


PRIMARY RESEARCH

Open Access



# Link between short tandem repeats and translation initiation site selection

Masoud Arabfard<sup>1,2</sup>, Kaveh Kavousi<sup>2\*</sup>, Ahmad Delbari<sup>3</sup> and Mina Ohadi<sup>3\*</sup> 

## Abstract

**Background:** Despite their vast biological implication, the relevance of short tandem repeats (STRs)/microsatellites to the protein-coding gene translation initiation sites (TISs) remains largely unknown.

**Methods:** We performed an Ensembl-based comparative genomics study of all annotated orthologous TIS-flanking sequences in human and 46 other species across vertebrates, on the genomic DNA and cDNA platforms (755,956 TISs), aimed at identifying human-specific STRs in this interval. The collected data were used to examine the hypothesis of a link between STRs and TISs. BLAST was used to compare the initial five amino acids (excluding the initial methionine), codons of which were flanked by STRs in human, with the initial five amino acids of all annotated proteins for the orthologous genes in other vertebrates (total of 5,314,979 pair-wise TIS comparisons on the genomic DNA and cDNA platforms) in order to compare the number of events in which human-specific and non-specific STRs occurred with homologous and non-homologous TISs (i.e.,  $\geq 50\%$  and  $< 50\%$  similarity of the five amino acids).

**Results:** We detected differential distribution of the human-specific STRs in comparison to the overall distribution of STRs on the genomic DNA and cDNA platforms (Mann Whitney  $U$  test  $p = 1.4 \times 10^{-11}$  and  $p < 7.9 \times 10^{-11}$ , respectively). We also found excess occurrence of non-homologous TISs with human-specific STRs and excess occurrence of homologous TISs with non-specific STRs on both platforms ( $p < 0.00001$ ).

**Conclusion:** We propose a link between STRs and TIS selection, based on the differential co-occurrence rate of human-specific STRs with non-homologous TISs and non-specific STRs with homologous TISs.

**Keywords:** Translation initiation site, Short tandem repeat, Genome-scale, Human-specific, Selection

## Introduction

An increasing number of human protein-coding genes are unraveled to consist of alternative translation initiation sites (TISs), which are selected based on complex and yet not fully known scanning mechanisms [1, 2]. The alternative TISs result in various protein structures and functions [3, 4]. Selection of TISs and the level of translation and protein synthesis depend partially on the *cis*-regulatory elements in the mRNA sequence and its secondary structure such as the formation of hair-pins and thermal stability [5–7]. Genomic DNA *cis*-elements

can also affect translation and TISs through various mechanisms (for a review see [8]).

One of the important and understudied *cis*-regulatory elements affecting translation is short tandem repeats (STRs)/microsatellites. In physiological terms, STRs can dramatically influence TIS and the amount of protein synthesis. Poly(A) tracts in the 5'-untranslated region (UTR) are important sites for translation regulation in yeast. These poly(A) tracts can interact with translation initiation factors or poly(A) binding proteins (PABP) to either increase or decrease translation efficiency. Pre-AUG A<sub>N</sub> can enhance internal ribosomal entry both in the presence of PABP and eIF-4G in *Saccharomyces cerevisiae* [9], and in the complete absence of PABP and eIF-4G [10]. Biased distribution of dinucleotide repeats is a known phenomenon in the region immediately upstream of the TISs in *Escherichia coli* [11]. In pathological instances,

\* Correspondence: [kkavousi@ut.ac.ir](mailto:kkavousi@ut.ac.ir); [mi.ohadi@uswr.ac.ir](mailto:mi.ohadi@uswr.ac.ir); [ohadi.mina@yahoo.com](mailto:ohadi.mina@yahoo.com)

<sup>2</sup>Laboratory of Complex Biological Systems and Bioinformatics (CBB), Department of Bioinformatics, Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran

<sup>3</sup>Iranian Research Center on Aging, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran

Full list of author information is available at the end of the article



expansion of STRs in the RNA structure results in toxic RNAs and non-AUG translation and the development of several human-specific neurological [12–14].

Genomic DNA STRs can affect TISs through their effect on alternative splicing and shuffling of novel ATG translation sites in novel exons [15]. Genome-scale findings of the evolutionary trend of a number of STRs have begun to unfold their implication in respect to speciation and species-specific characteristics/phenotypes [16–27]. The hypermutable nature of STRs and their large unascertained reservoir of functionality make them an ideal source of evolutionary adaptation, speciation, and disease [28–35]. In line with the above, recent reports indicate a role of repetitive sequences in the creation of new transcription start sites (TSSs) in human [36–39].

This research aimed to examine a possible link between STRs and TIS selection through studying the occurrence rate of TIS-flanking human-specific and non-specific STRs with homologous and non-homologous proteins.

## Methods

### Data collection

Forty-seven species encompassing major classes of vertebrates were selected, and in each species, the 120 bp upstream genomic DNA and cDNA sequences flanking all annotated protein-coding TISs ( $n = 755,956$ ) were downloaded based on the Ensembl database version 90 (<https://asia.ensembl.org>). The species studied are alphabetically listed as follows: anole lizard, armadillo, bush baby, cat, chicken, chimpanzee, cow, dog, dolphin, duck, ferret, fugu, gibbon, golden hamster, gorilla, guinea pig, hedgehog, human, horse, kangaroo rat, lamprey, lesser hedgehog tenrec, macaque, marmoset, megabat, microbat, mouse, mouse lemur, olive baboon, opossum, orangutan, pig, platypus, prairie vole, rabbit, rat, sheep, shrew, shrew mouse, squirrel, tarsier, Tasmanian devil, tree shrew, turkey, vervet-AGM, *Xenopus*, and zebrafish.

For each gene in each species, its Ensembl ID, all the annotated transcript IDs, the genomic DNA sequence, the cDNA, and the coding DNA sequence (CDS) were retrieved (the list of genes is available upon request). The genomic DNA, CDS, and the annotated cDNAs were downloaded using REST API from the Ensembl database. The first start codon for each transcript was determined using BLAST between the CDS and cDNA. The 120 bp genomic DNA and cDNA interval upstream of the start codon (ATG) were investigated for the presence of STRs of  $\geq 3$ -repeats (Additional file 1).

### Retrieval of gene IDs across species

Using the Enhanced REST API tools, a set of functions were developed to analyze genes and their transcripts information, including *func\_get\_ensemblID* and *func\_get\_TranscriptsID*. The genomic DNA, cDNA, and CDS

sequences of genes and their respective transcripts were obtained using *func\_get\_GenomicSequence*, *func\_get\_cDNASequence*, and *func\_get\_CDSSequence* functions.

### Identification of STRs in the human TIS-flanking genomic and cDNA intervals

A general method of finding human-specific and non-specific STRs ( $\geq 3$ -repeats in all classes of STRs, except the mononucleotide repeats, in which STRs of  $\geq 6$ -repeats were studied) for each individual gene was developed and applied as follows: the 120 bp genomic DNA and cDNA sequence upstream of the TISs of all annotated protein-coding gene transcripts was screened in human and 46 other species across vertebrates for the presence of STRs. A list of all STRs and their abundance was prepared for each gene in every species. The data obtained on the human STRs was compared to those of other species, and the term “human-specific” was applied to STRs that were not detected at  $\geq 3$ -repeats in any other species. Exceptionally, in the mononucleotide category, the threshold of repeats for “human-specificity” was set at  $> 6$ -repeats. The relevant pseudo-code for the identification of repeated substrings was used for STR identification (Additional file 2).

