BIOINFORMATICS DISCOVERY NOTE

Vol. 19 no. 6 2003, pages 681–685
DOI: 10.1093/bioinformatics/btg067

Genome-wide analysis of Bkm sequences (GATA repeats): predominant association with sex chromosomes and potential role in higher order chromatin organization and function

Subbaya Subramanian, Rakesh K. Mishra and Lalji Singh*

Centre for Cellular and Molecular Biology, Uppal Road, Hyderabad 500 007, India

Received on October 25, 2002; revised on November 27, 2002; accepted on December 11, 2002

ABSTRACT

Motivation: Bkm (Banded krait minor) satellite DNA sequences (GATA repeats) have been shown to be associated with the sex determining chromosomes of various eukaryotes and have been implicated in the evolution and differentiation of sex chromosomes in snakes. The objective of the study is to analyze the GATA repeats of human genome specifically, the Y-chromosome, and other model organisms to understand the possible function and potential role in higher order chromatin organization.

Results: Our extensive analysis of GATA repeats in the prokaryotic and eukaryotic genomes, which have been completely sequenced so far, has revealed that GATA repeats are absent in prokaryotes and have been gradually accumulated in higher organisms during the course of evolution. In human, the Y-chromosome has the highest GATA repeat density, which predominantly exists in the Yq centromeric region. Generally, occurrence of repeats in the genomes decreases steadily as the length of the repeat increases. In contrast, we report, that the occurrence of GATA repeats increases as the length of the repeat increases from six tandem repeats onwards and peaks at $(GATA)_{10–12}$. This has not been observed with any other simple repeat. Distribution of $(GATA)_{10–12}$ along the chromosome and their close proximity to Matrix Associated Regions (GATA-MAR) suggests that it may be demarking chromatin domains for a coordinated expression of genes residing in these domains. Supplementary data is available at http://www.ccmb.res.in/bkm/gata.htm.

Contact: lalji@ccmb.res.in

INTRODUCTION

After the sequencing of the human genome, one of the challenges is to find functions of the non-coding part, which constitutes the major portion (about 98%) of the genome (International Human Genome Sequencing Consortium, 2001). A large proportion of the non-coding genome, which is repetitive in nature, is often referred to as ‘junk’ DNA as their function is not obvious. However, now it has been generally accepted that at least some repeats are essential components of genomic organization and function. Among different types of repeats, transposable elements are the most abundant and make up to ∼45% of the total human genome (International Human Genome Sequencing Consortium, 2001). The other major type of repeats are 1–6 bp simple sequence repeats (SSRs) which occupy ∼3% of the human genome (International Human Genome Sequencing Consortium, 2001). Recent studies showed that allelic variations of HUMTH01 (TCAT repeats) correlate with quantitative and qualitative changes in the binding of ZNF191 protein, which contributes significantly to the control of expression of quantitative genetic traits (Albanese *et al.*, 2001). One possible role of such non-coding DNA may be to regulate expression of neighbouring genes by influencing the chromatin organization.

Bkm satellite DNA is one such simple sequence repeat, which is highly conserved in eukaryotes (Singh and Jones, 1982). The major component of Bkm is a tetranucleotide repeat of GATA (Singh and Jones, 1982). By using Bkm probe in Southern and *in situ* hybridization, the W chromosome of different species of snake can be identified except the primitive snakes having undifferentiated sex chromosomes (Singh *et al.*, 1980). In mouse the Bkm sequences are predominantly concentrated in the short arm of the Y-chromosome (Singh and Jones, 1982). This suggests that GATA repeats may have functional significance in the biology of sex chromosomes. In order to understand the function of such repeats we have undertaken an extensive analysis of the abundance and distribution patterns of GATA repeats. Here we report our analysis of the distribution patterns of GATA repeats in the human genome with special focus on the Y-chromosome.

*To whom correspondence should be addressed.

METHODS

The genome sequences were downloaded from ftp sites of NCBI, Sanger Centre, and Stanford University (online Supplementary Data 1: http://www.ccmb.res.in/bkm/gata.htm). The locations of GATA repeats on the Y-chromosomes were identified using tandem repeat finder program (Benson, 1991). We used the cutoff score as 50 and the alignment criteria as '2,7,7'. We have analysed the distribution of perfect GATA repeats of the length $\geqslant$12 base pairs (at least three GATA repeats) to estimate the GATA repeat density in all the human chromosomes. For this, a Java based program was written to identify the perfect tandem repeats of GATA in the whole human genome. Analysis of the matrix associated region (MAR) potential of the GATA flanking sequences was done using MAR finder (Singh *et al.*, 1997). For this analysis, the sequence window length of 200 was set and scan length was set as 10. The cutoff score for the matrix association potential was set to 0.65. We also used Excep, a program, which identifies the repeat clusters in a given sequence and allows us to find the probability of occurrence of a repeat (Klaerr-Blanchard *et al.*, 2000). The parameters set to find the ratio of expected to observed 'GATA' occurrence on a given sequence were: word length of four, window size of 50 000, and probability 0.01. The programs available at NCBI web page were used for other kinds of sequence analysis such as BLAST, ORF finder and mapping of the sequences to the cytogenetic locations.

