INVESTIGATION OF THE CANDIDATE TUMOR SUPPRESSOR GENE CTCF USING MULTI-OMICS DATA MINING

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF
ENGINEERING AND NATURAL SCIENCES
OF ISTANBUL MEDIPOL UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
BIOMEDICAL ENGINEERING AND BIOINFORMATICS

By
Esra DURSUN
February, 2021

INVESTIGATION OF THE CANDIDATE TUMOR SUPPRESSOR GENE CTCF USING MULTI-OMICS DATA MINING
By Esra Dursun

February, 2021

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Kıvanç Kök (Advisor)

Assist. Prof. Dr. Cüneyd Parlayan

Assist. Prof. Dr. Özge Şensoy

Approved by the Graduate School of Engineering and Natural Sciences:

Assoc. Prof. Dr. Yasemin Yüksel Durmaz

Director of the Graduate School of Engineering and Natural Sciences

Foreword

While doing this study, I set out to examine CTCF, a candidate tumor suppressor gene that plays a putative regulatory role in the formation of cancer, which is a major burden of disease worldwide. For this purpose, my thesis work first examined the comprehensive protein-protein interaction network of CTCF and revealed its diverse potential functions. Second part of this study which covered 12 different cancer tissues, uncovered methylation-specific biomarker regions that are expected to illuminate future studies. As a technical novelty, a specific multi-omics approach, based on cutting-edge data mining techniques, was applied to the CTCF research field in this study. All in all, this academic effort demonstrated that the advanced data mining approach we used in this study complements previous findings and is applicable for future studies by providing new insights on CTCF's tumor suppressor candidacy. In truth, I couldn't complete this study without a strong support group. Firstly, to my thesis advisor Assist. Prof. Dr. Kıvanç Kök, who passionately guides every stage of this study with his extensive knowledge and guidance. Secondly the thesis defense jury members and instructors who contributed to my education, since they provide an enlightening learning environment throughout my graduate education. Finally, my dear family and friends who always support me with deep love and understanding. Thank you all for your unwavering support. I dedicate this thesis to my beloved father, the person with the most beautiful soul.

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: ESRA DURSUN
Signature:

Acknowledgment

First, I want to thank my thesis advisor, Assist. Prof. Dr. Kıvanç Kök, for the guidance and mentorship in every stage of this study. I could not finish this work without his mentorship, vast academic knowledge and full support. I owe my special thanks to Assist. Prof. Dr. Cüneyd Parlayan, due to his significant contribution to my education and progress in the field of Bioinformatics. I also acknowledge other instructors of the graduate program for providing the inspiring, enlightening learning environment.

I express my endless thanks to members of the Genomed AR&GE, Tanju Basmacı and Simge Sayın Erdoğan, who have always been very understanding towards me. They allocated me the enough time to accomplish this thesis and created family peace in my workplace.

I would like to express my gratitude to my colleague Özge Anaç, one of my greatest supporters, who has been with me in all kinds of good, bad and difficult moments since the first moment we met. My sincere thanks to my childhood friend Sertab Bosnalıer, with whom I grew up, who always kept her emotional support at the highest level. I wholeheartedly thank to my dear friend and supporter Dr. Halil İbrahim Çolak, whose consultancy in the medical field was very beneficial for this thesis. I appreciate all kinds of his emotional and academic assistance.

My sincere thanks go to my life partner Nuri Yener Torlak, who is always by my side in every aspect of my life, who is always behind me in every difficulty I encounter. I am especially grateful to him for trusting me more than I do and for always supporting me. His encouragement was very reassuring, especially when situations got tough, so I could always dare new things. My heartfelt thanks.

My deepest gratitude is to my beloved mother and siblings who brought me to this day, supported me in all my choices, from the profession I chose to the work I did. Also, I am so happy that my sweet nephew and his dear father are in our family. I am very grateful to all of them because I have never felt alone and I am really lucky to be with them for the rest of my life.

And finally, I dedicate this thesis to my dear father, who made me who I am now, whom I constantly feel by my side in every moment of my life. I always strive to make him proud of everything I do and will do.

Contents

List of Figures

List of Tables

Abbreviations

GBA	Guilt by Association
TSS	Transcription Start Site
ZF	Zinc-Finger
RNAPII	RNA polymerase II
RNAPIII	RNA polymerase III
PPI	Protein-Protein Interaction
FDR	False Discovery Rate
GO	Gene Ontology
BP	Biological Process
MF	Molecular Function
CC	Cellular Component
PCA	Principal Component Analysis
TGFβ	Transforming Growth Factor Beta
BMP	Bone Morphogenetic Proteins

ÖZET

CTCF ADAY TÜMÖR SÜPRESÖR GENİNİN MULTİOMİK VERİ MADENCİLİĞİ İLE ARAŞTIRILMASI

Esra DURSUN
Biyomedikal Mühendisliği ve Biyoinformatik Yüksek Lisans
Tez Danışmanı: Dr. Öğr. Üye. Kıvanç KÖK
Şubat, 2021

CCCTC bağlanma faktörü (CTCF), omurgalılarda bulunan bir 11 çinko parmak proteinidir. Her dokuda ifade edilen bu çok işlevli transkripsiyon faktörü, genomdaki sayısız hedef bölgeye bağlanabilmektedir. Bu proteinin diğer ilişkilerinin yanısıra; transkripsiyon aktivitesinin düzenlenmesi, kromatin yapısının kontrolü ve embriyonik gelişimle alakalı olduğu görünmüştür. Önemli bir diğer husus ise, gün geçtikçe CTCF'in bir aday tümör süpresör gen olarak tanınmasının artmasıdır. Bu adaylık meselesini ele almak için yapılan çalışmalar artmasına rağmen, mevcut kanıtlar hala yetersizdir. Bunun esas sebebi deneysel bulguların eksikliğidir. Multi-omik verilerinin ortaya çıkışı, gelişi ve veri madenciliği tekniklerindeki gelişmeler, bu boşluğu doldurmak için benzeri görülmemiş bir fırsat sunmaktadır. Bununla ilgili olarak bu çalışma, çoklu omik veri madenciliği yaklaşımını kullanarak, CTCF'nin kanserle ilişkisine yeni bakış kazandırmayı amaçlamaktadır. Buna göre metilom, transkriptom ve interaktom olmak üzere üç moleküler karmaşıklık düzeyindeki alakalı veriler ilgili veritabanlarından elde edilmiştir. CTCF odaklı interaktomik seviyedeki analiz, CTCF'nin yaygın olarak bilinen kromatin organizasyonundaki rolünü doğrulamış ve yeni fonksiyonlarını tahmin etmiştir. Daha detaya inecek olursak, tamamlanan proteinprotein etkileşim ağ analizinde bağlantı yoluyla suç (GBA) yaklaşımı kullanılarak CTCF kanserle ilişkilendirilmiştir. DNA metilomu ve transkriptomdan hareketle yapılan analiz, CTCF regülasyonunun gelişim ve kanserle ilgili yönlerine yoğunlaşmıştır. İntegratif veri madenciliği yaklaşımı gelişim boyunca CTCF ile ilişkili belirgin prob seviyesindeki metilasyon-gen ifadesi korelasyon paternlerini ortaya çıkarmıştır. Ayrıca birçok kanser çeşidinin incelenmesi, CTCF ile bağlantılı, belirli metilasyon yerleri için doku çeşidine has bir şekilde kanser ve normal dokular arasında diferansiyel metilasyon-ekspresyon korelasyonunu ortaya çıkarmıştır. Bu yerler yeni aday kanser biyobelirteçler olarak belirlenmiştir. Prob seviyesinde hem yoğunluk hem de genomik konum farklılıklarını aynı anda hesaba katan bu tür bir korelasyon raporu CTCF üzerine yapılan böyle ilk çalışmadır. Genel olarak bu öncü çalışma, kullanılan çoklu omik veri madenciliği yaklaşımının uygulanabilirliğini göstermiş, ilgili deneysel bulguları tamamlamış ve CTCF'nin kanserdeki varsayılan rolünün açıklığa kavuşturulmasına katkı sağlamıştır.

Anahtar Sözcükler: CTCF aday tümör süpresör geni, Multi-omik veri madenciliği, GBA paradigması, Gen fonksiyonunun tahmini, Diferansiyel korelasyon analizi, Aday metilasyon biyobelirteçler

ABSTRACT

INVESTIGATION OF THE CANDIDATE TUMOR SUPPRESSOR GENE CTCF USING MULTI-OMICS DATA MINING

Esra DURSUN
M.S. in Biomedical Engineering and Bioinformatics
Advisor: Assist. Prof. Dr. Kıvanç KÖK
February, 2021

CCCTC binding factor (CTCF) is a vertebrate, 11-zinc-finger protein. This ubiquitously expressed multifunctional transcription factor can bind to a myriad of target sites in the genome. It is involved, among others, in regulation of transcription activity, chromatin structure control, embryonic development. Importantly, CTCF has been increasingly recognized as a candidate tumor suppressor gene. Despite the growing body of research addressing this candidacy issue, the current evidence is still inconclusive. This is mainly due to the lacking experimental findings. The emergence of multi-omics data and recent advances in data mining techniques offer an unprecedented opportunity to fill this gap. Relatedly, this study aimed provide a novel insight into the CTCF's implication in cancer using a multi-omics data mining approach. Accordingly, the relevant data on three levels of molecular complexity, namely the interactome, methylome, the transcriptome were retrieved from related repositories and explored. The CTCF-focused interactomics level analysis confirmed its widely known biological roles in chromatin organization and predicted novel functions. More specifically, the accomplished in-depth investigation of the protein-protein interaction network implicates CTCF in cancer using the guilt-by-association (GBA) approach. The performed DNA methylome and transcriptome-driven analysis concentrated on the developmental and cancer-related aspects of CTCF regulation. The integrative data mining approach uncovered CTCF related distinct probe-level methylation-gene expression correlation patterns across the development. Furthermore, examination of multiple cancer types revealed tissue-specific differential correlation between cancer and normal tissues for certain CTCF-associated methylation sites. These sites were proposed as novel candidate cancer biomarkers. This is the first such correlation report on CTCF, which simultaneously takes into account probe-level differences in both intensity and genomic location. Overall, this pioneering work demonstrates the applicability of the employed multi-omics data mining approach, complements related experimental findings and paves the way for the clarification of CTCF's putative role in cancer.

Keywords: Candidate tumor suppressor protein CTCF, Multi-omics data mining, GBA paradigm, Gene function prediction, Differential correlation analysis, Candidate methylation biomarkers

CHAPTER 1

INTRODUCTION

1.1 Protein-protein Interactions and Interactome

1.1.1 Protein-protein interaction networks

Interactions networks have been extensively utilized in biomedical research to unravel, model and represent complex biological processes. The advent of Systems Biology accelerated research on interaction networks. The rapid proliferation in this domain has led to rise of novel research fields. One of such prominent and promising fields is Network medicine [1]. This field benefits from network techniques with the ultimate goal to address complex biomedical questions. Idea of projecting interactions among the components of complicated biological systems into network provided a useful platform not only for modelling biological mechanisms, but also for integrating diverse types of biomedical data in a combinatory manner. Protein-protein interactions network represent a specific type of interactions, which covers connections on protein level (Figure 1). Here, proteins (gene products) are represented as nodes and links between proteins are depicted as edges. Protein-protein interactions be reconstructed using, among others, co-expression or physical interaction as data input. Proteins in the network can be automatically annotated using overrepresentation algorithms, based on functional annotation resources. This allows biologically meaning interpretation of the findings. Relatedly, other data types, such as differential expression values, can be mapped onto the network, enabling integration of multiple data types in a unified manner.

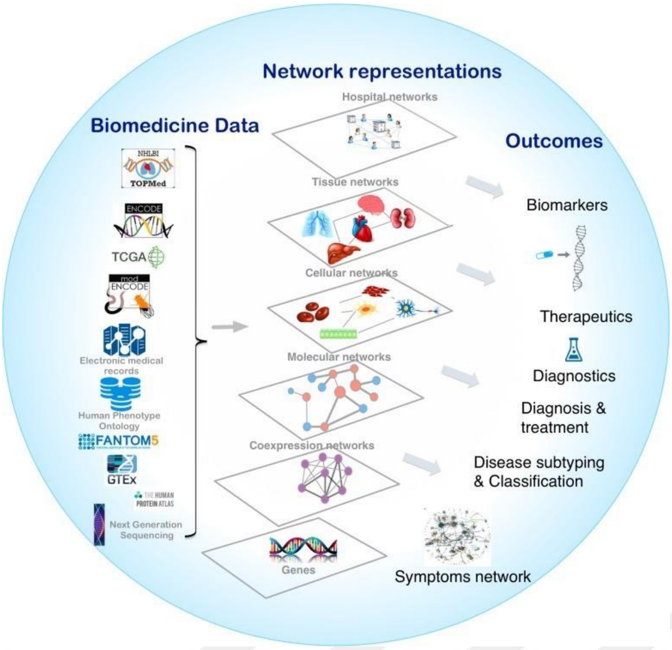


Figure 1: Application layers in Network Medicine, corresponding to levels of biological complexity [1]

1.1.2 Guilt by association paradigm

The guilt by association paradigm (GBA) is used to predict a previously unknown function of a gene [2]. At the basis of this approach, networks created based on interactions between any two genes are used. By examining the functions of the genes with which this gene interacts in the network, it is thought to be a potential candidate for new roles that it was previously not associated with [3] (Figure 2).

The GBA method is frequently used across many fields. An example of using this method to find new genes associated with known diseases is the study of Ke-Jun Ye et al. In China in 2018, using this paradigm to find new optimal genes for Fetal Growth Restriction (FGR). As a result of the study, they found new promising biomarkers to be used in the diagnosis of FGR [2].

Similarly, in a study conducted in 2019, the GBA method was used to unravel new genes that participate in the development of glucose homeostasis Type 1 diabetes mellitus (T1DM). As a result of this study, new potential cell pathways for the therapeutic or preventive treatment of the disease were found [4].

As a different approach of the GBA paradigm instead of finding new candidate genes for diseases, examining the network of a gene, implicated in a disease, and trying to find potential new functions of this gene can be given as an example. An example of this approach can be given in a study conducted by Gloeckner et al. in 2020, to examine the widespread protein protein network of Leucine-rich repeat kinase 2 (LRRK2) protein. LRRK2 is a member of G protein family, associated with Parkinson's disease. In this study, they were found that the LRRK2 protein functions as a major scaffold protein, which is related to cytoskeletal dynamics and associated with vesicular transport [5].

Modelling of co-expression networks with the objective of unravelling novel disease biomarkers is among three common approaches for effectively combining biological interaction maps with phenotype-level biomedical data [1]. Network modelling can also be used to identify and prioritize novel candidate disease genes. The notion of GBA is central to modelling of the protein-protein interaction map, irregardless of the data source. In another words, this universal methodology can also be applied to other types of networks, such as physical interaction networks. The GBA frequently integrates functional annotation data and the knowledge of putative functional modules within the network in order to better predict gene function.

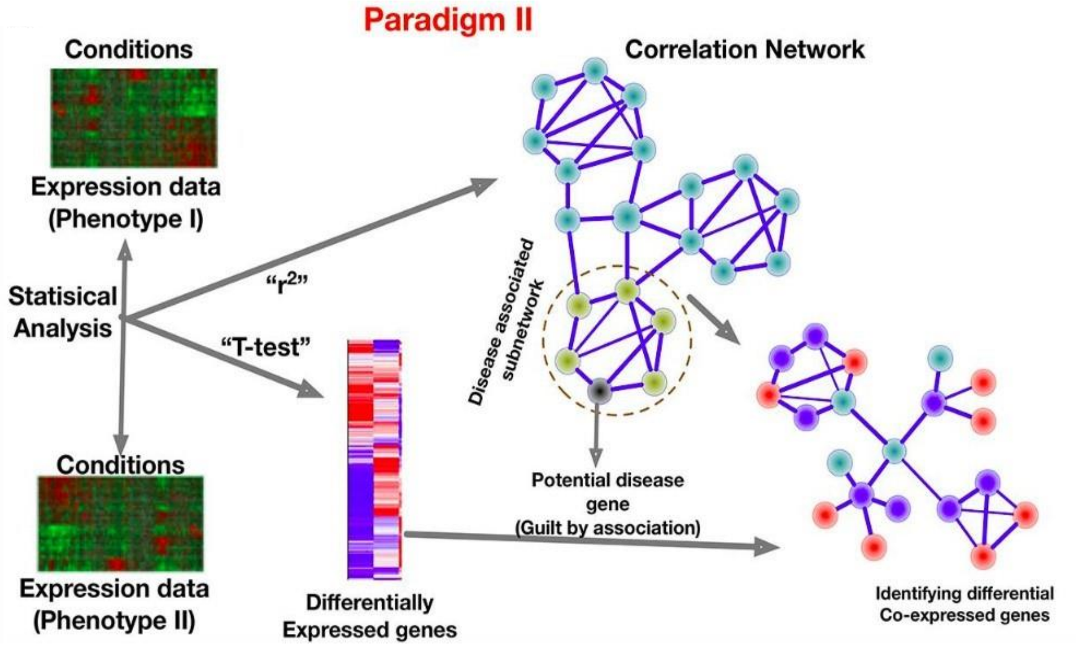


Figure 2: Network modeling and the Guilt-by-association (GBA) approach [1]

1.2 Methylation and Methylome

1.2.1 The biochemical basis of DNA methylation

Changes in gene expression and function, which are driven by direct alterations of DNA sequence in an organism, are within the research scope of Genetics studies. Abnormalities, such as deletions and duplications in the genome, are some of among these changes, which can result into various diseases, such as cancer [6]. However, genetic studies alone are not sufficient to explain the basis of diseases. Apart from, genetic factors, epigenetic mechanisms, which modify gene activity without directly changing the DNA sequence, perform crucial regulatory function (Figure 3A). Among them, DNA methylation has been the most studied mechanism [7].

While all cells within the organism possess the identical genetic information, gene expression in these cells is not static. This variation is partially mediated by epigenetic changes. On the one hand, DNA Methylation occurs by direct chemical modification of DNA. On the other hand, other epigenetic modifications, such as histone modifications, the length of the nucleosome along the DNA, ensure the balance of induction and reduction of gene expression by altering the structure of chromatin. In this way, regulation of DNA packaging plays a crucial role in DNA replication and in control of gene function [8]. The focus of this thesis is on DNA methylation, which is a type of epigenetic modification.

Historically, in mammals, discovery of methylation concided approximately with the discovery of DNA inheritance unit [9]. In later studies, although most researchers suggested that methylation had an important effect on changing gene activity, this notion remained as a hypothesis until 1980 [10]. While modulation of mRNA expression level by DNA methylation is an established fact today [11] (Figure 3B), this research topic still maintains its popularity and methylation has been the most studied epigenetic mechanism [6]. Covalent binding of a methyl group to the 5th carbon of cytosine gives rise to DNA methylation.

Most DNA methylation happens in cytosines, which precede a guanine nucleotide. These two nucleotides are collectively known as CpG sites. DNA methylation can affect gene activity differently based on the type of the genomic location, such as CpG regions, gene body and intergenic region. In another words, effect of DNA methylation is location-dependent. Thus, it is an advantage to take into account such aspects by examining the mentioned regions separately in methylation studies.

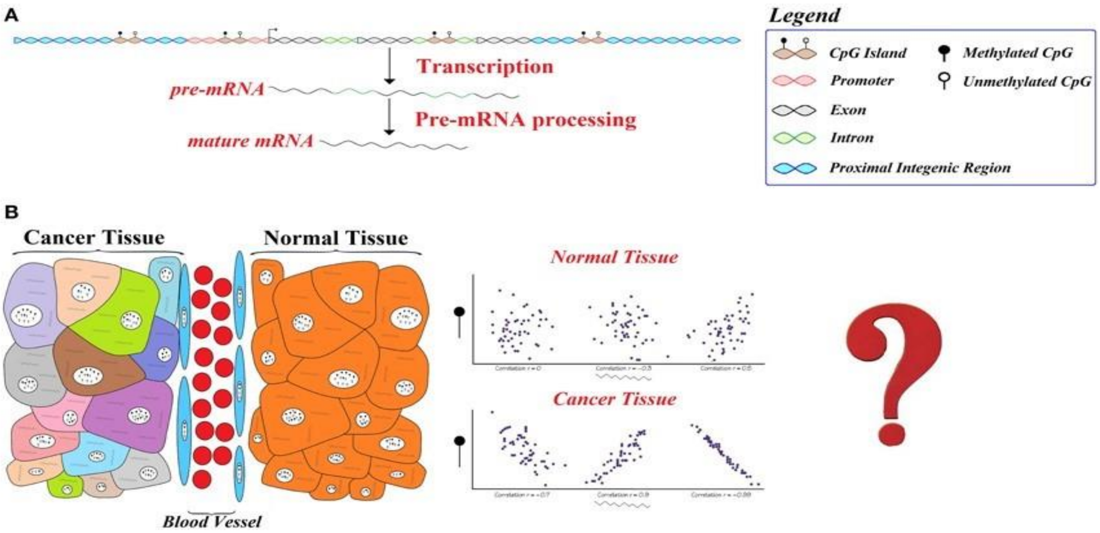


Figure 3: The reasoning behind differential DNA methylation-expression correlations in cancer A. Methylation as epigenetic regulatory mechanism. B. Identification of differential correlation of methylation and gene expression.[11]

1.2.2 CpG islands

The regions formed by the succession of cytosine and guanine nucleotides are called. CpG regions, and the regions where these regions are dense called CGI (CpG islands). Although these CpG islands are present in numerous locations throughout the genome, they are generally positioned in the promoter regions of genes. These promoter regions contain housekeeping genes and are thought to be crucial for the function of the genes as they are conserved throughout the course of evolution [12]. For this reason, in methylation studies, instead of examining the whole genome, important CpG regions in the promoter regions of the genes are focused.

In terms of tissue specificity, CpG islands show low tissue specificity, whereas CpG shore regions 2 kb from the TSS region and CpG shelves regions 4 kb from the TSS region have high tissue specificity [13]. In addition, it appears that methylation in these regions have an effect on the reduction of gene expression [14].

1.2.3 Gene body

The gene body is viewed as be the part where the gene crosses the first exon. Like promoter methylation, gene body methylation leads to gene silencing. However, studies have shown that methylation in the dividing cells in the gene body increases gene expression. For this reason, methylation appears to be positively correlated with gene expression [15].

However, in later studies, it has been observed that methylations in the gene body also reduce gene expression in the cells that divide slowly or do not divide. As a result, the effect of methylation on the body expression is still uncertain [16], [17].

1.2.4 DNA methylation abnormalities

Proper DNA methylation is critical for accurate gene expression and function. Any abnormation in methylome can cause a variety of diseases, (for instance, cancer). With hypermethylation, which means that methylation is more than normal, hypomethylation, which means that methylation is normally low, can both cause diseases.

Hypomethylation seems to be especially high in cancerous cells. Oncogenes, which should normally be silenced by methylation, can be activated by hypomethylation and causing cancer. Similarly, by hypermethylation of tumor suppressor genes that prevent cancer formation, these genes are silenced and cause cancer [18].

Most of the promoters, which are influenced by loss of DNA methylation correspond to tissue-specific genes. For example, antigens normally found only in testicular cells appear to be actuated in other types of cancer by hypomethylation [19].

In addition to DNA hypomethylation, another methylation balance change is called as hypermethylation. It has been shown in many studies that tumor suppressor genes are inactivated by hypermethylation in tumors, causing cancer. This was first demonstrated in 1989 with the discovery of hypermethylation in the promoter of the retinoblastoma (RB1) tumor suppressor gene [20].

In this way, these genes can be used as markers in tumors formed by hypermethylation of genes that normally take part in regulating gene expression. For instance, cancer can be observed by hypermethylation of genes implicated in the mechanism of DNA repair. As an example, hypermethylation of BRCA1 gene, which is participates in DNA repair, has been frequently studied in ovarian cancer and breast cancer [21].

Silencing genes involved in cell adhesion by DNA hypermethylation, for example CDH1 (E-cadherin) and CDH13 (H-cadherin) can cause tumor development. Genes such as apoptose-related protein kinase 1 (DAPK1), which mediate apoptotic signaling pathways (Gordian), can be inactivated by hypermethylation, leading to immortalization of the cancer cell [22]. Each type of cancer creates a separate methylome, so that hypermethylation can be monitored to differentiate cancer types [7].

1.3 Gene Expression and Transcriptome

1.3.1 mRNA expression level

The RNA molecule is the main component of all living creatures. RNA research is essential for examining the functions of genes, which were previously shown with DNA level studies [23]. The RNA content, which varies between cell types, and is directly related to protein formation for many genes.

The total number of all RNAs coding and non-coding in a cell is called transcriptome. In recent years, transcriptome studies have been carried out frequently to better examine the structure and function of the genes [24]. Transcriptome analysis is very important in protecting and maintaining the identity of cells, in the formation of gene expression regulatory complexes and especially in terms of being the main component of housekeeping complexes. In addition, misregulation on the gene expression level causes various developmental diseases. With these analyzes, questions that have been sought for years about the cell nature and tumorigenesis have been addressed [25].

Analysis of mRNA allows us to examine tissue and cell-dependent gene expression characteristics. In this way, we can better examine the dissimilarities in the cells with the variation in the transcriptome profile and we can evaluate the causes of the diseases in detail and finally take a step towards treatment [24]. Much of the previously unknown point about eukaryotic genome and transcriptome has been clarified by conducting transcriptome studies, and it has been found that protein-encoding genes make up just 2% of the genome [26]. Regulation of gene expression is an essential mechanism, which is central to the functioning of our entire body, in diverse biological contexts [27].

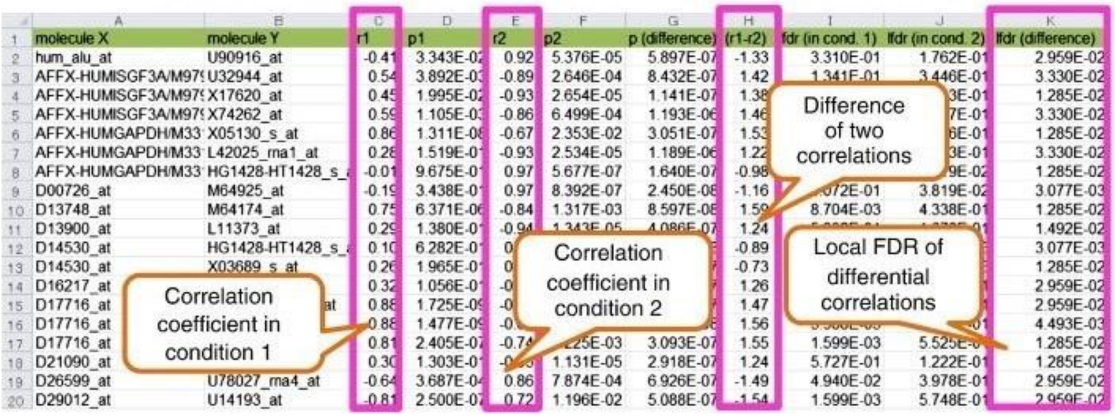
1.3.2 Gene expression microarrays

It is now possible to capture almost all of the transcripts using the microarray or sequence method, which are used to display DNA. The microarray method is less costly than RNA sequencing [28]. However, it has various limitations as it relies on the predetermined knowledge of the genome as a disadvantage. RNA sequencing, on the other hand, allows more comprehensive studies.

In the gene expression microarray method, transcripts that are not covered by the designed probe sets will not be recognized because these microarrays are based on predetermined probes. However, the number of these probes is updated and expanded day by day. Although it contains probes needed to identify a large number of important transcript sites, it is not as comprehensive (in terms of transcriptome coverage) as RNA sequencing [24]. However, in terms of its cheap cost and speed, it is still used more than the transcriptome sequence method. Especially in online databases, they have more microarray data than sequence data. In this study, microarray data will be used since it covers larger datasets instead of RNA sequence data.

1.3.3 Differential DNA methylation-gene expression correlation

Differential DNA methylation-gene expression correlation, is a recently designed statistical method to decipther regulatory differences between different tissue types or conditions. Genes, which show such differential correlations are referred shortly as “differentially correlated genes”[11](Figure 4). To the best of the author’s knowledge, this methodology has not been applied to investigate CTCF up until this thesis. Thus, this work represents the first such study. A straightforward method to compute Differential DNA methylation-gene expression correlation is to substracted two obtained correlation coefficients [29]. The most commonly used coefficient is the Pearson correlation coefficient. Other common methods, such as Spearman's Rank Correlation and Kendall Rank Correlation, are also applicable. Altogether, differences between tissues/conditions in their respective mRNA level correlation with DNA methylation level is a promising research avenue.



	A	B	C	D	E	F	G	H	I	J	K
1	molecule X	molecule Y	r1	p1	r2	p2	p (difference)	(r1-r2)	fdr (in cond. 1)	lfdr (in cond. 2)	lfdr (difference)
2	hum_alu_at	U90916_at	-0.41	3.343E-02	0.92	5.376E-05	5.897E-07	-1.33	3.310E-01	1.762E-01	2.959E-02
3	AFFX-HUMISGF3A/M97	U32944_at	0.54	3.892E-03	-0.89	2.646E-04	8.432E-07	1.42	1.341E-01	3.446E-01	3.330E-02
4	AFFX-HUMISGF3A/M97	X17620_at	0.45	1.995E-02	-0.93	2.654E-05	1.141E-07	1.38	[illegible]	3E-01	1.285E-02
5	AFFX-HUMISGF3A/M97	X74262_at	0.59	1.105E-03	-0.86	6.499E-04	1.193E-06	1.46	[illegible]	7E-01	3.330E-02
6	AFFX-HUMGAPDH/M33	X05130_s_at	0.86	1.311E-08	-0.67	2.353E-02	3.051E-07	1.53	[illegible]	6E-01	1.285E-02
7	AFFX-HUMGAPDH/M33	L42025_rna1_at	0.28	1.519E-01	-0.93	2.534E-05	1.189E-06	1.22	[illegible]	3E-01	3.330E-02
8	AFFX-HUMGAPDH/M33	HG1428-HT1428_s_a	-0.01	9.675E-01	0.97	5.677E-07	1.640E-07	-0.98	[illegible]	9E-02	1.285E-02
9	D00726_at	M64925_at	-0.19	3.438E-01	0.97	8.392E-07	2.450E-08	-1.16	.072E-01	3.819E-02	3.077E-03
10	D13748_at	M64174_at	0.75	6.371E-06	-0.84	1.317E-03	8.597E-08	1.59	8.704E-03	4.338E-01	1.285E-02
11	D13900_at	L11373_at	0.29	1.380E-01	-0.94	1.343E-05	4.086E-07	1.24	[illegible]	[illegible]	1.492E-02
12	D14530_at	HG1428-HT1428_s_a	0.10	6.282E-01	0	[illegible]	[illegible]	-0.89	[illegible]	[illegible]	3.077E-03
13	D14530_at	X03689_s_at	0.26	1.965E-01	0	[illegible]	[illegible]	-0.73	[illegible]	[illegible]	1.285E-02
14	D16217_at	[illegible]	0.32	1.056E-01	-0	[illegible]	[illegible]	1.26	[illegible]	[illegible]	2.959E-02
15	D17716_at	at	0.88	1.725E-09	-0	[illegible]	[illegible]	1.47	[illegible]	[illegible]	2.959E-02
16	D17716_at	[illegible]	0.88	1.477E-09	-0.	[illegible]	[illegible]	1.56	[illegible]	[illegible]	4.493E-03
17	D17716_at	[illegible]	0.81	2.405E-07	-0.74	.225E-03	3.093E-07	1.55	1.599E-03	5.525E-0	1.285E-02
18	D21090_at	[illegible]	0.30	1.303E-01	[illegible]	1.131E-05	2.918E-07	1.24	5.727E-01	1.222E-01	1.285E-02
19	D26599_at	U78027_rna4_at	-0.64	3.687E-04	0.86	7.874E-04	6.926E-07	-1.49	4.940E-02	3.978E-01	2.959E-02
20	D29012_at	U14193_at	-0.81	2.500E-07	0.72	1.196E-02	5.088E-07	-1.54	1.599E-03	5.748E-01	2.959E-02

Difference of two correlations
Correlation coefficient in condition 2
Local FDR of differential correlations
Correlation coefficient in condition 1

Figure 4: Example for computing differential correlation between DNA methylation and gene expression [29]

1.4 Regulatory Protein CTCF

1.4.1 Multifunctional protein CTCF

The CCCTC binding factor (CTCF) is a ubiquitous, highly conserved 11-zinc-finger vertebrate protein. It performs multiple functions and binds to myriad of target sites in the genome (Figure 5). Phylogenetic analysis shows that CTCF protein emerged in the early evolution of Metazoa and is shared among bilaterians as a conserved protein [30]. CTCF owes its multifunctional to diverse molecular interactions, which are mediated by distinct CTCF domains. According to the current understanding, each function is mediated by a distinct set of interaction partners [31].Thus, knowledge, which available for the interaction partners, sheds light on CTCF functions. Some of these protein partners, also interact among each others, forming protein complexes. Notably, the notion of putative function modules is pivotal for untagling these complexes.

In the early studies, CTCF protein was identified as a transcriptional suppressor of the Myc gene. In the subsequent inquiries, other notable functions such as enhancer blockage, inactivation of X-chromosome, gene expression suppression and activation or suppression of promoter were found [32], [33].

There are domains in the structure of CTCF enable its binding to different DNA motifs and to a large number of various regulatory proteins [34]. These different roles of the CTCF are generally thought to be the result of regulating of chromatin loop establishment in collaboration with the cohesine, which is involved in controlling the chromatin structure. Relatedly, although the studies CTCF research has frequently concentrated on the CTCF’s importance in the chromatin organization, more studies are needed on what role it plays in the developmental process [35].

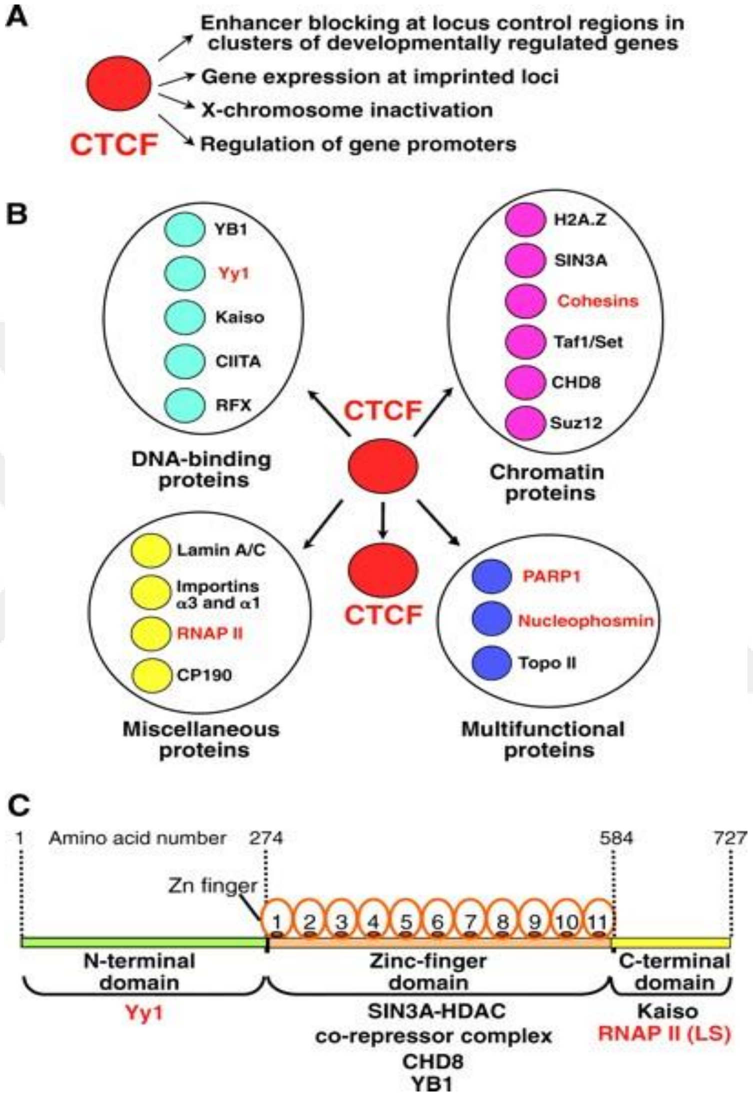


Figure 5: Summary of CTCF (A) functions, (B) network and (C) structure [31]

1.4.2 CTCF's domain structure and binding to DNA

CTCF consists of 3 main regions; It consists of a N-terminal site C-terminal site and a central zinc-finger (ZF) field, consisting of 11 C2H2. The ZFs have different roles. The ZF 3-7 binds to a 15 bp DNA motif. Other ZFs interact with adjacent DNA modules to regulate CTCF-binding stability [36]. All three areas of CTCF are open to rearrangement of these interactions, with post-translational changes by interacting with other proteins or RNA [37].

Studies conducted on breast cancer cells, demonstrated that abnormal polyADPribocylation of CTCF causes separation from the CDKN2A locus, leading to the tumor suppressor gene silencing in an abnormal manner and causing cancer [38].

Based on performed studies, transcriptional regulation of genes also seems appears to be cell type dependent CTCF. For instance, in primary mouse embryonic fibroblasts, misregulation of the 698 gene was shown to be caused by inactivation of the CTCF gene [39]. As a similar example, inactivation of the CTCF gene was shown to alter the expression level of circa 400 and 800 genes in mitotic embryonic and postnatal neurons, respectively [40].

1.4.3 CTCF protein-protein interactiome

Under certain conditions, such as changes in the cell, environmental conditions, the proteins which interacts with CTCF can change. Since the researchers have so far discovered new proteins that are reported to interact with CTCF every day, they have classified them into 4 main groups to make it easier to examine: DNA-binding proteins, chromatin proteins, multifunctional proteins, and miscellaneous proteins that not included in these groups [31].

If we first mention the group of DNA binding proteins, the proteins included in this group are: YB1, Yy1, KAISO, CIITA, RFX. The YB1 protein was shown to play crucial role in DNA replication, repair and transcription as DNA binding protein by interacting with Yy1 protein [41].

Yy1 is a ZF transcription factor. In the Tsix center of the X chromosome inactivation, the CTCF-Yy1 binding sites are clustered. CTCF appeared to interact with the N terminal to transactivate Tsix of Yy1 [42].

KAISO protein, as a ZF transcription factors, is an important DNA binding protein in development and cancerogenesis process. It acts to reduce the blocking effect of the enhancer by binding to the unmethylated sequence region near the 5' β-globulin via the CTCF-C domain [43].

The transcription factor RFX binds to the proximal promoters of the MHCII gene. CIITA Protein, on the other hand, is a transcriptional coactivator protein, which regulates expression by interacting with chromatin reformers and transcription factors. CTCF interacts directly with RFX and CIITA proteins, forming a triplex complex. This complex permits the expression of certain genes in a CTCF binding site-dependent manner [44].

Secondly, some of the proteins included in the group of chromatin proteins are H2A.Z SIN3A, Cohesins, Taf1/Set, CHD8, Suz12 proteins. H2A.Z and H2A protein are structural components of the nucleosome. H2A.Z protein, identified as CTCF cofactors by CTCF-affinity chromatography, has been found to be localized with CTCF across the genome [45]

Suz12 protein is the main component of polycomb repressor complex 2 (PRC2), which provides the methylation of histone H3 in the Lysine 27 region. It binds to the maternal allele of the P2 and P3 promoters of suppressed Igf2 allele in the Igf2/H19 locus, and interacts directly with CTCF [46].

SIN3A is a transcriptional co-repressor protein. It plays a role in activation of histone deacetylase by connecting to CTCF via zinc finger domain [47].

CHD8 protein is a member of the chromodomain family, which plays a role in the control of chromatin formation and gene expression. This protein has been found to bind to known CTCF zinc finger domains such as the promoter of H19 ICR, BRCA1 and MYC, HS5 of the LCR of the 5'β-globulin gene family. Silencing of CTCF or CHD8 appeared to result in loss of ICR isolator activity in luciferase reporter plasmids.

It also appeared that removal of CHD8 induced CpG hypermethylation and histone hypoacetylation in the BRCA1 and Myc promoters near the CTCF binding sites [48].

Taf1/Set proteins are molecular chaperones that are components of the INHAX complex that inhibits histone acetyltransferas. They have been identified as CTCF cofactors [45].

The cohesin protein, consisting of 4 subunits (Smc1, Smc3, Scc1 and Scc3), which together form the ring-like structure in sister chromatids, plays an important role in homologous recombination due to proper separation of chromosomes and DNA repair. Cohesin forms the control region of the main latent-associated transcript (LAT) gene of the herpesvirus associated with Kaposi's sarcoma, colocalized with CTCF [49]. Cohesin also functions as an isolator in H19 ICR and is required for the human p-globin locus in reporter plasmids [50]. Cohesin and CTCF bind to maternal DNA molecules, controlling transcription in both the G1 and G2 cells at the Igf2 / H19 suppressed locus. The cohesin interacts with CTCF in the Myc isolator, and it appears that the uptake of the cohesin to the chromosomal regions (Igf2 / H19 and DM locus) depends on the presence of CTCF [51].

A third group of multifunctional proteins are PARP1, Nucleophosmin and Topo II. Finally, the protein group that cannot be included in any group is called miscellaneous proteins, and examples include Lamin A / C, Importins, RNAP II and CP190 [31].

CP190 is the centrosome binding protein. Studies in Drosophila have been shown to be essential for life but not essential for cell division. CP190 protein is required for direct functioning of CTCF binding to chromatin by directly interacting with CTCF [52].

DNA binding transcription factors (TFs) are among the most important regulators of gene expression. Even though TFs are usually classified into distinct categories to facilitate understanding, in reality they are constantly interacting and interconnected with each other [53].

The analysis of the molecular mechanisms interacting through CTCF (c-myc, hTERT, RB, RBL2, CDKN2A and TP53) found that the tumor suppressor role of CTCF depends on the cell type. As mentioned earlier, the ability of the CTCF to bind CTSs has been shown to depend on cell type-specific factors such as DNA methylation, BORIS binding, and CTCF PARylation [54].

When the interaction of CTCF with other proteins is examined, an additional strategy has been introduced that this protein has an important function during cell differentiation and this interaction can be regulated in various genomic regions. Although CTCF-mediated chromosome interaction has been found to contain many proteins such as Yin-Yang 1 (YY1), Kaiso, CHD8, PARP1, Maz, JunD, ZNF143, Prdm5 and Nucleophosmin, to function properly and stably, only the cohesin has been shown to be required [49], [55], [56]. When the molecular mechanism of the interaction between the cohesin, which is shown to be important in this way, with the CTCF was examined, it was found that this interaction occurred through the carboxy terminal region of the CTCF and the SA2 subunit of the cohesin. Just like CTCF, cohesin has been observed to be present in intergenase-regulating regions, promoters, introns, and 5'UTRs of genes during interphase. Depending on the cell type, 50-80% of CTCF regions also appeared to be coated by cohesin, and down regulation of the cohine using RNAi was found to cause disruption of CTCF-mediated intra-chromosomal interactions [57].

Another factor that interacts with the CTCF is TFIIIC. This TFIIIC is a necessary factor for transcription of tRNAs, 5S rRNA, SINE B2 elements and other non-coding RNAs by RNA polymerase III (RNAPIII) [58]. TFIIIC appears to bind to many genomic regions that do not have RNAPIII and are called extra TFIIIC (VB) loci. This regulatory factor has been shown to cluster and bind DNA sequences of both tRNA genes and ETC loci to the nuclear environment in yeasts. As a result of all genome studies, it has been shown that CTCF and its binding partner, cohesin, are found in mouse and human cells near many tRNA genes and ETC loci [59], [60].

1.4.4 CTCF function in developmental process

In mice, the lack of CTCF in the oocytes causes embryo mortality in the morula stage, while homozygous null mutant embryos cannot be implanted in the pre-implantation stage, resulting in their death [61]. In zebrafish, the deficiency of CTCF in the unicellular phase results in a fatal outcome 24 hours after embryo fertilization [62]. When the reason of these remedies are examined in detail in both organisms, it is seen that there is common apoptosis mediated by the p53 down regulation and the Puma up regulation. Both of these genes are the direct targets of CTCF. Based on these results, it can be concluded that the CTCF played an active role in the very early development period [63].

CTCF is also involved in brain and neural development. When levels of gene expression were examined during development, it was observed that the level of CTCF expression decreased in the process from birth to adulthood [64].

CTCF is also known to play a pivotal role in limb development. In a mouse study, genes involved in limb development have been shown to have a large number of CTCF binding sites within their promoter regions hinting for CTCF’s function in establishing a regulatory complex [65]. The presence of CTCF has also been found in HOX genes, which coordinate limb development and generally participate in transcription regulation [66].

Based on these studies, examining the CTCF in the course development is a a rational initial step to unravel how the 3D genome organization incorporates numerous stimuli involved in the initiation and proper execution of transcription [67].

1.4.5 Diseases linked to CTCF

Knowing that CTCF has such important functions, it is estimated that changes in the chromatin structure can lead to various pathologies due to mutations in mediated loop formation. It is anticipated that various diseases may occur due to the change of the CTCF profile. As an example, studies have found that polymorphisms that alter the CTCF binding motif increase the susceptibility to autoimmune diseases such as vitiligo, multiple sclerosis and systemic lupus erythematosus, by altering the expression of human leukocyte antigens (HLA). As a similar example, SNP (single nucleotide exchange) encoded with rs34481144 has been found to affect the severity of a flu virus infection [68]. When the molecular mechanism of this is examined, with the change of C to T in Chr11, this SNP changes the CTCF binding motif, which increases the CTCF’s interaction with the promoter of Interferon-Induced Transmembrane Protein 3 and consequently a decrease in the expression of this gene. found to be [69].

Mutations in CTCF are have mostly been associated with and studied in context of the following diseases: mental retardation, Wiedemann syndrome, Silver-Russell syndrome and a variety of cancers. Germline loss of CTCF appears cause syndromic intellectual disability and autosomal dominant mental retardation 21 (MRD21). It has been observed that frame shift mutations can cause poor binding of CTCF to DNA [70].

1.4.6 Candidate tumor suppressor protein CTCF

The involvement of CTCF in multiple cancer types has been demonstrated. Studies have shown that mutations in the CTCF sequence lead to truncated protein formation, which influences the progression of head and neck cancer. Similarly, changes in CTCF binding motif were observed in colorectal cancer. Changes in the organization of CTCF mediated chromatin have been causally linked to mutations of some cancer genes [71].

