THE UNIVERSITY OF TEXAS AT SAN ANTONIO, COLLEGE OF BUSINESS

# Working Paper SERIES

May 30, 2008

Wp# 0053MSS-301-2008

## Over-represented sequences located on UTRs are potentially involved in regulatory functions

Kihoon Yoon
Department of Epidemiology and Biostatistics, The University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Dr., San Antonio, TX 78229-3900, USA

Daijin Ko
Department of Management Science and Statistics, The University of Texas at San Antonio, 6900 N. Loop 1604 West, San Antonio, TX 78249-0632, USA

Carolina B. Livi
Computational Biology Initiative, The University of Texas at San Antonio, 6900 N. Loop 1604 West, San Antonio, TX 78249-0667, USA

Nathan Trinklein
SwitchGear Genomics, 1455 Adams Drive #1317, Menlo Park, CA 94025, USA

Mark Doderer
Department of Computer Science, The University of Texas at San Antonio, 6900 N. Loop 1604 West, San Antonio, TX 78249-0667, USA

Stephen Kwek
Department of Computer Science, The University of Texas at San Antonio, 6900 N. Loop 1604 West, San Antonio, TX 78249-0667, USA

Luiz O. F. Penalva
Department of Cellular and Structural Biology, The University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Dr., San Antonio, TX 78229-3900, USA

*Department of Marketing,*
*University of Texas at San Antonio,*
*San Antonio, TX 78249, U.S.A*



UTSA COLLEGE OF BUSINESS KNOWLEDGE FOR A NEW WORLD

ONE UTSA CIRCLE
SAN ANTONIO, TEXAS 78249-0631
210 458-4317 | BUSINESS.UTSA.EDU

# Over-represented sequences located on UTRs are potentially involved in regulatory functions

Kihoon Yoon[1], Daijin Ko[2], Carolina B. Livi[1,3], Nathan Trinklein[4], Mark Doderer[5], Stephen Kwek[5], Luiz O. F. Penalva[6*]

[1]Department of Epidemiology and Biostatistics, The University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Dr., San Antonio, TX 78229-3900, USA

[2]Department of Management Science and Statistics, The University of Texas at San Antonio, 6900 N. Loop 1604 West, San Antonio, TX 78249-0632, USA

[3]Computational Biology Initiative, The University of Texas at San Antonio, 6900 N. Loop 1604 West, San Antonio, TX 78249-0667, USA

[4]SwitchGear Genomics, 1455 Adams Drive #1317, Menlo Park, CA 94025, USA

[5]Department of Computer Science, The University of Texas at San Antonio, 6900 N. Loop 1604 West, San Antonio, TX 78249-0667, USA

[6]Department of Cellular and Structural Biology, The University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Dr., San Antonio, TX 78229-3900, USA

*Corresponding author - E-mail: penalva@uthscsa.edu, Telephone: +1 (210) 562-9049, Fax: +1 (210) 562-9014

**ABSTRACT**

Eukaryotic gene expression must be coordinated for the proper functioning of biological processes. This coordination can be achieved both at the transcriptional and post-transcriptional levels. In both cases, regulatory sequences placed at either promoter regions or on UTRs function as markers recognized by regulators that can then activate or repress different groups of genes according to necessity. While regulatory sequences involved in transcription are quite well documented, there is a lack of information on sequence elements involved in post-transcriptional regulation. We used a statistical over-representation method to identify novel regulatory elements located on UTRs. An exhaustive search approach was used to calculate the frequency of all possible *n-mers* (short nucleotide sequences) in 16,160 human genes of NCBI RefSeq sequences and to identify any peculiar usage of *n-mers* on UTRs. After a stringent filtering process, we identified *circa* 4,000 highly over-represented *n-mers* on UTRs. We provide evidence that these *n-mers* are potentially involved in regulatory functions. Identified *n-mers* overlap with previously identified binding sites for HuR and Tia1 and, AU-rich and GU-rich sequences. We determined also that over-represented *n-mers* are particularly enriched in a group of 159 genes directly involved in tumor formation. Finally, a method to cluster *n-mer* groups allowed the identification of putative gene networks.

JEL Code: C10

## INTRODUCTION

Post-transcriptional regulation plays a very important role in many biological processes such as embryogenesis, stem cell proliferation, spermatogenesis, sex determination, neurogenesis, erythropoiesis, etc (reviewed in Kuersten and Goodwin, 2003 (1)). The impact it has on the final protein outcome of a cell can be appreciated through studies that compare the steady state levels of mRNAs (transcriptome) and proteins (proteome) in the same cell population (2-6). In the case of some genes, substantial differences were found with accumulated levels of the protein and its corresponding message varying by as much as 30-fold (2). Unfortunately, despite its importance, post-transcriptional regulation continues to be a poorly understood subject.

There are essentially four cytoplasmic processes that can be modulated in eukaryotic cells, ultimately leading to changes in protein production: RNA transport/localization, degradation/stability and RNA translation. Most of the elements necessary for proper regulation of the former three processes are located in the 5' and 3' untranslated regions (UTR) of mRNAs. UTR sequences involved in regulation can be grouped into different categories. The most common are short sequence motifs that function as binding sites for RNA binding proteins and/or non-coding RNAs. Repetitive sequence elements, such as CUG repeats, have also been documented to function as a target of RNA binding proteins. Finally, some UTR sequences interfere with gene expression independently of the action of a regulator; their structural features pose a barrier, potentially influencing the translation of mRNAs. This is the case of moderately stable secondary structures that are typically located in the 5' UTR relatively close to the AUG start codon (reviewed in Mignone et al., 2002 (7)).

There are several examples of UTR mediated regulation in connection to health related issues. For instance, Iron Regulatory Protein (IRP) controls the expression of several mRNAs (ferritin,

transferrin, mitochondrial aconitase, etc) that have a regulatory element named iron responsive element (IRE). IRPs bind to IREs in situations of iron deprivation and inhibit mRNA translation. Mutations that affect the IRE can lead to human disease such as hereditary hyperferritinemia-cataract (reviewed in Rouault, 2006 (8)). Another good example is the amyloid-β precurson protein (APP) implicated in Alzheimer's and Down syndrome. Translation of APP mRNA is up-regulated by interleukin-1 through 5' UTR sequences (reviewed in Pickering and Willis, 2005 (9)). UTR-mediated regulation is also associated with cancer. For instance, approximately 10% of all mRNAs have atypically long 5' UTRs, in many cases containing a variety of regulatory elements. 75% of genes with long 5' UTRs encode oncogenes and genes implicated in cell growth, death and proliferation (9).

Only a small fraction of the regulatory elements located on human UTRs are currently known. In most cases, the described elements were derived from studies of individual genes and their specific regulators. Unfortunately, current engines that predict putative UTR regulatory elements do not produce the expected results in high throughput searches; there are not sufficient labeled instances to allow the employment of machine learning techniques (for example, neural network or Markov models) to construct predictive models. Current UTR search/prediction tools are very rudimentary and cannot be compared to the sophisticated ones that predict transcription factor binding sites (e.g. TRANSFAC) (10). Therefore, novel alternative computational approaches that do not rely exclusively on previously described elements are needed.

Recently, a computational method was used to identify short sequence motifs (named *pyknons*) that are over-represented in the genome. After analyzing the distribution of *pyknons*, it was observed that there is a bias towards UTRs (11). *Pyknons* can constitute a valuable resource in terms of providing new lists of putative regulatory elements. Another recent study used the

power of evolutionary biology to map novel putative regulatory sequences via sequence alignment on promoters and 3' UTRs. This study successfully predicted new miRNA target sequences (12) and constitutes another useful source for the identification of UTR regulatory elements. Finally, an over representation method was used recently to predict target sites of miRNAs on 3' UTR of human genes (13). Our work goes a step beyond these analyses by using a method that specifically calculates over-represented sequences on human UTRs. Our approaches led to the identification of approximately 4000 highly over-represented sequence elements (*n-mers*). In agreement with the idea that these *n-mers* potentially function as regulatory elements, comparisons between them and mapped binding sites for the RNA binding proteins HuR and Tia1 and between them and AU rich sequences (ARE) and GU-Rich elements (GRE) showed statistically significant overlap. Very importantly, a subset of 5' UTR *n-mers* was tested *in vivo* leading to the identification of a family of repressors. Moreover, we managed to identify putative post-transcriptional "operons" by performing a cluster analysis to group genes that share the same set of *n-mers*. In several occasions, strong biological correlation was observed amongst genes present in a cluster.

