Nazaripanah *et al. Human Genomics* (2018) 12:17
https://doi.org/10.1186/s40246-018-0149-3

Human Genomics

PRIMARY RESEARCH Open Access

Genome-scale portrait and evolutionary significance of human-specific core promoter tri- and tetranucleotide short tandem repeats

N. Nazaripanah[1], F. Adelirad[1], A. Delbari[1], R. Sahaf[1], T. Abbasi-Asl[2] and M. Ohadi[1*]

Abstract

Background: While there is an ongoing trend to identify single nucleotide substitutions (SNSs) that are linked to inter/intra-species differences and disease phenotypes, short tandem repeats (STRs)/microsatellites may be of equal (if not more) importance in the above processes. Genes that contain STRs in their promoters have higher expression divergence compared to genes with fixed or no STRs in the gene promoters. In line with the above, recent reports indicate a role of repetitive sequences in the rise of young transcription start sites (TSSs) in human evolution.

Results: Following a comparative genomics study of all human protein-coding genes annotated in the GeneCards database, here we provide a genome-scale portrait of human-specific short- and medium-size (≥ 3-repeats) tri- and tetranucleotide STRs and STR motifs in the critical core promoter region between − 120 and + 1 to the TSS and evidence of skewing of this compartment in reference to the STRs that are not human-specific (Levene's test $p < 0.001$). Twenty-five percent and 26% enrichment of human-specific transcripts was detected in the tri and tetra human-specific compartments (mid-$p < 0.00002$ and mid-$p < 0.002$, respectively).

Conclusion: Our findings provide the first evidence of genome-scale skewing of STRs at a specific region of the human genome and a link between a number of these STRs and TSS selection/transcript specificity. The STRs and genes listed here may have a role in the evolution and development of characteristics and phenotypes that are unique to the human species.

Keywords: Short tandem repeat, Core promoter, Human-specific, Trinucleotide, Tetranucleotide

Introduction

Speciation and evolution are, at least in part, due to the plasticity (expansion or contraction) of short tandem repeats (STRs)/microsatellites, which can function as "tuning knobs" in response to the environment or other genes [1–3]. In line with the above, certain STRs are directionally expanded in the human species or co-occur identically in related taxa such as primates [4–8]. Genes that contain STRs in their promoters have higher expression divergence compared to genes with fixed or no STRs in the gene promoters [9]. Recent reports indicate a role of repetitive sequences in the rise of young transcription start sites (TSSs) in human evolution [10–12].

Preliminary data on the sequencing of a number of "exceptionally long" STRs (≥ 6-repeats), which compose 1–2% of all human core promoter STRs [3], support critical evolutionary adaptive roles for a number of these STRs. Human specificity of the predominant allele of the *RIT2* core promoter STR in the human species, the presence of the shortest allele of this STR (5-repeat) in hunter-gatherer humans (BUSHMAN KB1: rs113265205), the lack of this allele in the agricultural modern humans (Genome Aggregation database: gnomad.broadinstitute.org), and its co-occurrence with schizophrenia provide the

---

* Correspondence: mi.ohadi@uswr.ac.ir; ohadi.mina@yahoo.com
[1] 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

© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

first indication of STR allele selection in humans [13]. A link between the *CYTH4* core promoter STR (the longest tetranucleotide STR identified in a human gene core promoter) with the Old World monkeys and Apes and evidence of extreme "disease-only" genotypes at this STR with schizophrenia [14] provide the first link between a primate-specific STR and higher-order brain functions in human. The "exceptionally long" CA-repeat in the core promoter of *SCGB2B2* is another example of directional STR expansion in the Old World monkeys and Apes [5]. The *PAXBP1* gene is an extreme example in which expansion of a core promoter CT-repeat occurs in the Old World monkeys and reaches maximum length and complexity in human; OMIM: 617621 [4].

As "exceptionally long" STRs may be subject to natural selection, short- and medium-size alleles (≥ 3-repeats) might have had similar fate. This is indicated by the predominance of specific short- and medium-size penta- and hexanucleotide STRs and their cognate transcription factors (TFs) in the critical core promoter interval [15]. Indeed, shortening of a number of STRs and their identical co-occurrence is linked to the evolution of primates [8]. In line with the above findings, repeats associated with younger human TSSs tend to be shorter than those in older TSSs [10]. In the study reported here, we present genome-scale data on two categories of STRs, i.e., tri- and tetranucleotide STRs, and their implication in human evolution.

Materials and methods

The interval between − 120 and + 1 to the TSS of all human protein-coding genes annotated in the GeneCards database (version 3.0) (www.genecards.org) was screened for tri- and tetranucleotide STRs of ≥ 3-repeats, based on the Ensembl database (versions 87-91) (asia.ensembl.org) and using the Microsatellite Repeats Finder at the following link: http://insilico.ehu.es/mini_tools/microsatellites/

The evolutionary status of the identified STRs was analyzed in 25 species (N), including primates ($N = 5$), non-primate mammals ($N = 12$), birds and reptiles ($N = 5$), amphibians ($N = 1$), and fish ($N = 2$), based on the Ensembl database.

Human specificity of transcripts was evaluated based on the multiple and pair-wise %identity scoring of the TSS-flanking 5′ untranslated region (UTR), using the sequence alignment program Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo), and the overall composition of the transcript and encoded protein (i.e., length of the transcript, number of exons and amino acids). The threshold of sequence identity was set at 50%, which was based on the comparison of two randomly selected and unrelated sequences in the human genome.