CTCF gene has been associated with the following cancer types: testicular, colorectal , bladder and ovarian due to abnormal DNA methylation at the same loci. Similarly, the dysfunction of CTCF in the paternal allele resulted into the formation of different various cancer types, including (SRS), hepatocellular carcinoma [72], Wilms tumor [73], testicular cancer [74]and craniopharyngioma [75]. As a result of investigating CTCF-cohesin binding sites (CBS) abnormalities in cancers, mutations have been found in gastrointestinal cancer and skin cancer, and these mutations have been associated with late replication [76]. Considering the important roles of CTCF and its relationship with diseases, it is thought that it may be a factor in tumorigenesis since it interacts with many tumor related genes. When the mechanisms related to tumorigenesis are examined, CTCF is directly involved in the transcription control of diverse critical factors of cellular growth, apoptosis, silenceing, aging and differentiation. In this way, studies have been carried out with the idea that CTCF can interact with genes involved in these mechanisms and act as a tumor suppressor [77].

When the NCBI Pubmed literature on “tumor suppressor CTCF” was searched, it was seen that the studies increased over the course of the last years. The table, which included the search results (the number of PubMed records so far) was downloaded and updated to show the cumulative total and moving average by years (Table A.1). In order to analyze these results better, 2 different bar charts were created (Figure A.1). Part A and part B of this figure focus on the number of publication per year and the cumulative sum of publication (per year), respectively.

1.5 Aims

The aim of this study was to adopt a data mining approach, which was applied before to other field(s), for the research on CTCF. Previously designed methodologies were slightly modified and reformulated to suit the objectives of this study. The ultimate goal was to test the applicability of the employed approach and obtain novel insights in the context of the tumor suppressor protein CTCF. The first step to explore in details the CTCF protein-protein (PPI) interaction network using a series bioinformatics tools. These tools implement a range of data mining techniques and provide comprehensive visualizations for network analysis. This was an attempted to validate already known biological roles and uncover the unknown functions of this gene. The ultimate goal was to obtain new indications implicating this regulatory protein in cancer. This functional prediction relied on the GBA paradigm and the notion of the putative functional modules. The topology-based clustering and functional annotation analysis was essential for prediction of CTCF's function in the reconstructed protein-protein interaction network.

Secondly, impact of DNA methylation on CTCF expression was assessed by through computing the correlation of CTCF-associated methylation and gene expression data. Considerably strong correlation is indicative of methylation's impact on CTCF function. In contrast to the common methodology of previous studies, which relied on averaging probe-level methylation levels, such intensities were considered separately in this comprehensive study. The rationale for this careful examination was to obtain a higher resolution view on methylation patterns and to identify novel candidate probe-level cancer markers. As such, all these aspects resulted into unprecedented level of scrutiny, which adds to the novelty of the employed approach the change of CTCF in the developmental process and subsequently the change of the CTCF profile in cancer cells compared to normal cells.

CHAPTER 2

MATERIALS AND METHODS

2.1 Overall Design of the Multi-omics Data Mining Research

This thesis utilized 3 different data types: interactome data, methylome data and transcriptome data (Figure 6). Of these, methylation and transcript level data were taken from the NCBI GEO repository. Interactome data was retrieved using the Cytoscape Genemania App. The cross-omic nature of the performed research and the selected data computational solution to the name of this thesis.

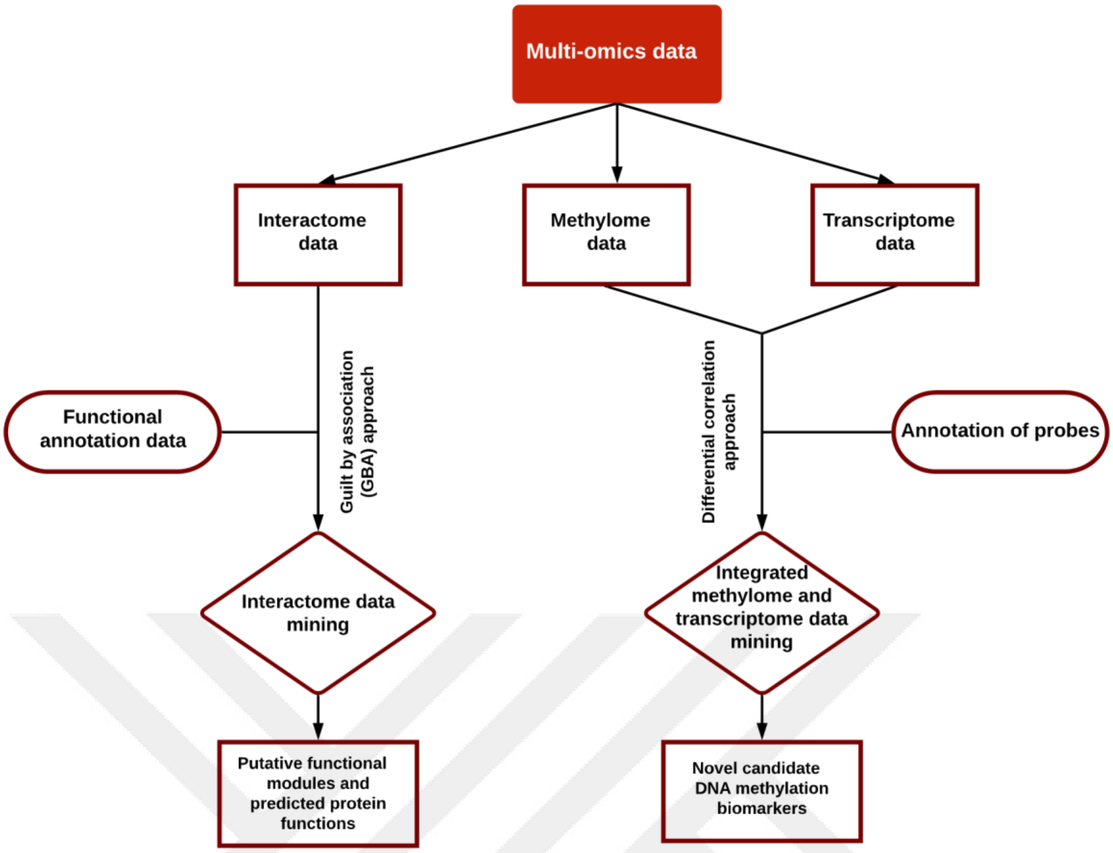

Multi-omics data
Interactome data
Methylome data
Transcriptome data
Functional annotation data
Guilt by association (GBA) approach
Differential correlation approach
Annotation of probes
Interactome data mining
Integrated methylome and transcriptome data mining
Putative functional modules and predicted protein functions
Novel candidate DNA methylation biomarkers

Figure 6: Multi-omics data mining workflow. Overview of the omic data types and data mining approaches used in this study

2.2 Interactomics Analysis

2.2.1 Cytoscape-based network analysis

The Cytoscape tool (v.3.8.1) was used to examine the genes with which the CTCF interacts and to observe the biological pathways in which it is involved. Cytoscape is the most commonly used open source software platform for studying complex molecular networks [78]. This tool also includes additional plugins that provide a very comprehensive data visualization.

GeneMANIA, a Cytoscape plug-in tool, was used for protein interaction analysis. GeneMANIA uses available genomic and proteomics data to best estimate the function of the protein in question and make a list of associated proteins and genes. Data for six different organism (human and five commonly used model organisms) is available in GeneMANIA. GeneMANIA collects the results of previous studies and presents the network analysis, along with the detailed list of the functions of the associated proteins [79].

Through the high accuracy of the GeneMANIA estimation algorithm, extensive database analysis and the additional tool integrated into the Cytoscape program, it is used in many studies [80]. This thesis implemented GeneMANIA App was used to retrieve and comprehensively examine genes related to the CTCF gene.

As a first step of network analysis, GeneMania App (v.3.5.2) was used to determine the gene list with which the CTCF gene interacts. As an alternative, an online version of this applications is also available. Gene full names, as further information about the members of the network, were obtained from the UniProt databases. Using the guilt-by association approach, Cytoscape enables to finding the putative functional modules, associated with the gene under investigation by applying topology clustering and functional annotation [80].

Then, by using the NetworkAnalyzer App (v.4.4.6), the topology parameters of this network (including node degrees, average clustering coefficients, topological coefficients) were calculated. NetworkAnalyzer is a plug-in commonly used to calculate network topology parameters, including the mean number of interaction partners and the number of linked node pairs [81].

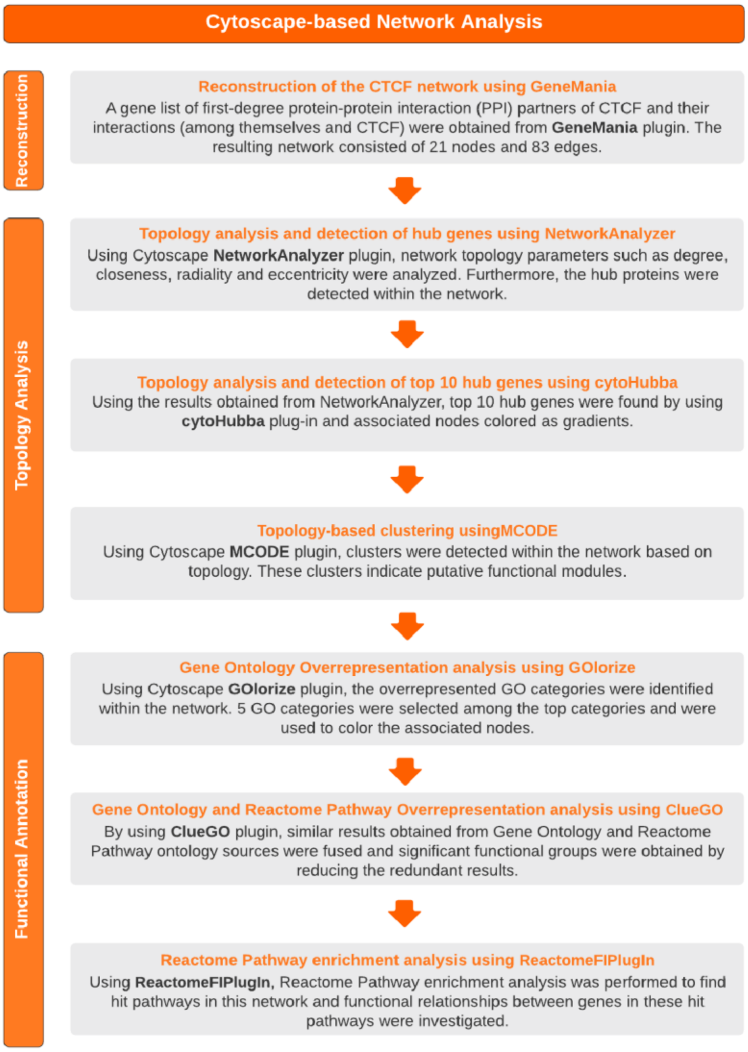
With the cytoHubba App (v.0.1), the most important nodes in this network were found using the degree algorithm from the previously determined topology parameters. The top10 gene with the highest degree value was selected and visualized on the network with a color gradient.

Then, MCODE App (v.1.6.1) was used to find gene clusters that are in similar pathways in a network entered as input and have similar functional roles in metabolism [82].

The GOlorize App (v1.0.0.beta1) has been used to observe overrepresented gene ontologies in this network. The GOlorize plugin uses BINGO as the first step. It then highlights proteins related to the same category using color coding and creates an advanced representation of the interaction map [83]. By using Gene Ontology (GO) database, the most represented 5 GO categories were selected and the genes on the network were colored according to their inclusion of these categories.

In the results obtained by using the GO database, the ClueGO App (v.2.5.7) was used to filter the redundant biological terms and to determine the most important pathways and functions. This plugin defines the edges that show the interactions of terms in the network using kappa statistics. In addition to the GO source, this plugin also uses databases such as KEGG, WikiPathways and Reactome. In this plug-in, in addition to filtering redundant terms, similar results are fused to increase the significance of the results [84].

In addition to over representation analysis, ReactomeFI App (v.7.2.3) was used to perform pathway enrichment analysis. This plugin enables pathway enrichment analysis for the entered gene list by accessing Reactome pathways in databases and enables the investigation of the functional relationships of genes involved in hit pathways in this network [85]. The workflow (Figure 7) showing the summary of the tools used in this network analysis is shown below.



Reconstruction of the CTCF network using GeneMania
A gene list of first-degree protein-protein interaction (PPI) partners of CTCF and their interactions (among themselves and CTCF) were obtained from **GeneMania** plugin. The resulting network consisted of 21 nodes and 83 edges.

Topology analysis and detection of hub genes using NetworkAnalyzer
Using Cytoscape **NetworkAnalyzer** plugin, network topology parameters such as degree, closeness, radiality and eccentricity were analyzed. Furthermore, the hub proteins were detected within the network.

Topology analysis and detection of top 10 hub genes using cytoHubba
Using the results obtained from NetworkAnalyzer, top 10 hub genes were found by using **cytoHubba** plug-in and associated nodes colored as gradients.

Topology-based clustering usingMCODE
Using Cytoscape **MCODE** plugin, clusters were detected within the network based on topology. These clusters indicate putative functional modules.

Gene Ontology Overrepresentation analysis using GOlorize
Using Cytoscape **GOlorize** plugin, the overrepresented GO categories were identified within the network. 5 GO categories were selected among the top categories and were used to color the associated nodes.

Gene Ontology and Reactome Pathway Overrepresentation analysis using ClueGO
By using **ClueGO** plugin, similar results obtained from Gene Ontology and Reactome Pathway ontology sources were fused and significant functional groups were obtained by reducing the redundant results.

Reactome Pathway enrichment analysis using ReactomeFIPlugIn
Using **ReactomeFIPlugIn,** Reactome Pathway enrichment analysis was performed to find hit pathways in this network and functional relationships between genes in these hit pathways were investigated.

Figure 7: Workflow of the Cytoscape based network analysis.

2.2.2 Complementary (online tools-based) network analysis

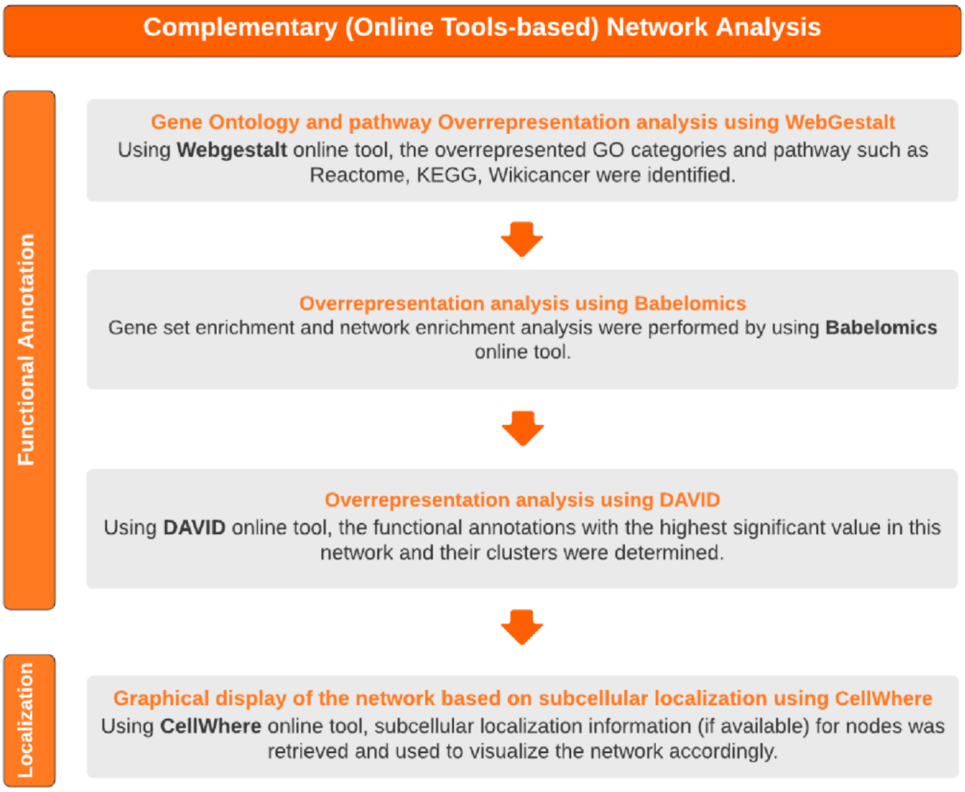
Webgestalt (WEB-based Gene Set Analysis Toolkit) (accessed on 11 November 2020) is an online application, which provides functional enrichment analysis functionality. A commonly methods, known as Representation Analysis (ORA), was selected for enrichment analysis [86]. The ORA analyzes were performed using the Gene ontology and Reactome pathway databases of this tool. GO-related findings were compare the results obtained with Cytoscape.

Redundancy reduction option, which is an extra feature of this tool, has been used especially to reduce excessive results and to examine the most meaningful results. This redundancy reduction option is based on 2 methods, affinity propagation and weighted set cover. Shortly, the affinity propagation method uses Jaccard index for similarity measurement and creates a representative gene set containing significant p value for each cluster. Similarly, weighted set cover method creates the minimum gene subset that can cover all of the enriched genes by using significant p values [86].

Babelomics web tool (accessed on 12 November 2020) was used to do functional enrichment analyze. This web tool, is used for the analysis of the genomic data, which includes steps such as normalization, clustering and differential gene expression [87]. This tool was used to validate the ORA results obtained with Webgestalt and thus to check the consistency of the results.

As the last functional annotation tool, DAVID, which stands for “The Database for Annotation, Visualization and Integrated Discovery”, v.6.8 web tool (accessed on 12 November 2020) was used. It performs comprehensive functional annotation analysis using the input gene list [88]. The results obtained were displayed in 3 different ways: functional clustering, functional chart and functional table. According to the strength of the cluster, the most hit gene clusters were obtained.

Finally, the CellWhere online tool (accessed on 24 October 2020) was used to examine the sub-cell localizations of the genes in the network [89]. For this, the file in xgmml format obtained from GeneMania was imported into the program. Then, firstly, with the default option, that is, "Screening by flavor" option was selected as muscle, and the analysis was made. Using "annotation frequency" as a second option, two figures showing the sub cell localizations of the genes are obtained. The workflow (Figure 8) showing the summary of the tools used in this network analysis is shown below.



Complementary (Online Tools-based) Network Analysis

Gene Ontology and pathway Overrepresentation analysis using WebGestalt

Using **Webgestalt** online tool, the overrepresented GO categories and pathway such as Reactome, KEGG, Wikicancer were identified.

Overrepresentation analysis using Babelomics

Gene set enrichment and network enrichment analysis were performed by using **Babelomics** online tool.

Overrepresentation analysis using DAVID

Using **DAVID** online tool, the functional annotations with the highest significant value in this network and their clusters were determined.

Graphical display of the network based on subcellular localization using CellWhere

Using **CellWhere** online tool, subcellular localization information (if available) for nodes was retrieved and used to visualize the network accordingly.

Figure 8: Workflow of the Complementary (Online Tools-based) Network Analysis

2.3 Integrated Mining of Methylation and Gene Expression Data

2.3.1 Data selection and the integrative data mining approach

Integration of methylation and gene expression data was achieved using the integrative data mining approach, which was mainly based on methylation-gene expression correlation analysis. Both types of data were retrieved from the related areas in the GEO database and match the same sample. The dataset covers specific representative tissue/condition type in order to demonstrate the flexible applicability of the employed approach. The overview of the selection procedure is summarized in the figure 9. The compiled dataset is summarized in the table 1. Detailed description is available in Appendix.

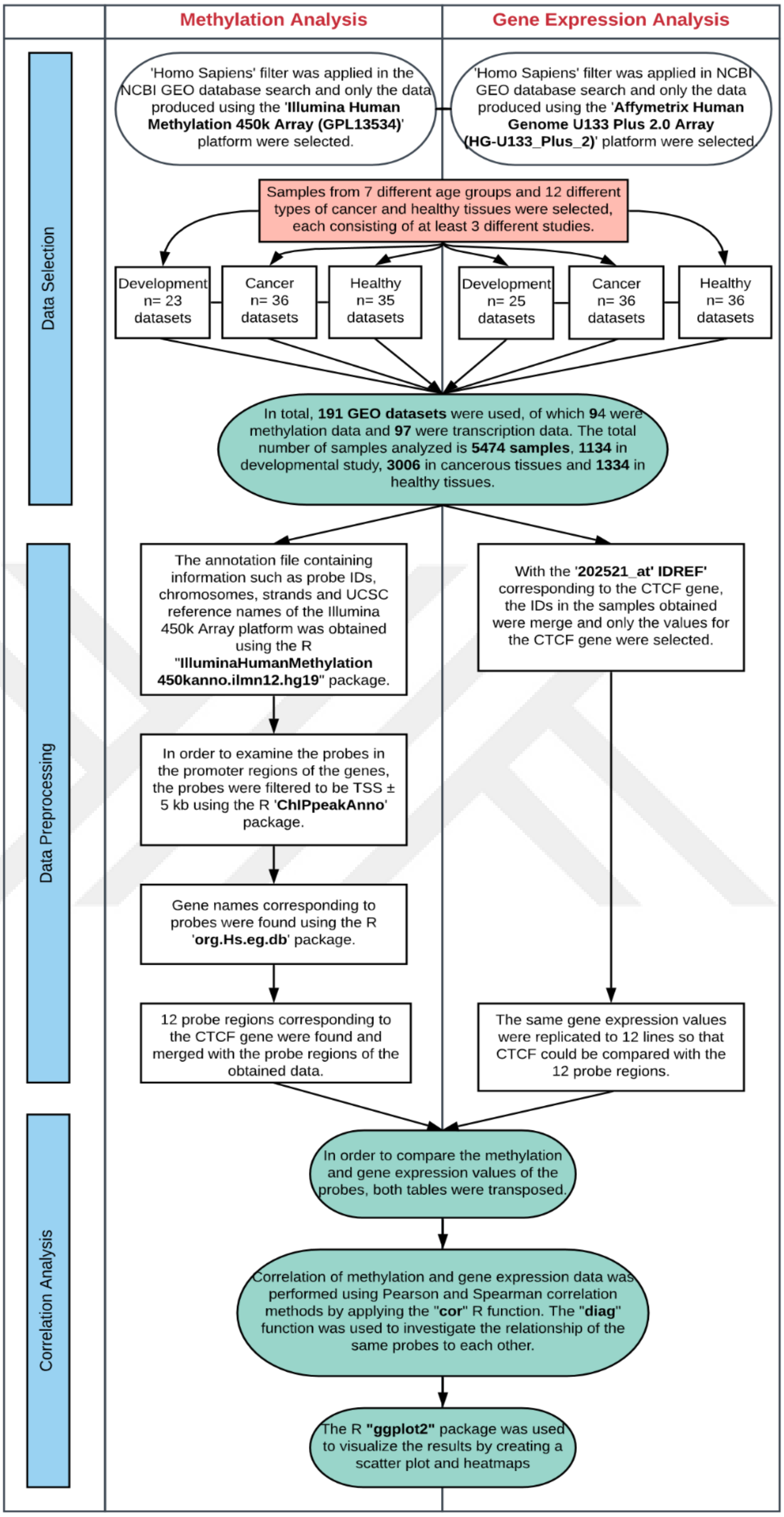

Methylation Analysis
Gene Expression Analysis
Data Selection
'Homo Sapiens' filter was applied in the NCBI GEO database search and only the data produced using the 'Illumina Human Methylation 450k Array (GPL13534)' platform were selected.
'Homo Sapiens' filter was applied in NCBI GEO database search and only the data produced using the 'Affymetrix Human Genome U133 Plus 2.0 Array (HG-U133_Plus_2)' platform were selected.
Samples from 7 different age groups and 12 different types of cancer and healthy tissues were selected, each consisting of at least 3 different studies.
Development n= 23 datasets
Cancer n= 36 datasets
Healthy n= 35 datasets
Development n= 25 datasets
Cancer n= 36 datasets
Healthy n= 36 datasets
In total, 191 GEO datasets were used, of which 94 were methylation data and 97 were transcription data. The total number of samples analyzed is 5474 samples, 1134 in developmental study, 3006 in cancerous tissues and 1334 in healthy tissues.
Data Preprocessing
The annotation file containing information such as probe IDs, chromosomes, strands and UCSC reference names of the Illumina 450k Array platform was obtained using the R "IlluminaHumanMethylation450kanno.ilmn12.hg19" package.
With the '202521_at' IDREF' corresponding to the CTCF gene, the IDs in the samples obtained were merge and only the values for the CTCF gene were selected.
In order to examine the probes in the promoter regions of the genes, the probes were filtered to be TSS ± 5 kb using the R 'ChIPpeakAnno' package.
Gene names corresponding to probes were found using the R 'org.Hs.eg.db' package.
12 probe regions corresponding to the CTCF gene were found and merged with the probe regions of the obtained data.
The same gene expression values were replicated to 12 lines so that CTCF could be compared with the 12 probe regions.
Correlation Analysis
In order to compare the methylation and gene expression values of the probes, both tables were transposed.
Correlation of methylation and gene expression data was performed using Pearson and Spearman correlation methods by applying the "cor" R function. The "diag" function was used to investigate the relationship of the same probes to each other.
The R "ggplot2" package was used to visualize the results by creating a scatter plot and heatmaps

Figure 9: Combined methylation and gene expression analyis. Modified PRISMA flow chart including selection of omic data and detailed steps of data analysis.

Research area	**Organ/phenotype**	**Selected datasets**	**Analyzed samples**	**Table ID in Appendix**
Development	fetal	9	196	B.1.1
Development	newborn	7	160	B.1.2
Development	infant	6	120	B.1.3
Development	childhood (5-17)	6	302	B.1.4
Development	early adulthood (18-40)	7	140	B.1.5
Development	late adulthood (41-80)	7	164	B.1.6
Development	senescence (80+)	6	52	B.1.7
Cancer Tissue	bladder cancer	6	60	B.2.1
Cancer Tissue	bone cancer	6	98	B.2.2
Cancer Tissue	brain cancer	6	506	B.2.3
Cancer Tissue	breast cancer	6	274	B.2.4
Cancer Tissue	colon cancer	6	72	B.2.5
Cancer Tissue	gastric cancer	6	288	B.2.6
Cancer Tissue	kidney cancer	6	406	B.2.7
Cancer Tissue	liver cancer	6	262	B.2.8
Cancer Tissue	lung cancer	6	406	B.2.9
Cancer Tissue	pancreas cancer	6	126	B.2.10
Cancer Tissue	prostate cancer	6	380	B.2.11
Cancer Tissue	small intestine cancer	6	128	B.2.12
Healthy Tissue	bladder	6	22	B.3.1
Healthy Tissue	bone	6	124	B.3.2
Healthy Tissue	brain	6	392	B.3.3
Healthy Tissue	breast	6	80	B.3.4
Healthy Tissue	colon	6	70	B.3.5
Healthy Tissue	gastric	6	116	B.3.6
Healthy Tissue	kidney	4	96	B.3.7
Healthy Tissue	liver	6	110	B.3.8
Healthy Tissue	lung	6	266	B.3.9
Healthy Tissue	pancreas	6	14	B.3.10
Healthy Tissue	prostate	6	26	B.3.11
Healthy Tissue	small intestine	7	18	B.3.12
TOTAL (SUM)		**191**	**5474**	

Table 1: Summary of dataset annotations

2.3.2 Mining CTCF methylation data

For methylation data, Illumina HumanMethylation450K BeadChip (HumanMethylation450_15017482) platform coded GPL13534 from NCBI GEO database was used. PRISMA protocol was applied while selecting methylome data to be used in this study. Accordingly, the PRISMA protocol is used to narrow the research area to the purpose, as databases contain a large number of data, including primary studies, clinical trials, and meta-analyzes, and most of them involve waste of information such as duplicate studies, incomplete or incorrect results [90].

Since this study is a multi-omics study and includes many sub-research steps, the modified version of the PRISMA flowchart was used instead of the classical version. Likewise, in accordance with this protocol, duplicated and incomplete studies and studies with incomplete information about the relevant data were not included in the study. See figure 9 for a detailed illustration of how many datasets were included in the study.

Illumina HumanMethylation450k BeadChip Array contains more than 480,000 CpG site probes. It was was made available by Illumina (CA, USA) in 2011 [91].

It has numerous advantages over other methylation platforms. The first advantage that there will be no selective bias towards shorter fragments due to the lack of PCR. As another advantage is that it is much cheaper than other methylation platforms such as whole bisulfite sequencing. And through these advantages, a lot of scientific research has been done using this platform, and thus, the number of data obtained with this platform in online databases is very high [92]. Since we use online databases in this study and we do not want batch effects caused by platform differences, we only used the data obtained with this platform.

2.3.3 Illumina Human Methylation 450k Array Analysis to Find Probe Regions of the CTCF Gene

To create an annotation file, all informations about the cg probes such as; name, which chromosome is located, starting region, strand etc. was obtained using the "IlluminaHumanMethylation450kanno.ilmn12.hg19" package (v.0.6.0) and converted to Granges object by selecting only required informations [93]. The annotation file named as “TSS.human.GRCh37” found in the “ChIPpeakAnno” package (v.3.20.1) which contains start and end region of all genes according to the hg19 reference genome was imported to R and the distance of all probes to TSS regions was calculated by "annotatePeakInBatch" function [94].

In this function, "output=both" option applied to find the distance of all probes to nearest gene instead of only distance of overlapping ones. The function that calculates the distance of the genes to the TSS region was updated to calculate the distance from the end points of the genes in the reverse strand and the starting points of the genes in the forward strand. Gene symbols corresponding to ensembl ids were obtained using the "org.Hs.eg.db" package (v.3.10.0) [95] .

After finding the probe regions corresponding to all genes, probes in the CTCF gene were selected. Likewise, in the expression data, the ID corresponding to the CTCF gene was found and the results were filtered to include only the CTCF gene.

Clustvis (accessed on 7 August 2020), an online web application, was used to create the PCA and heatmaps [96].

2.3.4 Mining CTCF gene expression data

For the expression data, Affymetrix chip technology, referred as "Human Genome U133 Plus 2.0 Array (HG-U133_Plus_2)". This platform is coded as "GPL570" in the NCBI GEO database was used. As mentioned above, the PRISMA protocol was followed while selecting the transcriptome data.

This platform covers 47,000 transcripts [97]. Using the Power of the Probe Set, it includes multiple independent measurements for each transcript, providing the highest accuracy and repeatability of any microarray platform.

In this study, the same platform selected for all gene expression trancript-level data to ensure consistency between samples. Similar, like the Illumina Human Methylation 450k array platform, this platform includes more probe sets than other gene expression platforms and enables the screening of approximately 14500 genes. It is much cheaper than whole genome sequencing platforms, so there are large numbers of samples in online databases.

2.3.5 Methylation- gene expression correlation analysis

Simple correlation is a statistical method, which used to measure relation between two variables. Two basic methods are used to measure correlation: Pearson and Spearman. Pearson's correlation is the most commonly used method for investigating correlation. It is a parametric method. Spearman is a nonparametric correlation method. Accordingly, the first method to be used to measure the relationship of two variables, showing a normal distribution, is Pearson correlation. Spearman correlation is used among others, more often, in case of non-normal distribution and in examining outliers [98]. In this work, the correlation between methylation and gene expression were measured directly using the cor function in R(v.3.6.3) by using R Studio (v.1.1.456). This function uses the Pearson correlation method by default. The correlation coefficient shows the level and direction of the correlation. For cancer-related research steps, differential correlation was assessed, The studied difference was computed by substracting two correlation coefficients. The correlation results obtained in this study were visualized in the form of a scatter plot using the R ggplot2 (v.2) package.

CHAPTER 3

RESULTS

3.1 Mining CTCF Interact

3.1.1 Cytoscape-based network analysis

3.1.1.1 Reconstruction of the CTCF network using GeneMania

In this study, in order to comprehensively explore the CTCF protein-protein interaction.

(PPI) network, we first identified the primary protein partners of the CTCF gene using Cytoscape's GeneMania plugin. According to the result, a network with 21 nodes (Table 2; Figure 10) and 83 edges was obtained.

Gene Symbol	Gene Full Name
CTCF	“CCCTC-binding factor”
CHD8	“chromodomain helicase DNA binding protein 8”
POU5F1	“POU class 5 homeobox 1”
SMAD6	“SMAD family member 6”
KANSL1	“KAT8 regulatory NSL complex subunit 1”
ZMYM2	“zinc finger MYM-type containing 2”
SMAD4	“SMAD family member 4”
SMAD5	“SMAD family member 5”
RPS6KB1	“ribosomal protein S6 kinase B1”
IRAK2	“interleukin 1 receptor associated kinase 2”
SET	“SET nuclear proto-oncogene”
ZMYM1	“zinc finger MYM-type containing 1”
YBX1	“Y-box binding protein 1”
NPM1	“nucleophosmin (nucleolar phosphoprotein B23 numatrin)”
ADNP	“activity dependent neuroprotector homeobox”
LLPH	“LLP homolog long-term synaptic facilitation”
FAU	“Finkel-Biskis-Reilly murine sarcoma virus (FBR-MuSV) ubiquitously expressed”
IFNG	“interferon gamma”
SMAD1	“SMAD family member 1”
THRB	“thyroid hormone receptor beta”
EBNA1BP2	“EBNA1 binding protein 2”

Table 2: Nodes details in the network created with GeneMania

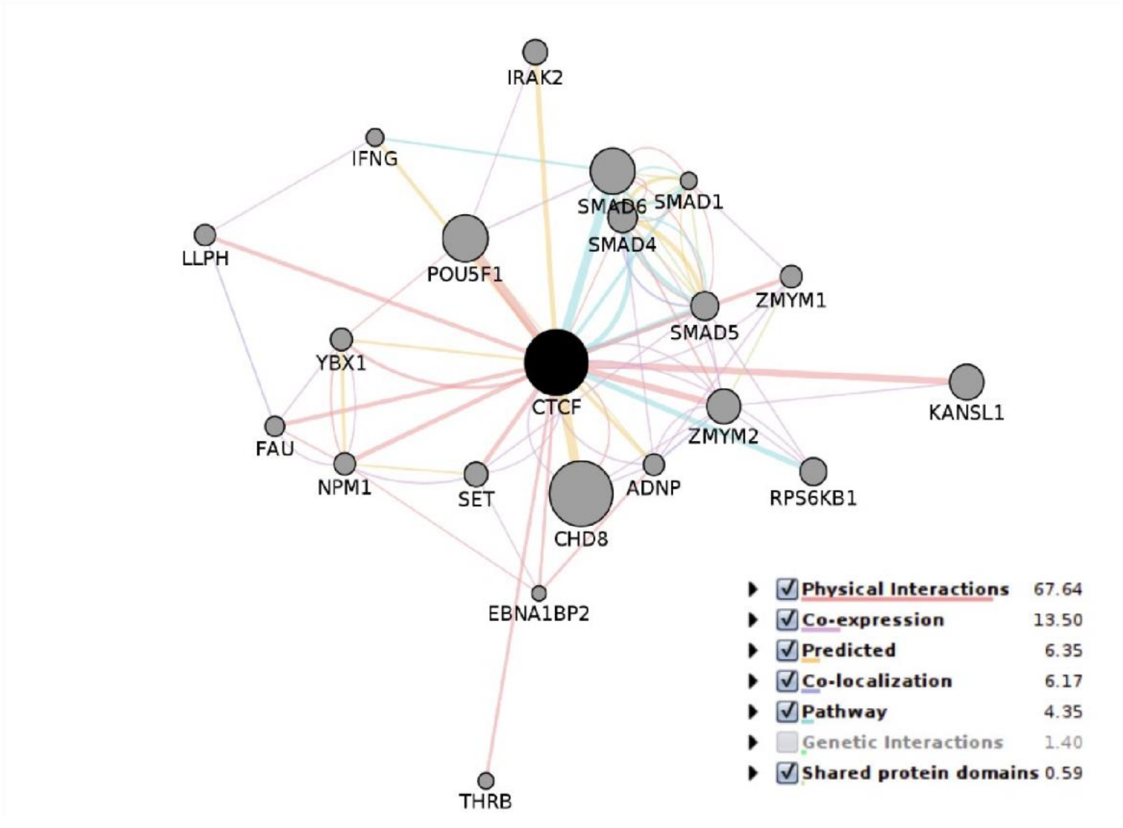
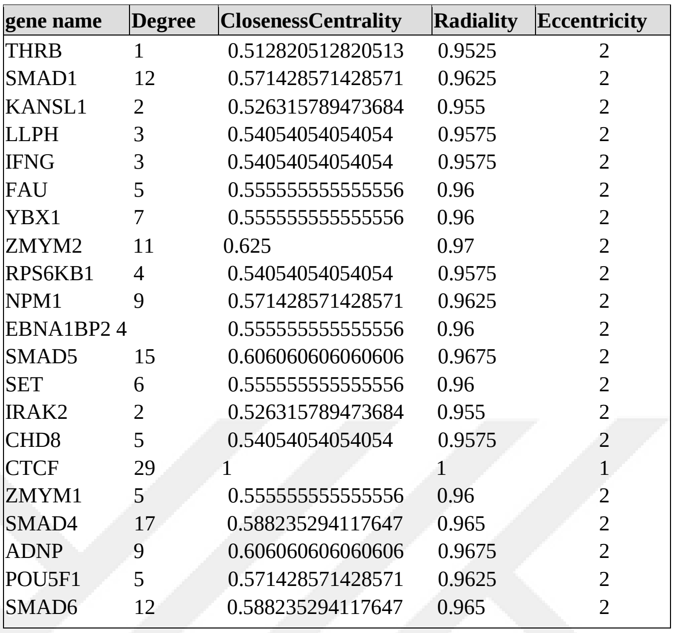


Figure 10: Gene network obtained using the Cytoscape GeneMania plugin

3.1.1.2 Topology analysis and detection of hub genes using NetworkAnalyzer

In order to find topology parameters such as degree, closeneess, radiality of this CTCF network, the NetworkAnalyzer plug-in was used (Table 3).

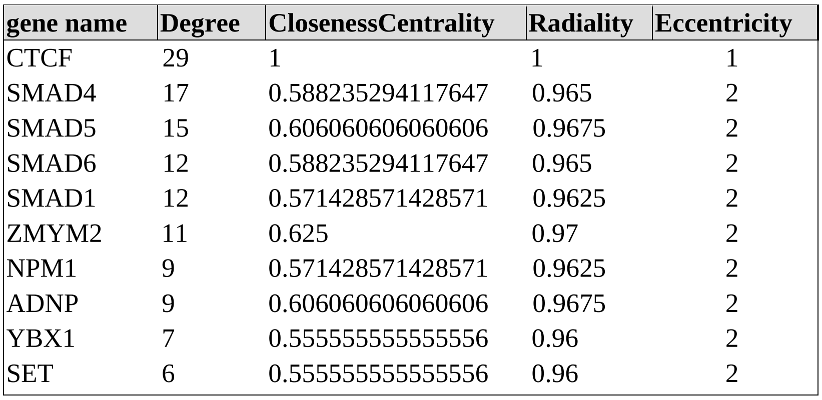


gene name	Degree	ClosenessCentrality	Radiality	Eccentricity
THRB	1	0.512820512820513	0.9525	2
SMAD1	12	0.571428571428571	0.9625	2
KANSL1	2	0.526315789473684	0.955	2
LLPH	3	0.54054054054054	0.9575	2
IFNG	3	0.54054054054054	0.9575	2
FAU	5	0.555555555555556	0.96	2
YBX1	7	0.555555555555556	0.96	2
ZMYM2	11	0.625	0.97	2
RPS6KB1	4	0.54054054054054	0.9575	2
NPM1	9	0.571428571428571	0.9625	2
EBNA1BP2	4	0.555555555555556	0.96	2
SMAD5	15	0.606060606060606	0.9675	2
SET	6	0.555555555555556	0.96	2
IRAK2	2	0.526315789473684	0.955	2
CHD8	5	0.54054054054054	0.9575	2
CTCF	29	1	1	1
ZMYM1	5	0.555555555555556	0.96	2
SMAD4	17	0.588235294117647	0.965	2
ADNP	9	0.606060606060606	0.9675	2
POU5F1	5	0.571428571428571	0.9625	2
SMAD6	12	0.588235294117647	0.965	2

Table 3: NetworkAnalyzer topology parameters

3.1.1.3 Topology analysis and detection of top 10 hub genes using cytoHubba

According to this result, top10 hub gene containing the highest degree result was determined (Table 4). These hub genes are respectively; CTCF, SMAD4, SMAD5, SMAD6, SMAD1, ZMYM2, NPM1, ADNP, YBX1, SET.



gene name	Degree	ClosenessCentrality	Radiality	Eccentricity
CTCF	29	1	1	1
SMAD4	17	0.588235294117647	0.965	2
SMAD5	15	0.606060606060606	0.9675	2
SMAD6	12	0.588235294117647	0.965	2
SMAD1	12	0.571428571428571	0.9625	2
ZMYM2	11	0.625	0.97	2
NPM1	9	0.571428571428571	0.9625	2
ADNP	9	0.606060606060606	0.9675	2
YBX1	7	0.555555555555556	0.96	2
SET	6	0.555555555555556	0.96	2

Table 4: Top10 hub genes according to degree

Using the cytoHubba plugin, both the results obtained with the NetworkAnalyzer were verified and the top10 hub genes were visualized (Figure 11).

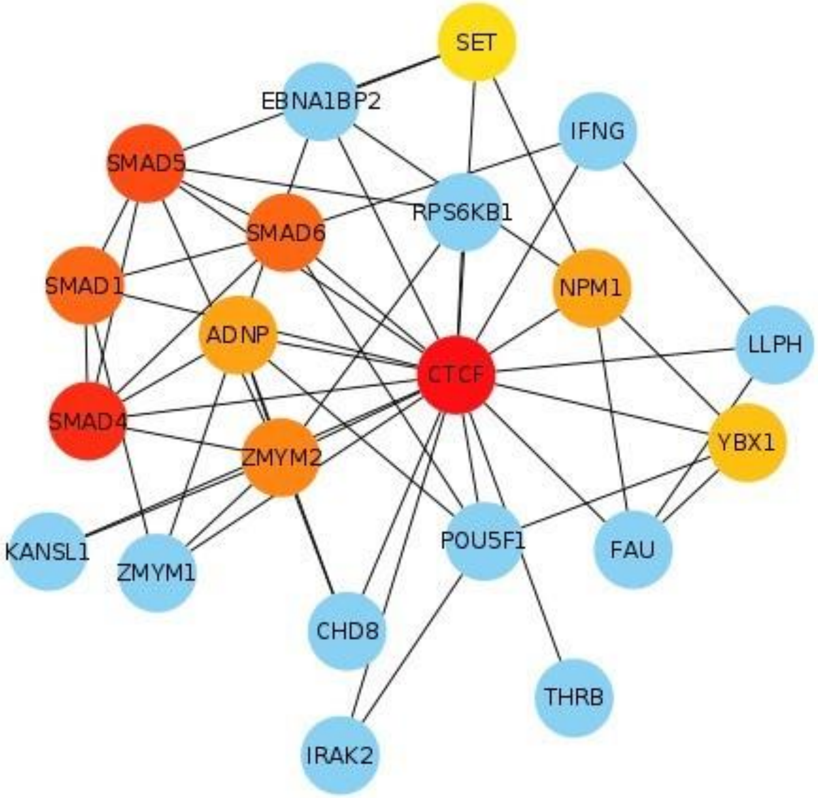
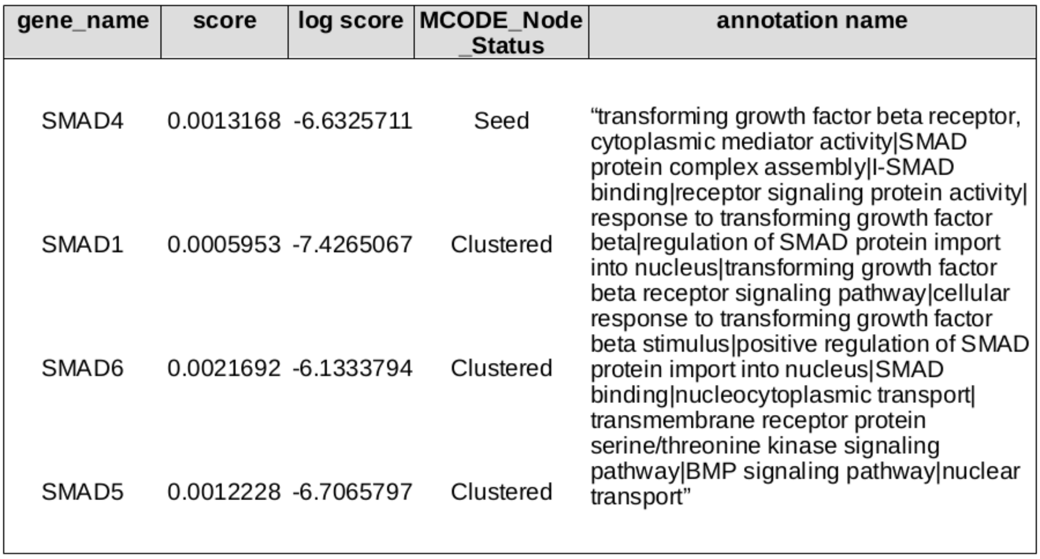


Figure 11: Top10 hub genes colored by degree level

3.1.1.4 Topology-based clustering using MCODE

According to topology results, clusters in this network were detected using the MCODE plug-in. Accordingly, 2 different clusters were obtained in this network.

