Dissertation

submitted to

The Combined Faculties for the Natural Sciences
and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany

for the degree of
Doctor of Natural Sciences

Presented by
Master Sci. Reham Helwa
born in Cairo, Egypt

Oral examination:

Functional studies on the galectin-4 promoter and its use for establishing a transcription factors array assay

Refrees: PD. Dr. Stefan Wiemann

Prof. Dr. Ruediger Hell

To my wonderful parent who were always behind me from my childhood till this moment.

To my great love, Egypt.

Table of contents

List of tables

List of figures

Abbreviations

ChIP Chromatin immunoprecipitation

CIMP CpG island methylator phenotype

CIN chromosomal instability

CRC colorectal cancer

CRD carbohydrate recognition domain

DamID DNA adenine methyltransferase identification

DMS Dimethyl sulphate

DNase I deoxyribonuclease I

DTT dithiothreitol

E. coli *Escherichia coli*

EDTA ethylendiamintetraacetic acid

EMSA Electrophoretic mobility shift assay

FAP familial adenomatous polyposis

g gram

h hour

HNPCC hereditary non-polyposis colon cancer

kb kilobase

m milli (10^{-3})

M molar = mol/l

min minute

MMR mismatch repair

MSI microsatellite instability

OD optical density

ORF Open Reading Frames

PCR polymerase chain reaction

RNA ribonucleic acid

RT reverse transcription

SDS sodium dodecyl sulphate

SELEX Systematic evolution of ligands by exponential enrichment

SNP single nucleotide polymorphism

SPR Surface plasmon resonance

TF transcription factor

TSGs tumor suppressor genes

TSS transcriptional start site

UC ulcerative colitis

UV ultraviolet

v/v volume/volume

w/v weight/volume

Abstract

Colorectal cancer (CRC) is one of the most common cancers among men and women and accounts for 10% of all new cancer cases and cancer deaths each year. Galectin-4 is expressed in gastrointestinal tissues but not in kidney, brain, skeletal muscles, heart, liver or lung tissues. The main aim of the present study was to define candidate gene(s) for colorectal cancer prognosis and to study its regulation. Using galectin-4 promoter, another aim for this study was establishing protein microarray conditions to investigate DNA-protein interaction.

In the present study, I profiled the expression pattern of different stages of six colorectal cancer and adenoma cell lines (SW1116, SW480, SW620, Co115, KM20L2, and LT97 in comparison to normal colon cell line CCD-18co). In our expression profiling data, we have found that galectin-4 was among the genes that are significantly upregulated in LT97 and KM20L2. We investigated galectin-4 upregulation as a signature of bad prognosis in colorectal cancer cell lines. In addition, we identified one possible mechanism of galectin-4 upregulation which is associated with a twin SNPs in the upstream sequence and the first intron. From the sequencing results of 115 patients, 26 out of the 115 (22.6%) were found to have the two SNPs together (ss217320370 and rs73933062), while the other patients did not carry any of these two SNPs. Therefore, we firstly showed that ss217320370 C>A at position 11572034 (a novel SNP was identified in the present study) and G>A at position 11571652 (rs73933062) are always together in the same individual (twin SNPs). Since the twin SNPs could potentially be associated with galectin-4 upregulation via deletion and insertion of new transcription factor binding sites, the twin SNPs may have medical impact in colorectal cancer patient.

In parallel, another project was carried out to establish transcription factor protein array for DNA-protein interaction detection. For this purpose, fifty transcription factors (TF) proteins were expressed and purified. Then, epoxy surface was optimised to immobilize TF proteins at 37ºC overnight. The DNA-protein interaction was optimised on-chip using retrieved short polynucleotide sequences from TransFac database and PCR product of galectin-4 promoter. Using galectin-4 promoter and its SNPs, a suitable condition for DNA-protein was established on microarray surface. Interestingly, the results of pull down- mass spectrometry were compatible with the TF-array results.

In conclusion, upregulation of galectin-4 was found to be associated with bad prognosis of colorectal cancer. In addition, two SNPs (ss217320370 and rs73933062) were found to be associated with galectin-4 upregulation, which might be referred to the changing in the binding sequence of the regulatory elements. Using galectin-4 promoter with the twin SNPs was useful in setting up TFs-array to detect DNA-protein interactions, since the microarray result was concordant with pull down-mass spectrometry analysis of galectin-4 promoter.

Zusammenfassung

Kolorektales Karzinom (CRC) gehört zu einer der meist häufigsten Krebsart bei Männern und Frauen und zählt jedes Jahr zu 10% aller neuen Krebsfälle und Krebs-bedingten Todesfällen. Galectin-4 wird in gastrointestinalen Geweben exprimiert und nicht in Niere, Hirn, Skelettmuskulatur, Herz, Leber oder Lunge.
Ziel dieser Arbeit war es, Kandidatengene für die Prognose von kolorektalem Krebs zu finden und deren Regulation zu untersuchen. Ein weiteres Ziel war die Etablierung eines Protein-Microarrays, um die DNA-Protein-Interaktion unter Verwendung des Gal-4 Promoters zu erforschen.

In der vorliegenden Studie habe ich das Expressionsmuster von sechs verschiedenen Stadien von kolorektalem Karzinom und Adenoma Zelllinien (SW1116, SW480, SW620, Co115, KM20L2 und LT97) im Vergleich zur normalen kolorektalen Zellinie CCD-18co profiliert. In den Expressionsprofilierungsdaten fanden wir Galectin-4 unter denjenigen Genen, die signifikant in LT97 und KM20L2 hoch reguliert waren. Wir stellten die Hochregulation von Galectin-4 in kolorektalen Krebszelllinien als ein Zeichen für schlechte Prognose fest. Außerdem identifizierten wir einen möglichen Mechanismus der Hochregulation von Galectin-4 mit einem assoziierten „Zwillings-SNP“ in der upstream Sequenz und dem ersten Intron. Von 115 Patienten-Sequenzierergebnissen wurden in 26 Patienten (22,6%) die zwei SNPs (ss217320370 und rs73933062) gefunden. Wir haben zunächst gezeigt, dass ss217320370 C>A an der Position 11572034 (ein neues SNP wurde in der vorliegenden Studie identifiziert) und G>A an der Position 11571652 (rs73933062) immer im gleichen Individuum vorkommen („Zwillings-SNP“). Da diese Zwillings-SNPs mit der Hochregulation von Galectin-4 via Deletion und Insertion von neuen Transkriptionsfaktoren Bindestellen assoziiert sind, könnten die Zwillings-SNPs eine mögliche medizinischen Auswirkung bei Patienten mit kolorektalem Krebs haben.

Parallel wurde ein weiteres Projekt durchgeführt, um einen Transkriptionsfaktor-Array für die Detektion von DNA-Protein Interaktion zu etablieren. Zu diesem Zweck wurden 50 Transkriptionsfaktoren (TF)-Proteine exprimiert und aufgereinigt. Daraufhin wurde eine Epoxy-Oberfläche optimiert, um die TF-Proteine bei 37°C übernacht zu immobilisieren. Die DNA-Protein Interaktion wurde auf dem Array optimiert unter Verwendung von kurzen Polynukleotid Sequenzen aus der TransFac Datenbank und PCR-Produkten des Galectin-4

Promoters. Unter Verwendung des Galectin-4 Promoters und deren SNPs wurden geeignete Bedingungen für DNA und Proteine auf der Microarray Oberfläche etabliert. Interessanterweise waren die Ergebnisse der Pull-down-Massen Spektrometrie Ergebnisse kompatibel mit denen der Transkriptionsfaktoren-Arrays.

Schließlich wurde eine Hochregulation von Gal-4 assoziiert mit einer schlechten Prognose von kolorektalem Krebs gefunden. Außerdem wurde heraus gefunden, dass zwei SNPs (ss217320370 und rs73933062) mit der Gal-4 Hochregulation assoziiert sind, was der Veränderung in der Bindesequenz des regulatorischen Elements zugeschrieben werden könnte. Die Verwendung des Galectin-4 Promoters mit den Zwillings-SNPs war nützlich, um einen Transkriptionfaktor-Array für die Detektion von DNA-Protein Interaktionen einzurichten, da die Microarray-Ergebnisse einstimmig mit den Pull-down Massenspektrometrie Analysen des Galectin-4 Promoters waren.

Part I: Expression Profiling Analysis of Colorectal Cancer Cell Lines: Reveals a Twin SNPs in Galectin-4 Promoter Associated with its Upregulation

Introduction

Colorectal cancer (CRC) is the third common cancer in North America in both males and females and accounts for 10% of all new cancer cases and cancer deaths each year [1,2]. Improvement of the overall 5-year survival rate from colon cancer is due to the early detection from increased screening. CRCs arise mostly from adenomatous precursors, and accumulation of mutations in proto-oncogenes and tumor suppressor genes (TSGs) leads to progression of adenomatous lesions to carcinoma [3,4,5]. High-throughput expression and genotyping arrays are starting to generate novel markers and gene signatures that may be of use in the management of colorectal cancer. At present, these are not sufficiently validated to be clinically useful [6].

Galectins are family of animal lectins [7]. They are characterized by their ability to recognizes β-galactose and their consensus amino acids sequences [8]. All galectins contain highly conserved carbohydrate-recognition domains (CRDs) of 130 amino acids responsible for carbohydrate binding [9,10]. In gastrointestinal tract, RNAse protection assays showed that galectin-4 is expressed in gastrointestinal tissues but not in kidney, brain, skeletal muscles, heart, liver or lung tissues. Upon expression of galectin-4, epithelial cells acquired a phenotype characterized by the ability to survive lack of nutrients and growth factors for the prolonged period of time. Thus, this cellular phenotype is likely to be advantageous in hyperplastic tissues of premalignant and malignant tumors [11]. Galectin-4 has been reported to be downregulated in colon cancers, suggesting an implication in early colorectal cancer carcinogenesis [12].

In the present study, dependent on our expression profiling results, we have proceeded to focus on galectin-4. We have shown galectin-4 upregulation in two cell lines with adenoma and Duke's D colon cancer which was a matter of interest. Therefore, the microarray results were validated and the upstream genotyping was studied. We have found two SNPs, one in the promoter region and the potential regulatory sequence after the transcription start site. Moreover, the two SNPs are always together (a twin SNP). Interestingly, the presence of the twin SNPs was associated with galectin-4 upregulation in the cell lines which we also confirmed it by the promoter reporter assay.

1.1. Colorectal cancer (CRC)

Colorectal cancer is a major public health problem in the developed countries, since there are nearly one million new cases diagnosed worldwide each year and half a million deaths. Recent reports show that CRC is the third most frequent cancer in the Western world. In the USA it is the most frequent form of cancer among people aged 75 years and older [13,14,15].

1.1.1. Molecular carcinogenesis

Although there are many alternative pathways to develop colorectal cancer [16], the classic model of colorectal cancer is the multistep adenoma-carcinoma pathway that is determined by gatekeeper and caretaker molecular pathways [17]. In the adenoma-carcinoma model, colorectal carcinoma arises through a series of well-characterized histopathological changes as a result of specific genetic hits in oncogenes and tumor suppressor genes. The analysis of the genetic alteration in parallel to the histopathological changes lead to the development of the model (figure 1) for the clonal evolution of colorectal tumor by acquisition of sequential mutations [4].

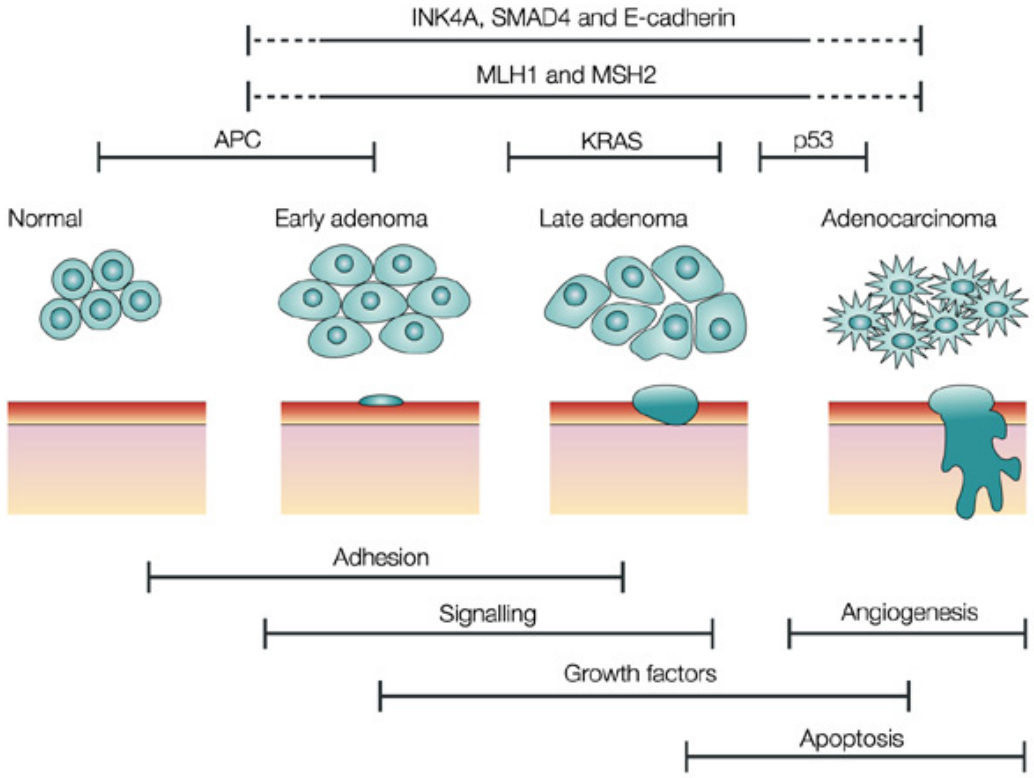


Figure 1. Adenoma-carcinoma multistep model. The transition from normal colon epithelium to premalignant adenoma and then invasive adenocarcinoma is associated in parallel with genetic alterations in several genes [18].

About 85% of sporadic colorectal cancer cases have chromosomal instability (CIN), an allelic imbalance at several chromosomal loci (including 5q, 8p, 17p, and 18q), chromosomal amplification, and chromosomal translocation, which together contribute to tumor aneuploidy [17,19,20,21]. On the other hand, the rest of cases 15% are showing microsatellite instability (MSI) [22,23], which is indicating the instability of DNA mismatch repair (MMR) system. MSI is controlled by several genes (e.g. MLH1, MSH2, and MSH6) and characterized by the accumulation of single nucleotide mutation and length alterations in the microsatellite sequence [24]. Tumors with MSI are characterized by proximal location, mucinous histology, poor differentiation, and lymphocytic infiltration. The molecular mechanism underlying the MSI phenotype is loss of mismatch repair genes function through mutation or epigenetic silencing [25,26]. CpG island methylator phenotype (CIMP) overlap with MSI and seems to have association with MSI in patients who don't have germline mutation in the MMR genes, which are characteristic of hereditary non-polyposis colon cancer HNPCC (Lynch syndrome) [27].

1.1.2. Risk factors

The most colorectal cancers arise are sporadic. Several risk factors are increasing the predisposition to colorectal cancer. Among these factors are increasing the age, gender (higher risk in males), previous colonic polyps, and environmental factors (e.g. red meat, high fat, diabetes mellitus, obesity, low fibers intake, smoking, and high alcohol consumption) [28]. Moreover, inflammatory bowel diseases (ulcerative colitis and Crohn's disease) represent roughly two-thirds of colorectal cancer incidence [29,30], and the risk is proportional to the duration of illness [30] and severity of inflammation [31]. In the hereditary syndrome, the HNPCC occurs in one in 300 CRC patients and one familial adenomatous polyposis (FAP) arises in 7000 of CRC cases [32].

1.1.3. Colorectal cancer screening

When screening is recommended, it generally starts with fecal occult blood testing (FOBT). Many colorectal cancers bleed into intestinal lumen, and fecal occult blood tests can detect the presence of blood that is otherwise unapparent through simple inspection. As blood passes through the gastrointestinal tract, it becomes degraded, and depending upon the site at which the hemorrhage occurs, blood products in the stool will vary. For example, if a lesion is located in the proximal colon, hemoglobin will be completely digested and will be metabolized by bacteria into porphyrins. However, if the lesion is in the distal colon, hemoglobin and heme will remain mostly intact and unchanged. The design of the FOBT, therefore, must take these

variations into consideration. There are several different FOBTs currently available in the market. The most commonly used is the Hemocult II test [33]. Stool is placed on guaiac impregnated test pads that detect the peroxidase-like activity of heme by changing the color of the pad from white to blue. At least three samples must be collected since most colorectal lesions have an intermittent bleeding pattern. Red meat and certain fresh fruits and vegetables can lead to false positive results. These food sources, therefore, should be avoided for three to five days prior to and during testing. Additionally, vitamin C intake should be avoided over the same time period because it can inhibit the guaiac reaction leading to false negative results. Nonsteroidal anti-inflammatory drugs, alcohol, or any other gastric irritants should also be avoided for a three-day period prior to collection, as they may increase gastrointestinal bleeding [34].

Since more than 60% of early lesions were believed to arise in the rectosegmoid areas of the large intestine, rigid Sigmoidoscopy was routinely used for screening in the past. Over the past several decades, however, there has been an increase in the number of lesions arising from more proximal regions of the colon, necessitating the use of flexible, fiberoptic sigmoidosopes. Although these methods are very effective and offer a means of removing neoplastic polyps, they still leave all lesions beyond the reach of the scope (estimated between 25% and 34%) undetected. Colonoscopy, based upon the same principle as sigmoidoscopy, allows visualization of the entire colon. Although, it is the gold standard in colorectal cancer screening, it is very expensive, requires cathartic preparation, and has an increased risk of morbidity and mortality [35].

The feasibility of fecal DNA screening as an approach to colorectal cancer detection was first shown in 1992 with successful assay of mutant k-ras in stool [36]. Stool-based DNA testing is a promising new diagnostic tool with potential to improve the overall effectiveness of colorectal cancer screening. Early clinical studies suggest that multi-target, DNA-based stool test is capable of detecting both premalignant adenomas and cancers with high sensitivity and specificity. Screening compliance could be enhanced by the user-friendly features of stool-based DNA testing which include: the testing is noninvasive, requires no cathartic bowel preparation or diet or medication restrictions, and requires just a single specimen per screen [37].

1.1.4. Colorectal Cancer staging

Currently, the recommended staging system is TNM classification, but also Duke's classification is being used.

1.1.4.1. Duke's classification

Duke's system was originally described to be used in rectal carcinoma, but it is also applicable in colon cancer. It classifies cancer in the following stages: 1) Duke's A, in which tumor is confined to the intestinal wall, 2) Duke's B; tumor invading the intestinal wall (serosa and/or subserosa), 3) stage C is characterized by lymph node infiltration, and 4) the tumor cells are migrating and forming secondary tumor (metastasis) in Duke's D [38].

1.1.4.2. TNM staging

TNM (tumor/ node/ metastasis) is the most common system that used to classify colorectal cancer. T is indicating the degree of tumor invasion to the intestinal wall. Lymph node involvement is denoted by N. While M represents the metastasis status. According to the degree of the involvement, it is again subgrouped as the following:

- Tis: tumor is restricted to the intestinal wall (tumor *in situ*).
- T1: tumor invades the submucosa, but is not penetrating muscularis propria.
- T2: tumor is invading muscularis propria.
- T3: tumor invades subserosa, but not into peritoneal cavity or other organs.
- T4: tumor invades other organ(s) or perforates the perineal cavity.
- Nx: nodal invasion can not be assessed.
- N0: there is no nodal metastasis.
- N1: 1-3 pericolic/perirectal nodes involved.
- N2: 4 or more pericolic/perirectal nodes involved.
- Mx: Distant metastasis can not be assessed.
- M0: No distant metastases.
- M1: Distant metastases.

1.1.5. Prognosis

The clinicopathological staging is still the gold standard methods for prognosis. Histopathological examination of the tumor materials could define the tumor status. Mostly, TNM classification is used, but also Duke's staging system is applied using lymphovascular

invasion and tumor grade [38,39]. Some other factors could also influence the clinical outcome like obstruction and perforation and pre-operative carcinoembryonic antigen levels [40,41].

Several proteins and genetic markers have been described. The prognostic markers are associated with survival that is apart from the treatment effect. Predictive markers are treatment indicators that could help to avoid toxicities associated with systemic therapy in patients who will not benefit from this treatment [42]. *KRAS* mutations as a predictive marker in EGFR-targeted treatment are the most important marker for metastatic colorectal cancer. Mutations in exon 2 (codons 12 and 13) and exon 3 (codon 61) of *KRAS* are associated with the onset of the adenocarcinoma pathway in approximately one-third of CRC patients [43]. Mutations in one of the three codons compromise the ability of GTPase-activating proteins to affect the inactivating hydrolysis of Ras-bound GTP to GDP [44]. Consequently, 99% of patients with mutated KRAS do not respond to EGFR inhibition [45].

1.2. Galectins

Galectins are a family of animal lectins. They are characterized by their ability to recognize β-galactose and by their consensus sequences [8]. Galectins show a high level of evolutionary conservation. Members of this family are present in organisms from nematodes to mammals [46]. All galectins have conserved carbohydrate recognition domain (CRD) of about 130 amino acids, which is responsible for carbohydrate binding. Fifteen different galectins in mammals have been identified and classified according to the CRD to 1) prototype galectins (galectin-1, -2, -5, -7, -10, -11, -13, -14, and -15) with one CRD, 2) tandem-repeat type galectins (galectin-4, -6, -8, -9, and -12) that consist of two CRDs in single polypeptide chain, and 3) chimera type like galectin-3 that contains one CRD and tandem repeats of short amino acids stretches fused to the CRD [9,10].

Galectins are involved in several cellular functions. They play different roles in cell-cell and cell-matrix adhesion [47]. Some galectins are secreted from the cell through non-classical pathway, since galectins a signal sequence that is required for the classical secretion pathway [48,49]. Also, galectins could be found as intracellular proteins in different subcellular locations [50]. Moreover, galectins play a rule in regulating cell survival and signaling, chemotaxis, and cytokines secretion and consequently acting as a regulator of immune cell homeostasis and inflammation [51,52].

Some galectins are distributed in a wide variety of tissues, whereas others are more tissue specific. The biological rules of galectins are dependent on cellular localization, and cell type involved as well [53,54]. Galectins display specific expression pattern in normal as well as in pathological situations. Therefore, they could act as interesting biomarkers for diagnosis and prognosis as well as targeted therapies [53].

1.2.1. Galectins expression in normal alimentary canal

Nine galectin subtypes have been identified in the mammalian digestive tract (galectin-1, -2, -3, -4, -6, -7, -8, -9, and -15). Among those, galectin-2, -6, -7, -9, and -15 were detected only in the non-human species. In humans, galectin-1 and -8 are located in both nucleus and cytoplasm of colonic epithelial calls and stromal cells [55,56,57]. Galectin-4 is only expressed in the epithelium of gastrointestinal tract from tongue to colon in the adult and also during the development [11,58].

1.2.2. Galectins in colorectal cancer

Several galectins are involved in the pathogenesis of colorectal cancer, since they are contributed in several cellular processes. Therefore, according the rule of galectin(s) in tumorigenesis process, they could act as cancer prognostic and predictive markers.

Galectin-1 overexpression is associated with dysplastic transformation [55,56] and tumor progression [55,56,59]. This overexpression is mainly detected in stromal cells [60]. Previously, galectin-1 overexpression is shown to have a prognostic value [59]. Moreover, a previous gene expression profile study found that the responders to pre-operative radiotherapy show high level of galectin-1 comparable to the non-responder rectal cancer patients. Therefore, it could be used as a therapeutic indicator [60].

Increased galectin-3 expression in colorectal cancer is associated with neoplastic progression [56,59,61,62,63,64,65]. Additionally, marked changes in the subcellular location of galectin-3 occur during tumor progression, with loss of nuclear galectin-3 in colon cancer cells [66]. At the beginning of tumor progression, nuclear galectin-3 is downregulated. On the other hand, in the late stage, the cytoplasmic expression increases [55]. Several studies considered that galectin-3 upregulation is associated with poor prognosis [55,59,62,63,64]. Interestingly, it has been found that galectin-3 upregulates MUC2 transcription through AP-1 activation [67]. MUC2 is strongly expressed in mucinous carcinomas [68]; with consideration that mucinous colorectal cancer represents advanced disease stage [69]. Elevated level of serum galectin-3

comparable to healthy controls is proportional to tumor progression with the maximum level of galectin-3 concentration in the metastatic patients. Thus, galectin levels in the serum samples could be interesting as a clinical tool [70].

Galectin-4 upregulation is also associated with poor prognosis. The combination of galectin-4 and galectin-1 expression was shown to provide more useful indication for colorectal cancer prognosis than galectin-3 [59]. It has been shown previously that T84 human colon adenocarcinoma cells exhibiting galectin-3 and galectin-4 localization in lamellipodia domains, which could indicate their function in cell-substrate interaction [71].

1.2.3. Galectin-4

Galectin-4 and an independently discovered 32 kD galectin from *C. elegans* were the first defined galectins with two CRD in the same polypeptide chain [72]. Rat and human galectin-4 consist of CRDs of 133aa and 130aa respectively, connected by a link peptide of 34aa and preceded by 17aa [71,73]. Galectin-4 has two distinctive properties: it tends to be associated with the insoluble cellular components, and the link peptide is sensitive to proteolysis [74]. In a previous study, RNA protection assay showed that galectin-4 is expressed in gastrointestinal tissues, but not in brain, kidney, skeletal muscles, heart, liver or lung tissues [11]. Through the nonclassical secretory pathway, galectin-4 can be secreted from basolateral and apical sides of the intestinal epithelial cells [58,75].

From the function point of view, the localization and relative insolubility of galectin-4 in normal epithelia indicate roles in stabilization of cellular junctions and membranes. As the other galectins, galectin-4 mediates cell adhesion [71]. In a previous study, wild type and mock transfected MDCK cells were unable to grow further and after 7-10 days, they underwent apoptosis. On the other hand, MDCK cells that were transfected with the full-length cDNA of galectin-4 and expressing this protein were able to survive. Accordingly, the expression of galectin-4 could offer the epithelial cells a chance to survive under lack of nutrients and growth factors for prolonged period of time. Therefore, such a cellular phenotype is likely to be advantageous in hyperplastic tissues of premalignant and malignant tumors [11].

Previous studies referred the expression of high galectin in the intestinal mucosa to the rule of galectin in the immune system, since it serves as a pathogen recognition protein [76,77]. Previous studies suggest that several galectins may recognize blood group antigens [78].

Recently, it has been reported that galectin-4 and galectin-8 recognize and kill human blood group antigen-expressing *E. coli* while failed to alter the viability of other *E. coli* strains or other Gram-negative or Gram-positive organisms both *in vivo* and *in vitro* [79].

The aim of this part of the thesis was according to the experiment as the following:

1. The expression profiling of colorectal cancer cell lines was to find the proper gene candidate(s) according to the stage which could have diagnostic and prognostic value.
2. Galectin-4 promoter study was to understand the gene regulation and the impact of SNPs on promoter activity. Meanwhile, the twin SNPs and their effect on the binding activity of galectin-4's promoter was valuable in TFs-array application.

Material and Methods

2.1. Materials

2.1.1. Chemicals

Name	Manufacturer
2′-Deoxyadenosine 5′-Triphosphate	MBI Fermentas, St. Leon-Rot
2′-Deoxycytidine 5′-Triphosphate	MBI Fermentas, St. Leon-Rot
2′-Deoxyguanosine 5′-Triphosphate	MBI Fermentas, St. Leon-Rot
2′-Deoxythymidine 5′-Triphosphate	MBI Fermentas, St. Leon-Rot
Agar	Roth, Karlsruhe, Germany
Agarose	Invitrogen, Carlsbad, USA
Ammoniumpersulfate (APS)	Roth, Karlsruhe, Germany
Ampicillin (sodium salt)	Genaxxon, Munich, Germany
Dimethylsulfoxide (DMSO)	Sigma-Aldrich, Deisenhofen
EDTA	Sigma-Aldrich, St.Louis, USA
Ethanol, absolute	Riedel de Haën Sigma, Seelze, Germany
Ethidium bromide	AppliChem, Darmstadt, Germany
First strand-buffer	Invitrogen, Karlsruhe, Germany
FluoroLink™ Cy3-dCTP. Cy5-dCTP	Amersham-Pharmacia-Biotech, Freiburg, Germany
Glycerol	J.T.Baker, Deventer, Holland
PCR- buffer, 10x	Qiagen, Hilden, Germany
Penicillin/Streptomycin	Invitrogen, Karlsruhe, Germany
Random Hexamer Primer (pd(N)6)	Amersham-Pharmacia-Biotech, Freiburg, Germany
SlideHyb™-Hybridization buffer 1	Ambion, Huntingdon, Cambridgeshire, UK
Sodium chloride	Riedel de Haën Sigma, Seelze, Germany
TEMED	Roth, Karlsruhe, Germany
Tetracyclin	Genaxxon, München, Germany
Tris-base	Sigma-Aldrich, St.Louis, USA
Tris-HCl	Sigma-Aldrich, St.Louis, USA

Triton X 100	Gerbu, Gaiberg, Germany
Triton- X 100	Sigma-Aldrich, Deisenhofen
Trypsin/ EDTA	Invitrogen, Karlsruhe , Germany
Tryptone/Peptone	Roth, Karlsruhe, Germany
Tween 20	Sigma-Aldrich, St.Louis, USA
Yeast extract	Gerbu, Gaiberg, Germany
β-Mercaptoethanol	Sigma-Aldrich, Deisenhofen

2.1.2. Enzymes

Name	**Manufacturer**
HindIII	NEB, Frankfurt, Germany
NheI	NEB, Frankfurt, Germany
RNAse A, lyophilized	Sigma-Aldrich, Deisenhofen
RNAse H (2 U/µl)	Invitrogen, Karlsruhe, Germany
RNAse Out–Inhibitor (40 U/µl)	Invitrogen, Karlsruhe, Germany
Superscript ™ III RT (200 U/µl)	Invitrogen, Karlsruhe, Germany
XhoI	NEB, Frankfurt, Germany

2.1.3. Kits

Name	**Manufacturer**
Allprep DNA/RNA/Protein minikit	Qiagen, Hilden, Germany
Dual-luciferase reporter assay system	Promega, Madison, WI
ECL Western Blotting Detection System	Amersham Pharmacia Biotech, Buckinghamshire, UK
Effectene transfection reagent	Qiagen, Hilden, Germany
Pierce Pull-Down Biotinylated Protein:Protein Interaction kit	Thermo Scientific, Rockford, USA
QIAquick PCR-Purification Kit	Qiagen, Hilden, Germany
Qproteome Mammalian Protein Kit	Qiagen, Hilden, Germany
QuantiFast SYBR Green RT-PCR Kit	Qiagen, Hilden, Germany
QuantiFast SYBR Green RT-PCR Kit	Qiagen, Hilden, Germany

2.1.4. Buffers

Name	Composition
Binding buffer 5X	0.1M Hepes pH 8.0, 0.25M KCl, 25mM DTT, 0.25mM EDTA, 5mM $MgCl_2$, 25% v/v glycerol
Ethidium bromide solution for staining	0,5 µg/ml End conzentration
Laemmli buffer	30,1 g Tris Base, 144,2 g Glycine, 50 ml SDS (20%), ad 1 l dH2O
LB-Agar	LB-Medium + 1,5% (w/v) Agar
LB-Medium (1 liter)	10 g Tryptone/Pepton, 5 g yeast extract, 10 g NaCl, pH 7.2
SSC (20x)	15,36% (w/w) NaCl, 7,73% (w/w) Tri-sodium citrate, 76,91% (w/w) d H_2O
TAE buffer (50x)	0.4 M Tris Base, 0.4 M acetic acid, 20 mM EDTA (pH 8)
TBS 10x (1 liter)	50 mM Tris, 150 mM NaCl mit HCl, pH 7.5
TBST	1X TBS/ 0,05% Tween20
Transfer buffer	150 mM Glycine, 25 mM Tris-base, 20% Ethanol

2.1.5. Consumables

Name	Manufacturer
6-well cell culture plate	Griener Bio One, Frickenhausen
Adhesive foil for microtiter plates	Nalge Nunc Int., Rochester, NY
Cell culture flask, 250 ml	Griener Bio One, Frickenhausen
Cell culture flask, 75 ml	Griener Bio One, Frickenhausen
Cell culture Petri dishes 96x20mm	stock
Cover slip	Erie Scientific Company, Portsmouth, USA
Falcon 15 ml and 50 ml	Becton Dickinson, Heidelberg
Lazy-L-Spreaders	Sigma-Aldrich, St.Louis, USA
Nitril gloves	Nitril, Microflex, Wien, Austria
Parafilm	PM 996, Pechiney Plastic packaging

PCR-plate, 96er	Steinbrenner, Neckargemünd
Reaction tubes 1,5 ml und 2 ml	Eppendorf, Hamburg
Sterile filter 500 ml	Nalgene, Rochester, USA
UV cuvette 220-1600 nm	Eppendorf, Hamburg
Lazy-L-Spreaders	Sigma-Aldrich, St.Louis, USA

2.1.6. Equipment

Name	**Manufacturer**
Centrifuge	5810R, Eppendorf, Hamburg Germany
Centrifuge	5415D, Eppendorf, Hamburg Germany
Hoefer TE 70 (Semi dry Wester-Blot)	Amersham Bioscience, Piscataway, USA
LightCycler 480	Roche Diagnostics, Mannheim
MicroGrid spotter	BioRobotics, Apogent Discoveries, USA
Mini-Protean®3 electrophoresis chamber and casting unit	BioRad, Waltham, USA
ND 1000, NanoDrop, Spectrophotometer	NanoDrop, Wilmington, DE USA
Oven	Haereus, Hanau, Germany
pH-Meter	MP 230, Mettler Toledo, Germany
Photometer Nanodrop ND-1000	
ScanArray 4000XL	Perkin Elmer, Boston, MA, USA
ScanArray 5000	Perkin Elmer, Boston, MA, USA
Vacuum-concentrator	H.Saur Laborbedarf, Reutlingen

2.1.7. Antibodies and electrophoresis ladders

Name	**Manufacturer**
GeneRuler 1 kb DNA marker	MBI Fermentas, St. Leon-Roth
GeneRuler™ 100 bp DNA marker	MBI Fermentas, St. Leon-Roth
Anti human galectin-4	R&D system, Minneapolis, MN, USA

2.1.8. Primers

Primer	**Sequence**
qRT-PCR *LGALS4*	Hs_LGALS4_1_SG QuantiTect Primer Assay, Qiagen, Hilden, Germany

LGALS4-700F	TTCACAGTTGCTGGGAGAGG.
LGALS4-700R	GATGACGAGGGCCAACAGTTAGACGTG
LGALS4 XhoI F	AAAAAAACTCGAGTTCACAGTTGCTGGGAGAGG
LGALS4 HindIII R	GGGGGGAAGCTTGCTGCGCTAGTGGCTGGTC
LGALS4 var2 NheI F	AAAAAAAGCTAGCCCACCATCTCCCACTCCTG
LGALS4 var2 HindIII R	TTTGGGGAAGCTTGATGACGAGGGCCAACAGTTAGA
LGALS4 2var NheI F	AAAAAAAGCTAGCTTCACAGTTGCTGGGAGAGG
LGALS4 2var HindIII R	TTTGGGGAAGCTTGATGACGAGGGCCAACAGTTAGA
ACO1 F	gcaggcaccacagactatcc
ACO1 R	cagcagcatcaaacacatca
ENO1 F	tcccaacatcctggagaataa
ENO1 R	atgccgatgaccaccttatc
FUBP1 F	gggctgcttattacgctcac
FUBP1 R	tggattctgctgatctccttg
HNRNPK F	cagacgccattatcctctgtt
HNRNPK R	cccagtgctgcagtagcc
FUBP2 F	gccgcttactacggacagac
FUBP2 R	acattcattcgattcattgagc
MYBBP1a F	ggcatccacctcctcaagt
MYBBP1a R	agacttttcctgatgcaggtct
PRDX1 F	cactgacaaacatggggaagt
PRDX1 R	tttgctcttttggacatcagg
PUF60 F	ggcgaccatagctctcca
PUF60 R	cctgtggaggtttccatttg
RBBP7 F	acgcaagatggcgagtaaag
RBBP7 R	cggtgtattcttcttccagatttt
FUS F	ggccagagcagctattcttc
FUS R	ggggagttgactgagttcca

2.2. Methodology

2.2.1. Cell culture

Seven cell lines have been used in the present study, a normal colon cell line CCD-18Co (ATCC) and LT97 as adenoma cell line (as a gift from Brigitte Marian) and SW1116, Sw480, Co115, SW620, and KM20L2 from Duke's stage A, B, C primary tumor, C lymph node infiltration, and D respectively (as a gift from Gabriela Aust, Heike Allgayer, Richard Iggo, Francis RAUL, Øystein Fodstad). The maintenance of the cells was carried out in the proper medium for each cell line at 37°C in a humidified atmosphere of 5% CO_2 and 95% air.

