Tandem repeats ubiquitously flank and contribute to translation initiation sites

Ali M.A. Maddi

University of Tehran

Kaveh Kavousi

University of Tehran

Masoud Arabfard

Baqiyatallah University of Medical Sciences

Hamid Ohadi

University of St. Andrews

Mina Ohadi (✉ mi.ohadi@uswr.ac.ir)

University of Social Welfare and Rehabilitation Sciences

Research Article

Keywords: translation initiation site, tandem repeat, genome-scale, TIS selection, homology

Posted Date: May 11th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1380780/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Tandem repeats ubiquitously flank and contribute to translation initiation sites.

Ali M.A. Maddi¹, Kaveh Kavousi^{1,*}, Masoud Arabfard², Hamid Ohadi³ and Mina Ohadi^{4,*}

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, United Kingdom.

⁴ Iranian Research Center on Aging, University of Social Welfare and Rehabilitation Sciences, Tehran, Tehran, 1985713871, Iran

To whom correspondence should be addressed. Tel: +98 21 71732803; Email: mi.ohadi@uswr.ac.ir

***Joint correspondence**

Email Addresses:

Ali M.A. Maddi: ali.maddi@ut.ac.ir

Kaveh Kavousi: kkavousi@ut.ac.ir

Masoud Arabfard: arabfard@gmail.com

Hamid Ohadi: hamid.ohadi@gmail.com

Mina Ohadi: mi.ohadi@uswr.ac.ir

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 elusive for the most part. Hence, here we employed the TIS homology concept to study the co-occurrence patterns of all categories of TRs 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. Two different weighing vectors were employed to assign TIS homology, and the results were assessed in 10-fold validation. **Results:** On average, every TIS was flanked by 1.19 TRs of various categories within the 120 bp upstream sequence. We detected statistically significant enrichment of non-homologous TISs co-occurring with human-specific TRs. On the contrary, homologous TISs co-occurred significantly with non-human-specific TRs. 2,991 human genes had at least one transcript flanked by a human-specific TR in their upstream flanking region, and nervous system development was the top enriched ontology term across those genes. Text mining of a number of the identified genes such as *MYH2*, *TTN*, *SLC6A8*, *CACNA1A*, and *EIF5AL1* yielded predominant expression and functions in the human brain and skeletal muscle. **Conclusion:** We conclude that TRs are abundant *cis* elements in the upstream sequences of TISs across species, and there may be a link between all categories of TRs and TIS selection. The potential biological consequences of this link are discussed in human as the reference sequence in this study.

Keywords: translation initiation site; tandem repeat; genome-scale; TIS selection; homology

INTRODUCTION

Translational regulation can be global or mRNA 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].

Whereas lack of sufficient knowledge has led to the conclusion that TISs are stochastic for the most part, the probability of using a particular TIS differs among mRNA molecules, and can be dynamically regulated over time[9]. Selection of TISs and the level of translation and protein synthesis depend on the *cis* regulatory elements in the mRNA sequence and its secondary structure such as the formation of hair-pins, stem loop, and thermal stability [10-15]. In fact, the ribosomal machinery has the potential to scan and use several ORFs at a particular mRNA species[16].

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[17-23], 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' UTR, short tandem repeats (STRs) (core units of 1-6 bp) can modulate translation, the effect of which has biological and pathological implications [24-28]. Abnormal STR expansions impact TIS selection in a number of neurological disorders [29, 30].

Based on a TIS homology approach, we recently reported a link between STRs and TIS selection[31]. Here, we extended our weighing methods, and developed a comprehensive software package, to study a possible link between all categories of TRs and TIS selection across 84 species.

RESULTS AND DISCUSSION

TRs are ubiquitous *cis* elements flanking TISs.

A total of 1,611,368 protein-coding genes and 2,730,515 transcripts were investigated across the 84 selected species, which resulted in the extraction of 3,283,771 TRs, 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, at 4.11 and 4.38, respectively, belonged to human. Human ranked 59th among 84 species in respect of the TR/transcript ratio (Additional Table 1). Out of the 93,706 identified protein-coding transcripts in the human genome, there were 50,169 transcripts which had at least one TR in the upstream flanking sequence (53.54% of protein-coding transcripts). At a similarly high rate, from the 22,791 identified protein-coding genes in the human genome, there were 15,256 genes which had at least one transcript flanked by a TR in its upstream flanking region (66.94% of human protein-coding genes). From the 2,850 different types of TRs in the human genome, there were 1,504 different types of TRs that were human-specific (52.77% of the TRs).