The p value for the skewing of the human-specific STR compartment was calculated using Levene's equality of variances test.

The p values for transcript enrichment were calculated using the two by two table analysis;

the human-specific tri- and tetranucleotide STR groups were compared against corresponding randomly selected STRs from the non-human-specific STRs. The comparison was set based on the sample size of the human-specific STRs (n) and the sample size of the non-human compartments (1.5n).

Results

Overall prevalence of tri- and tetranucleotide STR motifs across human protein-coding core promoter sequences

In total, 56 and 82 STR motifs were detected for the tri- and tetranucleotide repeats, respectively (Figs. 1 and 2). The most prevalent tri- and tetranucleotide STR motifs across the human protein-coding gene core promoters were GGC and GGGC, respectively (Figs. 1 and 2). In the category of non-GC STRs, GGA and TCCC were the most prevalent tri- and tetranucleotides, respectively.

Skewing of the human-specific core promoter tri- and tetranucleotide STRs

A significant skewing of the tri- and tetranucleotide STR distribution was found in the human-specific tri- (Fig. 1) and tetranucleotide (Fig. 2) compartments (Levene's $p < 0.001$). While the most prevalent tri- and tetranucleotide repeats in the non-human-specific category were the GGC- and GGGC-repeats, respectively, the most prevalent human-specific STRs were of the GCC and CTCC motifs, respectively. Disproportionate distribution of human-specific STRs was also detected in other STRs such as CCT, GAA, CTCC, GTTT, and GAAA.

The human-specific tri- and tetranucleotide STRs were of a wide range of motifs, e.g., the CCA motif in *ADCY6*, the TCCC motif in *ARHGEF35*, GCCC in *DRD2*, and GTTT in *MCTP2* (Tables 1 and 2).

In a number of instances, not only the STR, but also the genes containing those STRs, were human-specific (e.g., *ARHGEF35*, *AMY1C*, and *C1orf204*). Furthermore, a number of the tri- and tetranucleotide STRs were found to be unique to the human species at the specified interval of − 120 to + 1 TSS. For example, in the tetranucleotide compartment, CACC, GACA, CCGG, GATA, TCTG, GGCT, and TTTA STRs were detected in human only.

Enrichment of human-specific transcripts at the human-specific STR compartment

Based on sequence comparison and the overall composition of the transcript and encoded protein, 25

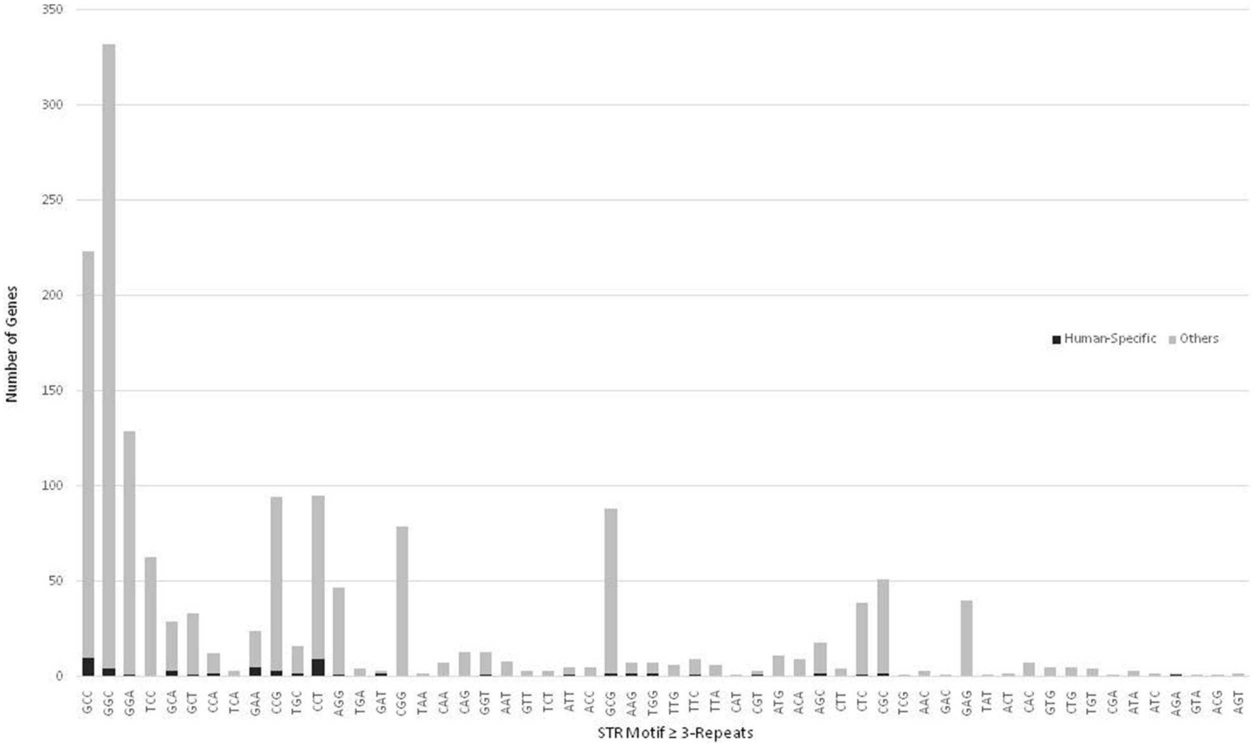


Fig. 1 Genome-scale prevalence of human protein-coding core promoter trinucleotide STRs and significant skewing of the human-specific STR compartment

