Gene 628 (2017) 156–161

Contents lists available at ScienceDirect

Gene

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

Research paper

Simple sequence repeats showing 'length preference' have regulatory functions in humans

Jaya Krishnan[1], Fathima Athar[1], Tirupaati Swaroopa Rani, Rakesh Kumar Mishra*

Stowers Institute for Medical Research, MO, United States
International Crops Research Institute for the Semi-Arid Tropics, Hyderabad, India
CSIR-Centre for Cellular and Molecular Biology, Hyderabad, India

ARTICLE INFO

Keywords:
Simple sequence repeats
Microsatellites
Simple tandem repeats
Cis-regulatory activity
Gene regulation
Boundary elements
Barrier elements

ABSTRACT

Simple sequence repeats (SSRs), simple tandem repeats (STRs) or microsatellites are short tandem repeats of 1–6 nucleotide motifs. They are twice as abundant as the protein coding DNA in the human genome and yet little is known about their functional relevance. Analysis of genomes across various taxa show that despite the instability associated with longer stretches of repeats, few SSRs with specific longer repeat lengths are enriched in the genomes indicating a positive selection. This conserved feature of length dependent enrichment hints at not only sequence but also length dependent functionality for SSRs. In the present study, we selected 23 SSRs of the human genome that show specific repeat length dependent enrichment and analysed their cis-regulatory potential using promoter modulation, boundary and barrier assays. We find that the 23 SSR sequences, which are mostly intergenic and intronic, possess distinct cis-regulatory potential. They modulate minimal promoter activity in transient luciferase assays and are capable of functioning as enhancer-blockers and barrier elements. The results of our functional assays propose cis-gene regulatory roles for these specific length enriched SSRs and opens avenues for further investigations.

1. Introduction

Simple sequence repeats (SSRs), also known as microsatellites or simple tandem repeats (STRs), are short stretches of 1–6 nucleotide motifs repeated in tandem, with the repetitive unit generally occurring anywhere between 10 and 20 times. SSRs are present in both vertebrates and invertebrates and occur throughout the genome, in coding as well as the non-coding regions. About 3% of the human genome is comprised of the SSRs, which is almost twice the amount of protein coding DNA. Depending on the repeat, long stretches of SSRs are highly unstable once the repeat length threshold of 60–150 bp is reached. This instability, causing repeat length polymorphisms, is explained by strand slippage replication and faulty recombination which ultimately contribute to high mutation rates in SSR ($\sim 10^{-2}$–10^{-7} per cell division) (Kim and Mirkin, 2013). Repeat length polymorphisms are significant as they accelerate the rate of evolution of genes and are also useful in genetic mapping and linkage analysis studies (Moxon et al., 1994; Kashi and King, 2006). Aberrant expansions of SSRs are cause of nearly thirty hereditary neurodegenerative and developmental diseases, among which the triplet repeat expansion disorders are well studied (La Spada and Taylor, 2010).

Though little is known about the functions of SSRs, studies have proposed roles for this class of repetitive DNA in regulation of gene expression, DNA replication and repair, recombination, genome organisation and evolution (Field and Wills, 1996; Li et al., 2004; Kumar et al., 2010; Kumar et al., 2013). Genome-wide analysis of SSRs and gene expression changes in lymphoblastoid cells report that more than two thousand SSRs within 100 kb of transcriptional start and end sites of gene transcripts are capable of affecting gene expression. These SSRs acting as expression quantitative trait loci (eQTLs) are enriched in conserved regions, regulatory elements and regions marked with certain epigenetic marks. Few expression STRs (eSTRs) also associated with clinically relevant conditions (Gymrek et al., 2016). Another study showed that $^{>}$100 SSRs within ± 1 kb of transcription start site of genes in the HapMap population were associated with changes in expression (e)/methylation (m) levels of adjacent genes. These eSTRs/mSTRs overlapped with transcription factor binding and DNaseI hypersensitive sites (Quilez et al., 2016). It has been suggested that inter-species

Abbreviation: SSRs, simple sequence repeats; STRs, simple tandem repeats; eQTLs, expression quantitative trait loci; SNPs, single nucleotide polymorphisms; DMEM, Dulbecco's Modified Eagle Medium; FCS, fetal calf serum; hCALM1, human calmodulin; EGFR, epidermal growth factor receptor

* Corresponding author at: CSIR-Centre for Cellular and Molecular Biology, Uppal Road, Habsiguda, Hyderabad 500007, India.
E-mail address: mishra@ccmb.res.in (R.K. Mishra).