The first of these, with a score number of 4, contains the following genes, respectively; SMAD4, SMAD1, SMAD6, SMAD5. (Table 5; Figure 12).



gene_name	score	log score	MCODE_Node _Status	annotation name
SMAD4	0.0013168	-6.6325711	Seed	“transforming growth factor beta receptor, cytoplasmic mediator activity|SMAD protein complex assembly|I-SMAD binding|receptor signaling protein activity|response to transforming growth factor beta|regulation of SMAD protein import into nucleus|transforming growth factor beta receptor signaling pathway|cellular response to transforming growth factor beta stimulus|positive regulation of SMAD protein import into nucleus|SMAD binding|nucleocytoplasmic transport|transmembrane receptor protein serine/threonine kinase signaling pathway|BMP signaling pathway|nuclear transport”
SMAD1	0.0005953	-7.4265067	Clustered	
SMAD6	0.0021692	-6.1333794	Clustered	
SMAD5	0.0012228	-6.7065797	Clustered	

Table 5: Cluster1 determined by topology table

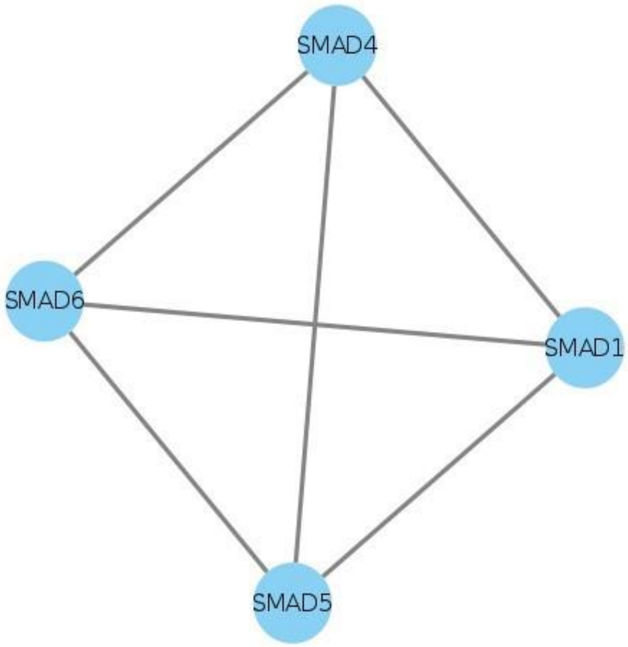


Figure 12: Cluster1 determined by topology figure

The second cluster with a score number of 3.429 contains the following 8 nodes: ADNP, ZMYM1, CTCF, CHD8, SET, EBNA1BP2, YBX1, FAU. (Table 6; Figure 13).

gene name	score	log score	MCODE_Node _Status	annotation name
ADNP	0.00084208	-7.07963477	Clustered	
ZMYM1	0.00088526	-7.02962701	Seed	
CTCF	0.87243231	-0.13647021	Clustered	“chromatin assembly or disassembly| nucleosome organization”
CHD8	0.00322622	-5.7364432	Clustered	
SET	0.00098875	-6.91906442	Clustered	“nucleocytoplasmic transport|chromatin assembly or disassembly|nucleosome organization|nuclear transport”
EBNA1BP2	0.00048159	-7.63842265	Clustered	
YBX1	0.00088386	-7.03120752	Clustered	“posttranscriptional regulation of gene expression”
FAU	0.00075337	-7.19096025	Clustered	

Table 6: Cluster2 determined by topology table

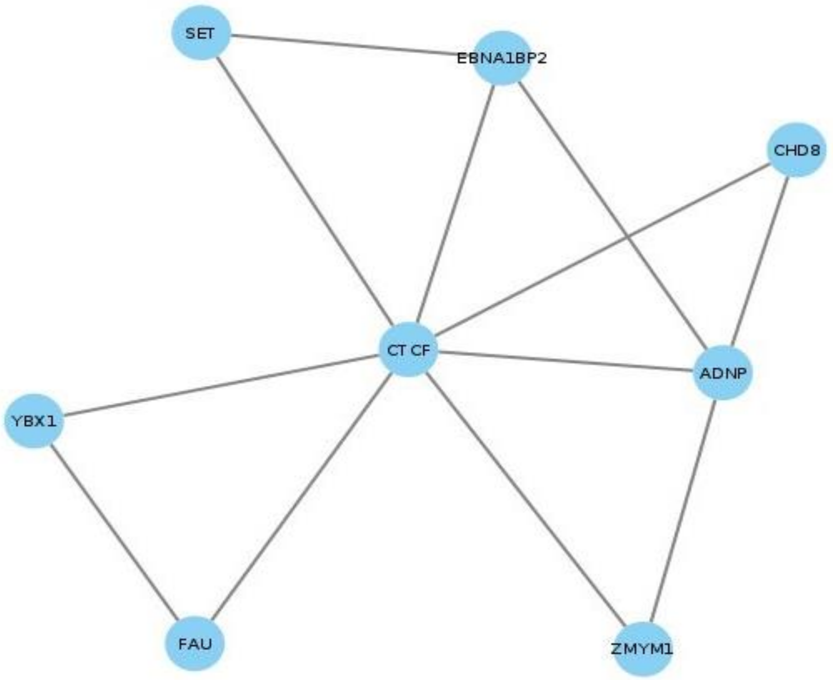
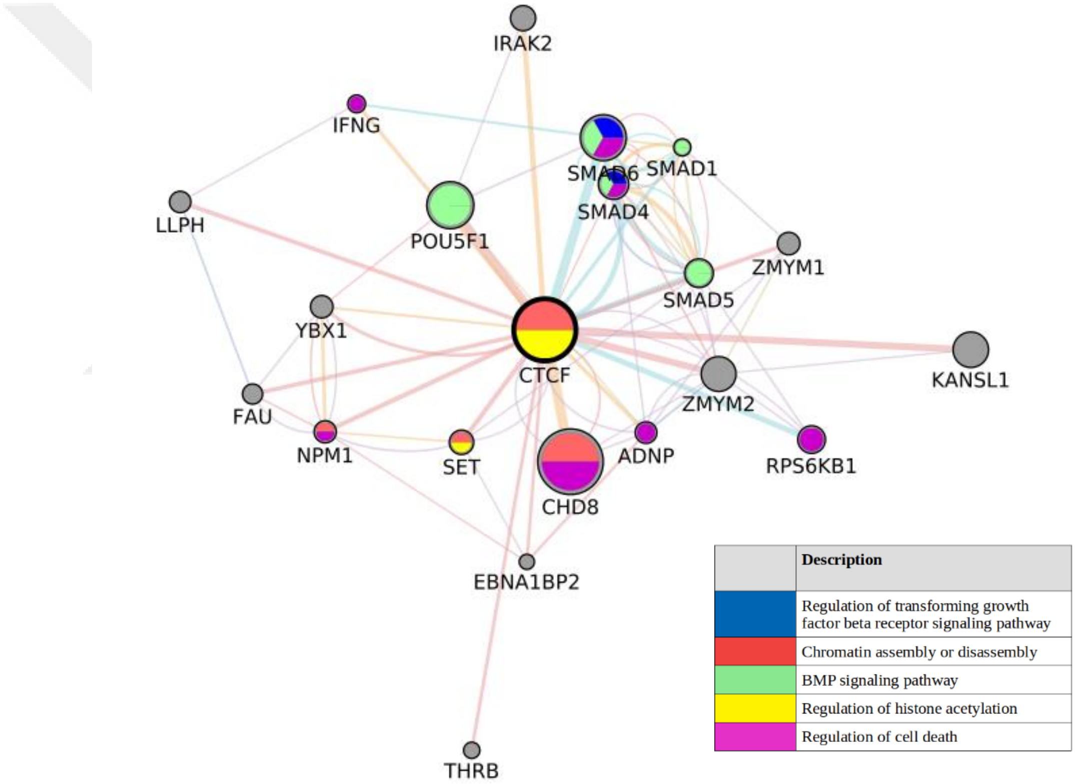


Figure 13: Cluster2 determined by topology figure

3.1.1.5 Gene Ontology overrepresentation analysis using GOlorize

Over-represented GO categories in this network were found using the GOlorize plug-in. In this plug-in as a first step, the most represented GO categories are determined by using the BINGO plug-in and additionally, the GOlorize plug-in provides coloring on the network.

Overrepresentation analysis, based on the GO-BP (biological process) database, identified 463 enriched terms (FDR<0.05). 5 terms were selected (Figure 14) from this list and mapped on the network (GOlorozing the nodes) with the purpose of functional annotation. Selection was performed using the following criteria: some of these are the same or related to the categories obtained with the MCODE, since the focus of the thesis is cancer, categories related to cancer, categories related to regulation.



	Description
🟦	Regulation of transforming growth factor beta receptor signaling pathway
🟥	Chromatin assembly or disassembly
🟩	BMP signaling pathway
🟨	Regulation of histone acetylation
🟪	Regulation of cell death

Figure 14: 5 overrepresented GO categories by BP (Biological Process)

Then, Overrepresented 5 GO-MF (molecular function) categories were selected among the 58 overrepresented terms, since they provide a comprehensive overview of the CTCF in terms of molecular function (Figure 15).

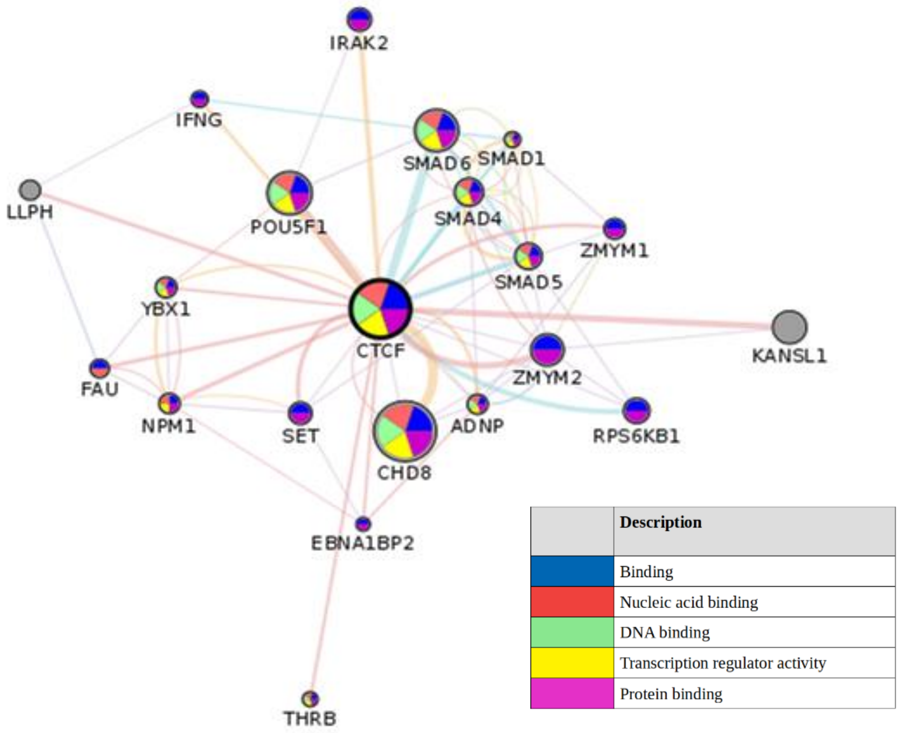


Figure 15: 5 overrepresented GO categories by MF (Molecular Function)

3.1.1.6 Gene Ontology and Reactome pathway overrepresentation analysis using ClueGO

The ClueGO plug-in was used to highlight specific biological roles and common approaches by reducing redundant repetitive results in the detailed CTCF network, including the GO annotations and clusters obtained by MCODE and GOlorize. The ClueGO network created with Kappa statistics uses ontology sources such as Gene Ontology, Reactome, KEGG, Wikipathways to find clusters in the input gene list and reveals the most important common biological roles by fusing similar results in selected categories.

Using this plug-in, we have demonstrated the significant biological roles of CTCF on a single network, using the GO Biological Process and Molecular Function categories, which we previously created with two different analyzes with GOlorize (Figure 16). Gene list obtained from GeneMania was used as input to obtain this network.

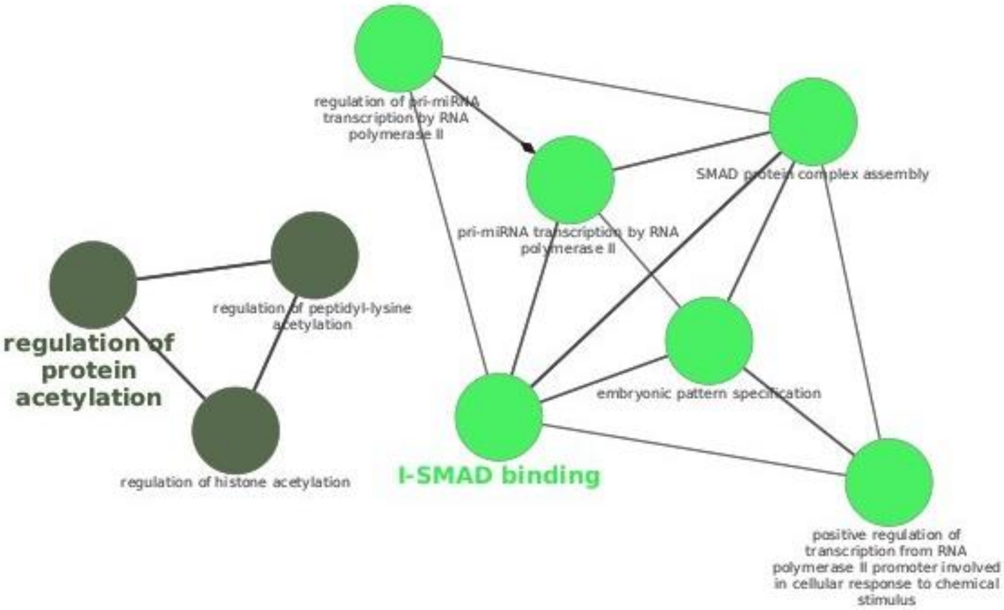
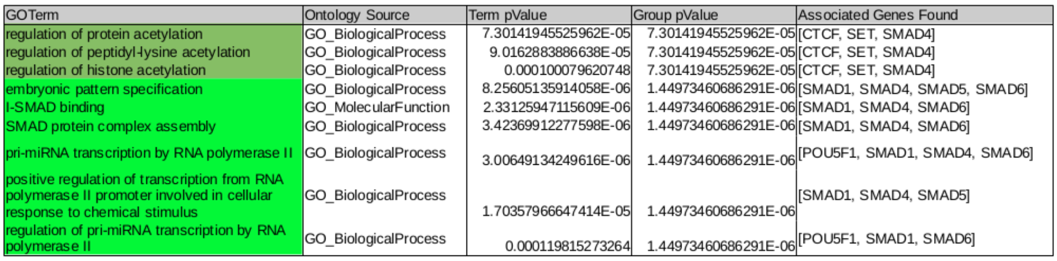

regulation of pri-miRNA transcription by RNA polymerase II
SMAD protein complex assembly
pri-miRNA transcription by RNA polymerase II
regulation of peptidyl-lysine acetylation
regulation of protein acetylation
embryonic pattern specification
regulation of histone acetylation
I-SMAD binding
positive regulation of transcription from RNA polymerase II promoter involved in cellular response to chemical stimulus

Figure 16: Network created by using Gene Ontology Biological Process and Molecular Function options with ClueGO plugin. The most specific terms in these groups was considered as the group leader and highlighted in larger and bold font.

As seen in this network, two functional groups/pathways are shown with two different colors. The table below (Table 7) contains detailed information about these functional groups, such as the detailed gene information, which biological function they roled in, and their p-values.



GOTerm	Ontology Source	Term pValue	Group pValue	Associated Genes Found
regulation of protein acetylation	GO_BiologicalProcess	7.30141945525962E-05	7.30141945525962E-05	[CTCF, SET, SMAD4]
regulation of peptidyl-lysine acetylation	GO_BiologicalProcess	9.0162883886638E-05	7.30141945525962E-05	[CTCF, SET, SMAD4]
regulation of histone acetylation	GO_BiologicalProcess	0.000100079620748	7.30141945525962E-05	[CTCF, SET, SMAD4]
embryonic pattern specification	GO_BiologicalProcess	8.25605135914058E-06	1.44973460686291E-06	[SMAD1, SMAD4, SMAD5, SMAD6]
I-SMAD binding	GO_MolecularFunction	2.33125947115609E-06	1.44973460686291E-06	[SMAD1, SMAD4, SMAD6]
SMAD protein complex assembly	GO_BiologicalProcess	3.42369912277598E-06	1.44973460686291E-06	[SMAD1, SMAD4, SMAD6]
pri-miRNA transcription by RNA polymerase II	GO_BiologicalProcess	3.00649134249616E-06	1.44973460686291E-06	[POU5F1, SMAD1, SMAD4, SMAD6]
positive regulation of transcription from RNA polymerase II promoter involved in cellular response to chemical stimulus	GO_BiologicalProcess	1.70357966647414E-05	1.44973460686291E-06	[SMAD1, SMAD4, SMAD5]
regulation of pri-miRNA transcription by RNA polymerase II	GO_BiologicalProcess	0.000119815273264	1.44973460686291E-06	[POU5F1, SMAD1, SMAD6]

Table 7: Detailed information of the network created using GO-BP and MF ontology sources from ClueGO plugin

In addition, the obtained functional terms are shown as bar charts and the obtained groups as pie charts (Figure 17). On the bar chart, terms that belong to more than one group are marked with an "*" and how many number of times they were repeated is also shown as numbers at the end of the bars. The pie chart shows how many percent of terms each group contains.

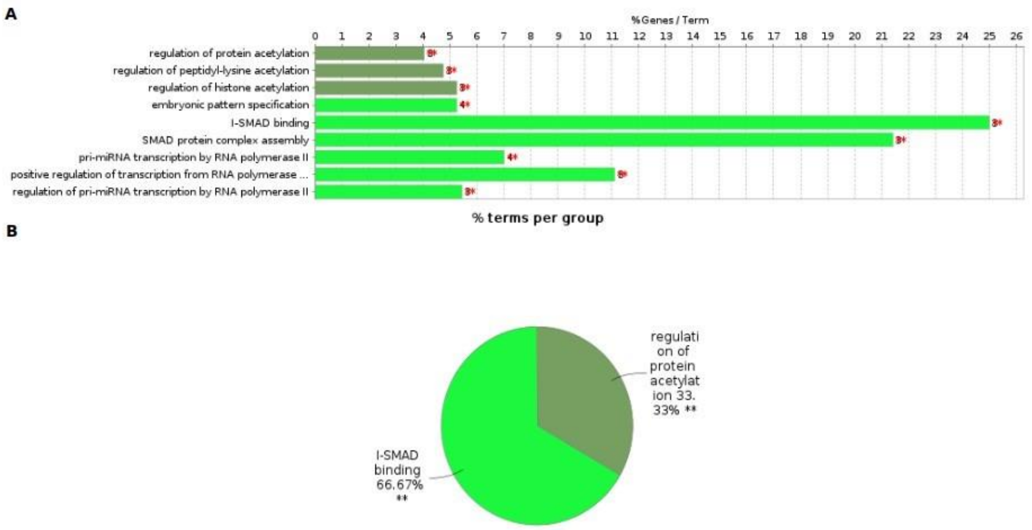


Figure 17: A. Significant Reactome pathway and reactions terms identified with the Kappa score B. Percent number of terms of functional pathway

With the ClueGO plugin, a network was created using Reactome Reactions and Reactome pathway options, a database that is frequently used in analyzing and displaying biological pathways (Figure 18).

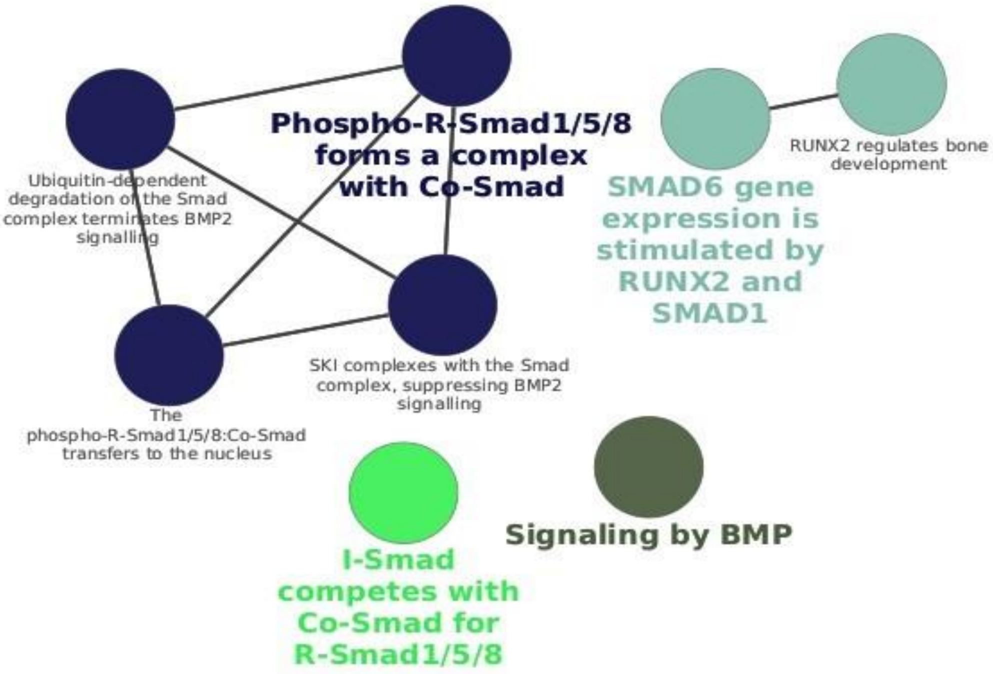
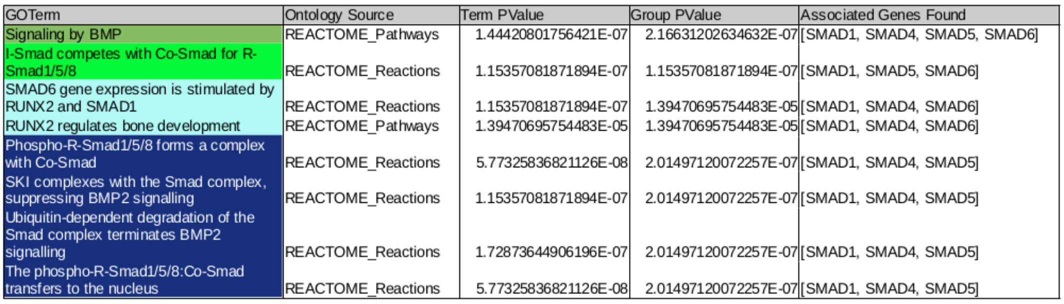

Phospho-R-Smad1/5/8 forms a complex with Co-Smad
Ubiquitin-dependent degradation of the Smad complex terminates BMP2 signalling
SKI complexes with the Smad complex, suppressing BMP2 signalling
The phospho-R-Smad1/5/8:Co-Smad transfers to the nucleus
RUNX2 regulates bone development
SMAD6 gene expression is stimulated by RUNX2 and SMAD1
Signaling by BMP
I-Smad competes with Co-Smad for R-Smad1/5/8

Figure 18: Network created using Reactome Reactions and Reactome Pathway options

As seen in this network, different pathways and reaction results are displayed in 4 different colors. Similar results are combined as in the network created using Gene Ontology options, and the most important biological pathways are shown in highlighted and bold font as the group leader. The table below (Table 8) contains detailed information about these biological pathway, such as the detailed gene information, which pathway and reactions they belong to, and their p-values.



GOTerm	Ontology Source	Term PValue	Group PValue	Associated Genes Found
Signaling by BMP	REACTOME_Pathways	1.44420801756421E-07	2.16631202634632E-07	[SMAD1, SMAD4, SMAD5, SMAD6]
I-Smad competes with Co-Smad for R-Smad1/5/8	REACTOME_Reactions	1.15357081871894E-07	1.15357081871894E-07	[SMAD1, SMAD5, SMAD6]
SMAD6 gene expression is stimulated by RUNX2 and SMAD1	REACTOME_Reactions	1.15357081871894E-07	1.39470695754483E-05	[SMAD1, SMAD4, SMAD6]
RUNX2 regulates bone development	REACTOME_Pathways	1.39470695754483E-05	1.39470695754483E-05	[SMAD1, SMAD4, SMAD6]
Phospho-R-Smad1/5/8 forms a complex with Co-Smad	REACTOME_Reactions	5.77325836821126E-08	2.01497120072257E-07	[SMAD1, SMAD4, SMAD5]
SKI complexes with the Smad complex, suppressing BMP2 signalling	REACTOME_Reactions	1.15357081871894E-07	2.01497120072257E-07	[SMAD1, SMAD4, SMAD5]
Ubiquitin-dependent degradation of the Smad complex terminates BMP2 signalling	REACTOME_Reactions	1.72873644906196E-07	2.01497120072257E-07	[SMAD1, SMAD4, SMAD5]
The phospho-R-Smad1/5/8:Co-Smad transfers to the nucleus	REACTOME_Reactions	5.77325836821126E-08	2.01497120072257E-07	[SMAD1, SMAD4, SMAD5]

Table 8: Detailed information of the network created using Reactome reactions and Reactome Pathway from ClueGO plugin

Similar to the network prepared for gene ontology, the results are also shown as a bar chart showing the terms of Reactome pathway and reactions (Figure 19), and a pie chart showing how many percent of terms the groups contain. On the bar chart, terms that belong to more than one group are marked with an “*” and how many number of times they were repeated is also shown as numbers at the end of the bars.

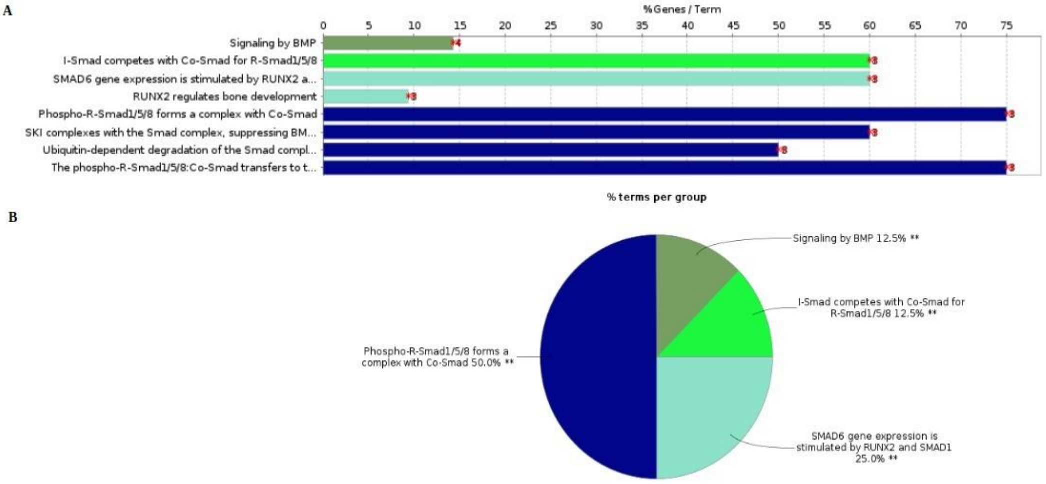
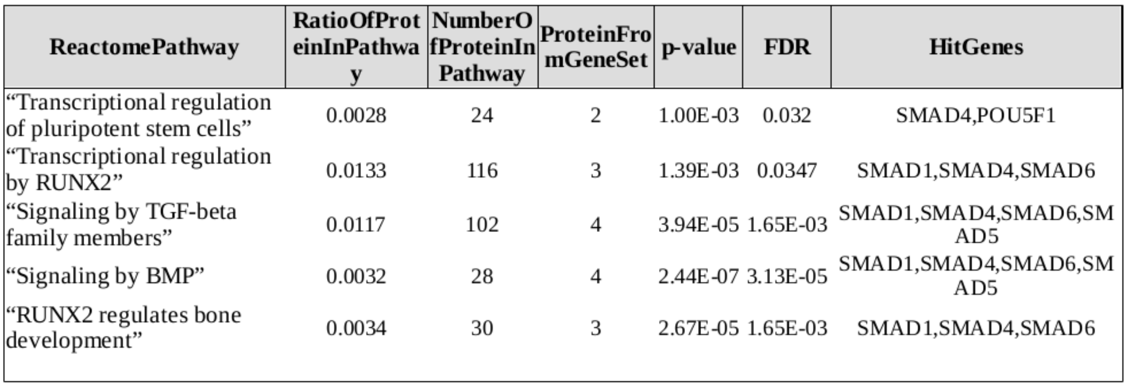


Figure 19: A. Significant Reactome pathway and reactions terms identified with the Kappa score B. Percent number of terms of functional pathways

3.1.1.7 Reactome pathway enrichment analysis using ReactomeFIPlugIn

Reactome pathway enrichment was done using ReactomeFI plugin. In total, 130 different pathways and hit genes included in them were obtained. Moreover, the following details for each term were obtained: “ratio of protein in pathway”, “number of protein in pathway”, “protein from gene set”, “p-value” and “FDR (false discovery rate) values”. The latter is a related with p value adjustment for multiple testing. The significant results (terms) were singled out using FDR 0.05 threshold. As a result, 5 significant pathways were obtained and are shown in the table 9.



ReactomePathway	RatioOfProteinInPathway	NumberOfProteinInPathway	ProteinFromGeneSet	p-value	FDR	HitGenes
“Transcriptional regulation of pluripotent stem cells”	0.0028	24	2	1.00E-03	0.032	SMAD4,POU5F1
“Transcriptional regulation by RUNX2”	0.0133	116	3	1.39E-03	0.0347	SMAD1,SMAD4,SMAD6
“Signaling by TGF-beta family members”	0.0117	102	4	3.94E-05	1.65E-03	SMAD1,SMAD4,SMAD6,SMAD5
“Signaling by BMP”	0.0032	28	4	2.44E-07	3.13E-05	SMAD1,SMAD4,SMAD6,SMAD5
“RUNX2 regulates bone development”	0.0034	30	3	2.67E-05	1.65E-03	SMAD1,SMAD4,SMAD6

Table 9: FDR <0.05 Reactome pathways prepared by using ReactomeFI plugin

By using the option of this plugin that enables to visualize the pathways, the hit pathways were displayed. Accordingly, the diagram of the main metabolic pathways was created. To obtain information about these pathways containing detailed gene clusters, they need to be examined on the Reactome web page. Below are the main diagrams that include these pathways. The figures of the main diagrams of these pathways are shown in the order below.

Firstly, the diagram of the developmental biology pathway, which is the main title of the “transcriptional regulation of pluripotent stem cells pathway” was created (Figure 20). By looking at this, the sub pathways under the developmental biology pathway can also be examined.

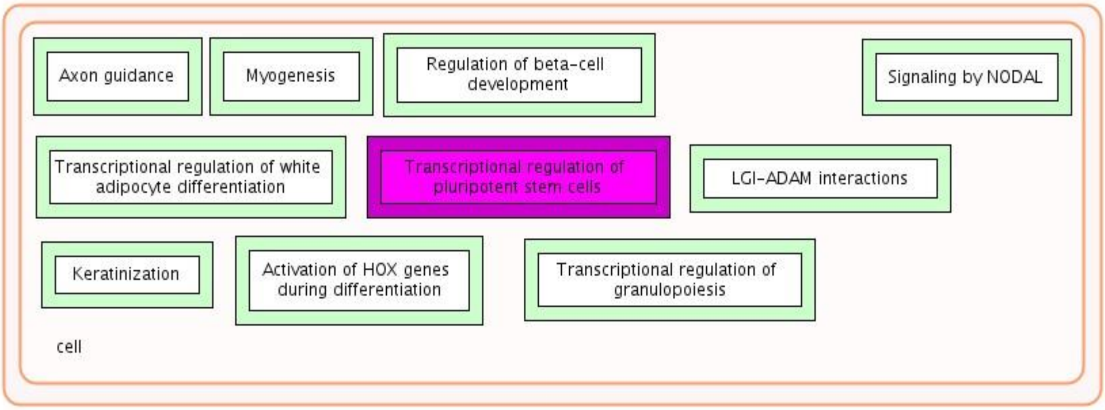

Axon guidance
Myogenesis
Regulation of beta-cell development
Signaling by NODAL
Transcriptional regulation of white adipocyte differentiation
Transcriptional regulation of pluripotent stem cells
LGI-ADAM interactions
Keratinization
Activation of HOX genes during differentiation
Transcriptional regulation of granulopoiesis
cell

Figure 20: Transcriptional regulation of pluripotent stem cells, the sub pathway under the developmental biology main pathway

The transcriptional regulation by RUNX2 pathway shown in the second row in the table is examined under the gene expression main diagram. When a detailed pathway examination is performed from the Reactome web page, the metabolic pathways in which SMAD1, SMAD4, SMAD6 genes take part appear in detail. Also similarly, other pathways under the main title of gene expression can be examined from this figure (Figure 21).

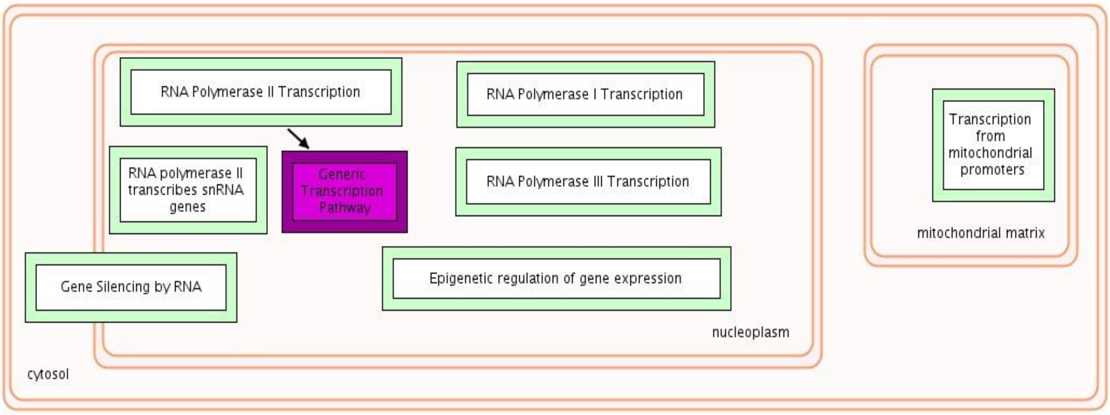

RNA Polymerase II Transcription
RNA Polymerase I Transcription
Transcription from mitochondrial promoters
RNA polymerase II transcribes snRNA genes
Generic Transcription Pathway
RNA Polymerase III Transcription
mitochondrial matrix
Gene Silencing by RNA
Epigenetic regulation of gene expression
nucleoplasm
cytosol

Figure 21: Transcriptional regulation by RUNX2, the sub pathway under the gene expression main pathway

Similarly, the RUNX2 regulates bone development pathway, shown in the 5th row in the table, also appears to be under the gene expression main diagram. The role of genes from the SMAD family in this pathway can be examined in detail from the Reactome Pathway Browser web page.

Signaling by TGF-beta family members shown in the 3rd row in the table are examined under the main title of pathway signal transduction. Figure showing detailed functions of genes included in this pathway is obtained (Figure 22).

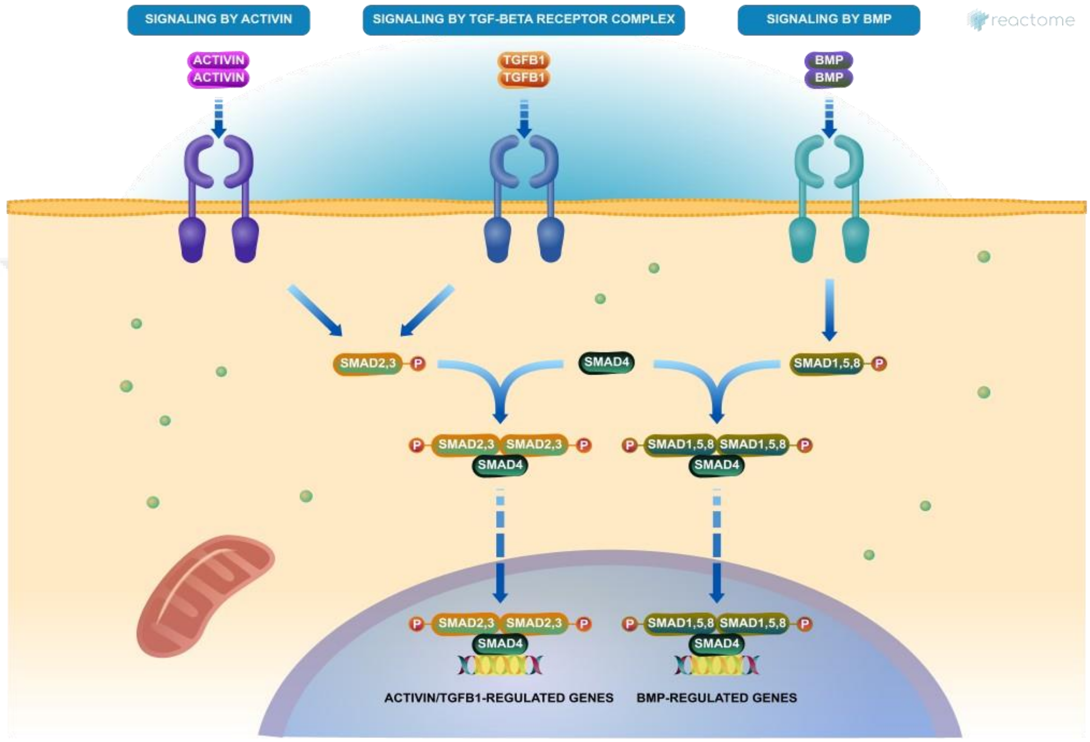


Figure 22: “Signaling by TGF-beta family members pathway” from Reactome Pathway Browser

The roles of SMAD genes in signal transduction can be examined in this figure. Detailed pathway information of Signaling by BMP, which is in the 4th row in the table, is also shown in this figure.

In addition, the detailed diagram of the Signaling by activin pathway has been obtained directly from the ReactomeFI plug-in (Figure 23).

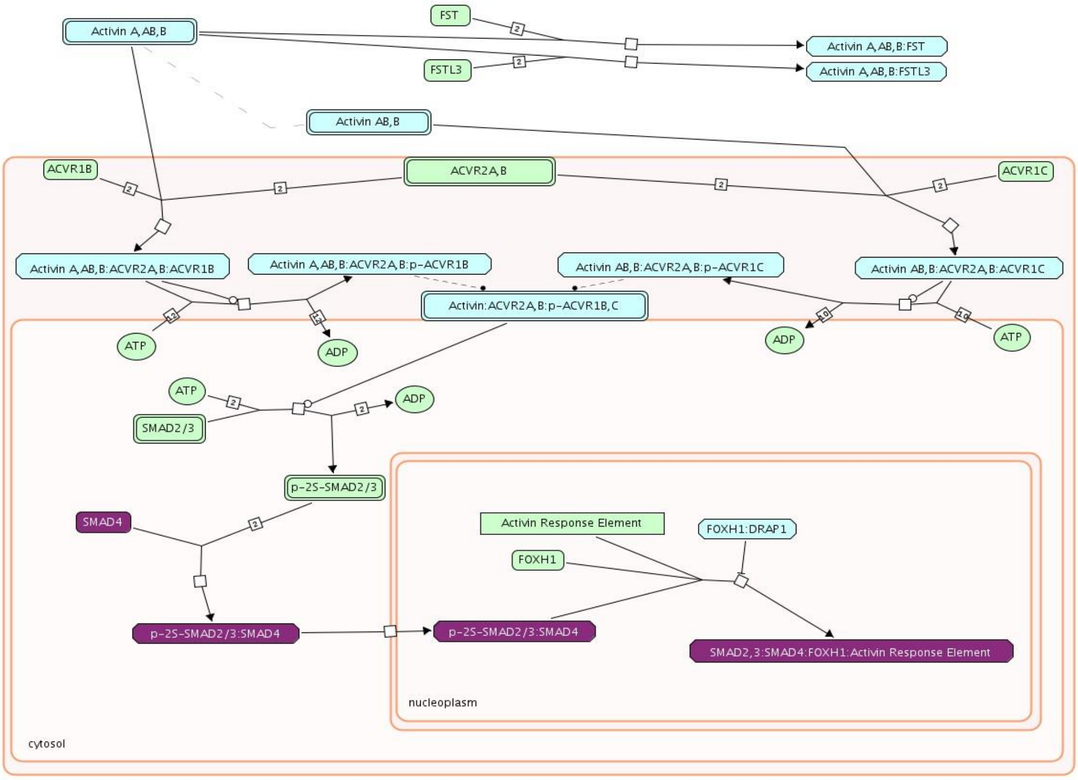


Figure 23: Signaling by activin diagram created using ReactomeFI plugin

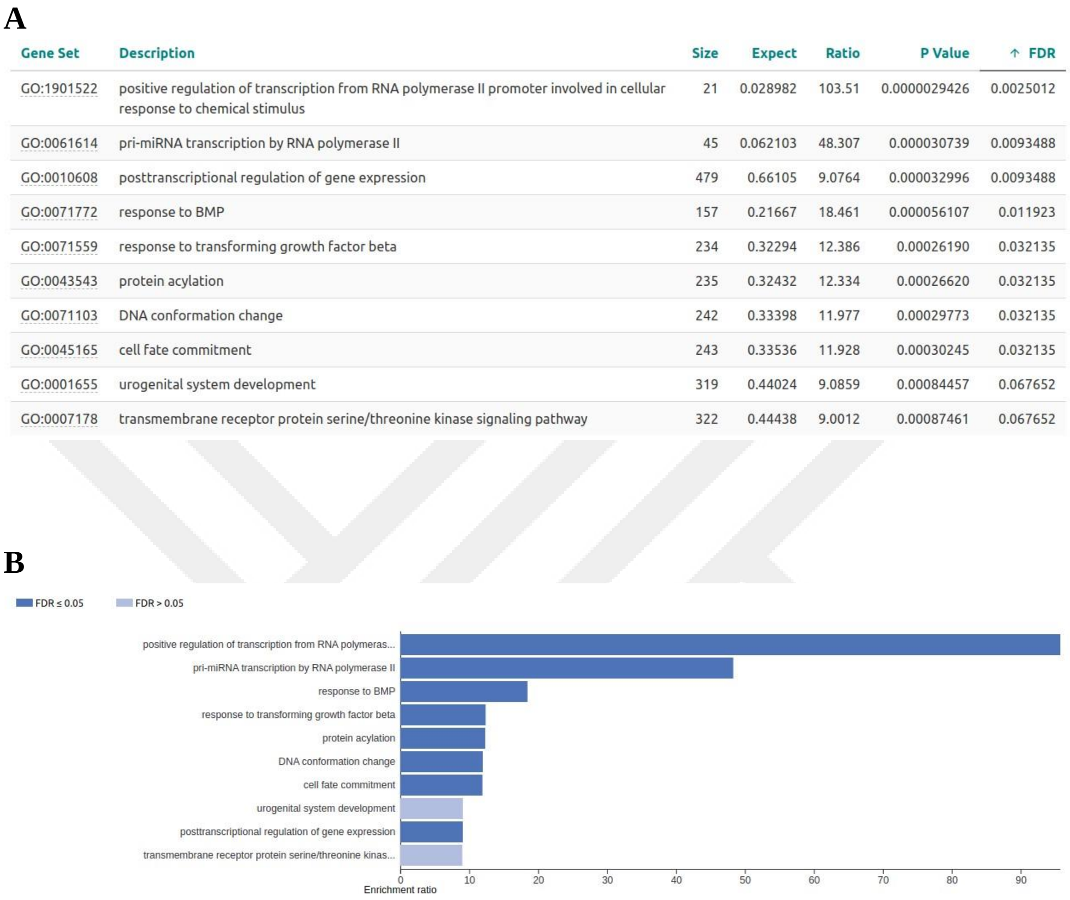
Among the genes in the CTCF network, those involved in the Signalization pathway by activin are highlighted in purple in the figure above.

3.1.2 Complementary (Online Tools-based) network analysis

The CTCF network, which is functionally annotated with Cytoscape plugins, was analyzed with frequently used web tools for this purpose, thus ensuring the comparison of the accuracy of the results and the completion of the missing parts.

3.1.2.1 Gene Ontology and pathway overrepresentation analysis using WebGestalt

First of all, over representation analysis was performed by using Gene ontology database in WEbGestalt online tool. The noRedundant option was used to reduce redundant results as used in the ClueGO plugin. First by using the Biological Process noRedundat option the table containing the p value and FDR values of each functional annotation and the bar graph showing the FDR significance value of the enriched gene sets are created (Figure 24).

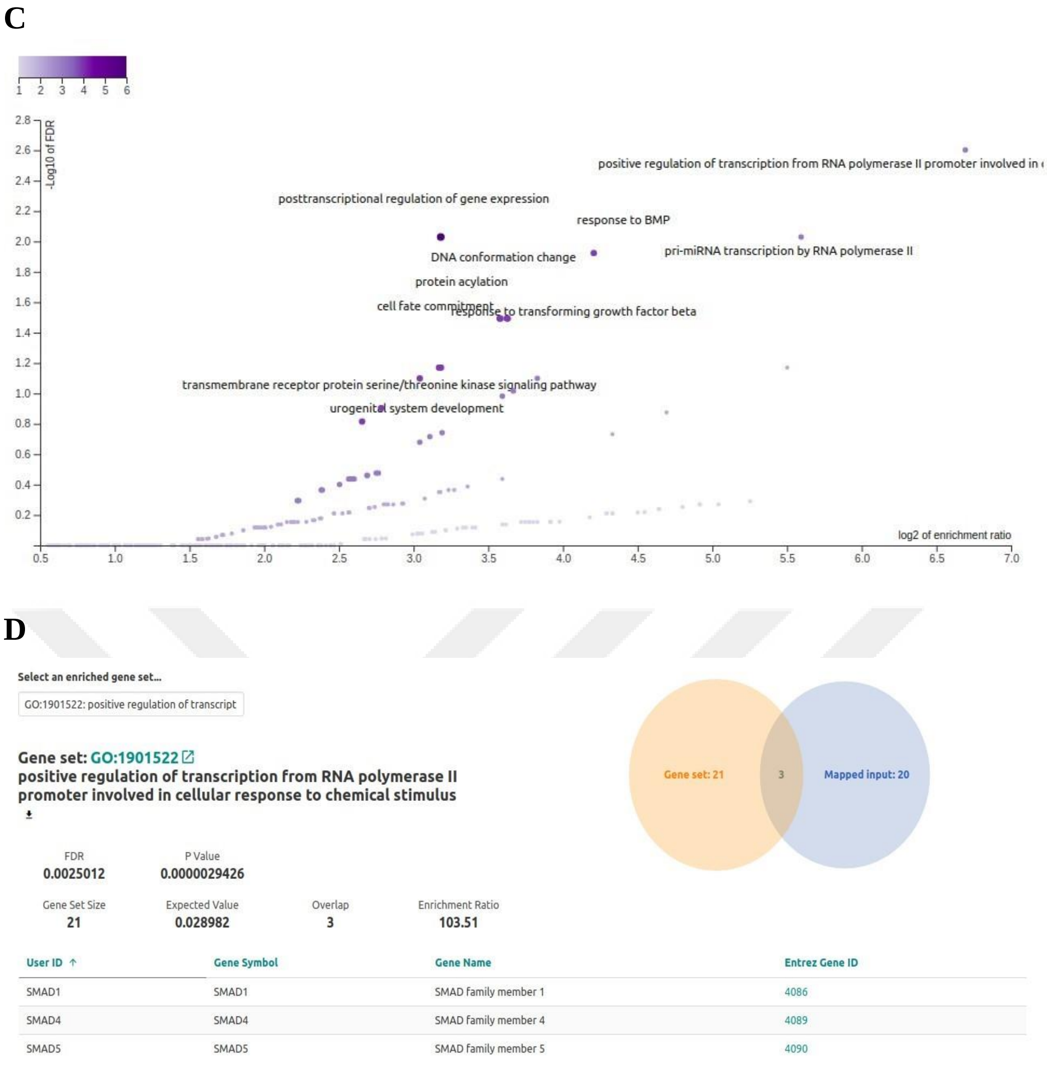


Gene Set	Description	Size	Expect	Ratio	P Value	↑ FDR
GO:1901522	positive regulation of transcription from RNA polymerase II promoter involved in cellular response to chemical stimulus	21	0.028982	103.51	0.0000029426	0.0025012
GO:0061614	pri-miRNA transcription by RNA polymerase II	45	0.062103	48.307	0.000030739	0.0093488
GO:0010608	posttranscriptional regulation of gene expression	479	0.66105	9.0764	0.000032996	0.0093488
GO:0071772	response to BMP	157	0.21667	18.461	0.000056107	0.011923
GO:0071559	response to transforming growth factor beta	234	0.32294	12.386	0.00026190	0.032135
GO:0043543	protein acylation	235	0.32432	12.334	0.00026620	0.032135
GO:0071103	DNA conformation change	242	0.33398	11.977	0.00029773	0.032135
GO:0045165	cell fate commitment	243	0.33536	11.928	0.00030245	0.032135
GO:0001655	urogenital system development	319	0.44024	9.0859	0.00084457	0.067652
GO:0007178	transmembrane receptor protein serine/threonine kinase signaling pathway	322	0.44438	9.0012	0.00087461	0.067652

A
B
FDR ≤ 0.05
FDR > 0.05
positive regulation of transcription from RNA polymeras...
pri-miRNA transcription by RNA polymerase II
response to BMP
response to transforming growth factor beta
protein acylation
DNA conformation change
cell fate commitment
urogenital system development
posttranscriptional regulation of gene expression
transmembrane receptor protein serine/threonine kinas...
0
10
20
30
40
50
60
70
80
90
Enrichment ratio

Figure 24: A. GO Biological process reduncany functional annotation containing reduced p value and FDR values B. GO Biological process reduncany reduced, enriched gene sets bar chart sorted by FDR values

Similarly, functional annotation results based on FDR values were displayed on a Volcano plot. In addition, the summary table containing the GO name of the enriched annotation result with the highest significance and the genes involved in this role in the network is shown (Figure 25). This tool includes an interface suitable for selecting the gene sets whose details are to be viewed.



