Gene 587 (2016) 83–90

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Genome-wide identification of human- and primate-specific core promoter short tandem repeats

A. Bushehri [a], M.R. Mashhoudi Barez [c], S.K. Mansouri [d], A. Biglarian [e], M. Ohadi [b,*]

[a] Genetics Research Center, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran
[b] Iranian Research Center on Aging, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran
[c] Cell and Molecular Biology Research Center, Department of Anatomy and Biology, Faculty of Medicine, Shahid Beheshti University, Velenjak, Tehran, Iran
[d] Clinical Psychology Department, Faculty of Science and Research, Qazvin Azad University, Qazvin, Iran
[e] Department of Biostatistics, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran

ARTICLE INFO

Article history:
Received 30 January 2016
Received in revised form 23 March 2016
Accepted 19 April 2016
Available online 22 April 2016

Keywords:
Short tandem repeat
Core promoter
Human-specific
Primate-specific
AP-2alpha
Sp1
MZF
Genome-wide

ABSTRACT

Recent reports of a link between human- and primate-specific genetic factors and human/primate-specific characteristics and diseases necessitate genome-wide identification of those factors. We have previously reported core promoter short tandem repeats (STRs) of extreme length (≥6-repeats) that have expanded exceptionally in primates vs. non-primates, and may have a function in adaptive evolution. In the study reported here, we extended our study to the human STRs of ≥3-repeats in the category of penta and hexaucleotide STRs, across the entire human protein coding gene core promoters, and analyzed their status in several superorders and orders of vertebrates, using the Ensembl database. The ConSite software was used to identify the transcription factor (TF) sets binding to those STRs. STR specificity was observed at different levels of human and non-human primate (NHP) evolution. 73% of the pentanucleotide STRs and 68% of the hexanucleotide STRs were found to be specific to human and NHPs. AP-2alpha, Sp1, and MZF were the predominantly selected TFs (90%) binding to the human-specific STRs. Furthermore, the number of TF sets binding to a given STR was found to be a selection factor for that STR. Our findings indicate that selected STRs, the cognate binding TFs, and the number of TF set binding to those STRs function as switch codes at different levels of human and NHP evolution and speciation.

1. Introduction

Recent reports of the involvement of human- and primate-specific genetic factors in characteristics and diseases that are unique to those species (Schwarz et al., 2016; Aghajanirefah et al., 2016; Rezazadeh et al., 2015; Prescott et al., 2015; Reilly et al., 2015; Drews et al., 2013; Ohadi et al., 2012b; Esmaeilzadeh-Gharehdaghi et al., 2011; Farokhashtiani et al., 2011) necessitate whole-genome identification of those factors. Short tandem repeats (STRs), by virtue of their special mutational and functional qualities, have a major role in generating the genetic variation underlying adaptive evolution (For a review see King et al., 2006). Indeed, STRs in gene regulatory regions provide a large store of phenotypic variation, genotype–phenotype variation, and genotype-by-environment interactions (Gymrek et al., 2016; Press et al., 2014; Gemayel et al., 2010; Vinces et al., 2009; Hammock and Young, 2005). Following genome-wide comparative genomics studies of STRs that have reached exceptional length (≥6-repeats) in the critical core promoter and 5′ untranslated region (UTR) across human protein-coding genes, we have previously reported functional STRs that are differentially expanded in human and non-human primates (NHPs) vs. non-primates (Nikkhah et al., 2016; Namdar-Aligoodarzi et al., 2015; Rezazadeh et al., 2014; Mohammadparast et al., 2014). On the other end of the spectrum, certain STRs might have been subject to selective contraction in the process of human and NHP evolution (Ohadi et al., 2015). Through their effect on gene expression, certain STRs may function as evolutionary switch codes for adaptive evolution and speciation. This notion is strengthened by the observations that a number of STRs are identical in primates vs. non-primates, and the genes linked to those STRs are involved in characteristics that have diverged primates from other mammals (e.g. craniofacial development, neurogenesis, and spine morphogenesis) (Ohadi et al., 2015). The GTTT-repeat in the core promoter of the *CYTH4* gene is the first example of the association of a primate-specific STR with major neuropsychiatric disorders in human (Rezazadeh et al., 2015; Rezazadeh et al., 2014). In the study reported here, we extend our analysis from human core promoter STRs that are ≥6-repeats (Ohadi et al., 2012a), to STRs of ≥3-repeats, in the penta and hexa category, and report STRs that are specifically linked to human, or NHP evolution.

Abbreviations: MZF, myeloid zinc finger; NHP, non-human primate; STR, short tandem repeat; TF, transcription factor; TSS, transcription start site; UTR, untranslated region.

[*] Corresponding author at: The Iranian Research Center on Aging, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran.

http://dx.doi.org/10.1016/j.gene.2016.04.041
0378-1119/

Table 1
Density of pentanucleotide STRs of ≥3-repeats across human protein-coding gene core promoters.

