Published online 30 October 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D75–D80***
doi:10.1093/nar/gkp902

UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs

Giorgio Grillo[1], Antonio Turi[2], Flavio Licciulli[1], Flavio Mignone[3], Sabino Liuni[1], Sandro Banfi[4], Vincenzo Alessandro Gennarino[4], David S. Horner[5], Giulio Pavesi[5], Ernesto Picardi[2] and Graziano Pesole[1,2,*]

[1]Istituto Tecnologie Biomediche del Consiglio Nazionale delle Ricerche (CNR), via Amendola 122/D, 70126 Bari, [2]Dipartimento di Biochimica e Biologia Molecolare 'E. Quagliariello', Università di Bari, via Orabona 4, 70126 Bari, [3]Dipartimento di Chimica Strutturale e Stereochimica Inorganica, Università degli Studi di Milano, 20133 Milano, [4]Telethon Institute of Genetics and Medicine, via Pietro Castellino 111, 80131 Naples and [5]Dipartimento di Scienze Biomolecolari e Biotecnologie, Università di Milano, via Celoria 26, 20133 Milano, Italy

Received September 1, 2009; Revised September 29, 2009; Accepted October 6, 2009

ABSTRACT

The 5′ and 3′ untranslated regions of eukaryotic mRNAs (UTRs) play crucial roles in the post-transcriptional regulation of gene expression through the modulation of nucleo-cytoplasmic mRNA transport, translation efficiency, subcellular localization and message stability. UTRdb is a curated database of 5′ and 3′ untranslated sequences of eukaryotic mRNAs, derived from several sources of primary data. Experimentally validated functional motifs are annotated and also collated as the UTRsite database where more specific information on the functional motifs and cross-links to interacting regulatory protein are provided. In the current update, the UTR entries have been organized in a gene-centric structure to better visualize and retrieve 5′ and 3′UTR variants generated by alternative initiation and termination of transcription and alternative splicing. Experimentally validated miRNA targets and conserved sequence elements are also annotated. The integration of UTRdb with genomic data has allowed the implementation of an efficient annotation system and a powerful retrieval resource for the selection and extraction of specific UTR subsets. All internet resources implemented for retrieval and functional analysis of 5′ and 3′ untranslated regions of eukaryotic mRNAs are accessible at http://utrdb.ba.itb.cnr.it/.

INTRODUCTION

One of the main challenges of the post-genomic era is the understanding of the mechanisms that control the spatio-temporal regulation of gene expression. The fate of newly synthesized mRNA with respect to its nucleo-cytoplasmic transport, stability, translation efficiency and subcellular localization is determined at the post-transcriptional level. Such regulation is mostly mediated by *cis*-acting elements located in the 5′ and 3′ untranslated regions of mRNAs (5′UTR and 3′UTR) (1) and miRNAs interacting with their specific targets in 3′UTRs (2,3).

Various specific functional sequence elements and miRNA targets have been identified and characterized in mRNA UTRs. These elements usually correspond to short oligonucleotide tracts whose biological activity relies on a combination of their primary sequence and specific secondary structure. These motifs act either as target sites for RNA binding factors or interact directly with the translation machinery. Additionally, miRNA targets, usually located in the 3′UTR, present a very degenerate complementarity with the miRNAs, tolerating several mismatches, gaps and G–U pairings, outside of 6–8 bp continuous seed region at the 5′-end of the miRNA. Additionally, some UTRs may be targeted by complementary natural antisense transcripts masking RNA binding protein or miRNA binding sites (4).

Notably, it is now clear that the same gene may generate several transcript variants, through the use of alternative sites for the initiation and termination of transcription and through alternative splicing. Alternative transcripts can differ both in the coding and in the untranslated regions

*To whom correspondence should be addressed. Tel: +39 080 5443588; Fax; +39 080 5443317; Email: graziano.pesole@biologia.uniba.it

(5). Specifically, alternative 5′ and 3′UTRs may differentially modulate the gene expression due to the presence of different combinations of functional motifs and miRNA targets.

The availability of a large collection of functionally related sequences—such as UTRs—is invaluable for structural and functional analyses and for a better understanding of the specific role of different variants. To address this issue we have developed a new version of UTRdb, a collection of 5′ and 3′ UTR sequences derived from eukaryotic mRNAs, where the entries have been organized in a gene-centric structure in order to provide relevant information about splicing variants. Sequences collated in UTRdb were recovered from the National Center for Biotechnology Information (NCBI) RefSeq transcripts (6) using custom software. For human genes, a more comprehensive collection of UTRs is available [derived from the full set of over 300 000 alternative full-length transcripts collected in ASPicDB (7)] generated by a thorough analysis of all available EST/mRNA data.

