*Mutagenesis*, 2016, 31, 401–408
doi:10.1093/mutage/gev086
Original Manuscript
Advance Access publication 16 December 2015

OXFORD

Original Manuscript

# Promoter SNPs rs116896264 and rs73933062 form a distinct haplotype and are associated with galectin-4 overexpression in colorectal cancer

**Reham Helwa[1,2,7,*], Mohamed Ramadan[3], Abdel-Hady A. Abdel-Wahab[4], Stian Knappskog[5,6] and Andrea S. Bauer[2]**

[1]Molecular Cell Biology Lab, Zoology Department, Faculty of Science, Ain Shams University, Cairo, Egypt, [2]Division of Functional Genome Analysis, Deutsche Krebsforschungszentrum (DKFZ), Heidelberg, Germany, [3]Surgical Oncology Department and [4]Cancer Biology Department, Egyptian National Cancer Institute, Cairo University, Cairo, Egypt, [5]Section of Oncology, Department of Clinical Science, University of Bergen, Bergen, Norway and [6]Department of Oncology, Haukeland University Hospital, Bergen, Norway

[7]Present address: Section of Oncology, Department of Clinical Science, University of Bergen, Bergen, Norway.

*To whom correspondence should be addressed. Tel: +20 1 288057593; Fax: +20 2 24662917; Email: rehamhg@yahoo.com

Received 2 September 2015; Revised 9 November 2015; Accepted 25 November 2015.

## Abstract

Galectin-4 is a member of the galectin family which consists of 15 galactoside-binding proteins. Previously, galectin-4 has been shown to have a role in cancer progression and metastasis and it is found upregulated in many solid tumours, including colorectal cancer (CRC). Recently, the role in the metastatic process was suggested to be *via* promoting cancer cells to adhere to blood vascular endothelium. In the present study, the regulatory region of *LGALS4* (galectin-4) in seven colon cell lines was investigated with respect to genetic variation that could be linked to expression levels and therefore a tumourigenic effect. Interestingly, qRT-PCR and sequencing results revealed that galectin-4 upregulation is associated with SNPs rs116896264 and rs73933062. By use of luciferase reporter- and pull-down assays, we confirmed the association between the gene upregulation and the two SNPs. Also, using pull-down assay followed by mass spectrometry, we found that the presence rs116896264 and rs73933062 is changing transcription factors binding sites. In order to assess the frequencies of the two SNPs among colon cancer patients and healthy individuals, we genotyped 75 colon cancer patients, 12 patients with adenomatous polyposis and 17 patients with ulcerative colitis and we performed data mining in the 1000 genomes databank. We found the two SNPs co-occuring in 21% of 75 CRC patients, 0 out of 12 patients of adenomatous polyposis, and 6 out of 17 patients (35%) with ulcerative colitis. Both in the patient samples and in the 1000 genomes project, the two SNPs were found to co-occur whenever present ($D' = 1$).

## Introduction

Colorectal cancer (CRC) is one of the most common cancers and accounts for 10% of all new cancer cases and cancer deaths each year (1,2) with 1.2 million diagnosed cases every year worldwide (3). Approximately half of CRC patients develop metastasis. The occurrence of metastatic disease decreases the 5-years survival rate and subsequently increases the rate of mortality (4). A recent 30-year follow-up study showed that improved surveillance and early diagnosis of disease have reduced the mortality rate (5).

Galectins are a family of animal lectins (6). So far, 15 mammalian galectins have been identified. They are found to be involved in many functions, ranging from the mediation of cell adhesion and the

promotion of cell–cell interactions to the recognition of pathogens. From the structural point of view, prototypic galectins (galectin-1, -2, -5, -7, -10,-11, -13, -14, -15) consist entirely of one carbohydrate recognition domain (CRD) containing 130 amino acids. Tandem repeat galectins (galectin-4, -6, -8, -9 and -12) have two homologous CRDs separated by a linker of up to 70 amino acids in a single polypeptide chain. A third type of galectin is the chimera type (galectin-3) that consists of an N-terminal region connected to a CRD (7–9).

Galectin-4 was found to be expressed only in the alimentary canal from the tongue to the large intestine (10,11). As reviewed elsewhere, the expression of galectin-4 in epithelial cells enable these cells to survive lack of nutrients and growth factors for prolonged time of period. Also, within a collection of human epithelial cancer cell lines, overexpression of soluble galectin-4 was found in the differentiated cells (11).

Although some investigators have found galectin-4 downregulated in colon cancers (12) and AML (13), most studies have found that upregulation of galectins are involved in cancer progression and metastasis (14,15). This also includes CRC, where several studies have shown that circulating galectins, including galectin-4, are notably increased and potentially promoting angiogenesis and metastasis (16–19). Chen *et al.* (16) assigned the importance of circulating galectins to their contribution to increased circulation of several cytokines and chemokines (G-SCF, IL-6 and MCP-1) which in turn promote the angiogenesis and metastasis processes. Also, in lung cancer, galectin-4 expression was considered as an independent predictor for lymph node metastasis and poor survival (20). Nagy *et al.* (21) have found that increased percentage of cells expressing galectin-4 is correlated with poor prognosis in colon carcinoma. *LGALS4* was also shown to be overexpressed in mucinous epithelial ovarian cancers and absent in normal ovaries (22).

In an expression profiling data set, we have previously observed that galctin-4 is upregulated in a cell line with Duke's D colon cancer (KM20L2; unpublished data). Moreover, it was also shown in the same data set that it is also overexpressed in a cell line derived from familial adenomatous polyposis patients (LT97). From the expression profiling data, we concluded that galectin-4 upregulation might be involved in cancer progression and metastasis, since the upregulation was not shown in other cell lines with different Duke's stages.

