BMC Genomic Data

RESEARCH

Open Access

Tandem repeats ubiquitously flank and contribute to translation initiation sites

Ali M. A. Maddi[1], Kaveh Kavousi[1*], Masoud Arabfard[2], Hamid Ohadi[3] and Mina Ohadi[4*]

Abstract

Background: While the evolutionary divergence of *cis*-regulatory sequences impacts translation initiation sites (TISs), the implication of tandem repeats (TRs) in TIS selection remains largely elusive. Here, we employed the TIS homology concept to study a possible link between TRs of all core lengths and repeats with TISs.

Methods: Human, as reference sequence, and 83 other species were selected, and data was extracted on the entire protein-coding genes ($n = 1{,}611{,}368$) and transcripts ($n = 2{,}730{,}515$) annotated for those species from Ensembl 102. Following TIS identification, two different weighing vectors were employed to assign TIS homology, and the co-occurrence pattern of TISs with the upstream flanking TRs was studied in the selected species. The results were assessed in 10-fold cross-validation.

Results: On average, every TIS was flanked by 1.19 TRs of various categories within its 120 bp upstream sequence, per species. We detected statistically significant enrichment of non-homologous human TISs co-occurring with human-specific TRs. On the contrary, homologous human TISs co-occurred significantly with non-human-specific TRs. 2991 human genes had at least one transcript, TIS of which was flanked by a human-specific TR. Text mining of a number of the identified genes, such as *CACNA1A, EIF5AL1, FOXK1, GABRB2, MYH2, SLC6A8,* and *TTN*, yielded predominant expression and functions in the human brain and/or skeletal muscle.

Conclusion: We conclude that TRs ubiquitously flank and contribute to TIS selection at the trans-species level. Future functional analyses, such as a combination of genome editing strategies and in vitro protein synthesis may be employed to further investigate the impact of TRs on TIS selection.

Keywords: Genome-scale, Tandem repeat, Translation initiation site, Homology, TIS selection

Introduction

Translational regulation can be global or gene-specific, and most instances of translational regulation affect the rate-limiting initiation step [1, 2]. While mechanisms that result in the selection of translation initiation sites (TISs) are largely unknown, conservation of the alternative TIS positions and the associated open reading frames (ORFs) between human and mouse cells [3] implies physiological significance of alternative translation. A vast number of human protein-coding genes consist of alternative TISs, which are selected based on complex and yet not fully understood scanning mechanisms [3–6]. The alternative TISs can result in various protein structures and functions [7, 8].

While recent findings indicate that TISs are predominantly a result of molecular error [9], the probability of using a particular TIS differs among mRNA molecules, and can be dynamically regulated over time [10]. Selection of TISs and the level of translation and protein synthesis depend on the *cis* regulatory elements in the

*Correspondence: kkavousi@ut.ac.ir; mi.ohadi@uswr.ac.ir

[1] Laboratory of Complex Biological systems and Bioinformatics (CBB), Department of Bioinformatics, Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Tehran 1417614411, Iran
[4] Iranian Research Center on Aging, University of Social Welfare and Rehabilitation Sciences, Tehran, Tehran 1985713871, Iran
Full list of author information is available at the end of the article

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

mRNA sequence and its secondary structure such as the formation of hair-pins, stem loops, and thermal stability [11–16]. In fact, the ribosomal machinery has the potential to scan and use several ORFs at a particular mRNA species [17].

A *tandem repeat* (TR) is a sequence of one or more DNA base pairs (bp) that is *repeated* on a DNA stretch. While TRs have profound biological effects in evolutionary, biological, and pathological terms [18–24], the effect of these intriguing elements on protein translation remains largely (if not totally) unknown. There are limited publications indicating that when located at the 5′ or 3′ untranslated region (UTR), short tandem repeats (STRs) (core units of 1–6 bp) can modulate translation, the effect of which has biological and pathological implications [25–29]. For example, eukaryotic initiation factors are clamped onto polypurine and polypyrimidine motifs in the 5′ UTRs of target RNAs, and influence translation [30]. Abnormal STR expansions impact TIS selection in a number of neurological disorders [31, 32].

Based on a TIS homology approach, we previously reported a link between STRs and TIS selection [33]. Here, we extend our study to TRs of all core lengths and repeats, an additional weighing vector (vector W_2), several additional species, improved sequence retrieval methods, and a newly developed software and database for data collection and storage.

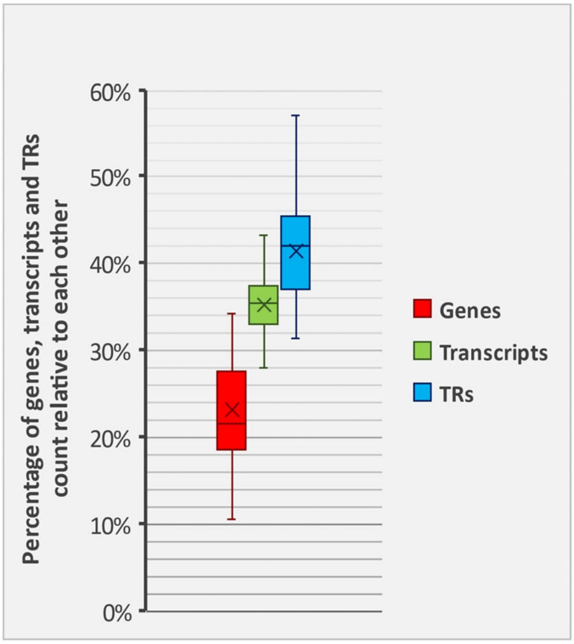

Fig. 1 Abundance interval of the genes, transcripts, and TRs to each other. In this chart, we compared variations in the number of genes, transcripts, and TRs in different species relative to each other. The vertical axis shows what percentage of the total number of genes plus transcripts plus TRs in different species belong to the genes, transcripts, or TRs

Results and discussion

TRs are ubiquitous *cis* elements flanking TISs

A total of 1,611,368 protein-coding genes, 2,730,515 transcripts and 3,283,771 TRs were investigated across the 84 selected species, of which 22,791 genes, 93,706 transcripts, and 99,818 TRs belonged to the human species (Additional Table 1). On average, there were 1.64 transcripts and 1.97 TRs per gene, and 1.19 TRs, per transcript, per species (Fig. 1). The highest ratios of transcripts and TRs per gene (4.11 and 4.38, respectively) belonged to human. Human ranked 59th among 84 species in respect of the TR/transcript ratio (Fig. 2) (Additional Table 1).

