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

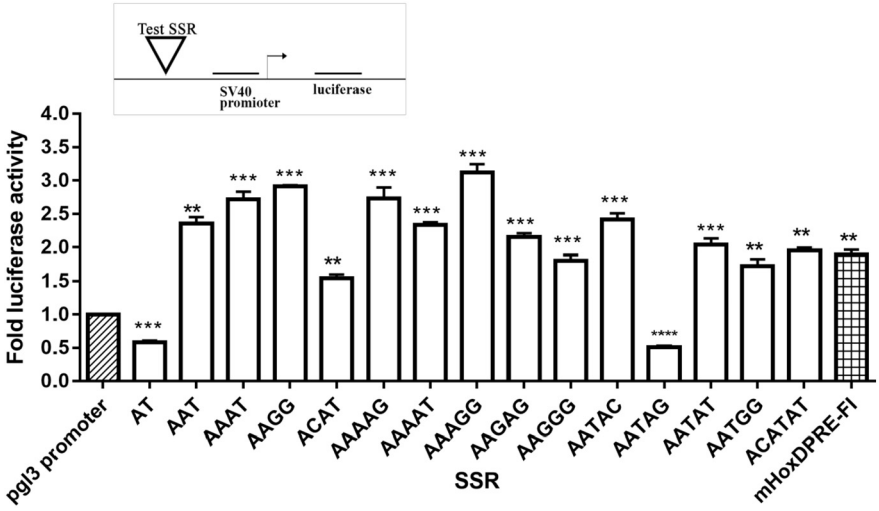

Fig. 1. Promoter modulation assay for the 23 human SSRs in K562 cell line.
Promoter modulation assay was carried out in K562 cell line. SSRs were cloned upstream of SV40 promoter which regulates the luciferase gene. Luciferase activity was determined, normalised to vector control. Statistically significant results are represented ($p < 0.01$ **, < 0.001 ***, < 0.0001 ****).

in each well. Surviving colonies in G418 medium for each construct relative to the empty vector, normalised for transfection differences (using the number of blasticidin resistant colonies) are represented. β-Globin boundary element was taken as a positive control. Test constructs showing values equal to or below a cut off of 0.7 were concluded as possessing boundary activity. Average of results from at least two independent experiments and their standard errors of mean are shown. Statistical significance calculated using Students *t*-test is represented ($p < 0.05$ *, < 0.01 **, < 0.001 ***, < 0.0001 ****).

2.4. Barrier assay

The vector used for barrier assays, pFSBS was generated by modifying the vector LMBP4800 (Heyninck and Beyaert, 1999). pFSBS contains a GFP reporter gene driven by CAGG promoter under control of a CMV-IE enhancer. A β-globin insulator is cloned upstream of the enhancer while the test SSRs are cloned in the MCS downstream of the GFP reporter. The vector also contains a blasticidin drug resistance gene after the MCS for selection of stable transfectants. K562 cells were used for this assay. Cells were transfected with vector control or test constructs. Stable transfectants were selected for about 2–3 weeks. Following this, drug was withdrawn and disappearance of GFP was monitored by FACS (fluorescence assisted cell sorting) using MoFlo (DakoCytomation) at regular intervals. Percent GFP positive cells were normalised to the empty vector control at each time point. A construct having chicken β-globin insulator elements cloned on either side was used as a positive control (B3). Results from at least two independent experiments along with their standard deviations are expressed. Statistical significance is calculated using Students *t*-test ($p < 0.05$ *, < 0.01 **, < 0.001 ***, < 0.0001 ****).