Due to the potential contribution of galectin-4 to cancer progression and metastasis, the present study aimed to assess regulatory mechanism(s) of this gene. We found two SNPs (rs116896264 and rs73933062) to be associated with gene upregulation in cell lines. Therefore, functional experiments were performed in order to evaluate the effect of these two SNPs on the promoter activity. In parallel, the frequency of rs116896264 and rs73933062 was also investigated in 104 patients with CRC and premalignant lesions. A linkage disequilibrium for the two SNPs was found and was consistent with data from 1000 Genomes Project Phase 3.

## Material and methods

### Cell lines and cell culture

Seven colon cell lines were used in this study: a normal colon cell line CCD-18Co (ATCC), and an adenoma cell line (LT97; gift from Brigitte Marian) as well as SW1116, Sw480, Co115, SW620 and KM20L2 from Duke's stage A, B, C primary tumour, C lymph node infiltration and D, respectively (gifts from Gabriela Aust, Heike Allgayer, Richard Iggo, Francis RAUL, Øystein Fodstad). The cultivation and maintenance of the cells was carried out in the proper medium (RPMI and DMEM according to ATCC recommendation for each cell line) at 37°C in a humidified atmosphere of 5% $CO_2$. LT97 was cultivated in MCDB 302 containing 20% of L15 Leibovitz medium, 2% FCS (foetal calf serum), 0.2 nM triiodo-l-thyronine, 1 µg/ml hydrocortisone (302 basic medium) supplemented with 10 µg/ml insulin, 2 µg/ml transferrin, 5 nM sodium selenite and 30 ng/ml EGF (epidermal growth factor).

### RNA and DNA purification

DNA, RNA and protein samples were isolated from cell pellets with Allprep DNA/RNA/Protein mini kit (Qiagen, Hilden, Germany), following the protocol suggested by the manufacturer. RNA integrity was evaluated using an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA). Only samples with an RNA Integrity Number of at least eight were used for microarray analysis.

### qRT-PCR for galectin-4 in cell lines and immunoblotting

Expression of galectin-4 was assessed by qRT-PCR using Hs_LGALS4_1_SG QuantiTect Primer Assay, one-step QuantiFast SYBR Green RT-PCR Kit (Qiagen, Hilden, Germany) and the Light Cycler 480 instrument (Roche Diagnostics). Starting with 10 ng of RNA, following the manufacturer's instructions, a reverse transcription step was done at 50°C for 10 min followed by PCR initial activation step for 5 min at 95°C. Afterwards, 40 cycles of 10 s at 95°C, 20 s annealing at 55°C and extension at 68°C for 20 s were performed. A final melting curve to check fidelity was done from 95°C for 5 s, 65°C 1 min with 5–10 signal acquisitions every 1°C up to 97°C. Expression levels were normalised relative to the transcription level of α-tubulin. All samples were run in triplicate.

Cell pellets were mixed with lysis buffer. Protein samples were resolved on SDS-polyacrylamide gel. Then, proteins were electrophoretically transferred to nitrocellulose membranes (Amersham Pharmacia Biotech, Buckinghamshire, UK) followed by blocking for 1 h in 3% non-fat milk in TBST. Subsequently, the membrane was incubated in the primary antibody against Galectin-4 (R&D system) for 1 h. After incubation with HRP conjugated secondary antibody (antigoat), bands were detected by chemiluminescence using the ECL Western Blotting Detection System (GE Healthcare, Amersham, Buckinghamshire, UK).

### Promoter sequencing

The promoter sequence of *LGALS4* was retrieved using Transcriptional Regulatory Element Database (23,24). The CpG island was predicted using Methprimer in the first exon and first intron. The promoter was sequenced using LGALS4-700F: TTCACAGTTGCTGGGAGAGG and LGALS4-700R: GATGACGAGGGCCAACAGTTAGACGTG. The primers were designed using Primer3 (25). The primers were designed to cover 700 nucleotide (244 nucleotide before TSS to nucleotide 242 after the start codon). The amplified products from the seven cell lines were Sanger sequenced at the GATC Biotech, Konstanz, Germany.

### Luciferase reporter assay

According to the results of promoter sequencing, two fragments of the *LGALS4* promoter were cloned into pGL4.1 firefly-expressing reporter vector (Promega, Madison, WI). A fragment containing the two SNPs found in our study and another one with wild-type sequence. PCR fragments were amplified from the wild-type cell line (Co115) and the variant cell line (KM20L2) using the following primers: LGALS4 2var F:

AAAAAAAGCTAGCTTCACAGTTGCTGGGAGAGGandLGALS4 2var R: TTTGGGGAAGCTTGATGACGAGGGCCAACAGTTAGA. Following the PCR amplification, the products were digested with Hind III and NheI and ligated into pGL4.1 plasmid. Then, the cell lines were co-transfected with constructed plasmid and pRL-TK using effectene kit (Qiagen, Hilden, Germany). After 24 h, the cells were lysed and the cell lysates were analysed for firefly luciferase activity with the dual-luciferase reporter assay system (Promega). Firefly activity was normalized to Renilla (pRL-TK).

The Co115 cell line was selected for transfection because it expresses galectin-4 in low but detectable levels (based on qRT-PCR results) and therefore it should contain balanced activators/repressors of galectin-4 promoter. Co115 cells were transfected with each construct (which contains Firefly luciferase as a reporter gene) and pRL-TK (which has a Renilla luciferase). The Firefly/Renilla ratio was calculated for transfection efficiency control and the fold changes were calculated by dividing Firefly/Renilla ratio from the variant construct over Firefly/Renilla ratio of the wild-type construct.

