Multifaceted roles of 5'-regulatory region of the cancer associated gene *B4GALT1* and its comparison with the gene family

MOHAMMED A. IBRAHIM AL-OBAIDE[1], HYTHAM ALOBYDI[2], ABDELSALAM G. ABDELSALAM[3], RUIWEN ZHANG[4] and KALKUNTE S. SRIVENUGOPAL[1]

[1]Department of Biomedical Sciences, School of Pharmacy, Texas Tech University Health Sciences Center, Amarillo, TX 79106; [2]Biomedica, LLC, Sterling Heights, MI, USA; [3]Department of Mathematics, Statistics and Physics, College of Arts and Sciences, Qatar University, Doha, Qatar; [4]Department of Pharmaceutical Sciences, School of Pharmacy, Texas Tech University Health Sciences Center, Amarillo, TX 79106, USA

Received June 25, 2015; Accepted August 5, 2015

DOI: 10.3892/ijo.2015.3136

Abstract. β1,4-Galactosylransferases are a family of enzymes encoded by seven *B4GALT* genes and are involved in the development of anticancer drug resistance and metastasis. Among these genes, the *B4GALT1* shows significant variations in the transcript origination sites in different cell types/tissues and encodes an interesting dually partitioning β-1, 4-galactosyltransferase protein. We identified at 5'-end of *B4GALT1* a 1.454 kb sequence forming a transcription regulatory region, referred to by us as the TR1-PE1, had all characteristics of a bidirectional promoter directing the transcription of *B4GALT1* in a divergent manner along with its long non-coding RNA (lncRNA) antisense counterpart *B4GALT1-AS1*. The TR1-PE1 showed unique dinucleotide base-stacking energy values specific to transcription factor binding sites (TFBSs), INR and BRE, and harbored CpG Island (CGI) that showed GC skew with potential for R-loop formation at the transcription starting sites (TSSs). The 5'-regulatory axis of *B4GALT1* also included five more novel TFBSs for CTCF, GLI1, TCF7L2, GATA3 and SOX5, in addition to unique $(TG)_{18}$ repeats in conjunction with 22 nucleotide TG-associated sequence (TGAS). The five lncRNA *B4GALT1-AS1* transcripts showed significant complementarity with *B4GALT1* mRNA. In contrast, the rest of *B4GALT* genes showed fewer lncRNAs, and all lacked the $(TG)_{18}$ and TGAS. Our results are strongly supported by the FANTOM5 study which showed tissue-specific variations in transcript origination sites for this gene. We suggest that the unique expression patterns for the *B4GALT1* in normal and malignant tissues are controlled by a differential usage of 5'-*B4GALT1* regulatory units along with a post-transcriptional regulation by the antisense RNA, which in turn govern the cell-matrix interactions, neoplastic progression, anticancer drug sensitivity, and could be utilized in personalized therapy.

Correspondence to: Dr Kalkunte Srivenugopal, Department of Biomedical Sciences, School of Pharmacy, Texas Tech University Health Sciences Center, 1406 S. Coulter Drive, Amarillo, TX 79106, USA
E-mail: kalkunte.srivenugopal@ttuhsc.edu

Key words: bidirectional promoter, drug resistance, lncRNA, metastasis, personalized therapy, transcription factor binding sites, TG repeats, TG-associated sequences

Introduction

The enzymatic glycosylation of proteins and lipids is a fundamental process in biology. β1, 4-Galactosylransferases are a family of enzymes that catalyze the formation of β4-N-acetyllactosamine linkages in extracellular matrices by transferring the UDP-bound galactose to terminal N-acetylglucosamines in carbohydrate chains (1,2). At least 7 members of this family (Gal T1 to Gal T7) have been characterized, reflecting the redundancy and a fine regulation of the glycoforms synthesized by these enzymes in specific cell-cell and cell-matrix adhesions. Of these, *B4GALT1* is the most studied and its cDNA was first cloned in 1986 (3); since then, it has served as a prototype for characterization of several galatosyltrasferases across the mammalian species (4). The *B4GALT1* gene generates two types of protein isoforms termed the long (399 amino acids) and short (386 amino acids). The two isoforms are type II membrane-bound glycoproteins and reside in Golgi apparatus and a portion of the longer protein functions on the cell surface as a recognition molecule by binding with appropriate glycoside substrates (5). Consequently, the *B4GALT1* plays important roles in numerous physiological and pathological processes such as inflammation, sperm-egg interaction, embryogenesis, and morphogenesis, development of the central nervous system, cell migration, cancer progression and metastasis (6-8).

Although considered mainly as a housekeeping gene, a large number of studies have shown a differential expression of *B4GALT1* in mammary gland, brain and other normal tissues such as the cartilage (7,9,10). On the other hand, this gene has been highly implicated in oncogenesis and tumor progression (6,11-13). Leukemia, melanoma and cancers of the breast, lung, ovary, liver, prostate and their metastatic counterparts have been shown to possess elevated levels of expression and cell-surface *B4GALT1* (11,12,14). Thus, its expression levels may alter the distribution and profile of cancer antigens. Also,

in a nude mouse model, the number of peritoneal dissemination foci of the antisense *B4GALT1*-transfected ovarian tumor cells was smaller than that of the control cells, suggesting the involvement of this enzyme in the invasive and metastatic ovarian cancer (15). Furthermore, *B4GALT1* has been strongly linked with multidrug resistance (8,16), tumor sensitivity to cisplatin (17,18) and promotion of cell death signaling pathways (19). Additionally, the gene undergoes promoter methylation in several cancer types and consequent silencing of the gene (13,20,21).

Little information is available on the regulation of *B4GALT1* transcription and the involvement of regulatory elements in the process. A single promoter is believed to mediate the varied and tissue-specific expression of the *B4GALT1* (6,13,20,21). However, a recent landmark analysis of promoters called the FANTOM5 (functional annotation of the mammalian genome 5) chose the *B4GALT1* as a representative gene and showed significant variations in the transcript origination sites in different cell types and patient specimens (22). They reported a *B4GALT1* promoter at 5' end and observed 266 bp CGI associated with *B4GALT1* transcription. Encouraged by these observations, we undertook a systematic study to characterize the alternative promoters of *B4GALT1*. In the present study, we report the characteristics of the *B4GALT1* alternative promoters and molecular features of the 5' regulatory region of the gene, which revealed multiple regulatory sequences, composed of 1.454 kb and its association with the long non-coding RNA gene, *B4GALT1-AS1*. Given the redundant and compensatory roles of β1, 4-galactosylransferases in glycan synthesis and extracellular matrix interactions (1), we also screened and compared the regulatory elements in other genes of the *B4GALT* family and compared them with the prototype *B4GALT1*.

Materials and methods

Genomic databases. The genomic criteria and alternative promoters of the divergent loci, *B4GALT1* and *B4GALT1-AS1* and the other *B4GALT* genes were searched during January-December 2014 in seven genomics databases, NCBI-GenBank, Transcriptional Regulatory Element Database (TRED), Mammalian Promoter/Enhancer Database (PEDB), Eukaryotic promoter database (EPD), Mammalian Promoter Database (MPromDb), Ensembl and UCSC Genome Browser. Several tools in the databases were used to retrieve the sequences, identify the strand (forward or reverse), flip the strand and in search for a specific sequence. NCBI-dbSNP was used to search for SNPs in the *B4GALT1* regulatory region.

Verification of map locations. The precise genomic map locations of the identified sequences of the *B4GALT1* alternative promoters were verified and updated to hg38 by using the BLAT tool, Gene Sorter and Table Browser tools of UCSC Genome Bioinformatics database.

Search for regulatory elements in the alternative promoters. The identified sequences of the *B4GALT1* alternative promoters were analyzed for TFBSs, namely, TATA-8 (TATAWAWR) and TATA-532 (HWHWWWWR, excluding: HTYTTTWR, CAYTTTWR, MAMAAAAR and CTYAAAAR), INR (YYANWYY), CCAAT and its inverted sequence TAACC, BRE (SSRCGCC), and DPE (RGWCGTG) binding sites (23-26). The TG tandem repeats were identified by Blat and Blast tools. Identification of CGIs in alternative promoters was searched at 100-bp window (N=100) moving across the sequence at 1-bp intervals, parameter sets used to search for CGIs in the alternative promoters: Observed/Expected CpG ≥0.6 and %G+C >55% (27,28). The ratio Observed/Expected (O/E) CpG was calculated according to the equation reported by Gardiner-Garden and Frommer (27).

Structural features of TR1-PE1 regulatory sequence. The dinucleotide base-stacking energy values were derived from values of dinucleotide base-stacking energy provided by Ornstein *et al* (29). According to the scale, which is in kcal/mol, the range of values from -3.82 kcal/mol (unstack easily) to -14.59 kcal/mol (difficult to unstack), thus, the obtained values show the relative dissociation stability of the double helix structure. Whereas, GC skew is calculated as [(G-C)/(G+C)], where C and G denote the numbers of cytosine and guanine (30,31). GC skew is useful for predicting the R loop formation and the origin and terminus of replication. Programs written in R were used to analyze the sequences and to plot the data of dinucleotide base-stacking energy values (Kcal/mol/dimer) and GC skew along the length of the TR1-PE1 sequence.

