Published online 2 July 2019

Nucleic Acids Research, 2019, Vol. 47, No. 14 ***7247–7261***
doi: 10.1093/nar/gkz562

Mapping of scaffold/matrix attachment regions in human genome: a data mining exercise

Nitin Narwade[1,†], Sonal Patel[2,†], Aftab Alam[2], Samit Chattopadhyay[2,3,*], Smriti Mittal[4,*] and Abhijeet Kulkarni[1,*]

[1]Bioinformatics Centre, Savitribai Phule Pune University, Pune - 411 007, Maharashtra, India, [2]Chromatin and Disease Biology Lab, National Centre for Cell Science, Pune - 411 007, Maharashtra, India., [3]Indian Institute of Chemical Biology, 4, Raja S.C. Mullick Road, Jadavpur, Kolkata - 700 032, West Bengal, India. and [4]Department of Biotechnology, Savitribai Phule Pune University, Pune - 411 007, Maharashtra, India.

Received April 25, 2019; Revised June 08, 2019; Editorial Decision June 13, 2019; Accepted June 27, 2019

ABSTRACT

Scaffold/matrix attachment regions (S/MARs) are DNA elements that serve to compartmentalize the chromatin into structural and functional domains. These elements are involved in control of gene expression which governs the phenotype and also plays role in disease biology. Therefore, genome-wide understanding of these elements holds great therapeutic promise. Several attempts have been made toward identification of S/MARs in genomes of various organisms including human. However, a comprehensive genome-wide map of human S/MARs is yet not available. Toward this objective, ChIP-Seq data of 14 S/MAR binding proteins were analyzed and the binding site coordinates of these proteins were used to prepare a non-redundant S/MAR dataset of human genome. Along with co-ordinate (location) details of S/MARs, the dataset also revealed details of S/MAR features, namely, length, inter-SMAR length (the chromatin loop size), nucleotide repeats, motif abundance, chromosomal distribution and genomic context. S/MARs identified in present study and their subsequent analysis also suggests that these elements act as hotspots for integration of retroviruses. Therefore, these data will help toward better understanding of genome functioning and designing effective anti-viral therapeutics. In order to facilitate user friendly browsing and retrieval of the data obtained in present study, a web interface, MARome (http://bioinfo.net.in/MARome), has been developed.

INTRODUCTION

Eukaryotic cell is compartmentalized into several organelles and a well-defined nucleus that harbors the genetic material. The human DNA with an approximate length of 3 m is highly compacted to fit into relatively small nucleus. This compaction, however, does not render the DNA inactive. Rather, DNA is accessed in a tightly controlled and dynamic manner to facilitate regulated gene expression. The nuclear matrix, a three-dimensional filamentous RNA–protein meshwork, forms the basis of structural support for orderly compaction of DNA (1). The chromatin is organized into loops by virtue of DNA sequences that tether the chromatin to the nuclear matrix (2). These anchor sequences are known as scaffold/matrix attachment regions (S/MARs). Various proteins, called S/MAR binding proteins (S/MARBPs), are known to interact with S/MARs to facilitate chromatin looping (2). Such looping of DNA has been proved to be crucial for many cellular processes like DNA replication, transcription, chromatin to chromosome transition and DNA repair (3,4). Interestingly, the S/MARs that tether these loops to the nuclear matrix lacks sequence conservation (5,6). However, features related to their secondary structure appear to be conserved and functionally relevant (5,7). S/MAR sequences are thus known to possess features such as origin of replication (OriC), AT richness, kinked and curved DNA, TG richness, MAR signature and Topoisomerase-II sites (7–9).

The human genome comprehends about 3.2 billion base pairs organized into 23 pairs of chromosomes. It is esti-

*To whom correspondence should be addressed. Tel: +91 020 2569 0195; Fax: +91 020 2569 0087; Email: abhijeet@bioinfo.net.in
Correspondence may also be addressed to Smriti Mittal. Tel: +91 020 2569 4952; Fax: +91 020 2569 1821; Email: spmittal@unipune.ac.in
Correspondence may also be addressed to Samit Chattopadhyay. Tel: +91 33 2413 1157; Fax: +91 33 2473 5197; Email: samit@iicb.res.in
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
Present address: Nitin Narwade, National Centre for Microbial Resources, National Centre for Cell Science, Pashan, Pune - 411 021, Maharashtra, India.

mated to contain 20 000 protein coding genes. Each chromosome thus harbors several genes that are transcribed in highly regulated manner under a well-studied spatio-temporal control. Croft *et al*., in 1999, reported importance of nuclear matrix in regulation of expression of genes on chromosome 18 and 19. The study indicated that genes located on chromosome 19, that occupies an internal position in the nucleus and has close association with nuclear matrix, are transcribed actively. Whereas, chromosome 18, which preferentially occupies peripheral position in nucleus, shows lesser gene expression (10). Similarly, S/MARs have been shown to increase the expression and stability of the transgene in various organisms (5,11–13). Thus, the crucial role of S/MARs and nuclear matrix in organization and functioning of the genetic material is evident. Further, interplay between S/MARs and nuclear matrix has been well studied in various conditions including diseases (14–17). Therefore, these two important players that control genome topology and function appears to be lucrative targets for therapeutic interventions. However, even after significant efforts toward better understanding of chromatin biology, a comprehensive genome-wide map of S/MARs is not yet available for human genome.

Advancements in DNA sequencing technologies, the next generation sequencing (NGS) has made it possible to generate a large amount of sequence data in high-throughput manner. Chromatin pull down using antibodies specific to chromatin binding proteins followed by sequencing of enriched DNA fragments (ChIP-Seq) is one such NGS application. ChIP-Seq experiments for various S/MARBPs have also been performed in independent attempts by various laboratories and the data is available in public repositories (18–21). In the present study, we reanalyzed ChIP-Seq data of 14 different human S/MARBPs to understand their genome-wide binding patterns. This information was then used to make a comprehensive S/MAR dataset that is genome-wide and non-redundant across selected proteins.

The dataset thus provides genomic co-ordinates of human S/MARs. It also reveals S/MAR details such as length, chromatin loop size, nucleotide repeats, abundant motifs, chromosomal distribution and genomic context. Further analysis of this dataset also indicates that the identified S/MARs indeed act as hotspots for integration of retroviruses. Therefore, the data presented herewith gives a better insight of chromatin organization occurring by S/MARs and its implication in diseases.

MATERIALS AND METHODS

Dataset preparation

The ChIP-Seq data for 14 selected S/MARBPs, namely, BRCA1, BRIGHT, SMAR1, CEBPB, CUX1/CDP, CTCF, Fast1/FOXH1, HoxC11, Ku autoantigen, NMP4, Mut-p53, SAF-A/hnRNPU, SATB1 and YY1 were retrieved from ENCODE and NCBI-SRA database with their appropriate controls in FASTQ format (18–23). If available, sequence data for experimental replicates were also retrieved. The data generated from a single sequencing platform i.e. Illumina genome analyser having single-end read layout, only for untreated human samples were considered for the study. These sequence files were then analyzed by using the standard ChIP-Seq data analysis pipeline as described below.

Raw data quality control

The raw data quality of individual samples was assessed using FastQC tool v0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and then reads were trimmed using NGSQC toolkit v2.3.3 (http://www.nipgr.res.in/ngsqctoolkit.html) (24) for retaining good quality adapter free reads with average phred score ≥ 20.

Raw read alignment

The high-quality reads from individual control and pull down samples were aligned to the human genome GhCR38/hg38 assembly in independent attempts using bowtie aligner v1.0.0 (25) with default parameters. A pre-built bowtie genome index available at http://bowtie-bio.sourceforge.net/tutorial.shtml#preb was used for performing these alignments. The SAM files generated after alignment were converted in to binary alignment format i.e. BAM using *view* utility provided by SAMtools v1.3.1 (26). Polymerase chain reaction (PCR) duplicates from the obtained alignment files were removed using *rmdup* utility of SAMtools with default parameters.

Peak calling

Peak calling was carried out for BAM files of 14 S/MARBPs (control and pull down) using MACS v1.4.2 with default parameters. The obtained BED files were concatenated into single file for each S/MAR binding protein and then subjected to the sortBed utility. These sorted BED files were merged using mergeBed in independent attempts for different S/MARBPs to get unique peaks within the replicates (if available). This resulted in generation of 14 different BED files. These were further merged by subjecting them to Bedtools' multiIntersect utility, thereby generating a single bed file with intersect peak coordinates across all S/MARBPs. At last, bedtools' merge utility was used with default parameters to merge the overlapping peaks in this file. The genomic DNA sequences corresponding to these coordinates were fetched from UCSC-DAS server (http://genome.ucsc.edu/cgi-bin/das/hg38/dna?segment=chr:start,end) and saved as a multi-fasta file. These obtained sequence and BED coordinates were used for subsequent analysis.

Motif and nucleotide repeat analysis

The extracted DNA sequences were analyzed for presence of motifs using Linux-compatible, standalone MEME-ChIP v4.10.1 tool (27). The motif analysis was carried out using default parameters of MEME-ChIP program. Abundance of mono-, di-, tri-, tetra-, penta- and hexa-nucleotide repeats in these sequences were estimated using standalone MISA v1.0 microsatellite finding PERL program.

Annotation of peak coordinates

The peak coordinates were annotated using R package called ChIPseeker v1.12.1 (28). The tool annotates ChIP-Seq peaks and reports nearest downstream gene and peak distribution in different genomic elements like promoter, untranslated regions, intron, exon and intergenic regions. The pathways associated with the nearest downstream gene were retrieved using KEGGREST R package and gene ontologies were retrieved using UniProt/SwissProt database (https://www.uniprot.org/).

S/MAR-associated features

S/MARs are characterized by presence of features like OriC, AT richness, kinked and curved DNA, TG richness, MAR signature and Topoisomerase-II sites. Therefore, the extracted DNA sequences were verified for the presence of one or more of these features. The motifs that defines these features have been described earlier (8,9). Therefore, presence of these features in sequences were determined by presence of such specific motifs. In brief, presence of OriC was determined by detecting presence of ATTA or ATTTA or ATTTTA motif, AT richness by presence of two WWWWWW (where W is A or T) motifs intervened by 8–12 nt, Kinked DNA by the presence of TAN_3TGN_3CA or TAN_3CAN_3TG or TGN_3TAN_3CA or TGN_3CAN_3TA or CAN_3TAN_3TG or CAN_3TGN_3TA motif (where N is any nucleotide), Curved DNA by presence of $AAAAN_7AAAAN_7AAAA$ or $TTTTN_7TTTTN_7TTTT$ or TTTAAA (where N is any nucleotide), TG richness by the presence of TGTTTTG or TGTTTTTTG or TTTTGGGG motifs, MAR signature by presence of a bipartite sequence containing AATAAYAA and AWWRTAANNWWGNNNC (where W is A or T, Y is pyrimidine, R is purine and N is any nucleotide) and Topoisomerase II binding site by the presence of RNYNNCNNGYNGKT NYNY or GTNWAYATTNATNNR (where W is A or T, Y is pyrimidine, R is purine and N is any nucleotide) consensus.