2.2.2. mRNA Expression Profiling

2.2.2.1. Microarray production

Amino-modified PCR products of 3872 human cDNAs have been arrayed in duplicate on epoxysilane surface (Schohtt Nexterion AG, Jena, Germany) with a MicroGridII arrayer (BioRobotics, Cambridge, UK) using SMP3 pins (TeleChem International Inc., Sunnyvale, Calif., USA).

2.2.2.2. Sample preparation and hybridization

DNA, RNA, and protein samples have been isolated from cell pellets with Allprep DNA/RNA/Protein minikit (Qiagen, Hilden, Germany). Flourescently labeled cDNAs were prepared from 10µg total RNA by incorporation of Cy3- or Cy5- labeled dCTP (Amersham Bioscience, Freiburg, Germany) during the first strand synthesis. The hybridization has been done in SlideHyb buffer 1 (Ambion Inc., Austin, Tex., USA) under glass cover slips at 62°C in water bath overnight.

2.2.2.3. Detection and analysis

Following the overnight hybridization, the slides have been washed in 0.1X SSC (15mM sodium chloride and 1.5mM sodium citrate) for 3 minutes and dried by nitrogen. Subsequently, the fluorescence signals have been detected using ScanArray5000 confocal laser scanner (Packard, Billerica, Mass., USA). Each cell line cDNA was hybridised versus the normal colon cells cDNA four times. Moreover, the fluorescent labelling has been switched every hybridization. Finally, at least eight data points per gene and individual experiment condition (duplicate spots on each array, four hybridizations per sample inclusive dye swap) were obtained. The signal intensities have been quantified by GenePix Pro 4.1 analysis

software (Axon Instruments Inc., Union City, Calif., USA). Data quality assessment, normalization and correspondence cluster analysis were performed with the MIAME-compatible [80] analysis and data warehouse software M-ChiPS [81].

Cluster analyses have been performed using correspondence analysis [82], which is an explorative computational method for the investigation of associations between variables, such as genes and patient samples, in a multidimensional space. It simultaneously displays data for two or more variables in a low-dimensional projection, thus revealing associations between them. Additionally, the cluster analysis has been repeated using MultiExperiment viewer (MeV_4). MeV is a desktop application for the analysis, visualization and data mining of large-scale genomic data. It is a versatile microarray tool, incorporating sophisticated algorithms for clustering, visualization, and classification [83,84].

2.2.3. Galectin-4 validation by qRT-PCR

Since galectin-4 was significantly upregulated in LT97 (adenoma cell line) and KM20L2 (Duke's D cell line) while not regulated in the other cell lines, galectin-4 was one of the interesting differentially regulated genes in the expression profiling. Therefore, we chose galectin-4 for further validation by qRT-PCR and subsequently detailed study of promoter variations using Hs_LGALS4_1_SG QuantiTect Primer Assay and one-step QuantiFast SYBR Green RT-PCR Kit (Qiagen, Hilden, Germany) and the Light Cycler 480 instrument (Roche Diagnostics) have been used. We started with 10ng of RNA, following the manufacturer's instructions, a reverse transcription step has been done at 50°C for 10 minutes followed by PCR initial activation step for 5 minutes at 95°C. Afterward, 40 cycles of 10 seconds denatuaration at 95°C, 20 seconds annealing at 55°C, and extension at 68°C for 20 seconds. A final melting curve to check fidelity was done from 95°C for 5 seconds, 65°C 1 minute 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.

2.2.4. Western blotting for galectin-4

Total protein was extracted using Qproteome mammalian protein preparation kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. Protein samples were resolved on SDS-polyacrylamide gel. Then, proteins have been electrophoretically transferred to nitrocellulose membranes (Amersham Pharmacia Biotech, Buckinghamshire, UK) followed by blocking for 1 h in 3% nonfat milk in TBST. Subsequently, the membrane has been

incubated in the primary antibody against galectin-4 (R&D system, Minneapolis, MN, USA) for 1 h. After incubation with secondary antibody, bands were detected by chemiluminescence using the ECL Western Blotting Detection System (GE Healthcare, Amersham, Buckinghamshire, UK).

2.2.5. Galectin-4 promoter analysis

2.2.5.1. Promoter sequencing

The promoter sequence has been retrieved using Transcriptional Regulatory Element Database [84,85]. We have sequenced the promoter from –400 to +300 using LGALS4-700F: TTCACAGTTGCTGGGAGAGG and LGALS4-700R: GATGACGAGGGCCAACAGTTAGACGTG. The primers have been designed using Primer3 (2). The amplified products from the seven cell lines were Sanger sequenced at the GATC Biotech, Konstanz, Germany.

2.2.5.2. Luciferase reporter assay

According to the results from the promoter sequencing, three fragments of the galectin-4 promoters have been cloned in pGL4.1 (Promega, Madison, WI). In details, one large fragment contains the site of two variations found in our study and each of the other two fragments contain one site of variation. Simply, the PCR fragments were amplified from the wild and the mutant cell lines using the following primers: LGALS4 XhoI F: AAAAAAACTCGAGTTCACAGTTGCTGGGAGAGG and LGALS4 HindIII R: GGGGGGAAGCTTGCTGCGCTAGTGGCTGGTC for –233 variation, for the 2nd variation at +150: LGALS4 var2 NheI F: AAAAAAAGCTAGCCCACCATCTCCCACTCCTG andLGALS4 var2 HindIII R: TTTGGGGAAGCTTGATGACGAGGGCCAACAGTTAGA and for the two variation: LGALS4 2var NheI F: AAAAAAAGCTAGCTTCACAGTTGCTGGGAGAGG and LGALS4 2var HindIII R: TTTGGGGAAGCTTGATGACGAGGGCCAACAGTTAGA. Following the PCR amplification, the products have been digested and ligated into pGL4.1 plasmid. The cell lines were cotransfected with constructed plasmid and pRL-TK using effectene kit (Qiagen, Hilden, Germany). After 24 hours, the cells were lysed and the cell lysates were analysed for firefly luciferase activity with the dual-luciferase reporter assay system (Promega, Madison, WI). Firefly activity was normalized to Renilla (pRL-TK).

2.2.5.3. Pull down of the binding proteins

The PCR fragments of galectin-4 promoter (the wild and the variants) were used as baits for the DNA binding proteins using Pierce Pull-Down Biotinylated Protein:Protein Interaction kit (Thermo Scientific, Rockford, USA). The kit had been modified to suit DNA:Protein interactions by changing the binding buffer. The promoter fragments were amplified using Biotinylated primers and reacted with immobilized streptavidin column. The CO115 nuclear proteins were extracted by NE-PER Nuclear and Cytoplasmic Extraction Reagents (Thermo Scientific, Rockford, USA). Then diluted in 5X binding buffer (0.1M Hepes pH 8.0, 0.25M KCl, 25mM DTT, 0.25mM EDTA, 5mM $MgCl_2$, 25% v/v glycerol) and incubated on the column 30 minutes at room temperature. Afterward, the columns were washed with TBST and the binding proteins eluted with Lamelli buffer at 95°C. Subsequently, the eluted protein samples were resolved on SDS-polyacrylamide gel. Subsequently, the bands were cut for mass spectrometry analysis (core facility, DKFZ, Heidelberg, Germany).

2.2.5.4. Methylation status

The CpG island was identified and the proper primers for the bisulfite sequencing were designed using MethPrimer [86]. Genomic DNA was treated with EpiTect DNA Bisulphite Modification Kit (Qiagen, Hilden, Germany). The bisulfite treated DNAs were amplified using LGALS4-cpg F: TTTTTATTTTGGGTATAAGAGTTATTAT and LGALS4-cpg R: ATAAAACCTAAACTAACATCTCACC and subsequently were Sanger sequenced at the GATC Biotech, Konstanz, Germany.

2.2.5.5. Promoter genotyping in colorectal cancer patient samples

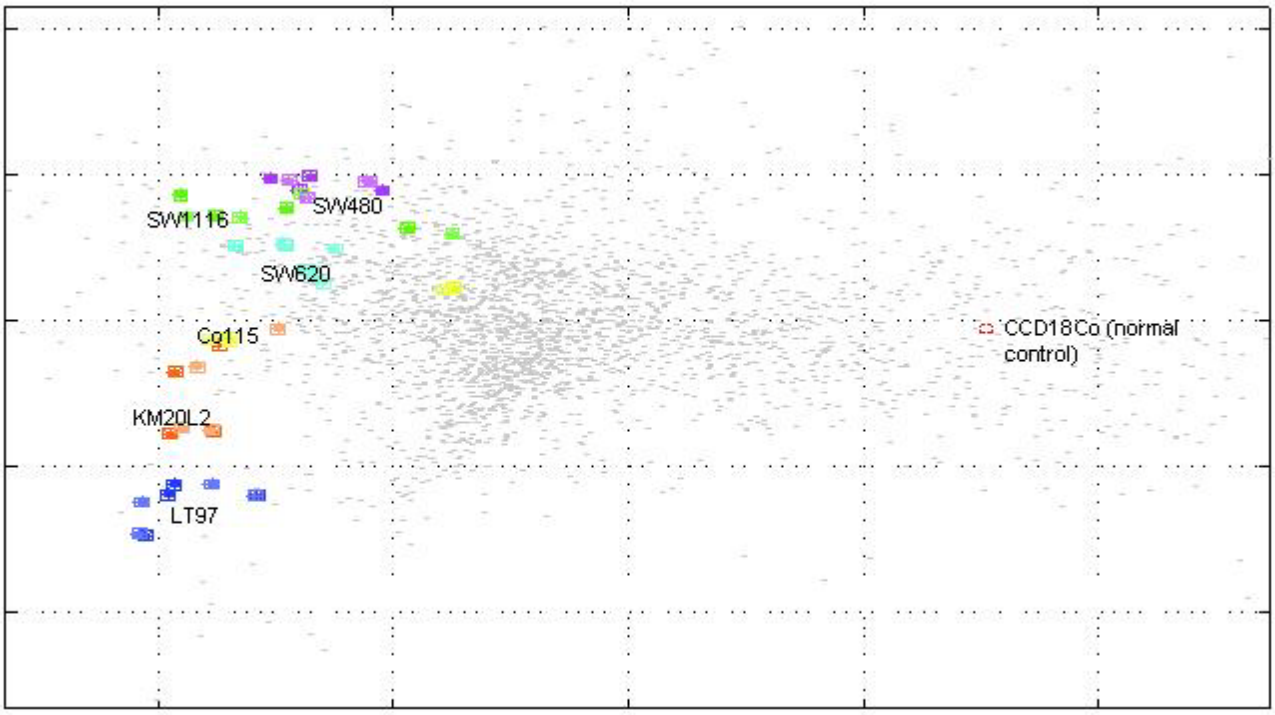
We screened the promoter sequences of galectin-4 for the two SNPs in colorectal lesions patients using LGALS4-700F: TTCACAGTTGCTGGGAGAGG and LGALS4-700R: GATGACGAGGGCCAACAGTTAGACGTG. 115 samples were collected as the following: 1) 18 colorectal tumors from the National Center of Tumor Diseases (NCT), Heidelberg, Germany, 2) DNA isolated from lesions and their adjacent normal tissues from 54 patients (27 colorectal cancer patients, 15 ulcerative colitis patients, and 12 patients with adenomatous polyps) which are collected from the Egyptian National Cancer Institute (ENCI), Cairo, Egypt. 3) 15 colorectal tumor samples from the Egyptian National Cancer Institute (ENCI), Cairo, and 4) 28 DNA samples isolated from blood of gastrointestinal tumors (the Egyptian National Cancer Institute (ENCI), Cairo).

3. Results

3.1. Expression profiling classifies colorectal cancer cell lines to bad and good prognosis rather than tumor stages

Six colorectal cancer cell lines that represent different Duke's stages were profiled for gene expression versus a normal colon cell line. Clustering of the samples using correspondence and MeV analyses indicate an association between them. As shown in figure 1 a and b, the colorectal cancer cell lines are classified nicely according to their stage and degree of aggressiveness except LT97 (benign adenoma cell line) that is classified with grade D cell line (KM20L2). The convergent clustering of LT97 and KM20L2 was the main motivation to pursue the present study. Since LT97 is a cell line derived from hereditary polyposis patient, it could be an indication for bad prognosis. Accordingly, we have selected galectin-4 gene for further analyses because it was significantly upregulated in LT97 and KM20L2 but not the rest of the cell lines (figure 2).

a



b

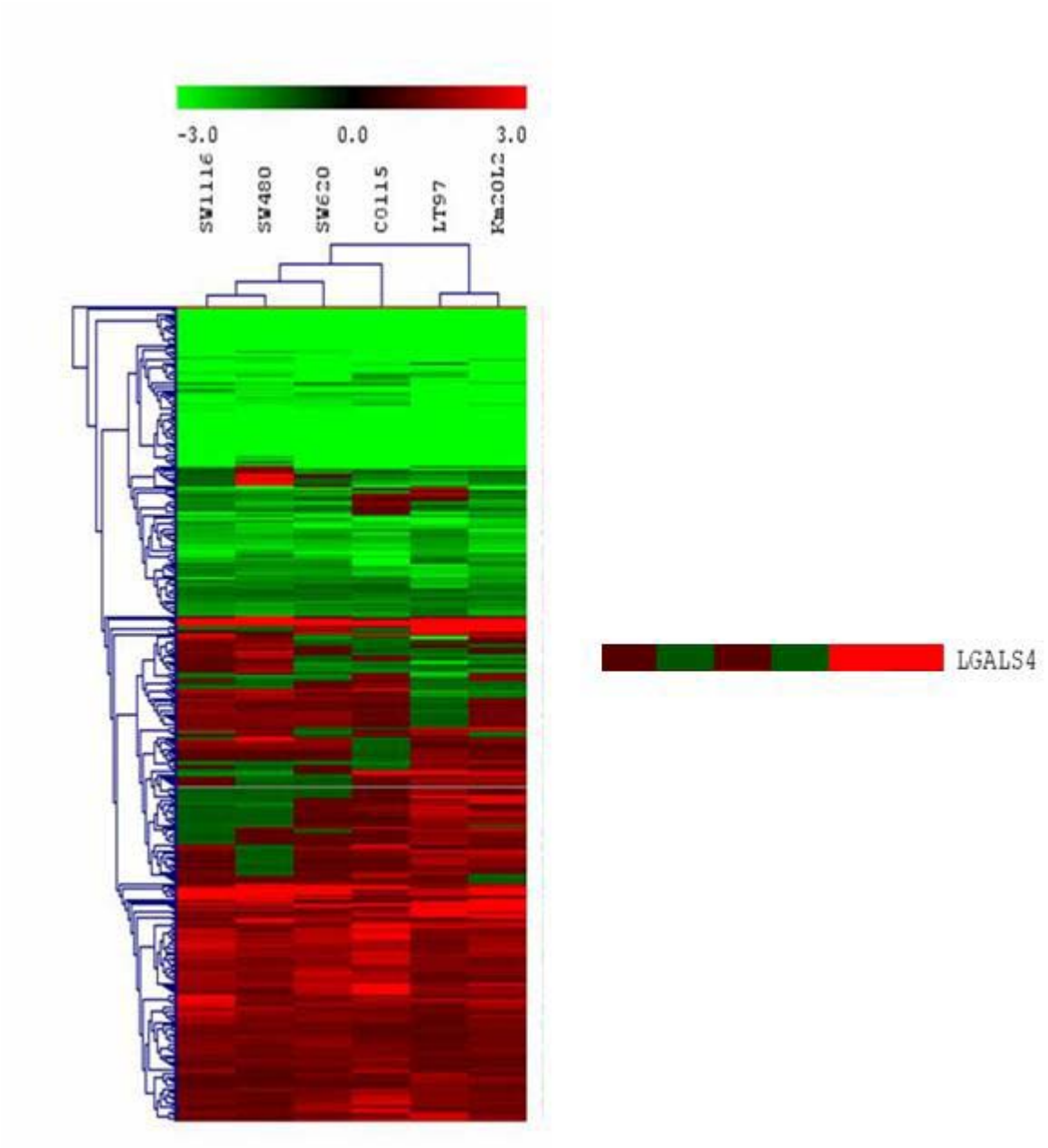


Figure 2. The expression profiling results of the six colon cell lines compared to the normal colon cell lines. a. the correspondence cluster analysis of transcript profiles. In the resulting biplot, each hybridization of a cell line is depicted as a colored square. Genes that exhibited significantly differential transcription level are shown as black dots. The closer the colocalizaion of two spots (both genes and cell lines), the higher the degree of association between them. b. The heatmap shows the hierarchical clustering of six cell lines based on expression profiling. In the heatmap, red represents high expression, black represents median expression, green represents low expression.

3.2. Galectin-4 is significantly upregulated in LT97 and KM20L2

Among 417 filtered differentially regulated genes from the microarray results, galectin-4 showed significant upregulation in LT97 and KM20L2 (benign adenoma and Duke's D cell lines), which was interesting to pursue more investigations and to focus on the upregulation mechanism of galectin-4. Subsequently, qRT-PCR and Western blotting were used to validate the upregulation that has been shown in the microarray data. The results of the qRT-PCR showed huge upregulation of galectin-4 in LT97 and Km20l2 which exceed 350 and 700 fold change respectively as shown in (figure 3). Moreover, CO115 showed a slight upregulation in the mRNA level, unless it didn't show protein expression in the Western-blotting.

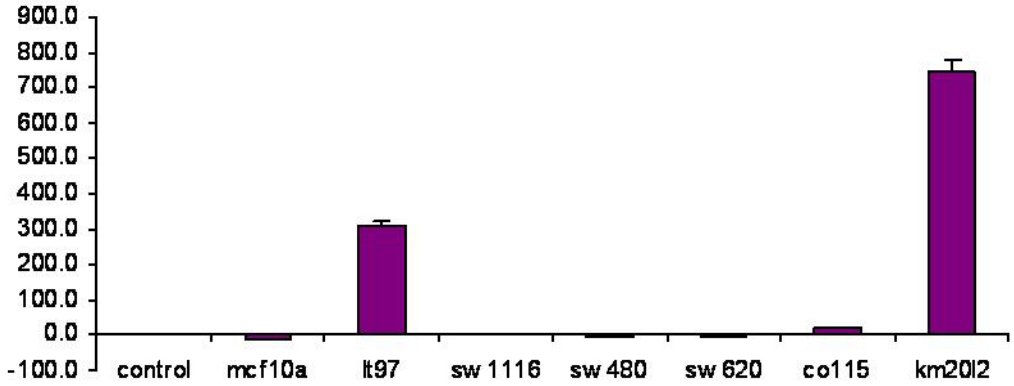


Figure 3. qRT-PCR of galectin-4 for six cell lines (LT97, SW1116, SW480, SW620, Co115, and KM20L2) in comparison to normal colon cell line as a control. As shown in figure, LT97 and KM20L2 exhibit significant upregulation (350, 700 fold change respectively). Also, CO115 showed a slight upregulation in the mRNA level (4 fold change).

3.3. A twin SNPs in the regulatory region are associated with galectin-4 upregulation in LT97 and KM20L2

Upon the sequencing results of the *LGALS4* upstream sequence (700bp including the first exon and part of the first intron), two SNPs were found in LT97 and KM20L2 cell lines, one before the transcription start site TSS and the second in the first intron (in a potential regulatory region) as shown in figure 4. The first exon SNP is previously known as rs73933062, but it is not known for medical impact. The upstream SNP was not recorded before and it is at contig position 11572034 C>A or at 39303816 C>A chromosomal position

and accordingly. We submitted this SNP and it got submission SNP access number ss217320370. In both cell lines, the two SNPs were observed together in the same sequence, and that why we called them a twin SNP.

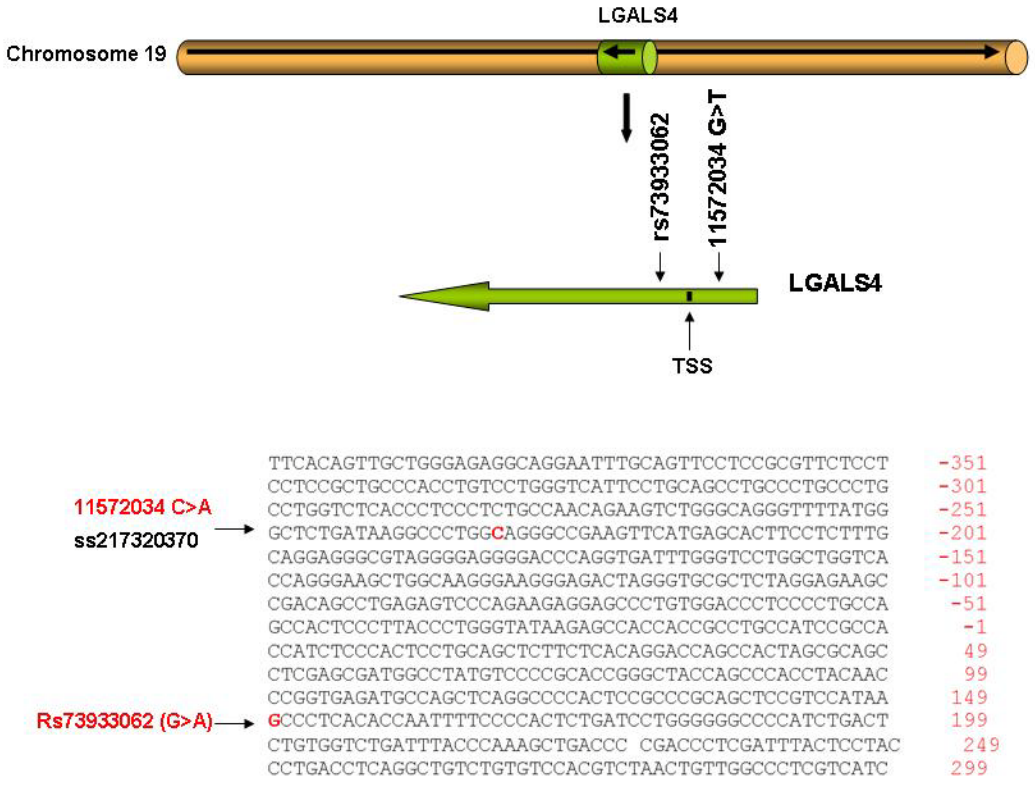


Figure 4. *LGALS4* is expressed from the minus strand and located on chromosome 19. The position of the two SNPs in the upstream sequence and the first intron of galectin-4 is showed in this figure comparable to transcription start site (TSS). Also the promoter sequence that was retrieved Transcriptional Regulatory Element Database.

3.4. Effect of the two SNPs activity on promoter

The two SNPs are located in the regulatory region of galectin-4: 1) ss217320370 is in the promoter sequence before the TSS and 2) rs73933062 is in a putative enhancer region, after the TSS but not in the translated sequence. Therefore, the luciferase activity of the promoter has been evaluated using pGL4.10 constructs for each SNP and the two SNPs as the following: 1) pGL4.10-wt1 and pGL4.10-mt1 which contains respectively the A or A at position 11572034, 2) pGL4.10-wt2 and pGL4.10-mt2 represents rs73933062 with G or A, and 3) two

constructs include the SNPs named as pGL4.10-2wt and pGL4.10-2mt. Co115 cell line was transfected with each construct (which contains Firefly luciferase as a reporter gene) and pRL-TK (which has a Renilla luciferase). Then, the Firefly/Renilla ratio was calculated for transfection efficiency control and the fold changes were calculated by dividing Firefly/Renilla ratio from variant-2 construct over Firefly/Renilla ratio of variant-1 construct. We have got the following results for the 3 different constructs: 1) a slight change (1.3 fold change) in the construct pGL4-mt1, which contains C/A at 11572034-contig position, 2) interestingly, the construct pGL4-mt2 that has rs73933062 showed 1500 folds change over pGL4-wt2, and 3) nearly 4 folds change has been obtained in the case of pGL4-2mt relative to pGL4-2wt that represent the construct that contains the two SNPs site (figure 5).

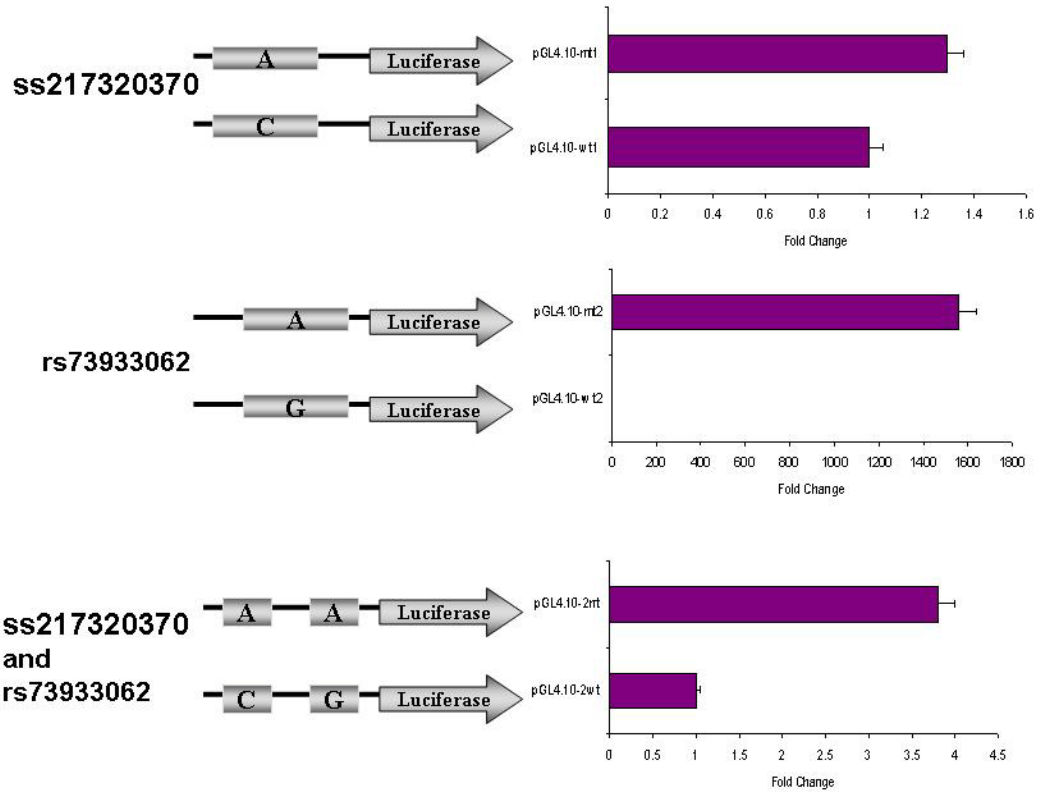


Figure 5. The transient transfection of luciferase reporter construct in Co115 cell line results: it shows the variation of the presence of each SNP individually comparable to the predominant genotype and also the combination of the two SNPs in the same construct. The presence of a mere ss217320370 in the construct mildly increased the luciferase activity 1.3 fold change in comparison to the normal construct. On the other hand, pGL4-mt2 that has rs73933062 showed 1500 folds change over pGL4-wt2. The combination of the two SNPs in one construct increased the luciferase activity four times.

To confirm rs73933062 luciferase activity results, SW1116 and SW620 were transfected with the construct, which contains the target SNP and the experiment was repeated twice independently. Interestingly, 36 and 37-fold change in the luciferase activity were obtained respectively for SW620 and SW1116 (figure 6).

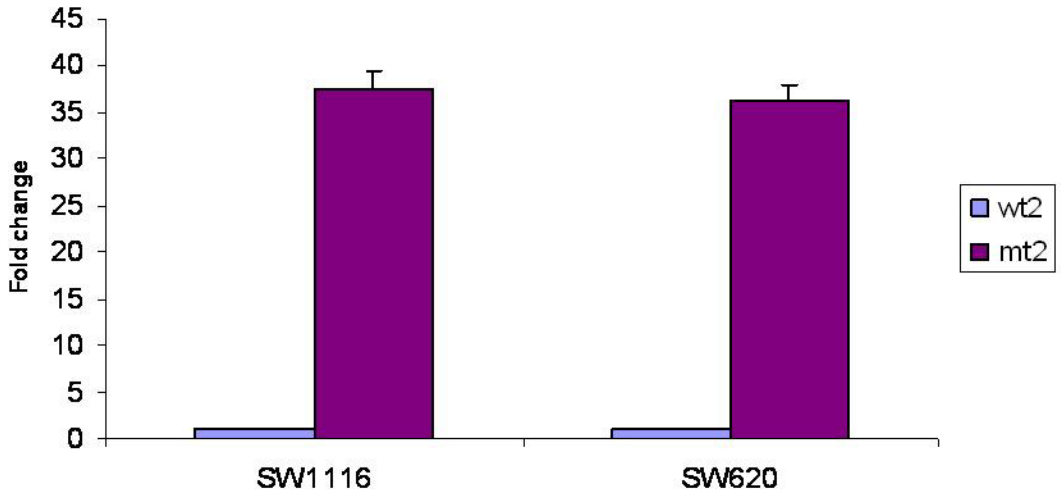


Figure 6. The transfection results of SW1116 and SW620 with the luciferase construct that contains rs73933062 SNP. In comparison to the wild type construct, respectively, 37 and 36 fold change increasing in the luciferase activity in SW1116 and SW620.

3.5. The two SNPs are affecting the protein binding sites

In order to understand the mechanism underlying the upregulation effect associated with the presence of the two SNPs, a pull down experiment was done to investigate the DNA-binding proteins that tethered to each variant. Interestingly, we have found that both rs73933062 and contig 11572034 C/A are introducing and deleting binding sites. Among the response element that have been omitted: Aconitase 1 (ACO1), Retinoblastoma Binding Protein-7 (RBBP7), siah binding protein 1 (PUF60), and DAZ associated protein. On the other hand, the two SNPs offered response elements for some interacting proteins such as MYB binding protein 1a (MYBB1a), fatty acid binding protein 5 (FABP5) and peoxiredoxin (PRDX1). Moreover, the pull-down experiment manifested the regulatory proteins that bind to

galectin-4 promoter like enolase-1 (ENO1), fuse-binding protein-1 (FUBP1), fuse-binding protein-2 (FUBP2 or KSRP), FUS, and LBP1a.

2mt (ss217320370 C>A and rs73933062 G>A)
54 kDa protein
90kDa heat shock protein
adenomatosis polyposis coli
alpha enolase
alpha-tubulin
annexin A2 isoform 2
apolipoprotein D, apoD
apoptosis inhibitor 5 isoform b
arginase (EC 3.5.3.1)
ATP-dependent DNA helicase II
ATP-dependent DNA helicase II, 70 kDa subunit
caspase 14 precursor
cathepsin D preproprotein
CCT8 protein
Chain A, High Resolution Solution Nmr Structure Of Mixed Disulfide Intermediate Between Mutant Human Thioredoxin And A 13 Residue Peptide Comprising Its Target Site In Human Nfkb
Chain A, Single Stranded Dna-Binding Domain Of Human Replication Protein A Bound To Single Stranded Dna, Rpa70 Subunit, Residues 183-420
chaperonin (HSP60)
chaperonin containing TCP1, subunit 6A isoform a
chaperonin containing TCP1, subunit 7 isoform a
colonic and hepatic tumor over-expressed protein isoform a
cystic fibrosis antigen
cytoplasmic chaperonin hTRiC5
desmoplakin I
DNA helicase Q1
DNA topoisomerase I
DNA topoisomerase II
DNA-binding protein
Dsc1a precursor
elongation factor-1 alpha
enhancer protein
enolase
eukaryotic initiation factor 5A
eukaryotic translation elongation factor 1 alpha 1
eukaryotic translation elongation factor 1 gamma
eukaryotic translation elongation factor 2
eukaryotic translation initiation factor 4A isoform 1
fatty acid binding protein 5 (psoriasis-associated)
filaggrin family member 2
FUBP1 protein
FUSE binding protein 2
gastric cancer-related protein FKSG9
glyceraldehyde-3-phosphate dehydrogenase
growth regulated nuclear 68 protein

heat shock 70kDa protein 8 isoform 1
heat shock protein
heat shock protein 90
heterogeneous nuclear ribonucleoprotein A/B isoform a
heterogeneous nuclear ribonucleoprotein L
histone H4
hnRNP U protein
Hsp89-alpha-delta-N
insulin-like growth factor 2 mRNA binding protein 1 isoform 1
K channel:SUBUNIT=beta
LBP-1a
lysozyme precursor (EC 3.2.1.17)
M4 protein
mitochondrial ATP synthase beta subunit precursor
mutant beta-actin (beta~-actin)
MYB binding protein (P160) 1a
myoblast antigen 24.1D5
nuclear DNA helicase II
nuclear factor IV
nuclear respiratory factor 1
nucleolin
phosphoglycerate dehydrogenase
phospholipase C-alpha
poly(ADP-ribose) polymerase
poly(rC) binding protein 2 isoform b
polypyrimidine tract-binding protein 1 isoform a
prolactin-induced protein precursor
RecName: Full=Desmoglein-1; AltName: Full=Desmosomal glycoprotein 1; Short=DG1; Short=DGI; AltName: Full=Pemphigus foliaceus antigen; Flags: Precursor
RecName: Full=Probable ATP-dependent RNA helicase DDX17; AltName: Full=DEAD box protein 17; AltName: Full=RNA-dependent helicase p72; AltName: Full=DEAD box protein p72
replication protein A1
ribophorin I precursor
Ro ribonucleoprotein
serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 12
serum albumin
spliceosomal protein SAP 155
splicing factor proline/glutamine rich (polypyrimidine tract binding protein associated)
splicing factor SF3a60
squamous cell carcinoma antigen
STE20-like kinase
structural maintenance of chromosomes 3
suprabasin isoform 1 precursor
t-complex polypeptide 1
testicular H1 histone
The deletion results in premature stop
thiol-specific antioxidant protein
TLS protein
topoisomerase I
TR4 orphan receptor
transcription factor LSF
transformation upregulated nuclear protein

transglutaminase E3
TUBB protein
ubiquitin
UV excision repair protein RAD23 homolog B
vimentin
Zn-alpha2-glycoprotein

Table 1. The list of proteins which bound to *LGALS4* upstream sequence, which contains the twin SNP (A instead of C and G). White box indicates common protein for wild and mutant types and orange boxes represent the unique proteins for each genotype.