and 26% of the transcripts in the tri and tetra human-specific compartments were found to be human-specific (mid-$p < 0.00002$ and mid-$p < 0.002$), respectively). The %identity score of multiple sequence alignment for the human-specific transcripts was 0 (exemplified in Fig. 3), and pair-wise analysis (exemplified in Fig. 4) resulted in %identity scores ranging from 37 to 48%. In the trinucleotide category, 14 genes, *MPRIP*, *NPAS1*, *PAQR9*, *PRSS1*, *R3HDM2*, *TMEM99*, *ZSCAN30*, *C22orf24*, *ECSCR*, *AMY1C*, *DDX58*, *C1orf204*, *RGPD6*, and *LCE2B*, contained human-specific transcripts. In the tetranucleotide category, five genes, *DRD2*, *DUX4*, *TEAD4*, *ARL17B*, and *ARHGEF35*, contained human-specific transcripts.

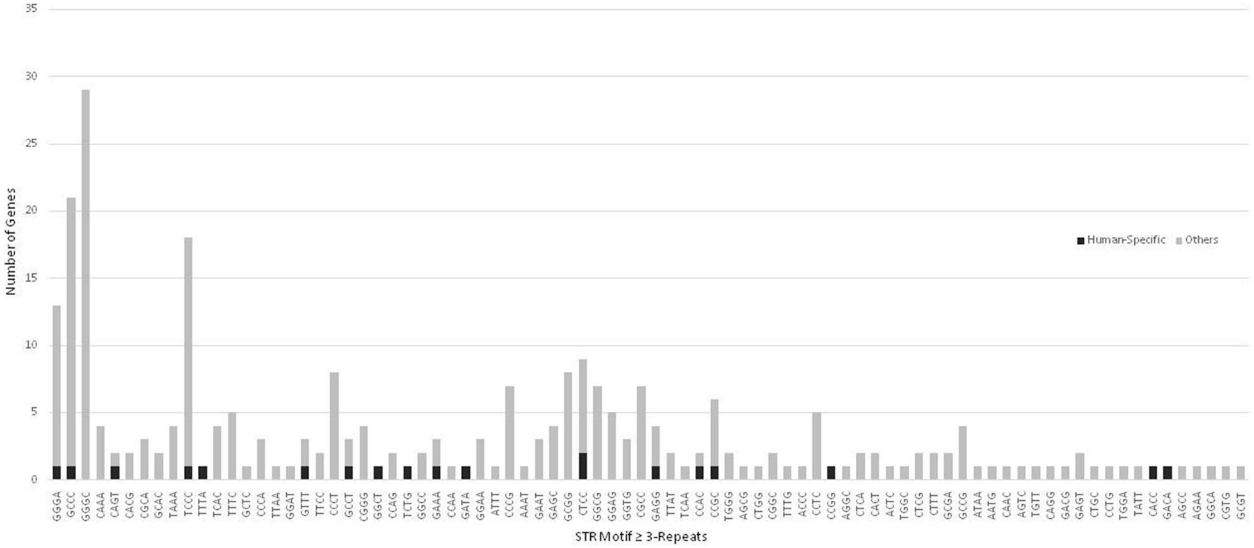


Fig. 2 Genome-scale prevalence of human protein-coding core promoter tetranucleotide STRs and significant skewing of the human-specific STR compartment

Table 1 Genome-scale human-specific core promoter trinucleotide STRs

Human gene symbol	Ensembl transcript ID	Variant no.	STR formula	
ADCY6	ENST00000307885.4	201	– 48 (CCA)3	
AMY1C	ENST00000370079.3	201	– 79 (ATT)3	
APBB1	ENST00000311051.7	202	– 11 (GCC)3	
BRINP2 (*FAM5B*)	ENST00000361539.4	201	– 26 (GGC)5	
BVES	ENST00000446408.2	203	– 35 (CCT)3	
CCDC178 (*C18orf34*)	ENST00000300227.12	201	– 79 (AGC)3	– 57 (CGC)5
CDH4	ENST00000543233.2	201	– 31 (CCT)3	
CIAPIN1	ENST00000563341.1	202	– 39 (CCT)3	
CNTNAP2	ENST00000361727.7	201	– 98 (TGC)3	
CST4	ENST00000217423.3	201	– 83 (GGA)3	
CYP4A11	ENST00000310638.8	201	– 86 (CCT)3	
C1orf204	ENST00000368102.5	201	– 11 (AAG)3	
C22orf24	NM_015372.2.1		– 98 (GCA)3	
C3AR1	ENST00000546241.1	202	– 70 (AGA)3	
DDX58	ENST00000379868.5	201	– 33 (CCT)3	
ECSCR	NM_001077693.3.1		– 121 (CCA)3	
GRIN2D	ENST00000263269.3	201	– 67 (GCC)3	
GSDMB	ENST00000394175.6	203	– 55 (GGC)3	
INPP4B	ENST00000503927.5	202	– 88 (CGC)3	
KBTBD12	ENST00000407609.7	204	– 57 (CCT)3	
KIAA1211	ENST00000504228.5	203	– 81 (AAG)3	
KRAS	ENST00000311936.7	202	– 89 (GAA)3	
KTN1	ENST00000395308.5	202	– 70 (GCG)9	
LACTBL1	ENST00000426928.6	201	– 58 (GAA)3	
LCE2B	ENST00000368780.3	201	– 79 (CCT)3	
LCOR (*C10orf12*)	ENST00000356016.7	202	– 97 (CCT)3	– 108 (GCC)3
MPRIP	ENST00000466186.2	209	– 88 (GCA)11	
MSANTD3 (*C9orf30*)	ENST00000374885.5	201	– 25 (GCC)7	
NPAS1	ENST00000439365.6	201	– 135 (GAA)9	
OR4X1	ENST00000320048.1	201	– 100 (GAT)3	
PABPC1L2B	ENST00000373521.3	201	– 60 (GCC)3	
PAQR9	ENST00000498470.1	203	– 29 (TGC)3	
PRSS1	ENST00000492062.1	205	– 96 (GAT)3	
RGPD6	ENST00000455695.1	205	– 70 (GGC)5	
RNF215	ENST00000215798.10	201	– 123 (GCT)5	
R3HDM2	ENST00000448732.1	208	– 27 (GCC)3	
SCN3B	ENST00000299333.7	201	– 29 (GGT)3	
SERPINB9	ENST00000380698.4	201	– 12 (GCA)3	
SIGLEC7	ENST00000305628.7	201	– 80 (TTC)3	
SPATC1L (*C21orf56*)	ENST00000330205.10	202	– 37 (TGG)4	–90 (TGG)4
STUB1	ENST00000219548.8	201	– 91 (GCC)3	
SUMF1	ENST00000272902.9	201	– 102 (AGC)3	
TEX12	ENST00000280358.4	201	– 105 (TGG)3	