Blast tree map tool. NCBI standard nucleotide BLAST tool was used to search by pairwise alignments for similar sequences to the regulatory sequences identified in the study. The distance tree of the obtained pairwise comparisons was produced to show evolutionary relatedness of regulatory sequences among species.

Statistical analysis. The regulatory sequences were analyzed and achieved using Excel software and programs written in R. The independence of each promoter element was examined using Fisher's exact probability test.

Results

Promoters and antisense lncRNA loci of B4GALT family members. Search in the databases showed that the expression of the seven genes of the *B4GALT* family is controlled by alternative promoters. The numbers of identified alternative promoters for each gene varied from nine to sixteen and were located on six chromosomes (Table I). In addition, antisense lncRNA divergent loci were identified for four *B4GALT* genes, namely; *B4GALT1*, *B4GALT4*, *B4GALT6* and *B4GALT7* (Table II). Data search showed five species of antisense lncRNA transcripts for *B4GALT1*. In contrast, the three other *B4GALT* genes, *B4GALT4*, *B4GALT6* and *B4GALT7* were each associated with a single lncRNA transcript. These observations encouraged us to perform a comprehensive search of the regulatory regions of *B4GALT1* locus and characterize the functional elements and structural features controlling the transcription of this gene.

Genomic context of B4GALT1 alternative promoters. The human *B4GALT1* (uc003zsg.2) is located on the negative strand of the short arm of chromosome 9: 33110641-33167358

Table I. Map locations and number of promoters of seven *B4GALT* genes.

B4GALT genes	GenBank ID	Map location	Promoters
B4GALT1	2683	chr9: 33,104,08233,167,356 Reverse strand	13
B4GALT2	8704	chr1: 43,978,94343,991,170 Forward strand	16
B4GALT3	8703	chr1: 161,171,310161,177,968 Reverse strand	11
B4GALT4	8702	chr3: 119,211,732119,241,103 Reverse strand	11
B4GALT5	9334	chr20: 49,632,94549,713,878 Reverse strand	11
B4GALT6	9331	chr18: 31,622,24731,685,836 Reverse strand	9
B4GALT7	11285	chr5: 177,600,100177,610,347 Forward strand	9

Table II. Genomic contexts of *B4GALT1-AS1*, *B4GALT4-AS1*, *B4GALT6-AS1* and *B4GALT7-AS1* transcripts.

B4GALT genes	Location	*B4GALT* antisense	Location	Locus space (bp)	Transcript ID	Length (bp)
B4GALT1	Chr9: 33110641-33167358 Reverse strand	*B4GALT1-AS1*	Chr9: 33166946-33179981 Forward strand	13036	NR_108110.1	808
B4GALT1	Chr9: 33110641-33167358 Reverse strand	*B4GALT1-AS1*	Chr9: 33166946-33179981 Forward strand	13036	NR_108109.1	932
B4GALT1	Chr9: 33110641-33167358 Reverse strand	*B4GALT1-AS1*	Chr9: 33167857-33179981 Forward strand	12125	NR_108108.1	1185
B4GALT1	Chr9: 33,104,082-33,167,356 Reverse strand	*B4GALT1-AS*1	Chr9: 33166975-33179710 Forward strand	12736	ENST00000426270	508
B4GALT1	Chr9: 33104082-33167356 Reverse strand	*B4GALT1-AS1*	Chr9: 33166975-33179983 Forward strand	13009	ENST00000442432	843
B4GALT4	Chr3: 119211732-119241103 Reverse strand	*B4GALT4-AS*	Chr3: 119226486-119290666 Forward strand	64181	ENST00000470790	544
B4GALT6	Chr18: 31622247-31685836 Reverse strand	*B4GALT6-AS*	Chr18: 31685655-31686823 Forward strand	1169	ENSG00000259985	1169
B4GALT7	Chr5: 177600100-177610347 Forward strand	*B4GALT7-AS*	Chr5: 177611253-177619754 Forward strand	8502	ENST00000499900	1751

The data were mined from NCBI-Gene, NCBI-Nucleotide, UCSC and Ensembl databases.

at 9p13 and was found to be divergently paired head to head with *B4GALT1-AS1* (uc033cop.1). *B4GALT1* has been assumed to harbor a single promoter of 500 bp (6,13,20). In this context, we were, however, surprised to find a large number of alternative promoters for *B4GALT1*. Our rigorous analysis revealed a total of thirteen alternative promoters for *B4GALT1*, although many of them were overlapping with TR1, HP2 and HP3 sequences (Table III and Fig. 1). The map location of HP1 (chr9: 33103518-33104577) indicated it is located outside of the *B4GALT1* locus, whereas three other promoters, HP2 (chr9: 33124856-33126364), HP3 (chr9: 33156756-33158354), and ES1 (chr9: 33128098-33128697) were found located within the space of *B4GALT1* locus. HP4 (chr9: 33125681-33126181) and HP5 (chr9: 33157101-33157692) sequences were located within HP2 and HP3 sequences, respectively. Our analysis also showed that one third of TR2 sequence (chr9: 3110342-33111341) was located outside the 3' end of the gene locus. Whereas, CP1, EP1, ES2 and ES3 were located within the vicinity of TR1 sequence (Table III) and this was verified by sequence alignment analyses (data not shown). Furthermore, approximately half (458 bp) of TR1 sequence (chr9: 33166901-33167900) was located inside *B4GALT1* locus at 5' side and 42 bp overlapped the sequence at 3' side of *B4GALT1-AS1* reverse complement (chr9: 33167859-33179983), the remaining of TR1 sequence, 500 bp, is the space between the two divergent genes *B4GALT1* and *B4GALT1-AS1* (Fig. 2). Furthermore, we observed five nucleotides at the 3' side of PE1 (chr9: 33167354-33168354) overlapped with the *B4GALT1* sequence at the 5' end and more than half of PE1 sequence overlapped TR1 sequence outside the 5' side of *B4GALT1* (Fig. 2). The outcome of map location analysis of the six overlapping alternative promoter sequences (TR1, CP1, EP1, ES2, ES3 and PE1) at 5'-end of *B4GALT1* revealed the presence of a complex regulatory unit, which we designated TR1-PE1 and is located at chr9: 33166901-33168354.

Table III. Map locations and CGIs of thirteen *B4GALT1* alternative promoters.

Alternative promoter source	Symbol	Map locations	Span (bp)	Obs/Exp CpG	%G+C	CGI
TRED-42895	TR1	chr9: 33166901-33167900	1000	0.88	73.3	+
TRED-113955	TR2	chr9: 33110342-33111341	1000	0.22	45.3	-
MPromDB- HG_ACW: 80083	HP1	chr9: 33103518-33104577	1060	0.13	49.57	-
MPromDB-HG_ACW: 80090	HP2	chr9: 33124856-33126364	1509	0.19	49.86	-
MPromDB-HG_KWN: 62900	HP3	chr9: 33156756-33158354	1599	0.15	39.59	-
MPromDB-HG_ACW:80090	HP4	chr9: 33125681-33126181	501	0.13	50.8	-
MPromDB-HG_KWN:62900	HP5	chr9: 33157101-33157692	592	0.15	37.4	-
Choi *et al* (6)	CP1	chr9: 33167181-33167680	500	0.89	76.6	+
EPD	EP1	chr9: 33167235-33167834	600	0.87	74.2	+
PEPED	PE1	chr9: 33167354-33168354	1001	0.73	69.2	+
Ensembl	ES1	chr9: 33128098-33128697	600	0.11	49.3	-
Ensembl	ES2	chr9: 33167085-33167684	600	0.91	76	+
Ensembl	ES3	chr9: 33167255-33167854	600	0.91	74.2	+
(This study)	TR1-PE1	chr9: 33166901-33168354	1454	0.78	67.9	+

+ and - indicate presence and absence of CGI, respectively, in the alternative promoters.