Mann-Whitney *U* test was used to compare the distribution of human-specific vs. the overall (specific and non-specific) STR distribution in human.

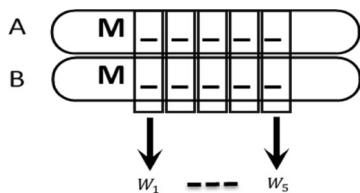
### TIS homology threshold estimation

Weighted homology scoring was performed, in which the initial five amino acids (excluding the initial methionine) of the human protein-coding TISs, codons of which were flanked by STRs, were BLASTed (compared using BLAST) against all initial protein-coding five amino acids annotated for the orthologous genes in 46 species across vertebrates [(3,872,779 pair-wise TIS comparisons on the genomic DNA platform (Ensembl 91) and 1,442,200 pair-wise TIS comparisons on the cDNA platform (Ensembl 92)]. The above was aimed at comparing the number of events in which human-specific and non-specific STRs occurred with homologous and non-homologous TISs.

The following equation was developed for the weighted scoring of homology (Eq. 1), where *A* refers to the five amino acid sequence (excluding the initial methionine M), codons of which were flanked by a STR at the genomic DNA or cDNA sequence, *j* refers to the gene, and *B* refers to all the transcripts of the same gene that contain the STR in other species.

If M is the first methionine amino acid of two sequences, for all five successive positions represented by *i* in the equation, we defined five weight coefficients  $W_1$  to  $W_5$  based on the importance of the amino acid position, i.e., proximity to the methionine starting codon, observed in the *W* vector. The degree of homology between the two

sequences  $A$  and  $B$  was calculated using function  $\phi$  for all five positions with the operations  $\sum_{i=1}^{L=5} W_i \phi(A_{jik}, B_{jik'})$ . We repeated this operation for  $k$  transcripts, where  $k$  stands for the number of transcripts in human.  $k'$  refers to all transcripts of the gene  $j$  in other species.



$$H_k^j = \sum_{i=1}^{L=5} W_i \phi(A_{jik}, B_{jik'}) \text{ for all } k \text{ and } k' \quad (1)$$

$$\phi(x, y) = \begin{cases} 1; & \text{if } x \neq y \\ 0; & \text{otherwise} \end{cases}$$

$$W = \{25, 25, 25, 12.5, 12.5\}$$

Homology of the five amino acids, and therefore the TISs, was inferred based on the %similarity scoring. We validated the homology threshold by measuring the %similarity of 3000 random pairs of human proteins (the first five amino acids excluding the initial methionine),

where similarity of  $\geq 50\%$  was virtually non-existent in that sample (6 in 3000, false positive rate = 0.001) (Additional file 3).

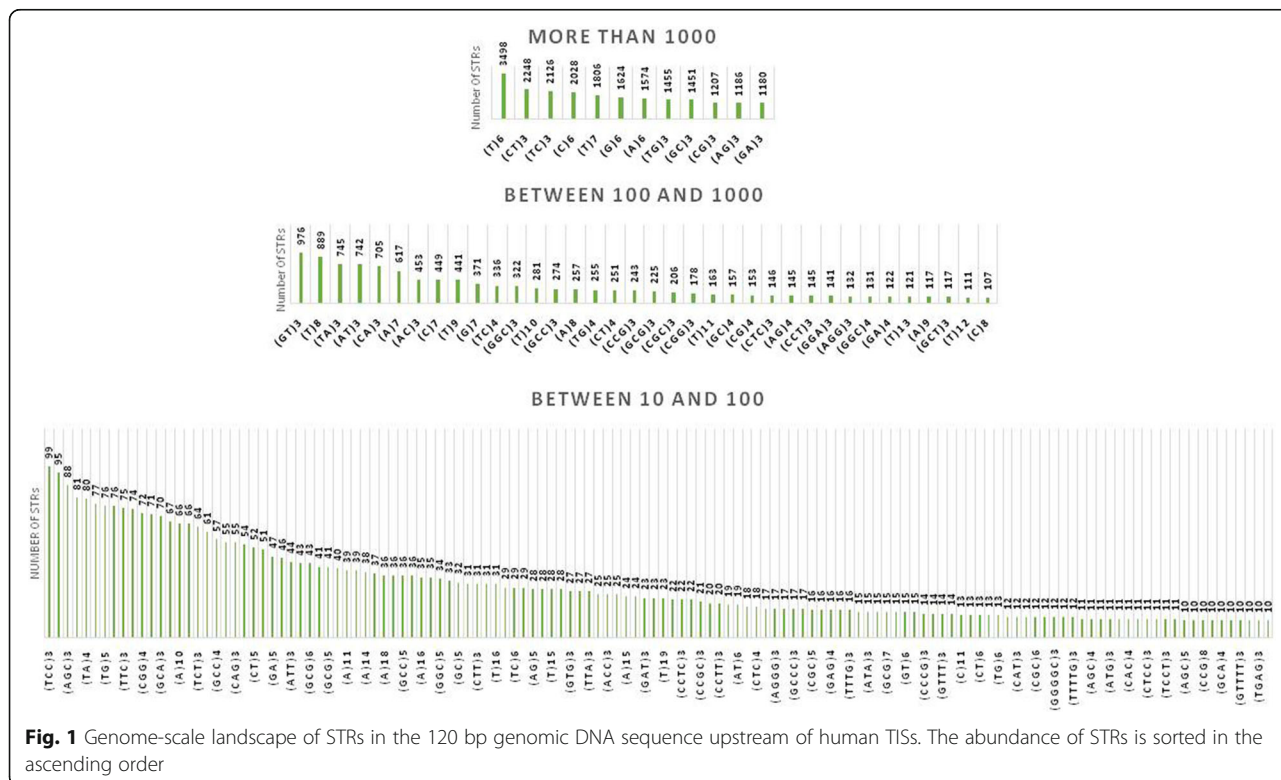
Finally, the two by two table and Fisher exact statistics were used to examine the link between STRs and TISs.

### Results

#### Genome-scale distribution of human STRs in the 120 bp upstream sequence of TISs

##### Genomic DNA platform

Mono- and dinucleotide STRs dominated STRs of  $> 1000$  counts, and the (T)6 mononucleotide repeat was the most abundant STR in this interval, succeeded by the (CT)3 and (TC)3 dinucleotide STRs (Fig. 1). Trinucleotide STRs were less abundant, observed at counts between 100 and 1000, and predominated by GC-rich composition such as (GGC)3, (GCC)3, (CCG)3, and (GCG)3. In the non-GC composition, (CTC)3 and (CCT)3 were the most common trinucleotide STRs. Tetra-, penta-, and hexanucleotide STRs were at lesser abundance than the above categories and observed at  $< 100$  counts, where (CCTC)3 and (CCGC)3 were the most abundant tetranucleotide STRs. Only three pentanucleotide STR classes, (GGGGC)3, (TTTTG)3, and (GTTTT)3, were observed at counts  $> 10$  in the screened interval.



**Fig. 1** Genome-scale landscape of STRs in the 120 bp genomic DNA sequence upstream of human TISs. The abundance of STRs is sorted in the ascending order

**cDNA platform**

The overall distribution of STRs in the 120 bp TIS-flanking cDNA sequences was significantly different in comparison to the genomic DNA STRs (Fig. 2). In comparison to the genomic DNA platform on which T(6) was the most abundant STR, GC-rich dinucleotide repeats were the most abundant on the cDNA platform. Numerous other instances at high, medium, and low abundance differentiated the genomic DNA vs. cDNA platforms (e.g., differential abundance of (T)8, (GT)3, (TA)3, and (CA)3 between the two platforms).