Across TR categories 1-4, we detected 660, 101, 339 and 404 different types of human-specific TRs, respectively, the top most abundant of which are represented in Table 1.

Link between TRs and TIS selection

We employed two weighing settings for designating homologous vs. non-homologous TISs, one of which was similar to our previous approach [31]. In both settings, there was significant co-occurrence

Fig. 1: Abundance interval of the genes, transcripts, and TRs to each other.

of human-specific TRs with non-homologous TISs, and non-human-specific TRs with homologous TISs (Fig. 2). The results were replicated in 10-fold validation (Fig. 3). In addition to the weighing vectors used, we employed the Needleman Wunsch algorithm [32] to check for the robustness of the proposed link. TIS homology was checked for the proteins encoded by the homologous genes between human and three species, mouse, macaque, and chimpanzee. A dramatically lower homology was observed for human proteins linked by human-specific TRs (Fig. 4). Similarity calculation between human proteins and the three other species was performed by RESTful API at: https://www.ebi.ac.uk/Tools/psa/emboss_needle [33].

Evolutionary and biological implications

Of the 15,256 human genes which had at least one transcript flanked by a TR in their upstream flanking region, there were 2,991 genes which had at least one transcript flanked by a human-specific TR in their upstream flanking region (Additional Tables 2 & 3). Text mining of a number of those

Fig. 2: Average of 10 experiments performed to evaluate the link between TRs and TISs in each of the four TR categories. Each histogram shows the number of homologous TISs against non-homologous TISs based on two different weighing methods.

Fig. 3: 10-fold cross-validation for evaluating the link between TRs and TISs in Categories 1-4. Each histogram shows the number of homologous TISs vs. nonhomologous TISs, based on two different weighing methods, as follows: category 1 (a), category 2 (b), category 3 (c), and category 4 (d) (Please see text for the description of categories 1 to 4).

Needleman Wunsch

Method One

Method Two

Fig. 4: 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, which were produced by human-specific TR (HS-TR) genes. Similarly, the second row of each panel shows the distribution that was constructed by BLASTing human proteins which were produced by non-human-specific TR (NHS-TR) genes. The Needleman Wunsch algorithm (upper panel) was used as a confirmatory measure to our two weighing methods (middle and lower panels). In each method, we detected a significant difference in the distribution of HS-TR genes vs. NHS-TR genes.

- 143 Our findings provide prime evidence of a link between all categories of TRs and TIS selection,
 144 mechanisms of which are virtually unknown at this time. Our approach was based on homology
 145 search, which reliably identifies "homologous" proteins by detecting excess similarity [40]. By

Fig. 5: Enrichment analysis of sample genes in the human-specific TR (HS-TR) compartment. Development of the central nervous system and calcium signaling pathways received the lowest p-values for enrichment significance in gene ontology a) and pathway b) analyses.

searching identical gene names across the selected species, our approach encompassed homologous genes. While homologous genes encompass orthologous and paralogous genes, the scope of our study was not to differentiate between the two latter categories, but rather, to compare the patterns of TIS co-occurrences with TRs in homologous genes, as an entity. This approach was executed using two weighing vectors and a confirmatory algorithm, which consistently supported the link.

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 [41]. It remains to be clarified how this intriguingly abundant reservoir of regulatory elements contributes to TIS selection across species.

Conclusion

We conclude and propose that TRs are abundant *cis* elements that flank TIS sequences, and contribute to TIS selection at the trans-species level. Future functional analyses are warranted to investigate the relevant downstream mechanisms involved.

MATERIAL 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>). 84 species were selected, which encompassed orders of vertebrates and non-vertebrates (Fig. 6). 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 [42, 43] 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 checked across the selected species to cover orthologous and homologous 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 the annotated transcript IDs were retrieved, and finally, for each transcript its coding DNA sequence, upstream flanking region and 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 - In the case of mono nucleotide cores, the number of repeats should be ≥ 6 . 2 - In the case of 2-9 bp cores, the number of repeats should be ≥ 3 . 3 - For other core lengths, the number of repetition of cores should be ≥ 2 . 4 - TRs of the same core sequence should not overlap if they were in the same upstream flanking region.

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.

194 **Retrieval of data across species**

195 Using the enhanced query (Additional Table 4) form on the Biomart Ensembl tool along with the
196 RESTful API tools, a Java package was developed to retrieve, store, and analyze the data and

197 information. The source codes and the Java package are available at:
 198 https://github.com/Yasilis/STRsMiner-JavaPackage_PaperSubmission/tree/develop .

Fig. 6: 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 Supplementary Table 3.

