Molecular Biology Open Access
ISSN: 2168-9547
Babu et al., Mol Biol 2016, 5:4
DOI: 10.4172/2168-9547.1000173
Research Article Open Access

Distribution and Associations of GATA Repeats in Rice Genome

Adari Prasad Babu[1], K Aruna Kumari[2] and Neelamraju Sarla[1]

[1]*Acharya N G Ranga Agriculture University, Hyderabad 500030, Andra Pradesh, India*
[2]*Rice Research, Rajendranagar, Hyderabad 500030, Andra Pradesh, India*

Abstract

GATA repeats are associated with sex differentiation in man, buffalo, mouse and even in plants such as papaya. Human X-chromosome region Xp22 that escapes inactivation is ten-fold enriched in GATA repeats suggesting a role in preventing heterochromatinization. The close proximity of GATA repeats to matrix-associated regions (MARs) indicates a role in chromatin organization and function. Chromosome-wise distribution and density of GATA repeats, neighboring genes and Matrix associated regions were analyzed in rice. $(GATA)_3$ and higher repeats were distributed non-randomly with the highest frequency on chromosome 11. About 60% of the repeats were found in intergenic regions flanked by regulatory genes involved in stress response or transposable elements. The GATA associated MAR sequences in rice had at least one or more of the consensus sequence to which GATA factors bind. The genomic milieu around GATA repeats suggests that their genomic context may determine their role in chromatin organization and gene regulation.

Keywords: Rice genome; GATA repeats; Matrix attachment regions; Genomics; Chromatin organization

Introduction

Genome sequences of eukaryotes reveal many non-coding sequences, a large proportion of which are repetitive in nature. Several sequenced genomes reveal absence of GATA repeats in prokaryotes and their accumulation in higher organisms during the course of evolution [1]. In animals, GATA repeats play an important role in differentiation of sex chromosomes. A striking ten-fold enrichment of $GATA_n$ was reported in the 10 Mb segment at Xp22 region of human X-chromosome that escapes inactivation and a similar enrichment was found in other eutherian genomes [2] indicating their possible role in regulation and formation of facultative heterochromatin. GATA/GACA repeat sequences are transcribed exclusively in Sertoli cells in addition to somatic tissues of normal rats but not infertile rats suggesting their regulatory role in male gonad [3]. Binding of a factor (BBP) to enriched stretches of GATA repeats (Bkm) in the heterogametic sex-specific chromosome of snakes, birds, mouse and man results in germ cell specific decondensation and transcriptional activation of these otherwise highly condensed chromosomes in the somatic tissues [4]. In man, nine GATA repeats in a microsatellite D17S1303 were associated with hypertension while 14 GATA repeats were associated with normal tension [5]. Thus, GATA repeats appear to have a role both in chromatin organization and function. GATA repeat containing markers are routinely used in forensic science and paternity testing because of their high polymorphism [6].

In plants, $GATA_4$ have been used in profiling rice germplasm [7,8], distinguishing various accessions within a single "Marzano" cv. of tomato and also individual plants of the same accession [9], fingerprinting wild and cultivated species of banana [10], *Brassica juncea* [11], cultivars of pearl millet [12], studying allelic diversity in sunflower [13]. GATA repeats reveal sex-specific differences even in plants. GATA containing sequences which were male specific were found useful in papaya where male and female plants do not show any specific morphological differences until flowering [14,15]. Earlier studies on genetic diversity in rice based on $(GAGA)_4$, $(AGAG)_4$ and $(GATA)_4$ primers were very informative in grouping based on genetic relationships and also traits such as drought, flood or salt tolerance [16,17]. $(GATA)_4$ containing sequences grouped bacterial leaf blight (BLB) susceptible and resistant rice varieties separately, indicating their association with BLB resistance [18]. However, the distribution and role of GATA repeats in plants has not been clearly demonstrated. MARs/SARs (Matrix attachment regions or Scaffold attachment regions) are DNA sequence elements that bind with some affinity to sites in the nuclear matrix. MAR from chicken lysozyme reduces variability in transgene expression and confers copy number dependence in transgenic rice plants [19]. Inclusion of MARs from soybean [20] in transgene cassettes reduces position effect variations and enhances the expression of transgenes in barley [21]. MARs can also act as boundary elements creating topologically isolated chromatin domains, which insulate genes located on the loop from cis-acting elements. Boundary elements were identified first in Drosophila and subsequently found to be present ubiquitously from yeast to humans. Predicted GATA-MAR regions on human Y-chromosome have been shown to function as enhancer blockers using transgenic assays in *D. melanogaster* [22]. Alternatively, the associated GATA repeats may be forming foci of transcription complex for the coordinated expression of spatially regulated genes [23].

Rice has one of the smallest genome among plants with a relatively lower frequency of repeated sequences among monocots. There are about 5251 hyper variable SSRs per Mb or 3 SSRs per gene in the rice genome [24]. Cues to the function of GATA repeats can emerge from analyzing their distribution in rice genome, the kinds of genes associated with them and their association with known boundary elements such as MARs. Additionally, analyses of the genes adjacent to $(GATA)_n$ or genes in which $(GATA)_n$ occur will also provide insight into the possible function of these repeats in rice. This study was undertaken to analyze *in silico,* the frequency and distribution of GATA repeats, the kind of associated genes and the association with putative S/MARs in the rice genome.