All UTRdb entries are further annotated for the occurrence of validated regulatory elements, conserved elements and structured RNAs, and miRNA targets (see below). Furthermore, the completeness of 5′UTRs is assessed by the occurrence of mapping CAGE tags (22) (if available) and that of 3′UTRs by the occurrence of a polyA signal and/or a polyA tail.

We have also further expanded UTRsite, a collection of regulatory elements located in 5′ and 3′ UTRs and whose function and structure have been experimentally determined and published. The UTRsite collection may prove useful in automatic annotation projects of unknown expressed sequences as well as for finding previously undetected signals in known sequences. In the present release, the information for each UTRsite entry has been further enriched including data on functional interacting RNA-binding proteins.

The gene-centric structure of UTRdb facilitates a full integration with all possible gene attributes collected in the NCBI Gene database (8) or other genomic resources such as the UCSC genome browser (9). In this way, the retrieval of specific UTR subsets is possible based on the features associated with each gene, for example a GO term (10), a MIM identifier (11) or a Unigene accession (12).

GENERATION OF UTRdb AND ITS INTEGRATION WITH OTHER DATABASES

UTRdb entries are automatically generated through the accurate parsing of the feature table of NCBI RefSeq and ASPicDB transcripts for the UTRef and UTRfull sections of UTRdb database, respectively. ASPicDB contains all possible transcript isoforms for a gene reconstructed by using all available transcript and EST sequences as described in (13). UTR entries are then annotated for the occurrence of tandem and interspersed repetitive elements by using RepeatMasker (v3.2.8, March 2009; A.F.A. Smit, R. Hubley & P. Green RepeatMasker at http://repeatmasker.org), and known regulatory motifs collected in the UTRsite database, as detailed in (14). Each UTRsite entry (Figure 1) is prepared/reviewed/updated by expert scientists (in many cases, those who performed the experimental analysis) by using a suitably developed submission tool (15).

UTRdb entries are also annotated for the occurrence of validated miRNA targets, collected in miRecords (16), a large, high-quality database of experimentally validated miRNA targets resulting from meticulous literature curation. Furthermore, we annotated a set of 3′UTR sequences that have a high likelihood to represent *bona fide* miRNA target recognition sites, as predicted by the HOCTAR tool (17).

For a subset of seven organisms, namely human, mouse, rat, cow, dog, chicken and *Arabidopsis*, for which a suitable genome assembly is available, we also determined the genomic coordinates of UTRs. For such species we were able to clean all redundancies based on the observation of coincident UTRs coordinates, arising from alternative mRNA isoforms.

Additional annotations are specifically provided for genome-linked UTRs. These include: (i) highly conserved sequence blocks from the 17-way PhastCons vertebrate conserved elements (18); (ii) significantly conserved tracts detected by Evofold (19); and (iii) structural conserved non-coding RNAs detected by RNAz (20).

PhastCons detects evolutionarily conserved elements using a genome-wide multiple alignment based on a phylogenetic hidden Markov model (21). Evofold is a general comparative genomics method based on phylogenetic stochastic context-free grammars for identifying RNA secondary structures encoded in the human genome and conserved in an eight-way genome-wide alignment of the human, chimpanzee, mouse, rat, dog, chicken, zebrafish and pufferfish genomes (19). RNAz evaluates conserved genomic DNA sequences for signatures of structural conservation of base pairing patterns and exceptional thermodynamic stability. We employed three sets retrieved from (20), with regions conserved with P-value >0.9 in human, mouse, rat and dog (Set 1); human, mouse, rat, dog and chicken (Set 2); in human, mouse, rat, dog, chicken and either fugu or zebrafish (Set 3).

To assess the completeness of 5′UTRs in human and mouse we used the mapping data of the CAGE tags indicating the location of the transcription start site (22). The 5′-end of a 5′UTR has been considered as complete when at least five CAGE tags map in a nearby position (a window of 5 bp around the mapping position of the 5′-end of the 5′UTR). Analogously, a 3′UTR is considered complete at its 3′-end if a polyA signal and/or a polyA tail is detected in the original transcript sequence.