**Human-specific STR fingerprints on the TIS-flanking genomic DNA and cDNA platforms and differential distribution of these compartments in comparison to the overall STR distribution**

**Genomic DNA platform**

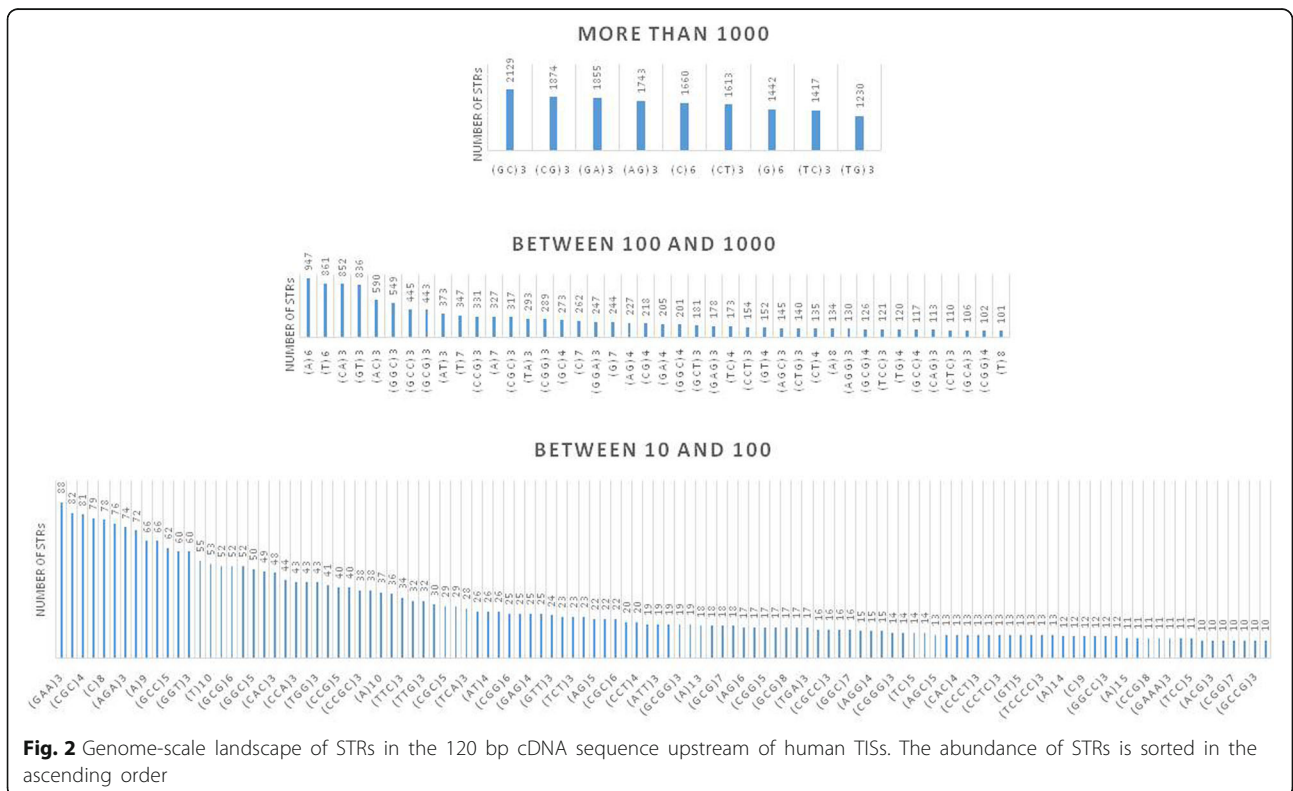
The flanking sequence of 755,956 TISs was screened in human and 46 other species in order to identify human-specific STRs. One thousand eight hundred eighty-seven genes contained human-specific TIS-flanking STRs on the genomic DNA platform, which were of a wide range of nucleotide compositions of mono-, di-, tri-, tetra-, penta-, and hexanucleotide repeats, of which poly(A) and poly(T) STRs were the longest (the 1st percentile, based on STR length, is listed in Table 1, and the list of all genes is provided as Additional file 4).

As an extreme example, the TIS of the *NVL* gene was flanked by a human-specific (T)22 STR, which was the longest STR detected in a human protein-coding gene TIS-flanking sequence. The TIS of the gene, *SULT1A3*, was flanked by the longest poly(A) at (A)18. Short- and medium-length STRs were also detected in the human-specific compartment (Additional file 4).

Significant skewing was observed in the distribution of human-specific STRs (Fig. 3) vs. the overall (i.e., human-specific and non-specific) STRs (Fig. 1) (Mann Whitney *U* test,  $p = 1 \times 10^{-5}$ ). While the (GC)3 and (CG)3 dinucleotide STRs were enriched in the overall STR compartment, their abundance was significantly lower in the human-specific compartment. Instead, (CA)3 and (AC)3 were significantly more abundant in the human-specific compartment. Differences in the distribution of tri- and tetranucleotide STRs were also observed between the two compartments. While trinucleotide and tetranucleotide STRs of GC composition were more abundant in the overall compartment, non-GC STR compositions (e.g., GGA, TTC, GCA, and ATAA) were more abundant in the human-specific compartment.

**cDNA platform**

Two thousand six hundred genes contained human-specific STRs in their TIS cDNA flanking sequence (the 1st percentile based on length is represented in Table 2 and



**Fig. 2** Genome-scale landscape of STRs in the 120 bp cDNA sequence upstream of human TISs. The abundance of STRs is sorted in the ascending order

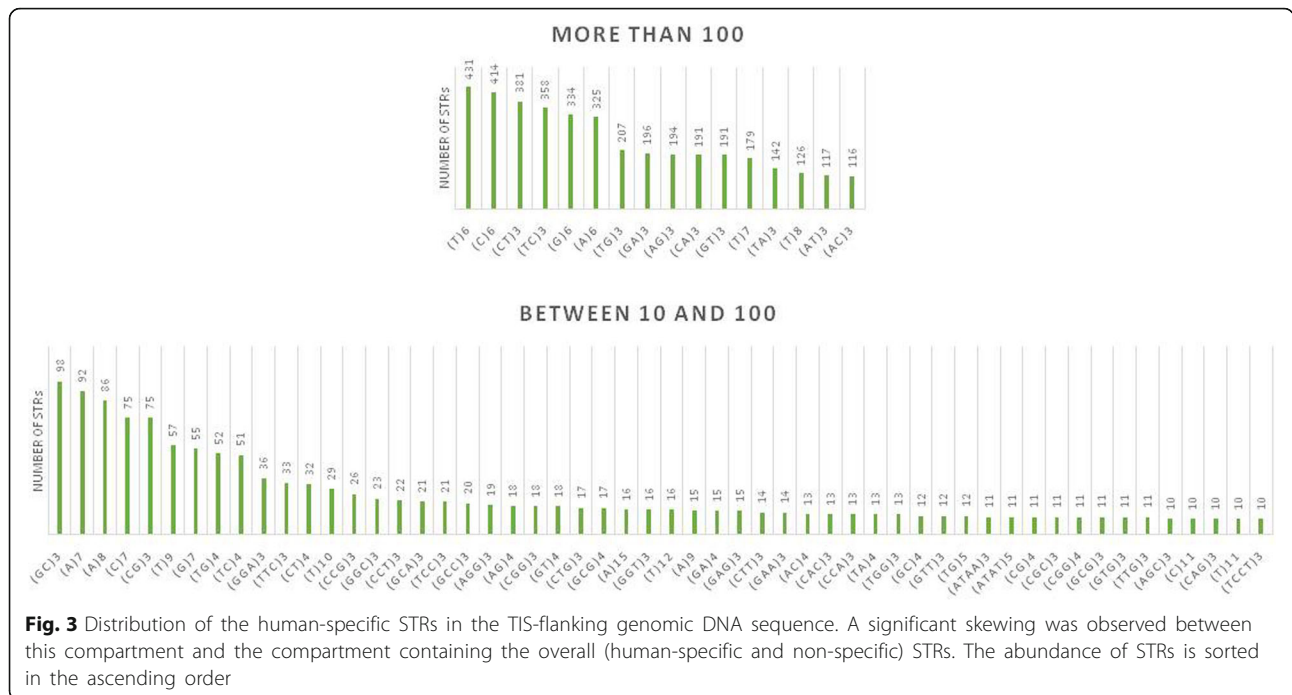
**Table 1** The 1st percentile of human protein-coding genes which contain human-specific STRs (length-wise) in their TIS-flanking genomic DNA sequence