3. Results

Our previous analysis of SSRs in 24 organisms across the evolutionary lineage identified 73 SSRs that showed specific repeat length dependent enrichment in the genomes (Ramamoorthy et al., 2014). We chose to study a subset of the 73 SSRs, specifically, the 23 SSRs occurring in the human genome. We determined the fraction of the total SSRs that is constituted by these 23 SSRs in the human genome. Our bioinformatics analysis revealed that the 23 SSRs showing length preference make up a major portion (almost 50%) of the SSRs in the human genome, while the rest of the 478 SSRs, constitute the other 50% (unpublished data). Thus, the 23 SSRs showing length preference are among the most abundant SSRs in the human genome. The 23 SSRs include one monomer (A), a dimer (AT) three trimers (AAG, AAT and ATC), six tetramers (AAAG, AAGG, ACAT, AGAT, AAAT and ATCC), ten pentamers (AAAAG, AAAGG, AAGGG, AAAAT, AACAT, AAGAG, AATAC, AATAG, AATAT and AATGG) and two hexamers (ACATAT and AGATAT). Majority of these SSRs are AT rich. Analysis of their distribution in the genome revealed that they are mostly in the intronic and intergenic regions. The optimally preferred repeat length as assessed from their repeat length *versus* abundance plots was ~45 bp (Ramamoorthy et al., 2014). Conservation of sequence and repeat length preferences suggests that these SSRs have functional significance. Many reports have suggested a transcriptional regulatory role for SSRs (Hoffman et al., 1990; Lafyatis et al., 1991; Sandaltzopoulos et al., 1995; Gymrek et al., 2016; Quilez et al., 2016). Hence, we analysed transcriptional cis-regulatory potential of these 23 SSRs using well-established promoter modulation, boundary and barrier assays. For each of the 23 SSRs, the repeat lengths used in the assays corresponded to their optimally preferred repeat lengths in the genome (Table 1).

The 23 SSRs are mostly intergenic or intronic. We hypothesised that these repeats are likely to influence promoter activities rather than act as promoters themselves. To test this, each of the 23 SSRs was cloned upstream of the SV40 promoter in pGL3-promoter vector. An increase or decrease in basal luciferase activity was used as an indicator of promoter modulating activity. SSRs showing luciferase activity of 1.5-fold and higher compared to the vector control were considered as positive modulators of promoter activity, while those showing an activity of 0.8-fold and lower were considered as modulating promoter activity negatively. The assay was performed in five human cells lines (IMR-32, MCF7, HeLa, HEK293 and K562) and results obtained in K562 cell line are shown (Fig. 1). Results indicate that all SSRs are capable of modulating promoter activity in at least one cell line (Supplementary Table 1). In the K562 cells, SSRs AT and AATAG influenced promoter activity negatively while AAG, AAAT, AAGG, AAAAG, AAAAT, AAAGG, AAGAG, AATAC and AATGG modulated promoter activity positively by ≥ 2-fold (Fig. 1).In addition to K562, AT and AATAG also decreased promoter activity while AAAT, AAAAG, AAAGG increased promoter activity in at least four other cell lines that we tested. Other SSRs including AAG, AAGG, AAAAT, AAGAG, AATAC and AATGG modulated promoter activity in a cell-type specific manner. For example, AAG, AAGG, AAAAT, AAGAG, AATAC and AATGG modulated promoter activity positively in K562 cells but did not show similar activity in other cell lines (Supplementary Table 1).

Enhancer-blocker or boundary elements prevent aberrant interactions between unrelated enhancers and promoters. Such elements can be identified by an assay based on the principle that when a boundary element is placed between an enhancer and a promoter driving a reporter gene, it can prevent communication between them and thus decrease the expression of the downstream reporter gene (Chung et al.,