RESULTS

Analysis of GATA repeats in model organisms

Analysis of the available sequences of complete and unfinished genomes of prokaryotes (archea and eubacteria) from the GenBank database revealed that tandem repeats of GATA sequences are not present in any of these genomes. We also extended the study to other model organisms such as *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Arabidopsis thaliana* and *Drosophila melanogaster*. In case of *S.cerevisiae*, the analysis revealed that the entire genome contained only two regions with a maximum of $(GATA)_7$ tandem repeats in chromosome VII. Similarly, *C.elegans* has only one $(GATA)_5$ and three $(GATA)_4$ repeat regions (Supplementary Data 2a). The *Drosophila* genome contains about ~86 GATA repeat regions, but only 11 of them contained GATA stretches of more than 10 tandem repeats (locations of GATA sequences and the genes flanking them are given in Supplementary Data 2b.

GATA repeat analysis in human genome

We analyzed the human genome for the occurrence of perfect tandem repeats of GATA. While Y-chromosome was found to have the highest GATA repeat density (222 bp/Mb), chromosomes 21 and X also showed above the average density found in other autosomes (Supplementary Fig. 1). Next, we looked for the repeat occurrence of different lengths of GATA repeats. As expected, starting from $(GATA)_4$, longer repeats showed a decreasing repeat occurrence. Surprisingly, however, this decrease in the repeat occurrence was not linear. We observed that repeat occurrence began to increase for longer repeats with repeat length from $(GATA)_9$, peaking at $(GATA)_{12}$ and then rapidly decreasing (Fig. 1). Analysis of autosomes also revealed a similar pattern. To test whether this pattern was specific for GATA repeats, we analysed the human genome for other tetranucleotide repeats such as GACA and GAAT in the whole genome and AAGT, AATG, GGGA and AAGG in the Y-chromosome. Supplementary Figure 2a and b show that this specific enrichment of $(GATA)_{10–12}$ is not found in these closely related repeats. Same trend was clearly observed in the case of mouse genome (Supplementary Fig. 2c). This suggests that $(GATA)_{10–12}$ repeats are unique and may have functional significance. With this rationale, we carried out a more detailed analysis of $(GATA)_{10–12}$ repeats in human Y-chromosome.

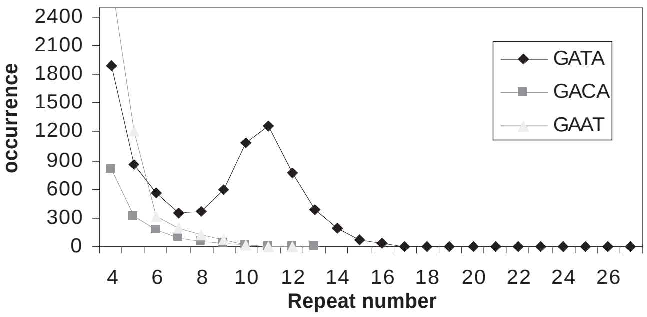

Fig. 1. Complete analysis of perfect tandem repeats of GATA in the human genome. The occurrence indicates number of times a repeat length occurs in the genome. Note the peaking of tandem repeats of $(GATA_{11–12})$. This selective accumulation of 11 to 12 GATA repeats is not seen even in the close tetranucleotide variants, suggesting that this trend is unique to GATA repeats.