Gene symbol	Gene Ensembl ID	Transcript ID	STR	GO term
<i>NVL</i>	ENSG00000143748	ENST00000436927	(T)22	ATP binding
<i>OR4K2</i>	ENSG00000165762	ENST00000641885	(T)20	Olfactory receptor activity
<i>MGRN1</i>	ENSG00000102858	ENST00000591895	(A)18	–
<i>SULT1A3</i>	ENSG00000261052	ENST00000338971 ENST00000395138		Sulfotransferase activity
<i>GDI2</i>	ENSG00000057608	ENST00000380127 ENST00000609712	(T)17	GTPase activator activity
<i>SULT1A4</i>	ENSG00000213648	ENST00000360423	(A)17	Sulfotransferase activity
<i>ZNF283</i>	ENSG00000167637	ENST00000618787 ENST00000593268	(T)17	Regulation of transcription
<i>ADAP2</i>	ENSG00000184060	ENST00000581548	(A)16	GTPase activator activity
<i>DDX20</i>	ENSG00000064703	ENST00000475700		Nucleic acid binding
<i>SGIP1</i>	ENSG00000118473	ENST00000435165	(A)16	Clathrin-dependent endocytosis
<i>LCA5L</i>	ENSG00000157578	ENST00000288350 ENST00000485895 ENST00000418018 ENST00000448288 ENST00000434281 ENST00000438404 ENST00000411566 ENST00000415863 ENST00000426783 ENST00000456017 ENST00000451131		Intracellular transport
<i>LRRC36</i>	ENSG00000159708	ENST00000569499 ENST00000568804	(T)14	–
<i>OR7A10</i>	ENSG00000127515	ENST00000641129	(CT)14	G protein-coupled receptor activity
<i>POLR2F</i>	ENSG00000100142	ENST00000492213	(T)14	Transcription, DNA templated
<i>SNX19</i>	ENSG00000120451	ENST00000528555 ENST00000530356	(T)14	Integral component of membrane
<i>TEX11</i>	ENSG00000120498	ENST00000395889	(TTCC)14	Meiotic cell cycle
<i>ACAT1</i>	ENSG00000075239	ENST00000527942	(T)13	Transferring acyl groups
<i>CHRFAM7A</i>	ENSG00000166664	ENST00000299847 ENST00000562729	(T)13	Ion transmembrane transport
<i>GALK2</i>	ENSG00000156958	ENST00000560654 ENST00000396509 ENST00000558145 ENST00000544523 ENST00000560138	(TG)13	Phosphotransferase activity

the complete list as Additional file 5). Similar to the genomic DNA platform, poly(A) and Poly(T) STRs were the longest STRs identified in the interval. The longest STR in this interval was (A)20 and belonged to *KCTD19*.

The distribution of human-specific STRs on the cDNA platform (Fig. 4) was unique to this platform and different from the overall STR distribution on the cDNA platform (Fig. 2) (Mann-Whitney *U* test  $p = 1 \times 10^{-5}$ ). While





the CG and CG dinucleotide STRs were more abundant in the overall distribution, CT, CA, and TG repeats were more abundant in the human-specific compartment. Various other differences were detected in other classes of STRs (e.g., more abundance of non-GC compositions in the trinucleotide STRs such as GGA, GCT, and CTG).

### STRs link to TIS selection on both genomic DNA and cDNA platforms

We examined the hypothesis that there may be a link between STRs and TIS selection. The initial five amino acids (excluding the initial methionine) of the human protein sequences, codons of which were flanked by STRs at the genomic DNA and cDNA, were BLASTed (compared using BLAST) against the initial five amino acids of all the proteins annotated for the orthologous genes in 46 species across vertebrates in order to compare the number of events in which human-specific and non-specific STRs occurred with homologous and non-homologous TISs ( $\geq 50\%$  and  $< 50\%$  similarity of the five amino acids). Total of 5,314,979 pair-wise TIS comparisons were performed, and significant correlation was observed between STRs and TIS selection both on the genomic DNA (Fig. 5a) and cDNA platforms (Fig. 5b) ( $p < 0.00001$ ), where there was excess occurrence of non-homologous TISs with human-specific STRs, and vice versa (i.e., excess occurrence of homologous TISs with non-specific STRs).

### Discussion

In this study, we characterized the genome-scale STR landscape of the immediate 120 bp upstream sequence

of human TISs on the genomic DNA and cDNA platforms, cataloged the human-specific compartment of these STRs, and investigated a possible link between STRs and TIS selection. Our findings provide the first evidence of a link between STRs and TIS selection on both platforms. This link is primarily deduced based on the differential co-occurrence rate of human-specific STRs with non-homologous TISs and non-specific STRs with homologous TISs.

Sequence similarity searches can reliably identify “homologous” proteins or genes by detecting excess similarity [40]. The TIS homology threshold of  $\geq 50\%$  was validated based on 3000 random similarity scorings of the initial protein-coding five amino acids (excluding the initiating methionine) of human proteins, in which that threshold was non-existent in effect (false-positive rate = 0.001). This scoring methodology was consistently applied to the TISs linked to human-specific and non-specific STRs.

We also observed differential distribution of the human-specific STRs vs. the overall distribution of STRs on both genomic and cDNA platforms. Importantly, each platform had a unique pattern of STR distribution, indicating differential selection of STRs based on their location and evolutionary course. Genome-scale skewing of STRs, albeit at a lesser scale of STR classes, was reported by our group in a preliminary study of the gene core promoter interval [36].