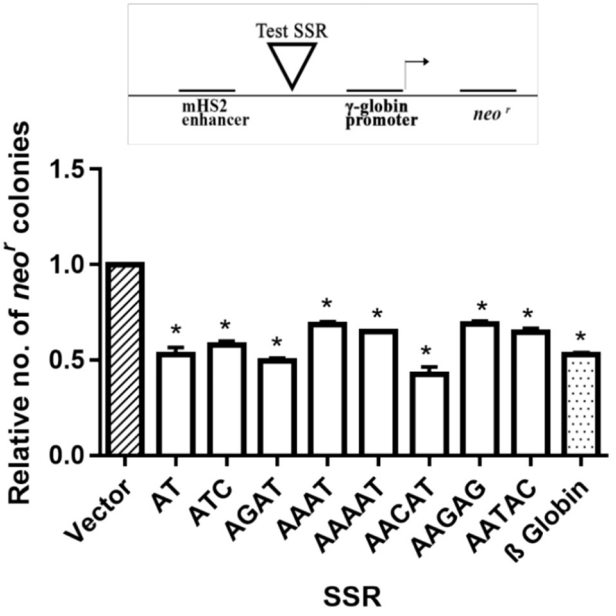

Fig. 2. Boundary assay for the 23 human SSRs in K562 cell line.
SSRs were cloned between the mHS2 enhancer and human γ-globin promoter controlling the expression of neomycin resistance gene in pJC54 vector. Numbers of neomycin resistant colonies were determined and normalised to vector control. Test constructs showing values equal to or below a cut off of 0.7 were considered as possessing boundary activity ($p < 0.05$ *).

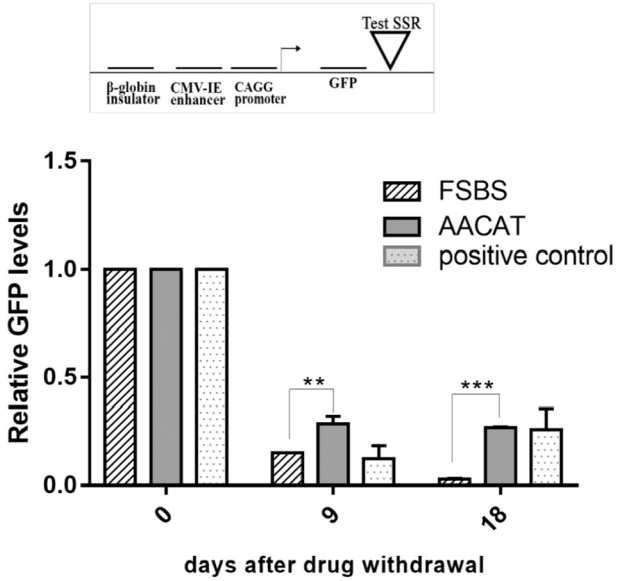

Fig. 3. Barrier assay in K562 cell line.
Barrier assay was carried out with select boundary positive SSRs in K562 cell line. GFP levels of stable transfectants were monitored by FACS at regular intervals following drug withdrawal. Relative GFP levels with their standard deviations at each time point are represented ($p < 0.01$ **, < 0.001 ***).

1993). SSRs were tested for their ability to abrogate communication between mHS2 enhancer and γ-globin promoter controlling neomycin resistance gene in an *in vitro* assay in K562 cells. Using the β-globin insulator element as a positive control, boundary assay was carried out. SSRs AT, ATC, AGAT, AAAT, AAAAT, AACAT, AAGAG and AATAC showed significant boundary activity in this assay (Fig. 2).

Barriers are elements that prevent spreading of heterochromatin into the euchromatin domains. Many known boundary elements like the β-globin boundary in mammals and t-DNA in yeast function as both boundaries and barriers (Chung et al., 1993; Donze et al., 1999). Since eight SSRs showed boundary function, they were tested for barrier activity. The barrier assay is based on the principle that transgenes in the genome are silenced over a period of time by the gradual spread of heterochromatin. However, a transgene flanked by barrier elements should prevent or delay this silencing. Using GFP as a reporter gene the ability of SSRs showing boundary activity to act as barrier elements was tested in K562 cell line (Fig. 3). Of the eight SSRs tested, AACAT showed weak barrier activity while others showed negligible barrier activity.

