Analysis of the proximal promoter of the human colon-specific B4GALNT2 (Sda synthase) gene: B4GALNT2 is transcriptionally regulated by ETS1

Cindy Wavelet-Vermuse, Sophie Groux-Degroote, Dorothée Vicogne, Virginie Cogez, Giulia Venturi, Marco Trinchera, Guillaume Brysbaert, Marie-Ange Krzewinski-Recchi, Elsa Hadj Bachir, Céline Schulz, et al.

BBA - Gene Regulatory Mechanisms 1864 (2021) 194747

Contents lists available at ScienceDirect

BBA - Gene Regulatory Mechanisms

journal homepage: www.elsevier.com/locate/bbagrm

Analysis of the proximal promoter of the human colon-specific *B4GALNT2* (Sd^a synthase) gene: *B4GALNT2* is transcriptionally regulated by ETS1

Cindy Wavelet-Vermuse [a], Sophie Groux-Degroote [a], Dorothée Vicogne [a], Virginie Cogez [a], Giulia Venturi [b], Marco Trinchera [c], Guillaume Brysbaert [a], Marie-Ange Krzewinski-Recchi [a], Elsa Hadj Bachir [d], Céline Schulz [a], Audrey Vincent [d], Isabelle Van Seuningen [d], Anne Harduin-Lepers [a,*]

[a] Univ. Lille, CNRS, UMR 8576 - UGSF - Unité de Glycobiologie Structurale et Fonctionnelle, F-59000 Lille, France
[b] Department of Experimental, Diagnostic and Specialty Medicine (DIMES), General Pathology Building, University of Bologna, 40126 Bologna, Italy
[c] Department of Medicine and Surgery, University of Insubria, 21100, Varese, Italy
[d] Univ. Lille, CNRS, Inserm, CHU Lille, UMR9020 - U1277 - CANTHER - Cancer Heterogeneity, Plasticity and Resistance to Therapies, F-59000 Lille, France

ARTICLE INFO

Keywords:
Transcriptional regulation
Core promoter
ETS1
Colon
B4GALNT2

ABSTRACT

Background: The Sd^a antigen and corresponding biosynthetic enzyme B4GALNT2 are primarily expressed in normal colonic mucosa and are down-regulated to a variable degree in colon cancer tissues. Although their expression profile is well studied, little is known about the underlying regulatory mechanisms. Methods: To clarify the molecular basis of Sd^a expression in the human gastrointestinal tract, we investigated the transcriptional regulation of the human *B4GALNT2* gene. The proximal promoter region was delineated using luciferase assays and essential *trans*-acting factors were identified through transient overexpression and silencing of several transcription factors. Results: A short *cis*-regulatory region restricted to the −72 to +12 area upstream of the *B4GALNT2* short-type transcript variant contained the essential promoter activity that drives the expression of the human *B4GALNT2* regardless of the cell type. We further showed that *B4GALNT2* transcriptional activation mostly requires ETS1 and to a lesser extent SP1. Conclusions: Results presented herein are expected to provide clues to better understand *B4GALNT2* regulatory mechanisms.

1. Introduction

The histo blood group antigen Sd^a was discovered long ago [1,2]. It is the only antigen of the Sid system [3]: mostly expressed on red blood cells, it is also detected in colon and kidney tissues as well as in urine and saliva, as reviewed recently [4]. In the human healthy colon, this sialylated trisaccharide (GalNAcβ1-4(NeuAcα2-3)Galβ1-) is primarily found on core 3 *O*-glycans of mucins [5] and, in colorectal cancer cell lines, it is described on core 1 and core 2 *O*-glycans of mucins [6]. It has long been observed that this Sd^a epitope is drastically down-regulated in colon cancer to the benefit of the sialyl Lewis x (sLe^x, NeuAcα2-3Galβ1-4 [Fucα1-3]GlcNAc-) epitope; however, the control mechanisms regulating this balanced expression remain unknown.

The human *B4GALNT2* gene encodes the Sd^a synthase (i.e. the β-1,4-*N*-acetylgalactosaminyltransferase 2), which was cloned concomitantly by two independent groups from the colonic Caco-2 cells [7,8]. The Sd^a synthase catalyzes the transfer of an *N*-acetylgalactosamine residue to an α2,3-sialylated galactose on either an *O*- or *N*-glycan in vitro [8] and in colon biopsies [9] or to an extended carbohydrate chain of glycolipid like sialylparagloboside in stomach [10]. The human *B4GALNT2* gene is located on chromosome 17q21.33 and contains 11 coding exons. Interestingly, previous studies using Northern blot [8] and PCR analyses ([7] suggested the existence of multiple transcripts diverging in their 5′- and 3′-untranslated region (UTR). Five *B4GALNT2* transcripts were expressed in the human colon and to a lower extent in ileum, stomach and kidney and 5′-rapid amplification of cDNA ends (5'-RACE) analysis conducted in differentiated Caco-2 cells demonstrated the existence of at least two transcriptional start sites (TSS) and two alternative first exons (AFE) in these cells. The long exon 1 noted exon 1 L encompasses 253 nucleotide (nt) (59 nt untranslated region (UTR) +194 nt coding region

* Corresponding author at: Unité de Glycobiologie Structurale et Fonctionnelle, UMR CNRS 8576, Université de Lille, Faculté des Sciences et Technologies, 59655 Villeneuve d'Ascq, France.
E-mail address: anne.harduin-lepers@univ-lille.fr (A. Harduin-Lepers).

https://doi.org/10.1016/j.bbagrm.2021.194747
Received 16 June 2021; Received in revised form 26 July 2021; Accepted 9 August 2021
Available online 7 September 2021

(CR), and the short exon 1 known as exon 1S is 38 nt long (24 nt UTR + 14 nt CR), leading to two transcript variants referred to as SF- and LF-type that differ at their 5′-end [8]. Since both exon 1 L and 1S contain a translational start site, *B4GALNT2* drives the expression of two distinct B4GALNT2 protein isoforms, a 506 amino-acids (aa) short form and a 566 aa long form exhibiting a very long cytoplasmic *N*-terminus. Interestingly, expression of these two protein isoforms in the colon cancer LS174T cells has shown that i) the shorter isoform has higher biosynthetic activity than the long form [11], ii) the longer isoform exhibits an additional post-Golgi subcellular localization [12] and iii) both the long and short B4GALNT2 isoforms are able to induce the expression of Sd^a and the inhibition of sLe^x antigens on glycoproteins from LS174T cells [11]. A third AFE and a 51 nt middle exon 1 (exon 1M) were originally predicted in GenBank (Accession number NM_001159388) [4] uncovering the existence of a third *B4GALNT2* transcript variant. This latter could be amplified using PCR, cloned and sequenced from CCD 841 CoN cells, stomach and colon samples [9]. No translational start codon (ATG) can be detected in exon 1 M, and the first in frame ATG is located in exon 2, downstream the transmembrane domain. Therefore, this transcript variant would drive the expression of a third B4GALNT2 isoform of 479 aa lacking its transmembrane domain, supporting the observations made by Western blot of the existence of shorter isoforms [9]. Nonetheless, PCR analyses of these transcript variants in colonic cells and human biopsies demonstrated that the short transcript variant was primarily expressed and dramatically down-regulated in cancer colon, whereas the long and middle variants were barely detectable [9]. The normal colon cells CCD 841 CoN expressed high levels of the short transcript variants and low levels of the long transcript variant whereas the cancer cell lines HT-116 did not express the *B4GALNT2* gene. These previous data strongly suggested an important point of *B4GALNT2* gene regulation at the transcriptional level. As for many other carbohydrate epitopes [13], Sd^a biosynthesis appears to be regulated at multiple levels in the gastrointestinal tract and regulation of B4GALNT2 is particularly complex. In the past, it has been shown that promoter hypermethylation of *B4GALNT2* silenced the expression of this gene in colon and gastric cancer cell lines and tissues, which might be related to aberrant expression of cancer-associated sLe^x antigen [14,15]. However, a recent study unveiled the important but unusual role of CpG sites methylation in the *B4GALNT2* gene since methylation of an intron-located open-sea site between exon 6 and 7 rather increased *B4GALNT2* expression [16] highlighting the complex regulatory mechanisms underpinning *B4GALNT2* expression in the digestive tract.

As a starting point, we aimed in this study to clarify the molecular basis of Sd^a expression in the human gastrointestinal tract. For that purpose, we explored transcriptional regulation of the human *B4GALNT2* gene in various gastro-intestinal cultured cells. We used 5′-RACE and a dual luciferase assay for sequential deletion and site-directed mutagenesis to identify transcriptional start site (TSS) and delineate the core promoter region. We determined the *cis*-acting elements and transcription factors (TF) that regulate *B4GALNT2* expression in CCD 841 CoN colon cells using transient overexpression and silencing of SP1, ETS1 and DMTF1, and chromatin immunoprecipitation (ChIP) assay. Our results indicating that 5′-flanking region at positions −72 to +12 nt relative to the *B4GALNT2* short variant start codon is critical for *B4GALNT2* transcription and mRNA expression in colon pave the way to unravel molecular mechanism underpinning *B4GALNT2* expression.

2. Material and methods

2.1. Plasmids construction and site-directed mutagenesis

Genomic DNA from HT-29 cells was prepared with Genomic DNA Mini Kit (Thermo Fisher Scientific Bioscience, Villebon sur-Yvette, France).) following manufacturer's instructions. The genomic sequence located between −1559 bp and + 675 bp was amplified by PCR using the primer pairs P1*Kpn*I and P1*Hind*III (Supplemental Table 1). The PCR product was cloned into the pCR®2.1 TOPO vector (Thermo Fisher Scientific Bioscience, Villebon sur-Yvette, France). The promoter sequence was then isolated by digestion of the vector using *Kpn*I and *Hind*III restriction sites. The purified fragment was sub-cloned into pGL3Basic vector (Promega, Madison, USA) upstream of the Firefly luciferase gene at *Kpn*I/*Hind*III sites. The resulting plasmid was designed pGL3(−1559/+675). Truncated promoter constructs were generated either by enzymatic digestions, ends blunting and ligation (Supplemental Table 2) or PCR (Supplemental Table 1) or substitution (Supplemental Table 3). All plasmid constructs were sequenced (GATC Biotech, Köln, Germany).

The SP1.1, SP1.2, SP1.3, DMTF1/ETS1, DMTF1 and ETS1 consensus sequences in the *B4GALNT2* promoter region present in the pGL3(−83/+72) plasmid were mutated by base substitution using a Quick-Change mutagenesis kit (Stratagene, La Jolla, CA) according to the manufacturer's protocol, using the mutagenetic oligonucleotide primers shown in Supplemental Table 1. The PCR conditions were a first denaturation step at 95 °C, followed by 12 cycles: 95 °C for 30 s, 55 °C for 1 min, 68 °C for 5 min. The PCR product was then digested with the *Dpn*I restriction enzyme for 1 h at 37 °C. After bacterial transformation using XL1-Blue competent cells and plasmid purification, all plasmids were sequenced to check the mutation of each site.

2.2. Cell culture, transient transfections (Luciferase assays, overexpression and siRNA) and luciferase assays

The human colon carcinoma cell line HCT-116 (ATCC, CCL-247) and the human cervical cancer cells HeLa (ATCC CCL-2) were routinely grown in Dulbecco's Modified Eagle Medium (DMEM) containing 2 mM glutamine and 10% fetal calf serum, at 37 °C in an atmosphere of 5% CO_2. The human normal colon epithelial cell line CCD 841 CoN (ATCC, CRL-1807) and the human normal colon fibroblast cells CCD 112 CoN (ATCC CRL-1541) were grown in Eagle's Minimal Essential Medium (EMEM) with 2 mM glutamine supplemented with 10% fetal calf serum. The human gastric carcinoma cells MKN-45 (DSMZ, ACC-409) was grown in Roswell Park Memorial Institute (RPMI) medium supplemented with 2 mM glutamine supplemented with 20% fetal calf serum. The cell lines were obtained by the ATCC (LGC Standards SARL, Molsheim, France).

2.3. Transient transfections

To assay Luciferase activity, HeLa, HCT-116, CCD 841 CoN and MKN-45 cells (70% confluency) were transfected using lipofectamine (Thermo Fisher Scientific Bioscience, Villebon sur-Yvette, France).) according to the manufacturer's instructions, with 1.5 μg of pGL3 constructions and 20 ng of control *Renilla* plasmid in Ultra MEM medium. After 6 h, the medium was replaced by fresh culture medium and further incubated for 48 h. Cells were then washed with Phosphate Buffered Saline (PBS), lysed with 80 μl of Passive Lysis Buffer (PLB, Dual Luciferase Reporter Assay System, Promega, Madison, USA) and 20 μl of lysate were used to measure the luciferase activity. Luminescence was measured with the Centro luminometer (Berthold Technologies, Bad Wildbad, Germany).

