Link Between Short tandem Repeats and Translation Initiation Site Selection.

Arabfard M[1,2], Kavousi K[2*], Delbari A[3], Ohadi M[3]*.

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

***Joint correspondence:**

Mina Ohadi, M.D., Ph.D
Email Address: ohadi.mina@yahoo.com
mi.ohadi@uswr.ac.ir

Kaveh Kavousi, Ph.D.
Email Address: kkavousi@ut.ac.ir

Keywords: translation initiation site; short tandem repeat; genome-scale; human-specific

Abbreviations

cDNA: Complementary DNA

CDS: Coding DNA sequence

STR: Short Tandem Repeat

TIS: Translation Initiation Site

TSS: Transcription Start Site

Abstract:

Recent work in yeast and humans suggest that evolutionary divergence in *cis*-regulatory sequences impact translation initiation sites (TISs). *Cis*-elements can also affect the efficacy and amount of protein synthesis. Despite their vast biological implication, the landscape and relevance of short tandem repeats (STRs)/microsatellites to the human protein-coding gene TISs remain largely unknown. Here we characterized the STR distribution at the 120 bp cDNA sequence upstream of all annotated human protein-coding gene TISs based on the Ensembl database. Furthermore, we performed a comparative genomics study of all annotated orthologous TIS-flanking sequences across 47 vertebrate species (755,956 transcripts), aimed at identifying human-specific STRs in this interval. We also hypothesized that STRs may be used as genetic codes for the initiation of translation. The initial five amino acid sequences (excluding the initial methionine) that were flanked by STRs in human were BLASTed against the initial orthologous five amino acids in other vertebrate species (2,025,817 pair-wise TIS comparisons) 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. ≥50% and <50% similarity of the five amino acids). We characterized human-specific STRs and a bias of this compartment in comparison to the overall (human-specific and non-specific) distribution of STRs (Mann Whitney $p=1.4 \times 10^{-11}$). We also found significant enrichment of non-homologous TISs flanked by human-specific STRs ($p<0.00001$). In conclusion, our data indicate a link between STRs and TIS selection, which is supported by differential evolution of the human-specific STRs in the TIS upstream flanking sequence.

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 (Andreev et al. 2017; Lee et al. 2012). The alternative TISs result in various protein structures and functions (Fukushima et al. 2012; Georgii et al. 2011). 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 and thermal stability (Cenik et al; Babendure et al. 2006; Master et al. 2016).

One of the important and understudied *cis*-regulatory elements affecting translation are 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_N can enhance internal ribosomal entry both in the presence of PABP and eIF-4G in *Saccharomyces cerevisiae* (Gilbert *et al.* 2007), and in the complete absence of PABP and eIF-4G (Shirokikh and Spirin 2008). Biased distribution of dinucleotide repeats is a known phenomenon in the region immediately upstream of the TISs in *E. coli* (Yamagishi et al. 2002). In pathological instances, expansion of STRs in the RNA structure result in toxic RNAs and non-AUG translation, and the development of several human-specific neurological disorders (Glineburg et al. 2018; Rovozzo et al. 2016; Krauss et al. 2013). Genome-scale findings of the evolutionary trend of a number of STRs has begun to unfold their implications in respect with speciation and species-specific characteristics/phenotypes (Yuan et al. 2018; Emamalizadeh et al. 2017; Abe and Gemmell 2016; Bushehri et al. 2016; Namdar-

Aligoodarzi et al. 2016; Nikkhah et al. 2016; Bilgin Sonay et al. 2015; Rezazadeh et al. 2014; Khademi et al. 2017; Mohammadparast et al. 2014; Ohadi et al. 2012; King et al. 2012). The hypermutable nature of STRs and their large unascertained reservoir of functionality make them an ideal source of evolutionary adaptation, speciation, and disease (Hannan et al. 2018; Bagshaw et al. 2017; Press et al. 2017; Ohadi et al.2015; Valipour et al. 2013; Heidari et al. 2012). In line with the above, recent reports indicate a role of repetitive sequences in the creation of new transcription start sites (TSSs) in human (Nazaripanah et al. 2018; Alizadeh et al. 2018; Li et al. 2018; Kramer et al. 2013).

Here, we performed a genome-scale screen of the upstream complementary DNA (cDNA) sequence flanking all human protein-coding TISs annotated in the Ensembl database and a comparative genomics analysis to examine a possible link between these STRs and TIS selection in 47 species encompassing the major classes of vertebrates.

Results

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

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).