4. Discussion

Though simple sequence repeats (SSRs) comprise about 3% of the human genome, very little is known about their functional relevance. These tandem repeats of motifs ranging from1–6 bp occur throughout the genome, both in coding and non-coding regions. Though longer repeat lengths are quite unstable, few SSRs showing preference for a specific optimal repeat length have been positively selected in various organisms. In order to decipher their functional relevance, the 23 SSRs of the human genome showing enrichment at specific repeat lengths were analysed in this study. The specific optimal length at which these SSRs are enriched in the genome was chosen as the length of the test SSRs in our assay constructs. Hypothesising a cis-regulatory function, investigations pertaining to the ability of these SSRs to modulate promoter activity and to function as boundary and barrier elements were carried out. Our results indicate that these 23 SSRs of the human genome have the potential to act as cis-regulatory elements.

SSR elements have been shown to regulate DNA replication and repair, recombination and gene expression. SSRs present in the regulatory regions of genes involved in the cell cycle and genes involved in correcting replication errors, the DNA MMR genes, influence their expression (Chang et al., 2001). SSRs are hotspots for DNA recombination. Dinucleotide repeats capable of forming stable secondary structures bind to Rad family of recombination proteins and influence the process of recombination (Biet et al., 1999). Significant association between frequency of recombination and the GT repeat in chromosome 22 is found in humans (Majewski and Ott, 2000). SSRs in or near promoter region of genes influence their gene expression. The homopurine/homopyrimidine tract in heat shock protein, Hsp26 (Sandaltzopoulos et al., 1995), $(TCCC)_n$ repeat in promoter of c-KI-ras and TGF-β3 (Hoffman et al., 1990; Lafyatis et al., 1991), $(CAG)_7$ repeat in 5′ UTR of human calmodulin (hCALM1) gene (Toutenhoofd et al., 1998), TCAT repeats in the first intron of tyrosine hydroxylase gene (Meloni et al., 1998) and the $(CA)_n$ tract in first intron of epidermal growth factor receptor (EGFR) gene (Gebhardt et al., 1999) influence the activities of their corresponding promoters. Not only repeat motifs but also repeat lengths influence SSR function. Frequency of RecA dependent homologous recombination *in vitro* decreases with increase in repeat length of GT SSR (Dutreix, 1997). Transcriptional activity of EGFR gene declines with increasing number of CA repeats (Gebhardt et al., 1999). Pax6 gene promoter activity in human brain is 8–10 fold higher for the > 29 repeat variant of $(AC)_m$ $(AG)_n$ when compared to its 26-repeat variant (Okladnova et al., 1998). SSRs also serve as binding sites for regulatory proteins. Many proteins/protein complexes from nuclear extracts of cells have been shown to bind SSRs *in vitro* (Aharoni et al., 1993; Epplen et al., 1993). Interestingly, SSR repeat numbers also influence protein binding (Solomon et al., 1986; Winter and Varshavsky, 1989). SSRs influence epigenetic regulation of genes. DNA methylation is observed at CCG repeat associated loci where the CCG repeat secondary structures serve as excellent substrates for DNA methyltransferases (Smith et al., 1994). The CCG repeat expansion of FXS allele is associated with increased H3K9me2 (Pietrobono et al., 2005) while AAG repeat has been found to be associated with silent histone modification marks (Greene et al., 2007; Al-Mahdawi et al., 2008). The GATA repeat functions as an enhancer blocker in *Drosophila* and human cells (Kumar et al., 2013) while AAGAG is essential for *Drosophila*

development and its transcripts are found to be associated with the nuclear matrix (Pathak et al., 2013). More recently, a genome-wide analysis of SSRs and gene expression changes in lymphoblastoid cells showed that more than two thousand SSRs within 100 kb of transcriptional start and end sites of gene transcripts are capable of affecting gene expression (Gymrek et al., 2016).