The UTRdb and UTRSite data have been organized into relational databases using MySQL as the Database Management System. A novel implementation detail in this new release is that several physical databases (containing UTR sequences and annotations from Refseq and ASPicDB transcripts, chromosome coordinates of source transcripts for the seven model organism, taxonomic data, etc.) are used to store all the information on UTRs and their annotations. The new search and retrieval system retrieves and integrates data contained in these different relational databases to give out the requested data on

UTRSite

Home | Signal Manager

user: guest

Signal Manager :: View

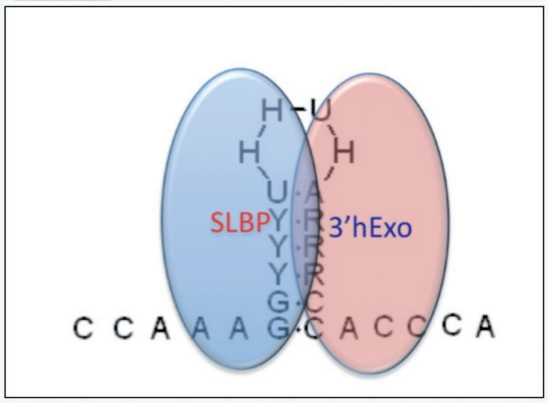
:: General Information

ID	**U0001**
Creation Date	30 July 1997
Updating Date	05 May 2008
Standard Name	Histone 3'UTR stem-loop structure (HSL3)
UTRSite Pattern	`rl={au,ua,gc,cg,gu,ug}` `0...1 mmmm pl=ggyyy u hhuh a rl-pl mm 0...3`
Random Expectation	0.0000192613 hits/kb

:: Taxonomy

UTR Region	3'
Taxon Range	Eukaryota

:: Description

Description	Metazoan histone 3'-UTR mRNAs, lacking a polyA tail, contain a highly conserved stem-loop structure with a six base stem and a four base loop (HSL3). This stem-loop structure plays a different role in the nucleus and in the cytoplasm. In the nucleus, it is involved in pre-mRNA processing and nucleocytoplasmic transport, whereas in the cytoplasm it enhances translation efficiency and regulates histone mRNA stability. The trans-acting factor which interacts with the 3'-UTR hai..... more ...
Image	

:: Features

Feature Key	HSL3

:: Database Cross-references

Link	RFAM: RF00032

:: Transcript(s) / Gene(s)

	Description	Species	Link	Ref.
UTR(s)	3'UTR in Mus musculus histone 2, H2aa1 (Hist2h2aa1), mRNA	Mus musculus	UTRef: CR073878	[U1]
Transcript(s)	Description	Species	Link	Ref.
	Mus musculus histone cluster 2, H2aa1 (Hist2h2aa1), mRNA	Mus musculus	RefSeq: NM_013549	[T1]
	Description	Species	Link	Ref.

:: Protein(s)

	Description	Species	Link	Ref.
Binding Protein(s)	Stem loop (histone) binding protein	Homo sapiens	RefSeq: NP_006518	[B1]
	three prime histone mRNA exonuclease 1	Homo sapiens	RefSeq: NP_699163	[B2]
Interactor(s) of Binding Protein(s)	Description	Species	Link	Ref.
	zinc finger protein 473 [ZPF100]	Homo sapiens	RefSeq: NP_056243	[P1]

:: References

Figure 1. Sample UTRsite entry. The general information section includes the pattern syntax of the regulatory motif in a format suitable for PatSearch software (23) and the number of hits/kb expected in a sequence collection of randomly generated sequences of the same nucleotide composition of UTRdb. The cross-link to the RFAM database (24), transcripts, genes and RNA binding proteins, if available are also provided, as well as all relevant references (not shown here).