These patterns were matched using custom PERL scripts written in house. Counts of sequences that have one or combination of these features are represented in the form of a venn diagram prepared using custom in house Javascript.

Nuclear matrix isolation

HCT116 cells were washed twice with phosphate-buffered saline. 5×10^6 cells were then lysed in extraction buffer (10 mM HEPES-KOH pH-7.2, 24 mM KCl, 10 mM $MgCl_2$, 1 mM phenylmethylsulfonyl fluoride (PMSF), 2 mM Dithiothreitol (DTT), 0.03% NP40 with protease inhibitors). The lysates were loaded on 0.8M sucrose bed and centrifuged at 6000 rpm for 20 min. The pellets containing nuclei were digested with DNase I for 30 min and then centrifuged at 6000 rpm for 10 min. The pellets were then washed with low salt buffer (10 mM HEPES-KOH, 0.2 mM $MgCl_2$ and 10 mM β-mercaptoethanol), high salt buffer (1.6M NaCl, 10 mM HEPES, 0.2 mM $MgCl_2$, 10 mM β-mercaptoethanol) and again low salt buffer, sequentially. EcoRI treatment was then given for 2 h at 37°C followed by centrifugation. The pellets were collected as nuclear matrix. DNA was purified using phenol-chloroform and precipitated using ethanol. The quality of the matrix was checked by agarose DNA electrophoresis and also by amplifying previously experimentally verified S/MARs (29,30). Two S/MARs from Girod *et al*., (29), namely, MAR 3–5 (P1) and MAR X-29 (P2) and three from Keaton *et al*., (30), namely, seq = 94 (P3) (chr18:23835886-23838503; Length = 2617), seq = 99 (P4) (chr18:24001839-24004790; Length = 2951) and seq = 1 (P5) (chr1:149425310-149430000; Length = 4690) were used as positive controls. The DNA was further used for amplifying S/MAR sequences using specific primers (Supplementary Table S3).

Mapping retroviral integration sites

Retrovirus Integration Database (RID) archives retroviral integration sites (IS) particularly, HIV-1 and HTLV-1. This information is archived in the form of genomic locus of integration (i.e. Chromosome and the coordinate as per hg19 genome build). RID archives 1 141 461 and 11 283 IS for HIV-1 and HTLV-1, respectively. In the present study, the S/MAR peak coordinates were deduced from hg38 assembly. Therefore, before mapping, all peak coordinates were converted to hg19 assembly using online version of UCSC liftover tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). HIV-1 and HTLV-1 IS were then mapped on to the converted peak coordinates. Number of IS residing within peak coordinates were then estimated. If the IS resides outside the peak coordinate, then its distance from nearest upstream and downstream S/MAR peak was determined. Only those IS that are flanked on either side of S/MAR peaks were considered for this analysis. All the mapping and distance estimations were carried out using custom PERL scripts written in house.

Development of web interface, MARome

The MARome web interface has been developed using Spring Framework - 1.2.1, Apache Maven, HTML5, JavaScript5, CSS3, Bootstrap3, Java - 1.8, PostgreSQL - 9.3.19. For automation/parsing, custom PERL scripts have been used wherever necessary. MARome is freely available at http://bioinfo.net.in/MARome.

RESULTS

Identification of S/MAR coordinates in the human genome: the dataset preparation

S/MARBPs are known to bind S/MAR regions. A non-redundant set of binding patterns of several SMARBPs can thus be used to trace S/MARs, in a genome-wide manner. Therefore, ChIP-Seq data of 14 different S/MARBPs, namely, BRCA1 (31), BRIGHT (32), SMAR1 (33), CEBPB (34,35), CUX1/CDP (36), CTCF (35,37,38), Fast1/FOXH1 (35), HoxC11 (35), Ku autoantigen (39), NMP4 (35,40), Mut-p53 (41), SAF-A/hnRNPU (35,42), SATB1 (35,36) and YY1 (40) were retrieved from public repositories. The accession numbers and other relevant information about the data used in present study is provided

in Supplementary Table S1. After quality assessment and filtering of raw data, the high-quality reads were aligned to the human genome hg38 assembly. The detailed alignment statistics is provided in Supplementary Table S2.

Peak calling using MACS14 resulted in a total of 452 881 peaks across all S/MARBPs which, also includes peaks resulted from their experimental replicates. At last, overlapping coordinates were merged resulting in a total of 298 443 peak coordinates. These peak coordinates are thus average representation of binding sites of one or more of the selected 14 S/MARBPs and are non-redundant.

Validation of dataset

In order to verify if the identified peak coordinates are indeed genomic locations for DNA sequences that resemble S/MARs, the nucleotide sequences corresponding to these coordinates were fetched from UCSC-DAS server. The nucleotide sequences were then analyzed for presence of S/MAR associated features such as OriC, AT richness, kinked and curved DNA, TG richness, MAR signature and Topoisomerase-II sites. The analysis revealed that, out of 298 443 curated sequences, 283 568 sequences show presence of at least one of these features indicating S/MAR like nature of these sequences. There were 14 857 sequences that lacked these features. OriC (272 016, ∼91%), AT richness (196 611, ∼66%) and Kinked DNA (178 960, ∼60%) were the most abundantly occurring features. The least represented feature was presence of Topoisomerase-II sites (9973, ∼3.3%). A total of 52 567 S/MARs showed presence of combinations of six features and only 190 S/MARs showed presence of all the seven features. (Figure 1A and B).

S/MARs and inferred topological details

In the present study, a total of 283 568 S/MARs were identified in human genome. The length of these S/MARs range from 33 to 61 755 bp with a median length of 596 bp. The aggregate length of all these sequences (230 177.6 kb) accounts for 7.4% of human genome. Out of these sequences, 269 046 i.e. 94.87% have length ≤2 Kb (Figure 2A).

The chromatin is tethered to the nuclear matrix by virtue of S/MARs thereby generating inter-S/MAR chromatin loops. We therefore, searched segments of genome that are flanked on either side by identified S/MAR coordinates/sequences. We identified a total of 283 453 inter-S/MAR regions or loops. Analysis of these loops revealed that their size ranges from 1 bp to 30 025.7 kb, with a median length of 4923 bp. Further, 267 096 number of chromatin loops, i.e. 94.23% of total identified loops have their length less than or equal to 31 Kb (Figure 2B).

Chromosome-wise distribution of S/MARs

In order to determine if S/MARs follow a random distribution or have preference for localization over specific chromosomes, the S/MARs coordinates obtained in the present study were visualized over chromosomes in the form of a circular plot Ideogram (Figure 3A). The S/MAR density per chromosome was also calculated. It was observed to be 95.74 S/MARs per Mb of genome for autosomes. Allosomes, however, showed a distinctly less S/MAR density as compared to autosomes. The Y and X chromosomes showed 10.8- and 1.7-fold lower densities of S/MARs compared to autosomes, respectively. On an average, presence of approximately 10 S/MARs per gene was detected. The S/MAR count per chromosome is represented in Figure 3B. Further, a positive correlation was observed between S/MAR density and gene density (Figure 3C). The details of gene number/density, S/MAR number/density for each human chromosome has been presented in Table 1.

Distribution of S/MARs in genomic elements

We determined distribution of S/MARs in various genomic elements. Approximately, 96.3% of S/MARs were found to be located in the non-coding region of genome. Out of them, 21% were found to be located in the promoter regions. Presence of S/MAR in promoter region is associated with transcriptional regulation of the downstream gene. Notably, miR-222, miR-34a, miR-371a, Bax, Cyclin D1, NFκB, CD40, FN1 and PDGFRB genes showed presence of S/MAR within 1 Kb region upstream to their transcription start sites (TSS). Presence of S/MARs in the promoters of these genes has already been demonstrated experimentally (21,43–49). Further, 35.57% of the total S/MARs were found to be located in the intergenic region (Figure 4A). It was also observed that 15 614 of the total identified S/MARs were present within −100 to +100 bp of TSS of 14 425 genes (Figure 4B). This accounts for 26.78% of total human genes (total number of genes is 58 288 as per GENCODE hg38 statistics https://www.gencodegenes.org/stats/current.html). Presence of S/MARs around TSS of such a high number of genes highlights essentiality of these elements for transcriptional regulation of genes.

Functional categorization of S/MAR-associated genes

It was observed that 20 905 of the total S/MARs overlap exactly with the TSS of 15 319 genes. Therefore, functional characterization of the genes containing S/MARs within 1.5 kb of their TSS was carried out. The genes were analyzed for enriched GO terms and pathways using UniProt/SwissProt and KEGG pathway analysis, respectively. The most represented molecular functions included transcription and post-translation; biological process included immune response, transcription and cell signaling; cellular components included extracellular regions, nucleus and extracellular space. This highlights the importance of S/MARs in overall gene expression program (Figure 5A). Pathway analysis of these genes revealed that 26% of these genes belong to metabolic pathways, 23% of them belong to signaling pathways, 16% of them belong to cancer related pathways, 7% belong to human papilloma virus infection related pathways and 5% are related to HTLV1 infection (Figure 5B). A high fraction of these S/MAR associated genes showed link with diseases (data not shown).

Nucleotide composition of S/MARs

Nucleotide sequence of the DNA is known to strongly influence its structure. Changes in nucleotide composition or order has been shown to influence DNA structure and