Given the various reports on the gene regulatory potential of SSRs, we sought to analyse if the 23 SSRs of the human genome, each showing specific length preference, have cis-regulatory potential. The 23 SSRs were intronic or intergenic and were not found very close to the transcription start sites of genes. We therefore hypothesised that these repeats were more likely to influence (enhance or repress) promoter activities rather than act as promoters themselves. The promoter modulating activity of the SSRs was assessed by their ability to increase or decrease the activity of a minimal promoter driving *luciferase* gene. Though transient assays may not be a true reflection of the SSR function in genomic context, this assay still has the advantage of exposing the minimal cis-regulatory activity associated with these DNA sequences. The promoter luciferase assay showed that almost all of the 23 SSRs possess cis-regulatory activity in at least one cell line (Supplementary Table 1). Some SSRs exclusively modulated promoter activity positively or negatively in more than one cell line while others did so in a cell-type specific manner. Eight of the SSRs were found to be positive for boundary activity in our boundary assay with 23 SSRs. We correlated promoter modulation assay and boundary assay carried out in the same cell line, K562. We observed that SSRs AAAT, AAAAT, AAGAG and AATAC which were positive for boundary activity also modulated promoter activity negatively, suggesting that these may be repressor elements rather than boundary elements. A, ATC, AGAT and AACAT did not alter promoter luciferase activity but were positive for boundary activity, suggesting that these are likely to be true boundary elements. Few other SSRs positive for boundary activity increased promoter activity in the luciferase assay. Though this is intriguing, it may be possible that the SSR elements may function to maintain an open chromatin structure and/or recruit DNA binding proteins which may facilitate transcription as previously reported (Soeller et al., 1993). However, further analysis would confirm if these SSRs indeed are true boundary elements. Few other SSRs decreased promoter activity of the SV40 promoter (promoter modulation assay) but not the γ-globin promoter (boundary assay) suggesting that SSRs may function not only in a cell-type specific but also in a promoter specific manner. Some of the established boundary elements like the β-globin boundary in mammals and t-DNA in yeast also function as barriers. Hence, we carried out barrier assay for the boundary positive SSRs. We analysed a population of stable transfectants to alleviate any bias arising due to insertions at active/inactive genomic loci which might lead to clonal variations in our observations. Results of the barrier assay indicated that most of the tested SSRs show negligible barrier activity albeit for one, AACAT, which seemed to show weak barrier activity in our study.

Thus, our analysis postulates a cis-regulatory potential for the 23 SSRs of the human genome that show specific length preference. To our knowledge this is the first study that systematically analyses 23 SSRs of the human genome occurring in intergenic and intronic regions for their biological function and opens avenues for further investigations. Our results suggest minimal and cell-type specific cis-regulatory potential for these DNA repeat sequences. However, it is of great interest and importance to study the functions of these SSRs and the relevance of their repeat lengths in their native genomic location and chromatin contexts. Most of these SSRs have been found to occur in nucleosome free regions (unpublished data), suggesting that they may have specific roles in regulation of DNA replication and transcription. It is also very likely that these SSRs have roles in genome organisation, alternative splicing, modulation of chromatin structure and transcription of non-coding RNAs. Discovery and analysis of specific proteins binding to these SSRs is also likely to provide valuable insights into their functions.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2017.07.022.

Acknowledgements

We thank Gary Felsenfeld for his kind gift of boundary assay plasmid and S. Krishnan for the barrier assay plasmid. We acknowledge financial support of Council of Scientific and Industrial Research (CSIR) through Network grants (BSC0118 and BSC0121).