Table 1 Genome-scale human-specific core promoter trinucleotide STRs *(Continued)*

Human gene symbol	Ensembl transcript ID	Variant no.	STR formula	
TMEM99	ENST00000301665.7	201	– 32 (CCG)3	– 47 (CCG)3
			– 59 (CCG)3	– 83 (CCG)3
			– 110 (CCG)3	– 125 (CCG)3
			– 48 (GCC)3	– 60 (GCC)3
			– 84 (GCC)3	– 126 (GCC)3
TNNC2	ENST00000372557.1	202	– 53 (GCC)3	
TPTE	ENST00000427445.6	201	– 110 (GCG)3	
TRBJ2-7	ENST00000390419.1	201	– 120 (GGC)3	
TRGV5	NC_000007.14:TRGV5:u_t_1.1		– 48 (CTC)3	
TRIM39	ENST00000376656.8	201	– 58 (CCT)4	
UAP1	ENST00000367926.8	204	– 9 (CGT)3	
VNN2	ENST00000326499.10	201	– 31 (GAA)10	
WRN	NM_000553.4.1		– 67 (GCC)3 – 69 (CCG)4	– 92 (GCC)3
WRNIP1	ENST00000618555.4	205	– 67 (CCG)3	
ZDHHC21	XM_006716760.1.1		– 32 (AGG)3	
ZSCAN30	ENST00000639929.1	212	– 57 (GAA)3	

The numbers before the brackets represent the start site of the STR in respect of the corresponding transcription start site. "Variant no" corresponds to the Ensembl isoform number

Table 2 Genome-scale human-specific core promoter tetranucleotide STRs

Human gene symbol	Ensembl transcript ID	Variant no.	STR formula
ARHGAP5	ENST00000345122.7	202	– 110 (GGGA)4
ARHGEF35	ENST00000378115.2	201	– 22 (TCCC)3
ARL17B	ENST00000622877.4	201	– 99 (CTCC)3
ATP7A	ENST00000343533.9	201	– 27 (GAGG)3
DRD2	ENST00000542616.1	207	– 54 (GCCC)3
DUX4	ENST00000565211.1	203	– 144 (GGCT)6
FAM83G	ENST00000388995.10	202	– 80 (TCTG)3
GTF2IRD2B	ENST00000614064.4	206	– 17 (GAAA)3
JCAD (*KIAA1462*)	ENST00000375377.1	201	– 13 (CCGG)3
MCTP2	ENST00000451018.7	203	– 102 (GTTT)3
METTL21C (*C13orf39*)	ENST00000267273.6	201	– 69 (CAGT)3
OR10G6	ENST00000307002.3	201	– 123 (GATA)13
PHYHD1	ENST00000308941.9	201	– 107 (TTTA)3
SAMD1	ENST00000269724.5	201	– 75 (CCGC)3
TEAD4	ENST00000540314.1	206	– 51 (CTCC)3
TRAJ49	ENST00000390488.1	201	– 124 (GCCT)7
TRDJ2	ENST00000390475.1	201	– 86 (CCAC)3
TRAV38-1	ENST00000390464.2	201	– 109 (CACC)3
TRAV7	ENST00000390429.3	201	– 111 (GACA)3

The numbers before the brackets represent the start site of the STR in respect of the corresponding transcription start site. "Variant no" corresponds to the Ensembl isoform number