### Pull down-assay and mass spectrometry

PCR fragments covering the two SNPs (244 nucleotide before TSS to nucleotide 242 after the start codon), with and without the SNPs, were used as baits for DNA binding proteins. Also shorter PCR fragments were used to cover each SNP independently. For rs11686264, biotinylated LGALS4 SNP1 F: AAAAAAACTCGAGTTCACAGTTGCTGGGAGAGGandLGALS4 SNP2 R: GGGGGGAAGCTTGCTGCGCTAGTGGCTGGTC were used. Also, biotinylated LGALS4 SNP2 F: AAAAAAAGCTAGCCCA CCATCTCCCACTCCTG and LGALS4 SNP2 R: TTTGGGGAAG CTTGATGACGAGGGCCAACAGTTAGA were used to amplify the sequence containing rs73933062. The pull down experiment was done using Pierce Pull-Down Biotinylated Protein:Protein Interaction kit (Thermo Scientific, Rockford, USA). The kit was modified to suit DNA:Protein interactions. The promoter fragments were amplified using biotinylated primers and linked to an immobilized streptavidin column. The nuclear proteins of Co115 (the same cell-line used for luciferase assays) were extracted by NE-PER Nuclear and Cytoplasmic Extraction Reagents (Thermo Scientific, Rockford, USA) and diluted in 5X binding buffer (0.1 M Hepes pH 8.0, 0.25 M KCl, 25 mM DTT, 0.25 mM EDTA, 5 mM $MgCl_2$, 25% v/v glycerol). The nuclear protein extract was incubated on the column for 30 min at room temperature. Then, columns were washed with TBST to remove unspecific bindings and then the bound proteins were eluted with Laemmli buffer at 95°C. Eluted protein samples were resolved on SDS-polyacrylamide gel. Subsequently, the unique bands were cut for mass spectrometry analysis. The mass spectrometry analysis was performed at core facility, DKFZ, Heidelberg, according to a protocol described in detail previously (26–28). Proteins with total score >50 were considered to be credible. The proteins with single peptide signal, or produced by trypsin digestion, and the proteins from skin contamination were omitted from the results.

### Promoter genotyping in CRC patient samples

The promoter sequences of galectin-4 were screened for the two SNPs in samples from CRC patients using LGALS4-700F: TTCACAGTTGCTGGGAGAGG and LGALS4-700R: GATGACGAGGGCCAACAGTTAGACGTG. DNA was isolated from 104 samples (Table 1): (i) 18 colorectal tumour tissues from the National Center of Tumor Diseases (NCT), Heidelberg, Germany, (ii) lesions and their adjacent normal tissues from 54 patients (27 CRC patients, 15 ulcerative colitis patients, and 12 patients with adenomatous polyps) from the Egyptian National Cancer Institute (ENCI), Cairo, Egypt, (iii) 15 colorectal tumour samples from the Egyptian National Cancer Institute (ENCI), Cairo and (iv) 15 DNA samples isolated from whole blood of CRC and two blood samples from ulcerative colitis patients (the Egyptian National Cancer Institute (ENCI), Cairo).

All the patient samples were obtained after ethical approval of the relevant review boards at the corresponding institutions; written informed consent had been obtained from all donors.

### Linkage disequilibrium calculation

The linkage disequilibrium was calculated using the equation: $D = p(AB) - p(A) \times p(B)$, where is p(AB) is the frequency of the SNPs pair, p(A) is frequency of 1st SNP, and p(B) is frequency of the 2nd SNP. Then, D′ was calculated using the equation: $D' = D/D_{max}$. where, if $D > 0$: $D_{max} = \min(p(A)p(b), p(a)p(B))$ or if $D < 0$: $D_{max} = \max(-p(A)p(B), -p(a)p(b))$.

## Results

### rs116896264 and rs73933062 SNPs are associated with galectin-4 upregulation in colon cell lines

In order to assess genetic variations potentially affecting Galectin-4 expression, we screened the regulatory region of this gene in seven colon cell lines. In parallel, qRT-PCR was performed for the same cells. Five of the cell lines did not reveal any differences from the reference sequence in the investigated region. However, we found two cell lines (LT97 and KM20L2) to harbour alternative nucleotides (A and A) in the polymorphic loci rs116896264 and rs73933062 (the wild-type alleles being C and G, respectively). Both cell lines (LT97 and KM20L2) carried both the two SNPs.

Interestingly, the two cell lines carrying the rs116896264 and rs73933062 variant alleles, revealed the by far the highest mRNA and protein levels of galectin-4 in the selected cell lines, with levels several 100-fold higher than the remaining five cell lines (Figure 1a and b).

**Table 1.** Source and types of patient samples

| Colorectal lesion | Source of samples | Collected sample | Samples number |
|---|---|---|---|
| **Colorectal cancer** | NCT, Heidelberg, Germany | Tumour tissues | 18 |
| | ENCI, Cairo, Egypt | Tumour tissues and adjacent normal samples | 27 |
| | | Tumour tissues | 15 |
| | | Blood | 15 |
| **Adenomatous polyposis** | ENCI, Cairo, Egypt | Polyps tissue and adjacent normal samples | 12 |
| **Ulcerative colitis** | ENCI, Cairo, Egypt | Tissue and adjacent normal | 15 |
| | | Blood | 2 |

The overexpression of galectin-4 in the present study was not correlated to degree of cellular differentiation, as we have used highly differentiated cell lines like SW1116 (29), moderately and less differentiated cells such as SW480 (30,31) and SW620 (31) and poorly differentiated as Co115 (32). All of these cells are not expressing galectin-4, regardless to the level of differentiation. On the other hand, Lt97 which represent colon microadenoma (33) and KM20L2 which exhibits lymph node metastasis (30) are highly over-expressing galectin-4 in mRNA and protein forms.