Materials and Methods

Repeat analysis

The rice genome sequences were downloaded from ftp sites of IRGSP, http://rgp.dna.affrc.go.jp/E/IRGSP/Build4/build4.html. A java

***Corresponding author:** K Aruna Kumari, Rice Research, Rajendranagar, Hyderabad 500030, Andra Pradesh, India, Tel: 919642222253; E-mail: arunaagbsc@gmail.com

Received July 22, 2016; **Accepted** September 07, 2016; **Published** September 14, 2016

Citation: Babu AP, Kumari KA, Sarla N (2016) Distribution and Associations of GATA Repeats in Rice Genome. Mol Biol 5: 173. doi: 10.4172/2168-9547.1000173

based program was written to analyze the distribution of the perfect tandem repeats of $(GATA)_3$ and higher repeats in the whole rice genome. Sequences 10 kb on either side of selected GATA repeats were analyzed for Matrix associated region (MAR) potential using the MAR-Wiz program. The analysis was carried out with the window width of 1000, slide distance of 100, scan length of 3 and cut off threshold score for the MAR potential set to 0.60. The core MAR rules included origin of replication rule, TG-richness, curved DNA, Topoisomearse II recognition, AT-richness and also plant MAR consensus.

Plant material and analysis of ISSRs

The cultivars of rice used for PCR analysis using $(GATA)_4$ primers were IR64, IR28, Swarna (control); Vandana, Rasi, Nagina 22 (drought tolerant); FR13A, Jalmagna, Sabita (flood tolerant); and CSR-30 (Yamini), Pokkali and Nonasail (salt tolerant). Selected $(GATA)_4$ containing ISSRs from these 12 rice cultivars [17] were cloned into TOPO-TA cloning vector (Invitrogen, CA, USA) and the sequences obtained using vector specific primers were analyzed.

Statistical analysis

Blast similarity search is a standard method for assessing the statistical significance of molecular sequences to ascertain whether an unusual pattern (perfect GATA repeats in this case) could have arisen simply by chance [25]. This is done by assigning appropriate scoring values to individual residues of sequences employing the formula: $E=K(mn)e^{-\lambda s}$, where K=0.035 and λ=0.252 are known empirical values, E is the expectation value (=p value) and m and n are length of sequences.

The Wilcoxon signed-rank test on the normalized proportion of perfect GATA repeats and GATA repeats with one mismatch and two mismatches was performed to check if there was any preference for perfect GATA repeats over GATA repeats with one and two allowed mismatches. Percent of AT content was determined as percent of A and T in a window of 250 bases on 5' and 3' of each GATA repeat. The method given in [26] was followed to calculate the correlation between occurrence of GATA repeats and number of genes on each chromosome while correcting for local AT content.

Results and Discussion

Analysis of GATA repeats in rice chromosomes

There were a total of 787 GATA repeats in the rice genome, with $(GATA)_3$ representing 50% and other higher repeats ranging from $(GATA)_{4-36}$ constituting the other 50% of the total repeats (Table 1). The frequency of the repeats decreased from 395 $(GATA)_3$ to 8 $(GATA)_{14}$ repeats. Five types of repeats (15, 17, 19, 27 and 36) appeared only once in the genome (Table 1). Repeats of $(GATA)_{3-5}$ were observed on all the chromosomes whereas $GATA_{6-7}$ repeats were absent on chromosome 5.

Score values calculated for various lengths of GATA repeats at E value of 0.01 indicated that no mismatches were tolerated upto 8 repeats of GATA and for repeats higher than $(GATA)_8$ the allowed mismatches vary depending on sequence length. The value of 8 repeats would be higher for E=0.001, which is the more stringent value used for comparison of nucleotide sequences.

If we look for the distribution of GATA repeats by allowing mismatches, we found 3112 GATA repeats ($GATA_{3-36}$) with one mismatch, with $(GATA)_{4-36}$ representing only 21% (659). When we allow 2 mismatches in the GATA repeats, we observe 15018 GATA repeats ($GATA_{3-36}$), with $(GATA)_{4-36}$ representing only 6% (909) including the perfect GATA repeat (Table 2). Considering that no mismatches are allowed till $GATA_8$, the representation of $GATA_{9-36}$ will be only 0.48% which does not appear significant. We performed the Wilcoxon signed-rank test on the normalized proportion of perfect GATA repeats and GATA repeats with one mismatch and two mismatches to check if there was any preference for perfect GATA repeats compared to GATA repeats with one or two allowed mismatches. The p-value indicated that perfect GATA repeats was significantly more preferred than GATA repeats with two allowed mismatches (p-value 0.0011). The preference of perfect GATA repeats over GATA repeats with one mismatch was less significant (p value 0.022).

Apart from GATA, another tetranucleotide repeat GACA has been shown to be enriched in the transcripts from somatic tissues of rat and buffalo, whereas the germ line transcripts in these organisms were enriched for GATA repeats [3,27]. When GACA/GATA repeats were analyzed *in silico* in six species, human, dog and *Arabidopsis thaliana* genomes were GATA-rich and chicken showed similar occurrence of GATA/GACA [27]. Therefore, in addition to GATA repeats, we also analyzed the distribution of GACA repeats and found their numbers insignificant as compared to $(GATA)_n$ in the rice genome (data not shown).

Chromosome-wise analysis of the GATA repeats from $(GATA)_{3-36}$ revealed a maximum of 104 repeats on chromosome 11 and a minimum of 42 repeats on chromosome 10 (Table 1). The frequency of repeats was highest in chromosome 11 with 34 repeats per 10Mb followed by chromosome 12 and 9, with 23 and 18 repeats per 10Mb, respectively. In others it varied from 9 to 17 $(GATA)_n$ per 10 Mb. Chromosome 11 had the highest number of $(GATA)_n$ per 10Mb or per 1000 genes and chromosome 3 had the lowest. The GATA repeats were distributed along the entire length of chromosomes. Higher repeats were more and appeared to be clustered on Chromosomes 9, 11 and 12. Chromosomes 1 and 4 had more repeats around the centromere, whereas chromosome 5 showed more repeats at both the telomeric ends. Chromosomes 1 and 3 had the least number of higher repeats. The Pearson correlation coefficient between the perfect GATA repeats per chromosome and chromosome length/no. of genes per chromosome/ was not statistically significant indicating the genome-wide distribution of GATA repeats was non-random. Even when it was corrected for local AT content, i.e., 250 base pairs on either side of GATA repeats, the correlation (0.125) was not significant. The overall AT content per chromosome was 56.5% but in the 500 bp window around GATA repeats it was 63.7% (Table 3). The minimum AT content was 48.12 and the maximum was 74.26 around GATA repeats.