STR	Number	Percentage	3 repeats	4 repeats	5 repeats	6 repeats	7 repeats	Human specific	%[a]	Primate trace	%[a]	Primate specific	%[a]
GGGGC	54	62.1%	41	7	2	3	1	4	14%	1	2%	16	28%
GCCCC	64		49	11	3	0	1	12		1		17	
GGCGG	8	4.7%	8	0	0	0	0	2	22%	1	11%	3	33%
CCGCC	1		1	0	0	0	0	0		0		0	
GGGGA	3	3.7%	2	1	0	0	0	1	14%	1	29%	0	43%
TCCCC	4		4	0	0	0	0	0		1		3	
CGGGG	4	3.2%	4	0	0	0	0	0	17%	0	—	1	17%
CCCCG	2		2	0	0	0	0	1		0		0	
GCCCA	4	3.2%	3	1	0	0	0	0	17%	1	17%	1	17%
TGGGC	2		2	0	0	0	0	1		0		0	
CCCCT	5	2.6%	2	2	0	1	0	1	20%	1	20%	3	60%
GGG + CC	3	2.6%	2	1	0	0	0	0	20%	0	—	3	60%
GG + CCC	2		2	0	0	0	0	1		0		0	
TTTTG	4	2.6%	3	1	0	0	0	0	—	0	—	2	60%
CAAAA	1		0	0	1	0	0	0		0		1	
GGGCG	3	2.1%	3	0	0	0	0	1	25%	1	25%	0	25%
CGCCC	1		1	0	0	0	0	0		0		1	
GGGAG	3	2.1%	3	0	0	0	0	0	—	0	—	1	50%
CTCCC	1		1	0	0	0	0	0		0		1	
GGGCA	2	2.1%	1	1	0	0	0	1	50%	0	—	1	50%
TGCCC	2		1	1	0	0	0	1		0		1	
CGCGG	3	1.6%	1	2	0	0	0	1	33%	0	—	2	67%
N + AAAA	2	1%	2	0	0	0	0	0	—	0	—	2	100%
GGGAA	2	1%	1	1	0	0	0	0	—	1	50%	1	50%
Others TCGGC	1	5.3%	1	0	0	0	0	1	10%	0	—	0	90%
CCCCA	1		0	1	0	0	0	0		0		1	
CGCCT	1		0	1	0	0	0	0		0		1	
CCCTC	1		1	0	0	0	0	0		0		1	
GCGGG	1		1	0	0	0	0	0		0		1	
ACGTG	1		1	0	0	0	0	0		0		1	
CTCCT	1		1	0	0	0	0	0		0		1	
GGAGA	1		1	0	0	0	0	0		0		1	
ACCCC	1		1	0	0	0	0	0		0		1	
GCCTC	1		1	0	0	0	0	0		0		1	
Total	190	100%	147	31	6	4	2	28	15%	9	5%	69	58%

[a] The percentage rate among the subgroup.

2. Materials and methods

2.1. Bioinformatics

The interval between −120 to +1 to the transcription start site (TSS) of the entire human protein-coding genes annotated in the GeneCards database was screened for penta and hexanucleotide STRs of ≥3-repeats, based on the Ensembl database (http://asia.ensembl.org/index.html), and using the Microsatellite Repeats Finder at the following link: http://biophp.org/minitools/microsatellite_repeats_finder/demo.php. The evolutionary status of the identified STRs was analyzed in 12 additional species, belonging to eight major superorders and orders across vertebrates, including Primates, Rodentia, Scandentia, Laurasiatheria, Afrotheria, Xenarthra, Birds, and Fish (Suppl 1 &Suppl 2).

Table 2
Density of hexanucleotide STRs of ≥3-repeats in the human protein-coding gene core promoters.

STR	Number	Percentage	3 repeats	4 repeats	6 repeats	>6 repeats	Human specific	%[a]	Primate trace	%[a]	Primate specific	%[a]
GGGGGC	7	31.6%	3	3	1	0	3	42%	0	8%	3	25%
GCCCCC	5		4	0	0	1	2		1		0	
GGGGCC	4	21.1%	2	2	0	0	3	50%	0	13%	1	38%
GGCCCC	4		2	2	0	0	1		1		2	
GCGGGC	3	7.9%	3	0	0	0	1	33%	1	33%	1	33%
GGAGGC	1	5.3%	1	0	0	0	1	50%	0	–	0	–
GCCTCC	1		0	1	0	0	0		0		0	
Others CCCTCT	1	34.2%	1	0	0	0	1	15%	0	15%	0	46%
CCCTCG	1		1	0	0	0	1		0		0	
CTCCCG	1		0	1	0	0	0		1		0	
GGGCGG	1		1	0	0	0	0		1		0	
GGCGGT	1		0	1	0	0	0		0		1	
ACCCCC	1		1	0	0	0	0		0		1	
TGGGGC	1		1	0	0	0	0		0		1	
AGGCCC	1		1	0	0	0	0		0		1	
GGGAAG	1		1	0	0	0	0		0		1	
CGCCCC	1		1	0	0	0	0		0		1	
GTCCCC	1		0	0	1	0	0		0		0	
CGGGGG	1		1	0	0	0	0		0		0	
TCCCCC	1		1	0	0	0	0		0		0	
Total	38	100%	25	10	2	1	13	34%	5	13%	13	34%

[a] The percentage rate among the subgroup.