UNIVERSITY OF TRENTO

International Doctoral School in Biomolecular Sciences

XXV Cycle

"MAPPING OF POST-TRANSCRIPTIONAL REGULATORY NETWORKS BY MEANS OF MECHANISTIC AND HIGH-THROUGHPUT DATA"

Tutor

Alessandro QUATTRONE

CIBIO – University of Trento

Ph.D. Thesis of

Erik DASSI

CIBIO – University of Trento

ABSTRACT

Post-transcriptional regulation of gene expression (PTR) is the process responsible for modulating mRNA levels and the related amount of protein. Initially thought to have a limited impact on cell phenotype, it has become increasingly recognized as a strong determinant of the quantitative changes in proteomes, and therefore a driving force for cell phenotypes. Untranslated regions of mRNAs (UTRs) are the core mediator of this process, containing sequence and structural elements bound by various kind of regulators, which influence nuclear export, localization, stability of mRNAs and their translation rates, as well as capping, alternative splicing and polyadenylation of the transcribed pre-mRNA.

One of the most important classes of PTR factors are the RNA-binding proteins (RBPs), whose human genome complement is at least 800 genes, characterized by the presence of different functional domains. RBPs bind to the 5'UTR of a transcript often to modulate translation initiation, and to the 3'UTR usually to influence its stability or translatability. Another major group of actors in PTR are non-coding RNAs (ncRNAs). Among them are various classes of long ncRNAs (lncRNAs), the intensively studied microRNAs (miRNAs), siRNAs (small-interfering RNAs) and several other RNA types. miRNAs bind to 3'UTRs by means of short regions of perfect sequence complementation or with some mismatches. Both RBPs and ncRNAs bind mRNAs to the so-called cis-elements, found primarily in 5' and 3' UTRs. These elements can be represented as recurring RNA sequences or secondary structures to which the trans factors bind to exert a control over the mRNA.

In order to integrate the available experimental data, we have developed AURA, a database offering a comprehensive view of the phenomena through regulatory data including RBP and miRNA binding sites, cis-element annotations, secondary structures, phylogenetic conservation, SNPs, RNA-editing data, gene expression profiles and more. A dynamic graphical interface allows the user to browse through the UTRs in an easy and seamless way. To further enrich this body of data, we also implemented a pipeline for the identification of hyperconserved elements in human UTRs, which we applied to both 5' and 3'UTRs. We were thus able to recover known and novel PTR mechanisms involving RBPs, including an RBP network controlled by HuR. We are eventually applying the results of these works to infer altered, and thus potentially disease-related, PTR mechanisms in an high-throughput neuroblastoma dataset.

1. INTRODUCTION

Post-transcriptional regulation of gene expression (PTR) is the process responsible for modulating mRNA levels and the consequent amount of protein products. Initially thought to have a limited impact on cell phenotype, it has become increasingly recognized as a strong determinant of the quantitative changes in proteomes [1], and therefore a driving force for cell phenotypes. As shown by Figure 1, untranslated regions of mRNAs (UTRs) [2] are the two non-coding regions upstream (5'UTR) and downstream (3'UTR) of the coding sequence in the mRNA. They are the core mediator of this process, containing sequence and structural elements, called cis-elements, which are bound by various kind of regulators to influence nuclear export, localization, stability of mRNAs and their translation rates, as well as capping, alternative splicing and polyadenylation of the transcribed pre-mRNA.

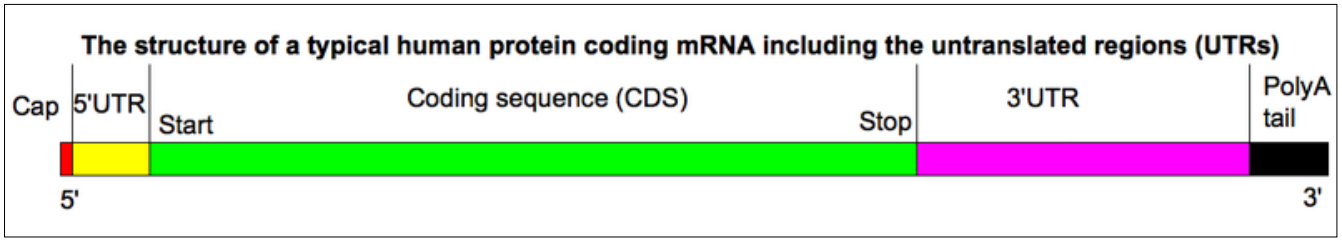


Figure 1: Structure of the human messenger RNA. *The human messenger RNA (mRNA) is composed by an upstream cap, which protects it from RNases and allows the recognition by the ribosome, the 5' untranslated region (important for modulation of translation initiation), the coding sequence which contains the protein sequence to be translated, the 3' untranslated region (which mediates stability and translatability of the messenger) and the Poly(A) tail, which protects the mRNA from degradation and promotes its export from the nucleus into the cytoplasm.*

1.1 RNA-binding proteins

The main role-players at this level of gene expression regulation are RNA-binding proteins (RBPs), non-coding RNAs (of which miRNAs are the most known and studied) and cis-elements. The human genome complement of RBPs is composed at least by 800 genes[3, 4, 5] which are characterized by the presence of different functional domains[6] among which the most represented are, according to the latest release of Ensembl (Ensembl 68), the zinc-finger C2H2 domain (787 genes), the RNA-recognition motif (RRM, 233 genes), the sterile alpha motif (SAM, 93 genes) and the K-homology domain (KH, 38 genes). The top ten RNA-binding domains, sorted according to the number of genes in which they are contained, are listed in Table 1. The most common domain, RRM, is about 90 amino acids long and contains a consensus sequence called RNP-1, which is eight amino acids long. The

typical RRM domain is composed by two alpha-helices with side chains stacking with RNA bases and by four anti-parallel beta-strands.

RBPs bind to the 5'UTR of a transcript often to modulate translation initiation, and to its 3'UTR often to influence its stability or translatability[3]; but they have also been well characterized for modulating splicing of the pre-mRNA, mRNA nuclear alternative polyadenylation, mRNA export, mRNA localization in the cytoplasm and mRNA cytoplasmic polyadenylation[7]. Target transcripts, sequence and secondary structure specificity are currently known just for a very small subset of this class of proteins. Experimental techniques such as SELEX[8], RIP-chip[9] and RNAcompete[10] were first developed in order to tackle this problem; nowadays, thanks to the advent of next-generation sequencing, we can exploit methods such as CLIP[11], PAR-CLIP[12] and iCLIP[13] to probe for all targets identities and binding sites of a specific RBP at once. Still, the fraction of RBPs for which these data are available is rather limited.

Domain	Description	Number of genes
ZNF C2H2	Zinc-finger C2H2	787
RRM	RNA-recognition motif	233
DEAD	DNA/RNA helicase	108
SAM	sterile alpha motif	93
KH	K-homology domain	38
G-patch	G-patch domain	30
DS_RBD	Double-stranded RNA binding	22
PAZ	Argonaute/Dicer protein domain	10
PIWI	Piwi proteins domain	8
PUM	Pumilio RNA-binding repeat	4

Table 1: Most frequent RNA-binding domains. *The table lists the ten more frequent RNA-binding domains in human genome proteins. Domain name, short description and number of genes in which it occurs are shown.*

1.2 Non-coding RNAs

MicroRNAs (miRNAs) are short single-stranded RNAs (around 21-23 nucleotides) which bind usually to the 3'UTR of a transcript (even though there is now some evidence indicating binding in the 5'UTR, see for instance [14]) by means of short regions of either perfect sequence complementation (which leads to increased transcript degradation) or with some mismatches (which promotes instead translational repression and increased degradation)[15]. Currently, around 1500 miRNAs are annotated in the human genome, a number being continuously refined by next-generation

sequencing experiments, which are uncovering new members of this class. As shown by Figure 2, pri-miRNA are transcribed in the nucleus as hairpins, which are then exported into the cytoplasm and processed to mature single-strand RNAs by Drosha, Pasha and Dicer proteins; mature miRNAs can then exert their repressive function by associating with Argonaute to form the RISC (RNA-Induced Silencing Complex). A lot of work has been devoted to miRNAs since their discovery in 1993: software tools able to predict miRNA-target interactions are many and employing the most different approaches. Among these, the most used are TargetScan[16], PITA[17] and miRanda[18]. Experimentally validated miRNA binding sites are less numerous, but still significant: sites for several hundreds of miRNAs are available through databases such as miRTarBase[19] and miRecords[20]. Several other types of non-coding RNAs exist, including various classes of long ncRNAs (lncRNAs, which involvement in PTR starts to be supported by several evidences), siRNAs (small-interfering RNAs) and then piRNAs (piwi-interacting RNAs), snoRNAs (small nucleolar RNAs), snRNAs (small nuclear RNAs) and more.

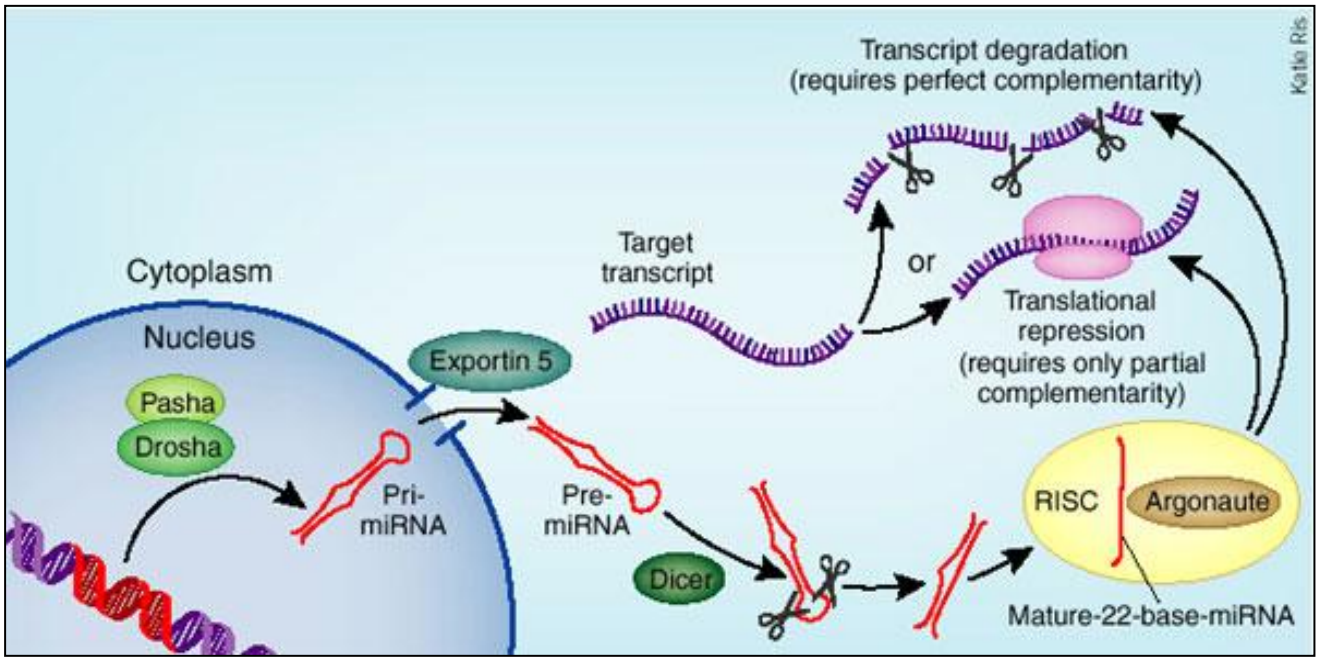


Figure 2: miRNA processing. *Pri-miRNA is transcribed in the nucleus, processed by Drosha and Pasha, exported in the cytoplasm and finally processed into single stranded miRNA by Dicer. At this point, the miRNA complexes into the RISC and can repress translation by hybridizing to the 3'UTR of the target transcript. If perfectly complementary binding occurs, the miRNA targets the mRNA for degradation [21].*

1.3 Cis-elements

Both RBPs and ncRNAs bind to mRNAs in the so-called cis-elements, found primarily in 5' and 3' UTRs. These elements can be represented as recurring RNA sequences or secondary structures shared by a number of transcripts and defined by a pattern, to which the trans factors bind to exert a control over the mRNA. A well-known example of cis-regulatory elements are the AU-Rich Elements (AREs)[29], motifs rich in Us with some interspersed As or Gs shared by several thousand 3'UTRs and bound by a large number of RBPs (the so-called ARE binding proteins, ARE-BP) of which at least 23 are known[29]. A number of tools are available to predict ARE presence in a transcript, exploiting the various identified patterns of ARE occurrence. Another well characterized class of UTR cis-elements are the Iron Response Elements (IREs), which help in coordinating cellular iron homeostasis at the translational level[30] by means of the Iron Response Proteins (IRP). Figure 3 details the IRP-mediated mechanism of translation inhibition by IREs. Various other classes of cis-elements have been characterized and experimentally validated in one or more transcripts: identification of all their occurrences throughout the genome is still for the most part achieved by the application of pattern-based predictive tools such as Transterm[31].

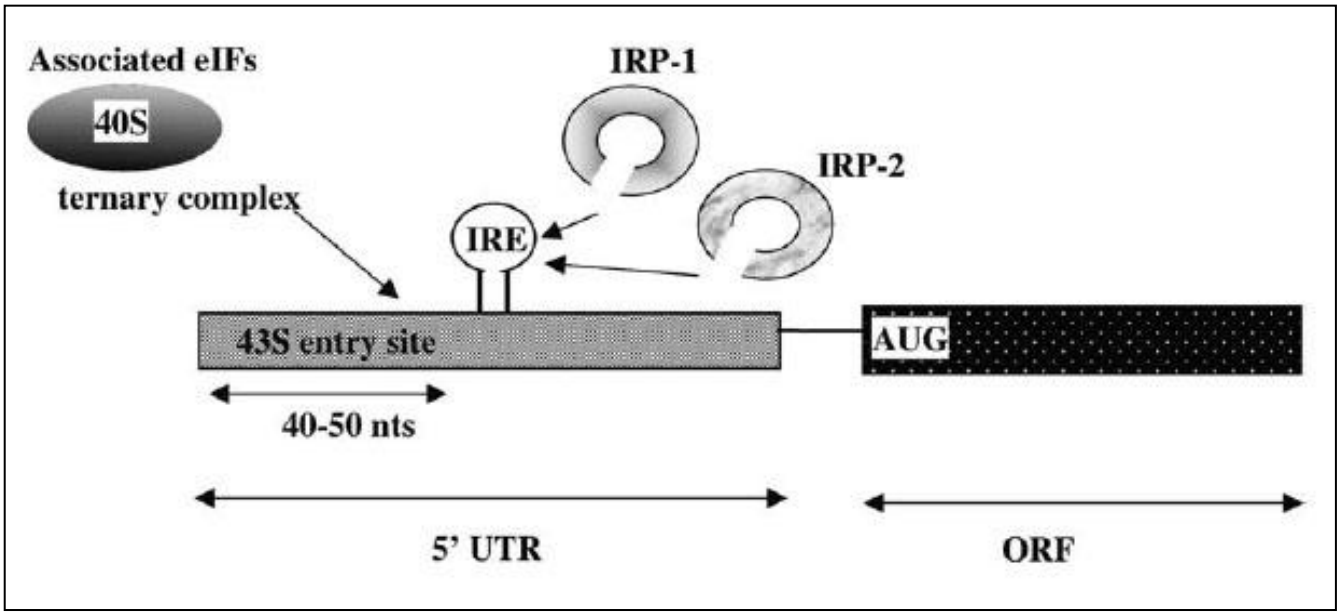


Figure 3: IRE-mediated translation inhibition. *Binding to the IRE in the transcript 5'UTR, IRP1 and IRP2 proteins prevents translation pre-initiation complex formation[32].*

1.4 The landscape of PTR data

The last years have seen a rapid increase in publications and resources dedicated to the analysis of PTR determinants, aimed at trying to unravel associated mechanisms of gene expression regulation. Of the more than twenty functioning and updated resources we identified, many are dedicated to non-coding RNAs [16,17,18,19,20,21,22,23,24,25,26,27] (to microRNAs in particular, with several prediction tools, but also to lncRNAs), only a couple databases focus on UTR annotations, and a comparable number deals with RBP, RBP-target interactions [33,34,35,36] and cis-elements [31,37,38,39,40,41,42]. Most of these resource do not proceed to the integration of the different types of determinants involved in PTR, thus providing only a very partial picture of the phenomena we are studying. Furthermore, while the limited number of high-throughput datasets is more visible (and most of the times inserted in a database as soon as it is generated), many mechanistic results still lie in the literature without being added to any database, thus loosing valuable pieces of information for a field in which the available data is quite limited. Our ability of tracing comprehensive networks of PTR, involving the different factors at play, and to precisely reconstruct the regulatory mechanisms acting on mRNAs is thus hampered by this lack of integration and scarcity of data. Fragmentation is therefore the dominant word in this field at present: this leads to a difficulty in handling the available information, both in terms of quickly finding data and actually being able to find it, preventing the PTR community to build on the amount of facts already established in the last years.

1.5 Our approach

In order to tackle this issue, we settled on implementing an integrative meta-database of post-transcriptional regulation: the Atlas of UTR Regulatory Activity (AURA). AURA is a manually curated and comprehensive catalog of human mRNA untranslated regions (UTRs) and UTR regulatory annotations; it records non-redundant, direct and experimentally assessed interactions of RNA binding proteins and microRNAs with human UTRs, along with cis-elements and several other types of annotations: among these are as SNPs, phylogenetic conservation, RNA secondary structure, gene expression profiles and RNA editing data. We focused on providing a dynamic and user-friendly graphical interface, accessible also to command-line averse biologists, which allows to perform complex queries and looking at the data both from an UTR-based or a trans factor-based point of view. Through the realization of a semi-automatic update pipeline and the availability of several ways to access the data, even in a programmatic fashion, we aim at providing a complete and effective tool which will allow and empower the discovery of novel PTR networks and mechanisms.

Another direction of our work is focused on discovering new cis-elements in UTRs and map the networks in which they are involved. In order to do so, we decided to focus on phylogenetic conservation: sequence evolutionary conservation in UTRs is indeed an aspect neglected by most works devoted to the identification of PTR-related cis-elements. Precedent works trying to identify

functional regions through phylogenetic conservation do exist but, excluding few works, none has focused on UTRs as interesting regions. Still, as no selective pressure on protein functionality applies to UTRs, these are unconstrained to change their sequence or structure just to fulfill their regulatory purpose: accordingly, highly conserved sequences or structures in orthologous genes would likely point to elements potentially endowed with regulatory activity. It is thus of remarkable interest to identify evolutionary highly conserved sequences in UTRs, which we called HCE (Hyper Conserved Elements). We therefore decided to implement a software pipeline allowing for such a search, both for 5' and 3'UTR, in a large set of vertebrate species on a wide phylogenetic distance. Once these regions were obtained, we proceeded to identify groups of related motifs, looked for a benchmark of correctness for our algorithm and a cluster of HCE-bearing mRNAs whose encoded proteins carry the same motif, so defining a translational network of RBPs controlled by HuR, another RBP.

Finally, we proceeded to apply the results of the previous work, AURA in particular, to discover altered, and thus potentially disease-related, PTR mechanisms in an high-throughput neuroblastoma dataset. Neuroblastoma is the most common extra-cranial solid cancer in childhood and the most common cancer in infancy, and arises from the neural crest of the sympathetic nervous system. It most frequently originates in adrenal glands. Its most aggressive form (high-risk) bear the genomic amplification of the MYCN gene locus, and its prognosis is extremely poor; low-risk neuroblastoma presents instead fewer genomic alterations and often has a good prognosis. Our dataset is composed by total and polysomal RNA profiling of thirteen neuroblastoma cell lines. We intersected factor-target relationships contained in AURA with the differentially expressed genes (DEGs) of this datasets composed by matched total and polysomal microarray samples. The histone genes theme emerged as the most enriched and the composing mRNAs were up-regulated. We believe that such an example clearly stands for the usefulness and power of an integrated data approach for the analysis of PTR, even in complex diseases such as cancer.

2. RESULTS

This section will present the results obtained in the three main works on which I focused during my doctoral period, linking them to the attached papers and highlighting my specific contributions to each of them. We will start by describing AURA, its implementation details, the different kind of data it contains and the features its interface offers to its users. We will then proceed to portray the HCEs (Hyper Conserved Elements) identified by our pipeline in the UTRs, detailing the various emerging functional themes and, in particular, a fully post-transcriptional network of mRNAs coding for RRM-type proteins we uncovered. Next, the first results of a total versus polysomal gene expression profiles comparison over a neuroblastoma dataset will be described, introducing an histones-related network resulting from the application of AURA capabilities to differentially expressed genes produced by this analysis. Eventually, we will briefly describe a review about tools and databases dedicated to PTR which I also wrote during my doctoral period.

2.1 The Atlas of UTR Regulatory Activity (AURA)

The Atlas of UTR Regulatory Activity (AURA, available at http://aura.science.unitn.it) is a database aiming at providing a comprehensive overview of currently available data on post-transcriptional regulation of gene expression. It is built in such a way to allow the simultaneous display of all annotations and regulatory events concerning an UTR, thus making possible to infer significant combination of events for the phenomena under study. We decided to consider and use only experimentally verified data (the only exception being the AREs); consequently, ten different databases have been integrated, partially or completely, into AURA: UCSC (UTR annotations, phastCons phylogenetic conservation and secondary structure folding only), AREsite, DARNED, dbSNP, miRTarBase, miRecords, RBPDB, starBase, UniprotKB (detailed genes description only) and ArrayExpress (gene expression profiles in various tissues and diseases are obtained in real-time through the GXA web programming interface). In addition to this amount of data, a thorough literature search returned 1200 more binding sites, which we also added to the database. Table 2 illustrates the most relevant figures for AURA at its current release, highlighting the fact that only a limited fraction of RBPs and miRNAs have been object of experiments aimed at discovering their targets and related binding sites.

Feature	Data quantity
5'UTRs	64550
3'UTRs	62973
UTR secondary structures	117119
Transcripts	63138
Genes	29345
Binding sites	406174
RBPs	100
miRNAs	311
Cis elements instances	19681
SNPs	775488
Transcripts halve-lives	31550
References	2171
Referenced databases	10

Table 2: Most relevant AURA figures. *The table lists the figures summarizing the data contained in AURA: in particular the number of binding sites, of RBP, miRNAs and cis-elements involved. Auxiliary annotations figures are also included. The references item represents the number of papers relating to data contained in the database.*

A great deal of attention has been placed into realizing a dynamic and user-friendly graphical interface. The website was implemented with the Django Python platform and is all AJAX-based, meaning it updates just the part of the pages which need to be, avoiding whole-page reload times and hassle. Two search modalities are available: 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 search 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. Figure 4 illustrates an example search results page for this modality, highlighting the intuitive interface and its various options. On the other hand, the latter query results in a disambiguation list where all the trans-factors, whose names or synonyms contain the search term, are shown; once the user selects the intended trans-factor, AURA returns the list of its target UTRs. These UTRs can be grouped by GO slim categories[1] or by chromosome mapping. Furthermore, before launching the search, the user can select to filter the results by a combination of supporting experimental evidences. Figure 5 shows the results page of this search type, with trans-factor details and target UTRs grouped by GO terms in this case. This kind of visualization allows a first inference on the role of the trans-factor by just considering the functional grouping of its target, thus empowering the selection of UTRs to be analyzed in detail.

Both search modes result, for the selected UTRs, in a page composed by a genome-browser like view for each of the UTRs: this type of display, highly dynamic, allows the user to explore the whole range

[1] http://www.geneontology.org/GO.slims.shtml

of interactions, or focus on a specific part of the sequence, or a kind of factor (data can even be hidden from the visualization) concerning the analyzed UTR. Selected UTRs are shown in an "UTR view", illustrated by Figure 6, consisting of two 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 Human Protein Atlas (HPA) database. Also shown are the overall conservation, which is the mean PhastCons single nucleotide conservation score for the UTR, and the corresponding transcript half-life according to a transcriptome-wide mRNA stability measurement assay.

- The AURA sequence browser, based on the JBrowse platform, 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. In the "SNP" track, AURA integrates the single nucleotide polymorphisms (SNPs) recorded in the dbSNP database allowing the user to combine with the other annotation tracks to look for variations of potential impact in PTR. The "RBP" track contains the RBP binding sites, while the "miR" track contains the microRNAs binding sites. The "RNA editing" track contains data about the UTR bases which have been found to be edited (mostly A>I conversions) after transcription. 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 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. All annotations are linked to their source, either a PMID indicating the publication or an ID relating to the original database (as in the case of dbSNP).

Furthermore, the predicted secondary structure of the UTR can be visualized through a button over the UTR view. An interactive viewer allows the user to zoom, tilt and move the secondary structure, in order to allow focusing on relevant details. Binding sites, SNPs and evolutionary conservation annotation are laid on top of the structure by means of a color scale (conservation) and color-coded highlighting. This particular view can immediately reveal if a binding site or a SNP is associated to a particular structure such as an hairpin or a bulge, and can help in guiding the investigation for further evidence. Eventually, AURA provides the user with multiple ways of grouping gene expression results, retrieved from the Gene Expression Atlas, and related to the gene locus of the selected UTR. Results are reported in tables where a row corresponds to a condition, while the columns, in order, show the number of times the gene was observed to be up- or down-regulated with respect to its mean expression value and the significance of the measure (log10 P-values). In case of a trans factor search, whenever data are available, 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.

Figure 4: Gene-based search in AURA. *The figure displays the interface presented to the user when searching for the UTRs of a specific gene. On the left one can see the panel describing the gene (name, synonyms and function) and depicting all its different UTRs along with the exon-intron structure of the related transcripts. UTRs can then be dragged onto the right panel which, as a cart-like feature, allows to include UTRs from different transcripts and genes at once. Selected UTRs can then be explored in detail through a genome-browser like view.*

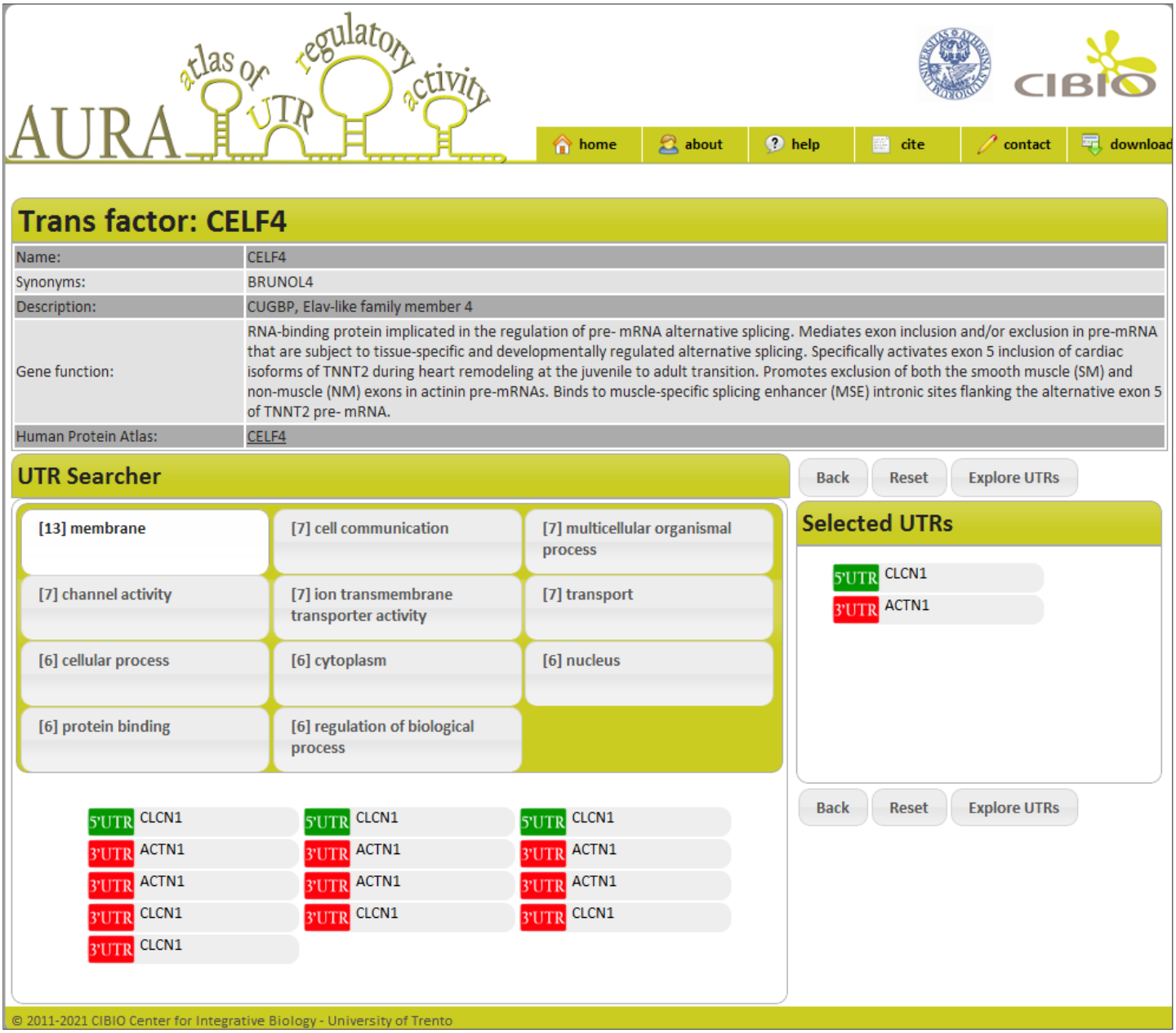


Figure 5: Trans factor-based search in AURA. *The figure displays the interface presented to the user when searching for targets of a trans-factor (either an RBP or a miRNA). Top panel gives the basic annotation for the trans-factor of interest, including the binding motif weblogo (computed from positional frequency matrixes) when available. The lower left part contains the list of target UTRs, grouped by Gene Ontology term, providing an indication of targets function. By clicking on each term, the list of belonging UTRs appear, allowing to select them. As previously stated, UTRs can be selected by dragging them onto the right-side panel: the selected ones will then be explored in detail.*

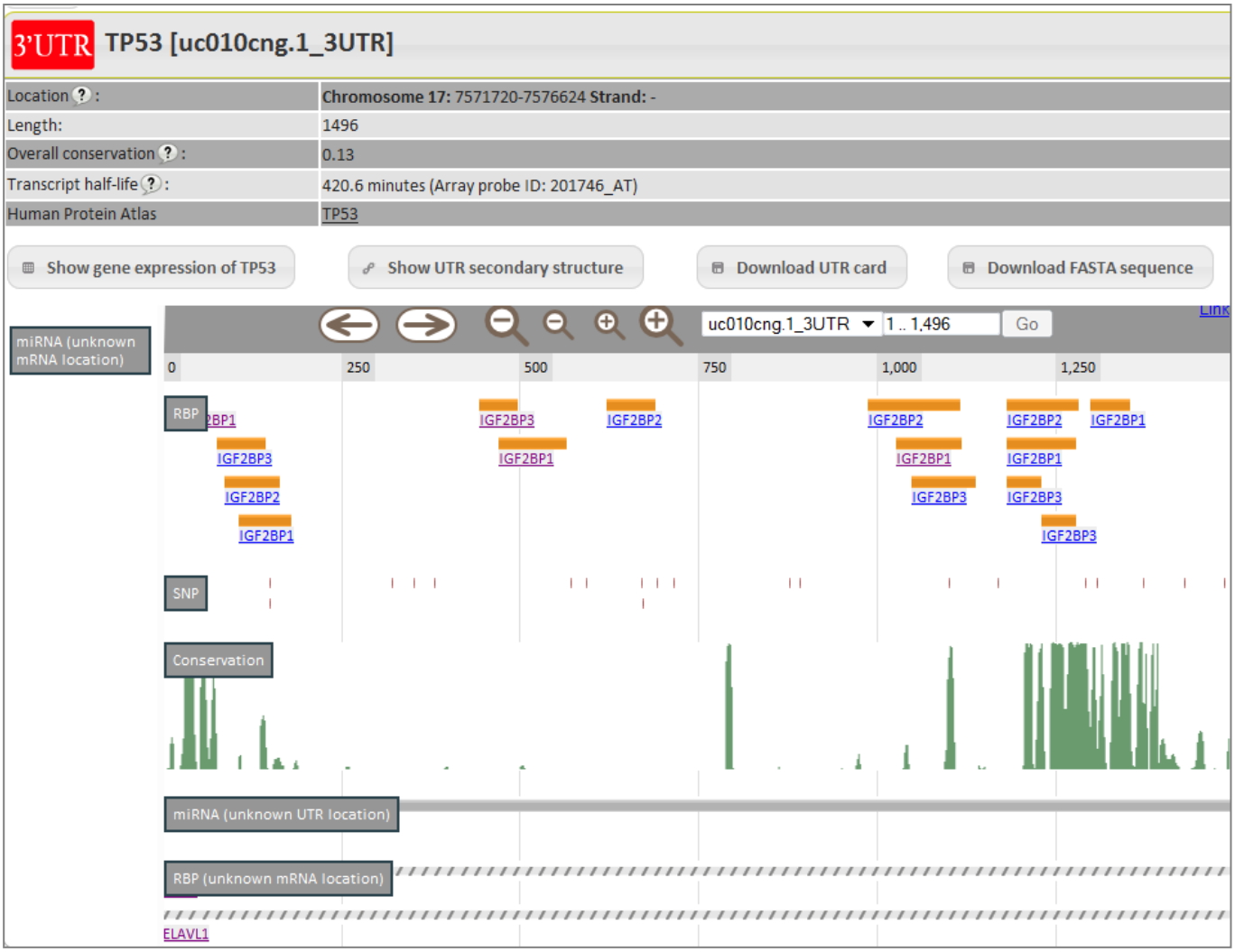


Figure 6: UTR exploration interface in AURA. *The figure shows the genome browser-like interface which the user can take advantage of to explore all data concerning an UTR at once. Top panel gives the basic annotation for the currently displayed UTR, including overall phylogenetic conservation and transcript half-life. The lower left part contains several track displaying RBP and miRNA binding sites, cis-elements, SNPs and evolutionary conservation tracks. The UTR can be zoomed and sequence can be scrolled to focus on the precise region of interest. Gene expression profiles and secondary structure of the UTR can be accessed via the buttons on top of these tracks.*

Aside from accessing and searching the database through its web interface, a more experienced or bioinformatics-oriented user can take advantage of the other options we provide to mine the data contained in AURA: first of all, the complete set of annotation of a single UTR, which we call UTRcard (including secondary structure, conservation, binding sites, and more), is downloadable from the UTR view, by composing the URL of the UTR in the browser or through a script; the whole MySQL database can also be downloaded and replicated on a local machine (schema description is provided); eventually, a BioMart, called AuraMart, is available and let users query the data in a simple and standard way through the BioMart platform (the same as used by Ensembl BioMart and many other major websites): having already used a Mart, an user will just need to know which data he or she wants to extract from the database, being able to exploit the query knowledge he or she has already acquired. Batch search and analysis tools are currently being developed and will soon be integrated in AURA.