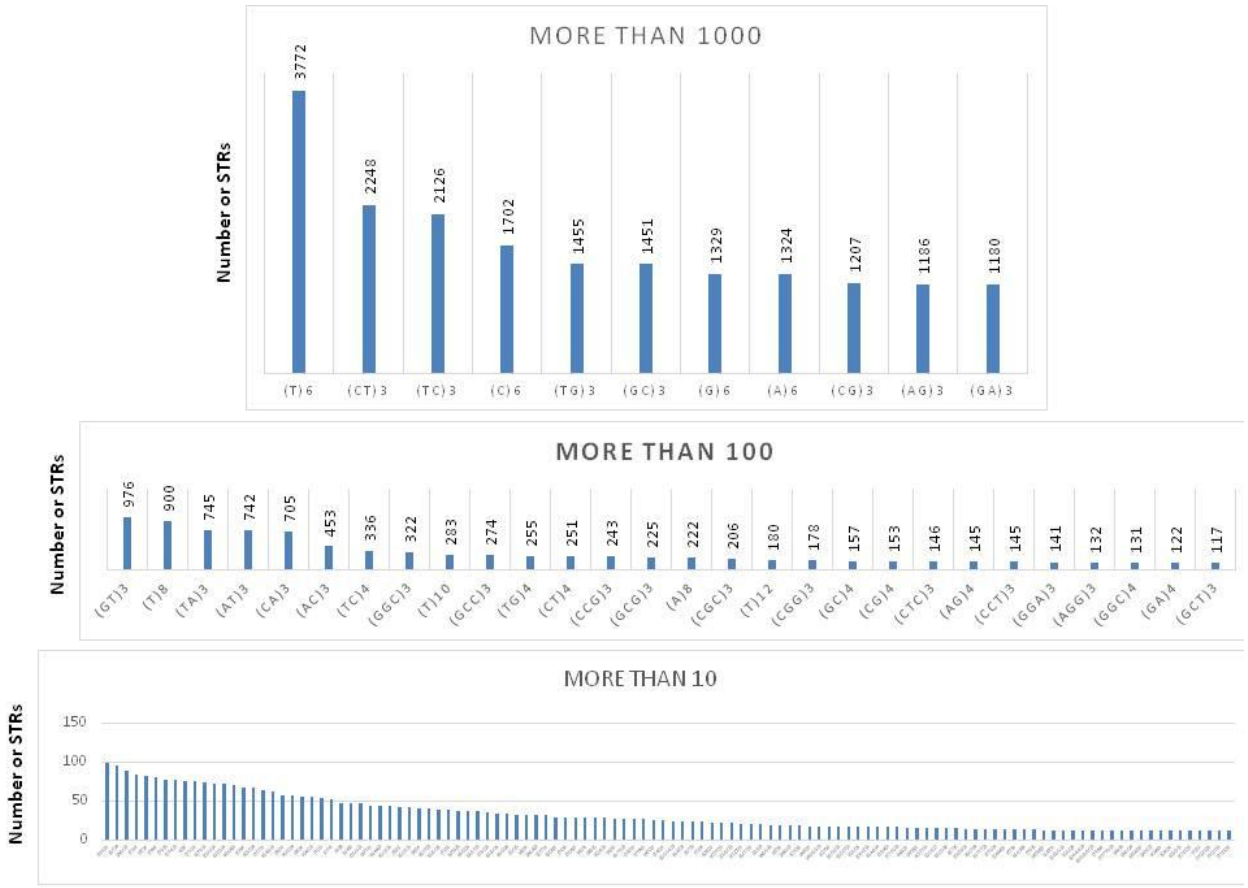


Fig. 1. Genome-scale landscape of the human STRs in the 120 bp cDNA sequence flanking TISs. The abundance of STRs is sorted in the ascending order.

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 (GGCG)3 and (CCTC)3 were the most abundant tetranucleotide STRs. Only three pentanucleotide STR classes, (GGGGC)3, (TTTTG)3 and (TCCCC)3 were observed at counts >10 in the screened interval.

Human-specific STR fingerprints in the TIS-flanking sequence and skewing of this compartment in comparison to the overall human STR distribution.

Two thousand two hundred sixty eight genes contained human-specific TIS-flanking STRs, which were of a wide range of nucleotide compositions of mono, di, tri, tetra, penta, and hexanucleotide repeats (Table 1).

The human-specific STRs were non-existent in the orthologous genes at ≥3-repeats in 46 species studied across major classes of vertebrates (a total of 755,956 transcripts). The 5th percentile of these genes is listed in Table 1 based on the length of the STRs (the entire list of the genes is presented as Suppl. 1). As an extreme example, the TIS of the *NVL* gene was flanked by a human-specific (T)22 STR, which is the longest STR detected in a human protein-coding gene TIS-flanking sequence. The TIS of the gene, *LRRC19*, was flanked by the longest poly-A at (A)20. Short and medium length STRs were also detected in the human-specific compartment (Suppl. 1).

Table 1. The 5th percentile of human protein-coding genes which contain human-specific STRs in their TIS-flanking sequence.

Gene Name	Gene Ensembl ID	Gene Transcript ID	STR
NVL	ENSG00000143748	ENST00000436927	(T)22
ELOF1	ENSG00000130165	ENST00000587806	(T)20
OR4K2	ENSG00000165762	ENST00000641885	(T)20
LRRC19	ENSG00000184434	ENST00000380055	(A)20
MGRN1	ENSG00000102858	ENST00000591895	(A)18
MGRN1	ENSG00000102858	ENST00000591895	(A)18
SULT1A3	ENSG00000261052	ENST00000338971	(A)18
SULT1A3	ENSG00000261052	ENST00000395138	(A)18
TBR1	ENSG00000136535	ENST00000410035	(TG)17
RORA	ENSG00000069667	ENST00000335670	(T)16
RORA	ENSG00000069667	ENST00000335670	(T)16
ADAP2	ENSG00000184060	ENST00000581548	(A)16
DDX20	ENSG00000064703	ENST00000475700	(A)16