In situ hybridization studies showed a preferential localization of GATA repeats in the heterochromatic and/or centromeric chromosomal areas in sugar beet [28], chickpea [29] and tomato [30]. Clustering of GATA was highest in chromosomes 9, 11 and 12, which are also the prominent chromosomes mapped with genes for tolerance to abiotic and biotic stresses in rice. QTLs for submergence tolerance and other biotic stresses have been reported on chromosome 9 [31,32] in the regions where the $(GATA)_n$ are found clustered. Chromosome 12 is also reported to have many significant QTLs/genes for tolerance to abiotic [33] and biotic stresses. Thus, there appears to be a correlation between the chromosomal distribution of $(GATA)_n$ clusters and that of genes/QTLs for tolerance to abiotic and biotic stresses [34]. It is also interesting to note a broad similarity of distribution of $(GATA)_n$ clusters to the heterochromatin distribution in the 12 chromosomes [35,36] and the distribution of disease resistance gene clusters [34,37] with chromosome 11 showing the maximum number of GATA repeats, largest heterochromatin region among all rice chromosomes

GATA repeat frequency distribution													
	Chromosome No												
No of GATA repeats	1	2	3	4	5	6	7	8	9	10	11	12	Total
3	23	41	27	32	35	32	39	35	25	22	46	38	395
4	17	9	8	16	9	8	12	8	16	10	14	15	142
5	8	7	2	11	4	7	7	6	9	4	14	10	89
6	4	5	1	1	0	3	4	4	2	2	6	5	37
7	4	2	2	2	0	2	0	3	2	2	8	0	27
8	2	3	0	1	2	0	2	1	2	0	4	3	20
9	1	2	0	0	2	1	1	2	1	1	3	2	16
10	0	1	1	1	1	0	2	0	0	1	4	1	12
11	1	0	1	4	0	0	1	1	2	0	0	2	12
12	1	3	1	1	1	0	0	0	0	0	1	2	10
13	1	1	0	2	1	0	0	1	1	0	2	2	11
14	0	4	0	0	0	1	0	1	0	0	2	0	8
15	1	0	0	0	0	0	0	0	0	0	0	0	1
17	0	0	0	0	0	0	0	0	0	0	0	1	1
19	0	1	0	0	0	0	0	0	0	0	0	0	1
21	0	1	1	1	0	0	0	0	0	0	0	0	3
27	0	0	0	0	0	0	1	0	0	0	0	0	1
36	1	0	0	0	0	0	0	0	0	0	0	0	1
Total	64	80	44	72	55	54	69	62	60	42	104	81	787

Table 1: Frequency distribution of (GATA)3 and higher perfect tandem repeats of GATA in the 12 chromosomes of rice.

	GATA distribution with 1 mismatch												
		chr01	**chr02**	**chr03**	**chr04**	**chr05**	**chr06**	**chr07**	**chr08**	**chr09**	**chr10**	**chr11**	**chr12**	**Total**
1	(GATA)1:	2018511	1694211	1693947	1640112	1389079	1466234	1397908	1343062	1082722	1073974	1365727	1318551	17484038
2	(GATA)2:	17574	15266	14797	14184	11893	12919	12173	12229	9309	9482	12420	11800	154046
3	(GATA)3:	270	262	230	192	200	200	192	174	152	150	224	207	2453
4	(GATA)4:	23	29	18	28	27	23	26	31	17	19	30	28	299
5	(GATA)5:	14	7	10	15	6	7	14	6	15	6	16	13	129
6	(GATA)6:	7	7	2	10	3	6	9	7	6	3	11	9	80
7	(GATA)7:	5	2	2	1		4	1	1	1	4	9	3	33
8	(GATA)8:	3	4		2	1	1	2	2		1	7		23
9	(GATA)9:	3	2	1		3	1		3	2		4	3	22
10	(GATA)10:		2	1	1		1	3	1	1	1	2	3	16
11	(GATA)11:	1	2		3	1			1		1	3	1	13
12	(GATA)12:		1	1	3	1		1		2		1	2	12
13	(GATA)13:	2	1	1						1			1	6
14	(GATA)14:		2		2	1	1		1			2	2	11
15	(GATA)15:		3						1	1		2		7
16	(GATA)16:	1												1
18	(GATA)18:												1	1
20	(GATA)20:		1											1
21	(GATA)21:			1	1									2
22	(GATA)22:		1											1
28	(GATA)28:							1						1
36	(GATA)36:	1												1
	sum 3-36	330	326	267	258	243	244	249	228	198	185	311	273	3112

Table 2: GATA repeat distribution considering one mismatch.

in Giemsa stained prometaphase mitotic chromosomes [35] and also the maximum number of genes involved in biotic and abiotic stress resistance [2] reported a striking ten-fold enrichment of $(GATA)_n$ in the 10 Mb segment at Xp22 region of human X-chromosome that escapes inactivation. In this classic paper they clearly demonstrated that presence of $(GATA)_n$ prevents heterochromatinization. Similarity of $(GATA)_n$ clusters distribution to the heterochromatin distribution in the 12 chromosomes at mitotic stage [35] and meiotic stage [36] suggests that $(GATA)_n$ in association with MARs may be protecting or shielding these genes from the negative effect that heterochromatinization may have on transcription of neighbouring genes [38].

Distribution of GATA repeats in various genomic regions of rice

In the rice genome, 673 GATA repeats were intergenic and 114 were intragenic. The details of the genes flanking intergenic GATA repeats. The details of genes which had GATA repeats within them. Most of the GATA repeats in the X and Y human chromosomes also were intergenic [1]. Intergenic repeats were more at a distance greater than 5 kb downstream or upstream of genes compared to those within 5 kb of a gene. The frequency of repeats within 1 to 3 kb region of a gene was greater when occurring upstream of the gene, but with increasing

distance; the repeats were more downstream of the gene indicating they may have a role in promoter regions of genes in plants.