2wt (wild type)
54 kDa protein
90kDa heat shock protein
aconitase 1
adenomatosis polyposis coli
alpha-tubulin
annexin A2 isoform 2
apoptosis inhibitor 5 isoform b
arginase (EC 3.5.3.1)
ATP-dependent DNA helicase II
ATP-dependent DNA helicase II, 70 kDa subunit
caspase 14 precursor
cathepsin D preproprotein
CCT8 protein
cell cycle protein p38-2G4 homolog
CGI-46 protein
Chain A, High Resolution Solution Nmr Structure Of Mixed Disulfide Intermediate Between Mutant Human Thioredoxin And A 13 Residue Peptide Comprising Its Target Site In Human Nfkb
Chain A, Single Stranded Dna-Binding Domain Of Human Replication Protein A Bound To Single Stranded Dna, Rpa70 Subunit, Residues 183-420
chaperonin (HSP60)
chaperonin containing TCP1, subunit 6A isoform a
chaperonin containing TCP1, subunit 7 isoform a
cystic fibrosis antigen
cytoplasmic chaperonin hTRiC5
damage-specific DNA binding protein 2
DAZ associated protein 1
DEK oncogene isoform 1
desmoplakin I
DNA helicase Q1
DNA topoisomerase I
DNA-binding protein
Dsc1a precursor
enolase 1
eukaryotic initiation factor 5A
eukaryotic translation elongation factor 1 gamma
eukaryotic translation elongation factor 2
filaggrin family member 2
FUBP1 protein

FUSE binding protein 2
glyceraldehyde-3-phosphate dehydrogenase
growth regulated nuclear 68 protein
GRP78 precursor
heat shock 70kDa protein 8 isoform 1
heat shock protein
heat shock protein 90
Heat shock protein HSP 90-alpha 2
heterogeneous nuclear ribonucleoprotein A/B isoform a
heterogeneous nuclear ribonucleoprotein L
histone H4
hnRNP U protein
insulin-like growth factor 2 mRNA binding protein 1 isoform 1
K channel:SUBUNIT=beta
LBP-1a=transcription factor binding to initiation site of HIV-1 {alternatively spliced} (human, Namalwa cells, Peptide, 504 aa)
lysozyme precursor (EC 3.2.1.17)
M4 protein
mitochondrial ATP synthase beta subunit precursor
mutant beta-actin (beta~-actin)
nuclear factor IV
nuclear respiratory factor 1
nucleolin
p53 cellular tumor antigen
phosphoglycerate dehydrogenase
phospholipase C-alpha
poly(ADP-ribose) polymerase
poly(rC) binding protein 2 isoform b
polypyrimidine tract-binding protein 1 isoform a
prolactin-induced protein precursor
RecName: Full=Desmoglein-1; AltName: Full=Desmosomal glycoprotein 1; Short=DG1; Short=DGI; AltName: Full=Pemphigus foliaceus antigen; Flags: Precursor
RecName: Full=Probable ATP-dependent RNA helicase DDX17; AltName: Full=DEAD box protein 17; AltName: Full=RNA-dependent helicase p72; AltName: Full=DEAD box protein p72
replication protein A1
retinoblastoma binding protein 7
ribophorin I precursor
RNA binding motif protein 39 isoform b
Ro ribonucleoprotein
serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 12
serum albumin
siah binding protein 1
splicing factor proline/glutamine rich (polypyrimidine tract binding protein associated)
splicing factor SF3a60
suprabasin isoform 1 precursor
SWI/SNF-related matrix-associated actin-dependent regulator of chromatin a5
t-complex polypeptide 1
testicular H1 histone
thiol-specific antioxidant protein
TLS protein
topoisomerase I
TR4 orphan receptor

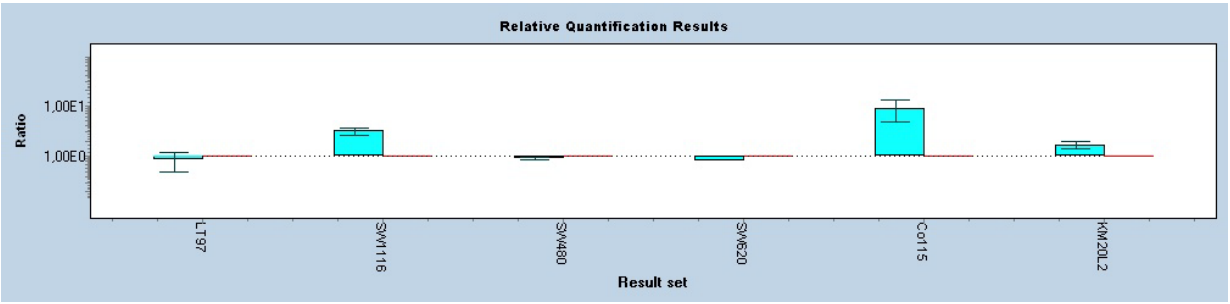
transcription factor LSF
transformation upregulated nuclear protein
transglutaminase E3
ubiquitin
UV excision repair protein RAD23 homolog B
vimentin
Zn-alpha2-glycoprotein

Table 2. Pull down- mass spectrometry results: the list of proteins which bound to the PCR product of *LGALS4* upstream sequence, which contains C and G genotype (wild type). White box indicates common protein for wild and mutant types and orange boxes represent the unique proteins for each genotype.

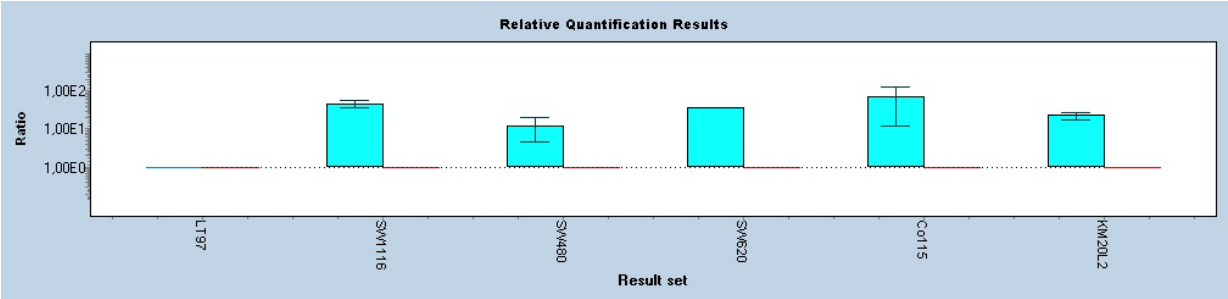
3.6. The expression of the binding proteins in different cell lines

As a trial to correlate the presence of the SNPs, expression level of the binding proteins, and the expression of *LGALS4*, qRT-PCR was carried out for some selected genes in which their proteins are binding to *LGALS4* upstream sequence. Therefore, 10 genes that are encoding the binding proteins (which were obtained by pull down-mass spectrometry results) were selected, either bound to the wild, the mutant-type, or both genotype of *LGALS4* promoter as the following: *ACO1*, *ENO1*, *FUBP1*, *FUBP2*, *HNRNPK*, *MYBBP1a*, *PRDX1*, *PUF60*, *RBBP7*, and *FUS*. The qRT-PCR was done for the six cell lines (LT97, SW1116, SW480, SW620, CO115, and KM20L2) comparable to CCD-18Co (normal colon cell line). The experiment was done in three repeats and the results were normalized to GAPDH (figure 7). White box indicates common protein for wild and mutant types and orange boxes represent the unique proteins for each genotype. All of the mRNA of the investigated gene were expressed in CO115, which was consistent with the pull down-mass spectrometry results, since the CO115 nuclear proteins were used.

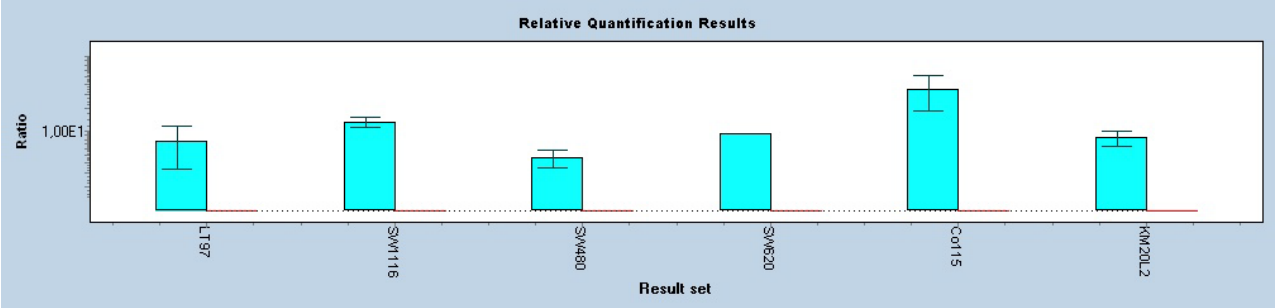
ACO1



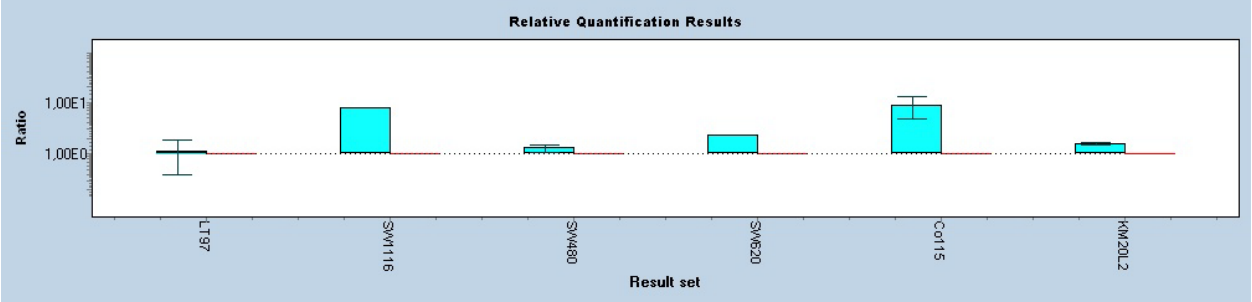
ENO1



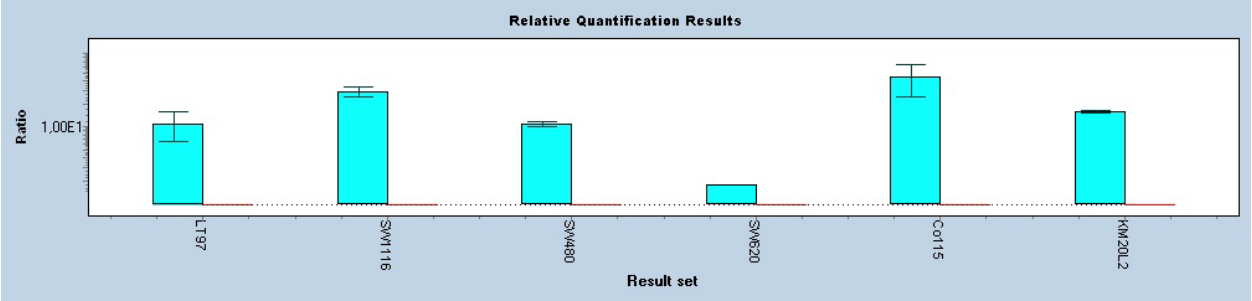
FUBP1



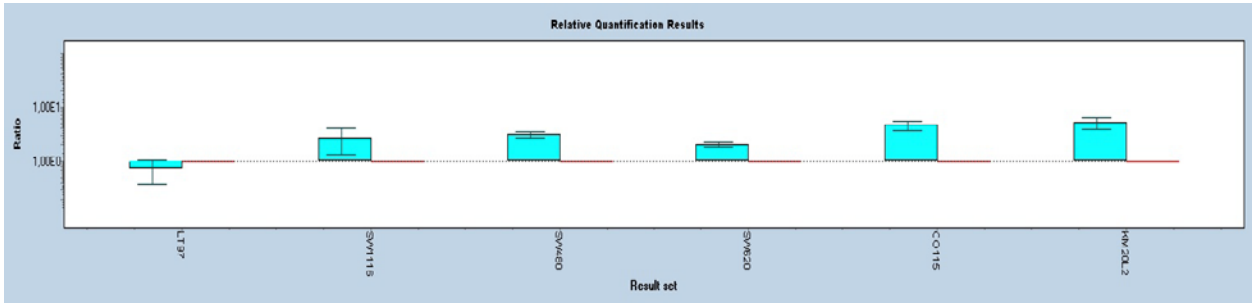
FUBP2



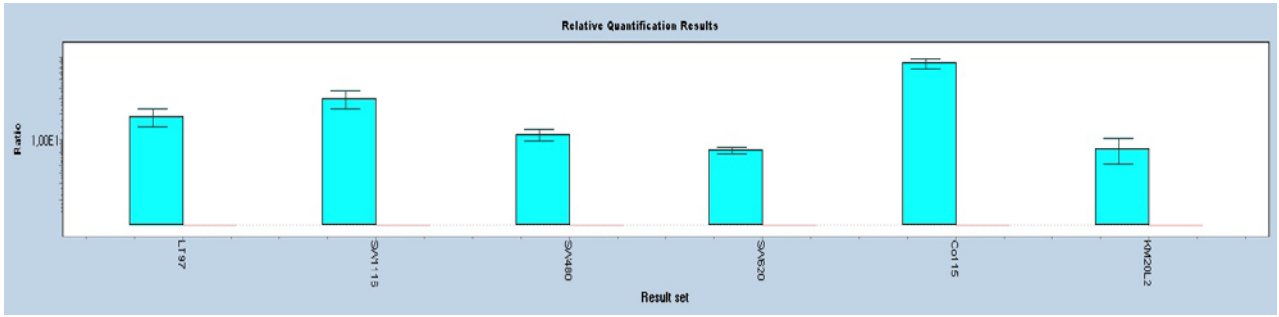
HNRNPK



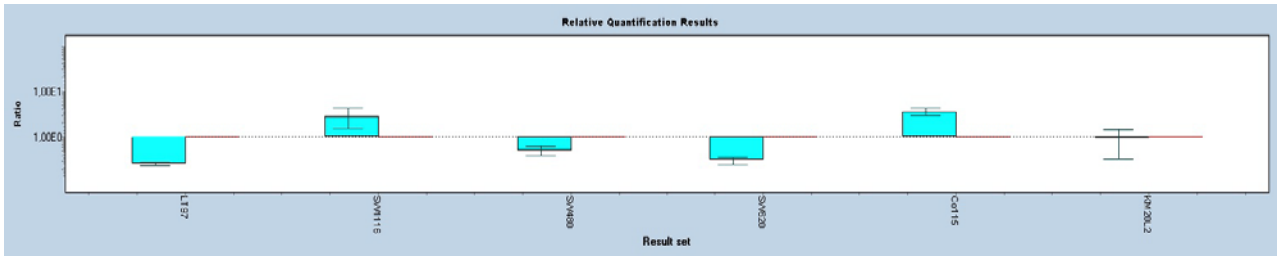
MYBBP1a



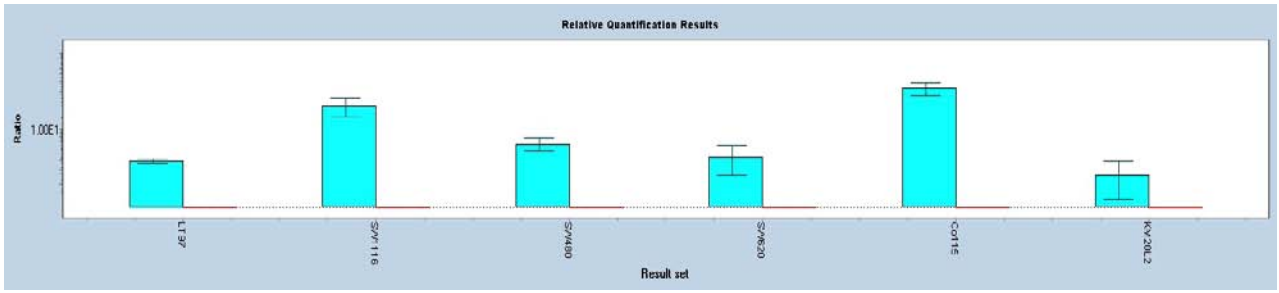
PRDX1



PUF60



RBBP7



FUS

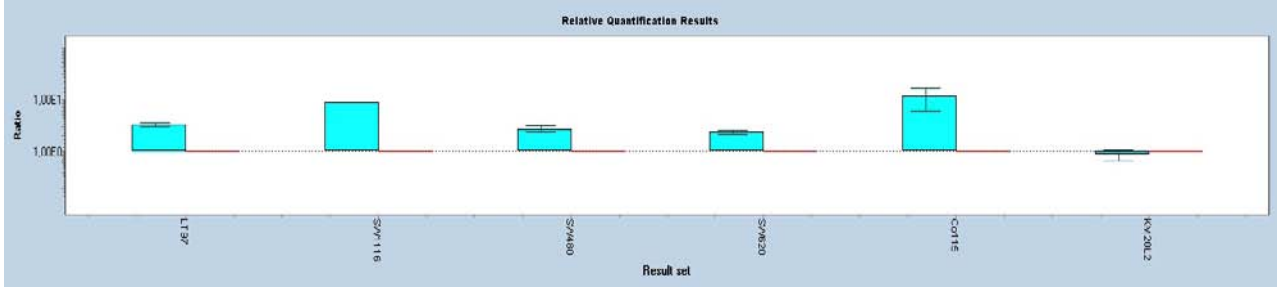


Figure 7. The qRT-PCR results of the genes, which encoding the binding proteins that bound to *LGALS4* promoter sequence.

3.7. Methylation status

In the present study, the CpG island was detected using Methprimer in the first exon and first intron. Accordingly, the primers for bisulfite sequencing were designed. After the bisulfite treatment, the PCR was performed and the methylation was detected by direct sequencing. Since the bisulfite treatment will convert the unmethylated cytocine to uracil, in the sequencing result the methylated cytocine will remain as it is, while the unmethylated cytocine will be thymine. As shown in figure 8b, the predicted CpG island of *LGALS4* contains 15 CG. The methylation status of six CG out of the 15 CGs were changed in colorectal cancer patients and cell lines samples, regarding the methylation status. In the cell lines exhibiting galectin-4 upregulation, LT97 and KM20L2 showed unmethylated pattern of the promoter region. Whereas, CCD18Co, SW1116, SW480, SW620, and CO115 exhibited methylated promoter (figure 8b).

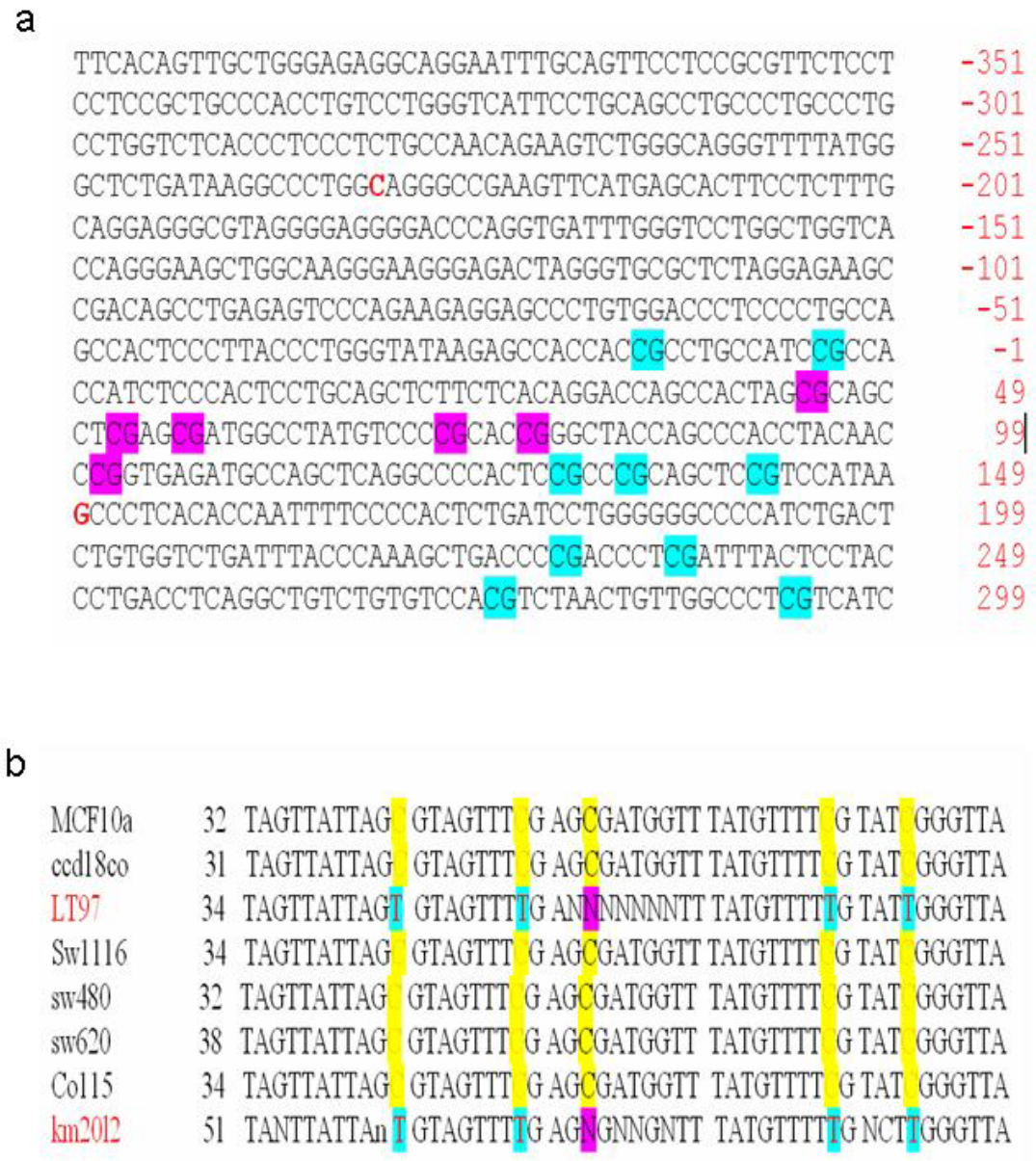


Figure 8. Methylation status of *LGALS4* promoter. a) The CpG island prediction and the primers for bisulfite sequencing using Methprimer. CpG island of *LGALS4* contains 15 CG. The methylation status of six CG out of the 15 CGs were changing b) The bisulfite treatment will convert the unmethylated cytocine to uracil, in the sequencing result the methylated cytocine will remain as it is, while the unmethylated cytocine will be thymine. Accordingly, the bisulfite sequencing results of CpG island in galectin-4 promoter, LT97 and KM20L2 showed unmethylated promoter. While the other cell lines exhibited methylated upstream sequence. N represents heterogeneous peaks of C and T.

3.8. Two SNPs are shown together in patient samples

Among 115 samples collected from Germany and Egypt, we have found ss217320370 C>A and G>A (rs73933062), both SNPs are present together in the same patient. Therefore, we called them twin SNPs. The direct sequencing of the upstream sequence of galectin-4

manifested the twin SNPs in 26 samples out of 115 patients (22.6 %) which are detailed as the following:

1) Four out of 18 (22.2%) rectal cancer patients (table 3) from NCT, Heidelberg, Germany.

Sample no.	ss217320370	rs73933062	age	gender	Grading	T	N	M
1	C/A	G/A	66	M	2	4	0	X
2	wt	wt	71	M	2	3	2	1 (liver)
3	wt	wt	71	F	2	3	0	X
4	wt	wt	58	M	2	2	0	X
5	wt	wt						
6	wt	wt	53	M	3	4	1	1 liver
7	wt	wt	56	M	2	2	0	X
8	C/A	G/A	64	F	2	2	0	X
9	C/A	G/A	88	F	2	4	2	X
10	wt	wt	41	M	3	3	2	1 (Liver, Lung)
11	wt	wt	60	M	2	3	1	X
12	wt	wt	53	F	2	3	1	X
13	wt	wt	67	M	2	3	0	X
14	wt	wt	57	M	2	3	0	X
15	wt	wt	78	M	2	2	0	X
16	C/A	G/A	67	M	2	2	1	X
17	wt	wt	49	F	2	two tumors 3/1	2/2	X
18	wt	wt	68	F	X	3	2	X

Table 3. The sequencing results of galectin-4 upstream sequence in 18 rectal cancer patients, which were collected from tissue bank, NCT, Heidelberg. The results show, four patients display ss217320370 and rs73933062 together in the upstream sequence.

2) 12 out of 54 patients (22.2%) showed the two SNPs in the lesion tissues and their adjacent normal tissue. The detailed results are as the following: 8 out of 27 (29.6%) tissue from colorectal cancer patients, 4/15 (26.7%) patients with ulcerative colitis, and 0/12 with adenomatous polyps (ENCI, Cairo, Egypt).

No.	lesion	normal	lesion	normal					
	ss217320370		rs73933062		age	gender	location	grade/grade of dysplasia	
1	C/A	C/A	G/A	G/A	48	M	Rectum	II	CRC
2	wt	wt	wt	wt	35	F	rectum	II	
3	wt	wt	wt	wt	37	M	rectum	II	
4	wt	wt	wt	wt	34	M	sigmoid	III	

5	wt	wt	wt	wt	52	F	Ascending	n/a	
6	wt	wt	wt	wt	38	F	rectum	II	
7	wt	wt	wt	wt	47	F	Ascending	II	
8	wt	wt	wt	wt	28	F	rectum	II	
9	C/A	C/A	G/A	G/A	60	M	rectum	I	
10	wt	wt	wt	wt	38	F	rectum	I	
11	C/A	C/A	G/A	G/A	55	F	Descnding	I	
12	C/A	C/A	G/A	G/A	56	M	rectum	II	
13	C/A	C/A	G/A	G/A	59	F	rectum	II	
14	wt	wt	wt	wt	55	F	rectum	II	
15	wt	wt	wt	wt	52	M	caecum	II	
16	wt	wt	wt	wt	37	F	rectum	II	
17	wt	wt	wt	wt	29	M	sigmoid	II	
18	wt	wt	wt	wt	48	F	Descnding	III	
19	C/A	C/A	G/A	G/A	64	F	rectum	II	
20	wt	wt	wt	wt	30	F	sigmoid	n/a	
21	wt	wt	wt	wt	40	F	rectum	II	
22	wt	wt	wt	wt	41	F	sigmoid	n/a	
23	C/A	C/A	G/A	G/A	46	M	sigmoid	n/a	
24	wt	wt			50		rectum	n/a	
25	C/A	C/A	G/A	G/A	58	M	Descnding	n/a	
26	wt	wt	wt	wt	13	F	rectum	n/a	
27	wt	wt	wt	wt	35	M	sigmoid	n/a	
28	C/A	C/A	G/A	G/A	28	M	UC	No dysplasia	UC
29	wt	wt	wt	wt	37	M	recto-sigmoid UC	No dysplasia	
30	wt	wt	wt	wt	74	M	UC	No dysplasia	
31	wt	wt	wt	wt	45	M	recto-sigmoid UC	No dysplasia	
32	C/A	C/A	G/A	G/A	30	M	UC	till splenic mild dysplasia	
33	C/A	C/A	G/A	G/A	69	M	whole splenic UC	No dysplasia	
34	wt	wt	wt	wt	27	F	recto-ileum UC	No dysplasia	
35	wt	wt	wt	wt	17	F	whole UC	dysplasia	
36	C/A	C/A	G/A	G/A	38	M	whole UC	No dysplasia	
37	wt	wt	wt	wt	50	M	whole UC	y moderate	
38	wt	wt	wt	wt	43	M	left UC	No dysplasia	
39	wt	wt	wt	wt	60	F	rectum UC	dysplasia	
40	wt	wt	wt	wt	65	F	Up to transvrese UC	Marked dysplasia	
41	wt	wt	wt	wt	30	M	recto-sigmoid UC	No dysplasia	
42	wt	wt	wt	wt	55	M	left UC	No dysplasia	
43	wt	wt	wt	wt	60	M	Transverese		polyps
44	wt	wt	wt	wt	69	M	transverese polyps		
45	wt	wt	wt	wt	50	F	hepatic polyps		
46	wt	wt	wt	wt	39	F	Transverse polyps		
47	wt	wt	wt	wt	45	F	Ascending polyps		

48	wt	wt	wt	wt	50	M	sigmoid polyps		
49	wt	wt	wt	wt	35	M	sigmoid polyps		
50	wt	wt	wt	wt	54	M	Descending polyps		
51	wt	wt	wt	wt	56	M	sigmoid polyps		
52	wt	wt	wt	wt	52	M	rectal		
53	wt	wt	wt	wt	50	M	cecum polyps		
54	wt	wt	wt	wt	45	M	sigmoid polyps		

Table 4. The sequencing results of galectin-4 upstream sequence in 54 patients with colorectal cancer, ulcerative colitis (UC), or colorectal polyps, which were collected from ENCI. Twelve out of 54 showed ss217320370 and rs73933062 together in galectin-4 promoter. Interestingly, the two SNPs were found in the lesion tissue (biopsy from tumor, UC area, or polyps) and the adjacent normal in all of the twelve patients.

3) Two out of 15 (13.3%) colorectal tumors and premalignant lesions from ENCI, Cairo, Egypt. The DNA samples were obtained from the lesion (tumor, or the premalignant lesion), but the adjacent normal tissues were unavailable.

no.	**ss217320370**	**rs73933062**	**Age**	**gender**	**Type of cancer**
1	wt	wt	30	M	Known patient for cancer rectum with suspecious recurrence. Mild non specific inflammation
2	wt	wt	57	F	anorectal cancer (malignant primary site , single tumor) grade 3. Poorly differentiated adenocarcinoma. Patient has metastasis in the adrenal gland
3	wt	wt	29	F	rectal adenocarcinoma grade 2 (lymph node -ve)
4	wt	wt	65	M	Recto segmoid adenocarcinoma, grade 2 (Duke's D)
5	wt	wt	58	F	rectal adenocarcinoma grade 2
6	wt	wt	43	F	rectosegmoid junction adenocarcinoma grade 3
7a,b	wt	wt	64	F	colon rectal multiple tumors, adenocarcinoma grade 2
8	wt	wt	43	M	rectal adenocarcinoma grade 2
9	wt	wt	58	M	rectal, tubulovilous adenoma with severe dysplasia
10	wt	wt	31	F	colon adenocarcinoma grade 2
11	wt	wt	61	M	peritoneum, metastastatic recurrence of mucinous adenocarcinoma
12	wt	wt	35	M	ulcerative colitis with inflammatory pseudo polyp
13	C/A	G/A	57	F	anal canal, adenocarcinoma, grade 2

14	wt	wt	57	M	rectal adenocarcinoma grade 2
15	C/A	G/A	46	F	rectum, multicentric adenocarcinoma, grade 2

Table 5. The sequencing results of *LGALS4* promoter in 15 tissue samples of colorectal diseases. Two out of 15 tissue samples showed ss217320370 and rs73933062 together in the upstream sequence.

4) Eight out of 28 (28.6%) in blood samples of gastrointestinal tumors.

no.	ss217320370	rs73933062	Age	Diagnosis
1	C/A	G/A	57	rectal carcinoma grade II
2	wt	wt	67	colon adenocarcinoma grade II
3	wt +G/A	G/A	36	pancreatic pseudo papillary tumor
4	wt	wt	57	multiple tumors in ovaries, mesentry, bladder
5	C/A	G/A	58	Scanty, fragmented consistent with metastatic carcinoma in lymph node
6	wt	wt	53	hepatocellular carcinoma
7	wt	wt	51	sigmoid colon adenocarcinoma grade
8	wt	wt	58	rectal invasive adenocarcinoma grade 2 negative
9	wt	wt	62	gall badder adenocarcinoma and Overctomy, bilateral ovarian serous cyst adenocarcinoma
10	C/A	G/A	54	rectal cancer grade II
11	C/A	G/A	52	mild chronic colitis
12	wt	wt	63	hepatocellular carcinoma
13	wt	wt	43	colon adenocarcinoma recurrence
14	wt	wt	57	stomach carcinoma
15	wt	wt	52	oesophagus, hyperplastic squamous epithelium
16	C/A	G/A	62	hepatocellular carcinoma
17	wt	wt	64	rectum, tubulovillous adenoma
18	wt	wt	37	desending colon adenocarcinoma grade
19	wt	wt	30	recto-sigmoid junction, mild chronic non specific colitis
20	wt	wt	43	colon, multicentric adenocarcinoma
21	wt	wt	40	rectum, mucinous adenocarcinoma
22	C/A	G/A	44	colon ulcerative colitis
23			57	anal canal, adenocarcinoma grade 2
24	wt	wt	58	rectum, tubulovillous adenoma with severe dysplasia
25	C/A	G/A	30	rectal adenocarcinoma
26	wt	wt	46	rectum, multicentric adenocarcinoma grade 2
27	wt	wt	61	colon inflammation and peritoneum metastatic recurrence of mucinous adenocarcinoma
28	wt	wt	69	metastatic adenocarcinoma of lung, breast, gastrointestinal tract, female genital

Table 6. The sequencing results of *LGALS4* promoter in blood samples of gastrointestinal tumors. Eight out of 28 blood samples showed ss217320370 and rs73933062 together in the upstream sequence.