GDI2	ENSG00000057608	ENST00000380127	(T)16
GDI2	ENSG00000057608	ENST00000609712	(T)16
GSTA4	ENSG00000170899	ENST00000370960	(A)16
SULT1A4	ENSG00000213648	ENST00000360423	(A)16
ZNF283	ENSG00000167637	ENST00000618787	(T)16
ZNF283	ENSG00000167637	ENST00000593268	(T)16
SGIP1	ENSG00000118473	ENST00000435165	(A)16
OR1C1	ENSG00000221888	ENST00000641256	(CA)15
CNGA1	ENSG00000198515	ENST00000402813	(CA)15
POLR2F	ENSG00000100142	ENST00000492213	(T)14
SNX19	ENSG00000120451	ENST00000528555	(T)14
SNX19	ENSG00000120451	ENST00000530356	(T)14
OR7A10	ENSG00000127515	ENST00000641129	(CT)14
HIGD2B	ENSG00000175202	ENST00000311755	(A)14
LCA5L	ENSG00000157578	ENST00000288350	(A)14
LCA5L	ENSG00000157578	ENST00000358268	(A)14
LCA5L	ENSG00000157578	ENST00000485895	(A)14
LCA5L	ENSG00000157578	ENST00000418018	(A)14
LCA5L	ENSG00000157578	ENST00000448288	(A)14
LCA5L	ENSG00000157578	ENST00000434281	(A)14
LCA5L	ENSG00000157578	ENST00000438404	(A)14
LCA5L	ENSG00000157578	ENST00000411566	(A)14
LCA5L	ENSG00000157578	ENST00000415863	(A)14
LCA5L	ENSG00000157578	ENST00000426783	(A)14
LCA5L	ENSG00000157578	ENST00000456017	(A)14
LRRC36	ENSG00000159708	ENST00000569499	(T)14
LRRC36	ENSG00000159708	ENST00000568804	(T)14
MRPL13	ENSG00000172172	ENST00000518918	(T)14
TEX11	ENSG00000120498	ENST00000395889	(TTCC)14
GALK2	ENSG00000156958	ENST00000560654	(TG)13
GALK2	ENSG00000156958	ENST00000396509	(TG)13
GALK2	ENSG00000156958	ENST00000558145	(TG)13
GALK2	ENSG00000156958	ENST00000544523	(TG)13
GALK2	ENSG00000156958	ENST00000560138	(TG)13
GALK2	ENSG00000156958	ENST00000559454	(TG)13
RFX2	ENSG00000087903	ENST00000586806	(T)12
RFX2	ENSG00000087903	ENST00000586806	(T)12
RFC5	ENSG00000111445	ENST00000484086	(T)12
PYCR1	ENSG00000183010	ENST00000582198	(A)12
PYCR1	ENSG00000183010	ENST00000579366	(A)12
DAGLB	ENSG00000164535	ENST00000436575	(T)12
NPTN	ENSG00000156642	ENST00000565282	(T)12

GLOD4	ENSG00000167699	ENST00000536578	(A)12
EFHB	ENSG00000163576	ENST00000344838	(T)12
ITGB1BP2	ENSG00000147166	ENST00000538820	(T)12
KCNN4	ENSG00000104783	ENST00000615047	(A)12
ACAT1	ENSG00000075239	ENST00000527942	(T)12
MRPS36	ENSG00000134056	ENST00000512880	(T)12
MRPS36	ENSG00000134056	ENST00000602380	(T)12
MYO5C	ENSG00000128833	ENST00000558479	(A)12
CHMP2B	ENSG00000083937	ENST00000472024	(T)12
CHRFAM7A	ENSG00000166664	ENST00000299847	(T)12
CHRFAM7A	ENSG00000166664	ENST00000562729	(T)12
TTC14	ENSG00000163728	ENST00000492617	(T)12
TTC14	ENSG00000163728	ENST00000495660	(T)12
SLC6A9	ENSG00000196517	ENST00000475075	(A)12
SNX1	ENSG00000028528	ENST00000560829	(A)12
HS3ST4	ENSG00000182601	ENST00000331351	(GCG)11
APOA2	ENSG00000158874	ENST00000468465	(GT)11
APOA2	ENSG00000158874	ENST00000463812	(GT)11
C1QB	ENSG00000173369	ENST00000510260	(TGGA)11
C1QB	ENSG00000173369	ENST00000509305	(TGGA)11
C1QB	ENSG00000173369	ENST00000432749	(TGGA)11
RAPGEF5	ENSG00000136237	ENST00000458533	(T)10
POLR3E	ENSG00000058600	ENST00000565358	(T)10
PDXDC1	ENSG00000179889	ENST00000563522	(T)10
PDXDC1	ENSG00000179889	ENST00000566426	(T)10
PDXDC1	ENSG00000179889	ENST00000567306	(T)10
PDE4DIP	ENSG00000178104	ENST00000479408	(A)10
SRPK1	ENSG00000096063	ENST00000512445	(T)10
TMBIM4	ENSG00000155957	ENST00000544599	(T)10
DCTN4	ENSG00000132912	ENST00000424236	(T)10
DCTN4	ENSG00000132912	ENST00000518015	(T)10
DCTN4	ENSG00000132912	ENST00000521533	(T)10
FZD8	ENSG00000177283	ENST00000374694	(C)10
NPM1	ENSG00000181163	ENST00000521672	(T)10
KITLG	ENSG00000049130	ENST00000552044	(T)10
OR7A17	ENSG00000185385	ENST00000642123	(AATA)10
OR7A17	ENSG00000185385	ENST00000641113	(AATA)10
OR52N4	ENSG00000181074	ENST00000641350	(T)10
UHRF1BP1L	ENSG00000111647	ENST00000545232	(C)10
UHRF1BP1L	ENSG00000111647	ENST00000548045	(C)10
UHRF1BP1L	ENSG00000111647	ENST00000551973	(C)10
UHRF1BP1L	ENSG00000111647	ENST00000550544	(C)10