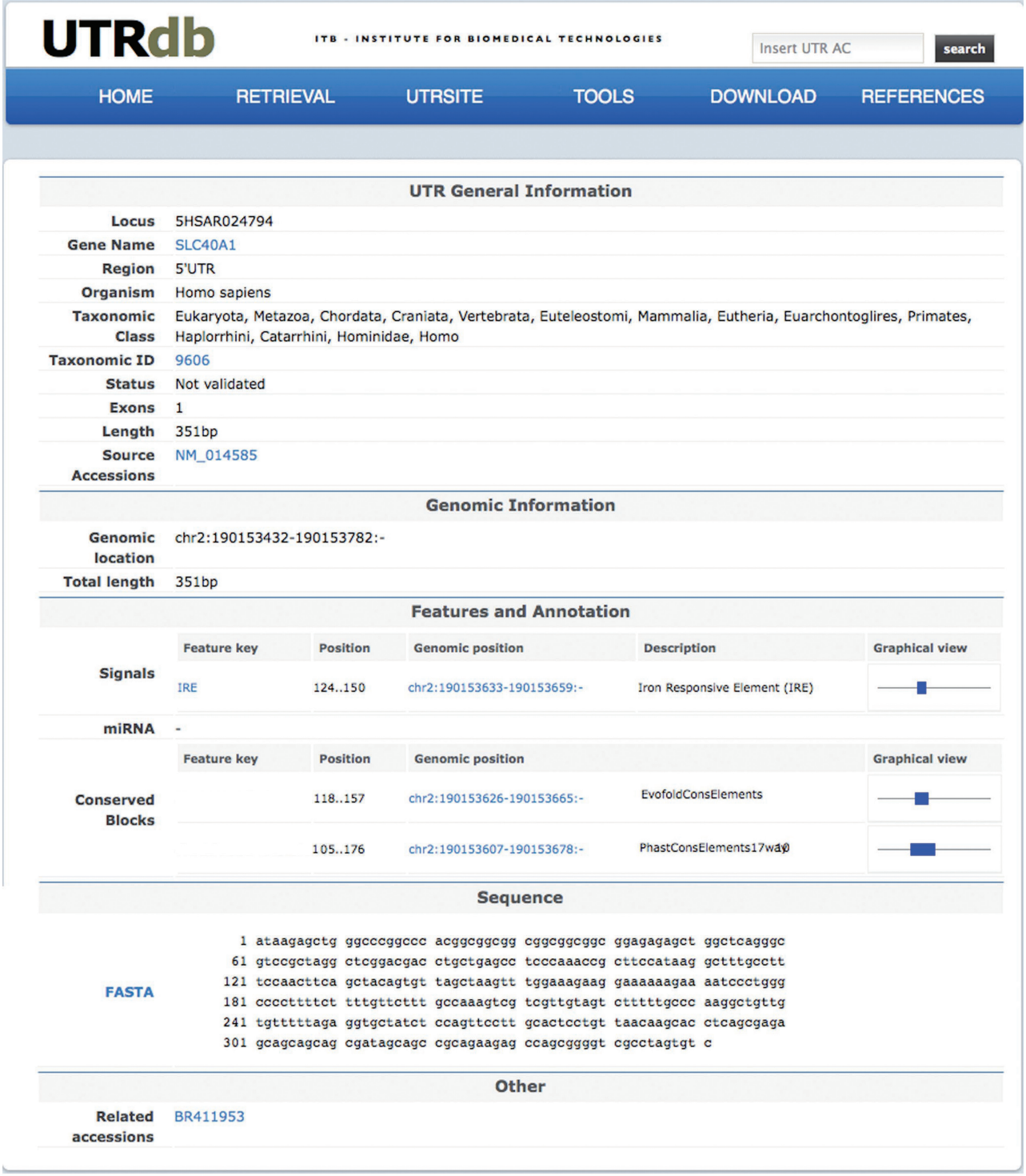


Figure 2. Sample entry of the UTRdb database. The 'Genomic Information' and 'Features and Annotation' sections report information on genome mapping coordinates and on the localization of UTRsite elements, miRNA targets and conserved elements, respectively.

UTRs and related annotations [such as the database from which a UTR was recovered (Refseq or ASPicDB), genomic coordinates and structure, miRNA targets and conserved elements localization, functional elements, etc.].

An exemplar entry of UTRdb is shown in Figure 2.

UTRdb CONTENT

UTRdb (UTRef section, release 2010) contains a total of 473 330 5′UTR and 527 323 3′UTR entries, respectively, from 483 605 genes in 79 species (see the Supplementary Data for more information).

A total of 788 370 UTRsite motifs are annotated (317 767 in the 5′UTRs and 470 603 in the 3′UTRs), 20 191 experimentally validated miRNA targets, and 242 773 conserved regions.

For human, the UTRfull section is also available, including UTRs deriving from full length transcripts collected in ASPicDB (7). Overall, UTRfull contains 124 345 and 194 503 5′ and 3′UTRs respectively (3.37/gene) and 3′UTRs (5.18/gene), with 348 412 annotated UTRsite motifs, 649 679 conserved elements and 105 209 experimentally validated miRNA targets.

AVAILABILITY OF UTRdb

UTRdb and UTRsite are accessible through a newly developed retrieval system where simple and advanced search forms are available. UTRs can be retrieved by several accession IDs, GO terms and MIM identifiers. Additionally, the advanced form permits a further refinement of the UTR subset to be retrieved using several criteria including the number of CAGE mapping tags (for 5′UTRs), the length of the UTR, the number of spanning exons, the occurrence of UTRsite motifs, conserved elements and miRNA targets.

A download facility for selected UTR entries in FASTA format is also available.

Further online utilities are UTRscan and UTRblast. The UTRscan feature allows the enquirer to search user-submitted sequences for any of the motifs collected in UTRsite. The UTRblast utility allows database searches against any of the UTRdb sections.