Occurrence of 'GATA' units on sex chromosomes is not random

We analysed the complete DNA sequences of human X- and Y-chromosomes using the Excep program to find the expected/observed ratio of 'GATA' units. In case of the X-chromosome, out of 2547 windows analyzed, only 36 had higher observed/expected ratio. Similarly, in Y-chromosome, out of 460 windows, three showed higher observed value (data not shown). Of the 457 windows, the ones coming from a ~4 Mb region next to the

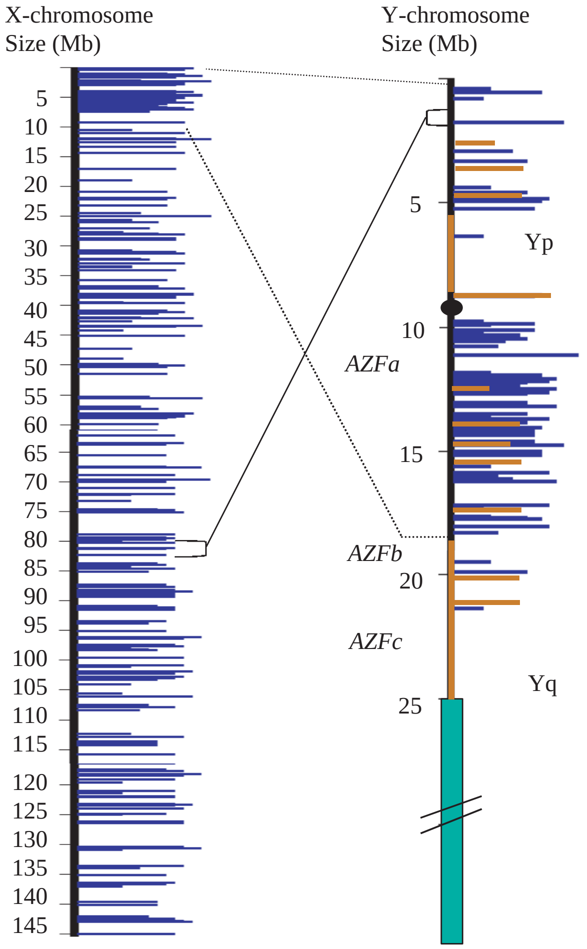

Fig. 2. Diagrammatic representation showing the distribution of GATA repeats along the X- and NRY of Y chromosomes: Connecting lines between the two homologous regions on the chromosome indicates the X-Y homologous regions. The GATA repeat regions, which are indicated as blue lines on the Y chromosome, show 84 to 98 % homology with that of X-chromosome, in their flanking sequence, which are concentrated in Xp22.3. A small segment of proximal Yp region shows homology to Xq21.3. Also note under representation of GATA in the Y segmentally duplicated regions. The distribution of GATA repeats on the X-chromosome shows the even distribution. This observation indicates that the proximal Xp arm, which is about 10Mb, contributes for the major portion of Y chromosome. The GATA flanking sequences that are specific to Y-chromosome are indicated as red lines. The green block indicates the hetrochromatic region on the Y-chromosome.

Yq heterochromatin were found to be almost devoid of GATA repeats (Fig. 2). Similarly, a ∼1.5 Mb region next to centromere in Yp, was GATA deficient. These observations indicate that the overall high density of GATA repeats on the sex chromosomes compared to the rest of the genome is in a non-random fashion.

GATA repeat analysis in human Y-chromosome

We could identify ∼89 imperfect regions of GATA repeats on the human Y-chromosome. Distribution of these repeats is not uniform along the length of the chromosome. The Y-specific regions (see below) were generally poor in GATA repeats while Yp arm and the Yq centromeric regions having homology with Xp arm were rich in these repeats (Fig. 2). After identifying the locations of tandem GATA repeats on the Y-chromosome, we took about 2.0 kb flanking sequences of each repeat region and BLASTed it against the human genome database from NCBI. There are ∼66 (60%), GATA repeat regions along with its flanking sequences on Y-chromosome, which show about 84–99% sequence similarity to that of the X- chromosome (Fig. 2). Sequences showing similarity to GATA repeat regions of both p and q arm of the Y-chromosome mapped to Xp 22.3. There were instances where the flanking sequences of GATA repeats were Y-specific and located in multiple copies (Fig. 2). Based on this analysis we could classify the Y-chromosome sequences into two types, viz. X–Y homologous sequences and Y-specific sequences. Finally, we did not find any significant ORFs within the GATA flanking sequences.

Association of sequences flanking the GATA repeats with MARs

The distribution pattern of GATA repeats on the Y-chromosome (Fig. 2) suggested that they might be marking the chromatin domain boundaries. To test this, we carried out Matrix Associated Region (MAR) analysis of the GATA flanking sequences. The results revealed that the sequences flanking GATA repeats had on both sides a very high matrix association potential (Supplementary Fig. 3), and the position of association with respect to GATA repeat location varied among the sequences analyzed. Moreover, the high level of conservation of GATA repeat flanking sequences between X- and the Y- chromosomes implies the functional significance of these sequences as they have been retained for millions of years since the evolution of Y. It was also observed that the GATA repeat regions present near the Yq centromeric region and in the proximal Yp region shared more homologies with that of the Xp22.33 and a small region at Xq21.3, respectively (Supplementary Data 3 and Fig. 2). Though the X–Y homologous genes are reported to be present in these regions of Xp and Yq, we report here that the intergenic sequences are also highly conserved. Interestingly, most of the duplicated segments were Y-specific having no counterpart on the X; and GATA tandem repeats were under represented in these regions (Fig. 2 and Supplementary Fig. 4). Such a high MAR association was not observed in case of other A/T rich tetramer repeats (Supplementary Data 4).