UHRF1BP1L	ENSG00000111647	ENST00000551980	(C)10
UGT2B7	ENSG00000171234	ENST00000502942	(T)10
ABCF1	ENSG00000204574	ENST00000468958	(A)10
ACAP2	ENSG00000114331	ENST00000439666	(T)10
ARHGEF18	ENSG00000104880	ENST00000359920	(T)10
ARL14	ENSG00000179674	ENST00000320767	(A)10
ATXN3	ENSG00000066427	ENST00000502250	(T)10
ATXN3	ENSG00000066427	ENST00000557311	(T)10
EXTL3	ENSG00000012232	ENST00000523149	(T)10
RAPGEF6	ENSG00000158987	ENST00000513227	(T)10
NFIA	ENSG00000162599	ENST00000371191	(T)10
ZNF91	ENSG00000167232	ENST00000595533	(T)10
ZNF480	ENSG00000198464	ENST00000335090	(T)10
TLR2	ENSG00000137462	ENST00000260010	(GT)10
TSNAXIP1	ENSG00000102904	ENST00000388833	(A)10

Significant skewing was observed in the distribution of human-specific STRs vs. the overall (i.e. human-specific and non-specific) distribution of STRs in the human TIS-flanking sequence interval (Mann Whitney W=62532, $p=1.4 \times 10^{-11}$) (Fig. 2). For example, while the most abundant STR in the overall compartment was (T)6, the most abundant STR in the human-specific compartment was (CT)3. 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. Difference in the distribution of tri and tetranucleotide STRs was also observed between the two compartments. For example, while trinucleotide and tetranucleotide STRs of GC composition were more abundant in the overall compartment, non-GC STR compositions were more abundant in the human-specific compartment.

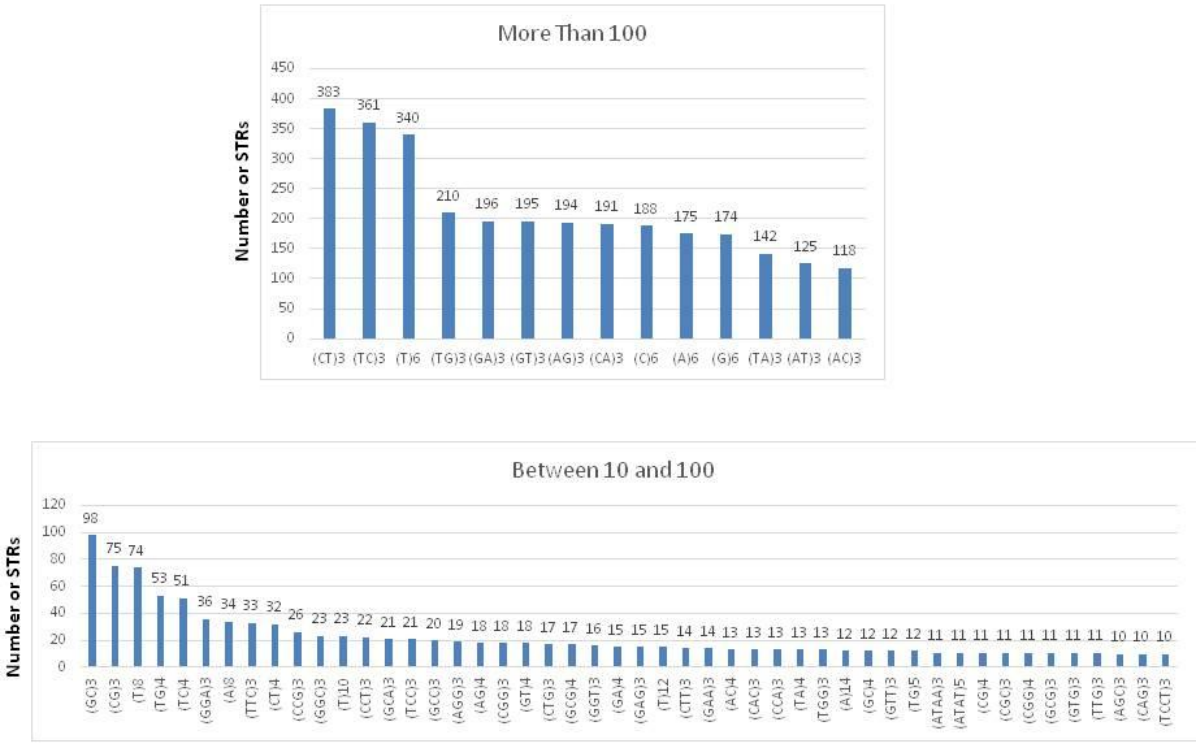


Fig. 2. Distribution of the human-specific STR compartment in the TIS-flanking cDNA sequence. A significant skewing was observed between this compartment and the compartment containing the overall (human-specific and non-specific) STRs. The abundance of STRs is sorted in the ascending order.