Across the 93,706 identified protein-coding transcripts in the human genome, there were 50,169 transcripts, in which TISs were flanked by at least one TR (53.54% of protein-coding transcripts). At a similarly high rate, from the 22,791 identified protein-coding genes in the human genome, 15,256 genes had at least one transcript, in which TISs were flanked by a TR (66.94% of human protein-coding genes). 2850 different types of TRs were identified in the human genome, of which 1504 types (52.77%) were human-specific; across TR categories 1–4, we detected 660, 101, 339 and 404 types of human-specific TRs, respectively, the top most abundant of which are represented in Table 1.

TRs differentially co-occur with TISs

We employed two weighing settings (vectors) for designating homologous vs. non-homologous TISs in human vs. other species. One of those settings was the same as in our previous approach (vector W_1) [33]. In both settings, there was significant co-occurrence of human-specific TRs with non-homologous human TISs, and non-human-specific TRs with homologous human TISs (Fisher's exact $p < 0.01$) (Fig. 3). The results were replicated in 10-fold cross-validation (Fig. 4) (Additional Table 2).

Biological and evolutionary implications

In 15,256 human genes, at least one TIS was flanked by a TR, of which in 2991 genes those TRs were human-specific (Additional Tables 3 & 4). A sample of those genes is listed in Table 2, text mining [34] of a number of which yielded predominant expression and functions in the human brain and/or skeletal muscle, such as *CACNA1A*, *EIF5AL1*, *FOXK1*, *GABRB2*, *MYH2*, *SLC6A8*, and *TTN*. These are examples of expression enrichment in tissues that are frequently subject to human-specific evolutionary processes. However, the nervous system and skeletal

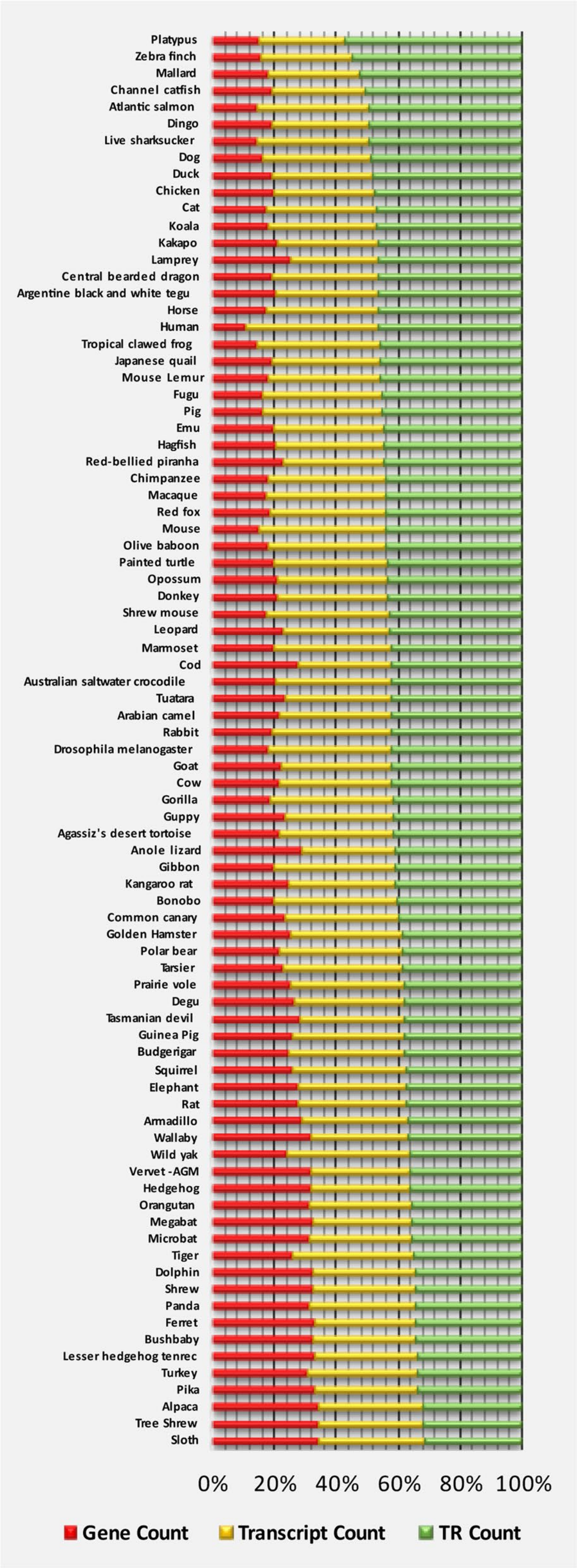

Fig. 2 Ratios of genes, transcripts, and TR counts for each species. The horizontal axis shows the percentage of each entity, and the vertical axis shows each species. Species can be cross-referenced in Additional Table 1

muscle may not be the only tissues, gene functions in which are associated with human-specific characteristics.

We employed the Needleman Wunsch algorithm [35] to further examine the relevance of our findings. To that end, comparison of proteins between human and three other species, consisting of chimpanzee, macaque, and mouse (RESTful API at: https://www.ebi.ac.uk/Tools/psa/emboss_needle [36]), revealed significantly lower homology for the human proteins, in which TISs were flanked by human-specific TRs (Fig. 5).

Our findings provide prime evidence of a link between TRs of all core lengths and repeats, and TIS selection, mechanisms of which are virtually unknown currently. Our approach was based on homology search, which reliably identifies" homologous" TISs by detecting excess similarity [37]. By searching identical gene names across the selected species, our approach encompassed orthologous and paralogous genes.

While the scope of our previous publication [33] was limited to the STRs, in the current study, we investigated TRs of all core lengths (ranging from 1 to 60 nucleotides) and repeats. Another advantage was employment of an improved method for retrieving the upstream flanking sequences. Moreover, whereas BLAST of CDS and cDNA sequences were used to extract the TISs and upstream flanking sequences in the previous study, here we used script programming on the Biomart web application, which is more reliable and accurate. In this method, we specified the gene name, transcript, and length of the upstream flanking sequence for the Biomart web application [38], by using an automated script. In comparison with our previously implemented methods, the result of the automated script is more accurate and comprehensive. An additional weighing method was also implemented in the current study to further examine the relevance of our homology assignment approach.