3.9. Galectin-4 is upregulated in colorectal cancer patients

The mRNA of 26 patients was available in good quality. Thus, the qRT-PCR was carried out for these samples to investigate galectin-4 regulation in CRC patients in comparison

to normal colon. The results were normalised to GAPDH as a housekeeping gene. Surprisingly, 25 cases out of 26 showed upregulation in the mRNA level of *LGALS4* with different range of fold change (figure 9). Mere one samples exhibited downregulation in galectin-4 comparable to the normal control.

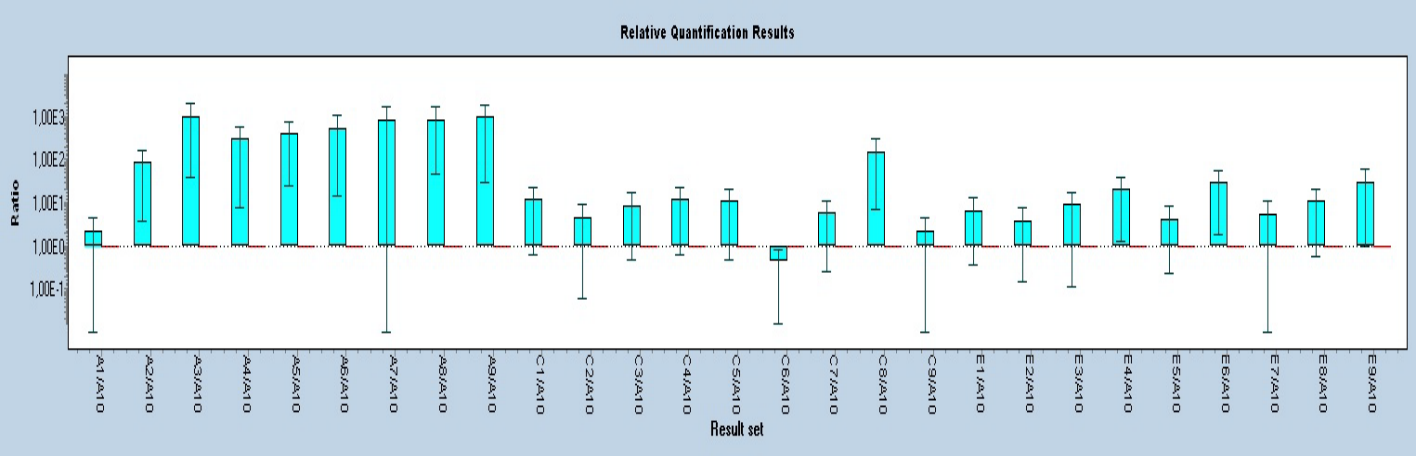


Figure 9. qRT-PCR results of 26 CRC patients. Out of the twenty-six patient, 25 cases showed upregulation in the mRNA level of galectin-4. Only one sample was downregulated among the twenty-six patients.

3.10. galectin-4 upregulation is associated with promoter methylation in the colorectal cancer patients

Only twenty samples were with enough quantity of DNA for bisulfite treatment. Therefore, in order to correlate mRNA expression and methylation status of CpG island galectin-4, twenty samples out of 26 with known galectin-4 expression were sequenced after the bisulfite treatment. All of those twenty samples showed mRNA upregulation of galectin-4. Also, the two SNPs were detected in three of these samples. Surprisingly, 17 patients showed methylated CpG island, which was associated with mRNA upregulation. Just three samples out of the twenty showed unmethylated promoter. Among the three unmethylated samples, one patient (number 7 as shown in table 5) has two tumors, one of them showed unmethylated CpG and second tumor exhibited methylated promoter of galectin-4. whereas, both tumor with unmethylated and methylated CpG islan in this patient showed upregulated mRNA level of *LGALS4* with 789 and 841 fold change respectively.

rs73933062	ss217320370	rs73933062	fold change		gender	age	Diagnosis
1	wt		2.205	methylated	M	30	Known patient for cancer rectum with suspecious recurrence. Mild non specific inflammation
2	wt	wt	85.46	methylated	F	57	anorectal cancer (malignant primary site , single tumor) grade 3. Poorly differentiated adenocarcinoma. Patient has metastasis in the adrenal vgland
3	wt	wt	1008	methylated	F	29	rectal adenocarcinoma grade 2 (lymph node -ve)
4	wt	wt	294.7	unmeth	M	65	Recto segmoid adenocarcinoma, grade 2 (Duke's D)
5	wt	wt	384.5	methylated	F	58	rectal adenocarcinoma grade 2
6	wt	wt	513.3	methylated	F	43	rectosegmoid junction adenocarcinoma grade 3
7a	wt	wt	789.4	unmeth			
7b	wt	wt	841.4	methylated	F	64	colon rectal multiple tumors, adenocarcinoma grade 2
8	wt	wt	930.7	methylated	M	43	rectal adenocarcinoma grade 2

	ss217320370	rs73933062					**Grading**	**T**	**N**	**M**
NCT 1	C/A at -233	G/A	9.017	methylated	M	66	2	4	0	X
NCT 2	wt	wt	21.12	methylated	M	71	2	3	2	1 (Hep)
NCT 3	wt	wt	29.76	methylated	F	71	2	3	0	X
NCT 4	wt	wt	4.096	methylated	M	58	2	2	0	X
NCT 5	wt	wt	28.64	unmeth	M	53	3	4	1	1 hepatic
NCT 6	wt	wt	5.446	methylated						
NCT 7	wt	wt	11.61	methylated	M	56	2	2	0	X
NCT 8	C/A at -233	G/A	4.581	methylated	F	64	2	2	0	X
NCT 9	C/A at -233	G/A	8.703	methylated	F	88	2	4	2	X
NCT 10	wt	wt	5.906	methylated	M	41	3	3	2	1 (Liver, Lung)
NCT 11	wt	wt	151.7	methylated	M	60	2	3	1	X
NCT 12	wt	wt	2.257	methylated	F	53	2	3	1	X

Table 7. The results of promoter sequencing, methylation status, and qRT-PCR of galectin-4 of twenty colorectal cancer patients.

Discussion

The overall 5-year survival rate from colon cancer has increased due to the early detection from increased screening. Most CRCs arise from adenomatous precursors and accumulation of mutations in proto-oncogenes and tumor suppressor genes (TSGs) leads to progression of adenomatous lesions to carcinoma [3,4,5]. High-throughput expression and genotyping arrays are starting to generate novel markers and gene signatures that may be of use in the management of colorectal cancer. However, at present, these are not sufficiently validated to be clinically useful [6].

In the present study, in order to get a gene expression profile overview for early and late stages of colorectal cancer, we have profiled the expression of different Duke's stages of colorectal cancer cell lines. Surprisingly, 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 tumor development and derived from early adenoma cells from microadenomas of patient suffering from hereditary familiar polyposis [87]. Thus, far from tumor classification, the expression profiling results differentiate between good and bad prognosis. Therefore, tracing the differentially regulated genes in LT97 and KM20L2 could be a promise to establish new prognostic markers. Therefore, it was interesting to focus on a gene, which is regulated in the cell lines with proposed bad prognosis aspects (in this study, LT97 and KM20L2 were the candidate of poor prognosis).

Among more than 400 genes that were filtered from analysis of the microarray results, there were several genes were showing differential expression in LT97 and KM20L2. From the literature, interestingly, galectin-4 is expressed in gastrointestinal tissues but not in kidney, brain, skeletal muscles, heart, liver or lung tissues [11]. In our expression profiling data, we have found that galectin-4 was among the genes that are significantly upregulated in LT97 and KM20L2 and not in the rest of cell line, which may give an indication that galectin-4 could have a role in tumor progression or invasion. Consistently, in a previous study, upon expression of galectin-4, epithelial cells acquired a phenotype characterized by the ability to survive lack of nutrients and growth factors for the prolonged period of time. Thus, this cellular phenotype is likely to be advantageous in hyperplastic tissues of premalignant and malignant tumors [11]. The association of poor prognosis of cancer and galectin-4 upregulation is previously reported

[59], which was consistent with our results and interpretation. Moreover, in the present study, among the mRNA of 26 colorectal cancer patient samples that were investigated by qRT-PCR, galectin-4 was found to be upregulated in 25 patients. Therefore, it was interesting to go deeply in galectin-4 regulation via promoter study (from the genetic and the epigenetic aspects).

The direct sequencing results of 0.7 kb of galectin-4 upstream sequence revealed two SNPs; a previously known SNP rs73933062 and SNP at contig 11572034 C/A. Moreover, the two SNPs were found together in the two cell lines, which are upregulating galectin-4 (LT97 and KM20L2). Hence, this gave us the suggestion that two SNPs could influence the promoter activity. rs73933062 is located in the first intron after the TSS, which could be in the putative enhancer sequence. It is encoded in the mRNA but not translated to a protein. However, rs73933062 is not previously reported for clinical association. SNP at contig 11572034 C/A is not previously annotated and we have submitted it as ss217320370. It is located before the TSS; therefore, it could be in the putative distal promoter.

To investigate the frequency of the two SNPs in colorectal cancer patient and also whether the two SNPs are together as a genotype in the same individual, we have sequenced the upstream sequence of 115 patients collected from NCT, Heidelberg, Germany and ENCI, Cairo, Egypt. In total 26 out 115 patients showed the two SNPs together (ss217320370 (11572034 C>A) and G>A which is known as rs73933062) as the following: Four out of 18 (22.2%) colorectal cancer patients from NCT, Heidelberg, Germany, 12 out of 54 patients (22.2%) showed the two SNPs in the lesion tissues and their adjacent normal tissue (8 out of 27 (29.6%) tissue from colorectal cancer patients, 4/15 (26.7%) patients with ulcerative colitis, and 0/12 with adenomatous polyps from ENCI, Cairo, Egypt), Two out of 15 (13.3%) colorectal tumors from ENCI, Cairo, Egypt, and 8/28 (28.6%) in blood samples of gastrointestinal tumors. Conclusively, the two SNPs were showed always together in 26 out of the 115 (22.6%) patient samples.

The two SNPs influenced the promoter activity. To conclude that, the upstream sequence has been cloned in luciferase reporter plasmid (pGL4.10) and transfected to Co115 cell line. As a result, the luciferase reporter of the upstream sequence containing either one of the two SNPs or both showed significant upregulation of the promoter activity comparable to normal genotype. The presence of mere rs73933062 in the luciferase construct showed 1500 fold change in the luciferase activity with p-value 0.002. To validate the huge fold change

increase of luciferase activity in case of rs73933062, two more cell lines were transfected (SW1116 and Sw620) with the luciferase construct with the investigated SNP. Interestingly, the presence of rs73933062 also caused an increase in the luciferase activity with 37 and 36 fold changes in SW1116 and SW620 respectively. Transfecting CO115 with luciferase construct that contains ss217320370 showed a mild change in the luciferase activity, 1.3 fold change with 0.03 p-value. The combination of ss217320370 and rs73933062 increased the promoter activity four times with 0.004 p-value. From the insight, we thought that the combination of the two SNPs will increase the promoter activity by value bigger than the presence of solitary SNP. However, the results showed that the combination of the two SNPs increased the promoter activity with a value less than 1500 that caused by rs73933062 alone, for instance. That could be referred to the distance between the two SNPs, which could contain silencer binding site.

Our results showed that the presence of ss217320370 and rs73933062 changed the transcription factors binding site. The verification of this result was carried out by using the upstream sequence as bait for protein pull down followed by mass spectrometry. The results showed the general binding proteins that bind to both genotype and some unique proteins for each genotype. Therefore, pull down results support the luciferase reporter assay results. Thus, the two variations caused deletion and insertion of regulatory protein binding sites. Basically, the upstream sequence of galectin-4, regardless to the genotype does it have, binds to transcription factors such as FUBP1, FUBP2 , FUS, LBP-1a, ENO1, and TFCP2.

As mentioned above, each genotype variant binds to some unique proteins in addition to basic binding regulatory proteins. Among the distinctive proteins of those binders that tethered to genotype containing C and G at the contig position 11572034 and 11571652 respectively: 1) Aconitase 1 (ACO1), also known as iron regulatory element binding protein 1 (IREB1), is a cytosolic protein which regulate ferritin mRNA via its binding to iron-responsive elements (IREs). Therefore, its binding to IREs results in repression of ferritin 5`-UTR mRNA [88]. Since G at the contig position 11571652 is transcribed and included in the mRNA sequence but not translated, the pull down of ACO1, in the case of G at the contig position 11571652, could indicate that galectin-4 is negatively regulated by ACO1 via pre-mRNA. 2) RBBP7 (retinoblastoma binding protein-7) which is found previously among several proteins that binds directly to retinoblastoma proteins that regulate the cell proliferation. The encoded protein is found in many histone deacetylase complexes [89,90]. It is also known by limiting the expression of estrogen-responsive genes [91]. 3) PUF60 (poly-U binding splicing factor 60

Kda) or also known as FIR (FBP Interacting Repressor): the protein forms a tenary 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 [92]. Since the upstream sequence of galectin-4 binds to FUBP1 and FUBP2 regardless to the variations, binding to PUF60 (in the case of C and G at the contig position 11572034 and 11571652 and not A and A at those two position) is supporting the luciferase reporter assay. In other word, PUF60 could repress galectin-4 promoter via binding FUSE-binding proteins. 4) Several other proteins with numerous functions ranging from transcription activation to repression such as BAT1, CGI-46, chromatin assembly factor-1 subunit B, DAZ associated protein-1.

On the other hand, in the case of C>A at position 11572034 and G>A at position 11571652, 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 [93]. However, it is also previously investigated as a novel co-repressor of NF-κB [94]. In addition, some other proteins are binding to the two SNPs (C>A at position 11572034 and G>A at position 11571652). Some of those proteins are not previously known as DNA-interacting proteins such as PRDX1, fatty acid binding protein-5 (FABP5), apolipoprotein D, and squamous cell carcinoma antigen.

In conclusion, from pull down - mass spectrometry results, it seems that changing of single nucleotide, is mostly deleting repressor binding site(s) rather than inserting new effectively transcriptional activating proteins. Activating or repressing function should be more investigated in the future research.

Since the two SNPs are influencing the binding sites for DNA-binding proteins, also the regulation of the TFs are influencing the promoter activity. For instance, if a given transcription silencer binds to a promoter and this silencer is downregulated or not expressed in the cells. In this case, it makes no sense if this silencer is binding to the promoter or not, since it is not expressed in the cells. To correlate the expression of the transcription factors (which bind to the either wild or mutant type promoter) and the expression of galectin-4, qRT-PCR was done for ten genes that are encoding the binding proteins. Among those proteins, for instance, ENO1 which we found to bind to galectin-4 promoter in the presence or the absence of the twin SNPs, is known previously, in terms of function, as transcription repressor [95,96]. The mRNA

expression of ENO1 was upregulated in five out of six cell lines. The only cell line which didn't show ENO1 upregulation is LT97, which could be correlated with the upregulation of galectin-4 in this cell line, since the ENO1 is a transcriptional repressor. Another example, RBBP7, is found in many histone deacetylase complexes, including mSin3 co-repressor complex. It is also present in protein complexes involved in chromatin assembly [97]. Also, in the qRT-PCR results, RBBP7 was found to be upregulated in all cell lines. RBBP7 bound to the wild type promoter and not in the presence of the SNPs which showed by pull down-mass spectrometry results. Therefore, the expression of RBBP7 in LT97 and KM20L2 might not influence the promoter activity and hence the gene expression, since the two cell line have ss217320370 and rs73933062. However, the overexpression of RBBP7 could give a clue to interpret the downregulation of galectin-4 in SW1116, SW480, SW620, and CO115, since it could function as co-repressor. The qRT-PCR results for ten selected genes, which are encoding the binding proteins, didn't give a concrete base to understand the mechanism of each binding protein whether they activate or suppress galectin-4 expression, even with the previous knowledge of each protein function. However, since we used CO115 nuclear protein in the pull down experiment, the qRT-PCR results for the selected genes in CO115 cell line could be positive control for pull down-mass spectrometry analysis, since all genes were upregulated in CO115.

The CpG island was predicted in the first exon and the first intron. The methylation status of galectin-4 promoter in the cell lines was consistent with the qRT-PCR results. So, LT97 and KM20L2 with upregulated galectin-4 showed unmethylated promoter. Whereas, the rest of the cell lines (SW1116, SW480, SW620, and CO115) showed methylated promoter. In colorectal cancer patients, the situation is much more complicated. While the twenty samples that have been investigated for methylation status were upregulating galectin-4, just three samples exhibited unmethylated promoter and the rest were methylated. Three out of the 17 methylated samples have the two SNPs. In these three samples with the two SNPs and methylated promoter, the galectin-4 upregulation could be referred to the presence of the two SNPs as discussed earlier regarding the effect of ss217320370 and rs73933062 on the promoter activity. On the other hand, 14 samples were found with upregulated galectin-4 and methylated promoter, in parallel, they don't contain ss217320370 and rs73933062. In this case (promoter methylation and no SNPs to activate the promoter), the methylation might play a role in preventing repressor(s) proteins from the binding to their response element in the promoter sequence. Therefore, methylation could be a mechanism to activate the galectin-4 promoter and

subsequently increase the gene expression, but it needs more investigation in the future research.

In conclusion, in the present study, galectin-4 upregulation was detected in most of colorectal cancer patient and two cell lines with potential susceptibility to develop colorectal cancer (LT97) and the Duke's D cell line (KM20L2), which could indicate a prognostic value of galectin-4 in colorectal cancer. The twin SNPs (ss217320370 and rs73933062) in the upstream sequence and the first intron were associated with upregulation of the galectin-4, which could be referred to changing in the protein binding site according to our pull down-mass spectrometry results. Therefore, the twin SNPs could have a medical impact in cancer via galectin-4 upregulation. Also, methylation was not found to be a mechanism for galectin-4 downregulation.

Outlook

Recently, an interaction was found between galectins and the blood group antigens. ABO(H) blood group antigens are composed of carbohydrate structures that only differ by distinct monosaccharides on the terminal structures of glycans [98]. Human galectin-3, -4, and -8 were shown to recognize multiple glycan structures [78,99]. In a recent study, it was also shown that galectins recognize blood group-positive bacteria and hence play a role in the innate immunity. Moreover, it has been reported that galectin-3, -4 and -8 bind specifically to blood group A and B at submicromolar concentrations but not to blood group O (H) [79].

By combination of the above-mentioned facts, we postulate that galectins enable circulating tumor cells, which are responsible for metastasis formation, to interact with red blood cells that are carrying a blood group antigen. We performed pilot experiments to check the validity of our assumption and the results support the hypothesis, although they are only preliminary in nature. Thus, we are aiming at an investigation of the functional aspects of galectin upregulation in cancers with bad prognosis and an analysis of their role in stabilizing circulating tumor cells by interaction with erythrocytes (figure 10).

In a preliminary analysis, we have looked at the interaction of blood and tumor cells and its effect on the tumor cells. In previous studies, we had found that galectin-4 showed

significant up-regulation in KM20L2, a Duke's D tumor cell line. KM20L2 was then used to study possible interactions of tumor cells that up-regulate galectin-4 and erythrocytes.

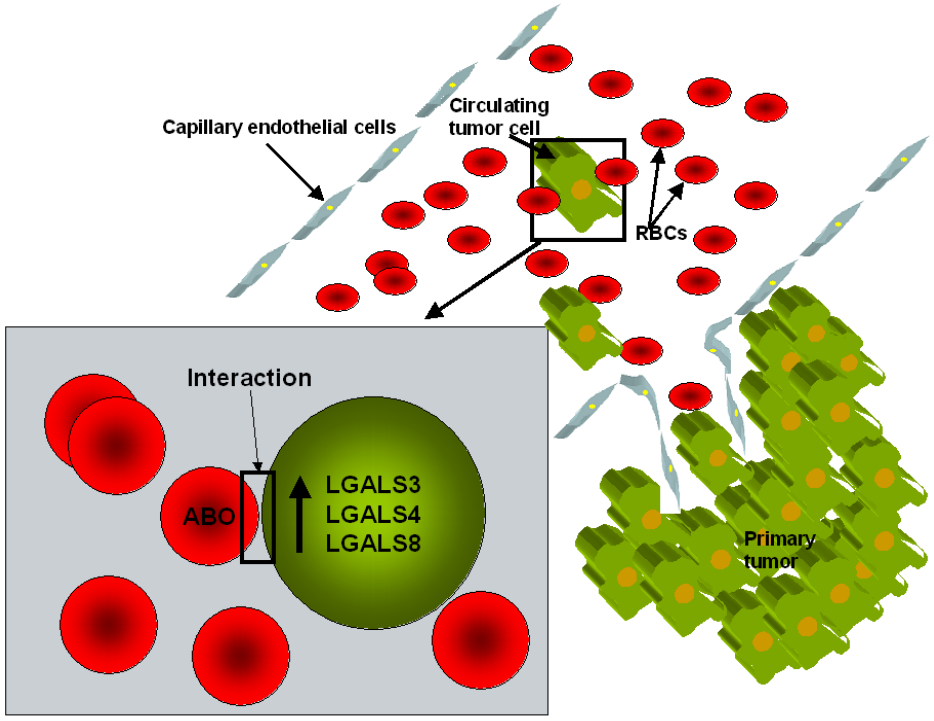


Figure 10. Schematic representation of the results of the pilot study. Tumor cells that exhibit up-regulation of galectins (LGALS) 3, 4, and 8 interact with red blood cells (RBCs), if blood of antigen group AB is used. Galectins accumulate at the sites of attachment to the erythrocytes.

We obtained blood from healthy individuals with blood groups AB and O. About 100,000 erythrocytes were added to 10,000 KM20L2 cells and incubated. Then, the respective mixture was smeared on glass slides and stained with haematoxylin-eosin or immunostained using a galectin-4 antibody. In the immunostaining, galectin-4 was found to occur predominantly at the cell membrane locations, which were in contact with the erythrocytes (figure 11).

In another experiment, KM20L2 cells were incubated with erythrocytes from blood group AB, in parallel to KM20L2 cells without blood cells. The live cells were observed with a Zeiss cell observer for five hours. Every two minutes, a picture was taken. At the end of the experiment, the recorded pictures were analyzed using the Axiovision documentation (Zeiss) and a video of the cellular interactions was obtained. The interactions of tumor cells and

erythrocytes started during the first 30 mins. Eventually, the cells separated again after 90-120 min; this is probably due to galectin-4 secretion and saturation of the erythrocyte membrane with the secreted galectin-4.

When KM20L2 cells were incubated without erythrocytes, the tumor cells underwent apoptosis. From these experiments and the known functions and association of galectin-4 with tumors, we postulate that the erythrocytes are serving as a substrate for tumor cell anchorage and that this interaction prevents their apoptosis while circulating in the blood. This phenomenon could be of great interest for understanding how to destabilize circulating tumor cells and thus to target metastasis formation.

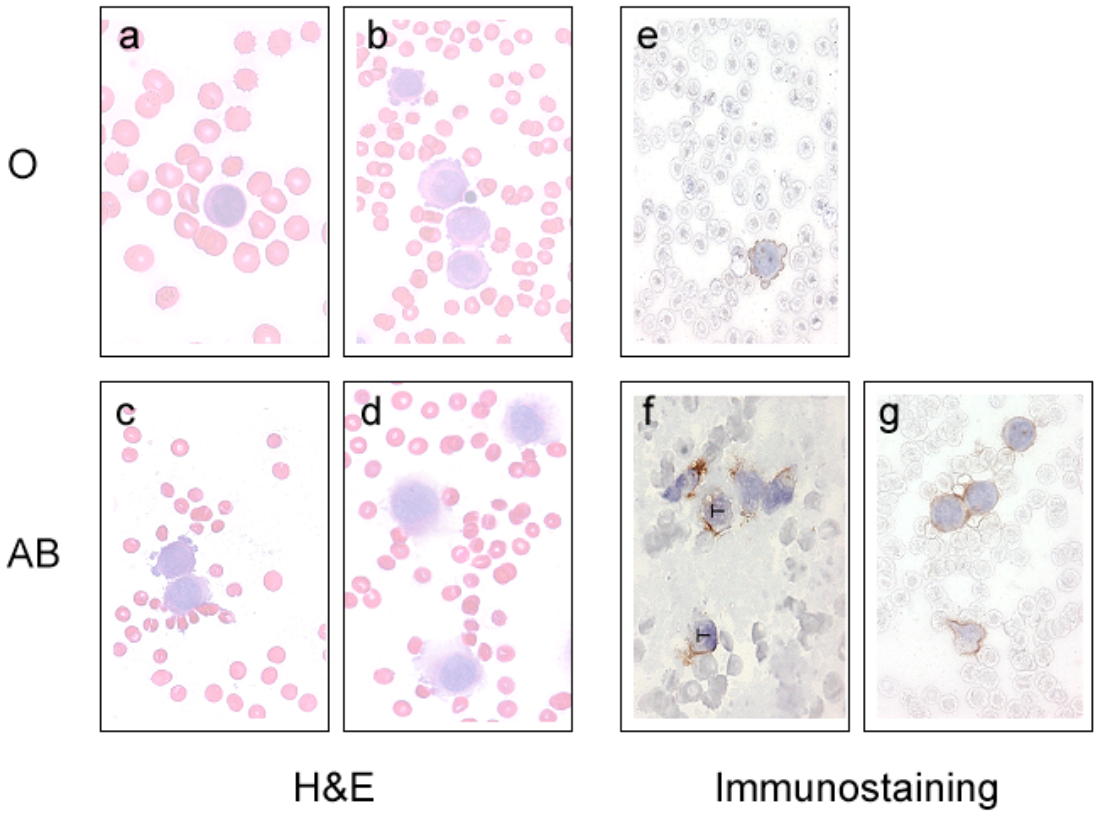


Figure 11. Staining of mixtures of tumor cell and erythrocytes. Staining was with haematoxylin & eosin (H&E, a-d) and by galectin-4 immunostaining (e-g) using blood of AB and O groups. In the case of AB-blood, there was an obvious binding between the tumor cells and the erythrocytes. Interestingly, galectin-4 protein was accumulated at the cell membrane locations of the tumor cells that were in contact to erythrocytes. This could indicate chemotaxis to blood group antigen. In case of blood group O, galectin-4 protein was evenly distributed in the cell membrane.

Part II: Setting up Transcription Factor Protein Array Detecting DNA-Protein Interactions

1. Introduction

Determination of the sequence of the human genome and knowledge of the genetic code has allowed rapid progress in the identification of mammalian proteins. However, far less is known about the molecular mechanisms that control expression of human genes and about the variations in gene expression that underlie many pathological states, including cancer. This is caused in part by lack of information about the binding specificities of DNA-binding proteins and particularly regulative important molecules such as transcription factors (TFs). It is consequently crucial to develop technologies for the detection of DNA-protein interaction in order to identify DNA response elements and the related transcription factors or other DNA-binding proteins. The techniques vary with respect to the type of result that can be expected from the assay: a mere qualitative demonstration of binding; the identification of response element sequences at high-throughput; or a quantitative characterisation of affinities.

For many years, biochemical processes were used to characterise DNA-protein interactions. However, such approaches are generally laborious and slow. Recent decades have witnessed the development of technologies that permit analyses at a larger scale and in a more unbiased manner. These approaches are often either gene-centred or protein-centred [100]. In gene-centred approaches, an individual protein is used to identify DNA target sequences. Inversely, an individual DNA sequence is used to identify and study the relevant DNA binding proteins in a protein-centred assay. Recently, microarray and sequencing technology were employed toward genome-wide analyses in both formats. As a result of these developments, more and more regulatory elements and factors are uncovered. In consequence, there is a continuously growing understanding of many basic regulatory processes that involve DNA-protein interaction.

In the present study, a suitable condition was optimised to pursue DNA-protein interaction on chip. For this purpose, the transcription factor proteins were expressed, purified, and spotted on appropriate surface and subsequently incubated with target DNA sequences.

1.1. DNA-protein interaction

Wide ranges of proteins are crucial for successful transcription by RNA polymerase in eukaryotic cells via binding to the cis-regulatory elements, such as promoters and enhancers. These proteins include general transcription factors, co-factors, histones, chromatin remodelling proteins, and sequence-specific transcription factors that direct transcription initiation to specific promoters [101,102,103].

Promoters consist of common sequence elements, such as a TATA box and an initiator sequence, and binding sites for other transcription factors, which work together to recruit the general transcriptional machinery to the transcriptional start site (TSS). Enhancers are located some distance from the site of transcription initiation and also contain binding sites for transcription factors. Transcriptional activity from general factors binding to the core promoter is usually low, but the binding of site-specific factors to proximal promoter regions can increase it. Promoter activity can be further stimulated by the binding of factors to distal enhancer regions and the subsequently recruit histone-modifying enzyme that creates a more the proper chromatin environment for transcription or of a kinase that induces a bound initiation complex to begin elongation. Repressive factors bind also to silencing sequence far from the TSS, which can interfere with activator binding [102,104,105].

The location relative to the TSS, at which a factor binds, is of interest, since it can provide insight into the mechanisms by which the factor regulates transcription. In other words, factors that bind close to TSS have been proposed to regulate transcription by stabilizing general transcription factors at the core promoter elements. On the other hand, factors that bind to distal regions, either upstream or downstream of a gene, may regulate transcription by mediating, through a looping mechanism, the protein–protein contacts between distal complexes and the general transcriptional machinery bound at TSS. Therefore, comprehensive analysis of the binding locations of a factor not only allows the development of a genomic map but also provides insight into the mechanisms for how this factor regulates transcription [104].

1.2. Analysis of DNA-protein interactions; from nitrocellulose filter-binding assays to microarray studies

1.2.1. Nitrocellulose filter-binding assay

This assay was developed in the early stages of molecular biology in the 1970s. The manipulations are rapid enough to allow kinetic studies as well as equilibrium measurements [106,107,108]. The process is based on the fact that proteins bind to nitrocellulose without loosing their DNA-binding capacity, while double-stranded DNA alone is not retained. For analysis, DNA and proteins are mixed and incubated under appropriate conditions. The mixture is then separated by electrophoresis and subsequently blotted onto nitrocellulose. Since only proteins bind, DNA remains on the membrane only if in complex with a protein. The exact amount of DNA that is stuck to the nitrocellulose can be quantified by measuring a label that is introduced to the DNA prior to incubation with the protein. The amount of information obtained from this kind of assay is limited, however. Only the mere retention of labelled nucleic acid is detected and not the identity of the proteins involved or the proportion of binding activity attributable to an individual protein, if more than one protein in the mixture exhibits DNA-binding capacity [109]. Moreover, as a technical complication, single-stranded nucleic acids are retained at nitrocellulose filters under particular conditions, resulting in background that can obscure the measurement [110].

1.2.2. DNase I fingerprinting

Fingerprinting assays (figure 12) exploit the fact that a protein, which is bound to a specific nucleic acid, will interfere with a chemical or enzymatic modification of that DNA-fragment. Thus, the modification can be used to localise the contact area between protein and DNA [111]. For the DNase I fingerprinting assay [112,113], a particular DNA-fragment is labelled at one end and mixed with the protein of interest. Following binding, the DNA is treated with the enzyme deoxyribonuclease I (DNase I), which digests DNA that is not in close contact to a protein and thus not protected from digestion. Performing a partial cleavage without protein produces labelled DNA-fragments that – because of the random nature of cleavage – cover the entire size range of the original DNA. In the presence of a binding protein, however, protection occurs in a particular region and labelled DNA-fragments of the respective length are not produced, while all longer or shorter ones are still present. Resolving the two samples on a polyacrylamide gel side by side, the differences in the resulting ladders of DNA

bands is visualised via the incorporated label. Gaps in the band ladder of the sample, in which protein had been present, indicate binding sites. Comparison of the patterns with sequencing reactions allows the identification of protected sequences with single-nucleotide resolution.

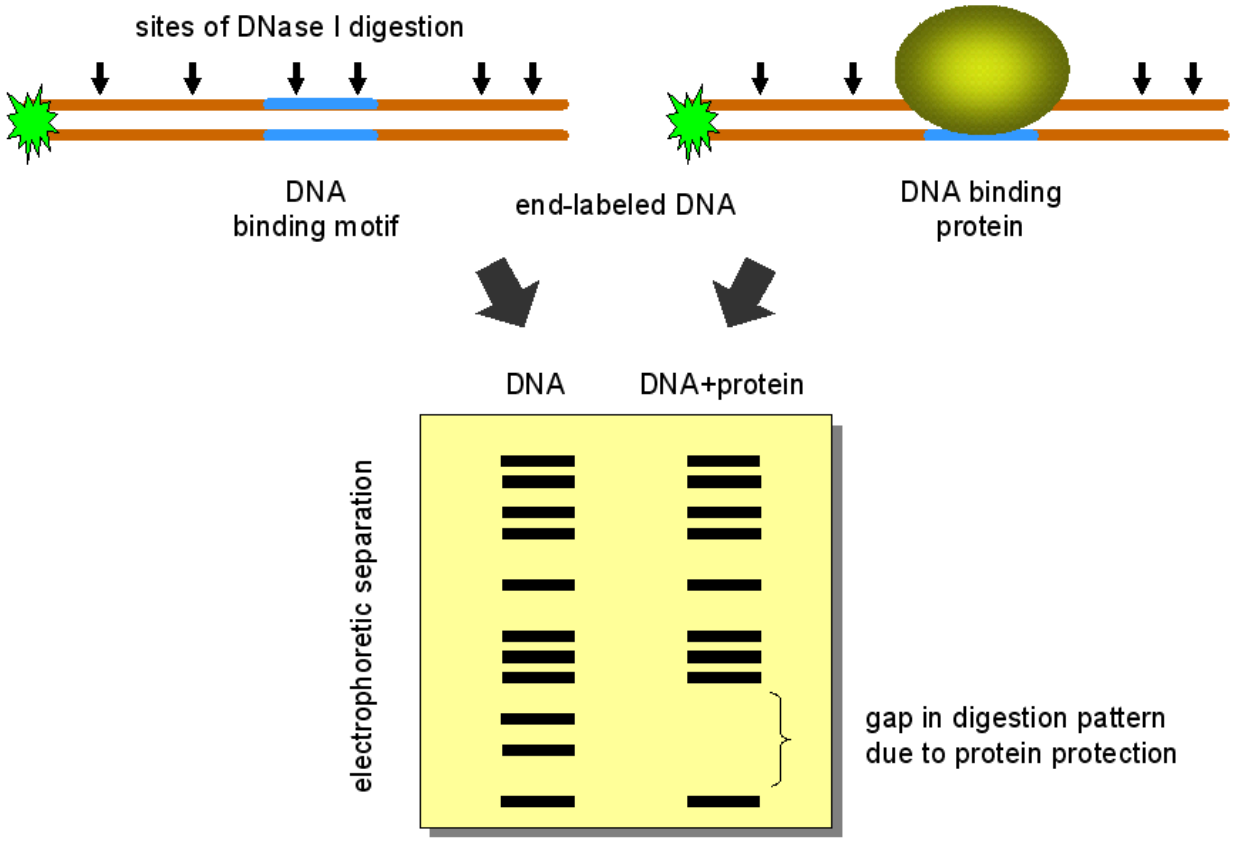


Figure 12. Schematic representation of DNase I footprinting. Further details on the process are given in the text.