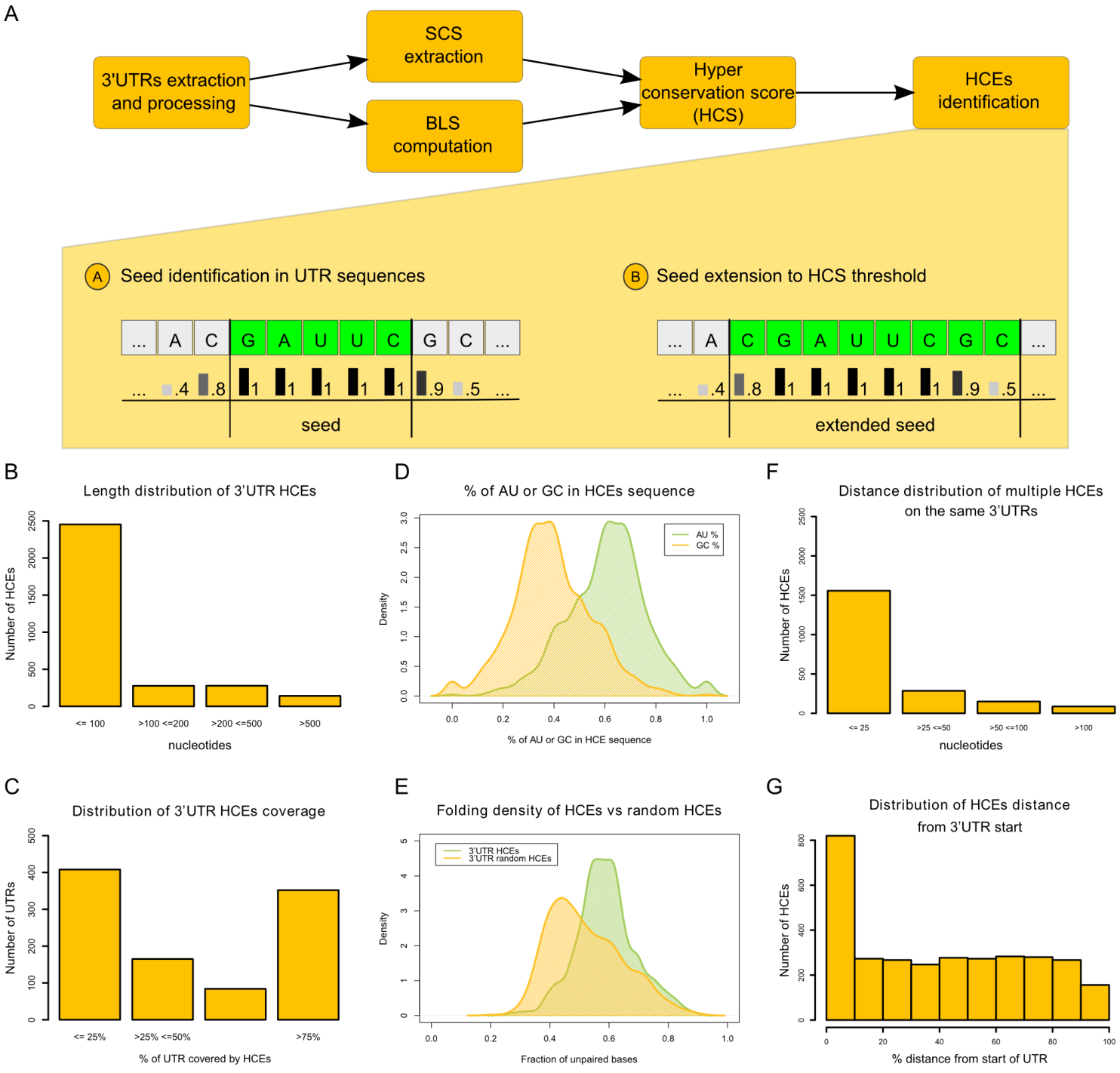
My contribution in the realization of AURA started with the implementation of the underlying database schema, including the design of entities and relationships to accommodate all data now present in the database and the server setup. I then realized a significant part of the graphical interface and of the underlying features, setting up and maintaining the website server. Eventually, I collected data from some of the ten integrated databases (all basic annotation from UCSC including conservation, SNPs) and performed a literature search for binding sites which was split in equal parts between all authors. I am currently managing and keeping AURA updated with new data.

2.2 Hyper Conserved Elements (HCE) identification

We firstly aimed at identifying HCEs in the 3' UTRs of the human exome by a seed extension strategy: these were derived from the human genome, by a custom pipeline (Figure 7A). We took advantage of the 44-way vertebrate UCSC alignment, from which we derived the phastCons sequence conservation score ([43], which we call SCS) for each base of the exon sequences annotated as 3'UTRs. We also computed, for each base, the Branch Length Score (BLS), defining the degree of sharing of conservation among the vertebrate species considered [44]. We firstly obtained short footprints of very high phylogenetic invariance represented by fully conserved 5-bases seeds (SCS >= 0.95 and BLS >= 0.85); we then extended these seeds upstream and downstream until they reached a preset threshold on our conservation score, which we called HCS (Hyper Conservation Score, computed for each base of the UTRs as the weighted average of SCS and BLS). The 3'UTR-HCE identification algorithm produced 3149 HCEs, belonging to 1010 3'UTRs, which corresponded to 877 genes. At least one 3'UTR HCE is thus present in only 1,8% of the total human 3'UTRs, and collectively HCEs cover only 0.47% of the 3'UTR space, making them extremely rare. They have an average length of 100 bases, but their length distribution (Figure 7B) is such that more than 77% of their total number is shorter, being only 4.5% of them over 500 bases. Their UTR coverage (Figure 7C) is instead prevalently low (25% or less of the 3'UTR) or high (75% or more of the 3'UTR). Together, these distributions show that 3'UTR-HCEs are relatively short and that they either occupy a small portion of a 3'UTR or the most of it. When multiple HCEs are present on an UTR, these have a clear tendency to localize in clusters, as indicated by the very small inter-HCE distance, 25 bases or less (Figure 7D), and to be spread along the 3'UTR, with 25% of the HCEs start nucleotides concentrated on the first 10% of the 3'UTR (Figure 7E). These elements are much richer in AU than in GC bases (Figure 7F, p-value 2.2E-16), and are by far more highly structured than random 3'UTR sequences of the same length, being structural density defined by the fraction of unpaired bases in the HCEs secondary structure(Figure 7G, p-value 1.2E-13). To provide a snapshot on HCE architecture diversity, we distributed HCE-bearing 3'UTRs into four classes, depending on their number and coverage. These classes, reported in Figure 7H, efficiently represent this diversity.

We then sought to understand what types of potentially functional cis-acting elements are found in 3'UTR-HCEs. To test for ncRNAs, we compared HCEs sites on 3'UTRs with a set of 15560 experimentally validated microRNAs binding sites extracted from AURA [34] and concerning 88 miRNAs. Only 51 HCEs (1.6%) were found to contain one or more microRNA binding sites, which are 60 in total and involve 33 different microRNAs. We also intersected 3'UTR-HCEs with lncRNAdb [25], a catalog of eukaryotic long non-coding RNAs. Performing a BLAST search yielded 151 statistically significant putative binding sites, at least 12 nucleotides long, involving 132 unique HCEs (4.2%) and 32 different lncRNAs. We performed the same procedure on a set of randomly derived 3'UTR sequences (random HCEs) with the same length distribution as our HCEs and being the set the same size as the total number of HCEs: only 19 (0.6%) random HCEs were found to contain one or more microRNA. Concerning lncRNAs, the blast search yielded 207 statistically significant putative binding

sites, at least 12 nucleotides long, involving 169 unique random HCEs (5.37%) and 39 different lncRNAs. We eventually scanned HCEs and random HCEs for matches with the position-frequency matrixes extracted from RBPDB [35]. Considering only matches with a minimum score of 80% and a matrix length greater than 4, we obtained 1.8 times more matches in the HCEs than in random HCEs (17173 matches for HCEs versus 9443 matches for random HCEs).



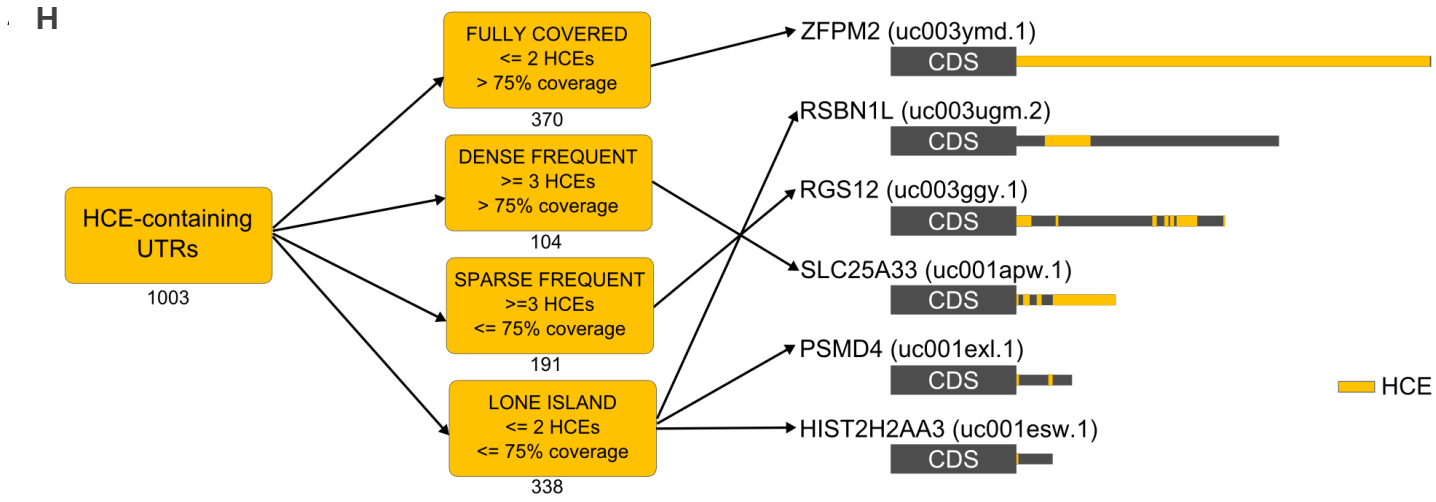


Figure 7: HCEs in 3'UTRs of the human exome are short, scattered and highly structured. *The overall HCE identification pipeline is shown in a), with the lower part detailing the algorithm searching for seeds and extending them to lead to the final HCEs. b-g) highlights the most relevant features of the HCEs: b) shows the length distribution of HCEs and c) the percent coverage of 3'UTRs by these d) displays the AU predominance over CG in HCE base composition and e) the prevalence of highly-structured HCEs; f) displays the distribution of distances between HCEs on the same UTRs and g) the HCEs distance distribution from UTR start, indicated in percent over the UTR length. h) shows the classification of 3'UTRs in four classes according to their HCE content on the right. Numbers below each class box indicates the number of HCE-containing 3'UTRs belonging to the class. On the right, a sample of six HCE-containing 3'UTRs: HCEs are mapped onto their UTR and represented as yellow areas in a grey rectangle representing the full-length 3'UTR. Arrows from class boxes to UTRs indicates which UTR belongs to which class.*

In order to appreciate the whole spectrum of biological functions expressed by 3'UTR-HCE containing genes, we performed an ontological enrichment by means of DAVID[50] (using Gene Ontology, InterPro, Smart, PFAM and KEGG ontologies) on the 877 genes bearing at least one HCE in their 3'UTR. We identified three gene groups endowed with high significance (Figure 8).

The first group is composed by 78 genes involved in chromatin structure (terms "nucleosome", "chromatin assembly", "DNA packaging"), including 51 (53.6%) of the 95 histone genes present in the human genome. It is well known that histone gene mRNAs all have a short 3'UTR, lacking a poly(A) tail, which is bound by the stem-loop binding protein (SLBP) in the cytoplasm to stabilize histone mRNAs and mediate their translation [45]. Alternative to polyadenylation, this mechanism is very ancient and is conserved over a wide evolutionary distance. We therefore hypothesized that the HCEs in the histone 3'UTRs were SLBP binding sites. In order to verify this, we aligned the SLBP binding motif to these HCEs and found a considerable fraction of these to contain a close, where not perfect, match to the known SLBP motif. Therefore, the algorithm we derived to select for HCEs is able to precisely identify cis-elements involved in a conserved and well demonstrated post-transcriptional

regulatory process: we therefore assumed this finding as an effective benchmark for the ability of 3'UTR HCEs to point to circuitries of phylogenetically old post-transcriptional control.
The second highly significant gene set is about the broad activity of transcription and mainly composed by genes involved in its repression. The 137 identified genes suggest that transcription factors like EPC1, TFAP2D and YY1 and cotranscriptional repressors such as FOXP2, MEIS2 and EZH2 can be heavily controlled at the post-transcriptional level, being their 3'UTR almost entirely highly conserved.

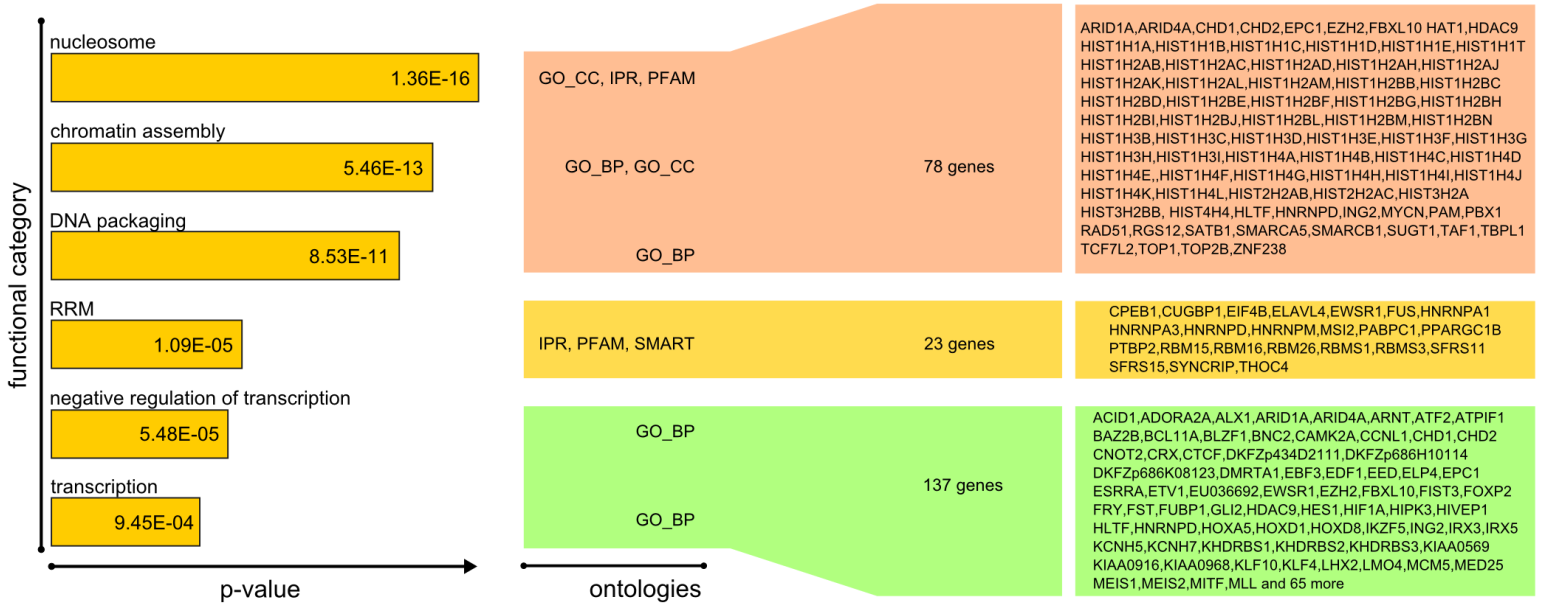


Figure 8: HCEs clusters in genes belonging to three different biological functions. *Ontology enrichment analysis of HCEs-containing genes highlights three groups of genes corresponding to three different biological functions. Multiple ontologies were used to infer possible functional groupings: the results exposed a most significant group composed of genes involved in chromosome assembly, a significant set consisting of 23 genes coding for RRM-containing proteins and a third, less significant group of genes playing a role in transcription. Here are shown the ontology terms clusters giving rise to these groups, along with their enrichment p-value and the final list of involved genes.*

When protein domains enrichment was computed over the 3'UTR-HCE containing genes, the most significant outcome resulted to be the RNA Recognition Motif, the RRM. Of the 23 enriched genes whose protein product contains RRM domains, 17 are experimentally verified RBPs and 14 have an RRM-only architecture, and their mRNA is characterized by 3'UTRs of all four classes, with a prevalence of full (66.7%) and dense frequent (19%). We therefore focused on this protein group to predict a possible RBP regulating some of them, through analysis of their 3'UTR-HCEs. We scanned the HCEs for hidden common elements by the Weeder algorithm, searching for six to twelve bases

long motifs, allowing one to four mutations and observed in at least 25% of the HCEs [46]. The scan produced two reliable 12-bases motifs that can be considered variants of the same motif, as they differ only in two positions. We speculated that this motif could represent an RBP binding site, since a number of these proteins are known to have a preference for 7-8 bases stem-loops [10] We then searched for secondary structure motifs in the same 3'UTR-HCEs by means of the RNAfold [47] and RNAforester [48] algorithms. This analysis resulted in a 17-bases structural motif in the form of a hairpin, whose core loop had a good correspondence (7/12 bases with both sequence motifs; 9 bases out of 12 for sequence motif 2) with the previously identified sequence motifs. Combining the results of both sequence and structure motif searches produced a remarkable concordance, as shown by motifs alignment in Figure 9A, leading us to a hairpin motif shared by 18 out of the 23 RRM genes reported in Figure 9B. Instances of the hairpin motifs in the mapped 3'UTRs are shown in Figure 9C.
After having identified this motif, we noticed its sequence was quite similar to an already known binding sequence for HuR (ELAVL1) protein [10]. In order to verify that our motif was effectively interacting with HuR, we performed a protein pulldown assay, followed by a western blot with anti-HuR antibody. Along with the putative HuR motif, we employed two positive controls for the technique (YB1 and PTB), two mutated and one degenerate loop probes, which design is shown in Figure 10A. As shown in Figure 10B, HuR indeed binds to the probe corresponding to our shared motif. Mutated and degenerated probes show very little recovery of HuR, suggesting that the interaction is specific and depending on the loop sequence and size. Positive controls western blots are shown in Figure 10C and 10D.

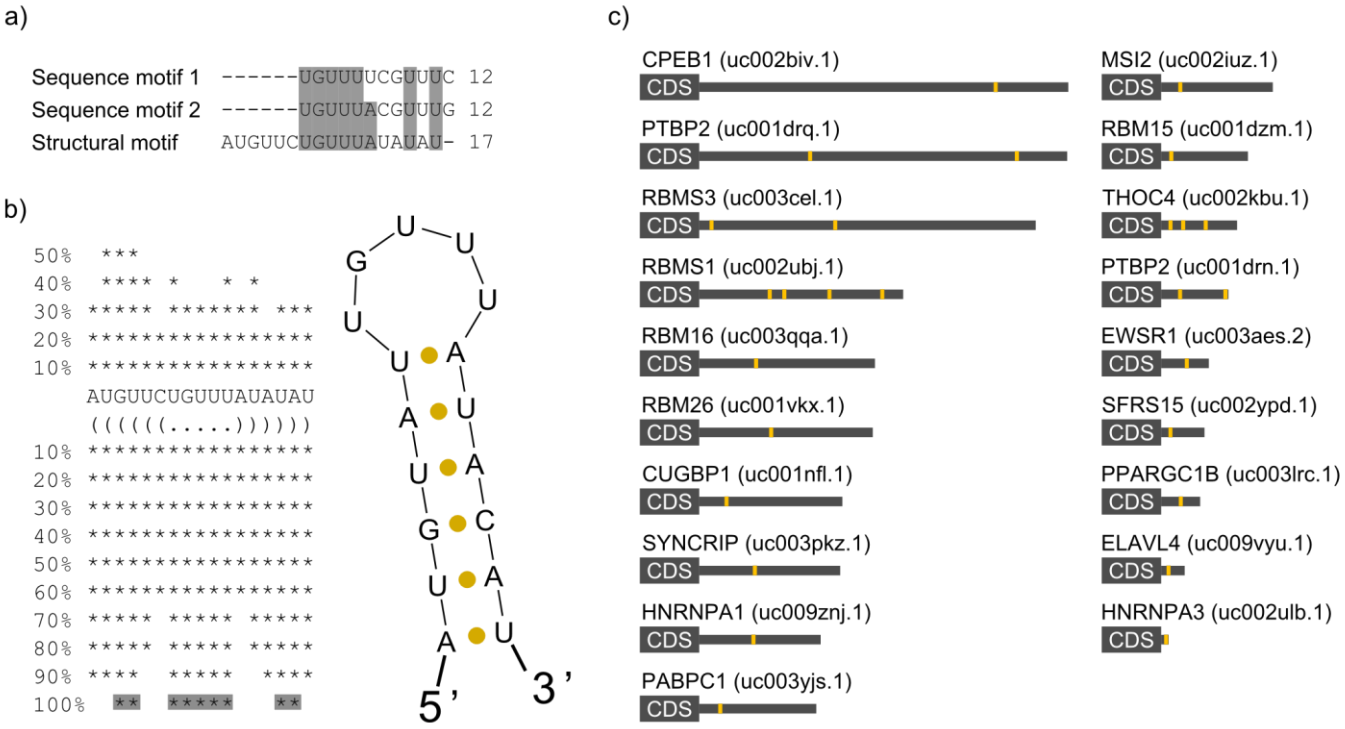


Figure 9: HCE-containing 3'UTRs of the RRM genes subset share a sequence and secondary structure motif. *HCEs contained in the group of RRM genes 3'UTRs were scanned for both sequence and secondary structure motifs. The first search returned two, almost identical, 12-bases motifs; the second one produced a 17 bases hairpin which, after examination by means of a multiple alignment, emerged to contain a 12-bases core markedly similar to previously identified sequence motifs. This core represents the loop part of the hairpin which, as the two searches are quite concordant on it, may indeed represent a binding motif for the key actor of the regulatory network we are trying to uncover. a) shows the alignment between sequence and secondary structure motifs b) shows the secondary structure motif and its sequence/structure motif. c) motif instances (yellow areas) mapped on their respective full length UTR (grey rectangle).*

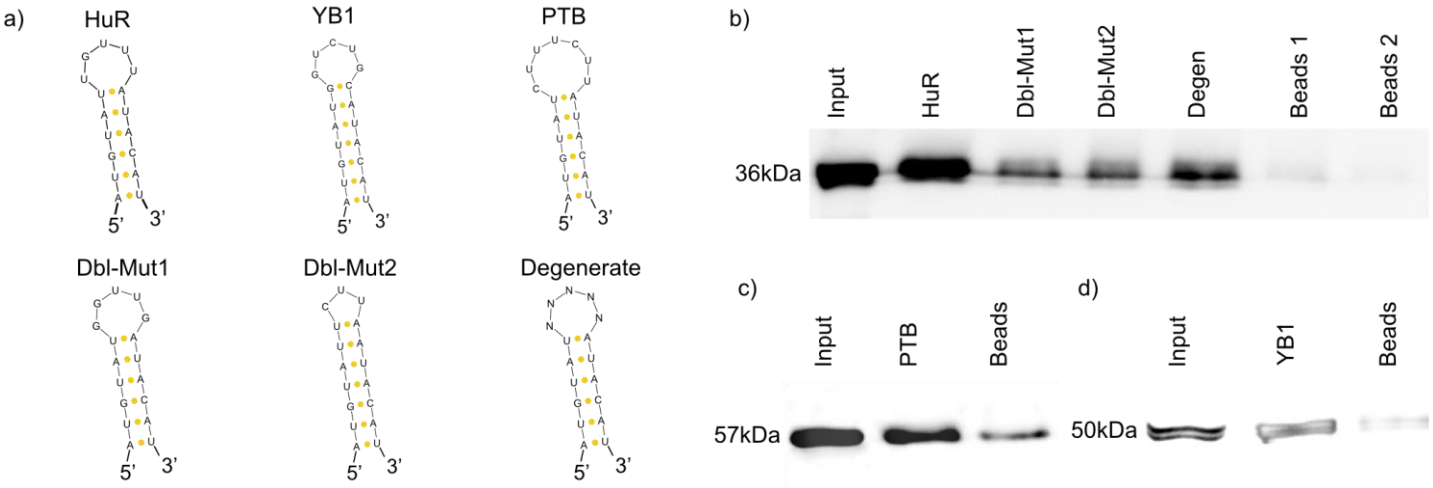


Figure 10: A protein pulldown experiments indicates HuR as the trans-factor binding to the shared motif. *The various RNA probes employed for the protein pulldown experiment are shown in a). HuR pulldown probe: this probe was designed by using the secondary structure motif shown in Figure 9, slightly modifying the lowest part*

of the hairpin so as to make it fold correctly when not in context. The loop part was designed by employing the most probable nucleotides of sequence and structure motifs. Positive controls pulldown probes are YB1 and PTB: their known binding motifs where obtained from the RNAcompete paper [10]. Again, the lowest part of the stem was slightly modified so as to make it fold as desired. Negative controls HuR probes are Dbl-Mut1, Dbl-Mut2 and Degenerate. The Degenerate probe was synthesized by allowing all four nucleotides to be present at each loop position, so to obtain a mixture of probes bearing all the possible 5-mers loops. The Dbl-Mut1 and Dbl-Mut2 probes were obtained by mutating two nucleotides of the original probe loop, in a way to preserve it in the first case and to obtain a 3-mer loop instead of a 5-mer in the second one. b) shows the HuR pulldown western blots. From the leftmost to the rightmost band: input, HuR probe, Dbl-Mut1,Dbl-Mut2, Degenerate probe and denaturized beads bands. As can be readily seen, the hairpin probes bind to HuR with a marked specificity for the correct probe with respect to degenerate and mutated probes. c) - d)PTB and YB1 pulldown. From the leftmost band to the rightmost: input, YB1/PTB probe, and denaturized beads. As shown by Western Blot images, the hairpin probes bind to PTB and YB1 respectively, thus confirming that the pulldown protocol works as expected.