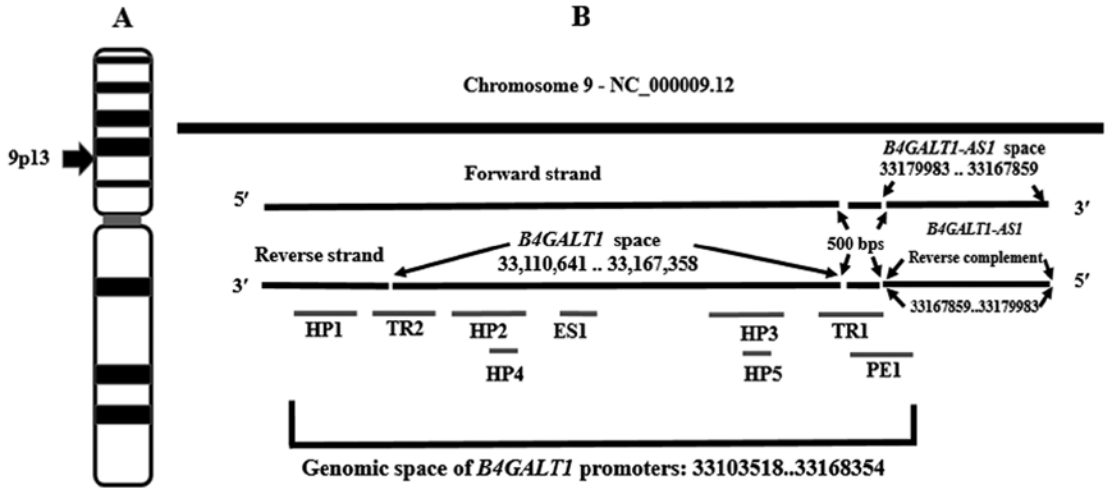

Figure 1. Map locations of alternative promoters of divergently-paired head to head genes, *B4GALT1* (uc003zsg.2) and *B4GALT1-AS1* (uc033cop.1). (A) Chromosomal locations of the two paired genes at 9p13 are shown. (B) Intergenic and intragenic locations of alternative promoters along *B4GALT1* and *B4GALT1-AS1* space at chromosome 9-NC_000009.12 is displayed.

Molecular and structural characteristics of the TR1-PE1. Alternative promoters of *B4GALT1* could be divided into two groups according to %G+C and Observed/Expected CpG values, which are indicators of CGIs (Table III). Accordingly, alternative promoters with CGIs were: TR1, CP1, EP1, ES2, ES3 and PE1, which form the TR1-PE1 regulatory complex; whereas the rest of alternative promoters lack CGIs. TR1-PE1 sequence was found rich with BRE and INR sequences. The clustering pattern of INR and BRE was unique, seven out of nine INR sequences were clustered in the 5'-side of TR1-PE1 region and the four identified BRE sequences were located in the 3'-TR1-PE1 side (Fig. 3). In contrast, TR1-PE1 sequence does not contain any of TATA-8 and GC-box sequences, but it does contain two TAACC in the upstream region and one TATA-532, TTCTTAAA, along with two INR sequences downstream TR1-PE1. In addition, the TR1-PE1 sequence harbored other regulatory sequences in the BRE region, such as three estrogen response elements (ERE) and a muscle actin promoter factor (MAPF) binding element, which were found within the CP1 alternative promoter reported by Choi *et al* (6).

Our next effort was to identify the TSSs within the regulatory complex. As expected the regulatory unit TR1-PE1 with an overlapping six alternative promoter sequences contained six TSSs along a sequence of 156 bp at chr9: 33167351-33167506 (Fig. 3). This region is rich in the BRE sequences and is comparable to the reported 266 bp CGI (chr9: 33167138-33167403) associated with *B4GALT1* transcription (22).

To find out the possible correlation between structural characteristics and distribution of TFBSs along TR1-PE1

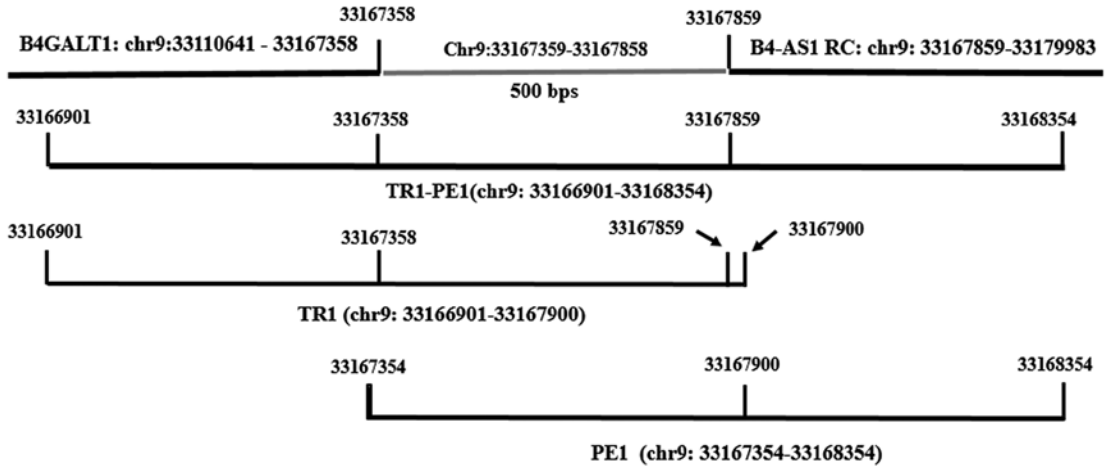

Figure 2. Map location of TR1-PE1 regulatory complex compared with locations of *B4GALT1* and *B4GALT1-AS1* reverse complement (B4-AS1 RC), TR1 and PE1.

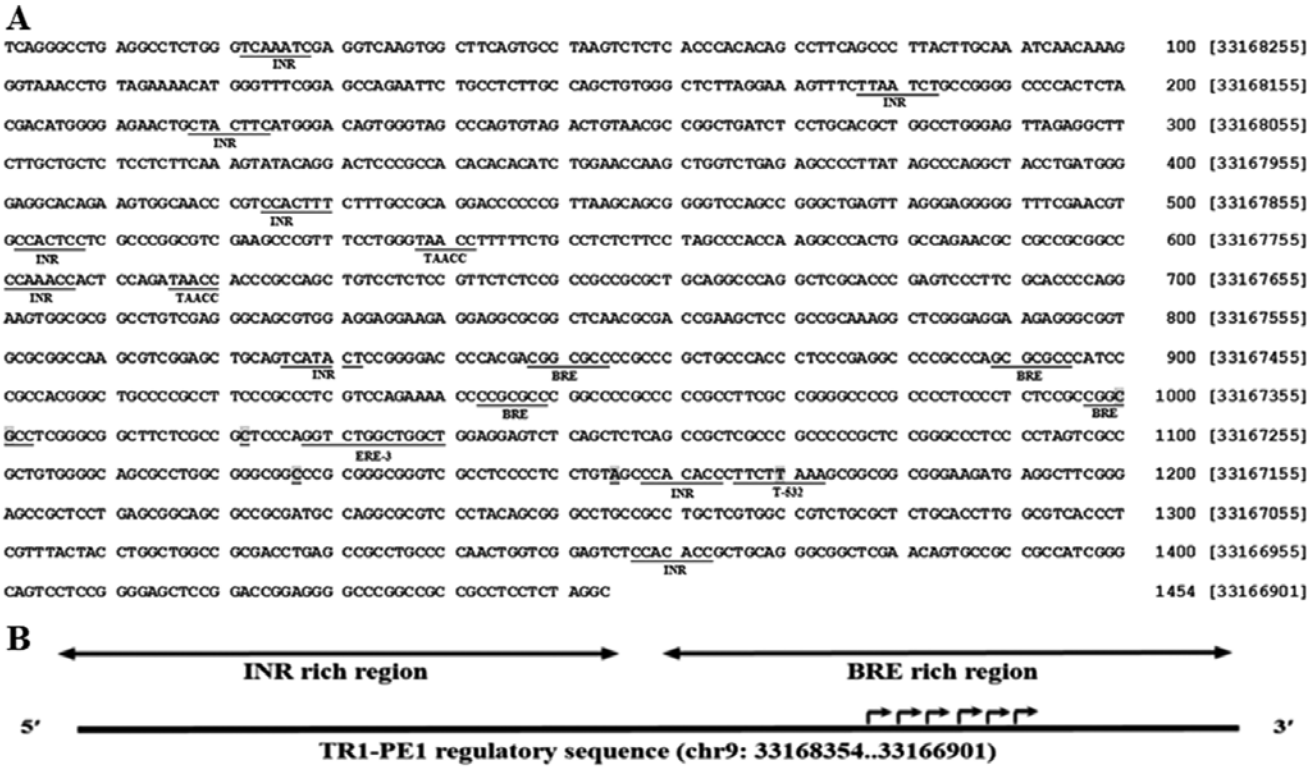

Figure 3. Distribution of TFBS and TSS along the TR1-PE1 regulatory complex. (A) The TR1-PE1 sequence, the underlined sequences and highlighted bases indicate TFBS and TSS, respectively. (B) Schematic presentation of BRE, INR rich regions and the six TSS (bent arrows) along TR1-PE1 regulatory sequence.

sequence, we investigated the dinucleotide base-stacking energy values of this sequence. The INR sites were found mainly along sequences easily unstack, whereas BRE sequences were observed at sequences comparatively difficult to unstack (Fig. 4). Three regions termed A, B and C with calculated stacking energy values of -7.661, -8.448 and -8.524 kcal/mol harbored INR, INR and TATA-532, respectively. Whereas, BRE sequences were clustered within region D composed of 156 bp at 33167351-33167506, which showed -9 to -9.75 kcal/mol. Also, we observed GC skew in the TR1-PE1 CGI region at TSS vicinity (Fig. 5). GC skew is a result of strand asymmetry down-stream TSSs and therefore it is an indication of possible formation of R-loops that are correlated with un-methylated status of CGI (see Discussion).