STRs link to TIS selection.

We examined the hypothesis that there may be a link between STRs and TIS selection in human. The initial five amino acids (excluding the initial methionine) of the human protein sequences that were flanked by STRs were checked against all the initial orthologous five amino acids in 46 species across vertebrates in order to compare the number of events where human-specific and

non-specific STRs occurred with homologous and non-homologous TISs (≥50% and <50% similarity of the five amino acids). Total of 2,025,817 pair-wise TIS comparisons were performed through the BLAST pipeline, and significant correlation was observed between STRs and TIS selection ($p<0.00001$) (Table 2), where there was a significant enrichment (2.98-fold) of non-homologous TISs co-occurring with human-specific STRs.

Table 2. Evaluation of a link between STRs and TIS selection

	Homology^{-} (Similarity <50%)*	Homology^{+} (Similarity ≥50%)	Marginal Row Total
TISs flanked by human-specific STRs	84,006	23,058	107,064
TISs flanked by non-specific STRs	1,057,463	861,290	1,918,753
Marginal Column Total	1,141,469	884,348	2,025,817 (Grand Total)

The Fisher exact test statistic value is < 0.00001. The result is significant at p < .05.

*Similarity was checked for the first five amino acids (excluding the initiating methionine) of the orthologous TISs. TIS=Translation initiation site. STR=Short tandem repeat

Discussion

The goal of this study was to characterize the STR landscape of the immediate 120 bp upstream sequence of human TISs at the whole-genome scale, to catalog the human-specific compartment of these STRs, and to investigate a possible link between STRs and TIS selection. Our findings provide the first indication of a link between STRs and TIS selection. The basis of this link was an excess of human-specific STRs co-occurring with non-homologous human TISs.

Sequence similarity searches can reliably identify "homologous" proteins or genes by detecting excess similarity (Pearson 2013). In our study, homology of the TISs was inferred based on

three thousand random similarity scorings of the initial protein-coding five amino acids (excluding the initiating methionine), in which a similarity of ≥50% was considered as "homology". This scoring methodology was consistently applied to the TISs linked to human-specific and non-specific STRs.

We also observed significant skewing of the human-specific STRs vs. the overall distribution of STRs. Genome-scale skewing of STRs, albeit at a lesser scale of STR classes, has been reported by our group in a preliminary study of the gene core promoter interval (Nazaripanah et al. 2018). The RNA structure influences recruitment of various RNA binding proteins, and determines alternative TISs (Martinez-Salas et al. 2013). Indeed, the ribosomal machinery has the potential to scan and use several open reading frames (ORFs) at a particular mRNA species (Kochetov et al. 2017). It is reasonable to envision that human-specific *cis* elements at the mRNA may result in the production of proteins that are specific to humans.

When located at the 5' or 3' UTR, STRs can modulate translation, the effect of which has biological and pathological implications (Rovozzo et al. 2016; Usdin 2008; Kumari et al. 2007). Remarkably, the disorders linked to the 5'UTR STRs encompass a number of human-specific neurological disorders. A number of the genes identified through our TIS-flanked STRs analysis (e.g. *NVL),* confer risk for diseases that are predominantly specific to the human species, such as schizophrenia and bipolar disorder (Wang et al. 2015). Neurodegeneration is another example linked to genes such as *SULT1A3* (Butcher et al. 2017). In another remarkable example, *TBR1* is involved in *FOXP2* gene expression, which has pivotal role in speech and language in human (Becker et al. 2018). The above are a few examples of how the identified genes and their potential human-specific translation may be linked 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 present the landscape of STRs at the immediate upstream cDNA sequence flanking the human protein-coding gene TISs. Further, we found a link between STRs and TIS selection, which is supported by the skewing of the human-specific STR compartment. The data presented here have implications at the inter- and intraspecies levels and warrant further functional and evolutionary studies.

Methods

Data collection

Forty seven species encompassing major classes of vertebrates were selected, and in each species, the 120 bp upstream cDNA sequence flanking all annotated protein-coding TISs (n=755,956 transcripts) were analyzed based on the Ensembl database versions 90 and 91 (https://asia.ensembl.org). For each gene in each species, its Ensembl ID, the annotated transcript IDs, the coding DNA sequence (CDS) and the annotated cDNA sequences were retrieved (the list of species and investigated genes are available upon request). The CDS sequences and their 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 cDNA interval upstream of the start codon (ATG) was investigated for the presence of STRs (Fig. 3).