## MATERIALS AND METHODS

### Preparation of mRNA sequence lists

In order to prepare a reliable list of human mRNA sequences, we started by conducting a feasibility study on 40,874 sequences obtained from NCBI Human Genome FTP site (ftp://ftp.ncbi.nih.gov/genomes/). To ensure the quality of data used, all mRNAs were constructed from chromosome sequences (Build 36.2) based on gene information from RefSeq. Only 'validated' or 'reviewed' gene information was used for mRNA construction. Subsequently, coding regions were extracted from the constructed mRNAs and BLASTed against the entire nucleotide database to confirm that the gene information from RefSeq was correct. If BLAST returned the identical gene ID with a perfect sequence match for a queried coding region, we retained the queried gene in the valid set of mRNAs. After filtering the data, we were left with 20,840 human mRNA sequences corresponding to 16,160 genes. This subset of sequences was used in our analysis.

### *n-mer* counting

An exhaustive search approach measured the appearance of all possible *n-mers* ($2 \leq n \leq 21$) in the mRNA data set. Appearances were counted on the 5' UTR, coding region and 3' UTR individually. In order to handle large number of possible *n-mers* within optimal time and space, we used a suffix tree counter, which is a type of data structure that allows efficient string matching and searching. Although there are many different flavors of suffix tree implementations, we used a straight forward implementation without data compression functions since our sole purpose is simple counting rather than string searching or matching. The counting

procedure collected *n-mer* information such as the list of associated mRNAs and locations of *n-mers* on the mRNA sequences.

To determine if a given *n-mer* is over-represented in a particular section of the mRNA, we first calculated their total lengths. We summed in each case (5' UTRs, coding regions and 3' UTRs) the individual values determined for the 20,840 transcripts present in our mRNA set. The numbers obtained were 4,626,913 nucleotides for 5' UTR, 37,499,577 nucleotides for coding region and 22,871,121 nucleotides for 3' UTR. We then constructed a conversion table in order to perform a balanced analysis that allows the comparison among *n-mers* of different sizes. Table S1 in Supplementary data shows the adjusted mathematical expected appearance value for each *n-mer* based on its length. These values were then used to determine if a particular *n-mer* is over-represented in 5' or 3 UTRs or coding region.

**Parameters to identify highly significant over-represented *n-mers***

An over-represented *n-mer* was considered highly significant in the following cases. A highly significant over-represented 5' UTR *n-mer* was defined by an adjusted P-value less than or equal to 0.01 and should appear in 5 or more 5' UTRs of different genes. However, this *n-mer* could appear in other regions in small numbers – at most 4 times in coding regions and at most twice in 3' UTRs. These numbers are arbitrary and they were selected based on the total length of each section of the mRNA, as described above. This under-representation in other regions as well as over-representation in 5' UTR was counted in calculating the adjusted P-values. For an over-represented 3' UTR *n-mer*, the adjusted P-value is set to less than or equal to 0.01 and 20 or more genes must contain it on their 3' UTR regions. It is also allowed to appear at most 4 times in coding regions and at most once in 5' UTRs. As in 5' UTR, this under-representation in these

regions was also counted in the adjusted P-values. These values were selected based on the average length of 5' UTR, coding region and 3' UTR.

**Statistical analysis to identify over-represented sequences**

Under the assumption that all four nucleotides are independent and distributed in equal proportions in our mRNA dataset, we estimate the probability P of finding a specific pattern of L-mer to be $P = 4^{-L}$. Hence in the data base of total length D of UTR or coding regions of length at least L, the expected number of the given pattern of L-mer is $\lambda_L = (D-n*(L-1))* 4^{-L}$ where n is the number of RNAs whose UTR or coding regions are of length at least L. If the motifs were randomly distributed over the different sections of the mRNA, the distribution of the number of RNAs with a specific motif would be Poisson with mean rate $\lambda_L$. We used the Poisson distribution to calculate the probability of observing *k* RNAs with the specific motif pattern, which is $(\lambda_L)^k e^{-k}/k!$. So with the expected number $\lambda_L$ of the given pattern of L-mer, the probability of observing *k* or more RNAs with the L-mer is 1- P(k-1, $\lambda_L$ ) and *k* or less instances is P(k, $\lambda_L$) where $P(k,\lambda_L) = \sum_{x=o}^{k} \frac{1}{x!} (\lambda_L)^x e^{-\lambda_L}$. For example, the probability of observing *k* or more instances in 5' UTR and at the same time, 0 instances in both 3' UTR and coding region is therefore P= (1- P(k-1, $\lambda_{L5}$ ))*P(0, $\lambda_{L\ 3}$)*P(0, $\lambda_{LC}$) where $\lambda_{L5}$ , $\lambda_{L3}$ , *and* $\lambda_{LC}$ are the expected numbers in 5' UTR, 3' UTR and coding region respectively. When several RNAs come from a single gene, dependence among the RNAs is expected. To achieve the conservative P-values, we use the number of gene-instances instead of the number of RNA-instances for the over-counting (*k* or more instances). For under-counting (*k* or less RNA instances), we used the number of RNA-instances to make P-values more conservative. Since we were testing the significance of a specific pattern for all the patterns, we used Bonferroni-Correction to adjust P-values for

multiple testing. The total number (TN) of the patterns from 2-mer through 21-mer is $\sum_{n=2}^{21} 4^n$ =5.864062e+12 and adjusted P value is min(TN*P, 1).

**Preparation of random samples**

One thousand sets of random *n-mer* sequences were generated from the sequences present in our list of mRNAs. To construct the random sets, we took into consideration the size and the number of over-represented *n-mers* present in our final data set.

**Comparison between *n-mers* and HuR and Tia1 binding sites**

The data provided by Dr. Isabel Lopez de Silanes contains binding sites for the RNA binding proteins HuR and Tia1 obtained via RIP-Chip and computational methods (14-16). We located all these binding sites on the mRNA sequences present in our list. These locations were then compared to the positions of the over-represented and random *n-mers*. Since the length of binding sites obtained for HuR and Tia-1 is longer than 21 nucleotides, we only considered two sequences to be a 'match' when an over-represented *n-mer* or a random *n-mer* appears within a HuR or Tia1 binding site. The numbers of matches obtained for the over-represented *n-mer* list was compared to the numbers obtained for a 1,000 sets of random *n-mers*.

**Search for *n-mers* containing the AUUUA (UAUUUAU) motif**

We searched for AUUUA and UAUUUAU motifs (ARE core sequences) in the 5' and 3' UTR over-represented *n-mer* sets and in random *n-mer* sets. Initially, the total number of AUUUA (or UAUUUAU) appearances in a data set was simply counted. However, it is possible that a small number of AU-rich *n-mers* in a given data set contribute to 2 or more motif counts. We counted

then in each data set the total number of *n-mers* containing one or more AUUUA (or UAUUUAU) motif. All the counting results were compared between the over-represented *n-mer* set and random *n-mer* sets.

**Search for *n-mers* containing GU-Rich elements GRE**

We searched for the UGUUUGUUUGU motif (GRE consensus sequence) in the 5' and 3' UTR over-represented *n-mer* sets and in random *n-mer* sets. The total number of GRE appearances in a data as well as the number of *n-mers* containing one or more GRE motifs was counted. All the counting results were compared between the over-represented *n-mer* set and random *n-mer* sets.

**Calculation of UTR length vs *n-mer* number**

5' UTR and 3' UTR length of each mRNA in the dataset were measured individually. Each UTR length was correlated to the number of over-represented *n-mers* appeared on the UTR regions to characterize the length effect on the number of *n-mers*. We also examined the lengths of UTRs and the number of over-represented *n-mers* in 159 cancer related genes. The gene list (Table S2) and the detailed method used can be obtained from Supplementary data as well as the supporting material website (http://gccri.uthscsa.edu/sequences.html).

**Cluster analysis on over-represented *n-mers***

We performed cluster analysis based on functional similarity. Functional Similarity between two *n-mers*, say *n-mer* 1 and *n-mer* 2, is defined as the number of genes that have both *n-mers* divided by the number of genes that have either one or both. The dissimilarity (distance) between two *n-mers* is defined as 1 minus the similarity. Using this dissimilarity measure and Kaufman

and Rousseeuw' s Partitioning Around Medoids (PAM) algorithm (17), we organized 5' UTR *n-mers* and 3' UTR *n-mers* into clusters. Average silhouette lengths were used to order the clusters. We represent in our supporting material website (http://gccri.uthscsa.edu/sequences.html) the top 25 5' UTR *n-mer* clusters (Table S3) and the top 100 3' UTR clusters (Table S4).

After grouping the over-represented *n-mers* into clusters, gene members in a cluster were analyzed by using 'Pathway Studio 5' (http://www.ariadnegenomics.com/) in order to identify known functional relationships among them.