Experimental evidence supporting the validity of our results on the presence of multiple TSSs within the TR1-PE1 sequence has been obtained as a part of the FANTOM5 consortium (22). These researchers used Cap analysis of gene expression (CAGE) to map the promoters and sets of transcripts for *B4GALT1* in numerous human tissues and cell lines. The CAGE patterns obtained for the 266 bases *B4GALT1* transcription initiation region located at chr9:

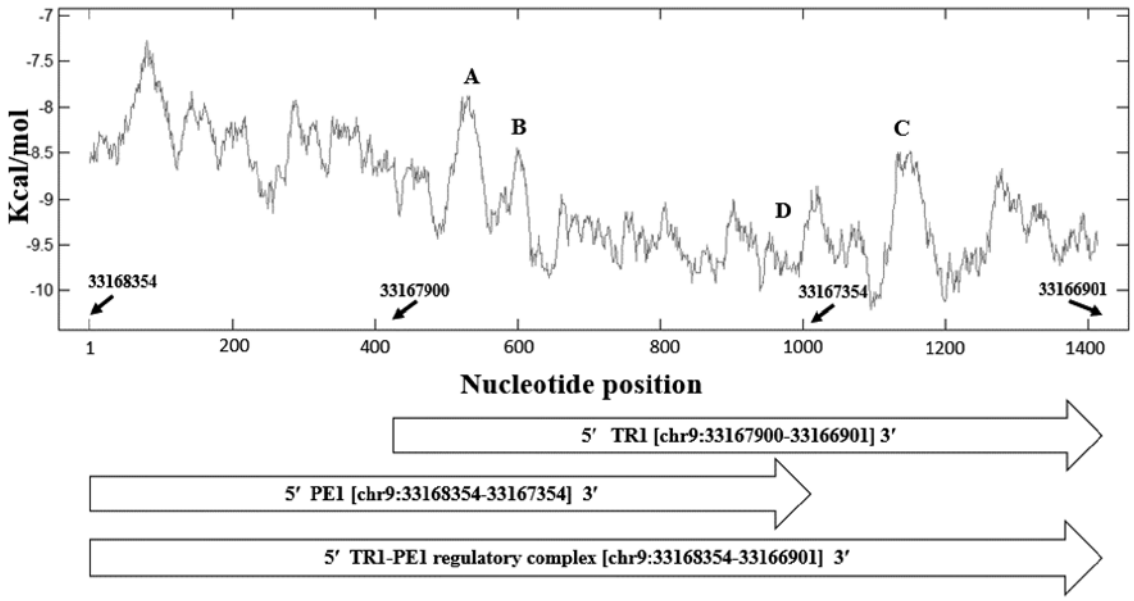

Figure 4. Plot of the dinucleotide base-stacking energy values (y-axis) along the TR1-PE1 regulatory sequence. A, B, C and D in the inset represent the locations of INR, INR, TATA-532 and BRE, respectively.

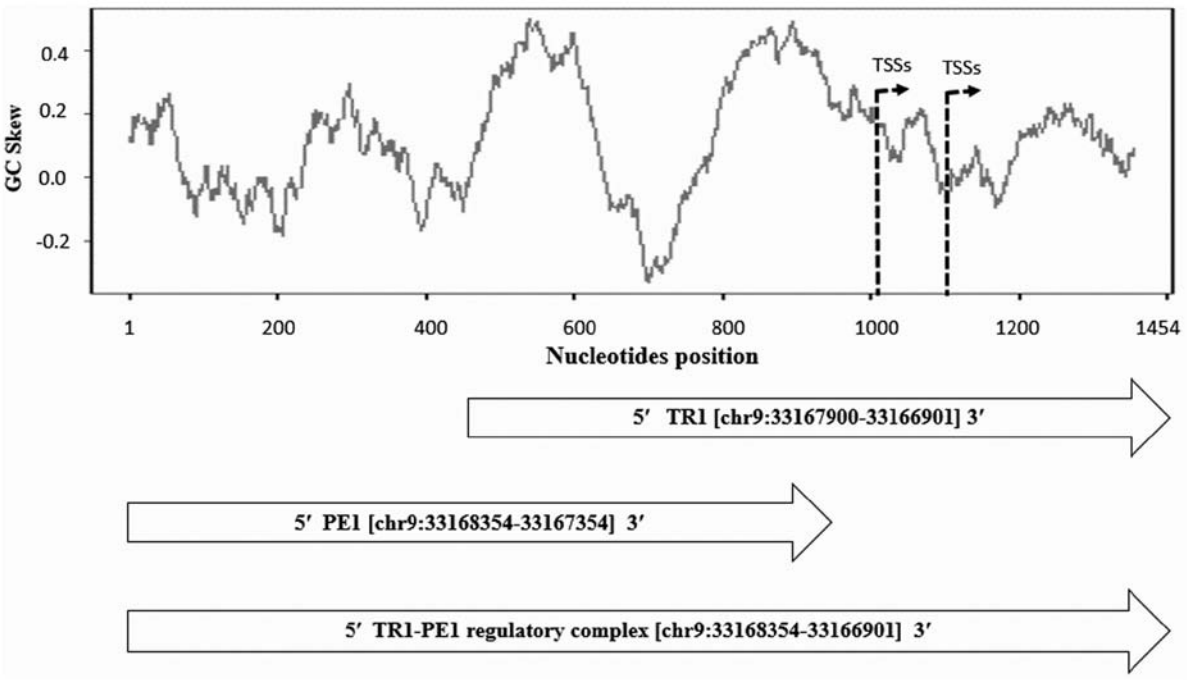

Figure 5. The GC skew (y-axis) along the *B4GALT1* TR1-PE1 promoter sequence (x-axis). The two bent arrows represent six transcription start sites in the bidirectional promoter.

33167138-33167403, which lies within the TR1-PE1 identified here (Fig. 1 in ref. 22) clearly suggested the presence of several sites for transcript initiation, very similar to the putative start sites mapped by us (Fig. 3). Thus, the present study verifies and lends credence to the observed data and likely to be helpful in corroborating the promoter databases, some of which are being constantly revised.

SNPs in the TR1-PE1 regulatory region. Search in the NCBI-dbSNP showed ten SNPs in the TR1-PE1 regulatory region at chr9: 33166911-33167159. Map locations and the type of SNPs are: chr9: 33166911 (missense), chr9: 33166941 (missense), chr9: 33166981 (synonymous), chr9: 33167008 (synonymous), chr9: 33167049 (missense), chr9: 33167109 (missense), chr9: 33167131 (synonymous), chr9: 33167135 (missense), chr9: 33167143 (missense) and chr9: 33167159 (missense). These results highlight another layer of regulation in the TR1-PE1 axis because the SNPs have the potential for altering the binding affinities of transcription factors and changing the transcription efficiency and levels.

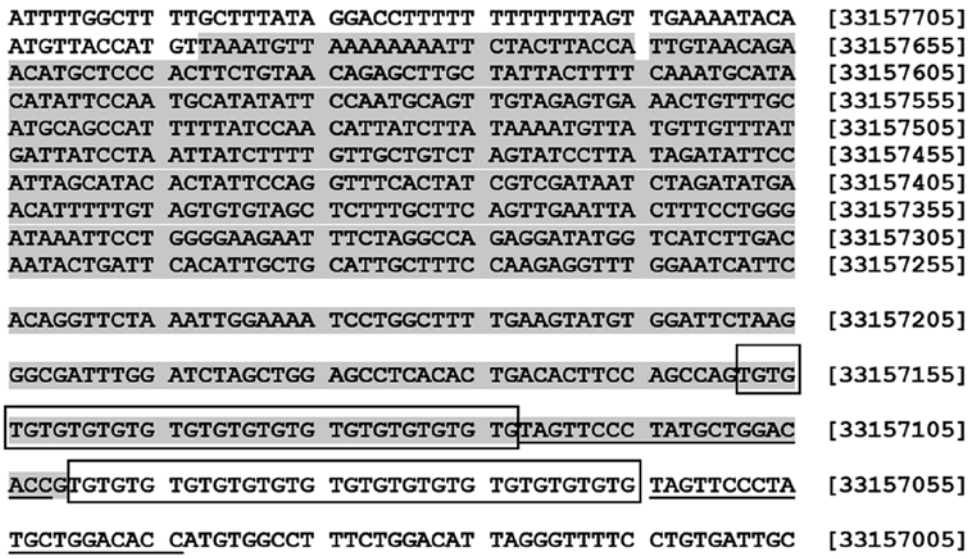

Figure 6. The two $(TG)_{18}$ tandem repeats and TGAS sequence at the 5' regulatory region of *B4GALT1*. The two $(TG)_{18}$ sequences are identified in HP3 intronic alternative promoter located at chr9: 33156756-33158354 close to TR1-PE1 (chr9: 33166901-33168354). Highlighted region is overlapping HP5 sequence, the two TGAS sequences are underlined, the numbers in square brackets show map location of part of the HP3 sequence, which harbor the TG repeats and TGAS sequence.