It is possible that asymmetric and stem-loop structures, which are inherent properties of repeat sequences result in genetic marks that enhance TIS selection. Asymmetric structures have recently been reported to be linked to various biological functions, such as replication and initiation of transcription start sites [39]. Recent studies implicate that the local folding and co-folding energy of the ribosomal RNA (rRNA) and the mRNA correlates with codon usage estimators of expression levels in model organisms such as

Table 1 The top most abundant human-specific TRs flanking TISs. It should be noted that human-specificity applied in the context of the relevant TISs

	Tandem Repeat	Core Length
Category 1	(CT)3	2
	(TC)3	2
	(GC)3	2
	(T)6	1
	(CG)3	2
	(GGC)3	3
	(CTG)3	3
	(CGCC)3	4
	(GGGGC)3	5
	(TGTTTT)3	6
	(CGCGCC)3	6
Category 2	(GGGGCGC)3	7
	(CCCGCCG)4	7
	(GCTGCGGG)3	8
	(AGGGGCGGG)4	9
	(CCTCCCG)4	7
	(CCGGGGG)3	7
	(TTTTTTG)3	7
	(AGCCCAGC)3	8
	(CCCCCGC)3	7
	(ACCCCTCC)3	8
	(AGCCCACGG)3	9
Category 3	(GTGTGTGTTT)2	10
	(ATTTTAAAATT)2	11
	(AAAATAAATAA)2	11
	(TGGCGGCGGCGG)2	12
	(CCCAGCCCCA)2	10
	(CCCCGCCCGCG)2	11
	(CGGGAGTGAGAG)2	12
	(AAGTGGGAAACTGG)2	14
	(TTCATAGATGTTC)2	13
	(ATAGATGTTC)2	10
	(CCCCGCCCCT)2	10
Category 4	(CCCCGAGGTCTCCGCG)2	16
	(CCGGCGTGTACCGAGAGACTGGCGT)2	25
	(ACCTGGAGGGCTGGGG)2	16
	(CCCTGCCCTGTCCTGTCCTGCCCTG)2	25
	(ACCCATCCCCACCTCCCT)3	18
	(CCCTGCCCTGTCCTGTCCTG)2	20
	(ACCCATCCCCACCTCCCT)3	18
	(CCCCACCTCCCTACCCAT)4	18
	(ACAGCGAGGTCGGCAGCGGCAGCGAGGTCGGCAGCGGC)2	38
	(TGAGTCGCAGGCCGAGGAGACAGTGAGTGCGCGCCC)2	36
	(ACTCTCTCTCTTTCTCGGGCTGCAGGTGCACCAGGCCGTCC)2	41

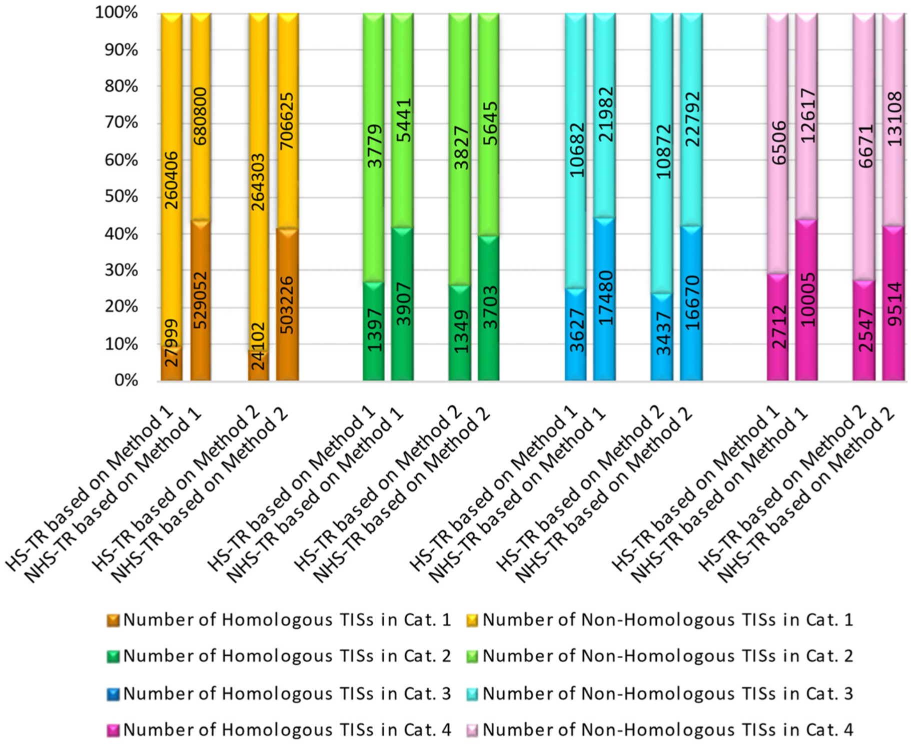

Fig. 3 Average of 10 experiments to examine co-occurrence patterns between TRs and TISs in each of the four TR categories. Each histogram shows the number of homologous vs. non-homologous TISs, based on two different weighing methods (vectors). HS-TR = human-specific tandem repeat, NHS-TR = non-human-specific tandem repeat, TIS = translation initiation site

chloroplast [40]. It may be speculated that RNA structures formed as a result of folding in the TR regions function as marks for TISs.

Among a number of options for future studies, genome editing strategies such as CRISPR/Cas9 [41] in combination with in vitro translation engineering, using cell-free protein synthesis (also known as *in vitro* protein synthesis or CFPS) and/or PURE system (i.e. protein synthesis using purified recombinant elements) [42, 43] may be useful to investigate the impact of TRs on TIS selection and protein synthesis.

Conclusion

We conclude that TRs ubiquitously flank TIS sequences and contribute to TIS selection at the trans-species level. Future functional analyses, such as a combination of genome editing strategies and in vitro protein synthesis are warranted to investigate the impact of TRs on TIS selection.

Materials and methods

Data collection

All sequences, species, and gene datasets collected in this study were based on Ensembl 102 (http://nov2020.archive.ensembl.org/index.html), scheme of which is depicted in Fig. 6 .

84 species were selected, which encompassed orders of vertebrates and one non-vertebrate species (*D. melanogaster*) (Fig. 2). Throughout the study, all species were compared with the human sequence, as reference. The list of species was extracted via RESTful API, in Java language. In parallel, a list of available gene datasets of the selected species was collected by using the "biomaRt" package [44, 45] in R language. In the next step, in each selected species, all protein-coding transcripts of protein-coding genes were extracted. To that end, identical gene names were used across the selected species to group orthologous/paralogous genes in those species.