**Clone preparation and luciferase assays**

A total of 30 5' UTR *n-mer* sequences were cloned into the 5' UTR of the actinβ gene present in the vector pSGG_5UTR (Switchgear Genomics). We determined how these sequences influenced gene expression using a luciferase assay. The resulting constructs as well as the empty vector control and two other controls containing the iron responsive element (IRE) in sense and anti-sense orientations in the 5' UTR were transfected into HT1080 and HeLa cells. HT1080 cells were selected to be the primary source for the assays based on its consistency when compared to other cell lines tested by Switchgear Genomics. ~5000 cells were plated in 96-well plates in OPTI-MEM (Invitrogen). Cells were transfected 16 hours later with FuGene (Roche) following the manufacture's protocol. 24 hours after transfection, 100µl of Steady-glo reagent (Promega) were added to each well. Samples were covered with foil, incubated at room temperature for 30 minutes and read in a luminomenter. All the experiments were performed in quadruplicate.

A more detailed version of the Methods section is provided in our supporting material website (http://gccri.uthscsa.edu/sequences.html).

## RESULTS AND DISCUSSION

### Identification of over-represented sequences on UTRs of human genes

We used over-representation as a strategy to map putative regulatory sequences on UTRs. Over-represented sequences are frequently used as point of reference to identify putative regulatory elements (18). It assumes that an observed sequence bias can indicate the presence of a regulatory element. A given sequence motif (*n-mer*) is considered over-represented if it appears more frequently than its statistically expected frequency. Contrary to previous studies, we employed a more elaborate method specifically designed for a UTR study. First, having in mind that transcribed and non-transcribed sequences are under different selective pressure and that repetitive sequences present in intergenic regions can create a bias and alter the final results; we used mRNA sequences instead of genomic sequences as the sample for the counting process. Second, since regulatory elements can vary in size, we opted not to restrict the size of the *n-mers* to be counted.

Our strategy to identify highly significant over-represented *n-mers* located on UTRs was divided into two steps. Initially, we calculated over-representation taking into consideration mathematical expected frequencies, size of the *n-mer* and average length of the different portions of the mRNA (5' UTR, coding region and 3' UTR) in our sample pool. It is worth noticing that 3' UTRs are on average 4 times longer than 5' UTRs. The result of this analysis led to the identification of approximately 43 million over-represented sequences. This number of sequences is too large to allow us to extract meaningful biological data. Moreover, in this first analysis, we did not compare the *n-mers* to each other. As was expected, the first list of over–represented sequences contains a lot of redundancy. Thus we carried out a second analysis step to organize and reduce the data. Briefly, in this second step, we eliminated *n-mers* that are not

totally contained in UTRs and *n-mers* that are part of a longer one (the shorter *n-mers* were eliminated only in case the corresponding longer ones are contained in the exact same subset of mRNAs). Moreover, we established additional parameters: 1) we initially took into consideration only P-values to list a given sequence as over-represented; in this second step, we also took into account the number of genes in which an individual *n-mer* appears; 2) based on average length of each section of the mRNA, we established minimum and maximum appearance values for 5' UTR, coding regions and 3' UTR; in this new scenario, *n-mers* are selected only if they are over-represented in either the 5' or 3' UTR and at the same time have low counts in the other two sections of the mRNA. From these criteria, we identified **1124** and **2772** over-represented motifs from 5' UTR and 3' UTR region, respectively (see supplementary data file 4). A summary of our strategy to identify UTR over-represented elements is represented in **Figure 1 and 2,** and details are given in the Methods section.

The data was organized into two *n-mer* sets (*n-mers* over-represented in 5' UTR only and *n-mers* over-represented in 3' UTR only). **Table 1** shows examples of identified over-represented *n-mers* located on 5' or 3' UTR. In order to facilitate future analyses, we ranked the *n-mers* according to the adjusted P-value that indicates the statistical significance of the "fold increase" in relation to the adjusted expected frequency. P-values are reported in log units and for $P<10^{-200}$ we set it to be $10^{-200}$. The entire list of over-represented *n-mers* is present in supplementary data (see the supporting material website).

## The nature of *n-mer* sequences

We performed an initial analysis with the two lists of over-represented *n-mers* to identify particular features and commonalties shared by the *n-mers*. First, one can notice the absence of

short *n-mers* in both 5' and 3' UTR lists. In the case of 5' UTR, the *n-mers* range between 9 and 32 nucleotides while in the case of 3' UTR, they range between 9 and 38 nucleotides. The absence of *n-mers* shorter than 9 nucleotides can find its explanation in the parameters and cut-offs we used and established to generate the lists. First, we were looking for *n-mers* that are over-represented in only one section of the mRNA. A large number of *n-mers* were discarded because we determined they are over-represented also in other sections of the mRNA. Second, as explained above, *n-mers* that are part of a longer one were also eliminated.

We determined the GC and AU content of *n-mers* present in the 5' and 3' UTR lists. The detailed tables are provided in our supplementary data website. 5' UTR *n-mers* are extremely rich in GC. Circa 20% of all *n-mers* identified are 100% GC; ~75% of them have a content of 80% GC or higher. If we calculate the GC content in the entire 5' UTR dataset, we will notice a slight bias. However, it is not, in any case, sufficient to justify the high number of GC-rich containing *n-mers*. We suggest that these *n-mers* might correspond to large families of regulatory elements and we discuss some possibilities. The first possibility is that a portion of these sequences are the target of RNA binding proteins that have preferences for GC sequences. This is the case for members of the CUGBP family (19). For instance, CUGBP1 binds a sequence with several copies of the GC dinucleotide in the 5' UTR of the p21 mRNA and enhances its translation. We observed that part of the regulatory element located in the 5' UTR of the p21 mRNA (CTGCCGCCGCCG) is present in some *n-mers* of our list (20). Moreover, a substantial number of *n-mers* have the format $GCN_{(1-3)}GCN_{(1-3)}GC$…, similar to p21 regulatory element. Another possibility is that some *n-mers* with high GC content could be part of highly complex secondary structure that could interfere with translation (21). In this particular case, it is worth mentioning the recently described RNA G-quadruplex. This element initially identified in the

NRAS proto oncogene functions as a translator repressor (22). The authors observed that 2,992 mRNAs contain one ore more G-quadruplex similar to the one of the NRAS 5' UTR. In our dataset, we also identified over-represented *n-mers* containing a G-quadruplex or part of it, for instance: GGGCCGGGCCGGGCCGGG, GGCGGGCGGGCGGGC and CGGGCGGGGCGGGGC. Finally, we should mention a family of elements named Simple Sequence Repeats (SSRs); they are tandem repeats of sequences between 1 and 6 nucleotides. In a recent study, it was observed that they are present in great number in 5' UTRs, especially amongst house-keeping genes (23). The most frequent repeat is CGG. We checked our data set and identified several *n-mers* containing $(CGG)_N$ sequences. The function of SSRs in 5' UTRs still needs to be determined, but there is data showing that in some cases they play a role in gene expression (24). The situation regarding the 3'UTR *n-mers* is the opposite; we observed that *n-mers* are rich in AU sequences. ~5% of 3' UTR *n-mers* are 100% AU; ~39% of them have a content of 80% AU or higher. The strong presence of AU rich sequences in our 3' UTR *n-mer* set did not come as a surprise. ARE sequences (AU rich) are probably the most relevant group of regulatory elements located on 3' UTRs (reviewed in Lopez de Silanes et al., 2007 (25)). Numerous mRNAs have been identified to be regulated at the post-transcriptional level by RNA binding proteins that recognize these sequences. We describe below an analysis that establishes a correlation between the identified 3' UTR *n-mers* and ARE.

**_n-mer_ sequences overlap with previously identified regulatory motifs**

If over-represented *n-mers* are indicative of the presence of regulatory sequences, one would expect to see an overlap between them and already mapped RNA binding protein recognition sites. In order to test this hypothesis, we compared our *n-mer* list to binding sites of the RNA

binding proteins HuR and Tia1 obtained via RIP-Chip analysis. These binding sites were deduced with computational methods based on commonalities at the level of RNA sequence and structure and information from previously characterized HuR and Tia1 sites (14,15). Detailed information was kindly provided by Dr. Isabel Lopez de Silanes. To determine if our results are statistically significant, we generated a total of 1000 random sequence sets from actual human UTR sequences; the length of individual sequences present in the over-represented *n-mer* lists was considered when preparing those lists. Finally, we compared the lists of Tia1 and HuR binding sites to the lists of random sequences to determine the number of overlaps. The results we obtained are summarized in **Table 2.** In agreement with the idea that there is a correlation between over-representation and biological function, the number of over-represented sequences (*n-mers*) in 3' UTRs matching either HuR or Tia1 binding sites is significantly higher than the numbers obtained from the comparison with the random sets (P-values < 0.001). The list of *n-mers* match to HuR and Tia1 binding sites are shown in **Table S5** and **S6**, respectively.