For transient transfection with small interfering RNA (siRNA), oligonucleotide duplexes (ON-TARGETplus SMARTpool; Horizon Discovery; Cambridge; UK) targeting ETS1, DMTF1, SP1 and a nontargeted oligonucleotide duplex (ON-TARGET plus siCONTROL; Horizon Discovery; Cambridge; UK) as a negative control, were used in the transfection experiments with Lipofectamine 2000 Transfection Reagent (Thermo Fisher Scientific; Waltham; MA; USA). Briefly, CCD 841 CoN cells were grown to 80% confluence. ETS1, DMTF1 and SP1 siRNAs (200 pmol) and control siRNA (200 pmol) were incubated with the transfection reagent at room temperature for 25 min, to form complexes, which then were added to 6-well plates containing cells and medium.

After 4 h in a 5% CO_2 atmosphere at 37 °C, the delivery mix was replaced with EMEM supplemented with 10% fetal bovine serum (FBS). The following day, the cultures were split and the delivery mix containing 100 pmol siRNAs was re-applied. The cells were harvested 24 h later and used for further analysis.

To overexpress ETS1 and DMTF1, the full-length sequence of ETS1 and DMTF1 cDNA was cloned into the pcDNA3.1 vector (pcDNA3.1-ETS1 kindly provided by Dr. M. Aumercier UMR CNRS 8576, Villeneuve d'Ascq and pcDNA3.1-DMTF1α kindly provided by Dr. B. Torbett, Scripps Research Institute, La Jolla, CA). An empty pcDNA3.1 vector was used as a negative control. To overexpress SP1, the full-length sequence of SP1 cDNA was cloned into the pCMV4 vector (pCMV4-Sp1 [17].

CCD 841 CoN cells were cultured to 80% confluence and then transfected with the various vectors (2 μg) using Lipofectamine 2000 Transfection Reagent (Thermo Fisher Scientific; Waltham; MA; USA) according to the manufacturer's instructions. After 24 h, the cells were harvested and used for further analysis.

2.4. Preparation of cell lysate and electrophoretic mobility shift assay (EMSA)

EMSAs were performed with the Light Shift Chemiluminescent EMSA Kit (Thermo Scientific, Illkirch, France) following the manufacturer's recommendations, but the binding reaction volume was scaled down to 10 μl. In the case of G/C-rich sequences, the binding reactions contained 50 mM KCl, 5% glycerol, 5 ng/μl of poly(dI·dC), and no detergent NP-40 or $MgCl_2$. Nuclear extracts were prepared using the NE-PER extraction kit (Thermo Fisher Scientific Bioscience, Villebon sur-Yvette, France), and 1 μl of extract (3–4 μg of protein) was added to each binding reaction. DNA probes were biotinylated at the 5′-end (Eurogentec, Belgium) (Supplemental Table 4). For the control supershift assay, the anti-SP1 (PEP2; sc-59) polyclonal antibody was added in the binding reaction before addition of labeled DNA probes. DNA/protein complexes were separated by 6% native PAGE, transferred to nylon membrane, cross-linked under UV light, and detected according to the kit protocol.

2.5. Chromatin immunoprecipitation ChIP assay

Chromatin immunoprecipitation assays (ChIP) were achieved basically as previously described [18]. Briefly, CCD 841 CoN and HCT-116 cells (1×10^6 cells per antibody) were treated with 1% (v/v) formaldehyde for 10 min at room temperature and cross-links were quenched with glycine at a final concentration of 0.125 M for 5 min. Cells were rapidly rinsed with ice-cold Dulbecco-PBS (D-PBS) containing a cocktail of protease inhibitors (Roche Diagnostics, Penzberg, Germany) and scraped off and collected by centrifugation at 700 ×*g* for 5 min at 4 °C, before being resuspended in lysis buffer (10 mM Hepes-KOH, pH 7.9, 10 mM KCl, 1.5 mM $MgCl_2$ and 0.1% Nonidet P40) containing protease inhibitors and incubated for 10 min on ice. Chromatin was sheared with the Bioruptor system (Bioruptor® Plus, Diagenode, Seraing, Belgium). The extracts were sonicated for 10 pulses of 30 s each with a 30 s rest between each pulse at 200 W at 4 °C. After clearing by centrifugation at 10,000 ×*g* for 10 min at 4 °C, the supernatant was fractionated and precipitated with either 3 μg of the specific antibody anti-ETS1 (C-20, sc-350), anti-DMTF1 (S-19, sc-6552) or the anti-SP1 at 4 μg (PEP2, sc-59) or normal goat IgGs (Sc-2028) from Santa Cruz biotechnology (Santa Cruz biotechnology, Dallas, Texas, USA) and normal Rabbit IgG at 4 μg (12–370, Millipore, Burlington, Massachusetts, United-States). An aliquot of the total supernatant was removed as input control. Immunoprecipitation was performed overnight on a rotating platform at 4 °C, a Protein A/G magnetic bead mix (Invitrogen) was then added and left for another 3 h as previously described [19]. Magnetic beads were collected and washed 5 times sequentially in Low Salt Immune Complex wash buffer (20 mM Tris/HCl (pH 8.0), 150 mM NaCl, 2 mM EDTA, 0.1% SDS and 1% Triton X-100), High Salt buffer (20 mM Tris/HCl (pH 8.0), 500 mM NaCl, 2 mM EDTA, 0.1% SDS and 1% Triton X-100) and TE buffer (10 mM Tris/HCl (pH 8.0) and 1 mM EDTA). The complexes were eluted with 210 μl of elution buffer [0.05 M Tris/HCl (pH 8.0), 10 mM EDTA and 1% SDS] after a 15 min incubation at 65 °C. Formaldehyde cross-links were reversed with 0.2 M NaCl at 65 °C overnight. Chromatin-associated proteins were digested with Proteinase K at 37 °C for 1 h and the DNA was purified with the Wizard® DNA Clean-up (Promega, Madison, USA). Samples were then subjected to qPCR analysis as described previously using ChIP primers pairs (Supplemental Table 5). ChiP-qPCR data were normalized using the percent input method where Ct values obtained from the ChIP samples are divided by signals obtained from the adjusted input signal. The percentage of input was therefore calculated using this formula: 100*2^(adjusted input - Ct(IP). The percent input from relevant antibodies was then compared to control IgG.

2.6. RNA extraction, cDNA synthesis, determination of transcription start site in cells and biopsies (5'-RACE) and Q-PCR

The Rapid Amplification of cDNA 5′ ends (5′-RACE) was performed with the First Choice® RLM-RACE kit (Ambion) according to the protocol provided by the manufacturer. Briefly, 10 μg of total ARN were treated with Calf Intestine Alkaline Phosphatase (CIP) and then with Tobacco Acid Pyrophosphatase (TAP). A 45 base RNA Adapter oligonucleotide (Supplemental Table 1) was ligated to the RNA using T4 RNA ligase. A reverse transcription was performed with random decamers. After synthesis of the first strand cDNA, nested PCR was performed using AccuTaq TMLA DNA polymerase (Sigma-Aldrich, Saint Quentin Fallavier, France). 5'-RACE Outer Primer/5'RACE B4GALNT2 Outer Primer and 5'RACE Inner Primer/5'RACE B4GALNT2 Inner Primer pairs were used for a first and second PCR round, respectively. PCR products were size-separated by agarose gel electrophoresis, subcloned into pCR2.1-TOPO vector (Invitrogen) and sequenced by GATC Biotech (Köln, Germany).

Gene expression was analyzed by quantitative PCR (qPCR) using the Mx3005p Quantitative System (Stratagene, La Jolla, CA, USA). PCR reaction (25 μL) contained 12.5 μL of the 2× Brilliant SYBR Green qPCR Mastermix (Thermo Fischer Scientific, Rockford, USA), 300 nM of primers and 4 μL of cDNA (dilution 1:40). Thermal cycling profiles analysis of amplification performed using Mx3005p software. All experiments were performed in triplicate using three different biological samples. The quantification was achieved by the method described by Pfaffl [20].

2.7. Bioinformatic analyses: genomatix and genetic regulatory network of transcription and pathway crosstalk

In silico analysis of the promoter was performed with BLAST analysis of the human genome of the NCBI database. The core promoter sequence was analyzed with MatInspector 8.0 (www.genomatix.de) using TRANSFAC 8.4 matrices with "core similarity 0.95" and "matrice similarity: optimized".

For the gene regulatory network, a list of 45 genes including the three transcription factors ETS1, SP1 and DMTF1 and 42 glyco-genes involved in *N*-glycan, *O*-glycan and glycolipid biosynthesis pathways (list available in supplementary data) was first created. The master regulators of these 45 genes were thus searched with the iRegulon v1.3 app [21] for the Cytoscape network visualization and analysis software v3.8.2 [22,23] in the 500 bp upstream the transcription start site (results and parameters available in supplemental data). For the creation of the regulatory network, a minimum NES score = 2 was considered for the identification of targets of the three input transcription factors and a NES score = 3 was used for the identification of master regulators that target *B4GALNT2*. The lowest threshold was used in order to maximize the possible knowledge of targets of the input transcription factors whereas the highest was used in order to limit the number of other

regulators and keep the network readable. The regulations described in this article were added to the network. The glyco-genes that were identified as co-differentially expressed with *B4GALNT2* in the work of Pucci et al. [4] were annotated in the network as elliptic nodes. Self-loops were removed.

COlorectal ADenocarcinoma (COAD) data of The Cancer Genome Atlas (TCGA) database were used through the UALCAN webtool [24]. Individual expression levels of *B4GALNT2*, *SP1*, *DMTF1* and *ETS1* for normal ($n = 41$) and primary tumor ($n = 286$) samples were extracted and represented as box-whisker plots in Fig. S3A. Heatmaps of each gene of the regulatory network except *MUC5B* (data not available in UALCAN) were extracted as well and depicted in Fig. S3B. The same data were used and $\log_2$ fold change of medians of primary tumor vs normal conditions were computed for each gene of the regulatory network and integrated as a gradient from blue ($\log_2FC \leq -3$) to white ($\log_2FC = 0$) and red ($\log_2FC \geq 3$) in Fig. 8; only signals with statistical significance ≤ 0.05 were colored, others are white. For correlation analysis data, gene expression of *B4GALNT2*, *ETS1*, *SP1* and *DMTF1* were extracted from The Cancer Genome Atlas (TCGA COADREAD, https://cancergenome.nih.gov/) for 524 colorectal cancer patients using the cBioPortal website. Transcript expression values (z-scores) relative to normal samples were retrieved as RSEM (Batch normalized from Illumina HiSeq_RNASeqV2). For expression analysis in normal samples, *B4GALNT2*, *ETS1*, *SP1* and *DMTF1* gene expression profiles were downloaded from the NCBI Gene Expression Omnibus (GEO) database with accession number GSE44076 (https://www.ncbi.nlm.nih.gov/geo/). GSE44076 is a dataset containing gene expression profiles from the Colonomics project that includes expression data of paired normal adjacent mucosa and colorectal tumor samples from 98 individuals and 50 healthy colon mucosae from healthy donors. Prior to analysis, probe identification in the GPL13667 platform (Affymetrix Human Genome U219 Arrays) were converted into standard gene symbols. For genes with more than one probe set in the array, average values were used. TCGA and Colonomics data analyses were processed using R studio (https://rstudio.com/). Pearson's R correlation coefficient was used to assess relationships between transcript expression levels of *B4GALNT2* and *ETS1*, *SP1* and *DMTF1* into the two data series. The r values and p values were calculated for each combination of genes in all the samples of the TCGA COADREAD dataset and in healthy colon mucosa samples in the Colonomics dataset. $p < 0.05$ was considered as statistically significant. Heat maps were generated to visualize gene expression data of the four genes of interest across all Colonomics samples. Expression levels were standardized (centered and scaled) within rows for visualization.

2.8. Statistical analyses

Statistical analyses were performed with Graphpad Prism software version 4.0 (Graphpad softwares Inc., La Jolla, CA, USA) using the ANOVA test. Results were deemed significant for p values less than 0.05 ($p < 0.05$). * $p < 0.05$; ** $p < 0.01$ *** $p < 0.001$.