Subsequently, the 120 bp upstream flanking sequence of all annotated protein coding TISs were retrieved and analyzed. All steps of data collection were performed by querying on the Biomart Ensembl tool via RESTful API, which was implemented in the Java language, except fetching the primary list of available species and gene datasets. For each species, its name, common name and display name were retrieved. For each gene in each species, its gene name, Ensembl ID and

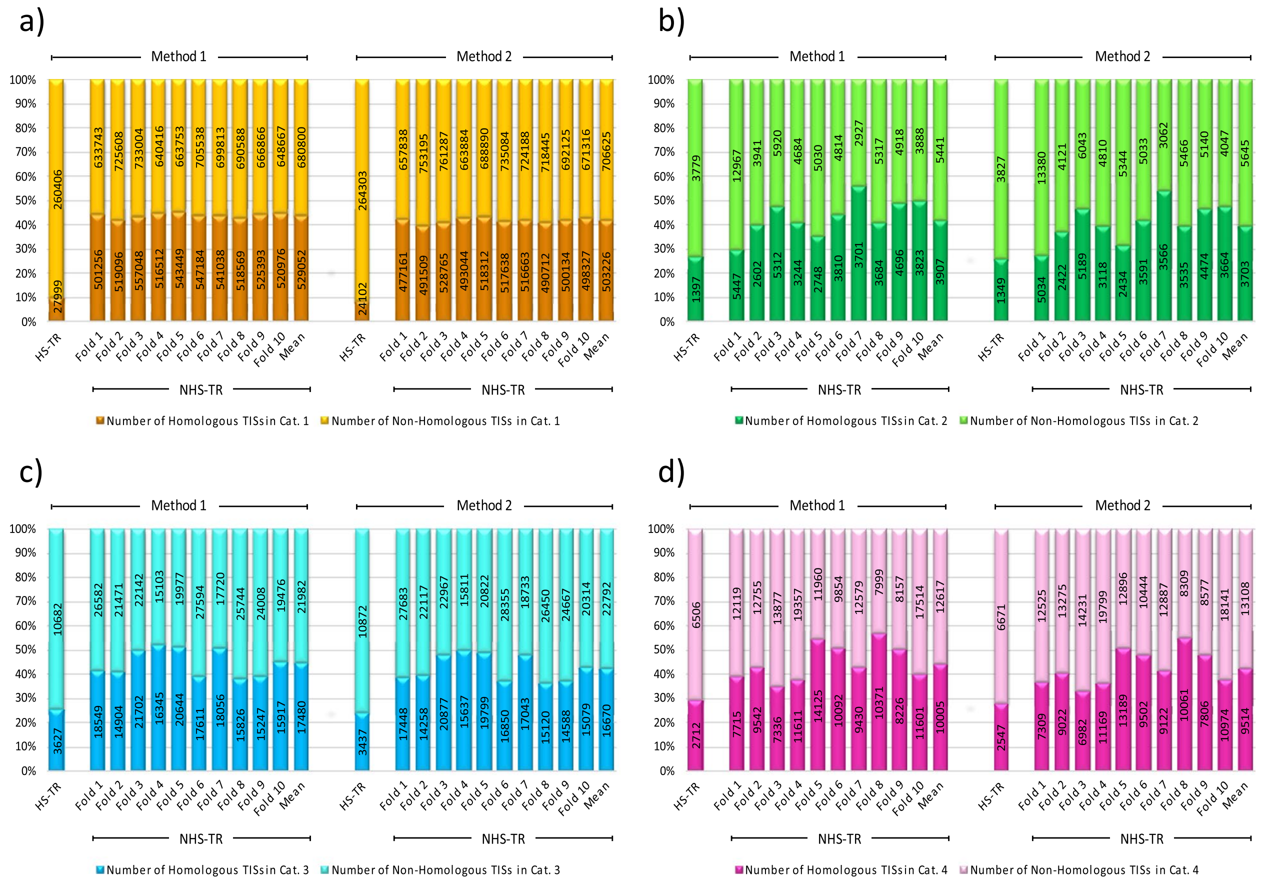

Fig. 4 10-fold cross-validation of co-occurrence patterns between TRs and TISs in TR Categories 1–4. Each histogram shows the number of homologous vs. non-homologous TISs, based on two different weighing methods (vectors), as follows: category 1 (a), category 2 (b), category 3 (c), and category 4 (d) (Please see text for the description of TR categories 1 to 4). HS-TR = human-specific tandem repeat, NHS-TR = non-human-specific tandem repeat, TIS = translation initiation site

the annotated transcript IDs were retrieved, and finally, for each transcript, the coding sequence, the TIS, the upstream flanking sequence of the TIS, and the protein sequence were retrieved.

All collected data was stored in a MySQL database which is accessible at https://figshare.com/search?q=10.6084%2Fm9.figshare.15405267 .

A candidate sequence was considered a TR if it complied with the following four rules: (1) for mononucleotide cores, the number of repeats should be ≥6. (2) for 2–9 bp cores, the number of repeats should be ≥3. (3) for other core lengths, the number of repeats should be ≥2. (4) TRs of the same core sequence should not overlap if they were in the same upstream flanking sequence.

We categorized the TRs based on the core lengths as follows: Category 1: 1–6 bp, Category 2: 7–9 bp, Category 3: 10–15 bp, and Category 4: ≥16 bp. This was an arbitrary classification to allow for possible differential effect of various core length ranges in evolutionary and biological terms.

Retrieval of data across species

Using the enhanced query (Additional Table 5) form on the Biomart Ensembl tool along with the RESTful API tools, a Java package was developed to retrieve, store, and analyze the data and information. The source codes and the Java package are available at: https://github.com/Yasilis/STRsMiner-JavaPackage_PaperSubmission/tree/develop .