It was found that the same types of genes (most of them coding for conserved hypothetical protein) flanked GATA repeats (present on both 5' and 3' of repeats) in 70 cases. Of these as many as 20 genes were in chromosome 11 and chromosome 10 had none. The details of such flanking genes are given in Table 4 also reported that the DNA flanking the GATA probe in tomato were highly homologous to each other [30].

Chromosome 11 was enriched in $GATA_n$ and also had the highest instances of the same gene present closest on either side of the repeats. In rice many disease and pest resistance genes map to chromosome 11, which has the highest frequency of resistance gene analogues [34]. On the other hand, the absence of instances of same gene present on both sides of $(GATA)_n$ in chromosome 10 was striking. Wild species derived yield QTLs were also reported to be absent on chromosome 10 [39]. Also microRNA clusters were reported in all rice chromosomes except chromosome 10 [40]. It remains to be seen whether there exists a link between these observations or they are only anecdotal.

The presence of GATA repeats in rice transcriptome has not been reported but in papaya, a GATA repeat containing 0.8 kb sequence was transcribed only in the male plant indicating sex specific expression [15]. The specificity was maintained even in sex reversed (female to male) plants. The differential expression of GACA/GATA tagged transcripts has also been shown in buffalo, where all GATA-tagged transcripts were unique to testes or spermatozoa [27]. It is significant that all GATA-tagged transcripts showed highest expression and were conserved across species.

Association of GATA repeats with MARs

GATA-MAR sequences from human Y-chromosome were shown to function as boundary elements in enhancer blocking assays in *D. melanogaster* and human cells [41]. It is possible that intergenic GATA-MAR sequences could be acting as insulators thereby regulating genes which have to be expressed at different stages of development or in different tissues. Therefore, 97 intergenic $(GATA)_n$ flanked by regulatory genes known to be involved in temporal or spatial expression were analyzed for MAR association (Supplementary File 3). All such GATA repeats were found to be associated with MARs with very high scores of 0.7 and above. Some GATA repeats were within the MAR sequences. The length of the MAR sequences ranged from 500 bp to 1 kb, a few was less than 200 bp and a few over 1 kb. Some MAR sequences also included the plant MAR consensus sequence. All the GATA associated MARs were found to have one or more of the GATA factor binding consensus sequences A/T GATA A/G.

The genomes of Arabidopsis and rice show 29 and 28 loci respectively, that encode for putative GATA factors [42]. These proteins bind strongly to GATA motifs and regulate transcription of the neighboring genes. A 40 kb DNA (5a) containing 7 tandem repeats of GATA located at the 5'boundary area of Locus Control Region (LCR) of human β-globin gene exhibited a silencer activity in erythroid cells upon binding to GATA-1 protein [38]. GAGA-binding protein was shown to bind specifically to GAGA elements in the promoter of gene encoding chlorophyll and heme synthesis enzyme [43]. Duplication of 305 bp element containing GAGA repeat in the promoter of barley gene activates gene expression in tobacco by binding to BBC, a GAGA-binding factor [44]. Such examples for GATA repeats in promoters having activating or silencing effects are not reported in plants. It appears that GATA-MARs association ensures a certain degree of transcription of genes important for survival and adaptability and such genes are probably shielded from repressing influences.

Experimental validation of the presence of $(GATA)_4$ in rice cultivars

Analysis of fingerprinting patterns in 12 rice cultivars using $GATA_4$ primers showed the presence of a 1111 bp unique band in cultivar Vandana, 630-664 bp band in Swarna, Rasi, FR13A, Jalmagna, Pokkali and Nonasail, and 285-320 bp band common to IR28, Swarna, Rasi, FR13A and Jalmagna. The sequences of all three polymorphic bands were intergenic and as expected showed 4 repeats of GATA at the 5' and 3' ends (Supplementary File 4). BLAST analysis of the sequences identified the gene found upstream of 1110 kb sequence as NADH ubiquinone oxidoreductase chain4 on chromosome 12 and similar to cell cycle control protein cwf14 on chromosome 4. The genes downstream of the sequence were cyclin like F box domain containing protein and ubiquinone specific protease 22, respectively (Sup File 5). Genes coding for impedance induced protein and HD-Zip protein were upstream of the 664 bp and the 316 kb sequences, respectively while the downstream genes were plant lipid transfer/seed storage/trypsin-alpha amylase domain containing protein and inositol 1,3,4-triphosphate 5/6kinase family protein, respectively. All the three ISSR sequences were associated with MARs (the 316kb sequence was included within the MAR sequence) and appear to be associated with genes involved in one or more stress-responses [45-48]. The role of the various genes in stress responses. The PCR products specifically amplified in a given group of cultivars using $(GATA)_4$ primers could therefore, reflect loci

Chromosome	GATA repeats	AT content % around GATA repeats 500bp window	AT content % for each chromosome
chr01	64	63.19	56.23
chr02	80	62.14	56.67
chr03	44	63.48	56.31
chr04	72	64.65	55.75
chr05	55	63.44	56.05
chr06	54	63.55	56.39
chr07	69	63.84	56.50
chr08	62	64.00	56.62
chr09	60	63.00	56.50
chr10	42	62.94	56.44
chr11	104	65.29	57.10
chr12	81	65.05	56.99
mean	65.58	63.71	56.46
sum	787		
correlation	0.27673781		

Table 3: AT content around GATA repeats in each chromosome.

