BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 1 2012, pages 142–144
doi:10.1093/bioinformatics/btr608

Databases and ontologies

Advance Access publication November 4, 2011

AURA: Atlas of UTR Regulatory Activity

E. Dassi*,†, A. Malossini†, A. Re†, T. Mazza, T. Tebaldi, L. Caputi and A. Quattrone

Laboratory of Translational Genomics - Centre for Integrative Biology, University of Trento, Via delle Regole, 101, 38123 Mattarello (TN), Italy

Associate Editor: Mario Albrecht

ABSTRACT

Summary: The Atlas of UTR Regulatory Activity (AURA) is a manually curated and comprehensive catalog of human mRNA untranslated regions (UTRs) and UTR regulatory annotations. Through its intuitive web interface, it provides full access to a wealth of information on UTRs that integrates phylogenetic conservation, RNA sequence and structure data, single nucleotide variation, gene expression and gene functional descriptions from literature and specialized databases.

Availability: http://aura.science.unitn.it

Contact: aura@science.unitn.it; dassi@science.unitn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 11, 2011; revised on October 27, 2011; accepted on October 28, 2011

1 INTRODUCTION

The 5′ and 3′ *untranslated regions* (UTRs) are the portions of an mRNA located at each side of the coding sequence. UTRs contain information for post-transcriptional regulation of mRNA, including transport, stability, localization and access to translation, and hence they largely determine the fate of mature mRNAs in the cell (Keene, 2007). Such events are mediated by hundreds of *trans*-acting factors: primarily RNA binding proteins (RBPs), associated with all cellular mRNAs to form ribonucleoprotein complexes (RNPs), but also non-coding RNAs, of which the microRNA (miRNA) class has a clear functional role.

The experimentally determined sequence and structure binding constraints of UTRs vary widely between and within RBPs and non-coding RNAs, and the regulatory interactions are globally characterized by extreme complexity, since a regulator can bind to multiple UTRs in multiple sites and vice versa. Moreover, the mRNA *trans–cis* interaction network undergoes remarkable plasticity, since the fate of an mRNA is determined by its temporally and spatially dependent association to several regulators (Anderson *et al.*, 2009). Unraveling the molecular code behind this sophisticated process is the key for: (i) understanding to what extent cell programs are regulated by the degree of mRNA abundance, localization and translation; (ii) deciphering how malfunction of *trans*-acting factors or mutation of target sites is at the root of some severely altered cellular phenotypes; (iii) identifying novel therapeutics aimed at modulating mRNA dynamics in the window between transport and translation. With this aim, a growing number of studies, both mechanistic and systems-based, provide information on factors binding to UTRs. Nevertheless, integration of these data and annotation of UTRs in genome browsers are lacking or insufficient.

The *Atlas of UTR Regulatory Activity* (AURA) fills this gap with unprecedented richness and coverage, by collecting and combining human UTR annotation and binding data from several sources.

2 DESCRIPTION AND USAGE

The increasing centrality of post-transcriptional regulation among gene expression studies is witnessed by the recent release of several specialized databases. RBPDB focuses on *trans*-acting proteins by collecting semi-manually curated literature data about RBPs and their demonstrated or predicted binding motifs (Cook *et al.*, 2011); Transterm is a regulatory sequence database that aggregates heterogeneous lists of *cis*-acting motifs relevant for post-transcriptional regulation (Jacobs *et al.*, 2009); starBase and CLIPZ store primary data of *trans–cis* interactions obtained by next-generation high-throughput technologies (Khorshid *et al.*, 2011). In addition, more specialized resources allow the user to search and analyze a limited number of particularly well-known regulatory elements in greater detail (e.g. AREsite, Gruber *et al.*, 2010, UTRdb and UTRsite, Grillo *et al.*, 2010).