3. Results

3.1. Identification of the B4GALNT2 gene transcriptional start sites (TSS)

We have previously reported the existence of at least two TSS and two AFE giving rise to two transcript variants (Short Form (SF) and Long Form (LF)) in the colon cancer cell line Caco-2 [8] and that the CCD 841 CoN and in CCD 112 CoN cell lines as well as healthy colon biopsy expressed the highest level of *B4GALNT2* transcripts [9]. To locate the transcriptional start sites in healthy colon, the 5′-UTR of *B4GALNT2* transcripts were analyzed using 5′-RACE in the 2 cell lines and in healthy colon as described in the Material and Methods section. Different 5′-RACE products were observed on agarose gel (Fig. 1A) and their sequencing demonstrated the existence of multiple TSS and AFE in healthy colon. The majority of the amplified sequences from CCD 841 CoN cells and healthy colon biopsy were of the SF-type indicating that *B4GALNT2* transcription primarily started within exon 1S. Two different TSS were positioned at −95 and -24 nt relative to the first in frame start codon ATG in exon 1S designated as (+1) (Fig. 1B, transcripts 1, 2). Another transcript of the LF-type with a unique TSS located −154 nt upstream the start codon in Exon 1 L was also detected in the cell lines (Fig. 1B, transcript 3) and more specifically in the CCD 112 CoN cells. Interestingly, a few sequences corresponding to a transcript of the middle form (MF)-type with a TSS located −81 nt upstream the start codon in Exon 2 (Fig. 1B, transcript 4) were detected in the healthy colon biopsy. These data established the existence of three different sets

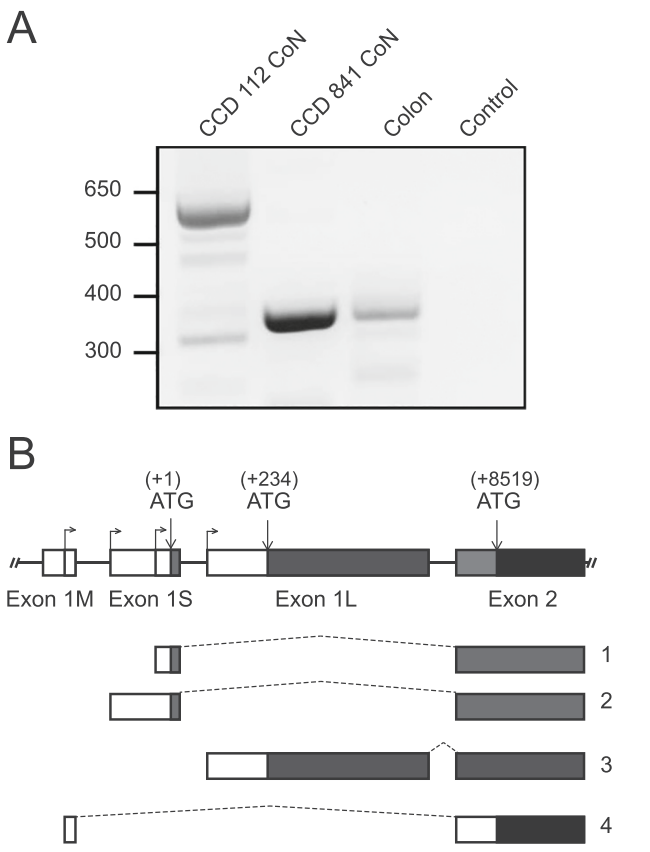

Fig. 1. Mapping the transcriptional start sites (TSS) of the human *B4GALNT2* gene in colon cells using 5'-RACE. A) Agarose gel electrophoresis of 5'-RACE amplification products shows products at approximately 600 bp primarily in CCD 112 CoN cells and barely detected in CCD 841 CoN cells. This corresponds to a transcript of the LF-type with a unique TSS located −154 nt upstream the start codon in Exon 1 L corresponding to transcript 3 depicted below. Less intense bands seen below were also extracted and sequenced and found to correspond to incomplete synthesis by the reverse transcriptase. Amplification products visualized at approximately 340 bp in CCD 841 CoN cells and colon sample were mostly of the SF-type transcript variant and correspond to transcripts 1 and 2 shown below. These two transcripts correspond to two different TSS positioned at −95 and -24 nt relative to the first in frame start codon ATG in exon 1S, designated as (+1). Interestingly, these amplification products detected in the healthy colon biopsy also include a few sequences corresponding to a transcript of the MF-type with a TSS located −81 nt upstream the start codon in Exon 2, corresponding to transcript 5 described below. A negative control of the 5′-RACE with no ARN is shown on the right side of the gel. B) Schematic representation of the 5′-flancking region of *B4GALNT2* gene and four transcripts variants (1–4) found in colonic cells. This scheme indicates the differences in the 5′UTR of *B4GALNT2* transcripts in colon. The bent arrows indicate the various transcription start positions. The arrows indicate the three ATG in each alternative first exon (AFE) 1S and 1L and in exon 2 and their position relative to the first methionine codon is indicated above.

of TSS located on three different AFE of the *B4GALNT2* gene in healthy colon and normal colon cell lines corroborating previous observations and confirming the predominant expression of SF-type transcripts [9].

3.2. Delineation of the B4GALNT2 promoter region

We next sought to determine the promoter region of the *B4GALNT2* gene. For that purpose, the genomic sequence located between −1559 bp and + 675 bp encompassing the three AFE was cloned into the promoter-less pGL3Basic upstream the firefly luciferase gene to generate the pGL3(−1559/+675) reporter plasmid (Fig. 2A). The normal colon CCD 841 CoN cells were chosen for initial transient transfections since they express both the SF- and LF-type transcript variants of *B4GALNT2* and the colon cancer HCT-116 because they do not express *B4GALNT2* [9]. In addition, the gastric carcinoma MKN-45 cells and cervix cancer HeLa cells which express very low levels or no B4GALNT2 [9] were also used as control recipient cells (Fig. 2). However, not much luciferase activity could be detected for the pGL3(−1559/+675) plasmid (Fig. 2B). Therefore, several 5′- or 3′-deleted constructs were made that were transfected into the different cell types for subsequent luciferase assays. The results presented in Fig. 2B show that all constructs lacking the +72/+675 genomic region exhibit increased activities compared to the pGL3Basic plasmid used as a reference. Maximal promoter activities were obtained for pGL3(−83/+72) construct, which proved to be the most active whatever the cell line used for transfection (17.7-, 15-, 14.8- and 7.1-fold increased activity into CCD 841 CoN, HCT 116, HeLa and MKN-45, respectively) (Fig. 2B). These first experiments enabled us to delineate a *B4GALNT2* promoter region (−83/+72) containing the positive regulatory elements responsible for constitutive activity regardless of the cell type.

3.3. Characterization of the proximal promoter region of B4GALNT2

The genomic region −277/+236 of the human *B4GALNT2* gene encompassing exon 1S and part of exon 1 L and the putative promoter region −83/+72 delineated above was analyzed with the MatInspector 8.0 program (Genomatix) with core similarity of 0.95 and an optimized matrix similarity. This bioinformatic analysis did not identify TATA- nor CAAT-box in the delineated *B4GALNT2* promoter, but several GC-boxes were observed. These features and the heterogeneous TSS identified for *B4GALNT2* transcripts in colon, are structural features typical of the so-called housekeeping genes and most glycosyltransferase genes described up to now. However, it led to the identification of several putative transcriptional factor binding sites (Supplemental Table 6 in supplemental data) including those for factors belonging to the specificity protein (SP) transcription factor family like SP1 (Stimulating protein 1) and for transcription factors involved in various cellular processes like cell proliferation, cell cycle or growth such as DMTF1 (Cyclin D binding myb-like protein 1) and ETS1 (E26-transformation specific), illustrated in Fig. 3A. Three SP1 binding sites (consensus sequence G/TGGGCGGG/AG/AG/T) were found at positions -72/-63 and -18/-9 upstream the SF-variant start codon and at +8/+17, upstream the LF-variant start codon and one DMTF1 and/or ETS binding site (consensus sequence CCCG(G/T)ATGT) [25] was found at -33/-25 (Fig. 3A).

We then assessed the possibility of two promoters driving the expression of the two major types of *B4GALNT2* transcripts (SF- and LF-variants). For that purpose, we examined in more details the promoter activity of the genomic regions upstream the two major AFE and generated additional constructs spanning ~250 bp of each genomic region. The pGL3(+1/+234), pGL3(+14/+234) and pGL3(+1/+72) constructs including genomic sequences upstream the LF-translational start site were transfected into CCD 841 CoN or HCT-116 cells and

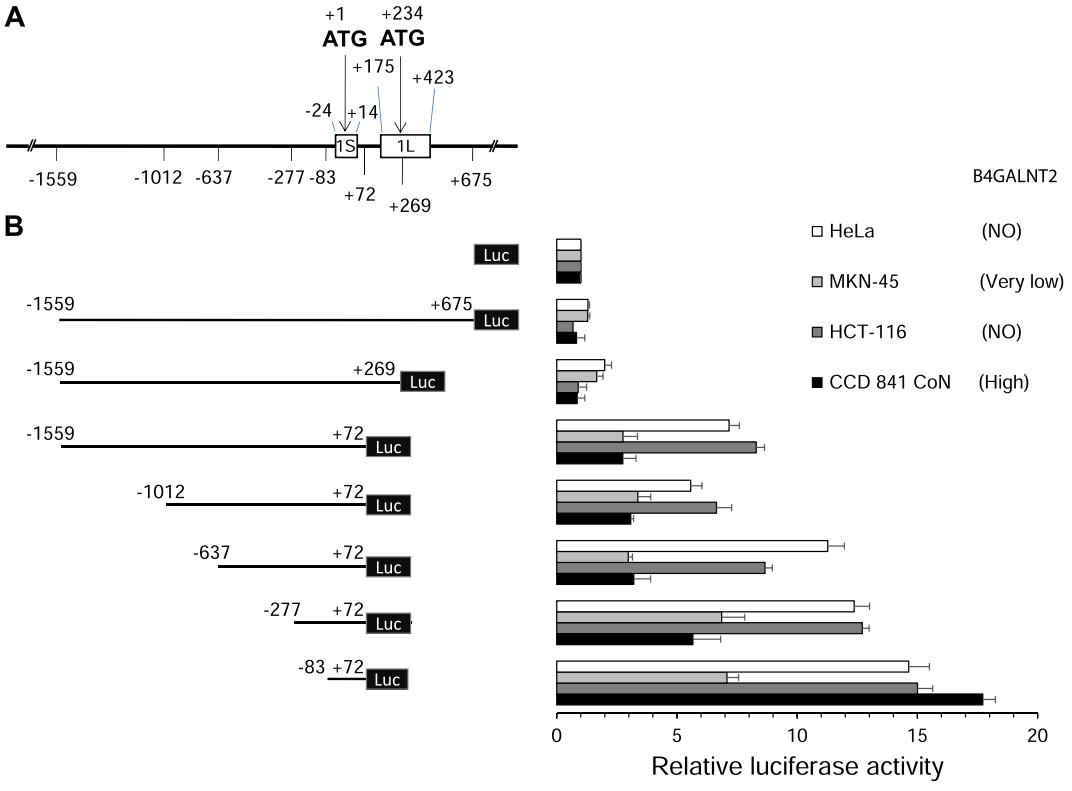

Fig. 2. Promoter activity of serial deletion constructs of the human *B4GALNT2* gene. A) Schematic representation of the cloned *B4GALNT2* genomic region (−1559/+675 nt region) cloned in the luciferase (LUC) reporter plasmid. Nucleotides are numbered relative to the first base of the methionine codon of the short transcript variant and B) luciferase activities of the *B4GALNT2* deletion mutants relative to the mock (pGL3Basic) in four cell lines CCD 841 CoN, HCT-116, MNK-45 and HeLa cells. The experiments were carried out in triplicates and data are the mean +/− S.D.