With the motif confirmed to be recognized by HuR, we next sought to understand whether HuR had a marked preference for RRM-containing genes with respect to RNA-binding domains and the most frequent domains in the genome. To compute this enrichment, we took advantage of a HuR PAR-CLIP dataset published by Lebedeva et al. [49]. We extracted all HuR 3'UTR binding sites from this dataset and obtained the genes to which these UTRs belonged. We then computed, by means of the Fisher test, the enrichment of genes containing the most common RNA binding domains (Zinc Finger, RRM, KH,SAM) with respect to the most frequent domains in the genome (IG-like, GPCR superfamily, Serine Threonine kinase and Olfactory Receptor) and to the complete set of RBPs. Results are shown in Figure 11a): RRM domain resulted to be significantly enriched with respect to all these domains and RBPs, being the only RNA-binding domain having a significant enrichment in all cases. This suggests, as was our hypothesis, that HuR has a marked preference for RRM-bearing genes regulation. We then plotted all 3'UTR HuR targets identified by Lebedeva along with our group of RRMs, to highlight overlapping and unique genes of the two sets. The resulting intersection counts are shown in figure 11b, while the network is shown in Figure 11c) and discriminates between genes categories by means of shapes and colors, as shown in the bottom part legend. Nine out of the 23 HCE-containing RRMs are not identified by Lebedeva as being bound by HuR and in particular, 4 of them are among the ones we validated by RT-PCRs. Figure 12a) shows the results of the RIP validation of interaction between HuR and four HCE-containing target mRNAs; Figure 12b) displays the western blot confirming HuR silencing in the cell line we used for the last step, the RT-PCR polysomal validation we performed on four of the 23 genes (shown in Figure 12c)): all of them show a translational repression effect, suggesting a stabilizing effect for HuR when bound to these mRNAs.

My contribution to this work consisted in the realization of the pipeline for the identification of the HCEs: in particular, retrieving the conservation data, writing the source code which computes conservation scores and output these regions, functional analysis of the results and identification of interesting groups of HCEs. I then identified the relevant sequence motif for the RRM group of

proteins and isolated the secondary structure used for the pulldown experiment. After the experimental part verifying HuR binding and the specificity of these interactions, I crossed the group of RRM genes with data in AURA and the other high-throughput works on HuR. Once completely mined, I will insert HCEs into AURA as additional cis-elements, able to provide even more clues on the post-transcriptional regulatory events involving a given UTR under study.

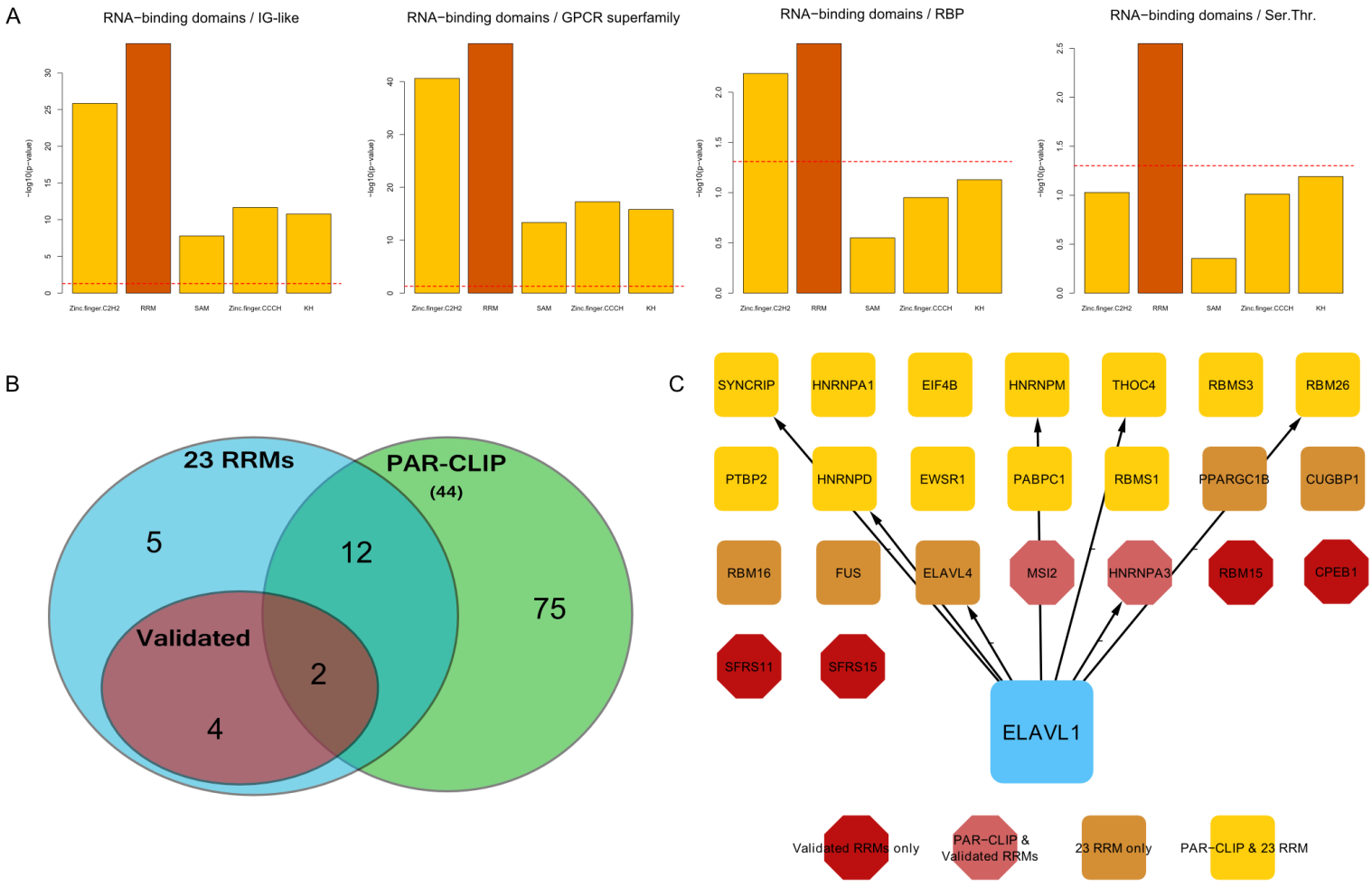


Figure 11: HuR has a preference for binding to the 3'UTR of RRM-type RBPs.

a) shows the enrichment of HuR 3'UTR binding sites for several RNA-binding domains with respect to the most frequent human protein domains and to RBPs as a whole. Data is extracted by the PAR-CLIP experiment published in (44). b) shows a Venn diagram indicating the overlap between our HuR RRM-type mRNA targets and the experimentally identified HuR PAR-CLIP RRM-type mRNA targets. c) displays HuR 3'UTR RRM-type mRNA targets, highlighted in different colors and shapes according to their belonging to our set of 23 mRNAs, to mRNAs we validated by RIP-qPCR and their intersection with the RRM-type mRNA targets from the PAR-CLIP dataset.

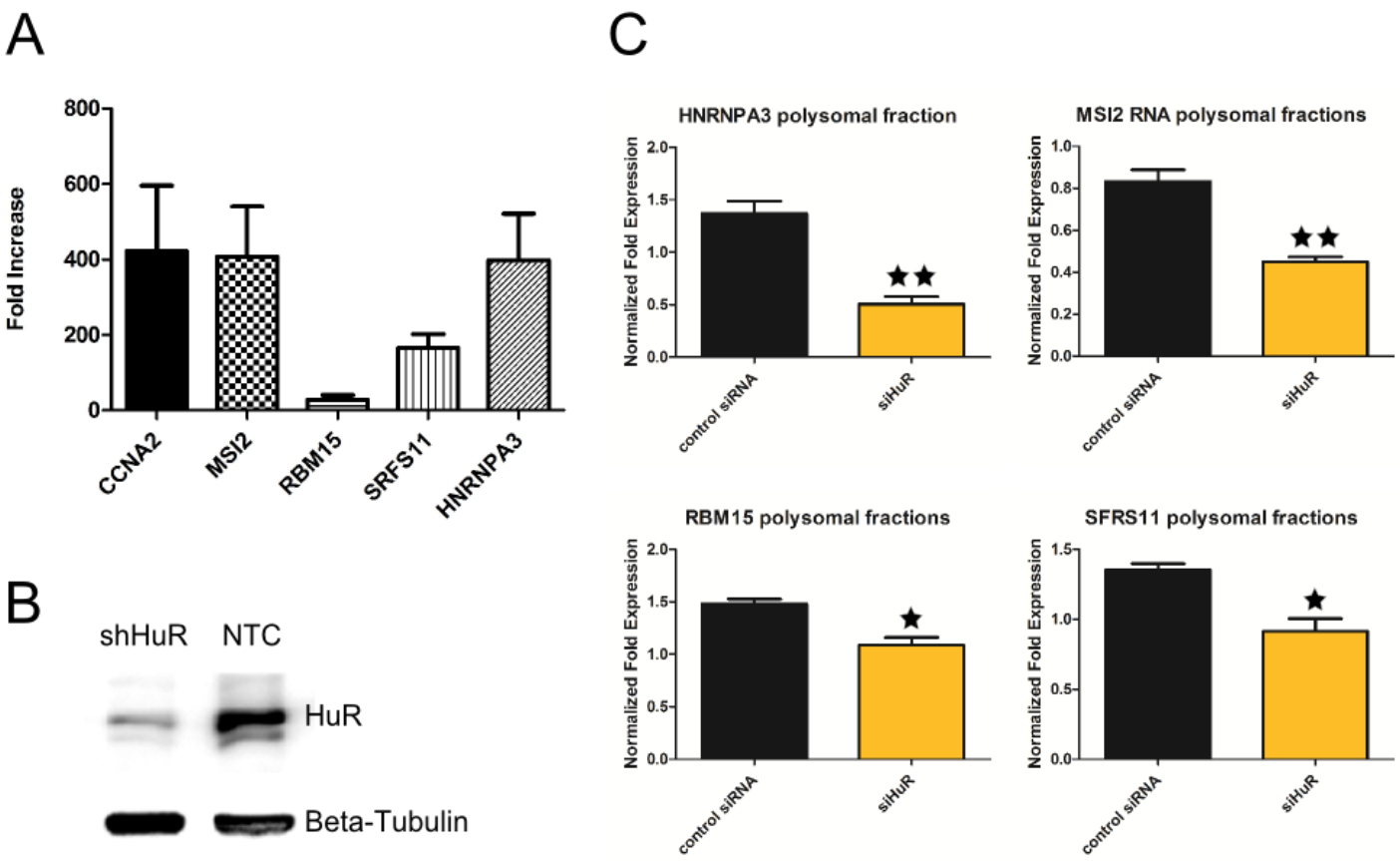


Figure 12: HuR and RRM 3'UTRs interaction properties. The network of HuR binding to mRNAs for RRM-type RBPs is a functional translational network.

a) shows the fold enrichment results (with respect to control) for four predicted RBP mRNAs (plus the CCNA2 mRNA as control) subjected to ribonucleoprotein immunoprecipitation (RIP) from lysates of HuR overexpressing MCF-7 cells and quantitative RT-PCR, demonstrating interaction of HuR with these mRNAs. b) reports the western blot confirming HuR silencing in MCF-7 cell line. Beta-tubulin is used as housekeeping gene. c) shows the statistically significant decrease of mRNA levels for the same four RRM-type RBP mRNAs, indicating a translational enhancing effect of HuR on these mRNAs. Increasing level of significance (<= 0.05, ** <= 0.01) is indicated by one or two stars.*

2.3 PTR networks in neuroblastoma

We eventually proceeded to analyze a set of microarrays performed on neuroblastoma cell lines. Our dataset was composed by 13 samples, profiled at both the total and the polysomal RNA levels by means of Agilent Human 44k microarrays. All samples bear the MYCN gene locus amplification, marker of the most aggressive form of the disease; other genomic alterations are also present but not uniformly across the cell lines. We started by quality filtering and quantile normalizing the 26 arrays: this was done by means of the R software and the Bioconductor package. We then proceeded by employing four algorithms to compute the differentially expressed genes (DEGs), namely PCA, RankProd, SAM and the T-test (again by means of R and by selecting a p-value threshold of 0.01 for all four algorithms). The resulting proportion of DEGs (illustrated by Figure 13) varies widely between the methods, with PCA producing just 118 genes as differentially expressed and SAM returning as much as 2743 genes. We thus selected RankProd-derived DEGs as our reference list of up- and down-regulated genes (1335 genes).

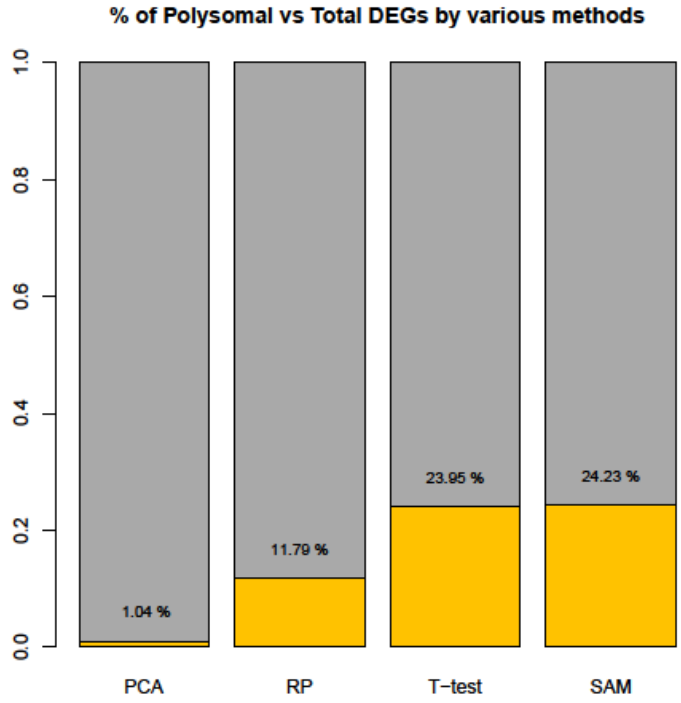


Figure 13: Different methods identify largely varying degrees of DEGs in the comparison in polysomal versus total mRNAs. *We performed DEGs selection with four different methods, perceived as progressively more stringent. Indeed, while the t-test and SAM identify around 24% of the genes as significant DEGs, RankProd falls down to 12% and the PCA calls little more than 1% of the genes as differentially expressed. We selected RankProd DEGs as our reference genes for subsequent analysis.*

In order to understand which processes and functions were represented in the DEGs groups, we subjected the up- and down-regulated lists to functional enrichment analysis by means of DAVID [50], employing the Gene Ontology, InterPro, Smart, PFAM and KEGG ontologies. Enrichment p-value were corrected for multiple testing by the Benjamini-Hochberg correction. This analysis highlighted several themes (coherent grouping of terms) as significant: in particular, a theme we called Histone, composed by terms such as "*Histone*", "*nucleosome*", "*chromatin assembly*", "*Histone-fold*" and many others, was found to be highlighted by most of the employed ontologies with a consistently low p-value (lowest is 5.5E-28) and including 64 genes. The other themes, as

shown by Table 3, were either composed by a low number of genes or supported by just one ontology, making them less interesting to focus on. We thus decided to pursue the histone theme and perform further analysis on it.

Functional theme	Status	Ontologies	# of genes	Average theme p-value (-log10)
Histone	Up-regulated	GO, Smart, Interpro, PFAM	64	15.659
Mitochondrion	Up-regulated	GO	31	2.537
ATP-binding	Down-regulated	GO	88	13.661
mRNA processing	Down-regulated	GO, KEGG	29	5.674

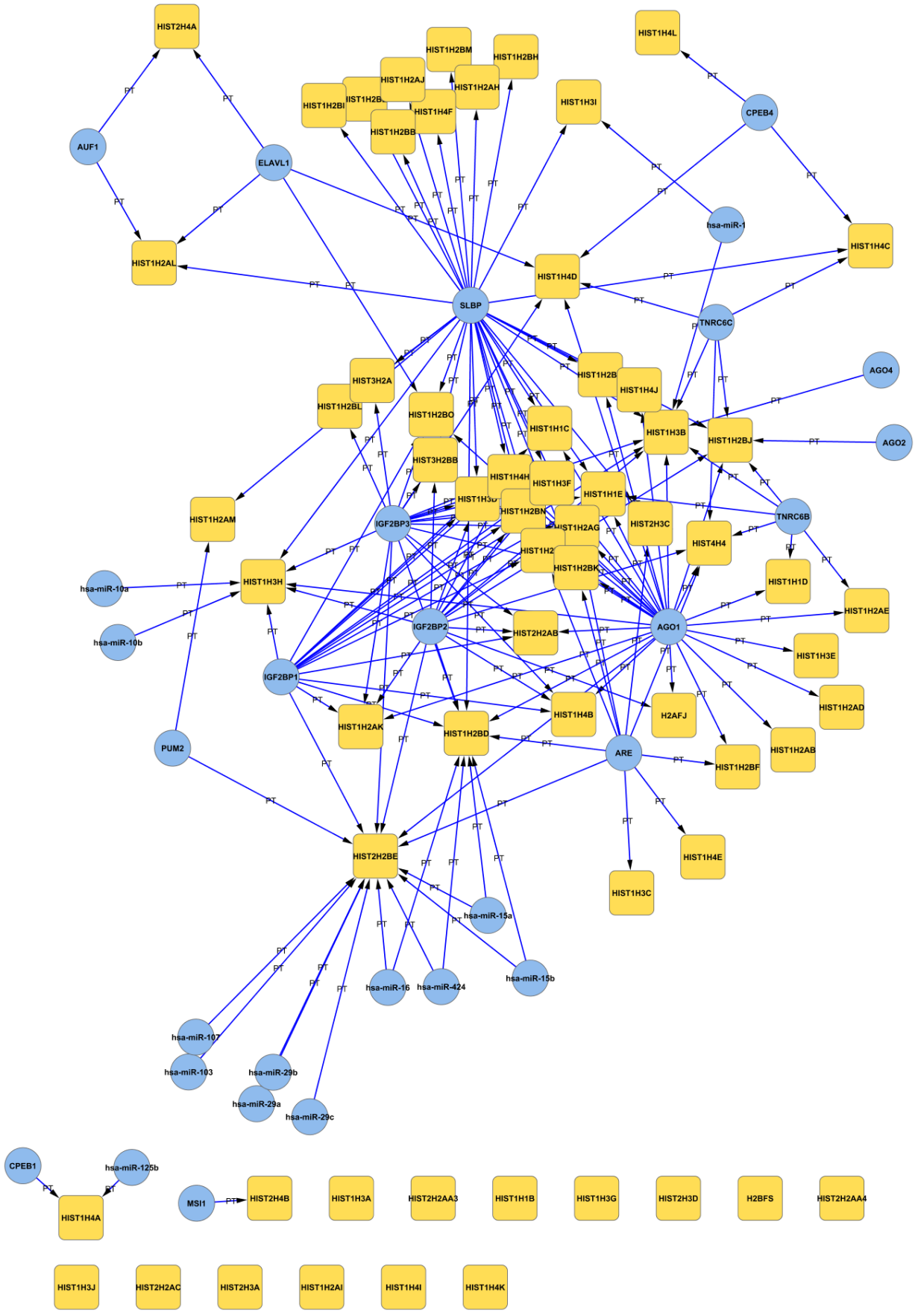
Table 3: Significant up- and down-regulated functional themes in neuroblastoma polysomal versus total comparison. *The table shows the significant functional themes (grouping of ontology terms with coherent functional meaning) emerging from DEGs lists in our neuroblastoma dataset. The status column indicates whether genes composing the theme were up- or down-regulated, while average theme p-value was computed as the –log10 geometric mean of the single terms p-values. While all four themes are statistically significant, it is immediately clear that the histones theme is stronger both in terms of significance and of being highlighted by both gene-based and protein-based ontologies.*

In order to highlight the possible post-transcriptional interactions mediating this up-regulation of histone-related genes at the translational level (with the respect to the transcriptional level) we proceeded by intersecting this genes with the data contained in AURA, by means of a script able to produce the list of regulators and target of a given gene, outputting its binding sites and the eventual co-localized SNPs or RNA editing events. This feature will soon be made available as a batch search modality in AURA. Then, by means of another script exploiting the Cytoscape [51] programming interface, we generated a network based on this list, in which directional edges indicates a post-transcriptional regulator role for a source node with respect to the target node. In the network, shown in Figure 14, DEGs are highlighted with different color and shape in order to distinguish them from the other involved genes returned by AURA. Also this network-building capability, with export to a graphical format, will be soon made available on the website, coupled to the feature described above. By detailed examination of the network, shown in Figure 14, we can obtain some evidence to guide further analysis and experiments: aside from SLBP regulation of a number of histones (fact already known and described in the HCE results section above) we notice the involvement of several microRNAs and of various genes of interest, such as the two ARE binding-proteins AUF1 and HuR and

the two TNRC (6B and 6C), known to have a role in miRNA-mediated mRNA repression. On the same line, various AGO family proteins have been found to bind different histone genes. Eventually, ARE cis-elements are shared by many histones, fact that can be crossed with the binding of AUF1 and HuR that we discussed above. No data is contained in AURA for 12 histones UTRs, suggesting the need to study the post-transcriptional interaction of these genes in a more complete way, possibly by applying techniques such as protein pulldown (to understand regulators of a given mRNA) or PAR-CLIP (to discover all target genes of a possible histone regulator protein).

My contribution in this work consisted in the analysis of the microarrays data produced out of our samples, the identification of differentially expressed genes and the consequent functional analysis of these gene lists; then I performed the intersection with AURA of the functionally coherent groups of DEGs and the construction of the PTR network shown in Figure 14.

Figure 14: Histones PTR network. *The figure displays the histone-centered post-transcriptional regulatory network emerged as up-regulated at the polysomal level in our neuroblastoma microarray analysis. DEGs correspond to square-shaped yellow nodes, while other interacting factors are represented by blue, circle-shaped nodes. An edge between two nodes indicates a verified post-transcriptional interaction extracted from AURA.*


HIST2H4A
HIST1H4L
HIST1H2BM
HIST1H2BH
HIST1H2AJ
HIST1H2AH
HIST1H2BI
HIST1H4F
HIST1H3I
CPEB4
HIST1H2BB
AUF1
ELAVL1
hsa-miR-1
HIST1H4C
HIST1H2AL
HIST1H4D
SLBP
TNRC6C
HIST3H2A
HIST1H2BL
HIST1H2BO
HIST1H2B
HIST1H4J
AGO4
HIST1H1C
HIST1H3B
HIST1H2BJ
AGO2
HIST3H2BB
HIST1H4H
HIST1H3F
HIST1H1E
HIST1H2AM
IGF2BP3
HIST1H2BN
HIST2H3C
TNRC6B
HIST4H4
HIST1H3H
HIST1H1D
hsa-miR-10a
HIST1H2AE
IGF2BP2
HIST2H2AB
AGO1
hsa-miR-10b
HIST1H3E
IGF2BP1
HIST1H2AD
HIST1H4B
H2AFJ
HIST1H2AK
HIST1H2BD
PUM2
ARE
HIST1H2AB
HIST1H2BF
HIST2H2BE
HIST1H4E
HIST1H3C
hsa-miR-15a
hsa-miR-16
hsa-miR-424
hsa-miR-15b
hsa-miR-107
hsa-miR-103
hsa-miR-29b
hsa-miR-29a
hsa-miR-29c
CPEB1
hsa-miR-125b
HIST1H4A
MSI1
HIST2H4B
HIST1H3A
HIST2H2AA3
HIST1H1B
HIST1H3G
HIST2H3D
H2BFS
HIST2H2AA4
HIST1H3J
HIST2H2AC
HIST2H3A
HIST1H2AI
HIST1H4I
HIST1H4K

2.4 PTR tools and database review

In the frame of my doctoral work, identifying and characterizing the available resources on PTR has been a necessary task. Obtaining an overview of data types, amounts and the way in which these were accessible has been a prerequisite to develop AURA and to proceed with the other parts of my work. Thus, writing a review that would serve both as a catalog and as an initial "PTR toolbox" fitted naturally in the context of my activities. The review, recently published by *RNA Biology* starts by classifying the resources (both databases and software tools) according to their biological focus (RBPs, ncRNAs, cis-elements): on top of these foundations, we propose a PTR analysis pipeline which we eventually apply to a breast cancer microarray dataset in order to exemplify its operation and usefulness.

My contribution to this work consisted in collecting data about the resources, defining the pipeline, applying it to the example dataset and eventually writing the manuscript.

3. DISCUSSION

The post-transcriptional regulation of gene expression field has witnessed a lot of developments in the last few years. Still, the amount of work dedicated to it lies far behind that devoted to, for instance, transcriptional regulation. Aside from the lack of mechanistic studies, which can be compensated only by laboratory activity, the main issue we see in the field lies in the fragmentation and dispersion of currently available data: this makes even more difficult obtaining a global and comprehensive picture of this layer of gene expression regulation, let alone identifying new regulatory mechanisms by leveraging on the existing amount of data. Indeed, only three tools among the available ones attempt to integrate different component of PTR networks, such as RBP, miRNA or cis-elements.

AURA can be considered as a "meta-database" integrating for the first time several useful and reliable sources. Differently from all the other available resources (UTRdb/site, RBPdb, doRiNA), AURA integrates the most informative UTR annotations generated by other databases and genome browsers with sequence-based general information (exon-intron structure, evolutionary conservation, intraspecies variation) and with gene- and transcript-centered annotations, such as ontological hierarchies, variability of protein levels in different tissues and transcript stability.

UTRdb[33] is the only other resource to be UTR-centered as AURA. Along with the basic annotations it offers a good amount of data, providing cis-elements prediction through the cognate site (UTRsite). The UTR annotation by phylogenetic conservation is available by both AURA and the last release of UTRdb; however AURA relies on a broader and more updated set of multiple species alignments (phastCons46way, Fujita et al., 2011), as compared to UTRdb (phastCons17way, Fujita et al., 2011). Both RBP and miRNA binding sites datasets are more complete and obtained through more sources in AURA than in UTRdb. Furthermore, whereas the latter provides only conserved elements, the former displays the direct base-wise conservation scores in order to allow a more flexible reuse of this information. On the other side, UTRdb provides structural conservation scores absent in AURA. Furthermore, AURA uniquely collects experimental estimations of transcript stability and of transcript abundances, and the levels of proteins in different tissues. These indicators may result essential when needing to embed the regulatory interactions stored in AURA in a meaningful biological context.

A query to RBPDB[35] on any RBP of interest returns the list of sequences which have been experimentally determined to bind that RBP, together with the RNA-binding specificity consensus (where experimentally obtained). However, it does not directly link the RBP to the targeted transcripts. Thanks to the manual refinement we carried out on the experimental data collected by RBPDB, a similar query to AURA directly shows the target transcripts, the positional information within each transcript UTR as well as the RNA binding specificity logo, with a net gain in terms of completeness of information for the "wet biology oriented" user.

With respect to AURA, doRiNA[36] contains just RBP and miRNA binding sites information: the former are collected only from high-throughput experiments such as PAR-CLIP for RBP and the latter originate from a set of predictions. AURA is more complete in including also experimentally validated miRNA binding sites and mechanistic assays-derived sites for RBP. As a consequence, while providing this data is useful and interesting, also doRiNA lacks the integrative approach necessary to provide a global overview of PTR.

AURA does not yet offer analytical tools (although some are in preparation); this is in contrast with the other databases which offer them to various extents. However, AURA is the only resource of the lot to provide a BioMart query system. BioMart is a standard platform allowing to query various databases from the same interface and in the same way: this is a very powerful feature, as anyone used to query through BioMart will be able to extract any data from AURA in a matter of minutes, without having to learn a new system from scratch.

Nevertheless, In order for AURA to be complete, a thorough literature search would be necessary to retrieve and insert all past PTR data available, resulting from mechanistic experiments: however, while text mining tools may help in reducing complexity and the number of articles to be examined, this task is extremely time-consuming and would need the dedicated effort of more than one individual to be accomplished. Moreover, in order to be even more effective, AURA needs to offer batch analysis tools to its users. Some of these are already being developed and will be ready for AURA 2.0. In particular, these will include the network-generating scripts presented in the results section, a regulator enrichment computation (through Fisher tests and similar) tool and more. Eventually, there are now a number of additional UTRs extracted from next-generation sequencing experiments: adding these isoforms to the standard set of UTR annotation would enrich the database and its completeness (even though these UTRs are currently annotated in a very limited way). NGS-derived tissue-specific expression profiles are also available now: additional mRNA and possibly protein expression profiling dataset would further facilitate the integrated inference of regulatory mechanisms. Eventually, perfecting an automated data update pipeline and continuing to add new data types will be essential for keeping the usefulness of AURA at its top.

More in general, future tool developments should point towards providing a one-stop, truly integrated, comprehensive and multi-faceted PTR analysis toolset. Availability of such a tool will consistently empower the mapping of post-transcriptional and specifically translational networks, reaching the level of service already offered by resources focusing on the analysis of transcriptional regulation. The consequent implementation and update effort could be eased by coordination with major genome databases such as the UCSC Genome Browser and Ensembl. Furthermore, two additional features are currently missing: first of all, a systematic literature-derived annotation of the molecular downstream and phenotypic effects of a given interaction would provide more grounded clues, orienting the experimental validation; then, tailored statistical methods for enrichment of cis-elements or trans-factor, as those for ontology terms enrichment, would be beneficial to avoid

generation of a large number of false positives as an effect of the high multiplicity of action of several studied trans-factors.

Concerning the second part of my work, we have demonstrated the HCE identification algorithm to be sensitive and specific enough to retrieve both already known (histones – SLBP) and novel post-transcriptional regulatory mechanisms (RRM – HuR). In order to extract as much information as possible from these HCEs we will need to analyze them one by one (excluding the ones included in the above groups): a possible way would be to setup an high-throughput luciferase screening to understand the role of these region in modulating protein levels (and the responsible of this modulation). That amounts to testing around 3000 regions, which could be long and time consuming: to focus on the most interesting candidates we will need to devise a prioritization strategy that could be based on the pathway or processes in which HCE-containing genes are involved. This needs to consider the fact that, as shown in the results section, a number of HCEs corresponds to the entire 3'UTR: in these cases we can affirm that conservation concerns the whole regulatory factors for that transcript, and not single binding sites. These HCEs will thus need to be treated in a separate way to isolate relevant subparts of the sequence. Another aspect currently not taken into account by our algorithm is secondary structure conservation: sequence with an higher degree of variation may lead to the same structural element (for instance an hairpin, or a bulge), and in a number of cases the conformation may be more important than the sequence for protein recognition of the binding site. We would thus need to define a structural conservation measure and identify these elements to complete our picture of conservation-based post-transcriptional functional elements in UTRs. Eventually, we will then proceed to the analysis of 5'UTRs HCEs, identified by our pipeline but not studied in this work: more than mRNA stability or localization or polyadenylation, as in the 3'UTR, translation initiation regulation will most probably be the process influenced by these regions in 5'UTRs.