1.2.3. Dimethyl sulphate (DMS) protection fingerprinting

DMS protection fingerprinting is a chemical variant of the enzymatic DNase I fingerprinting and relies on the ability of DMS to methylate specifically guanine residues in DNA. The methylated G residues are cleaved by exposure to piperidine, whereas no cleavage occurs at unmethylated bases [114]. A protein bound to a DNA will protect the G residues from methylation and hence from cleavage by piperidine.

1.2.4. Electrophoretic mobility shift assay (EMSA)

In comparison to the assays described above, EMSA is a relatively rapid method to detect particular DNA-protein interactions [115,116]. It relies on the fact that the electrophoretic mobility of a complex of nucleic acid and protein is less than that of the free nucleic acid. Mobility-shift assays are often used for qualitative purposes. However, under

appropriate conditions, they can even provide quantitative data for the determination of binding stoichiometries, affinities and kinetics. The exact methods used differ for each purpose as reviewed elsewhere [117,118]. Despite of the fact that the technique is widely used to detect DNA-protein interactions, it has several limitations. Rapid dissociation during electrophoresis can prevent detection of complexes. Also, many complexes are significantly more stable in a gel than they are in free solutions [119]. In addition, EMSA is not informative about the sequences of the nucleic acids that are bound by the proteins. Another limitation is the fact that the observed mobility shift does not provide a direct measure of the molecular weight or identity of the proteins responsible. However, modifications such as an electrophoretic supershift assay and procedures that combine EMSA with Western blotting or mass spectrometry have been designed to identify the DNA-binding proteins [118].

1.2.5. Methylation interference assay

This procedure is also based on the ability of dimethyl sulphate (DMS) to methylate G residues, which in turn are cleaved by piperidine [114] and the fact that methylation of purine residues in a DNA sequence inhibits formation of a DNA-protein complex [120,121]. For the assay, an end-labelled DNA probe is partially methylated with DMS and incubated with a nuclear protein extract. The protein-DNA complexes formed during the incubation are then separated from the free DNA using EMSA. Both the protein-bound DNA and the free DNA are eluted from the gel, cleaved with piperidine and again resolved by denaturing polyacrylamide gel electrophoresis. If methylation had occurred at a particular guanine residue that is critical for the DNA-protein interaction, the binding of the protein to that DNA was inhibited, resulting in the recovery of the DNA only from the free DNA fraction. The presence of particular DNA-fragments in the free DNA fraction and their concomitant absence from the DNA-protein fraction indicates that those nucleotides are contact points of proteins.

1.2.6. Chromatin immunoprecipitation (ChIP)

ChIP is a method to identify DNA-protein complexes that occur *in vivo* and currently the standard method for the identification of histone modification locations and transcription factor binding sites [122]. Cells are initially treated with a cross-linking agent to link covalently any DNA-binding protein to the chromatin. Then, the cells are lysed and the genomic DNA is isolated and sonicated to produce sheared chromatin. An antibody specific to the protein of interest is added to the sonicated material and used to isolate the protein with all attached DNA via immunoprecipitation. The DNA is released by reversing the cross-linking and purified

subsequently. Classically, the DNA obtained from ChIP reactions was assessed by PCR, details of which are reviewed elsewhere [123]. In combination with hybridisation microarrays (ChIP-on-chip), the assay became much more powerful, identifying the binding sequence and its location in the genome by hybridisation to a particular array feature [124]. ChIP-seq is a more recent alternative, using second generation DNA sequencing instead of microarray technology for sequence identification [125]. In any combination, ChIP has several inherent features that can make the identification of DNA-binding sites difficult. Particularly lacking specificity of the DNA-binding protein or the antibody used for precipitation can result in an experiment with insufficient enrichment [126].

1.2.7. DNA adenine methyltransferase identification (DamID)

Like ChIP, DamID is an assay to detect DNA-protein interactions *in vivo*. The target protein is expressed as a fusion molecule with DNA adenine methyltransferase (Dam) originally expressed in *E. coli*. Dam methylates the adenine of a GATC site. Therefore, when the fusion-protein binds to DNA, the Dam portion will methylate GATC sites that are located in the vicinity of the binding sites. The methylated sites in the target DNA and a control sample (expression of Dam alone) are detected by digestion with methyl-specific restriction enzymes, followed by amplification, labelling and hybridisation to a microarray as for a ChIP-on-chip assay [127]. DamID has been used to identify binding sites of proteins in different organisms and targeted the binding sites for sequence-specific TFs, DNA methyltransferase and chromatin-associated proteins using PCR-amplicon arrays or 60-mer oligonucleotide tiling arrays.

DamID has some advantages compared to ChIP. For example, DamID is not dependent on the availability of high-quality antibodies. Moreover, there is no need to use crosslinking reagents, eliminating the risk of crosslinking artefacts. Third, DamID can be performed on about 10^6 cells, which is 10- to 100-fold less material than typically used in ChIP experiments [128]. However, DamID has also limitations. It requires an exogenous fusion protein, whereas ChIP can be performed with the endogenous protein; some proteins lose their genomic binding specificity when fused to Dam. DamID is also less suitable for the detection of rapid changes in protein binding (e.g., during the cell cycle), because the methylation patterns obtained in a typical DamID experiment represent the average of a time period of about 24 h or more. Last, DamID cannot be used to map post-translational modifications such as histone modifications; for such applications, ChIP-on-chip is more suitable.

1.2.8. Surface plasmon resonance (SPR) measurement

Surface plasmon resonance measurement [129,130] is an optical technique that allows studying the interaction between an immobilised molecule and an analyte that is in solution. It relies on the change in the refractive index of solutions adjacent to a surface upon an increase in surface thickness, which is caused by analyte binding. In studies aimed at DNA-protein interactions, either the DNA-molecule is attached, for instance by means of biotinylation, or a protein is immobilised by tags, such as poly-histidine or glutathione-S-transferase. The system offers real-time recording of the association and dissociation of the analyte at the immobilised ligand, thus permitting rapid and accurate stoichiometric kinetic, affinity, and thermodynamic measurements [131]. With regard to quantitative measurements, this procedure is currently the gold standard for protein interaction analysis.

1.2.9. Systematic evolution of ligands by exponential enrichment (SELEX)

SELEX has been used to identify high-affinity nucleic acid ligands for a large number of proteins. It exploits the power of genetic selection and the advantages of *in vitro* biochemistry. The assay is done by selecting a subset of oligonucleotides from a complex mixture of nucleic acid sequences. This is achieved by incubation of the DNA with the investigated protein, separation of bound molecules from the unbound fraction, release and amplification. This process is repeated iteratively until the nucleic acid sequences that bind to the protein with high affinity are enriched significantly [132]. The SELEX method enabled *in vitro* selection of the optimal binding sites of several TFs [133], for example. A limitation is the fact that the system does not allow to define the exact *in vivo* selectivity of proteins but is aiming at the identification of the best binding DNA-targets.

1.2.10. Yeast one-hybrid system

The yeast one-hybrid system [134] is conceptually similarly designed to the yeast two-hybrid system that is used for the detection of protein-protein interactions. A DNA sequence of interest (the DNA bait) is cloned upstream of a reporter gene and integrated into the yeast genome by site-specific recombination. At the protein side, a hybrid protein is generated by fusion of the prey protein to a transcription activation domain. When the prey protein and the DNA bait physically interact with each other, the reporter gene expression is activated. By generating libraries of DNA-fragments and fusion proteins, complex analyses can be performed. With this system, 283 interactions between 72 *C. elegans* digestive tract gene

promoters and 117 proteins could be identified [135], for example. A similar system but based on bacteria rather than yeast has also been described [136].

1.2.11. Proximity ligation

A technique that permits an analysis even at single-molecule level is proximity ligation [137]. It is a method for very sensitive solution-phase detection of interaction partners. While mostly useful for studying protein-protein interactions, also the DNA-binding of proteins can be investigated. To this end, three probe molecules are required; one is an antibody against the investigated protein, the second is the DNA recognition sequence in form of a double-stranded DNA. Both molecules have attached a single-stranded DNA-tag segment. The third molecule needed is an oligonucleotide that is complementary to the ends of both DNA-segments. The target protein binds to the DNA probe and is subsequently detected by the antibody. In consequence, the two DNA-segments attached to antibody and DNA-fragment will come in close proximity. Only then they can hybridise to the oligonucleotide, which acts as a connector (figure 13). In the resulting double-stranded sequence, the DNA-segments of antibody and recognition sequence can be joined by an enzymatic ligation. The resulting sequence can be specifically amplified and detected in real-time by PCR or rolling circle amplification [138].

While this technique is powerful in terms of specificity and sensitivity, it is very dependent on the availability of specific antibodies. Also, extensive preparative steps are required in order to perform the analysis. In principle, the approach permits a parallel investigation, as long as specific sequences are attached to each pair of probe molecules. Also, the method is less suited for the identification of new binding sites, since no connector molecule would be available for their detection. For *in vivo* studies of defined protein-DNA interactions at the level of single-molecule sensitivity, however, the method could become enormously useful for its sensitivity and specificity.

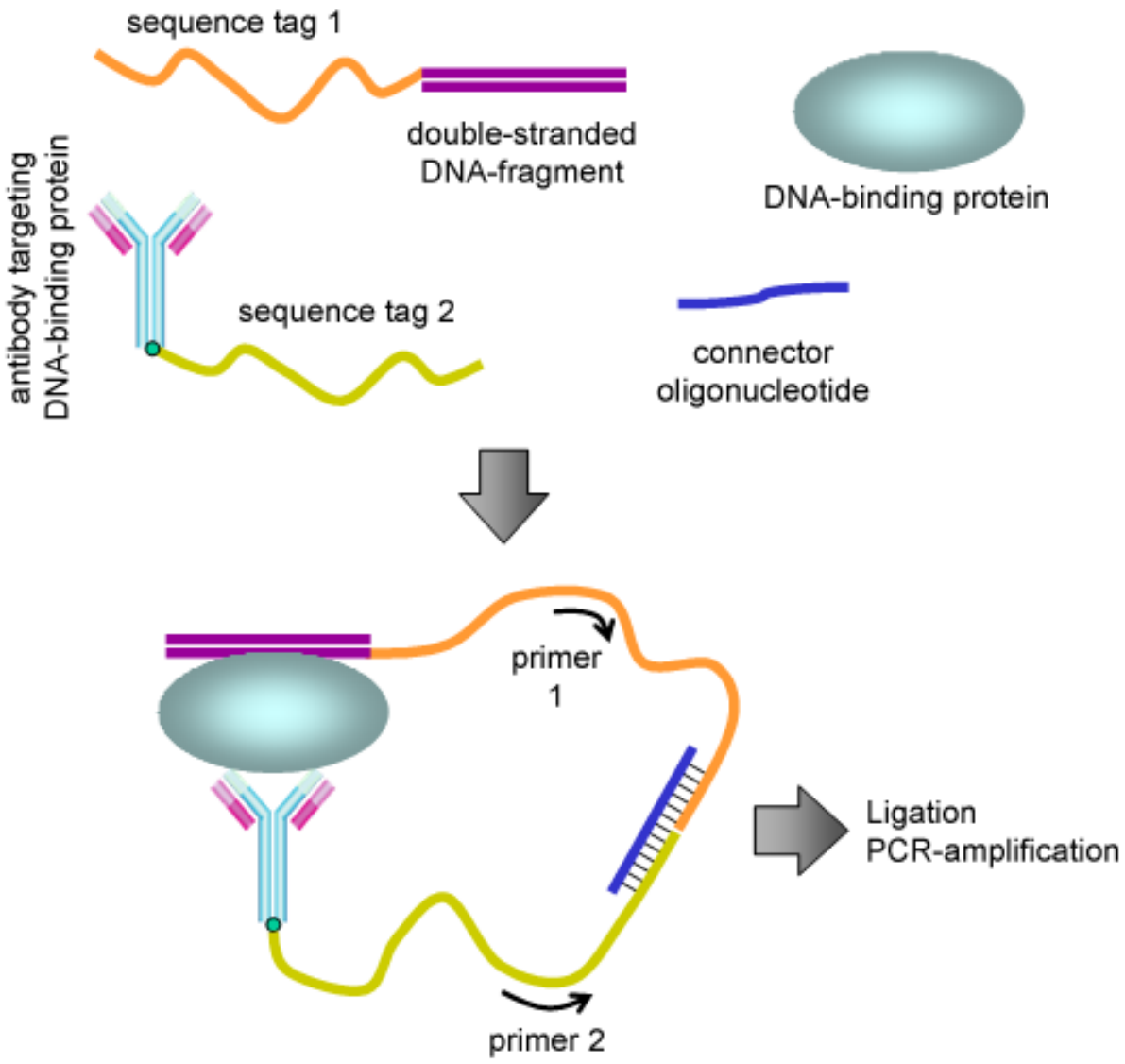


Figure 13. Schematic presentation of a proximity ligation assay. Upon incubation of a DNA-fragment, an appropriate protein and an antibody that binds the protein, a complex of the three molecules is formed. Both the DNA-fragment and the antibody are tagged with single-stranded DNA-segments. After addition of an oligonucleotide that is complementary to both ends of the tag sequences, the DNA molecules are joined by ligation and can then be amplified by PCR.

1.2.12. Microarray-based assays

For the achievement of high-throughput analysis, microarray technology was an obvious candidate. In consequence, studies have been performed that employed microarrays for the detection of DNA-protein interactions. Either DNA or protein microarrays were utilised to this end.

1.2.11.1. *DNA microarrays*

As already reported above, ChIP-on-chip and microarray-based DamID improved the throughput of the basic methods considerably and are currently still the most widely used array-based methods for the identification of transcription factor binding sites. However, DNA-microarrays can also be used for a direct analysis of DNA-protein interactions. Yet, since most sequence-specific DNA-binding proteins bind to double-stranded DNA, processes had to be developed to create double-stranded DNA-molecules on chip surfaces. DNA-microarrays used for other purposes are usually made of single-stranded DNA.

In some studies, PCR-products were printed on microarrays [139]. As an alternative, and superior in terms of investigating the specificity of binding, analyses were performed on oligonucleotide arrays. Double-strand formation can be achieved by synthesising long oligonucleotides, which consist of self-complementary sequences, therefore forming hairpin structures. Alternatively, the second DNA-strand can be produced enzymatically by means of a sequence that is common to the 3'-end of all single-stranded oligonucleotides on the array. Upon addition of a complementary primer molecule and an appropriate polymerase, the initially single-stranded oligonucleotides are converted to DNA duplexes [140].

For monitoring protein binding, first the protein of interest is expressed. Frequently, an expression system has been used that adds an epitope tag. Such a tag has two functions: first, it is used to isolate the protein by affinity-purification and, second, it permits detection by means of an epitope-specific reporter, such as an antibody. Alternatively, directly labelled proteins can be used in the assay. The protein is incubated on the microarray and the signal intensities obtained at the various array features are measured. The method was applied successfully in several studies for an analysis of proteins that had been uncharacterised before [141,142].

In terms of coverage, microarrays were used that contained PCR-products, which covered entire genomes [142,143]. Microarrays made of long PCR-products have the advantage that they cover much sequence space with relatively few microarray features. At the same time, however, they have the disadvantage that the probability of calling a binding event correctly is less for a single binding site, which is embedded in a long rather than a short sequence. Moreover, depending on the number and types of candidate binding sites within a single region, interaction may occur once or several times by one or several proteins at various degrees of affinity [144]. From the blend of information gained from such measurements, it

may well be impossible to extract how many interactions are involved, let alone any accurate information about strength and specificity. For such ends, arrays that consist of short synthetic double-stranded oligonucleotides exhibit superior performance [145,146]. While coverage of all binding sites in a genome is unlikely to be achieved with oligonucleotide arrays – and unnecessary because of the advent of sequencing techniques that are better suited for this purpose – arrays have been produced that are comprehensive with respect to the binding sequences, presenting all 10 bp sequences possible for example [147,148].

1.2.11.2. *Protein microarrays*

While analyses on DNA-microarrays promise a rather comprehensive identification of the target sequences of a specific transcription factor or other DNA-binding protein, they are not able to identify the various proteins, which recognize a particular sequence of interest. This is made possible, however, by reversing the assay format. Studies on protein arrays that represented 282 potential yeast TFs [149], 802 *Arabidopsis* TFs [150], or 4191 non-redundant human DNA-binding proteins [100], for example, demonstrated the potential of this approach. The arrays were created using epitope-tagged proteins, mostly by fusion to glutathione-S-transferase. Subsequent to purification by affinity chromatography, the proteins were attached to the chip surface and incubated with double-stranded oligonucleotides. Each oligomer was made of three or four repeats of the relevant sequence motif. In human, for example, a total of 17,718 DNA-protein interactions between 460 DNA-motifs and 4,191 human proteins were identified [100]. Among them was a large number of interactions of TFs with DNA sequences, which had not been anticipated before. Also, the binding characteristics of TFs were identified this way. Recent developments toward the production of comprehensive protein arrays [151,152,153] should permit an extension of the format. However, the structural integrity of the proteins or at least of their DNA-binding domains is of critical importance for a successful application and is unlikely to be conserved on a microarray surface for all molecules. In addition, problems could occur with proteins that need to form multimers or complexes with other proteins in order to exhibit their binding activity.

2. Material and methods

2.1. Materials

2.1.1. Equipment

Name	Manufacturer
PCR-Maschine PTC200	M J Research BioRad, Waltham, USA
Photometer (UV/Vis)	Ultrospec 2000, Pharmacia Biotech, Freiburg Germany
pH-Meter	MP 230, Mettler Toledo, Germany
Mini-Protean®3 electrophoresis chamber and casting unit	BioRad, Waltham, USA
Photometer (UV/Vis)	Ultrospec 2000, Pharmacia Biotech, Freiburg, Germany
Hoefer TE 70 (Semi dry Wester-Blot)	Amersham Bioscience, Piscataway, USA
ND 1000, NanoDrop, Spectrophotometer	NanoDrop, Wilmington, DE USA
Oven	Haereus, Hanau, Germany
Centrifuge	5810R, Eppendorf, Hamburg Germany
Centrifuge	5415D, Eppendorf, Hamburg Germany
ESI spotter	Eurogebtec, Biorad, Munich, Germany
MicroGrid II spotter	Genomic Solution, Huntington, UK
SMP3 pins	TeleChem International, Sunnyvale, USA
ScanArray 4000XL	Perkin Elmer, Boston, MA, USA
ScanArray 5000	Perkin Elmer, Boston, MA, USA

2.1.2. Chemicals

Name	**Manufacturer**
Acrylamide/Bisacrylamide	Roth, Karlsruhe, Germany
Agar	Roth, Karlsruhe, Germany
Agarose	Invitrogen, Carlsbad, USA
Ammoniumpersulfate (APS)	Roth, Karlsruhe, Germany
Ampicillin (sodium salt)	Genaxxon, Munich, Germany
BSA (bovine serum albumin)	Sigma-Aldrich, Schnelldorf Germany
Chloramphenicol	Genaxxon, Munich, Germany

Dithiothreitol (DTT)	Invitrogen, Carlsbad, USA
Ethanol, absolute	Riedel de Haën Sigma, Seelze, Germany
Ethidium bromide	AppliChem, Darmstadt, Germany
(EDTA)	Sigma-Aldrich, St.Louis, USA
Glycerol	J.T.Baker, Deventer, Holland
Hepes	Roth, Karlsruhe, Germany
Imidazole	Roth, Karlsruhe Germany
Isopropyl-β-D-thiogalactopyranoside (IPTG)	Roth, Karlsruhe, Germany
Kanamycin	Genaxxon, München, Germany
Methanol	VWR, Darmstadt, Germany
Protein-Marker: Broad Range	NEB, Frankfurt Germany
Sodium dihydrogenphosphat (NaH2PO4 x H2O)	Roth, Karlsruhe, Germany
Sodium chloride	Riedel de Haën Sigma, Seelze, Germany
TEMED	Roth, Karlsruhe, Germany
Tetracyclin	Genaxxon, München, Germany
Triton X 100	Gerbu, Gaiberg, Germany
Tris-base	Sigma-Aldrich, St.Louis, USA
Tris-HCl	Sigma-Aldrich, St.Louis, USA
Tryptone/Peptone	Roth, Karlsruhe, Germany
Tween 20	Sigma-Aldrich, St.Louis, USA
Yeast extract	Gerbu, Gaiberg, Germany

2.1.3. Kits

Name	**Manufacturer**
Qiagen QIAprep Spin Miniprep Kit	Qiagen, Hilden, Germany
Qiagen QIAquick PCR Purification Kit	Qiagen, Hilden, Germany
Pierce HisPur cobalt spin columns	Thermo Fisher Scientific, Waltham, USA
Synthesis-Kit für Geniom 2x	Sigma, Seelze, Germany

2.1.4. Buffers and mediums

Buffer	**Composition**
Binding buffer 5X	0.1M Hepes pH 8.0, 0.25M KCl, 25mM DTT,

	0.25mM EDTA, 5mM $MgCl_2$, 25% v/v glycerol
Ethidium bromide solution for staining	0,5 µg/ml Endkonzentration
Laemmli buffer	30,1 g Tris Base, 144,2 g Glycine, 50 ml SDS (20%), ad 1 l dH2O
LB-Agar	LB-Medium + 1,5% (w/v) Agar
LB-Medium (1 liter)	10 g Tryptone/Pepton, 5 g yeast extract, 10 g NaCl, pH 7.2
PBS 10x (1 liter)	80 g NaCl, 2 g KCl, 26.8 g Na2HPO4, 2.4 g KH2PO4, pH7.4
Spotting buffer (buffer E)	20mM HEPES-KOH, 0.4mM EDTA, 1mM DTT, 100mM NaCl, 25% v/v glycerol, pH 8
TBE 10x (1 liter)	108 g Tris, 55 g Boric acid, 40 ml 0,5 M Na2EDTA (pH8)
TBS 10x (1 liter)	50 mM Tris, 150 mM NaCl mit HCl, pH 7.5
TBST	1X TBS/ 0,05% Tween20
Transfer buffer	150 mM Glycine, 25 mM Tris-base, 20% Ethanol

2.1.5. Labware

Name	**Manufacturer**
Falcon 15 ml and 50 ml	Becton Dickinson, Heidelberg
Polyfiltronics TM Uniplate, 384er, V-bottom	Nunc GmbH & Co.KG
Microtiter cover film	Nunc GmbH & Co.KG
Geniom - Biochip	Sigma, Seelze, Germany
Lazy-L-Spreaders	Sigma-Aldrich, St.Louis, USA
Disposable cuvettes	UVette, Eppendorf, Hamburg Germany
Latex gloves	Latex, Blossom Mexpo, Hayward, USA
Nitril gloves	Nitril, Microflex, Wien, Österreich
Parafilm	PM 996, Pechiney Plastic packaging
Sterile filter 500 ml	Nalgene, Rochester, USA
Cell culture Petri dishes 96x20mm	stock
Eppendorf reaction vessels	Safe Lock Tubes 1,5ml und 2ml, Eppendorf,

	Hamburg Germany

2.1.6. Enzymes, vectors, bacterial strains

Name	**Supplier**
LR clonase	Invitrogen, Karlsruhe, Germany
BssHII	NEB, Frankfurt, Germany
pDEST17	Invitrogen, Karlsruhe, Germany
BL21-AI One Shot	Invitrogen, Karlsruhe, Germany
Rosetta-Gami2 (DE3)	Novagen, Darmstadt, Germany

2.1.7. Software and web pages

Software	
GenePix Pro 6.0	Axon Instrument, Inc., Union City, USA
Search of rare codons in nucleotide sequence	http://molbiol.edu.ru/eng/scripts/01_11.html
PROTEIN CALCULATOR v3.3	http://www.scripps.edu/~cdputnam/protcalc.html
Recombinant Protein Solubility Prediction	http://www.biotech.ou.edu/
Protein molecular weight	http://www.bioinformatics.org/sms/prot_mw.html
TESS : Transcription Element Search System	http://www.cbil.upenn.edu/cgi-bin/tess/tess

2.2. Methodology

Setting up TFs-chip

The aim of this part was the establishment of protein microarray for studying DNA-protein interactions. Practically in the lab, getting specific signal resulted from DNA-protein interactions on the microarray surface was the goal in the lab. Therefore, we followed several steps to pursue our objective as the following:

2.2.1. DNA-binding protein expression and purification.

2.2.2. Protein spotting and immobilization on the proper surface chemistry keeping the protein functional.

2.2.3. On-chip DNA-protein interactions.

2.2.4. Verifying Transfac database consensus sequences using DNA-microarray.

2.2.5. Applying oligos, PCR products of promoter regions, and genomic DNA to TFs-chip.

2.2.1. DNA-binding protein expression and purification

2.2.1.1. Protein expression

In the present study, in total fifty proteins were expressed and purified. To express the proteins and to purify them, several troubleshooting were faced which is covered in details the discussion section. The Open Reading Frames (ORFs) of the DNA-binding proteins were obtained from Prof. Jussi Taipale, Helsinki, Finland. Therefore, the protein expression was started from shuttling the ORF into the gateway destination expression vector (pDEST-17) using the LR clonase (Invitrogen, Karlsruhe, Germany). The LR recombination reaction was performed overnight at 25°C in the Thermal cycler with heated lid. Then, DH5α library efficiency was transformed according the manufacturer's protocol. In the next day, the colonies were picked up in 2ml LB medium (with 100µg Ampecillin/ml) and incubated overnight at 37°C with shaking at 240rpm and the plasmid was isolated using Qiagen miniprep kit (Qiagen, Hilden, Germany). Afterward, Rosetta-Gami2 (DE3) (Novagen, Darmstadt, Germany) was transformed according the manufacturer's recommendations. In 10ml LB medium (with 100µg ampecillin/ml, 12.5µg/ml tetracycline, and 36µg/ml chloramphenicol), the colonies were picked up and incubated overnight at 37°C with shaking at 240rpm. In order to induce the protein expression, the bacteria were diluted 1:5 times with LB medium (100µg ampecillin/ml, 12.5µg/ml tetracycline, and 36µg/ml chloramphenicol) in a flask. The bacteria were incubated at 37°C with shaking at 240rpm till reaching OD 600 of 0.6. At OD600 of 0.6, protein expression was induced by adding 0.1mM IPTG and incubated overnight at 37°C with shaking at 240 rpm. Finally, the bacterial pellets were collected by centrifugation at 4000 rpm at 4°C and stored at –20°C for further investigations and purification.

2.2.1.2. Protein detection and purification

The expressed proteins were detected by Western blotting using anti-6His antibody-peroxidase (Roche Diagnostics, Mannheim, Germany). Protein samples were resolved on SDS-polyacrylamide gel and electrophoretically transferred using semidry blotting blot to nitrocellulose membranes (Amersham Pharmacia Biotech, Buckinghamshire, UK). After blocking, membranes were incubated with anti-6His antibody-peroxidase for one hour at room temperature. Afterward, bands were detected by chemiluminescence using the ECL Western Blotting Detection System (Amersham Pharmacia Biotech, Buckinghamshire, UK).

In the case of successful protein expression, the recombinant proteins were purified using Pierce HisPur cobalt spin columns (Thermo Fisher Scientific, Waltham, USA) using the denaturing protocol according to the manufacture instructions. To reach the optimum conditions, several protocols were used as detailed later in the discussion section.

2.2.2. Protein spotting and immobilization

After testing many conditions for the surface, spotting buffer, and immobilization temperatures, the proteins were spotted at a concentration (20-100ng) in Buffer E (20mM HEPES-KOH, 0.4mM EDTA, 1mM DTT, 100mM NaCl, 25% v/v glycerol, pH 8) on epoxy-coated glass slides (Schott-Nexterion, Jena, Germany) using a MicroGrid II (Genomic Solutions, Huntington, UK) equipped with SMP3 pins (TeleChem International, Sunnyvale, USA). After printing the slides, the proteins were immobilized at 37°C overnight.

2.2.3. On-chip DNA-protein interactions

Prior to hybridization, the slides were washed in 1X binding buffer for 30 min. The Hybridization was carried out by incubating the labeled DNA (oligos or PCR product from promoter sequences) using LifterSlips (Erie Scientific, Portsmouth, USA) for one hour at room temperature. Subsequently, the slides were washed in 1X binding buffer for 10 min followed by another 10 min in TBST and briefly with distilled water. Then, the slides were dried with N_2 and the signals were detected immediately.

The DNA-protein interaction signals were detected by confocal four-color laser scanners (ScanArray 4000XL and ScanArray 5000, Perkin Elmer, Boston, MA, USA). The quantification of the signal intensities was done with GenePix Pro 6.0 analysis software (Axon Instruments, Inc., Union City, USA).

Furthermore, for more confirmation, to avoid the doubt that the signals obtained by the microarray could be referred to fluorechrome-protein interaction and not DNA-protein interaction, the DNA was digested by DNase I prior to the microarray incubation.

2.2.4. Electrophoretic Mobility Shift Assay (EMSA)

The assessment of the DNA-protein interactions was done by comparing the results of microarray interaction signals with the results using EMSA. In EMSA, the biotinylated DNA was incubated with 50ng of the target proteins in 1X binding buffer for 30 min at room temperature. Then the bands were resolved on 10% non-denaturing polyacrylamide gel. The

bands were blotted on Nylon membrane and immobilized by UV-crosslinking. The membrane was incubated subsequently with streptavidin-POD for 30 min at room temperature. Finally, the bands were detected by chemiluminescence using the ECL Western Blotting Detection System (Amersham Pharmacia Biotech, Buckinghamshire, UK).

2.2.6. Verifying Transfac database consensus sequences using DNA-microarray

Several consensus sequences could be obtained from Transfac database with several qualities that based on the method of the DNA-binding protein verification. Therefore, we have applied a DNA array with sequences derived from the Transfac consensus sequence to investigate the protein preference and subsequently the sequences could be useful for TF-chip application.

The DNA-microarray was designed with Microsoft Excel on Linux program prepared Mutscan and the Geniom software, Febit. All possible permutations of one or two nucleotides were generated from the Transfac consensus sequence with the program Mutscan. The expansion on 125,000 spots per array enabled the investigation of all possible 9 mer binding sequence. The necessary consensus sequences and protein information were related of following source:
http://www.biobase.de/pages/index.php?id=transfacdatabases

The oligonucleotides synthesis was executed in the Geniom device (Febit, Heidelberg, Germany) through chemical synthesis using commercially available synthesis kits. Each spotted sequence was complementary. Therefore, they could form the double-stranded DNA through hairpin formation. After the DNA synthesis, all of the washing steps and protein reaction were carried out in an external unit. Firstly, the chip was blocked for an hour with 2% BSA in 1x PBS. Then, the TF-protein in 1X binding buffer with 2% BSA was incubated 1hour. Subsequently, the 8 arrays of the chips were washed with 1X binding buffer with 2% BSA. The further incubations were performed in the Geniom 2X system as the following:

- Protein detection by the anti-His antibody: 10 ml, 1X binding buffer + 10 µl biotinylated anti-His antibody.
- SAPE-solution (streptavidin-phycoerythrin): 9 ml 1x Binding buffer + 44 µl SAPE + 2% w/v BSA
- Wasch-solution: 80 ml 1x Binding buffer + 2% w/v BSA

The Phycoerythrin dye is detected by means of an integrated Cy3-Filtersets. Typical exposure times lay in the area of 1000-5000 ms. The data (in the form of signal intensity) were analyzed using Microsoft Excel and Sigmaplot from Systat software.

2.2.7. Applying oligos, PCR products of promoter regions, and genomic DNA to TFs-chip

Using accordant consensus sequences from the literature, DNA-chip results of ETS1 and SPI1, and galectin-4 promoter study, we pursued to validate and optimize the TF-chip.

Protein name	**DNA response sequence**
KLF8	caccctagagCCACACCCTggtaag
NFKB1 (1)	ctgtAGTTGAGGGGACTTTCCCAGGCactg
NFKB1 (2)	agccggtagggaagcccccaggaagcgcctg
NFKB1 (3)	ggcgcttcctgggggcttccctaccggctc
ETS1 (DNA array results)	AGGGGAAGGAGGGGAAGGAGGGGAAGGAGGGGAAGG
SPI1 (DNA array results)	GGAGAAAGTGGAGAAAGTGGAGAAAGTGGAGAAAGT

3. Results

3.1. Protein expression and purification

In the present study, 50 proteins were successfully expressed and purified with end concentrations ranging from 0.02 mg/ml to 0.2 mg/ml. The detection of the expression was carried out using Western blotting against the 6his-tag. The successfully expressed proteins were then purified and the quality was checked using SDS-PAGE followed by comassie blue staining and Agilent 2100 Bioanalyzer (figure 14 and 15).

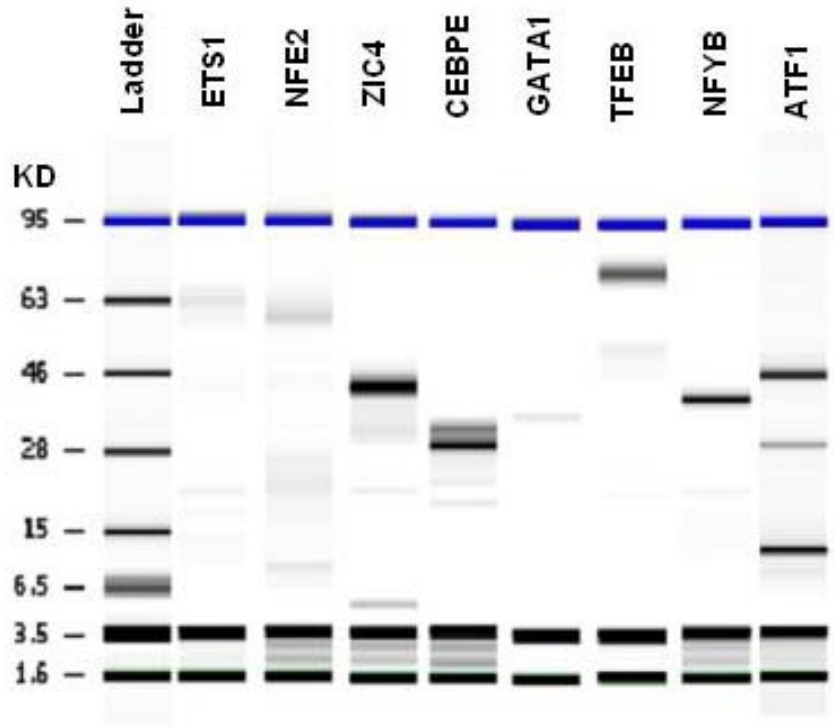


Figure 14. Nine different purified transcription factors proteins detected by Agilent 2100 Bioanalyzer.