## rs116896264 and rs73933062 increase promoter activity

The two SNPs are located in the regulatory region of galectin-4: rs116896264 is in the promoter sequence before the transcription start site (TSS) and rs73933062 is in a putative enhancer region, after the TSS. The impact of the two SNPs on transcriptional activation was assessed by luciferase activity assay. Promoter sequence with or without the two SNPs was evaluated using pGL4.10 constructs and transfection of Co115 cells.

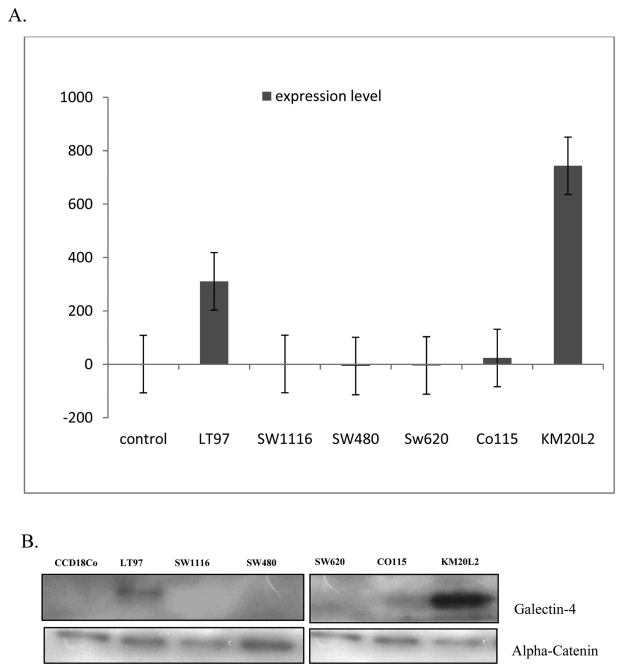


**Figure 1.** (**A**) Galectin-4 mRNA expression in colorectal cancer cell lines as observed by qRT-PCR. Expression levels are shown as fold changes relative to a normal colon cell line (CCD18Co). All CT values were normalized to alpha tubulin. (**B**) Immunoblotting of galectin-4 in the seven cell lines used in the present study and the corresponding alpha-tubulin in each cell line.

Upon transfection with the construct containing the promoter sequence with the two SNPs, we observed a 4-fold increase in the luciferase activity as compared to the wild type, strongly indicating the haplotype carrying the two SNPs to cause elevated galectin-4 levels (Figure 2).

## rs116896264 and rs73933062 influence the protein binding sites

In order to understand the mechanism underlying the transcriptional upregulation associated with the presence of the two SNPs, a pull down experiment followed by mass spectrometry analysis was performed. In this experiment, biotinylated PCR fragments representing the *LGALS4* promoter region were used as baits for potential interaction partners in nuclear extracts from Co155 cells. The pull-down was performed using short PCR fragments containing each of the two SNPs (rs116896264 and rs73933062) and their corresponding wild-type sequences. Also, a longer fragment containing the two SNPs together (and a corresponding wild-type fragment) was analysed.

The presence of rs116896264 influenced the binding affinity by generating new binding sites for factors like CSTF3, PRKDC, EEF1A, NRF1 and NUP54 and deleting other response elements such as FABP5, FLG2 and LBP-1a. The same effect was also found by rs73933062: It introduced CSTF3, EEF1A1 and NR6A1 binding sites and deleted API5, FBXL7, TP63, TLS and TRAP1 (details are shown in Table 2). Also, the pull-down experiment manifested the regulatory proteins that bind to galectin-4 promoter regardless to the presence or absence of the two SNP. These proteins include enolase-1 (ENO1), fuse-binding protein-1 (FUBP1), fuse-binding protein-2 (FUBP2 or KSRP), FUS, NCL, AZGP and LBP1a.

In addition to the effects seen when the SNPs were introduced one by one, interestingly, we also found that concomitant presence of rs116896264 and rs73933062 introduced new, as well as removed several protein binding sites: Among the response element that were deleted were TP53, Aconitase 1 (ACO1), Retinoblastoma Binding Protein-7 (RBBP7), siah binding protein 1 (PUF60) and DAZ associated protein. On the other hand, the two SNPs introduced new response elements for proteins such as MYB binding protein 1a (MYBB1a), fatty acid binding protein 5 (FABP5) and peoxiredoxin (PRDX1).

## The two SNPs are shown together in patient samples

Among 104 patients which were collected from Germany and Egypt, rs116896264 and rs73933062 SNPs were found concomitantly in four out of 18 (22.2%) German CRC samples, 13 out of 57 (22.8%), Egyptian CRC patients, 0 of 12 of Egyptian adenomatous polyps patients and 6 out of 17 (35.29%) Egyptian ulcerative colitis patients. Of note, the genotype of the two SNPs were assessed in 54

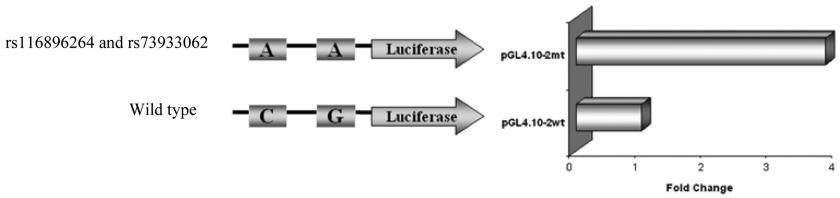


**Figure 2.** Luciferase reporter assay for galectin-4 upstream sequence: the presence of the two SNPs rs116896264 and rs73933062 in the upstream sequence increased the promoter activity by 4-fold. Bars indicate transcription activity as Firefly luciferase (vectors containing SNP-variant or wild-type galectin-4 sequence) relative to renilla luciferase transfection efficiency control, in Co115 cells.