It is imperative to envision that human-specific *cis* elements at the mRNA and DNA may result in the production of proteins that are specific to humans. The RNA structure influences recruitment of various RNA binding

**Table 2** The 1st percentile of human protein-coding genes which contain human-specific STRs (length-wise) in their TIS-flanking cDNA sequence

Gene symbol	Gene Ensembl ID	Transcript ID	STR	GO term
<i>KCTD19</i>	ENSG00000168676	ENST00000566295	(A)20	Protein homooligomerization
<i>ATP8B1</i>	ENSG00000081923	ENST00000585322	(A)17	Magnesium ion binding
<i>C1QTNF1</i>	ENSG00000173918	ENST00000583904		Collagen trimer
<i>SEC11A</i>	ENSG00000140612	ENST00000558196		Peptidase activity
<i>SHQ1</i>	ENSG00000144736	ENST00000463369		–
<i>SPRY1</i>	ENSG00000164056	ENST00000505319		Multicellular organism development
		ENST00000610581		
<i>DDX20</i>	ENSG00000064703	ENST00000475700	(A)16	ATP binding
<i>NAB1</i>	ENSG00000138386	ENST00000409641	(T)16	Negative regulation of transcription
<i>SGIP1</i>	ENSG00000118473	ENST00000435165	(A)16	–
<i>SOX6</i>	ENSG00000110693	ENST00000528252	(A)14	Multicellular organism development
<i>DPP6</i>	ENSG00000130226	ENST00000406326	(T)13	Proteolysis
		ENST00000377770		
<i>MLF1</i>	ENSG00000178053	ENST00000482628	(G)13	–
<i>SHC4</i>	ENSG00000185634	ENST00000558220	(T)13	Stem cell differentiation
<i>ITGB1BP2</i>	ENSG00000147166	ENST00000538820	(T)12	Calcium ion binding
<i>NELL2</i>	ENSG00000184613	ENST00000548531		Calcium ion binding
<i>GIPC1</i>	ENSG00000123159	ENST00000393028	(GCG)11	–
		ENST00000345425		
		ENST00000587210		
<i>HBS1L</i>	ENSG00000112339	ENST00000527578	(T)11	GTPase activity
<i>HOPX</i>	ENSG00000171476	ENST00000556376	(A)11	Cell differentiation
<i>OR7D2</i>	ENSG00000188000	ENST00000642043	(T)11	G protein-coupled receptor activity
<i>RNF145</i>	ENSG00000145860	ENST00000520638		Integral component of membrane
<i>TXNL4A</i>	ENSG00000141759	ENST00000592837		mRNA splicing, via spliceosome
<i>ABCF1</i>	ENSG00000204574	ENST00000468958	(A)10	ATP binding
<i>ARHGEF18</i>	ENSG00000104880	ENST00000359920	(T)10	Rho guanyl-nucleotide exchange factor activity
<i>ARL14</i>	ENSG00000179674	ENST00000320767	(A)10	GTP binding
<i>ASNS</i>	ENSG00000070669	ENST00000448127	(T)10	Asparagine biosynthetic process
<i>EIF2S1</i>	ENSG00000134001	ENST00000466499		Translation initiation factor activity

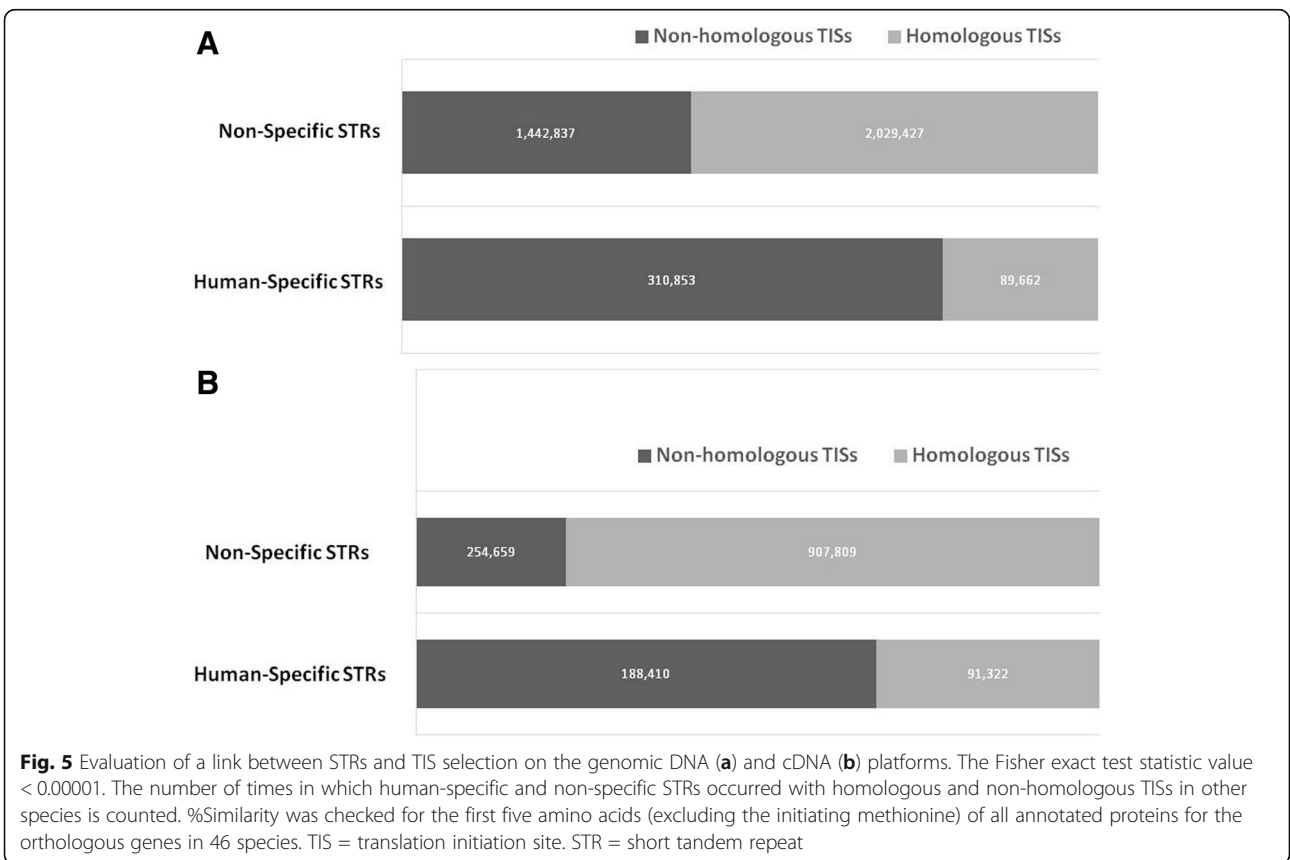
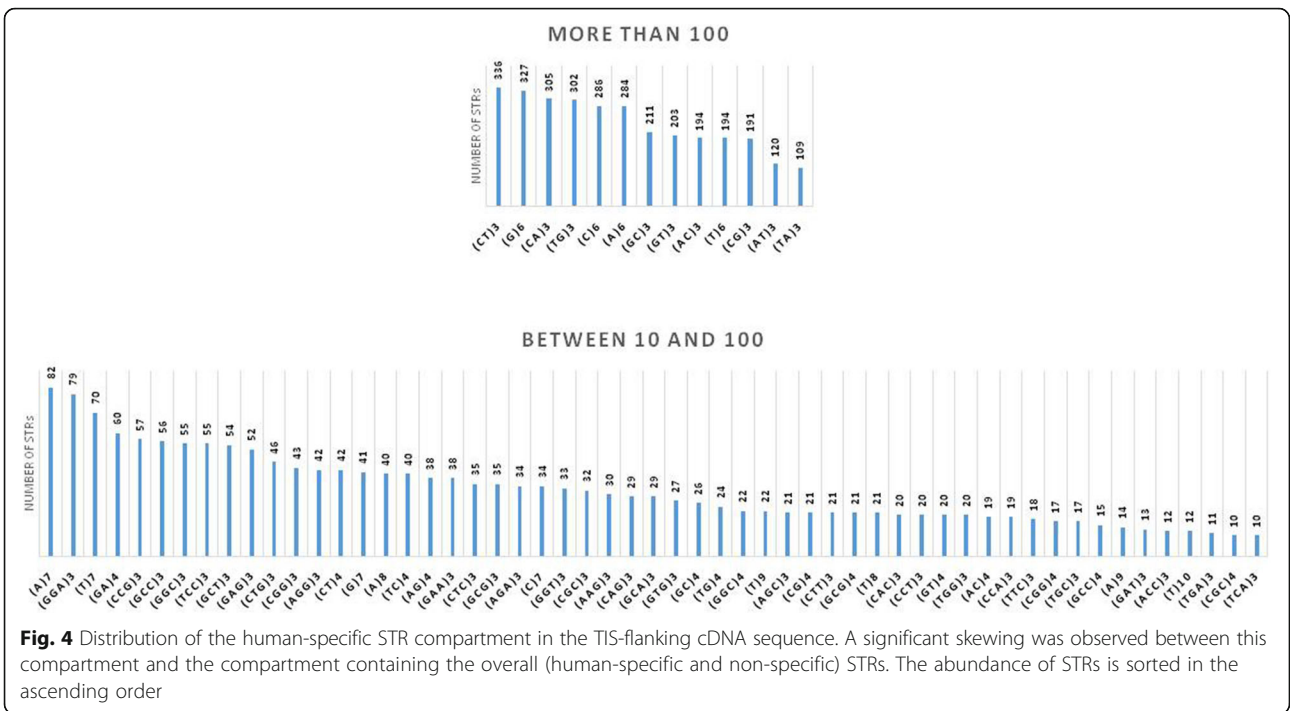
proteins and determines alternative TISs [41]. Indeed, the ribosomal machinery has the potential to scan and use several open reading frames (ORFs) at a particular mRNA species [42]. When located at the 5' or 3' UTR, STRs can modulate translation, the effect of which has biological and pathological implications [13, 43, 44]. The disorders linked to the 5' UTR STRs encompass a number of human-specific neurological disorders.