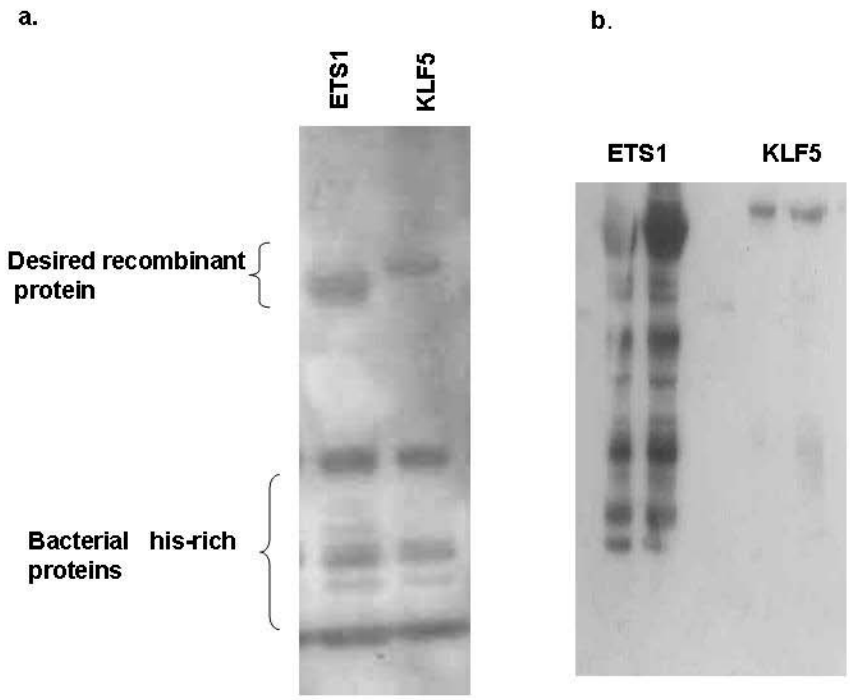


Figure 15. Western blotting of recombinant unpurified ETS1 and KLF5. In the left panel (a), the bacterial lysate was incubated with anti-his antibody, which shows the band of the target recombinant protein and also showing some his-rich proteins from the bacterial source. In the right panel (b), the bacterial pellets were incubated with the specific antibodies (anti-ETS1 and anti KLF5). As it is clear, the bacterial his-rich proteins are not detected, since the specific antibodies were used. In the case of ETS1, the main protein band is shown with extra bands that could be either degraded ETS1 protein or truncated protein.

At the beginning of this project, the protein expression experiment was started with two proteins (JUN and NFKB1) in BL21 bacteria strain. Figure 16 shows the MALDI-TOF analysis of the JUN protein which was obtained after protein expression and purification and as it is clear in the figure, truncated JUN protein (28.9 KD instead of 35.7 KD). In case of NFKB1, no protein was obtained.

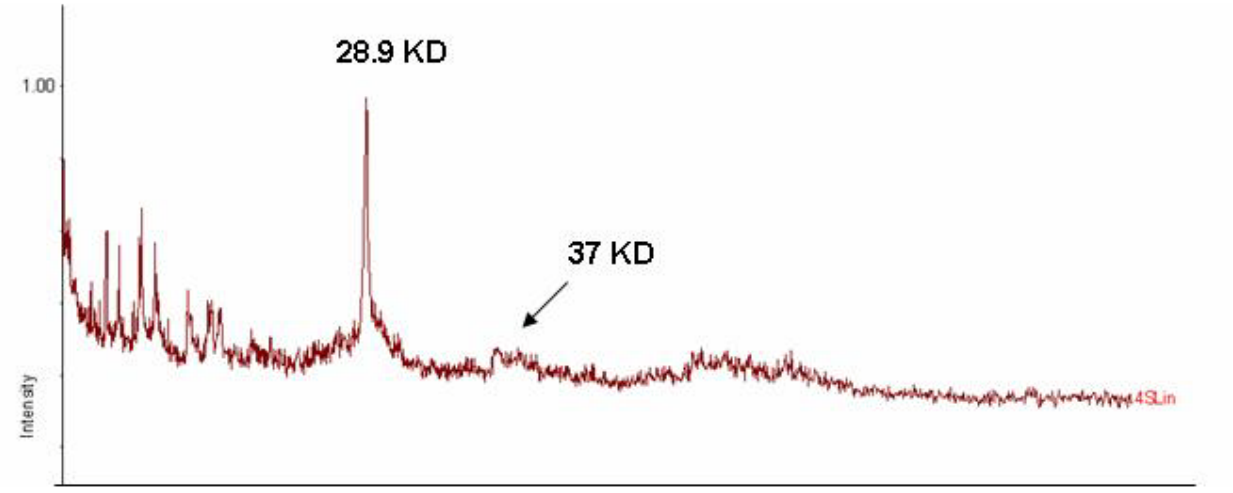


Figure 16. MALDI-TOF analysis of the recombinant JUN protein that was expressed in BL21 bacteria strain: The figure shows a high peak intensity of 28.9 KD that represents the truncated protein.

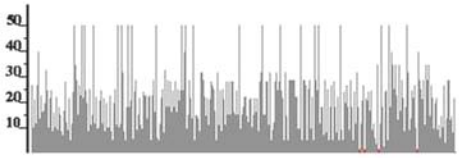
One of the factors that could influence protein expression is rare codons. Therefore, the in case NFKB1 and JUN, the presence of rare codons was investigated using a rare codons calculator (http://www.bioline.com/calculator/01_11.html). Subsequently, several rare codons were found in JUN and NFKB1. In the case of *JUN*, the rare codons were located the end of the ORF (figure 17a), which could explain the abundance of 28.9KD protein product. On the other hand, *NFKB1* ORF sequence was overloaded with rare codons (figure 17b) and subsequently NFKB1 was not expressed at all. Therefore, using Rosetta-gami 2 (DE3) that contain tRNA for rare codons, the full-length proteins were obtained.

a

> ### translated in organism: Escherichia coli

```
  1  mtakmettfy ddalnasflp sesgpygysn pkilkqsmtl nladpvgslk
 51  phlraknsdl ltspdvgllk laspelerli iqssnghitt tptptqflcp
101  knvtdeqegf aegfvralae lhsqntlpsv tsaaqpvnga gmvapavasv
151  aggsgsggfs aslhseppvy anlsnfnpga lssgggapsy gaaglafpaq
201  pqqqqqpphh lpqqmpvqhp rlqalkeepq tvpempgetp plspidmesq
251  erikaeRkrm Rnriaaskcr kRkleriarl eekvktlkaq nselastanm
301  lReqvaqlkq kvmnhvnsgc qlmltqqlqt f*
```

Bar chart of codon frequences



b

> ### translated in organism: Escherichia coli

```
  1  maeddpylgR peqmfhldps lthtifnpev fqpqmalpta dgpylqileq
 51  pkqrgfrfry vcegpshggl pgasseknkk sypqvkicny vgpakvivql
101  vtngknihlh ahslvgkhce dgictvtagp kdmvvgfanl gilhvtkkkv
151  fetlearmte aciRgynpgl lvhpdlaylq aegggdrqlg drekelirqa
201  alqqtkemdl svvrlmftaf lpdstgsftR rlepvvsdai ydskapnasn
251  lkivrmdRta gcvtggeeiy llcdkvqkdd iqirfyeeee nggvwegfgd
301  fsptdvhrqf aivfktpkyk dinitkpasv fvqlrRksdl etsepkpfly
351  ypeikdkeev qRkrqklmpn fsdsfgggsg agaggggmfg sgggggtgs
401  tgpgysfphy gfptyggitf hpgttksnag mkhgtmdtes kkdpegcdks
451  ddkntvnlfg kvietteqdq epseatvgng evtltyatgt keesagvqdn
501  lflekamqla kRhanalfdy avtgdvkmll avqrhltavq dengdsvlhl
551  aiihlhsqlv Rdllevtsgl isddiinmrn dlyqtplhla vitkqedvve
601  dllRagadls lldrlgnsvl hlaakeghdk vlsillkhkk aallldhpng
651  dglnaihlam msnslpclll lvaagadvna qeqksgrtal hlavehdnis
701  lagclllegd ahvdsttydg ttplhiaagr gstRlaallk aagadplven
751  feplydldds wenagedegv vpgttpldma tswqvfdiln gkpyepefts
801  ddllaqgdmk qlaedvklql yklleipdpd knwatlaqkl glgilnnafr
851  lspapsktlm dnyevsggtv relvealrqm gyteaieviq aasspvktts
901  qahslplspa stRqqidelr dsdsvcdsgv etsfrklsft esltsgasll
951  tlnkmphdyg qegplegki*
```

Bar chart of codon frequences

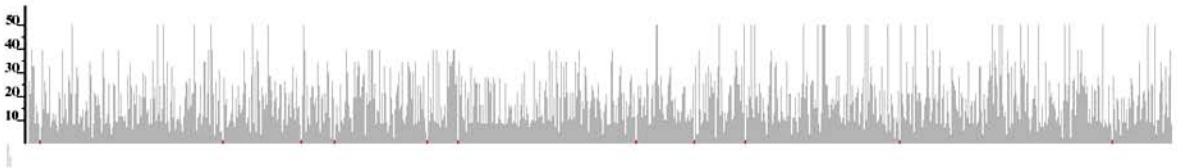


Figure 17. The rare codons distribution among *JUN* and *NFKB1* ORFs: Using the rare codons calculator (http://www.bioline.com/calculator/01_11.html), the rare codons were clustered at the end of *JUN* ORF sequence(a). While in the NFKB1 case (b), the rare codons were distributed along the whole sequence. R is referred to rare codon.

As in details later in the discussion section, using Ni-NTA Agarose for His-tagged protein purification, on column depending on the gravity sedimentation, bacterial proteins contamination was obtained. Therefore, we switched to cobalt spin column and hence the protein purity was increased.

3.2. Protein spotting and immobilization

Six TF-proteins were immobilized on several types of coated surfaces (Ni, 2D-Epoxy, 3D- Epoxy (glass and COC slides), and 3D- aldehyde (glass and COC slides)) at room temperature and 60°C. Subsequently, for the quality control, the immobilized proteins were stained using Sypro-Ruby. In order to increase the immobilized protein quantity and eliminate the high background obtained using the 3D coated slides as shown in figure 18, 2D-Epoxy coated glass surface was selected to pursue the TF-chip further optimization.

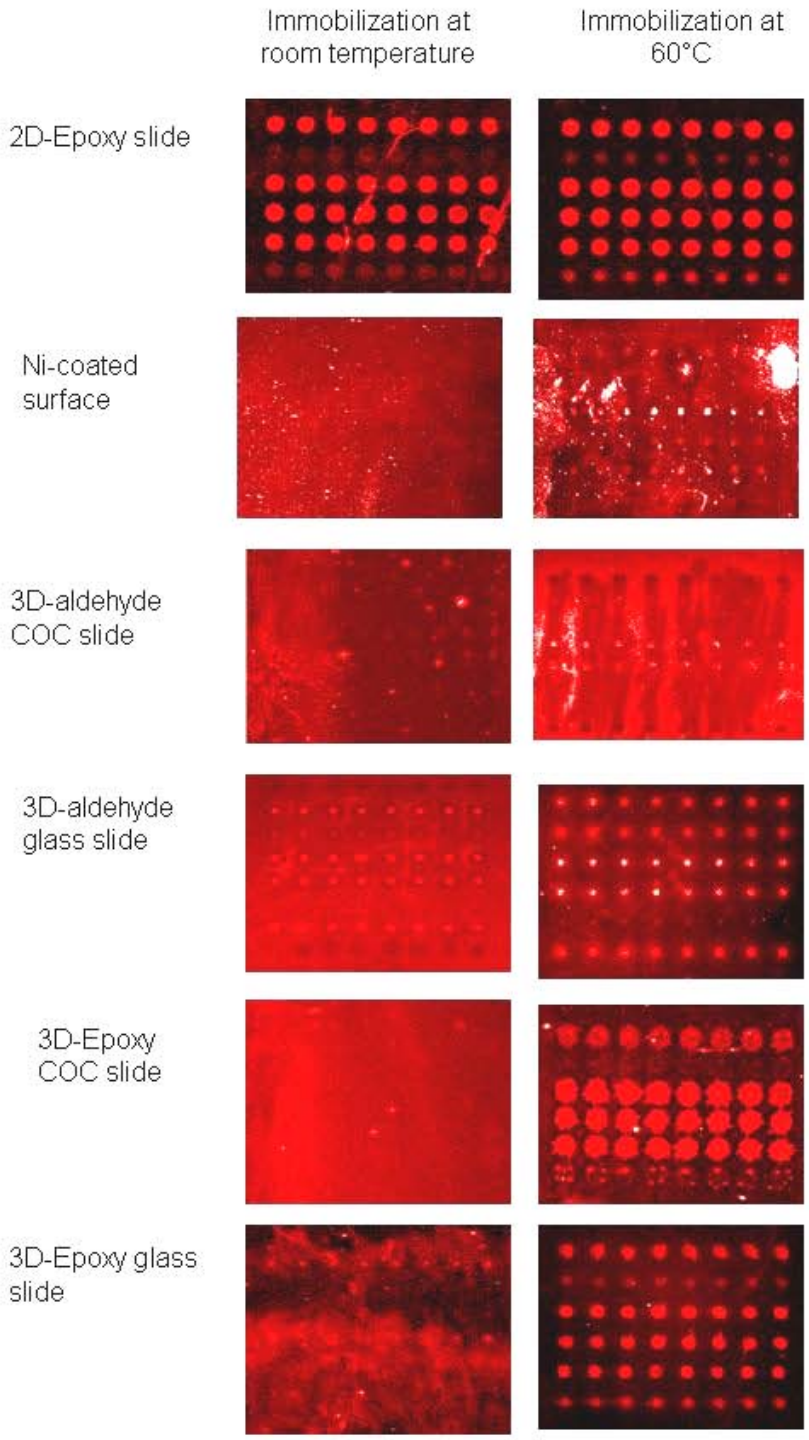


Figure 18. Optimization of the surface chemistry to immobilize TFs proteins. Six different proteins were immobilized on different types of coated slides and at two different temperature. Then, in order to detect the immobilization quality, the slides was stained with sypro ruby and scanned.

In parallel, several spotting buffer have been investigated (PBS, adding Tween 20, verifying the glycerol concentration). Figure 19 shows the results of using five different spotting buffers. A protein solubility problem was probably obtained with PBS and subsequently the proteins were washed away and therefore are not shown on the array. Adding detergent (tween 20) was enhanced the protein solubility but in parallel decreased the surface tension of the spots. Using another buffer caused donuts-shape spots. On the other hand, adding glycerol to the spotting buffer increased the surface tension of the spots and consequently adjusted the shape of the spot. However, 40% v/v glycerol caused a problem while spotting process due to sticking to the pins (detected by spotting water that showed signals from contamination).

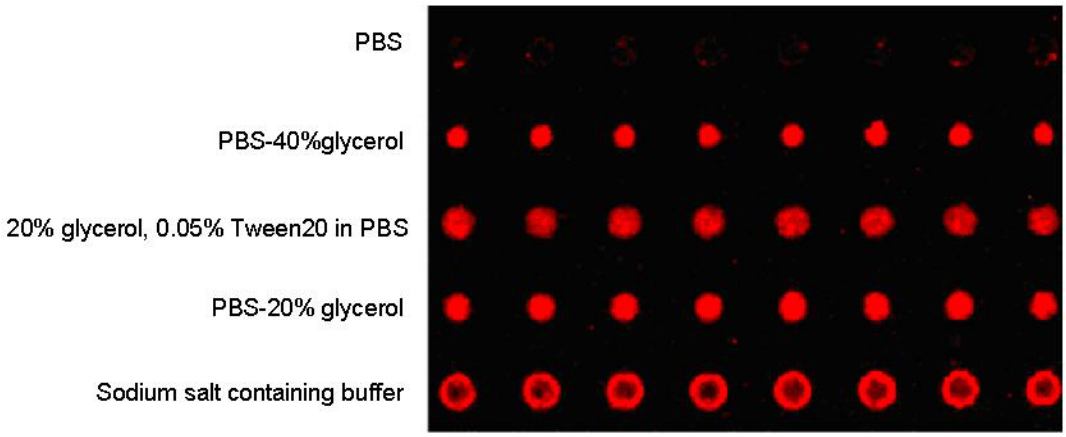


Figure 19. The results of testing different spotting buffer: 150ng NFKB1 protein dissolved in five different buffers and hybridized by Texas Red labeled NFKB1 consensus sequence. Using mere PBS caused evaporation of the spotting buffer and subsequently protein aggregation and therefore proteins were washed away afterward. Adding glycerol increased the surface tension and therefore enhanced the spot shape. Including Tween20 in the spotting buffer increased the protein solubility, but decreased the surface tension of the protein and then increased the spot size. Sodium salt rich buffer greatly decreased the surface tension and caused donuts-shape spots.

In the next step, when many proteins were applied and after storing the recombinant proteins more than two months, the spotting was not working properly which could be referred to protein aggregation. Several suggestions were considered like adding detergent or imidazole to the spotted proteins. In order to investigate whether adding detergent will increase the

protein solubility or not, two detergents were applied to the proteins (10% SDS and 4M urea). Comparable to the proteins without detergent, SDS and urea were increased the protein solubility in different grade (figure 20). While SDS was enhancing the solubility, it was also impairing the surface tension of the spots. As a result from this experiment, urea was selected for further investigation to increase the protein solubility, but at the same time to check the functionality of the proteins under such denaturing conditions. For that purpose, two different concentrations of NFKB1 protein was calibrated against three different concentrations of urea (figure 21). After comparing protein concentration, urea molarity, and the signals intensity, we decided to use 1M urea that increased the protein solubility and didn't interfere with the protein function.

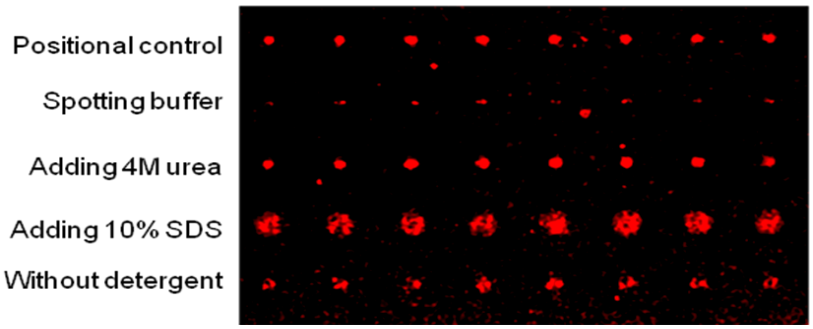


Figure 20. Adding denaturant to the spotted protein: 150ng NFKB1 protein dissolved using 20% glycerol/PBS with two different denaturing agents (10% SDS, 4M urea) and hybridized by Texas Red labelled NFKB1 consensus sequence. Moreover, the different solutions were spotted alone as negative controls.

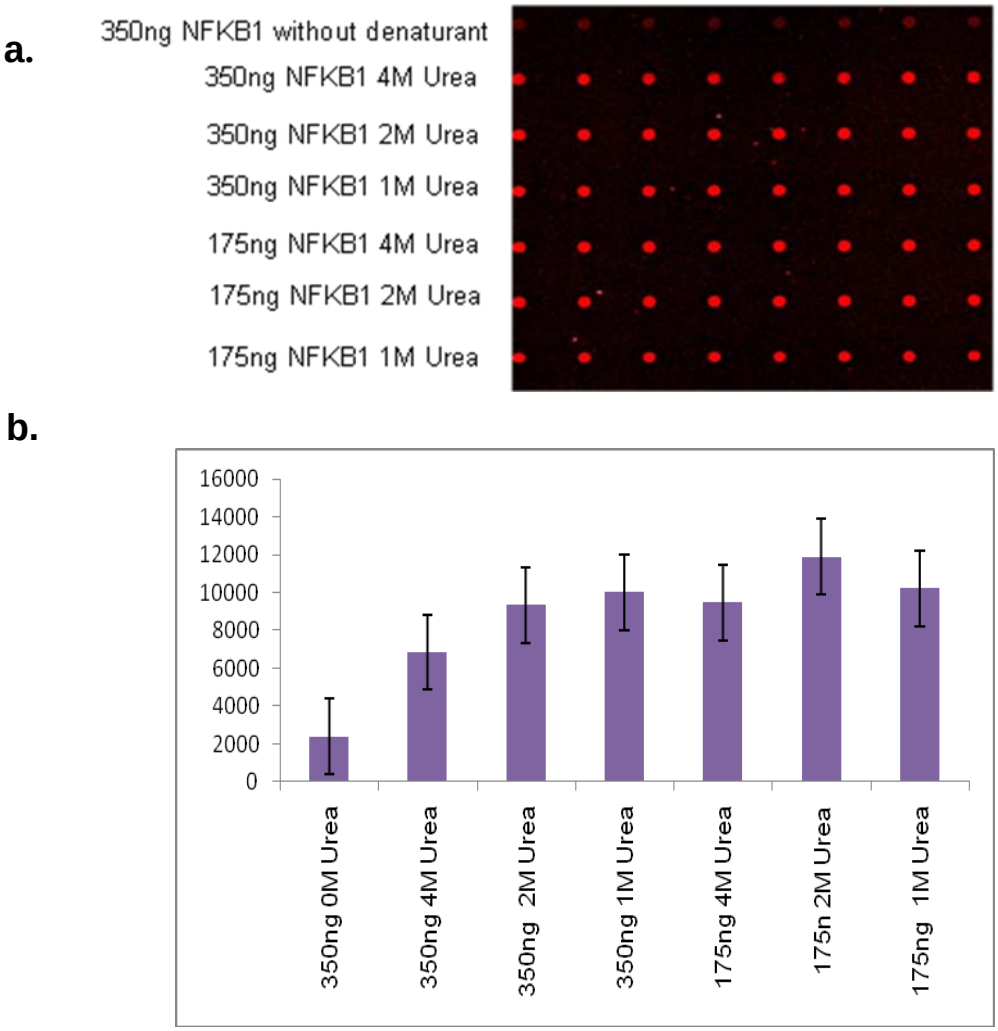


Figure 21. The calibration of urea concentration: a. NFKB1 protein was spotted in two different concentrations (350 ng and 175 ng) and three different urea concentrations and b. is the corresponding signal intensity.

Adding imidazole to his-tagged proteins was another solution to avoid their aggregation. Therefore, 120mM imidazole was added to the TF-proteins and spotted using the previous mentioned spotting and immobilization conditions. In addition to TF-proteins, spotting buffer and BSA were spotted as negative controls and hybridised by a labeled PCR product of IL-8 promoter (since IL-8 promoter has NFKB1 binding site). The results of this experiment showed increased solubility of the proteins, but on the other hand a strong signal was obtained from the spotting buffer (the negative control) as shown in figure 22.

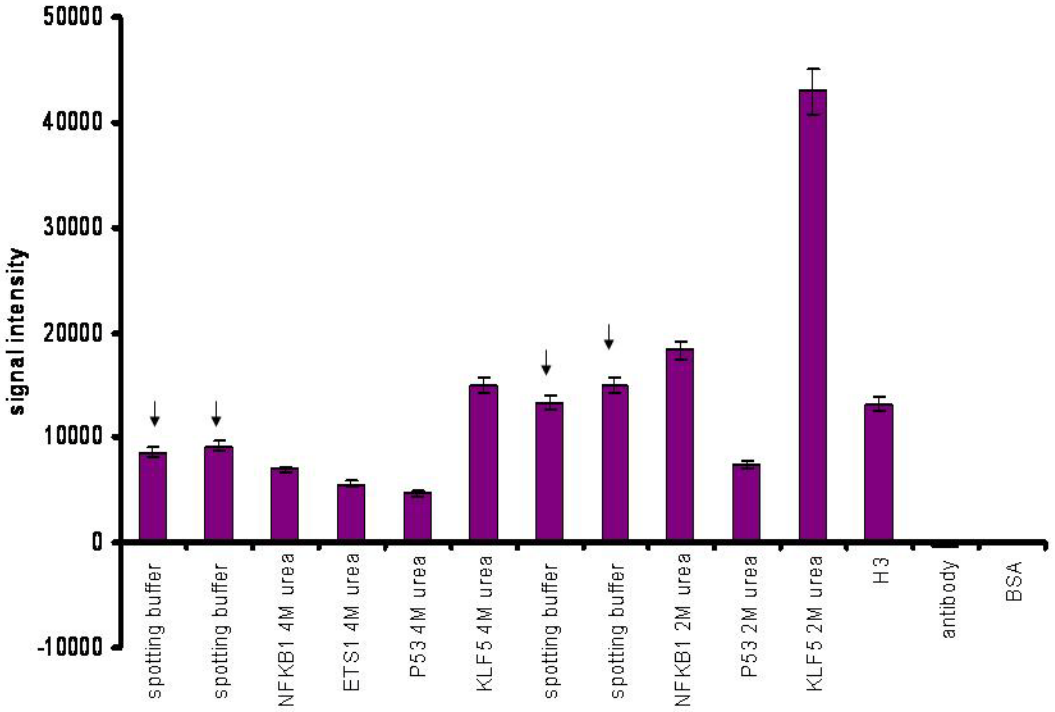


Figure 22. Adding 120mM Imidazole to the spotting buffer showed increasing in the recombinant protein solubility. However, on the other hand, in this histogram, as indicated with the arrowheads, the spotting buffer with imidazole gave strong signals while scanning. In this experiment, the TFs-chip was hybridized with a PCR product of IL-8 promoter.

Another point to control the solubility and the DNA-protein interactions on chip was the spotted protein concentration. In that regards, different NFKB1 protein concentrations (300 ng, 200 ng, 100 ng, and 50 ng) were spotted and hybridized by labeled NFKB1 binding-sequence. As shown in figure 23, the less the concentration (100 ng or 50 ng) the better the obtained signals. Therefore, 50-100 ng of the TF-protein was used in the spotting process.

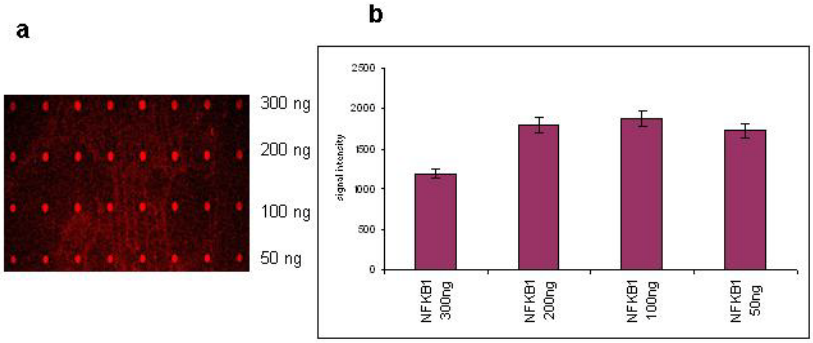


Figure 23. Optimizing the spotted protein concentration: NFKB1 protein was spotted in four different concentrations and hybridized with the labeled NFKB1 response element sequence (a) and the signal intensities were calculated and compared as shown in the histogram (b).

From the previous results, we decided to use the following conditions: 1) spotting on 2D-epoxy surface, 2) immobilizing at 60°C for one hour, 3) adding 1M urea to the spotting buffer, and 4) using 50-100 ng of the recombinant protein. However, using labeled genomic DNA as positive control, applying the 50 proteins to the chip, the conditions were not compatible with every protein for the DNA-protein interaction. Thus, playing with all of alternative (spotting buffer, adding detergent, and immobilization temperature), the conditions were changed again to suit all of the proteins to: 1) spotting on 2D-Epoxy surface, 2) immobilization at 37°C overnight, 3) dissolving the proteins in buffer E which contains DTT as a detergent, and 4) using 50-100 ng of the TF-proteins.

3.3. On-chip DNA-protein interaction

After getting DNA-protein interaction on the TF-chip, the determination of the reaction specificity was next step to execute. Therefore, two experiments were done: 1) investigating whether it is DNA-protein interaction or Fluorochrome-protein interaction, and 2) validation of the microarray results using EMSA.

3.3.1. DNA-protein interaction and not Fluorochrome-protein interaction

To investigate whether it is DNA-protein interaction or Fluorochrome-protein interaction, the PCR product of *IL-8* promoter was digested using DNaseI prior to the hybridization. Since it is known from the literature that *IL-8* promoter binds to NFKB1, the NFKB1 protein was spotted on two slides and subsequently two chips were hybridized by digested and undigested PCR product. Indicating that it is DNA-protein interaction and not Fluorochrome-protein interaction, the digested PCR product did not show any signal comparable to the strong signals obtained by the undigested PCR product (figure 24).

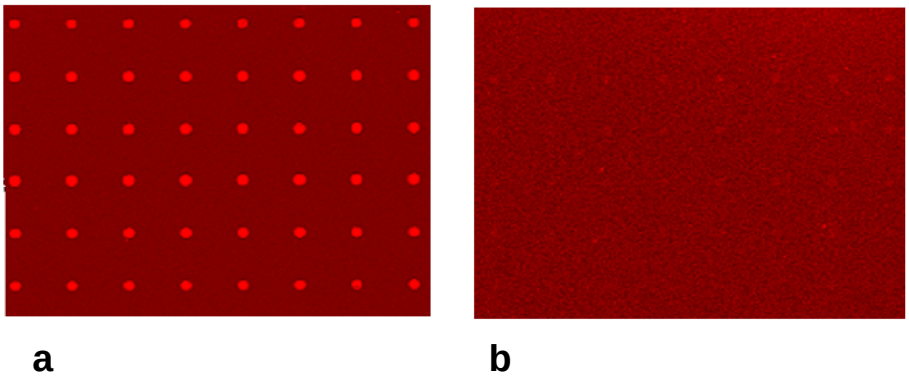


Figure 24. NFKB1 protein was spotted on two chips and hybridized using: **(a)** 206bp from *IL-8* promoter and **(b)** 5 min DNaseI digested PCR product of *IL-8*.

3.3.2. TFs-array validation by EMSA

In order to verify the TFs-chip results, in parallel to the protein array hybridization with the target DNA sequence, EMSA was done. As shown in figure 25, upon incubation of TFs-array with 30-mer double-stranded oligonucleotides from the *NFKB1* promoter sequence and the corresponding parallel EMSA experiment results, the two results (TFs-chip and EMSA) were concordant in terms of binding strength.

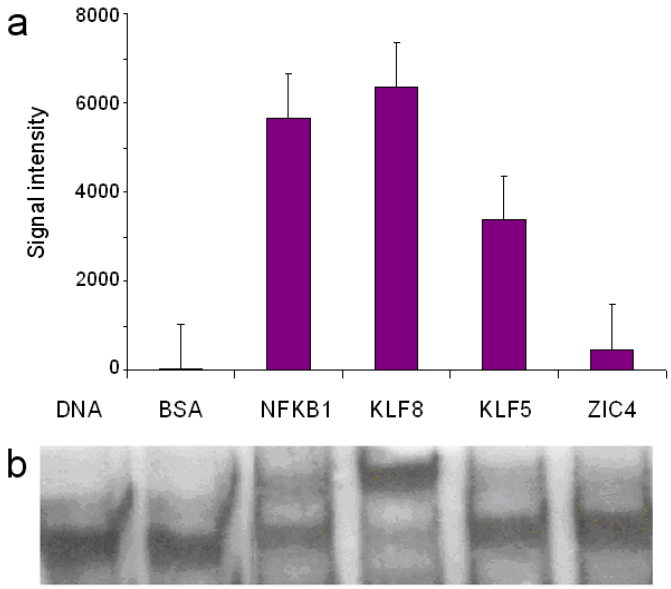


Figure 25. Comparison of results from a protein array measurement and a related EMSA experiment. Transcription factors were used to produce a protein microarray. Panel (a) shows signal intensities obtained upon incubation with a 30-mer double-stranded oligonucleotide, whose sequence had been selected from the promoter region of *NFKB*1. Results at four transcription factors and a BSA negative control are shown, clearly illustrating the fact that particular sequences may be shared between transcription factors, especially if members of a family. In panel (b), the corresponding results of an EMSA experiment are presented. In the lane labelled with DNA, no protein had been added. The extent of the band shifts corresponds with the array data.

3.4. Verifying Transfac database consensus sequences using DNA-microarray

One of the important points that was considered in this study is to select the DNA response element sequence, in order to investigate the specificity of TF-array. Therefore, we used Transfac database to retrieve the consensus sequences from the previous studies. Several consensus sequences were annotated with different quality ranging from 1 to 6. In this study, quality 2 was selected, which was in previous studies validated using recombinant proteins. In this experiment, a DNA-array was synthesized as mentioned before in the method section and SPI1 and ETS1 proteins were applied. After getting the results in the form of signal intensities, the highest signals were selected (table 8 and 9) and analysed to get the consensus sequences (figure 26).