Identification of human-specific TRs

The 120 bp upstream flanking sequence of TISs of all annotated protein-coding transcripts of protein-coding genes were screened in 84 species for the presence of TRs in four categories based on the TR core length. The data obtained on the human TRs was compared to those of

Table 2 Example of human genes (represented by *gene symbol*), which contain human-specific TRs

No.	Gene Symbol	No.	Gene Symbol	No.	Gene Symbol	No.	Gene Symbol
1	*ACSL6*	28	*DMPK*	55	*KRT23*	82	*PHF8*
2	*ADAM22*	29	*DOK6*	56	*KRT73*	83	*PLEC*
3	*ADSSL1*	30	*EFHC1*	57	*KRT8*	84	*PPP1CC*
4	*AKAP7*	31	*EIF3K*	58	*L3MBTL1*	85	*PPP1R14A*
5	*ARHGAP42*	32	*EIF5AL1*	59	*LCAT*	86	*PRIMA1*
6	*ASIC1*	33	*ELMO1*	60	*LMNA*	87	*PTBP1*
7	*ASRGL1*	34	*ENSG00000258947*	61	*MBNL1*	88	*REG1B*
8	*ATXN10*	35	*EPB41L4B*	62	*MPRIP*	89	*RYR1*
9	*C11orf63*	36	*EXTL3*	63	*MTDH*	90	*RYR3*
10	*C19orf12*	37	*FAM101B*	64	*MYH2*	91	*SCIN*
11	*CACNA1A*	38	*FMNL3*	65	*NEK3*	92	*SERHL2*
12	*CACNA1F*	39	*FOXK1*	66	*NOL3*	93	*SERPINB6*
13	*CACNA1G*	40	*FOXP1*	67	*OBSCN*	94	*SIPA1L3*
14	*CAPNS2*	41	*GABRB2*	68	*OLIG1*	95	*SLC25A27*
15	*CDK16*	42	*GDF11*	69	*PAMR1*	96	*SLC4A1*
16	*CELF4*	43	*GSK3A*	70	*PANK2*	97	*SLC6A8*
17	*CELF6*	44	*GSTM2*	71	*PCDH7*	98	*SLIT2*
18	*CEP55*	45	*HCN2*	72	*PCDHA10*	99	*SPEG*
19	*CERCAM*	46	*HDAC4*	73	*PCDHA12*	100	*SYN1*
20	*CKB*	47	*HDAC8*	74	*PCDHA13*	101	*SYNGAP1*
21	*CLIP2*	48	*HRC*	75	*PCDHA7*	102	*TCF3*
22	*COL3A1*	49	*INPP5K*	76	*PCDHB14*	103	*TMEM132A*
23	*COPRS*	50	*ITSN1*	77	*PCDHB5*	104	*TMEM59L*
24	*CRIPT*	51	*KCNA2*	78	*PCDHB6*	105	*TRNP1*
25	*CROCC*	52	*KCNC1*	79	*PCDHB9*	106	*TTN*
26	*DAO*	53	*KIAA1191*	80	*PCDHGC4*	107	*ZFHX3*
27	*DCTN2*	54	*KRT10*	81	*PDLIM4*		

other species, and the TRs which were specific to human were identified.

To identify human-specific TRs, in the first step, the selected genes of all species were grouped based on gene name. Therefore, all homologous genes, consisting of orthologous and paralogous genes, were placed in one group. In each group, all the TRs located in the upstream flanking sequence of every transcript were extracted. In the next step, the extracted TRs were grouped and specified according to the species. All the TRs that were detected in more than one species were removed. The remaining TRs belonged to only one species and were specific to that species. Subsequently, we identified the human-specific TRs for a specific gene name by selecting the human species. This process was repeated for each group of genes and the results were aggregated together to identify all the TRs which were specific and non-specific in reference to human.

Evaluation of TIS homology

Identifying the degree of homology between two transcripts requires assigning a weight value to each position of the sequence. Weighted homology scoring was performed in two different weight settings, as weighing vectors W_1 (originally used by our group for studying a link between STRs and TIS selection) [33] and W_2, which can

(See figure on next page.)

Fig. 5 Protein homology check of TISs flanked by human-specific and non-specific TRs. Every chart shows the distribution of similarity abundance between human proteins and three species, mouse, macaque, and chimpanzee, in the same gene. For each panel, the first row shows the distribution that was constructed by BLASTing human proteins, TISs of which were flanked by human-specific TRs. Similarly, the second row of each panel shows the distribution that was constructed by BLASTing human proteins, TISs of which were flanked by non-human-specific TRs. The Needleman Wunsch algorithm (upper panel) was used as a complementary measure to our two weighing methods (methods 1 and 2). In each method, we detected a significant difference in the distribution. TIS = translation initiation site, TR = tandem repeat