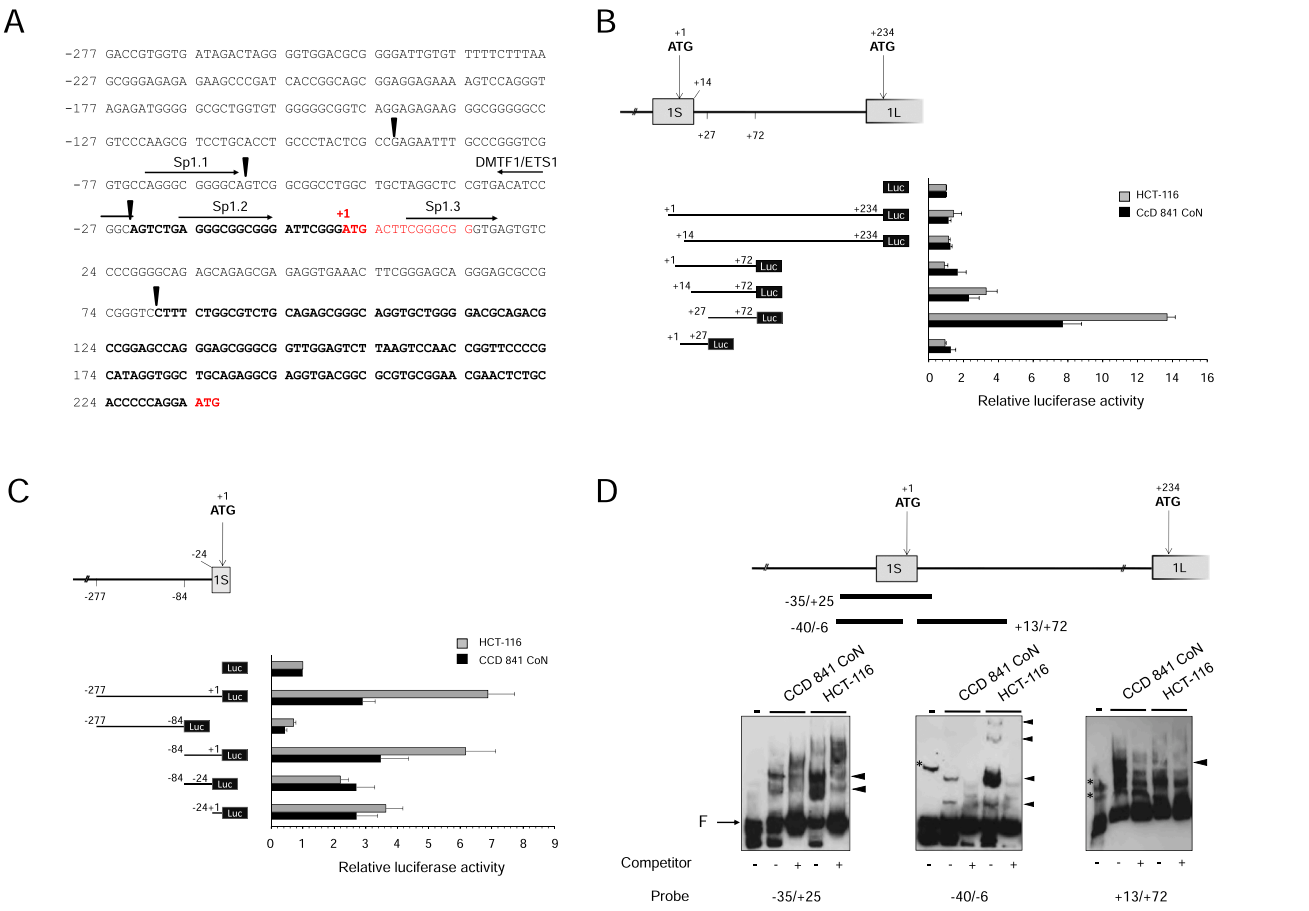

Fig. 3. Nucleotide sequence of the 5′-flanking region of the *B4GALNT2* gene and promoter activities A) Genomatix analysis of the −277/+234 nt genomic region. Consensus TF binding motifs for SP1, ETS1, DMTF1 were identified in the 5′-flanking region of the *B4GALNT2* gene. Nucleotides are numbered relative to the first base of the methionine codon of the short transcript variant. The 5′-untranslated region (5'-UTR) of each transcript variant is indicated with bold characters and the translated regions are indicated in red. Arrowheads indicate the location of transcriptional sites. B) Promoter activity of the long *B4GALNT2* transcript promoter (+1/+234). A schematic illustrating this *B4GALNT2* gene region including the two alternative first exons (AFE) 1S and 1 L is given above the representation of promoter deletion constructs corresponding to the long *B4GALNT2* transcript variant cloned in the luciferase (LUC) reporter vector. Their corresponding luciferase activities (relative to the mock pGL3Basic) in CCD 841 CoN and HCT-116 cells are represented on the right. C) Promoter activity of the short *B4GALNT2* transcript promoter region (−277/+1). A schematic illustrating this *B4GALNT2* gene region upstream of the first exon 1S is given above the representation of the promoter deletion constructs corresponding to the short *B4GALNT2* transcript variant cloned in the luciferase (LUC) reporter vector. Their corresponding luciferase activities (relative to the mock pGL3Basic) in CCD 841 CoN and HCT-116 cells are represented on the right. D) Electrophoresis mobility shift assay (EMSA) of the *B4GALNT2* promoter region. The three 5′-biotinyated-labeled probes (−35/+25, −40/−6 and + 13/+72) containing SP1-binding sites 2 and 3, the ETS1-binding site and the DMTF1-binding site are positioned below the *B4GALNT2* gene in the region corresponding to the promoter of the short transcript. The labeled probes were incubated with either CCD 841 CoN or HCT-116 nuclear extracts. For each panel, lane 1 correspond to the negative control with no nuclear extract. The free labeled probe (F) is indicated by an arrow on the left side and the specific DNA-protein complexes are pointed by arrowheads on the right side of the gel. (*) indicate non-specific complexes. (−) indicate no competitor and (+) indicate the presence of an excess of the unlabeled probe competing with the formation of the specific DNA-protein complexes.

showed almost no luciferase activity compared to the pGL3 basic plasmid (Fig. 3B). The 3′-deleted construct pGL3(+14/+72) showed increased luciferase activity. Interestingly the pGL3(+27/+72) construct showed the highest luciferase activity, whereas the pGL3(+1/+27) had no activity. These results further suggested the presence of *cis*-acting motifs enhancing promoter activity in the +27/+72 region and the presence of a negative regulatory element within the +1/+27 region. After deletion of the −277/−84 region, the luciferase activity of pGL3(−84/+1) construct is comparable to the luciferase activity of pGL3(−277/+1), suggesting that a core promoter region for the *B4GALNT2* gene is located within the sequence −84/+1 upstream of the exon 1S. The pGL3(−277/+1) construct including genomic sequences upstream the SF-translational start site transfected into CCD 841 CoN or HCT-116 cells showed the highest activity (Fig. 3C). The pGL3(−84/+1) showed similar level of activity compared to the pGL3(−277/+1) whereas the 3′-deleted construct pGL3(−277/−84) showed less or no activity compared to pGL3basic. These data further suggested that the *cis*-acting elements enhancing promoter activity of the SF-variant are located in the −84/+1 region. These conclusions are supported by synergistic activity of the pGL3(−84/−24) and pGL3(−24/+1) constructs. Since the major form expressed in healthy colon is the SF-type transcript variant [9], we next focused on the regulation of this variant.

Gel electrophoresis mobility shift assay (EMSA) was used to detect protein complexes with nucleic acids in the various regions of interest (Fig. 3D). EMSA experiments carried out using CCD 841 CoN and HCT-116 nuclear extracts and the 5′-biotinyated −35/+25 probe, which encompasses the DMTF1, ETS1, Sp1.2 and Sp1.3 binding sites revealed two major complexes (left panel, Fig. 3D). These complexes were competed with 200-fold excess of unlabeled and specific consensus sequence oligonucleotides for competition and were more detectable with the protein extracted from the HCT-116 cells. Supershift assays using antibody against SP1 and CCD 841 CoN nuclear extracts confirmed implication of this TF (Supplemental Fig. S1). The potential implication of the DMTF1, ETS1 and Sp1.2 binding sites positioned upstream the

AFE 1S was confirmed using the 5′-biotinyated −40/−6 probe, with the formation of four protein-DNA complexes (middle panel, Fig. 3D). One complex was equally detectable with the protein extracted from the CCD-841 CoN and HCT 116 cells, one was more detectable with the protein extracted from the HCT-116 cells and two complexes were detectable only with the protein extracted from the HCT-116 cells. Finally, the 5′-biotinyated +13/+72 probe containing no major binding sites upstream the AFE 1 L revealed only minor complexes with proteins extracted from HCT-116 cells (right panel, Fig. 3D). The other complexes were not outcompeted by the unlabeled probe and therefore are considered as non-specific complexes. Taken together, these data indicate the presence of several specific complexes, that are more readily detectable with proteins extracted from the HCT-116 cells, within the −84/+1 core promoter region. In addition, one complex more detectable with proteins extracted from CCD 841 CoN cells was identified within the +27/+72 region.

3.4. Contribution of ETS1, SP1 and DMTF1 elements to human *B4GALNT2* promoter activity

To get insights into the role of each identified TF binding site and further define the specific elements that contribute to the proximal promoter activity, we performed site-directed mutagenesis. We generated a series of *B4GALNT2* promoter constructs with mutations in the ETS1, DMTF1 and SP1 elements of the pGL3 (−83/+72) construct, which served as a reference. DMTF1 and ETS1 share the same binding site GC**CGGAT**GT in the *B4GALNT2* gene (Fig. 2). However, Hirai and Sherr reported that CC**CGTAT**GT could specifically bind DMTF1 whereas CC**CGGAA**GT binds specifically ETS1 [25]. Therefore, the AC**TTCCG**GC (ETS1* only) and the AC**ATACG**GC (DMTF1* only) mutations were brought in the pGL3(−83/+72) plasmid to abrogate DMTF1 and ETS1 binding, respectively and the AC**ATCTA**GC mutation to abrogate binding of both TFs. The various mutated pGL3(−83/+72) constructs were transfected in CCD 841 CoN and HCT-116, as previously described. Interestingly, mutation of the ETS1 binding site where only DMTF1 can bind did not seem to have an impact on the promoter activity, but the combined mutation of the ETS1 and DMTF1 sites led to relevant 61 to 66% promoter inhibition in CCD 841 CoN and HCT-116, respectively, (Fig. 4). Finally, mutation of the DMTF1 binding site, where only ETS1 can bind increased significantly *B4GALNT2* promoter activity (73–128% in CCD 841 CoN and HCT-116, respectively) (Fig. 4). Mutation of the SP1.1 binding site resulted in a significant decrease of luciferase activity (~16–40% in HCT-116 and CCD 841 CoN), whereas mutations of SP1.2 and SP1.3 binding sites led to a significant increase of luciferase activity (~27–34% for SP1.2 in CCD 841 CoN and HCT-116 and 18–41% SP1.3 in CCD 841 CoN and HCT-116) (Fig. 4). These data further suggest the implication of SP1 (at the SP1.1 binding site) and a combination of ETS1 and DMTF1 TFs to enhance the *B4GALNT2* promoter activity.

3.5. Identification of transcription factors essential for regulation of *B4GALNT2* gene

Firstly, we checked endogenous ETS1, SP1 and DMTF1 TF expression in HCT-116 and CCD 841 CoN cell lines by RT-qPCR: all TF were expressed in both cell lines, with a higher expression of DMTF1 in both cell types (Supplemental Fig. S2A). The overexpression of ETS1, SP1, DMTF1 was achieved through transient transfections of expression vectors in CCD 841 CoN cells, with a resulting 50 to 100-fold increase of TF expression levels (Supplemental Fig. S2B). Our RT-qPCR analysis showed that the overexpression of ETS1 and DMTF1 increased the expression of the SF *B4GALNT2* transcript (50% and 25% increase, respectively), whereas SP1 overexpression had no effect (Fig. 5A). The same experiments were carried out in HCT-116 cells, which do not express endogenous *B4GALNT2* [9]. However, overexpression of these TFs did not induce *B4GALNT2* expression (data not shown) further suggesting more complex regulatory mechanisms. Additionally, the inhibition of ETS1, SP1, DMTF1 expression was performed using siRNA expression in CCD 841 CoN cells (Supplemental Fig. S2C), which resulted in decreased expression of SF *B4GALNT2* transcript for the three TFs of about 70% for SP1 and 50% for both ETS1 and DMTF1 (Fig. 5B). These results suggest that ETS1, SP1 and DMTF1 can modulate *B4GALNT2* gene expression in normal colon cells.

Secondly, to confirm direct involvement of these TF, Chromatin Immuno-Precipitation (ChIP) experiments were carried out. After checking the relative expression of the three TF in CCD 841 CoN using qPCR (Supplemental Fig. 2A), chromatin from CCD 841 CoN was immunoprecipitated with either anti-ETS1, anti-SP1 or anti-DMTF1 antibody and isotype IgG as controls. DNA was isolated from each immunoprecipitate and the ETS1, SP1 and DMTF1 binding regions in *B4GALNT2* promoter region were amplified by PCR. Five experiments were conducted and representative results are shown in Fig. 6B. Our data indicate that ETS1 and SP1 transcription factors bind to *B4GALNT2* promoter in CCD 841 CoN cells. However, no direct binding of DMTF1 to the promoter could be detected.