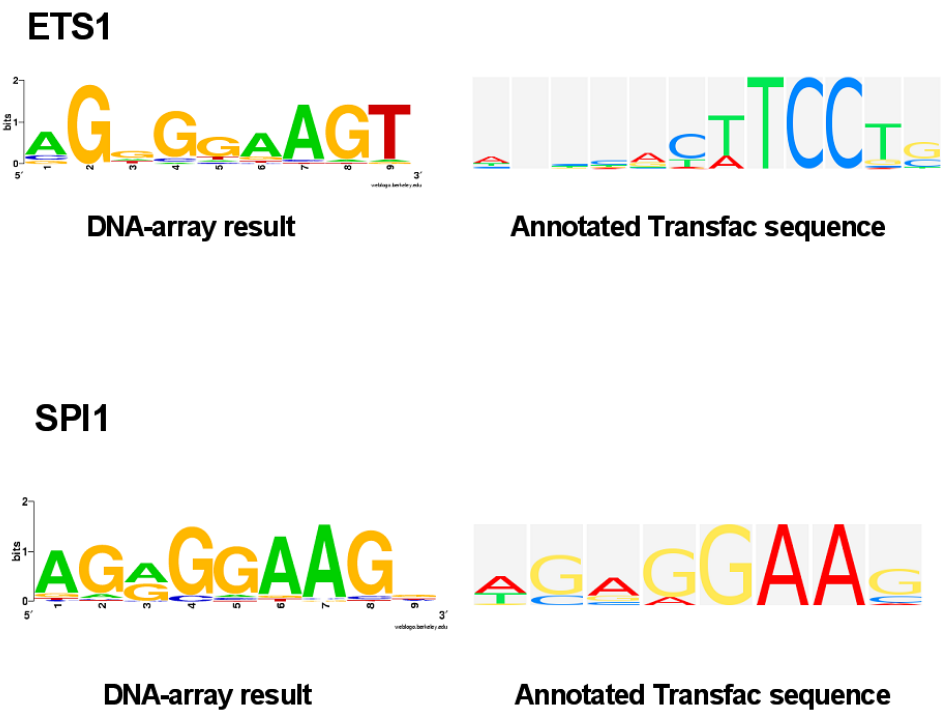


Figure 26. DNA-microarray results for ETS1 and SPI1 transcriptional proteins: the figure shows the consensus sequences of the two proteins that were obtained after the selection of the best signal intensities.

sequence	Signal Mean
GGAGGAAGT	15473
AGGGGAAGT	15499
AGAGGAAGC	15066
AGAGGGAGT	15140
CGAGGGAGT	15075
CGAGGAATT	16684
GGGGGAAGT	16400
GGAGAAAGT	16270
AGGAGAAGT	15118
AGGCGAAGT	15186
AGGGTAAGT	15610
AGGGAAAGT	15359
AGGGCAAGT	15448
AGGGGCAGT	16675
AGGGGTAGT	15757
AGGGGACGT	15865
AGGGGATGT	15490
AGGGGAAAT	15380
AGGGGAAGA	17425
AGGGGAAGG	18433
AGTCGAAGT	15227
AGTGTAAGT	15194
AGTGCAAGT	15104
AGTGGGAGT	15260
AGTGGTAGT	15366
CGCGGAAGT	15123
CGAGTAAGT	15174
GGGGGAAGT	17157
GGAGAAAGT	16730
GGAGGAAGG	15095
AGGGGAAGA	16181
AGGGGAAGG	17700
GGGGGAAGT	15390
GGAGAAAGT	15355
AGGGGAAGG	15297
AGACGACGT	15054
CGGGGAAGT	18309
AGGGGAAGG	16135
AGAGTTAGT	26615
AGAGTACGT	25800
ATAGGACGT	19228
ATAGGAGGT	17270
AGGGGAAGG	15275
AGGGGAAGA	17353
AGGGGAAGG	15162
AGGGGAAGA	17972
AGGGGAAGG	20186
AGGGGAAGG	15158

Table 8. ETS1 response element obtained from the verification of Transfac database consensus sequence. The retrieved Transfac sequence subjected to *in silico* permutation. Then, the permutated sequences were synthesized using Geniom software. Subsequently, the DNA-chip was hybridized by ETS1 protein and the DNA-protein interaction was detected by anti-his tag antibody. The highest 50 signals were then selected and their sequences were analyzed to get a common consensus sequence for ETS1

Sequence	Signal Mean
AGGGGAAGG	20630
GGAGAAAGT	20393
TGAGGGAGT	20516
TGAGGTAGT	20794
AGGGGAACT	20344
AGGGGAAGC	26756
AGGGGAAGG	20441
GGAGAAAGT	20866
AGGGGAAGG	21353
AGAGGAAGG	21544
CTAGGAAGT	28175
AGGGGAAGG	20857
AGACGAAGA	25092
AGACGAAGC	32457
AGACGAAGG	25013
AGGGGAAGG	22115
AGGGGAAGC	20153
AGGGGAAGG	22939
AGGGGAAGG	20420
AGAGGAGGG	20046
AGGGGAAGA	24787
GGAGTAAGT	20158
AGGGGAAGG	20838
AAAGGAAGA	21100
AAAGGAAGC	21840
AAAGGAAGG	22411
ACCGGAAGT	20279
AGGGGGAGT	27548
AGGGGTAGT	29007
AGGGGACGT	28166
AGGGGAAGG	20087
AGAGCAAGC	22913
AGAGCAAGG	22645
AGAGGAATG	20713
AGAGGAACC	22800
AGAGGTAGT	22463
AGAGGACGT	22739
CGAGTAAGT	28362
GGAGAAAGT	20138
AGGGGGAGT	20378
AGGGGAAGA	25316
AGGGGAAGG	27922
ATAGGAAGC	38265
AGACGAAGC	20660
AGAGTCAGT	20795
AGAGTATGT	21617
AGAGTAATT	21398
AGAGTAAAT	21434
AGAGAGAGT	22113,39062

Table 9. SPI1 response element obtained from the verification of Transfac database consensus sequence. The retrieved Transfac sequence subjected to *in silico* permutation. Then, the permutated sequences were synthesized using Geniom software. Subsequently, the DNA-chip was hybridized by SPI1 protein and the DNA-protein interaction was detected by anti-his tag antibody. The highest 50 signals were then selected and their sequences were analyzed to get a common consensus sequence for SPI1.

3.5. Applying oligos and PCR product of promoter region

3.5.1. Oligos from Transfac database and DNA-array results

Incubating the labeled sequences that were obtained from Transfac didn't work specifically with the target transcription factor protein. For example, the consensus binding sequence that was obtained for KLF8 protein was applied to TFs-chip. The results, as shown in figure 27, showed weak binding to KLF8. On the other hand, the sequence bound strongly to several other proteins like LEF1, ZNF140, and KLF5.

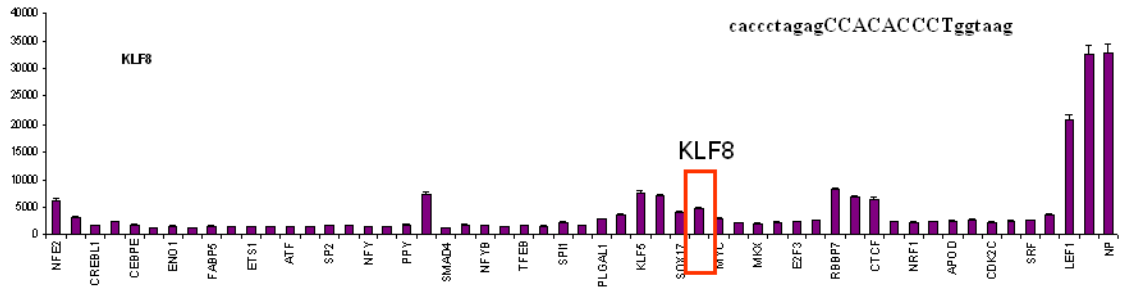


Figure 27. Applying Transfac consensus sequence of KLF8 protein on TF-chip: the KLF8 protein showed a binding signal with the obtained Transfac sequence as defined by the red rectangle. However, he sequence bound strongly to several other proteins like LEF1, ZNF140, and KLF5.

The next clue to verify the specificity of our TF-chip is utilizing the resulted consensus sequences of the DNA array for ETS1 and SPI1. Therefore, we used the obtained sequences as one unit of the binding sequence or as a four repeats of the binding motif as in the previous publication [100,149,150]. Using the single unit of the consensus sequence to hybridize the TF-chip, which is composed of 9mers, didn't show any signal. Using the four repeats of the binding motif sequence, the sequences bound to ETS1 and SPI1. However, the results showed even stronger binding to other proteins. Interestingly, since the obtained sequences from the results of DNA-microarray of ETS1 and SPI1 were different in one nucleotide, the resulted binding signals were more or less the same as shown in figure 28.

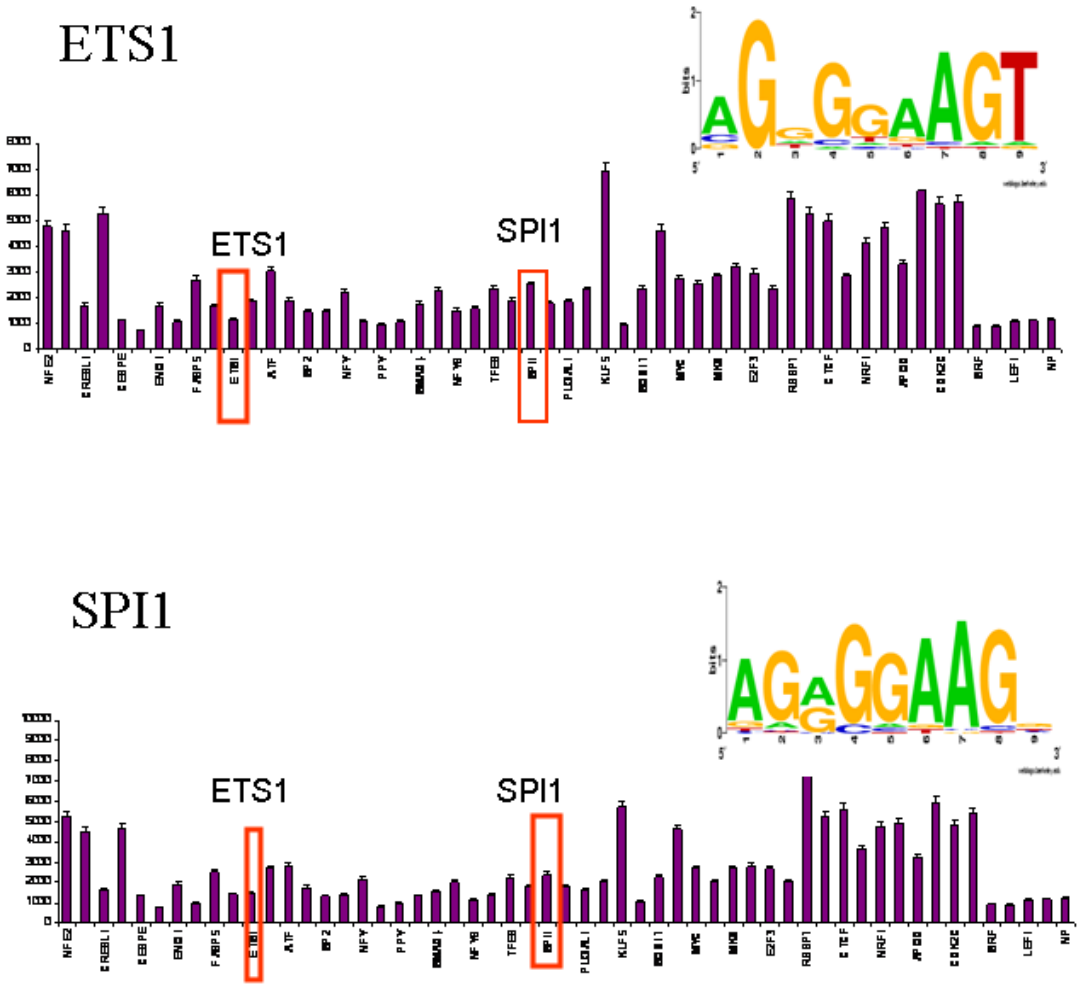


Figure 28. Upon the incubation of TF-chip with the consensus binding sequences of ETS1 and SPI1 proteins obtained from the DNA-array results in the form of four repeats, the sequences showed binding with ETS1 and SPI1 and some other proteins as well. The two sequences applied which are different in single nucleotide showed the same binding pattern with minute differences in the signal intensities.

3.5.2. PCR products from promoter sequences

From the first part of the thesis, utilizing the results of SNPs in *LGALS4* promoter and their effects on the protein binding, the labeled PCR products of the promoter region with two SNPs in different genotypes were co-incubated on TFs-chip. Interestingly, the result of TFs-chip was concordant with pull down – mass spectrometry results of galectin-4 promoter. In details, for example, P53 was detected as one hit in the mass spectrometry results in the

proteins that are binding to the wild type but not the mutant. In TFs-chip results as shown in figure 29, the P53 protein bound with signal intensity twice as the mutant type. The same result was encountered with RBBP7 that bound to wild but not the mutant type *LGALS4* promoter in both pull down-mass spectrometry and TFs-chip experiments results. ENO1, FUBP1, and FUS were defined to bind both genotype sequences and that was the same on chip results.

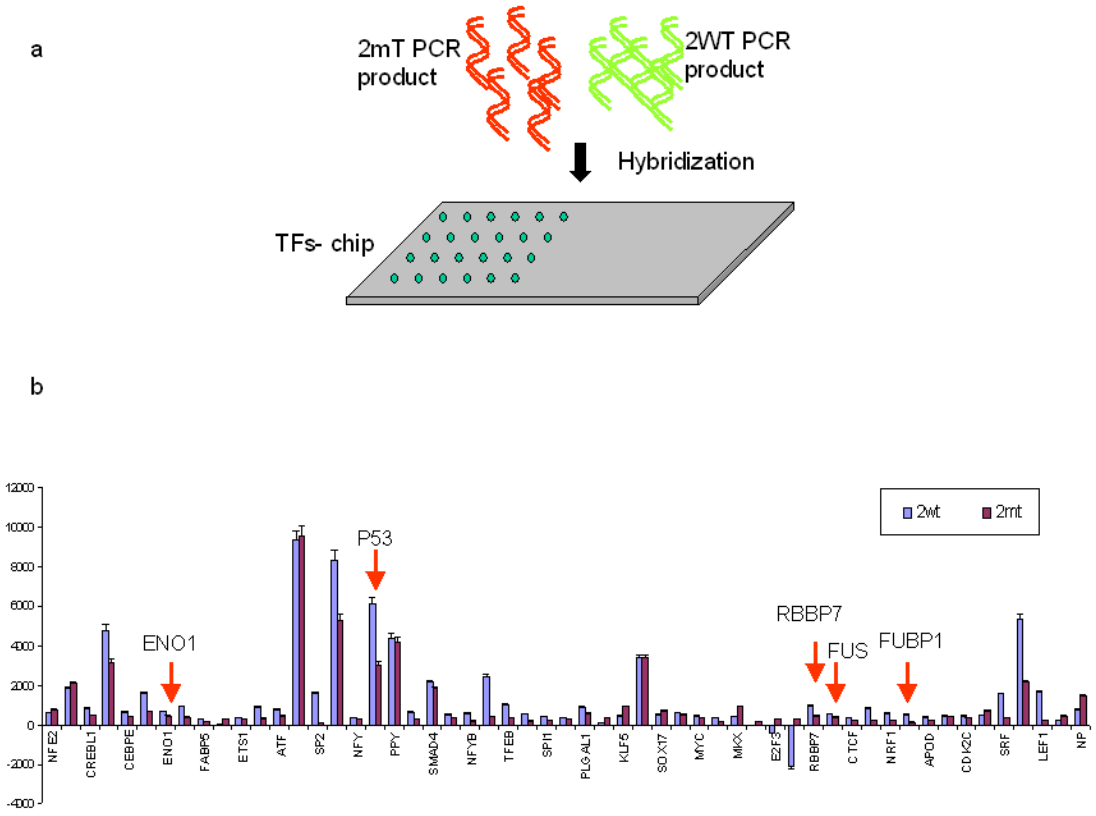


Figure 29. *LGALS4* promoter with different SNPs on TFs-chip: a) scheme represents the outline of the experiment. The PCR product of the *LGALS4* promoter was amplified with different SNPs. Each genotype was labeled with different fluorescent. Then, both PCR products were co-incubated on chip. B) The signal intensities results of both wild- and mutant- types. As indicated with arrowheads, both genotypes were shown to bind with ENO1, P53, RBBP7, FUS, and FUBP1 that was confirmed before with pull down – mass spectrometry experiments.

4. Discussion

4.1. Protein expression, purification, and storage troubleshooting

It was crucial for the present study to express pure, sufficient amount of, many, and functional proteins. Starting from the gateway cloning to protein purification, several troubleshooting were faced and subsequently the conditions were modified to get the target proteins in the desired form. In this section, the technical problems are discussed sequentially as faced during the experiment.

4.1.1. Truncated protein expression

At the beginning of protein expression experiment, we either did not get the proteins or truncated proteins were obtained. To overcome this problem, we prolonged the induction time in order to allow the full-length protein translation from four hours to overnight incubation. After time extension, still, there were some truncated proteins. Back to the fact that rare codons and rare codon clusters are capable of causing qualitative and quantitative expression problems in *E. coli* or other organisms [154,155]. These problems mainly occur on translation level rather than on transcription level. The main translational problems caused by rare codons or rare codon clusters include (a) that rare codons reduce the translation rate of the target gene, (b) low or undetectable amount of the expressed target protein, (c) misincorporation of amino acids into the target protein, (d) truncated or amino acids-deleted peptides or proteins are synthesized, and (e) frame-shifted peptides or proteins are synthesized [154,155]. Rosetta-gami 2 host strains combine the advantages of Rosetta 2 and Origami 2 strains to alleviate codon bias and enhance disulfide bond formation in the cytoplasm when heterologous proteins are expressed in *E. coli*. These *trxB*/*gor* mutants are compatible with kanamycin-resistant vectors, and carry the chloramphenicol-resistant pRARE2 plasmid, which supplies seven rare tRNAs. In Rosetta-gami 2(DE3) pLacI, the rare tRNA genes are present on the same plasmid that carries the *lac* repressor gene. DE3 indicates that the host is a lysogen of λDE3, and therefore carries a chromosomal copy of the T7 RNA polymerase gene under control of the *lacUV5* promoter. Such strains are suitable for production of protein from target genes cloned in pET vectors by induction with IPTG. Accordingly, we switched from BL21 competent cells to Rosetta-Gami2 (DE3). As result of changing the bacterial strain, we started to get the first protein (NFKB1) that was not expressed at all BL21 strain.

4.1.2. Inclusion body and protein solubility

In the bacterial cells, recombinant proteins usually fail to fold properly and accumulate as refractive, insoluble particles called inclusion bodies. In the previous studies, higher yields of soluble proteins have been pursued either by reducing the culture temperature, engineering the protein sequence, adding fusion partners or coproducing selected chaperones [156]. Under native conditions, we got a problem in purification. So, it was either no protein yield or less amount of protein comparable to the detected amount in the Western blotting. Therefore, we tried to minimize the induction temperature and time. Also, we changed the imidazole concentration in order to elute more amount of the His-tagged protein. None of those modifications enhanced the purified protein yield significantly. Under denaturating conditions the His tag is completely exposed, it will facilitate the binding to metal ions columns[157]. Thus, we switched to the denaturing conditions for dissolving the inclusion bodies, followed by protein refolding after the purification.

4.1.3. Bacterial proteins contamination

Technically, we faced bacterial proteins contamination. Regarding the 6His-tagged proteins, from the previous studies, it is recommended to optimise the imidazole concentration for the washing in order to minimize the contaminant proteins concentration or to increase the washing steps to get rid from the undesired bacterial proteins [157]. However, in our experiment, neither changing the imidazole concentration, nor the increasing of the washing steps enhances the situation. It was applicable for decreasing the bacterial proteins concentration. But, on the other side, that decreased the concentration of the recombinant protein as well in the same proportion. Nevertheless, changing the purification method was helpful to overcome this troubleshoot. So, changing from the Ni-column that depends on the gravity to spin column successfully reduced the amount of bacterial proteins. From these results, we have suggested that with spin column, all of the solutions surrounding the metal ion and the recombinant protein are disposed completely during the centrifugation between each washing step. Consequently, it reduces the chance of bacterial protein contamination from the incomplete disposed solution. On the other hand, using column that depends on the gravity to get rid from the solution of each step, it is technically hard to get rid from the complete solution on the metal beads. Interestingly, also, switching from BL21 strain to Rosetta-Gami2 (DE3) strain reduced the amount bacterial protein contamination. That could be referred to the His-rich protein amount in the different strain. Also, enhancing the amount of the 6 His-tagged will

increase the desired protein binding to the metal beads and subsequently will decrease the competition of bacterial proteins on the metal beads.

4.1.4. Storage

At the beginning of our experiment, after getting the recombinant proteins, we stored the proteins in PBS/20%glycerol at –20°C. After nearly two months, the proteins started to aggregate. From the literature, it was found that the histidine residues are implicated in triggering recombinant protein aggregation. At a physiological pH, histidine residues can form salt bridges with negatively charged amino acids and thereby mediating protein aggregation [158]. In a previous study, it was found that adding imidazole to the 6His-tagged recombinant proteins enhanced their solubility and stability [159]. Regarding the imidazole effect on 6His-tagged proteins solubility, our result was concordant with the previous observations. So, adding imidazole to the storing buffer prevent protein aggregation. Unfortunately, on the other hand, the buffer was not consistent with the microarray downstream application. In other words, with adding imidazole to the protein buffer, we got unspecific fluorescent signal from the negative control, which was the buffer alone. From this point, we decided to store the protein in the elution buffer/8M urea at 4°C. Then, the proteins were desalted in the proper working buffer.

4.2. Protein spotting and immobilization

In order to carry out DNA-protein interaction on-chip, several factors had to be optimised. Therefore, we went through several setting up steps from optimising the surface chemistry, spotting buffer, and protein immobilization, till optimising DNA-protein interaction on-chip.

4.2.1. Surface chemistry and protein immobilization

Selection of the proper surface chemistry is one of the crucial steps to optimize microarray in general. Since surface chemistry is the interface between the immobilized molecules and the solid surface, it could affect the molecules functionality, spots morphology, background, hybridization, and subsequently the on-chip interactions [160,161]. In case of protein array, it is much more laborious to optimize a surface chemistry, since the proteins are more complicated and different molecules than nucleic acid. Proteins are easily losing their structure and biochemical activity due to denaturation, dehydration, or oxidation [162]. In principle, protein immobilization could be done using two strategies; 1) randomly oriented

proteins by means of noncovalent or covalent attachment, and 2) uniformly oriented proteins [163].

In this study many surfaces were investigated to get the optimal conditions for spots morphology, proper protein immobilization on the surface, and less background noise. Thus, the immobilization was carried out using the mean of covalent random immobilization (aldehyde and epoxy surfaces in the forms of 2D and 3D chemistry) and also uniformly oriented proteins immobilization strategy was also investigated (Ni-coated slides). As shown in the results, using the 3D epoxy or aldehyde surfaces showed an increase of the immobilized protein. However, it caused a proportional increase of the background noise. On the other hand, using the Ni-coated slides, which is specific to immobilize his-tagged proteins, didn't show proper amount of the immobilized proteins. Practically, the tertiary structure of the his-tagged proteins in solution could hide the 6His. Therefore, the improper immobilization of his-tagged proteins on Ni-coated slides in our experiment results might refer to the inadequate exposure of the his-tag to the Ni-surface. In terms of background noise and immobilized protein amount, the 2D-epoxy surface was the optimum surface used in our experiment to immobilize TF proteins.

Another important factor that was considered in our experiment is the immobilization temperature. Two temperatures were investigated, which are 60°C for one hour and 37°C overnight. In terms of the amount of the immobilized proteins, the TF proteins showed better immobilization at 60°C. Interestingly, for the Ni-coated slides, since the immobilization at 60°C (denaturing condition) was obviously better, that could confirm our suggestion that the hidden his-tag is the cause of the improper immobilization and increasing the temperature denatures the protein and consequently expose the his-tag to the surface. Nevertheless, from the protein functionality side of view, using 60°C as immobilization temperature was working with TF proteins but not all when the TF-chip incubated with genomic DNA to check whether the proteins are functional or not. Instead, using the combined conditions of 2D-epoxy surface and 37°C overnight for the protein immobilization kept the TF proteins functional when it was incubated with labeled genomic DNA.

4.2.2. Spot morphology

Spot quality influences the subsequent evaluation of array data and therefore has a direct effect on the results reliability. Getting homogeneous uniform spots on the array depends

on the surface chemistry and the spotting solution [162,164]. In our experiment, we didn't experience a negative effect of the investigated surfaces on the spots morphology. However, the spotting solutions showed a great effect on the spots uniformity and homogeneity. For instance, in a preliminary experiment, five different conditions for spotting buffer were used (PBS, PBS/20% glycerol, 20% glycerol, 0.05% Tween 20 in PBS, PBS/40%, and sodium salt rich buffer). As a result, using PBS caused protein drought during the immobilization due to the evaporation of the spotting solution and consequently protein aggregation and meanwhile protein loss during the washing steps. Adding glycerol either in 20% or 40% v/v increased the surface tension of the spots, prevented the evaporation, and subsequently enhanced the spot morphology. On the other hand, the presence of high concentration of glycerol in the spotting buffer caused a technical problem during the spotting due to sticking of glycerol to the pins and subsequently a contamination was experienced. The addition of 0.05% Tween 20 to the spotting buffer decreased the surface tension of the spots and therefore increased the spots size. Concerning the protein functionality, several solutions were investigated (adding different denaturants, adding imidazole, and using the binding buffer) in combination with adding 20% glycerol. Adding denaturants was recommended in the previous studies [165]. Therefore, in order to solve protein aggregation problem, several denaturants were investigated. Among denaturants were urea, SDS, and Tween 20. Interestingly, urea as shown in the result section increased the quality of the spots morphology, protein solubility, and even some proteins showed better functionality in denaturing solution due to increasing the protein solubility. However, again denaturing conditions killed the function of several proteins. Imidazole in the spotting buffer resulted in unspecific fluorescent signal during the scanning, which was detected in the negative control spots (mere spotting buffer). Therefore, the spotting in the 1X binding buffer offered protein solubility (it does have DTT) and the buffer was consistent with DNA-protein interaction (the same as hybridization buffer).

4.2.3. Spotted protein concentration

In order to reach the optimum conditions of interaction array, the proper spotted protein concentration was another factor to adjust. In our results, it was shown that the lower the concentration the better the signals. Thus, in our experiments ≤ 100ng of the TF proteins was applied to the array surface. Moreover, in a previous study [164], the higher protein concentration was not recommended due to the smearing effects in proximity of the spots or across the whole slide (comet tail).

4.3. On-chip DNA-protein interaction

The last step to set up TFs-array was applying DNA-protein interaction on-chip and checking reaction specificity. To execute this objective, two steps were followed: 1) determination of the reaction specificity via investigating whether the resulted signals on chip are produced by DNA-protein interaction or Flourochrome-protein interaction and verifying the microarray results with EMSA, and 2) applying TF-response sequences obtained from Transfac database or the verified ETS1 and SPI1 Transfac sequence using DNA-array, and applying the *LGALS4* promoter study with the detected SNPs.

4.3.1. Interaction specificity validation

In the present study, the determination of the reaction specificity was carried out by two means: 1) the general concept of DNA-protein interaction on-chip, and 2) validation of the microarray results with EMSA.

The obtained signals from the on-chip DNA-protein interaction were the first concern. So, the question was whether the fluorescent signals are the result of the interaction of labeled DNA with the TFs protein or just an unspecific interaction between the cyanine3 and cyanine5 and TF-proteins. Thus, in order to exclude the factor that the resulted fluorescent signals are unspecific reaction between the fluorescent dye labels and the transcriptional factors proteins, the labeled DNA (PCR product of IL-8) was digested prior to the incubation with TFs-chip. In parallel, another TFs-chip was incubated with undigested PCR product of IL-8 promoter. As shown in the results, we didn't obtain any signal in the case of the digested DNA incubation on the TFs-array. Therefore, accordingly, the assumption of dye-proteins interaction was excluded.

EMSA [118] was used to validate the microarray results. In a comparison experiment between on-chip DNA-protein interaction results and EMSA results, as described earlier in the results, a labeled 30-mer double-stranded oligonucleotide from the promoter region of *NFKB1* was incubated with TFs-chip and in parallel separately with each TF-protein in EMSA experiment. The obtained results of the two parallel experiments were concordat that indicates the specificity of the TF-array results. Interestingly, the microarray results in this experiment was much more sensitive. In other words, the resolution of the shifted bands that indicate the DNA-protein interaction was too faint (it is difficult to detect). Therefore, microarray results were showing advantages over EMSA in being quantitative, sensitive, and specific.

4.3.2. Applying oligos and PCR products to TFs-chip

Another strategy to confirm the specificity of DNA-protein interaction on the array surface was applying previously studied sequences. Therefore, the sequences used in this study were obtained as the following: 1) retrieved sequences from Transfac database and labeled sequences were applied directly on chip, 2) verified Transfac sequences (which underwent DNA-microarray study to confirm the binding affinity) in two forms; single binding motif sequence and four repeats of the binding sequences according to the previous studies [100,149,150], and 3) *LGALS4* promoter with the emphasis on the effect of the two SNPs on the binding affinity, which is also done by us and explained in details in the first part of this thesis.

The obtained short oligos from the Transfac database bound successfully to the target proteins. However, on the other hand, the retrieved sequences showed also binding to some other proteins and even with stronger signals than the investigated target protein. That could indicate unspecific reaction. But, the microarray results were validated using EMSA that was completely concordant with the microarray results. Indeed, getting back to the derivation of consensus binding motifs for some transcription factors, they were concluded from *in vitro* studies and sequence comparisons of small sets of promoters that are known to bind the investigated factor [166]. Subsequently, the identification of the consensus binding sequence of the target transcription factor is executed using bioinformatic analyses that search the human genome using the consensus motifs or position weight matrices [167,168]. Thus, this approach provides all possible locations for a target transcription factor. However, there is many more occurrence of a consensus motif for a given factor than there are binding site in the mammalian genome [169,170]. Chen *et al*. found that some regions, called multiple transcription factor-binding loci (MTLs), were bound by several factors [171]. So, keeping in mind the differences between the results of the in vivo and in vitro studies, also the other facts from the previous studies (for example, multiple transcription factor loci and the occurrence of consensus sequence and the binding sites of the transcription factor in a mammalian genome), could strength the suggestion of the inaccuracy of the consensus sequence for a transcription factor. In another word, the consensus sequence for a given factor could contain overlapping binding sequence(s) for other factors and also the occurrence of this sequence in a promoter region doesn't mean that this promoter will be regulated with that transcription factor. Accordingly, we could refer the multiple binding of the Transfac retrieved sequences to several proteins in addition to the target protein to this suggestion.

In another experiment, the Transfac consensus sequence were validated using DNA microarray synthesized using the facility of Geniom and permutation bioinformatic tool to investigate the binding affinity of ETS1 and SPI1 under the influence of changing single or double nucleotide of their consensus sequence. The highest 50 signals obtained from DNA-protein interaction were then selected and further analyzed to get the consensus sequence. Then, the obtained sequences were applied to TFs protein array. In order to execute DNA-protein interaction on our protein chip, the obtained consensus sequences were in two forms: either single sequence of the motif sequence or four repeats of the binding sequence as used in several previous studies [100,149,150]. Using single motif sequence (8-9mers), no signal obtained from the DNA-protein interaction on-chip. Weakness and reversibility of the binding could interpret the absence of interaction signals, since the binding could be lost during the washing steps. In the opposite side, using four repeats of the binding motif sequence resulted in not only the binding of the target proteins, but also several other proteins. Actually, that could be because the repetition of the polynucleotide sequence could create new binding site(s) for other transcription factors. For example, if we have GATC as a consensus sequence. Then we will synthesize three repetetion of it GATCGATCGATC. Consequently, for example a sequence of CGAT or several other potential possiblities will be created and could be a motif for another TF protein. here, an example from *in silico* analysis of a target sequence is shown in figure 30.

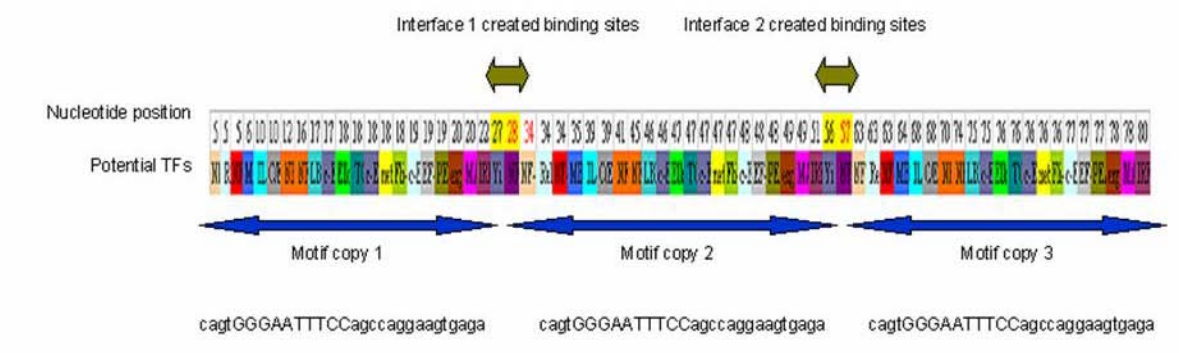


Figure 30. An example of the troubleshooting using a sequence which contains 3-4 repeats of the binding motifs. A NFKB1 binding sequence has been retieved from Transfac database [172]. Then, the three repeats of the sequence have been analyzed by TESS. As obviously shown from the figure, some transcription factor binding sites were created in the interface between the successive sequences.

Utilizing the promoter study of *LGALS4* and the two SNPs effect on its binding in combination with TFs-array gave more convincing results regarding the DNA-protein interaction specificity on-chip. Since we got concordance between the binding preferences in presence or absence of the two SNPs using pull down-mass spectrometry analyses and TFs-array results. The most obvious result was the case of P53, since in the pull down- mass spectrometry analyses showed one hit polypeptide sequence of the protein that bound to the wild type but not the mutant. Interestingly, in the microarray results, the interaction of the labeled PCR product of *LGALS4* promoter with P53 protein on chip showed more than double signal intensity with wild type over than mutant type. In addition, concordant results were also obtained with RBBP7, ENO1, FUS, and FUBP1 proteins.

Getting the results together, obtaining DNA-protein interaction on-chip was feasible. In terms of specificity, several results were obtained: 1) the detected interaction signals were the results of DNA-protein interaction and not fluorescent dye-protein interaction, 2) short oligonucleotides were confusing, since they bound to the target protein as well as several others, and 3) using PCR product of *LGALS4* promoter was the most solid results in terms of concordance with our *LGALS4* promoter study. In conclusion, it is more realistic to use TFs-array as a DNA-protein assay to study the binding affinity of genes promoters rather than discovering the binding motif sequences of transcription factors proteins.