Select an enriched gene set...

GO:1901522: positive regulation of transcript

Gene set: GO:1901522
positive regulation of transcription from RNA polymerase II promoter involved in cellular response to chemical stimulus

FDR	P Value		
0.0025012	0.0000029426		

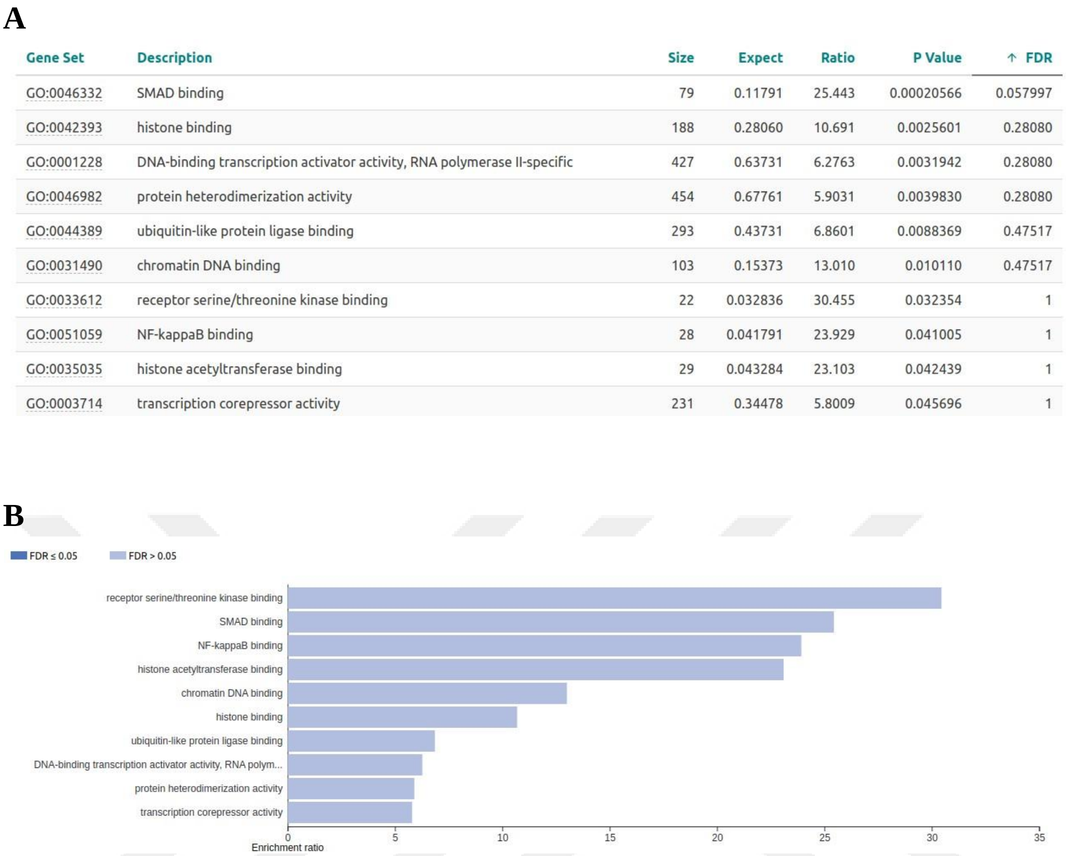
Gene Set Size	Expected Value	Overlap	Enrichment Ratio
21	0.028982	3	103.51

User ID ↑	Gene Symbol	Gene Name	Entrez Gene ID
SMAD1	SMAD1	SMAD family member 1	4086
SMAD4	SMAD4	SMAD family member 4	4089
SMAD5	SMAD5	SMAD family member 5	4090

Figure 25: C. Visualization of functional annotation results of GO-BP on Volcano plot D. Detailed information of the most significant enriched GO set

When these results were compared with the results obtained with ClueGO, the following GO categories appeared to be the same for the GO-BP database: "Positive regulation of transcription from RNA Polymerase II promoter involved in cellular response to chemical stimulus”, “Pri mRNA transcription by RNA Polymerase II”, “Regulation of protein acetylation”.

Second, by using the GO Molecular function noRedundant option the table containing the p value and FDR values of each functional annotation and the bar graph showing the FDR significance value of the enriched gene sets are created (Figure 26).

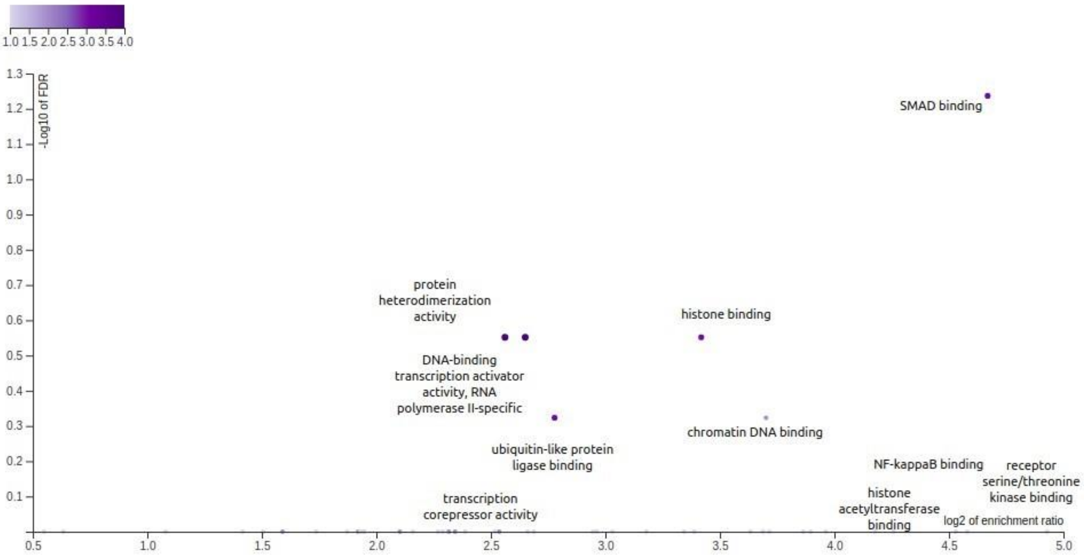


Gene Set	Description	Size	Expect	Ratio	P Value	↑ FDR
GO:0046332	SMAD binding	79	0.11791	25.443	0.00020566	0.057997
GO:0042393	histone binding	188	0.28060	10.691	0.0025601	0.28080
GO:0001228	DNA-binding transcription activator activity, RNA polymerase II-specific	427	0.63731	6.2763	0.0031942	0.28080
GO:0046982	protein heterodimerization activity	454	0.67761	5.9031	0.0039830	0.28080
GO:0044389	ubiquitin-like protein ligase binding	293	0.43731	6.8601	0.0088369	0.47517
GO:0031490	chromatin DNA binding	103	0.15373	13.010	0.010110	0.47517
GO:0033612	receptor serine/threonine kinase binding	22	0.032836	30.455	0.032354	1
GO:0051059	NF-kappaB binding	28	0.041791	23.929	0.041005	1
GO:0035035	histone acetyltransferase binding	29	0.043284	23.103	0.042439	1
GO:0003714	transcription corepressor activity	231	0.34478	5.8009	0.045696	1

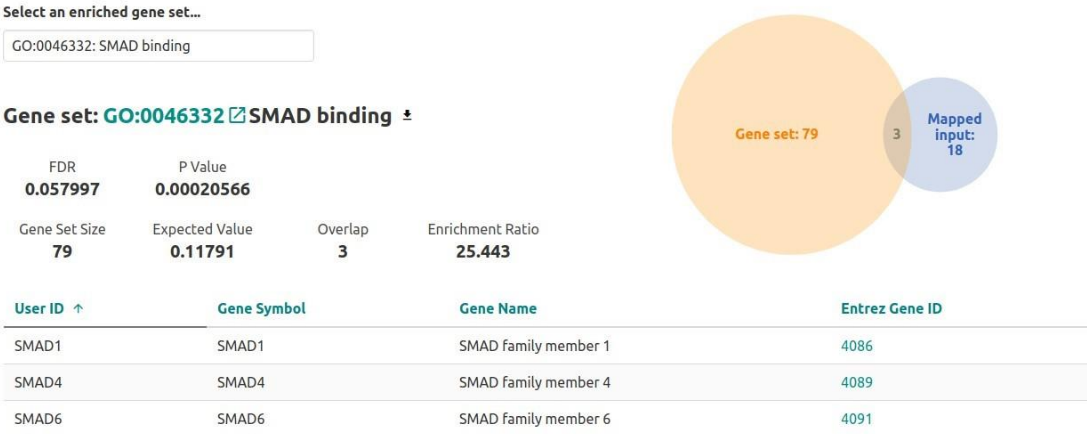
Figure 26: A. GO Molecular function redundancy functional annotation containing reduced p value and FDR values B. GO Molecular function redundancy reduced, enriched gene sets bar chart sorted by FDR values

As performed in GO-MF analysis, functional annotation results based on FDR values were displayed on a Volcano plot. In addition, the summary table containing the GO name of the enriched annotation result with the highest significance and the genes involved in this role in the network is shown (Figure 27).

C



D


Select an enriched gene set...

GO:0046332: SMAD binding

Gene set: GO:0046332 SMAD binding

FDR	P Value
0.057997	0.00020566

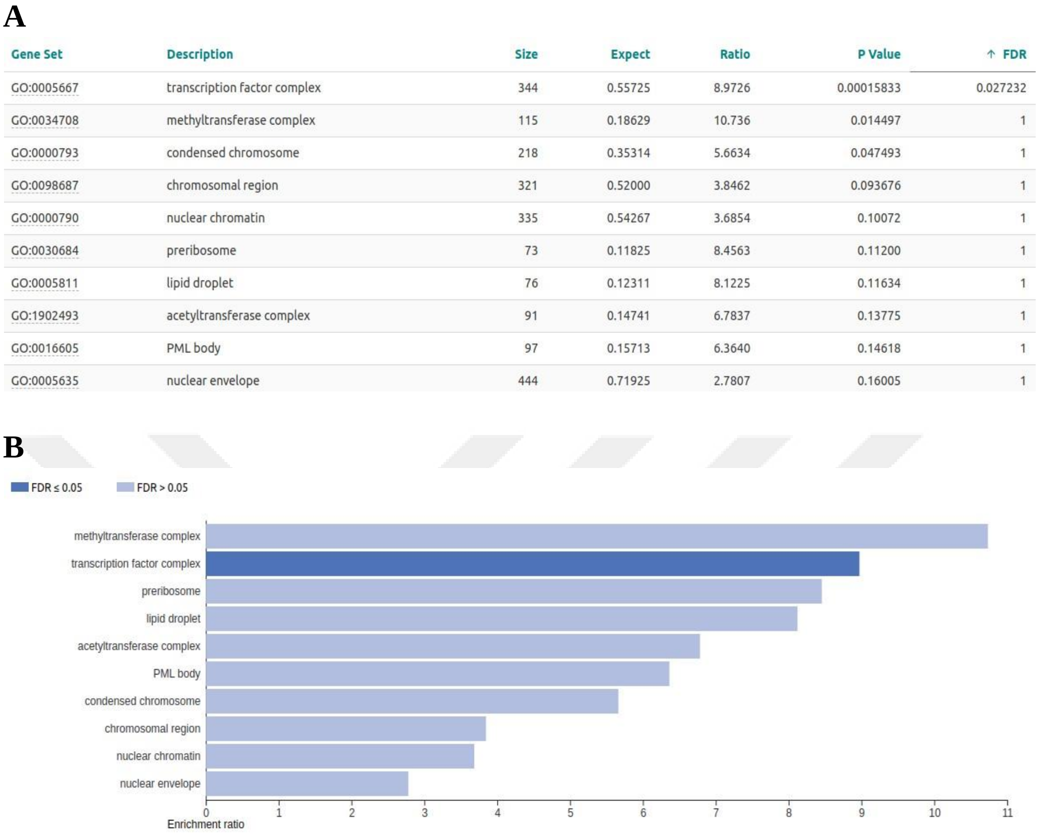
Gene Set Size	Expected Value	Overlap	Enrichment Ratio
79	0.11791	3	25.443

User ID ↑	Gene Symbol	Gene Name	Entrez Gene ID
SMAD1	SMAD1	SMAD family member 1	4086
SMAD4	SMAD4	SMAD family member 4	4089
SMAD6	SMAD6	SMAD family member 6	4091

Figure 27: C. Visualization of functional annotation results of GO-MF on Volcano plot D. Detailed information of the most significant enriched GO set

When these results were compared with the results obtained with ClueGO, the following GO set appeared to be the same for the GO-MF database: "SMAD binding”. It is also understood from the relevant figures that this is the most significant gene set for GO-MF.

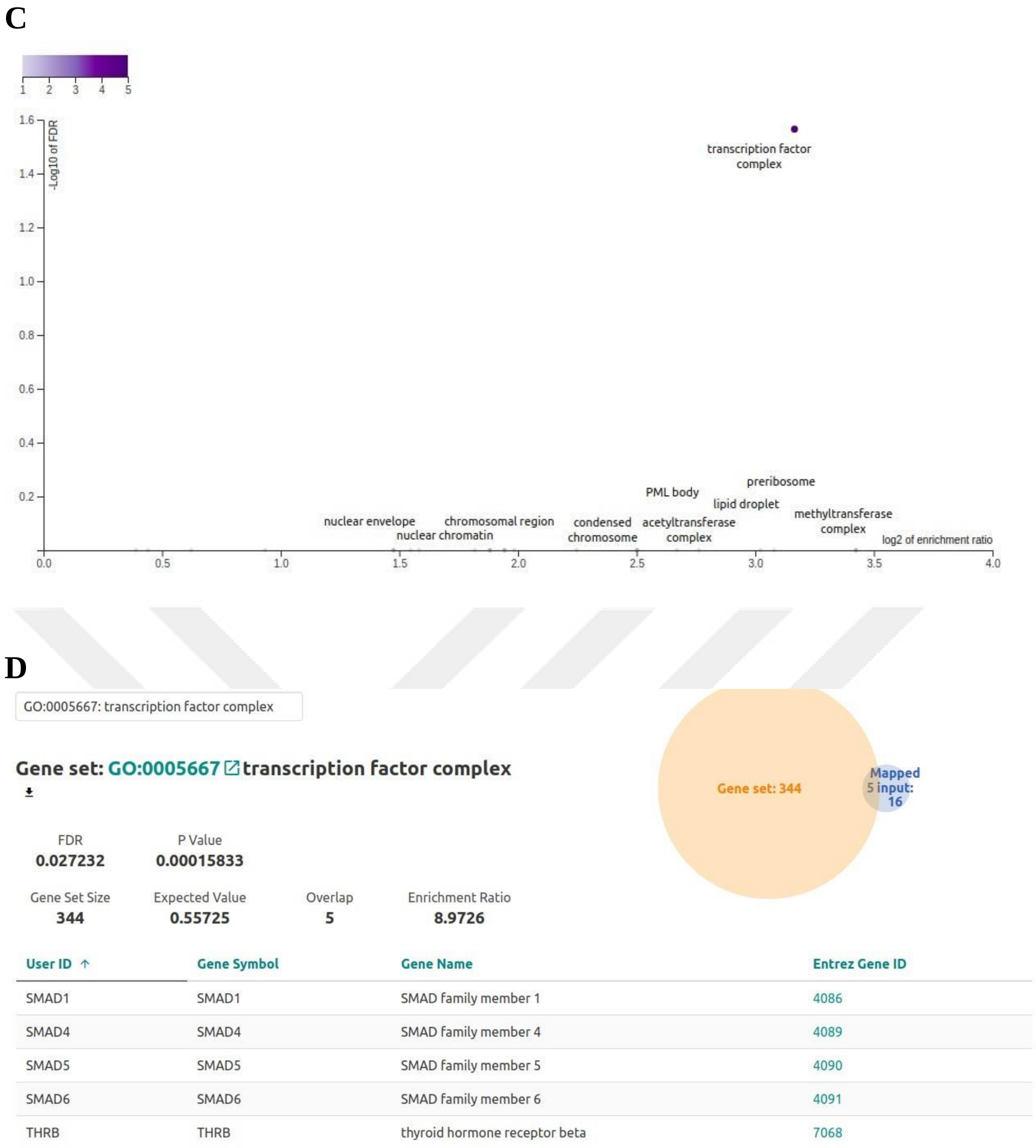
As a last analysis generated using the GO database, by using the GO Cellular components noRedundant option the table containing the p value and FDR values of each functional annotation and the bar graph showing the FDR significance value of the enriched gene sets are created (Figure 28).



Gene Set	Description	Size	Expect	Ratio	P Value	↑ FDR
GO:0005667	transcription factor complex	344	0.55725	8.9726	0.00015833	0.027232
GO:0034708	methyltransferase complex	115	0.18629	10.736	0.014497	1
GO:0000793	condensed chromosome	218	0.35314	5.6634	0.047493	1
GO:0098687	chromosomal region	321	0.52000	3.8462	0.093676	1
GO:0000790	nuclear chromatin	335	0.54267	3.6854	0.10072	1
GO:0030684	preribosome	73	0.11825	8.4563	0.11200	1
GO:0005811	lipid droplet	76	0.12311	8.1225	0.11634	1
GO:1902493	acetyltransferase complex	91	0.14741	6.7837	0.13775	1
GO:0016605	PML body	97	0.15713	6.3640	0.14618	1
GO:0005635	nuclear envelope	444	0.71925	2.7807	0.16005	1

Figure 28: A. GO Cellular component redundancy functional annotation containing reduced p value and FDR values B. GO Cellular component redundancy reduced, enriched gene sets bar chart sorted by FDR values

As performed in GO-CC analysis, functional annotation results based on FDR values were displayed on a Volcano plot (Figure29-C). In addition, the summary table containing the GO name of the enriched annotation result with the highest significance and the genes involved in this role in the network is shown (Figure29-D).



GO:0005667: transcription factor complex

Gene set: GO:0005667 transcription factor complex

FDR	P Value
0.027232	0.00015833

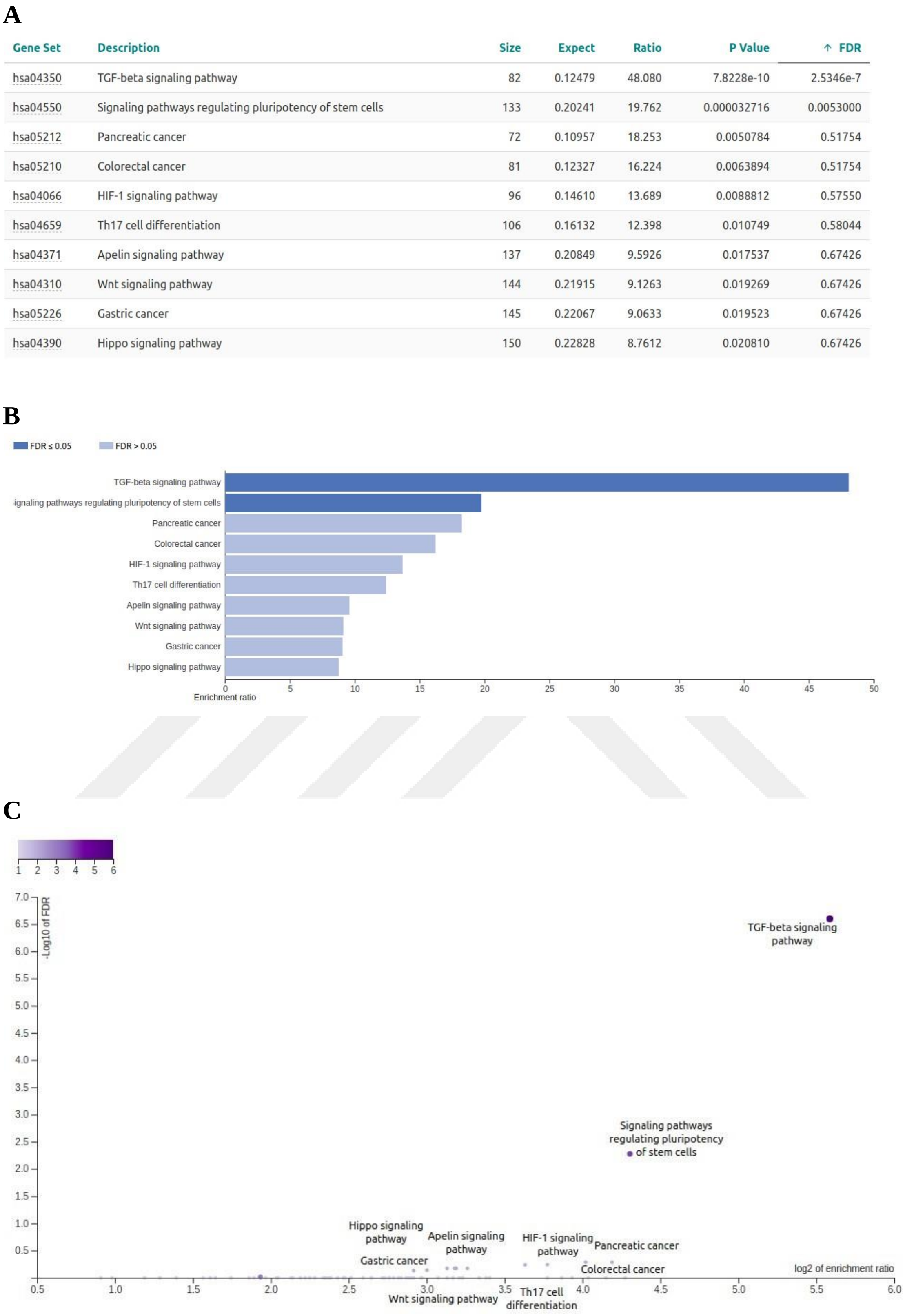
Gene Set Size	Expected Value	Overlap	Enrichment Ratio
344	0.55725	5	8.9726

User ID ↑	Gene Symbol	Gene Name	Entrez Gene ID
SMAD1	SMAD1	SMAD family member 1	4086
SMAD4	SMAD4	SMAD family member 4	4089
SMAD5	SMAD5	SMAD family member 5	4090
SMAD6	SMAD6	SMAD family member 6	4091
THRB	THRB	thyroid hormone receptor beta	7068

Figure 29: C. Visualization of functional annotation results of GO-CC on Volcano plot D. Detailed information of the most significant enriched GO set.

When these results were compared with the results obtained with ClueGO, it appears that there are no results for this GO category in ClueGO.

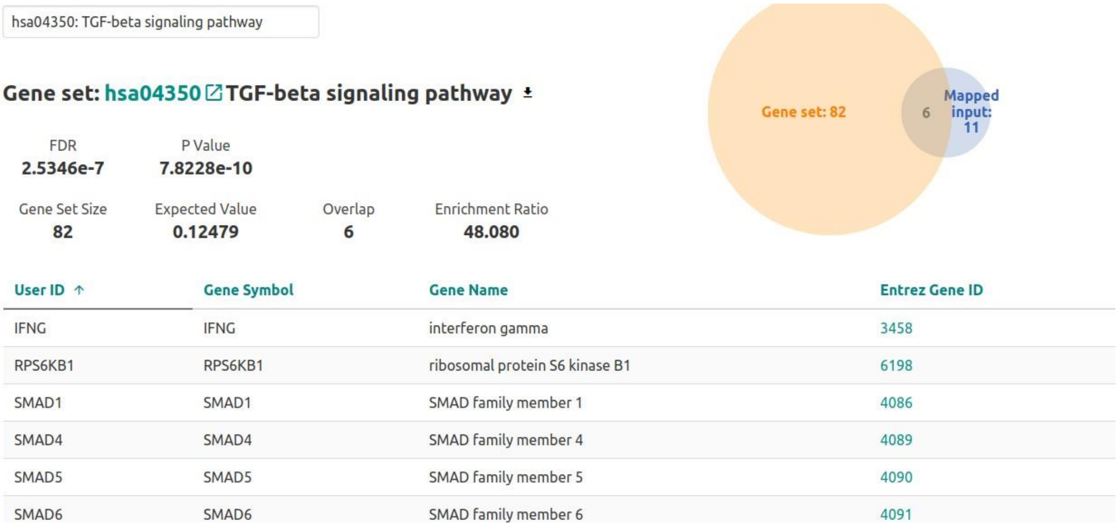
In addition to the Gene Ontology database, the pathway databases KEGG and Reactome were used for functional annotation analysis. Similar to the GO results, the table and bar chart containing p-value and FDR values are shown (Figure 30). Then the Volcano plot and figure showing the details of the most significant gene set result (Figure 31) are shown.



Gene Set	Description	Size	Expect	Ratio	P Value	↑ FDR
hsa04350	TGF-beta signaling pathway	82	0.12479	48.080	7.8228e-10	2.5346e-7
hsa04550	Signaling pathways regulating pluripotency of stem cells	133	0.20241	19.762	0.000032716	0.0053000
hsa05212	Pancreatic cancer	72	0.10957	18.253	0.0050784	0.51754
hsa05210	Colorectal cancer	81	0.12327	16.224	0.0063894	0.51754
hsa04066	HIF-1 signaling pathway	96	0.14610	13.689	0.0088812	0.57550
hsa04659	Th17 cell differentiation	106	0.16132	12.398	0.010749	0.58044
hsa04371	Apelin signaling pathway	137	0.20849	9.5926	0.017537	0.67426
hsa04310	Wnt signaling pathway	144	0.21915	9.1263	0.019269	0.67426
hsa05226	Gastric cancer	145	0.22067	9.0633	0.019523	0.67426
hsa04390	Hippo signaling pathway	150	0.22828	8.7612	0.020810	0.67426

Figure 30: A. KEGG functional annotation containing p value and FDR values B. KEGG pathway, enriched gene sets bar chart sorted by FDR values C. Visualization of functional annotation results of KEGG on Volcano plot

D



hsa04350: TGF-beta signaling pathway

Gene set: hsa04350 TGF-beta signaling pathway

FDR	P Value
2.5346e-7	7.8228e-10

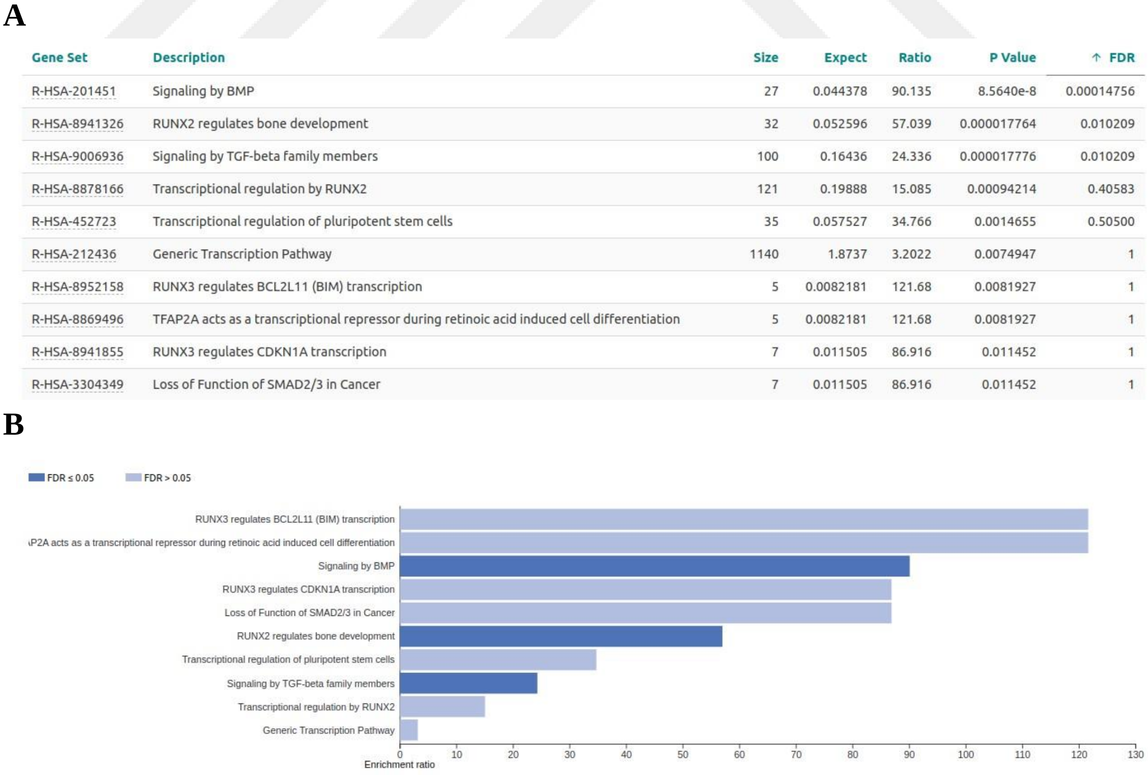
Gene Set Size	Expected Value	Overlap	Enrichment Ratio
82	0.12479	6	48.080

User ID ↑	Gene Symbol	Gene Name	Entrez Gene ID
IFNG	IFNG	interferon gamma	3458
RPS6KB1	RPS6KB1	ribosomal protein S6 kinase B1	6198
SMAD1	SMAD1	SMAD family member 1	4086
SMAD4	SMAD4	SMAD family member 4	4089
SMAD5	SMAD5	SMAD family member 5	4090
SMAD6	SMAD6	SMAD family member 6	4091

Figure 31: D. Detailed information of the most significant enriched KEGG set.

When these results were compared with the results obtained with ClueGO, it appears that there are no results for KEGG pathway in ClueGO.

In addition to the KEGG pathway, the same analyzes were performed for the Reactome pathway (Figure 32; Figure 33).



A

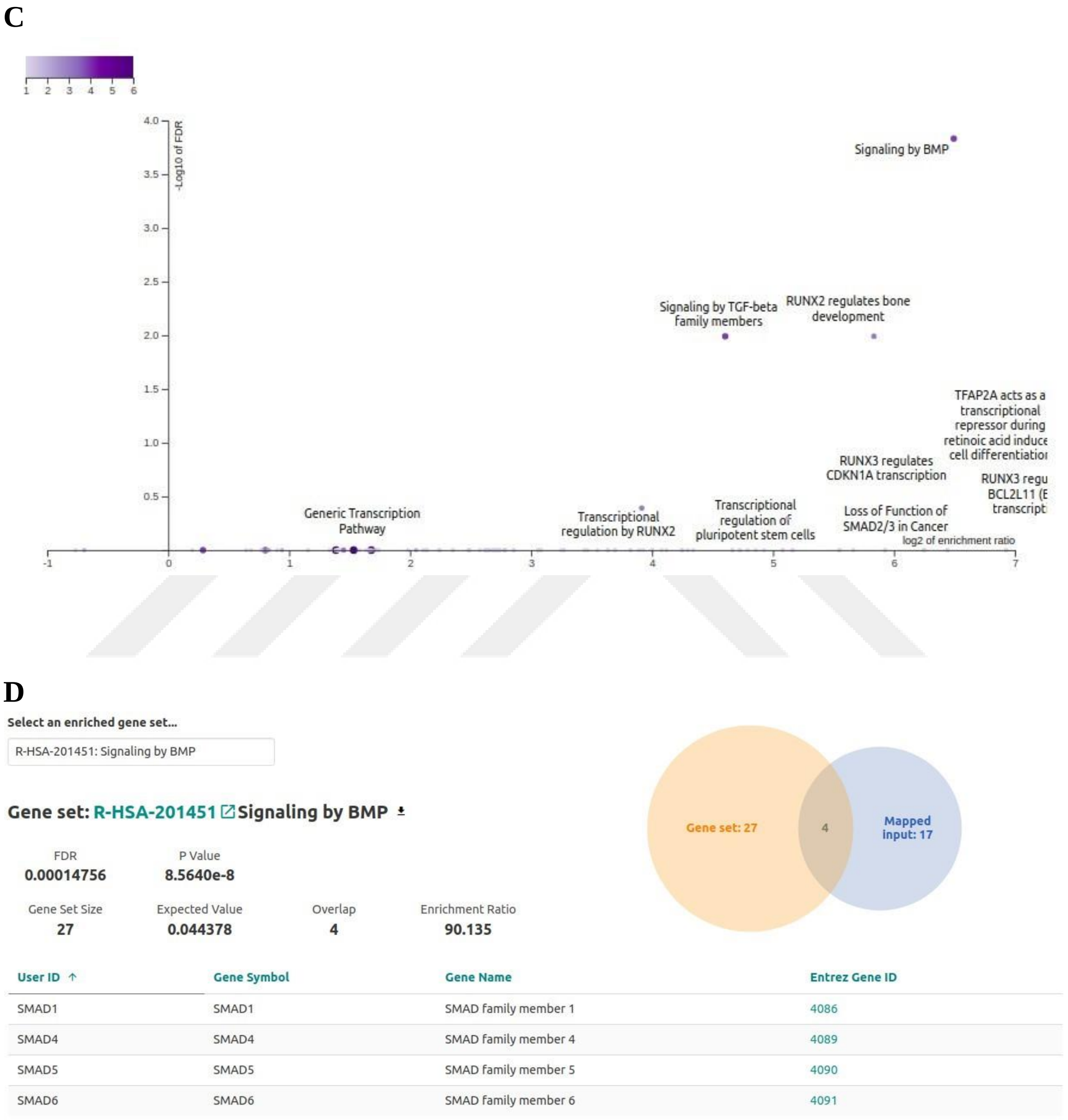
Gene Set	Description	Size	Expect	Ratio	P Value	↑ FDR
R-HSA-201451	Signaling by BMP	27	0.044378	90.135	8.5640e-8	0.00014756
R-HSA-8941326	RUNX2 regulates bone development	32	0.052596	57.039	0.000017764	0.010209
R-HSA-9006936	Signaling by TGF-beta family members	100	0.16436	24.336	0.000017776	0.010209
R-HSA-8878166	Transcriptional regulation by RUNX2	121	0.19888	15.085	0.00094214	0.40583
R-HSA-452723	Transcriptional regulation of pluripotent stem cells	35	0.057527	34.766	0.0014655	0.50500
R-HSA-212436	Generic Transcription Pathway	1140	1.8737	3.2022	0.0074947	1
R-HSA-8952158	RUNX3 regulates BCL2L11 (BIM) transcription	5	0.0082181	121.68	0.0081927	1
R-HSA-8869496	TFAP2A acts as a transcriptional repressor during retinoic acid induced cell differentiation	5	0.0082181	121.68	0.0081927	1
R-HSA-8941855	RUNX3 regulates CDKN1A transcription	7	0.011505	86.916	0.011452	1
R-HSA-3304349	Loss of Function of SMAD2/3 in Cancer	7	0.011505	86.916	0.011452	1

B

FDR ≤ 0.05
FDR > 0.05
RUNX3 regulates BCL2L11 (BIM) transcription
P2A acts as a transcriptional repressor during retinoic acid induced cell differentiation
Signaling by BMP
RUNX3 regulates CDKN1A transcription
Loss of Function of SMAD2/3 in Cancer
RUNX2 regulates bone development
Transcriptional regulation of pluripotent stem cells
Signaling by TGF-beta family members
Transcriptional regulation by RUNX2
Generic Transcription Pathway
0
10
20
30
40
50
60
70
80
90
100
110
120
130
Enrichment ratio

Figure 32: A. Reactome pathway functional annotation containing p value and FDR values B. Reactome pathway, enriched gene sets bar chart sorted by FDR values

When these results were compared with the results obtained with ClueGO, the following Reactome pathways appeared to be the same: "Signaling by BMP, RUNX2 regulates bone development”. On the other hand, when compared with the results obtained with the ReactomeFI plugin, it was seen that the results completely overlap.



Select an enriched gene set...

R-HSA-201451: Signaling by BMP

Gene set: R-HSA-201451 Signaling by BMP

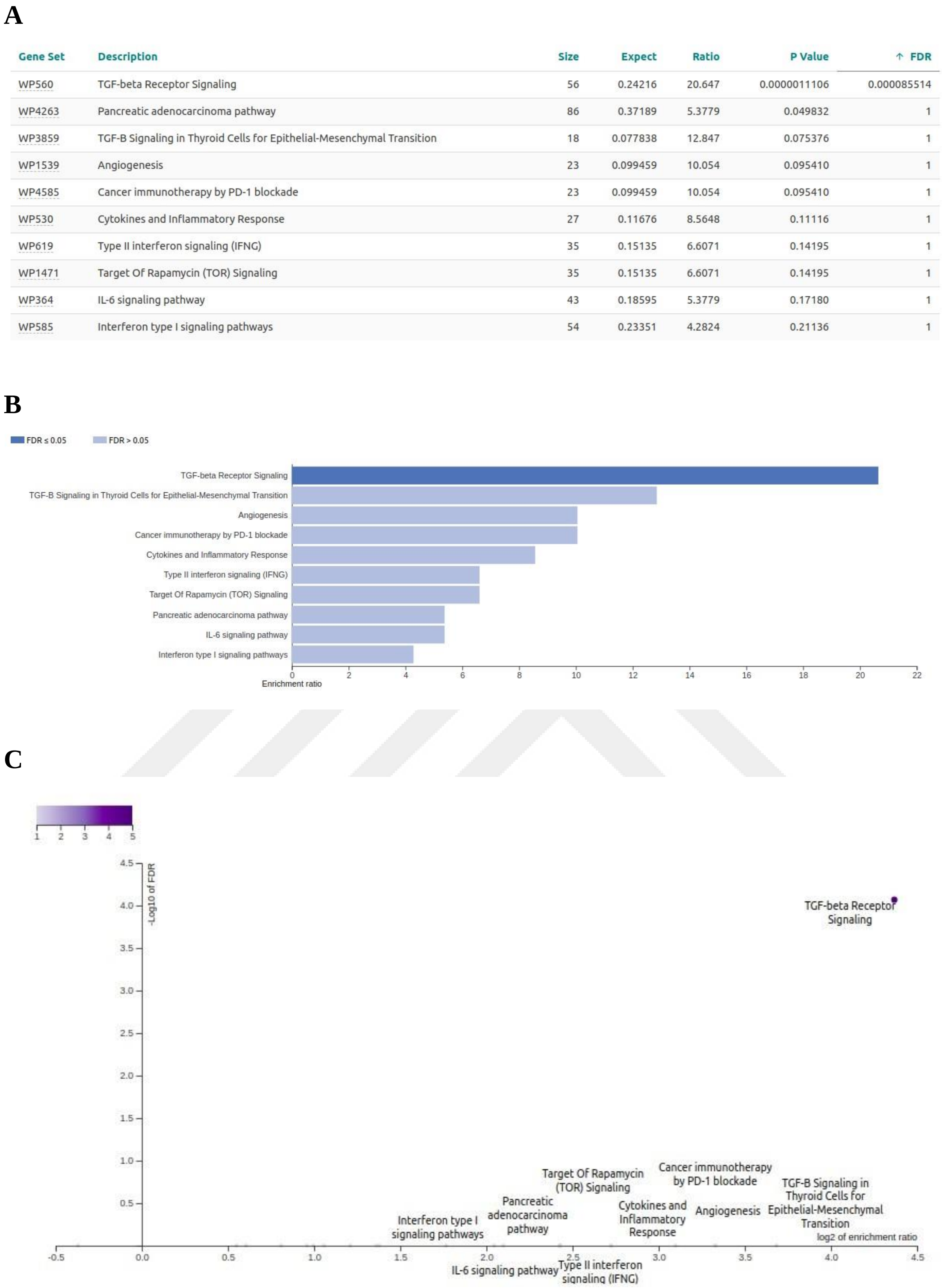
FDR	P Value
0.00014756	8.5640e-8

Gene Set Size	Expected Value	Overlap	Enrichment Ratio
27	0.044378	4	90.135

User ID ↑	Gene Symbol	Gene Name	Entrez Gene ID
SMAD1	SMAD1	SMAD family member 1	4086
SMAD4	SMAD4	SMAD family member 4	4089
SMAD5	SMAD5	SMAD family member 5	4090
SMAD6	SMAD6	SMAD family member 6	4091

Figure 33: C. Visualization of functional annotation results of Reactome pathway on Volcano plot D. Detailed information of the most significant enriched Reactome pathway gene sets

In addition to the KEGG and Reactome pathway, the same analyzes were performed by using the Wikipathway cancer database (Figure 34; Figure 35).

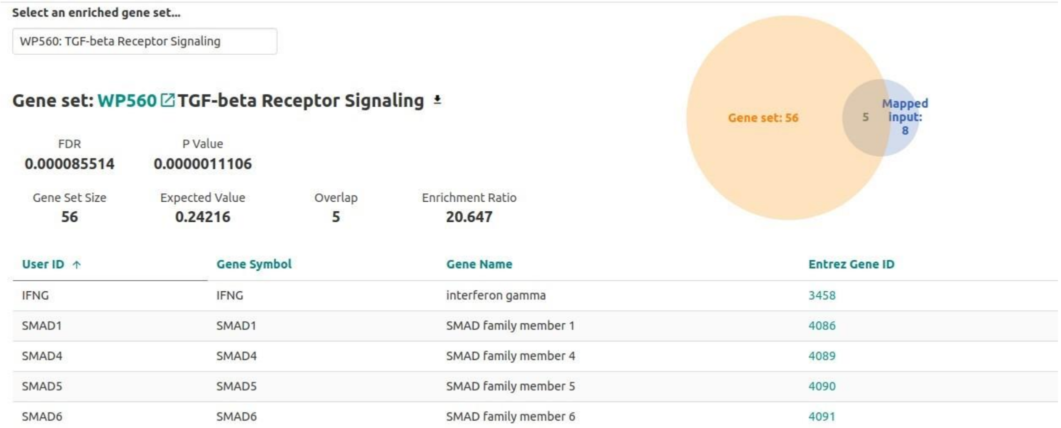


Gene Set	Description	Size	Expect	Ratio	P Value	↑ FDR
WP560	TGF-beta Receptor Signaling	56	0.24216	20.647	0.0000011106	0.000085514
WP4263	Pancreatic adenocarcinoma pathway	86	0.37189	5.3779	0.049832	1
WP3859	TGF-B Signaling in Thyroid Cells for Epithelial-Mesenchymal Transition	18	0.077838	12.847	0.075376	1
WP1539	Angiogenesis	23	0.099459	10.054	0.095410	1
WP4585	Cancer immunotherapy by PD-1 blockade	23	0.099459	10.054	0.095410	1
WP530	Cytokines and Inflammatory Response	27	0.11676	8.5648	0.11116	1
WP619	Type II interferon signaling (IFNG)	35	0.15135	6.6071	0.14195	1
WP1471	Target Of Rapamycin (TOR) Signaling	35	0.15135	6.6071	0.14195	1
WP364	IL-6 signaling pathway	43	0.18595	5.3779	0.17180	1
WP585	Interferon type I signaling pathways	54	0.23351	4.2824	0.21136	1

B
FDR ≤ 0.05
FDR > 0.05
TGF-beta Receptor Signaling
TGF-B Signaling in Thyroid Cells for Epithelial-Mesenchymal Transition
Angiogenesis
Cancer immunotherapy by PD-1 blockade
Cytokines and Inflammatory Response
Type II interferon signaling (IFNG)
Target Of Rapamycin (TOR) Signaling
Pancreatic adenocarcinoma pathway
IL-6 signaling pathway
Interferon type I signaling pathways
0
2
4
6
8
10
12
14
16
18
20
22
Enrichment ratio
C
1
2
3
4
5
-Log10 of FDR
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
TGF-beta Receptor Signaling
Target Of Rapamycin (TOR) Signaling
Cancer immunotherapy by PD-1 blockade
TGF-B Signaling in Thyroid Cells for Epithelial-Mesenchymal Transition
Pancreatic adenocarcinoma pathway
Cytokines and Inflammatory Response
Angiogenesis
Interferon type I signaling pathways
log2 of enrichment ratio
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
IL-6 signaling pathway
Type II interferon signaling (IFNG)

Figure 34: A. Wiki cancer pathway functional annotation containing p value and FDR values B. Wiki cancer pathway, enriched gene sets bar chart sorted by FDR values C. Visualization of functional annotation results of Wiki cancer pathway on Volcano plot

D



Select an enriched gene set...

WP560: TGF-beta Receptor Signaling

Gene set: WP560 TGF-beta Receptor Signaling

FDR: 0.000085514

P Value: 0.0000011106

Gene Set Size: 56

Expected Value: 0.24216

Overlap: 5

Enrichment Ratio: 20.647

User ID	Gene Symbol	Gene Name	Entrez Gene ID
IFNG	IFNG	interferon gamma	3458
SMAD1	SMAD1	SMAD family member 1	4086
SMAD4	SMAD4	SMAD family member 4	4089
SMAD5	SMAD5	SMAD family member 5	4090
SMAD6	SMAD6	SMAD family member 6	4091

Figure 35: D. Detailed information of the most significant enriched Wiki cancer pathway gene sets

3.1.2.2 Overrepresentation analysis using Babelomics

Using the Babelomics web tool, GO-BP analysis has been performed in order to compare it with other results obtained. A table showing functional annotation terms and genes involved was created (Table 10).