[1] Equal contribution.

http://dx.doi.org/10.1016/j.gene.2017.07.022
Received 30 December 2016; Received in revised form 18 May 2017; Accepted 10 July 2017
Available online 13 July 2017
0378-1119/

variations not completely explained by single nucleotide polymorphisms (SNPs) may be explained by polymorphisms in STR regions, and more significantly by the promoter associated STRs. 25% of protein coding genes in humans have STRs of ≥ 3 repeats in their core promoters. A small fraction of these promoters also have longer repeats of ≥ 6 which are evolutionarily conserved, functionally significant and contribute to variability among organisms (Ohadi et al., 2012).Variations in lengths of SSRs in promoter regions of Pax3/7 binding protein, PAXBP1 (CT-repeat), SGB2B2, a member of secretoglobins (CA-repeat) and cytohyesin-4 (CYTH4) (GTTT-repeat) have been shown to influence their gene expression and may have played vital role in evolution of primate species (Mohammadparast et al., 2014; Rezazadeh et al., 2014; Nikkhah et al., 2015).

Our previous analysis of abundance of each of 501 theoretically possible SSRs of different repeat unit lengths, across genomes of 24 organisms showed that the SSR abundance generally decreases with increase in their repetitive units. However, of the 501, 73 SSRs though following the same trend were additionally and unusually enriched at specific repeat lengths in different organisms (45 bp was optimally preferred).We referred to this feature as 'length preference'. This feature of preferential length dependent enrichment was conserved across taxa suggesting that these SSRs have been positively selected for by nature not only based on their sequence but also repeat lengths (Ramamoorthy et al., 2014). We hypothesised that these SSRs may have functional significance and maybe involved in gene regulation. In the present study, we systematically analysed 23 SSRs occurring in the human genome, which are a subset of the 73 SSRs identified across various organisms. These 23 SSRs show specific length preference and constitute about half of the total SSRs of the human genome (unpublished data). We analysed these SSRs for cis-regulatory potential using well established assays in cell lines of human origin. Our results reveal that many of the 23 SSRs are capable of modulating promoter activity in transient assays and are also capable of showing boundary activity.

2. Materials and methods

2.1. SSR oligonucleotides, cell culture and transfection

SSRs were synthesised (Eurofins Genomics, Bangalore, India) as oligonucleotides of specified lengths (Table 1) and were cloned into respective assay plasmids by ligation. The following human cell lines were used in the study: IMR-32 (neuroblastoma cell line), MCF-7 (breast adenocarcinoma cell line), HeLa (cervix adenocarcinoma cell line), HEK293T (human embryonic kidney cell line) and K562 (erythromyeloblastoidleukemia cell line). Cells were grown in Dulbecco's Modified Eagle Medium (DMEM) (Invitrogen) supplemented with 10% FCS (fetal calf serum) and antibiotics (penicillin, streptomycin and kanamycin) and were maintained in a humidified incubator at 37 °C and 5% CO_2. For the assays, lipid mediated transfection was carried out using Lipofectamine® 2000 (Invitrogen). Briefly, cells were seeded 18–24 h before transfection in 24 or 6-well plates. 400 ng or 1 μg of DNA was used for transfection and cells were incubated in the DNA-lipid transfection mix. After 3–4 h the transfection mix was replaced with fresh medium. Cells were harvested after 24 h for luciferase assays or transferred to fresh medium containing drugs (G418 or blasticidin) for drug selection.