4. Discussion

The human *B4GALNT2* gene is expressed under various biological conditions and most abundantly in healthy colon. However, its expression is dramatically downregulated in colon cancer [9,26–28] leading to disappearance of the Sd^a determinant and a concomitant increased expression of sLe^x structures on glycoproteins [9,11]. Interestingly, recent analysis of transcriptomic data from The Cancer Genome Atlas

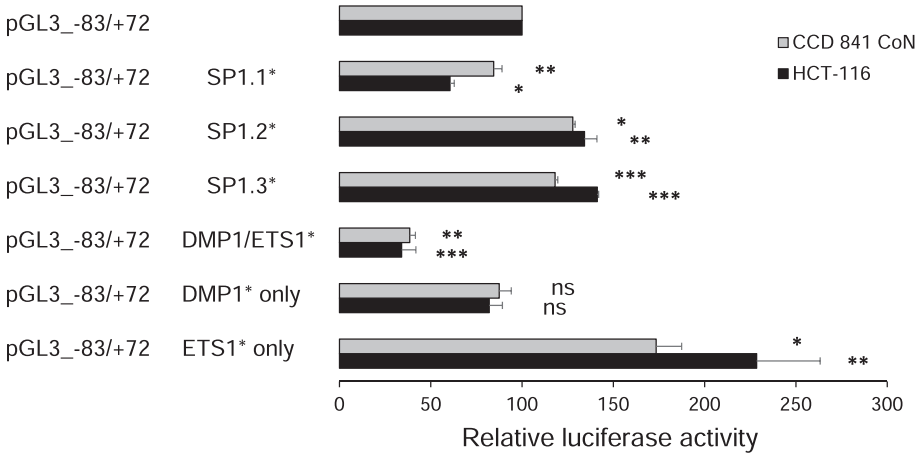

Fig. 4. Mutation analysis of the *B4GALNT2* promoter activity. The *B4GALNT2* transcript promoter region (−83/+72) cloned into the luciferase (LUC) reporter vector was used to generate mutated promoter fragments by site-directed mutagenesis of each TF binding sites SP1 (SP1.1*; SP1.2*; SP1.3*) and both ETS1 and DMTF1 (DMTF1/ETS1*) binding site; DMTF1*only and ETS1* only means that only DMTF1 and ETS1 can bind to the ETS1 and DMTF1 mutated binding site, respectively. Transient transfections of these various plasmids were achieved in CCD 841 CoN and in HCT-116 cells and Luciferase assays were performed. Their corresponding luciferase activities (relative to the pGL3(−83/+72)) are represented on the right. Data represent the mean +/− S.D. of three independent experiments. *: $p \leq 0.05$; **: $p \leq 0.01$; ***: $p < 0.001$; ns: non-significant (vs pGL3(−83/+72)).

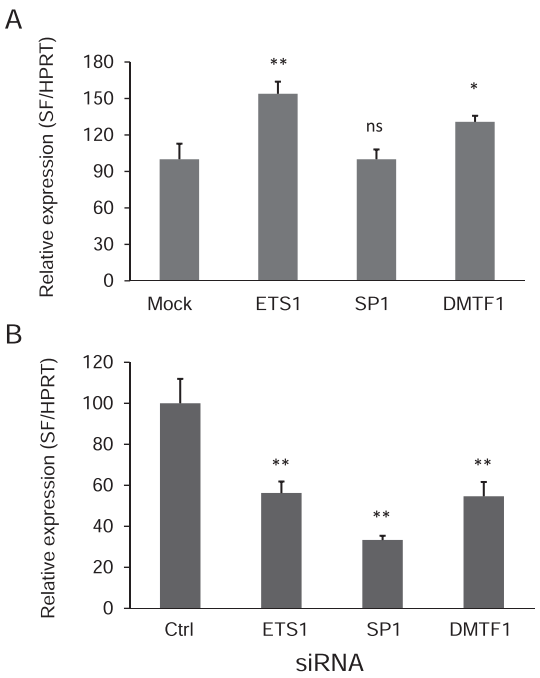

Fig. 5. Effect of ETS1, SP1 and DMTF1 overexpression (A) or silencing (B) on the relative expression of *B4GALNT2* SF transcripts in CCD-841-CoN cells. (A) CCD-841-CoN cells were transiently transfected with an empty vector (Mock) or pcDNA3-ETS1, pCMV4-SP1, pcDNA3-DMTF1 expression vector. After 48 h, total RNA was extracted and the short *B4GALNT2* transcript variants were quantified by qPCR as described previously [9]. (B) CCD-841-CoN cells were transiently transfected with control siRNA or specific siRNA targeting ETS1, SP1, or DMTF1. After 48 h, total RNA was extracted and the short *B4GALNT2* transcript variants were quantified by qPCR as described previously [9]. Data are means +/- SD of three independent experiments. *: $p \leq 0.05$; **: $p \leq 0.01$; ***: $p < 0.001$; ns: non-significant (vs Mock).

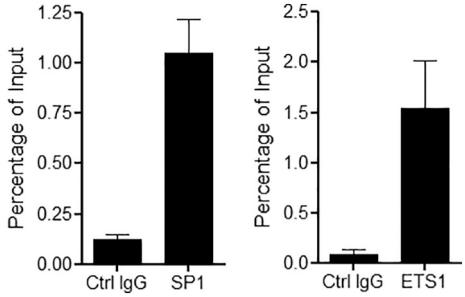

Fig. 6. Endogenous binding of SP1 and ETS1 on the *B4GALNT2* promoter in CCD 841 CoN cells. Chromatin immunoprecipitation was performed as described in the materials and methods section from CCD 841 CoN cells. Quantitative PCR were carried out with a specific pair of primers covering either the SP1 or the ETS1 binding sites (Supplemental Table 5). The percentage of input was calculated from ChIP samples obtained with irrelevant (Ctrl IgG) or specific antibodies (either anti-SP1 or anti-ETS1) normalized to the adjusted input values.

(TCGA) database uncovered that high expression of the *B4GALNT2* gene was associated to good prognosis value in colorectal cancer patients [16,29]. In addition, it has long been known that the Sd^a-glycosylation profile of the digestive tract underpins host-microbe interactions and modulates susceptibility to infectious microbes as reviewed recently [30]. Therefore, deciphering the regulatory molecular mechanism for this gene expression is of utmost interest in the field of colon cancer biology.

4.1. Three AFE and three sets of TSS for the B4GALNT2 gene

Previous studies have shown that the human *B4GALNT2* gene coding sequence is scattered over 11 exons and that it drives the expression of at least 5 distinct transcripts in the digestive tract [8]. Interestingly, the mouse *b4galnt2* gene, previously known as CT-GalNAc transferase [31] differs in its first exon and drives the expression of a unique transcript. In this study, we identified the multiple *B4GALNT2* TSS using PCR analyses, as shown to occur for several other glycosyltransferase genes including the *B4GALNT2* paralog known as *B4GALNT1* encoding the GM2/GD2 synthase [32], the murine *B4GALT1* [33,34] and the human sialyltransferase *ST3GAL5* encoding the GM3 synthase [35]. Moreover, we showed the existence of three sets of TSS located in three distinct AFE for the human *B4GALNT2* gene, further suggesting the existence of multiple promoter regions as described previously for the human *ST6GAL1* gene [36,37] an unusual feature of the mammalian genes gained during the time course of vertebrates' evolution [38]. In fact, most glycosyltransferase genes show a unique first exon with either one TSS as for the human *ST8SIA1* (e.g. GD3 synthase) [39,40] or the *B4GALT3* genes [41], or several TSS as for the β1,4-galactosyltransferase genes *B4GALT4* and *B4GALT5* [42,43], the *ABO* gene [44], the sialyltransferase *ST6GAL2* [45] and *ST6GALNAC1* genes [46] and the β1,6-*N*-acetylglucosaminyltransferase *GCNT1* gene ([47].

4.2. Delineation of the core promoter and identification of TF implicated in B4GALNT2 gene expression

The 5′-flanking region of the human *B4GALNT2* gene was cloned and characterized, and a GC-rich promoter region (−83/+72) lacking the canonical TATA and CCAAT boxes was delineated. This genomic region contained the positive regulatory elements responsible for constitutive activity, regardless of the recipient cell. Furthermore, we defined the main promoter orchestrating the expression of the short *B4GALNT2* variant and identified major *Cis*-acting regulatory elements required for accurate and efficient initiation of *B4GALNT2* transcription.

This region was analyzed by MatInspector program (Genomatix) revealing three SP1-binding sites (SP1.1 (−73/−64); SP1.2 (−18/−8); SP1.3 (+8/+17)) (Fig. 3A). Site directed mutagenesis of each site uncovered that SP1.1 could be involved in promoter activation, whereas SP1.2 and SP1.3 could be involved in promoter repression. SP1 is a well-known zinc finger TF that belongs to the Specificity Protein/Krüppel-like Factor (SP/KLF) TF family. It is widely expressed in all mammalian cells and is characterized by the highly conserved (sequence identity more than 65%) GC-rich DNA binding domain [48]. UALCAN portal analysis [24] of Colorectal Adenocarcinoma (COAD) samples from TCGA database showed no significant changes in the SP1 expression levels (Supplemental Fig. S3). However, using TCGA and Colonomics database, we were not able to show a significantly positive correlation between *B4GALNT2* and *SP1* expression ($r = 0.29$, $p = 0.04386$) in healthy mucosa samples (Fig. 7). Other TF of the SP/KLF family sharing the same DNA binding site like SP3 have comparable enhancing and inhibitory activity depending on cell context. *SP1* gene expression is regulated during development, cell differentiation, tumorigenesis and plays opposite roles in cancer [49]. The bioinformatic analysis of this region also led to the identification of ETS1 putative binding site at positions −33/−25 (Fig. 3A). ETS family members are evolutionarily conserved TF that share a highly conserved purine-rich DNA-binding motif (i.e. GGAA/T ETS domain) in the promoter/enhancer regions of many genes. ETS genes are multifaceted TFs since they have been implicated in both in cell proliferation and differentiation, and in tumor formation and metastasis and can be viewed as tumor-suppressor [50,51]. Indeed, overexpression of ETS1 in HCT-116 cells that do not express this TF (Supplemental Fig. S2) was shown to suppress tumorigenicity [52] and ETS1 levels of expression decreased slightly, but significantly (statistical significance = 8.89E-03) in the COAD primary tumor of the TCGA database (Supplemental Figs. S3A and S3B). This was