**Table 2.** Pull-down/mass spectrometry analysis result of binding proteins to *LGALS4* promoter

| | Binding to variant sequence | Binding to wild type | Binding to both genotypes |
|---|---|---|---|
| rs116896264 | CSTF3, EEF1A, HSPA8, LTF, NRF1, NUP54, PHGDH, PRKDC, RFC1, LSM8 | FLG2, nuclear factor IV, S100A9 UBP1 (LBP-1a) | AZGP1, CEP164, DNA helicase Q1, DNA replication ATP-dependent helicase, DSP, GTF2I, HNRNPU, HSP90A, HSPD1, NCL, NONO (P54), PARP1, PTBP1, RAD23B, SFPQ, FUBP2 |
| rs73933062 | CSTF3, EEF1A1, NR6A1 | A2LP, API5, EIF5A, FBXL7, HNRN-PK, NCL, TLS (FUS), TP63, TRAP1 | Fabp5, FLG2, GTF2I, HNRNPL, HNRNPU, HSP60, HSP90, HSPA8, KHSRP, LBP1A, NONO (P54), NR2C2, PARP1, RPA1, SFPQ, TFCP2, topoisomerase I, VIM |
| Both SNPs | DNMT1, EEF1A, EEF1A1, IF4A1, MYBB1a, NUP54, PRKDC | ACO1, DAZAP1, DDB2, FBRNP, FEN1, GTF2I, LSM2, NUP93, PA2G4, PUF60, RBBP7, RBM39, RCC1, RU-VBL1, TP53, TP63 | AZGP1, DNA helicase Q1, DNA replication ATP-dependent helicase, EIF5A, ENO1, FLG2, FUBP1, HNRNPA, HNRNPL, HNRNPU, HSP60, HSP90, HSP90A, HSPA8, KHSRP, LBP1A, NCL, NONO (P54), NR2C2, NRF1, PCBP2, PTBP1, RAD23B, SF3A3, SFPQ, TFCP2, TLS, TOP1, TROVE2, VIM |

cases in both colorectal lesion and adjacent normal tissues (27 CRC, 12 polyps and 15 ulcerative colitis patients); for all these samples the genotyping of normal tissue was in concordance with the original genotyping from tumour tissue of the same patient.

### rs116896264 and rs73933062 are in linkage disequilibrium in multiple populations

Since the two SNPs in the present study were only observed concomitantly, D′ is equal to 1, reflecting the strong linkage disequilibrium.

By mining the 1000 Genomes Project Phase 3 (www.1000genomes.org) (34), in order to assess allele frequencies in different populations, we found rs116896264 and rs73933062 have the identical allele frequencies within each population except from the South Asian (SAS; Figure 3), where there is a slight difference: here, the frequency of the rs116896264 t-allele is 21%, whereas, the corresponding frequency of the rs73933062 T-allele is 22%. Assessing the data of the 1000 genomes project data base on the individual genotype level, only 7 individuals out of 2504 cases harboured one of the SNPs but not the other. So, rs116896264 and rs73933062 are showing strong linkage disequilibrium in the 1000 Genomes project as well, with D′ approximately equal to 1.

## Discussion

In an independent unpublished study, in order to compare gene expression patterns of early and late stages of CRC, we have profiled the expression of different Duke's stages of CRC cell lines and a normal colon cell lines. The correspondence analysis of the expression profiling showed a convergence between the benign polyposis cell line (LT97) and Duke's D stage cell line (KM20L2). LT97 is an early stage of tumour development and derived from early adenoma cells from microadenomas of a patient suffering from hereditary familiar polyposis (33). One of the genes which were disregulated in common between familial adenomatous polyposis cell lines and Duke's D cell line was galectin-4. So, further promoter sequence analysis was performed in order to outline the potential mechanism(s) of galectin-4 upregulation. In parallel, the sequencing analysis of 0.7 KB from *LGALS4* upstream sequence revealed two SNPs (rs116896264 and rs73933062) in LT97 and KM20L2 which are overexpressing galectin-4. Hence, we extended our study to learn more about promoter variation(s) and their influence on the galectin-4 transcription activity. The luciferase reporter of the upstream sequence containing the two SNPs displayed significant upregulation of the promoter activity as compared to the wild-type genotype.

Furthermore, using the galectin-4 regulatory sequence as bait, protein pull down followed by mass spectrometry was done to investigate the regulatory proteins that bind to the promoter with or without the SNPs present. The results showed that the two variations caused deletion and insertion of several regulatory protein binding sites.

The presence of rs116896264 caused generation of several new binding sites like CSTF3, PRKDC, EEF1A, NRF1 and NUP54 and deletion of other response elements such as FLG2 and LBP-1a (UBP1). Similar effects were also found by rs73933062. It introduced CSTF3, EEF1A1 and NR6A1 binding sites and deleted API5, FBXL7, TP63, TLS and TRAP1.