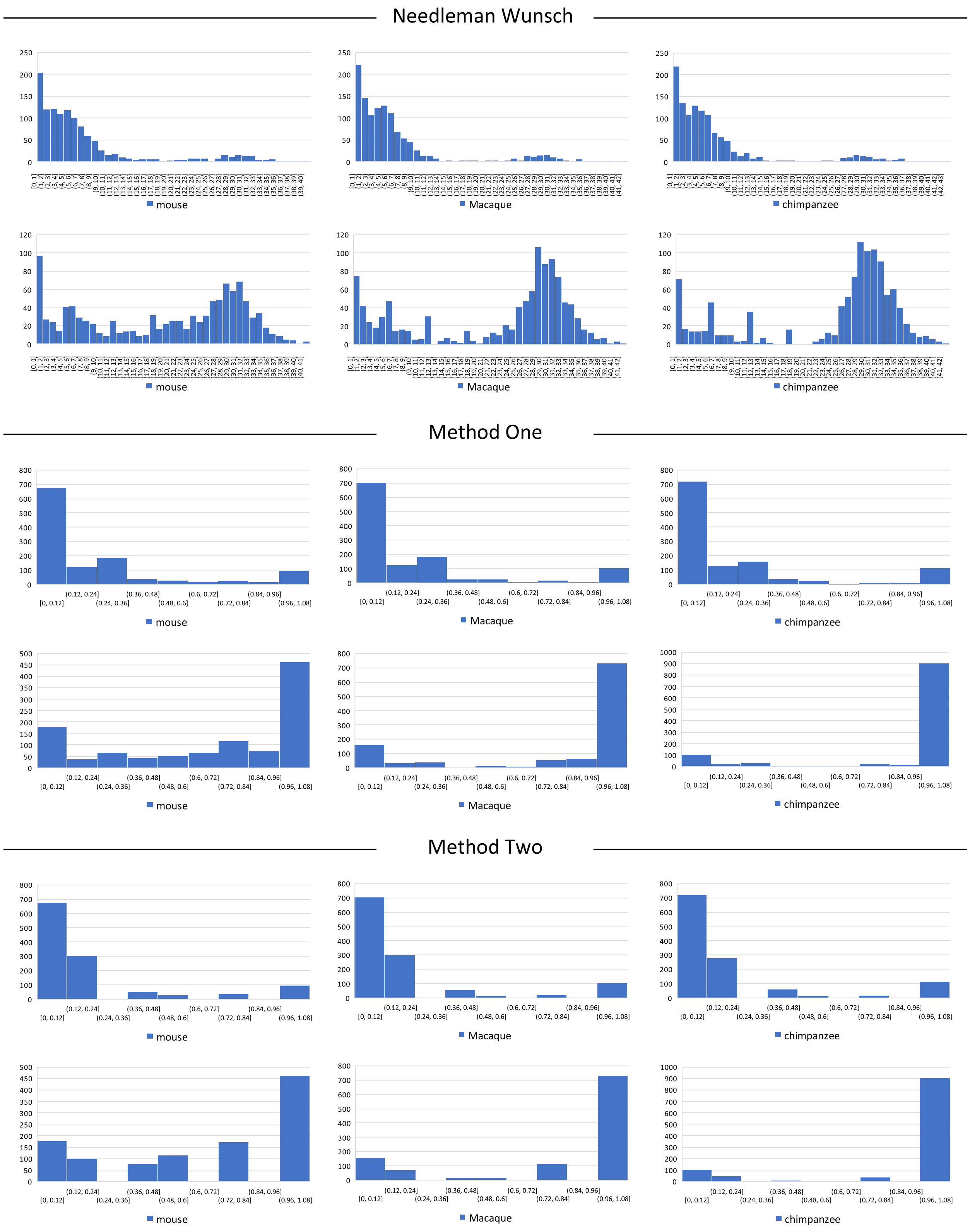

Fig. 5 (See legend on previous page.)

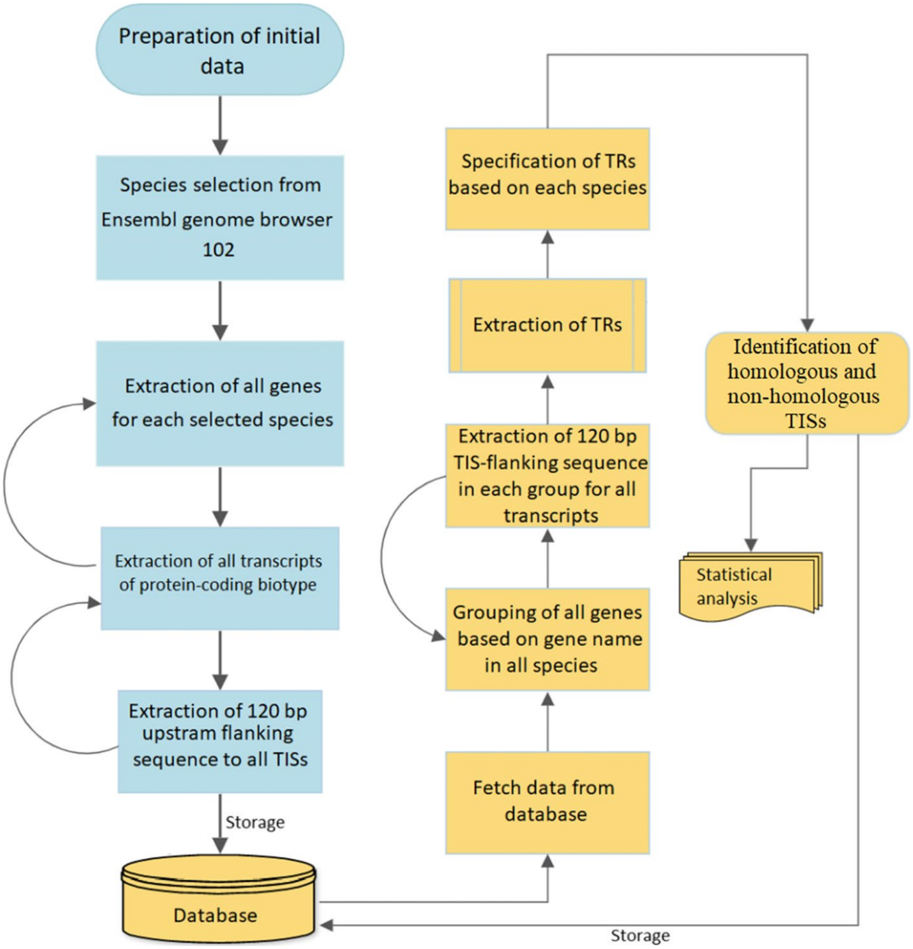
Preparation of initial data
Species selection from Ensembl genome browser 102
Extraction of all genes for each selected species
Extraction of all transcripts of protein-coding biotype
Extraction of 120 bp upstram flanking sequence to all TISs
Storage
Database
Fetch data from database
Grouping of all genes based on gene name in all species
Extraction of 120 bp TIS-flanking sequence in each group for all transcripts
Extraction of TRs
Specification of TRs based on each species
Identification of homologous and non-homologous TISs
Statistical analysis
Storage

Fig. 6 Scheme representing the steps taken for data collection and analysis

be distinguished by $k = \{1, 2\}$. These two weighing vectors are defined as follow (Eq. 1, 2):

$$W_1 = \left\{0, 25, 25, 25, 12.5, 12.5\right\} \quad (1)$$

$$W_2 = \left\{0, 20, 20, 20, 20, 20\right\} \quad (2)$$

If M is the first methionine amino acid of the two peptide sequences (position of 0 in the two weighing vectors), for all next five successive positions represented by i in the formula (Eq. 9), we defined five weight coefficients $w_{k,1}$ to $w_{k,5}$, observed in the W_k vector.

Homology of the first five amino acids (excluding the initial methionine), and, therefore the TIS, was inferred based on the value of pair-wise similarity scoring between human, as reference, and other species. A similarity of ≥50% was considered "homology". This threshold was achieved following BLASTing three thousand random pair-wise similarity checks of the initial five amino acids of randomly selected proteins as previously described [33].