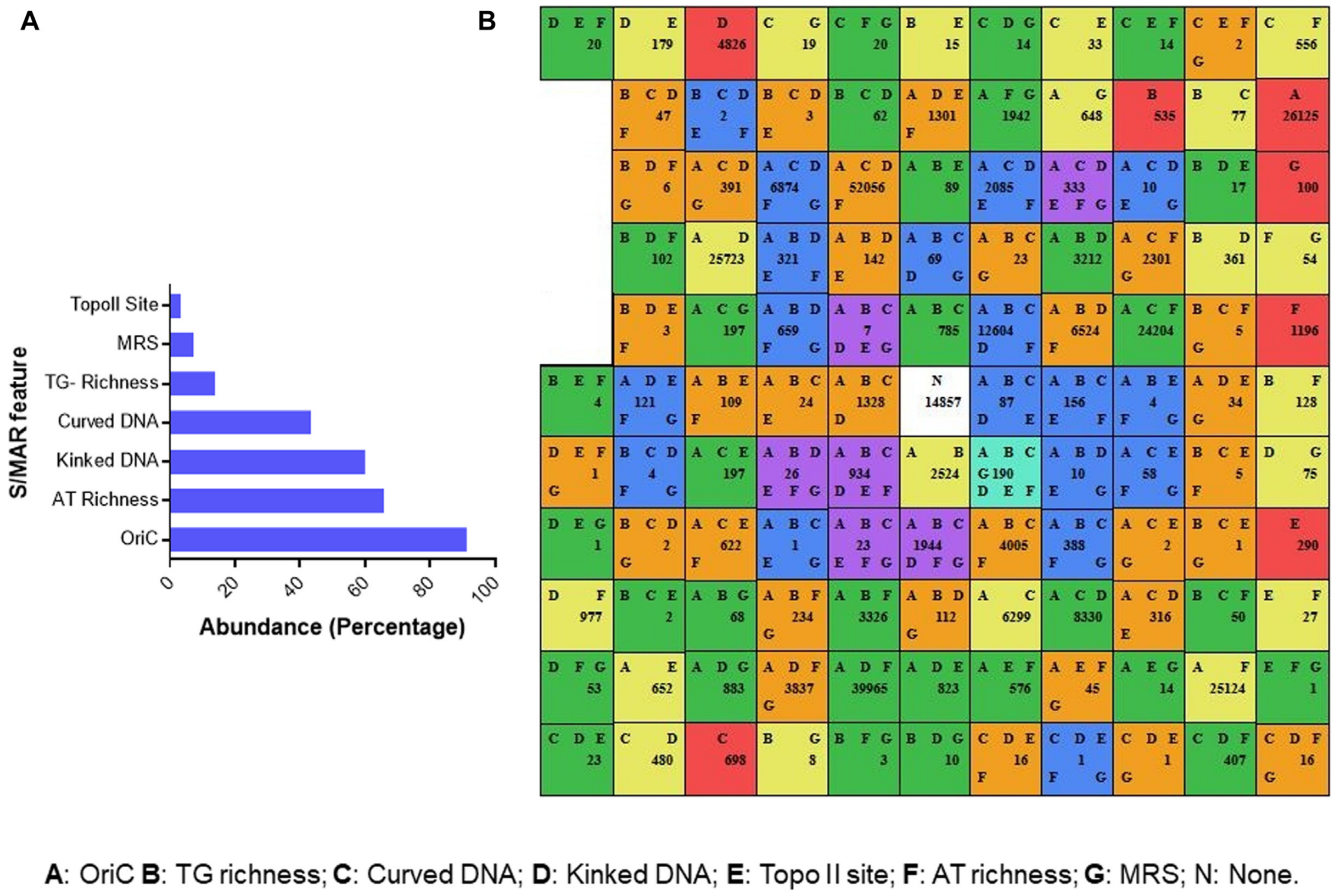

Figure 1. Validation of dataset by determining presence of S/MAR-associated features. (**A**) Abundance (in percentage) of seven S/MAR features including OriC, TG richness, curved DNA, kinked DNA, Topo II site, AT richness and MRS in the dataset. (**B**) Venn diagram depicting number of S/MAR sequences having one or more features.

Table 1. Distribution of genes and S/MARs on human chromosomes

Chromosome	Size (Mb)	S/MAR Count	S/MAR density/Mb	Number of Genes	Gene density/Mb
chr1	248.9564	25 689	103.1867	2785	11.1867
chr2	242.1935	24 405	100.7665	1791	7.394913
chr3	198.2956	18 543	93.51193	1541	7.771228
chr4	190.2146	14 907	78.3694	1066	5.604198
chr5	181.5383	16 524	91.02214	1288	7.094923
chr6	170.806	16 841	98.59725	1416	8.290108
chr7	159.346	15 428	96.82077	1318	8.27131
chr8	145.1386	13 440	92.60112	1008	6.945084
chr9	138.3947	11 982	86.57845	1105	7.984409
chr10	133.7974	13 264	99.13494	1084	8.1018
chr11	135.0866	13 270	98.23327	1658	12.27361
chr12	133.2753	13 964	104.7756	1369	10.27197
chr13	114.3643	7703	67.35492	619	5.412527
chr14	107.0437	8567	80.03272	931	8.697381
chr15	101.9912	9017	88.4096	988	9.687111
chr16	90.33835	9427	104.3521	1125	12.45318
chr17	83.25744	10 989	131.9882	1556	18.68902
chr18	80.37329	6487	80.7109	425	5.287827
chr19	58.61762	7813	133.2876	1774	30.26394
chr20	64.44417	7349	114.0367	772	11.97936
chr21	46.70998	3428	73.38902	410	8.777567
chr22	50.81847	4527	89.08179	633	12.4561
chrX	156.0409	8773	56.22244	1151	7.376271
chrY	57.22742	509	8.894338	141	2.463854

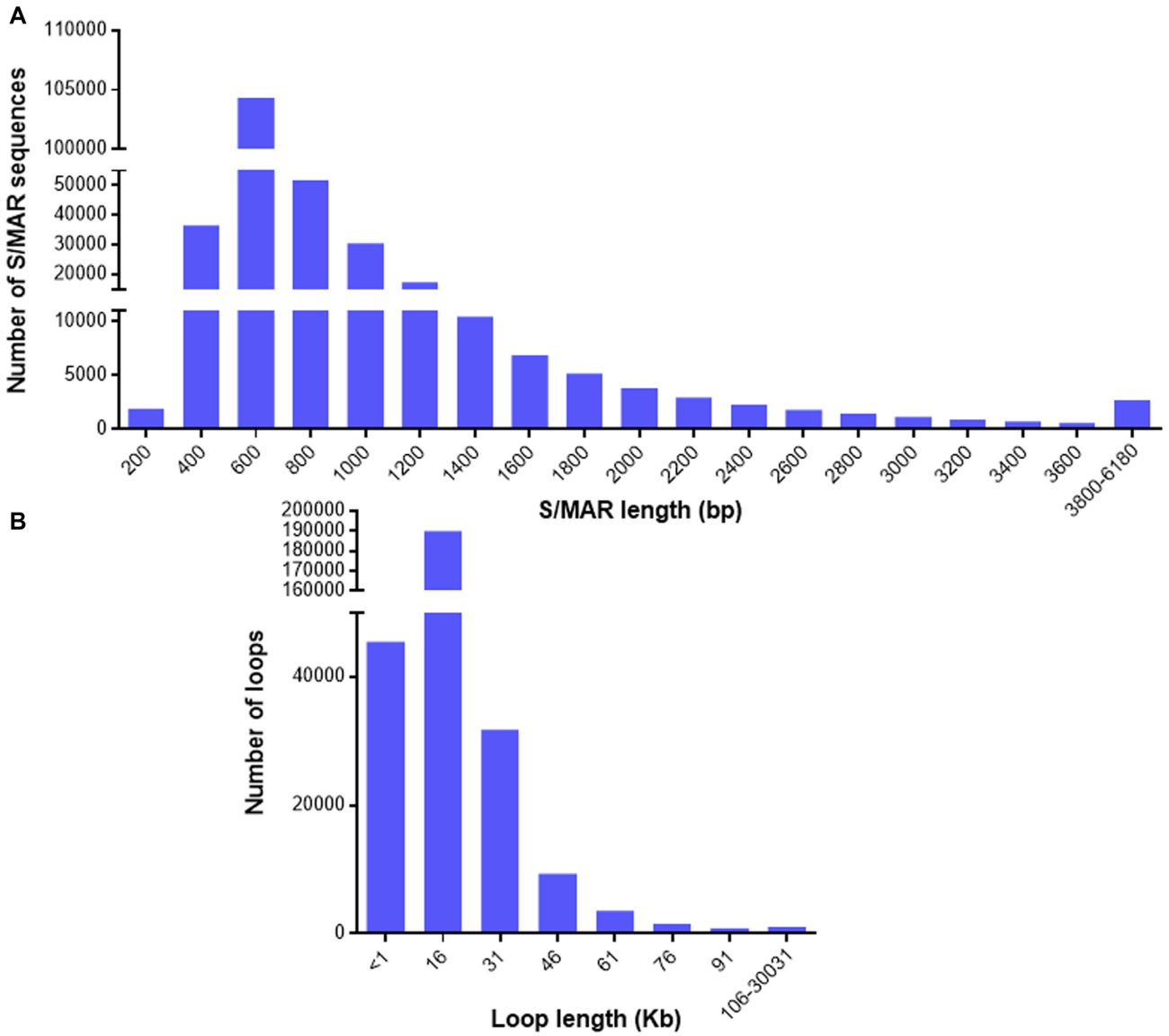

Figure 2. Length distribution of S/MARs and chromatin loops. (**A**) Length of S/MARs (in bp) was plotted against their occurrence. (**B**) Inter-S/MAR distance or chromatin loop size (in Kb) was plotted against their occurrence.

DNA–protein interaction that regulate vital cellular process (50,51). Function of S/MARs also associates with structural features such as kinks and curves in DNA and thus these elements also have characteristic nucleotide composition. Therefore, nucleotide repeat and motif analysis of S/MAR sequences was carried out. Abundance of various mono-, di-, tri-, tetra-, penta-, hexa-nucleotide repeats was determined (Figure 6A). The analysis revealed that $[A]_{\geq 10}/[T]_{\geq 10}$ repeat was the most abundant pattern (75 023 times) in the dataset indicating A/T richness of these sequences. The same was also evident from motif analysis done using MEME-ChIP program. Motif 1 with pattern GAGGYRGAGGTTGCAGTGAGC occurred in 7161 S/MARs. Motif 2 with A/T rich TTTTTTTTTTTG AGAYRGAGTYTYRCTCT occurred in 4055 S/MARs. Details of other nucleotide repeats and motifs predicted by MEME has been shown (Figure 6B–D). Abundance of different types of repeat patterns were also checked. Tandem repeats, direct repeats and palindromes were found to be most represented in S/MAR dataset (Figure 6E).

Experimental validation of human S/MARs

To experimentally validate the identified S/MAR sequences, the nuclear matrix DNA from human colon cancer cell line, HCT116, was isolated and used as template. The matrix DNA quality was determined by agarose gel electrophoresis and also by amplifying five previously experimentally proven S/MARs (29,30) (Figure 7A). Thirty representative S/MAR sequences from the entire dataset were chosen randomly and amplified using specific primers. Two randomly chosen inter-S/MAR sequences were used as negative controls (Figure 7B). It was observed that all 30 S/MARs showed specific amplification (Although, sequence number 19 amplified in less amount) (Figure 7B–D).