Application of AURA to our neuroblastoma list of differentially expressed genes resulted in a post-transcriptional network of factor-target interactions which lends further evidence to the usefulness of such an integrated database. It is known that histone mRNAs are heavily controlled at the post-transcriptional level, mainly through the SLBP protein and many components of the polyadenylation machinery. The analysis will proceed by first determining the expression patterns of histone genes across the 13 employed cell lines: as the nucleosome composition is stoichiometric, the protein levels of the five histone types (H1, H2a, H2b, H3, H4) must be tightly regulated to guarantee proper assembly of this complex. We may thus devise a mechanism by which one or more genomic alterations acting on histone transcription are subsequently compensated by a post-transcriptional mechanism re-coupling and enhancing protein translation to yield precise quantities of these proteins. Efficiently proliferating tumor cells would thus have evolved or enhanced a way of compensating an unfavorable (from the tumor point of view) alteration. Other tumor datasets (containing both total and polysomal profiling) will be investigated in order to understand whether this finding is neuroblastoma-specific or rather could be shared by various tumor types. To further

complement this data, a sequence and structure motif search could be executed: even though, given the number of genes involved, it is likely that multiple mechanisms are at play, which will make it difficult to identify one or even a few shared motifs on which to focus our subsequent analysis. Aside from validating, via RT-PCR, the fold change values of these genes obtained by microarray, we could also direct our attention towards the phenotype resulting by reversing this up-regulation: targeting the SLBP, which stabilizes histone genes, could allow us to reduce levels of histone mRNAs and observe whether the produced phenotype significantly impacts tumor properties. Further work would then be needed to identify the factors responsible for this effect, both among the interactions found in AURA and other likely regulators of these histone genes.

4. REFERENCES

1. Moore MJ. From birth to death: the complex lives of eukaryotic mRNAs. Science 2005; 309(5740):1514-18.
2. Mignone F, Gissi C, Liuni S, Pesole G. Untranslated regions of mRNAs. Genome Biol 2002; 3(3):REVIEWS0004.
3. Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett 2008; 582(14):1977-86.
4. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C et al. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. Cell 2012; 149(6):1393-406.
5. Baltz AG, Munschauer M, Schwanhäusser B, Vasile A, Murakawa Y, Schueler M et al. The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. Mol Cell 2012; 46(5):674-90.
6. Lunde BM, Moore C, Varani G. RNA-binding proteins: modular design for efficient function. Nat Rev Mol Cell Biol 2007; 8(6):479–90.
7. Andreassi C, Riccio A. To localize or not to localize: mRNA fate is in 3'UTR ends. Trends Cell Biol 2009; 19(9):465-74.
8. Ellington AD, Szostak JW. In vitro selection of RNA molecules that bind specific ligands. Nature 1990; 346:818–822.
9. Tenenbaum SA. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. Proc. Nat. Acad. Sci. 2000;97(26):14085–14085.
10. Ray D, Kazan H, Chan ET, Peña Castillo L, Chaudhry S, Talukder S, Blencowe BJ, Morris Q, Hughes TR. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat Biotechnol. 2009;27(7):667-70.
11. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. CLIP Identifies Nova-Regulated RNA Networks in the Brain. Science 2003; 302(5648):1212-5.
12. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 2010; 141(1):129-41.
13. König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol 2010; 17(7):909-15.
14. Ørom UA, Nielsen FC, Lund AH. MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. Mol Cell. 2008;30(4):460-71.
15. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet 200 2008; 9(2):102-14.
16. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 2005; 120(1):15-20.
17. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat. Genet 2007; 39(10):1278-1284.

18. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human microRNA Targets. PLoS Biol. 2004; 2(11):1862-1879.
19. Hsu SD, Lin FM, Wu WY, Liang C, Huang WC, Chan WL et al. miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res 2011; 39(Database issue):D163-D169.
20. Xiao F, Zuo Z, Cai G, Kang S, Gaso X, Li T. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res 2009; 37(Database issue):D105-110.
21. Yang JH, Li JH, Shao P, Zhou H, Chen YQ, Qu LH. starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. Nucleic Acids Res. 2001; 39(Database issue):D202-D209.
22. Hsu SD, Chu CH, Tsou AP, Chen SJ, Chen HC, Hsu PW et al. miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes. Nucleic Acids Res 2008; 36(Database issue):D165-169.
23. Cho S, Jun Y, Lee S, Choi HS, Jung S, Jang Y et al. miRGator v2.0: an integrated system for functional investigation of microRNAs. Nucleic Acids Res. 2001; 39(Database issue):D158-D162.
24. Bu D, Yu K, Sun S, Xie C, Skogerbø G, Miao R et al. NONCODE v3.0: integrative annotation of long noncoding RNAs.Nucleic Acids Res. 2012 Jan;40(Database issue):D210-5.
25. Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res. 2011; 39(Database issue):D146-D151.
26. Dinger ME, Pang KC, Mercer TR, Crowe ML, Grimmond SM, Mattick JS. NRED: a database of long noncoding RNA expression. Nucleic Acids Res 2009; 37(Database issue):D122-126.
27. Bisognin A, Sales G, Coppe A, Bortoluzzi S, Romualdi C. MAGIA2: from miRNA and genes expression data integrative analysis to microRNA-transcription factor mixed regulatory circuits (2012 update). Nucleic Acids Res. 2012; 40(Web Server issue):W13-21.
28. Mack S.G. MicroRNA gets down to business. Nature Biotechnology 2007 ; 25: 631 - 638.
29. Barreau C, Paillard L, Osborne HB. AU-rich elements and associated factors: are there unifying principles? Nucleic Acids Res 2005; 33(22):7138-7150.
30. Wang J, Pantopoulos K. Regulation of cellular iron metabolism. Biochem J 2011; 434(3):365-81.
31. Jacobs GH, Chen A, Stevens SG, Stockwell PA, Black MA, Tate WP et al. Transterm: a database to aid the analysis of regulatory sequences in mRNAs. Nucleic Acids Res 2009; 37(Database issue):D72-76.
32. Pickeringa M.B, Willisb A.E. The implications of structured 5' untranslated regions on translation and disease. Semin.Cell.Dev. Biol. 2005; 16:39-47.
33. Grillo G, Turi A, Licciulli F, Mignone F, Liuni S, Banfi S et al. UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res 2010; 38(Database issue):D75-80.
34. Dassi E, Malossini A, Re A, Mazza T, Tebaldi T, Caputi L et al. AURA: Atlas of UTR Regulatory Activity. Bioinformatics 2011; doi: 10.1093/bioinformatics/btr608

35. Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. RBPDB: a database of RNA-binding specificities. Nucleic Acids Res. 2011; 39(suppl 1):D301-D308.
36. Anders G, Mackowiak SD, Jens M, Maaskola J, Kuntzagk A, Rajewsky N et al. doRiNA: a database of RNA interactions in post-transcriptional regulation. Nucleic Acids Res 2011; doi:10.1093/nar/gkr1007.
37. Bakheet T, Williams BR, Khabar KS. ARED 3.0: the large and diverse AU-rich transcriptome. Nucleic Acids Res 2006; 34(Database issue):D111-114.
38. Gruber AR, Fallmann J, Kratochvill F, Kovarik P, Hofacker IL. AREsite: a database for the comprehensive investigation of AU-rich elements. Nucleic Acids Res. 2001; 39(Database issue):D66-D69.
39. Mokrej M, Masek T, Vopálensky V, Hlubucek P, Delbos P, Pospisek M. IRESite – a tool for the examination of viral and cellular internal ribosome entry sites. Nucleic Acids Res 2010; 38(Database issue):D131-D136.
40. Castellano S, Gladyshev VN, Guigó R, Berry MJ . SelenoDB 1.0: a database of selenoprotein genes, proteins and SECIS elements. Nucleic Acids Res 2008; 36(Database issue):D332-D338.
41. Andken BB, Lim I, Benson G, Vincent JJ, Ferenc MT, Heinrich B et al. 3'-UTR SIRF: A database for identifying clusters of short interspersed repeats in 3'untranslated regions. BMC Bioinformatics 2007; 8:274
42. Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S et al. Rfam: Wikipedia, clans and the decimal release. Nucleic Acids Res. 2011; 39(Database issue):D141-D145.
43. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res., 15(8), 1034-50.
44. Stark, A. et al. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures.Nature, 450(7167),219-232.
45. Marzluff WF, Wagner EJ, Duronio RJ.(2008) Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. Nat Rev Genet., 9(11), 843-54.
46. Pavesi, G. et al. (2006) MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes.Nucleic Acids Res., 34(Web Server issue), W566-570.
47. Lorenz R, Bernhart SH, Hoener Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. (2011) ViennaRNA Package 2.0. Algorithms Mol Biol., 6(1), 26.
48. Höchsmann M, Töller T, Giegerich R, Kurtz S. (2003) Local similarity in RNA secondary structures. Proc IEEE Comput Soc Bioinform Conf, 2, 159-68.
49. Lebedeva, S., *et al*.(2011) Transcriptome-wide Analysis of Regulatory Interactions of the RNA-binding protein HuR. Molecular Cell, 43, doi:10.1016/j.molcel.2011.06.008.
50. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009; 4(1):44-57.
51. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 2011; 27(3):431-432.

5. PAPERS

-> Dassi E, Malossini A, Re A, Mazza T, Tebaldi T, Caputi L et al. **AURA: Atlas of UTR Regulatory Activity**. Bioinformatics 2011; doi: 10.1093/bioinformatics/btr608

-> Dassi E, Zuccotti P, Leo S, Provenzani A, Riva P, Quattrone A. **Hyper conserved elements in 3'UTRs reveal a translational network of RNA binding proteins controlled by HuR.** [*Submitted to Nucleic Acids Research*]

-> Dassi E and Quattrone A. **Tuning the engine: an introduction to resources on post-transcriptional regulation of gene expression.** RNA Biology 2012; 9(10): 1-9.

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.

HYPER CONSERVED ELEMENTS IN VERTEBRATE mRNA 3'-UTRs REVEAL A TRANSLATIONAL NETWORK OF RNA BINDING PROTEINS CONTROLLED BY HUR

Erik Dassi[1], Paola Zuccotti[2], Sara Leo[1], Alessandro Provenzani[3], Paola Riva[2] and Alessandro Quattrone[1,*]

[1]Laboratory of Translational Genomics, Centre for Integrative Biology, University of Trento

[2] Department of Medical Biotechnology and Translational Medicine, University of Milan

[3]Laboratory of Genomic Screening, Centre for Integrative Biology, University of Trento

* To whom correspondence should be addressed.

ABSTRACT

Almost unknown is the map of the posttranscriptional networks controlling gene expression in eukaryotes, and unclear is its evolution and the relative role in it of RNA-based and protein-based regulative factors. Here we introduce a simple approach relying on both phylogenetic sequence sharing and conservation in the whole mapped 3'UTRs of vertebrate species to gain knowledge on core posttranscriptional networks. The identified human Hyper Conserved Elements (HCEs) were predicted to be preferred binding sites for RNA binding proteins (RBPs) and not for non coding RNAs (ncRNAs), namely microRNAs and long ncRNAs. To test for exploitation of the HCE map, we found that it identified a well-known network posttranscriptionally regulating histone mRNAs, and that promoted the discovery of a previously unknown translational network. We experimentally verified this last network, composed of RRM-type RBP mRNAs positively controlled by the RRM-type RBP HuR. Analysis of HCE distribution in the validated HCE 3'UTR map shows a profile of prevalently small clusters separated by unconserved intercluster RNA stretches, predicting the formation in 3'UTRs of discrete small ribonucleoprotein complexes.

We therefore suggest RBP-mRNA networks at the root of posttranscriptional control of gene expression in vertebrate cells, and provide a means to get insights into their structure.

INTRODUCTION

The 3' untranslated region (3'UTR) of mRNAs is a fundamental mediator of the processes affecting posttranscriptional regulation of gene expression (1, 2), exerted through the binding of RNA binding proteins (RBPs) and non-coding RNAs (ncRNAs). While a subclass of ncRNAs, the microRNAs (miRNAs), bind the mRNA 3'UTR in a ribonucleoprotein complex with AGO proteins to always negatively control target mRNAs (3, 4, reviewed in 5), 3'UTR interacting RBPs can exert complex effects, influencing mRNA transport, localization, polyadenylation state, rate of degradation, and finally rate of translation through regulated assembly/disassembly of actively recycling polysomes (6). In this way, RBPs behave as topological controllers of gene expression and can influence it both negatively and positively.

Mechanistic studies have helped to identify dozens of single cis-elements in 3'UTRs bound by specific RBPs and miRNAs (7, 8), sometimes with defined consequences on gene expression and cell phenotypes. In vitro (9, 10, 11) or in vivo (12, 13, 14) high-throughput approaches are also starting to provide transcriptome-wide maps of RBP and miRNA regions of interaction with mRNAs, allowing us to trace the first mRNP networks in yeast (15,16,17) and in vertebrates (18,19,20).

Trans-factors bind to mRNA UTRs in short continuous regions, often corresponding to a defined secondary structure and a recurrent consensus sequence. If the same among species, these trans-factor footprints should determine a local increase in sequence homology. On the assumption that in a purifying (negative) selection context highly conserved noncoding sequences in orthologous protein-coding genes would point to elements potentially endowed with regulatory activity, it would be possible to obtain information regarding the core networks involved in mRNA regulation by isolating the regions bearing an high degree of sequence evolutionary conservation in UTRs. This holds also because no selective pressure for protein functionality applies to UTRs, which are thus unconstrained to change sequence or structure just to fulfill their regulatory purpose.

On a genomic scale, the identification of putative functional elements on the basis of evolutionary conservation has been mostly based on the comparison between human,

rat and mouse genomes, with the definition of the so-called Ultra Conserved Regions (UCRs) as 200bp identical DNA stretches. This procedure selects for mostly nonexonic portions of the genome (21, 22, 23, 24), now collected in a database (25). Only a very limited number of these UCRs lies in mRNA UTRs. The same approach has been recently applied to the transcriptome (26) as defined by a library of expressed sequence tags. The identified 3096 sequences clustered in 96 segments, of which 23 were fully in the CDS and 80 overlapped or were entirely in UTRs. Out of UCRs, specific mining of UTRs for regions of high conservation has been pioneered almost ten years ago (27) by identifying conserved motif cores and extending them up to a defined threshold, or by computing a motif conservation degree based on pairwise alignment homology frequency (28). In each of these two studies four mammalian species were compared for the small number of UTRs known at that time. Genome-wide multiple alignments of several species has been rendered possible in recent years by the increased sequencing capabilities (29, 30), but they have never been applied to specifically address the identification of potentially functional sites in UTRs. In vertebrates, 3'UTRs are longer and less conserved than 5'UTRs, and surprisingly they are modestly variable in length between species with respect to the observed intraspecies length distribution (31). This could suggest the existence of unknown phylogenetic constraints acting on their length, like long-range interactions among functional elements.

We introduce here an approach for identifying hyper conserved elements (HCEs) in 3'UTRs of mRNAs, weighting sequence conservation information and phylogenetic distance on 44 vertebrate species, from human to lamprey. The approach does not require the assumption of an a priori sequence length, takes limited computation time and can be used for any desired reference species and species subgroup. Its application to human 3'UTRs led us to the mapping of more than three thousand HCEs, which occupy less than 0.5% of the total 3'UTR sequence space. These regions have peculiar properties, including a clustered pattern of recurrence, and show a potential to localize functional cis elements belonging to highly conserved mRNA control networks. To demonstrate the usefulness of HCEs in prioritizing sequences for further analysis, we used them to identify a network of mRNAs coding for RBPs whose 3'UTRs are

bound by the HuR RBP, and we proved this network to be functional in translational regulation of gene expression.

MATERIALS AND METHODS

HCE identification pipeline.

Human 3' UTR sequences were fetched from the hg18 assembly in the UCSC database (32) and all UTRs shorter than 5 bases were filtered out, as they are likely to derive from annotation error. The Sequence Conservation Score (SCS) for each base of the UTRs, as computed by phastCons (33), was retrieved from the same source along with the 44-way Multiz alignment in MAF format for the relevant regions of the genome. We computed the Branch Length Score (BLS) (34) as the fraction of the length of the total phylogenetic tree branches covered by the alignment of each exon composing an UTR, employing the lowest BLS of all exons as the BLS for the whole UTR. The final conservation score, which we term hyper conservation score (HCS), was computed for each base of the UTRs as the weighted average of SCS and BLS. Weight for both components was set at 0.5, even though our pipeline allows changing these weights to obtain a different combination of the two features. A schematic view of the pipeline can be found in Figure 1A.

A threshold was set on average HCSs under which sequences should not be considered as hyper conserved. The threshold was chosen to be 0.85 as, by weighting SCS and BLS equally, that would require one part of the score to be at least 0.7 when the other part is 1.0 and vice versa. This stringent constraint guarantees that only the most conserved regions of the UTRs are actually selected as Hyper Conserved elements (HCEs).

HCEs were identified in 3'UTRs by means of a two-step algorithm:

1. First, a search was run in every UTR for five-base seeds which have an almost complete conservation sequence-wise (SCS greater or equal than 0.95) and which average HCS is not less than 0.85.

2. Then, these seeds were extended upstream and downstream into the UTR, one base at a time, for as long as the average HCS of the HCE did not fall below the preset threshold.

Resulting HCEs were eventually merged to remove overlaps and duplicates, which could occur in the case of very high conservation spanning a substantial part, if not the whole, UTR. A schematic view of the algorithm can be found in Figure 1A.

Construction of the non-HCEs datasets.

In order to compare HCEs properties with respect to non-HCE UTR portions, we built 1000 datasets composed by an equal number of non-HCE sequence elements. Via a Python script we randomly chose UTR and start position; the region length was drawn from the HCE length distribution, in order to mimic the HCE size ranges.

HCE intersection with binding sites of ncRNAs.

Experimentally validated binding sites of miRNAs were extracted from the SQL version of AURA (18), available on the download page of the website. The dataset contained 15560 binding sites regarding a total of 88 distinct miRNAs. Coordinates of these sites were intersected with HCEs and only sites falling completely inside an HCE were kept. HCEs and non-HCEs sites were also intersected with miRNA binding sites predicted by three popular tools, miRanda (35), PicTar (36) and PITA (37). The content of lncRNAdb (38) was downloaded from the website and filtered to keep only human lncRNAs. A BLAST (39) database was built with these sequences and a search was performed with HCEs as query, with the BLAST "task" parameter set as "blastn-short"; only matches with a maximum e-value of 0.05 were considered as true positives.

HCE intersection with RBP Position-Frequency Matrices.

Position-Frequency Matrices (PFMs) for 69 RBPs were extracted from the RBPDB database (40). HCE and non-HCE sequences were matched against these PFMs via the BioPython functions dedicated to this task. We retained only matrices longer than 4 bases (for a total of 29 matrices) and filtered out all matches with score lower than 80%.

HCE intersection with the mRNA-protein occupancy profile.

T>C conversion profiles were downloaded from the GEO database (series GSE38355) and filtered to include only bases falling into 3'UTRs. HCEs and non-HCEs bases were intersected with the conversion profiles, quantiles were computed and distributions of scores were tested for significant differences by means of a t-test. For the non-HCE case, the iteration giving the best results was used to compare with HCE scores distribution.

Overrepresentation analysis.

All genes which UTRs contained at least one HCEs were extracted and input to the DAVID Functional Annotation tool (41) to identify the overrepresentation of functional terms contained in various ontologies (selected resources were Gene Ontology (GO) Molecular Function, Biological Process and Cellular Component; IPR; SMART; PFAM, SP_PIR_keywords, Biocarta, KEGG and OMIM disease). Estimation for the terms p-value was Bonferroni corrected and only terms for which the p-value was under 0.05 were included in final results; terms were grouped according to their similarity via the DAVID Functional Clustering tool, using high-stringency clustering criteria.

Identification of the SLBP binding sites.

Sequences of the HCEs belonging to genes annotated to be part of the chromosome assembly functional group were aligned by means of ClustalW2 (42) along with the canonical SLBP binding motif to detect if these HCEs actually contained the latter. The multiple alignment algorithm was run with its default set of parameters.

Sequence motif search.

Sequence motif search inside HCEs was performed by means of the Weeder algorithm (43). Motif length was set to be 6, 8, 10 or 12 nucleotides and the minimum occurrence frequency of the motif was set to 25% of the sequences composing the dataset. We considered as relevant all the motifs reported by Weeder as highest ranking.

Secondary structure motif search.

The secondary structure folding of the HCEs contained in the RRM-type RBP mRNA group were predicted via the RNAfold program of the Vienna RNA package (44). Motifs were searched over these structures by means of the RNAforester tool (45), run in the local, multiple alignment mode.

HuR overexpression and silencing.

MCF-7 cells were transiently transfected using Lipofectamine2000 (Invitrogen, Carlsbad, CA, USA) with a pT-REX mammalian expression vector coding fror human HuR (55) and with the mock empty vector as control. The same cells were infected with lentiviral transduction particles bearing shRNAs (Sigma Aldrich, Mission shRNA) against the HuR sequence, following the manifacturer protocol and testing four different shRNA sequences. Non-target control transduction particles were used to infect MCF-7 cells as negative controls. Stably silenced clones were selected with puromycin. The most effective pool, KD1, was derived from the TRCN0000017273 shRNA. Sequences are reported in supplementary material.

Cell culture and treatments.

Human breast cancer MCF-7 and MCF-7 shHuR cells were cultured in DMEM with 10% FBS, 100 U/ml penicillin-streptomycin and 0.01 mM L-glutamine (all media ingredients were obtained from Sigma-Aldrich, AS, Oslo, Norway). Cultures were maintained at 37°C in a 5% CO2 incubator. Puromycin (final concentration 2.5µg/µl) was used for selection and maintenance of stable short hairpin RNA (shRNA) transfectants. All reagents were purchased from Sigma. 1,5 $x10^6$ MCF-7 and MCF-7 shHuR cells were seeded into two 10cm Petri dishes for polysomal RNA extractions and into one 10cm Petri dish for total RNA extractions. Total RNA and polysomal extractions were performed 72 hrs after seeding; all the experiments were in biological triplicate.

RNA-Protein pull-down assay.

RNA probes for HuR (AUGUAUUGUUUAUACAU), Degenerated (AUGUAUNNNNNAUACAU), Dblmut1

(AUGUAUGGUUGAUACAU), Dblmut2 (AUGUAUUCUUAAUACAU), YB1 (AUGUAUGGUCUGCAUACAU) and PTB (AUGUAUCUUUCUUAUACAU) have been synthesized by Sigma using 0,05μmol Synthesis Scale and HPLC purification with a 5' biotinilated DNA polyC linker. Their predicted secondary structure folding is shown in Figure 3. Biotin pull-down assays were performed by incubating 40μg of MCF-7 cell lysates with 1μg of biotinylated probes for 1 hr at room temperature. The complexes were isolated using 100μl of paramagnetic streptavidin-conjugated Dynabeads (Dynal®, Invitrogen, Carlsbad, CA, USA), and bound proteins in the pull-down material were analyzed by Western blotting using antibodies recognizing HuR (Santa Cruz, CA, USA), YB1 (Abcam, Cambridge, UK) and PTB (Santa Cruz, CA, USA). After secondary-antibody incubations, the signals were visualized by chemiluminescence (Amersham Biosciences, GE Healthcare, UK).

Total RNA extraction.

Total RNAs from treated and non-treated cells was isolated using the TRIzol reagent (Invitrogen, Carlsbad, CA, USA), according to the manufacturer's instructions. Purity of RNAs (A260/A280 value of 1.8–2.1) and concentration were measured using the Nanodrop spectrophotometer. To eliminate DNA contamination, total RNA was treated with DNase I (RNase-Free DNase Set, Qiagen) and then purified with RNeasy kit (Qiagen, Hilden, Germany).

Polysomal RNA extraction.

MCF-7 cells, treated as described above, were incubated for 3 minutes with 0.01mg/ml cycloheximide at 37°C, then the plates were put on ice. The media was removed and the cells were washed twice with cold phosphate buffer saline (PBS) + cycloheximide 0.01mg/ml. Cells were directly lysed on the plate with 300μl cold lysis buffer [10mM NaCl, 10mM $MgCl_2$, 10mM Tris–HCl, pH 7.5, 1% Triton X-100, 1% sodium deoxycholate, 0.2U/ml RNase inhibitor (Fermentas, Burlington, CA), 1mM dithiothreitol and 0.01mg/ml cycloheximide], scraped and transferred to an Eppendorf tube. The extracts were centrifuged for 5 min at 12000g at 4°C. The supernatant was frozen in liquid nitrogen and stored at -80°C or loaded directly onto a 15–50% linear sucrose

gradient containing 30mM Tris–HCl, pH 7.5, 100mM NaCl, 10mM $MgCl_2$, and centrifuged in an SW41 rotor for 100 min at 180000g. Fractions (polysomal and sub-polysomal) were collected monitoring the absorbance at 254 nm and treated directly with 0.1 mg/ml proteinase K for 2 hours at 37°C. After phenol–chloroform extraction and isopropanol precipitation, polysomal and sub-polysomal RNAs were resuspended in 30µl of RNAse free water and then repurified with RNeasy kit (Qiagen, Hilden, Germany).

Quantitative RT-PCRs.

For quantification of mRNAs, a two-step Taq-Man real-time PCR analysis was performed, using probes obtained from Applied Biosystems (Foster, CA, USA). cDNA was synthesized from total and polysomal RNA (1µg) in 20 µl reactions, using the iScript cDNA Synthesis Kit from BioRad (cat n°#170-8891). The reverse transcriptase reaction was performed by incubating the samples at 25°C for 5 min, 42°C for 30 min, and 85°C for 5 min. The PCR reactions (10µl) were performed on 20ng of cDNA, the mix were prepared with 5X KAPA FAST probe (cat n° KK4702, Kapa Biosystems, (Boston, MA, USA) and the 20X appropriate Taq-Man probe. The PCR mixtures were incubated at 95°C for 3 min, followed by 39 cycles of 95°C for 30 s and 60°C for 20 s and 72°C for 60 s. mRNA levels were calculated based on the ∆CT method, using RPL0 and HPRT1 as reference genes. All PCRs were performed in triplicate using an iQ5 RealTime PCR detection system (Bio-Rad, Hercules, CA, USA).

Ribonucleoprotein Immunoprecipitation.

Ribonucleoprotein Immunoprecipitation (RIP) was performed using human HuR overexpressing MCF-7 cell line lysates. Cell extracts were resuspended in NT2 buffer (50mM Tris HCl pH=7.5, 150mM NaCl, 1mM MgCl2, 0,05% NP40, 1U/µl Ribolock (Fermentas, Glen Burnie, MD, USA), 2mM DTT, 30mM EDTA) supplemented with a protease inhibitor cocktail (P8340, Sigma), chilled at 4°C. The cell lysates were added to the Protein G Dynabeads (Dynal®, Invitrogen, Carlsbad, CA, USA) at 50µl beads/250µl lysate. Beads were previously incubated with cell extracts and then bound with 5µg of mouse monoclonal anti-HuR antibody (Santa Cruz, sc-71290, CA, USA) or

mouse IgG (Millipore, NI03-100UG). Associated RNA was extracted using phenol:chloroform:isoamyl-alcohol (25:24:1) and precipitated with ethanol. RNA pellets were resuspended in 10µl RNA-grade water and, after DNAse treatment (Fermentas, Glen Burnie, MD, USA), cDNA was obtained from each samples as previously detailed. Real Time quantitative PCR was performed in duplicate using the C1000 (Bio-Rad, Hercules, CA, USA) thermal cycler for 40 cycles, and results were evaluated by cycle threshold (Ct) values. Cyclin A mRNA was quantified as positive control, being a known HuR target. Obtained data were the average of at least three independent experiments.

Construction of the HuR / RRM-type RBP mRNA network.

HuR binding sites as identified in HEK293 cells by a recent PAR-CLIP study (46) were downloaded from GEO, accession number GSE29943. Sites were intersected with UCSC 3'UTR coordinates (hg18 assembly) and extracted along with the genes mapping to these 3'UTRs. Enrichment was computed by counting the number of genes for each domain found in the resulting genes list and by performing a Fisher test by means of the R statistical environment. The HuR RRM-type RBP target mRNA network was built by adding all RRM-type RBP mRNAs found to be bound by HuR in the PAR-CLIP study to our HCE-containing 23 RRM-type RBP mRNAs. An edge was added between HuR and its target mRNA to indicate the regulatory relationship. Intersections between the PAR-CLIP-derived 89, our 23 and the 6 validated by us RRM-type RBP mRNAs were computed and highlighted by employing different colors and shapes of the nodes, as shown in Figures 5B and 5C.

RESULTS

HCEs in the mRNA 3'UTRs are rare, short, highly structured and organized in clusters.

We aimed at identifying regions of exceptional evolutionary conservation in the 3' UTRs of the human exome by a seed extension strategy. 3'UTR HCEs (3'UTR Hyper Conserved Elements) were derived from the hg18 assembly of the human genome (hg18, The Genome Sequencing Consortium) as reported by the UCSC database (32) by a custom pipeline (Figure 1A). We took advantage of the 44-way vertebrate UCSC alignment (32), generated by first computing pairwise alignments for each species using BLASTZ (47) and then merging them with MULTIZ (48). From this alignment we derived the phastCons sequence conservation score (SCS, 33) for each base of the exome annotated as 3'UTR. We also calculated for each base the Branch Length Score (BLS), defining the degree of sharing of the conservation among the vertebrate species considered (34), in our case of entire 3'UTRs. We firstly restricted our analysis to short footprints of very high phylogenetic invariance, represented by fully conserved 5-bases seeds (SCS >= 0.95 and BLS >= 0.85). We then extended these seeds upstream and downstream until they reached a preset threshold (0.85) on the conservation measure we called HCS (Hyper Conservation Score, computed for each base of the 3'UTRs as the weighted average of SCS and BLS). Weight for both components was set at 0.5, which we identify as the best measure (changing these weights would change the relative importance of one of the two features, see Supplementary Material). After preliminary filtering, the dataset obtained from the UCSC database contained 55444 3'UTRs, each one corresponding to a different transcript (including all annotated mRNA splicing variants). The 3'UTR HCE identification algorithm gave 3149 HCEs, belonging to 1010 3'UTRs, which corresponded to 877 genes. At least one 3'UTR HCE is present in only 1,8% of the total human 3'UTRs, and collectively HCEs cover only 0.47% of the 3'UTR space, making them extremely rare. 3'UTR HCEs have an average length of 100 bases, but their length distribution (Figure 1B) is such that more than 77% of their total number is shorter, being only 4.5% of them over 500 bases. The subset of HCEs shorter than 100 bases have an average length of 23 bases, with 25% of them at most 8 bases long. Their UTR coverage (Figure 1C) is instead prevalently low (25% or less of

each 3'UTR) or high (75% or more of the 3'UTR). Together, these distributions show that 3'UTR HCEs are relatively short and that they either occupy a small portion of a 3'UTR or the most of it. These elements are much richer in AU than in GC bases (Figure 1D, p-value 2.2E-16), and are by far more highly structured than random 3'UTR sequences of the same length, being the structural density defined by the fraction of unpaired bases in the HCE secondary structure (Figure 1E, p-value 1.2E-13). Also their localization in the 3'UTRs has interesting properties: when multiple HCEs are present on an UTR, these have a clear tendency to localize in clusters, as indicated by the very small inter-HCE distance, 25 bases or less (Figure 1F), and to be distributed along the 3'UTR with a preference for its beginning, with 25% of the HCEs starting on the 3'UTR 10% initial bases (Figure 1G). To provide a snapshot on HCE architecture diversity, we distributed HCE-bearing 3'UTRs into four classes, depending on their number and coverage. The classes reported in Figure 2A efficiently represent this diversity. We then focused on the HCE clustered pattern because it could be an effect of an higher order structure of trans-factors. We thus computed the amount of HCEs lying in clusters with intracluster distances (maximum distance between two contiguous HCEs in a cluster) ranging from 5 to 40 bases. As shown in Figure 2B, a plateau starts at 20 bases, setting therefore a threshold. At this distance, 81% of the HCEs belong to clusters of 2 or more elements (the figure already excluded the 577 HCEs which are unique on their 3'UTR). We thus propose a model, reported in Figure 2C, for which 3'UTRs contain clusters of binding sites separated by each other, possibly delineating a scenario in which groups of trans-factors interact with each other in complexes spaced by unconserved regions of unbound 3'UTR.