NPAS1

```
Prairie      -----------------CACCTGCGACTCCTTCGCGCCTGAAT-------------TCTC
Dog          ----------------------CGGAGACCTGGGTTCCTGGCCTTGAGGGAGGGCAGCTC
Opossum      --------------------------CTCCTGGGTTCCTG--------------------
Zebrafish    ---CT---------CTTTTACTCTGTGTGCTAATCAGCATT-------T---------GA
Human        --AATGGGAGTCATGGCAGCCATGGTGGTCAG----------------------------
Gorilla      -GTCTCCAAGTCTGCGAGGCCGAGGTGG-----GCGCCGAG-------A---------GC
Mouse        GGTCTCAAAGTCTGCGAGGCCGCCGTGG-----GCGCCGAG-------A---------GC
Macaque      --TCTCCAAGTCTGCGAGGCCGAGGTGG-----GCGCCGAG-------A---------GC
Chimpanzee   -----------GGGCGGAGACGCGGCGGGCGGGGCTCCGGC-------A---------GA
Horse        ------------------------------------------------------------
```

DDX58

```
Gorilla      --------------------GCACTTTCGATTTTCCCTTTAGTTATTAAAGTTCCTATGC
Mouse        -------------------GGCAGTTTCGATTTCCTATGGC--CA--GC--CGGCTAGGC
Human        -------ACAGCCTGCGGGG--AACGTAGCTAGCTGCAAGC--AGAGGC--CGG------
Chimpanzee   ---------CTCCAGCTTGGGCAACATAGTGAGAC------------TC--CGT------
Prairie      -------------GAACTGTGCGGACTTTA-TAATACAAAC--CG---------------
Horse        GGCTTTAAAGGCAGCGGGGTGCGGGTGAG----------GC--TGGGAC--TCG------
Guinea Pig   -GC---CACCGCGGCCAGTTGCAGTTTCGGTTTCTCC--TC--TGT------AG------
Rabbit       -------------ATGAAGTGCAGTA------TATTCAGGC--TGAGAA--AAA------
```

DRD2

```
Mouse        -ACGGCTGCCGGAGGGGC------GGCCGTGC----GTGGA-----TGCGGCGGGAGCTG
Anole        GCCGGCTCT-----TGGCGCTTCATGCCAC--TAACTCCTCCA----CCTGCGTGACCTC
Armadillo    ---------------------------------------------CCTCCGTGGGGGCAG
Zebrafish    CGCTG---------CGCCTGG-CATACAGTACGCGCTCCTCTGCTCAC------------
Chicken      -----------CCGGTGCTGAATCCGCCGC--TCGCTCCGCAGCC--CCGGGG----CGG
Human        -ACTGCTCCCCGCGGGCCAGAGCCGGCCG------AGCTGCTGCCCGCCGGGG----CTC
Dog          ----------CGAAAGCCTCGAGCAGCCGG--CGCCGTCTCTGCC--CCCGGG----CGC
Bushbaby     -----------------CTGG-CCGCCCCG--TGGCTCCGCGGCC--CCG----------
Gibbon       ------------AGAGCCTGG-CCACCCAG--TGGCTCCACCGCC--CCG----------
Rabbit       --------------AGCCTGG-CCATCCAG--TGGCCCCGCCGCC--CCG----------
```

TEAD4

```
Mouse        ----------------------------------------------GCGAGGCGCCCCAC
Human        ---ATTCCAGCGTCCTCCTCGCACACTCGAGGCCAGGGG-GCGGGA-GGGCC-GCAGCTC
Dog          ACAGGTCCAACGGGCGCTCCTC-CAAGCGGAGCCTTGG-------AGGGCAA--------
Bushbaby     ------------GTTGGAGCG--CCGGCGGACTCCTTGGAACTGGCGTAGAG-CACCCAC
Xenopus      -----------GAGCAGAGCTT-TCAGG-AAGCCT-------------------------
Flycatcher   ---------------------------CGAATCCTC------TGTAGTAGCA-GT-----
Chicken      ----------------------------------AT------TAGATGGGAG-TCCCCTC
Armadillo    -------------------CTC-CCCGCGGGGCCGG------TGGCGGGGCC-GCCACTC
Gibbon       --------------CAGAGAGA-TCGTCGCGGCCGG------AAGAGTTGGC-GC---T-
Zebrafish    ----------------------------TTGGC--------------ATGGA-GC---T-
```