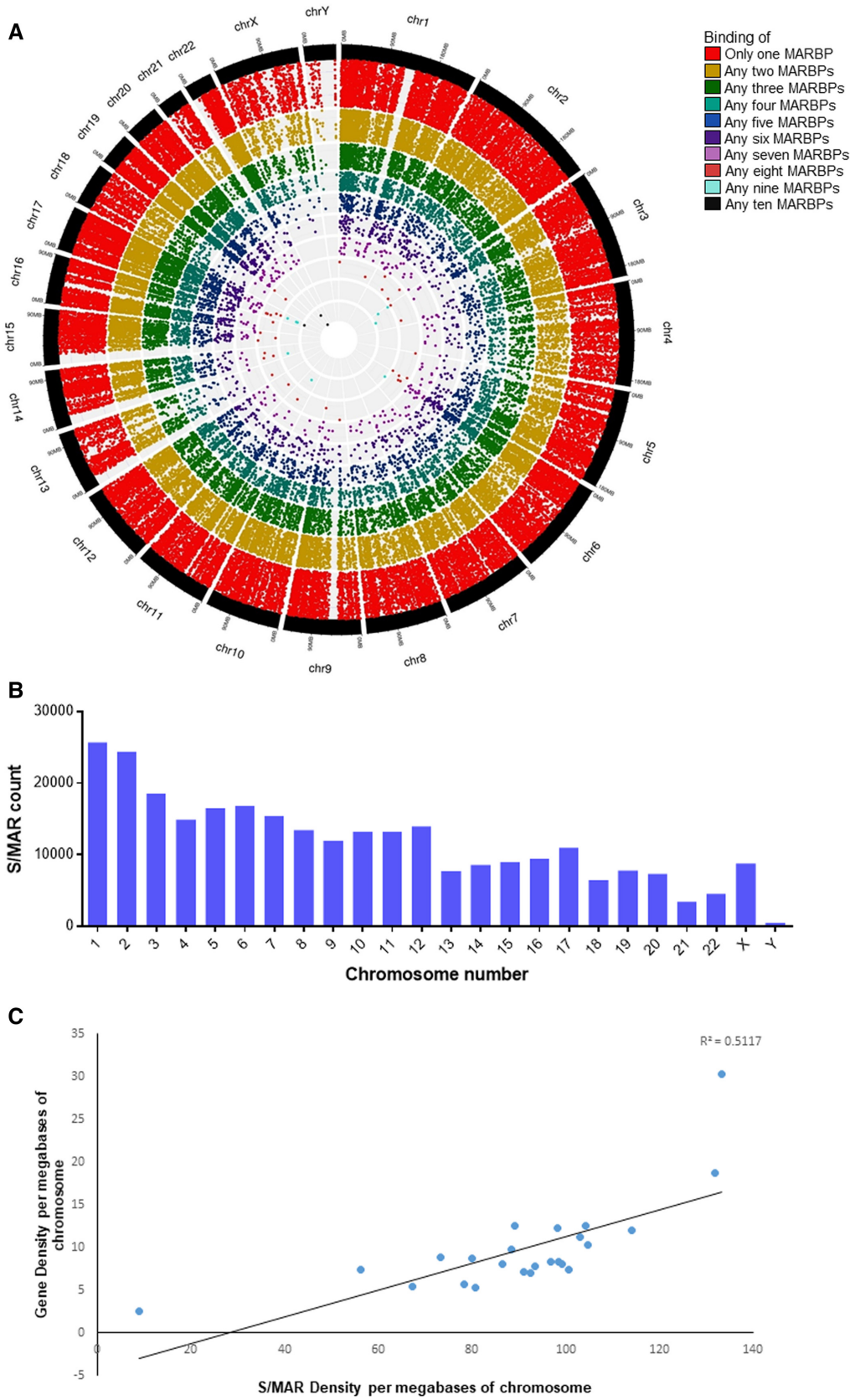

Figure 3. Distribution of S/MARs on human chromosomes. (**A**) Visualization of S/MARs on all human chromosomes. (**B**) Number of S/MARs present on each human chromosome. (**C**) Gene density and S/MAR density correlation graph for each human chromosome.

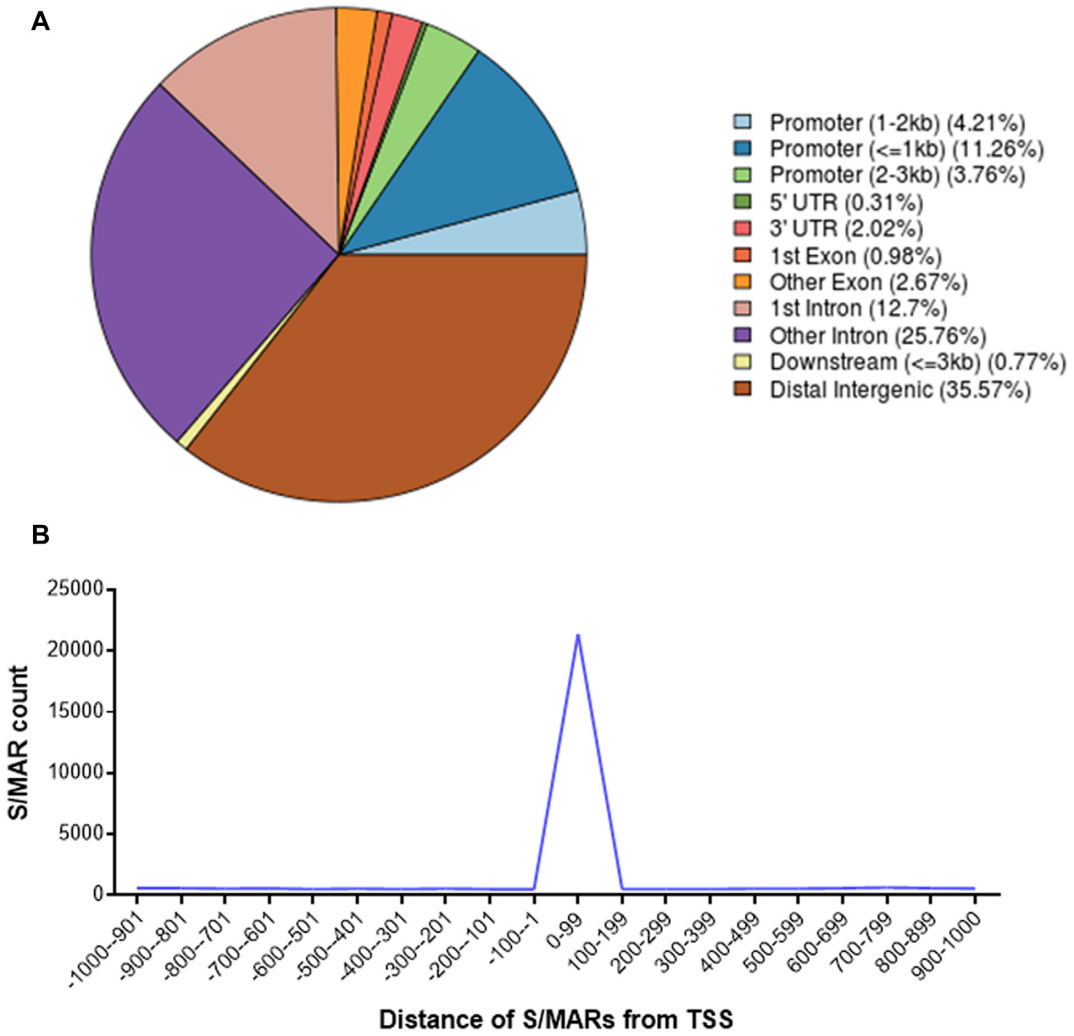

Figure 4. Genomic context of S/MARs: (**A**) Percentage distribution of S/MARs in different genomic regions. (**B**) Distance of S/MARs from the TSS of nearest downstream gene versus S/MAR count.

Thus, randomly chosen 30 S/MAR sequences were experimentally proved to be part of nuclear matrix.

S/MARs: hotspots of retroviral integration

Retrovirus integration is not a random event, various viral and host factors are known to mediate this process. One such factor discussed earlier is the S/MARs of the host genome (17). In order to determine whether S/MARs identified in the present study has any correlation with retrovirus integration event, HIV-1 and HTLV-1 insertion sites (IS) were mapped on to the identified S/MAR coordinates. A very strong correlation was observed between 'presence of S/MAR' and 'presence of IS' for HIV-1 and HTLV-1. Out of total mapped 1 141 899 HIV-1 IS, 102 408 IS were present exactly within S/MAR coordinates. Further, 599 389 (52.5%) IS were present within 5 kb and 956 873 (84%) IS were present within 15 kb region of identified S/MARs (Figure 8A). In case of HTLV-1, out of total 11 286 mapped IS, 1059 were located exactly within S/MAR coordinates. A total of 4986 (44%) IS were present within 5 kb of S/MAR sites and 8169 (72%) IS were present within 15 kb region around S/MARs (Figure 8B).

MARome web interface

Using MARome, S/MARs identified in the present study and related annotation (both for hg19 and hg38 assemblies) can easily be browsed using various search strategies. MARome provides search options by unique IDs, genomic coordinates, query sequences and gene ID/symbol. In MARome, every S/MAR entry is represented by unique identifier. With prior knowledge of these identifiers, user can browse particular S/MAR using search by ID strategy. Users can submit genomic coordinates of their interest in standard bed format to retrieve S/MARs available at and around loci of their interest. Search by sequence strategy provided by MARome allow users to search S/MARs similar to query sequence of their interest. This strategy internally runs NCBI-blast+ blastn against identified S/MAR sequences and returns the best hit along with top 10 align-

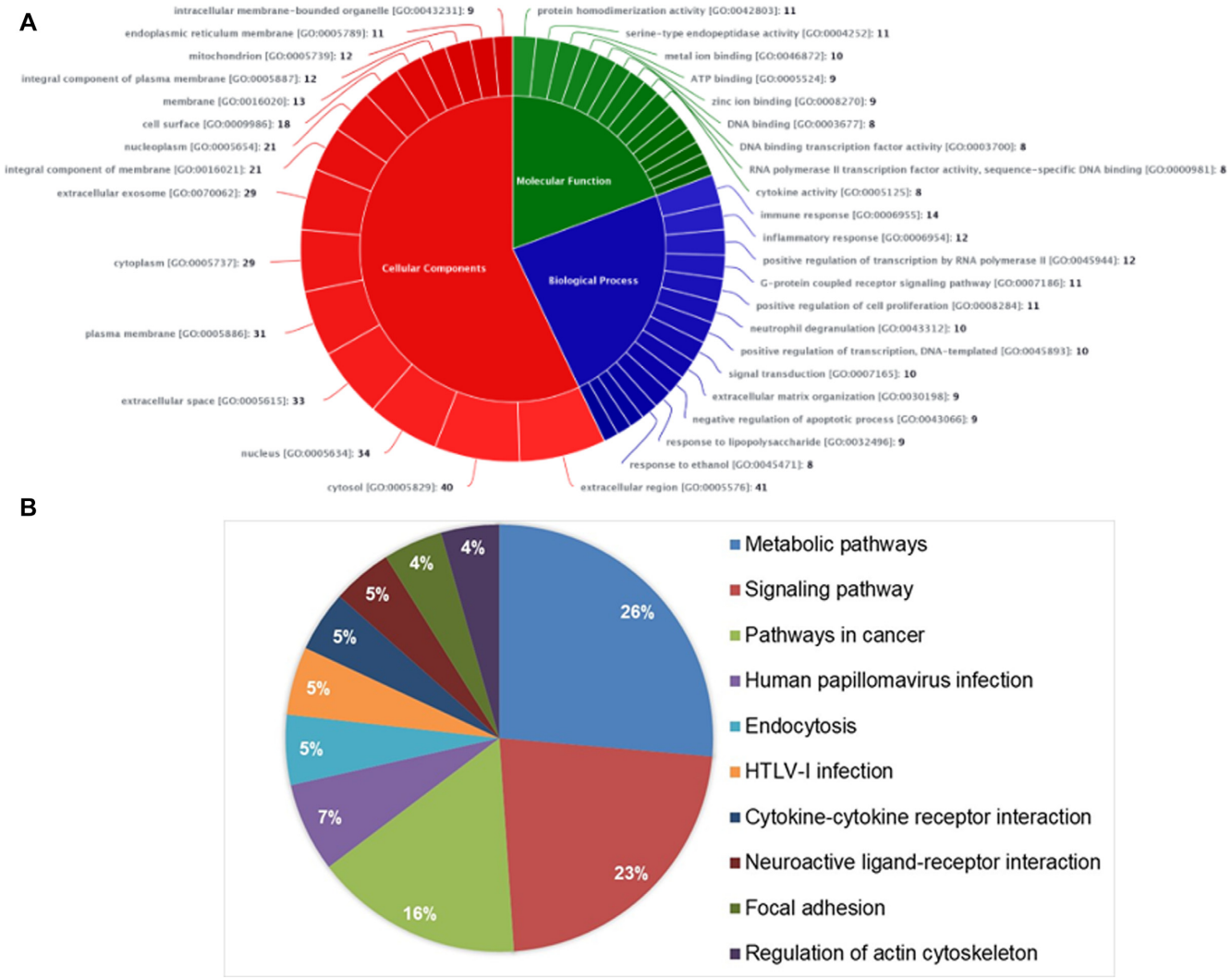

Figure 5. Functional classification of S/MAR associated genes. (**A**) Classification of genes based on gene ontology; Biological Processes. (**B**) Classification of genes based on their involvement in different pathways.