Sequence alignment of B4GALT1 and B4GALT1-AS1 transcripts. The possible association of *B4GALT1-AS1*, which encodes an lncRNA in the regulation of *B4GALT1* expression, motivated us to explore the sequence complementarity of divergent transcripts. As mentioned in Table II, five *B4GALT1-AS1* transcripts were identified compared to one antisense transcript for each of *B4GALT4*, *B4GALT6* and *B4GALT7* loci. The sequence alignment analysis between primary *B4GALT1* transcript of 4124 bp (NM_001497.3) and predicted *B4GALT1* transcript (XM_005251440.2) with the five *B4GALT1-AS1* transcripts showed complementarity in the range 59.1 and 94.7%, suggesting the potential for duplex formation. Consistent with this finding, the TR1-PE1 has all features and properties of a bidirectional promoter including a divergent or 'head-to-head' configuration of *B4GALT1* and *B4GALT1-AS1* with a 500-bp intervening sequence between them. In addition, the G+C contents, CGIs and TFBSs of TR1-PE1 confirm to the properties found in bidirectional promoters (32-35). We propose that such bidirectional multifaceted regulatory region might be present in the genomic space of the other three *B4GALT* family members with antisense divergent loci, *B4GALT4*, *B4GALT6* and *B4GALT7*. It is worth mentioning that the promoters of *B4GALT* family members are of different types and sequences, and may contain specific regulatory features in *B4GALT4*, *B4GALT6* and *B4GALT7* loci.

Identification of TG repeats in HP3 promoter close to TR1-PE1 regulatory sequence. We investigated the possible involvement of other regulatory sequences near the TR1-PE1 region in the expression of *B4GALT1*, for example TG repeats (36). We found two $(TG)_{18}$ tandem repeats at chr9: 33157065-33157158 located in the intronic alternative promoter HP3 placed 10.146 kb from 3'-TR1-PE1 side (Fig. 6). Our analysis also showed presence of a tag sequence next to $(TG)_{18}$ sequence. The sequence is composed of 22 bases (TAGTTCCCTATGCTGGACACCG), located at the 3' side of both $(TG)_{18}$ repeats, we refer to as the TG Associated Sequence, TGAS. The two sequences were not observed in other promoters of *B4GALT* family members, but they were observed in other loci located in five other chromosomes (Table IV). Notably, these TG repeats and TGAS sequences were linked with cancer genes, for example SOX5, which is reported to be associated with glioma, prostate, testicular seminomas and colorectal cancer. Thus, our data suggest possible regulatory roles for the TG tandem repeats and TGAS in tumorigenesis. In support of this assumption, we identified four novel types of TFBS for SOX5, GLI1, TCF7L2 and GATA3 in the 14.151 kb (chr9: 33167360-33153209) at 5'-side of *B4GALT1* regulatory region that includes TR1-PE1 bidirectional promoter, intronic HP3 promoter, TG repeats and TGAS. The number of identified TFBSs for SOX5 (five), GLI1 (one), TCF7L2 (six) and GATA3 (six); we noted that TCF sites are tied to GATA sites. These TFBSs are known to be associated with cancer and cell identity (see Discussion).

Enhancers, open chromatin and CTCF binding sites. The discovery of *B4GALT1* bidirectional promoter, TG repeats and TGAS sequence enthused us to search for other regulatory sequences that might regulate the expression of *B4GALT1* and other family members. Our search identified following sites in the regulatory regions of *B4GALT* family members: enhancers, open chromatin and CTCF binding sites (Fig. 7). CTCF (CCCTC-binding transcription factor) is a well-studied multifunctional protein involved not only in DNA methylation but also associated with transcriptional activation/repression, and chromatin looping (37). While *B4GALT1*, *B4GALT3* and *B4GALT4* contain a single CTCF binding site, the other members, *B4GALT5*, *B4GALT6* and *B4GALT7* showed 3 to 5 binding sites for CTCF. We found one to three open chromatin sites in *B4GALT1*, *B4GALT4*, *B4GALT5* and *B4GALT6*, which are indicative of active transcription. On the other hand, three genes *B4GALT1*, *B4GALT5* and *B4GALT6* were found

Table IV. The map locations of 116-bp sequence[a] composed of two $(TG)_{18}$ repeats and two TGAS sequences in six human chromosomes.

Chr	Strand	Map location	Identified sequences[b]	Span	Score
9	Reverse	chr9: 33157043-33157158	**TGTGTGTGTGTGTGTGTGTGTGTGTGT GTGTGTG**TAGTTCCCTATGCTGGACACC*G* **TGTGTGTGTGTGTGTGTGTGTGTGTGT GTGTGTG**TAGTTCCCTATGCTGGACACC*A*	116	116/116
1	Forward	chr1: 5387048-5387087	**TGTG*G*TGT*C*TGTGT*GG*TGTGTGTG*GG*TGT GTGTGTG**TAGT	40	33/116
5	Forward	chr5: 26315000-26315046	**GTGTGTGTGTGTGTGTGTGTGTGTGTGT GTGTGTG**TAGTTCCCTA	47	47/116
12	Reverse	chr12: 96164336-96164418	**TGTG*G*TGTGTGTGTGTG*G*TGTGTGTG*AA*TG *CG*GTGTGTGTGT*AG*TGTGTGTG*AA*TGTG *G*TGTGTGTGTGTGTGTGTG**TAGT	83	67/116
12	Reverse	chr12: 23831951-23831984	**TGTGTGTGTGTGTGTGTGTGTGTGTG**TAGTTCCC	34	34/116
12	Reverse	chr12: 96164379-96164418	**TGTG*G*TGTGTGTGTGTG*G*TGTGTGTG*AA*TG*CG*GTGTGTGT**	40	29/116
13	Reverse	chr13: 20498045-20498093	**TGTGTGT*A*TGTGTG*G*TGTGTGTGT*AG*TGTG*TATAG GG*TGTGTGTG**TAGTT	49	37/116
14	Forward	chr14:90274778-90274824	**TGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTG TGTGTGTGTGTGTGTG**TAGTTCCCTAT	61	61/116

[a]The 116-bp sequence was originally identified and located at 5'-*B4GALT1* regulatory region in reverse strand of chromosome 9. [b]Blat tool was used to search for the 116 nucleotides sequence in the human genome hg38. The $(TG)_{18}$ repeats are shown in bold letters, the TGAS sequences are underlined, the italicezed letters show the polymorphisms, insertion and gaps in the sequences.

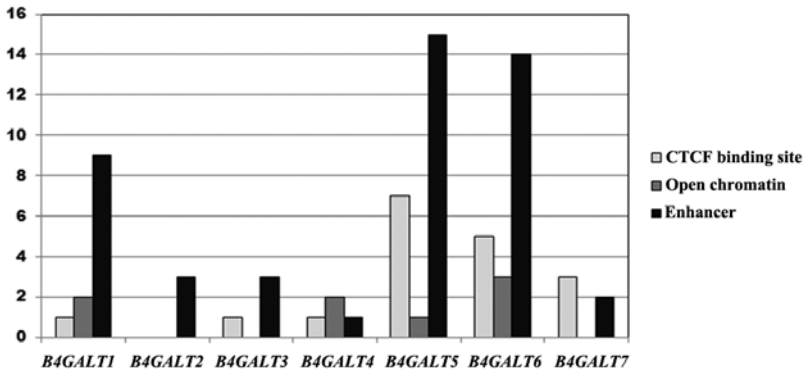

Figure 7. Enhancers, open chromatin and CTCF binding sites in the regulatory regions of *B4GALT* family. The data were extracted from the Ensembl-regulation database of multi-cell regulatory features.

rich in enhancer sequences that are associated with TCF7L2 and GATA3 in unique cell identity regulatory programs for specific expression of these genes in normal and cancer cells. Our analysis showed that the majority of *B4GALT1* enhancers are located within and toward 5'-*B4GALT1* regulatory region.

Evolutionary conservation of regulatory sequences identified in this study. To explore the evolutionary relationships of our discovered regulatory sequences, we generated blast tree maps. The results showed that among the *B4GALT* family, the TR1-PE1 promoter structure was exclusively found in *B4GALT1* and/or *B4GALT1-AS1*. Nevertheless, the TE1-PE1 sequences were highly prevalent and conserved in the *B4GALT1* gene across the animal kingdom. For example, structures similar to TR1-PE1 were observed in *B4GALT1* locus in primates including gorilla (Sequence ID: XM_004047921.1), common chimpanzee (sequence ID: XM_003312037.2) Sumatran orangutan, *Pongo abelii* (Sequence ID: XM_002819701.2), and humans (sequence ID: NG_008919).