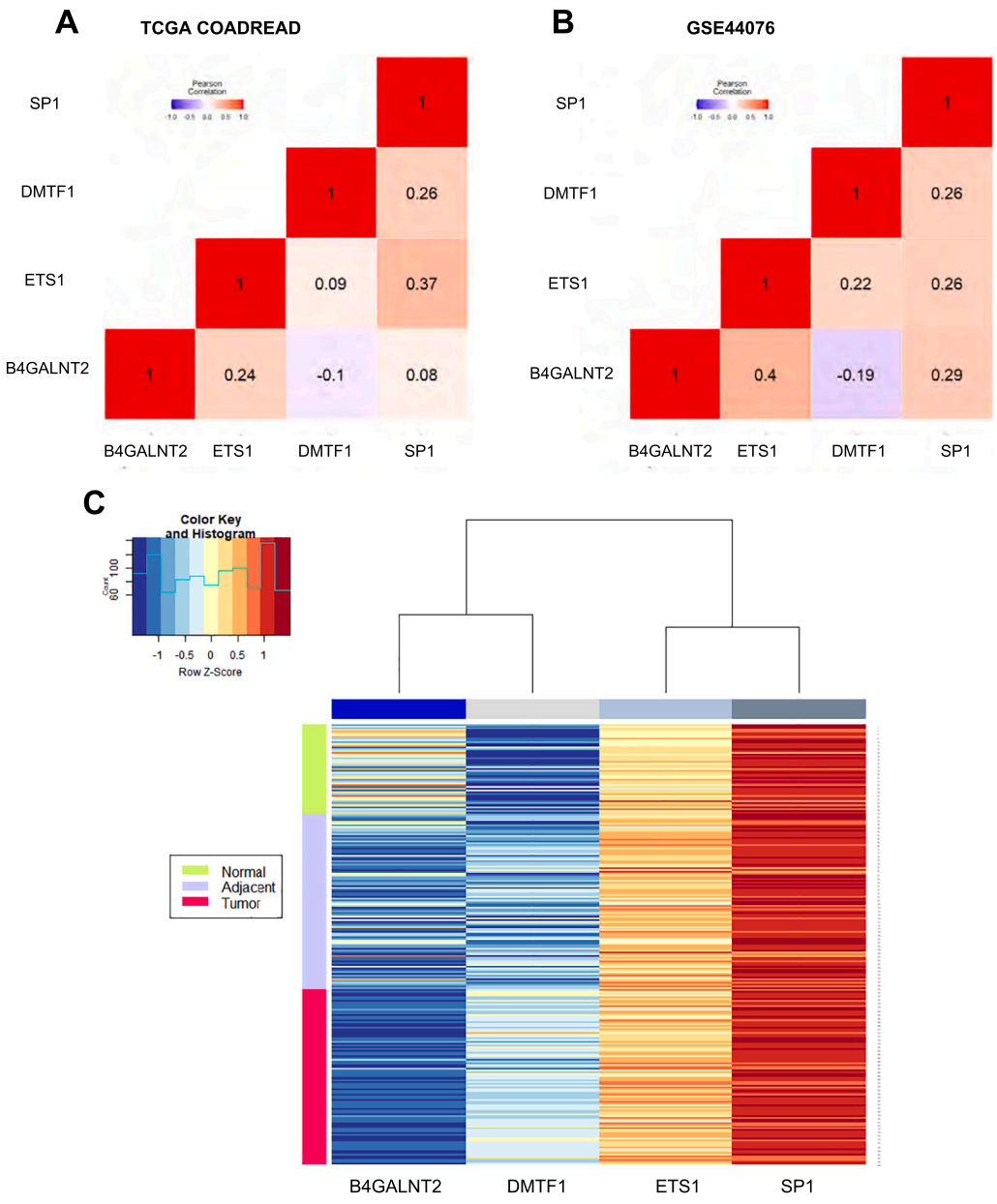

Fig. 7. Gene expression correlation between *B4GALNT2* and *ETS1*, *SP1* and *DMTF1* transcription factor encoding genes in normal and tumor patient samples. (A-B) Triangular heat maps representing gene correlation analysis in normal and cancer colon tissues. (A) Correlation Pearson r values were calculated for each transcript combination from the COADREAD cohort (n = 524). (B) Correlation Pearson r values were calculated for each transcript combination from normal tissues (healthy mucosa) of the Colonomics cohort (n = 50). Cut off values for positive or negative correlation were set at 0.2 and − 0.1, respectively. Correlation coefficient are depicted according to the color scale: the strongest positive and the strongest negative correlations are shown in red and blue, respectively. (C) Heat map showing the expression of the *B4GALNT2* gene and *ETS1*, *SP1* and *DMTF1* transcription factor encoding genes in all Colonomics samples (GSE44076). Columns of the matrix correspond to the expression of each gene of interest and rows correspond to normal (in green), adjacent (in blue) and tumor (in pink) samples. The individual tiles in the heat map are scaled by rows with a range of colors proportionate to gene expression values in each sample (row z-score). The red and blue colors in tiles reflect relatively high and low expression levels, respectively, as indicated by the scale bar.

accompanied with a positive and significant correlation between *B4GALNT2* and *ETS1* gene expression both in the TCGA COADREAD (r = 0.24, $p < 0.00001$) and healthy mucosa from Colonomics (r = 0.40, p = 0.00413) datasets (Fig. 7). Of particular interest is the observation that in many cases, the affinity of the ETS proteins to their DNA targets is increased by association with other transcription factors including, among others, AP1, SP1, CBF, C/EBP and SRF [53,54]. In this study, we identified a putative DMTF1 (cyclin D binding myb-like protein 1) binding site that overlapped the ETS1 binding site in the *B4GALNT2* promoter region (Fig. 3A). DMTF1 also known as DMP1 is a TF that was isolated in a yeast two-hybrid screen through its binding property to cyclin D2 [55] and its expression is ambiguous because shown as decreased in colorectal cancer cells (CRC) cells in [56], but shown as slightly increasing in TCGA data (supplemental Fig. S3A), which can be due to the fact that a high heterogeneity is observed for DMTF1 expression in tumorous tissues (Supplemental Fig. S3A and S3B). DMTF1 specifically binds to the DNA consensus sequences CCCG(G/T)ATGT to activate transcription. As identified in the *B4GALNT2* promoter, some DMTF1 binding sites contain a GGA core that could serve as a ETS1-responsive element [25,57]. Our mutagenesis analysis on each ETS1 and DMTF1 potential binding site showed a significant effect on the promoter activity. When DMTF1 binding site was deleted our data suggested that ETS1-binding only significantly increased the *B4GALNT2* promoter activity, whereas mutation of the ETS1 site did not change significantly its activity (Fig. 4). This observation was supported by our ChIP data carried out in CCD 841 CoN showing fixation of ETS1 (Fig. 6) but not of DMTF1 (data not shown). In accordance with these results, we were not able to identify positive correlations of expression between *B4GALNT2* and *DMTF1* in our analysis (Fig. 7). Finally, mutation of both TF binding sites led to a significant silencing of the promoter activity (Fig. 4), further suggesting that the combination of both ETS1 and DMTF1 is necessary to activate the *B4GALNT2* promoter. Further silencing and overexpression of these TF in CCD 841 CoN showed that a decreased expression of ETS1 and SP1 conducted to a decreased expression of the *B4GALNT2* SF-type transcript variant, whereas an increased expression of ETS1, likely combined to endogenous DMTF1 in CCD 841 CoN led to an increased expression of the *B4GALNT2* SF-type variant (Fig. 5).

4.3. Glyco-gene network regulated by ETS1/SP1/DMTF1/CREB3L1

Despite the fact that transcriptional regulation is recognized as a major regulator of glycan expression [13,58], very little is known about TF-mediated control of glyco-genes. In an attempt to obtain a general picture for gene regulatory network centered on *B4GALNT2*, a list of 45 genes including the three transcription factors ETS1, SP1 and DMTF1 identified in our study (Supplemental data) and 42 glyco-genes belonging to the same pathways as *B4GALNT2* and involved in the terminal biosynthetic steps of the Sd^a epitope on *N*-glycans, *O*-glycans and glycolipids [59] was created. Since the major components of the mucus layer in the gastrointestinal tract are heavily *O*-glycosylated mucin proteins like MUC2 and MUC5B [60,61] these genes were also included in the list. In addition, several glycosyltransferase genes with variable expression levels in lower and higher *B4GALNT2* expresser CRC patients [16] including *ST6GAL1*, *ST6GAL2*, *B3GNT7*, *ST6GALNAC1*, *ST6GALNAC6*, *B3GNT6* were added to the list. Their master TF regulators were then searched with the iRegulon program (see material and method section) which is based on the enrichment of binding motifs of regulatory factors. We chose the option to search in the 500 bp upstream the TSS including the regulatory region we identified in our work. A regulatory network considering only the targets of the three TFs and the master regulators of *B4GALNT2* was visualized with Cytoscape (Fig. 8). We also integrated differentially expressed signals between primary tumor and normal tissues of the COAD data from UALCAN in the network and colored the genes as a gradient from blue to white and red, accordingly. Even if not significantly differentially expressed, it mainly shows CREB3L1 (cAMP responsive element binding protein 3 like 1), a TF that represses expression of genes regulating metastasis, invasion, and angiogenesis and regulates the unfolded protein response (UPR) in the ER [62] (Fig. 8). DMTF1 was not identified as a regulator likely because of a lack of knowledge concerning this TF. However, SP1 shows a wide putative regulation in the three biosynthetic pathways, in particular in the *O*-Glycans synthesis pathway. The network also shows potential regulation by ETS1, making it a very central TF. ETS1 is slightly but significantly lowered in primary tumors and motifs for the binding of ETS1 appear in the promoter of several glyco-genes involved in *N*-Glycans and glycolipids synthesis like *B4GALT1*, which is also repressed in primary tumors. Some motifs also appear at various levels of *O*-Glycans synthesis (*ST3GAL2* and *MUC1*). However, they show different variations in their signals depicting a more advanced regulation than simple direct TF-target one. The figure shows that the regulations of *B4GALNT2* by SP1, ETS1 and potentially DMTF1 may not be limited to this gene, but likely affect the expression of other genes of these three biosynthetic pathways, like *MUC1*, eventually regulated by ETS1 and which was shown by [16] to exhibit co-expression with *B4GALNT2*, although not differentially expressed in TCGA. Interestingly, this later work highlighted other co-expressed glyco-genes like

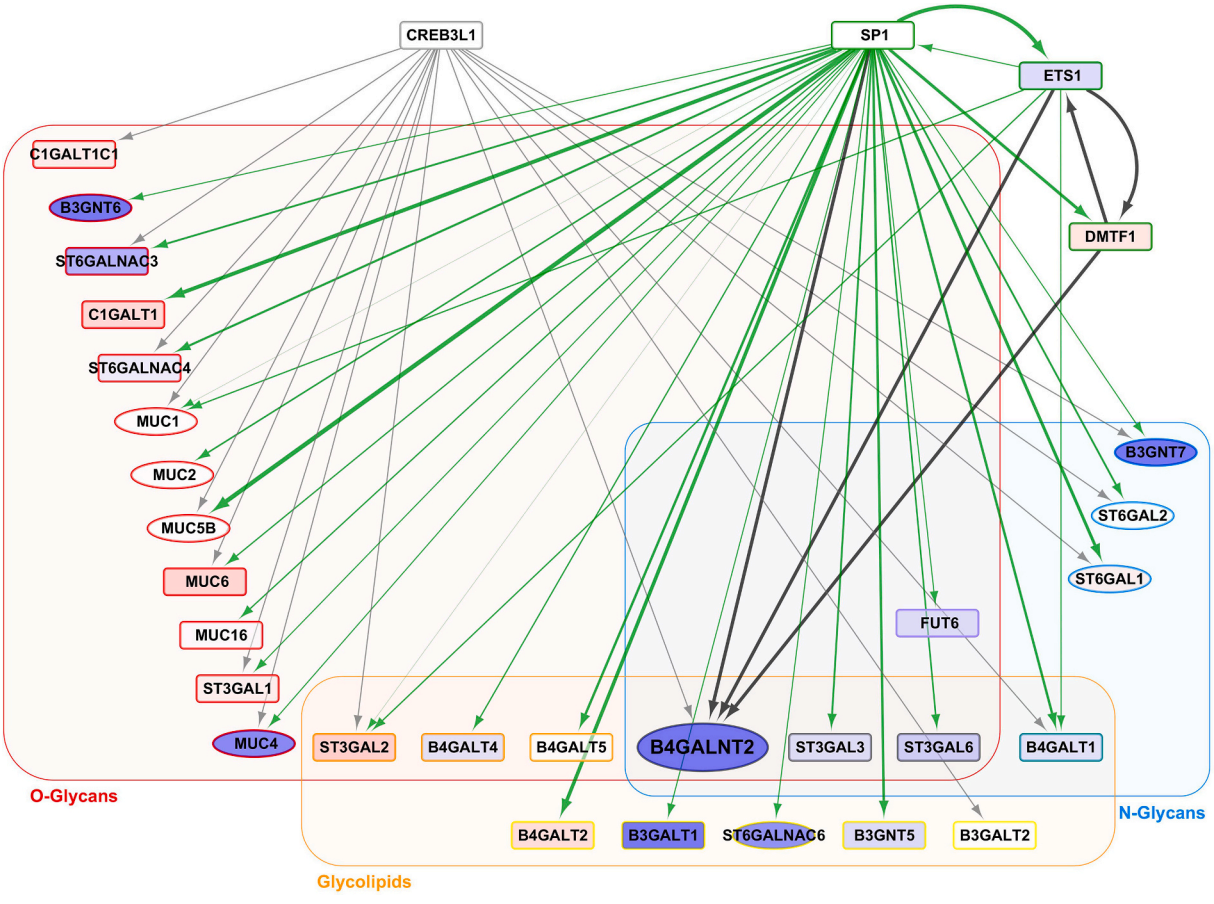

Fig. 8. Regulatory network of glyco-genes involved in the *N*-Glycans, *O*-Glycans and glycolipids biosynthesis pathways; the network was built with the iRegulon app for Cytoscape, keeping the regulators of *B4GALNT2* (grey) and the targets of ETS1, SP1 and DMTF1 (green); the regulations identified experimentally in our work are colored black; thickness of the edges corresponds to the number of times a motif of the transcription factor was found by iRegulon in the 500 bp upstream the transcription start site; glyco-genes are colored in function of their membership to one or several biosynthetic pathways: *O*-Glycans (red), *N*-Glycans (blue) or glycolipids (yellow), the ones that belong to several pathways are colored with a mixture of these primary colors; genes found as co-differentially expressed with *B4GALNT2* in the work of Pucci et al. [4] are shown as elliptic nodes; $\log_2$ fold change between primary tumor (COAD) and normal tissue from the UALCAN Portal were integrated in the network. Genes are colored as a gradient from blue ($\log_2 FC \leq -3$) to white ($\log_2 FC = 0$) and red ($\log_2 FC \geq 3$), only signals with statistical significance ≤ 0.05 were colored, others are white.