On the genomic DNA platform, proximity to the splice sites may increase the biological/pathological implication of repeats [45]. Similar to the cDNA STRs, we observed significant enrichment of non-homologous TISs co-occurring with human-specific genomic STRs,

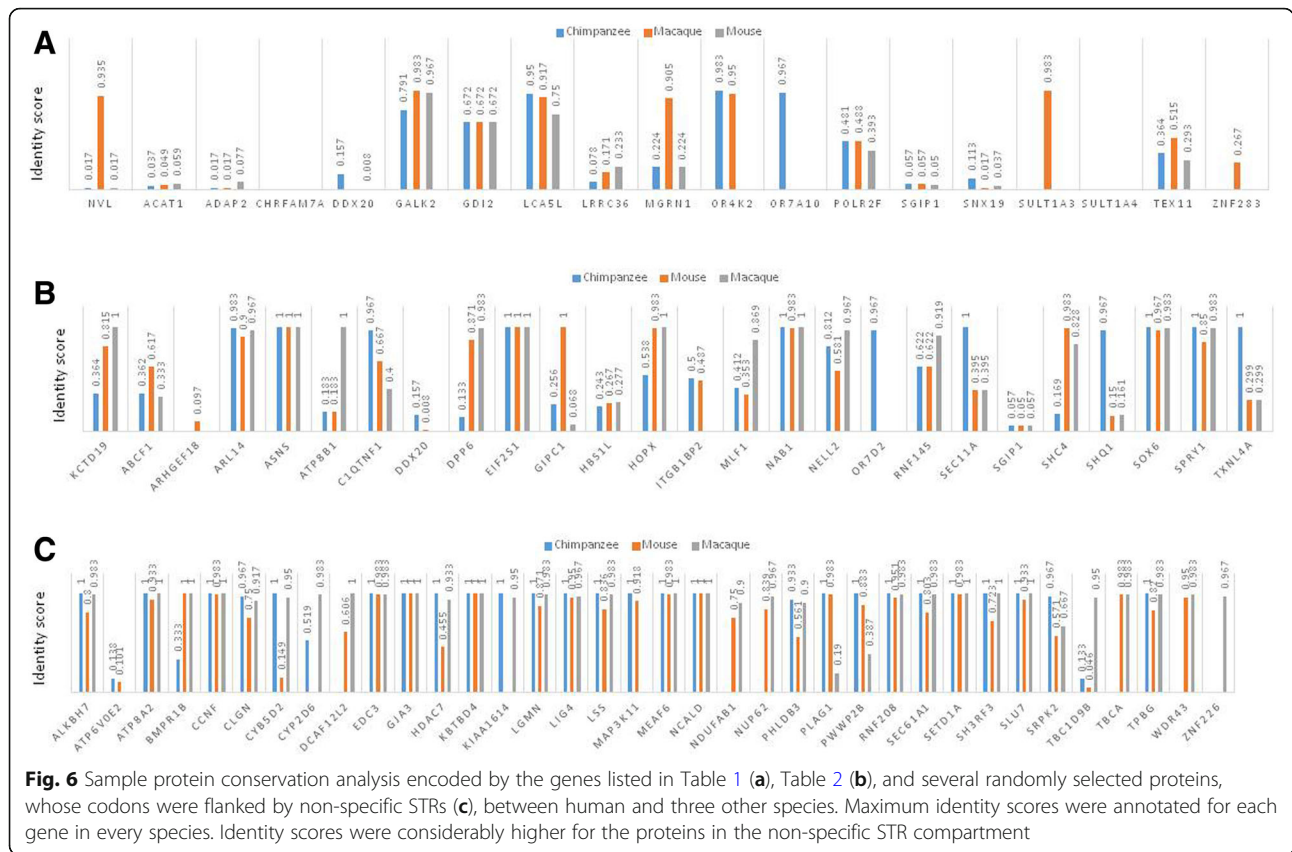
which were substantially near the exons (within 120 bp upstream of the TISs).

Gene Ontology (GO) search yielded a variety of terms across the identified genes, including neuron cell fate specification, multicellular organism development, translation initiation factor activity, and cell differentiation (<https://www.ebi.ac.uk/QuickGO/>), examples of which are presented in Tables 1 and 2.

EMBOSS Needle ([https://www.ebi.ac.uk/Tools/psa/emboss\\_needle](https://www.ebi.ac.uk/Tools/psa/emboss_needle)) pair-wise comparison was performed between human and three other species (chimpanzee, macaque, and mouse), of the proteins encoded by the transcripts in Table 1 (Fig. 6a), Table 2 (Fig. 6b), and several randomly selected proteins, codons of which were







flanked by non-specific STRs (Fig. 6c). Identity scores were considerably lower across the three species for the genes in Tables 1 and 2, compared to the identity scores for the genes in the non-specific STR compartment.

A number of the identified genes such as *ACAT1* (Table 1) and *SGIP1* (Table 2) confer risk for diseases or endophenotypes that are predominantly specific to the human species, such as complex psychiatric disorders [46, 47]. *SULT1A4* (Table 1) plays a critical role in neurotransmitter metabolism in the human brain and is also linked to neurodegeneration [48]. *APOA2* (Additional file 4) along with several other lipoproteins is linked to cognitive health [49]. In another remarkable example, *TBR1* (Additional file 5) is involved in *FOXP2* gene expression, which has pivotal role in speech and language in human [50]. *SRGAP2* (Additional file 5) family proteins may have increased the density of dendritic spines and promoted neoteny of the human brain during crucial periods of human evolution [51].

GO terms, protein conservation comparisons, and phenotypes stated above are only a few examples of the identified genes, in which human-specific STRs and the linked TISs may contribute to human evolution and disease. Future studies are warranted to examine the implication of the identified STRs and genes at the inter- and intra-species levels.