3'UTR HCEs contain putative binding sites for RBPs and not for ncRNAs.

The main question now was what types of potentially functional cis-acting elements are found in 3'UTR HCEs. To test for miRNAs, we compared the 3'UTR HCEs with a set of 15560 experimentally determined 3'UTR miRNA binding sites (produced by 88 miRNAs and involving 2232 3'UTRs) extracted from the AURA database (18). Out of the total 3149 HCEs, only 51 (1.6%) of them was found to contain one or more miRNA binding sites, which were 60 in total involving 33 different miRNAs. These data resulted in whole

3’UTRs being more enriched in miRNA binding sites than HCEs (Fisher test p-value = 2.37E-10). To verify if this small number was close to random occurrence, we performed the same procedure on 1000 sets of randomly derived 3’UTR segments, which we call non-HCEs, with the same length distribution and of the same size as the HCEs. The maximum of the distribution of these iterations gave 40 unique miRNA binding sites involving 47 different miRNAs, which confirms our hypothesis. We eventually proceeded to predict miRNA binding sites in HCEs and non-HCEs by means of three popular prediction tools [miRanda (35), PicTar (36), PITA (37)]. Compared to the best non-HCE iteration, the number of miRNA binding sites in HCEs is always heavily depleted (Fisher test reports enrichment of non-HCEs sites with p-value lower than 2.2E-16 in all three cases). To check also for other ncRNAs we intersected 3’UTR HCEs with lncRNAdb (38), a catalog of eukaryotic long non-coding RNAs (lncRNAs). A BLAST search yielded 151 statistically significant putative binding sites at least 12 bases long, involving 132 unique HCEs (4.2%) and 32 different lncRNAs. Again among the 1000 non-HCEs iterations, the BLAST search yielded, for the iteration giving the best results, 209 statistically significant putative lncRNAs binding sites at least 12 nucleotides long, involving 167 unique non-HCEs (5.30%) and 39 different lncRNAs. Therefore, HCEs are unlikely to be preferred sites for miRNAs and lncRNAs.

We then scanned the HCE and the non-HCE lists for matches with the position-frequency matrixes (PFMs) extracted from the RBPDB resource (40), which collects the known experimental consensi for RBP binding to mRNAs. Considering only matches with a minimum score of 80% and a matrix length greater than 4 (leaving us with 29 matrices), we consistently obtained at least 1.8 times more matches in the HCE than in the non-HCE sets (17173 matches for HCEs vs 9443 matches for the best iteration of non-HCE sequences). Enrichment of RBP sites in HCEs with respect to non-HCEs is also suggested by the Fisher test (p-value=5.85E-11). If really 3’UTR HCEs identify mainly RBP binding sites, they should at least partially span an experimentally determined RBP mRNA occupancy profile. A recent PAR-CLIP study defines, as T>C conversion scores (14), contact sites for RNA-interacting proteins, including RBPs, in the mRNA transcriptome of proliferating HEK293 cells (49). The distributions of T>C conversion scores for each base falling in 3’UTR HCEs and non-HCEs were tested

against each other for statistically significant differences. Indeed, HCEs were found to have a significantly higher level of T>C scores than non-HCEs, with the performed t-test producing a p-value lower than 2.2E-16, and with a median T>C score of HCEs of 5.5 versus 4.5 of non-HCEs. This suggests that 3'UTR HCEs are enriched for RBP binding sites.

3'UTR HCEs identify the ancient control mediating histone mRNA fate.

In order to appreciate the spectrum of biological functions expressed by 3'UTR HCE containing genes, we performed an ontological enrichment on the 877 genes bearing at least one HCE in the 3'UTR. We identified three gene groups endowed with high significance (Supplementary Figure S1). The first group is composed by 78 genes involved in chromatin structure (terms "nucleosome", "chromatin assembly", "DNA packaging"), including 51 (53.6%) of the 95 histone genes present in the human genome. This wide histone component of the signature is that producing the strongest over-representation signal, because the terms remain highly significant even when performing the ontological enrichment after having removed the non-histone genes. It is well known that all histone gene mRNAs have a short 3'UTR, lacking a poly(A) tail, which is bound by the stem-loop binding protein (SLBP) in the cytoplasm to stabilize these mRNAs and mediate their nuclear processing and their translation (50). Alternative to polyadenylation, this mechanism is very ancient and is conserved over a wide evolutionary distance (51). We therefore hypothesized that the HCEs in the histone 3'UTRs were SLBP binding sites. In order to verify this conjecture, we aligned the known SLBP binding motif (52) to these HCEs and found that a considerable fraction of the HCEs (75 out of 127) contain a close, if not perfect, match to the SLBP motif (Supplementary Figure S2). Therefore, the metrics we devised to select for HCEs precisely identifies cis-elements involved in a conserved and well demonstrated posttranscriptional regulatory process. We assumed this finding as an effective benchmark for the ability of 3'UTR HCEs to point to circuitries of phylogenetically old posttranscriptional control.

The second gene set of high statistical significances is about the broad activity of transcription, being prevalent in the signature its repression. The 137 identified genes

suggest that transcription factors as EPC1, TFAP2D and YY1 and co-transcriptional repressors as FOXP2, MEIS2 and EZH2 can be heavily controlled post-transcriptionally, being their 3'UTR almost entirely highly conserved. Finally, the third emerging gene set came from the protein domain annotation, giving the RNA Recognition Motif, the RRM.
We also divided the HCEs on the basis of the four classes identified, to see again if they had a preferential representation of themes. We found that the "chromatine structure" theme is enriched only in the "lone island" category (Figure 1H), further confirming that it emerges from the histone mRNA SLBP binding site (51). Transcriptional regulation terms appear instead enriched in the "sparse frequent" and "fully covered" groups, while both the "dense frequent" and "fully covered" 3'UTR groups, i.e. those mostly HCE-rich, point to a significant over-representation (p-value = 1.09E-05) for mRNA-related activities (GO terms: "RNA binding", "mRNA processing"; domains: KH, RRM).

A hyperconserved motif in the 3'UTR of 19 RRM-type RBP mRNAs bound by HuR.

Given the recurrent tendency of the enrichment analysis to select the mRNAs of RRM-type RBPs as preferred sites for 3'UTR HCEs, we further explored these mRNAs. The RRM is the evolutionarily most successful among the solutions appeared to mediate interaction between RNA and proteins (53). Of the 23 enriched genes whose mRNA bears at least one HCE and whose protein product contains RRM domains, 17 were experimentally verified RBPs and 16 had an RRM-only architecture (Supplementary Table 1). Their mRNAs are characterized by 3'UTRs of all four types, with a prevalence of full (66.7%) and sparse frequent (19%) types, with lone island and dense frequent types representing respectively just 9.5% and 4.7% of the 3'UTRs. RBPs have been shown in the yeast to be nodes of highly interconnected networks of posttranscriptional regulation (15, 17), but very few is known about vertebrate RBP networks. We therefore focused on the mRNA 3'UTR HCEs of this protein group, to predict RBPs coregulating them. We scanned the HCEs for hidden common elements by the Weeder algorithm (43), searching for 6-to-12 bases long motifs with the tolerance of 1-to-4 mismatches which are observed in at least 25% of the HCEs. The scan produced as best score two 12 bases sequences that can be considered variants of the same sequence motif, as

they differ only in two positions. We speculated that this sequence motif could represent an RBP binding site, since a number of these proteins are known to have a preference for short unstructured sequences or loops in stem-loop secondary structures (53). We then searched for secondary structure motifs in the same 3'UTR HCEs with the RNAfold (44) and the RNAforester (45) algorithms. This analysis resulted in a 17-bases structural motif in the form of a hairpin, whose core loop had a good correspondence (7 out of 12 bases for both sequence motifs; 9 out of 12 bases for sequence motif 2) with the previously identified sequence motifs. Combining the results of both sequence and structure motif searches produced a remarkable concordance, as shown by the alignment in Figure 3A, eventually leading us to a hairpin motif shared by 18 out of the 23 RRM genes reported in Figure 3B. The instances of the hairpin motifs in the mapped 3'UTRs of the 18 genes resulted to be up to four per 3'UTR, with 13 of them harboring only one instance (Figure 3C). We then noticed that this motif had a sequence quite similar to an already known binding site for the HuR (ELAVL1) protein (11). In order to verify that our motif was effectively interacting with HuR, we performed a protein pulldown assay, followed by a western blot with an anti-HuR antibody. Along with the putative HuR motif, we adopted two positive controls (the YB1 and PTB known binding sites), and two mutated and one degenerated loop probes for assaying specificity. The probe design is exemplified in Figure 4A. As reported in Figure 4B, HuR indeed binds to the probe corresponding to our shared motif. Mutated and degenerated probes show very little recovery of HuR, suggesting that the interaction is specific and depending on the loop sequence and size. The positive controls, testifying the correctness of the procedure, are shown in Figures 4C and 4D.

HuR controls a translational network of RRM-type RBPs.

With the motif confirmed to be recognized by HuR, we next sought to understand whether HuR really had a preference for RRM-containing RBP mRNAs, with respect to mRNAs of RBPs bearing other types of RNA binding domains and to mRNAs of proteins bearing the most frequent domains in the genome. To calculate enrichments we took advantage of a recently published HuR PAR-CLIP, therefore unbiased, dataset

(49). We extracted all HuR 3'UTR binding sites from this dataset and derived the corresponding genes. We then computed, by means of the Fisher test, the enrichment in this gene set for: (a) proteins containing the most common experimentally verified RNA binding domains (zinc finger C2H2, KH, SAM, RRM); (b) proteins containing the three absolute most frequent domains in the human genome (IG-like, GPCR superfamily and serine threonine kinase); (c) the complete set of RBPs irrespective of the RNA binding domain. Figure 5A shows that the RRM domain containing gene set resulted to be the only one significantly enriched. This confirms that HuR has a marked preference for binding to the 3'UTR of RRM-bearing mRNAs. We then plotted all 3'UTR HuR targets identified by the PAR-CLIP study along with our group of RRMs, to highlight overlapping and unique genes of the two sets. The resulting intersection is shown in the Venn diagram of Figure 5B and in the network in Figure 5C, which discriminates between gene categories by means of shapes and colors. Fourteen out of the 23 HCE-containing mRNAs for RRM-type proteins are identified as HuR binding, and in particular 4 of them are among the ones we checked by quantitative RIP-PCR (54), see Figure 6A. This last Figure reports the results of a validation sampling of the identified network, both in structural (Figure 6A) and in functional (Figures 6B and 6C) terms. We used HuR overexpressing MCF7 cells, already employed for high throughput studies on HuR (55, 56), firstly to perform 5 quantitative RIP-qPCR assays on the MSI2, RBM15, SRFS11, HNRNPA3 RBP mRNAs (predicted for being bound by HuR), and the CCNA2 (cyclin A) mRNA as a positive control (57). Three RBP mRNAs showed a strong enrichment in the immunoprecipitated pellets, ranging from 200 to 400 fold, with the exception of RBM15 which reported a more modest, but still significant, enrichment (28.3 fold). This proved that these mRNAs are indeed interacting with the HuR RBP in exponentially growing MCF-7 cells. We subsequently infected the same MCF-7 cells with a number of lentiviral silencing shRNAs, and selected those infectants with the strongest HuR inhibition as seen by western blotting (Figure 6B). We then measured the level of polysomally loaded mRNAs for the same 4 RBP genes after sucrose gradient centrifugation (58) and collected the polysomal fractions of both the wild type and the HuR silenced MCF-7 cells. For all the RBP mRNAs tested we found a statistically significant decrease of their localization on polysomes, which demonstrates that binding

of HuR to these 4 RBP mRNAs has a functional effect in promoting their inclusion in polysomes, and therefore their translation. At least for this sample of the network, therefore, we were able to show that HuR acts as a translational enhancer.

DISCUSSION

Despite its widespread role in heavily reprogramming mRNA transcriptome variations (58,59), posttranscriptional control of gene expression has been object of few systematic attempts to map and study the involved circuits. A large number of prediction algorithms and of experimental work has focused on the identification of miRNA/mRNA target sites and of the corresponding inhibitory networks (reviewed respectively in 60 and in 61), while for the RBP/mRNA networks the only available information derives from some high throughput yeast studies (15, 62), suggesting interesting preliminary principles (16, 63). We reasoned that a simple starting point to deal in an unbiased way with core posttranscriptional networks in human cells would be to exploit data on vertebrate phylogenetic conservation by genome-wide alignments, available at the UCSC Genome Browser (32). The original release of this dataset has been already employed by the authors to derive interesting information about, among several other things, UTR conservation for some model genomes (33). We added to the original phastCons (33) algorithm a stronger dependence on completeness of the species tree, in order to increase the sensitivity for really hyper conserved DNA regions. We also restricted the analysis to 3'UTRs for their known regulative power on gene expression (1, 2), and because in the original cited genome-wide comparative study some of the absolute extreme conservation in vertebrates was seen exactly in 3'UTRs of genes regulating other genes, already suggesting widespread posttranscriptional regulation (33). Interestingly, the same trend seemed not to be present in *Drosophila* and *Caenorhabditis* (33). Our derived HCEs were found only in less than 2% of the total 3'UTRs and in a tiny fraction, less than 0.5%, of the total 3'UTR space, being also very short, since 77% of them have an average length of 23 bases. We had therefore the impression to have really sieved a limited number of small RNA stretches with exceptional integrity and permanence through the vertebrates clade, and with potential

biological activity as cis-elements. But for what trans-factors? Using the available information, we showed that these trans factors most likely are not miRNAs or lncRNAs. Instead, several clues bring to the hypothesis that mainly RBP binding sites nest in HCEs. First, many of their most common HCE dimensions are compatible with RNA stretches necessary for interacting with RBP domains (11,53); second, known RBP binding sites are represented with a double density in HCEs with respect to the best scoring comparable random sampling of 3'UTR stretches; third, experimental mRNA-protein interactome signals by PAR-CLIP data (49) are also enriched in HCEs; fourth, HCEs allowed to identify by a simple ontological over-representation analysis the SLBP binding site on histone 3'UTRs, possibly the most unconventional cis-element bound by an RBP identified to date (64), confined to a specific gene class. That the more ancient posttranscriptional networks in vertebrates could involve the action of RBPs on mRNAs is of great interest. We know that RBPs can act both negatively and positively on gene expression, and therefore their combination can build different types of circuits in posttranscriptional networks (16). The yeast genome, devoid of miRNAs (65), contain about 561 RBPs (15), which are presumably the primary actors of the posttranscriptional controls exerted in a concerted way to coordinate topological localization and translation of mRNAs (17). Two recent studies (49, 66) experimentally identify, with comparable methods, the RBP complement of human cells, which appear to consist in about 800 genes whose biological activity is largely still unexplored. We predict from our study that a fraction of these RBPs could be involved in gene expression regulative circuitries appeared at the root of the phylogenesis of vertebrate genomes, and preserved till now in an evolutionary history of more than 500 million years. Given their complete or almost complete sharing in the tested 44 species analyzed, these RBP-based networks are possibly essential in the cell architecture of the bearing organisms, being each of them endowed with unknown but possibly essential biological activities. It would be interesting to assay the degree of persistence of vertebrate HCEs in several invertebrate model genomes, to confirm or deny the suggested lack of conservation (33).

A simple way of getting some information on the possible function of the networks of which HCEs were cis components was to observe the functional polarization, when

gene function was known, of the genes bearing them. This immediately provided us a proof of the good sensitivity of the approach, since the strongest signal detected was the well-known and highly conserved network between the SLBP RBP and the histone gene mRNAs (51). The other most interesting signal found was the tendency of HCEs to be enriched in the 3'UTR of mRNAs of RBPs, especially of those RBPs bearing the RRM as interface with the bound RNAs. Therefore, HCEs not only bore cis elements which were potentially mainly RBP binding sites, but also were enriched in the 3'UTR of mRNAs coding for RBPs. Given previous suggestions about the tendency of posttranscriptional networks to establish short regulative and autoregulative feedbacks both in yeast (67,15) and in mammals (68, 72,58), we were especially intrigued by this finding. Building on it, we thought to be in a good position to reach the main goal of the study, the proof-of-principle of phylogenesis-assisted identification and demonstration of new posttranscriptional networks in human cells, rendered possible by the current wide and detailed genome sequence and annotation in vertebrates. Scanning of the 23 3'UTR HCEs of the selected mRNAs coding for RRM-type RBPs, we found a sequence and structure defined motif which we experimentally demonstrated to be binding site of the HuR RBP (Figure 6). By developing a cell-based inducible model of HuR overexpression, we also showed that the network HuR RBP / RRM-type RBP mRNAs was at least for the four assayed mRNAs, a translation enhancing network, bearing to HuR-induced increase in polysomal localization of the target mRNAs. This finding is compatible with the mRNA stabilizing and translation-promoting function already well documented for HuR (69). Moreover, exploiting unbiased PAR-CLIP interaction data, we confirmed that HuR has a clear preference, at least among vertebrates, for binding mRNAs of RRM-type RBPs (Figure 5). HuR is an essential (70), ubiquitous and intensely studied RBP (71,72), whose nuclear and cytoplasmic action seems to be subsequent to energy metabolism (73,74,75) and cell damage induced stresses (76,77,78), and which has been found to positively regulate a large number of bound mRNAs. Here we add that the RRM-type HuR has an evolutionarily ancient propensity to positively control the translatability of a set of mRNAs coding for other RBPs bearing RRM-type domains. Taken together with the known ability of HuR to bind and regulate its own mRNA (79,80,81), we predict HuR to be a posttranscriptional hub protein

exerting wide and marked effects, both directly and indirectly, through the action of several RRM-type RBPs which on turn control many other mRNAs. Added to the known HuR capability to bind and affect the mRNAs of many transcription factors (46), this finding predicts its ability to heavily influence both posttranscriptional and transcriptional networks, as key ”regulator of regulators”, in vertebrate cells. Interestingly, a HuR orthologue is absent in invertebrate model genomes, and probably arose in vertebrates as duplication of one of the neuron-specific members of the ELAV family (HuB, HuC, HuD), establishing its new role that became essential in all cells (82).

But, on a more general ground, how are these RBP-based posttranscriptional networks physically structured in vertebrates? While by CLIP data RBPs appear to bind, sometimes in a preferential fashion (14, 43), 5’ and 3’ UTRs, nothing is known about their supramolecular organization, if any, on the bound mRNAs. We provide here a first clue on this organization, which from our analysis of HCEs in vertebrate 3’UTRs could result in patterns of small clusters of 3-4 stretches on average (but with a variability from 2 to 28) of continuous sequence, each of them being a potential binding site for one or more contiguous RBPs (Figure 2C). Increase in resolution power of the newly introduced mRNA transcriptome-wide clipping technique (46) could provide a detailed enough map of RNA-protein contact points to confirm or deny this model. It is likely that the HCE length and cluster organization could derive from RNA-dependent and RNA-independent RBP interactions on the 3’UTRs. Several RBPs are known to undergo homo or hetero dimerization and oligomerization (83, 84, 85), which could represent the structural basis for the formation of complexes.

We finally note that, being our signals coming from exceptional conservation of collinear RNA sequences in the vertebrate clade, it implies a strict coevolution pattern of the trans-factors involved. If the trans-factors result to be orthologue, this could suggest to attribute the essential biological activities responsible for high conservation much less to single trans-factors than to the supramolecular complexes they form. Which means, in other words, that the possible presence of small intermittent ribonucleoprotein clusters as preferred organization scheme along the 3’UTR length could impose the study of these clusters instead of the single forming RBPs to understand function.

With this work, we provide evidence that tailored phylogenetic analyses based on genome sequence information can allow us to prioritize potential cis-element in posttranscriptional networks, providing a way for their experimental identification and suggesting clues for the definition of their topology.

FUNDING

This work was funded by Fondazione CARITRO, Trento, Italy, in the frame of the TRADENT project (2010-2012).

ACKNOWLEDGEMENTS

We would like to thank Toma Tebaldi and Angela Re for the precious comments and suggestions at the initial stages of this work.

REFERENCES

1. Mazumder B, Seshadri V, & Fox PL. (2003) Translational control by the 3'-UTR: the ends specify the means. Trends Biochem. Sci, 28, 91-98.
2. Andreassi C et al. (2009) To localize or not to localize: mRNA fate is in 3'UTR ends. Trends in cell biology, 19(9), 465-474.
3. Guo H, Ingolia NT, Weissman JS, Bartel DP. (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature, 466(7308), 835-40.
4. Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. (2008) Widespread changes in protein synthesis induced by microRNAs. Nature, 455(7209), 58-63.
5. Filipowicz W, Bhattacharyya SN, Sonenberg N. (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat. Rev. Genet., 9(2), 102-14.
6. Glisovic T et al. (2008) RNA-binding proteins and post-transcriptional gene regulation. FEBS letters, 582(14), 1977-1986.
7. Jacobs GH, Chen A, Stevens SG et al. (2009) Transterm: a database to aid the analysis of regulatory sequences in mRNAs. Nucleic Acids Res., 37(Database issue), D72-76.

8. Grillo G, Turi A, Licciulli F, 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(Database issue), D75-80.
9. Ellington A, Szostak J. (1990) In vitro selection of RNA molecules that bind specific ligands. Nature 346, 818 - 822.
10. Tuerk C, Gold L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science, 249(4968), 505-10.
11. Ray D, Kazan H, Chan ET, Peña Castillo L, Chaudhry S, Talukder S, Blencowe BJ et al. (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat Biotechnol., 27(7), 667-70.
12. Keene JD, Komisarow JM, Friedersdorf MB. (2006) RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nat Protoc., 1(1), 302-7.
13. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. (2003) CLIP Identifies Nova-Regulated RNA Networks in the Brain. Science, 302(5648), 1212-5.
14. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A et al. (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell, 141(1), 129-41.
15. Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO. (2008) Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. PLoS Biol, 6(10), e255.
16. Kanitz A, Gerber AP. (2010) Circuitry of mRNA regulation. Wiley Interdiscip Rev Syst Biol Med., 2(2), 245-51.
17. Gerber AP, Herschlag D, Brown PO. (2004) Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. PLoS Biol., 2(3), E79.
18. Dassi E, Malossini A, Re A, Tebaldi T, Mazza T, Caputi L, Quattrone A. (2011) AURA: Atlas of UTR Regulatory Activity. Bioinformatics, doi: 10.1093/bioinformatics/btr608.

19. Khorshid M, Rodak C, Zavolan M. (2011) CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. Nucleic Acids Res., 39(suppl 1), D245-D252.
20. Anders G, Mackowiak SD, Jens M, Maaskola J, Kuntzagk A, Rajewsky N, Landthaler M et al. (2012) doRiNA: a database of RNA interactions in post-transcriptional regulation. Nucleic Acids Res., 40(Database issue), D180-6.
21. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. (2004) Ultraconserved elements in the human genome. Science., 304(5675), 1321-5.
22. Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR et al. (2007) Human genome ultraconserved elements are ultraselected. Science, 317(5840), 915.
23. McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC. (2012) Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res., 22(4), 746-54.
24. Reneker J, Lyons E, Conant GC, Pires JC, Freeling M, Shyu CR, Korkin D. (2012) Long identical multispecies elements in plant and animal genomes. Proc Natl Acad Sci, 109(19), E1183-91.
25. Taccioli C, Fabbri E, Visone R, Volinia S, Calin GA, Fong LY, Gambari R et al. (2009) UCbase & miRfunc: a database of ultraconserved sequences and microRNA function. Nucleic Acids Res., 37(Database issue), D41-8.
26. Sathirapongsasuti JF, Sathira N, Suzuki Y, Huttenhower C, Sugano S. (2011) Ultraconserved cDNA segments in the human transcriptome exhibit resistance to folding and implicate function in translation and alternative splicing. Nucleic Acids Res., 39(6), 1967–1979.
27. Shabalina SA, Ogurtsov AY, Lipman DJ, Kondrashov AS. (2003) Patterns in interspecies similarity correlate with nucleotide composition in mammalian 3’UTRs. Nucleic Acids Res., 31(18), 5433-5439.
28. Shabalina SA, Ogurtsov AY, Rogozin IB, Koonin EV, Lipman DJ. (2004) Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. Nucleic Acids Res., 32(5), 1774-1782.

29. Christley S, Lobo NF, Madey G (2008). Multiple organism algorithm for finding ultraconserved elements. BMC Bioinformatics, 9, 15.
30. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P et al. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature, 478(7370), 476-82.
31. Mignone F, Gissi C, Liuni S, Pesole G. (2002) Untranslated regions of mRNAs. Genome Biology, 3(3), reviews0004.1–0004.10.
32. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M et al. (2011) The UCSC Genome Browser database: update 2011.Nucleic Acids Res., 39(suppl 1), D876-D882.
33. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res., 15(8), 1034-50.
34. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA et al. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures.Nature, 450(7167),219-232.
35. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human microRNA Targets. PLoS Biol. 2004; 2(11):1862-1879.
36. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ et al. Combinatorial microRNA target predictions. Nat. Genet 2005; 37:495-500.
37. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat. Genet 2007; 39(10):1278-1284.
38. Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. (2011) lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res, 39, D146-151.
39. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. (2008) BLAST+: architecture and applications. BMC Bioinformatics, 10, 421.
40. Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. (2011) RBPDB: a database of RNA-binding specificities. Nucleic Acids Res., 39(suppl 1), D301-D308.
41. Huang DW, Sherman BT, Lempicki RA. (2009) Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc., 4 (1), 44-57.

42. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F et al. (2007) ClustalW and ClustalX version 2 (2007). Bioinformatics, 23(21), 2947-2948.
43. Pavesi G, Mereghetti P, Zambelli F, Stefani M, Mauri G, Pesole G. (2006) MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. Nucleic Acids Res., 34(Web Server issue), W566-570.
44. Lorenz R, Bernhart SH, Hoener Zu Siederdissen C, Tafer H, Flamm C, Stadler PF et al. (2011) ViennaRNA Package 2.0. Algorithms Mol Biol., 6(1), 26.
45. Höchsmann M, Töller T, Giegerich R, Kurtz S. (2003) Local similarity in RNA secondary structures. Proc IEEE Comput Soc Bioinform Conf, 2, 159-68.
46. Lebedeva S, Jens M, Theil K, Schwanhäusser B, Selbach M, Landthaler M, Rajewsky N. (2011) Transcriptome-wide Analysis of Regulatory Interactions of the RNA-binding protein HuR. Molecular Cell, 43, doi:10.1016/j.molcel.2011.06.008.
47. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D et al. (2003) Human-mouse alignments with BLASTZ. Genome Res., 13(1), 103-7.
48. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res., 14(4), 708-15.
49. Baltz, AG, Munschauer M, Schwanhäusser B, Vasile A, Murakawa Y, Schueler M, Youngs N, et al. (2012) The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. Molecular Cell, 46(5), 674–690.
50. Marzluff WF, Wagner EJ, Duronio RJ. (2008) Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. Nat Rev Genet., 9(11), 843-54.
51. Dávila López M, Samuelsson T. (2008) Early evolution of histone mRNA 3' end processing. RNA, 14(1), 1-10.
52. Davin Townley-Tilson WH, Pendergrass SA, Marzluff WF, Whitfield ML. (2006) Genome-wide analysis of mRNAs bound to the histone stem-loop binding protein. RNA, 12, 1853-1867.
53. Auweter SD, Oberstrass FC, Allain FH. (2006) Sequence-specific binding of single-stranded RNA: is there a code for recognition? Nucleic Acids Res., 34(17), 4943-59.

54. Gilbert C, Svejstrup JQ. (2006) RNA immunoprecipitation for determining RNA-protein associations in vivo. Curr Protoc Mol Biol., Chapter 27, Unit 27.4.
55. Latorre E, Tebaldi T, Viero G, Spartà AM, Quattrone A, Provenzani A. (2012) Downregulation of HuR as a new mechanism of doxorubicin resistance in breast cancer cells. Mol Cancer, 11, 13.
56. Mazan-Mamczarz K, Hagner PR, Corl S, Srikantan S, Wood WH, Becker KG, Gorospe M et al. (2008) Post-transcriptional gene regulation by HuR promotes a more tumorigenic phenotype. Oncogene, 27(47), 6151-63.
57. Wang W, Caldwell MC, Lin S, Furneaux H, Gorospe M. HuR regulates cyclin A and cyclin B1 mRNA stability during cell proliferation. EMBO J 2000; 19:2340-50.
58. Tebaldi T, Re A, Viero G, Pegoretti I, Passerini A, Blanzieri E, Quattrone A. (2012) Widespread uncoupling between transcriptome and translatome variations after a stimulus in mammalian cells. BMC Genomics, 13(1), 220.
59. Bates JG, Salzman J, May D, Garcia PB, Hogan GJ, McIntosh M, Schlissel MS et al. (2012) Extensive gene-specific translational reprogramming in a model of B cell differentiation and Abl-dependent transformation. PLoS One, 7(5), e37108.
60. Saito T, Saetrom P. (2010) MicroRNAs--targeting and target prediction. Nat Biotechnol., 27(3), 243-9.
61. Ebert MS, Sharp PA. (2012) Roles for microRNAs in conferring robustness to biological processes. Cell, 149(3), 515-24.
62. Scherrer T, Mittal N, Janga SC, Gerber AP. (2010) A screen for RNA-binding proteins in yeast indicates dual functions for many enzymes. PLoS One, 5(11), e15499.
63. Mittal N, Roy N, Babu MM, Janga SC. (2009) Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. Proc Natl Acad Sci U S A., 106(48), 20300-5.
64. Battle DJ, Doudna JA. (2001) The stem-loop binding protein forms a highly stable and specific complex with the 3' stem-loop of histone mRNAs. RNA, 7(1), 123-132.
65. Aravind L, Watanabe H, Lipman DJ, Koonin EV. (2000) Lineage-specific loss and divergence of functionally linked genes in eukaryotes. Proc Natl Acad Sci U S A, 97(21), 11319–11324.

66. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE et al. (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell, 149(6), 1393-406.
67. Mittal N, Scherrer T, Gerber AP, Janga SC. (2011) Interplay between posttranscriptional and posttranslational interactions of RNA-binding proteins. J Mol Biol., 409(3), 466-79.
68. Meyuhas O. (2000) Synthesis of the translational apparatus is regulated at the translational level. Eur J Biochem., 267(21), 6321-30.
69. Brennan CM, Steitz JA. (2001) HuR and mRNA stability. Cell Mol Life Sci., 58(2), 266-77.
70. Katsanou V, Milatos S, Yiakouvaki A, Sgantzis N, Kotsoni A, Alexiou M, Harokopos V et al. (2009) The RNA-binding protein Elavl1/HuR is essential for placental branching morphogenesis and embryonic development. Mol Cell Biol., 29(10), 2762-76.
71. Abdelmohsen K, Gorospe M. (2010) Posttranscriptional regulation of cancer traits by HuR. Wiley Interdiscip Rev RNA, 1(2), 214-29.
72. Srikantan S, Gorospe M. (2012) HuR function in disease. Front Biosci., 17, 189-205.
73. Wang W, Fan J, Yang X, Fürer-Galban S, Lopez de Silanes I, von Kobbe C, Guo J et al. (2002) AMP-activated kinase regulates cytoplasmic HuR. Mol Cell Biol., 22(10), 3425-36.
74. Wang W, Yang X, Kawai T, López de Silanes I, Mazan-Mamczarz K, Chen P, Chook YM et al. (2004) AMP-activated protein kinase-regulated phosphorylation and acetylation of importin alpha1: involvement in the nuclear import of RNA-binding protein HuR. J Biol Chem., 279(46), 48376-88.
75. Zou T, Liu L, Rao JN, Marasa BS, Chen J, Xiao L, Zhou H et al. (2008) Polyamines modulate the subcellular localization of RNA-binding protein HuR through AMP-activated protein kinase-regulated phosphorylation and acetylation of importin alpha1.
Biochem J., 409(2), 389-98.
76. Kim HH, Abdelmohsen K, Gorospe M. (2010) Regulation of HuR by DNA Damage Response Kinases. J Nucleic Acids, pii: 981487.

77. Rhee WJ, Ni CW, Zheng Z, Chang K, Jo H, Bao G. (2010) HuR regulates the expression of stress-sensitive genes and mediates inflammatory response in human umbilical vein endothelial cells. Proc Natl Acad Sci U S A, 107(15), 6858-63.
78. Von Roretz C, Di Marco S, Mazroui R, Gallouzi IE. (2011) Turnover of AU-rich-containing mRNAs during stress: a matter of survival. Wiley Interdiscip Rev RNA., 2(3), 336-47.
79. Pullmann R Jr, Kim HH, Abdelmohsen K, Lal A, Martindale JL, Yang X, Gorospe M. (2007) Analysis of turnover and translation regulatory RNA-binding protein expression through binding to cognate mRNAs. Mol Cell Biol., 27(18), 6265-78.
80. Al-Ahmadi W, Al-Ghamdi M, Al-Haj L, Al-Saif M, Khabar KS. (2009) Alternative polyadenylation variants of the RNA binding protein, HuR: abundance, role of AU-rich elements and auto-Regulation. Nucleic Acids Res., 37(11), 3612-24.
81. Dai W, Zhang G, Makeyev EV. (2012) RNA-binding protein HuR autoregulates its expression by promoting alternative polyadenylation site usage. Nucleic Acids Res., 40(2), 787-800.
82. Good PJ. (1995) A conserved family of elav-like genes in vertebrates. Proc Natl Acad Sci U S A., 92(10), 4557-61.
83. Fialcowitz-White EJ, Brewer BY, Ballin JD, Willis CD, Toth EA, Wilson GM. (2007) Specific protein domains mediate cooperative assembly of HuR oligomers on AU-rich mRNA-destabilizing sequences. J. Biol. Chem., 282, 20948-20959.
84. Pedrotti S, Busà R, Compagnucci C, Sette C. (2012) The RNA recognition motif protein RBM11 is a novel tissue-specific splicing regulator. Nucleic Acids Res., 40(3), 1021-32.
85. Martel C, Dugré-Brisson S, Boulay K, Breton B, Lapointe G, Armando S, Trépanier V et al. (2010) Multimerization of Staufen1 in live cells. RNA, 16(3), 585-97.

FIGURE LEGENDS

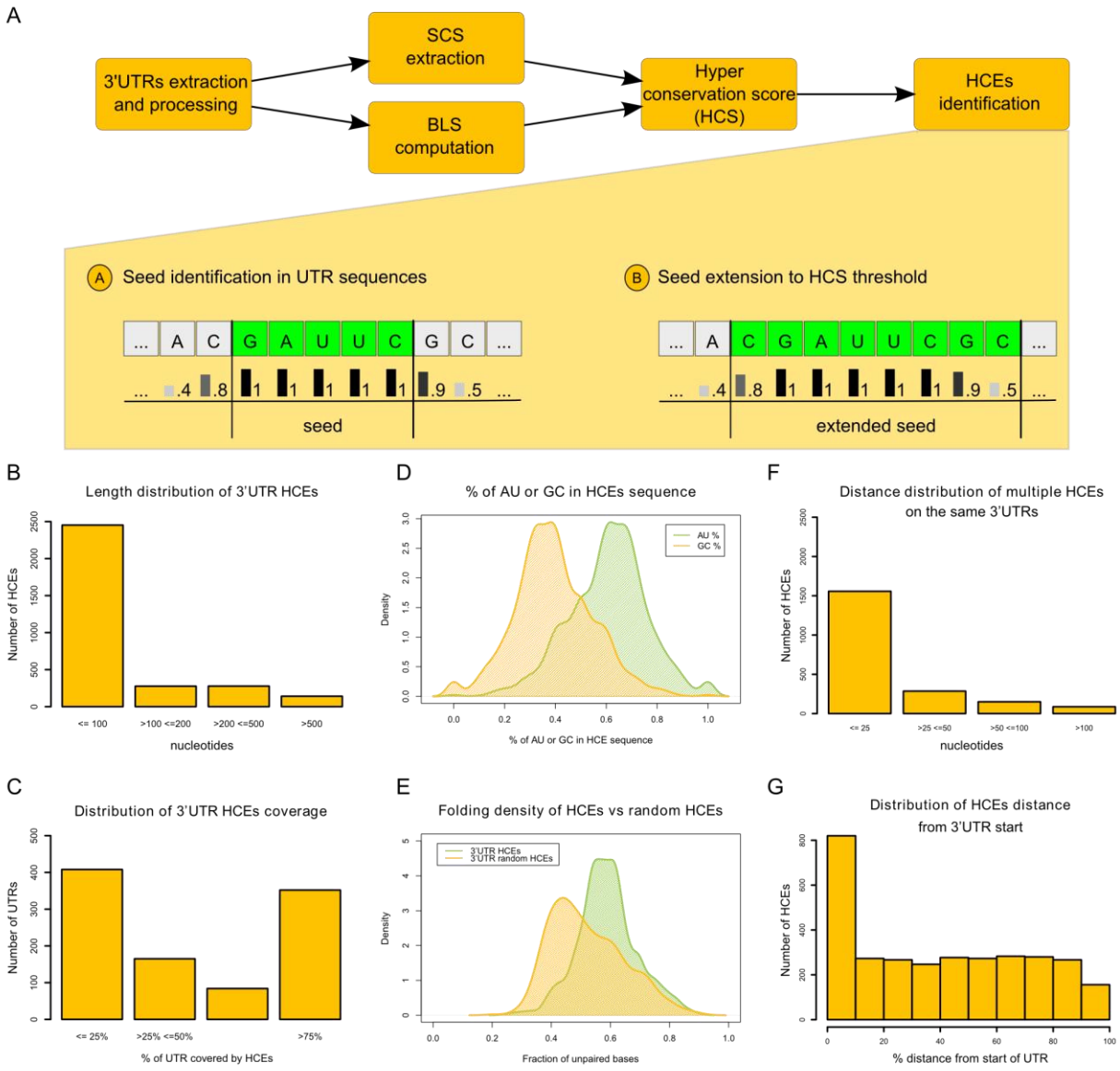


Figure 1. HCEs are short, scattered and highly structured.

The overall HCE identification pipeline is shown in **A.**, with the lower part detailing the algorithm used searching for seeds and extending them to lead to the final HCEs. B.-G. highlights the most relevant features of the HCEs. **B.** shows the length distribution of HCEs and **C.** their percent coverage of 3'UTRs; **D.** displays the predominance of AU base pairs content over CG base pairs in HCE bases composition and **E.** the prevalence of highly structured HCEs, as indicated by the shown distribution of secondary structure density in HCEs. **F.** displays the distribution of distances between

HCEs on the same 3'UTRs and **G.** the HCE distance distribution from the 3'UTR start, indicated in percent over the 3'UTR length.

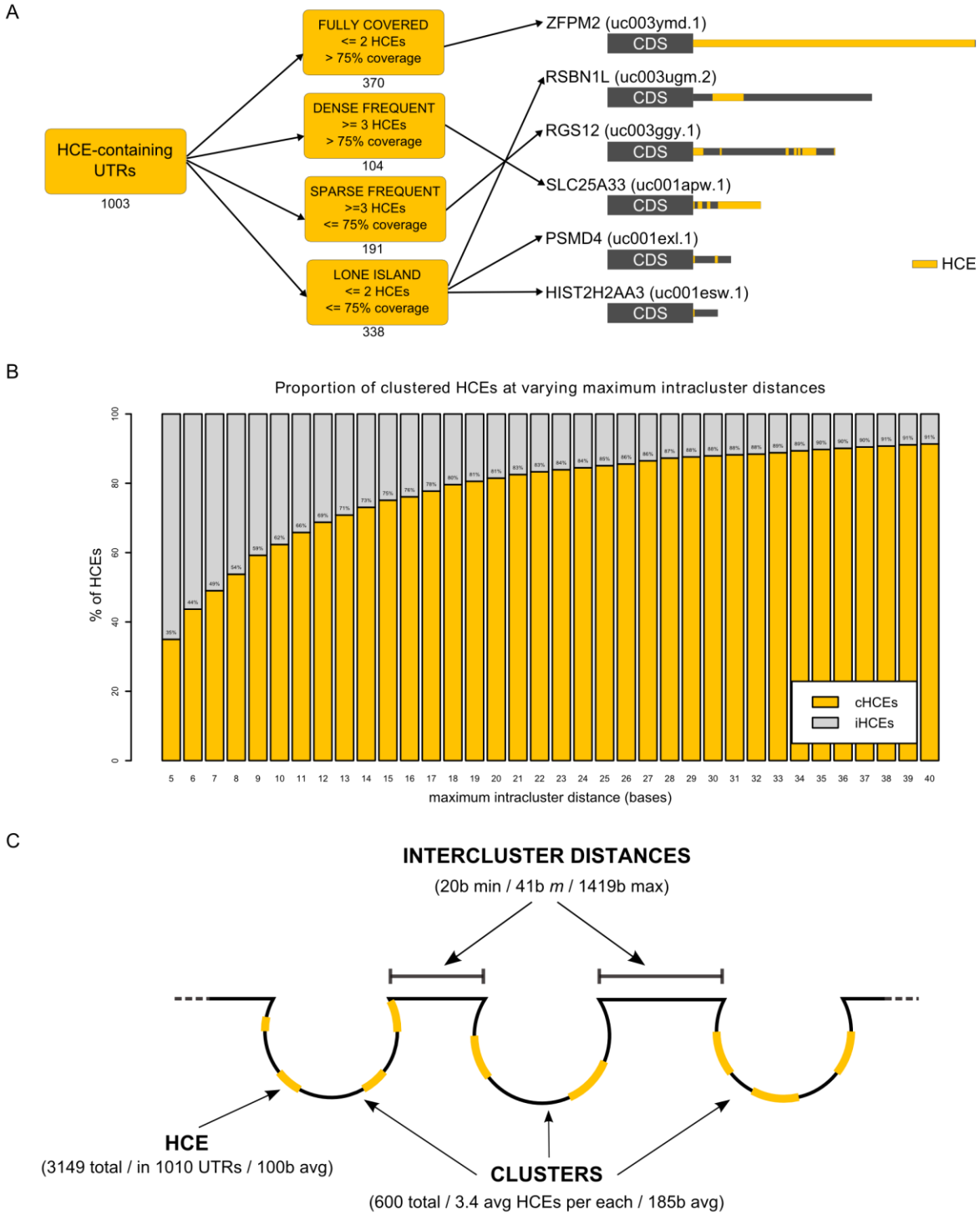


Figure 2. HCEs can be classified according to their pattern of occurrence in 3'UTRs and are organized in clusters.

A. shows the classification of 3'UTRs in four classes, according to their HCE content (on the left). Numbers below each class box are the number of HCE-containing 3'UTRs belonging to the class. On the right, a sample of six HCE-containing 3'UTRs: HCEs are mapped onto their 3'UTR and represented as yellow areas, being a grey rectangle the full-length 3'UTR. Arrows from class boxes to UTRs indicate which UTR belongs to which class. **B.** displays the increasing percentage of clustered HCEs when increasing the maximum intracluster distance allowed for an HCE to be considered part of a cluster. We span from 5 to 40 bases, and at 20 bases we can observe the beginning of a plateau. We therefore chose 20 bases as the maximum intracluster distance to consider. The graph is drawn excluding the 577 HCEs which are unique on their respective 3'UTR. **C.** Graphical representation of the proposed model of trans-factor binding to 3'UTRs, assuming that HCEs are binding sites of one or more trans-factors. by intercluster RNA stretches of variable length (from 20 to 1419 bases), suggesting a coordinated action on the 3'UTR.

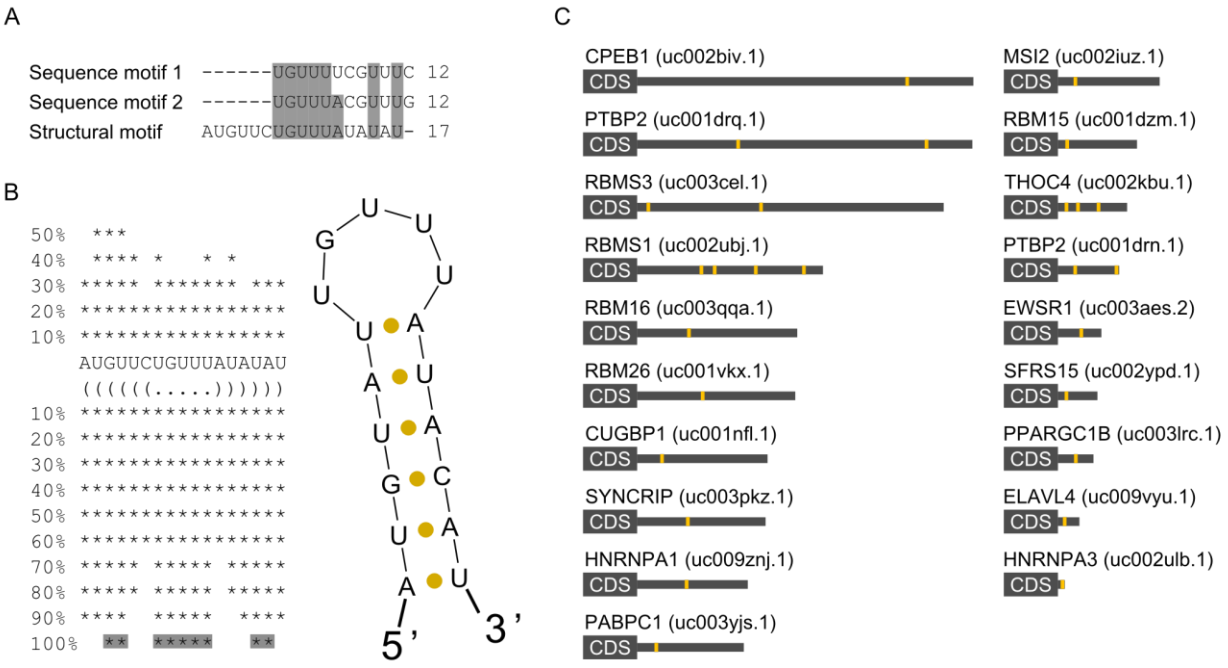


Figure 3. HCEs in mRNAs encoding RRM-type RBPs share a sequence and secondary structure motif.

HCEs contained in the group of RRM-type RBP genes 3'UTRs were scanned for both sequence and secondary structure motifs. The first search returned two, almost identical, 12-bases motifs; the second one produced a 17 bases hairpin which, after multiple alignment, emerged to contain a 12-bases core markedly similar to previously identified sequence motifs. This core represents the loop part of the hairpin which, as the two searches are quite concordant on it, may indeed represent a binding motif for the trans-factor of the regulatory network we are trying to identify. **A.** shows the alignment between sequence and secondary structure motifs. **B.** shows the secondary structure motif and its bidimensional structure. **C.** Motif instances (yellow areas) mapped on the full length 3'UTR (grey rectangle) of the 19 RRM-type RBP mRNAs. HGNC gene names are on the left, UCSC UTR names are on the right in parenthesis.

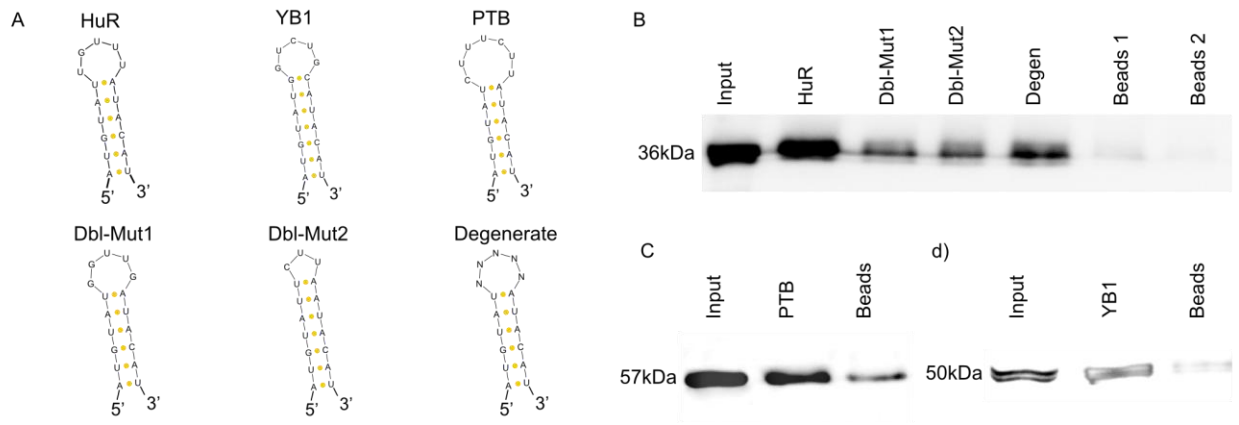


Figure 4. HuR is a trans-factor binding in vitro to the HCE motif shared by mRNAs encoding RRM-type RBPs.

The different RNA probes employed for the protein pull-down experiment are shown in **A.** HuR pulldown probe: this probe was designed by using the secondary structure motif reported in Figure 3, slightly modifying the lowest part of the hairpin so as to make it fold correctly when not in context. The loop part was designed by employing the most probable bases of the sequence and structure motifs. Positive controls probes are the known binding sites for the YB1 and PTB RBPs, experimentally obtained (11). Again, the lowest part of the stem was slightly modified so as to make it fold as desired. Negative controls HuR probes are Dbl-Mut1, Dbl-Mut2 and Degenerate. The Degenerate probe was synthesized by allowing all four nucleotides to be present at each loop position, so to obtain a mixture of probes bearing all the possible 5-mers loops. The Dbl-Mut1 and Dbl-Mut2 probes were obtained by mutating two bases of the original probe loop, in a way to preserve it in the first case and to obtain a 3-mer loop instead of a 5-mer loop in the second case. **B.** shows the HuR pull-down western blots. From the leftmost band to the rightmost: Input, HuR probe, Dbl-Mut1, Dbl-Mut2, Degenerate probe and denaturated beads bands. As can be readily seen, the hairpin probes bind to HuR with a marked specificity for the correct one**. C.-D.** YB1 and PTB RBPs pull-down. From the leftmost band to the rightmost: input, YB1/PTB probe, and denaturated beads. As shown by western blotting, the hairpin probes bind to PTB and YB1 respectively, thus confirming that the pull-down protocol works as expected.

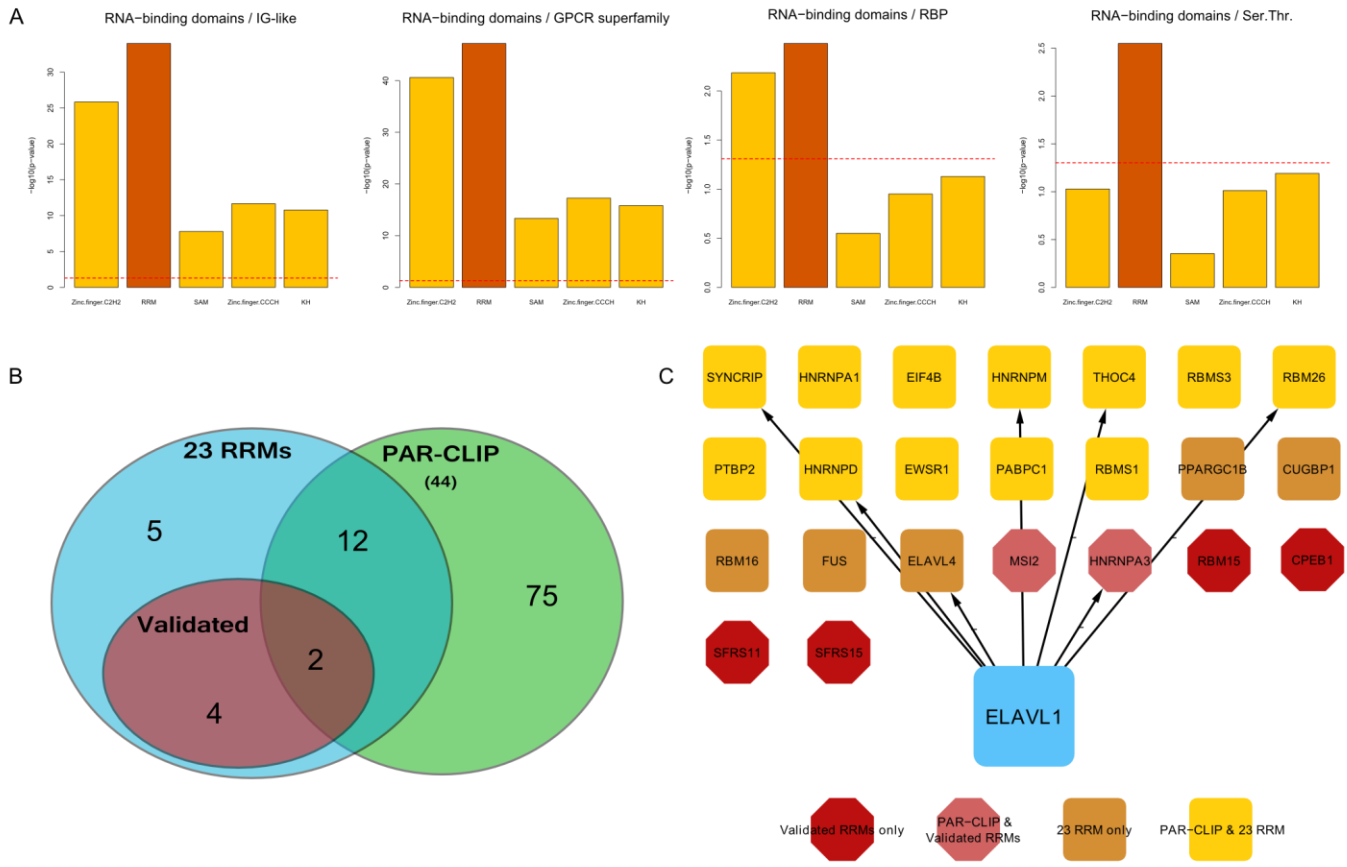


Figure 5. HuR has a preference for the binding of the 3'UTR of RRM-type RBPs.

A. shows the enrichment of HuR 3'UTR binding sites for several RNA-binding domains with respect to the most frequent human protein domains and to RBPs as a whole. Data is extracted by the PAR-CLIP experiment published in (44). **B.** shows a Venn diagram indicating the overlap between our HuR RRM-type mRNA targets and the experimentally identified HuR PAR-CLIP RRM-type mRNA targets. **C.** displays HuR 3'UTR RRM-type mRNA targets, highlighted in different colors and shapes according to their belonging to our set of 23 mRNAs, to mRNAs we validated by RIP-qPCR and their intersection with the RRM-type mRNA targets from the PAR-CLIP dataset.

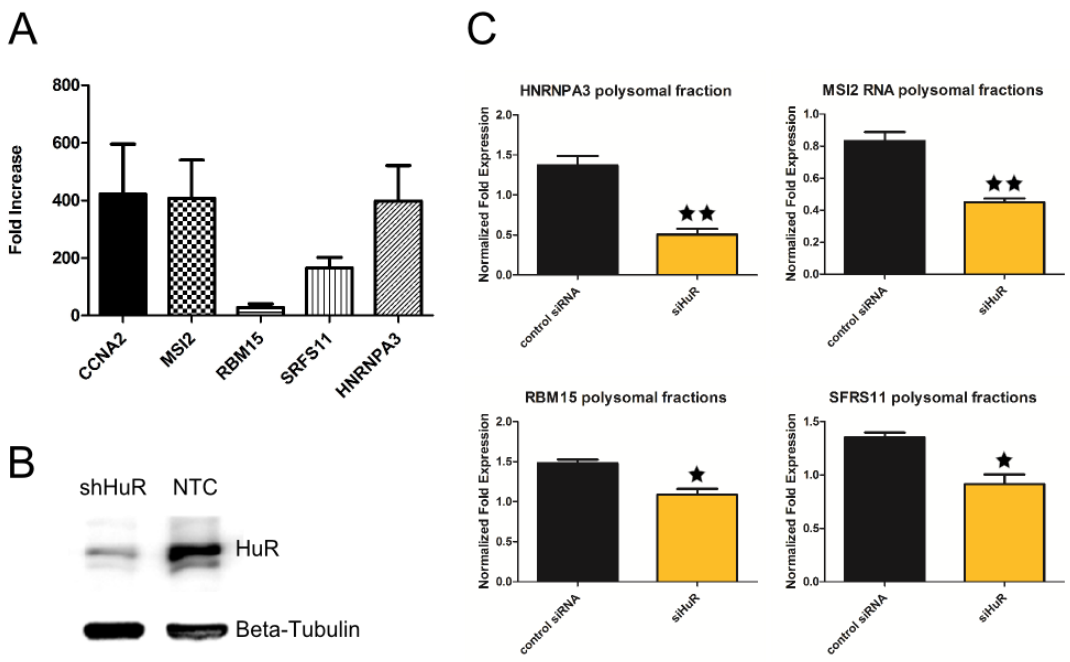


Figure 6. The network of HuR binding to mRNAs for RRM-type RBPs is a functional translational network.

A. shows the fold enrichment results (with respect to control) for four predicted RBP mRNAs (plus the CCNA2 mRNA as control) subjected to ribonucleoprotein immunoprecipitation (RIP) from lysates of HuR overexpressing MCF-7 cells and quantitative RT-PCR, demonstrating interaction of HuR with these mRNAs. **B.** reports the western blot confirming HuR silencing in MCF-7 cell line. Beta-tubulin is used as housekeeping gene. **C.** shows the statistically significant decrease of mRNA levels for the same four RRM-type RBP mRNAs, indicating a translational enhancing effect of HuR on these mRNAs. Increasing level of significance (0.05, 0.01) is indicated by one or two stars.

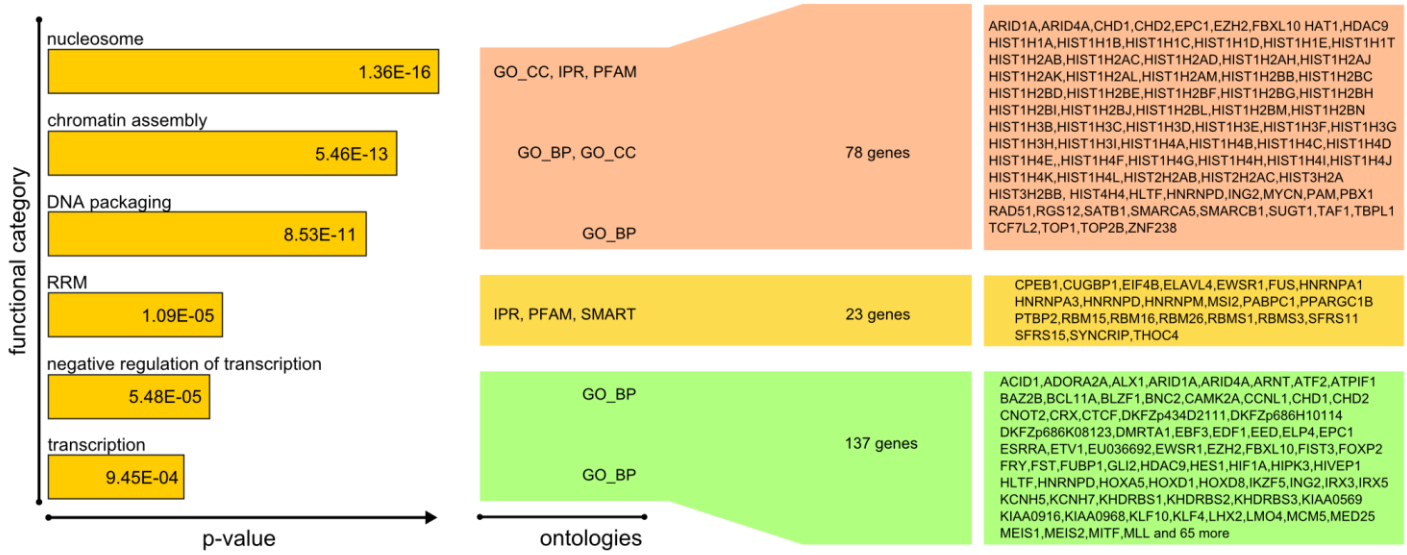


Figure S1. HCEs cluster in genes belonging to three different biological functions. Ontology enrichment analysis of HCEs-containing genes highlights three groups of genes corresponding to three different biological functions. Multiple ontologies were used to infer possible functional groupings: the top results are a most significant group composed of genes involved in chromosomes assembly, a significant set consisting of 23 genes coding for RRM-containing genes for RBPs and a third, less significant group of genes playing a role in transcription. Here the ontology terms clusters giving rise to these groups are shown, along with their enrichment p-value and the final list of involved genes.