MUC4, *B3GNT7*, *MUC5B*, *B3GNT6*, *ST6GALNAC6*, *ST6GAL1*, *ST6GAL2* and *MUC2* also identified here as regulated by SP1 and/or CREB3L1 TFs (see ellipses in Fig. 8) and repressed for a majority of them, making in particular the *N*-glycans biosynthesis pathway globally repressed in the network (see also supplemental Fig. S3B). SP1 and/or ETS1 regulation was already described for *B4GALT5* in cancer cells [42,63] and for *B4GALT4* [43], two β1,4-galactosyltransferases involved in the *N*- and *O*-glycan elongation. Previous studies also demonstrated implication of SP1 in the transcriptional regulation of the α2,3-sialyltransferase *ST3GAL1* [64] involved in *O*-glycans sialylation and *ST3GAL3* [65] involved in the sialylation of *O*- and *N*-glycans and of glycolipids. Altogether, these observations further nurture the idea of a complex regulation of the glyco-genes involved in the Sd^a biosynthetic pathways by mechanisms that involve the three TFs.

4.4. Other transcriptional regulation, surrounding genomic region

However, our data do not fully explain the molecular mechanisms underpinning Sd^a/Cad expression in the healthy colon and its disappearance in the cancerous colon and beyond TFs, other regulatory mechanisms are likely involved in *B4GALNT2* transcriptional regulation.

Epigenetic modifications like DNA methylation in the promoter region of the *B4GALNT2* gene were reported to downregulate expression of this gene in cancer cell lines like HCT-116 and gastric Kato III cells [14,15]. In this regard, we obtained evidences that overexpression of ETS1 in HCT-116 did not induced expression of the *B4GALNT2* SF-type transcript (data not shown). This observation could be explained by CpG island methylation of the *B4GALNT2* promoter in these cells that would impair ETS1 binding. Of particular interest is the recent study of Pucci et al. reporting the existence of DNA methylation of an intronic site located between exon 6 and exon 7 [16] that drives increased expression of the *B4GALNT2* gene. Other epigenetic modifications include histone modifications (i.e. acetylation and methylation). In the past, Kawamura and collaborators obtained evidences using the histone deacetylase inhibitor butyrate that these histone modifications would not be primarily involved in *B4GALNT2* gene regulation [14].

Additional long-range regulatory elements with enhancer motifs could be located far upstream the *B4GALNT2* gene and could account for the tissue specificity expression of the B4GALNT2 gene. Indeed, in the RIIIS/J mouse strain, a conserved region located 30 kb upstream of the gene was described for the *b4galnt2* locus in mice, where a mutation is responsible for low plasma levels of the von Willebrand factor (VWF). This mutation causes a regulatory switch in the *b4galnt2* expression leading to its expression in the vascular endothelium and its concomitant disappearance in colon [66]. We conducted BLAST analysis in various mammalian genomes and our preliminary data led to the identification of a ~ 400 bp highly conserved genomic region located 9623, 19,046, 22,671, 27,654, 47,610, 50,276 and 30,571 bp upstream the *b4galnt2* gene in *Otolemur garnettii*, *Bos taurus*, *Loxodonta Africana, Ovis aries*, *Macaca mulatta*, *Homo sapiens* and *Mus musculus* (data not shown). Further experimental data are needed to support this finding.

Downstream genomic region could be also implicated in the *B4GALNT2* transcriptional regulation. Beside the well-established role of TF, noncoding RNAs (ncRNAs) are new regulators of great interest and among these, microRNAs (miRNAs) are the most studied. Long noncoding RNAs (lncRNAs) are non-translated RNA transcripts of more than 200 nucleotides [67,68]. Their expression is tissue-specific and changes observed in a tissue have been correlated with various human pathologies like cancer [68]. Of particular interest is the lncRNA RP11-708H21.4, a locus found at 17q21. It corresponds to a 1352 nt sequence in the 3′-untranslated region of the *B4GALNT2* gene. It was found to be downregulated in colorectal cancer and could inhibit tumorigenesis through CyclinD1 and p27 expression regulation [68]. Furthermore, the upregulation of lncRNA RP11-708H21.4 inhibits migration and invasion, induces apoptosis, and enhances 5-FU sensitivity in CRC cells by inactivating mTOR signaling.

In summary, this study shed light on the molecular basis of Sd^a expression in the human gastrointestinal tract. We identified TSS and delineate the −72/+12 nt core promoter region of the *B4GALNT2* gene relative to the short-type transcript variant. We identified ETS1 as a major regulator of its expression in colon although further studies are required to elucidate precisely how *B4GALNT2* expression is controlled in healthy colon and deregulated in colon cancer.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbagrm.2021.194747.

Credit authorship contribution statement

Cindy Wavelet-Vermuse: Formal analysis, Investigation, Methodology, Writing – review & editing. **Sophie Groux-Degroote:** Data curation, Formal analysis, Investigation, Methodology, Supervision, Writing – review & editing. **Dorothée Vicogne:** Investigation, Methodology. **Virginie Cogez:** Investigation, Methodology. **Giulia Venturi:** Investigation, Methodology. **Marco Trinchera:** Data curation, Methodology, Supervision, Writing – review & editing. **Guillaume Brysbaert:** Data curation, Investigation, Methodology, Software, Writing – original draft. **Marie-Ange Krzewinski-Recchi:** Investigation, Methodology. **Elsa Hadj Bachir:** Investigation, Methodology. **Céline Schulz:** Investigation, Methodology. **Audrey Vincent:** Investigation, Methodology. **Isabelle Van Seuningen:** Data curation, Supervision. **Anne Harduin-Lepers:** Data curation, Formal analysis, Funding acquisition, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank former students Sherazade Sebda and Sarah Calin for their excellent technical assistance, Marc Aumercier for the generous gift of ETS1 expression vector, Mario Tschan and Bruce Torbett for the generous gift of DMTF1 expression vector.

The contribution of the COST Action CA18103-INNOGLY supported by the European Cooperation in Science and Technology (COST) is greatly acknowledged.

The results shown here are in whole or in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. The authors are indebted to Monsieur Gérard Peze, président du comité du Pas de Calais and la Ligue contre le cancer for their interest in this work and financial support.

Fundings

This work was supported in part by the Centre National de la Recherche Scientifique (CNRS), the University of Lille. La Ligue contre le cancer (2016) and the Agence Nationale de la Recherche (ANR) financial support (ANR-2010-BLAN-120401) are acknowledged.

References

[1] S.I. Macvie, J.A. Morton, M.M. Pickles, The reactions and inheritance of a new blood group antigen, Sda, Vox Sang. 13 (1967) 485–492.
[2] P.H. Renton, P. Howell, E.W. IKIN, C.M. Giles, K.L.G. Goldsmith, Anti-Sda, a new blood group antibody, Vox Sang. 13 (1967) 493–501.
[3] L. Stenfelt, Å. Hellberg, M. Möller, N. Thornton, G. Larson, M.L. Olsson, Missense mutations in the C-terminal portion of the B4GALNT2-encoded glycosyltransferase underlying the Sd(a−) phenotype, Biochem. Biophys. Rep. 19 (2019), 100659.

[4] F. Dall'Olio, N. Malagolini, M. Chiricolo, M. Trinchera, A. Harduin-Lepers, The expanding roles of the Sd(a)/Cad carbohydrate antigen and its cognate glycosyltransferase B4GALNT2, Biochim. Biophys. Acta 1840 (2014) 443–453.

[5] C. Capon, E. Maes, J.C. Michalski, H. Leffler, Y.S. Kim, Sd(a)-antigen-like structures carried on core 3 are prominent features of glycans from the mucin of normal human descending colon, Biochem. J. 358 (2001) 657–664.

[6] K. Madunic, T. Zhang, O.A. Mayboroda, S. Holst, K. Stavenhagen, C. Jin, N. G. Karlsson, G.S.M. Lageveen-Kammeijer, M. Wuhrer, Colorectal cancer cell lines show striking diversity of their O-glycome reflecting the cellular differentiation phenotype, Cell. Mol. Life Sci. 78 (2021) 337–350, https://doi.org/10.1007/s00018-020-03504-z.

[7] L. Lo Presti, E. Cabuy, M. Chiricolo, F. Dall'Olio, Molecular cloning of the human beta1,4 N-acetylgalactosaminyltransferase responsible for the biosynthesis of the Sd(a) histo-blood group antigen: the sequence predicts a very long cytoplasmic domain, J. Biochem. 134 (2003) 675–682.

[8] M.D. Montiel, M.A. Krzewinski-Recchi, P. Delannoy, A. Harduin-Lepers, Molecular cloning, gene organization and expression of the human UDP-GalNAc: Neu5Acalpha2-3Galbeta-R beta1,4-N-acetylgalactosaminyltransferase responsible for the biosynthesis of the blood group Sda/Cad antigen: evidence for an unusual extended cytoplasmic domain, Biochem. J. 373 (2003) 369–379.

[9] S. Groux-Degroote, C. Wavelet, M.A. Krzewinski-Recchi, L. Portier, M. Mortuaire, A. Mihalache, M. Trinchera, P. Delannoy, N. Malagolini, M. Chiricolo, F. Dall'Olio, A. Harduin-Lepers, B4GALNT2 gene expression controls the biosynthesis of Sda and sialyl Lewis X antigens in healthy and cancer human gastrointestinal tract, Int. J. Biochem. Cell Biol. 53 (2014) 442–449.

[10] T. Dohi, A. Nishikawa, I. Ishizuka, M. Totani, K. Yamaguchi, K. Nakagawa, O. Saitoh, S. Ohshiba, M. Oshima, Substrate specificity and distribution of UDP-GalNAc:sialylparagloboside N-acetylgalactosaminyltransferase in the human stomach, Biochem. J. 288 (Pt 1) (1992) 161–165.

[11] N. Malagolini, D. Santini, M. Chiricolo, F. Dall'Olio, Biosynthesis and expression of the Sda and sialyl Lewis x antigens in normal and cancer colon, Glycobiology 17 (2007) 688–697.

[12] S. Groux-Degroote, C. Schulz, V. Cogez, M. Noel, L. Portier, D. Vicogne, C. Solorzano, F. Dall'Olio, A. Steenackers, M. Mortuaire, M. Gonzalez-Pisfil, M. Henry, F. Foulquier, L. Heliot, A. Harduin-Lepers, The extended cytoplasmic tail of the human B4GALNT2 is critical for its Golgi targeting and post-Golgi sorting, FEBS J. 285 (2018) 3442–3463.

[13] S. Neelamegham, L.K. Mahal, Multi-level regulation of cellular glycosylation: from genes to transcript to enzyme to structure, Curr. Opin. Struct. Biol. 40 (2016) 145–152.

[14] Y.I. Kawamura, M. Toyota, R. Kawashima, T. Hagiwara, H. Suzuki, K. Imai, Y. Shinomura, T. Tokino, R. Kannagi, T. Dohi, DNA hypermethylation contributes to incomplete synthesis of carbohydrate determinants in gastrointestinal cancer, Gastroenterology 135 (2008) (142-151 e143).

[15] H.R. Wang, C.Y. Hsieh, Y.C. Twu, L.C. Yu, Expression of the human Sd(a) beta-1,4-N-acetylgalactosaminyltransferase II gene is dependent on the promoter methylation status, Glycobiology 18 (2008) 104–113.

[16] M. Pucci, N. Malagolini, F. Dall'Olio, Glycosyltransferase B4GALNT2 as a predictor of good prognosis in colon cancer: lessons from databases, Int. J. Mol. Sci. 22 (2021) 4331.

[17] I. Van Seuningen, M. Perrais, P. Pigny, N. Porchet, J.P. Aubert, Sequence of the 5′-flanking region and promoter activity of the human mucin gene MUC5B in different phenotypes of colon cancer cells, Biochem. J., 348 Pt 3 (2000) 675–686.

[18] F. Colomb, M.A. Krzewinski-Recchi, A. Steenackers, A. Vincent, A. Harduin-Lepers, P. Delannoy, S. Groux-Degroote, TNF up-regulates ST3GAL4 and sialyl-Lewisx expression in lung epithelial cells through an intronic ATF2-responsive element, Biochem. J. 474 (2017) 65–78.

[19] N. Jonckheere, A. Velghe, M.P. Ducourouble, M.C. Copin, I.B. Renes, I. Van Seuningen, The mouse Muc5b mucin gene is transcriptionally regulated by thyroid transcription factor-1 (TTF-1) and GATA-6 transcription factors, FEBS J. 278 (2011) 282–294.

[20] M.W. Pfaffl, A new mathematical model for relative quantification in real-time RT-PCR, Nucleic Acids Res. 29 (2001), e45.

[21] R. Janky, A. Verfaillie, H. Imrichova, B. Van de Sande, L. Standaert, V. Christiaens, G. Hulselmans, K. Herten, M. Naval Sanchez, D. Potier, D. Svetlichnyy, Z. Kalender Atak, M. Fiers, J.C. Marine, S. Aerts, iRegulon: from a gene list to a gene regulatory network using large motif and track collections, PLoS Comput. Biol. 10 (2014), e1003731.