UTRdb, UTRsite and other related resources are publicly available at http://utrdb.ba.itb.cnr.it/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Fatima Gebauer for helpful comments and suggestion on the UTRsite structure.

FUNDING

Ministero dell'Istruzione, dell'Università e della Ricerca, Italy: Fondo Italiano Ricerca di Base, Italy: 'Laboratorio Internazionale di Bioinformatica' (LIBI); Laboratorio di Bioinformatica per la Biodiversità Molecolare (MBLAB). Funding for open access charge: Ministero dell'Istruzione, Università e Ricerca, Italy.

REFERENCES

1. Mignone,F., Gissi,C., Liuni,S. and Pesole,G. (2002) Untranslated regions of mRNAs. *Genome Biol.*, **3**, REVIEWS0004.
2. Flynt,A.S. and Lai,E.C. (2008) Biological principles of microRNA-mediated regulation: shared themes amid diversity. *Nat. Rev. Genet.*, **9**, 831–842.
3. Rana,T.M. (2007) Illuminating the silence: understanding the structure and function of small RNAs. *Nat. Rev. Mol. Cell Biol.*, **8**, 23–36.
4. Faghihi,M.A. and Wahlestedt,C. (2009) Regulatory roles of natural antisense transcripts. *Nat. Rev. Mol. Cell Biol.*, **10**, 637–643.
5. Kim,E., Goren,A. and Ast,G. (2008) Alternative splicing: current perspectives. *Bioessays*, **30**, 38–47.
6. 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.
7. Castrignano,T., D'Antonio,M., Anselmo,A., Carrabino,D., D'Onorio De Meo,A., D'Erchia,A.M., Licciulli,F., Mangiulli,M., Mignone,F., Pavesi,G. *et al.* (2008) ASPicDB: a database resource for alternative splicing analysis. *Bioinformatics*, **24**, 1300–1304.
8. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
9. Kuhn,R.M., Karolchik,D., Zweig,A.S., Wang,T., Smith,K.E., Rosenbloom,K.R., Rhead,B., Raney,B.J., Pohl,A., Pheasant,M. *et al.* (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.*, **37**, D755–D761.
10. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
11. Amberger,J., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.
12. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
13. Bonizzoni,P., Mauri,G., Pesole,G., Picardi,E., Pirola,Y. and Rizzi,R. (2009) Detecting alternative gene structures from spliced ESTs: a computational approach. *J. Comput. Biol.*, **16**, 43–66.
14. Pesole,G., Liuni,S., Grillo,G., Licciulli,F., Mignone,F., Gissi,C. and Saccone,C. (2002) UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. Update 2002. *Nucleic Acids Res.*, **30**, 335–340.
15. Mignone,F., Grillo,G., Licciulli,F., Iacono,M., Liuni,S., Kersey,P.J., Duarte,J., Saccone,C. and Pesole,G. (2005) UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.*, **33**, D141–D146.
16. Xiao,F., Zuo,Z., Cai,G., Kang,S., Gao,X. and Li,T. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res*, **37**, D105–D110.
17. Gennarino,V.A., Sardiello,M., Avellino,R., Meola,N., Maselli,V., Anand,S., Cutillo,L., Ballabio,A. and Banfi,S. (2009) MicroRNA target prediction by expression analysis of host genes. *Genome Res.*, **19**, 481–490.
18. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
19. Pedersen,J.S., Bejerano,G., Siepel,A., Rosenbloom,K., Lindblad-Toh,K., Lander,E.S., Kent,J., Miller,W. and Haussler,D. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, **2**, e33.
20. 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.
21. King,D.C., Taylor,J., Elnitski,L., Chiaromonte,F., Miller,W. and Hardison,R.C. (2005) Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome Res.*, **15**, 1051–1060.
22. Severin,J., Waterhouse,A.M., Kawaji,H., Lassmann,T., van Nimwegen,E., Balwierz,P.J., de Hoon,M.J., Hume,D.A.,

Carninci,P., Hayashizaki,Y. *et al.* (2009) FANTOM4 EdgeExpressDB: an integrated database of promoters, genes, microRNAs, expression dynamics and regulatory interactions. *Genome Biol.*, **10**, R39.
23. Grillo,G., Licciulli,F., Liuni,S., Sbisa,E. and Pesole,G. (2003) PatSearch: a program for the detection of patterns and structural motifs in nucleotide sequences. *Nucleic Acids Res.*, **31**, 3608–3612.
24. Gardner,P.P., Daub,J., Tate,J.G., Nawrocki,E.P., Kolbe,D.L., Lindgreen,S., Wilkinson,A.C., Finn,R.D., Griffiths-Jones,S., Eddy,S.R. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37**, D136–D140.