S. No.	Chr	GATA repeat	GATA start	GATA end	Gene start	Gene end	gene -GATA distance		strand	5' gene - GATA 3' Gene family	Gene start	Gene end	GATA -gene distance	strand	5' GATA - gene 3' Gene family
1	1	4	24701786	24701801	24687098	24687846	13940	.	+	Conserved hypothetical protein	24730425	24736622	28624	+	Conserved hypothetical protein
2	1	7	33652066	33652093	33645557	33647067	4999	.	-	Conserved hypothetical protein	33665405	33667496	13312	+	Conserved hypothetical protein
3	2	8	6195108	6195139	6185578	6187201	7907	.	-	Conserved hypothetical protein	6212115	6217584	16976	-	Conserved hypothetical protein
4	2	6	6198384	6198407	6185578	6187201	11183	.	-	Conserved hypothetical protein	6212115	6217584	13708	-	Conserved hypothetical protein
5	2	12	11274860	11274907	11238034	11246765	28095	.	-	Conserved hypothetical protein	11321705	11329300	46798	-	Conserved hypothetical protein
6	2	13	23380532	23380583	23372705	23373388	7144	.	-	Conserved hypothetical protein	23386917	23389250	6334	-	Conserved hypothetical protein
7	2	8	24364590	24364621	24353515	24354170	10420	.	-	Conserved hypothetical protein	24370958	24371568	6337	+	Conserved hypothetical protein
8	4	5	6694998	6695017	6675795	6685701	9297	.	+	Conserved hypothetical protein	6862048	6865249	167031	+	Conserved hypothetical protein
9	5	3	4030322	4030333	3953409	3957864	72458	.	-	Conserved hypothetical protein	4054054	4056450	23721	+	Conserved hypothetical protein
10	5	3	17643571	17643582	17637814	17640132	3439	.	-	Conserved hypothetical protein	17644625	17648132	1043	-	Conserved hypothetical protein
11	5	5	26109190	26109209	26089436	26090412	18778	.	-	Conserved hypothetical protein	26111120	26117661	1911	+	Conserved hypothetical protein
12	5	3	26109220	26109231	26089436	26090412	18808	.	-	Conserved hypothetical protein	26111120	26117661	1889	+	Conserved hypothetical protein
13	5	12	26109235	26109282	26089436	26090412	18823	.	-	Conserved hypothetical protein	26111120	26117661	1838	+	Conserved hypothetical protein
14	5	5	26109286	26109305	26089436	26090412	18874	.	-	Conserved hypothetical protein	26111120	26117661	1815	+	Conserved hypothetical protein
15	5	8	27095951	27095982	27056983	27060414	35537	.	-	Conserved hypothetical protein	27098165	27100217	2183	-	Conserved hypothetical protein
16	6	5	20606674	20606693	20578313	20579327	27347	.	+	Conserved hypothetical protein	20607668	20612259	975	+	Conserved hypothetical protein
17	6	4	20606697	20606712	20578313	20579327	27370	.	+	Conserved hypothetical protein	20607668	20612259	956	+	Conserved hypothetical protein
18	6	3	24828024	24828035	24826822	24827697	327	.	+	Conserved hypothetical protein	24882102	24887029	54067	-	Conserved hypothetical protein
19	6	5	26448319	26448338	26421483	26423481	24838	.	+	Conserved hypothetical protein	26460154	26462091	11816	+	Conserved hypothetical protein
20	6	3	26448343	26448354	26421483	26423481	24862	.	+	Conserved hypothetical protein	26460154	26462091	11800	+	Conserved hypothetical protein
21	7	3	2421358	2421369	2401320	2402899	18459	.	+	Conserved hypothetical protein	2421820	2423024	451	+	Conserved hypothetical protein
22	7	3	9328743	9328754	9310983	9315142	13601	.	-	Conserved hypothetical protein	9376788	9381618	48034	+	Conserved hypothetical protein
23	7	4	9374660	9374675	9310983	9315142	59518	.	-	Conserved hypothetical protein	9376788	9381618	2113	+	Conserved hypothetical protein
24	7	4	12043464	12043479	12003709	12005437	38027	.	+	Conserved hypothetical protein	12045905	12048421	2426	-	Conserved hypothetical protein
25	7	5	16613269	16613288	16598453	16601216	12053	.	-	Conserved hypothetical protein	16651040	16654265	37752	+	Conserved hypothetical protein
26	8	3	2682883	2682894	2575590	2580750	102133	.	-	Conserved hypothetical protein	2751832	2759584	68938	+	Conserved hypothetical protein
27	8	7	2682907	2682934	2575590	2580750	102157	.	-	Conserved hypothetical protein	2751832	2759584	68898	+	Conserved hypothetical protein
28	8	3	7935443	7935454	7910803	7915433	20010	.	+	Conserved hypothetical protein	7938709	7943310	3255	+	Conserved hypothetical protein
29	8	5	9983163	9983182	9975980	9978979	4184	.	-	Conserved hypothetical protein	9983349	9984269	167	+	Conserved hypothetical protein
30	8	4	25428761	25428776	25421015	25422438	6323	.	-	Conserved hypothetical protein	25429105	25431954	329	+	Conserved hypothetical protein
31	9	5	6163699	6163718	6122922	6123584	40115	.	-	Conserved hypothetical protein	6164444	6165552	726	+	Conserved hypothetical protein
32	11	5	6840485	6840504	6835300	6836819	3666	.	+	Conserved hypothetical protein	6844680	6848435	4176	-	Conserved hypothetical protein
33	11	3	9500298	9500309	9473972	9477991	22307	.	-	Conserved hypothetical protein	9508399	9510985	8090	-	Conserved hypothetical protein
34	11	3	24637726	24637737	24541941	24543463	94263	.	+	Conserved hypothetical protein	24641828	24651369	4091	-	Conserved hypothetical protein
35	11	3	28646652	28646663	28633396	28636229	10423	.	-	Conserved hypothetical protein	28674086	28675560	27423	-	Conserved hypothetical protein
36	11	7	28984443	28984470	28959489	28960386	24057	.	+	Conserved hypothetical protein	28995282	28995889	10812	+	Conserved hypothetical protein
37	11	5	28984474	28984493	28959489	28960386	24088	.	+	Conserved hypothetical protein	28995282	28995889	10789	+	Conserved hypothetical protein
38	11	7	28984497	28984524	28959489	28960386	24111	.	+	Conserved hypothetical protein	28995282	28995889	10758	+	Conserved hypothetical protein
39	11	4	28984539	28984554	28959489	28960386	24153	.	+	Conserved hypothetical protein	28995282	28995889	10728	+	Conserved hypothetical protein
40	11	7	28984572	28984599	28959489	28960386	24186	.	+	Conserved hypothetical protein	28995282	28995889	10683	+	Conserved hypothetical protein
41	11	3	29975190	29975201	29875114	29877003	98187	.	+	Conserved hypothetical protein	30056649	30060682	81448	+	Conserved hypothetical protein
42	11	6	29981910	29981933	29875114	29877003	104907	.	+	Conserved hypothetical protein	30056649	30060682	74716	+	Conserved hypothetical protein
43	11	3	29994118	29994129	29875114	29877003	117115	.	+	Conserved hypothetical protein	30056649	30060682	62520	+	Conserved hypothetical protein
44	11	3	30025776	30025787	29875114	29877003	148773	.	+	Conserved hypothetical protein	30056649	30060682	30862	+	Conserved hypothetical protein
45	11	13	30172348	30172399	30108631	30109372	62976	.	-	Conserved hypothetical protein	30176581	30180156	4182	-	Conserved hypothetical protein
46	12	3	6490991	6491002	6488583	6489594	1397	.	+	Conserved hypothetical protein	6562463	6566699	71461	-	Conserved hypothetical protein
47	12	9	7357642	7357677	7339923	7343904	13738	.	-	Conserved hypothetical protein	7388706	7395059	31029	+	Conserved hypothetical protein
48	1	4	16066316	16066331	16008960	16012655	53661	.	+	Cyclin-like F-box domain containing protein	16082626	16086208	16295	+	Cyclin-like F-box domain containing protein
49	4	4	6424281	6424296	6377017	6381325	42956	.	-	Cyclin-like F-box domain containing protein	6446988	6448425	22692	-	Cyclin-like F-box domain containing protein
50	1	5	35861380	35861399	35845307	35847004	14376	.	+	Cytochrome P450 family protein	35870848	35872605	9449	+	Cytochrome P450 family protein
51	1	5	35861403	35861422	35845307	35847004	14399	.	+	Cytochrome P450 family protein	35870848	35872605	9426	+	Cytochrome P450 family protein
52	7	3	3389045	3389056	3383460	3384824	4221	.	-	Esterase/lipase/thioesterase domain containing protein	3392153	3393711	3097	-	Esterase/lipase/thioesterase domain containing protein
53	1	3	15090454	15090465	15067582	15070637	19817	.	+	Hypothetical protein	15098359	15102028	7894	+	Hypothetical protein
54	8	3	1253157	1253168	1238099	1249360	3797	.	+	Hypothetical protein	1281397	1285548	28229	-	Hypothetical protein