Identification of Human-specific TRs

The TIS-flanking upstream 120 bp 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 core length. The data obtained on the human TRs was compared to those of other species, and the TRs which were specific to human were identified. The selected genes were clustered based on the gene names (homologous genes were placed in a cluster). All TRs of each cluster were extracted and categorized based on the species. In the next step, in each cluster, only TRs that were specific to human (not detected in other species) were retained and set as reference.

Evaluation of TIS homology

Identifying the degree of homology between two transcripts requires assigning a weight value to each position of the sequences. Weighted homology scoring was performed in two different weight settings, as weighing vectors W_1 and W_2 , which can be distinguished by $k = \{1,2\}$. These two weighing vectors are defined as follow (Eq. 1, 2):

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

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

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 five amino acids and, therefore, the TIS was inferred based on the value of similarity scoring, in which a similarity of $\geq 50\%$ 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 [31].

Scoring human-specific and non-specific TR co-occurrences with 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 TRs and non-human-specific TRs were BLASTed against all the initial five amino acids (excluding the initial methionine) of the homologous genes in the remaining 83 species. The above was aimed at comparing the number of events in which human-specific and non-specific TRs occurred with homologous and non-homologous (TISs). 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).

$$G = \{g | g \text{ is a human protein coding gene}\} \quad (3)$$

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

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

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

Moreover, T^* denoted all filtered transcripts of T which had at least one human-specific TR at the 120 bp genomic DNA interval upstream of the TIS, while, T^+ denoted all filtered transcripts of T , which had at least one TR at the 120 bp genomic DNA 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).

$$H_k = \sum_{g \in G} \sum_{t_a \in T_H(g)} \sum_{t_b \in T_{\bar{H}}(g)} \Theta_k(t_a, t_b) \quad (6)$$

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):

$$\Theta_k(t_a, t_b) = \begin{cases} 1, & \text{if } S_k(t_a, t_b) \geq 50 \\ 0, & \text{o. w.} \end{cases} \quad (7)$$

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

$$\Phi(x, y) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{o. w.} \end{cases} \quad (8)$$

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):

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

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 in the human non-specific TR group were randomly selected according to the number of genes in the human-specific TRs group. 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 exact test was run for each-fold (Additional Table 5).

REFERENCES

1. Sonenberg N, Hinnebusch AG: **Regulation of translation initiation in eukaryotes: mechanisms and biological targets.** *Cell* 2009, **136**(4):731-745.
2. Gebauer F, Hentze MW: **Molecular mechanisms of translational control.** *Nature reviews Molecular cell biology* 2004, **5**(10):827-835.
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.** *Proceedings of the National Academy of Sciences* 2012, **109**(37):E2424-E2432.
4. Na CH, Barbhuiya MA, Kim M-S, Verbruggen S, Eacker SM, Pletnikova O, Troncoso JC, Halushka MK, Menschaert G, Overall CM: **Discovery of noncanonical translation initiation sites through mass spectrometric analysis of protein N termini.** *Genome research* 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 research* 2017, **45**(2):513-526.
6. Studtmann K, Ölschläger-Schütt J, Buck F, Richter D, Sala C, Bockmann J, Kindler S, Kreienkamp H-J: **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):4354-4361.
8. Bazykin GA, Kochetov AV: **Alternative translation start sites are conserved in eukaryotic genomes.** *Nucleic acids research* 2011, **39**(2):567-577.
9. Boersma S, Khuperkar D, Verhagen BMP, Sonneveld S, Grimm JB, Lavis LD, Tanenbaum ME: **Multi-Color Single-Molecule Imaging Uncovers Extensive Heterogeneity in mRNA Decoding.** *Cell* 2019, **178**(2):458-472 e419.
10. Li JJ, Chew G-L, Biggin MD: **Quantitative principles of cis-translational control by general mRNA sequence features in eukaryotes.** *Genome biology* 2019, **20**(1):1-24.
11. 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):21705-21726.