Unlike these catalogs, AURA is designed to be a comprehensive and centralized warehouse of human UTR mapped annotations, both in terms of regulatory macromolecules and their site of binding. AURA records non-redundant, direct and experimentally assessed interactions of RNA binding proteins and microRNAs with human UTRs. It contains an updated set of annotated human UTRs (except those <5 bases) from the UCSC Genome Browser (GRCh37/hg19 assembly), experimental literature data (1041 publications) and consolidated information from several specialized databases, including miRTarBase (Hsu *et al.*, 2011), miRecords (Xiao *et al.*, 2009) and the aforementioned AREsite and RBPDB resources. Currently, it covers 127 523 human UTRs, corresponding to 63 138 transcripts encoded by 19 364 protein coding genes. An extensive comparison between AURA and related resources can be found in File S2 in Supplementary Material.

AURA is developed according to the convention that an RBP is a protein showing a reviewed RNA binding domain, and according to the rule that whenever positional data on mRNA regulatory binding sites are made available, the coordinates of each binding site are evaluated against the current genome annotation to verify the site lies within or overlaps the spliced UTR of a transcript.

The current AURA release provides a checked evidence of 299 393 interactions between 100 RBPs and 33 836 UTRs, of 28 351 interactions between 303 miRNAs and 5885 UTRs and collectively

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

of 56 910 *cis*-sites over 11 559 UTRs. Additional major attributes enabling the characterization and/or assessment of the interactions between UTRs and *trans*-acting factors include synteny information and joint visualization of gene expression profiles for the interacting partners. Furthermore, the assessment of an interaction between an RBP and an UTR is improved by the cross-reference to the Protein Human Atlas database (Berglund *et al.*, 2008). A high-level schema of the database can be found in Supplementary Figure S1.

2.1 Search

To account for the observation that a transcript can interact with multiple RBPs as well as an RBP can interact with multiple transcripts, AURA exhibits an intuitive interface through which the user can query a 'target locus' or a 'trans factor', respectively. The former query returns a list of genes whose HGNC gene symbol or synonyms contain the searched term; each gene in the list is annotated with its functional description, synonyms and UTRs. Furthermore, an exon–intron map of the UTRs is provided in order to allow proper discrimination between the different transcripts of a gene. On the other hand, the latter query results in a disambiguation list where all the *trans*-factors, whose names or synonyms contain the searching term, are shown. To select the *trans*-factor of interest, the user might benefit from genes' short descriptions and functional summaries. Upon selection, AURA returns the list of its target UTRs. These UTRs can be grouped by gene ontology (GO) slim categories (http://www.geneontology.org/GO.slims.shtml) or by chromosome mapping. Furthermore, the user can filter the results by selecting a combination of supporting experimental evidences.

2.2 UTR view

Selected UTRs are shown in an 'UTR view', consisting of two standard elements:

- The textual header containing: the chromosomal position and length of the spliced UTR, the HGNC name and UniProt description of the gene the UTR belongs to, and the link to the HPA database. Also shown are the overall conservation, which is the mean PhastCons single nucleotide conservation score for the UTR (Fujita *et al.*, 2011), and the corresponding transcript half-life according to a transcriptome-wide stability measurement (Friedel *et al.*, 2009).
- The AURA sequence browser, based on the JBrowse architecture (Skinner *et al.*, 2009), contains all the annotations related to a specific UTR, i.e. multiple tracks annotating the UTR by evolutionary conservation, single nucleotide variation and *cis*-regulatory binding sites. The 'Conservation' track displays the score calculated for each nucleotide in the UCSC 46 species alignment (Fujita *et al.*, 2011). In the 'SNP' track, AURA integrates the single nucleotide polymorphisms (SNPs) recorded in the dbSNP database (Sherry *et al.*, 2001), allowing the user to combine with the other annotation tracks to look for variations of potential impact in post-transcriptional regulation. The 'RBP' track contains the RBP binding sites, whereas the 'miR' track contains the microRNAs binding sites. Two further tracks are provided to show the *trans*-factors for which only partial information is available. The 'unknown mRNA location' track denotes the *trans*-factors known to bind a transcript without any further mapping information. Instead, the 'unknown UTR location' track indicates the *trans*-factors whose UTR binding site is unknown. All the annotations in the tracks are clickable: whenever the user clicks on an annotation, a description page containing binding sites and cross-references is shown. In this view, the minimal energy predicted secondary structure (Fujita *et al.*, 2011) together with the color-coded nucleotide phylogenetic conservation, SNP locations and *trans*-factor binding sites of the selected UTR can be optionally drawn through VARNA (Darty *et al.*, 2009).