We employed another approach to determine if over-represented *n-mers* coincide with previously described UTR regulatory elements. We compared our dataset to ARE sequences (described in the previous section) and to the recently identified GU-Rich elements (GRE) (26). The AUUUA and the UAUUUAU motifs have been described as the basic core of ARE sequences. We expected to see a large portion of the 3' UTR *n-mers* that contained the core sequence as well as a bias towards the 3' UTR since ARE sequences have not been assigned for 5' UTRs. Indeed, the number of over-represented sequences (*n-mers*) that have core ARE sequences is significantly higher than that obtained from random sets. Moreover, a 3' UTR bias was observed (P-values < 0.001) - **Table 3.** The GU-rich element (GRE), whose consensus is UGUUUGUUUGU, was identified via computational methods to find conserved sequences in

the 3' UTR of genes that exhibited rapid decay in primary human T-cells. These sequences were determined to be involved in mRNA stability and to be regulated by the CUG-binding protein 1 (26). Identically to what was observed for the ARE sequences, we determined that the number of over- *n-mers* containing a GRE is significantly higher than that obtained from random sets; a 3' UTR bias (P-values < 0.001) was observed as well - **Table 4.** In conclusion, the results of this section indicate that *n-mers* do overlap with regulatory elements. The lists of *n-mers* containing ARE and GRE sequences are in **Table S7**, **S8** and **S9**. Detailed results of both analyses of this section are provided as supporting materials.

## ***In vivo*** **analysis of 5' UTR** ***n-mers*** **identifies a highly conserved negative regulator associated with uORFs**

The top ranked 30 5' UTR *n-mers* (corresponding to those with higher gene counts) were cloned into a 5' UTR luciferase-reporter vector (pSGG-5UTR from Switchgear Genomics) and used in transient transfections; their impact on gene expression was measured via luciferase activity. The use of this vector to evaluate 5' UTR regulatory sequences was tested successfully with a characterized iron responsive element (IRE) cloned in sense and anti-sense orientation (data not shown). Experiments were primary done in HT1080 cells and results confirmed in HeLa cells. Our results indicate that the *n-mer* sequences we selected repress gene expression at different levels – **Figure 3A and B**. A group of very similar *n-mers* stood out among the ones that caused a very strong negative effect. These are five *n-mer* sequences (5, 12, 13, 16, 21) containing the core element AUGGCGG – **Figure 3A**. We examined our dataset to encounter other *n-mers* with the same core sequence; a total of 67 *n-mers* covering 101 genes were identified. The 5' UTRs of these genes were compared and a larger consensus sequence was determined: an AUG placed

within a GC-rich sequence - **Figure 3C**. A closer look at all implicated genes pointed out that the *n-mers* always overlap with the start codon of a uORF (**Table S10**). In agreement with our results, uORFs have been shown to affect mRNA translation in a negative fashion. 31% of uORFs contained in human genes are also present in rodents, highlighting its importance as regulatory elements (27). Based on its high conservation, we suggest that the GC-rich context in which the uAUGs are located contributes substantially to the observed regulatory effect. Further experiments are necessary to dissect this conserved element and check how the GC-rich component contributes to regulation.

A second group of *n-mers* that produced strong negative effect (10, 11, 15, 25) has in common a high G content. When examining the results of all constructs as a whole, we observed that comparatively, *n-mers* rich in Gs tend to produce a stronger regulatory effect. We discussed in the previous section the participation of G and GC rich elements in 5' UTR mediated regulation. The four *n-mer* sequences listed above are very similar and could be the target of a RBP with preferential binding for G rich sequences.

**Cancer related genes and *n-mer* frequency**

We calculated the number of *n-mers* located on the 5' and 3' UTRs for each gene present in our list. Although a certain variation was observed, we can affirm that there is a direct correlation between the length of the UTR and the number of *n-mers* identified **(Figure 4)**. We expect that genes with a high number of UTR regulatory sequences to be tightly regulated and/or to present a restricted pattern of expression. Assuming that the number of *n-mers* correlates with the number of regulatory sequences, we anticipate that genes falling into this category contain more *n-mers* than the average. This should be the case of genes directly involved in tumor formation,

whose expression has to be tightly regulated and, in the case of oncogenes, restricted to particular developmental stages. Mapping UTR regulatory elements in cancer related genes is a very important issue, since it can lead to the discovery of novel pathways involved in tumorigenesis as well as novel alternatives for cancer therapy. To test our hypothesis, we used a Poisson regression model and compared the number of *n-mers* present in the 5' and 3' UTRs of 159 genes that were directly implicated in tumor formation, mainly oncogenes, to the number of the *n-mers* present in the UTRs of the entire mRNA list we generated – see **Figure 5** for results. In both cases (5' and 3' UTRs), cancer related genes generally have more *n-mers* than other genes. These differences are statistically significant in both the 3' UTR and 5' UTR ($P<0.001$). The difference appears more dramatic in the case of the 5' UTR than in the case of 3' UTR. This data is in agreement with the fact that 5' UTR regulatory elements have been reported for numerous oncogenes (28). Moreover, several oncogenes are known to have their translation initiated by internal ribosome entry sites (IRES) in a cap-independent mechanism. These IREs are in general contained in long 5' UTRs with high GC content (29,30). Regarding cap-dependent translation, it is known that the P13K/AKT/mTOR signaling pathway regulates the translation of genes involved in cell proliferation and growth, among them several oncogenes (31,32).

## Cluster analysis and identification of putative gene networks regulated at the post-transcriptional level

All biological processes depend on the coordinated activity of a selected group of proteins. Before a given biological process like cell division takes place, it is necessary to synchronize the expression of genes that code for the set of implicated proteins. This synchronization can be

achieved at the post-transcriptional level through the action of specific RNA binding proteins and non-coding RNAs that recognize UTR sequences shared by the gene group. Regulators and their corresponding target mRNAs form the so-called post-transcriptional operons (33,34).

In order to identify genes that could be potentially co-regulated, forming a functional post-transcriptional operon, we employed a method to identify gene clusters that share sets of *n-mers*. Briefly, we considered that two *n-mers* are 'similar' if these *n-mers* are frequently appearing in the same genes. Unlike a clustering method based on the sequence similarity, a good cluster is defined as a group of *n-mers* sharing nearly identical gene lists. **Figure 6** illustrates the clustering analysis procedure. A more detailed explanation about the cluster is described in the Methods section. We generated 25 5' UTR clusters and 100 3' UTR clusters. We observed that in general the *n-mers* present in a cluster have similarities in terms of sequence. This is an agreement with the idea that they constitute variations of a functional element recognized by the same gene regulator. We then performed multiple sequence alignments for each set of similar *n-mers* present in a given cluster to identify a core element. If our cluster analysis functions as a method to identify gene networks, we should be able to identify strong biological associations among genes in the same cluster at least in some of the cases. To identify these possible associations, we analyzed the gene clusters using 'Pathway Studio 5'. This analysis indicated that several sets of gene clusters share commonalities in terms of pathway and function. Moreover, interacting proteins turned out to be present in numerous clusters. **Figure 7** shows examples of strong biological associations as well as core sequence elements identified in two different gene clusters. The 5' UTR cluster represented in the figure shows a group of genes linked to the TGFβ signaling pathway while the represented 3' UTR cluster shows a group of genes implicated in the VEGF signaling pathway. The remaining clusters that turned out to show positive results and *n-*

*mer* comparisons are described in the supporting material website. To check that the identified relations in **Figure 7** are not a random incident, we performed 100 cluster analyses with sets of random genes. We concluded that the same type of direct correlations as exemplified in **Figure 7** cannot be obtained by chance alone ($P < 0.01$) (see the supporting materials website for details).

## CONCLUSION

We designed a specific method based on over-representation to map putative regulatory sequences present on UTRs. A very strict filter consisting of minimum and maximum values of appearances for each region of the mRNA (5' UTR, coding region and 3' UTR) was used to select a group of approximately 4,000 highly relevant over-represented sequences (*n-mers*). The evidence strongly indicates a correlation between over-representation and function. The identified *n-mers* overlap with previously identified binding sites for HuR and Tia1 and AU-rich and GU-rich sequences. Moreover, a group of selected 5' UTR *n-mers* proved to affect gene expression. In particular, we identified a set of *n-mers* containing a highly conserved sequence that that turned out to be part of uORFs and functions as a negative regulator. We also determined that over-represented *n-mers* are particularly enriched in a group of 159 cancer related genes. Finally, a method to cluster *n-mer* groups allowed the identification of putative post-transcriptional gene networks.