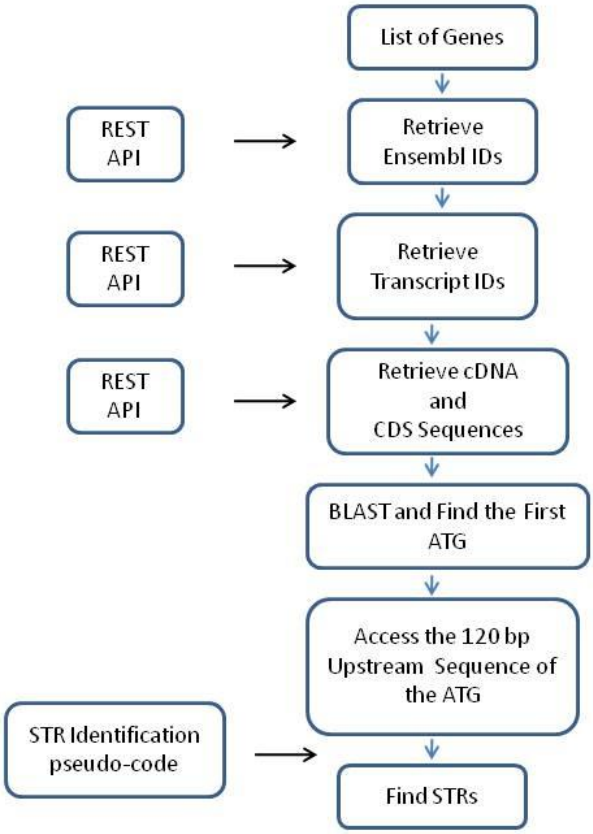


Fig. 3. Workflow of STR identification in the TIS-flanking upstream sequence.

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 cDNA and CDS sequences of genes and their respective transcripts were obtained using *func_get_GenomicSequence* and *func_get_CDSSequence* functions.

Identification of STRs in the human TIS-flanking interval

A general method of finding human-specific and non-specific STRs for each individual gene was developed and applied as follows: The 120 bp cDNA sequence flanking the TIS of all annotated

protein-coding gene transcripts was screened in 47 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 STRs which were "specific" to human (i.e. not present at ≥3-repeats in any other species) were identified. The relevant pseudo-code for the identification of repeated substrings (STRs) is illustrated in Fig. 4.

```
Input: A string Sequence.
Output: All tandem repeats in Sequence.
int len_seq = sequence.Length;
int min_repeats = 3;
int min_length = 2;
int max_length = 9;
for (int i = 0; i < len_seq- min_length; i++)
{
    for (int j = min_length; j < max_length + 1; j++)
    {
        if ((i + j) > len_seq) { break; }
        string sub_seq = sequence.Substring(i, j);
        double len_sub_seq = sub_seq.Length;
        string sub_seq_pattern;
        sub_seq_pattern = sub_seq;
        int matches = 1;
        while (IsMatch(sub_seq_pattern, sequence.Substring(i + j * matches, j)))
            matches++;
        if (matches >= min_repeats && (j * matches) >= min_length)
        {
            Output.Add(i, matches, j, sequence.Substring(i, j * matches));
            i += j * matches;
        }
    }
}
```

Fig.4. STR identification Pseudo-code

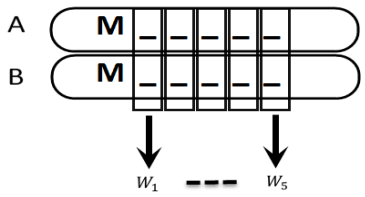
Mann-Whitney U test was used to compare the distribution of human-specific vs. the overall (specific and non-specific) STR distribution in human.

Evaluation of a link between STRs and TIS selection

We hypothesized that there may be a link between STRs and TIS selection. Weighted homology scoring was performed, in which the initial five amino acid sequence (excluding the initial methionine) of the human TISs that were linked to STRs were BLASTed against all initial protein-coding five amino acids in 46 species across vertebrates (2,025,817 pair-wise TIS comparisons). The above was aimed at comparing the number of events where 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) that are flanked by a STR at the 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 5 successive positions represented by i in the equation, we defined 5 weight coefficients W_1 to W_5 based on the importance of the amino acid position, observed in the W vector. The degree of homology between the two sequences A and B was calculated using function φ for all 5 positions with the operations $\sum_{i=1}^{L=5} W_i \varphi\left(A_{jik}, B_{jik'}\right)$. 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.



(1) $\mathrm{H}_{\mathrm{k}}^{\mathrm{j}} = \sum_{i=1}^{L=5} W_i \varphi\left(A_{jik}, B_{jik'}\right); for\ all\ k\ and\ k'$

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

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

Homology of the five amino acids and, therefore, the TIS was inferred based on the %similarity scoring, in which a similarity of ≥50% was considered as "homology". This threshold was achieved following BLASTing three thousand random pair-wise TIS five amino acids, which yielded a threshold of ≥50% (i.e. similarity scores ≥50% were considered "homology"). Finally, the two by two table and Fisher exact statistics were used to examine the link between STRs and TISs.

Disclosure Declaration

The authors have no conflict of interest to declare.

References:

Abe H, Gemmell NJ. 2016. Evolutionary Footprints of Short Tandem Repeats in Avian Promoters. *Sci Rep* **6:**19421.

Alizadeh F, Bozorgmehr A, Tavakkoly-Bazzaz J, Ohadi M. 2018. Skewing of the genetic architecture at the ZMYM3 human-specific 5' UTR short tandem repeat in schizophrenia. *Mol Genet Genomics* doi: 10.1007/s00438-018-1415-8.