TG repeats have been implicated in gene expression, but not widely investigated for their significance in evolution of gene expression. Here, we determined their abundance across the genomes of various species. The 36 bp of the $(TG)_{18}$ repeats were found widespread starting from protozoan

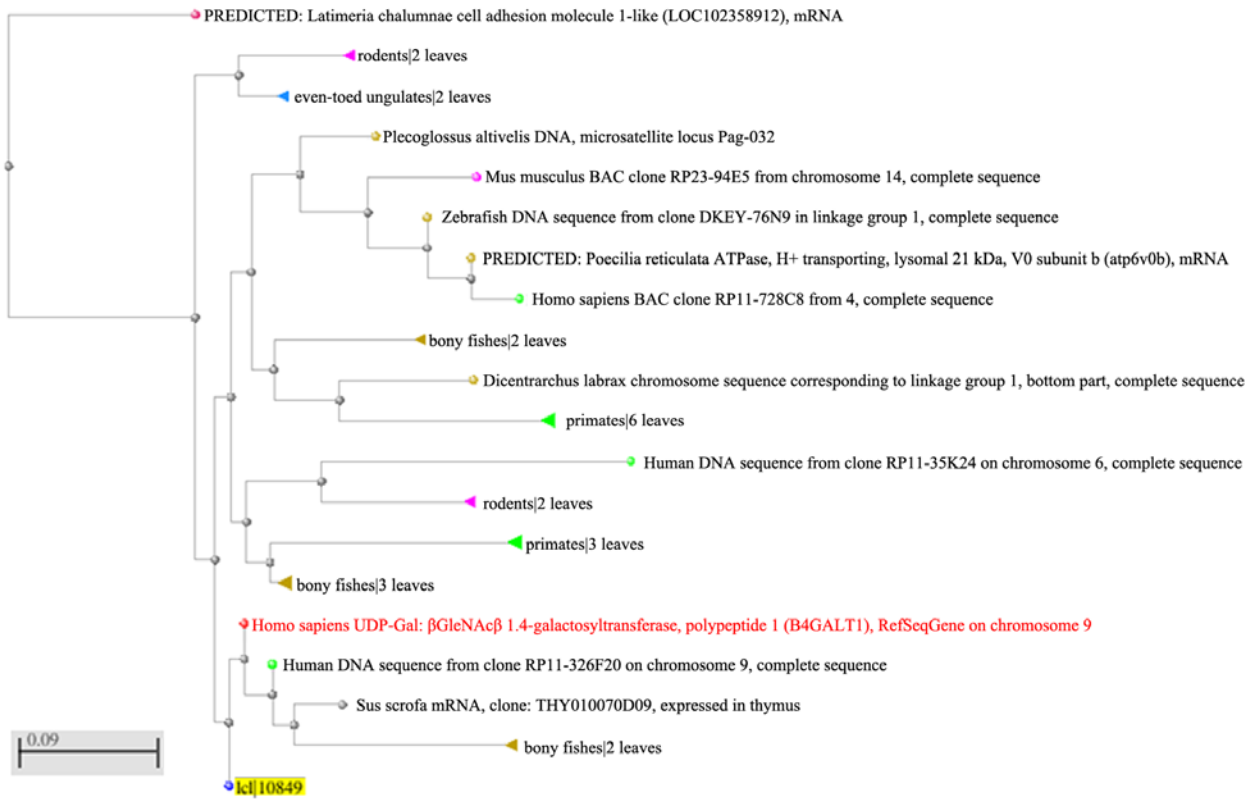

Figure 8. Blast tree map of 58-bp sequence composed from $(TG)_{18}$ and the TGAS of 22 bp. The data imply that the two sequences $(TG)_{18}$ and the TGAS are highly conserved during evolution.

(*trypanosomes*), helminth (*Onchocerca*), higher plants (*Oryza*, *Triticum*) to humans. Then, we searched for the presence and distribution of the 22 bp of TGAS. The sequence was found conserved in *B4GALT1* gene but not in any other *B4GALT* genes. Also, we noted that the 36 bp of $(TG)_{18}$ repeats showed genetic evolutionary genetic relatedness with *B4GALT1* when investigated with first 10 bp of TGAS sequence (TAGTTCCCTA), results in Table IV support this conclusion. These data indicated the critical role of this ancient 10 bp motif in evolution of *B4GALT1* regulatory sequence at 5'-side. Furthermore, the $(TG)_{18}$ sequence showed evolutionary genetic relatedness with fewer species, ~30, when analyzed by blast tool in combination with TGAS sequence (Fig. 8) and their duplicate of 116 bp showed similar results. This remains the first report of TGAS in literature. The results imply that when the $(TG)_{18}$ sequence is tied to TGAS, they become highly conserved during evolution and suggest a specific mechanistic role in gene expression.

Discussion

B4GALT1 has emerged as a model gene for deciphering the complex transcriptional regulation underlying its cell type- and tissue-specific gene expression. As an enzyme and cell recognition molecule mediating cell adhesion and signaling, *B4GALT1* serves many important roles in both physiological and pathological settings. Although, a survey of databases indicated the presence of more than one promoter for *B4GALT1*, before the present study, there was no definitive mapping of the transcription start sites for this gene. Our extensive and careful analysis revealed a large number of alternative promoters, thirteen to be exact, involved in human *B4GALT1* expression. Of these six overlapped and resided within 1.454 kb forming a complex transcriptional regulatory unit referred to as the TR1-PE1, which contains six TSS and apparently mediates most of *B4GALT1* expression. A recent publication by the FANTOM Consortium (22) showed multiple transcription start sites for the *B4GALT1* promoter at 33167138-33167403 are located within the 266 base CGI, which is part of TR1-PE1 regulatory complex sequence characterized in the present study. This region was also shown to be associated with *B4GALT1* cell type-specific transcription. These data support our findings that the six TSSs present in the TR1-PE1 complex harboring six overlapping alternative promoters may all be involved in the expression of *B4GALT1* in different tissues.

A most significant finding of the present study is the revelation that the *B4GALT1* gene is paired with the production of a long non-coding RNA, which can potentially function as an antisense transcript, called *B4GALT1-AS1* via the TR1-PE1 regulatory complex. More specifically, the TR1-PE1 is a bidirectional promoter that divergently directs the transcription of *B4GALT1* and *B4GALT1-AS1* genes in a head to head orientation. The configuration of any paired genes can be co-directional (either ← ← or → →), convergent (→ ←) or divergent (← →). Many promoters show divergent transcription, and it is estimated that more than 10% of the genes in the human genome are divergently transcribed wherein the genes share a single promoter with their transcription start sites separated by <1,000 base pairs. Examples include the

DNA repair genes, such as the BRCA1, BRCA2, CKEK1 and FANC family members (32,33). Many bidirectional pairs are co-expressed, but some are anti-regulated. Furthermore, the promoter segments between two bidirectional genes initiate transcription in both directions and contain shared elements that regulate both genes, thus, providing a unique mechanism of regulation for numerous genes (34). Our characterization of TR1-PE1 regulatory complex showed that it bridges the two adjacent head to head genes, *B4GALT1* and *B4GALT1-AS1*, which are transcribed in opposite directions. Several types of TFBSs and CGI identified in TR1-PE1 are indeed found in bidirectional promoters (34,35). In this respect, the reported promoters by Poeta *et al* (13) and Kim *et al* (20), which are more likely similar to one of overlapping alternative promoters composed TR1-PE1 regulatory sequence, were found to contain methylated CGIs, which highlight the role of CGIs in *B4GALT1* bidirectional promoter and link them to cancer and drug resistance.

Another salient finding here pertains to the possible generation of five lncRNA transcript species under the control of TR1-PE1 bidirectional promoter (Table II); these natural and putative antisense transcripts arising from the forward strand showed significant potential for forming RNA-RNA duplexes. The lncRNAs are abundant in the human genome and the FANTOM3 project identified ~35,000 non-coding transcripts from ~10,000 distinct loci (38) that bear many signatures of mRNAs, including 5' capping, splicing, and poly-adenylation, but have little or no open reading frame (ORF). Most lncRNAs are >200 nucleotides long and recent evidence points to a variety of functions for them in cellular processes. These include the activation or inhibition of transcription, organization of nuclear sub-structures, alteration of chromatin state, and regulation of gene expression through the interaction with effector proteins and modulation of their activity (39-42). Our data highlight the complexity underlying the transcriptional and possible post-transcriptional (by the lncRNA) regulation of the *B4GALT1* gene in its differential expression in human tissues.

Also, we considered possible involvement of other elements in modulating *B4GALT1* expression. In the present study, we identified $(TG)_{18}$ tandem repeats placed 10.146 kb from TR1-PE1 regulatory sequence. It has been reported that TG tandem repeats, which are highly conserved throughout eukaryotic genome evolution, enhance transcription especially when they are located closer to the promoter (36,43,44). Although, the potential roles of these repeats have been demonstrated in recombination (45), DNA repair (46), alternative splicing (47) and breast cancer (48,49), some observations do not support a function for these repeats in gene expression (36).