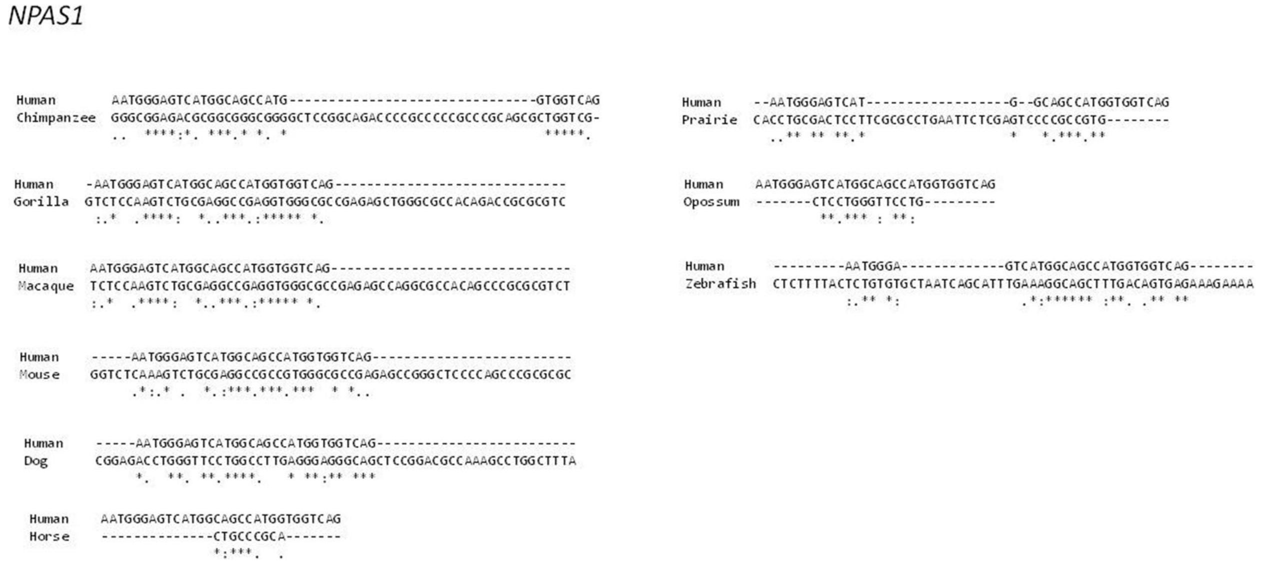
Fig. 3 Multiple sequence alignment of the TSS-flanking 5′UTRs. Examples of ClustAl Omega sequence alignment are represented in the tri- (**a**) and tetranucleotide (**b**) categories. Species inclusion was based on the information available in the Ensembl database

A number of the identified STRs were linked to non-canonical translation in the following genes, *TEAD4*, *ECSCR*, *MPRIP*, *PAQR9*, *PRSS1*, and *ZSCAN30.*

Discussion

There is an ever-growing literature on the biological and pathological implications of STRs at the inter- and intraspecies levels [16–27]. The STRs listed in the present study are genetic codes that are unique to humans and are likely to be responsible for the human-specific regulation of the relevant genes. The significant enrichment of human-specific transcripts at the human-specific STR compartment indicates a link to a mechanism for TSS selection and transcript specificity.

A number of the identified STRs such as GTTT have established repressor activity [6, 28, 29] and are differentially expanded in certain genes in the Old World



NPAS1

```
Human        AATGGGAGTCATGGCAGCCATG-----------------------------GTGGTCAG
Chimpanzee   GGCGGAGACGCGGCGGGCGGGGCTCCGGCAGACCCGCCCCCGCCCGCAGCGCTGGTCG-
             ..  ****:*. ***.* *. *                               *****.

Human        -AATGGGAGTCATGGCAGCCATGGTGGTCAG-----------------------------
Gorilla      GTCTCCAAGTCTGCGAGGCCGAGGTGGGCGCCGAGAGCTGGGCGCCACAGACCGCGCGTC
              :.*  .****:  *..***.:***** *.

Human        AATGGGAGTCATGGCAGCCATGGTGGTCAG------------------------------
Macaque      TCTCCAAGTCTGCGAGGCCGAGGTGGGCGCCGAGAGCCAGGCGCCACAGCCCGCGCGTCT
             :.*  .****:  *..***.:***** *.

Human        -----AATGGGAGTCATGGCAGCCATGGTGGTCAG-------------------------
Mouse        GGTCTCAAAGTCTGCGAGGCCGCCGTGGGCGCCGAGAGCCGGGCTCCCCAGCCCGCGCGC
                  .*:.* .  *.:***.***.***  * *..

Human        -----AATGGGAGTCATGGCAGCCATGGTGGTCAG-------------------------
Dog          CGGAGACCTGGGTTCCTGGCCTTGAGGGAGGGCAGCTCCGGACGCCAAAGCCTGGCTTTA
                  *.  **. **.****.   * **:** ***

Human        AATGGGAGTCATGGCAGCCATGGTGGTCAG
Horse        --------------CTGCCCGCA-------
                           *:***.  .

Human        --AATGGGAGTCAT------------------G--GCAGCCATGGTGGTCAG
Prairie      CACCTGCGACTCCTTCGCGCCTGAATTCTCGAGTCCCCGCCGTG--------
               ..** ** **.*                  *   *.***.**

Human        AATGGGAGTCATGGCAGCCATGGTGGTCAG
Opossum      -------CTCCTGGGTTCCTG---------
                    **.*** : **:

Human        ---------AATGGGA-------------GTCATGGCAGCCATGGTGGTCAG--------
Zebrafish    CTCTTTTACTCTGTGTGCTAATCAGCATTTGAAAGGCAGCTTTGACAGTGAGAAAGAAAA
                      :.** *:              .*:****** :**. .** **
```

Fig. 4 Pair-wise sequence comparison of the TSS-flanking 5′UTRs. %identity scoring was performed between human and other species. Asterisks represent sequence identity monkeys and Apes [14]. Purine STRs such as GAAA repeats are also functional in gene expression regulation, and their link to certain diseases unique to humans were previously reported [30, 31]. While the CG-rich STRs (e.g., CCG, GGGC) are subject to DNA methylation and can repress gene expression activity [32], they can also form G4 quadruplex structures, which have significant functions in gene expression regulation [33]. Several other identified STRs can form G4 structures with high overlap fraction (e.g., AGGG/CCCT, GCCC/GGGC).