55	11	8	7841748	7841779	7832542	7834380	7368	.	-	Hypothetical protein	7862028	7866571	20249	+	Hypothetical protein
56	11	3	7841783	7841794	7832542	7834380	7403	.	-	Hypothetical protein	7862028	7866571	20234	+	Hypothetical protein
57	11	9	7841798	7841833	7832542	7834380	7418	.	-	Hypothetical protein	7862028	7866571	20195	+	Hypothetical protein
58	11	3	11393282	11393293	11386764	11393097	185	.	+	Hypothetical protein	11395179	11398769	1886	-	Hypothetical protein
59	11	3	16826438	16826449	16821123	16821846	4592	.	+	Hypothetical protein	16898905	16899628	72456	+	Hypothetical protein
60	7	11	21559924	21559967	21544775	21546312	13612	.	+	Peptidase A1, pepsin family protein	21587010	21588539	27043	-	Peptidase A1, pepsin family protein
61	3	3	16374955	16374966	16314465	16318269	56686	.	-	Peptidase S10, serine carboxypeptidase family protein	16375058	16380059	92	-	Peptidase S10, serine carboxypeptidase family protein
62	11	3	5945273	5945284	5930290	5931158	14115	.	+	Plant disease resistance response protein family protein	5979435	5980319	34151	+	Plant disease resistance response protein family protein
63	1	3	30693667	30693678	30689681	30691036	2631	.	+	Protein kinase-like domain containing protein	30695502	30696854	1824	-	Protein kinase-like domain containing protein
64	6	6	7352392	7352415	7340810	7343914	8478	.	-	SAM dependent carboxyl methyltransferase family protein	7368138	7370783	15723	-	SAM dependent carboxyl methyltransferase family protein
65	6	3	7352644	7352655	7340810	7343914	8730	.	-	SAM dependent carboxyl methyltransferase family protein	7368138	7370783	15483	-	SAM dependent carboxyl methyltransferase family protein
66	6	3	7352676	7352687	7340810	7343914	8762	.	-	SAM dependent carboxyl methyltransferase family protein	7368138	7370783	15451	-	SAM dependent carboxyl methyltransferase family protein
67	1	3	12906680	12906691	12884715	12888201	18479	.	+	Similar to Lipase homolog (Fragment)	12947238	12952863	40547	+	Similar to Lipase homolog (Fragment)
68	2	8	22740753	22740784	22711442	22721453	19300	.	-	Terpenoid cylases/protein prenyltransferase alpha-alpha toroid domain containing protein	22744171	22757091	3387	-	Terpenoid cylases/protein prenyltransferase alpha-alpha toroid domain containing protein
69	2	9	22740788	22740823	22711442	22721453	19335	.	-	Terpenoid cylases/protein prenyltransferase alpha-alpha toroid domain containing protein	22744171	22757091	3348	-	Terpenoid cylases/protein prenyltransferase alpha-alpha toroid domain containing protein
70	1	4	32420354	32420369	32411152	32412950	7404	.	+	UDP-glucuronosyl/UDP-glucosyltransferase family protein	32422485	32424160	2116	+	UDP-glucuronosyl/UDP-glucosyltransferase family protein

Table 4: Details of 70 instances of same gene/ gene family flanking a particular GATA repeat.

which are transcriptionally active and associated with stress adaptive functions [49].