Andreev DE, O'Connor PB, Loughran G, Dmitriev SE, Baranov PV, Shatsky IN. 2017.

Insights into the mechanisms of eukaryotic translation gained withribosome profiling. *Nucleic Acids Res.* **45**: 513-526.

Babendure JR, Babendure JL, Ding JH, Tsien RY. 2006. Control of mammalian translation by mRNA structure near caps. *RNA* **12:** 851-61.

Bagshaw ATM. 2017. Functional Mechanisms of Microsatellite DNA in Eukaryotic Genomes. *Genome Biol Evol.* **9:** 2428-2443.

Bilgin Sonay T, Carvalho T, Robinson MD, Greminger MP, Krützen M, Comas D, Highnam G, Mittelman D, Sharp A, Marques-Bonet T, Wagner A. 2015. Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. *Genome Res* **25:**1591-9.

Becker M, Devanna P, Fisher SE, Vernes SC. 2018. Mapping of Human *FOXP2* Enhancers Reveals Complex Regulation. *Front Mol Neurosci*. **11:**47.

Butcher NJ, Horne MK, Mellick GD, Fowler CJ, Masters CL; AIBL research group, Minchin RF. 2017. Sulfotransferase 1A3/4 copy number variation is associated with neurodegenerative disease. *Pharmacogenomics J.* 2017 Apr 4. doi: 10.1038/tpj.2017.4.

Cenik, Can; Cenik, Elif Sarinay; Byeon, Gun W; Candille, Sophie P.; Spacek, Damek; Araya, Carlos L; Tang, Hua; Ricci, Emiliano; Snyder, Michael P. 2015. Integrative analysis of RNA,

translation, and protein levels reveals distinct regulatory variation across humans. *Genome Res* **25**: 1610–21.

Emamalizadeh B, Movafagh A, Darvish H, Kazeminasab S, Andarva M, Namdar-Aligoodarzi P, Ohadi M. 2017. The human RIT2 core promoter short tandem repeat predominant allele is species-specific in length: a selective advantage for human evolution? *Mol Genet Genomics* **292:** 611-617.

Fukushima M, Tomita T, Janoshazi A, Putney JW. 2012. Alternative translation initiation gives rise to two isoforms of Orai1 with distinct plasma membrane mobilities. *J Cell Sci*. 125(Pt 18):4354-61.

Georgii A. Bazykin and Alex V. Kochetov. 2011. Alternative translation start sites are conserved in eukaryotic genomes. *Nucleic Acids Res* **39:** 567–577.

Gilbert W. V., Zhou K., Butler T. K., Doudna J. A., 2007. Cap-independent translation Is required for starvation-induced differentiation in yeast. *Science* **317:** 1224–1227

Glineburg MR, Todd PK, Charlet-Berguerand N, Sellier C. 2018. Repeat-associated non-AUG (RAN) translation and other molecular mechanisms in Fragile X Tremor Ataxia Syndrome. *Brain Res* pii: S0006-8993(18)30064-7.

Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. *Nat Rev Genet* **19:** 286-298. doi: 10.1038/nrg.2017.115. Review.

Heidari A., et al., Core promoter STRs: novel mechanism for inter-individual variation in gene expression in humans. *Gene* 2012. **492:** 195-8.

Kochetov AV, Allmer J, Klimenko AI, Zuraev BS, Matushkin YG, Lashin SA. 2017. AltORFev facilitates the prediction of alternative open reading frames in eukaryotic mRNAs. Bioinformatics. **33:** 923-925.

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, Shokraeian P, Akhavan-Niaki H, Paisán-Ruiz C, Darvish H, Ohadi M. 2017. Support for "Disease-Only" Genotypes and Excess of Homozygosity at the CYTH4 Primate-Specific GTTT-Repeat in Schizophrenia. *Genet Test Mol Biomarkers* **21:** 485-490.

King DG. 2012. Evolution of simple sequence repeats as mutable sites. *Adv Exp Med Biol.* **769:**10-25.

Kramer M, et al. 2013. Alternative 50 untranslated regions are involved in expression regulation of human heme oxygenase-1. *PLoS ONE*.**8:** e77224.

Krauss S, Griesche N, Jastrzebska E, Chen C, Rutschow D, Achmüller C, Dorn S, Boesch SM, Lalowski M, Wanker E, Schneider R, Schweiger S. 2013. Translation of HTT mRNA with expanded CAG repeats is regulated by the MID1-PP2A protein complex. *Nat Commun.* **4:**1511.

Kumari S, Bugaut A, Huppert JL, Balasubramanian S. 2007. An RNA G-quadruplex in the 5′ UTR of the NRAS proto-oncogene modulates translation. *Nat Chem Biol* **3:** 218–221.

Li C, Lenhard B, Luscombe NM. 2018. Integrated analysis sheds light on evolutionary trajectories of young transcription start sites in the human genome. *Genome Res.* 2018 Apr 4. doi: 10.1101/gr.231449.117

Lee S, Liu B, Lee S, Huang SX, Shen B, Qian SB. 2012. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci U S A.* **109:** E2424-32.

Li C, Lenhard B, Luscombe N.M. 2018. Integrated analysis sheds light on evolutionary trajectories of young transcription start sites in the human genome. *Genome Res*. pii: gr.231449.117. doi: 10.1101/gr.231449.117.