The method we employed differs from previous analysis of UTR motifs in several ways. First, the choice of sequence data for UTR analysis is different from others' published efforts. Most previous UTR analyses dealt with the entire chromosome sequences to construct their models while we used only transcript sequences for our analysis in order to build a more accurate background model. It is also notable that we explicitly handled the number of *n-mer* appearances in non-target regions with the consideration of length effects. Moreover, our clustering approach was based on functional relations, not sequence similarities, which has more biological sense. Last, the compilation of *n-mers* present in a given identified cluster allowed us to construct 'core *n-mers*'. The sequence

variation observed among members of a "core *n-mer"* resembles what is observed for actual binding sites targeted by the same regulator. When all the evidence is combined, we believe this dataset contains information that will guide the discovery of novel functional elements. All data is available online to the scientific community.

**ACKNOWLEDGEMENTS**

The authors would like to thank Dr. Cristine Vogel, Dr. John Cornell, Dr. Jonathan Gelfond, Dr. Yufei Xiao, Dr. Ulus Atasoy, Dr. Susan Naylor, Dr. Jack Keene, Dr. Fatima Gebauer, Tarea Burton and Suzanne Burns for comments and Dr. Isabel Lopez de Silanes for comments and providing the list of binding sites for HuR and Tia1 and for comments and Dr. Brad Windle for advice regarding the *n-mer* analysis.

This work was sponsored by a grant from the SempRuss Foundation and Melvin Leazar Memorial Fund of the San Antonio Area Foundation (SAAF) and supported by the Computational Biology Initiative (UTSA/UTHSCSA).

Dr. Ko's research was supported in part by the College of Business Summer Research Grant at the University of Texas at San Antonio.

## REFERENCES


1. Kuersten, S. and Goodwin, E.B. (2003) The power of the 3' UTR: translational control and development. *Nat Rev Genet*, **4**, 626-637.
2. Gygi, S.P., Rochon, Y., Franza, B.R. and Aebersold, R. (1999) Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol*, **19**, 1720-1730.
3. Knoll-Gellida, A., Andre, M., Gattegno, T., Forgue, J., Admon, A. and Babin, P.J. (2006) Molecular phenotype of zebrafish ovarian follicle by serial analysis of gene expression and proteomic profiling, and comparison with the transcriptomes of other animals. *BMC genomics*, **7**, 46.
4. Unwin, R.D. and Whetton, A.D. (2006) Systematic proteome and transcriptome analysis of stem cell populations. *Cell cycle (Georgetown, Tex )*, **5**, 1587-1591.
5. Habermann, J.K., Paulsen, U., Roblick, U.J., Upender, M.B., McShane, L.M., Korn, E.L., Wangsa, D., Kruger, S., Duchrow, M., Bruch, H.-P. *et al.* (2007) Stage-specific alterations of the genome, transcriptome, and proteome during colorectal carcinogenesis. *Genes, chromosomes & cancer*, **46**, 10-26.
6. Lu, P., Vogel, C., Wang, R., Yao, X. and Marcotte, E.M. (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol*, **25**, 117-124.
7. Mignone, F., Gissi, C., Liuni, S. and Pesole, G. (2002) Untranslated regions of mRNAs. *Genome Biol*, **3**, REVIEWS0004.
8. Rouault, T.A. (2006) The role of iron regulatory proteins in mammalian iron homeostasis and disease. *Nat Chem Biol*, **2**, 406-414.
9. Pickering, B.M. and Willis, A.E. (2005) The implications of structured 5' untranslated regions on translation and disease. *Seminars in cell & developmental biology*, **16**, 39-47.
10. Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K. *et al.* (2006) TRANSFAC (R) and its module TRANSCompel (R): transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, **34**, D108-D110.
11. Rigoutsos, I., Huynh, T., Miranda, K., Tsirigos, A., McHardy, A. and Platt, D. (2006) Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 6605-6610.
12. Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S. and Kellis, M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**, 338-345.
13. Cora, D., Di Cunto, F., Caselle, M. and Provero, P. (2007) Identification of candidate regulatory sequences in mammalian 3' UTRs by statistical analysis of oligonucleotide distributions. *BMC bioinformatics*, **8**, 174.
14. Lopez de Silanes, I., Fan, J., Galban, C.J., Spencer, R.G., Becker, K.G. and Gorospe, M. (2004) Global analysis of HuR-regulated gene expression in colon cancer systems of reducing complexity. *Gene expression*, **12**, 49-59.

15. Lopez de Silanes, I., Galban, S., Martindale, J.L., Yang, X., Mazan-Mamczarz, K., Indig, F.E., Falco, G., Zhan, M. and Gorospe, M. (2005) Identification and functional outcome of mRNAs associated with RNA-binding protein TIA-1. *Mol Cell Biol*, **25**, 9520-9531.
16. Lopez de Silanes, I., Lal, A. and Gorospe, M. (2005) HuR: post-transcriptional paths to malignancy. *RNA biology*, **2**, 11-13.
17. Kaufman, L. and Rousseeuw, P. (1990) *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.
18. Defrance, M. and Touzet, H. (2006) Predicting transcription factor binding sites using local over-representation and comparative genomics. *BMC bioinformatics*, **7**, 396.
19. Barreau, C., Paillard, L., Mereau, A. and Osborne, H.B. (2006) Mammalian CELF/Bruno-like RNA-binding proteins: molecular characteristics and biological functions. *Biochimie*, **88**, 515-525.
20. Timchenko, N.A., Iakova, P., Cai, Z.J., Smith, J.R. and Timchenko, L.T. (2001) Molecular basis for impaired muscle differentiation in myotonic dystrophy. *Mol Cell Biol*, **21**, 6927-6938.
21. Kozak, M. (1991) Structural Features in Eukaryotic Messenger-Rnas That Modulate the Initiation of Translation. *J Biol Chem*, **266**, 19867-19870.
22. Kumari, S., Bugaut, A., Huppert, J.L. and Balasubramanian, S. (2007) An RNA G-quadruplex in the 5 ' UTR of the NRAS proto-oncogene modulates translation. *Nat Chem Biol*, **3**, 218-221.
23. Lawson, M. and Zhang, L. (2008) Housekeeping and tissue-specific genes differ in simple sequence repeats in the 5'-UTR region. *Gene*, **407**, 54-62.
24. Toutenhoofd, S.L., Garcia, F., Zacharias, D.A., Wilson, R.A. and Strehler, E.E. (1998) Minimum CAG repeat in the human calmodulin-1 gene 5 ' untranslated region is required for full expression. *Bba-Gene Struct Expr*, **1398**, 315-320.
25. Lopez de Silanes, I., Quesada, M.P. and Esteller, M. (2007) Aberrant regulation of messenger RNA 3'-untranslated region in human cancer. *Cellular oncology : the official journal of the International Society for Cellular Oncology*, **29**, 1-17.
26. Vlasova, I.A., Tahoe, N.M., Fan, D., Larsson, O., Rattenbacher, B., John, J.R.S., Vasdewani, J., Karypis, G., Reilly, C.S., Bitterman, P.B. *et al.* (2008) Conserved GU-rich elements mediate mRNA decay by binding to CUG-binding protein 1. *Molecular Cell*, **29**, 263-270.
27. Iacono, M., Mignone, F. and Pesole, G. (2005) UAUG and uORFs in human and rodent 5 ' untranslated mRNAs. *Gene*, **349**, 97-105.
28. van der Velden, A.W. and Thomas, A.A. (1999) The role of the 5' untranslated region of an mRNA in translation regulation during development. *The international journal of biochemistry & cell biology*, **31**, 87-106.
29. Willis, A.E. (1999) Translational control of growth factor and proto-oncogene expression. *Int J Biochem Cell B*, **31**, 73-86.
30. Stoneley, M. and Willis, A.E. (2003) Aberrant regulation of translation initiation in tumorigenesis. *Curr Mol Med*, **3**, 597-603.
31. Mamane, Y., Petroulakis, E., LeBacquer, O. and Sonenberg, N. (2006) MTOR, translation initiation and cancer. *Oncogene*, **25**, 6416-6422.

32. Polunovsky, V.A. and Bitterman, P.B. (2006) The cap-dependent translation apparatus integrates and amplifies cancer pathways. *RNA Biol.*, **3**, 10-17.
33. Keene, J.D. and Tenenbaum, S.A. (2002) Eukaryotic mRNPs may represent posttranscriptional operons. *Molecular Cell*, **9**, 1161-1167.
34. Keene, J.D. (2007) RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet*, **8**, 533-543.

## FIGURES LEGENDS

**Figure 1. Summary of experimental procedures to identify short sequence elements (*n-mers*) that are over-represented in either 5' or 3' UTRs**. mRNA sequences were generated after a filtering process where mRNA and genomic sequence data from the NCBI website were compared. A total of 20,840 mRNA sequences were used in our counting process. Individual *n-mer* counts were performed for the coding region (CR), 5' and 3' UTRs. In this first step, the adjusted mathematical expected appearance value for each *n-mer* based on its size was used to determine if a given sequence is over-represented in either the 5' or 3' UTRs. A total of ~43 million over-represented *n-mers* were identified.