The longer PCR fragment which covers the two SNPs was also used to understand the effect of concomitant presence of rs116896264 and rs73933062 in comparison to the wild type sequence. Among the distinctive proteins found to bind the wild type sequence we observed: (i) Aconitase 1 (ACO1), also known as iron regulatory element binding protein 1 (IREB1), which is a cytosolic protein that regulate ferritin mRNA *via* its binding to iron-responsive elements (IREs). The binding to IREs results in repression of ferritin 5′-UTR mRNA (35). So, galectin-4 could be negatively regulated by ACO1. (ii) RBBP7 (retinoblastoma binding protein-7) which is found previously among several proteins that binds directly to retinoblastoma protein that regulates cell proliferation. The encoded protein is found in many histone deacetylase complexes (36,37). It is also known by limiting the expression of estrogen-responsive genes (38). (iii) PUF60 (poly-U binding splicing factor 60 Kda) also known as FIR (FBP interacting repressor): the protein forms a ternary complex with far upstream element (FUSE) and FUSE-binding proteins. It can repress a c-myc reporter *via* the FUSE. Also, it is known to target transcription factor IIH and inhibits activated transcription (39). PUF60 could be also a potential repressor for galectin-4 promoter *via* binding FUSE-binding proteins. (iv) P53 which is known to act as transcription activator or repressor (40). Also, there are several other proteins with numerous functions ranging from transcription activation to repression such as TP63, DDB2 and DAZ associated protein-1.

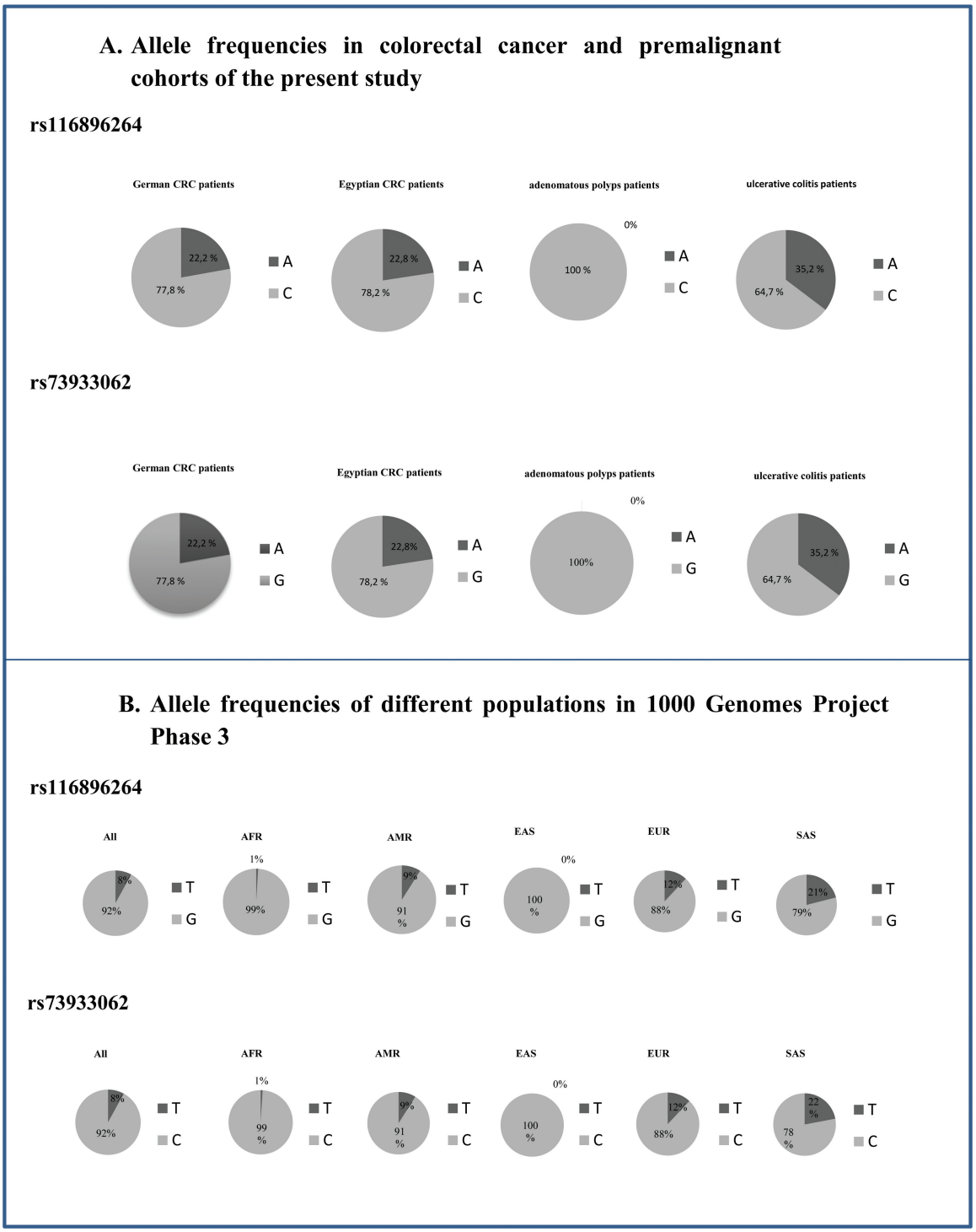


**Figure 3.** Allele frequencies of rs116896264 and rs73933062 in the present study cohort (**A**) and the frequencies of multiple populations, mined from 1000 Genomes Project Phase 3 (**B**).

On the other hand, in the case of concomitant rs116896264 and rs73933062 presence, the two SNPs are inserting new binding sites for regulatory proteins such as: MYB binding protein-1a which is a transcription factor. It is a member of SRC/p160 gene family and has been known as a coactivator (41). However, it is also previously investigated as a novel co-repressor of NF-κB (42).

