistics page, where frequently used keywords corresponding to canonical terms in various ontology sets are presented. These statistics are useful for providing an overview of the entire non-coding RNA sequences from multiple aspects using different ontologies such as SO, taxonomy and several ontologies of the Open Biomedical Ontologies (http://www.obofoundry.org/): human disease ontology and gene ontology (biological/molecular processes).

fRNAdb also provides sequence homology search using Blastn (15). In order to provide better usability, we divided our database in two parts: one contains sequences longer than 50 bases and the other contains sequences 50 bases or shorter since some users are not interested in small sequences that include a large number of deep sequencing products. fRNAdb::Blast automatically adjusts some parameters according to the length of a query sequence in order to improve performance for short (<50 bases) query sequences. The adaptive parameters are gap opening/extension cost, E-value, and word size. All Blast parameters can be overridden by users. More details about fRNAdb are provided on the fRNAdb::Help page.

UCSC GENOME BROWSER FOR FUNCTIONAL RNA

This database is an extended mirror of the UCSC Genome Browser (16) hosting genomes of humans (hg17 and hg18), mice (mm9), rats (rn3) and fruit flies (dm3). This database has been updated extensively. There were 15 original tracks in the previous version (6). We re-organized our tracks and added more custom tracks. For hg18, our extension includes 26 essential tracks for the ncRNA Prediction and Mapping Tracks group, five essential tracks for the Misc. Genomic Element Tracks group, and five essential tracks for the miRNA-related Tracks group. Tracks for the whole human tiling array of Affymetrix Transfrags (17) are available (currently supported only on hg17).

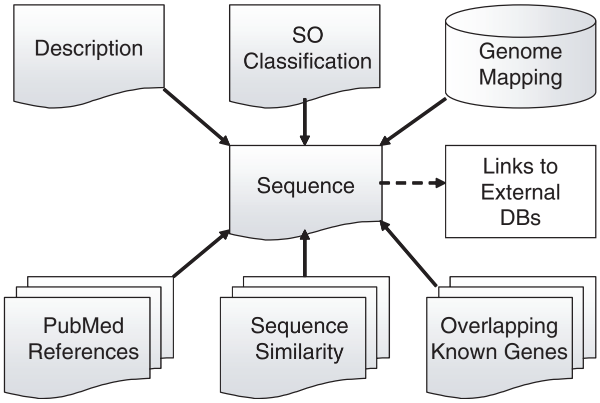

Figure 1. Diagram showing a registered sequence and its associations to other information.

We have developed several tracks to support an improved presentation. For example, the miRNA Atlas (18) track has a feature to present the expression profile of multiple miRNAs residing inside the GenomeBrowser window (Figure 2). Another example is tissue-specific enhancers and the target loci (19) track. This track indicates an enhancer region with an orange box and its associated gene locus with a green bar, which is rendered in darker green when the locus is activated in more tissues. Yet another extension is given to the conservation track, which shows not only a multiple genome alignment but also predicted common RNA secondary structures. When clicking on the conservation track in the window showing a genomic region ⩽500 bp, prediction is dynamically perfomed in both strands. Then, the browser presents a predicted secondary structure, minimum free energy and the number of base pairs per strand. The estimated secondary structure is downloadable as PDF graphics and in Stockholm format, which is a secondary structure annotated alignment file. This file can be used for determining homologous secondary structure in a database using Infernal software package (http://infernal.janelia.org). Complete listing and details of extension tracks are found in the Project Specific Custom Tracks page (http://www.ncrna.org/custom-tracks).

ACKNOWLEDGEMENTS

The authors thank everyone in the bioinformatics group of the Functional RNA Project for constructive criticisms and fruitful discussions.