Our analysis of dinucleotide base-stacking energy along TR1-PE1 showed variable values, which reflect the patterns of double helix dissociation in the regulatory sequence. Additionally, we were able to correlate and co-localize the values of dinucleotide base-stacking energy with the transcription factor binding sequences. As described by Ornstein *et al* (29), the relative stability of the double helix structure can be demonstrated according to the range of dinucleotide base-stacking energy values from -3.82 kcal/mol (unstack easily) to -14.59 kcal/mol (difficult to unstack), thus the obtained values for INR sites are at regions that dissociate easily in comparison with BRE and TSS sites. Also, we observed another interesting structural feature in the TR1-PE1 bidirectional promoter, which is related to the GC skew type at TSS sites and possible formation of R-loops. It has been reported that GC strand asymmetry downstream of TSSs is prone to R-loop formation that is correlated with unmethylated status of CGI (31). These data on structural features of TR1-PE1 regulatory region, in addition to other criteria of alternative promoters reported in present study, intergenic or intragenic locations, types of TFBSs, CGI and potential role of $(TG)_{18}$ tandem repeats are critical factors and have important consequences in the transcription process and cell type specific expression of *B4GALT1*.

The identification of cell type-specific TFBS (TCF7L2 and GATA3) in the vicinity of TR1-PE1 and HP3 intronic promoter provided further functions played by *B4GALT1* in the development of cancer, cancer stem cell-linked expression and relapse. The TCF7L2 transcription factor is linked to a variety of human diseases, including cancer (50). It plays a critical role in enhancer activity, especially super enhancer, which is known to be associated with cell identity and diseases (50-54). Furthermore, TCF7L2 is tied to the genome by association with GATA3 (50), which is the case shown in this study. GATA transcription factors are also involved in carcinogenesis. The roles of GATA factors in carcinogenesis *vis-à-vis* the normal functions is a result of malfunctions. GATA1 mutations are associated with megakaryoblastic leukemias in patients with Down syndrome; loss of GATA3 expression is involved in breast cancer; whereas silencing of GATA4 and GATA5 expression are reported in gastric, colorectal and lung cancer (53,54).

Another finding was the identification of the binding site for glioma-associated oncogene homolog 1 (GLI1) transcription factor adjacent to TG-TGAS sequence. GLI1 is the nuclear mediator of the Hedgehog pathway that regulates genes essential for various stages of tumor development and progression (55). Accordingly, it is proposed that GLI1 is a potential target for cancer therapy. Recently, four natural compounds of physalins family showed dose-dependent GLI1-transcriptional inhibitory activity (56).

Furthermore, we identified ten SNPs in the *B4GALT1* TR1-PE1 regulatory sequence; these can modify splicing and can alter *B4GALT1* expression. The SNPs are frequent in the human genome, the 1000 Genomes Project reported 38 million SNPs. SNPs that are located within the exons or exon-intron boundaries can modify the splicing sites and consequently the protein function resulting in development of many associated diseases including cancer (57). Many SNPs are located within the regulatory regions of genes, which may influence their expression (58).

Our studies on *B4GALT1* also enabled a comparison of regulatory elements present within the *B4GALT1* gene family. This is important, because the β-1, 4-galactosyltransferases perform redundant catalysis and functions, often compensating within and outside of this class of enzymes (1,4). For example, in *B4GALT1* knockout mice, there was a shift in the galactose linkages from the largely β-1,4 linkage to β-1,3 linkage, suggesting that *B4GALT1* deficiency was compensated for by β-1,3-galactosyltransferases (59). We observed both similarities and differences in the genomic regulatory features when

B4GALT1 was compared with the other members of the family. The similarities included the presence of multiple promoters and some members with long non-coding RNAs. Differences were that the TR1-PE1, $(TG)_{18}$ and TGAS sequences were restricted to *B4GALT1* and the epigenetic landscape appeared to be different as well.

In conclusion, this study provides further insight into the regulatory features that govern *B4GALT1* and reveal a novel bidirectional promoter and multifaceted regulatory region at the 5'-*B4GALT1* locus, which includes several genomic-epigenetic regulatory elements that control at least six transcription start sites embedded within regulatory sequence of 1.454 kb. The identified $(TG)_{18}$ and the TGAS sequences in the regulatory regions of *B4GALT1*, which were found conserved through evolution in many species, give further indication of involvement and evolution of several ancient sequences in the transcription process of specific genes. The data highlight the complexity and sophistication underlying the transcriptional and possible post-transcriptional regulation of the *B4GALT1* gene and its differential expression in human normal and cancer cells. The special genomic-epigenomic characteristics of *B4GALT1* gene expression is manifested by presence of several antisense lncRNA transcripts in comparison to other *B4GALT* members. There are five *B4GALT1-AS1* transcripts that have the potential to control the expression of two *B4GALT1* transcripts. The data are likely to advance and develop the role of this important enzyme in cancer pathophysiology, drug resistance and personalized therapy.

Acknowledgements

The present study was supported in part by a grant from the Cancer Prevention and Research Institute of Texas (RP130266) to KSS. We thank Ibtisam Ismael Alobaidi for technical assistance.

References

1. Amado M, Almeida R, Schwientek T and Clausen H: Identification and characterization of large galactosyltransferase gene families: Galactosyltransferases for all functions. Biochim Biophys Acta 1473: 35-53, 1999.
2. Qasba PK, Ramakrishnan B and Boeggeman E: Structure and function of beta -1,4-galactosyltransferase. Curr Drug Targets 9: 292-309, 2008.
3. Appert HE, Rutherford TJ, Tarr GE, Wiest JS, Thomford NR and McCorquodale DJ: Isolation of a cDNA coding for human galactosyltransferase. Biochem Biophys Res Commun 139: 163-168, 1986.
4. Hennet T: The galactosyltransferase family. Cell Mol Life Sci 59: 1081-1095, 2002.
5. Lopez LC, Youakim A, Evans SC and Shur BD: Evidence for a molecular distinction between Golgi and cell surface forms of beta 1,4-galactosyltransferase. J Biol Chem 266: 15984-15991, 1991.
6. Choi HJ, Chung TW, Kim CH, Jeong HS, Joo M, Youn B and Ha KT: Estrogen induced β-1,4-galactosyltransferase 1 expression regulates proliferation of human breast cancer MCF-7 cells. Biochem Biophys Res Commun 426: 620-625, 2012.
7. Liu W, Cui Z, Wang Y, Zhu X, Fan J, Bao G, Qiu J and Xu D: Elevated expression of β1,4-galactosyltransferase-I in cartilage and synovial tissue of patients with osteoarthritis. Inflammation 35: 647-655, 2012.
8. Zhou H, Ma H, Wei W, Ji D, Song X, Sun J, Zhang J and Jia L: B4GALT family mediates the multidrug resistance of human leukemia cells by regulating the hedgehog pathway and the expression of p-glycoprotein and multidrug resistance-associated protein 1. Cell Death Dis 4: e654, 2013.
9. Mengle-Gaw L, McCoy-Haman MF and Tiemeier DC: Genomic structure and expression of human beta-1,4-galactosyltransferase. Biochem Biophys Res Commun 176: 1269-1276, 1991.
10. Shaper NL, Charron M, Lo NW and Shaper JH: Beta1,4-galactosyltransferase and lactose biosynthesis: Recruitment of a housekeeping gene from the nonmammalian vertebrate gene pool for a mammary gland specific function. J Mammary Gland Biol Neoplasia 3: 315-324, 1998.
11. Zhang S, Cai M, Zhang SW, Hu Y and Gu JX: Involvement of beta 1,4 galactosyltransferase 1 and Gal beta1-->4GlcNAc groups in human hepatocarcinoma cell apoptosis. Mol Cell Biochem 243: 81-86, 2003.
12. Zhu X, Jiang J, Shen H, Wang H, Zong H, Li Z, Yang Y, Niu Z, Liu W, Chen X, *et al*: Elevated beta1,4-galactosyltransferase I in highly metastatic human lung cancer cells. Identification of E1AF as important transcription activator. J Biol Chem 280: 12503-12516, 2005.
13. Poeta ML, Massi E, Parrella P, Pellegrini P, De Robertis M, Copetti M, Rabitti C, Perrone G, Muda AO, Molinari F, *et al*: Aberrant promoter methylation of beta-1,4 galactosyltransferase 1 as potential cancer-specific biomarker of colorectal tumors. Genes Chromosomes Cancer 51: 1133-1143, 2012.
14. Radhakrishnan P, Chachadi V, Lin MF, Singh R, Kannagi R and Cheng PW: TNFα enhances the motility and invasiveness of prostatic cancer cells by stimulating the expression of selective glycosyl- and sulfotransferase genes involved in the synthesis of selectin ligands. Biochem Biophys Res Commun 409: 436-441, 2011.
15. Yamashita H, Kubushiro K, Ma J, Fujii T, Tsukazaki K, Iwamori M and Nozawa S: Alteration in the metastatic potential of ovarian cancer cells by transfection of the antisense gene of beta-1,4-galactosyltransferase. Oncol Rep 10: 1857-1862, 2003.
16. Zhou H, Zhang Z, Liu C, Jin C, Zhang J, Miao X and Jia L: B4GALT1 gene knockdown inhibits the hedgehog pathway and reverses multidrug resistance in the human leukemia K562/adriamycin-resistant cell line. IUBMB Life 64: 889-900, 2012.
17. Chang X, Monitto CL, Demokan S, Kim MS, Chang SS, Zhong X, Califano JA and Sidransky D: Identification of hypermethylated genes associated with cisplatin resistance in human cancers. Cancer Res 70: 2870-2879, 2010.
18. Helleman J, Jansen MP, Span PN, van Staveren IL, Massuger LF, Meijer-van Gelder ME, Sweep FC, Ewing PC, van der Burg ME, Stoter G, *et al*: Molecular profiling of platinum resistant ovarian cancer. Int J Cancer 118: 1963-1971, 2006.
19. Yuan Q, Yang H, Cheng C, Li C, Wu X, Huan W, Sun H, Zhou Z, Wang Y, Zhao Y, *et al*: β-1,4-Galactosyltransferase I involved in Schwann cells proliferation and apoptosis induced by tumor necrosis factor-alpha via the activation of MAP kinases signal pathways. Mol Cell Biochem 365: 149-158, 2012.
20. Kim MS, Louwagie J, Carvalho B, Terhaar Sive Droste JS, Park HL, Chae YK, Yamashita K, Liu J, Ostrow KL, Ling S, *et al*: Promoter DNA methylation of oncostatin m receptor-beta as a novel diagnostic and therapeutic marker in colon cancer. PLoS One 4: e6555, 2009.
21. Michailidi C, Soudry E, Brait M, Maldonado L, Jaffe A, Ili-Gangas C, Brebi-Mieville P, Perez J, Kim MS, Zhong X, *et al*: Genome-wide and gene-specific epigenomic platforms for hepatocellular carcinoma biomarker development trials. Gastroenterol Res Pract 2014: 597164, 2014.
22. Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T, Kulakovskiy IV, Lizio M, Itoh M, *et al*; FANTOM Consortium and the RIKEN PMI and CLST (DGT): A promoter-level mammalian expression atlas. Nature 507: 462-470, 2014.
23. Yang C, Bolotin E, Jiang T, Sladek FM and Martinez E: Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. Gene 389: 52-65, 2007.
24. Hsu PC, Chao CC, Yang CY, Ye YL, Liu FC, Chuang YJ and Lan CY: Diverse Hap43-independent functions of the Candida albicans CCAAT-binding complex. Eukaryot Cell 12: 804-815, 2013.
25. Ko LJ and Engel JD: DNA-binding specificities of the GATA transcription factor family. Mol Cell Biol 13: 4011-4022, 1993.
26. Blauwkamp TA, Chang MV and Cadigan KM: Novel TCF-binding sites specify transcriptional repression by Wnt signalling. EMBO J 27: 1436-1446, 2008.
27. Gardiner-Garden M and Frommer M: CpG islands in vertebrate genomes. J Mol Biol 196: 261-282, 1987.