It is not possible to estimate the number of crucial events that have led to the emergence of the human species. However, only a few genetic changes are needed to spur the evolution of new species in general, exemplified by the highly restricted initial divergence in butterfly hybridization models [34]. Accelerated evolution of a number of the identified genes in the present study (e.g., *DRD2*) has a well-established role in the origin of *Homo sapiens* [35]. Remarkably, a human-specific 7-amino acid transcript of this gene is flanked by a human-specific GCCC-repeat. Human-specific transcripts are increasingly recognized of having a role in the pathogenesis of diseases unique to the human species, such as schizophrenia [25, 36].

In a number of instances, not only the STR and the transcript, but also the gene containing these STRs and transcripts, were unique to humans, e.g., *AMY1C*, which is indicated in the evolution of the human phenotype during the Pleistocene [37].

For a number of the identified genes, sparse literature is available on the relevant function and pathways (e.g., *ARHGEF35*, *CXorf40A*, *C22orf24*, *TMEM99*, and *ARL17B*).

In a number of the identified genes, the STRs were linked to noncanonical (non-AUG) translation. Although the significance of this compartment is unknown for the most part, recent emerging data indicates likely biological functions [38].

The plasticity of STRs confers them unique ability to respond to adaptive evolutionary processes in a more efficient way than the quaternary codes provided by the SNSs. This potential aspect of STRs is vastly unknown at present, and it is expected that identification of STRs that have evolved differentially in humans vs. other species may pave the way for better understanding of the evolutionary implication of these highly mutable motifs.

This study warrants expansion to other vitally important gene regulatory sequences such as the distal promoter, 5′UTR, and 3′UTR. It is also necessary to sequence these STRs in characteristics and diseases that are unique to the human species. The recent reports of mass STR analysis using CRISPR/Cas9 [39] make it particularly more feasible to investigate STRs in the context of human evolution.

Conclusion

Our findings provide the first evidence of genome-scale skewing of STRs at a specific region of the human genome, and support a link between STRs and TSS selection/transcript specificity. The genes and STRs listed here may have a role in the divergence of humans from other species through the development of characteristics and phenotypes that are unique to the human species.

Abbreviations

SNS: Single nucleotide substitution; STR: Short tandem repeat; TF: Transcription factor; TSS: Transcription start site; UTR: Untranslated region

Funding

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

Availability of data and materials

Please contact author for data requests.

Authors' contributions

NN, FA, and TAA carried out the bioinformatics studies. AD and RS were the advisors of the project and 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

[1]Iranian Research Center on Aging, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran. [2]Department of Biostatistics, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran.

Received: 19 December 2017 Accepted: 20 March 2018
Published online: 05 April 2018

References

1. Press MO, Carlson KD, Queitsch C. The overdue promise of short tandem repeat variation for heritability. Trends Genet. 2014;30(11):504–12.
2. King DG. Evolution of simple sequence repeats as mutable sites. Adv Exp Med Biol. 2012;769:10-25.
3. Ohadi M, Mohammadparast S, Darvish H. Evolutionary trend of exceptionally long human core promoter short tandem repeats. Gene. 2012; 507(1):61–7.
4. Mohammadparast S, Bayat H, Biglarian A, Ohadi M. Exceptional expansion and conservation of a CT-repeat complex in the core promoter of PAXBP1 in primates. Am J Primatol. 2014;76(8):747–56.
5. Nikkhah M, Rezazadeh M, Khorshid HK, Biglarian A, Ohadi M. 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):109–14.
6. Rezazadeh M, Gharesouran J, Mirabzadeh A, Khorshid HK, 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.