ments. Similarly, users can search S/MAR associated genes of their interest using search by Gene Name/Symbol strategy. The tabular output obtained through every search strategy further provides, SMAR binding proteins targeting SMARs, SMAR associated features, location of SMARs in genome context/element, its distance from TSS of nearest gene and HTLV/HIV insertion sites associated with SMARs. The output data are also cross-linked to public databases like NCBI-gene and ENSEMBLE for further annotation details. It is also cross-linked to UCSC Genome browser for data visualization. The interface also allows complete and S/MARBP-wise download of S/MAR sequences, coordinate files, annotations, etc. in bed and tsv formats. Further, a scoring scheme (details provided in online help manual of MARome) that considers number of S/MARBPs, number of different 'S/MAR associated features' and number of times 'S/MAR associated features' appears in a particular S/MAR has been implemented in the database to score the S/MAR entries.

DISCUSSION

Spatio-temporal control of gene expression is a hallmark of multicellular organisms. Apart from the individual's genetic makeup, epigenetics also plays a vital role in shaping differential phenotypic traits. Epigenetic regulation occurs through histone modifications, DNA methylation, non-coding RNAs and regulatory elements such as Locus Control Regions (LCRs), S/MARs etc. Chromatin organization, an integral part of gene regulation is brought about by DNA sequences called S/MARs (1). These S/MARs act as topological sinks that hold the chromatin loops to nuclear matrix and are involved in context-dependent activation or repression of the surrounding genes. However, the molecular mechanism underlying this loop organization remains poorly characterized. Defects in S/MARs have also been implicated in various diseases like cancers, inflammatory diseases, facioscapulohumeral dystrophy and viral infections (14–16,52). In this context, a map of all the characterized S/MARs in human genome would be beneficial in understanding chromatin- and disease-biology. Toward

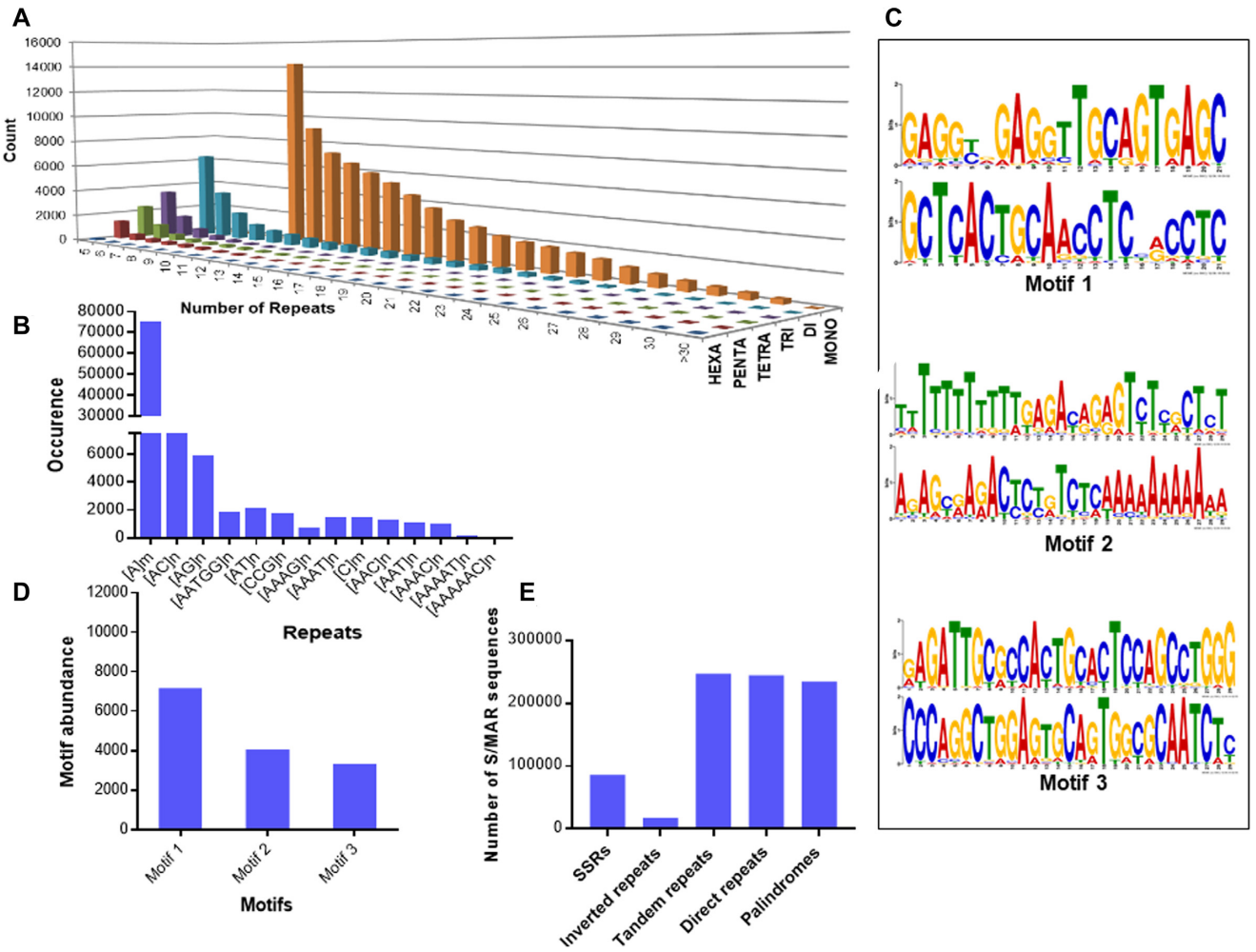

Figure 6. Repeats and motifs present in S/MAR sequences. (**A**) Graphical representation for number of various mono-, di-, tri-, tetra-, penta- and hexanucleotide repeats present in S/MARs. (**B**) Occurrence of 12 abundant nucleotide repeats in S/MAR sequences. (**C**) Three most abundant motifs as identified by MEME-ChIP program in the S/MARs. (**D**) Graphical representation of abundance of the identified motifs. (**E**) Abundance of various repeats in S/MAR dataset.

this objective, we reanalyzed ChIP-Seq data of 14 different human S/MARBPs, namely, BRCA1, BRIGHT, SMAR1, CEBPB, CUX1/CDP, CTCF, Fast1/FOXH1, HoxC11, Ku autoantigen, NMP4, Mut-p53, SAF-A/hnRNPU, SATB1 and YY1 to understand their genome-wide binding patterns. This information was then used to make a comprehensive S/MAR dataset that is genome-wide and non-redundant across selected proteins.

We obtained 452 881 peak coordinates by analyzing ChIP-Seq data of the selected S/MARBPs. The peak number reduced to 298 443 by drawing peak intersects and by merging the overlapping peaks. This indicates that there is ∼70% redundancy in identified binding sites and multiple S/MARBPs target same/adjacent genomic loci. Analysis of protein-protein interaction data available in 'Biological General Repository for Interaction Datasets' (BioGRID) indicates that the selected S/MARBPs interact with each other. Therefore, these proteins can form multiprotein complexes or co-localize together while targeting specific genomic loci. The same can account for the redundancy in their binding sites observed in the present study. It also confirms strong S/MAR potential of the identified coordinates. DNA sequences corresponding to these coordinates can thus be considered as S/MAR dataset.

Curves and kinks in DNA have been recognized as a vital structural feature that favors DNA–protein interactions. Sequences with kinked and curved DNA signatures are prone to undergo kinking and curving in response to binding of accessory factors that induce distortions in DNA. Such distortions, in turn favors binding of other protein factors to mediate biological processes (53–55). In present study, ∼60% and 43% of identified SMARs have kinked and curved DNA signatures, respectively. The ability of S/MARs to interact with a variety of regulatory proteins which, ultimately regulates gene expression can thus be explained.

Similarly, DNA molecules that are rich in AT stretches are flexible and are prone to strand separation. They are also susceptible to superhelical stress-induced duplex destabilization (56). OriC is one such element that contains AT stretches, making it prone to strand separation, thereby facilitating initiation of DNA replication (57). S/MARs are