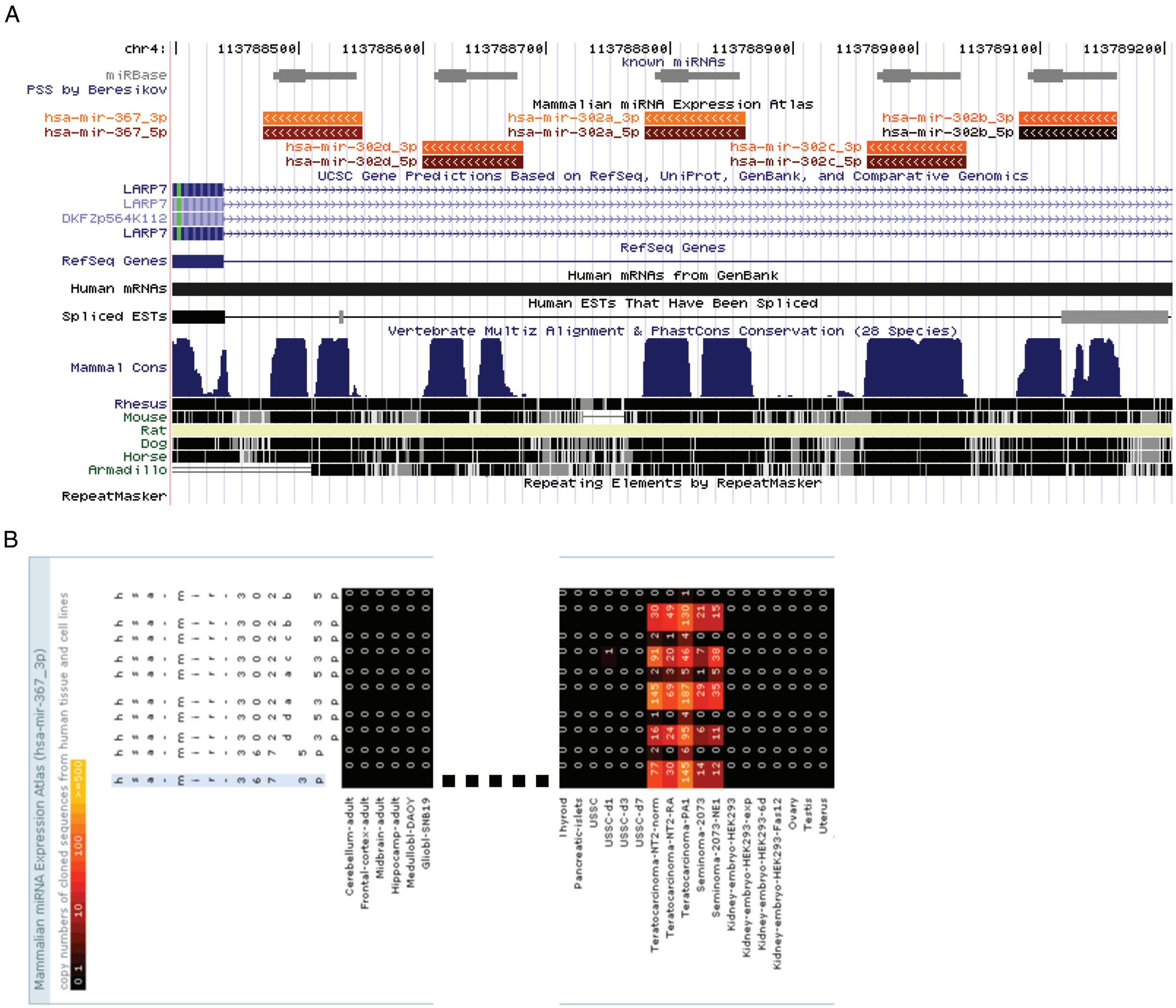

Figure 2. Mammalian miRNA Expression Atlas track showing miR-302a/b/c/d highly expressed at 3p (**A**). The detailed page shows expression profiles for these miRNAs with a heat map and actual read numbers previously reported by (20) (**B**).

FUNDING

This work was supported by the Functional RNA Project funded by New Energy and Industrial Technology Development Organization (NEDO). Funding for open access charge: Japan Biological Informatics Consortium (JBIC).

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. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
3. Inagaki,S., Numata,K., Kondo1,T., Tomita,M., Yasuda1,K., Kanai,A. and Kageyama,Y. (2005) Identification and expression analysis of putative mRNA-like non-coding RNA in Drosophila. *Genes Cell*, **10**, 1163–1173.
4. Sasaki,Y.T.F., Sano,M., Kin,T., Asai,K. and Hirose,T. (2007) Coordinated expression of ncRNAs and HOX mRNAs in the human HOXA locus. *Biochem. Biophys. Res. Comm.*, **357**, 724–730.
5. Xue,C., Li,F. and Li,F. (2008) Finding noncoding RNA transcripts from low abundance expressed sequence tags. *Cell Res.*, **18**, 695–700.
6. Kin,T., Yamada,K., Terai,G., Okida,H., Yoshinari,Y., Ono,Y., Kojima,A., Kimura,Y., Komori,T. and Asai,K. (2007) fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Res.*, **35**, D145–D148.
7. Kawamura,Y., Saito,K., Kin,T., Ono,Y., Asai,K., Sunohara,T., Okada,T.N., Siomi,M.C. and Siomi,H. (2008) Dropophila endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature*, **453**, 793–797.
8. Czech,B., Malone,C.D., Zhou,R., Stark,A., Schlingeheyde,C., Dus,M., Perrimon,N., Kellis,M., Wohlschlegel,J.A., Sachindanandam,R. *et al.* (2008) An endogenous small interfering RNA pathway in Drosophila. *Nature*, **453**, 798–802.
9. Okamura,K., Chung,W.J., Ruby,J.G., Guo,H., Bartel,D.P. and Lai,E.C. (2008) The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs. *Nature*, **453**, 803–806.

10. Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
11. He,S., Liu,C., Skogerbø,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.
12. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
13. Pang,K.C., Stephen,S., Dinger,M.E., Engström,P.G., Lenhard,B. and Mattick,J.S. (2007) RNAdb 2.0—an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res.*, **35**, D178–D182.
14. Lestrade,L. and Weber,M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.*, **34**, D158–D162.
15. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
16. Kuhn,R.M., Karolchik,D., Zweig,A.S., Trumbower,H., Thomas,D.J., Thakkapallayil,A., Sugnet,C.W., Stanke,M., Smith,K.E., Siepel,A. *et al.* (2007) The UCSC genome browser database: update 2007. *Nucleic Acids Res.*, **35**, D668–D673.
17. Kapranov,P., Cheng,J., Dike,S., Nix,D.A., Duttagupta,R., Willingham,A.T., Stadler,P.F., Hertel,J., Hackermüller,J., Hofacker,I.L. *et al.* (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, **316**, 1484–1488.
18. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O. and Landthaler,M. (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.
19. Pennacchio,L.A., Loots,G.G., Nobrega,M.A. and Ovcharenko,I. (2007) Predicting tissue-specific enhancers in the human genome. *Genome Res.*, **17**, 201–211.
20. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O., Landthaler,M. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**:1401–1414.