Both the SNPs investigated in the present study, were originally found together in both cell lines (LT97 and KM20L2). To get a rough estimate of the frequency of the two SNPs in patients, 104 patients with CRC, adenomatous polyps and ulcerative colitis were sequenced. Both rs116896264 and rs73933062 were found in roughly 22% if German and Egyptian CRC patients, 0 of 12 of adenomatous polyp's patients and in 35% ulcerative colitis patients. Both SNPs were found together in each patient. Although this strongly indicated linkage disequilibrium and a D′ close to 1, we further investigated the frequency and potential linkage of the two

SNPs in the published data of the 1000 genomes project (34). Here, we also found very similar allele frequencies of rs116896264 and rs73933062 and the raw data of these populations revealed only a handful of individuals harbouring one of the SNPs without harbouring the other. All in all, both our own data and the data set form the 1000 genomes project, strongly indicate the two SNPs to be linked, with a D′ very close to 1.

Checking the Catalogue of Published Genome-Wide Association Studies (www.genome.gov/gwastudies) (43), rs116896264 and rs73933062 are not included as high risk SNPs in these 23 published GWAS studies. This information could conclude that galectin-4 doesn't have a role in cancer initiation. But on the other hand, a bulk of recent publications highlighted galectin-4 and other galectins as metastatic contributors (16,17,44,45). So, the two SNPs might have a role in the gene upregulation during metastasis process and therefore could reflect prognostic value rather than risk factor.

Summing up our results together, rs116896264 and rs73933062 are linked SNPs and potentially activating galectin-4 expression *via* inserting/deleting new binding sites for transcription activators/repressors *in vitro*. Additional studies using biobank resources will be needed in the future research to conclude the impact of these SNPs in cancer prognosis.

## Funding

Reham Helwa has received Egyptian governmental scholarship. The experimental work was funded by Deutsche Krebsforschungszentrum (DKFZ), Heidelberg, Germany.
Conflict of interest statement: None declared.

## References

1. Shih, W., Chetty, R. and Tsao, M. S. (2005) Expression profiling by microarrays in colorectal cancer (Review). *Oncol. Rep.*, 13, 517–524.
2. Jemal, A., Murray, T., Ward, E., Samuels, A., Tiwari, R. C., Ghafoor, A., Feuer, E. J. and Thun, M. J. (2005) Cancer statistics, 2005. *CA Cancer J. Clin.*, 55, 10–30.
3. Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C. and Parkin, D. M. (2010) Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int. J. Cancer*, 127, 2893–2917.
4. Van Cutsem, E., Cervantes, A., Nordlinger, B., Arnold, D. and Group, E. G. W. (2014) Metastatic colorectal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.*, 25(Suppl 3), iii1–iii9.
5. Shaukat, A., Mongin, S. J., Geisser, M. S., Lederle, F. A., Bond, J. H., Mandel, J. S. and Church, T. R. (2013) Long-term mortality after screening for colorectal cancer. *N. Engl. J. Med.*, 369, 1106–1114.
6. Gabius, H. J. (1997) Animal lectins. *Eur. J. Biochem.*, 243, 543–576.
7. Dumic, J., Dabelic, S. and Flogel, M. (2006) Galectin-3: an open-ended story. *Biochim. Biophys. Acta*, 1760, 616–635.
8. Viguier, M., Advedissian, T., Delacour, D., Poirier, F. and Deshayes, F. (2014) Galectins in epithelial functions. *Tissue Barriers*, 2, e29103.
9. Vasta, G. R. (2009) Roles of galectins in infection. *Nat. Rev. Microbiol.*, 7, 424–438.
10. Gitt, M. A., Colnot, C., Poirier, F., Nani, K. J., Barondes, S. H. and Leffler, H. (1998) Galectin-4 and galectin-6 are two closely related lectins expressed in mouse gastrointestinal tract. *J. Biol. Chem.*, 273, 2954–2960.
11. Huflejt, M. E. and Leffler, H. (2004) Galectin-4 in normal tissues and cancer. *Glycoconj. J.*, 20, 247–255.
12. Rechreche, H., Mallo, G. V., Montalto, G., Dagorn, J. C. and Iovanna, J. L. (1997) Cloning and expression of the mRNA of human galectin-4, an S-type lectin down-regulated in colorectal cancer. *Eur. J. Biochem.*, 248, 225–230.
13. El Leithy, A. A., Helwa, R., Assem, M. M. and Hassan, N. H. (2015) Expression profiling of cancer-related galectins in acute myeloid leukemia. *Tumour Biol.* 36, 7929–7939.
14. Liu, F. T. and Rabinovich, G. A. (2005) Galectins as modulators of tumour progression. *Nat. Rev. Cancer*, 5, 29–41.
15. Newlaczyl, A. U. and Yu, L. G. (2011) Galectin-3—a jack-of-all-trades in cancer. *Cancer Lett.*, 313, 123–128.
16. Chen, C., Duckworth, C. A., Fu, B., Pritchard, D. M., Rhodes, J. M. and Yu, L. G. (2014) Circulating galectins -2, -4 and -8 in cancer patients make important contributions to the increased circulation of several cytokines and chemokines that promote angiogenesis and metastasis. *Br. J. Cancer*, 110, 741–752.
17. Barrow, H., Guo, X., Wandall, H. H., Pedersen, J. W., Fu, B., Zhao, Q., Chen, C., Rhodes, J. M. and Yu, L. G. (2011) Serum galectin-2, -4, and -8 are greatly increased in colon and breast cancer patients and promote cancer cell adhesion to blood vascular endothelium. *Clin. Cancer Res.*, 17, 7035–7046.
18. Barrow, H., Rhodes, J. M. and Yu, L. G. (2013) Simultaneous determination of serum galectin-3 and -4 levels detects metastases in colorectal cancer patients. *Cell Oncol.*, 36, 9–13.
19. Watanabe, M., Takemasa, I., Kaneko, N. *et al.* (2011) Clinical significance of circulating galectins as colorectal cancer markers. *Oncol. Rep.*, 25, 1217–1226.
20. Hayashi, T., Saito, T., Fujimura, T. *et al.* (2013) Galectin-4, a novel predictor for lymph node metastasis in lung adenocarcinoma. *PloS One*, 8, e81883.
21. Nagy, N., Legendre, H., Engels, O. *et al.* (2003) Refined prognostic evaluation in colon carcinoma using immunohistochemical galectin fingerprinting. *Cancer*, 97, 1849–1858.
22. Heinzelmann-Schwarz, V. A., Gardiner-Garden, M., Henshall, S. M. *et al.* (2006) A distinct molecular profile associated with mucinous epithelial ovarian cancer. *Br. J. Cancer*, 94, 904–913.
23. Jiang, C., Xuan, Z., Zhao, F. and Zhang, M. Q. (2007) TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res.*, 35, D137–D140.
24. Zhao, F., Xuan, Z., Liu, L. and Zhang, M. Q. (2005) TRED: a transcriptional regulatory element database and a platform for in silico gene regulation studies. *Nucleic Acids Res.*, 33, D103–D107.
25. Rozen, S. and Skaletsky, H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, 132, 365–386.
26. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V. and Mann, M. (2006) In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.*, 1, 2856–2860.
27. Lohr, J. M., Faissner, R., Koczan, D. *et al.* (2010) Autoantibodies against the exocrine pancreas in autoimmune pancreatitis: gene and protein expression profiling and immunoassays identify pancreatic enzymes as a major target of the inflammatory process. *Am. J. Gastroenterol.*, 105, 2060–2071.
28. Hellman, U., Wernstedt, C., Gonez, J. and Heldin, C. H. (1995) Improvement of an "In-Gel" digestion procedure for the micropreparation of internal protein fragments for amino acid sequencing. *Anal. Biochem.*, 224, 451–455.
29. Zou, J., Yu, X. F., Bao, Z. J. and Dong, J. (2011) Proteome of human colon cancer stem cells: a comparative analysis. *World J. Gastroenterol.*, 17, 1276–1285.
30. Flatmark, K., Maelandsmo, G. M., Martinsen, M., Rasmussen, H. and Fodstad, O. (2004) Twelve colorectal cancer cell lines exhibit highly variable growth and metastatic capacities in an orthotopic model in nude mice. *Eur. J. Cancer*, 40, 1593–1598.
31. Leibovitz, A., Stinson, J. C., McCombs, W. B. III, McCoy, C. E., Mazur, K. C. and Mabry, N. D. (1976) Classification of human colorectal adenocarcinoma cell lines. *Cancer Res.*, 36, 4562–4569.
32. Carrel, S., Sordat, B. and Merenda, C. (1976) Establishment of a cell line (Co-115) from a human colon carcinoma transplanted into nude mice. *Cancer Res.*, 36, 3978–3984.
33. Richter, M., Jurek, D., Wrba, F., Kaserer, K., Wurzer, G., Karner-Hanusch, J. and Marian, B. (2002) Cells obtained from colorectal microadenomas