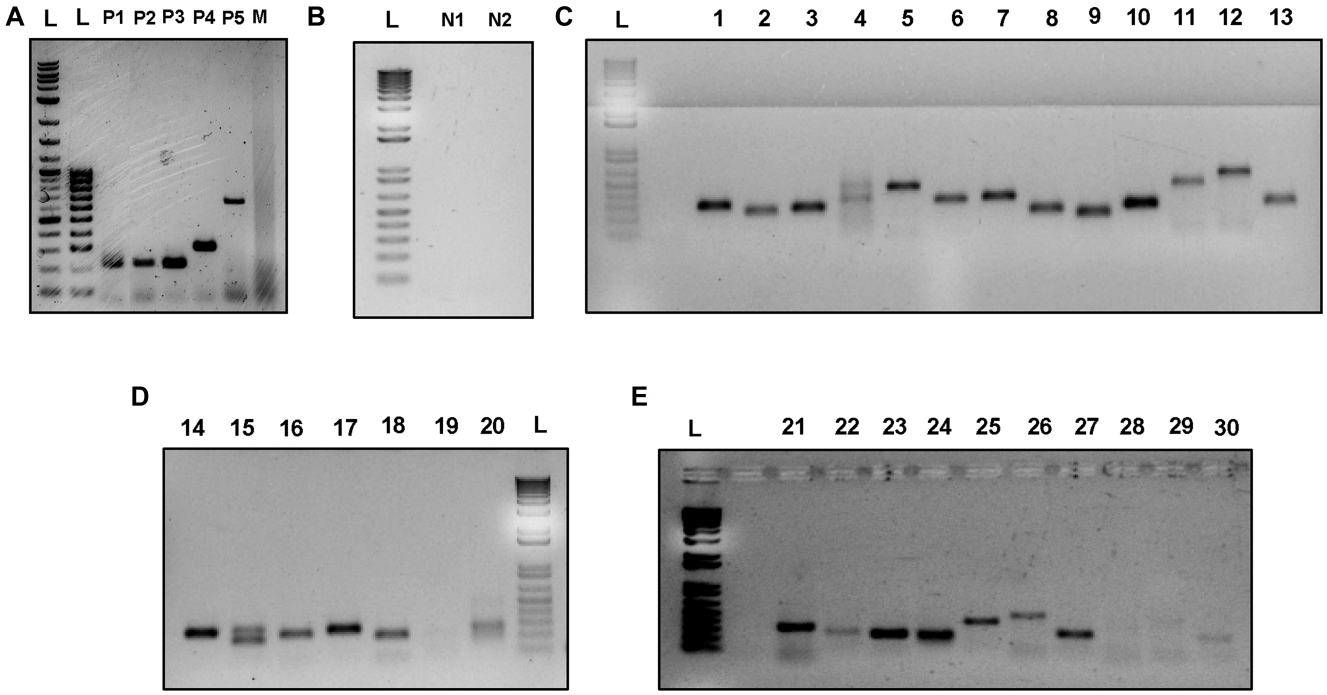

Figure 7. Experimental validation of S/MAR sequences by nuclear matrix DNA PCR. Matrix-DNA preparation: M; Semi-quantitative PCR for positive controls: P1-P5; (**A**), negative controls: N1 and N2 (**B**) and randomly selected 30 S/MAR sequences (**C**–**E**).

known to possess both these features. In present study, ∼91% of identified S/MARs have OriC signatures and ∼66% of them have signatures of AT richness. Thus, role played by S/MARs in biological processes such as replication, transcription and repair (viz., regulated DNA strand separation) can be supported.

The S/MAR length and the inter-S/MAR chromatin loop size are major determinants of chromatin structure and function. There is a lot of disparity about length of S/MARs in published literature and they are discussed to be 100 bp to several kb long (30,58,59). The median S/MAR length observed in the present study is 596 bp and 94.87% of identified S/MARs have length ≤2 kb. Thus in general S/MARs are small stretches of DNA having varied lengths. The dataset also contain small number of exceptional S/MARs that are longer or shorter than the observed median length. Similarly, the size of chromatin loop is reported to vary from 20 to 200 Kb (60,61). Functionally related genes tend to co-localize on same chromatin loop to facilitate their expression in a concomitant manner (45). In the present study, the median length of the chromatin loop was observed to be 4.923 kb and 94.23% of the identified chromatin loops have length ≤31 kb. The dataset also contain small number of exceptional chromatin loops that are longer or shorter than the observed median length accounting for the huge standard deviation of 76.35 kb. It has been reported that the chromatin loop size varies depending upon its position on the chromosome and correlates with size of replicon (62,63). Telomeric regions tend to have smaller loop size than the ones found away from the telomeres (64). Size of loops are also hypothesized to influence the biological state of the cell. Increase in the length of loops is linked with cellular differentiation whereas its decrease is associated with proliferation (65). Thus the observed chromatin loop lengths should be considered with a clear caveat that they can be influenced by various factors in dynamic cellular environment.

S/MARs found on different chromosomes have different structural as well as functional implications. Chromosome 18 and 19 are shown to have differential S/MAR densities that correlates well with expression profile of genes located on them (10). In the present study S/MAR density was determined for different chromosomes. Allosomes were observed to have lower S/MAR density as compared to autosomes. The data revealed a positive correlation between gene density and S/MAR density. It is known that chromosomes have preference for nuclear territories (66). It was observed that the chromosomes that occupy central position in nucleus (chr1, 16, 17 and 19) had higher S/MAR density than the chromosomes that occupy nuclear periphery (chr2, 4, 13, 18).

Anchorage of S/MARs to nuclear matrix is known to play a dual role. (i) Structural role to maintain the higher order chromatin confirmation and (ii) functional role in regulation of DNA replication and gene expression. The S/MAR size and loop length are responsible for up-keeping the structural domains of chromatin. The functional aspect of S/MARs can partly be answered on the basis of the genomic loci they occupy. Recent reports suggest that S/MARs can influence transcription by insulating nearby genes (67,68), thus making them act either as activator or repressor for the transgene in a context dependent manner (69). Localization of S/MARs in different genomic elements such as promoters, introns and intergenic regions has been demonstrated earlier (70,71). Differential distribution of S/MARs across various genomic elements, de-

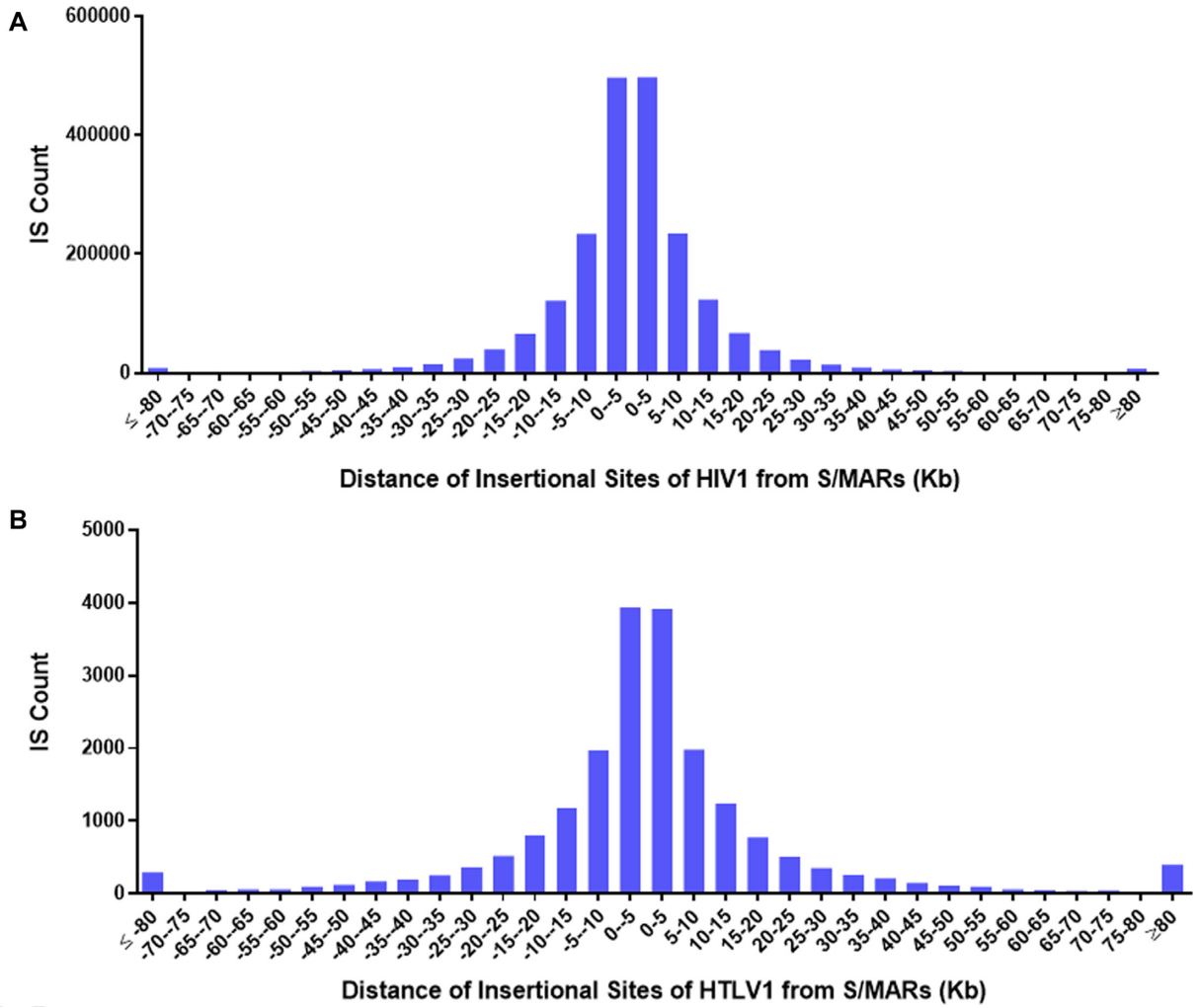

Figure 8. Correlation between S/MARs and retrovirus integration sites. (**A**) Distance of HIV integration sites from the nearest upstream and downstream S/MARs plotted against their count. (**B**) Distance of HTLV integration sites from the nearest upstream and downstream S/MARs plotted against their count.

termined in the present study, revealed an inverse correlation between coding regions of genome and the presence of S/MAR. Thus a majority of S/MARs were present in the non-coding region of genome indicating their regulatory functions. Also, S/MARs have been reported to be associated with the TSS, thereby influencing the transcription of downstream gene (72,73). In agreement with this, a number of S/MARs identified in the present study overlapped with TSS of high number of genes which, can be attributed to their role in transcriptional regulation.

S/MARs are known to physically associate with nuclear matrix, a three-dimensional filamentous RNA-protein meshwork. Therefore, the most direct and legitimate evidence for any sequence to be SMAR is its presence in nuclear matrix fraction. The matrix–DNA isolation method provides complete nucleic acid complement that is in close physical association with nuclear matrix. Therefore, matrix DNA-PCR has been used to validate identified S/MARs. This method is cost and time efficient over other laboratory methods and allows validation of multiple S/MARs. ChIP-PCR, S/MARBP-S/MAR co-localization studies and electrophoretic mobility shift assays that can also be used for validation purpose, need recombinant purified S/MARBPs and antibodies specific to the S/MARBPs making them time consuming and inefficient with respect to resources required. Similarly, the data used as starting point in the present study is based on ChIP experiments. Therefore, doing similar experiment for validation purpose is redundant.

Retrovirus infection is almost incurable due to stable integration of viral genome in to host genome. This event in viral life cycle makes the pathogen unique leading to lifelong infection escaping the immune system and anti-retroviral therapy regime. The integration of viral genome to host genome is known to occur only at the terminal end of viral DNA, however, for host genome, integration sites can be random. Decoding if this integration has a preferential inclination toward any specific site holds a great advantage in designing effective anti-retroviral therapy. It is believed that host *cis* elements and chromosomal topography plays an invincible role in viral integration and latency. Further, a large number of genes coding for inflammatory cytokines and transcriptional regulator also get disrupted by viral in-

tegration thereby providing favorable condition for its survival. S/MARs are predicted to be most potent sites for retroviral integration due to its structural features such as DNA bending, topoisomerase sites, DNA hypersensitivity, AT richness, kinked DNA etc. (17,74–77). Researchers all over world have contradictory assumption and hypothesis regarding retroviral integration into the host genome. To decipher whether it is a random event or a sequence/topology associated phenomenon, HIV-1 and HTLV-1 IS archived in RID database were mapped on to the identified S/MARs. It was observed that 84% and 72% of the total HIV-1 and HTLV-1 IS, respectively are located within 15 kb distance from their nearest S/MAR. Thus, a major fraction of known IS for these viruses are located within S/MARs and chromatin loop regions in its close proximity. In summary, closer the loci to the S/MARs, higher is the probability of retroviral integration. A number of reports have shown that HIV-1 prefers integration at the intronic regions as well as near highly expressed genes (78). HIV-1 tends to target active gene for its active transcription and viral propagation. A number of active genes with S/MAR regions around their TSS, were also identified in the present study that further highlights the importance of S/MAR sites in retroviral infection. Thus, HIV and HTLV integration is not a random event and S/MARs indeed act as hotspots for their integration into the human genome.