2.2. Luciferase assay for modulation of promoter activity

SSRs were cloned in pGL3-promoter vector which contains a SV40 promoter upstream of the luciferase gene. Cells were plated in 24-well plates 18–24 h before transfection. 200–400 ng of test constructs or empty vector control were co-transfected with ~10 ng or less of pRL-TK (Renilla luciferase). Cells were harvested after 24 h of transfection, washed twice with PBS and lysed using Passive lysis buffer (Dual Glo luciferase assay system, Invitrogen) and the assay was performed according to the manufacturer's instructions. Luciferase activity of pRL-TK was used to normalise transfection efficiency. Relative luciferase activity was determined by normalising the ratios of firefly to Renilla luciferase activities of test constructs to that of the empty vector. A fragment of the mouse HoxD PRE region (mHoxPRE-FI) which modulated promoter activity positively in our previous study was used as a positive control (Vasanthi et al., 2013). Averages of 3–4 independent experiments along with their standard errors of mean are expressed. Statistical significance calculated using Student's *t*-test is shown ($p < 0.05$ *, < 0.01 **, < 0.001 ***, < 0.0001 ****).

Table 1
Summary of promoter modulation and boundary assays in K562 cell line.

S. no.	SSR	Repeat units/size (bp)	Promoter modulation assay	Boundary assay	Barrier assay
1	A	36/36	-*	-	NA
2	AT	21/42	↓↓ 0.58***	√	-
3	AAG	19/57	-	-	NA
4	AAT	14/42	↑↑ 2.35**	-	NA
5	ATC	12/36	-	√	-
6	AGAT	10/40	↑ 1.54**	√	-
7	AAAG	13/52	-	-	NA
8	AAAT	10/40	-	√	-
9	AAGG	11/44	↑↑↑ 2.72***	-	NA
10	ACAT	10/40	↑↑↑ 2.91***	-	NA
11	ATCC	9/36	-	-	NA
12	AAAAG	11/55	↑↑↑ 2.73**	-	NA
13	AAAAT	8/40	↑↑ 2.34***	√	-
14	AAAGG	12/60	↑↑↑ 3.12***	-	NA
15	AACAT	10/50	-	√	w
16	AAGAG	12/60	↑↑ 2.12***	√	-
17	AAGGG	11/55	↑ 1.76***	-	-
18	AATAC	12/60	↑↑ 2.36***	√	-
19	AATAG	11/55	↓↓ 0.51****	-	NA
20	AATAT	9/45	↑↑ 2.05***	-	NA
21	AATGG	8/40	↑ 1.63**	-	-
22	ACATAT	8/48	↑ 1.95**	-	NA
23	AGATAT	7/42	-	-	NA
24	Positive controls[a]		↑ 1.84**	√	

Maximally enriched repeat number of the SSRs in the human genome as determined in the previous study (Ramamoorthy et al., 2014) and their corresponding length in bp are shown. Promoter modulation and boundary assays in K562 cell line are summarised. Upward arrows indicate basal promoter activity ≥ 1.5 fold while lower arrows indicate promoter activity ≤ 0.8 fold (↑ 1.5–2, ↑↑ 2–2.5, ↑↑↑ > 2.5, ↓ 0.8–0.6, ↓↓ 0.6–0.4, ↓↓↓ < 0.4-fold). √ - positive boundary activity; '-' no change when compared to control.

[a] Positive controls used in the assays - mHoxPRE-FI for promoter modulation assay and β-globin boundary element for boundary assay.

* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.
**** $p < 0.0001$.

2.3. Boundary assay

Enhancer blocker or boundary assay was carried out by colony formation assay using K562 cells as described previously (Chung et al., 1993). Briefly, vectors were generated each with an SSR oligonucleotide inserted between the mHS2 enhancer and human γ-globin promoter controlling the expression of neomycin resistance gene (*neo*^{r}) in a slightly modified form of the parent vector, pJC54. A chicken β-globin insulator element present downstream of the neomycin resistance gene protects it from the position effects in the genome. Equal numbers of K562 cells seeded in 6-well plates were transfected with 5 μg of each test construct along with 1 μg of vector expressing blasticidin drug resistance gene as a control for transfection. After 24–36 h of transfection, cells were plated in duplicates in soft agar medium. Number of neomycin-resistant colonies was determined from at least three view-fields