### Conclusion

We characterized the landscape of STRs at the immediate upstream genomic DNA and cDNA sequences flanking the human protein-coding gene TISs and found differential distribution of the human-specific STRs in comparison to the overall distribution of STRs on both platforms. Further, we propose a link between STRs and TIS selection, based on the differential co-occurrence rate of human-specific STRs with non-homologous TISs and non-specific STRs with homologous TISs. The data presented here have implications at the inter- and intraspecies levels, which warrant further functional and evolutionary studies.

### Additional files

- Additional file 1:** Workflow of STR identification in the TIS-flanking genomic DNA and cDNA upstream sequences. (JPG 37 kb)
- Additional file 2:** Pseudo-codes used for STR identification. (JPG 43 kb)
- Additional file 3:** Homology threshold validation. Three thousand random pair-wise similarity checks were performed on the five initial amino acids (excluding methionine) of human protein sequences. A similarity threshold of  $\geq 50\%$  was considered "homology." (JPG 24 kb)
- Additional file 4:** List of all human protein-coding genes which contain human-specific STRs in their TIS-flanking genomic DNA sequence. (DOCX 282 kb)
- Additional file 5:** List of all human protein-coding genes which contain human-specific STRs in their TIS-flanking cDNA sequence. (DOCX 393 kb)

**Abbreviations**

cDNA: Complementary DNA; CDS: Coding DNA sequence; GO: Gene Ontology; mRNA: Messenger RNA; STR: Short tandem repeat; TIS: Translation initiation site; TSS: Transcription start site

**Acknowledgements**

Not applicable

**Funding**

This research was funded jointly by the University of Social Welfare and Rehabilitation Sciences, Tehran, Iran, and University of Tehran, Iran.

**Availability of data and materials**

Please contact the author for data requests.

**Authors' contributions**

MA carried out the bioinformatics studies. KK participated in project supervision, data analysis, and co-ordination. AD helped in coordination. MO conceived the study, designed the project, supervised the analysis, and wrote the manuscript. All authors read and approved the final manuscript.

**Ethics approval and consent to participate**

Not applicable

**Consent for publication**

Not applicable

**Competing interests**

The authors declare that they have no competing interests.

**Publisher's Note**

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

**Author details**

<sup>1</sup>Department of Bioinformatics, Kish International Campus University of Tehran, Kish, Iran. <sup>2</sup>Laboratory of Complex Biological Systems and Bioinformatics (CBB), Department of Bioinformatics, Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran. <sup>3</sup>Iranian Research Center on Aging, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran.

Received: 16 August 2018 Accepted: 10 October 2018

Published online: 29 October 2018

**References**

- Andreev DE, O'Connor PBF, 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. <https://doi.org/10.1093/nar/gkw1190>.
- 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 U S A.* 2012;109(37):E2424–32. <https://doi.org/10.1073/pnas.1207846109>.
- 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(18):4354–61. <https://doi.org/10.1242/jcs.104919>.
- Bazykin GA, Kochetov AV. Alternative translation start sites are conserved in eukaryotic genomes. *Nucleic Acids Res.* 2011;39(2):567–77. <https://doi.org/10.1093/nar/gkq806>.
- Cenik C, Cenik ES, Byeon GW, Grubert F, Candille SI, Spacek D, Snyder MP. Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome Res.* 2015;25(11):1610–21. <https://doi.org/10.1101/gr.193342.115>.
- Babendure JR, Babendure JL, Ding J-H, Tsien RY. Control of mammalian translation by mRNA structure near caps. *RNA.* 2006;12(5):851–61. <https://doi.org/10.1261/rna.2309906>.
- Master A, Wójcicka A, Giżewska K, Poplawski 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. <https://doi.org/10.1371/journal.pone.0155359>.
- Park E, Pan Z, Zhang Z, Lin L, Xing Y. The expanding landscape of alternative splicing variation in human populations. *Am J Hum Genet.* 2018; 102(1):11–26. <https://doi.org/10.1016/j.ajhg.2017.11.002>.
- Gilbert WW, Zhou K, Butler TK, Doudna JA. Cap-independent translation is required for starvation-induced differentiation in yeast. *Science.* 2007;317: 1224–7. <https://doi.org/10.1126/science.1144467>.
- 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. <https://doi.org/10.1073/pnas.0804940105>.
- Yamagishi K, Oshima T, Masuda Y, Ara T, Kanaya S, Mori H. Conservation of translation initiation sites based on dinucleotide frequency and codon usage in *Escherichia coli* K-12 (W3110): non-random distribution of A/T-rich sequences immediately upstream of the translation initiation codon. *DNA Res.* 2002;9(1):19–24. <https://doi.org/10.1093/dnares/9.1.19>.
- 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. <https://doi.org/10.1016/j.brainres.2018.02.006>.
- 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. <https://doi.org/10.1371/journal.pone.0168204>.
- Krauss S, Griesche N, Jastrzebska E, Chen C, Rutschow D, Achmüller C, Dorn S, Boesch SM, Lalowski M, Wanker E, Schneider R, Schweiger S. Translation of HTT mRNA with expanded CAG repeats is regulated by the MID1-PP2A protein complex. *Nat Commun.* 2013;4:1511. <https://doi.org/10.1038/ncomms2514>.
- Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Erlich Y. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat Genet.* 2016;48(1):22–9. <https://doi.org/10.1038/ng.3461>.
- Yuan Z, Liu S, Zhou T, Tian C, Bao L, Dunham R, Liu Z. Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. *BMC Genomics.* 2018;19: 141. <https://doi.org/10.1186/s12864-018-4516-1>.
- Emamalizadeh B, Movafagh A, Darvish H, Kazeminasab S, Andarva M, Namdar-Aligoodarzi P, Ohadi M. The human RIT2 core promoter short tandem repeat predominant allele is species-specific in length: a selective advantage for human evolution? *Mol Genet Genomics.* 2017;292(3):611–7. <https://doi.org/10.1007/s00438-017-1294-4>.
- Abe H, Gemmell NJ. Evolutionary footprints of short tandem repeats in avian promoters. *Sci Rep.* 2016;6:19421. <https://doi.org/10.1038/srep19421>.
- Bushehri A, Barez MR, Mansouri SK, Biglarian A, Ohadi M. Genome-wide identification of human- and primate-specific core promoter short tandem repeats. *Gene.* 2016;587:83–90. <https://doi.org/10.1016/j.gene.2016.04.041>.
- Namdar-Aligoodarzi P, Mohammadparast S, Zaker-Kandjani B, Talebi Kakroodi S, Jafari Vesiehsari M, Ohadi M. Exceptionally long 5' UTR short tandem repeats specifically linked to primates. *Gene.* 2015;569:88–94. <https://doi.org/10.1016/j.gene.2015.05.053>.
- Nikkhah M, et al. An exceptionally long CA-repeat in the core promoter of SCGB2B2 links with the evolution of apes and Old World monkeys. *Gene.* 2016;576(1 Pt 1):109–14. <https://doi.org/10.1016/j.gene.2015.09.070>.
- Bilgin Sonay T, Carvalho T, Robinson MD, Greminger MP, Krützen M, Comas D, Wagner A. Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. *Genome Res.* 2015;25(11): 1591–9. <https://doi.org/10.1101/gr.190868.115>.
- Rezazadeh M, Gharesouran J, Mirabzadeh A, Khorram Khorshid HR, Biglarian A, Ohadi M. A primate-specific functional GTTT-repeat in the core promoter of CYTH4 is linked to bipolar disorder in human. *Prog Neuro-Psychopharmacol Biol Psychiatry.* 2015;56:161–7. <https://doi.org/10.1016/j.pnpbp.2014.09.001>.
- Khademi E, Alehabib E, Shandiz EE, Ahmadifard A, Andarva M, Jamshidi J, Rahimi-Aliabadi S, Pouriran R, Nejad FR, Mansoori N, Shahmohammadibeni N, Taghavi S, Shokraei P, Akhavan-Niaki H, Paisán-Ruiz C, Darvish H, Ohadi M. Support for "disease-only" genotypes and excess of homozygosity at the CYTH4 primate-specific GTTT-repeat in schizophrenia. *Genet Test Mol Biomarkers.* 2017;21:485–90. <https://doi.org/10.1089/gtmb.2016.0422>.
- Mohammadparast S, et al. Exceptional expansion and conservation of a CT-repeat complex in the core promoter of PAXBP1 in primates. *Am J Primatol.* 2014;76:747–56. <https://doi.org/10.1002/ajp.22266>.