#term	term_size	term_size_in_genome	annotated_genes	converged id list	lor	adj_p value
protein complex subunit organization(GO:0071822)	7	3783	SMAD6 SMAD4 CTCF SMAD1 NPM1 FAU	TRUE	0.2658980539	0.0430918011
cellular component assembly(GO:0022607)	6	5055	SMAD6 SMAD4 CTCF SMAD1 SET NPM1	TRUE	0.3353266537	0.0314226471
macromolecular complex assembly(GO:0065003)	6	3373	SMAD6 SMAD4 CTCF SMAD1 SET NPM1	TRUE	0.3353266537	0.0314226471
cellular component biogenesis(GO:0044085)	7	5405	SMAD6 SMAD4 CTCF SMAD1 EBNA1BP2 SET NPM1	TRUE	0.2492740075	0.0487167691
ıacromolecular complex subunit organization(GO:004393	7	4361	SMAD6 SMAD4 CTCF SMAD1 SET NPM1 FAU	TRUE	0.2658980539	0.0430918011

Table 10: Functional annotation results of GO-BP databases by using Babelomics

In addition to this table, GO-BP's network has been obtained (Figure 36).

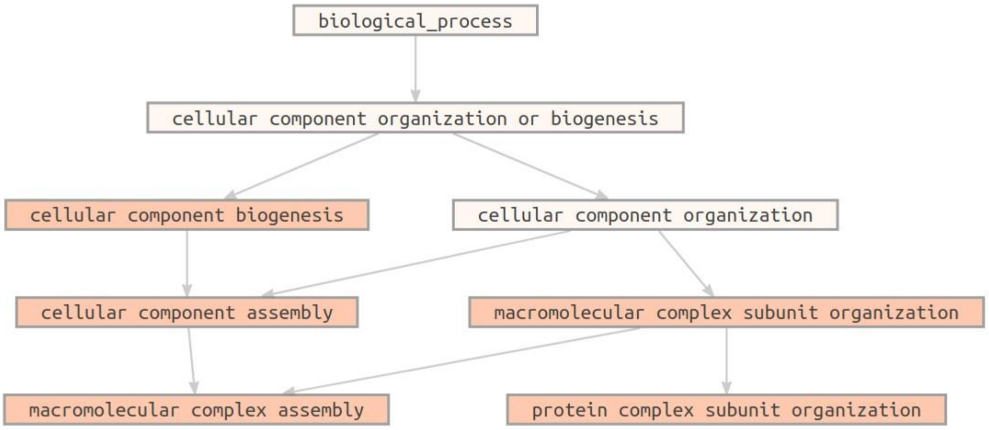
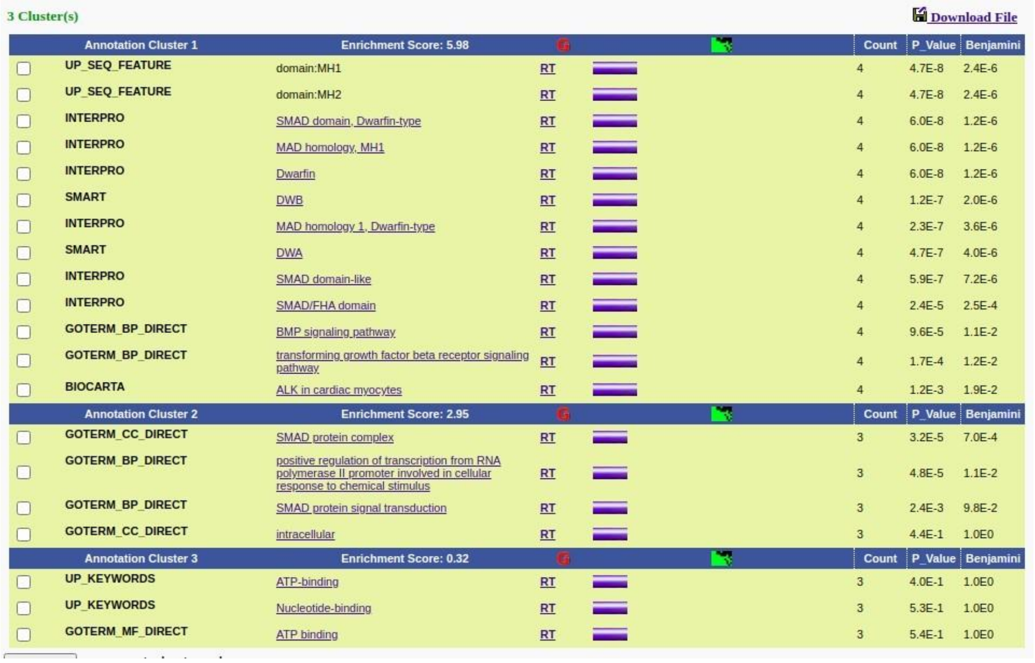

biological_process
cellular component organization or biogenesis
cellular component biogenesis
cellular component organization
cellular component assembly
macromolecular complex subunit organization
macromolecular complex assembly
protein complex subunit organization

Figure 36: GO-BP annotation network by using

Babelomics No results were obtained from Babelomics for other GO databases.

3.1.2.3 Overrepresentation analysis using DAVID

DAVID was implemented for functional annotation enrichment analysis. In order to observe the annotations better and to examine the most significant results, the enrichment score was filtered as the highest. The following table consisting of 3 clusters was obtained (Figure 37).



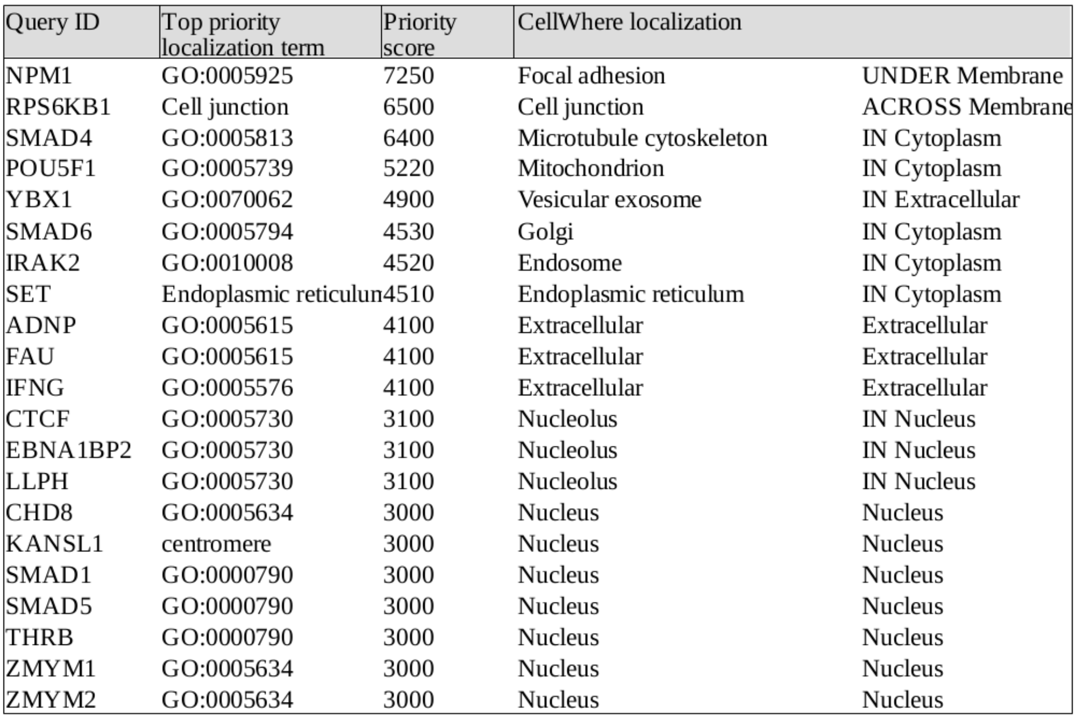
	Annotation Cluster 1	Enrichment Score: 5.98		Count	P_Value	Benjamini
□	UP_SEQ_FEATURE	domain:MH1	RT	4	4.7E-8	2.4E-6
□	UP_SEQ_FEATURE	domain:MH2	RT	4	4.7E-8	2.4E-6
□	INTERPRO	SMAD domain, Dwarfin-type	RT	4	6.0E-8	1.2E-6
□	INTERPRO	MAD homology, MH1	RT	4	6.0E-8	1.2E-6
□	INTERPRO	Dwarfin	RT	4	6.0E-8	1.2E-6
□	SMART	DWB	RT	4	1.2E-7	2.0E-6
□	INTERPRO	MAD homology 1, Dwarfin-type	RT	4	2.3E-7	3.6E-6
□	SMART	DWA	RT	4	4.7E-7	4.0E-6
□	INTERPRO	SMAD domain-like	RT	4	5.9E-7	7.2E-6
□	INTERPRO	SMAD/FHA domain	RT	4	2.4E-5	2.5E-4
□	GOTERM_BP_DIRECT	BMP signaling pathway	RT	4	9.6E-5	1.1E-2
□	GOTERM_BP_DIRECT	transforming growth factor beta receptor signaling pathway	RT	4	1.7E-4	1.2E-2
□	BIOCARTA	ALK in cardiac myocytes	RT	4	1.2E-3	1.9E-2
	Annotation Cluster 2	Enrichment Score: 2.95		Count	P_Value	Benjamini
□	GOTERM_CC_DIRECT	SMAD protein complex	RT	3	3.2E-5	7.0E-4
□	GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter involved in cellular response to chemical stimulus	RT	3	4.8E-5	1.1E-2
□	GOTERM_BP_DIRECT	SMAD protein signal transduction	RT	3	2.4E-3	9.8E-2
□	GOTERM_CC_DIRECT	intracellular	RT	3	4.4E-1	1.0E0
	Annotation Cluster 3	Enrichment Score: 0.32		Count	P_Value	Benjamini
□	UP_KEYWORDS	ATP-binding	RT	3	4.0E-1	1.0E0
□	UP_KEYWORDS	Nucleotide-binding	RT	3	5.3E-1	1.0E0
□	GOTERM_MF_DIRECT	ATP binding	RT	3	5.4E-1	1.0E0

Figure 37: Functional annotation clustering results by using DAVID online tool

As it appears in this table pathways such as "SMAD domain”, “BMP signaling pathway”, “transforming growth factor beta receptor signaling pathway”, “positive regulation of transcription from RNA polymerase II promoter involved in cellular response to chemical stimulus" were obtained using the DAVID online tool, similar to the results we have achieved so far.

3.1.2.4 Graphical display of the network based on subcellular localization using CellWhere

CellWhere, an online tool, is used to detect and visualize the subcellular localization of nodes in this network. In this tool, first visual graphic was created using default options (display localization based on: muscle) (Table 11; Figure 38).



Query ID	Top priority localization term	Priority score	CellWhere localization	
NPM1	GO:0005925	7250	Focal adhesion	UNDER Membrane
RPS6KB1	Cell junction	6500	Cell junction	ACROSS Membrane
SMAD4	GO:0005813	6400	Microtubule cytoskeleton	IN Cytoplasm
POU5F1	GO:0005739	5220	Mitochondrion	IN Cytoplasm
YBX1	GO:0070062	4900	Vesicular exosome	IN Extracellular
SMAD6	GO:0005794	4530	Golgi	IN Cytoplasm
IRAK2	GO:0010008	4520	Endosome	IN Cytoplasm
SET	Endoplasmic reticulun	4510	Endoplasmic reticulum	IN Cytoplasm
ADNP	GO:0005615	4100	Extracellular	Extracellular
FAU	GO:0005615	4100	Extracellular	Extracellular
IFNG	GO:0005576	4100	Extracellular	Extracellular
CTCF	GO:0005730	3100	Nucleolus	IN Nucleus
EBNA1BP2	GO:0005730	3100	Nucleolus	IN Nucleus
LLPH	GO:0005730	3100	Nucleolus	IN Nucleus
CHD8	GO:0005634	3000	Nucleus	Nucleus
KANSL1	centromere	3000	Nucleus	Nucleus
SMAD1	GO:0000790	3000	Nucleus	Nucleus
SMAD5	GO:0000790	3000	Nucleus	Nucleus
THRB	GO:0000790	3000	Nucleus	Nucleus
ZMYM1	GO:0005634	3000	Nucleus	Nucleus
ZMYM2	GO:0005634	3000	Nucleus	Nucleus

Table 11: Subcellular localization of the nodes table

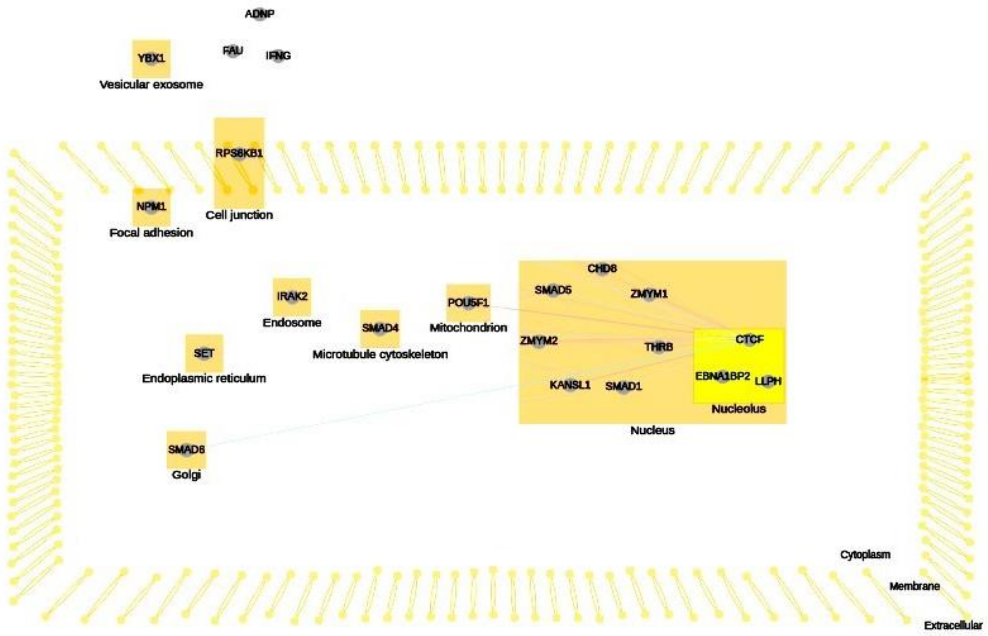


Figure 38: Sub localizations of CTCF obtained by selecting "muscle" from the CellWhere

Secondly, another graphical visual was created by selecting the display localization based on option as annotation frequency (Figure 39). According to the result, it was determined that most of the nodes in this network are located in the nucleus.

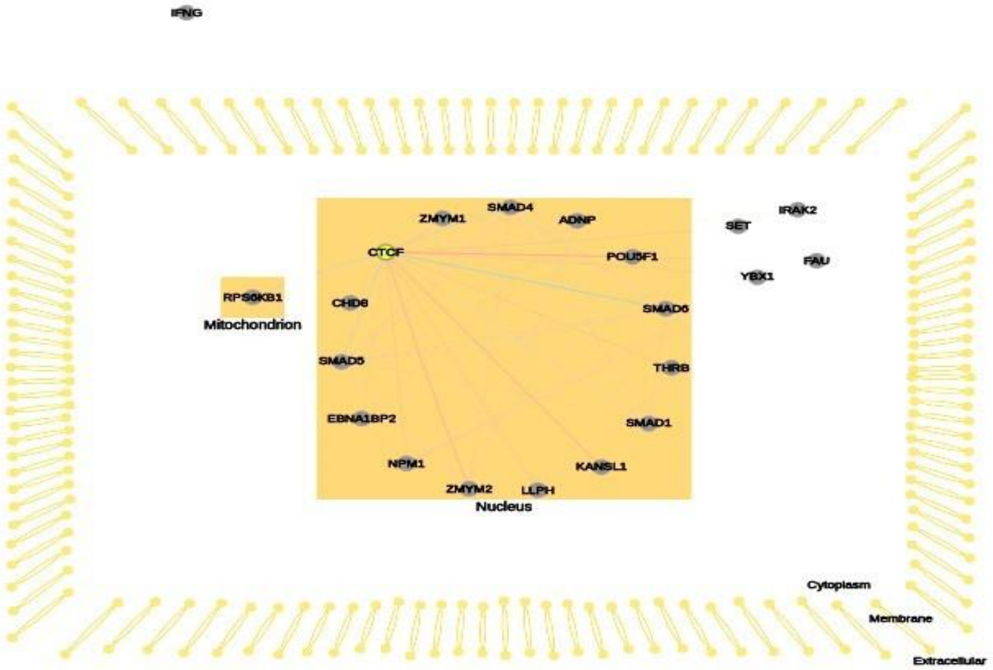


Figure 39: Sub localizations of CTCF obtained by selecting "annotation frequency"

3.2 Mining Methylation and Gene Expression Data

In order to perform combined methylation intensity and transcript abundance analysis, unlike previous studies, all probe regions that are methylation specific on the CTCF gene were analyzed separately in this study. The annotation information of these probes was determined using the "IlluminaHumanMethylation450kanno.ilmn12hg19" R package .

Then, using the ChIPpeakAnno package, all probe regions within 5 kb of upstream and downstream lengths according to the TSS region of the CTCF were determined. Accordingly, 12 probe regions, 6 of which are upstream and 6 of which are inside of the CTCF gene, were determined (Table 12). These 12 probe regions will be used in all future analyzes and "Inside Feature" column will be used as a reference in scatter plots.

CG probe ID	Ensembl ID	HGNC Symbol	Distance to TSS	insideFeature
cg23858565	ENSG00000102974	CTCF	-608	upstream
cg07967402	ENSG00000102974	CTCF	-468	upstream
cg08324636	ENSG00000102974	CTCF	-442	upstream
cg06241380	ENSG00000102974	CTCF	-357	upstream
cg10218542	ENSG00000102974	CTCF	-326	upstream
cg04487155	ENSG00000102974	CTCF	-32	upstream
cg10481400	ENSG00000102974	CTCF	126	inside
cg01866162	ENSG00000102974	CTCF	203	inside
cg27250362	ENSG00000102974	CTCF	1005	inside
cg02215945	ENSG00000102974	CTCF	1063	inside
cg16517579	ENSG00000102974	CTCF	1832	inside
cg04545079	ENSG00000102974	CTCF	3613	inside

Table 12: Probe regions on the CTCF gene

Correlation analysis was performed to observe the impact of methylation of CTCF gene on gene expression in healthy samples.

The same "Illumina 450K array (GPL13534) platform" was used for the methylation data and "Affymetrix Human Genome U133 Plus 2.0 Array (GPL570-HG- U133_Plus_2)" platform was used for the gene expression data for all different stages samples.

3.2.1 Age-Related Change of the CTCF Gene Methylation Profile and Its Effect on Gene Expression

To examine the effect of development-associated variation in the CTCF methylation profile for the CTCF gene on gene expression, samples taken from online databases are divided into 7 different developmental age categories: fetal (prenatal), newborn, infancy (1-4), childhood (5-17), early adulthood (18-40), late adulthood (41-80) and senescence (80+).

At least three different datasets were used for each stages sample, to minimize problems that may result from technical errors and to increase the consistency. The annotation files, which contain detailed information about which datasets are used and detailed information about this data set, has been prepared separately for each age group and is given in the appendices (Table:B.1.1-B1.7).

Pearson and Spearman correlations were performed to examine assess the impact of CTCF methylation changes on CTCF mRNA level in different developmental stages. Scatter plots where each point represents a separate probe and the tables showing correlation values corresponding to each probe site were created separately for each age category.

The scatter plots (Figure 42-48) and the correlation tables (Table 13-19) created for the categories in this developmental process are shown in the next section respectively. PCA plot (Figure 40) and heatmap (Figure 41) demonstrates which age group categories are closer to each other based on the methylation data.

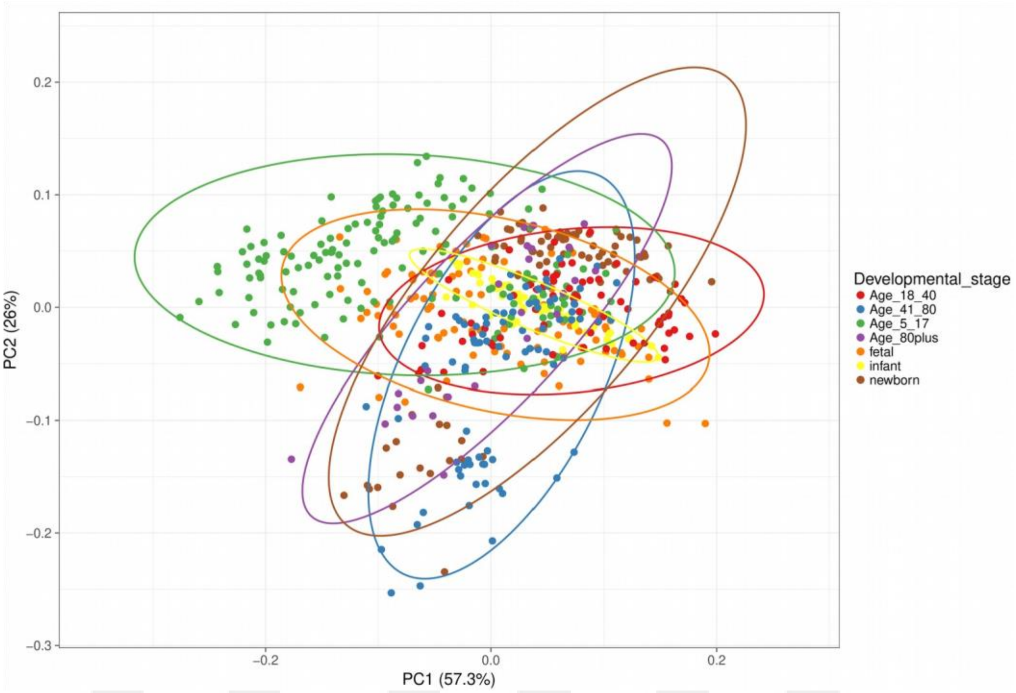


Figure 40: PCA plot of the all developmental age categories

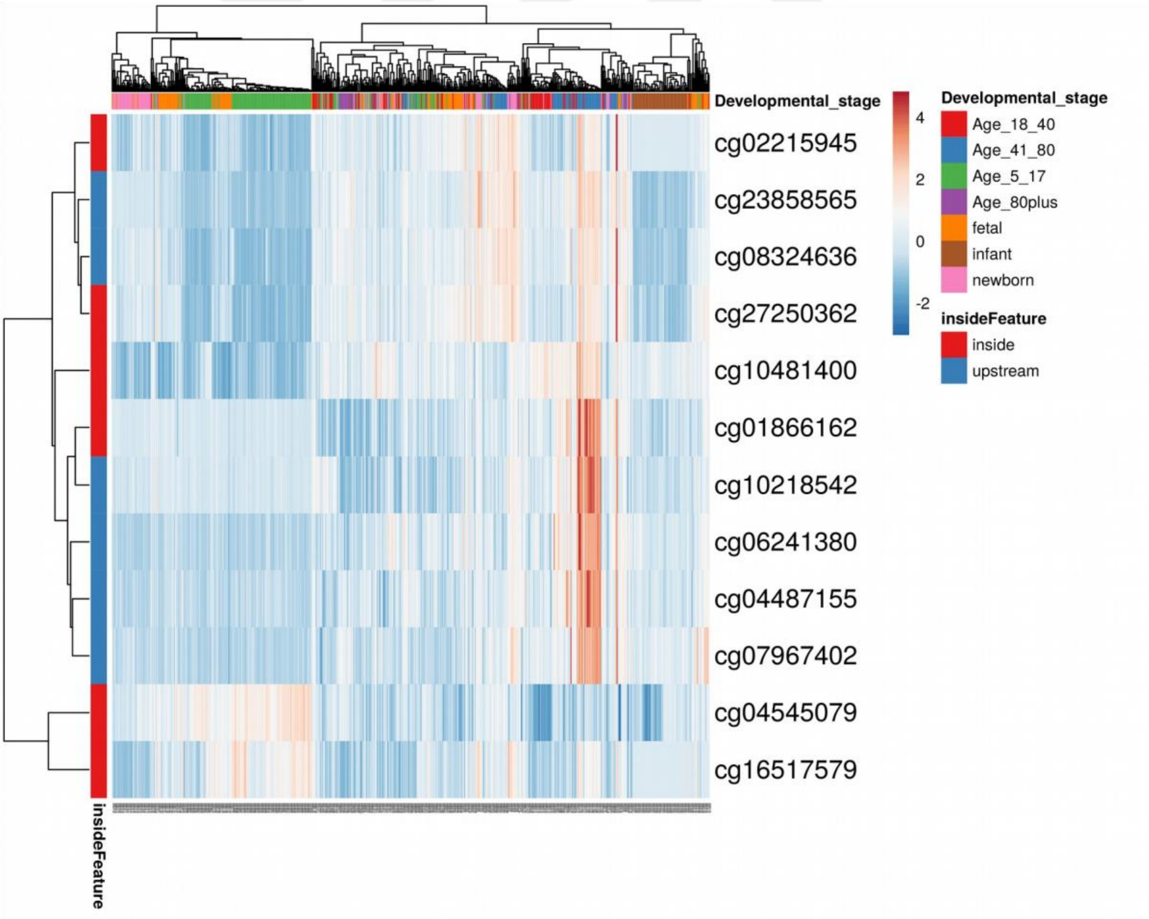


Figure 41: Heatmap of the all developmental age categories

3.2.1.1 Fetal

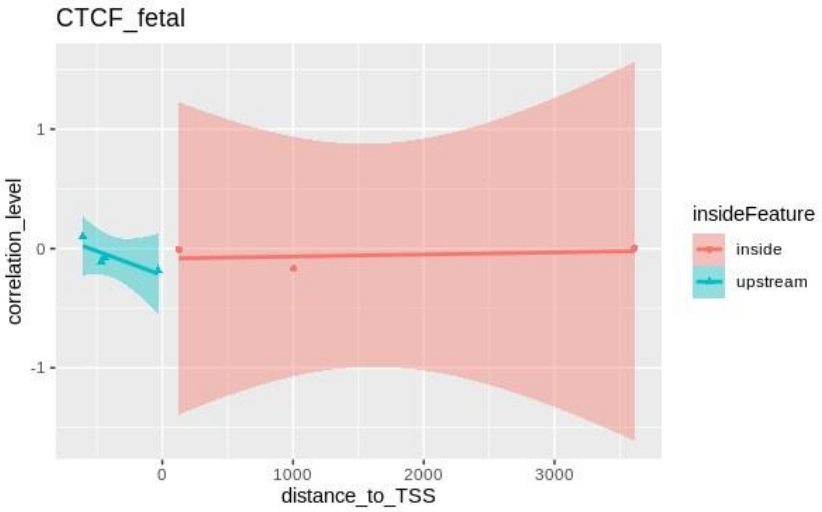


Figure 42: Scatter plot of probe sites in the CTCF gene of fetal samples

CG probe ID	Distance to TSS	insideFeature	Pearson correlation (r)	Spearman correlation (rho)
cg23858565	-608	upstream	0.10	0.10
cg07967402	-468	upstream	-0.11	0.03
cg08324636	-442	upstream	-0.07	-0.10
cg06241380	-357	upstream	NA	NA
cg10218542	-326	upstream	NA	NA
cg04487155	-32	upstream	-0.18	-0.12
cg10481400	126	inside	-0.01	-0.02
cg01866162	203	inside	NA	NA
cg27250362	1005	inside	-0.17	-0.19
cg02215945	1063	inside	NA	NA
cg16517579	1832	inside	NA	NA
cg04545079	3613	inside	0.00	-0.03

Table 13: Correlation of probe regions in the CTCF gene of fetal samples

3.2.1.2. Newborn

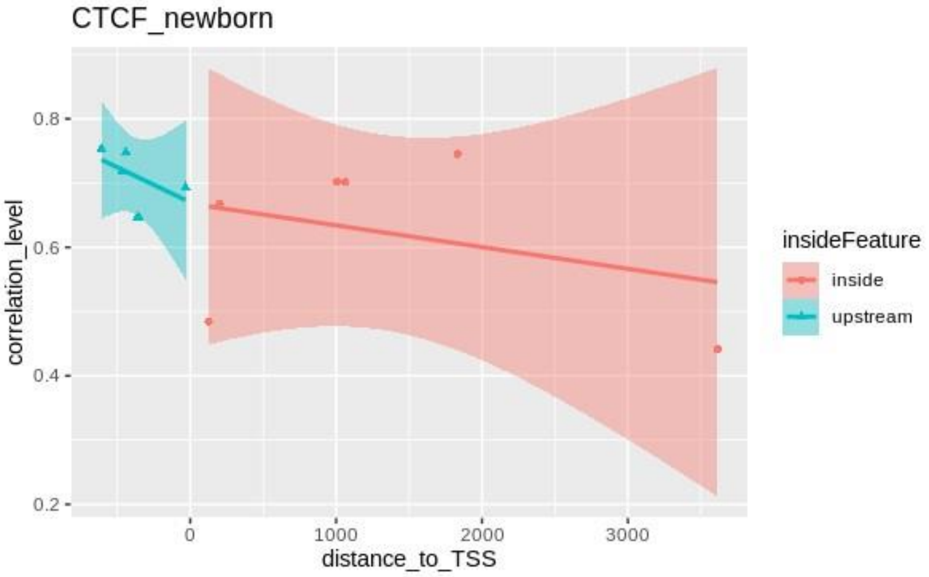


Figure 43: Scatter plot of probe sites in the CTCF gene of newborn samples

CG probe ID	Distance to TSS	insideFeature	Pearson correlation (r)	Spearman correlation (rho)
cg23858565	**-608**	**upstream**	**0.75**	**0.59**
cg07967402	-468	upstream	0.72	0.43
cg08324636	**-442**	**upstream**	**0.75**	**0.52**
cg06241380	**-357**	**upstream**	**0.65**	**0.52**
cg10218542	-326	upstream	NA	NA
cg04487155	**-32**	**upstream**	**0.69**	**0.57**
cg10481400	126	inside	0.48	0.40
cg01866162	**203**	**inside**	**0.67**	**0.60**
cg27250362	**1005**	**inside**	**0.70**	**0.55**
cg02215945	1063	inside	0.70	0.49
cg16517579	**1832**	**inside**	**0.75**	**0.50**
cg04545079	3613	inside	0.44	0.39

Table 14: Correlation of probe regions in the CTCF gene of newborn samples

CTCF methylation in newborn babies appeared to be positively correlated with all probes.

3.2.1.3 Infancy (1-5)

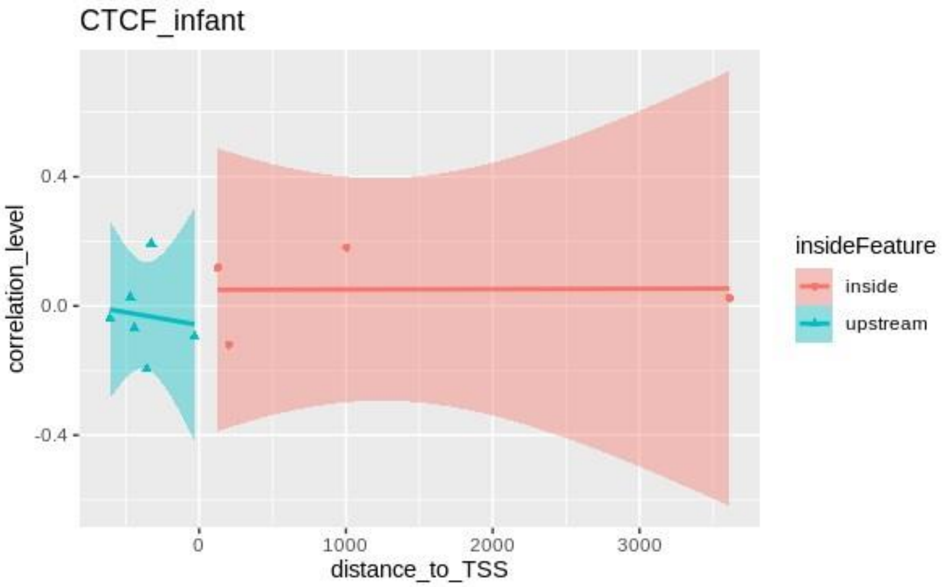


Figure 44: Scatter plot of probe sites in the CTCF gene of infant samples

CG probe ID	Distance to TSS	insideFeature	Pearson correlation (r)	Spearman correlation (rho)
cg23858565	-608	upstream	-0.04	-0.03
cg07967402	-468	upstream	0.03	0.06
cg08324636	-442	upstream	-0.07	-0.12
cg06241380	-357	upstream	-0.20	-0.17
cg10218542	-326	upstream	0.19	0.24
cg04487155	-32	upstream	-0.09	-0.10
cg10481400	126	inside	0.12	0.07
cg01866162	203	inside	-0.12	-0.10
cg27250362	1005	inside	0.18	0.26
cg02215945	1063	inside	NA	NA
cg16517579	1832	inside	NA	NA
cg04545079	3613	inside	0.03	-0.06

Table 15: Correlation of probe regions in the CTCF gene of infant samples

Very weak or no relationship between methylation and gene expression was observed.

3.2.1.4 Childhood (5-17)

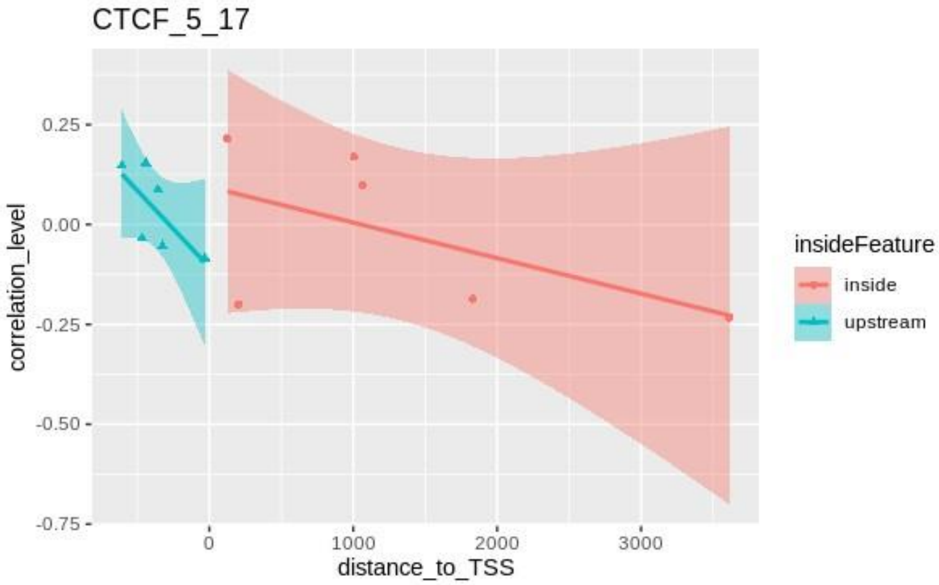


Figure 45: Scatter plot of probe sites in the CTCF gene of childhood sample

CG probe ID	Distance to TSS	insideFeature	Pearson correlation (r)	Spearman correlation (rho)
cg23858565	-608	upstream	0.15	0.12
cg07967402	-468	upstream	-0.03	-0.05
cg08324636	-442	upstream	0.15	0.13
cg06241380	-357	upstream	0.09	0.00
cg10218542	-326	upstream	-0.05	-0.11
cg04487155	-32	upstream	-0.08	-0.05
cg10481400	126	inside	0.22	0.16
cg01866162	203	inside	-0.20	-0.17
cg27250362	1005	inside	0.17	0.13
cg02215945	1063	inside	0.10	0.04
cg16517579	1832	inside	-0.19	-0.26
cg04545079	3613	inside	-0.23	-0.26

Table 16: Correlation of probe regions in the CTCF gene of childhood samples

Similar to infants, very week or no correlation of methylation and between gene expression in childhood was observed at the stage.

3.2.1.5 Early adulthood (18-40)

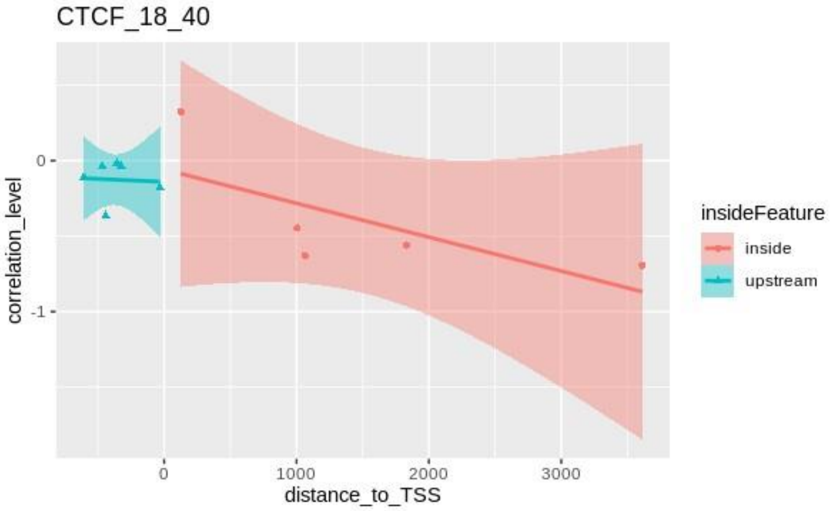


Figure 46: Scatter plot of probe sites in the CTCF gene of early adulthood samples

CG probe ID	Distance to TSS	insideFeature	Pearson correlation (r)	Spearman correlation (rho)
cg23858565	-608	upstream	-0.11	-0.05
cg07967402	-468	upstream	-0.04	0.01
cg08324636	-442	upstream	-0.37	-0.33
cg06241380	-357	upstream	-0.02	-0.02
cg10218542	-326	upstream	-0.04	-0.04
cg04487155	-32	upstream	-0.18	-0.10
cg10481400	126	inside	0.32	0.34
cg01866162	203	inside	NA	NA
cg27250362	1005	inside	-0.45	-0.42
cg02215945	**1063**	**inside**	**-0.63**	**-0.58**
cg16517579	1832	inside	-0.56	-0.48
cg04545079	**3613**	**inside**	**-0.69**	**-0.63**

Table 17: Correlation of probe regions in the CTCF gene of early adulthood samples

It was observed that almost all probes showed negative correlation in early adulthood.

3.2.1.6 Late Adulthood (41-80)

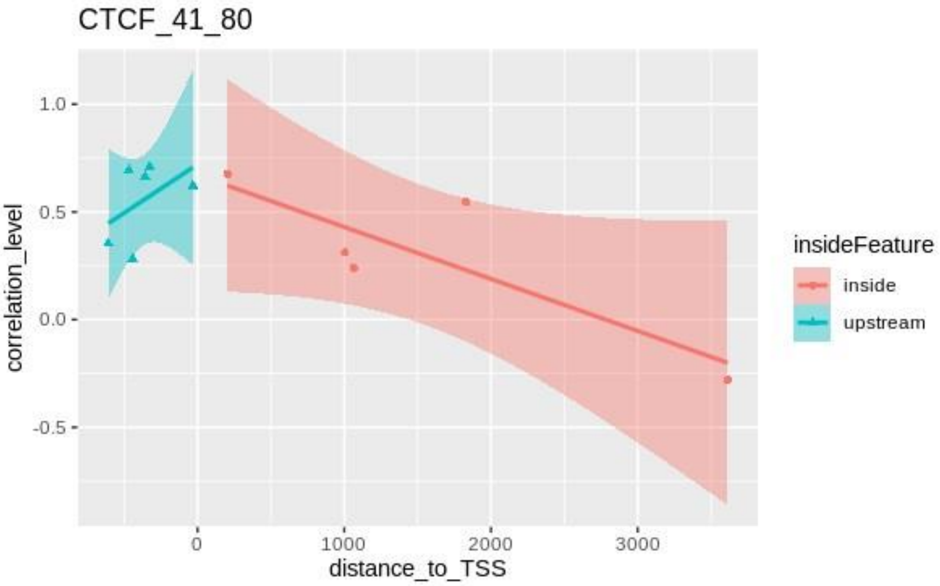


Figure 47: Scatter plot of probe sites in the CTCF gene of late adulthood samples

CG probe ID	Distance to TSS	insideFeature	Pearson correlation (r)	Spearman correlation (rho)
cg23858565	-608	upstream	0.35	0.40
cg07967402	**-468**	**upstream**	**0.69**	**0.62**
cg08324636	-442	upstream	0.28	0.33
cg06241380	**-357**	**upstream**	**0.66**	**0.66**
cg10218542	**-326**	**upstream**	**0.71**	**0.68**
cg04487155	**-32**	**upstream**	**0.62**	**0.58**
cg10481400	126	inside	NA	NA
cg01866162	**203**	**inside**	**0.68**	**0.67**
cg27250362	1005	inside	0.31	0.36
cg02215945	1063	inside	0.24	0.28
cg16517579	**1832**	**inside**	**0.55**	**0.52**
cg04545079	3613	inside	-0.28	-0.28

Table 18: Correlation of probe regions in the CTCF gene of late adulthood samples

In late adulthood, unlike early adulthood, almost all probes were found to show a high positive correlation.

3.2.1.7 Senescence (80+)

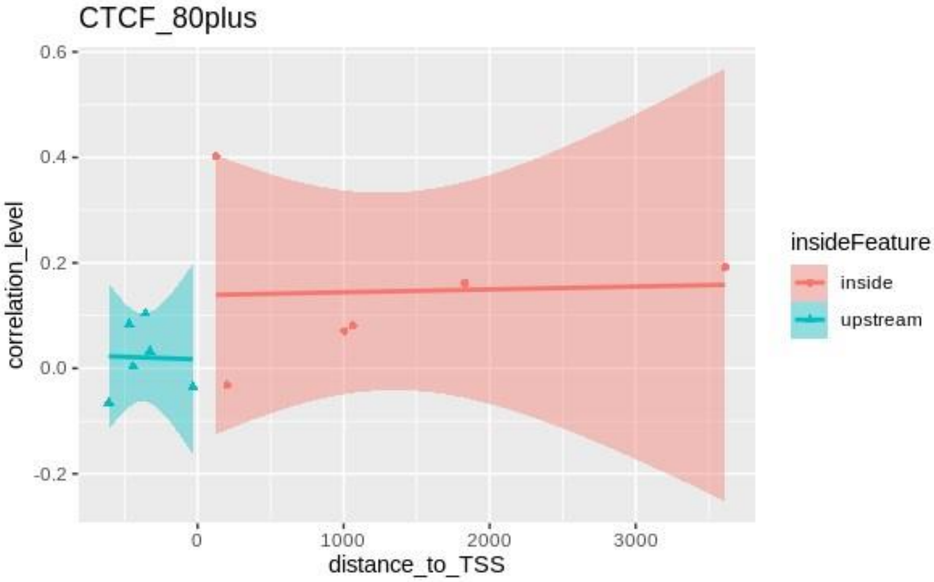


Figure 48: Scatter plot of probe sites in the CTCF gene of senescence samples

CG probe ID	Distance to TSS	insideFeature	Pearson correlation (r)	Spearman correlation (rho)
cg23858565	-608	upstream	-0.06	0.17
cg07967402	-468	upstream	0.08	0.26
cg08324636	-442	upstream	0.00	0.25
cg06241380	-357	upstream	0.10	0.29
cg10218542	-326	upstream	0.03	0.26
cg04487155	-32	upstream	-0.03	0.08
cg10481400	126	inside	0.40	0.61
cg01866162	203	inside	-0.03	0.12
cg27250362	1005	inside	0.07	0.27
cg02215945	1063	inside	0.08	0.29
cg16517579	1832	inside	0.16	0.42
cg04545079	3613	inside	0.19	0.23

Table 19: Correlation of probe regions in the CTCF gene of senescence samples

In senescence samples, weak relationship between methylation and gene expression of CTCF gene was observed.

3.2.1.8 Numerical summary of development-related correlation results

A single scatter plot that shows correlation values across all age ranges were created for better comparison of all results (Figure 49). In this scatter plot, different displays were provided according to the age range, and different colors were provided according to the probes.

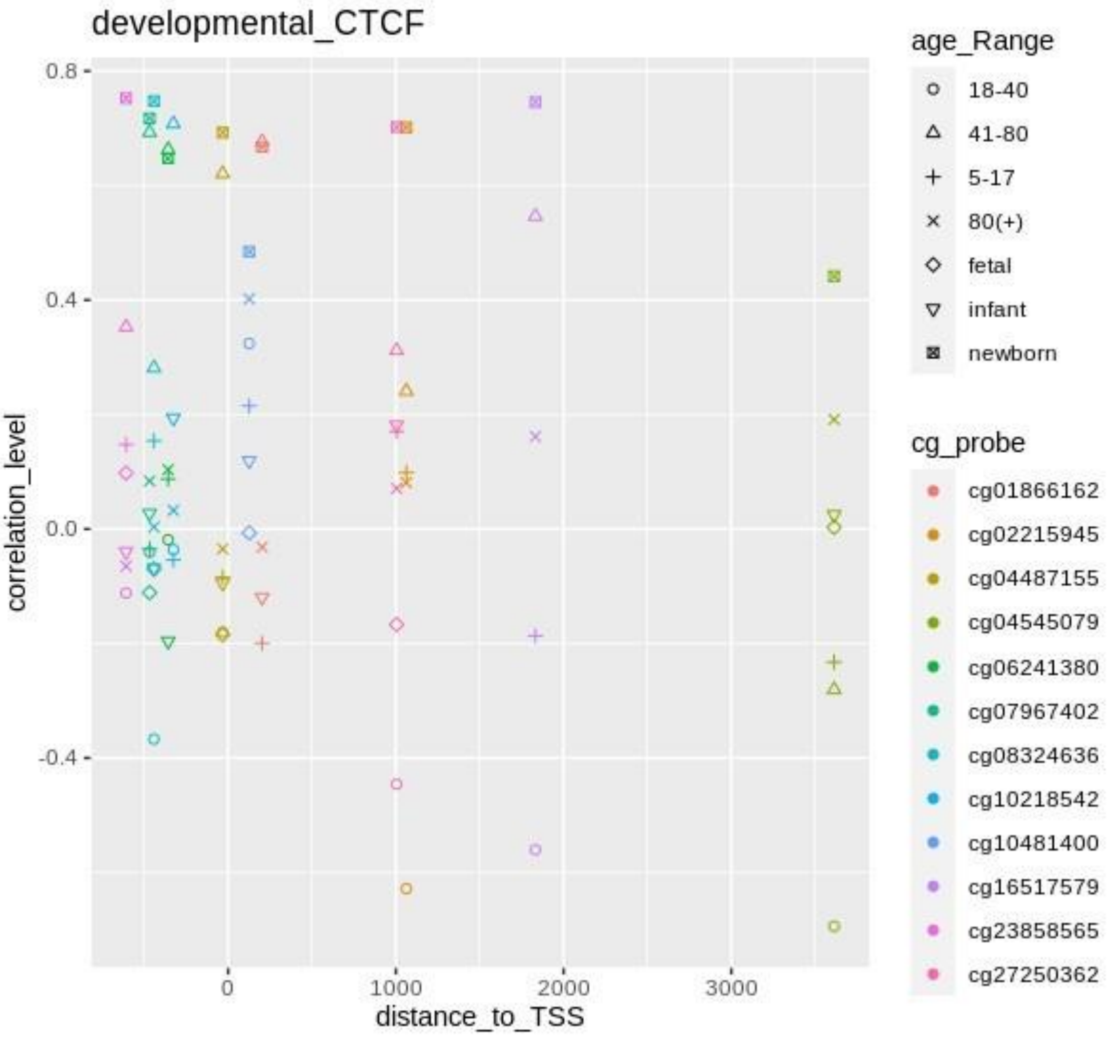


Figure 49: Scatter plot showing all age categories together

Looking at the scatter plot that included all age categories obtained, it was found that there was a similar positive correlation in the newborn and late adulthood, and conversely, there was a negative correlation in early adulthood.