28. Hackenberg M, Barturen G, Carpena P, Luque-Escamilla PL, Previti C and Oliver JL: Prediction of CpG-island function: CpG clustering vs. sliding-window methods. BMC Genomics 11: 327, 2010.
29. Ornstein RL, Rein R, Breen D and Macelroy R: An optimized potential function for the calculation of nucleic acid interaction energies I. base stacking. Biopolymers 17: 2341-2360, 1978.
30. Vesth T, Lagesen K, Acar Ö and Ussery D: CMG-biotools, a free workbench for basic comparative microbial genomics. PLoS One 8: e60120, 2013.
31. Ginno PA, Lott PL, Christensen HC, Korf I and Chédin F: R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell 45: 814-825, 2012.
32. Seila AC, Core LJ, Lis JT and Sharp PA: Divergent transcription: A new feature of active promoters. Cell Cycle 8: 2557-2564, 2009.
33. Yang MQ, Koehly LM and Elnitski LL: Comprehensive annotation of bidirectional promoters identifies co-regulation among breast and ovarian cancer genes. PLOS Comput Biol 3: e72, 2007.
34. Orekhova AS and Rubtsov PM: Bidirectional promoters in the transcription of mammalian genomes. Biochemistry (Mosc) 78: 335-341, 2013.
35. Yang MQ and Elnitski LL: Diversity of core promoter elements comprising human bidirectional promoters. BMC Genomics 9 (Suppl 2): S3, 2008.
36. Zhang W, He L, Liu W, Sun C and Ratain MJ: Exploring the relationship between polymorphic (TG/CA)n repeats in intron 1 regions and gene expression. Hum Genomics 3: 236-245, 2009.
37. De La Rosa-Velázquez IA, Rincón-Arano H, Benítez-Bribiesca L and Recillas-Targa F: Epigenetic regulation of the human retinoblastoma tumor suppressor gene promoter by CTCF. Cancer Res 67: 2577-2585, 2007.
38. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, *et al*; RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group): The transcriptional landscape of the mammalian genome. Science 309: 1559-1563, 2005.
39. Uesaka M, Nishimura O, Go Y, Nakashima K, Agata K and Imamura T: Bidirectional promoters are the major source of gene activation-associated non-coding RNAs in mammals. BMC Genomics 15: 35, 2014.
40. Lepoivre C, Belhocine M, Bergon A, Griffon A, Yammine M, Vanhille L, Zacarias-Cabeza J, Garibal MA, Koch F, Maqbool MA, *et al*: Divergent transcription is associated with promoters of transcriptional regulators. BMC Genomics 14: 914, 2013.
41. Magistri M, Faghihi MA, St Laurent G III and Wahlestedt C: Regulation of chromatin structure by long noncoding RNAs: Focus on natural antisense transcripts. Trends Genet 28: 389-396, 2012.
42. Kung JT, Colognori D and Lee JT: Long noncoding RNAs: Past, present, and future. Genetics 193: 651-669, 2013.
43. Hamada H, Petrino MG and Kakunaga T: A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. Proc Natl Acad Sci USA 79: 6465-6469, 1982.
44. Hamada H, Seidman M, Howard BH and Gorman CM: Enhanced gene expression by the poly(dT-dG).poly(dC-dA) sequence. Mol Cell Biol 4: 2622-2630, 1984.
45. Dutreix M: (GT)n repetitive tracts affect several stages of RecA-promoted recombination. J Mol Biol 273: 105-113, 1997.
46. Huang W, Zheng J, He Y and Luo C: Tandem repeat modification during double-strand break repair induced by an engineered TAL effector nuclease in zebrafish genome. PLoS One 8: e84176, 2013.
47. Hui J, Hung L-H, Heiner M, Schreiner S, Neumüller N, Reither G, Haas SA and Bindereif A: Intronic CA-repeat and CA-rich elements: A new class of regulators of mammalian alternative splicing. EMBO J 24: 1988-1998, 2005.
48. Zenklusen JC, Bièche I, Lidereau R and Conti CJ: (C-A)n microsatellite repeat D7S522 is the most commonly deleted region in human primary breast cancer. Proc Natl Acad Sci USA 91: 12155-12158, 1994.
49. Mukherjee B, Zhao H, Parashar B, Sood BM, Mahadevia PS, Klinger HP, Vikram B and Achary MP: Microsatellite dinucleotide (T-G) repeat: A candidate DNA marker for breast metastasis. Cancer Detect Prev 27: 19-23, 2003.
50. Frietze S, Wang R, Yao L, Tak YG, Ye Z, Gaddis M, Witt H, Farnham PJ and Jin VX: Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3. Genome Biol 13: R52, 2012.
51. Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-André V, Sigova AA, Hoke HA and Young RA: Super-enhancers in the control of cell identity and disease. Cell 155: 934-947, 2013.
52. Pott S and Lieb JD: What are super-enhancers? Nat Genet 47: 8-12, 2015.
53. Akiyama Y, Watkins N, Suzuki H, Jair KW, van Engeland M, Esteller M, Sakai H, Ren CY, Yuasa Y, Herman JG, *et al*: GATA-4 and GATA-5 transcription factor genes and potential downstream antitumor target genes are epigenetically silenced in colorectal and gastric cancer. Mol Cell Biol 23: 8429-8439, 2003.
54. Zheng R and Blobel GA: GATA transcription factors and cancer. Genes Cancer 1: 1178-1188, 2010.
55. Carpenter RL and Lo HW: Hedgehog pathway and GLI1 isoforms in human cancer. Discov Med 13: 105-113, 2012.
56. Arai MA, Uchida K, Sadhu SK, Ahmed F and Ishibashi M: Physalin H from Solanum nigrum as an Hh signaling inhibitor blocks GLI1-DNA-complex formation. Beilstein J Org Chem 10: 134-140, 2014.
57. Faber K, Glatting KH, Mueller PJ, Risch A and Hotz-Wagenblatt A: Genome-wide prediction of splice-modifying SNPs in human genes using a new analysis pipeline called AASsites. BMC Bioinformatics 12 (Suppl 4): S2, 2011.
58. Guo Y and Jamison DC: The distribution of SNPs in human gene regulatory regions. BMC Genomics 6: 140, 2005.
59. Kotani N, Asano M, Iwakura Y and Takasaki S: Knockout of mouse beta 1,4-galactosyltransferase-1 gene results in a dramatic shift of outer chain moieties of N-glycans from type 2 to type 1 chains in hepatic membrane and plasma glycoproteins. Biochem J 357: 827-834, 2001.