Scoring human-specific and non-specific TR co-occurrences with homologous and non-homologous TISs

In both weighing methods, the initial five amino acid sequence (excluding the initial methionine) of the human TISs that were flanked by human-specific and non-specific TRs were BLASTed against all the initial five amino acids (excluding the initial methionine) of the orthologous/paralogous genes in the remaining 83 species. The above was aimed at comparing the number of events in which human-specific and non-specific TRs co-occurred with homologous and non-homologous (TISs) in reference to human. For computing the number of homologous and non-homologous TISs, we needed to consider a number of assumptions. We defined G as the set of all human protein coding genes. Therefore, g denoted a gene that belonged to the G set (Eq. 3).

$$G = \{g | g\ is\ a\ human\ protein\ coding\ gene\} \tag{3}$$

We also defined $T_H(g)$ and $T_{\overline{H}}(g)$ as the set of all annotated transcripts in a gene g, which belonged to human and other species, respectively (Eqs. 4 and 5).

$$T_H(g) = \left\{ t \mid \begin{array}{c} t \text{ was a human protein coding transcript which} \\ \text{belonged to the gene}, g \end{array} \right\} \tag{4}$$

$$T_{\overline{H}}(g) = \left\{ t \mid \begin{array}{c} t \text{ was a protein coding transcript which belonged} \\ \text{to the gene}, g \text{ but, did not exist in human} \end{array} \right\} \tag{5}$$

Moreover, T^* denoted all filtered transcripts of T which had at least one human-.

specific TR at the 120 bp interval upstream of the TIS, while, T^+ denoted all filtered transcripts of T, which had at least one TR at the 120 bp interval upstream of the TIS.

The following formula was developed to measure the degree of similarity of two peptides in the two weighing settings (Eq. 6).

$$H_k = \sum_{g \epsilon G} \sum_{t_a \epsilon T^*_H(g)} \sum_{t_b \epsilon T^+_{\overline{H}}(g)} \Theta_k(t_a, t_b) \tag{6}$$

In this formula, Θ is a binary function that decides whether the transcripts are homologous or not, and $k = \{1, 2\}$ refer to each weight setting. If S function measures the similarity score, Θ can be defined as follow (Eq. 7):

$$\Theta_k(t_a, t_b) = \left\{ \begin{array}{c} 1, if\ S_k(t_a, t_b) \geq 50 \\ 0, o.w. \end{array} \right. \tag{7}$$

For calculating the similarity score, we used another binary function. We defined Φ as follows: (Eq. 8):

$$\Phi(x, y) = \left\{ \begin{array}{c} 1, if\ x = y \\ 0, o.w. \end{array} \right. \tag{8}$$

This function takes two amino acids as argument and returns 1 as output if they are the same, and zero if they are not the same. Therefore, $S(t_a, t_b)$ is defined by the following formula (Eq. 9):

$$S_k(t_a, t_b) = \sum_{i=2}^{6} w_{k,i} \Phi(P_i(t_a), P_i(t_b)) \tag{9}$$

In this function, the i^{th} amino acid in the sequence of the transcript t, is denoted by $P_i(t)$.

We replicated the comparisons in 10-fold cross-validation. In each-fold, genes with human non-specific TRs were randomly selected according to the number of genes in the group with human-specific TRs. This process was repeated for the two methods (two different weight vectors) and for each of the four categories of TRs. For each category and weighing method, the mean of the result of each round was calculated as a final result. Finally, the Fisher's exact test was run for each-fold (Additional Table 2).

Abbreviations

TIS: Translation initiation site; TR: Tandem repeat; STR: Short tandem repeat; ORF: Open reading frame; UTR: untranslated region; HS-TR: Human-specific TR; NHS-TR: Non-human-specific TR.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12863-022-01075-5.

Additional file 1 Additional Table 1. The number of genes, transcripts and extracted TRs for each species. The rows of the table are sorted from large to small, based on the ratio of the number of TRs to the number of genes and transcripts in each species.

Additional file 2 Additional Table 2. The number of events/co-occurrences of homologous and non-homologous TISs (in human as reference) with the two groups of human-specific and non-specific TRs and their *p*-values, calculated by Fisher's exact test in each method across TR categories 1, 2, 3 and 4.

Additional file 3 Additional Table 3. The list of all human genes and their Ensembl gene ID, which contained human-specific TRs in their TIS-flanking sequence for TR categories 1, 2, 3, and 4.

Additional file 4 Additional Table 4. The list of all human specific TRs and their abundance.

Additional file 5 Additional Table 5. The list of queries that were used to communicate with the Ensembl data repositories.

Acknowledgements

Not applicable.

Authors' contributions

A.M.A.M performed and analyzed the bioinformatics data. M.A. and H.O. contributed to data collection. K.K. and M.O. conceived, designed, and supervised the project. M.O. wrote the manuscript with input from all authors. The author(s) read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The datasets generated and analyzed during this study are available in the "figshare" repository, with the identifier "https://doi.org/10.6084/m9.figshare.15405267".

Also, other source code and software available in the GitHub repository. (https://github.com/Yasilis/STRsMiner-JavaPackage_PaperSubmission/tree/develop)

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interest.

Author details

[1] Laboratory of Complex Biological systems and Bioinformatics (CBB), Department of Bioinformatics, Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Tehran 1417614411, Iran. [2] Chemical Injuries Research

Center, Systems Biology and Poisonings Institute, Baqiyatallah University of Medical Sciences, Tehran, Tehran 1435916471, Iran. [3]School of Physics and Astronomy, University of St. Andrews, St. Andrews KY16 9SS, UK. [4]Iranian Research Center on Aging, University of Social Welfare and Rehabilitation Sciences, Tehran, Tehran 1985713871, Iran.