3.2.2 Change of the Methylation Profile of the CTCF Gene in Cancerous Tissues and Its Effect on Gene Expression

One of the purpose of this study was to asses the effect of methylation on CTCF expression in the context of cancer. To examine whether the methylations in methylation-specific probe sites had an effect on the CTCF gene, cancer and healthy samples were collected for 12 different tissues (for both normal and cancer status): bladder, bone, brain, breast, colon, gastric, kidney, liver, lung, pancreas, prostate, and small intestine.

At least three different datasets were used for each cancer tissue sample, to minimize problems that may result from technical errors and to increase the consistency. The annotation files, which contain detailed information about which datasets are used and detailed information about this data set, has been prepared separately for each tissue group and is given in the appendices (Table: B.2.1-B.2.12; Table: B.3.1-B.3.12).

In the next step, the PCA and heatmap created by using Clustvis online tool will be shown first, using methylation data of cancer and healthy tissue types. Then, to evaluate the influence of methylations on CTCF gene expression, Pearson and Spearman correlation analysis was conducted for each tissue type, first using healthy tissue and then cancerous tissue. Correlation analysis results are shown on the scatter plot in the same order.

In addition, in order to observe the correlation difference between cancerous tissues and healthy tissues more clearly, scatter plots showing these two results together were created for each tissue type. Then, the difference between healthy and cancerous correlation analysis was calculated for each probe region, and both “Pearson correlation difference” and “Spearman correlation difference” were shown on the table for each tissue type.

3.2.2.1 Colon

PCA (Figure 50) and heatmap (Figure 51) created using methylation data of healthy and cancerous tissues are shown below.

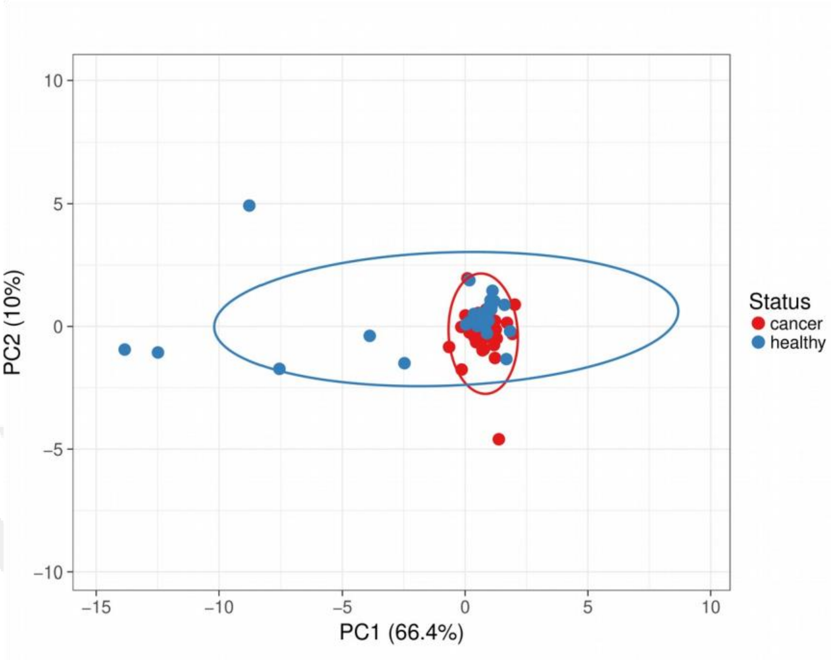


Figure 50: PCA of colon methylation data

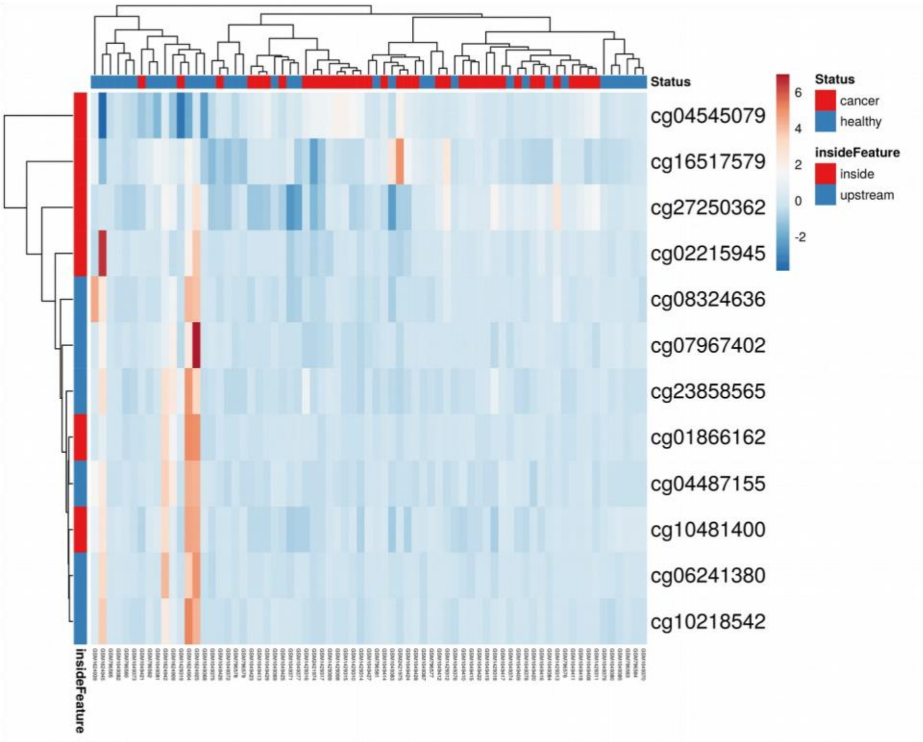


Figure 51: Heatmap of colon methylation data

Scatter plots showing correlation analysis of methylation-gene expression data of correlation in healthy (Figure 52) and cancerous colon samples (Figure 53) are depicted below.

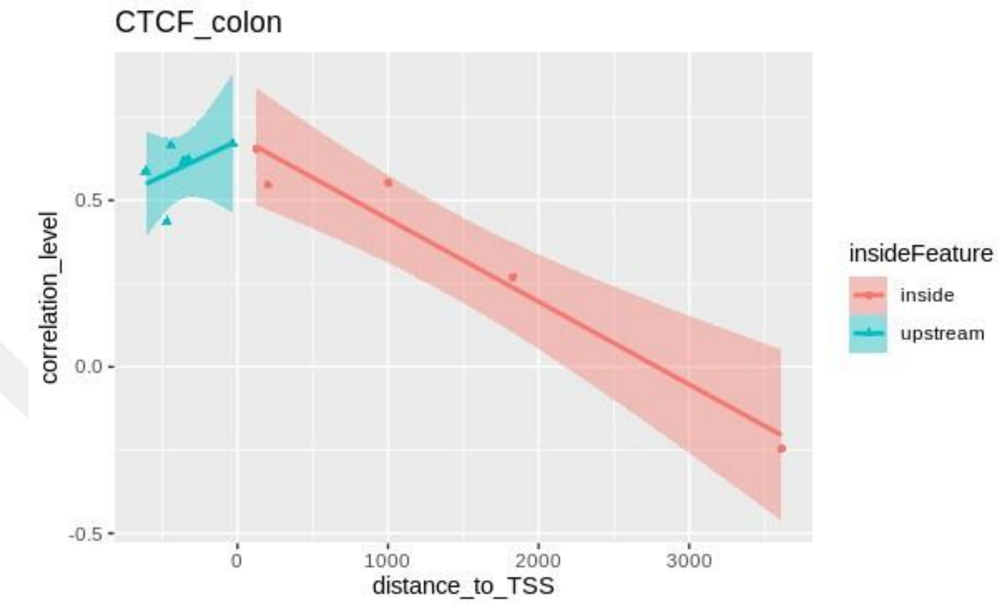


Figure 52: Scatter plot created using healthy colon data

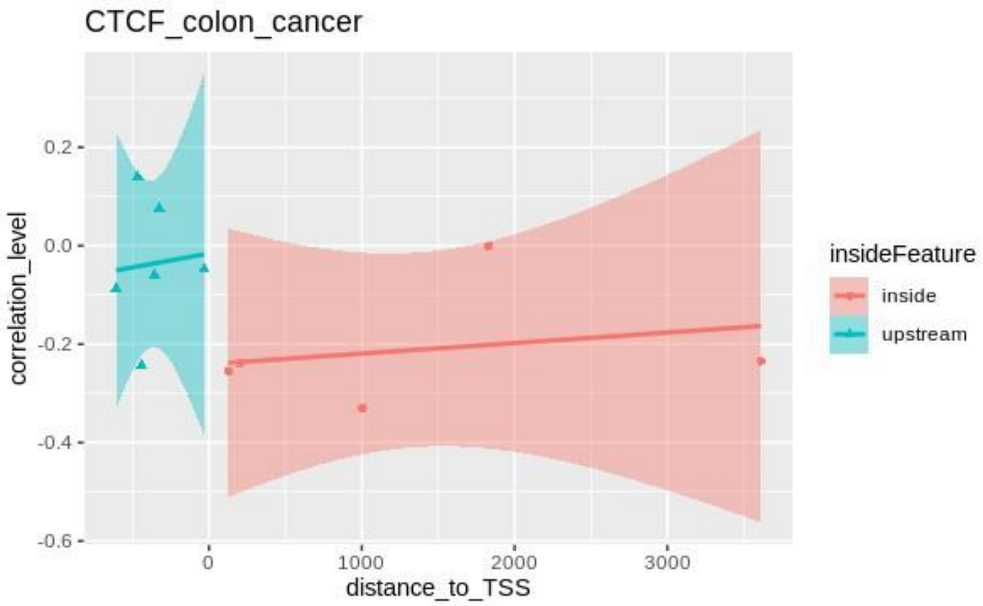


Figure 53: Scatter plot created using colon cancer data

The scatter plot (Figure 54) and table (Table 20) created to better observe the difference between healthy and cancerous tissues are shown below.

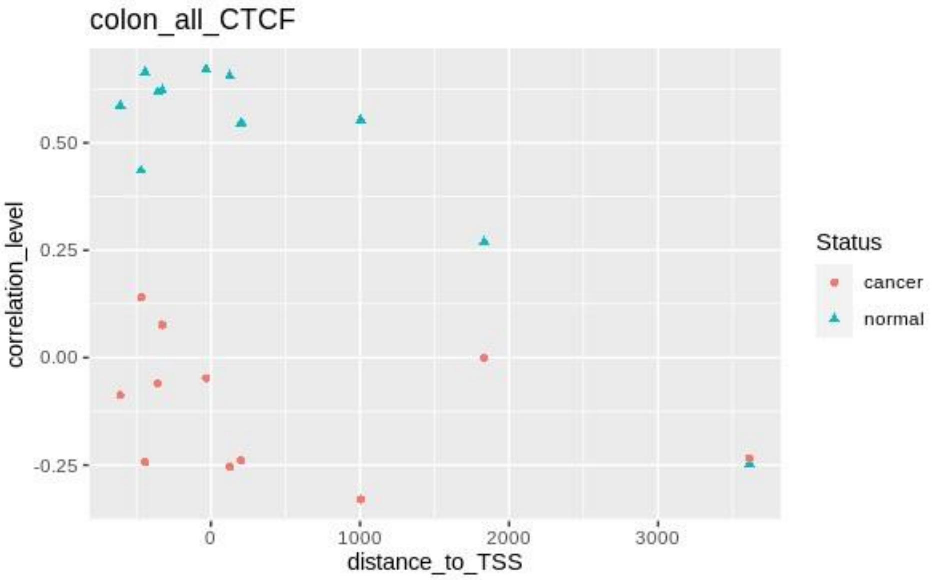


Figure 54: Scatter plot showing both healthy and cancerous colon samples

CG probe ID	Distance to TSS	insideFeature	r in healthy samples (r1)	r in cancer samples (r2)	(r1- r2)	rho in healthy samples (rho1)	rho in cancer samples (rho2)	(rho1- rho2)
cg23858565	-608	upstream	0.59	-0.09	0.67	0.30	-0.18	0.48
cg07967402	-468	upstream	0.44	0.14	0.29	0.22	0.40	-0.18
cg08324636	**-442**	**upstream**	**0.66**	**-0.24**	**0.91**	**0.47**	**-0.11**	**0.58**
cg06241380	**-357**	**upstream**	**0.62**	**-0.06**	**0.68**	**0.55**	**-0.04**	**0.59**
cg10218542	-326	upstream	0.62	0.08	0.55	0.60	0.12	0.48
cg04487155	-32	upstream	0.67	-0.05	0.72	0.51	0.32	0.20
cg10481400	**126**	**inside**	**0.66**	**-0.25**	**0.91**	**0.54**	**-0.06**	**0.60**
cg01866162	203	inside	0.55	-0.24	0.79	0.19	-0.01	0.20
cg27250362	**1005**	**inside**	**0.55**	**-0.33**	**0.88**	**0.46**	**-0.19**	**0.65**
cg02215945	1063	inside	NA	NA	NA	NA	NA	NA
cg16517579	1832	inside	0.27	0.00	0.27	0.37	0.04	0.32
cg04545079	3613	inside	-0.25	-0.23	-0.01	-0.05	-0.34	0.30

Table 20: Correlation of probe regions in the CTCF gene of colon cancer samples. In the columns showing the differences between healthy and cancerous correlation values, values exceeding the 0.5 cut-off threshold are highlighted. In both correlation analyzes, probes exceeding the cut-off threshold are shown as bold.

Looking at this scatter plot, it seems clear that there is an entirely different correlation in cancerous tissue, as cancerous and healthy samples are clustered separately. When the Pearson correlation difference was examined, it was seen that 8 probes exceeded the cut-off threshold, and when the Spearman correlation difference was examined, 4 probes exceeded the cut-off threshold. In addition, when probes that exceed the cut-off threshold were examined in both correlation analyzes, it was observed that 4 probes exceeded the cut-off threshold.

3.2.2.2 Bone

PCA (Figure 55) and heatmap (Figure 56) created using methylation data of healthy and cancerous tissues are shown below.

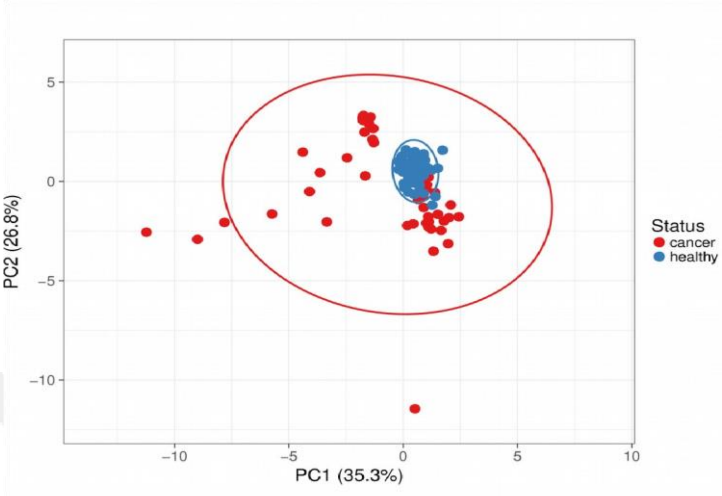


Figure 55: PCA of bone methylation data

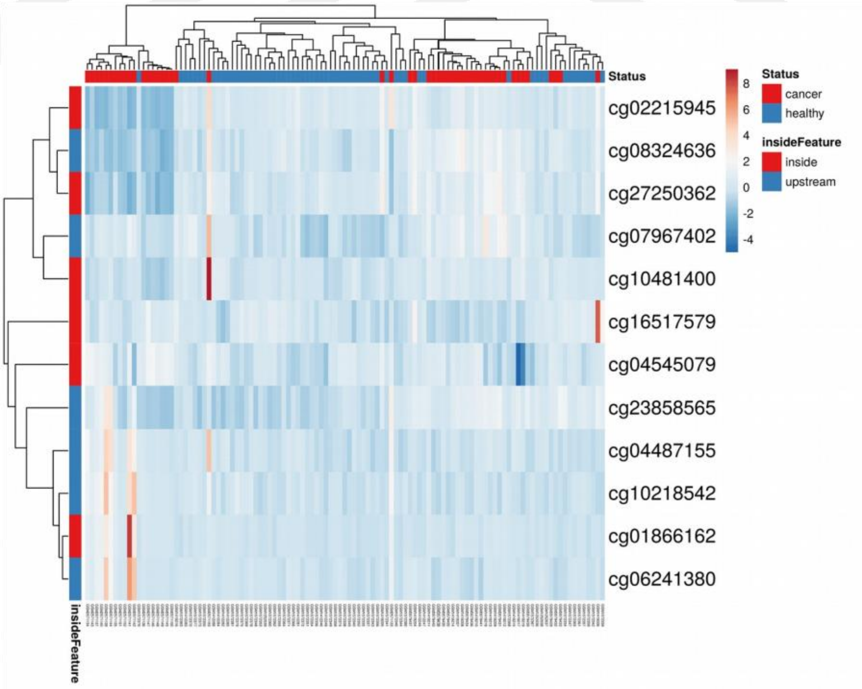


Figure 56: Heatmap of bone methylation data

Scatter plots showing correlation between methylation and gene expression in healthy (Figure 57) and cancerous bone samples (Figure 58) below.

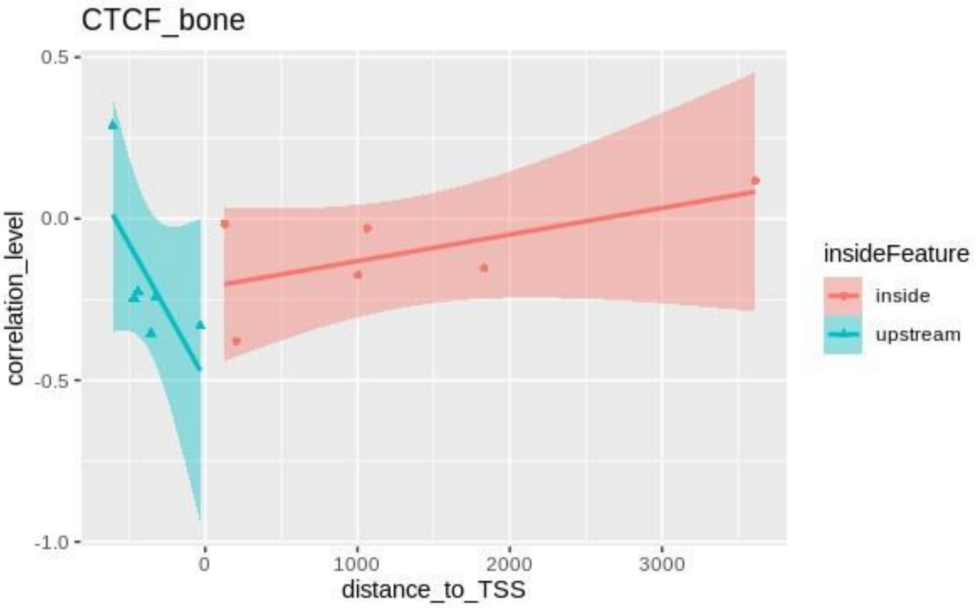


Figure 57: Scatter plot created using healthy bone data

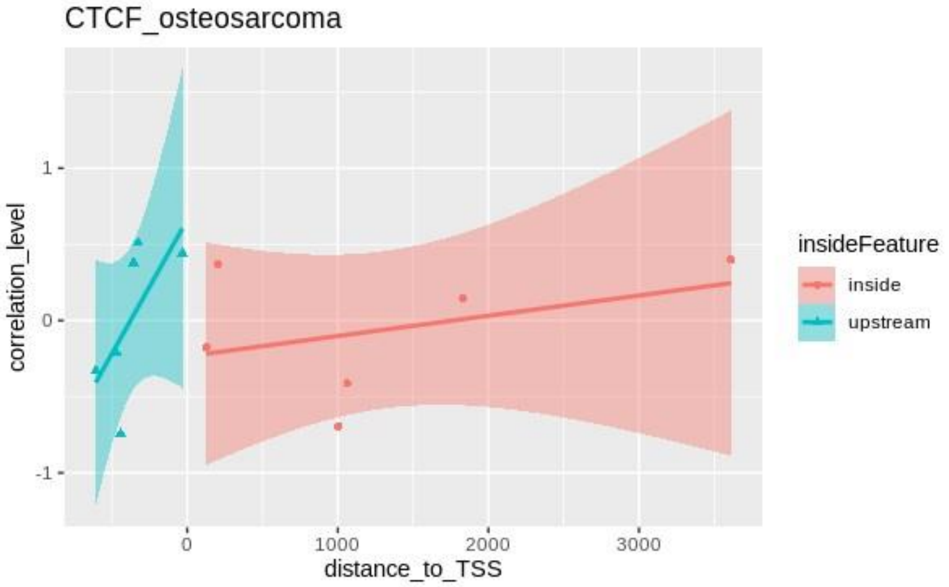


Figure 58: Scatter plot created using bone cancer data

The scatter plot (Figure 59) and table (Table 21) created to better observe the difference between healthy and cancerous tissues are shown below.

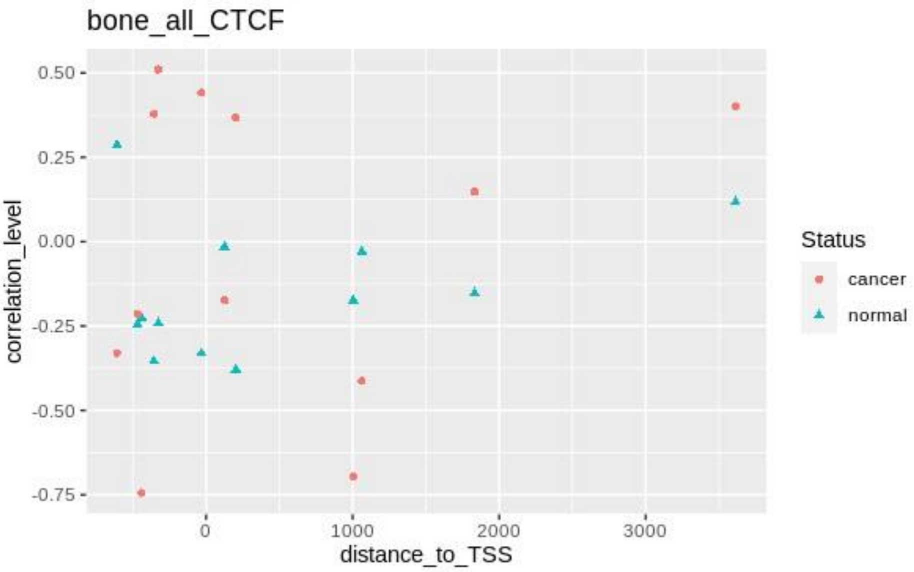


Figure 59: Scatter plot showing both healthy and cancerous bone samples

CG probe ID	Distance to TSS	insideFeature	r in healthy samples (r1)	r in cancer samples (r2)	(r1 - r2)	rho in healthy samples (rho1)	rho in cancer samples (rho2)	(rho1 - rho2)
cg23858565	**-608**	**upstream**	**0.29**	**-0.33**	**0.62**	**0.33**	**-0.37**	**0.70**
cg07967402	-468	upstream	-0.25	-0.21	-0.03	-0.25	-0.44	0.19
cg08324636	**-442**	**upstream**	**-0.22**	**-0.74**	**0.52**	**-0.02**	**-0.68**	**0.66**
cg06241380	**-357**	**upstream**	**-0.35**	**0.38**	**-0.73**	**-0.26**	**0.28**	**-0.54**
cg10218542	**-326**	**upstream**	**-0.24**	**0.51**	**-0.75**	**-0.24**	**0.55**	**-0.79**
cg04487155	-32	upstream	-0.33	0.44	-0.77	-0.20	0.29	-0.49
cg10481400	126	inside	-0.02	-0.17	0.16	-0.16	-0.69	0.53
cg01866162	**203**	**inside**	**-0.38**	**0.37**	**-0.75**	**-0.50**	**0.49**	**-1.00**
cg27250362	1005	inside	-0.17	-0.70	0.52	-0.30	-0.69	0.38
cg02215945	1063	inside	-0.03	-0.41	0.38	-0.04	-0.60	0.56
cg16517579	1832	inside	-0.15	0.15	-0.30	0.03	0.42	-0.39
cg04545079	3613	inside	0.12	0.40	-0.28	0.04	0.49	-0.45

Table 21: Correlation of probe regions in the CTCF gene of bone cancer samples. In the columns showing the differences between healthy and cancerous correlation values, values exceeding the 0.5 cut-off threshold are highlighted. In both correlation analyzes, probes exceeding the cut-off threshold are shown as bold.

While the correlation values of methylation in the CTCF gene in samples taken from healthy tissues in the bone appeared around 0, the correlation values were found to be high when looking at the cancerous tissue. When the Pearson correlation difference was examined, it was seen that 7 probes exceeded the cut-off threshold, and when the Spearman correlation difference was examined, 7 probes exceeded the cut-off threshold. In addition, when probes that exceed the cut-off threshold were examined in both correlation analyzes, it was observed that 5 probes exceeded the cut-off threshold.

3.2.2.3 Breast

PCA (Figure 60) and heatmap (Figure 61) created using methylation data of healthy and cancerous tissues are shown below.

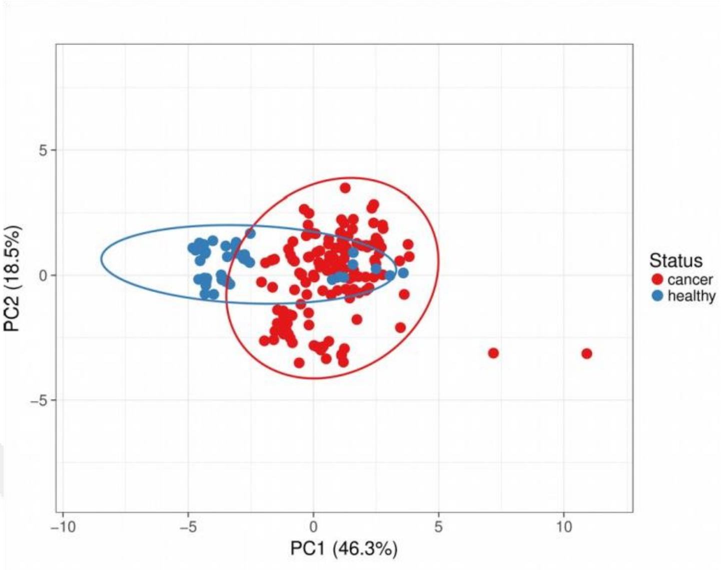


Figure 60: PCA of breast methylation data

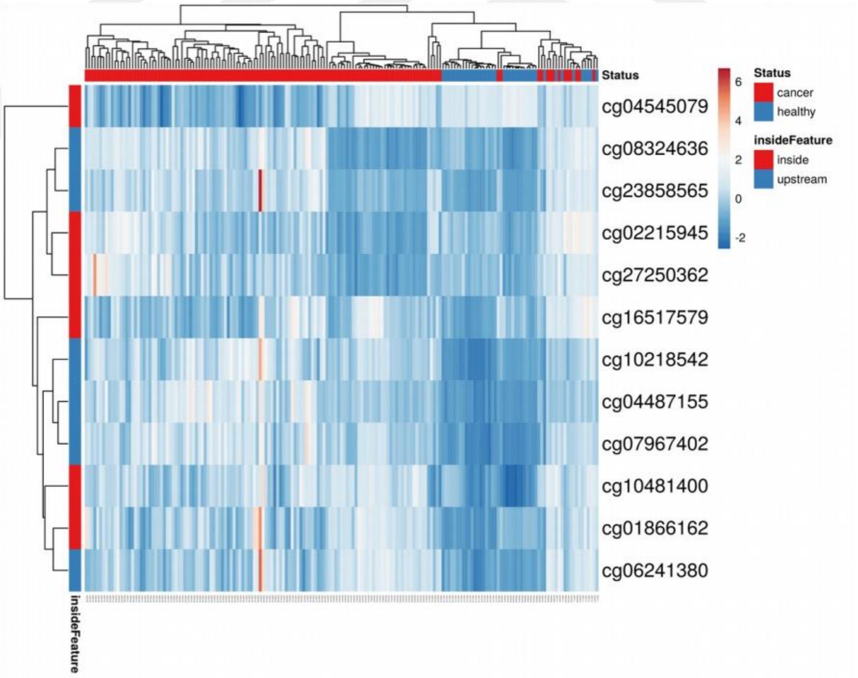


Figure 61: Heatmap of breast methylation data

Scatter plots showing correlation between methylation and gene expression in healthy (Figure 62) and cancerous breast samples (Figure 63) are presented below.

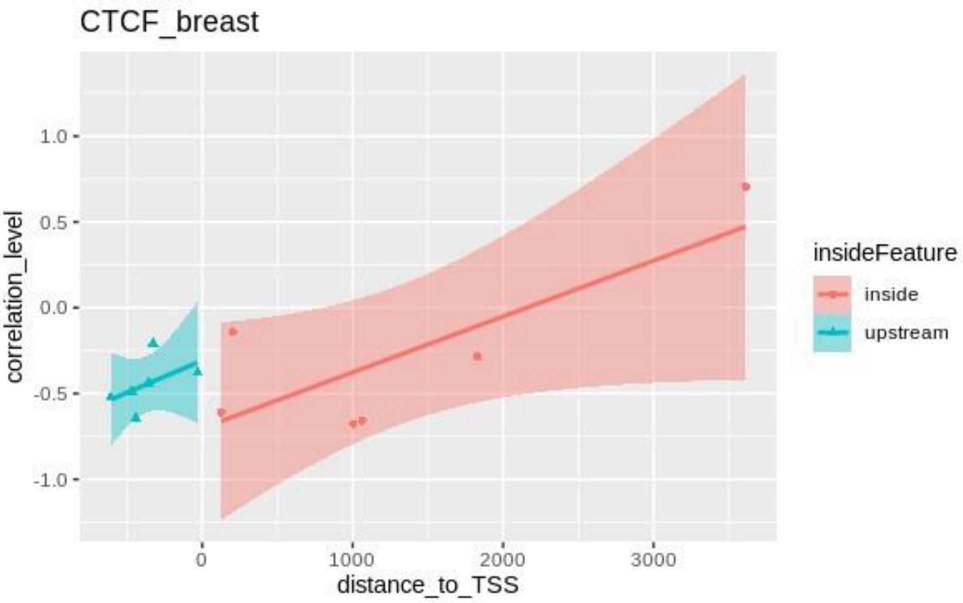


Figure 62: Scatter plot created using healthy breast data

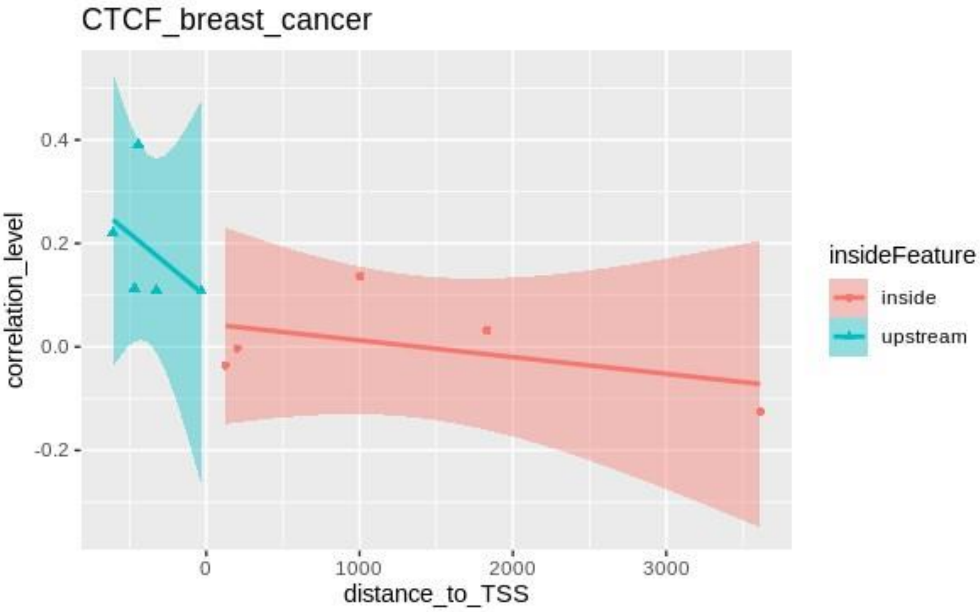


Figure 63: Scatter plot created using breast cancer data

The scatter plot (Figure 64) and table (Table 22) created to better observe the difference between healthy and cancerous tissues are shown below.

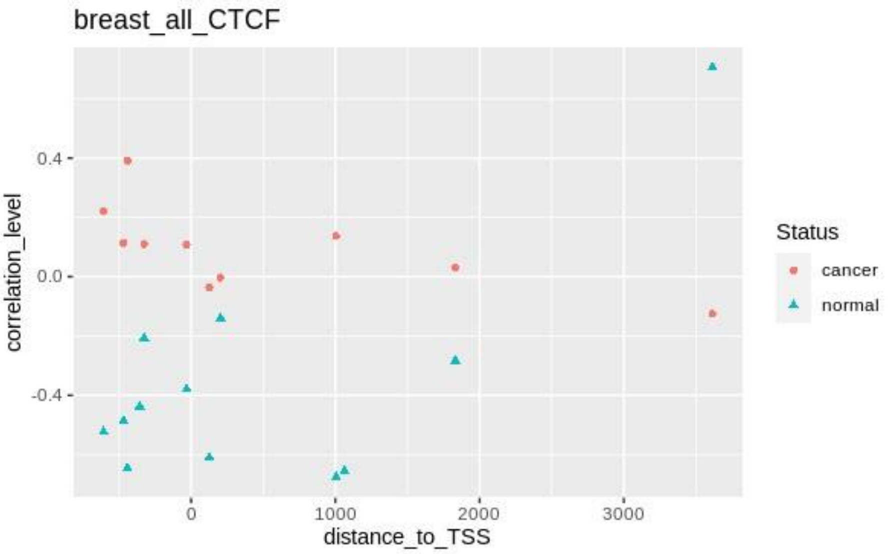


Figure 64: Scatter plot of probe sites in the CTCF gene of breast cancer and healthy samples

CG probe ID	Distance to TSS	insideFeature	r in healthy samples (r1)	r in cancer samples (r2)	(r1- r2)	rho in healthy samples (rho1)	rho in cancer samples (rho2)	(rho1- rho2)
cg23858565	**-608**	**upstream**	**-0.52**	**0.22**	**-0.74**	**-0.37**	**0.41**	**-0.78**
cg07967402	-468	upstream	-0.49	0.11	-0.60	-0.30	0.17	-0.48
cg08324636	**-442**	**upstream**	**-0.65**	**0.39**	**-1.04**	**-0.65**	**0.41**	**-1.06**
cg06241380	-357	upstream	-0.44	NA	NA	-0.30	NA	NA
cg10218542	-326	upstream	-0.21	0.11	-0.32	0.07	0.15	-0.08
cg04487155	-32	upstream	-0.38	0.11	-0.49	-0.24	0.16	-0.40
cg10481400	126	inside	-0.61	-0.04	-0.57	-0.57	-0.14	-0.43
cg01866162	203	inside	-0.14	0.00	-0.14	0.02	-0.15	0.17
cg27250362	**1005**	**inside**	**-0.68**	**0.14**	**-0.81**	**-0.65**	**0.23**	**-0.88**
cg02215945	1063	inside	-0.66	NA	NA	-0.69	NA	NA
cg16517579	1832	inside	-0.28	0.03	-0.32	-0.06	-0.02	-0.04
cg04545079	**3613**	**inside**	**0.71**	**-0.12**	**0.83**	**0.71**	**-0.26**	**0.98**

Table 22: Correlation of probe regions in the CTCF gene of breast cancer samples. In the columns showing the differences between healthy and cancerous correlation values, values exceeding the 0.5 cut-off threshold are highlighted. In both correlation analyzes, probes exceeding the cut-off threshold are shown as bold.

It was observed that cancer samples and healthy ones were clustered separately in the breast, similar to the colon. In the form of a serious distinction, positive correlation was seen in the cancerous tissue, whereas the healthy one had a negative correlation. When the Pearson correlation difference was examined, it was seen that 6 probes exceeded the cut-off threshold, and when the Spearman correlation difference was examined, 4 probes exceeded the cut-off threshold. In addition, when probes that exceed the cut-off threshold were examined in both correlation analyzes, it was observed that 4 probes exceeded the cut-off threshold.

3.2.2.4 Pancreas

PCA (Figure 65) and heatmap (Figure 66) created using methylation data of healthy and cancerous tissues are shown below.

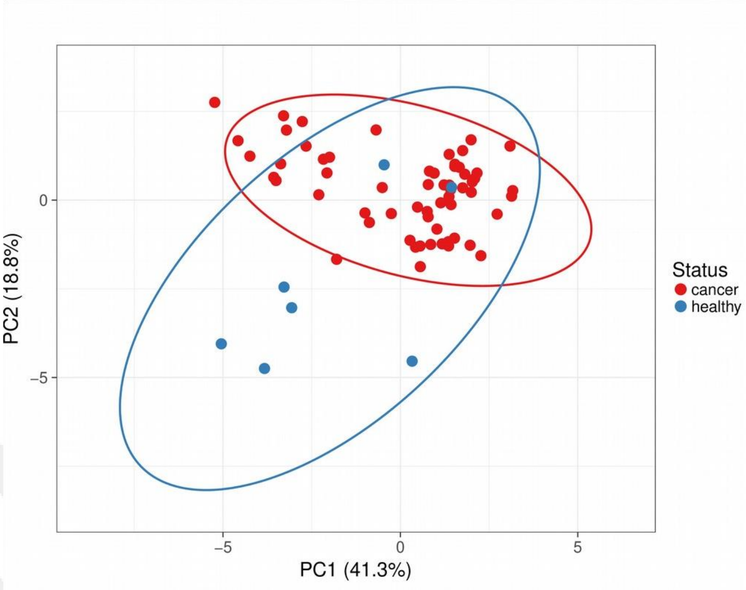


Figure 65: PCA of pancreas methylation data

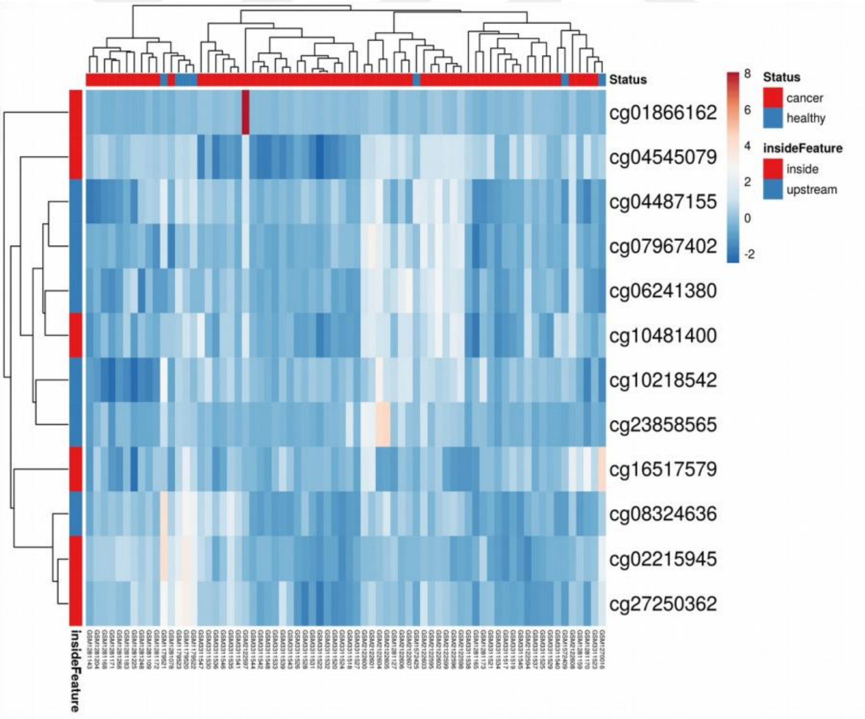


Figure 66: Heatmap of pancreas methylation data

Scatter plots showing correlation between methylation and gene expression in healthy (Figure 67) and cancerous tissues pancreas samples (Figure 68) are shown below.

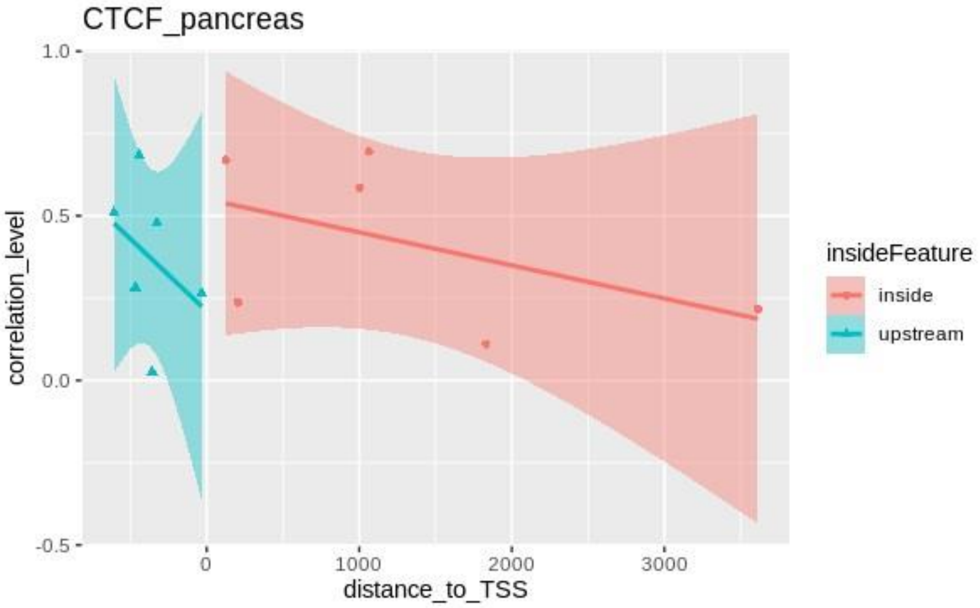


Figure 67: Scatter plot created using healthy pancreas data

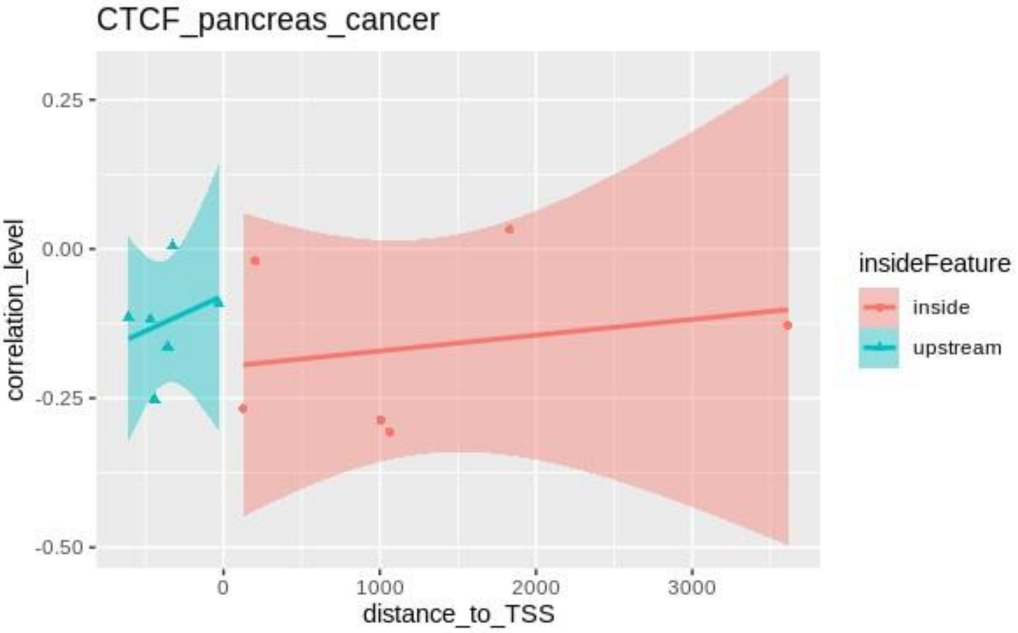


Figure 68: Scatter plot created using pancreas cancer data

The scatter plot (Figure 69) and table (Table 23) created to better observe the difference between healthy and cancerous pancreas samples are shown below.

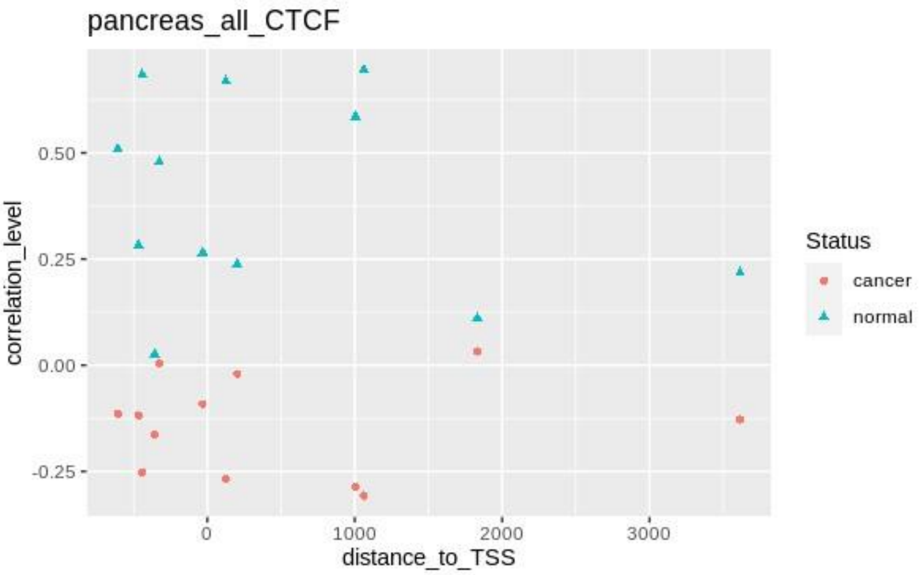


Figure 69: Scatter plot of probe sites in the CTCF gene of pancreas cancer and healthy samples

CG probe ID	Distance to TSS	insideFeature	r in healthy samples (r1)	r in cancer samples (r2)	(r1 - r2)	rho in healthy samples (rho1)	rho in cancer samples (rho2)	(rho1 - rho2)
cg23858565	-608	upstream	0.51	-0.11	0.62	0.46	0.15	0.31
cg07967402	-468	upstream	0.28	-0.12	0.40	0.18	0.10	0.08
cg08324636	**-442**	**upstream**	**0.68**	**-0.25**	**0.94**	**0.57**	**-0.02**	**0.60**
cg06241380	-357	upstream	0.03	-0.16	0.19	-0.07	-0.01	-0.06
cg10218542	-326	upstream	0.48	0.00	0.47	0.71	-0.06	0.77
cg04487155	-32	upstream	0.26	-0.09	0.36	-0.07	-0.16	0.09
cg10481400	**126**	**inside**	**0.67**	**-0.27**	**0.94**	**0.64**	**-0.03**	**0.67**
cg01866162	203	inside	0.24	-0.02	0.26	0.07	-0.17	0.24
cg27250362	**1005**	**inside**	**0.58**	**-0.29**	**0.87**	**0.64**	**-0.08**	**0.72**
cg02215945	1063	inside	0.70	-0.31	1.00	0.46	0.11	0.35
cg16517579	1832	inside	0.11	0.03	0.08	0.11	0.11	0.00

Table 23: Correlation of probe regions in the CTCF gene of pancreas cancer samples. In the columns showing the differences between healthy and cancerous correlation values, values exceeding the 0.5 cut-off threshold are highlighted. In both correlation analyzes, probes exceeding the cut-off threshold are shown as bold.