Conclusion

Thus, the genomic milieu around GATA repeats presented in our paper suggests that their genomic context may determine their role in chromatin organization and gene regulation. Based on the distribution of $(GATA)_{10\text{-}12}$ along the chromosome and their close proximity to Matrix Associated Regions (GATA-MAR) in man it was suggested that it may be delineating chromatin domains for their coordinated expression [1]. There is growing evidence that GATA repeat elements have chromatin domain boundary activity in *Drosophila melanogaster* as well as in human cells and play a role in packaging of genome and in regulatory mechanisms involving large regions of chromosomes [41].

The role of GATA-MARs associations may be more pronounced in plants as the need to ensure coordinated expression of genes when faced with an environmental stress is more as plants are sessile. The ISSR sequences of three of the fragments obtained following PCR analysis of stress-tolerant cultivars using $(GATA)_4$ primers were associated with genes involved in one or more stress-responses. The positional information of perfect GATA repeats provided in this paper and the adjacent genes and MARs can serve as a framework for further analysis of their biological meaning. Eppelen noted that "the question of the functional meaning, if any, of simple, tandemly repeated sequences such as GATA/GACA DNA remains unanswered" [50]. Evidence from several organisms points to a definite role of these repeats in regulation of gene functions and that is not restricted to sex specific differences in man, rat, buffalo or papaya. An overlay of GATA repeat distribution with the distribution of heterochromatin, nucleosome positioning, whole genome methylation, acetylation, AT content and several such features involved in chromatin structure and function may give deeper insights into their function.

Acknowledgement

APB and CSR were supported by the Department of Biotechnology, Government of India project BT/AB/03/FG-2/2003 as part of Network Project on Rice Functional Genomics and BPMS by CSIR-UGC fellowship. We thank Rakesh Mishra for useful discussions. The help of Senthilkumar Ramamurthy in the initial analysis of repeats in rice genome is gratefully acknowledged.

References

1. Apisitwanich S, Shishido R, Akiyama Y, Fukui K (2001) Quantitative chromosome map of representative indica rice. Euphytica 118:113-118.
2. Bernier J, Kumar A, Ramaiah V, Spaner D, Atlin G (2007) A large-effect QTL for grain yield under reproductive-stage drought stress in upland rice. Crop Sci 47: 507-518.
3. Bhatia S, Das S, Jain A, Lakshmikumaran M (1995) DNA fingerprinting of *Brassica juncea* cultivars using microsatellite probes. Electrophoresis16: 1750-1754.
4. Blair MW, Panaud O, McCouch SR (1999) Inter-simple sequence repeat (ISSR) amplification for analysis of microsatellite motif frequency and fingerprinting in rice (*Oryza sativa* L). Theor Appl Genet 98:780-792.
5. Breyne P, Van Montagu M, Depicker A, Gheysen G (1992) Characterization of a plant scaffold attachment region in a DNA fragment that normalizes transgene expression in tobacco. Plant Cell 4:463-471.
6. Buising CM, Benbow RM (1994) Molecular analysis of transgenic plants generated by microprojectile bombardment, effect of petunia transformation booster sequence. Mol Gen Genet 243: 71-81.
7. Cai HN, Shen P (2001) Effects of cis arrangement of chromatin insulators on enhancer-blocking activity. Science 291: 493-495.
8. Cheng Z, Buell RC, Wing RA (2001) Toward a cytological characterization of the rice genome. Genome Res 11: 2133-2141.

9. Cheung BMY, Leung RYH, Man YB, Wong LYF, Lau CP (2005). Association of essential hypertension with a microsatellite marker on chromosome 17. Journal of Human Hypertension 19 407-411.

10. Chowdari KV, Venkatachalam SR, Davierwala AP, Gupta VS, Ranjekar PK, et al.(1998) Hybrid performance and genetic distance as revealed by the (GATA) 4 microsatellite and RAPD markers in pearl millet. Theor Appl Genet 97: 163-169.

11. Cui X, Xu SM, Mu DS, Yang ZM (2009) Genomic analysis of rice micro RNA promoters and clusters. Gene 431: 61-66.

12. Davierwala AP, Ramakrishna W, Ranjekar PK, Gupta VS (2000) Sequence variations at a complex microsatellite locus in rice and its conservation in cereals. Theor Appl Genet 101:1291-1298.

13. Davierwala AP, Ramakrishna W, Chowdari V, Ranjekar PK, Gupta VS (2001) Potential of (GATA) n microsatellites from rice for inter- and intra-specific variability studies. BMC Evol Biol 1: 7.

14. Diniz FM, Iyengar A, da Costa Lima PS, Maclean N, et al. (2007) Application of a double enrichment procedure for microsatellite isolation and the use of tail primers for high throughput genotyping. Genetics and Mol Biol 30: 380-384.

15. Epplen JT (1988) On simple repeated GATA sequences in animal genomes, a critical reappraisal. Journal of Heredity 79: 409-417.

16. Gangadharan S, Kapur V, Ali S (2001) GATA/GACA repeat sequences are transcribed in the normal fertile rat *Rattus norvegicus* but not in the infertile ones. Curr Sci 81: 1320-1324.

17. Gangopadhyay G, Roy SK, Ghose K, Poddar R, Bandyopadhyay T, et al. (2007) Sex determination in *Carica papaya* and *Cycus circinalis* in pre-flowering stage by ISSR and RAPD. Curr Sci 92: 524-526.

18. Gortner G, Nenno M, Weising K, Zink D, Nagl W, et al. (1998) Chromosomal localization and distribution of simple sequence repeats and the Arabidopsis-type telomere sequence in the genome of **Cicer arietinum** L. Chrom Res 6: 97-104.

19. Henriksson E, Olsson ASB, Johannesson H, Johansson H, Hanson J, et al. (2005) Homeodomain leucine zipper class I genes in Arabidopsis expression patterns and phylogenetic relationships. Plant Physiol 139: 509-518.