Furthermore, AURA provides the user with multiple ways of grouping gene expression results retrieved from the Gene Expression Atlas (http://www.ebi.ac.uk/gxa/) and related to the gene locus of the selected UTR. Results are reported in tables where a row corresponds to a condition, whereas the columns, in order, show the number of times the gene was observed to be up- or downregulated with respect to its mean expression value and the significance of the measure ($\log_{10} P$-values). In case of *trans*-factor search, a joint table containing gene expression experiments for both the gene coding for the *trans*-factor and the gene bearing the bound UTR is shown. Moreover, significant differences in common between regulator and target are highlighted to emphasize possible correlations or anti-correlations between them. Annotations concerning an UTR can be extracted in textual format through the UTRCard feature; furthermore, the whole MySQL database can be downloaded from a dedicated page. A last way of mining the data contained in AURA is through the AURA Mart, which is available at the website and provides all query functionalities offered by the well-known BioMart platform (http://www.biomart.org).

3 FUTURE DEVELOPMENT

AURA gathers data by aggregation, integration and summarization of knowledge from scientific literature and specialized databases. Future developments include (i) the integration of the UTR mapping catalog according to RNA-Seq data; (ii) the enrichment of the *trans*-factor catalog with long non-coding RNAs; (iii) the expansion of the UTR regulatory annotations to include internal ribosomal entry sites and upstream open reading frame (ORFs); (iv) the inclusion of annotations coming from genome-wide RNAi-based gene silencing phenotypic screens; and (v) the improvement of the search engine as well as of the visualization and retrieval systems.

Funding: This work is supported by the University and Scientific Research Services of the Autonomous Province of Trento.

Conflict of Interest: none declared.

REFERENCES

Anderson,P. *et al*. (2009) RNA granules: post-transcriptional and epigenetic modulators of gene expression. *Nat. Rev. Mol. Cell Biol.*, **10**, 430–436.

Berglund,L. *et al*. (2008) A gene-centric human protein atlas for expression profiles based on antibodies. *Mol. Cell Proteomics*, **10**, 2019–2027.

Cook,K.B. *et al*. (2011) RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.*, **39** (Suppl. 1), D301–D308.

Darty,K. *et al*. (2009) VARNA: interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, **25**, 1974–1975.

Friedel,C.C. *et al*. (2009) Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. *Nucleic Acids Res.*, **37**, e115.

Fujita,P.A. *et al*. (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39** (Suppl. 1), D876–D882.

Grillo,G. *et al*. (2010) UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res*., **38**, D75–D80.

Gruber,A.R. *et al.* (2011) AREsite: a database for the comprehensive investigation of AU-rich elements. *Nucleic Acids Res.*, **39**, D66–D69.

Hsu,S.D. *et al*. (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res*., **39**, D163–D169.

Jacobs,G.H. *et al*. (2009) Transterm: a database to aid the analysis of regulatory sequences in mRNAs. *Nucleic Acids Res.*, **37**, D72–D76.

Keene,J.D. (2007) RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet.*, **8**, 533–543.

Khorshid,M. *et al*. (2011) CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Res.*, **39** (Suppl. 1), D245–D252.

Sherry,S.T. *et al*. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Skinner,M.E. *et al*. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.

Xiao,F. *et al.* (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, **37** (Suppl. 1), D105–D110.