As in the breast and colon, healthy and cancerous probes were clustered separately. While it normally showed a positive correlation, it showed a negative correlation in cancerous cells. When the Pearson correlation difference was examined, it was seen that 5 probes exceeded the cut-off threshold, and when the Spearman correlation difference was examined, 4 probes exceeded the cut-off threshold. In addition, when probes that exceed the cut-off threshold were examined in both correlation analyzes, it was observed that 3 probes exceeded the cut-off threshold.

3.2.2.5 Prostate

PCA (Figure 70) and heatmap (Figure 71) created using methylation data of healthy and cancerous tissues are shown below.

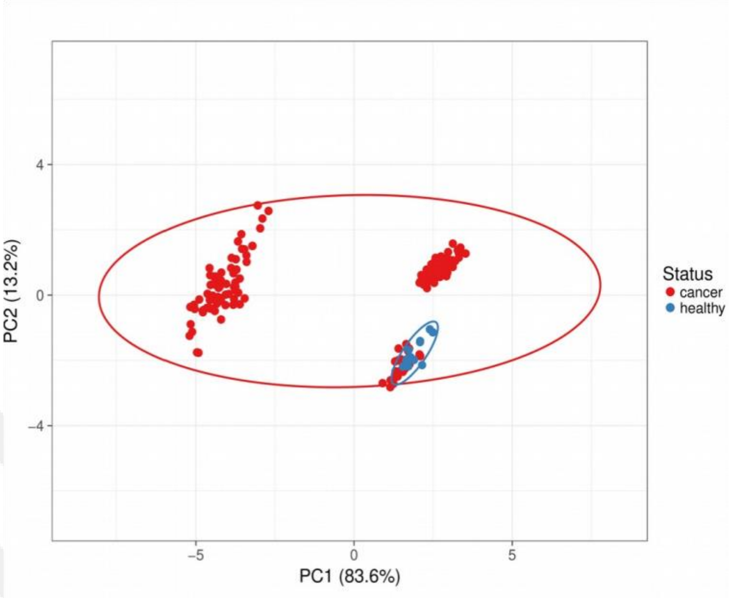


Figure 70: PCA of prostate methylation data

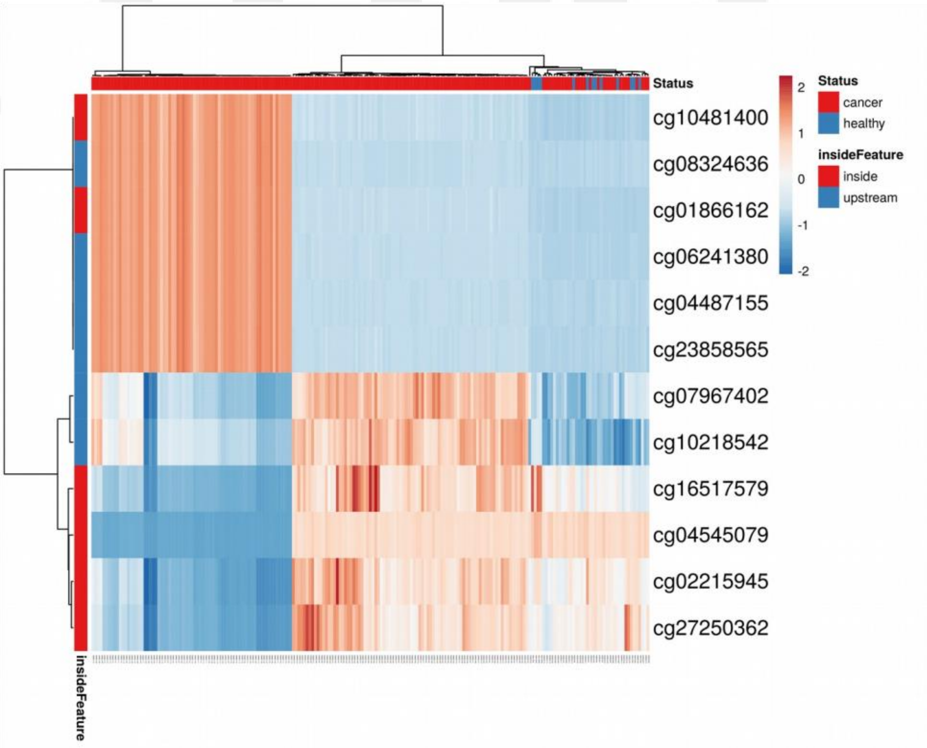


Figure 71: Heatmap of prostate methylation data

Scatter plot showing outcome of correlation analysis of methylation and gene expression in normal (Figure 72) and cancerous prostate samples (Figure 73) are included below.

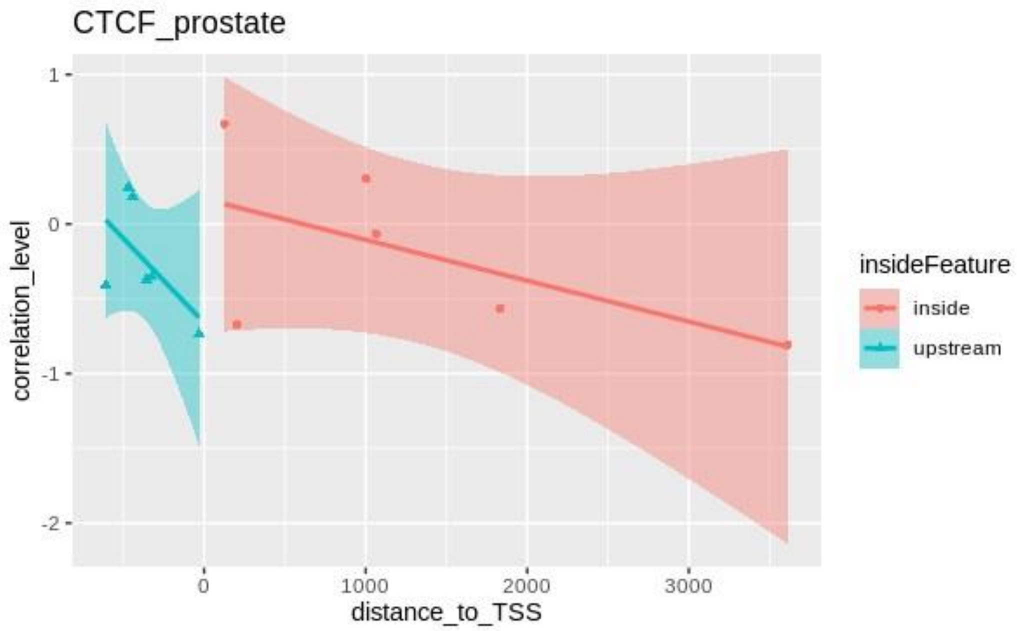


Figure 72: Scatter plot created using healthy prostate data

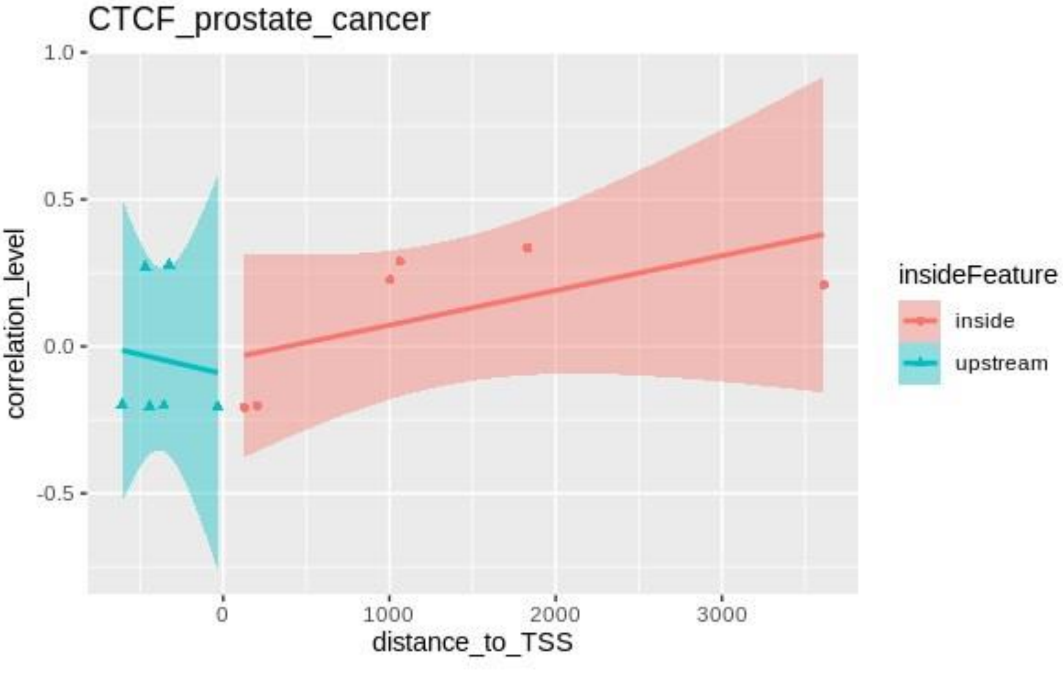


Figure 73: Scatter plot created using prostate cancer data

The scatter plot (Figure 74) and table (Table 24) created to better observe the difference between healthy and cancerous prostate samples are shown provided below.

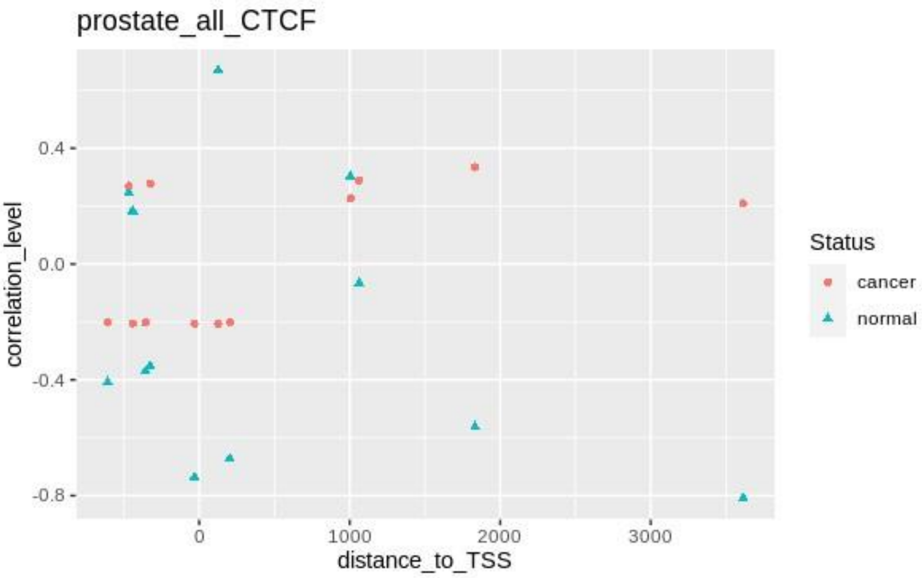


Figure 74: Correlation of probe regions in the CTCF gene of prostate cancer samples

CG probe ID	Distance to TSS	insideFeature	r in healthy samples (r1)	r in cancer samples (r2)	(r1 - r2)	rho in healthy samples (rho1)	rho in cancer samples (rho2)	(rho1 - rho2)
cg23858565	-608	upstream	-0.41	-0.20	-0.21	-0.25	0.29	-0.54
cg07967402	-468	upstream	0.25	0.27	-0.02	0.16	-0.19	0.36
cg08324636	-442	upstream	0.18	-0.21	0.39	0.38	0.30	0.08
cg06241380	-357	upstream	-0.37	-0.20	-0.17	-0.17	0.27	-0.44
cg10218542	-326	upstream	-0.35	0.28	-0.63	-0.66	-0.20	-0.46
cg04487155	**-32**	**upstream**	**-0.74**	**-0.21**	**-0.53**	**-0.75**	**0.25**	**-1.00**
cg10481400	126	inside	0.67	-0.21	0.88	0.59	0.19	0.40
cg01866162	203	inside	-0.67	-0.20	-0.47	-0.79	0.24	-1.03
cg27250362	1005	inside	0.30	0.23	0.08	0.70	-0.25	0.95
cg02215945	1063	inside	-0.07	0.29	-0.35	0.23	-0.23	0.46
cg16517579	1832	inside	-0.56	0.33	-0.90	-0.70	-0.25	-0.45
cg04545079	**3613**	**inside**	**-0.81**	**0.21**	**-1.02**	**-0.76**	**-0.21**	**-0.55**

Table 24: Correlation of probe regions in the CTCF gene of prostate cancer samples. In the columns showing the differences between healthy and cancerous correlation values, values exceeding the 0.5 cut-off threshold are highlighted. In both correlation analyzes, probes exceeding the cut-off threshold are shown as bold.

Negative correlation was seen in the healthy prostate, while positive correlation was observed in prostate cancer. When the Pearson correlation difference was examined, it was seen that 5 probes exceeded the cut-off threshold, and when the Spearman correlation difference was examined, 5 probes exceeded the cut-off threshold. In addition, when probes that exceed the cut-off threshold were examined in both correlation analyzes, it was observed that 2 probes exceeded the cut-off threshold.

3.2.2.6 Bladder

PCA (Figure 75) and heatmap (Figure 76) created using methylation data of healthy and cancerous tissues are shown below.

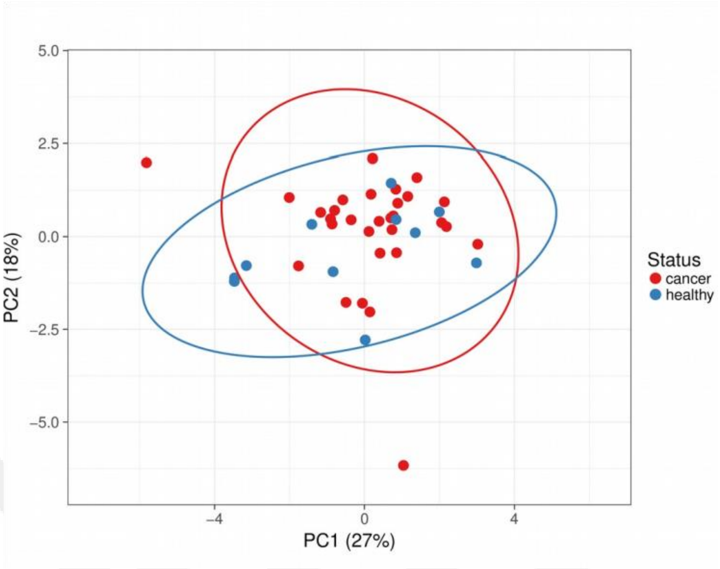


Figure 75: PCA of bladder methylation data

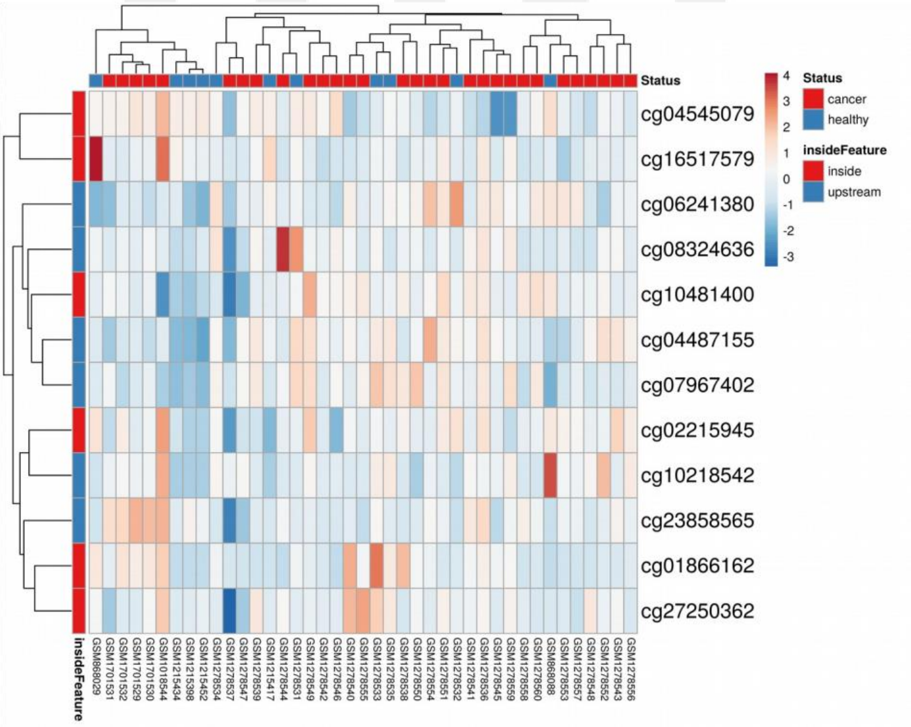


Figure 76: Heatmap of bladder methylation data

Scatter plots showing results of correlation analysis of methylation and gene expression in healthy (Figure 77) and cancerous bladder samples (Figure 78) are shown presented below.

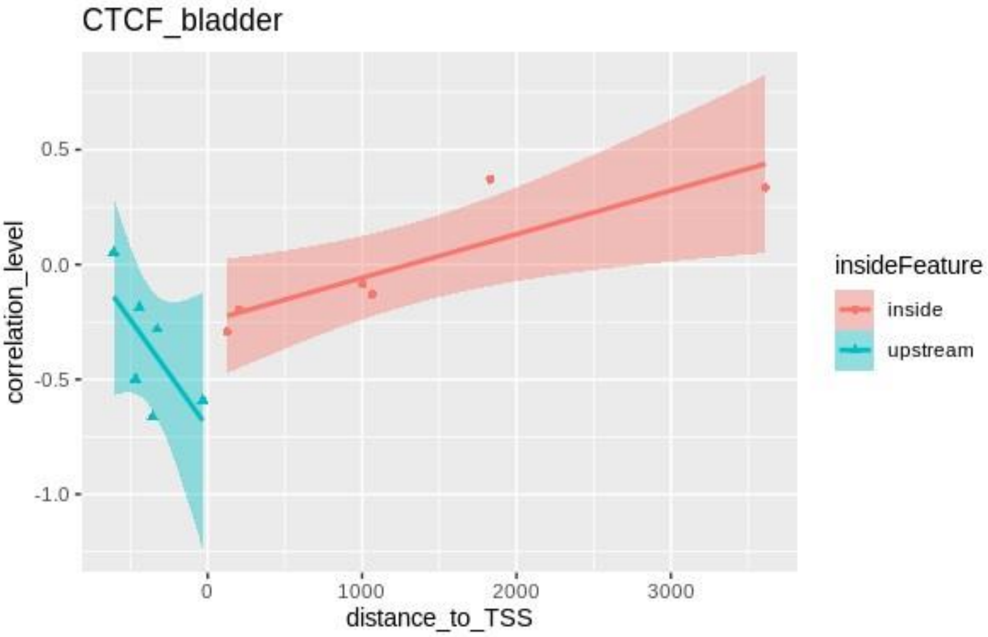


Figure 77: Scatter plot created using healthy bladder data

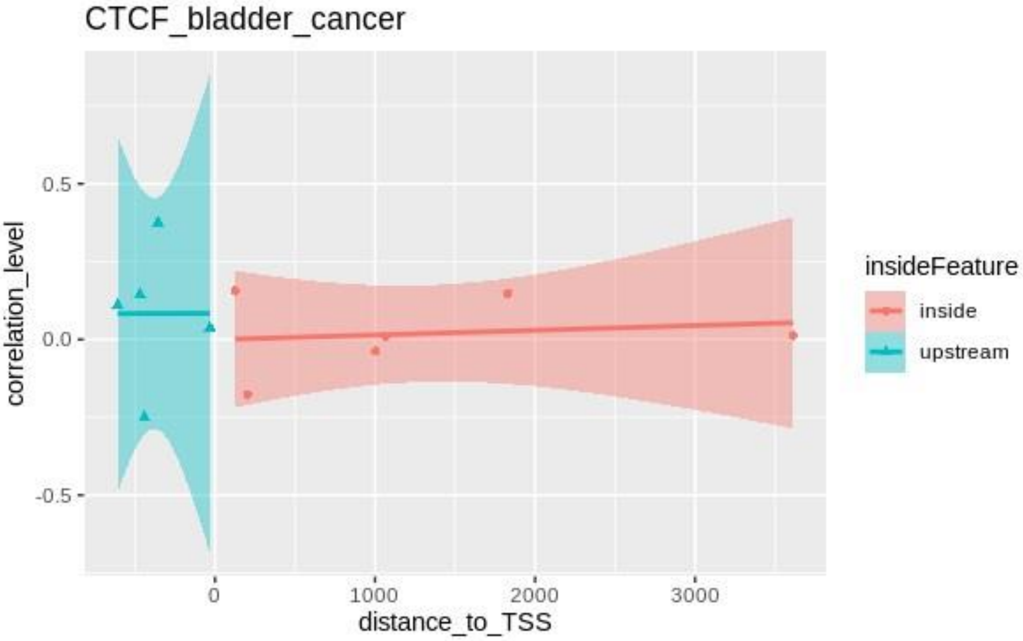


Figure 78: Scatter plot created using bladder cancer data

The scatter plot (Figure 79) and table (Table 25) created to better observe the difference between healthy and cancerous tissues are shown below.

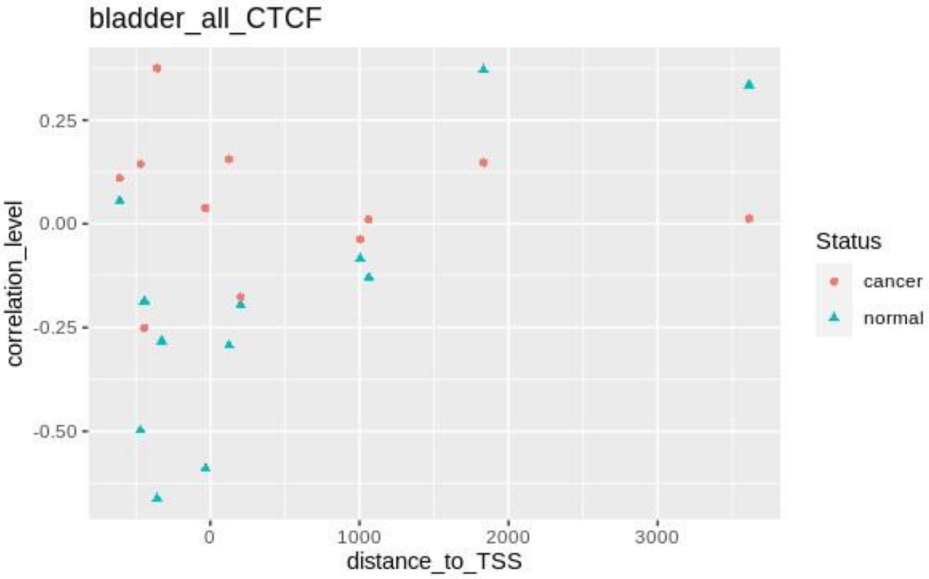


Figure 79: Correlation of probe regions in the CTCF gene of bladder cancer samples

CG probe ID	Distance to TSS	insideFeature	r in healthy samples (r1)	r in cancer samples (r2)	(r1 - r2)	rho in healthy samples (rho1)	rho in cancer samples (rho2)	(rho1 - rho2)
cg23858565	-608	upstream	0.05	0.11	-0.06	0.02	0.20	-0.19
cg07967402	**-468**	**upstream**	**-0.50**	**0.14**	**-0.64**	**-0.68**	**0.13**	**-0.81**
cg08324636	-442	upstream	-0.19	-0.25	0.06	-0.21	-0.48	0.28
cg06241380	**-357**	**upstream**	**-0.66**	**0.38**	**-1.04**	**-0.51**	**0.21**	**-0.71**
cg10218542	-326	upstream	-0.28	NA	NA	-0.25	NA	NA
cg04487155	**-32**	**upstream**	**-0.59**	**0.04**	**-0.63**	**-0.70**	**-0.11**	**-0.59**
cg10481400	126	inside	-0.29	0.16	-0.45	-0.43	0.11	-0.53
cg01866162	203	inside	-0.20	-0.18	-0.02	-0.33	-0.06	-0.27
cg27250362	1005	inside	-0.08	-0.04	-0.05	-0.32	-0.17	-0.15
cg02215945	1063	inside	-0.13	0.01	-0.14	-0.21	0.02	-0.23
cg16517579	1832	inside	0.37	0.15	0.22	0.30	0.26	0.04
cg04545079	3613	inside	0.33	0.01	0.32	0.54	0.05	0.49

Table 25: Correlation of probe regions in the CTCF gene of bladder cancer samples. In the columns showing the differences between healthy and cancerous correlation values, values exceeding the 0.5 cut-off threshold are highlighted. In both correlation analyzes, probes exceeding the cut-off threshold are shown as bold.

When the Pearson correlation difference was examined, it was seen that 3 probes exceeded the cut-off threshold, and when the Spearman correlation difference was examined, 4 probes exceeded the cut-off threshold. In addition, when probes that exceed the cut-off threshold were examined in both correlation analyzes, it was observed that 4 probes exceeded the cut-off threshold.

3.2.2.7 Gastric

PCA (Figure 80) and heatmap (Figure 81) created using methylation data of healthy and cancerous tissues are shown below.

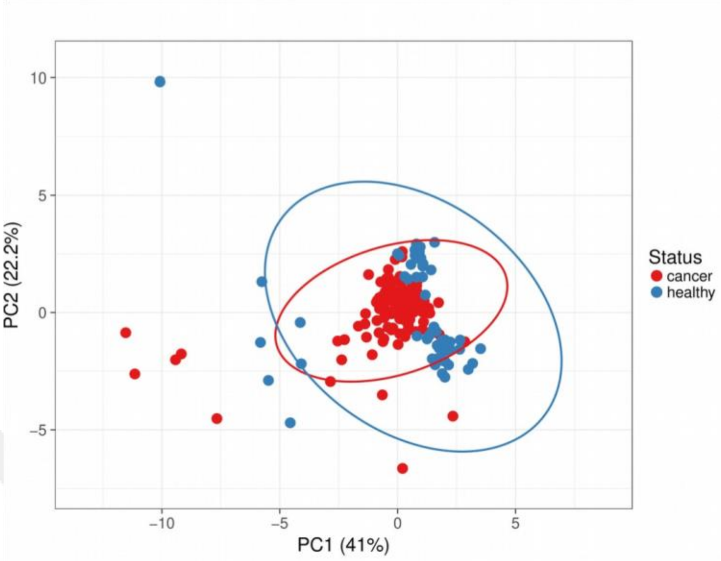


Figure 80: PCA of gastric methylation data

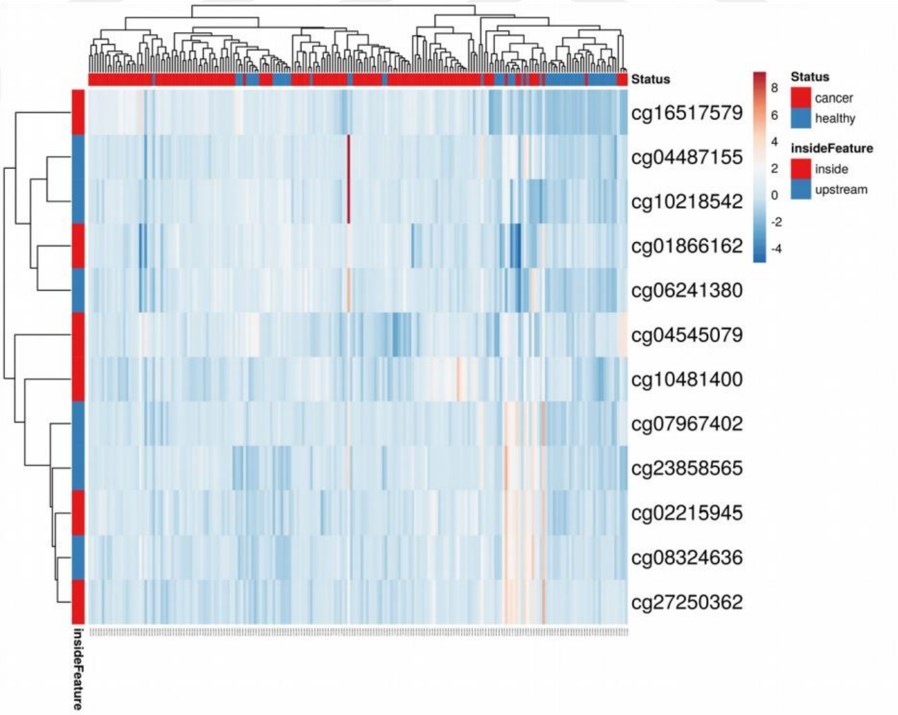


Figure 81: Heatmap of gastric methylation data

Scatter plots showing results of correlation analysis of methylation and gene expression data of healthy (Figure 82) and cancerous gastric samples (Figure 83) are included below.

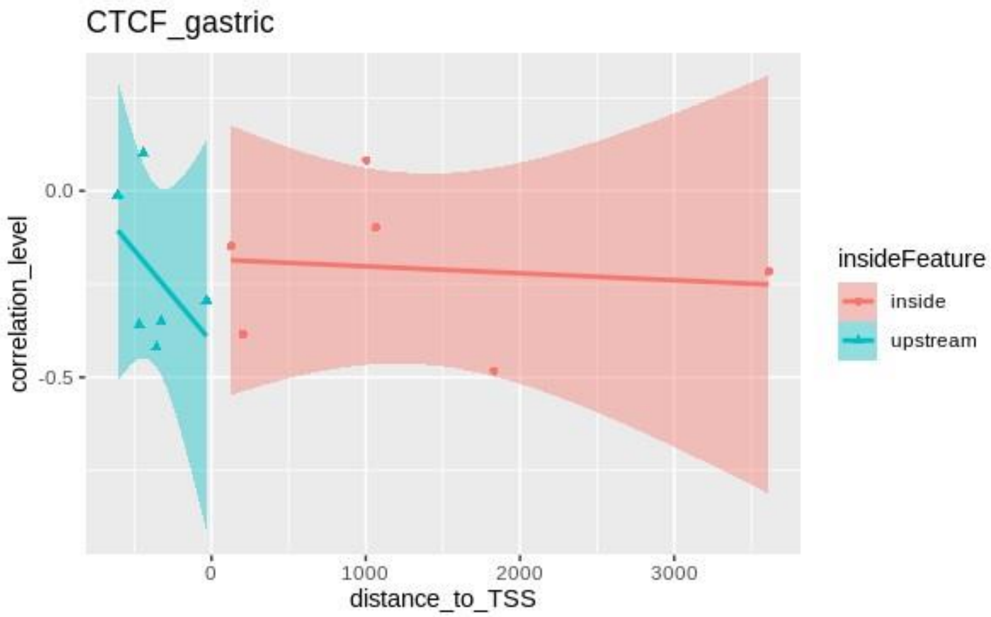


Figure 82: Scatter plot created using healthy gastric data

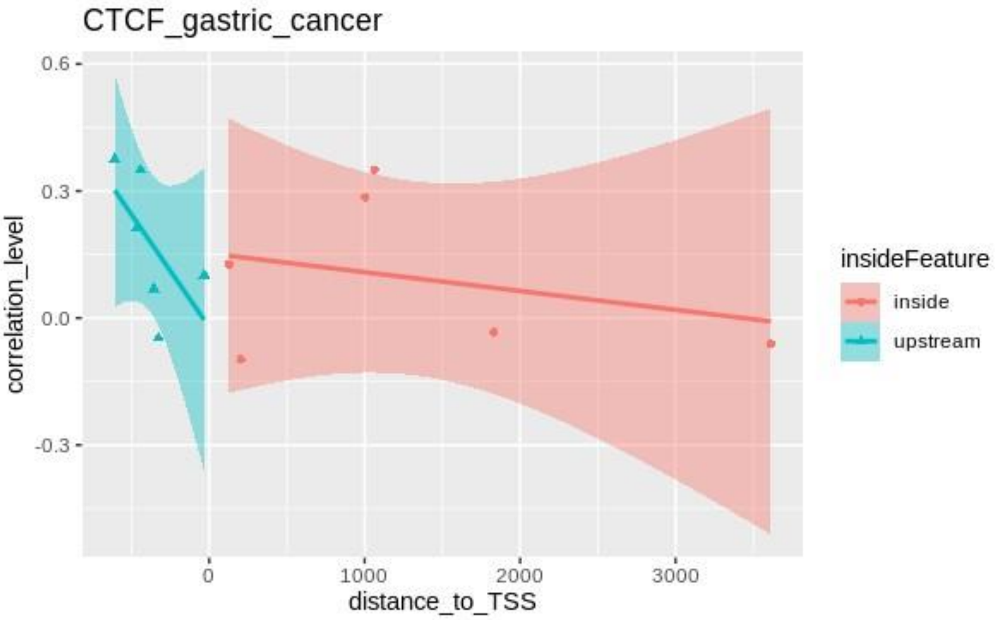


Figure 83: Scatter plot created using gastric cancer data

The scatter plot (Figure 84) and table (Table 26) created to better observe the difference between healthy and cancerous tissues are shown below.

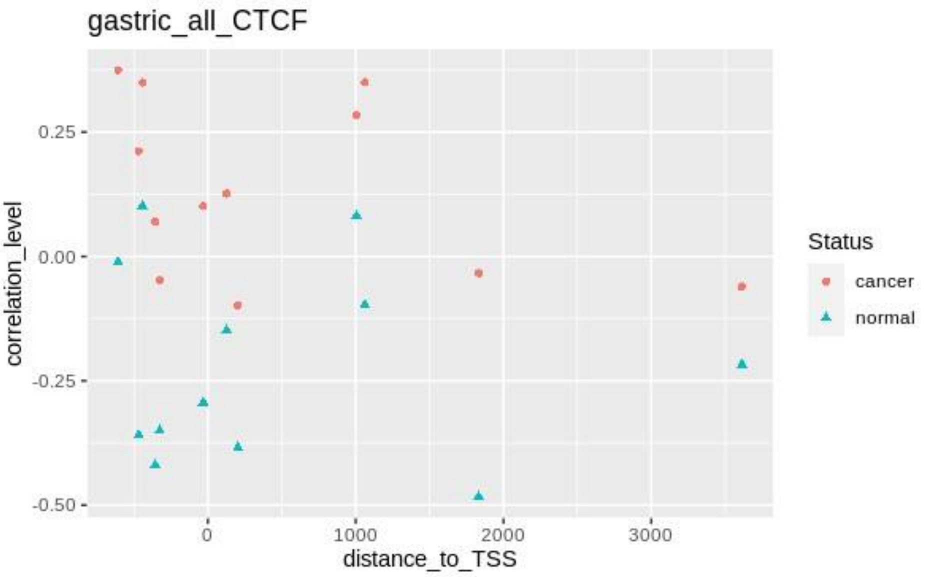


Figure 84: Correlation of probe regions in the CTCF gene of gastric cancer samples

CG probe ID	Distance to TSS	insideFeature	r in healthy samples (r1)	r in cancer samples (r2)	(r1 - r2)	rho in healthy samples (rho1)	rho in cancer samples (rho2)	(rho1 - rho2)
cg23858565	-608	upstream	-0.01	0.37	-0.39	0.17	0.21	-0.03
cg07967402	**-468**	**upstream**	**-0.36**	**0.21**	**-0.57**	**-0.53**	**0.06**	**-0.59**
cg08324636	-442	upstream	0.10	0.35	-0.25	0.39	0.17	0.22
cg06241380	-357	upstream	-0.42	0.07	-0.49	-0.57	0.12	-0.70
cg10218542	-326	upstream	-0.35	-0.05	-0.30	-0.56	0.01	-0.57
cg04487155	-32	upstream	-0.29	0.10	-0.39	-0.57	0.05	-0.61
cg10481400	126	inside	-0.15	0.13	-0.27	-0.11	0.12	-0.23
cg01866162	203	inside	-0.38	-0.10	-0.29	-0.51	-0.01	-0.51
cg27250362	1005	inside	0.08	0.28	-0.20	0.30	0.22	0.08
cg02215945	1063	inside	-0.10	0.35	-0.45	-0.06	0.23	-0.29
cg16517579	1832	inside	-0.48	-0.03	-0.45	-0.46	-0.11	-0.35
cg04545079	3613	inside	-0.22	-0.06	-0.16	-0.27	-0.14	-0.12

Table 26: Correlation of probe regions in the CTCF gene of gastric cancer samples. In the columns showing the differences between healthy and cancerous correlation values, values exceeding the 0.5 cut-off threshold are highlighted. In both correlation analyzes, probes exceeding the cut-off threshold are shown as bold.

When the Pearson correlation difference was examined, it was seen that a probe exceeded the cut-off threshold, and when the Spearman correlation difference was examined, 5 probes exceeded the cut-off threshold. In addition, when probes that exceed the cut-off threshold were examined in both correlation analyzes, it was observed that a probe exceeded the cut-off threshold.

3.2.2.8 Small intestine

PCA (Figure 85) and heatmap (Figure 86) created using methylation data of healthy and cancerous tissues are shown below.

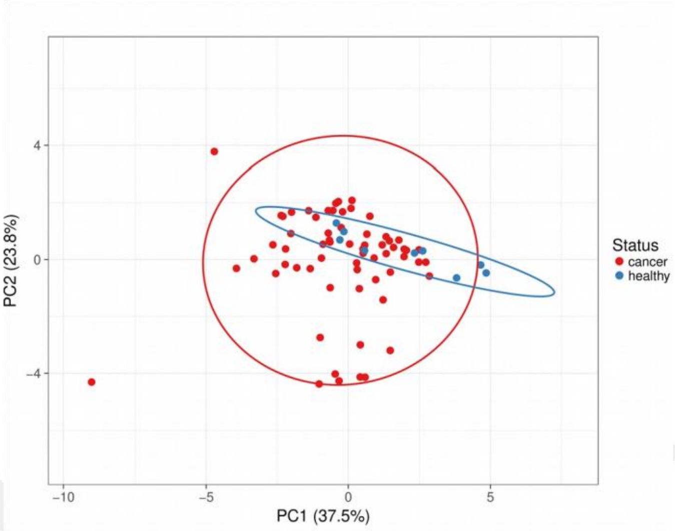


Figure 85: PCA of small intestine methylation data

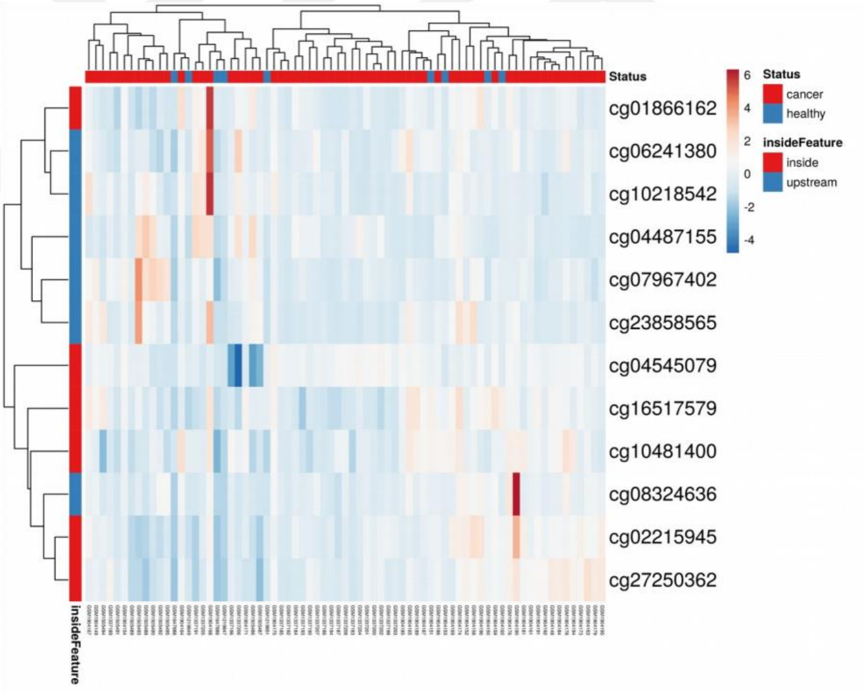


Figure 86: Heatmap of small intestine methylation data

The outcome of correlation analysis of methylation and gene expression data in healthy (Figure 87) and cancerous small intestine samples (Figure 88) are presented in the scatter plots below.

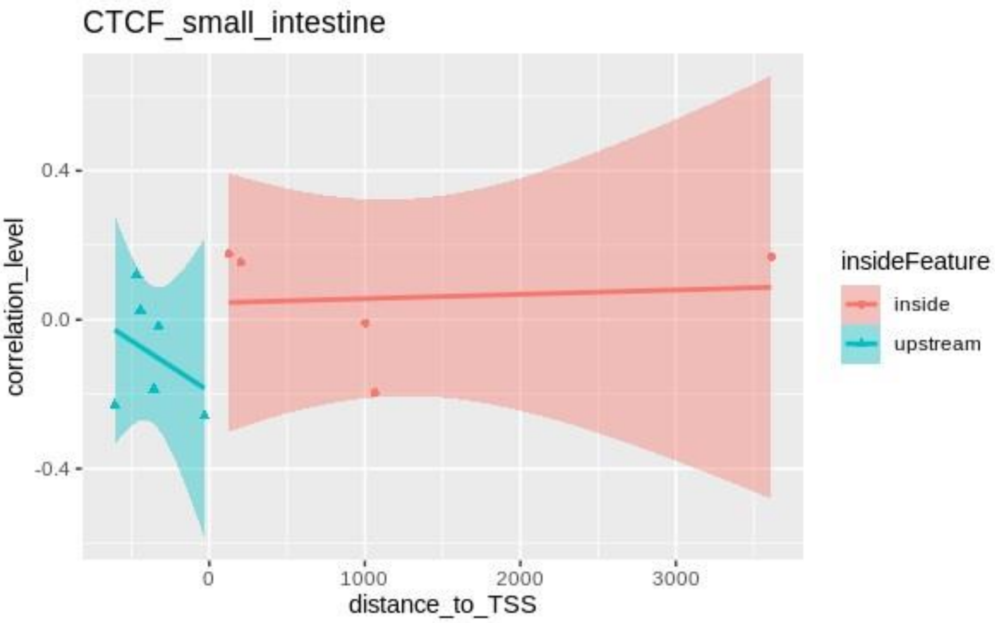


Figure 87: Scatter plot created using healthy small intestine data

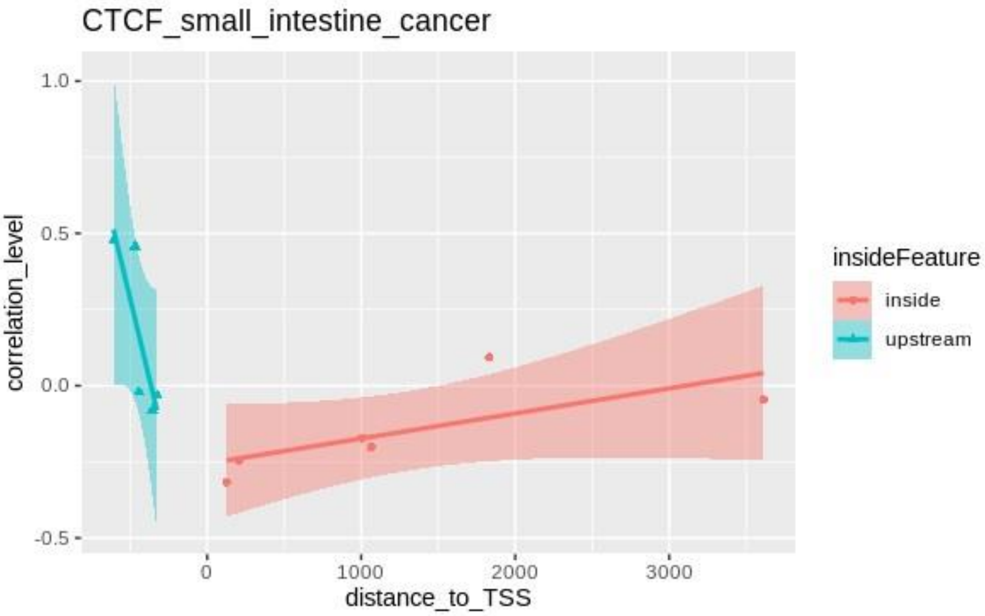


Figure 88: Scatter plot created using small intestine cancer data

The scatter plot (Figure 89) and table (Table 27) created to better observe the difference between healthy and cancerous tissues are shown below.