In the light of above observations, our study will facilitate a better understanding of the genome wide location data for S/MARs and help unravel the functional aspects of chromatin. Understanding of S/MARs as HIV integration site will greatly facilitate designing therapeutic arsenal against the latent infection. Targeted genome editing with new genetic engineering tools such as CRISPR/Cas9 can work as potential therapy against this deadly infection. The ability of retroviruses to stably integrate into the host genome has also been harnessed to use them as vehicles for transduction (79). Insertion of these retroviral vectors at wrong loci has been associated with activation of proto-oncogenes. In the view of this fact, a better understanding of the integration sites will help us in designing a suitable retroviral vector for treating and targeting various genetic disorders.

Several algorithms have been developed for *in silico* predictions of S/MAR elements. However, efficacy and predictive potential of these algorithms have so far been restricted due to limited number of sequences available for training the models and lack of features that defines S/MARs effectively. Our attempt to make a genome-wide map of S/MARs in human can complement the development of better performing predictive tool. A collection of experimentally proven S/MARs and nuclear matrix proteins of various organisms including human is available in the form of database (S/MAR transaction database, S/MARt DB) (80). This database however, is published in year 2002, a year before the release of first draft of human genome, which itself has now been extensively revised with respect to sequence information. Therefore, there is a need to revisit this problem and develop a database with updated human S/MAR sequence information. Further such data will be useful to researchers working in the field of computational biology, genomics, functional genomics and virology. Therefore, the web interface, MARome developed by us will facilitate such use of data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Authors thank the Director, Bioinformatics Centre, Savitribai Phule Pune University (SPPU) for providing infrastructural facilities. The Bioinformatics Centre, the Department of Biotechnology at SPPU and the National Centre for Cell Science, Pune are The Department of Biotechnology (DBT), Government of India supported Centres. DBT, DBT-SyMeC and Council of Scientific and Industrial Research, Government of India are duly acknowledged.

FUNDING

Departmental Research Development Programme (DRDP) Grant of Savitribai Phule Pune University, Pune (to A.K., S.M.); DBT Research project grant-BT/PR8749/BID/7/473/2013 (to A.K.); CSIR-Senior Research Fellowship (to S.P.); CSIR-UGC Senior Research Fellowship and DBT-SyMeC-Research Associateship (to A.A.). DBT-SyMeC projects and J.C. Bose Fellowship (to S.C.) Funding for open access charge: Corresponding Author's Institute/Grants.
Conflict of interest statement. None declared.

REFERENCES

1. Heng,H.H.Q. (2004) Chromatin loops are selectively anchored using scaffold/matrix-attachment regions. *J. Cell Sci.*, **117**, 999–1008.
2. Capco,D.G., Wan,K.M., Penman,S., Weber,K., Franke,W.W. and Fyne,C.-T. (1982) The nuclear matrix: three-dimensional architecture and protein composition. *Cell*, **29**, 847–858.
3. Razin,S. V, Gromova,I.I. and Iarovaia,O. V (1995) Specificity and functional significance of DNA interaction with the nuclear matrix: new approaches to clarify the old questions. *Int. Rev. Cytol.*, **162B**, 405–448.
4. Stein,G.S., Zaidi,S.K., Braastad,C.D., Montecino,M., van Wijnen,A.J., Choi,J.-Y., Stein,J.L., Lian,J.B. and Javed,A. (2003) Functional architecture of the nucleus: organizing the regulatory machinery for gene expression, replication and repair. *Trends Cell Biol.*, **13**, 584–592.
5. Breyne,P., van Montagu,M., Depicker,N. and 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. Laemmli,U.K., Käs,E., Poljak,L. and Adachi,Y. (1992) Scaffold-associated regions: cis-acting determinants of chromatin structural loops and functional domains. *Curr. Opin. Genet. Dev.*, **2**, 275–285.
7. Tikhonov,A.P., Bennetzen,J.L. and Avramova,Z. V (2000) Structural domains and matrix attachment regions along colinear chromosomal segments of maize and sorghum. *Plant Cell*, **12**, 249–264.
8. 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.
9. Van Drunen,C.M., Sewalt,R.G.A.B., Oosterling,R.W., Weisbeek,P.J., Smeekens,S.C.M. and Van Driel,R. (1999) A bipartite sequence element associated with matrix/ scaffold attachment regions. *Nucleic Acids Res.*, **27**, 2924–2930.
10. Croft,J.A., Bridger,J.M., Boyle,S., Perry,P., Teague,P. and Bickmore,W.A. (1999) Differences in the localization and morphology of chromosomes in the human nucleus. *J. Cell Biol.*, **145**, 1119–1131.

11. Allen,G.C., Spiker,S. and Thompson,W.F. (2000) Use of matrix attachment regions (MARs) to minimize transgene silencing. *Plant Mol. Biol.*, **43**, 361–376.
12. Zhao,C.-P., Guo,X., Chen,S.-J., Li,C.-Z., Yang,Y., Zhang,J.-H., Chen,S.-N., Jia,Y.-L. and Wang,T.-Y. (2017) Matrix attachment region combinations increase transgene expression in transfected Chinese hamster ovary cells. *Sci. Rep.*, **7**, 42805.
13. Vain,P., Worland,B., Kohli,A., Snape,J.W., Christou,P., Allen,G.C. and Thompson,W.F. (1999) Matrix attachment regions increase transgene expression levels and stability in transgenic rice plants and their progeny. *Plant J.*, **18**, 233–242.
14. Barboro,P., Repaci,E., D'Arrigo,C. and Balbi,C. (2012) The role of nuclear matrix proteins binding to matrix attachment regions (MARs) in prostate cancer cell differentiation. *PLoS One*, **7**, e40617.
15. Gluch,A., Vidakovic,M. and Bode,J. (2008) Scaffold/Matrix Attachment Regions (S/MARs): Relevance for Disease and Therapy. Klussmann,E. and Scott,J. (eds). *Protein-protein interactions as new drug targets. Handbook of experimental Pharmacology*. Springer, Berlin, Heidelberg, pp. 67–103.
16. Petrov,A., Pirozhkova,I., Carnac,G., Laoudj,D., Lipinski,M. and Vassetzky,Y.S. (2006) Chromatin loop domain organization within the 4q35 locus in facioscapulohumeral dystrophy patients versus normal human myoblasts. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 6982–6987.
17. Johnson,C.N. and Levy,L.S. (2005) Matrix attachment regions as targets for retroviral integration. *Virol. J.*, **2**, 68.
18. Kim,S.W., Yoon,S.-J., Chuong,E., Oyolu,C., Wills,A.E., Gupta,R. and Baker,J. (2011) Chromatin and transcriptional signatures for Nodal signaling during endoderm formation in hESCs. *Dev. Biol.*, **357**, 492–504.
19. Do,P.M., Varanasi,L., Fan,S., Li,C., Kubacka,I., Newman,V., Chauhan,K., Daniels,S.R., Boccetta,M., Garrett,M.R. *et al.* (2012) Mutant p53 cooperates with ETS2 to promote etoposide resistance. *Genes Dev.*, **26**, 830–845.
20. Walsh,C.A., Bolger,J.C., Byrne,C., Cocchiglia,S., Hao,Y., Fagan,A., Qin,L., Cahalin,A., McCartan,D., McIlroy,M. *et al.* (2014) Global gene repression by the steroid receptor coactivator SRC-1 promotes oncogenesis. *Cancer Res.*, **74**, 2533–2544.
21. Mathai,J., Mittal,S.P.K., Alam,A., Ranade,P., Mogare,D., Patel,S., Saxena,S., Ghorai,S., Kulkarni,A.P. and Chattopadhyay,S. (2016) SMAR1 binds to T(C/G) repeat and inhibits tumor progression by regulating miR-371-373 cluster. *Sci. Rep.*, **6**, 33779.
22. ENCODE Project Consortium, T.E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
23. Winick-Ng,W., Caetano,F.A., Winick-Ng,J., Morey,T.M., Heit,B. and Rylett,R.J. (2016) 82-kDa choline acetyltransferase and SATB1 localize to β-amyloid induced matrix attachment regions. *Sci. Rep.*, **6**, 23914.
24. Patel,R.K. and Jain,M. (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS One*, **7**, e30619.
25. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
26. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
27. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.
28. Yu,G., Wang,L.-G. and He,Q.-Y. (2015) ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**, 2382–2383.
29. Girod,P.-A., Nguyen,D.-Q., Calabrese,D., Puttini,S., Grandjean,M., Martinet,D., Regamey,A., Saugy,D., Beckmann,J.S., Bucher,P. *et al.* (2007) Genome-wide prediction of matrix attachment regions that increase gene expression in mammalian cells. *Nat. Methods*, **4**, 747–753.
30. Keaton,M.A., Taylor,C.M., Layer,R.M. and Dutta,A. (2011) Nuclear scaffold attachment sites within ENCODE regions associate with actively transcribed genes. *PLoS One*, **6**, e17912.
31. Huber,L.J. and Chodosh,L.A. (2005) Dynamics of DNA repair suggested by the subcellular localization of Brca1 and Brca2 proteins. *J. Cell Biochem.*, **96**, 47–55.
32. Herrscher,R.F., Kaplan,M.H., Lelsz,D.L., Das,C., Scheuermann,R. and Tucker,P.W. (1995) The immunoglobulin heavy-chain matrix-associating regions are bound by Bright: A B cell-specific trans-activator that describes a new DNA-binding protein family. *Genes Dev.*, **9**, 3067–3082.
33. Chattopadhyay,S., Kaul,R., Charest,A., Housman,D. and Chen,J. (2000) SMAR1, a novel, alternatively spliced gene product, binds the scaffold/matrix-associated region at the T cell receptor β locus. *Genomics*, **68**, 93–96.
34. van Wijnen,A.J., Bidwell,J.P., Fey,E.G., Penman,S., Lian,J.B., Stein,J.L. and Stein,G.S. (1993) Nuclear matrix association of multiple sequence-specific DNA binding activities related to SP-1, ATF, CCAAT, C/EBP, OCT-1, and AP-1. *Biochemistry*, **32**, 8397–8402.
35. Maksimenko,O., Gasanov,N.B. and Georgiev,P. (2015) Regulatory elements in vectors for efficient generation of cell lines producing target proteins. *Acta Nat.*, **7**, 15–26.
36. Chattopadhyay,S., Whitehurst,C.E. and Chen,J. (1998) A nuclear matrix attachment region upstream of the T cell receptor β gene enhancer binds Cux/CDP and SATB1 and modulates enhancer-dependent reporter gene expression but not endogenous gene expression. *J. Biol. Chem.*, **273**, 29838–29846.
37. Yusufzai,T.M. and Felsenfeld,G. (2004) The 5′-HS4 chicken -globin insulator is a CTCF-dependent nuclear matrix-associated element. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 8620–8624.
38. Dunn,K.L., Zhao,H. and Davie,J.R. (2003) The insulator binding protein CTCF associates with the nuclear matrix. *Exp. Cell Res.*, **288**, 218–223.
39. Galande,S. and Kohwi-Shigematsu,T. (1999) Poly(ADP-ribose) polymerase and Ku autoantigen form a complex and synergistically bind to matrix attachment sequences. *J. Biol. Chem.*, **274**, 20521–20528.
40. Torrungruang,K., Alvarez,M., Shah,R., Onyia,J.E., Rhodes,S.J. and Bidwell,J.P. (2002) DNA binding and gene activation properties of the Nmp4 nuclear matrix transcription factors. *J. Biol. Chem.*, **277**, 16153–16159.
41. Will,K., Warnecke,G., Wiesmüller,L. and Deppert,W. (1998) Specific interaction of mutant p53 with regions of matrix attachment region DNA elements (MARs) with a high potential for base-unpairing. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 13681–13686.
42. Göhring,F. and Fackelmayer,F.O. (1997) The scaffold/matrix attachment region binding protein hnRNP-U (SAF-A) is directly bound to chromosomal DNA in vivo: a chemical cross-linking study. *Biochemistry*, **36**, 8276–8283.
43. Mittal,S.P.K., Mathai,J., Kulkarni,A.P., Pal,J.K. and Chattopadhyay,S. (2013) miR-320a regulates erythroid differentiation through MAR binding protein SMAR1. *Int. J. Biochem. Cell Biol.*, **45**, 2519–2529.
44. Sinha,S., Malonia,S.K., Mittal,S.P.K., Mathai,J., Pal,J.K. and Chattopadhyay,S. (2012) Chromatin remodelling protein SMAR1 inhibits p53 dependent transactivation by regulating acetyl transferase p300. *Int. J. Biochem. Cell Biol.*, **44**, 46–52.
45. Sinha,S., Malonia,S.K., Mittal,S.P.K.K., Singh,K., Kadreppa,S., Kamat,R., Mukhopadhyaya,R., Pal,J.K. and Chattopadhyay,S. (2010) Coordinated regulation of p53 apoptotic targets BAX and PUMA by SMAR1 through an identical MAR element. *EMBO J.*, **29**, 830–842.
46. Rampalli,S., Pavithra,L., Bhatt,A., Kundu,T.K. and Chattopadhyay,S. (2005) Tumor suppressor SMAR1 mediates cyclin D1 repression by recruitment of the SIN3 / histone deacetylase 1 complex. *Mol. Cell Biol.*, **25**, 8415–8429.
47. Singh,K., Sinha,S., Malonia,S.K., Bist,P., Tergaonkar,V. and Chattopadhyay,S. (2009) Tumor suppressor SMAR1 represses IκBα expression and inhibits p65 transactivation through matrix attachment regions. *J. Biol. Chem.*, **284**, 1267–1278.
48. Chemmannur,S. V., Badhwar,A.J., Mirlekar,B., Malonia,S.K., Gupta,M., Wadhwa,N., Bopanna,R., Mabalirajan,U., Majumdar,S., Ghosh,B. *et al.* (2015) Nuclear matrix binding protein SMAR1 regulates T-cell differentiation and allergic airway disease. *Mucosal. Immunol.*, **8**, 1201–1211.