12. Cenik C, Cenik ES, Byeon GW, Grubert F, Candille SI, Spacek D, Alsallakh B, Tilgner H, Araya CL, Tang H *et al*: **Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans.** *Genome Res* 2015, **25**(11):1610-1621.
13. Babendure JR, Babendure JL, Ding J-H, Tsien RY: **Control of mammalian translation by mRNA structure near caps.** *Rna* 2006, **12**(5):851-861.
14. 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.
15. Jagodnik J, Chiaruttini C, Guillier M: **Stem-loop structures within mRNA coding sequences activate translation initiation and mediate control by small regulatory RNAs.** *Molecular cell* 2017, **68**(1):158-170. e153.
16. 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):923-925.
17. Hannan AJ: **Tandem repeats mediating genetic plasticity in health and disease.** *Nat Rev Genet* 2018, **19**(5):286-298.
18. Afshar H, Adelirad F, Kowsari A, Kalhor N, Delbari A, Najafipour R, Foroughan M, Bozorgmehr A, Khamse S, Nazariapanah N: **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):514-522.
19. 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):1169-1178.
20. Bagshaw ATM: **Functional Mechanisms of Microsatellite DNA in Eukaryotic Genomes.** *Genome Biol Evol* 2017, **9**(9):2428-2443.
21. Abe H, Gemmell NJ: **Evolutionary Footprints of Short Tandem Repeats in Avian Promoters.** *Sci Rep* 2016, **6**:19421.
22. Ohadi M, Valipour E, Ghadimi-Haddadan S, Namdar-Aligoodarzi P, Bagheri A, Kowsari A, Rezazadeh M, Darvish H, Kazeminasab S: **Core promoter short tandem repeats as evolutionary switch codes for primate speciation.** *American journal of primatology* 2015, **77**(1):34-43.
23. 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):747-756.
24. Rovozzo R, Korza G, Baker MW, Li M, Bhattacharyya A, Barbarese E, Carson JH: **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.
25. 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):408-411.
26. 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):10738-10743.
27. Usdin K: **The biological effects of simple tandem repeats: lessons from the repeat expansion diseases.** *Genome Res* 2008, **18**(7):1011-1019.
28. Kumari S, Bugaut A, Huppert JL, Balasubramanian S: **An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation.** *Nature chemical biology* 2007, **3**(4):218-221.
29. Krauß S, Griesche N, Jastrzebska E, Chen C, Rutschow D, Achmüller C, Dorn S, Boesch SM, Lalowski M, Wanker E: **Translation of HTT mRNA with expanded CAG repeats is regulated by the MID1–PP2A protein complex.** *Nature communications* 2013, **4**(1):1-9.

30. 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 research* 2018, **1693**:43-54.
31. 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.
32. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins.** *Journal of molecular biology* 1970, **48**(3):443-453.
33. Madeira F, Park YM, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey AR, Potter SC, Finn RD: **The EMBL-EBI search and sequence analysis tools APIs in 2019.** *Nucleic acids research* 2019, **47**(W1):W636-W641.
34. Thierry-Mieg D, Thierry-Mieg J: **AceView: a comprehensive cDNA-supported gene and transcripts annotation.** *Genome biology* 2006, **7**(1):1-14.
35. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P: **STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets.** *Nucleic acids research* 2019, **47**(D1):D607-D613.
36. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, Clark NR, Ma’ayan A: **Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.** *BMC bioinformatics* 2013, **14**(1):1-14.
37. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A: **Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.** *Nucleic acids research* 2016, **44**(W1):W90-W97.
38. Xie Z, Bailey A, Kuleshov MV, Clarke DJ, Evangelista JE, Jenkins SL, Lachmann A, Wojciechowicz ML, Kropiwnicki E, Jagodnik KM: **Gene set knowledge discovery with Enrichr.** *Current protocols* 2021, **1**(3):e90.
39. Suzuki IK, Gacquer D, Van Heurck R, Kumar D, Wojno M, Bilheu A, Herpoel A, Lambert N, Cheron J, Polleux F: **Human-specific NOTCH2NL genes expand cortical neurogenesis through Delta/Notch regulation.** *Cell* 2018, **173**(6):1370-1384. e1316.
40. Pearson WR: **An introduction to sequence similarity ("homology") searching.** *Curr Protoc Bioinformatics* 2013, **Chapter 3**:Unit3 1.
41. 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 Research* 2021, **49**(1):e4-e4.
42. Durinck S, Spellman PT, Birney E, Huber W: **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.** *Nature protocols* 2009, **4**(8):1184.
43. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.** *Bioinformatics* 2005, **21**(16):3439-3440.

379 **Abbreviations**

380 TIS: Translation initiation site

381 TR: Tandem repeat

382 STR: Short tandem repeat

383 ORF: Open reading frame

384 UTR: untranslated region

385 HS-TR: Human-specific TR

386 NHS-TR: Non-human-specific TR

387

388

389

390 **Declarations**

391 **Ethics approval and consent to participate**

392 Not applicable

393

394 **Consent for publication**

395 Not applicable

396

397 **Availability of data and material**

398 The datasets generated and analyzed during this study are available in the “figshare” repository, with
399 the identifier “10.6084/m9.figshare.15405267”

400 Also, other source code and software available in the GitHub repository

401 (https://github.com/Yasilis/STRsMiner-JavaPackage_PaperSubmission/tree/develop)

402

403 **Competing interests**

404 The authors declare no competing interest.

405

406 **Funding**

407 Not applicable

408

409 **Authors' contributions**

410 A.M.A.M performed and analyzed the bioinformatics data. M.A. and H.O. contributed to data
411 collection. K.K. and M.O. conceived, designed, supervised the project, and wrote the manuscript with
412 input from all authors.