References

Aharoni, A., Baran, N., Manor, H., 1993. Characterization of a multisubunit human protein which selectively binds single stranded d(GA)n and d(GT)n sequence repeats in DNA. Nucleic Acids Res. 21, 5221–5228.
Al-Mahdawi, S., Pinto, R.M., Ismail, O., Varshney, D., Lymperi, S., Sandi, C., Trabzuni, D., Pook, M., 2008. The Friedreich ataxia GAA repeat expansion mutation induces comparable epigenetic changes in human and transgenic mouse brain and heart tissues. Hum. Mol. Genet. 17, 735–746.
Biet, E., Sun, J., Dutreix, M., 1999. Conserved sequence preference in DNA binding among recombination proteins: an effect of ssDNA secondary structure. Nucleic Acids Res. 27, 596–600.
Chang, D.K., Metzgar, D., Wills, C., Boland, C.R., 2001. Microsatellites in the eukaryotic DNA mismatch repair genes as modulators of evolutionary mutation rate. Genome Res. 11, 1145–1146.
Chung, J.H., Whiteley, M., Felsenfeld, G., 1993. A 5′ element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in *Drosophila*. Cell 74, 505–514.
Donze, D., Adams, C.R., Rine, J., Kamakaka, R.T., 1999. The boundaries of the silenced HMR domain in *Saccharomyces cerevisiae*. Genes Dev. 13, 698–708.
Dutreix, M., 1997. (GT)n repetitive tracts affect several stages of RecA-promoted recombination. J. Mol. Biol. 273, 105–113.
Epplen, C., Melmer, G., Siedlaczck, I., Schwaiger, F.W., Maueler, W., Epplen, J.T., 1993. On the essence of “meaningless” simple repetitive DNA in eukaryote genomes. EXS 67, 29–45.
Field, D., Wills, C., 1996. Long, polymorphic microsatellites in simple organisms. Proc. Biol. Sci. 263, 209–215.
Gebhardt, F., Zanker, K.S., Brandt, B., 1999. Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. J. Biol. Chem. 274, 13176–13180.
Greene, E., Mahishi, L., Entezam, A., Kumari, D., Usdin, K., 2007. Repeat-induced epigenetic changes in intron 1 of the frataxin gene and its consequences in Friedreich ataxia. Nucleic Acids Res. 35, 3383–3390.
Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., Daly, M.J., Price, A.L., Pritchard, J.K., Sharp, A.J., Erlich, Y., 2016. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat. Genet. 48, 22–29.
Heyninck, K., Beyaert, R., 1999. The cytokine-inducible zinc finger protein A20 inhibits IL-1-induced NF-kB activation at the level of TRAF6. FEBS Lett. 442, 147–150.
Hoffman, E.K., Trusko, S.P., Murphy, M., George, D.L., 1990. An S1 nuclease-sensitive homopurine/homopyrimidine domain in the c-Ki-ras promoter interacts with a nuclear factor. Proc. Natl. Acad. Sci. U. S. A. 87, 2705–2709.
Kashi, Y., King, D.G., 2006. Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22, 253–259.
Kim, J.C., Mirkin, S.M., 2013. The balancing act of DNA repeat expansions. Curr. Opin. Genet. Dev. 23, 280–288.
Kumar, R.P., Senthilkumar, R., Singh, V., Mishra, R.K., 2010. Repeat performance: how do genome packaging and regulation depend on simple sequence repeats? BioEssays 32, 165–174.
Kumar, R.P., Krishnan, J., Pratap Singh, N., Singh, L., Mishra, R.K., 2013. GATA simple sequence repeats function as enhancer blocker boundaries. Nat. Commun. 4, 1844.
La Spada, A.R., Taylor, J.P., 2010. Repeat expansion disease: progress and puzzles in disease pathogenesis. Nat. Rev. Genet. 11, 247–258.
Lafyatis, R., Denhez, F., Williams, T., Sporn, M., Roberts, A., 1991. Sequence specific protein binding to and activation of the TGF-beta 3 promoter through a repeated TCCC motif. Nucleic Acids Res. 19, 6419–6425.
Li, Y.C., Korol, A.B., Fahima, T., Nevo, E., 2004. Microsatellites within genes: structure, function, and evolution. Mol. Biol. Evol. 21, 991–1007.
Majewski, J., Ott, J., 2000. GT repeats are associated with recombination on human chromosome 22. Genome Res. 10, 1108–1114.
Meloni, R., Albanese, V., Ravassard, P., Treilhou, F., Mallet, J., 1998. A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element in vitro. Hum. Mol. Genet. 7, 423–428.
Mohammadparast, S., Bayat, H., Biglarian, A., Ohadi, M., 2014. Exceptional expansion and conservation of a CT-repeat complex in the core promoter of PAXBP1 in primates. Am. J. Primatol. 76, 747–756.
Moxon, E.R., Rainey, P.B., Nowak, M.A., Lenski, R.E., 1994. Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr. Biol. 4, 24–33.
Nikkhah, M., Rezazadeh, M., Khorram Khorshid, H.R., Biglarian, A., Ohadi, M., 2015. An exceptionally long CA-repeat in the core promoter of SCGB2B2 links with the evolution of apes and Old World monkeys. Gene 576, 109–114.
Ohadi, M., Mohammadparast, S., Darvish, H., 2012. Evolutionary trend of exceptionally long human core promoter short tandem repeats. Gene 507, 61–67.