49. Song,G., Liu,K., Yang,X., Mu,B., Yang,J., He,L., Hu,X., Li,Q., Zhao,Y., Cai,X. *et al.* (2017) SATB1 plays an oncogenic role in esophageal cancer by up- regulation of FN1 and PDGFRB. *Oncotarget*, **8**, 17771–17784.
50. Travers,A.A. (1995) Reading the minor groove. *Nat. Struct. Biol.*, **2**, 615–618.
51. Steitz,T.A. (1990) Structural studies of protein-nucleic acid interaction: the sources of sequence-specific binding. *Q. Rev. Biophys.*, **23**, 205–280.
52. Zink,D., Fische,A.H., Nickerson,J.A., Lozano,M., Kobayashi,R., Ross,S., Dudley,J., Romeyn,L. and Copeland,N. (2004) Nuclear structure in cancer cells. *Nat. Rev. Cancer*, **4**, 677–687.
53. Han,W., Lindsay,S.M., Dlakic,M. and Harrington,R.E. (1997) Kinked DNA. *Nature*, **386**, 563–563.
54. Singh,R.K., Sasikala,W.D. and Mukherjee,A. (2015) Molecular origin of DNA kinking by transcription factors. *J. Phys. Chem. B*, **119**, 11590–11596.
55. Chen,C.-Y., Ko,T.-P., Lin,T.-W., Chou,C.-C., Chen,C.-J. and Wang,A.H.-J. (2005) Probing the DNA kink structure induced by the hyperthermophilic chromosomal protein Sac7d. *Nucleic Acids Res.*, **33**, 430–438.
56. Benham,C., Kohwi-Shigematsu,T. and Bode,J. (1997) Stress-induced duplex DNA destabilization in scaffold/matrix attachment regions. *J. Mol. Biol.*, **274**, 181–196.
57. Boulikas,T. (1993) Nature of DNA sequences at the attachment regions of genes to the nuclear matrix. *J. Cell. Biochem.*, **52**, 14–22.
58. Shaposhnikov,S.A., Akopov,S.B., Chernov,I.P., Thomsen,P.D., Joergensen,C., Collins,A.R., Frengen,E. and Nikolaev,L.G. (2007) A map of nuclear matrix attachment regions within the breast cancer loss-of-heterozygosity region on human chromosome 16q22.1. *Genomics*, **89**, 354–361.
59. Frisch,M., Frech,K., Klingenhoff,A., Cartharius,K., Liebich,I. and Werner,T. (2002) In silico prediction of scaffold/matrix attachment regions in large genomic sequences. *Genome Res.*, **12**, 349–354.
60. Razin,S. V (1999) Chromosomal DNA loops may constitute basic units of the eukaryotic genome organization and evolution. *Crit. Rev. Eukaryot. Gene Expr.*, **9**, 279–283.
61. Jackson,D.A., Dickinson,P. and Cook,P.R. (1990) The size of chromatin loops in HeLa cells. *EMBO J.*, **9**, 567–571.
62. Marilley,M. and Gassend-Bonnet,G. (1989) Supercoiled loop organization of genomic DNA: a close relationship between loop domains, expression units, and replicon organization in rDNA from Xenopus laevis. *Exp. Cell Res.*, **180**, 475–489.
63. Buongiorno-Nardelli,M., Micheli,G., Carri,M.T. and Marilley,M. (1982) A relationship between replicon size and supercoiled loop domains in the eukaryotic genome. *Nature*, **298**, 100–102.
64. Heng,H.H., Krawetz,S.A., Lu,W., Bremer,S., Liu,G. and Ye,C.J. (2001) Re-defining the chromatin loop domain. *Cytogenet. Cell Genet.*, **93**, 155–161.
65. Vassetzky,Y.S., Hair,A. and Razin,S. V (2000) Rearrangement of chromatin domains in cancer and development. *J. Cell Biochem. Suppl.*, **35**, 54–60.
66. Boyle,S., Gilchrist,S., Bridger,J.M., Mahy,N.L., Ellis,J.A. and Bickmore,W.A. (2001) The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.*, **10**, 211–219.
67. Bushey,A.M., Dorman,E.R. and Corces,V.G. (2008) Chromatin insulators: regulatory mechanisms and epigenetic inheritance. *Mol. Cell*, **32**, 1–9.
68. Namciu,S.J. and Fournier,R.E.K. (2004) Human matrix attachment regions are necessary for the establishment but not the maintenance of transgene insulation in Drosophila melanogaster. *Mol. Cell Biol.*, **24**, 10236–10245.
69. Brouwer,C., Bruce,W., Maddock,S., Avramova,Z. and Bowen,B. (2002) Suppression of transgene silencing by matrix attachment regions in maize: a dual role for the maize 5′ ADH1 matrix attachment region. *Plant Cell*, **14**, 2251–2264.
70. Pascuzzi,P.E., Flores-Vergara,M.A., Lee,T.-J., Sosinski,B., Vaughn,M.W., Hanley-Bowdoin,L., Thompson,W.F. and Allen,G.C. (2014) In vivo mapping of arabidopsis scaffold/matrix attachment regions reveals link to nucleosome-disfavoring poly(dA:dT) tracts. *Plant Cell*, **26**, 102–120.
71. Chattopadhyay,S. and Pavithra,L. (2007) MARs and MARBPs. *Chromatin Dis.*, **41**, 213–230.
72. Liebich,I., Bode,J., Reuter,I. and Wingender,E. (2002) Evaluation of sequence motifs found in scaffold/matrix-attached regions (S/MARs). *Nucleic Acids Res.*, **30**, 3433–3442.
73. Pathak,R.U., Srinivasan,A. and Mishra,R.K. (2014) Genome-wide mapping of matrix attachment regions in Drosophila melanogaster. *BMC Genomics*, **15**, 1022.
74. Mielke,C., Maass,K., Tümmler,M. and Bode,J. (1996) Anatomy of highly expressing chromosomal sites targeted by retroviral vectors. *Biochemistry*, **35**, 2239–2252.
75. D'ugo,E., Bruni,R., Argentini,C., Giuseppetti,R. and Rapicetta,M. (1998) Identification of scaffold/matrix attachment region in recurrent site of woodchuck hepatitis virus integration. *DNA Cell Biol.*, **17**, 519–527.
76. Shera,K.A., Shera,C.A. and James,K. (2001) Small tumor virus genomes are integrated near nuclear matrix attachment regions in transformed cells. *J. Virol.*, **75**, 12339–12346.
77. Kulkarni,A., Pavithra,L., Rampalli,S., Mogare,D., Babu,K., Shiekh,G., Ghosh,S. and Chattopadhyay,S. (2004) HIV-1 integration sites are flanked by potential MARs that alone can act as promoters. *Biochem. Biophys. Res. Commun.*, **322**, 672–677.
78. Craigie,R. and Bushman,F.D. (2012) HIV DNA integration. *Cold Spring Harb. Perspect. Med.*, **2**, a006890.
79. Barquinero,J., Eixarch,H. and Pérez-Melgosa,M. (2004) Retroviral vectors: new applications for an old tool. *Gene Ther.*, **11**, S3–S9.
80. Liebich,I., Bode,J., Frisch,M. and Wingender,E. (2002) S/MARt DB: a database on scaffold/matrix attached regions. *Nucleic Acids Res.*, **30**, 372–374.