413

414 **Acknowledgements**

415 Not applicable

416

417

418

TABLE AND FIGURES LEGENDS

Table 1: The top most abundant human-specific TRs flanking TISs.

Table 2: Example of genes enriched in the brain and skeletal muscle, based on STRING (<https://string-db.org/>)[35].

Fig. 6: 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 Supplementary Table 3. Abundance interval of the genes, transcripts, and TRs to each other.

Fig. 7: Average of 10 experiments performed to evaluate the link between TRs and TISs in each of the four TR categories. Each histogram shows the number of homologous TISs against non-homologous TISs based on two different weighing methods.

Fig. 8: 10-fold cross-validation for evaluating the link between TRs and TISs in Categories 1-4. Each histogram shows the number of homologous TISs vs. nonhomologous TISs, based on two different weighing methods, as follows: category 1 (a), category 2 (b), category 3 (c), and category 4 (d).

Fig. 9: 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, which were produced by human-specific TR (HS-TR) genes. Similarly, the second row of each panel shows the distribution that was constructed by BLASTing human proteins which were produced by non-human-specific TR (NHS-TR) genes. The Needleman Wunsch algorithm (upper panel) was used as a confirmatory measure to our two weighing methods (middle and lower panels). In each method, we detected a significant difference in the distribution of HS-TR genes vs. NHS-TR genes.

Fig. 10: Enrichment analysis of sample genes in the human-specific TR (HS-TR) compartment. Development of the central nervous system and calcium signaling pathways received the lowest p-values for enrichment significance in gene ontology a) and pathway b) analyses.

Fig. 11: 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 Supplementary Table 3.

Additional Table 1: The number of genes, transcripts and extracted TRs for each species in our study. 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 Table 2: The list of all human genes and their Ensembl gene ID, which contained human-specific TRs in their TIS-flanking sequence for categories 1, 2, 3, and 4.

Additional Table 3: The list of all human specific TRs and their abundance.

Additional Table 4: The list of queries that were used to communicate with the Ensembl data repositories.

Additional Table 5: The number of homologous and non-homologous events in the two groups of human specific TRs and non-human specific TRs and their P-values, calculated by Fisher exact test in each method across categories 1, 2, 3 and 4.

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.

	TR	Core Length
Cat 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
Cat 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
	(AGCCACGG)3	9
Cat 3	(GTGTGTGTTT)2	10
	(ATTTTAAATT)2	11
	(AAAATAAATAA)2	11
	(TGGCGGCGGCGG)2	12
	(CCCAGCCCCA)2	10
	(CCCCGCCCCGCG)2	11
	(CGGGAGTGAGAG)2	12
	(AAGTGGGAACTGG)2	14
	(TTCATAGATGTTT)2	13
	(ATAGATGTTT)2	10
	(CCCCGCCCT)2	10
Cat 4	(CCCCGAGGTCTCCGCG)2	16
	(CCGGCGGTGTACCGAGAGACTGGCGT)2	25
	(ACCTGGAGGGCTGGGG)2	16
	(CCCTGCCCTGTCCTGTCCTGCCCTG)2	25
	(ACCCATCCCCACCTCCCT)3	18
	(CCCTGCCCTGTCCTGTCCTG)2	20
	(ACCCATCCCCACCTCCCT)3	18
	(CCCCACCTCCCTACCCAT)4	18
	(ACAGCGAGGTCGGCAGCGGCAGCGAGGTCGGCAGCGGC)2	38
	(TGAGTCGACAGGCCGAGGAGACAGTGAGTGCGCGCCC)2	36

(ACTCTCTCTTTCTCGGGCTGCAGGTGCACCAGGCCGTCC)2	41
--	----

Table 2. Example of human genes enriched in the brain and skeletal muscle, based on STRING (<https://string-db.org/>)[35].

<i>Gene Symbol</i>							
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	EIF5A1	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	CRIP1	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		

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [AdditionalTable1.xlsx](#)
- [AdditionalTable2.xlsx](#)
- [AdditionalTable3.xlsx](#)
- [AdditionalTable4.docx](#)
- [AdditionalTable5.docx](#)