5. References

1. Shih W, Chetty R, Tsao MS (2005) Expression profiling by microarrays in colorectal cancer (Review). Oncol Rep 13: 517-524.
2. Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, et al. (2005) Cancer statistics, 2005. CA Cancer J Clin 55: 10-30.
3. Arnold CN, Goel A, Blum HE, Boland CR (2005) Molecular pathogenesis of colorectal cancer: implications for molecular diagnosis. Cancer 104: 2035-2047.
4. Fearon ER, Vogelstein B (1990) A genetic model for colorectal tumorigenesis. Cell 61: 759-767.
5. Kinzler KW, Vogelstein B (1996) Lessons from hereditary colorectal cancer. Cell 87: 159-170.
6. Walther A, Johnstone E, Swanton C, Midgley R, Tomlinson I, et al. (2009) Genetic prognostic and predictive markers in colorectal cancer. Nat Rev Cancer 9: 489-499.
7. Gabius HJ (1997) Animal lectins. Eur J Biochem 243: 543-576.
8. Barondes SH, Castronovo V, Cooper DN, Cummings RD, Drickamer K, et al. (1994) Galectins: a family of animal beta-galactoside-binding lectins. Cell 76: 597-598.
9. Cooper DN (2002) Galectinomics: finding themes in complexity. Biochim Biophys Acta 1572: 209-231.
10. Leffler H, Carlsson S, Hedlund M, Qian Y, Poirier F (2004) Introduction to galectins. Glycoconj J 19: 433-440.
11. Huflejt ME, Leffler H (2004) Galectin-4 in normal tissues and cancer. Glycoconj J 20: 247-255.
12. Rechreche H, Mallo GV, Montalto G, Dagorn JC, Iovanna JL (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. Boyle P, Leon ME (2002) Epidemiology of colorectal cancer. Br Med Bull 64: 1-25.
14. Wolpin BM, Mayer RJ (2008) Systemic treatment of colorectal cancer. Gastroenterology 134: 1296-1310.
15. Espey DK, Wu XC, Swan J, Wiggins C, Jim MA, et al. (2007) Annual report to the nation on the status of cancer, 1975-2004, featuring cancer in American Indians and Alaska Natives. Cancer 110: 2119-2152.
16. Jass JR, Whitehall VL, Young J, Leggett BA (2002) Emerging concepts in colorectal neoplasia. Gastroenterology 123: 862-876.
17. Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, et al. (1988) Genetic alterations during colorectal-tumor development. N Engl J Med 319: 525-532.
18. Kerr D (2003) Clinical development of gene therapy for colorectal cancer. Nat Rev Cancer 3: 615-622.
19. Lengauer C, Kinzler KW, Vogelstein B (1998) Genetic instabilities in human cancers. Nature 396: 643-649.
20. Lothe RA, Peltomaki P, Meling GI, Aaltonen LA, Nystrom-Lahti M, et al. (1993) Genomic instability in colorectal cancer: relationship to clinicopathological variables and family history. Cancer Res 53: 5849-5852.
21. Vogelstein B, Fearon ER, Kern SE, Hamilton SR, Preisinger AC, et al. (1989) Allelotype of colorectal carcinomas. Science 244: 207-211.
22. Aaltonen LA, Peltomaki P, Leach FS, Sistonen P, Pylkkanen L, et al. (1993) Clues to the pathogenesis of familial colorectal cancer. Science 260: 812-816.

23. Ionov Y, Peinado MA, Malkhosyan S, Shibata D, Perucho M (1993) Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. Nature 363: 558-561.
24. Peltomaki P (2003) Role of DNA mismatch repair defects in the pathogenesis of human cancer. J Clin Oncol 21: 1174-1179.
25. Kane MF, Loda M, Gaida GM, Lipman J, Mishra R, et al. (1997) Methylation of the hMLH1 promoter correlates with lack of expression of hMLH1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines. Cancer Res 57: 808-811.
26. Cunningham JM, Christensen ER, Tester DJ, Kim CY, Roche PC, et al. (1998) Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite instability. Cancer Res 58: 3455-3460.
27. Weisenberger DJ, Siegmund KD, Campan M, Young J, Long TI, et al. (2006) CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nat Genet 38: 787-793.
28. Reedy J, Wirfalt E, Flood A, Mitrou PN, Krebs-Smith SM, et al. Comparing 3 dietary pattern methods--cluster analysis, factor analysis, and index analysis--With colorectal cancer risk: The NIH-AARP Diet and Health Study. Am J Epidemiol 171: 479-487.
29. von Roon AC, Reese G, Teare J, Constantinides V, Darzi AW, et al. (2007) The risk of cancer in patients with Crohn's disease. Dis Colon Rectum 50: 839-855.
30. Eaden JA, Abrams KR, Mayberry JF (2001) The risk of colorectal cancer in ulcerative colitis: a meta-analysis. Gut 48: 526-535.
31. Itzkowitz SH, Harpaz N (2004) Diagnosis and management of dysplasia in patients with inflammatory bowel diseases. Gastroenterology 126: 1634-1648.
32. Cunningham D, Atkin W, Lenz HJ, Lynch HT, Minsky B, et al. Colorectal cancer. Lancet 375: 1030-1047.
33. Ahlquist DA, Shuber AP (2002) Stool screening for colorectal cancer: evolution from occult blood to molecular markers. Clin Chim Acta 315: 157-168.
34. Hewitson P, Glasziou P, Irwig L, Towler B, Watson E (2007) Screening for colorectal cancer using the faecal occult blood test, Hemoccult. Cochrane Database Syst Rev: CD001216.
35. Johnson CD, Ahlquist DA (1999) Computed tomography colonography (virtual colonoscopy): a new method for colorectal screening. Gut 44: 301-305.
36. Sidransky D, Tokino T, Hamilton SR, Kinzler KW, Levin B, et al. (1992) Identification of ras oncogene mutations in the stool of patients with curable colorectal tumors. Science 256: 102-105.
37. Ahlquist DA (2002) Stool-based DNA tests for colorectal cancer: clinical potential and early results. Rev Gastroenterol Disord 2 Suppl 1: S20-26.
38. Corman ML (1980) Classic articles in colonic and rectal surgery. The classification of cancer of the rectum. Dis Colon Rectum 23: 605-611.
39. Compton CC (1999) Pathology report in colon cancer: what is prognostically important? Dig Dis 17: 67-79.
40. Steinberg SM, Barkin JS, Kaplan RS, Stablein DM (1986) Prognostic indicators of colon tumors. The Gastrointestinal Tumor Study Group experience. Cancer 57: 1866-1870.
41. Wanebo HJ, Rao B, Pinsky CM, Hoffman RG, Stearns M, et al. (1978) Preoperative carcinoembryonic antigen level as a prognostic indicator in colorectal cancer. N Engl J Med 299: 448-451.
42. Locker GY, Hamilton S, Harris J, Jessup JM, Kemeny N, et al. (2006) ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer. J Clin Oncol 24: 5313-5327.

43. Andreyev HJ, Norman AR, Cunningham D, Oates JR, Clarke PA (1998) Kirsten ras mutations in patients with colorectal cancer: the multicenter "RASCAL" study. J Natl Cancer Inst 90: 675-684.
44. Downward J (2003) Targeting RAS signalling pathways in cancer therapy. Nat Rev Cancer 3: 11-22.
45. Karapetis CS, Khambata-Ford S, Jonker DJ, O'Callaghan CJ, Tu D, et al. (2008) K-ras mutations and benefit from cetuximab in advanced colorectal cancer. N Engl J Med 359: 1757-1765.
46. Houzelstein D, Goncalves IR, Fadden AJ, Sidhu SS, Cooper DN, et al. (2004) Phylogenetic analysis of the vertebrate galectin family. Mol Biol Evol 21: 1177-1187.
47. Patterson RJ, Wang W, Wang JL (2004) Understanding the biochemical activities of galectin-1 and galectin-3 in the nucleus. Glycoconj J 19: 499-506.
48. Cooper DN, Barondes SH (1990) Evidence for export of a muscle lectin from cytosol to extracellular matrix and for a novel secretory mechanism. J Cell Biol 110: 1681-1691.
49. Hughes RC (1999) Secretion of the galectin family of mammalian carbohydrate-binding proteins. Biochim Biophys Acta 1473: 172-185.
50. Liu FT, Patterson RJ, Wang JL (2002) Intracellular functions of galectins. Biochim Biophys Acta 1572: 263-273.
51. Toscano MA, Ilarregui JM, Bianco GA, Campagna L, Croci DO, et al. (2007) Dissecting the pathophysiologic role of endogenous lectins: glycan-binding proteins with cytokine-like activity? Cytokine Growth Factor Rev 18: 57-71.
52. Ilarregui JM, Bianco GA, Toscano MA, Rabinovich GA (2005) The coming of age of galectins as immunomodulatory agents: impact of these carbohydrate binding proteins in T cell physiology and chronic inflammatory disorders. Ann Rheum Dis 64 Suppl 4: iv96-103.
53. Demetter P, Nagy N, Martin B, Mathieu A, Dumont P, et al. (2008) The galectin family and digestive disease. J Pathol 215: 1-12.
54. Liu FT, Rabinovich GA (2005) Galectins as modulators of tumour progression. Nat Rev Cancer 5: 29-41.
55. Sanjuan X, Fernandez PL, Castells A, Castronovo V, van den Brule F, et al. (1997) Differential expression of galectin 3 and galectin 1 in colorectal cancer progression. Gastroenterology 113: 1906-1915.
56. Hittelet A, Legendre H, Nagy N, Bronckart Y, Pector JC, et al. (2003) Upregulation of galectins-1 and -3 in human colon cancer and their role in regulating cell migration. Int J Cancer 103: 370-379.
57. Nagy N, Bronckart Y, Camby I, Legendre H, Lahm H, et al. (2002) Galectin-8 expression decreases in cancer compared with normal and dysplastic human colon tissue and acts significantly on human colon cancer cell migration as a suppressor. Gut 50: 392-401.
58. Gitt MA, Colnot C, Poirier F, Nani KJ, Barondes SH, et al. (1998) Galectin-4 and galectin-6 are two closely related lectins expressed in mouse gastrointestinal tract. J Biol Chem 273: 2954-2960.
59. Nagy N, Legendre H, Engels O, Andre S, Kaltner H, et al. (2003) Refined prognostic evaluation in colon carcinoma using immunohistochemical galectin fingerprinting. Cancer 97: 1849-1858.
60. Watanabe T, Komuro Y, Kiyomatsu T, Kanazawa T, Kazama Y, et al. (2006) Prediction of sensitivity of rectal cancer cells in response to preoperative radiotherapy by DNA microarray analysis of gene expression profiles. Cancer Res 66: 3370-3374.
61. Irimura T, Matsushita Y, Sutton RC, Carralero D, Ohannesian DW, et al. (1991) Increased content of an endogenous lactose-binding lectin in human colorectal carcinoma progressed to metastatic stages. Cancer Res 51: 387-393.

62. Legendre H, Decaestecker C, Nagy N, Hendlisz A, Schuring MP, et al. (2003) Prognostic values of galectin-3 and the macrophage migration inhibitory factor (MIF) in human colorectal cancers. Mod Pathol 16: 491-504.
63. Nakamura M, Inufusa H, Adachi T, Aga M, Kurimoto M, et al. (1999) Involvement of galectin-3 expression in colorectal cancer progression and metastasis. Int J Oncol 15: 143-148.
64. Schoeppner HL, Raz A, Ho SB, Bresalier RS (1995) Expression of an endogenous galactose-binding lectin correlates with neoplastic progression in the colon. Cancer 75: 2818-2826.
65. Greco C, Vona R, Cosimelli M, Matarrese P, Straface E, et al. (2004) Cell surface overexpression of galectin-3 and the presence of its ligand 90k in the blood plasma as determinants in colon neoplastic lesions. Glycobiology 14: 783-792.
66. Lotz MM, Andrews CW, Jr., Korzelius CA, Lee EC, Steele GD, Jr., et al. (1993) Decreased expression of Mac-2 (carbohydrate binding protein 35) and loss of its nuclear localization are associated with the neoplastic progression of colon carcinoma. Proc Natl Acad Sci U S A 90: 3466-3470.
67. Song S, Byrd JC, Mazurek N, Liu K, Koo JS, et al. (2005) Galectin-3 modulates MUC2 mucin expression in human colon cancer cells at the level of transcription via AP-1 activation. Gastroenterology 129: 1581-1591.
68. Blank M, Klussmann E, Kruger-Krasagakes S, Schmitt-Graff A, Stolte M, et al. (1994) Expression of MUC2-mucin in colorectal adenomas and carcinomas of different histological types. Int J Cancer 59: 301-306.
69. Bresalier RS, Niv Y, Byrd JC, Duh QY, Toribara NW, et al. (1991) Mucin production by human colonic carcinoma cells correlates with their metastatic potential in animal models of colon cancer metastasis. J Clin Invest 87: 1037-1045.
70. Iurisci I, Tinari N, Natoli C, Angelucci D, Cianchetti E, et al. (2000) Concentrations of galectin-3 in the sera of normal controls and cancer patients. Clin Cancer Res 6: 1389-1393.
71. Huflejt ME, Jordan ET, Gitt MA, Barondes SH, Leffler H (1997) Strikingly different localization of galectin-3 and galectin-4 in human colon adenocarcinoma T84 cells. Galectin-4 is localized at sites of cell adhesion. J Biol Chem 272: 14294-14303.
72. Hirabayashi J, Satoh M, Kasai K (1992) Evidence that Caenorhabditis elegans 32-kDa beta-galactoside-binding protein is homologous to vertebrate beta-galactoside-binding lectins. cDNA cloning and deduced amino acid sequence. J Biol Chem 267: 15485-15490.
73. Oda Y, Herrmann J, Gitt MA, Turck CW, Burlingame AL, et al. (1993) Soluble lactose-binding lectin from rat intestine with two different carbohydrate-binding domains in the same peptide chain. J Biol Chem 268: 5929-5939.
74. Leffler H, Masiarz FR, Barondes SH (1989) Soluble lactose-binding vertebrate lectins: a growing family. Biochemistry 28: 9222-9229.
75. Hokama A, Mizoguchi E, Mizoguchi A (2008) Roles of galectins in inflammatory bowel disease. World J Gastroenterol 14: 5133-5137.
76. Wooters MA, Hildreth MB, Nelson EA, Erickson AK (2005) Immunohistochemical characterization of the distribution of galectin-4 in porcine small intestine. J Histochem Cytochem 53: 197-205.
77. Vasta GR (2009) Roles of galectins in infection. Nat Rev Microbiol 7: 424-438.
78. Stowell SR, Arthur CM, Mehta P, Slanina KA, Blixt O, et al. (2008) Galectin-1, -2, and -3 exhibit differential recognition of sialylated glycans and blood group antigens. J Biol Chem 283: 10109-10123.
79. Stowell SR, Arthur CM, Dias-Baruffi M, Rodrigues LC, Gourdine JP, et al. Innate immune lectins kill bacteria expressing blood group antigen. Nat Med 16: 295-301.

80. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, et al. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29: 365-371.
81. Fellenberg K, Hauser NC, Brors B, Hoheisel JD, Vingron M (2002) Microarray data warehouse allowing for inclusion of experiment annotations in statistical analysis. Bioinformatics 18: 423-433.
82. Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD, et al. (2001) Correspondence analysis applied to microarray data. Proc Natl Acad Sci U S A 98: 10781-10786.
83. Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, et al. (2006) TM4 microarray software suite. Methods Enzymol 411: 134-193.
84. Jiang C, Xuan Z, Zhao F, Zhang MQ (2007) TRED: a transcriptional regulatory element database, new entries and other development. Nucleic Acids Res 35: D137-140.
85. Zhao F, Xuan Z, Liu L, Zhang MQ (2005) TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. Nucleic Acids Res 33: D103-107.
86. Li LC, Dahiya R (2002) MethPrimer: designing primers for methylation PCRs. Bioinformatics 18: 1427-1431.
87. Richter M, Jurek D, Wrba F, Kaserer K, Wurzer G, et al. (2002) Cells obtained from colorectal microadenomas mirror early premalignant growth patterns in vitro. Eur J Cancer 38: 1937-1945.
88. Wang W, Di X, D'Agostino RB, Jr., Torti SV, Torti FM (2007) Excess capacity of the iron regulatory protein system. J Biol Chem 282: 24650-24659.
89. Nicolas E, Morales V, Magnaghi-Jaulin L, Harel-Bellan A, Richard-Foy H, et al. (2000) RbAp48 belongs to the histone deacetylase complex that associates with the retinoblastoma protein. J Biol Chem 275: 9797-9804.
90. Qian YW, Lee EY (1995) Dual retinoblastoma-binding proteins with properties related to a negative regulator of ras in yeast. J Biol Chem 270: 25507-25513.
91. Creekmore AL, Walt KA, Schultz-Norton JR, Ziegler YS, McLeod IX, et al. (2008) The role of retinoblastoma-associated proteins 46 and 48 in estrogen receptor alpha mediated gene expression. Mol Cell Endocrinol 291: 79-86.
92. Liu J, Kouzine F, Nie Z, Chung HJ, Elisha-Feil Z, et al. (2006) The FUSE/FBP/FIR/TFIIH system is a molecular machine programming a pulse of c-myc expression. EMBO J 25: 2119-2130.
93. Anzick SL, Kononen J, Walker RL, Azorsa DO, Tanner MM, et al. (1997) AIB1, a steroid receptor coactivator amplified in breast and ovarian cancer. Science 277: 965-968.
94. Owen HR, Elser M, Cheung E, Gersbach M, Kraus WL, et al. (2007) MYBBP1a is a novel repressor of NF-kappaB. J Mol Biol 366: 725-736.
95. Ghosh AK, Steele R, Ray RB (1999) Functional domains of c-myc promoter binding protein 1 involved in transcriptional repression and cell growth regulation. Mol Cell Biol 19: 2880-2886.
96. Ray R, Miller DM (1991) Cloning and characterization of a human c-myc promoter-binding protein. Mol Cell Biol 11: 2154-2161.
97. Guan LS, Rauchman M, Wang ZY (1998) Induction of Rb-associated protein (RbAp46) by Wilms' tumor suppressor WT1 mediates growth inhibition. J Biol Chem 273: 27047-27050.
98. Yamamoto F, Clausen H, White T, Marken J, Hakomori S (1990) Molecular genetic basis of the histo-blood group ABO system. Nature 345: 229-233.
99. Stowell SR, Arthur CM, Slanina KA, Horton JR, Smith DF, et al. (2008) Dimeric Galectin-8 induces phosphatidylserine exposure in leukocytes through polylactosamine recognition by the C-terminal domain. J Biol Chem 283: 20547-20559.

100. Hu S, Xie Z, Onishi A, Yu X, Jiang L, et al. (2009) Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling. Cell 139: 610-622.
101. Lemon B, Tjian R (2000) Orchestrated response: a symphony of transcription factors for gene control. Genes Dev 14: 2551-2569.
102. Lee TI, Young RA (2000) Transcription of eukaryotic protein-coding genes. Annu Rev Genet 34: 77-137.
103. Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, et al. (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. Nat Rev Genet 8: 424-436.
104. Farnham PJ (2009) Insights from genomic profiling of transcription factors. Nat Rev Genet 10: 605-616.
105. Blackwood EM, Kadonaga JT (1998) Going the distance: a current view of enhancer action. Science 281: 60-63.
106. Riggs AD, Bourgeois S, Newby RF, Cohn M (1968) DNA binding of the lac repressor. J Mol Biol 34: 365-368.
107. Riggs AD, Newby RF, Bourgeois S (1970) lac repressor--operator interaction. II. Effect of galactosides and other ligands. J Mol Biol 51: 303-314.
108. Towbin H, Staehelin T, Gordon J (1979) Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. Proc Natl Acad Sci U S A 76: 4350-4354.
109. Woodbury CP, Jr., von Hippel PH (1983) On the determination of deoxyribonucleic acid-protein interaction parameters using the nitrocellulose filter-binding assay. Biochemistry 22: 4730-4737.
110. Beattie KL, Wiegand RC, Radding CM (1977) Uptake of homologous single-stranded fragments by superhelical DNA. II. Characterization of the reaction. J Mol Biol 116: 783-803.
111. Tullius TD (1989) Physical studies of protein-DNA complexes by footprinting. Annu Rev Biophys Biophys Chem 18: 213-237.
112. Galas DJ, Schmitz A (1978) DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res 5: 3157-3170.
113. Brenowitz M, Senear DF, Shea MA, Ackers GK (1986) Quantitative DNase footprint titration: a method for studying protein-DNA interactions. Methods Enzymol 130: 132-181.
114. Maxam AM, Gilbert W (1980) Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol 65: 499-560.
115. Fried M, Crothers DM (1981) Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. Nucleic Acids Res 9: 6505-6525.
116. Garner MM, Revzin A (1981) A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. Nucleic Acids Res 9: 3047-3060.
117. Carey J (1991) Gel retardation. Methods Enzymol 208: 103-117.
118. Hellman LM, Fried MG (2007) Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. Nat Protoc 2: 1849-1861.
119. Fried MG, Bromberg JL (1997) Factors that affect the stability of protein-DNA complexes during gel electrophoresis. Electrophoresis 18: 6-11.
120. Siebenlist U, Gilbert W (1980) Contacts between Escherichia coli RNA polymerase and an early promoter of phage T7. Proc Natl Acad Sci U S A 77: 122-126.
121. Yang VW (1998) Eukaryotic transcription factors: identification, characterization and functions. J Nutr 128: 2045-2051.
122. Massie CE, Mills IG (2008) ChIPping away at gene regulation. EMBO Rep 9: 337-343.

123. Hoffman BG, Jones SJ (2009) Genome-wide identification of DNA-protein interactions using chromatin immunoprecipitation coupled with flow cell sequencing. J Endocrinol 201: 1-13.
124. Wu J, Smith LT, Plass C, Huang TH (2006) ChIP-chip comes of age for genome-wide functional analysis. Cancer Res 66: 6899-6902.
125. Mardis ER (2007) ChIP-seq: welcome to the new frontier. Nat Methods 4: 613-614.
126. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99-104.
127. Vogel MJ, Peric-Hupkes D, van Steensel B (2007) Detection of in vivo protein-DNA interactions using DamID in mammalian cells. Nat Protoc 2: 1467-1478.
128. Lee TI, Johnstone SE, Young RA (2006) Chromatin immunoprecipitation and microarray-based analysis of protein location. Nat Protoc 1: 729-748.
129. Johne B, Gadnell M, Hansen K (1993) Epitope mapping and binding kinetics of monoclonal antibodies studied by real time biospecific interaction analysis using surface plasmon resonance. J Immunol Methods 160: 191-198.
130. Buckle M, Williams RM, Negroni M, Buc H (1996) Real time measurements of elongation by a reverse transcriptase using surface plasmon resonance. Proc Natl Acad Sci U S A 93: 889-894.
131. Rich RL, Myszka DG (2006) Survey of the year 2005 commercial optical biosensor literature. J Mol Recognit 19: 478-534.
132. Oliphant AR, Brandl CJ, Struhl K (1989) Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. Mol Cell Biol 9: 2944-2949.
133. Roulet E, Busso S, Camargo AA, Simpson AJ, Mermod N, et al. (2002) High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. Nat Biotechnol 20: 831-835.
134. Deplancke B, Dupuy D, Vidal M, Walhout AJ (2004) A gateway-compatible yeast one-hybrid system. Genome Res 14: 2093-2101.
135. Deplancke B, Mukhopadhyay A, Ao W, Elewa AM, Grove CA, et al. (2006) A gene-centered C. elegans protein-DNA interaction network. Cell 125: 1193-1205.
136. Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, et al. (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell 133: 1277-1289.
137. Gustafsdottir SM, Schlingemann J, Rada-Iglesias A, Schallmeiner E, Kamali-Moghaddam M, et al. (2007) In vitro analysis of DNA-protein interactions by proximity ligation. Proc Natl Acad Sci U S A 104: 3067-3072.
138. Conze T, Shetye A, Tanaka Y, Gu J, Larsson C, et al. (2009) Analysis of genes, transcripts, and proteins via DNA ligation. Annu Rev Anal Chem (Palo Alto Calif) 2: 215-239.
139. Bulyk ML (2006) DNA microarray technologies for measuring protein-DNA interactions. Curr Opin Biotechnol 17: 422-430.
140. Bulyk ML, Gentalen E, Lockhart DJ, Church GM (1999) Quantifying DNA-protein interactions by double-stranded DNA arrays. Nat Biotechnol 17: 573-577.
141. Lieb JD, Liu X, Botstein D, Brown PO (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat Genet 28: 327-334.
142. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, et al. (2000) Genome-wide location and function of DNA binding proteins. Science 290: 2306-2309.
143. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, et al. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298: 799-804.

144. Berger MF, Bulyk ML (2006) Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. Methods Mol Biol 338: 245-260.
145. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, et al. (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat Genet 36: 1331-1339.
146. Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, et al. (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell 133: 1266-1276.
147. Warren CL, Kratochvil NC, Hauschild KE, Foister S, Brezinski ML, et al. (2006) Defining the sequence-recognition profile of DNA-binding molecules. Proc Natl Acad Sci U S A 103: 867-872.
148. Berger MF, Bulyk ML (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc 4: 393-411.
149. Ho SW, Jona G, Chen CT, Johnston M, Snyder M (2006) Linking DNA-binding proteins to their recognition sequences by using protein microarrays. Proc Natl Acad Sci U S A 103: 9940-9945.
150. Gong W, He K, Covington M, Dinesh-Kumar SP, Snyder M, et al. (2008) The development of protein microarrays and their applications in DNA-protein and protein-protein interaction analyses of Arabidopsis transcription factors. Mol Plant 1: 27-41.
151. He M, Taussig MJ (2001) Single step generation of protein arrays from DNA by cell-free expression and in situ immobilisation (PISA method). Nucleic Acids Res 29: E73-73.
152. Ramachandran N, Hainsworth E, Bhullar B, Eisenstein S, Rosen B, et al. (2004) Self-assembling protein microarrays. Science 305: 86-90.
153. Angenendt P, Kreutzberger J, Glokler J, Hoheisel JD (2006) Generation of high density protein microarrays by cell-free in situ expression of unpurified PCR products. Mol Cell Proteomics 5: 1658-1666.
154. Kane JF (1995) Effects of rare codon clusters on high-level expression of heterologous proteins in Escherichia coli. Curr Opin Biotechnol 6: 494-500.
155. Dong H, Nilsson L, Kurland CG (1996) Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. J Mol Biol 260: 649-663.
156. Sorensen HP, Mortensen KK (2005) Soluble expression of recombinant proteins in the cytoplasm of Escherichia coli. Microb Cell Fact 4: 1.
157. Middelberg AP (2002) Preparative protein refolding. Trends Biotechnol 20: 437-443.
158. Warwicker J, O'Connor J (1995) A model for vicilin solubility at mild acidic pH, based on homology modelling and electrostatics calculations. Protein Eng 8: 1243-1251.
159. Hamilton S, Odili J, Pacifico MD, Wilson GD, Kupsch JM (2003) Effect of imidazole on the solubility of a his-tagged antibody fragment. Hybrid Hybridomics 22: 347-355.
160. Peterson AW, Heaton RJ, Georgiadis RM (2001) The effect of surface probe density on DNA hybridization. Nucleic Acids Res 29: 5163-5168.
161. Vainrub A, Pettitt BM (2002) Coulomb blockage of hybridization in two-dimensional DNA arrays. Phys Rev E Stat Nonlin Soft Matter Phys 66: 041905.
162. Kusnezow W, Hoheisel JD (2003) Solid supports for microarray immunoassays. J Mol Recognit 16: 165-176.
163. Jonkheijm P, Weinrich D, Schroder H, Niemeyer CM, Waldmann H (2008) Chemical strategies for generating protein biochips. Angew Chem Int Ed Engl 47: 9618-9647.
164. Sobek J, Aquino C, Schlapbach R (2007) Optimization workflow for the processing of high quality glass-based microarrays: applications in DNA, peptide, antibody, and carbohydrate microarraying. Methods Mol Biol 382: 33-51.

165. Sobek J, Aquino C, Schlapbach R (2007) Processing protocols for high quality glass-based microarrays: applications in DNA, peptide, antibody, and carbohydrate microarraying. Methods Mol Biol 382: 53-66.
166. Wright WE, Funk WD (1993) CASTing for multicomponent DNA-binding complexes. Trends Biochem Sci 18: 77-80.
167. Elnitski L, Jin VX, Farnham PJ, Jones SJ (2006) Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. Genome Res 16: 1455-1464.
168. Morgan XC, Ni S, Miranker DP, Iyer VR (2007) Predicting combinatorial binding of transcription factors to regulatory elements in the human genome by association rule mining. BMC Bioinformatics 8: 445.
169. Yang A, Zhu Z, Kapranov P, McKeon F, Church GM, et al. (2006) Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells. Mol Cell 24: 593-602.
170. Rabinovich A, Jin VX, Rabinovich R, Xu X, Farnham PJ (2008) E2F in vivo binding specificity: comparison of consensus versus nonconsensus binding sites. Genome Res 18: 1763-1777.
171. Chen X, Xu H, Yuan P, Fang F, Huss M, et al. (2008) Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell 133: 1106-1117.
172. Wingender E, Dietze P, Karas H, Knuppel R (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res 24: 238-241.

Curriculum Vitae

Personal:

- **Name:** Reham Hassaan Gomaa Helwa
- **Date and place of birth:** 12 June 1979, Cairo
- **Sex:** female
- **Nationality:** Egyptian
- **E-mail:** rehamhg@hotmail.com , r.helwa@dkfz.de (until December 2010)

Education:

- M.SC. Zoology (molecular biology) from Ain Shams University (November 2004)
- B.SC. Zoology from Ain Shams University (June 2000).

Employment:

- **Lecturer Assistant (**22nd December 2004 – till now) in zoology department, Ain Shams University, Cairo, Egypt.
- **Demonstrator** (13th December 2000- 22nd December 2004) in zoology department, Ain Shams University, Cairo, Egypt.

Research training:

- **PhD thesis** in the division of functional Genome Analysis, DKFZ, Heidelberg, Germany (1st August 2007 to December 2010) in the field of DNA-protein interactions using microarray technology and expression profiling of colorectal cancer with focus on galectin-4 regulation via a promoter study.

- **Pre- PhD training:**
  1. 28th October 2006 to 30th July 2007 in the Institute of Pathology, Heidelberg, Germany, in the invasion and metastasis laboratory. The main area of research was about the molecular changes in the invasion front of liver metastasis of a xenograft model.
  2. October 2004 to October 2006 in the Cancer Biology Department, Egyptian National Cancer Institute, Cairo University, Cairo, Egypt, on the genetic changes in the mitochondrial genome in colorectal cancer.

- **M.Sc. thesis** (February 2002 to October 2004) in Ain Shams University in collaboration with Egyptian National Cancer Institute, Cairo University. The main area of research was about the detection of colorectal cancer in the stool using K-Ras, P53, and BAT26 as molecular markers.

Awards:

1. Prof. Dr. T.A. Mousa for distinguished students in graduation for the academic year 2000.
2. Prof. Dr. Nahed H. Reyad for distinguished students in graduation for the academic year 2000.
3. Prof. Dr. A. H. Helmy Mohammad for the best performance in the "Principles of Taxonomy" postgraduate course (2001).
4. Egyptian governmental scholarship for studying PhD in Germany.

Publications:

- **Reham Helwa**, Andrea Bauer, Abdel-Hady A. Abdel-Wahab, Mohamed Gameel, Joerg Hoheisel (submitted). Expression profiling of colorectal cancer cell lines reveals twin SNPs in the Galectin-4 promoter that are associated with its up-regulation.

- **Reham Helwa** & Joerg D. Hoheisel (2010). DNA-protein interactions; from nitrocellulose filter-binding assays to microarray studies. *Anal. Bioanal. Chem*., in revision.

- Obul R Bandapalli , Susanne Dihlmann , **Reham Helwa** , Stephan Macher-Goeppinger, Jurgen Weitz , Peter Schirmacher & Karsten Brand (2009). Transcriptional activation of the *beta-catenin* gene at the invasion front of colorectal liver metastases. *J. Pathol.* **218**, 370–379.

Posters:

- International Congress of Genetics. Berlin, Germany July 2008
  L-DNA universal platform for the Analysis of Breast Cancer.
  Martinez R, Jacob A, Hauser N, Helwa R, Hoheisel J.

- 11th status seminar, chip technologies (5-6 March 2009), DECGEMA-Haus, Frankfurt am Main
  Setting Up a Transcription Factor Proteins Array Technology Detecting DNA-Protein Interactions
  Reham Helwa, Raphael Martinez, Andrea Bauer, Jussi Taipale and Jörg Hoheisel

- DKFZ PhD retreat, Weil der Stadt, 2009
  Setting Up a Transcription Factor Proteins Array Technology Detecting DNA-Protein Interactions
  Reham Helwa, Raphael Martinez, Andrea Bauer, Jussi Taipale and Jörg Hoheisel

- DKFZ PhD poster session, Heidelberg, 2009
  Transcriptional factors proteins array detecting DNA-protein interactions.
  Reham Helwa, Rafael Martinez, Andrea Bauer, Jussi Taipale and Jörg Hoheisel

- 2nd annual meeting of NGFN-Plus and NGFN-Transfer (26th- 28th November 2009)
  Galectin-4 upregulation is associated with -233 point mutation and hypomethylation in colorectal cancer
  Reham Helwa, Andrea Bauer, Abdel-Hady A. Abdel-Wahab, Mohamed Gameel, Joerg Hoheisel

Acknowledgement

Three years of work in the division of functional genome analysis to execute this thesis. By that time, I learnt a lot from uncountable number of people whom I want to express my deepest and warmest thanks. In addition to learning science, it is unforgettable friendship and cultural exchange. So, hereby, hopefully, with my humble acknowledgment, I could express part of my feeling toward everybody in this group.

Firstly, I want to convey my sincere gratitude to my PhD supervisor, Dr. Joerg Hoheisel, the head of functional genome analysis division for everything. He was always the true scientist with opened door. I am really speechless toward his help from accepting me in his group, encouraging and valuable advices during the thesis work, till helping in writing a grant proposal to start my career in Egypt. Furthermore, I will never forget the wonderful environment of working conditions, which enriched my skills as a student and personally, I never felt homesick or got distracted. Therefore, I am really indebted to him.

I would like to express my deepest thanks to my doctor-father Dr. Stefan Wiemann for accepting me as student, the valuable discussions with him, and the constructive comments on this thesis.

Toward Prof. Dr. Ruediger Hell, I would like to express my appreciation for accepting me as student and his warm welcoming in his office.

I gratefully acknowledge Dr. Andrea Bauer for everything; planning the work schedule, following up my progress in the thesis work, discussing the daily troubleshooting, and revising my thesis. Even personally, I will never forget her listening to my daily life experiences and her valuable advices in that regards.

I would like also to express my deepest thanks to my dearest colleagues, who introduced me to the lab, their long discussions in science, and their valuable advices. Amongst, I would like to express special thanks to Rafael Martinez, Manuel Fugazza, Anette Jacob, Christian Krauss, Christoph Schroeder, Michaela Schanne, Anette Heller, Mahmoud Younis, Yasser Riazalhosseini, and Pedro Simonini.

This work would not be completed without our collaborations. Thereby, I would like thank Prof. Dr. Abdel-Hady Ali Abdel-Wahab for his help to get the tumor samples from the Egyptian National Cancer Institute. Also, for the ORF clones of the transcription factors genes, I would express my thanks to Prof. Dr. Jussi Taipale.

I am really indebted to my wonderful country Egypt for everything in my life and toward I would like to express all of my love and thanks. Moreover, hereby, I would mention that during this thesis I was financially supported by the Egyptian governmental scholarship.