Received: 4 April 2022 Accepted: 18 July 2022
Published online: 27 July 2022

References

1. Sonenberg N, Hinnebusch AG. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell. 2009;136(4):731–45.
2. Gebauer F, Hentze MW. Molecular mechanisms of translational control. Nat Rev Mol Cell Biol. 2004;5(10):827–35.
3. Lee S, Liu B, Lee S, Huang S-X, Shen B, Qian S-B. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci. 2012;109(37):E2424–32.
4. Na CH, Barbhuiya MA, Kim M-S, Verbruggen S, Eacker SM, Pletnikova O, et al. Discovery of noncanonical translation initiation sites through mass spectrometric analysis of protein N termini. Genome Res. 2018;28(1):25–36.
5. Andreev DE, O'Connor PB, Loughran G, Dmitriev SE, Baranov PV, Shatsky IN. Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. Nucleic Acids Res. 2017;45(2):513–26.
6. Studtmann K, Ölschläger-Schütt J, Buck F, Richter D, Sala C, Bockmann J, et al. A non-canonical initiation site is required for efficient translation of the dendritically localized Shank1 mRNA. PLoS One. 2014;9(2):e88518.
7. Fukushima M, Tomita T, Janoshazi A, Putney JW. Alternative translation initiation gives rise to two isoforms of Orai1 with distinct plasma membrane mobilities. J Cell Sci. 2012;125(Pt 18):4354–61.
8. Bazykin GA, Kochetov AV. Alternative translation start sites are conserved in eukaryotic genomes. Nucleic Acids Res. 2011;39(2):567–77.
9. Xu C, Zhang J. Mammalian alternative translation initiation is mostly nonadaptive. Mol Biol Evol. 2020;37(7):2015–28.
10. Boersma S, Khuperkar D, Verhagen BMP, Sonneveld S, Grimm JB, Lavis LD, et al. Multi-color single-molecule imaging uncovers extensive heterogeneity in mRNA decoding. Cell. 2019;178(2):458–472 e419.
11. Li JJ, Chew G-L, Biggin MD. Quantitative principles of cis-translational control by general mRNA sequence features in eukaryotes. Genome Biol. 2019;20(1):1–24.
12. Martinez-Salas E, Lozano G, Fernandez-Chamorro J, Francisco-Velilla R, Galan A, Diaz R. RNA-binding proteins impacting on internal initiation of translation. Int J Mol Sci. 2013;14(11):21705–26.
13. Cenik C, Cenik ES, Byeon GW, Grubert F, Candille SI, Spacek D, et al. Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. Genome Res. 2015;25(11):1610–21.
14. Babendure JR, Babendure JL, Ding J-H, Tsien RY. Control of mammalian translation by mRNA structure near caps. Rna. 2006;12(5):851–61.
15. Master A, Wójcicka A, Giżewska K, Popławski P, Williams GR, Nauman A. A novel method for gene-specific enhancement of protein translation by targeting 5'UTRs of selected tumor suppressors. PLoS One. 2016;11(5):e0155359.
16. Jagodnik J, Chiaruttini C, Guillier M. Stem-loop structures within mRNA coding sequences activate translation initiation and mediate control by small regulatory RNAs. Mol Cell. 2017;68(1):158–170. e153.
17. Kochetov AV, Allmer J, Klimenko AI, Zuraev BS, Matushkin YG, Lashin SA. AltORFev facilitates the prediction of alternative open reading frames in eukaryotic mRNAs. Bioinformatics. 2017;33(6):923–5.
18. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet. 2018;19(5):286–98.
19. Afshar H, Adelirad F, Kowsari A, Kalhor N, Delbari A, Najafipour R, et al. Natural selection at the NHLH2 core promoter exceptionally long CA-repeat in human and disease-only genotypes in late-onset neurocognitive disorder. Gerontology. 2020;66(5):514–22.
20. Press MO, McCoy RC, Hall AN, Akey JM, Queitsch C. Massive variation of short tandem repeats with functional consequences across strains of Arabidopsis thaliana. Genome Res. 2018;28(8):1169–78.
21. Bagshaw ATM. Functional mechanisms of microsatellite DNA in eukaryotic genomes. Genome Biol Evol. 2017;9(9):2428–43.
22. Abe H, Gemmell NJ. Evolutionary footprints of short tandem repeats in avian promoters. Sci Rep. 2016;6:19421.
23. Ohadi M, Valipour E, Ghadimi-Haddadan S, Namdar-Aligoodarzi P, Bagheri A, Kowsari A, et al. Core promoter short tandem repeats as evolutionary switch codes for primate speciation. Am J Primatol. 2015;77(1):34–43.
24. Mohammadparast S, Bayat H, Biglarian A, Ohadi M. Exceptional expansion and conservation of a CT-repeat complex in the core promoter of PAXBP1 in primates. Am J Primatol. 2014;76(8):747–56.
25. Rovozzo R, Korza G, Baker MW, Li M, Bhattacharyya A, Barbarese E, et al. CGG repeats in the 5'UTR of FMR1 RNA regulate translation of other RNAs localized in the same RNA granules. PLoS One. 2016;11(12):e0168204.
26. Todur SP, Ashavaid TF. Association of Sp1 tandem repeat polymorphism of ALOX5 with coronary artery disease in Indian subjects. Clin Transl Sci. 2012;5(5):408–11.
27. Shirokikh NE, Spirin AS. Poly(a) leader of eukaryotic mRNA bypasses the dependence of translation on initiation factors. Proc Natl Acad Sci U S A. 2008;105(31):10738–43.
28. Usdin K. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res. 2008;18(7):1011–9.
29. Kumari S, Bugaut A, Huppert JL, Balasubramanian S. An RNA G-quadruplex in the 5′ UTR of the NRAS proto-oncogene modulates translation. Nat Chem Biol. 2007;3(4):218–21.
30. Leppek K, Das R, Barna M. Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. Nat Rev Mol Cell Biol. 2018;19(3):158–74.
31. Krauß S, Griesche N, Jastrzebska E, Chen C, Rutschow D, Achmüller C, et al. Translation of HTT mRNA with expanded CAG repeats is regulated by the MID1–PP2A protein complex. Nat Commun. 2013;4(1):1–9.
32. Glineburg MR, Todd PK, Charlet-Berguerand N, Sellier C. Repeat-associated non-AUG (RAN) translation and other molecular mechanisms in fragile X tremor Ataxia syndrome. Brain Res. 2018;1693:43–54.
33. Arabfard M, Kavousi K, Delbari A, Ohadi M. Link between short tandem repeats and translation initiation site selection. Human genomics. 2018;12(1):1–11.
34. Thierry-Mieg D, Thierry-Mieg J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. Genome Biol. 2006;7(1):1–14.
35. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48(3):443–53.
36. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, et al. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 2019;47(W1):W636–41.
37. Pearson WR. An introduction to sequence similarity ("homology") searching. *Curr Protoc Bioinformatics*. 2013; Chapter 3:Unit3 1.
38. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database. 2011;2011.
39. Georgakopoulos-Soares I, Mouratidis I, Parada GE, Matharu N, Hemberg M, Ahituv N. Asymmetron: a toolkit for the identification of strand asymmetry patterns in biological sequences. Nucleic Acids Res. 2021;49(1):e4.
40. Ezra SC, Tuller T. Modeling the effect of rRNA-mRNA interactions and mRNA folding on mRNA translation in chloroplasts. Computational and structural Biotechnol J. 2022.
41. Ran F, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the CRISPR-Cas9 system. Nat Protoc. 2013;8(11):2281–308.
42. Gregorio NE, Levine MZ, Oza JP. A user's guide to cell-free protein synthesis. Methods and protocols. 2019;2(1):24.
43. Hammerling MJ, Krüger A, Jewett MC. Strategies for in vitro engineering of the translation machinery. Nucleic Acids Res. 2020;48(3):1068–83.
44. Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4(8):1184.
45. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21(16):3439–40.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.