Okladnova, O., Syagailo, Y.V., Tranitz, M., Stober, G., Riederer, P., Mossner, R., Lesch, K.P., 1998. A promoter-associated polymorphic repeat modulates PAX-6 expression in human brain. Biochem. Biophys. Res. Commun. 248, 402–405.

Pathak, R.U., Mamillapalli, A., Rangaraj, N., Kumar, R.P., Vasanthi, D., Mishra, K., Mishra, R.K., 2013. AAGAG repeat RNA is an essential component of nuclear matrix in *Drosophila*. RNA Biol. 10, 564–571.

Pietrobono, R., Tabolacci, E., Zalfa, F., Zito, I., Terracciano, A., Moscato, U., Bagni, C., Oostra, B., Chiurazzi, P., Neri, G., 2005. Molecular dissection of the events leading to inactivation of the FMR1 gene. Hum. Mol. Genet. 14, 267–277.

Quilez, J., Guilmatre, A., Garg, P., Highnam, G., Gymrek, M., Erlich, Y., Joshi, R.S., Mittelman, D., Sharp, A.J., 2016. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. Nucleic Acids Res. 44, 3750–3762.

Ramamoorthy, S., Garapati, H.S., Mishra, R.K., 2014. Length and sequence dependent accumulation of simple sequence repeats in vertebrates: potential role in genome organization and regulation. Gene 551, 167–175.

Rezazadeh, M., Gharesouran, J., Mirabzadeh, A., Khorram Khorshid, H.R., Biglarian, A., Ohadi, M., 2014. A primate-specific functional GTTT-repeat in the core promoter of CYTH4 is linked to bipolar disorder in human. Prog. Neuro-Psychopharmacol. Biol. Psychiatry 56, 161–167.

Sandaltzopoulos, R., Mitchelmore, C., Bonte, E., Wall, G., Becker, P.B., 1995. Dual regulation of the *Drosophila* hsp26 promoter in vitro. Nucleic Acids Res. 23, 2479–2487.

Smith, S.S., Laayoun, A., Lingeman, R.G., Baker, D.J., Riley, J., 1994. Hypermethylation of telomere-like foldbacks at codon 12 of the human c-Ha-ras gene and the trinucleotide repeat of the FMR-1 gene of fragile X. J. Mol. Biol. 243, 143–151.

Soeller, W.C., Oh, C.E., Kornberg, T.B., 1993. Isolation of cDNAs encoding the *Drosophila* GAGA transcription factor. Mol. Cell. Biol. 13, 7961–7970.

Solomon, M.J., Strauss, F., Varshavsky, A., 1986. A mammalian high mobility group protein recognizes any stretch of six A.T base pairs in duplex DNA. Proc. Natl. Acad. Sci. U. S. A. 83, 1276–1280.

Toutenhoofd, S.L., Garcia, F., Zacharias, D.A., Wilson, R.A., Strehler, E.E., 1998. Minimum CAG repeat in the human calmodulin-1 gene 5′ untranslated region is required for full expression. Biochim. Biophys. Acta 1398, 315–320.

Vasanthi, D., Nagabhushan, A., Matharu, N.K., Mishra, R.K., 2013. A functionally conserved Polycomb response element from mouse HoxD complex responds to heterochromatin factors. Sci Rep 3, 3011.

Winter, E., Varshavsky, A., 1989. A DNA binding protein that recognizes oligo(dA).oligo(dT) tracts. EMBO J. 8, 1867–1877.