DISCUSSION

Current models on higher order chromatin organization and its functional consequences generally have at least one common aspect which can be summarized as follows: the 30 nm chromatin fiber is secured on some kind of matrix or scaffold to package the genome in the form of loops of an average size of 60 kb (Paulson and Laemmli, 1977; Earnshaw and Laemmli, 1983). Interaction of promoter with appropriate regulatory elements (enhancers, silencers) is restricted within a domain that is defined by the presence of boundary or insulator elements (Gerasimova and Corces, 2001; Labrador and Corces, 2002). The DNA sequence associated with matrix, MAR, may also coincide with boundary like elements that define the domains of gene activity in the genome (Namciu *et al.*, 1998). The distribution pattern of GATA and their association with MAR seems to point that these combinations define functional /structural domains of the genes positioned between such combinations. MARs have been implicated to be associated with regulatory elements (Gasser and Laemmli, 1986). We propose that GATA repeats, along with associated MARs, define domains on Y-chromosome that are regulated in a concerted fashion. This model will have several predictions: (a) there should be a set of proteins that will drive the expression of genes associated with these domains; (b) genes located in the GATA-rich or GATA-devoid regions will be expressed differently; and finally, (c) GATA repeats should have a regulatory function of its own.

Indeed, in our earlier studies, we have reported the presence of sex and tissue-specific Bkm-binding protein (BBP) in the ovary of snake and testis of mouse (Singh *et al.*, 1994a). This protein is different from the GATA family of transcription factors as it binds to repeats of GATA sequences. Occurrence of similar BBP in mouse testis and the coincidence of its expression with the testis-specific decondensation of the Y-chromosome indicated that the BBP is involved in some modulatory aspects of sex determining chromosomes (Singh *et al.*, 1994a). The absence of GATA repeats in the organisms, which do not have chromosomal sex determination, and the presence of very few GATA repeats in *S.cerevisiae* and *C.elegans* further strengthens the view that the GATA sequences are important for the function of sex chromosomes. Studies have shown that the mouse Y-chromosome remains condensed and therefore inactive in somatic cells but decondenses specifically in germ cells (primary spermatocytes; Singh *et al.*, 1994b). Taken together, the association of GATA sequences with the MARs (present study) and the presence of BBP, which binds to the GATA sequences of Bkm, suggest a functional significance of this association of GATA repeats with MARs. Furthermore, GATA rich regions of the Y contains genes that are expressed during early developments while most of the genes in the GATA devoid region are expressed late during spermatogenesis, suggesting a functional role for coordinated gene regulation mediated by GATA repeats.

The similarity between the GATA flanking sequences of X- and Y-chromosomes and their association with MARs implies that these sequences are specific to sex chromosomes and have evolved for specific function. Most of the GATA repeats are found to be intergenic (Supplementary Data 5), suggesting that they may, in association with MAR, function to regulate the entire domain defined by these sequences. This model will suggest two states of these domains in terms of chromatin organization based upon the protein composition at the GATA repeats. Inactive or the default state will be defined by the absence of any specific high affinity GATA repeat binding protein. The other state will be dominated by the interaction of high affinity/cooperatively binding protein (like BBP) with the GATA repeats. Considering structural parameters such as base stacking, propeller twist and protein deformability for repeat sequences (Baldi and Baisnee, 2000), we find that among the tetramer repeats with the high occurrence, GATA has the maximum bendability and is generally more flexible. Such flexibility might provide a favorable architecture where a series of DNA–protein and protein–protein interactions are possible. These interactions, in turn, will facilitate formation of ‘GATA repeat-protein complex’ in a co-operative fashion. GATA repeats may be functioning at more local level to repress only the associated genes and not the entire domain. $(GATA)_7$ motif has been shown to exhibit silencer activity in erythroid cells (Ramchandran *et al.*, 2000). These regulations are most likely to be mediated by transcription factors and not by proteins like BBP. Interestingly, however, clusters of GATA related sequences have been shown to be essential component of chromatin domain boundary elements (Zhao *et al.*, 1995; Cuvier *et al.*, 1998) and GATA motif occurs twice and interacts with sequence specific binding protein in *Fab-7*, a boundary isolated from bithorax complex of *Drosophila* (Mishra, personal communication). Any direct link with these GATA related sequences and GATA repeats discussed in this paper remains to be established. We would like to emphasize that there may be similar mechanisms operating to regulate the expression of genes on the sex chromosomes which may not be using GATA repeats. Simple sequence repeats, like GATA, may be a component of the mechanism regulating expression of genes located in the specific chromatin domains by modulating the higher order chromatin organization in response to specific signals.