20. International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436: 793-800.

21. Kaemmer D, Afza R, Weising K, Kahl G, Novak FJ (1992) Oligonucleotide and amplification fingerprinting of wild species and cultivars of banana (*Musa* spp). Bio/Technology 10: 1030-1035.

22. Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. PNAS 87: 2264-2268.

23. Kim SH, Yi S (2007) Understanding relationship between sequence and functional evolution in yeast proteins. Genetica 131:151-156.

24. Kottapalli KR, Sarla N, Kikuchi S (2006) *In silico* insight into two rice chromosomal regions associated with submergence tolerance and resistance to bacterial leaf blight and gall midge. Biotech Advances 24: 561-589.

25. Kreps JA, Wu Y, Chang HS, Zhu T, Wang X, et al. (2002) Transcriptome changes for Arabidopsis in response to salt, osmotic and cold stress. Plant Physiol 130: 2129-2141.

26. Kumar RP (2007) Bkm (GATA-repeats) associated chromatin domain boundaries of human Y-chromosome. PhD Thesis Submitted to Jawaharlal Nehru University, New Delhi, India.

27. Lawson MJ, Zhang L (2006) Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. Genome Biology 7:R14.

28. Leroy XJ, Leon K, Branchard M (2000) Plant genomic instability detected by microsatellite-primers. Eur J Bot 3: 140-148.

29. McNeil JA, Smith KP, Hall LL, Lawrence JB (2006) Word frequency analysis reveals enrichment of dinucleotide region repeats on the human X chromosome and [GATA] n in the X escape. Genome Res 16: 477-484.

30. Michalak P (2008) Co-expression, co-regulation and co-functionality of neighboring genes in eukaryotic genomes. Genomics 91: 243-248.

31. Mishra R (2005) CCMB Annual Report 99.

32. Mishra R (2007) Genome chromatin and regulation of genetic information CCMB Annual Report 83.

33. Niu X, Chen Q, Wang X (2008) OsITL1 gene encoding an inositol 1,3,4-trisphosphate 5/6-kinase is a negative regulator of osmotic stress signaling. Biotechnol Lett 30: 1687-1692.

34. Norrgard K (2008) Forensics DNA fingerprinting and CODIS. Nature Education 1.

35. Oh SJ, Jeong JS, Kim EH, Yi NR, Yi SI, et al. (2005) Matrix attachment region from the chicken lysozyme locus reduces variability in transgene expression and confers copy number-dependence in transgenic rice plants. Plant Cell Rep 24: 145-154.

36. Parasnis AS, Ramakrishna W, Chowdari KV, Gupta VS, Ranjekar PK (1999) Microsatellite (GATA)n reveals sex-specific differences in papaya. Theor Appl Genet 99: 1047-1052.

37. Petersen K, Leah R, Knudsen S, Cameron-Mills V (2002) Matrix attachment regions (MARs) enhance transformation frequencies and reduce variance of transgene expression in barley. Plant Mol Biol 49: 45-58.

38. Rajendrakumar P, Biswal AK, Sakthivel K (2009) Development and validation of class I SSR markers targeting (GATA)n repeat motifs in rice. Euphytica 169: 263-271.

39. Ramchandran RR, Bengra C, Whitney B, Lanclos K, Tuan D (2000) A (GATA)7 Motif located in the 5' boundary area of the human β-globin locus control region exhibits silencer activity in erythroid cells. American Journal of Hematology 65:14-21.

40. Rao R, Corrado G, Bianchi M, Di Mauro A (2006) (GATA)4 DNA fingerprinting identifies morphologically characterized San Marzano tomato plants. Plant Breeding 125:173-176.

41. Reddy CS, Babu AP, Swamy BPM, Kaladhar K, Sarla N (2009) ISSR markers based on GA and AG repeats reveal genetic relationship among rice varieties tolerant to drought, flood or salinity. J Zhejiang Univ Sci B 10: 133-141.

42. Reyes JC, Muro-Pastor MI, Florencio FJ (2004) The GATA family of transcription factors in Arabidopsis and rice. Plant Physiology 134: 1718-1732.

43. Sangwan IA, O'Brian MR (2002) Identification of a soybean protein that interacts with GAGA element dinucleotide repeat DNA. Plant Physiology 129: 1788-1794.

44. Santi L, Wang Y, Stile MR, Berendzen K, Wanke D, et al. (2003) The GA octonucleotide repeat binding factor BBR participates in the transcriptional regulation of the homeobox gene Bkn3. The Plant J 34: 813-826.

45. Sarla N, Neeraja CN, Siddiq EA (2005) Use of anchored (AG)n and (GA)n primers to assess genetic diversity in rice landraces and varieties. Curr Sci 89: 1371-1381.

46. Schmidt T, Heslop-Harrison JS (1996) The physical and genomic organization of microsatellites in sugar beet. Proc Natl Acad Sci USA 93: 876-8765.

47. Silverstein KAT, Moskal Jr WA, Wu HC, Underwood BA, Graham MA, Town CD, et al. (2007) Small cysteine-rich peptides resembling antimicrobial peptides has been under-predicted in plant. The Plant Journal 51: 262-280.

48. Singh L, Wadhwan R, Naidull S, Nagaraj R, Ganesan M (1994) Sex- and tissue-specific Bkm(GATA)-binding protein in the germ cells of heterogametic sex. J Biol Chem 269: 25321-25327.

49. Srivastava J, Premi S, Kumar S, Ali S (2008) Organization and differential expression of the GACA/GATA tagged somatic and spermatozoal transcriptomes in buffalo *Bubalus bubalis*. BMC Genomics 9:132.

50. Subramanian S, Mishra RK, Singh L (2003) Genome-wide analysis of Bkm sequences (GATA repeats), predominant association with sex chromosomes and potential role in higher order chromatin organization and function. Bioinformatics 19: 681-685.