**Figure 2. Schematic representation of the filtering process to identify highly over-represented *n-mers* in either 5' or 3' UTRs**. We employed a two step process. In filter 1, we established minimum and maximum counts for the different sections of the mRNA [5' UTR, coding region (CR) and 3' UTR], the counting process was done with transcripts rather than genes. In filter 2, we established a minimum number of appearances in different genes to consider an *n-mer* over-represented.

**Figure 3. *In vivo* analysis of the top 30 5' UTR *n-mers*. A)** List of *n-mers* cloned into the 5' UTR of pSGG_5UTR. **B)** Resulting clones were transfected into HT1080 cells and the luciferase activity was measured. Values reflect the results of 4 independent experiments done for the 30 different constructs and the empty vector control. **C)** Consensus sequence determined for the initiation codon of uORFs associated with *n-mers* containing the conserved sequence AUGGCGG.

**Figure 4. UTR length vs number of *n-mers***. In this figure, the average number of *n-mers* identified per gene is plotted as a function of the length of their 5' or 3' UTR. To facilitate visualization, we created groups; for instance one group contains all genes containing between 1 and 5 *n-mers* in their 5' UTR. Yellow bars indicate variation observed in each "*n-mer* group".

**Figure 5. Cancer related genes and *n-mers***. The graphs illustrate comparisons between the average numbers of *n-mers* encountered for the 5' and 3' of UTRs of 159 cancer related genes and the average numbers of *n-mers* encountered for the 5' and 3' UTRs of the entire population of mRNAs used in our analysis.

**Figure 6. Schematic representation of the cluster analysis used to identify putative post-transcriptional operons**. First, a dissimilarity score matrix was constructed by comparing gene lists associated with each over-represented *n-mer* to all the others. Next, Partitioning Around Medoids (PAM) clustering algorithm starts with randomly selected arbitrary number of *n-mers* that serve as medoids (centers of clusters). The rest of the non-medoid *n-mers* are assigned to the nearest medoids according to their dissimilarity scores. After the initial partitioning, the algorithm swaps the current medoids with non-medoid *n-mers* and updates cluster memberships for non-medoid *n-mers* to check if new medoids lead to a better partition in term of the average dissimilarities in clusters. These steps are repeated until the average dissimilarities of clusters cannot be reduced further.

**Figure 7. Examples of gene clusters that show strong biological associations**. Most relevant gene clusters identified in our study were analyzed with the Pathway Studio software to identify possible biological interactions amongst the genes present in them. Only direct associations/interactions are illustrated in the figure. The results from

multiple sequence alignments represent possible ‘core *n-mers*’ that were built with *n-mers* present in the cluster. Multiple sequences alignments were performed by using Clustal X. The 5’ UTR cluster represented in the figure shows a group of genes linked to the TGFβ signaling pathway while the represented 3’ UTR cluster shows a group of genes implicated in the VEGF signaling pathway.

## TABLES

**Table 1**. **Example of over-represented *n-mers* for 5' UTR (A) and 3' UTR (B) identified after very stringent criteria**.

| 5' UTR *n-mers* | adj. P value | Gene count |
|---|---|---|
| CTCCCGCGCGC | -200 | 19 |
| GCGCCCCCTCCCC | -200 | 18 |
| GCCCGGCTCGGC | -200 | 15 |
| CCCCGCGCTCCC | -200 | 15 |
| CGGGCGCCCGCG | -200 | 15 |
| CCCGGCCCGCCCG | -200 | 15 |
| GCGGCGCTCGGG | -200 | 14 |
| TCTCCACAGAGGAG | -200 | 9 |
| ACCTGCAGGTATTG | -200 | 9 |
| GCAGGTATTGGGAGAT | -200 | 9 |
| AGAGGAAGAGGAAAG | -200 | 8 |
| AAGGAGAAGATCTGCC | -200 | 7 |
| ACCACTCAGGGTCCTGTGGACAGCTCACCTAG | -200 | 5 |

| 3' UTR *n-mers* | adj. P value | Gene count |
|---|---|---|
| CTGGCCAACATGGTGAAACCC | -200 | 169 |
| AGCCTGGCCAACATGGTGAAA | -200 | 144 |
| AACTCCTGACCTCAGGTGATC | -200 | 125 |
| AACCCCGTCTCTACTAAAAAT | -200 | 122 |
| GATCACCTGAGGTCAGGAGTT | -200 | 116 |
| CTGGCCAACATGGTGAAACCCC | -200 | 115 |
| GTGGCTCACACCTGTAATCCC | -200 | 113 |
| TCCCAGCTACTCAGGAGGCTG | -200 | 102 |
| TGGCTCACACCTGTAATCCCAG | -200 | 101 |
| ACTGCACTCCAGCCTGGGTGA | -200 | 100 |
| ACCTGTAATCCCAGCACTTTG | -200 | 98 |
| CACTGCACTCCAGCCTGGGTG | -200 | 93 |
| TTTTTTTTTTTTTTGAGACAG | -200 | 88 |

**Table 2. *n-mer* comparison to HuR and Tia1 binding sites. Each comparison is represented in two columns.** In the first column, the numbers reflect perfect overlaps between described HuR or Tia1 binding sites and over-represented *n-mers*. In the second column, the numbers reflect average values obtained from 1000 comparisons between described HuR or Tia1 binding sites and random *n-mer* sets generated from sequences present in our mRNA set. (SD: Standard Deviation)

| | | Comparison to HuR binding sites | | Comparison to Tia1 binding sites | |
|---|---|---|---|---|---|
| | | Over-represented *n-mers* | Mean (SD) of random *n-mer* samples | Over-represented *n-mers* | Mean (SD) of random *n-mer* samples |
| Number of RNAs with at least one mapped HuR/Tia1 binding site matching a *n-mer* | in 5' UTR | 0 | 0 | 0 | 0 |
| | in 3' UTR | 839 | 101.6 (27.7) | 314 | 41.8 (24.5) |
| Total number of mapped Hur/Tia1 binding site matching a *n-mer* | 5' UTR | 0 | 0 | 0 | 0 |
| | 3' UTR | 1078 | 108.1 (47.5) | 377 | 57.5 (83.2) |

**Table 3. *n-mer* comparison to ARE sequences**. The table contains the number of over-represented *n-mers* containing AUUUA and UAUUUAU sequences as well as average values obtained for 1000 comparisons with random *n-mer* sets generated from sequences present in our mRNA set. (SD: Standard Deviation)

| | | Over-represented *n-mer* | Mean (SD) of Random Samples |
|---|---|---|---|
| Number of appearances of the AUUUA motif in a *n-mer* set | 5' UTR | 0 | 6.5 (2.6) |
| | 3' UTR | 122 | 70.5 (8.3) |
| Number of appearances of the UAUUUAU motif in a *n-mer* set | 5' UTR | 0 | 0.4 (0.6) |
| | 3' UTR | 35 | 9.0 (2.9) |
| Number of *n-mers* containing one or more AUUUA motif | 5' UTR | 0 | 6.4 (2.5) |
| | 3' UTR | 116 | 67.1 (7.8) |
| Number of *n-mers* containing one or more UAUUUAU motif | 5' UTR | 0 | 0.4 (0.7) |
| | 3' UTR | 35 | 8.7 (3.0) |

**Table 4. *n-mer* comparison to GRE sequences**. The table contains the number of over-represented *n-mers* containing GRE sequences as well as average values obtained for 1000 comparisons with random *n-mer* sets generated from sequences present in our mRNA set. (SD: Standard Deviation)

| | | Over-represented *n-mer* | Mean (SD) of Random Samples |
|---|---|---|---|
| Number of appearances of the UGUUUGUUUGU motif in a *n-mer* set | 5' UTR | 0 | 0 (0) |
| | 3' UTR | 12 | 0.29 (0.79) |
| Number of *n-mers* containing one or more UGUUUGUUUGU motif | 5' UTR | 0 | 0 (0) |
| | 3' UTR | 5 | 0.15 (0.39) |