ACKNOWLEDGEMENTS

We thank Vamsi Madhav, Ranjan George, Satish Kumar, M.W. Pandit and the team and others at Ingenovis for help in developing repeat-search program, to U.N.Sinha of NAL, Bangalore for providing supercomputing facility and DBT supported Bioinformatics facility at CCMB. SS is thankful to CSIR for fellowship.

REFERENCES

Albanese,V., Biguet,N.F., Kiefer,H., Bayard,E., Mallet,J. and Meloni,R. (2001) Quantitative effects on gene silencing by allelic variation at a tetranucleotide microsatellite. *Hum. Mol. Genet.*, **10**, 1785–1792.

Baldi,P. and Baisnee,P.F. (2000) Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths. *Bioinformatics*, **16**, 865–889.

Benson,G. (1991) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.

Cuvier,O., Hart,C.M. and Laemmli,U.K (1998) Identification of a class of chromatin boundary elements. *Mol. Cell. Biol.*, **18**, 7478–7486.

Earnshaw,W.C. and Laemmli,U.K. (1983) Architecture of metaphase chromosomes and chromosome scaffolds. *J. Cell Biol.*, **96**, 84–93.

Gasser,S.M. and Laemmli,U.K. (1986) Cohabitation of scaffold binding regions with upstream/enhancer elements of three developmentally regulated genes of *D. melanogaster*. *Cell*, **46**, 521–530.

Gerasimova,T.I. and Corces,V.G. (2001) Chromatin insulators and boundaries: effects on transcription and nuclear organization. *Annu. Rev. Genet.*, **35**, 193–208.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Klaerr-Blanchard, Chiapello,H. and Coward,E. (2000) Detecting localized repeats in genomic sequences: a new strategy and its application to *Bacillus subtilis* and *Arabidopsis thaliana* sequences. *Comput. Chem.*, **24**, 57–70.

Labrador,M. and Corces,V.G. (2002) Getting the boundaries of chromatin domains and nuclear organization. *Cell*, **111**, 151–154.

Namciu,S.J., Blochlinger,K.B. and Fournier,R.E. (1998) Human matrix attachment regions insulate transgene expression from chromosomal position effects in *Drosophila melanogaster*. *Mol. Cell Biol.*, **18**, 2382–2391.

Paulson,J.R. and Laemmli,U.K. (1977) The structure of histone-depleted metaphase chromosomes. *Cell*, **12**, 817–828.

Ramchandran,R., Bengra,C., Whitney,B., Lanclos,K. and Tuan,D. (2000) A $(GATA)_7$ motif located in the $5'$ boundary area of the human beta-globin locus control region exhibits silencer activity in erythroid cells. *Am. J. Hematol.*, **65**, 14–24.

Singh,L., Purdom,I.F. and Jones,K.W. (1980) Sex chromosome associated satellite DNA: evolution and conservation. *Chromosoma (Berl)*, **79**, 137–157.

Singh,L. and Jones,K.W. (1982) Sex reversal in the mouse (*Mus musculus*) is caused by a recurrent nonreciprocal crossover involving the X and an aberrant Y-chromosome. *Cell*, **28**, 205–216.

Singh,L., Wadhwa,R., Naidu,S., Nagaraj,R. and Ganesan,M. (1994a) Sex- and tissue-specific Bkm (GATA)-binding protein in the germ cells of heterogametic sex. *J. Biol. Chem.*, **269**, 25,321–25,327.

Singh,L., Panicker,S.G., Nagaraj,R. and Majumdar,K.C. (1994b) Banded krait minor-satellite (Bkm)-associated Y-chromosome-specific repetitive DNA in mouse. *Nucleic Acids Res.*, **12**, 2289–2295.

Singh,G.B., Kramer,J.A. and Krawetz,S.A. (1997) Mathematical model to predict regions of chromatin attachment to the nuclear matrix. *Nucleic Acids Res.*, **25**, 1419–1425.

Zhao,K., Hart,C.M. and Laemmli,U.K. (1995) Visualization of chromosomal domains with boundary element-associated factor BEAF-32. *Cell*, **81**, 879–889.