7. Namdar-Aligoodarzi P, Mohammadparast S, Zaker-Kandjani B, Kakroodi ST, Vesiehsari MJ, Ohadi M. Exceptionally long 5′ UTR short tandem repeats specifically linked to primates. Gene. 2015;569(1):88–94.
8. Ohadi M, Valipour E, Ghadimi-Haddadan S, Namdar-Aligoodarzi P, Bagheri A, Kowsari A, et al. Core promoter short tandem repeats as evolutionary switch codes for primate speciation. Am J Primatol. 2015;77(1):34–43.
9. Sonay TB, Carvalho T, Robinson MD, Greminger MP, Krützen M, Comas D, et al. Tandem repeat variation in human and great ape populations and its impact on gene expression divergence. Genome Res. 2015;25(11):1591–9.
10. Li C, Lenhard B, Luscombe NM. Integrated analysis sheds light on evolutionary trajectories of young transcription start sites in the human genome. bioRxiv. 2017. https://doi.org/10.1101/192757.
11. Kramer M, Sponholz C, Slaba M, Wissuwa B, Claus RA, Menzel U, et al. Alternative 5′untranslated regions are involved in expression regulation of human heme oxygenase-1. PLoS One. 2013;8(10):e77224.
12. Li Y, Seidel K, Marschall P, Klein M, Hope A, Schacherl J, et al. A polymorphic microsatellite repeat within the ECE-1c promoter is involved in transcriptional start site determination, human evolution, and Alzheimer's disease. J Neurosci. 2012;32(47):16807–20.
13. Emamalizadeh B, Movafagh A, Darvish H, Kazeminasab S, Andarva M, Namdar-Aligoodarzi P, et al. The human RIT2 core promoter short tandem repeat predominant allele is species-specific in length: a selective advantage for human evolution? Mol Gen Genomics. 2017;292(3):611–7.
14. Khademi E, Alehabib E, Shandiz EE, Ahmadifard A, Andarva M, Jamshidi J, et al. Support for "disease-only" genotypes and excess of homozygosity at the CYTH4 primate-specific GTTT-repeat in schizophrenia. Genet Test Mol Biomarkers. 2017;21(8):485–90.
15. Bushehri A, Barez MM, Mansouri S, Biglarian A, Ohadi M. Genome-wide identification of human-and primate-specific core promoter short tandem repeats. Gene. 2016;587(1):83–90.
16. Valipour E, Kowsari A, Bayat H, Banan M, Kazeminasab S, Mohammadparast S, et al. Polymorphic core promoter GA-repeats alter gene expression of the early embryonic developmental genes. Gene. 2013;531(2):175–9.
17. Heidari A, Fam ZNS, Esmaeilzadeh-Gharehdaghi E, Banan M, Hosseinkhani S, Mohammadparast S, et al. Core promoter STRs: novel mechanism for inter-individual variation in gene expression in humans. Gene. 2012;492(1):195–8.
18. Bagshaw AT. Functional mechanisms of microsatellite DNA in eukaryotic genomes. Genome Biol Evol. 2017;9(9):2428–43.
19. Bagshaw AT, Horwood LJ, Fergusson DM, Gemmell NJ, Kennedy MA. Microsatellite polymorphisms associated with human behavioural and psychological phenotypes including a gene-environment interaction. BMC Med Genet. 2017;18(1):12.
20. Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet. 2016;48(1):22–9.
21. Hammock EA, Young LJ. Microsatellite instability generates diversity in brain and sociobehavioral traits. Science. 2005;308(5728):1630–4.
22. Carrat GR, Hu M, Nguyen-Tu M-S, Chabosseau P, Gaulton KJ, van de Bunt M, et al. Decreased STARD10 expression is associated with defective insulin secretion in humans and mice. Am J Hum Genet. 2017;100(2):238–56.
23. Abe H, Gemmell NJ. Evolutionary footprints of short tandem repeats in avian promoters. Sci Rep. 2016;6:19421.
24. Fondon JW, Hammock EA, Hannan AJ, King DG. Simple sequence repeats: genetic modulators of brain function and behavior. Trends Neurosci. 2008; 31(7):328–34.
25. 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; https://doi.org/10.1007/s00438-018-1415-8.
26. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet. 2018; https://doi.org/10.1038/nrg.2017.115.
27. Watts PC, Kallio ER, Koskela E, Lonn E, Mappes T, Mokkonen M. Stabilizing selection on microsatellite allele length at arginine vasopressin 1a receptor and oxytocin receptor loci. Proc Biol Sci. 2017;284(1869):20171896. https://doi.org/10.1098/rspb.2017.1896.
28. Andrioli LPM, Vasisht V, Theodosopoulou E, Oberstein A, Small S. Anterior repression of a Drosophila stripe enhancer requires three position-specific mechanisms. Development. 2002;129(21):4931–40.
29. Andrioli LP, Oberstein AL, Corado MS, Yu D, Small S. Groucho-dependent repression by sloppy-paired 1 differentially positions anterior pair-rule stripes in the Drosophila embryo. Dev Biol. 2004;276(2):541–51.
30. Darvish H, Heidari A, Hosseinkhani S, Movafagh A, Khaligh A, Jamshidi J, et al. Biased homozygous haplotypes across the human caveolin 1 upstream purine complex in Parkinson's disease. J Mol Neurosci. 2013;51(2):389–93.
31. Heidari A, Hosseinkhani S, Talebi S, Meshkani R, Esmaeilzadeh-Gharedaghi E, Banan M, et al. Haplotypes across the human caveolin 1 gene upstream purine complex significantly alter gene expression: implication in neurodegenerative disorders. Gene. 2012;505(1):186–9.
32. Quilez J, Guilmatre A, Garg P, Highnam G, Gymrek M, Erlich Y, et al. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. Nucleic Acids Res. 2016; 44(8):3750–62.
33. Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, Black MA, et al. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. PLoS One. 2013;8(2):e54710.
34. Kronforst MR, Hansen ME, Crawford NG, Gallant JR, Zhang W, Kulathinal RJ, et al. Hybridization reveals the evolving genomic architecture of speciation. Cell Rep. 2013;5(3):666–77.
35. Dorus S, Vallender EJ, Evans PD, Anderson JR, Gilbert SL, Mahowald M, et al. Accelerated evolution of nervous system genes in the origin of Homo sapiens. Cell. 2004;119(7):1027–40.
36. Li M, Jaffe AE, Straub RE, Tao R, Shin JH, Wang Y, et al. A human-specific AS3MT isoform and BORCS7 are molecular risk factors in the 10q24. 32 schizophrenia-associated locus. Nat Med. 2016;22(6):649–56.
37. Hardy K, Brand-Miller J, Brown KD, Thomas MG, Copeland L. The importance of dietary carbohydrate in human evolution. Q Rev Biol. 2015;90(3):251–68.
38. Na CH, Barbhuiya MA, Kim MS, Verbruggen S, Eacker SM, Pletnikova O, Troncoso JC, Halushka MK, Menschaert G, Overall CM, Pandey A. Discovery of noncanonical translation initiation sites through mass spectrometric analysis of protein N termini. Genome Res. 2018;28(1):25–36.
39. Shin G, Grimes SM, Lee H, Lau BT, Xia LC, Ji HP. CRISPR–Cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis. Nat Commun. 2017;8:14291.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central