**FIGURE 1**

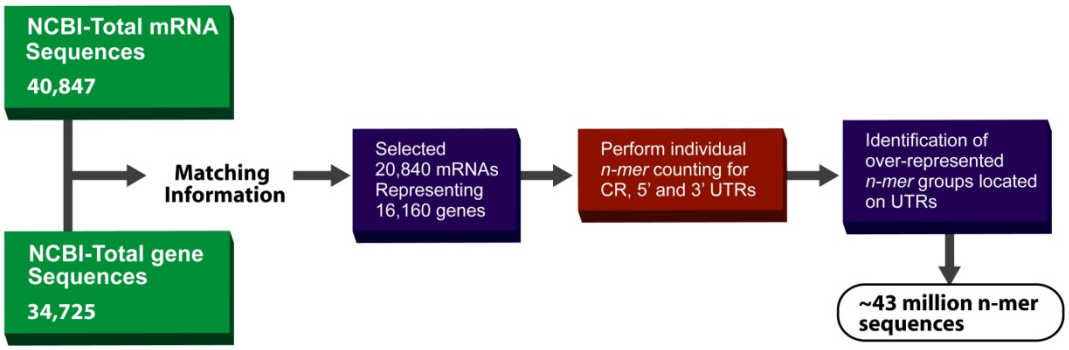
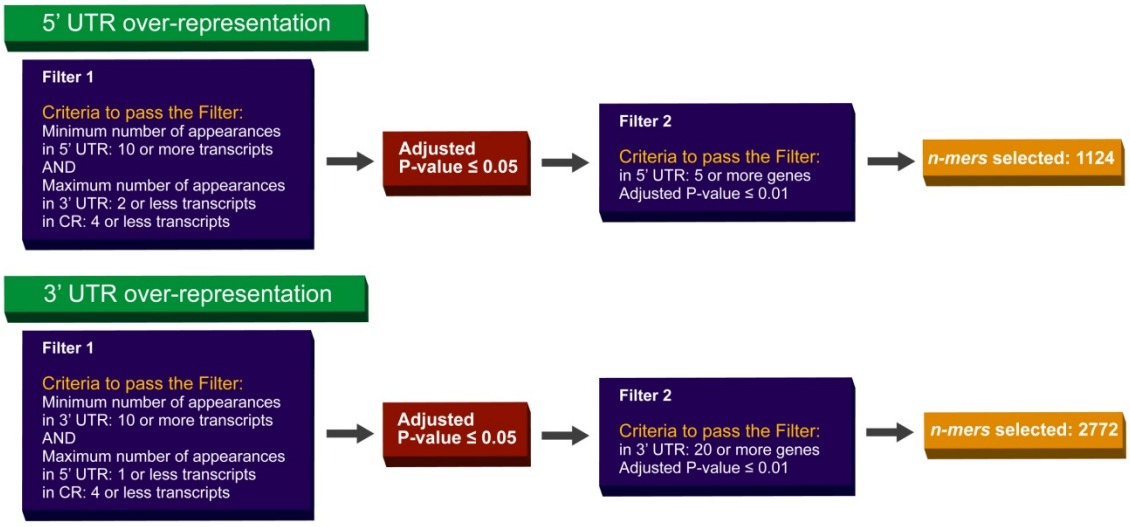



**FIGURE 2**




Over-represented sequences located on UTRs

**FIGURE 3**

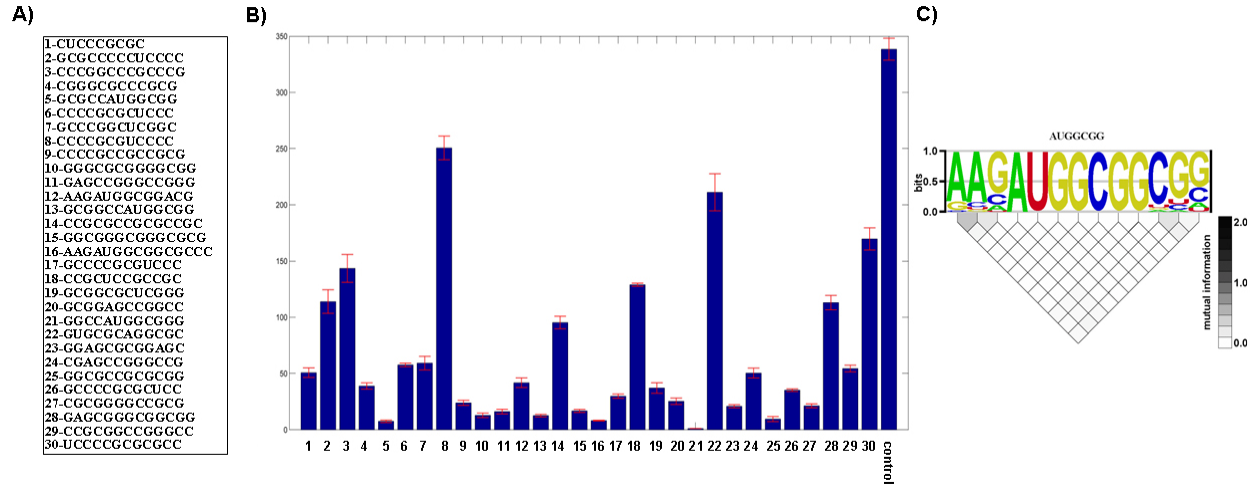
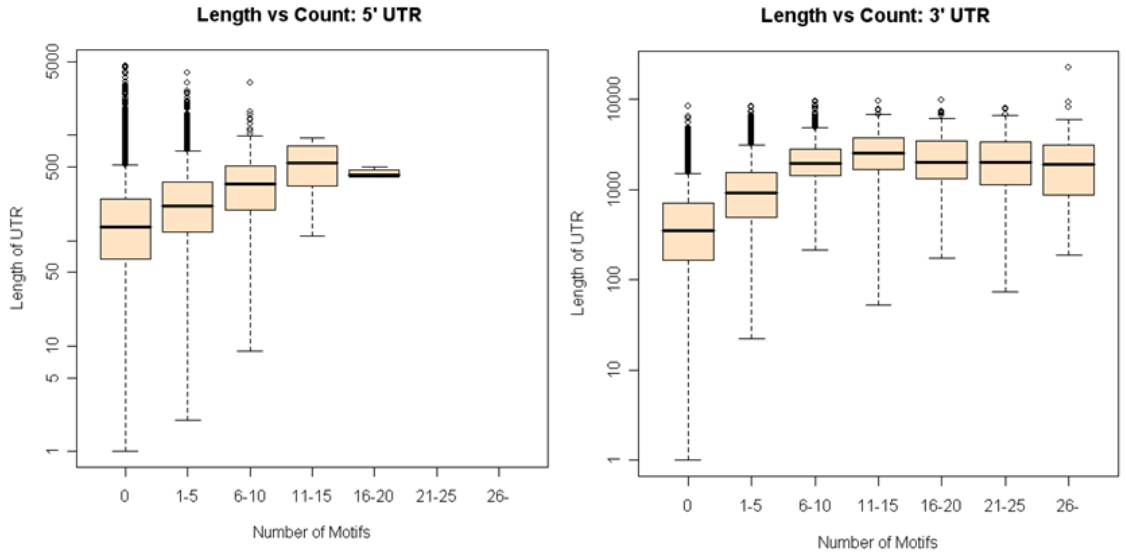
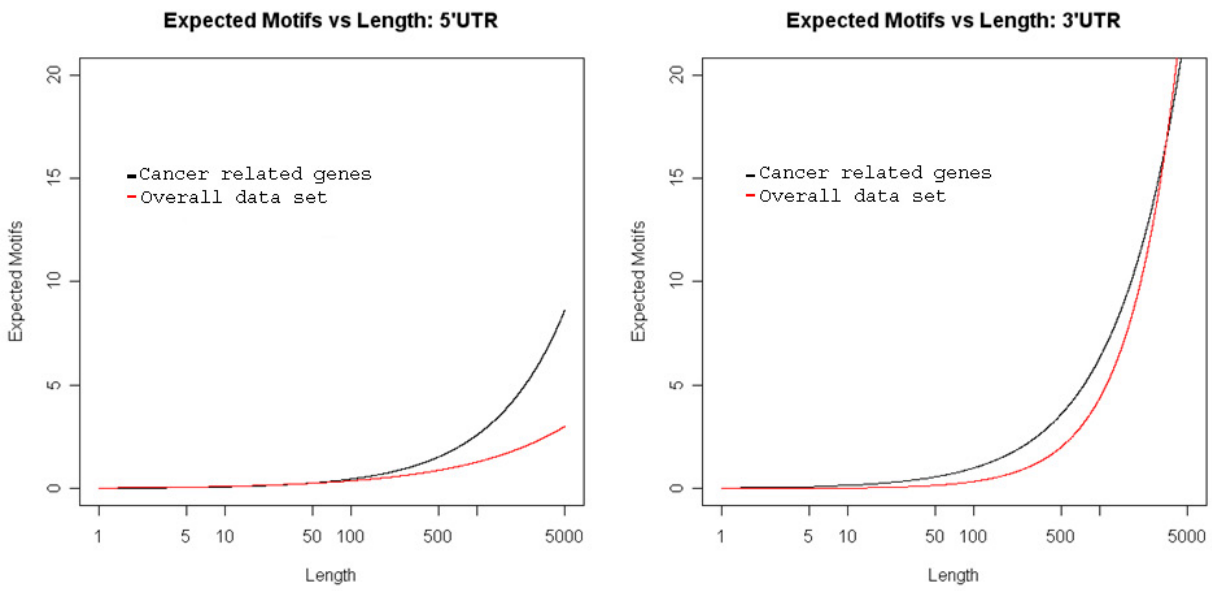