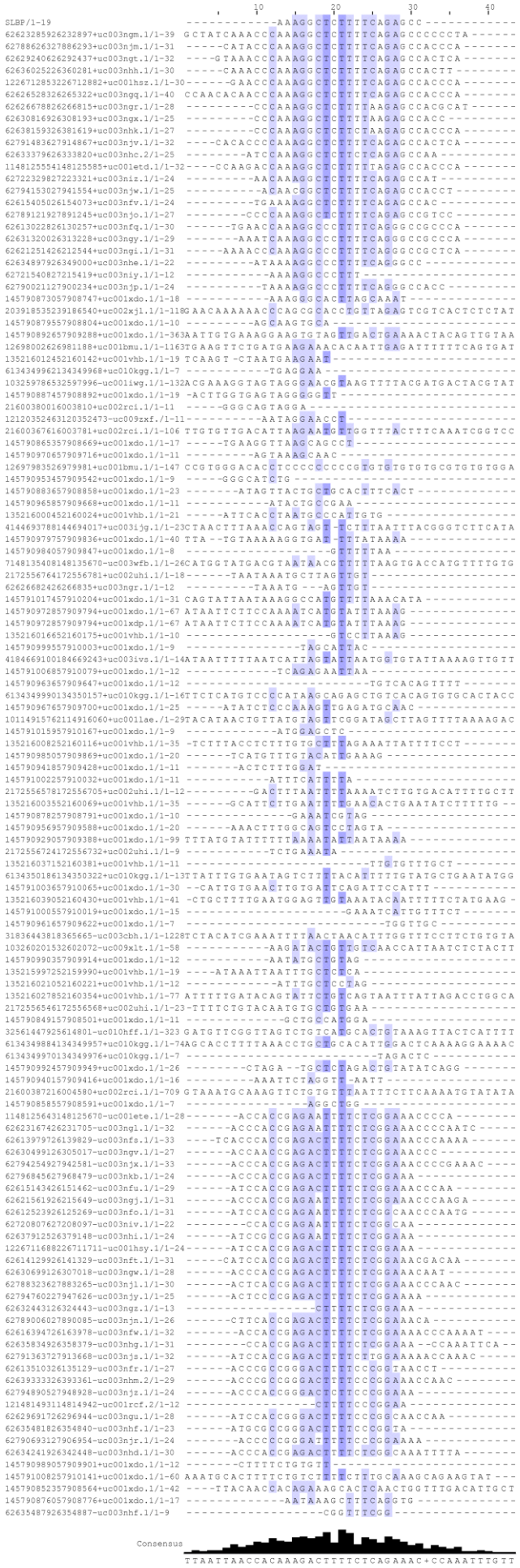


Figure S2. HCEs in 3'UTR of chromosome assembly genes identify SLBP binding sites.

A significant fraction of HCEs found in the 3'UTR of genes belonging to the chromosome assembly functional group was noticed to harbor a sequence corresponding to the binding motif of the stem-loop binding protein (SLBP), which is known to bind to the 3'UTR of histone genes and to stabilize the mRNA in order to compensate for the absence of a poly(A) tail. This stabilization mechanism is known to be heavily conserved and can thus be considered as a benchmark for our HCE identification method. Here the ClustalW2 alignment of these HCEs with the SLBP binding motif (the first sequence in the alignment) is displayed.

1. ON THE COMPOSITION OF HYPER-CONSERVATION SCORE (HCS)

We defined the sequence conservation measure, which we call Hyper Conservation Score (HCS), by first selecting one of the two conservation measures defined for the 44-way alignments available at the UCSC genome browser (1). We choose as Sequence Conservation Score (SCS) the *phastCons*-derived metric (2) instead of the *phyloP* one (3), as the former considers neighboring bases in determining a base score, being thus sensible to stretches of conserved bases: this fact makes it more suitable for identifying conserved elements than *phyloP*, which instead computes conservation independently at each position. *phastCons* takes into account the phylogenetic tree to estimate the probabilities for bases to be conserved or not in the HMM models it is based upon. Nevertheless, being our aim to identify exceptionally conserved sequence stretches because of their potential functional meaning as cis components of core posttranscriptional networks, we estimated as essential the requirement for sharing of the sequences among the different vertebrate species considered. To put more weight on the phylogenetic distance, we included in our metric the Branch Length Score (BLS) as introduced in a comparison between close *Drosophila* species (4). This measure is the proportion of the distance covered by the branches of the phylogenetic tree by the alignment of a particular sequence, thus giving more importance to elements conserved across a wide range of species than to the ones restricted to a group of closely related species. We argued that, while phylogenetic information are already included in SCS, BLS would have been not redundant. To verify this we computed the Pearson correlation coefficient between SCS and BLS, obtaining a value of 0.48, which indicates only a moderate correlation of the two components of our HCS. This result confirms that the BLS usefully complements the SCS.

We further had to find a convenient measure of relative weight of SCS and BLS in HCS. We performed several runs of our pipeline, varying the SCS-BLS score composition from SCS only (100%-0%) to BLS only (0%-100%), through five intermediate proportions (80%-20%; 60%-40%; 50%-50%; 40%-60%; 20%-80%). What we obtain as result is a progressive increase in HCE sizes in parallel with a marked reduction of their total number. While more than 120000 HCEs are produced in the first two runs (100%-

0%, 80%-20%), only 3149 are retained in the half-half proportion (50%-50%), and this number goes down to just 232 HCEs for the BLS-only run. Median and average HCE lengths increase respectively from 62 and 17 bases to 114 and 249 bases: the 50%-50% case has a median length of 23 bases and an average length of 100 bases. We selected the 50% SCS and 50% BLS composition as our final conservation measure, because of the number of selected HCEs identified a small percentage of the total UTR space (0.47%) and a corresponding small percentage of mRNAs (1.8%). With this choice we believed to have greatly reduced the number of false positives HCEs in our final dataset.

2. HuR SILENCING SEQUENCES

TRCN0000017274:
CCGGGAGAACGAATTTGATCGTCAACTCGAGTTGACGATCAAATTCGTTCTCTTTTT

TRCN0000017273:
CCGGCGTGGATCAGACTACAGGTTTCTCGAGAAACCTGTAGTCTGATCCACGTTTTT

TRCN0000017277:
CCGGGCAGCATTGGTGAAGTTGAATCTCGAGATTCAACTTCACCAATGCTGCTTTTT

TRCN0000017275:
CCGGACCATGACAAACTATGAAGAACTCGAGTTCTTCATAGTTTGTCATGGTTTTTT

REFERENCES

1. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M et al. (2011) The UCSC Genome Browser database: update 2011.Nucleic Acids Res., 39(suppl 1), D876-D882.

2. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res., 15(8), 1034-50.

3. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res., 20(1), 110-21.

3. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, Carlson JW, Crosby MA et al. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures.Nature, 450(7167),219-232.

Supplementary Table 1: List of the 23 HCE-containing RRM-type RBP identified by our pipeline. Listed are gene symbol, name, Uniprot gene function description and wether the protein contains only RRM or also other domains.

Gene symbol	Gene name	Uniprot description	RRM-only architecture
CPEB1	Cytoplasmic polyadenylation element-binding protein 1	Sequence-specific RNA-binding protein that regulates mRNA cytoplasmic polyadenylation and translation initiation during oocyte maturation, early development and at postsynapse sites of neurons. Binds to the cytoplasmic polyadenylation element (CPE), an uridine-rich sequence element (consensus sequence 5'-UUUUUAU-3') within the mRNA 3'-UTR. In absence of phosphorylation and in	**v**

		association with TACC3 is also involved as a repressor of translation of CPE-containing mRNA; a repression that is relieved by phosphorylation or degradation	
CUGBP1	CUGBP Elav-like family member 1	RNA-binding protein implicated in the regulation of several post-transcriptional events. Involved in pre-mRNA alternative splicing, mRNA translation and stability. Mediates exon inclusion and/or exclusion in pre-mRNA that are subject to tissue-specific and developmentally regulated alternative splicing.	**v**
EIF4B	Eukaryotic translation initiation factor 4B	Required for the binding of mRNA to ribosomes. Functions in close association with EIF4-F and EIF4-A. Binds near the 5'-terminal cap of mRNA in presence of EIF-4F and ATP. Promotes the ATPase activity and the ATP-dependent RNA unwinding activity of both EIF4-A and EIF4-F.	**v**
ELAVL4	ELAV-like protein 4	May play a role in neuron-specific RNA processing. Protects CDKN1A mRNA from decay by binding to its 3'-UTR. Binds to AU-rich sequences (AREs) of target mRNAs, including VEGF and FOS mRNA.	**v**
EWSR1	RNA-binding protein EWS	Might normally function as a repressor. EWS-fusion-proteins (EFPS) may play a role in the tumorigenic	**x**

		process. They may disturb gene expression by mimicking, or interfering with the normal function of CTD-POLII within the transcription initiation complex. They may also contribute to an aberrant activation of the fusion protein target genes.	
FUS	RNA-binding protein FUS	Binds both single-stranded and double-stranded DNA and promotes ATP-independent annealing of complementary single-stranded DNAs and D-loop formation in superhelical double-stranded DNA. May play a role in maintenance of genomic integrity.	**x**
HNRNPA1	Heterogeneous nuclear ribonucleoprotein A1	Involved in the packaging of pre-mRNA into hnRNP particles, transport of poly(A) mRNA from the nucleus to the cytoplasm and may modulate splice site selection. May play a role in HCV RNA replication.	**v**
HNRNPA3	Heterogeneous nuclear ribonucleoprotein A3	Plays a role in cytoplasmic trafficking of RNA. Binds to the cis-acting response element, A2RE. May be involved in pre-mRNA splicing.	**v**
HNRNPD	Heterogeneous nuclear ribonucleoprotein D0	Binds with high affinity to RNA molecules that contain AU-rich elements (AREs) found within the 3'-UTR of many proto-oncogenes and cytokine mRNAs. Also binds to double- and single-	**v**

		stranded DNA sequences in a specific manner and functions a transcription factor. Each of the RNA-binding domains specifically can bind solely to a single-stranded non-monotonous 5'-UUAG-3' sequence and also weaker to the single-stranded 5'-TTAGGG-3' telomeric DNA repeat. Binds RNA oligonucleotides with 5'-UUAGGG-3' repeats more tightly than the telomeric single-stranded DNA 5'-TTAGGG-3' repeats. Binding of RRM1 to DNA inhibits the formation of DNA quadruplex structure which may play a role in telomere elongation. May be involved in translationally coupled mRNA turnover. Implicated with other RNA-binding proteins in the cytoplasmic deadenylation/translational and decay interplay of the FOS mRNA mediated by the major coding-region determinant of instability (mCRD) domain.	
HNRNPM	Heterogeneous nuclear ribonucleoprotein M	Pre-mRNA binding protein in vivo, binds avidly to poly(G) and poly(U) RNA homopolymers in vitro. Involved in splicing. Acts as a receptor for carcinoembryonic antigen in Kupffer cells, may initiate a series of signaling events leading to	**v**

		tyrosine phosphorylation of proteins and induction of IL-1 alpha, IL-6, IL-10 and tumor necrosis factor alpha cytokines.	
MSI2	RNA-binding protein Musashi homolog 2	RNA binding protein that regulates the expression of target mRNAs at the translation level. May play a role in the proliferation and maintenance of stem cells in the central nervous system	**v**
PABPC1	Polyadenylate-binding protein 1	Binds the poly(A) tail of mRNA. May be involved in cytoplasmic regulatory processes of mRNA metabolism such as pre-mRNA splicing. Its function in translational initiation regulation can either be enhanced by PAIP1 or repressed by PAIP2. Can probably bind to cytoplasmic RNA sequences other than poly(A) in vivo. May be involved in translationally coupled mRNA turnover. Implicated with other RNA-binding proteins in the cytoplasmic deadenylation/translational and decay interplay of the FOS mRNA mediated by the major coding-region determinant of instability (mCRD) domain.	**x**
PPARGC1B	Peroxisome proliferator-activated receptor gamma coactivator 1-beta	Plays a role of stimulator of transcription factors and nuclear receptors activities. Activates transcritional activity of estrogen receptor alpha,	**v**

		nuclear respiratory factor 1 (NRF1) and glucocorticoid receptor in the presence of glucocorticoids. May play a role in constitutive non-adrenergic-mediated mitochondrial biogenesis as suggested by increased basal oxygen consumption and mitochondrial number when overexpressed. May be involved in fat oxidation and non-oxidative glucose metabolism and in the regulation of energy expenditure.	
PTBLP	Polypyrimidine tract-binding protein 2	RNA-binding protein which binds to intronic polypyrimidine tracts and mediates negative regulation of exons splicing. May antagonize in a tissue-specific manner the ability of NOVA1 to activate exon selection. Beside its function in pre-mRNA splicing, plays also a role in the regulation of translation. Isoform 5 has a reduced affinity for RNA.	**v**
RBM15	Putative RNA-binding protein 15	May be implicated in HOX gene regulation.	**x**
RBM16	Putative RNA-binding protein 16	May play a role in mRNA processing.	**x**
RBM26	RNA-binding protein 26		**x**
RBMS1	RNA-binding motif, single-stranded-interacting protein 1	Single-stranded DNA binding protein that interacts with the region upstream of the MYC gene. Binds specifically to the DNA sequence motif 5'-[AT]CT[AT][AT]T-3'. Probably has a role in DNA replication.	**v**

RBMS3	RNA-binding motif, single-stranded-interacting protein 3	Binds poly(A) and poly(U) oligoribonucleotides.	**v**
SFRS11	Splicing factor, arginine/serine-rich 11	May function in pre-mRNA splicing.	**v**
SFRS15	Splicing factor, arginine/serine-rich 15	May act to physically and functionally link transcription and pre-mRNA processing	**x**
SYNCRIP	synaptotagmin binding, cytoplasmic RNA interacting protein, hnRNPQ	Heterogenous nuclear ribonucleoprotein (hnRNP) implicated in mRNA processing mechanisms.Component of the CRD-mediated complex that promotes MYC mRNA stability.	**v**
THOC4	THO complex subunit 4	CRD-mediated complex that promotes MYC mRNA stability.	**v**

RNA Biology 9:10, 1–9; October 2012; © 2012 Landes Bioscience

Tuning the engine

An introduction to resources on post-transcriptional regulation of gene expression

Erik Dassi and Alessandro Quattrone

Laboratory of Translational Genomics; Centre for Integrative Biology; University of Trento; Trento, Italy

Keywords: post-transcriptional regulation, translation, UTR, database, tool, RBP, ncRNA, miRNA, cis-element, trans-factor

Abbreviations: PTR, post-transcriptional regulation; UTR, untranslated region of mRNA; RBP, RNA-binding protein; RRM, RNA-recognition motif; KH, K-homology domain; dsRBD, double strand RNA binding domain; ncRNA, non-coding RNA; lncRNA, long non-coding RNA; miRNA, micro-RNA; siRNA, small interfering RNA; piRNA, piwi-interacting RNA; snoRNA, small nucleolar RNA; snRNA, small nuclear RNA; ARE, AU-rich element; IRE, iron response element; IRES, internal ribosome entry site; SECIS, seleno-cysteine insertion sequence; SIRF, short interspersed repeats fragment; CLIP, cross-linking immunoPrecipitation; PAR-CLIP, photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation; iCLIP, individual-nucleotide resolution UV cross-linking and immunoprecipitation

In the last years post-transcriptional regulation (PTR) of gene expression has been increasingly recognized to be a powerful and general determinant of the quantitative changes in proteomes, and therefore a driving force for cell phenotypes. By means of networks of trans-factors on one hand, and cis-elements found primarily in untranslated regions (UTRs) of mRNA on the other hand, mRNA availability to translation and translation rates are tightly controlled and can be rapidly tuned according to the changing state of the cell. A number of dedicated resources and tools, including databases and predictive algorithms, have been proposed as bioinformatics aids for the study of this fundamental layer of gene expression regulation. Their use, however, is rendered difficult by heterogeneity and fragmentation.

This review aims to locate these resources in their proper space, classifying them according to their goals, limitations and integration capabilities and, in the end, to provide the user with an initial toolbox for the bioinformatic analysis of post-transcriptional regulation of gene expression. The accompanying website, available at www.ptrguide.org, lists all resources, provides summary and features for each one and will be regularly updated in the future.

Introduction

Post-transcriptional regulation (PTR) of gene expression is the process responsible for modulating mRNA levels and the related amount of protein. Initially thought to have a limited impact on cell phenotype, it has become increasingly recognized as a powerful and general determinant of the quantitative changes in proteomes.[1] Untranslated regions of mRNAs (UTRs) are the fundamental mediators of this process, because they bear sequence and structure patterns preferentially bound by regulators which influence nuclear export, localization, stability of mRNAs and ultimately their translation rates,[2] as well as capping, alternative splicing and polyadenylation of the transcribed pre-mRNA. One of the most important classes of post-transcriptional regulatory factors are the RNA-binding proteins (RBPs), whose human genome complement is at least 800 genes[3,4,5] and which are characterized by the presence of different functional domains[6] among which the most represented are, according to the latest release of Ensembl (Ensembl 68), the zinc-finger C2H2 domain (787 genes), the RNA-recognition motif (RRM, 233 genes), the sterile α motif (SAM, 93 genes) and the K-homology domain (KH, 38 genes). RBPs bind to the 5'UTR of a transcript often to modulate translation initiation, and to its 3'UTR usually to influence its stability or translatability;[3] but they have also being well characterized for modulating splicing of the pre-mRNA, mRNA export and mRNA localization in the cytoplasm.[7] Another major group of actors in PTR are non-coding RNAs (ncRNAs). Among them are various classes of long ncRNAs (lncRNAs), the intensively studied micro-RNAs (miRNAs), and then siRNAs (small-interfering RNAs), piRNAs (piwi-interacting RNAs), snoRNAs (small nucleolar RNAs), snRNAs (small nuclear RNAs), and several other types.[8] miRNAs (around 1,500 are currently annotated in the human genome) bind to the 3'UTR of a transcript by means of short regions of perfect sequence complementation (which leads to increased transcript degradation) or with some mismatches (which promotes translational repression and increased degradation).[9] Both RBPs and ncRNAs bind mRNAs in the so-called cis-elements, found primarily in 5' and 3' UTRs. These elements can be represented as recurring RNA sequences or secondary structures shared by a number of transcripts and defined by a pattern, to which the trans factors bind to exert a control over the mRNA. A well-known example of cis-regulatory

Correspondence to: Erik Dassi and Alessandro Quattrone; Email: dassi@science.unitn.it and alessandro.quattrone@unitn.it
Submitted: 08/03/12; Revised: 08/29/12; Accepted: 08/31/12
http://dx.doi.org/10.4161/rna.22035

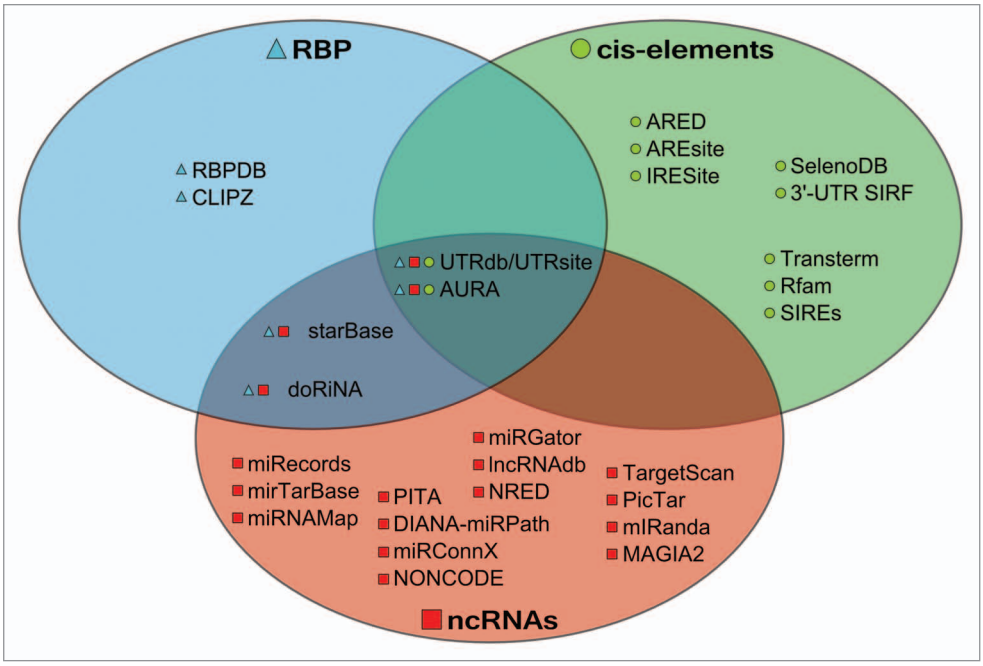


Figure 1. Venn diagram showing the classification of the analyzed resources according to their biological focus. Symbols next to the resource name correspond to each set (triangles for RBPs, squares for ncRNAs and circles for cis-elements) and further highlight the presence of a limited number of integrative tools, with most of the resources being confined to only one kind of regulatory element.

elements are the AU-Rich Elements (ARE),[10] motifs rich in Us with some interspersed As or Gs shared by several thousand 3'UTRs and bound by a large number of RBPs of which at least 23 are known.[10] Another well characterized class of UTR cis-elements are the Iron Response Elements (IREs), which help in coordinating cellular iron homeostasis at the translational level.[11]

The last years have seen a rapid increase in resources dedicated to the analysis of these factors and elements to unravel associated mechanisms of gene expression regulation. Available databases are focused mainly on UTRs annotation,[12,13] RBP-target interactions,[14,15] ncRNAs,[16,17,18,19,20,21,22,23,24] of which miRNA-target interactions are the greater part,[16,17,18,19,20] with a limited number of resources focusing on lncRNAs,[22,23] and cis-elements.[25,26,27,28,29,30] Furthermore, a small number of resources integrating different data types is available.[12,13,31] Predictive tools also exist, in particular for cis-elements pattern-based search[32,33] and ncRNAs.[34,35,36,37,38,39] This review will first introduce the foremost available resources, excluding those related to splicing and the no longer updated ones, and will catalog them in three categories: *RBP*-oriented, *ncRNA*-oriented and *cis-element*-oriented, with a number of resources falling in more than one category (**Fig. 1**). We will highlight also further features of these resources, as integrating different data types or being predictive. We will then proceed to illustrate a tentative pipeline combining several of these tools to enable the discovery of regulatory mechanisms. Eventually, we will present a biological use-case in which these resources are employed to identify potential regulatory circuits. We conclude with a short discussion on the future directions to be pursued in order to enhance the usefulness and completeness of this toolbox for the analysis of circuits of post-transcriptional regulation of gene expression.

We provide an accompanying website to this review, available at www.ptrguide.org. The website lists all the cited resources, providing a summary and details on features and availability of the resources; it will be regularly updated with new resources and updates of existing ones, with the aim of providing a one-stop catalog for available PTR mining tools (**Table 1**).

Resources

Databases and tools can be classified according to their main focus and purpose. They can be **RBP-oriented** when they deal with RBPs and the effect these exert on mRNAs, **ncRNA-oriented** when they analyze regulation by the various families of these RNAs (as miRNAs and lncRNAs); and **cis-oriented** whenever a cis element is annotated and characterized in its occurrences throughout expressed exons.

RBP-Oriented

Despite the increasingly recognized importance of these factors in PTR of gene expression, only five resources are available which focus on RBPs, completely or even only partially. *RBPDB*[12] and *CLIPZ*[15] are built exclusively around RBPs: RBPDB offers a literature-curated collection of RBP binding sites and motifs, searchable by species or by protein domain and including logos or position-weight matrices where available. It allows the user to input sequences that can be searched for the presence of binding sites of the included RBPs. It also has predictive capabilities, albeit limited: indeed, it allows the user to match an input sequence vs. position weighted matrices (PWMs) contained in the database to identify possible RBP binding sites. *CLIPZ* is instead an analysis environment of RNA binding sequences by RBPs derived from the high-throughput techniques for cross-linking based mRNA footprinting, including CLIP,[40] PAR-CLIP[41] and iCLIP[42] followed by RNA-seq. It contains analytical tools to let the user load and analyze the own CLIP-seq data, identify binding sites and annotate them on the reference genome. *UTRdb/UTRsite*,[12] *AURA*[13] and *doRiNA*[31] hold RBP-related data as the two resources described above, but they differ in still keeping a broader perspective on post-transcriptional regulation. *UTRdb/UTRsite* contains data about UTRs in a number of species, annotating them with a specific subset of RBP binding sites, cis-regulatory sequence patterns, miRNAs and SNP data. It provides UTR sequence data along with conserved elements, visually arranged in a linear fashion. *AURA* annotates human UTRs with RBPs, ncRNAs,

Table 1. PTR resources presented in the review

Name	Ref	Last update	Batch mode	Data download	Organisms	Link
ARED	25	Mar 2011	v	x	HSA, MMU	http://brp.kfshrc.edu.sa/ARED/
AREsite	26	Nov 2010	x	v	HSA, MMU	http://rna.tbi.univie.ac.at/AREsite
AURA	13	Nov 2011	x	v	HSA	http://aura.science.unitn.it/
CLIPZ	15	Jan 2011	v	v	HSA, MMU, CEL	http://www.clipz.unibas.ch/
DIANA-miRPath	38	Mar 2012	v	x	HSA, MMU	http://www.microrna.gr/miRPathv2
doRiNa	31	May 2012	v	v	HSA, MMU, DME, CEL	http://dorina.mdc-berlin.de
IRESite	27	Apr 2011	x	x	HSA, MMU, RNO, DME, SCE and 4 more	http://iresite.org/
lncRNAdb	22	Jul 2011	x	v	HSA, MMU, DME, CEL, ATH, XLA, SCE and 53 more	http://lncrnadb.com/
miRanda	36	Nov 2010	v	V	HSA, MMU, RNO, DME, CEL	http://www.microrna.org/microrna/home.do
MAGIA2	24	Apr 2012	v	X	HSA, MMU, RNO, DME	http://gencomp.bio.unipd.it/magia2
miRConnX	39	Jul 2011	v	v	HSA, MMU	http://mirconnx.csb.pitt.edu/
miRecords	18	Nov 2010	x	v	HSA, MMU, RNO, DME, CEL, GGA, DRE, OAR, CFA	http://mirecords.biolead.org/
miRGator	23	Jan 2011	v	x	HSA	http://mirgator.kobic.re.kr
miRNAMap	19	Jul 2007	x	v	HSA, MMU, RNO, DME, CEL, XTR and 4 more	http://mirnamap.mbc.nctu.edu.tw/
miRTarBase	17	Oct 2011	x	v	HSA, MMU, RNO, DME, CEL, ATH, XLA and 7 more	http://mirtarbase.mbc.nctu.edu.tw/
NONCODE	21	Jan 2012	x	v	HSA, MMU, DME, CEL, ATH, XLA and 1233 more	http://www.noncode.org
NRED	24	Sep 2008	x	v	HSA, MMU	http://jsm-research.imb.uq.edu.au/nred
PicTar	35	Mar 2007	x	v	HSA, MMU, DME, CEL	http://pictar.mdc-berlin.de/
PITA	37	Aug 2008	v	v	HSA, MMU, DME, CEL	http://genie.weizmann.ac.il/pubs/mir07/mir07_data.html
RBPDB	14	Jan 2011	x	v	HSA, MMU, DME, CEL	http://rbpdb.ccbr.utoronto.ca/
Rfam	30	Jun 2011	x	v	HSA, MMU, DME, CEL, ATH, SCE and 3104 more	http://rfam.sanger.ac.uk/
SelenoDB	28	Sep 2007	x	v	HSA, MMU, DME, CEL, SCE and 3 more	http://www.selenodb.org/
SIREs	33	Jan 2010	v	x	any	http://ccbg.imppc.org/sires/
starBase	16	Sep 2011	x	v	HSA, MMU, CEL, ATH, OSA, VME	http://starbase.sysu.edu.cn/
TargetScan	34	Mar 2012	x	v	HSA, MMU, CEL, DRE	http://www.targetscan.org/
Transterm	32	Oct 2011	v	x	any	http://mrna.otago.ac.nz/
UTRdb/UTRsite	12	Oct 2009	v	v	HSA, MMU, DME, CEL, ATH, XLA and 73 more	http://utrdb.ba.itb.cnr.it/

The table shows the list of databases and tools presented in the review: for each of them we report the last update (or publication date when the former is not available) along with the reference number in the manuscript, the resource website address, the organisms for which the resource provides data (listed by their three-letters code), the possibility to do a batch analysis (searching for more than one gene/element at a time) and to download the whole database.

cis-elements, phylogenetic conservation and sequence variation obtained from 10 different databases, and includes literature curation. This database has its strength in committing to experimentally inferred interactions; it allows displaying UTRs in a genome browser like view, with calculated UTR secondary structures, and experimental mRNA and protein levels; visualization of joint gene expression data of targets and associated regulators can also aid inference of regulatory events. *doRiNA*[31] integrates RBP and miRNA binding sites, by including only high-throughput assays-derived data sets for RBPs and a set of predictions for miRNAs. It exploits the UCSC database genome viewer annotated with binding sites, offering various query possibilities: by specifying a specific list of RBPs and miRNA one can obtain subsets of UTRs regulated by common groups of RBPs and miRNAs,

thus guiding the discovery of novel PTR networks. By including high-throughput techniques-derived data, AURA and doRiNA provide a great wealth of information on RBP binding sites: the user need however to be aware that, as these data are available for only a limited number of RBPs, the resulting network will be biased toward these factors, providing a potentially incomplete or misleading picture of the PTR phenomena at study.

ncRNA-Oriented

A wealth of resources focused on noncoding RNAs are available: these can be differentiated by the data they hold, either experimentally validated or predicted. *miRecords*,[17] *mirTarBase*[18] and *miRNAMap*[19] aim to collect miRNAs annotations and miRNA-target interactions. *miRecords* and *miRNAMap* contain both experimentally validated and predicted data (which are obtained by merging the output of 11 prediction algorithms for miRecords and 3 for miRNAmap), while *miRTarBase* includes only experimentally validated data. All three databases link out to various miRNA reference annotation sources such as *miRBase*, with *mirTarBase* and *miRNAMap* also displaying pre-miR secondary structure and miRNA expression levels in various normal and diseased tissues. *miRGator*[20] also focuses on this class of ncRNAs, trying to give a broader overlook on the miRNA functional role by means of several auxiliary annotations: it integrates predicted miRNA-mRNA interactions, paired miRNA-mRNA expression profiles and miRNA disease signatures. Through their association analysis feature, exploiting the various expression profiles contained in the database, a miRNA can be associated to a particular tissue, a disease state or to anti-coexpressed genes. User expression profiles cannot be uploaded, although miRNA sets can be tested for enrichments through the miR set analysis tool. *starBase*[16] is quite unique in its kind as it is dedicated to the annotation of experimentally validated Argonaute binding sites, derived from CLIP-seq and Degradome-seq[43] assays: these sites, hallmark of miRNA-mediated regulation, are then merged with the output of various miRNA-target prediction tools in order to infer several thousands of miRNA-mRNA relationships. The experimental data-based tool is MAGIA2,[24] an analysis platform allowing to upload your own miRNA and mRNA expression data sets, combine them with transcription factor binding sites and miRNA target predictions, and eventually infer regulatory networks from the integrated data. A wealth of tools is instead available to computationally predict miRNA targets: among these we consider *TargetScan*,[34] *PicTar*,[35] *miRanda*,[36] *PITA*,[37] *DIANA-mirPath*[38] and *miRConnX*.[39] *TargetScan* predicts interaction by requiring seed match conservation in five species and by filtering false positives through comparable abundance hexamers control; along the same line, based on sequence information, *DIANA-mirPath* combines predictions with experimentally verified targets, employing artificial neural networks or sequence-based 38-bases sliding windows to identify true positive miRNA binding sites in human and model organisms 3'UTRs. Users can also exploit data on SNPs in miRNA binding sites and related pathways information. *PicTar* instead identifies seed matches by keeping into account free energy of the miRNA-mRNA hybrid and by using a combination of scores to evaluate match goodness; *miRanda* also employs hybrid free energy but also phylogenetic conservation and seed matching, complemented with non-uniform distribution of target sites and 5'–3' asymmetry constraints. *PITA* is the last of the tools keeping secondary structure free energy into account: it scores the sequence seed matches according to the gain in free energy obtained when the miRNA binds to the target, compared with the energy needed to open the structure of the target in that portion and thus promoting binding. *miRConnX* takes advantage of a pre-computed network of predicted mRNA-miRNA, transcription factor-gene and transcription factor-miRNA relationships, supplemented with literature data, combined with a dynamic networks based on user-provided gene expression data (both mRNA and miRNA). The user data network is built by various correlation measures (following the guilt-by-association principle) and integrated with the pre-computed one through a weighted sum integration function. The resulting integrated network can then be browsed, exported or analyzed in several ways, such as searching for network motifs. Users need to keep in mind that the data set size required for such an approach to produce meaningful results is quite high (in the order of tens, if not hundreds, of samples). *NONCODE*,[21] *lncRNAdb*[22] and *NRED*[23] are reference databases for ncRNAs and related expression information. Long-noncoding RNAs have been mostly regarded as chromatin-associated, and thus transcription-related, factors. However, some evidence of their involvement in PTR of gene expression is emerging (for examples, see refs. 46 and 47), and we therefore include them in our review. *NONCODE* offers a wealth of expression and functional data concerning all kinds of ncRNAs: data are predominantly experimental, and the database includes a novel classification system based on cellular function. *lncRNAdb and NRED* are connected and aim, on one side to comprehensively list experimentally inferred lncRNAs described to have biological function in eukaryotes, and on the other side to provide gene expression information for thousands of these lncRNAs in human and mouse. *lncRNAdb* includes sequence and structure information along with links to the UCSC genome browser,[44] literature sources and data from the *NRED* database: these are obtained primarily by microarray or in situ hybridization analysis and are complemented by auxiliary annotations, such as phylogenetic conservation and secondary structure evidence.