mirror early premalignant growth patterns in vitro. *Eur J. Cancer*, 38, 1937–1945.
34. Genomes Project Consortium, Abecasis, G. R., Auton, A. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491, 56–65.
35. Wang, W., Di, X., D'Agostino, R. B. Jr, Torti, S. V. and Torti, F. M. (2007) Excess capacity of the iron regulatory protein system. *J. Biol. Chem*., 282, 24650–24659.
36. Nicolas, E., Morales, V., Magnaghi-Jaulin, L., Harel-Bellan, A., Richard-Foy, H. and Trouche, D. (2000) RbAp48 belongs to the histone deacetylase complex that associates with the retinoblastoma protein. *J. Biol. Chem*., 275, 9797–9804.
37. Qian, Y. W. and Lee, E. Y. (1995) Dual retinoblastoma-binding proteins with properties related to a negative regulator of ras in yeast. *J. Biol. Chem*., 270, 25507–25513.
38. Creekmore, A. L., Walt, K. A., Schultz-Norton, J. R., Ziegler, Y. S., McLeod, I. X., Yates, J. R. and Nardulli, A. M. (2008) The role of retinoblastoma-associated proteins 46 and 48 in estrogen receptor alpha mediated gene expression. *Mol. Cell. Endocrinol*., 291, 79–86.
39. Liu, J., Kouzine, F., Nie, Z., Chung, H. J., Elisha-Feil, Z., Weber, A., Zhao, K. and Levens, D. (2006) The FUSE/FBP/FIR/TFIIH system is a molecular machine programming a pulse of c-myc expression. *EMBO J*., 25, 2119–2130.
40. Fischer, M., Steiner, L. and Engeland, K. (2014) The transcription factor p53: not a repressor, solely an activator. *Cell Cycle*, 13, 3037–3058.
41. Anzick, S. L., Kononen, J., Walker, R. L. *et al.* (1997) AIB1, a steroid receptor coactivator amplified in breast and ovarian cancer. *Science*, 277, 965–968.
42. Owen, H. R., Elser, M., Cheung, E., Gersbach, M., Kraus, W. L. and Hottiger, M. O. (2007) MYBBP1a is a novel repressor of NF-kappaB. *J. Mol. Biol*., 366, 725–736.
43. Welter, D., MacArthur, J., Morales, J. *et al.* (2014) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*., 42, D1001–D1006.
44. Reticker-Flynn, N. E. and Bhatia, S. N. (2015) Aberrant glycosylation promotes lung cancer metastasis through adhesion to galectins in the metastatic niche. *Cancer Discov*., 5, 168–181.
45. Dimitroff, C. J. (2015) Galectin-binding O-glycosylations as regulators of malignancy. *Cancer Res*., 75, 3195–3202.