[22] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504.

[23] G. Su, J.H. Morris, B. Demchak, G.D. Bader, Biological network exploration with Cytoscape 3, Curr. Protoc. Bioinformatics 47 (2014) (8 13 11-24).

[24] D.S. Chandrashekar, B. Bashel, S.A.H. Balasubramanya, C.J. Creighton, I. Ponce-Rodriguez, B. Chakravarthi, S. Varambally, UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses, Neoplasia 19 (2017) 649–658.

[25] H. Hirai, C.J. Sherr, Interaction of D-type cyclins with a novel myb-like transcription factor, DMP1, Mol. Cell. Biol. 16 (1996) 6457–6467.

[26] T. Dohi, Y. Yuyama, Y. Natori, P.L. Smith, J.B. Lowe, M. Oshima, Detection of N-acetylgalactosaminyltransferase mRNA which determines expression of Sda blood group carbohydrate structure in human gastrointestinal mucosa and cancer, Int. J. Cancer 67 (1996) 626–631.

[27] N. Malagolini, F. Dall'Olio, G. Di Stefano, F. Minni, D. Marrano, F. Serafini-Cessi, Expression of UDP-GalNAc:NeuAc alpha 2,3Gal beta-R beta 1,4(GalNAc to Gal) N-acetylgalactosaminyltransferase involved in the synthesis of Sda antigen in human large intestine and colorectal carcinomas, Cancer Res. 49 (1989) 6466–6470.

[28] M. Pucci, I. Gomes Ferreira, N. Malagolini, M. Ferracin, F. Dall'Olio, The Sda synthase B4GALNT2 reduces malignancy and stemness in colon cancer cell lines independently of Sialyl Lewis X inhibition, Int. J. Mol. Sci. 21 (2020) 6558.

[29] M. Pucci, I. Gomes Ferreira, M. Orlandani, N. Malagolini, M. Ferracin, F. Dall'Olio, High expression of the Sda synthase B4GALNT2 associates with good prognosis and attenuates stemness in colon cancer, Cells 9 (2020) 948.

[30] A. Galeev, A. Suwandi, A. Cepic, M. Basu, J.F. Baines, G.A. Grassl, The role of the blood group-related glycosyltransferases FUT2 and B4GALNT2 in susceptibility to infectious disease, Int. J. Med. Microbiol. 311 (2021), 151487.

[31] P.L. Smith, J.B. Lowe, Molecular cloning of a murine N-acetylgalactosamine transferase cDNA that determines expression of the T lymphocyte-specific CT oligosaccharide differentiation antigen, J. Biol. Chem. 269 (1994) 15162–15171.

[32] K. Furukawa, H. Soejima, N. Niikawa, H. Shiku, Genomic organization and chromosomal assignment of the human beta1, 4-N-acetylgalactosaminyltransferase gene. Identification of multiple transcription units, J. Biol. Chem. 271 (1996) 20836–20844.

[33] A. Harduin-Lepers, J.H. Shaper, N.L. Shaper, Characterization of two cis-regulatory regions in the murine beta 1,4-galactosyltransferase gene. Evidence for a negative regulatory element that controls initiation at the proximal site, J. Biol. Chem. 268 (1993) 14348–14359.

[34] A. Harduin-Lepers, N.L. Shaper, J.A. Mahoney, J.H. Shaper, Murine beta 1,4-galactosyltransferase: round spermatid transcripts are characterized by an extended 5′-untranslated region, Glycobiology 2 (1992) 361–368.

[35] S.W. Kim, S.H. Lee, K.S. Kim, C.H. Kim, Y.K. Choo, Y.C. Lee, Isolation and characterization of the promoter region of the human GM3 synthase gene, Biochim. Biophys. Acta 1578 (2002) 84–89.

[36] E.C. Svensson, P.B. Conley, J.C. Paulson, Regulated expression of alpha 2,6-sialyltransferase by the liver-enriched transcription factors HNF-1, DBP, and LAP, J. Biol. Chem. 267 (1992) 3466–3472.

[37] E.C. Svensson, B. Soreghan, J.C. Paulson, Organization of the beta-galactoside alpha 2,6-sialyltransferase gene. Evidence for the transcriptional regulation of terminal glycosylation, J. Biol. Chem. 265 (1990) 20863–20868.

[38] D. Petit, A.M. Mir, J.M. Petit, C. Thisse, P. Delannoy, R. Oriol, B. Thisse, A. Harduin-Lepers, Molecular phylogeny and functional genomics of beta-galactoside alpha2,6-sialyltransferases that explain ubiquitous expression of st6gal1 gene in amniotes, J. Biol. Chem. 285 (2010) 38399–38414.

[39] M. Bobovski, A. Harduin-Lepers, P. Delannoy, ST8alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 1 (ST8SIA1), in: N. Taniguchi, K. Honke, M. Fukuda, H. Narimatsu, Y. Yamaguchi, T. Angata (Eds.), Handbook of Glycosyltransferases and Related Genes, Springer-Verlag GmbH, 2013.

[40] K. Furukawa, M. Horie, K. Okutomi, S. Sugano, K. Furukawa, Isolation and functional analysis of the melanoma specific promoter region of human GD3 synthase gene, Biochim. Biophys. Acta 1627 (2003) 71–78.

[41] R. Tange, T. Tomatsu, T. Sato, Transcription of human beta4-galactosyltransferase 3 is regulated by differential DNA binding of Sp1/Sp3 in SH-SY5Y human neuroblastoma and A549 human lung cancer cell lines, Glycobiology 29 (2019) 211–221.

[42] T. Sato, K. Furukawa, Transcriptional regulation of the human beta-1,4-galactosyltransferase V gene in cancer cells: essential role of transcription factor Sp1, J. Biol. Chem. 279 (2004) 39574–39583.

[43] A. Sugiyama, N. Fukushima, T. Sato, Transcriptional mechanism of the beta4-galactosyltransferase 4 gene in SW480 human colon cancer cell line, Biol. Pharm. Bull. 40 (2017) 733–737.

[44] F. Yamamoto, P.D. McNeill, S. Hakomori, Genomic organization of human histo-blood group ABO genes, Glycobiology 5 (1995) 51–58.

[45] S. Lehoux, S. Groux-Degroote, A. Cazet, C.M. Dhaenens, C.A. Maurage, M.L. Caillet-Boudin, P. Delannoy, M.A. Krzewinski-Recchi, Transcriptional regulation of the human ST6GAL2 gene in cerebral cortex and neuronal cells, Glycoconj. J. 27 (2010) 99–114, https://doi.org/10.1007/s10719-009-9260-y.

[46] N. Kurosawa, S. Takashima, M. Kono, Y. Ikehara, M. Inoue, Y. Tachida, H. Narimatsu, S. Tsuji, Molecular cloning and genomic analysis of mouse GalNAc alpha2, 6-sialyltransferase (ST6GalNAc I), J. Biochem. 127 (2000) 845–854.

[47] V.R. Falkenberg, K. Alvarez, C. Roman, N. Fregien, Multiple transcription initiation and alternative splicing in the 5′ untranslated region of the core 2 beta1-6 N-acetylglucosaminyltransferase I gene, Glycobiology 13 (2003) 411–418.

[48] L. Li, J.R. Davie, The role of Sp1 and Sp3 in normal and cancer cell biology, Ann. Anat. 192 (2010) 275–283.

[49] K. Beishline, J. Azizkhan-Clifford, Sp1 and the 'hallmarks of cancer', FEBS J. 282 (2015) 224–258.

[50] T. Hsu, M. Trojanowska, D.K. Watson, Ets proteins in biological control and cancer, J. Cell. Biochem. 91 (2004) 896–903.

[51] D.P. Turner, D.K. Watson, ETS transcription factors: oncogenes and tumor suppressor genes as therapeutic targets for prostate cancer, Expert. Rev. Anticancer. Ther. 8 (2008) 33–42.

[52] H. Suzuki, V. Romano-Spica, T.S. Papas, N.K. Bhat, ETS1 suppresses tumorigenicity of human colon cancer cells, Proc. Natl. Acad. Sci. U. S. A. 92 (1995) 4442–4446.

[53] B.J. Graves, J.M. Petersen, Specificity within the ets family of transcription factors, Adv. Cancer Res. 75 (1998) 1–55.

[54] F. Shirasaki, H.A. Makhluf, C. LeRoy, D.K. Watson, M. Trojanowska, Ets transcription factors cooperate with Sp1 to activate the human tenascin-C promoter, Oncogene 18 (1999) 7755–7764.

[55] M.P. Tschan, K.M. Fischer, V.S. Fung, F. Pirnia, M.M. Borner, M.F. Fey, A. Tobler, B.E. Torbett, Alternative splicing of the human cyclin D-binding Myb-like protein (hDMP1) yields a truncated protein isoform that alters macrophage differentiation patterns, J. Biol. Chem. 278 (2003) 42750–42760.

[56] X. Yang, Y. Lou, M. Wang, C. Liu, Y. Liu, W. Huang, miR675 promotes colorectal cancer cell growth dependent on tumor suppressor DMTF1, Mol. Med. Rep. 19 (2019) 1481–1490.
[57] K. Inoue, A. Mallakin, D.P. Frazier, Dmp1 and tumor suppression, Oncogene 26 (2007) 4329–4335.
[58] H. Guo, M.J. Pierce, Transcriptional regulation of glycan expression, in: N. Taniguchi, T. Endo, G.W. Hart, P.H. Seeberger, C.-H. Wong (Eds.), Glycoscience: Biology and Medicine vol. 2, Springer Japan, Japan, 2015, pp. 1173–1180.
[59] Y. Narimatsu, H.J. Joshi, R. Nason, J. Van Coillie, R. Karlsson, L. Sun, Z. Ye, Y.-H. Chen, K.T. Schjoldager, C. Steentoft, S. Furukawa, B.A. Bensing, P.M. Sullam, A. J. Thompson, J.C. Paulson, C. Büll, G.J. Adema, U. Mandel, L. Hansen, E. P. Bennett, A. Varki, S.Y. Vakhrushev, Z. Yang, H. Clausen, An atlas of human glycosylation pathways enables display of the human glycome by gene engineered cells, Mol. Cell 75 (2019) (394-407.e395).
[60] M. Andrianifahanana, N. Moniaux, S.K. Batra, Regulation of mucin expression: mechanistic aspects and implications for cancer and inflammatory diseases, Biochim. Biophys. Acta 1765 (2006) 189–222.
[61] N. Jonckheere, A. Vincent, B. Neve, I. Van Seuningen, Mucin expression, epigenetic regulation and patient survival: a toolkit of prognostic biomarkers in epithelial cancers, Biochim. Biophys. Acta Rev. Cancer 1876 (2021), 188538.
[62] P. Mellor, L. Deibert, B. Calvert, K. Bonham, S.A. Carlsen, D.H. Anderson, CREB3L1 is a metastasis suppressor that represses expression of genes regulating metastasis, invasion, and angiogenesis, Mol. Cell. Biol. 33 (2013) 4985–4995.
[63] T. Sato, K. Furukawa, Sequential action of Ets-1 and Sp1 in the activation of the human beta-1,4-galactosyltransferase V gene involved in abnormal glycosylation characteristic of cancer cells, J. Biol. Chem. 282 (2007) 27702–27712.
[64] A. Taniguchi, I. Yoshikawa, K. Matsumoto, Genomic structure and transcriptional regulation of human Galbeta1,3GalNAc alpha2,3-sialyltransferase (hST3Gal I) gene, Glycobiology 11 (2001) 241–247.
[65] A. Taniguchi, K. Saito, T. Kubota, K. Matsumoto, Characterization of the promoter region of the human Galbeta1,3(4)GlcNAc alpha2,3-sialyltransferase III (hST3Gal III) gene, Biochim. Biophys. Acta 1626 (2003) 92–96.
[66] K.L. Mohlke, A.A. Purkayastha, R.J. Westrick, P.L. Smith, B. Petryniak, J.B. Lowe, D. Ginsburg, Mvwf, a dominant modifier of murine von Willebrand factor, results from altered lineage-specific expression of a glycosyltransferase, Cell 96 (1999) 111–120.
[67] T. Derrien, R. Guigo, R. Johnson, The long non-coding RNAs: a new (P)layer in the "dark matter", Front. Genet. 2 (2011) 107.
[68] R. Zhang, F. Ni, B. Fu, Y. Wu, R. Sun, Z. Tian, H. Wei, A long noncoding RNA positively regulates CD56 in human natural killer cells, Oncotarget 7 (2016) 72546–72558.