26. Ohadi M, Mohammadparast S, Darvish H. Evolutionary trend of exceptionally long human core promoter short tandem repeats. *Gene*. 2012; 507(1):61–7. <https://doi.org/10.1016/j.gene.2012.07.001>.
27. King DG. Evolution of simple sequence repeats as mutable sites. *Adv Exp Med Biol*. 2012;769:10–25.
28. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet*. 2018;19:286–98.
29. Bagshaw ATM. Functional mechanisms of microsatellite DNA in eukaryotic genomes. *Genome Biol Evol*. 2017;9(9):2428–43. <https://doi.org/10.1093/gbe/evx164>.
30. 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;28:1169–78. <https://doi.org/10.1101/gr.231753.117>.
31. Press MO, Carlson KD, Queitsch C. The overdue promise of short tandem repeat variation for heritability. *Trends Genet*. 2014;30(11):504–12. <https://doi.org/10.1016/j.tig.2014.07.008>.
32. 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. *Am J Primatol*. 2015;77(1):34–43. <https://doi.org/10.1002/ajp.22308>.
33. Valipour E, et al. Polymorphic core promoter GA-repeats alter gene expression of the early embryonic developmental genes. *Gene*. 2013;531(2): 175–9. <https://doi.org/10.1016/j.gene.2013.09.032>.
34. Darvish H, Heidari A, Hosseinkhani S, Movafagh A, Khaligh A, Jamshidi J, Noorollahi-Moghaddam H, Heidari-Rostami HR, Karkheiran S, Shahidi GA, Togha M, Paknejad SM, Ashrafiyan H, Abdi S, Firouzabadi SG, Jamalidini SH, Ohadi M. Biased homozygous haplotypes across the human caveolin 1 upstream purine complex in Parkinson's disease. *J Mol Neurosci*. 2013;51(2): 389–93. <https://doi.org/10.1007/s12031-013-0021-9>.
35. Heidari A, Nariman Saleh Fam Z, Esmailzadeh-Gharehdaghi E, Banan M, Hosseinkhani S, Mohammadparast S, Oladnabi M, Ebrahimpour MR, Soosanabadi M, Farokhashtiani T, Darvish H, Firouzabadi SG, Farashi S, Najmabadi H, Ohadi M. Core promoter STRs: novel mechanism for inter-individual variation in gene expression in humans. *Gene*. 2012;492:195–8. <https://doi.org/10.1016/j.gene.2011.10.028>.
36. Nazariapanah N, Adelirad F, Delbari A, Sahaf R, Abbasi-Asl T, Ohadi M. Genome-scale portrait and evolutionary significance of human-specific core promoter tri- and tetranucleotide short tandem repeats. *Hum Genomics*. 2018;12:17 <https://doi.org/10.1186/s40246-018-0149-3>.
37. Alizadeh F, Bozorgmehr A, Tavakkoly-Bazzaz J, Ohadi M. Skewing of the genetic architecture at the ZMYM3 human-specific 5' UTR short tandem repeat in schizophrenia. *Mol Gen Genomics*. 2018;293:747–52. <https://doi.org/10.1007/s00438-018-1415-8>.
38. Li C, Lenhard B, Luscombe NM. Integrated analysis sheds light on evolutionary trajectories of young transcription start sites in the human genome. *Genome Res*. 2018. <https://doi.org/10.1101/gr.231449.117>.
39. Kramer M, Sponholz C, Slaba M, Wissuwa B, Claus RA, Menzel U, Bauer M. Alternative 5' untranslated regions are involved in expression regulation of human heme oxygenase-1. *PLoS One*. 2013;8(10):e77224. <https://doi.org/10.1371/journal.pone.0077224>.
40. Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics/editorial board*, Andreas D. Baxevanis. [et al.]. doi: <https://doi.org/10.1002/0471250953.bi0301s42>
41. Martínez-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. <https://doi.org/10.3390/ijms141121705>.
42. 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;33:923–5. <https://doi.org/10.1093/bioinformatics/btw736>.
43. Usdin K. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. *Genome Res*. 2008;18(7):1011–9. <https://doi.org/10.1101/gr.070409.107>.
44. 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. <https://doi.org/10.1038/nchembio864>.
45. Zhang X, Lin H, Zhao H, Hao Y, Mort M, Cooper DN, Liu Y. Impact of human pathogenic micro-insertions and micro-deletions on post-transcriptional regulation. *Hum Mol Genet*. 2014;23(11):3024–34. <https://doi.org/10.1093/hmg/ddu019>.
46. Shibuya Y, Niu Z, Bryleva EY, Harris BT, Murphy SR, Kheirollah A, Bowen ZD, Chang CCY, Chang TY. Acyl-coenzyme A: cholesterol acyltransferase 1 blockage enhances autophagy in the neurons of triple transgenic Alzheimer's disease mouse and reduces human P301L-tau content at the presymptomatic stage. *Neurobiol Aging*. 2015;36(7):2248–59. <https://doi.org/10.1016/j.neurobiolaging.2015.04.002>.
47. Hodgkinson CA, Enoch MA, Srivastava V, Cummins-Oman JS, Ferrier C, Iarikova P, Sankararaman S, Yamini G, Yuan Q, Zhou Z, Albaugh B, White KV, Shen PH, Goldman D. Genome-wide association identifies candidate genes that influence the human electroencephalogram. *Proc Natl Acad Sci U S A*. 2010;107(19):8695–700. <https://doi.org/10.1073/pnas.0908134107>.
48. Butcher NJ, Horne MK, Mellick GD, Fowler CJ, Masters CL, AIBL research group, Minchin RF. Sulfotransferase 1A3/4 copy number variation is associated with neurodegenerative disease. *Pharmacogenomics J*. 2017. <https://doi.org/10.1038/tpj.2017.4>.
49. Muenchhoff J, Song F, Poljak A, Crawford JD, Mather KA, Kochan NA, Yang Z, Trollor JN, Reppermund S, Maston K, Theobald A, Kirchner-Adelhardt S, Kwok JB, Richmond RL, McEvoy M, Attia J, Schofield PW, Brodaty H, Sachdev PS. Plasma apolipoproteins and physical and cognitive health in very old individuals. *Neurobiol Aging*. 2017;55:49–60. <https://doi.org/10.1016/j.neurobiolaging.2017.02.017>.
50. Becker M, Devanna P, Fisher SE, Vernes SC. Mapping of human *FOXP2* enhancers reveals complex regulation. *Front Mol Neurosci*. 2018;11:47. <https://doi.org/10.3389/fnmol.2018.00047>.
51. Lucas B, Hardin J. Mind the (sr)GAP - roles of Slit-Robo GAPs in neurons, brains and beyond. *J Cell Sci*. 2017;130:3965–74. <https://doi.org/10.1242/jcs.207456>.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