Martínez-Salas E, Lozano G, Fernandez-Chamorro J, Francisco-Velilla R, Galan A, Diaz R. 2013. RNA-binding proteins impacting on internal initiation of translation. 2013. *Int J Mol Sci.* 2013 **14:**21705-26.

Master A, Wójcicka A, Giżewska K, Popławski P, Williams GR, Nauman A. 2016. A Novel Method for Gene-Specific Enhancement of **ProteinTranslation** by Targeting 5'UTRs of Selected Tumor Suppressors. *PLoS One*. **11:**e0155359.

Mohammadparast, S., et al. 2014. Exceptional expansion and conservation of a CT-repeat complex in the core promoter of PAXBP1 in primates. *Am J Primatol* 2014. **76:** 747-56.

Namdar-Aligoodarzi P, Mohammadparast S, Zaker-Kandjani B, Talebi Kakroodi S, Jafari Vesiehsari M, Ohadi M. 2015. Exceptionally long 5' UTR short tandem repeats specifically linked to primates. Gene. **569:** 88-94.

Nazaripanah N, Adelirad F, Delbari A, Sahaf R, Abbasi-Asl T, **Ohadi M**. 2018. Genome-scale portrait and evolutionary significance of human-specific core promoter tri- and tetranucleotide short tandem repeats. *Hum Genomic*. **12**(1):17.

Nikkhah, M., et al., 2016. 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.

Ohadi M, Valipour E, Ghadimi-Haddadan S, Namdar-Aligoodarzi P, Bagheri A, Kowsari A, Rezazadeh M, Darvish H, Kazeminasab S. 2015. Core promoter short tandem repeats as evolutionary switch codes for primate speciation. *Am J Primatol* **77**(1):34-43.

Ohadi, M., S. Mohammadparast, and H. Darvish, 2012. Evolutionary trend of exceptionally long human core promoter short tandem repeats. *Gene* **507**(1): p. 61-7.

Press MO, Carlson KD, Queitsch C. 2014. The overdue promise of **short tandem repeat** variation for heritability. *Trends Genet* **30**(11):504-12.

Rezazadeh M, Gharesouran J, Mirabzadeh A, Khorram Khorshid HR, Biglarian A, **Ohadi** M. 2015. *Prog Neuropsychopharmacol Biol Psychiatry*. **56:**161-7.

Pearson WR. An **introduction** to **sequence similarity** ("**homology**")**searching**. Curr Protoc Bioinformatics. 2013 Jun;Chapter 3:Unit3.1. doi: 10.1002/0471250953.bi0301s42. Review.

Press MO, McCoy RC, Hall AN, Akey JM, Queitsch C. 2017. Massive variation of short tandem repeats with functional consequences across strains of Arabidopsis thaliana. doi:https://doi.org/10.1101/145128

Rezazadeh, M., et al., A primate-specific functional GTTT-repeat in the core promoter of CYTH4 is linked to bipolar disorder in human. Prog Neuropsychopharmacol Biol Psychiatry, 2015. 56: p. 161-7.

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 Dec 22;11(12):e0168204.

Rovozzo R, Korza G[2], Baker MW[3], Li M[4], Bhattacharyya A[4], Barbarese E[5], Carson JH[1]. CGG Repeats in the 5'UTR of FMR1 RNA Regulate Translation of Other RNAs Localized in the Same RNA Granules. PLoS One. 2016 Dec 22;11(12):e0168204. doi: 10.1371/journal.pone.0168204. eCollection 2016.

Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, et al. 2010. Complete Khoisan and Bantu genomes from southern Africa. *Nature*. **463**(7283):943-7.

Shirokikh N. E., Spirin A. S., 2008. Poly(A) leader of eukaryotic mRNA bypasses the dependence of translation on initiation factors. *Proc. Natl. Acad. Sci. USA* **105:** 10738–10743

Studtmann K, Olschläger-Schütt J, Buck F, Richter D, Sala C, Bockmann J, Kindler S, Kreienkamp HJ. 2014. A non-canonical initiation site is required for efficient translation of the dendritically localized Shank1 mRNA. *PLoS One*. **9**(2):e88518.

Usdin K 2008.The biological effects of simple tandem repeats: lessons from therepeat expansion diseases. *Genome Res* **18**(7):1011-9.

Valipour, E., et al. 2013. Polymorphic core promoter GA-repeats alter gene expression of the early embryonic developmental genes. *Gene* 531(2): p. 175-9.

Wang M, Chen J, He K, Wang Q, Li Z, Shen J, Wen Z, Song Z, Xu Y, Shi Y. 2015. The **NVL** gene confers risk for both major depressive disorder and schizophrenia in the Han Chinese population. *Prog Neuropsychopharmacol Biol Psychiatry* **62:**7-13.

Yamagishi K, Oshima T, Masuda Y, Ara T, Kanaya S, Mori H. 2002. 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* **9**(1):19-24.

Yuan Z, Liu S, Zhou T, Tian C, Bao L, Dunham R, Liu Z. 2018. Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. *BMC Genomics* **19**(1):141.

Suppl. 1. List of all human protein-coding genes which contain human-specific STRs in their TIS-flanking sequence.