**FIGURE 4**



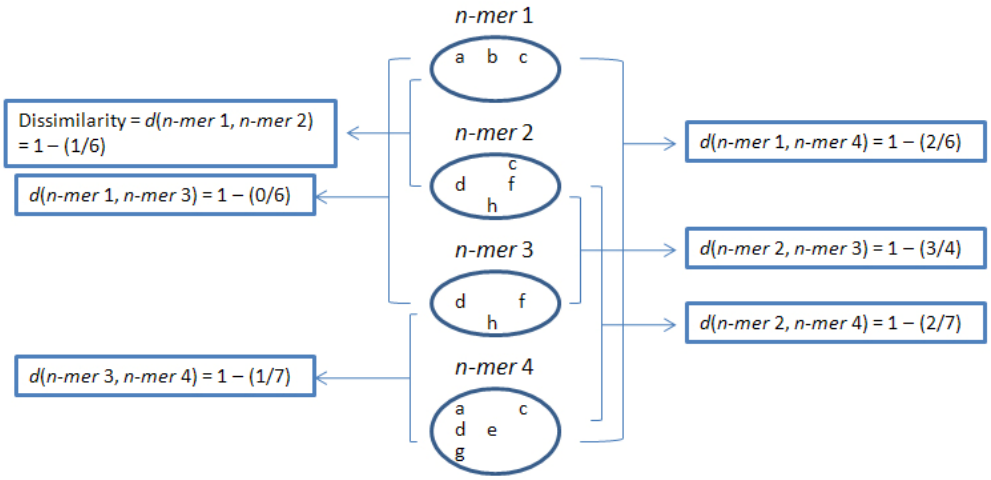

Over-represented sequences located on UTRs

**FIGURE 5**




Over-represented sequences located on UTRs

**FIGURE 6**




1. Dissimilarity Matrix Construction.

Dissimilarity Matrix

| d(n1,n2) | n-mer 1 | n-mer 2 | n-mer 3 | n-mer 4 |
|---|---|---|---|---|
| n-mer 1 | 0 | 0.83 | 1 | 0.67 |
| n-mer 2 | 0.83 | 0 | 0.25 | 0.71 |
| n-mer 3 | 1 | 0.25 | 0 | 0.86 |
| n-mer 4 | 0.67 | 0.71 | 0.86 | 0 |

**Clustering by PAM**

2. Select an initial set of medoids.

| d(n1,n2) | n-mer 1 | n-mer 2 | n-mer 3 | n-mer 4 |
|---|---|---|---|---|
| n-mer 2 (m1) | 0.83 | 0 | 0.25 | 0.71 |
| n-mer 4 (m2) | 0.67 | 0.71 | 0.86 | 0 |

*n-mer* 2 & 4 are selected as medoids, *m1* and *m2*.

3. Assign each *n-mer* to the nearest medoid.

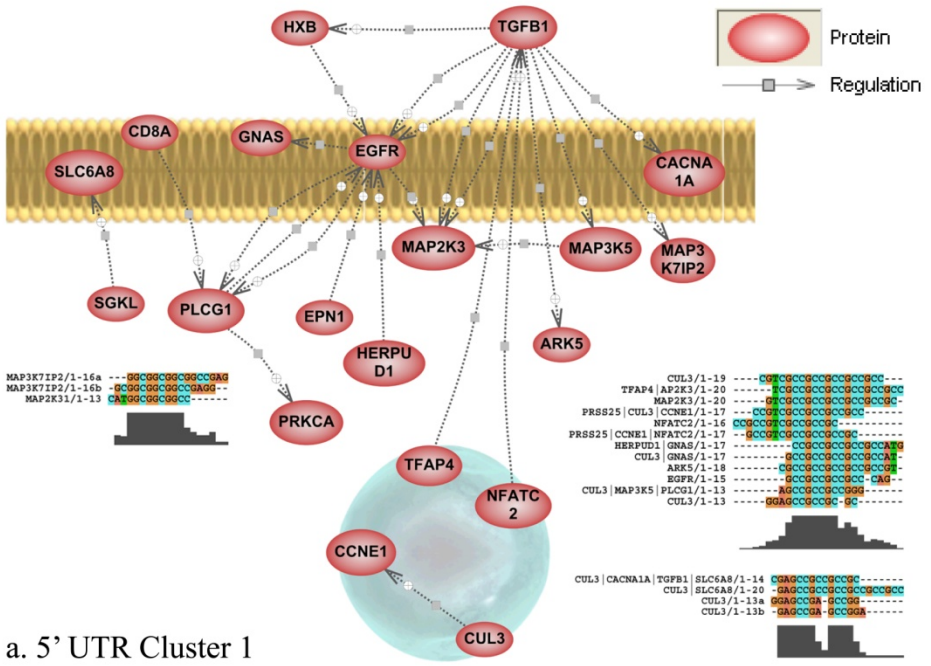
| d(n1,n2) | n-mer 1 | n-mer 2 | n-mer 3 | n-mer 4 |
|---|---|---|---|---|
| n-mer 2 (m1) | 0.83 | 0 | 0.25 | 0.71 |
| n-mer 4 (m2) | 0.67 | 0.71 | 0.86 | 0 |

- Assign *n-mer 1* and *n-mer 3* into *m1* and *m2* clusters: compare dissimilarity scores between medoids and non-medoids *n-mers*. For example, *n-mer* 1 is assigned to the cluster *m2* since 0.83 > 0.6. Thus, *n-mer 3* belongs to cluster *m1*.
- - Average dissimilarities are measured in each cluster ( Avg. dis(*m1*) = 0.25 and Avg.dis(*m2*) = 0.6).

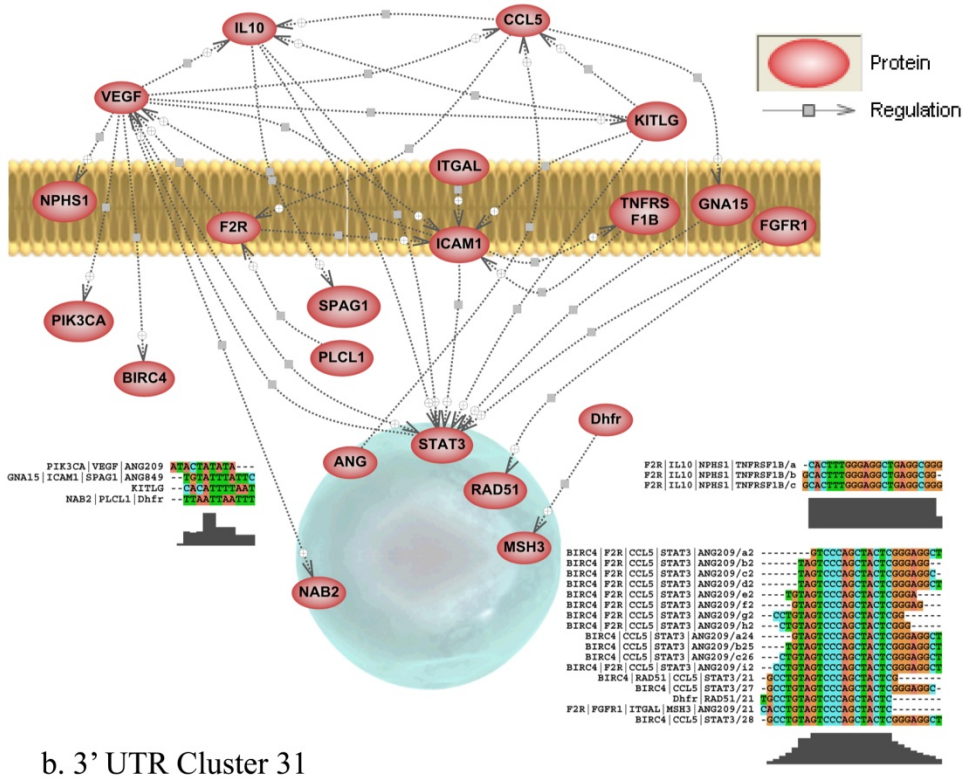
4. Swapping medoids with non-medoids *n-mers* randomly .
5. Repeat step 3 and 4 until find clusters with lower average dissimilarities or stop the iteration if the average dissimilarities are not reducible further.

**In this example, PAM converged into two optimally partitioned clusters, *m1* and *m2*. In terms of average dissimilarity, cluster *m1* is a better cluster than cluster *m2*.**

**FIGURE 7**




a. 5' UTR Cluster 1




b. 3' UTR Cluster 31

Over-represented sequences located on UTRs