Cis-Oriented

Most of the resources in this category are focused on one specific type of cis-elements; still, among them four databases are more general and aim at considering or predicting a great deal of these: *Transterm*[32] containing various patterns of cis-regulatory elements in mRNA UTRs: input sequences can be selected among the sets provided on the website or provided by the user: all instances matching the patterns or just the ones of the user-selected pattern will be reported. The *UTRscan* feature of *UTRdb/UTRsite*[12] works in the same way, predicting instances of cis-elements. *AURA*[13] contains instead annotated instances of elements like AREs (predicted) and mRNA-editing data (experimentally validated). The last general resource, *Rfam*,[30] annotates

and lists, organizing them in clans and families, currently known cis-elements found in 5'UTRs and 3'UTRs. On a wider perspective, this database also aims at cataloging all ncRNAs by means of sequence alignment and statistical profile models. *ARED*[24] and *AREsite*[25] are two databases devoted to AU-rich elements (AREs), a widely studied cis-element type found in 3' UTRs. *ARED* is built by searching in GenBank mRNA and EST records for a single 13-base pairs pattern, and the results are then classified according to ARE classes.[10] Every ARE-containing mRNA is then linked to the related UniGene and Gene Ontology annotation. *AREsite* works along the same line, but allows the user to screen UTRs for eight different ARE patterns, corresponding to types extracted from the literature. Along with ARE localization on the UTR, it displays information about the structural context of the motif and its level of phylogenetic conservation. *IRESite*[27] contains experimentally validated Internal Ribosome Entry Sites (IRESs) found in 5'UTRs. These are listed with related gene and mRNA details; furthermore, the user can input its own sequence or secondary structure to search for matches with all IRESs contained in the database. *SIREs*[33] is instead a web server for the prediction of IREs:[11] it takes into account both sequence and secondary structure constraints known to characterize this kind of elements. Structure analysis, folding data and quality indications are provided for each prediction output. *SelenoDB*[28] aims at annotating all selenoproteins and SECIS (SEleno Cysteine Insertion Sequence) elements[45] found in the 3'UTRs of the mRNAs coding for these proteins. These cis-elements are predicted in selenoprotein 3'UTRs by means of a computational tool, and annotated with sequence, position and related gene data. Finally, *3'-UTR SIRF*,[29] lists all computationally predicted short-interspersed repeats in 3'UTRs. Motifs can be searched alone or in combination to identify genes whose 3'UTRs bear these putatively co-occurring repeats.

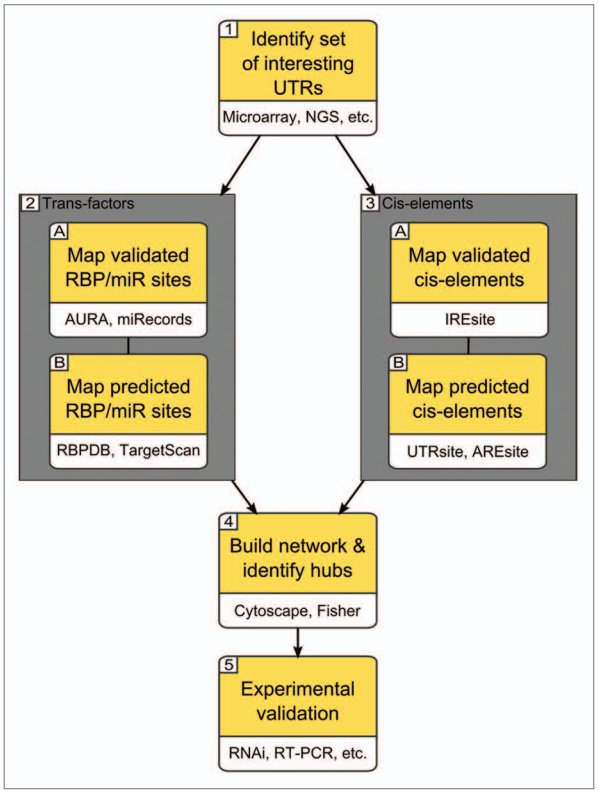


Figure 2. A possible discovery pipeline for post-transcriptional regulatory mechanisms. The workflow starts by the selection of interesting UTRs: these may, for instance, come from high-throughput experiments done by the microarrays or next generation sequencing technologies. The pipeline then proceeds by searching for both experimentally validated and computationally predicted trans-factors binding sites and cis-elements over these UTRs. The resulting interactions are then collected into a network: important nodes are identified by enrichment tests such as the Fisher test. Interesting leads are eventually subjected to experimental validation by various methods of targeted gene expression perturbation, as RNA interference.

Designing a Discovery Pipeline

Choosing which resources to use among the ones presented here may be far from trivial, especially for non-computational biologists. We thus propose a pipeline to empower the discovery of potential post-transcriptional regulatory mechanisms by exploiting some of the available tools. This is, of course, just one of the many possible combinations of instruments that can be used to reach this goal, and is offered as an example to illustrate the concepts behind an effective discovery workflow. **Figure 2** reports the steps composing our pipeline. It starts with the identification of a set of interesting genes or mRNAs (1) and related UTRs: in a common setting these may represent differentially expressed genes obtained through a case-control microarray or RNA-seq experiment, although the UTR list can come in whatever other way. In the next step the workflow splits in two parallel branches: on one side, UTRs are searched for known binding sites of trans-factors (2). These are both experimentally validated (A) for RBPs and miRNAs coming from *AURA*[13] and *miRecords*[18] respectively, and computationally predicted (B) by applying *RBPDB*[12] and *TargetScan*.[34] In the other branch (3) we scan our UTRs in order to identify cis-elements that may be contained therein. Again, we employ both experimentally validated data (A), coming from *IRESite*[27] and possibly other sources, and computationally predicted annotations, obtained through *UTRsite*,[12] *AREsite*,[26] *SIREs*[33] and others. Once data collection is completed, we can move to the next step (4): building a network including our initial genes and all the factors identified until this point as regulators. Such construction can be done by means of software like *Cytoscape*[48] and can be automated through a scripting language such as Python. Visual inspection of the resulting network will highlight hub nodes, that is, highly controlled mRNAs or widespread regulators of the mRNAs of interest. More rigorous statistical analysis can be performed on the network nodes. As regulatory factors like RBPs may post-transcriptionally control hundreds of different mRNAs, it is worth looking for enrichment of a potential regulator in the set of mRNAs under analysis: this may be done by applying a Fisher test, as it is commonly done for the over-representation of ontology terms in gene lists.[49] This test will be associated to a p-value testifying for the hypothesized enrichment. In order to discriminate between general factors and potential aspecific interactions, it can be useful to also generate a

control network to compare with the one under study: to do so, one can select a comparable number of UTRs at random (from the data produced by the same experiment) and reapply the pipeline to this new data set. The two resulting networks can then be compared, and factors present or enriched in both of these be excluded from further analysis: these may indeed represent widespread regulatory mechanisms, most probably not responsible for the differential expression of this group of genes and difficult to target. The last step of our pipeline leaves the in silico world and goes back to the bench: in order to understand and validate the regulatory mechanism we have hypothesized and prioritized, a classical array of methods are available. In case of cell studies, gene silencing through RNA interference, gene overexpression through transfection or viral infection, and target gene expression probing through real-time PCR or high throughput mRNA quantification methods are the most common choices. This will eventually provide data concerning the effect of the depletion or enrichment of our potential regulator(s) over target genes, and, on a wider perspective, over the network we are characterizing.

A Case

We now proceed to apply the proposed pipeline to a set of differentially expressed genes, in order to provide a practical example of how it could work. We downloaded the GSE11324 data set[51] from *GEO*[52]: in this data set, the transcriptome of MCF7 cells is profiled under estrogen stimulation at several time points (0 to 12hrs). By means of *GEO2R*,[52] we computed differentially expressed genes between 0hr and 12hr of estrogen stimulation. We then selected the 50 mRNAs with the highest absolute fold change, corresponding to 43 genes (obviously this is an arbitrary choice, we presume that the highest fold changes indicate the most relevant biological changes, even this cannot be necessarily the case, and there are other ways of prioritizing the genes). The gene list (a) and a summary of the most prominent findings (b-c) are shown in **Figure 3**. From (b), reporting only trans-factors and cis-elements shared by at least 10% of the mRNAs (in which the size of the circle is proportional to the percentage of controlled mRNAs), it is evident that the relevant genes share AU-rich elements and AGO binding sites: in particular, AREs are predicted to be present in the 3'UTR of 86% of the mRNAs (enrichment p-value = 1.3E-10), while AGO binding sites in 83% of these(enrichment p-value = 4.9E-10). Other potentially involved factors are the IGF-binding proteins IGFBP1/2/3, having experimentally determined binding sites in 80% of the mRNAs (enrichment p-value = 0.48); PUM2, whose binding elements are found for 44% of the mRNAs (enrichment p-value = 1.37E-08); along with predicted IRES (Internal Ribosome Entry Site, 88% of the mRNAs), MBEs (Musashi binding elements, 67% of the mRNAs) and K-Boxes, GY-Box and PAS (Poly-adenylation signal) at a lower frequency. Two families of miRNAs (mir-15 and mir-16) are predicted by *Targetscan*[34] to control at least ten genes of our set. In order to understand if these factors are specific to our DEGs network, we randomly selected another 50 mRNAs from the data set and reapplied the pipeline to these (network not shown): while AGO and IGFBP1/2/3 sites are again found in many UTRs (90% and 75% respectively), leading us to consider them non-relevant findings, ARE and PUM2 sites are found in lower proportions (54% and 28% of the UTRs); predicted IRES involve only 26% of the randomly selected mRNA, while MBEs are found in the same proportion as in the top DEGs (67%). Among microRNAs, mir-15 and mir-16 are not predicted to control many of our mRNAs: miR-590 and miR-30 seem to control instead 15 or more genes of our random set, with miR-23 predicted for 13 genes. Other elements are found with low frequencies (less than 10% of the mRNAs) and are thus not considered as relevant. We can thus confirm some of the involved factors as specific for our DEGs network, avoiding to focus on possibly general regulatory mechanisms. These findings are obviously biased by the still low number of available transcriptome-wide CLIP experiments, which provide much more data than literature annotations, and therefore emerge in the results. Enrichment p-values are computed for experimentally validated data by means of a Fisher test, as previously stated. The resulting post-transcriptional network of RBP, miRNAs and cis-elements, shown in (c) and built via a Python script into the *Cytoscape*[48] platform, offers a complex landscape for further validation.

Future Directions

This review has highlighted the main tools of the steadily increasing number of resources available on networks of regulation at the post-transcriptional level, as one of the indicators of the growing interest in the topic. In particular, a wealth of databases and algorithms is offered focusing on miRNA-mRNA interactions, both for experimentally validated data and computational prediction, mirroring the exceptional interest raised by these controllers of gene expression in the research community. A more limited variety of resources dedicated to RBPs, cis-elements and others ncRNAs is also available. Only three tools, among the ones we analyzed, attempt to integrate different component of these networks: RBPs and miRNA binding sites only,[31] or including also predicted RNA secondary structures and cis-elements.[12,13] While these resources considerably ease the task of hypothesizing the existence of new networks, they still contain just a fraction of the data really available in the literature, and obviously are affected by the small number of trans-factor experimentally tested in a high throughput way with respect to the annotated ones. Moreover, the majority of the tools still does not allow online batch or programmatic analysis, forcing the user willing to work on a medium-to-big sized data set to download and replicate the database locally, and write ad hoc scripts. Integrating these tools into an automatic or semi-automatic pipeline is thus time consuming, if not impossible. Future developments should go toward this direction, providing a one-stop, truly integrated, comprehensive and multi-faceted PTR analysis toolset. Availability of such a tool will consistently empower the mapping of post-transcriptional and specifically translational networks, reaching the level of service already offered by resources focusing on the analysis of transcriptional regulation. Nevertheless, this will require a substantial effort of implementation and update, which could

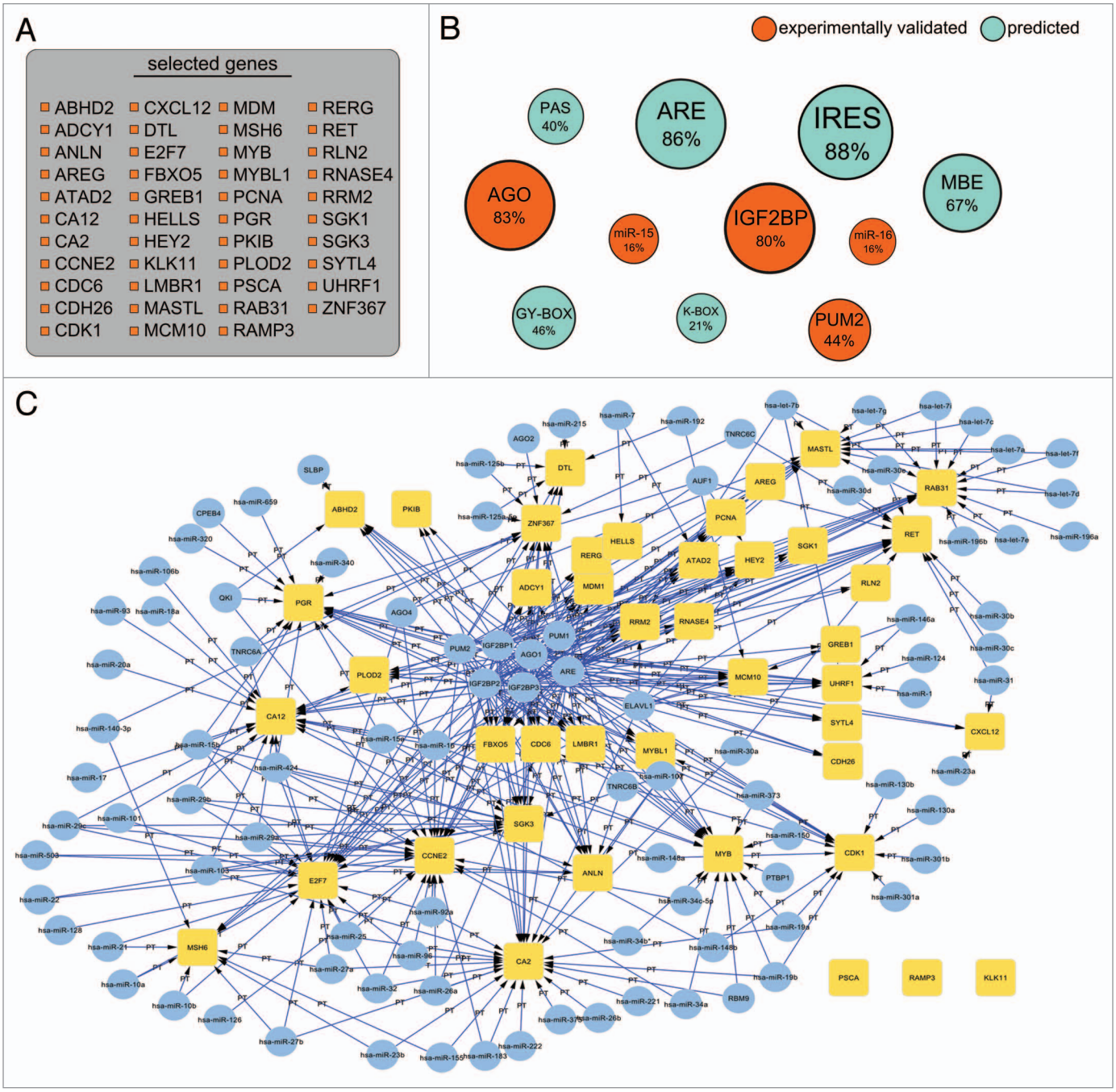


Figure 3. Selected genes and results obtained by the application of the proposed pipeline. (A) Is the list of genes selected for the case example. (B) Shows the post-transcriptional interactions prioritized through the pipeline: orange circles represent experimentally validated interactions while cyan circles represent predicted interactions. Size of the circles is proportional to the fraction of genes controlled by the element which name labels the circle (RBP, ncRNA or cis-element). Percentage of controlled genes is shown under the factor name. (C) Displays the post-transcriptional regulatory network composed of RBPs, miRNAs and cis-elements obtained by the application of the pipeline. Yellow squares represent our genes of interest, while light blue circles are the different factors controlling these genes. Oriented arrows pointing toward a gene represent an observed regulatory event (binding site or cis-element).

be eased by coordination between the available resources and integration with major genome databases such as the UCSC Genome Browser[44] and Ensembl.[50] Furthermore, we think that at least two additional features are currently missing but definitely needed. First, a systematic literature-derived annotation of the molecular downstream and phenotypic effects of a given interaction would provide more grounded clues, orienting the experimental validation. Second, more tailored statistical methods for enrichment of cis-elements or trans-factor, as those for ontology terms enrichment,[49] would be beneficial to avoid generation of a large number of false positives as an effect of the high multiplicity of action of several studied trans-factors.

References

1. Moore MJ. From birth to death: the complex lives of eukaryotic mRNAs. Science 2005; 309:1514-8; PMID:16141059; http://dx.doi.org/10.1126/science.1111443.
2. Mignone F, Gissi C, Liuni S, Pesole G. Untranslated regions of mRNAs. Genome Biol 2002; 3: reviews0004.1; PMID:11897027; http://dx.doi.org/10.1186/gb-2002-3-3-reviews0004.
3. Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett 2008; 582:1977-86; PMID:18342629; http://dx.doi.org/10.1016/j.febslet.2008.03.004.
4. Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, et al. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell 2012; 149:1393-406; PMID:22658674; http://dx.doi.org/10.1016/j.cell.2012.04.031.
5. Baltz AG, Munschauer M, Schwanhäusser B, Vasile A, Murakawa Y, Schueler M, et al. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Mol Cell 2012; 46:674-90; PMID:22681889; http://dx.doi.org/10.1016/j.molcel.2012.05.021.
6. Lunde BM, Moore C, Varani G. RNA-binding proteins: modular design for efficient function. Nat Rev Mol Cell Biol 2007; 8:479-90; PMID:17473849; http://dx.doi.org/10.1038/nrm2178.
7. Andreassi C, Riccio A. To localize or not to localize: mRNA fate is in 3'UTR ends. Trends Cell Biol 2009; 19:465-74; PMID:19716303; http://dx.doi.org/10.1016/j.tcb.2009.06.001.
8. Costa FF. Non-coding RNAs: Meet thy masters. Bioessays 2010; 32:599-608; PMID:20544733; http://dx.doi.org/10.1002/bies.200900112.
9. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet 2008; 9:102-14; PMID:18197166; http://dx.doi.org/10.1038/nrg2290.
10. Barreau C, Paillard L, Osborne HB. AU-rich elements and associated factors: are there unifying principles? Nucleic Acids Res 2005; 33:7138-50; PMID:16391004; http://dx.doi.org/10.1093/nar/gki1012.
11. Wang J, Pantopoulos K. Regulation of cellular iron metabolism. Biochem J 2011; 434:365-81; PMID:21348856; http://dx.doi.org/10.1042/BJ20101825.
12. Grillo G, Turi A, Licciulli F, Mignone F, Liuni S, Banfi S, et al. UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res 2010; 38(Database issue):D75-80; PMID:19880380; http://dx.doi.org/10.1093/nar/gkp902.
13. Dassi E, Malossini A, Re A, Mazza T, Tebaldi T, Caputi L, et al. AURA: Atlas of UTR Regulatory Activity. Bioinformatics 2011; http://dx.doi.org/10.1093/bioinformatics/btr608
14. Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. RBPDB: a database of RNA-binding specificities. Nucleic Acids Res 2011; 39(suppl 1):D301-8; PMID:21036867; http://dx.doi.org/10.1093/nar/gkq1069.
15. Khorshid M, Rodak C, Zavolan M. CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. Nucleic Acids Res 2011; 39(suppl 1):D245-52; PMID:21087992; http://dx.doi.org/10.1093/nar/gkq940.
16. Yang JH, Li JH, Shao P, Zhou H, Chen YQ, Qu LH. starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. Nucleic Acids Res 2011; 39(Database issue):D202-9; PMID:21037263; http://dx.doi.org/10.1093/nar/gkq1056.
17. Hsu SD, Lin FM, Wu WY, Liang C, Huang WC, Chan WL, et al. miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucleic Acids Res 2011; 39(Database issue):D163-9; PMID:21071411; http://dx.doi.org/10.1093/nar/gkq1107.
18. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res 2009; 37(Database issue):D105-10; PMID:18996891; http://dx.doi.org/10.1093/nar/gkn851.
19. Hsu SD, Chu CH, Tsou AP, Chen SJ, Chen HC, Hsu PW, et al. miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes. Nucleic Acids Res 2008; 36(Database issue):D165-9; PMID:18029362; http://dx.doi.org/10.1093/nar/gkm1012.
20. Cho S, Jun Y, Lee S, Choi HS, Jung S, Jang Y, et al. miRGator v2.0: an integrated system for functional investigation of microRNAs. Nucleic Acids Res 2011; 39(Database issue):D158-62; PMID:21062822; http://dx.doi.org/10.1093/nar/gkq1094.
21. Bu D, Yu K, Sun S, Xie C, Skogerbø G, Miao R, et al. NONCODE v3.0: integrative annotation of long noncoding RNAs. Nucleic Acids Res 2012; 40(Database issue):D210-5; PMID:22135294; http://dx.doi.org/10.1093/nar/gkr1175.
22. Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res 2011; 39(Database issue):D146-51; PMID:21112873; http://dx.doi.org/10.1093/nar/gkq1138.
23. Dinger ME, Pang KC, Mercer TR, Crowe ML, Grimmond SM, Mattick JS. NRED: a database of long noncoding RNA expression. Nucleic Acids Res 2009; 37(Database issue):D122-6; PMID:18829717; http://dx.doi.org/10.1093/nar/gkn617.
24. Bisognin A, Sales G, Coppe A, Bortoluzzi S, Romualdi C. MAGIA²: from miRNA and genes expression data integrative analysis to microRNA-transcription factor mixed regulatory circuits (2012 update). Nucleic Acids Res 2012; 40(Web Server issue):W13-21; PMID:22618880; http://dx.doi.org/10.1093/nar/gks460.
25. Bakheet T, Williams BR, Khabar KS. ARED 3.0: the large and diverse AU-rich transcriptome. Nucleic Acids Res 2006; 34(Database issue):D111-4; PMID:16381826; http://dx.doi.org/10.1093/nar/gkj052.
26. Gruber AR, Fallmann J, Kratochvill F, Kovarik P, Hofacker IL. AREsite: a database for the comprehensive investigation of AU-rich elements. Nucleic Acids Res 2011; 39(Database issue):D66-9; PMID:21071424; http://dx.doi.org/10.1093/nar/gkq990.
27. Mokrejs M, Masek T, Vopálensky V, Hlubucek P, Delbos P, Pospísek M. IRESite--a tool for the examination of viral and cellular internal ribosome entry sites. Nucleic Acids Res 2010; 38(Database issue):D131-6; PMID:19917642; http://dx.doi.org/10.1093/nar/gkp981.
28. Castellano S, Gladyshev VN, Guigó R, Berry MJ. SelenoDB 1.0 : a database of selenoprotein genes, proteins and SECIS elements. Nucleic Acids Res 2008; 36(Database issue):D332-8; PMID:18174224; http://dx.doi.org/10.1093/nar/gkm731.
29. Andken BB, Lim I, Benson G, Vincent JJ, Ferenc MT, Heinrich B, et al. 3'-UTR SIRF: a database for identifying clusters of whort interspersed repeats in 3' untranslated regions. BMC Bioinformatics 2007; 8:274; PMID:17663765; http://dx.doi.org/10.1186/1471-2105-8-274.
30. Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, et al. Rfam: Wikipedia, clans and the "decimal" release. Nucleic Acids Res 2011; 39(Database issue):D141-5; PMID:21062808; http://dx.doi.org/10.1093/nar/gkq1129.
31. Anders G, Mackowiak SD, Jens M, Maaskola J, Kuntzagk A, Rajewsky N, et al. doRiNA: a database of RNA interactions in post-transcriptional regulation. Nucleic Acids Res 2012; 40(Database issue):D180-6; PMID:22086949; http://dx.doi.org/10.1093/nar/gkr1007.
32. Jacobs GH, Chen A, Stevens SG, Stockwell PA, Black MA, Tate WP, et al. Transterm: a database to aid the analysis of regulatory sequences in mRNAs. Nucleic Acids Res 2009; 37(Database issue):D72-6; PMID:18984623; http://dx.doi.org/10.1093/nar/gkn763.
33. Campillos M, Cases I, Hentze MW, Sanchez M. SIREs: searching for iron-responsive elements. Nucleic Acids Res 2010; 38(Webserver issue):W360-7; PMID:20460462; http://dx.doi.org/10.1093/nar/gkq371.
34. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 2005; 120:15-20; PMID:15652477; http://dx.doi.org/10.1016/j.cell.2004.12.035.
35. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, et al. Combinatorial microRNA target predictions. Nat Genet 2005; 37:495-500; PMID:15806104; http://dx.doi.org/10.1038/ng1536.
36. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human MicroRNA targets. PLoS Biol 2004; 2:e363; PMID:15502875; http://dx.doi.org/10.1371/journal.pbio.0020363.
37. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet 2007; 39:1278-84; PMID:17893677; http://dx.doi.org/10.1038/ng2135.
38. Vlachos IS, Kostoulas N, Vergoulis T, Georgakilas G, Reczko M, Maragkakis M, et al. DIANA miRPath v.2.0: investigating the combinatorial effect of microRNAs in pathways. Nucleic Acids Res 2012; 40(Webserver issue):W498-504; PMID:22649059; http://dx.doi.org/10.1093/nar/gks494.
39. Huang GT, Athanassiou C, Benos PV. mirConnX: condition-specific mRNA-microRNA network integrator. Nucleic Acids Res 2011; 39(Webserver issue):W416-23; PMID:21558324; http://dx.doi.org/10.1093/nar/gkr276.
40. Ule J, Jensen KB, Ruggiu M, Mele A, Ule A, Darnell RB. CLIP identifies Nova-regulated RNA networks in the brain. Science 2003; 302:1212-5; PMID:14615540; http://dx.doi.org/10.1126/science.1090095.
41. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell 2010; 141:129-41; PMID:20371350; http://dx.doi.org/10.1016/j.cell.2010.03.009.
42. König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat Struct Mol Biol 2010; 17:909-15; PMID:20601959; http://dx.doi.org/10.1038/nsmb.1838.
43. German MA, Pillay M, Jeong DH, Hetawal A, Luo S, Janardhanan P, et al. Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. Nat Biotechnol 2008; 26:941-6; PMID:18542052; http://dx.doi.org/10.1038/nbt1417.
44. Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, et al. The UCSC Genome Browser database: extensions and updates 2011. Nucleic Acids Res 2012; 40(D1):D918-23; PMID:22086951; http://dx.doi.org/10.1093/nar/gkr1055.
45. Hoffmann PR, Berry MJ. Selenoprotein synthesis: a unique translational mechanism used by a diverse family of proteins. Thyroid 2005; 15:769-75; PMID:16131320; http://dx.doi.org/10.1089/thy.2005.15.769.

46. Gong C, Maquat LE. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. Nature 2011; 470:284-8; PMID:21307942; http://dx.doi.org/10.1038/nature09701.
47. Yoon JH, Abdelmohsen K, Srikantan S, Yang X, Martindale JL, De S, et al. LincRNA-p21 Suppresses Target mRNA Translation. Mol Cell 2012; 47:648-55; PMID:22841487; http://dx.doi.org/10.1016/j.molcel.2012.06.027.
48. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 2011; 27:431-2; PMID:21149340; http://dx.doi.org/10.1093/bioinformatics/btq675.
49. Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009; 4:44-57; PMID:19131956; http://dx.doi.org/10.1038/nprot.2008.211.
50. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, et al. Ensembl 2012. Nucleic Acids Res 2012; 40(D1):D84-90; PMID:22086963; http://dx.doi.org/10.1093/nar/gkr991.
51. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, et al. Genome-wide analysis of estrogen receptor binding sites. Nat Genet 2006; 38:1289-97; PMID:17013392; http://dx.doi.org/10.1038/ng1901.
52. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res 2011; 39(Database issue):D1005-10; PMID:21097893; http://dx.doi.org/10.1093/nar/gkq1184.