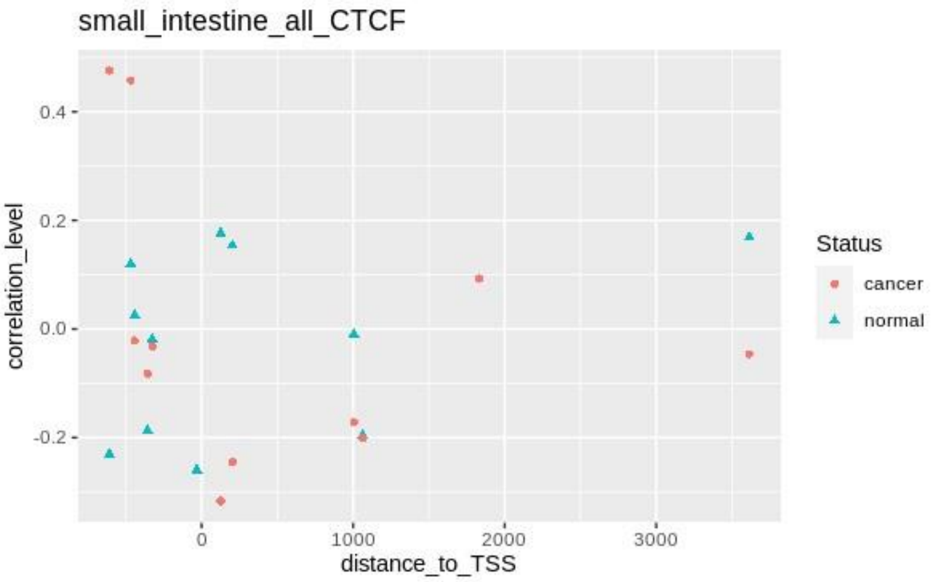


Figure 89: Correlation of probe regions in the CTCF gene of small intestine cancer samples

CG probe ID	Distance to TSS	insideFeature	r in healthy samples (r1)	r in cancer samples (r2)	(r1 - r2)	rho in healthy samples (rho1)	rho in cancer samples (rho2)	(rho1 - rho2)
cg23858565	-608	upstream	-0.23	0.48	-0.71	0.12	-0.13	0.25
cg07967402	-468	upstream	0.12	0.46	0.34	0.22	-0.06	0.28
cg08324636	-442	upstream	0.03	-0.02	-0.05	0.18	-0.18	0.37
cg06241380	-357	upstream	-0.19	-0.08	0.10	-0.10	-0.09	-0.01
cg10218542	-326	upstream	-0.02	-0.03	-0.01	-0.40	0.10	-0.50
cg04487155	-32	upstream	-0.26	NA	NA	-0.18	NA	NA
cg10481400	126	inside	0.18	-0.32	-0.49	0.48	-0.32	0.80
cg01866162	203	inside	0.15	-0.25	-0.40	-0.15	-0.10	-0.05
cg27250362	1005	inside	-0.01	-0.17	-0.16	-0.02	-0.25	0.23
cg02215945	1063	inside	-0.20	-0.20	-0.01	-0.15	-0.29	0.14
cg16517579	1832	inside	NA	0.09	NA	NA	-0.33	NA
cg04545079	3613	inside	0.17	-0.05	-0.22	0.10	-0.03	0.13

Table 27: Correlation of probe regions in the CTCF gene of small intestine cancer samples. In the columns showing the differences between healthy and cancerous correlation values, values exceeding the 0.5 cut-off threshold are highlighted

When the Pearson correlation difference was examined, it was seen that 2 probes exceeded the cut-off threshold, and when the Spearman correlation difference was examined, a probe exceeded the cut-off threshold. In addition, when probes exceeding the cut-off threshold were examined in both correlation analyzes, it was seen that no probe exceeded the cut-off threshold.

3.2.2.9 Brain

PCA (Figure 90) and heatmap (Figure 91) created using methylation data of healthy and cancerous tissues are shown below.

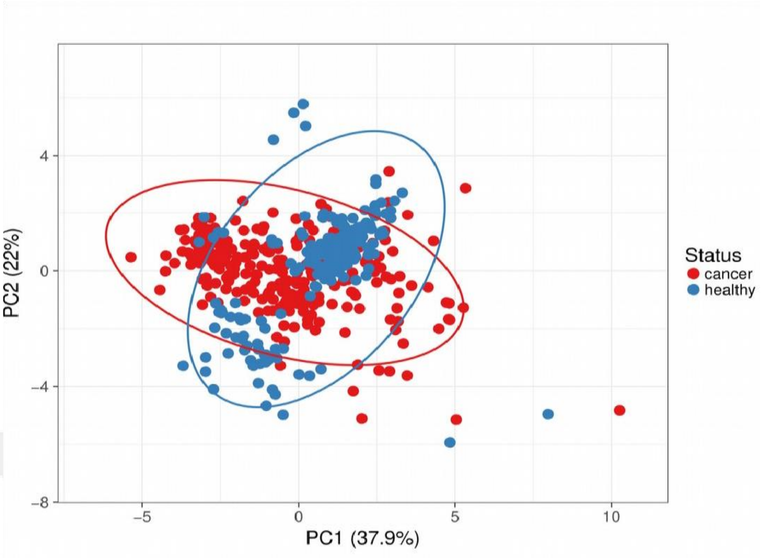


Figure 90: PCA of brain methylation data

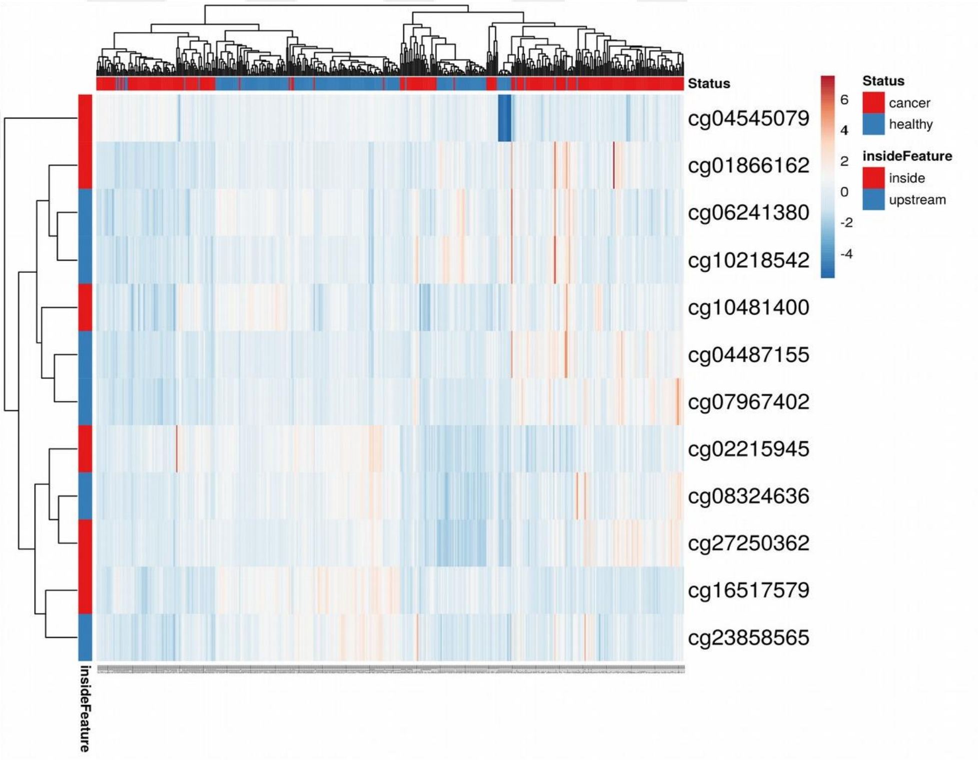


Figure 91: Heatmap of brain methylation data

Scatter plots for correlation analysis of methylation and gene expression in healthy (Figure 92) and cancerous brain samples (Figure 93) are provided below.

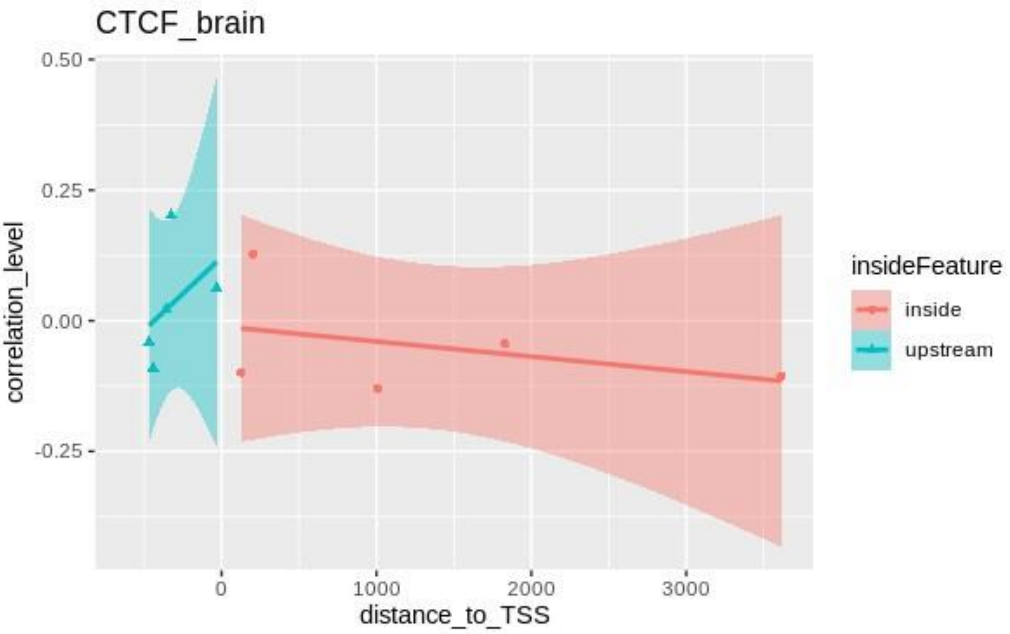


Figure 92: Scatter plot created using healthy brain data

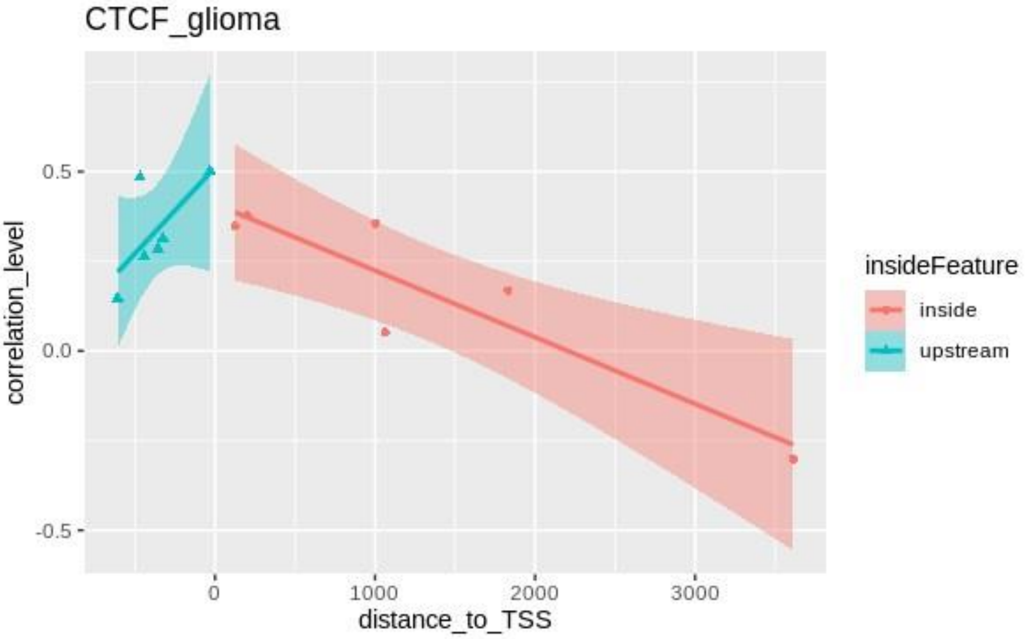


Figure 93: Scatter plot created using brain cancer data

The scatter plot (Figure 94) and table (Table 28) created to better observe the difference between healthy and cancerous tissues are shown below.

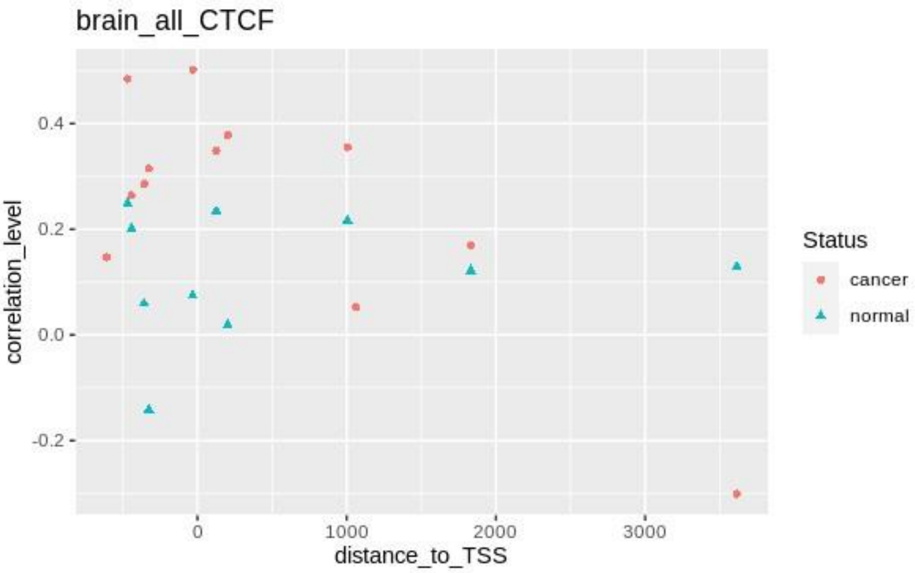


Figure 94: Correlation of probe regions in the CTCF gene of brain cancer samples

CG probe ID	Distance to TSS	insideFeature	r in healthy samples (r1)	r in cancer samples (r2)	(r1 - r2)	rho in healthy samples (rho1)	rho in cancer samples (rho2)	(rho1 - rho2)
cg23858565	-608	upstream	NA	0.15	NA	NA	0.14	NA
cg07967402	-468	upstream	0.25	0.48	-0.24	0.03	0.47	-0.45
cg08324636	-442	upstream	0.20	0.26	-0.06	0.11	0.22	-0.12
cg06241380	-357	upstream	0.06	0.29	-0.23	0.05	0.27	-0.22
cg10218542	-326	upstream	-0.14	0.31	-0.46	-0.18	0.34	-0.52
cg04487155	-32	upstream	0.07	0.50	-0.43	0.03	0.47	-0.43
cg10481400	126	inside	0.23	0.35	-0.11	0.21	0.31	-0.10
cg01866162	203	inside	0.02	0.38	-0.36	-0.03	0.41	-0.44
cg27250362	1005	inside	0.22	0.36	-0.14	0.07	0.34	-0.26
cg02215945	1063	inside	NA	0.05	NA	NA	0.03	NA
cg16517579	1832	inside	0.12	0.17	-0.05	0.05	0.15	-0.11
cg04545079	3613	inside	0.13	-0.30	0.43	-0.02	-0.29	0.27

Table 28: Correlation of probe regions in the CTCF gene of brain cancer samples. In the columns showing the differences between healthy and cancerous correlation values, values exceeding the 0.5 cut-off threshold are highlighted.

When the Pearson correlation difference was examined, it was seen that none of the probes exceeded the cut-off threshold, and when the Spearman correlation difference was examined, a probe exceeded the cut-off threshold. Therefore, there is no common probe crossing the cut-off threshold in both correlation analyzes.

3.2.2.10 Kidney

PCA (Figure 95) and heatmap (Figure 96) created using methylation data of healthy and cancerous tissues are shown below.

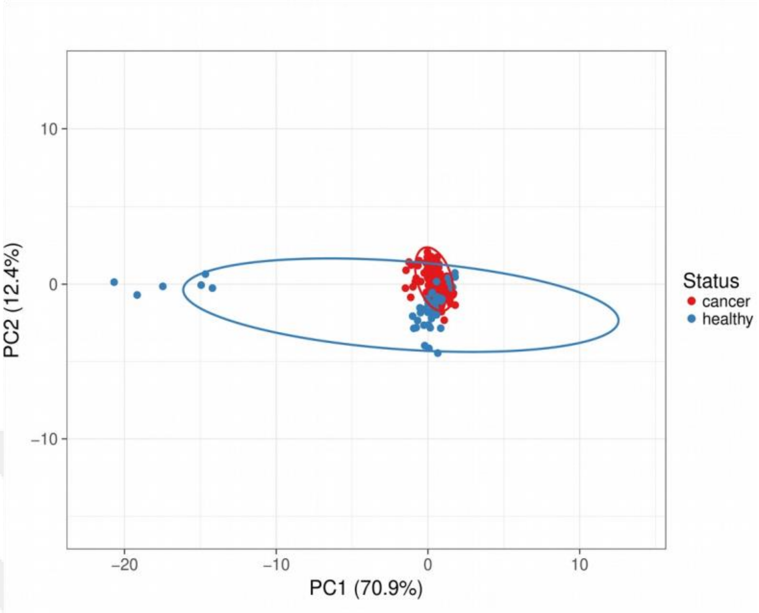


Figure 95: PCA of kidney methylation data

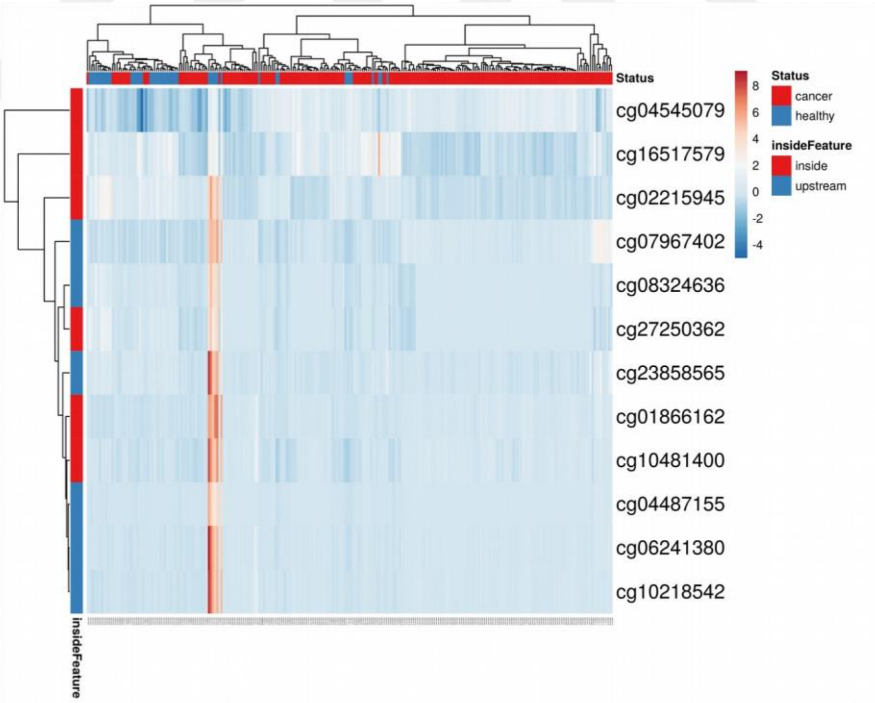


Figure 96: Heatmap of kidney methylation data

Scatter plots for correlation of methylation and gene expression in healthy (Figure 97) and cancerous kidney samples (Figure 98) are provided below.

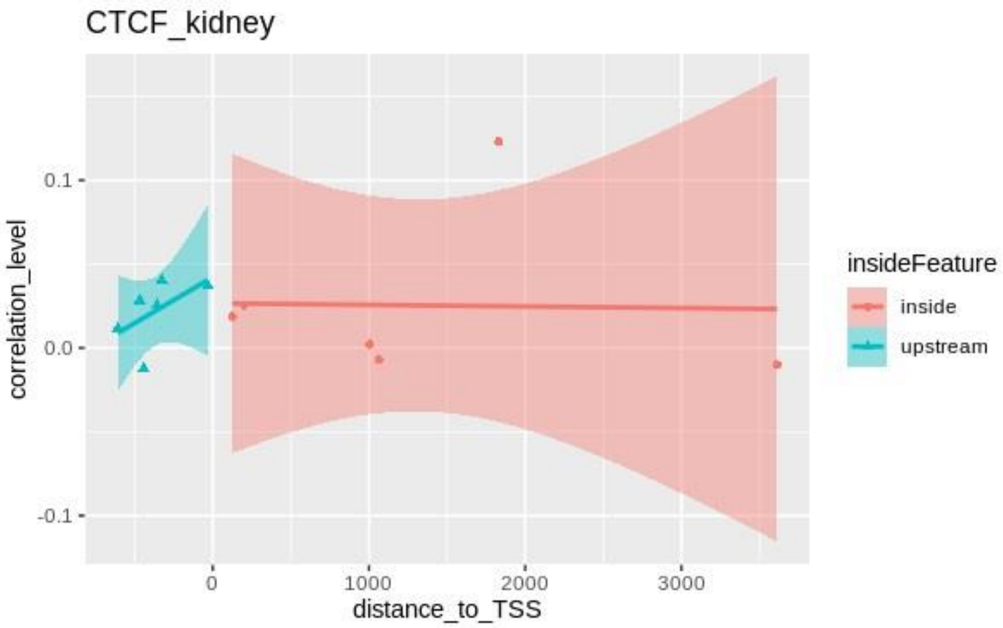


Figure 97: Scatter plot created using healthy kidney data

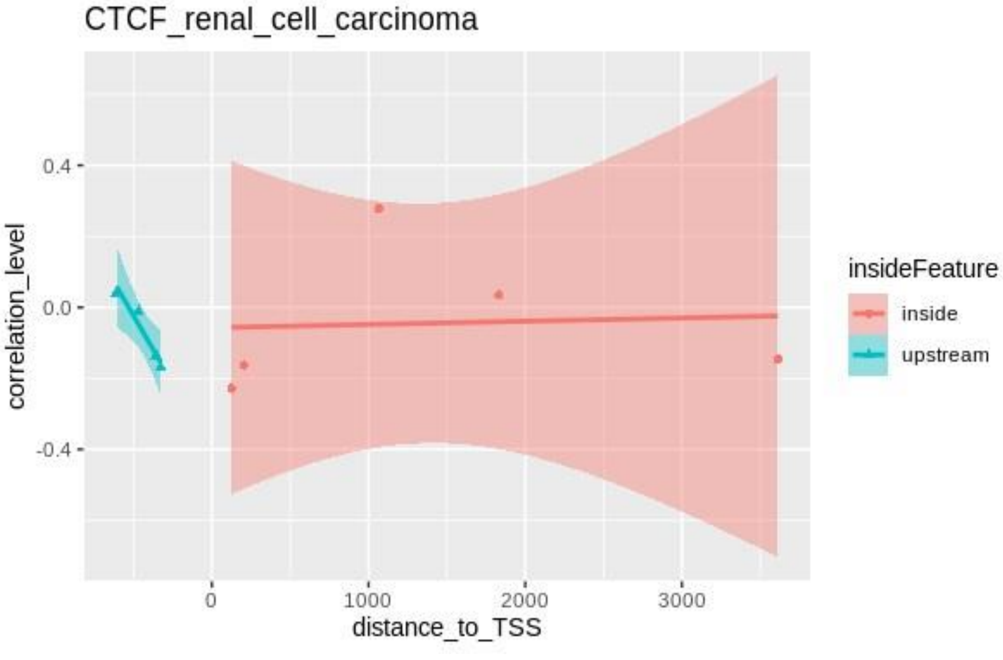


Figure 98: Scatter plot created using kidney cancer data

The scatter plot (Figure 99) and table (Figure 29) created to better observe the difference between healthy and cancerous tissues are shown below.

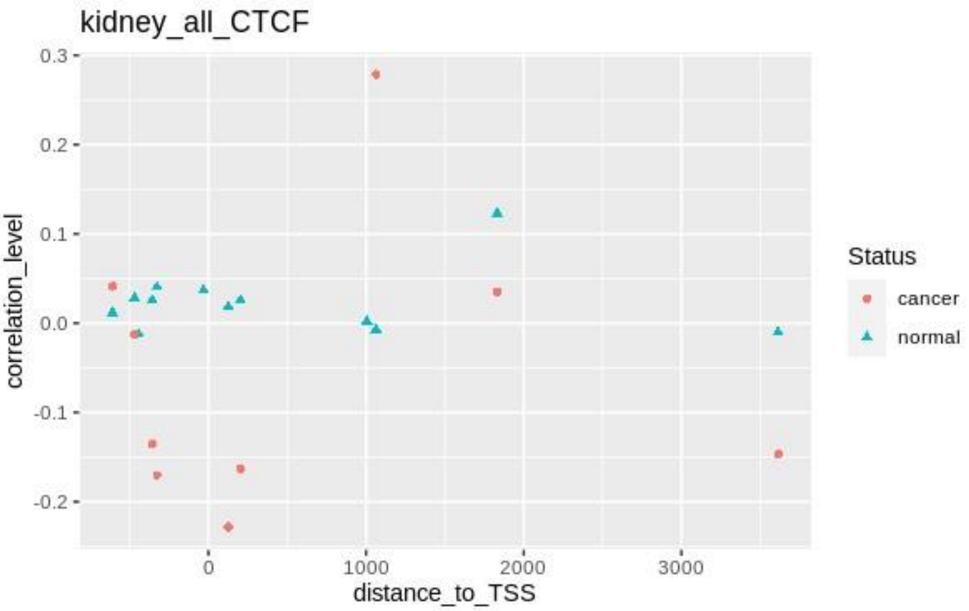


Figure 99: Correlation of probe regions in the CTCF gene of kidney cancer samples

CG probe ID	Distance to TSS	insideFeature	r in healthy samples (r1)	r in cancer samples (r2)	(r1 - r2)	rho in healthy samples (rho1)	rho in cancer samples (rho2)	(rho1 - rho2)
cg23858565	-608	upstream	0.01	0.04	-0.03	-0.05	0.09	-0.15
cg07967402	-468	upstream	0.03	-0.01	0.04	0.17	0.14	0.03
cg08324636	-442	upstream	-0.01	NA	NA	-0.02	NA	NA
cg06241380	-357	upstream	0.03	-0.14	0.16	-0.14	0.08	-0.22
cg10218542	-326	upstream	0.04	-0.17	0.21	0.25	0.16	0.09
cg04487155	-32	upstream	0.04	NA	NA	0.08	NA	NA
cg10481400	126	inside	0.02	-0.23	0.25	-0.10	0.17	-0.27
cg01866162	203	inside	0.03	-0.16	0.19	-0.01	0.12	-0.13
cg27250362	1005	inside	0.00	NA	NA	0.04	NA	NA
cg02215945	1063	inside	-0.01	0.28	-0.29	-0.09	-0.21	0.12
cg16517579	1832	inside	0.12	0.04	0.09	0.18	-0.08	0.26
cg04545079	3613	inside	-0.01	-0.15	0.14	0.00	0.01	-0.01

Table 29: Correlation of probe regions in the CTCF gene of kidney cancer samples.

When the Pearson and Spearman correlation differences were examined, it was seen that none of the probes exceeded the cut-off threshold.

3.2.2.11 Liver

PCA (Figure 100) and heatmap (Figure 101) created using methylation data of healthy and cancerous tissues are shown below.

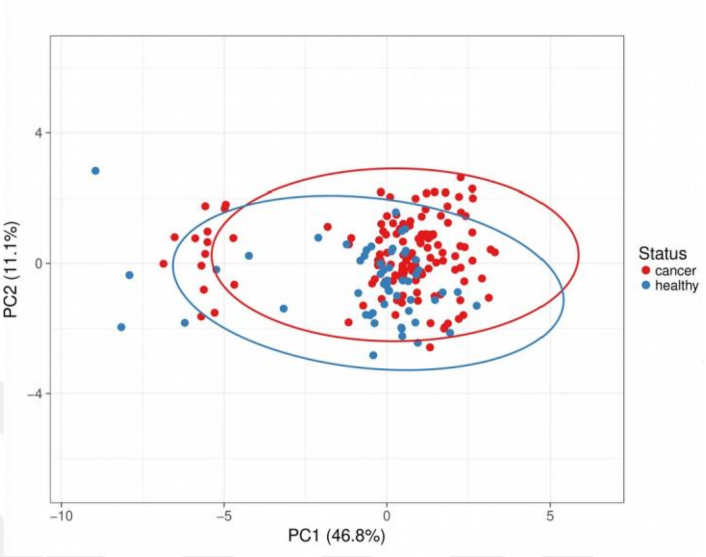


Figure 100: PCA of liver methylation data

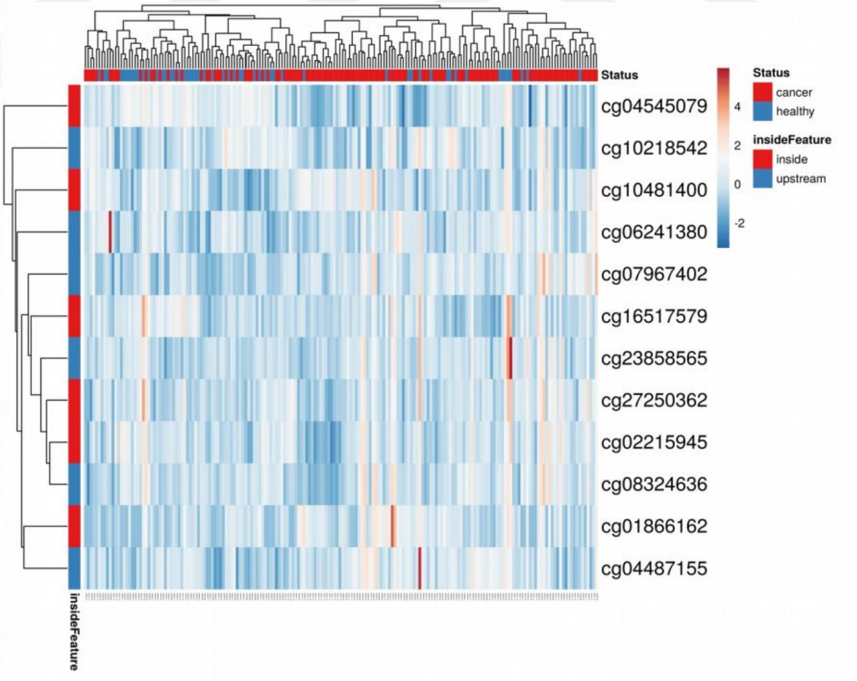


Figure 101: Heatmap of liver methylation data

Scatter plots showing results of correlation analysis of methylation and gene expression in healthy (Figure 102) and cancerous liver tissues (Figure 103) are provided below.

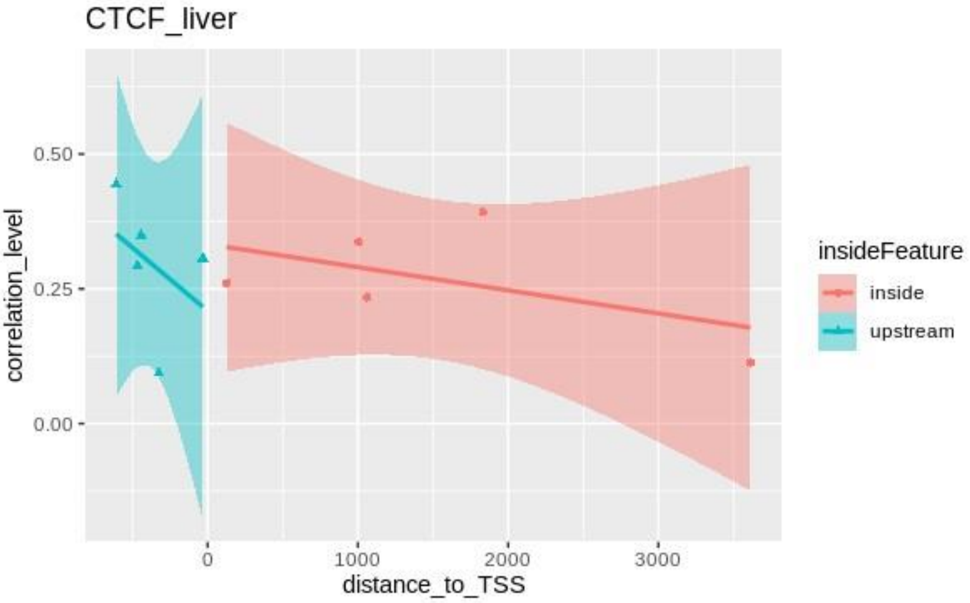


Figure 102: Scatter plot created using healthy liver data

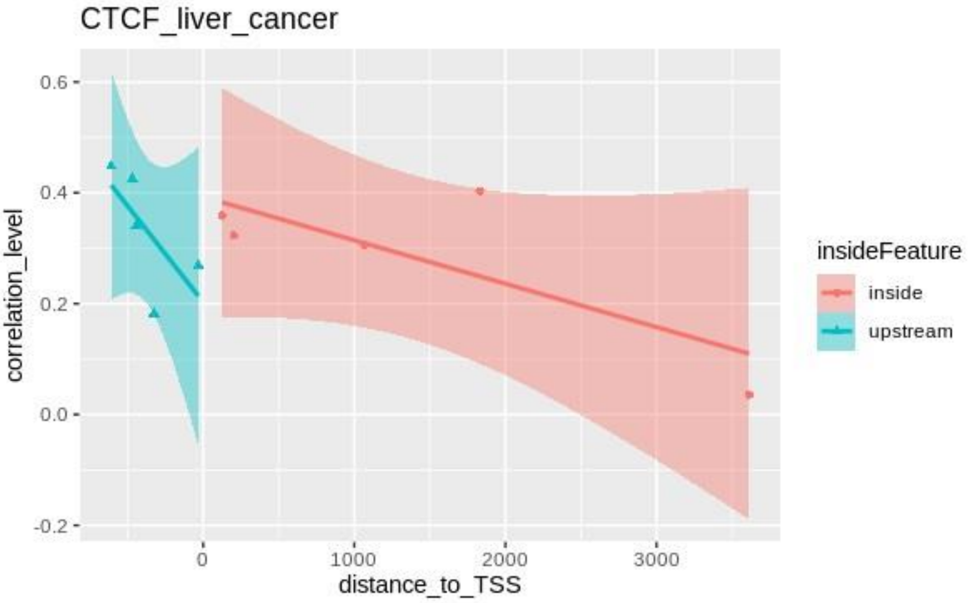


Figure 103: Scatter plot created using liver cancer data

The scatter plot (Figure 104) and table (Table 30) created to better observe the difference between healthy and cancerous tissues are shown below.

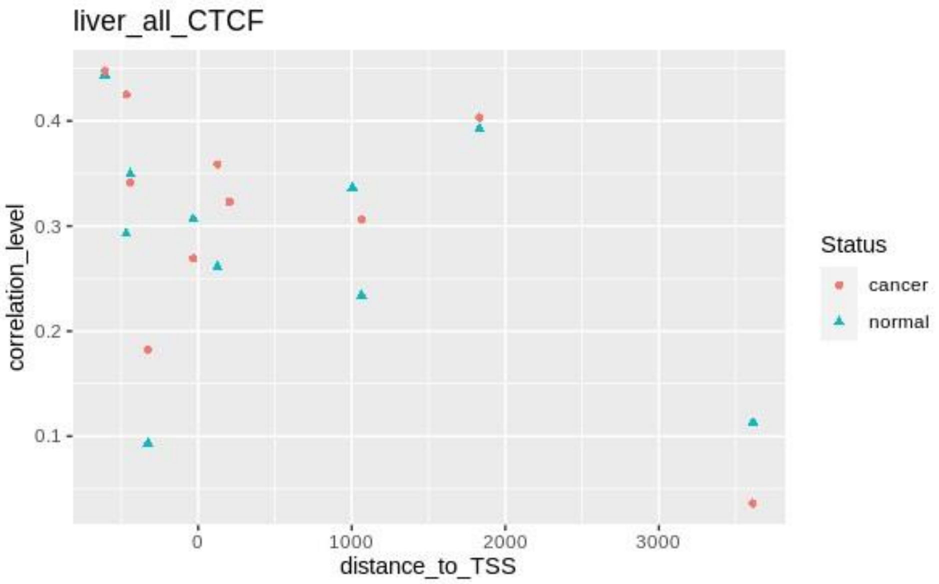


Figure 104: Correlation of probe regions in the CTCF gene of liver cancer samples

CG probe ID	Distance to TSS	insideFeature	r in healthy samples (r1)	r in cancer samples (r2)	(r1 - r2)	rho in healthy samples (rho1)	rho in cancer samples (rho2)	(rho1 - rho2)
cg23858565	-608	upstream	0.44	0.45	0.00	0.39	0.12	0.27
cg07967402	-468	upstream	0.29	0.42	-0.13	0.24	0.40	-0.17
cg08324636	-442	upstream	0.35	0.34	0.01	0.26	-0.03	0.29
cg06241380	-357	upstream	NA	NA	NA	NA	NA	NA
cg10218542	-326	upstream	0.09	0.18	-0.09	0.08	0.11	-0.03
cg04487155	-32	upstream	0.31	0.27	0.04	0.32	0.22	0.11
cg10481400	126	inside	0.26	0.36	-0.10	0.16	0.33	-0.16
cg01866162	203	inside	NA	0.32	NA	NA	0.17	NA
cg27250362	1005	inside	0.34	NA	NA	0.30	NA	NA
cg02215945	1063	inside	0.23	0.31	-0.07	0.22	0.18	0.05
cg16517579	1832	inside	0.39	0.40	-0.01	0.44	0.42	0.02
cg04545079	3613	inside	0.11	0.04	0.08	0.05	-0.06	0.11

Table 30: Correlation of probe regions in the CTCF gene of liver cancer samples.

When the Pearson and Spearman correlation differences were examined, it was seen that none of the probes exceeded the cut-off threshold.

3.2.2.12 Lung

PCA (Figure 105) and heatmap (Figure 106) created using methylation data of healthy and cancerous tissues are shown below.

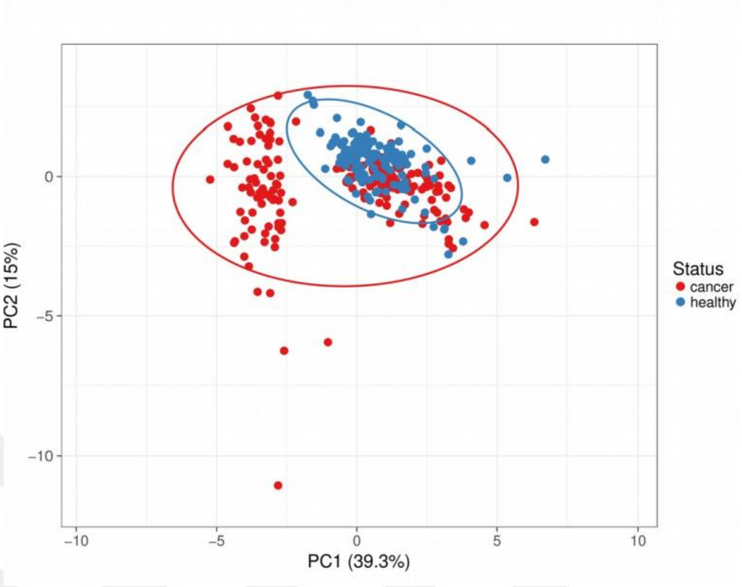


Figure 105: PCA of lung methylation data

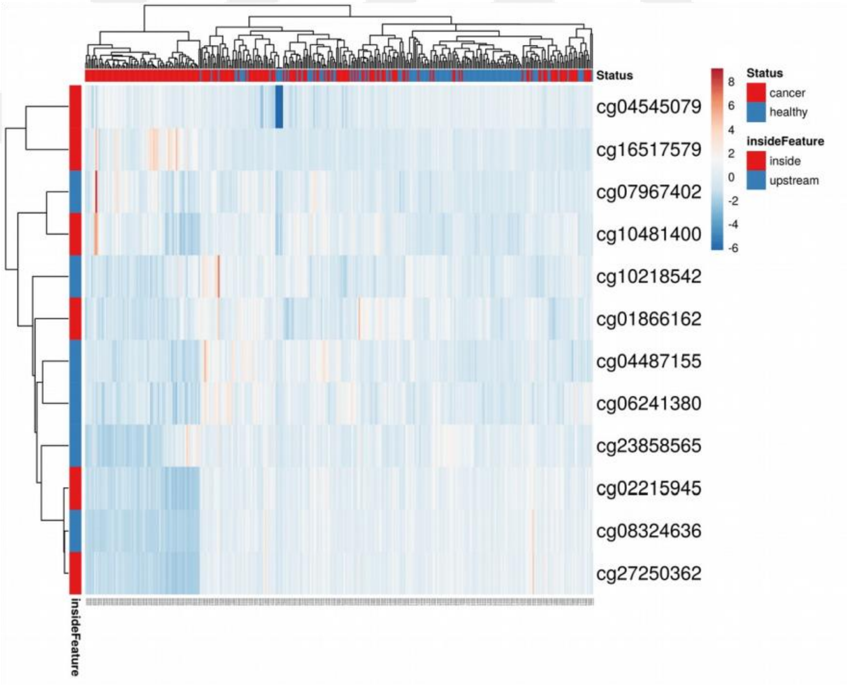


Figure 106: Heatmap of lung methylation data

Scatter plots visualizations of correlation between methylation and gene expression in healthy (Figure 107) and cancerous lung samples (Figure 108) are shown below.

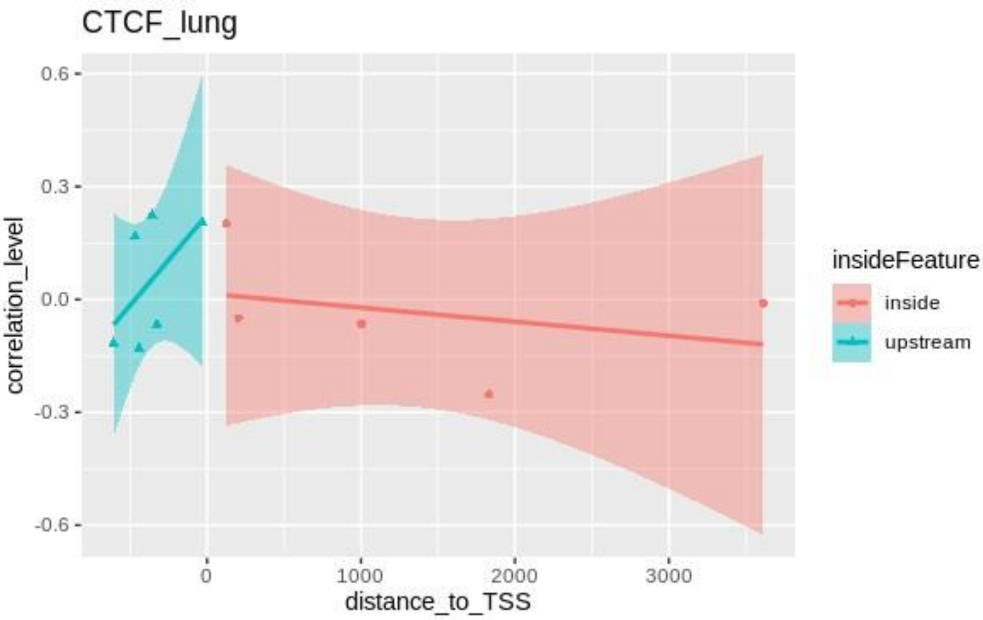


Figure 107: Scatter plot created using healthy lung data

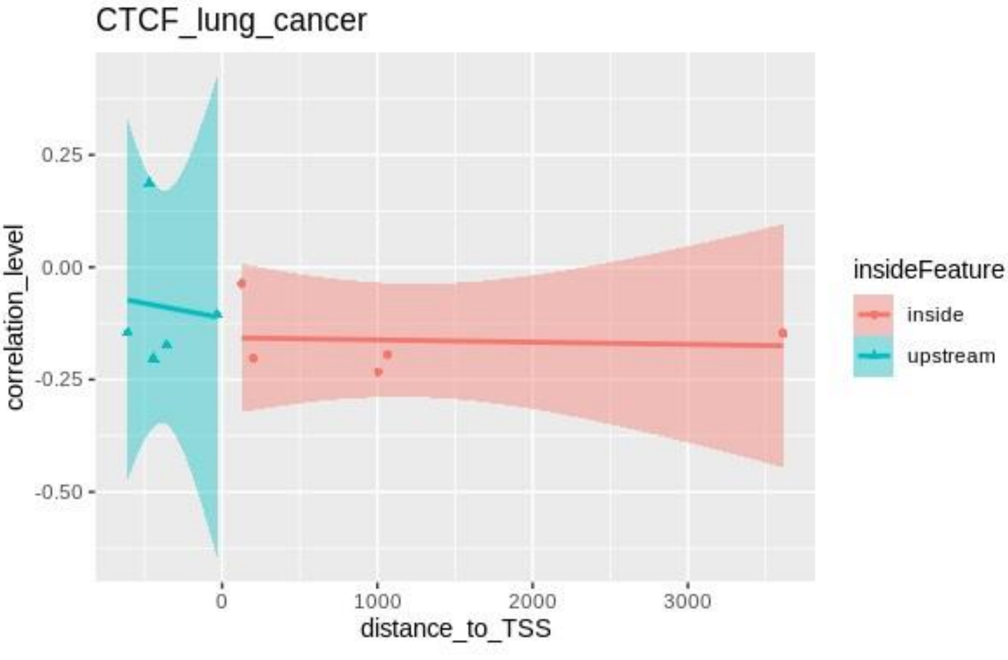


Figure 108: Scatter plot created using lung cancer data

The scatter plot (Figure 109) and table (Table 31) created to better observe the difference between healthy and cancerous tissues are shown below.

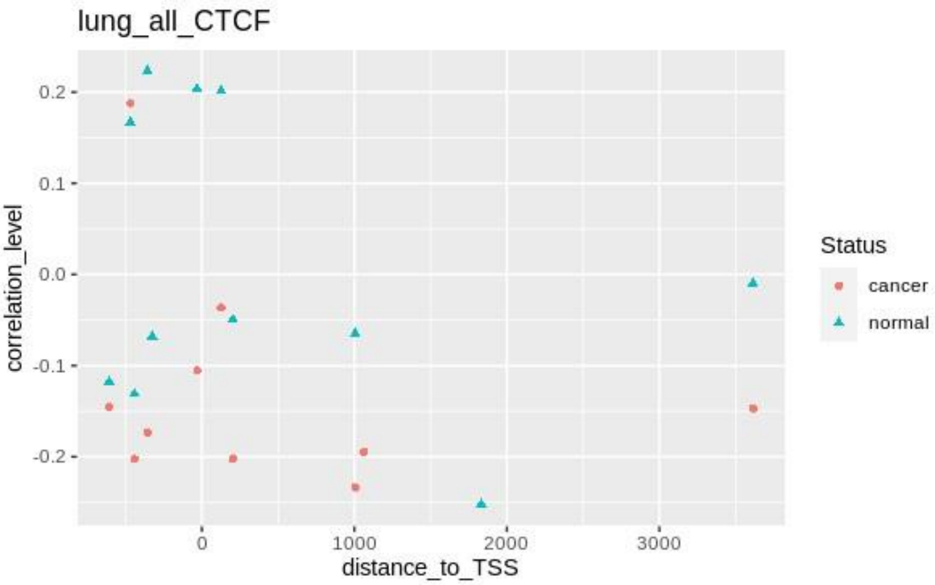


Figure 109: Correlation of probe regions in the CTCF gene of lung cancer samples

CG probe ID	Distance to TSS	insideFeature	r in healthy samples (r1)	r in cancer samples (r2)	(r1 - r2)	rho in healthy samples (rho1)	rho in cancer samples (rho2)	(rho1 - rho2)
cg23858565	-608	upstream	-0.12	-0.15	0.03	-0.11	-0.18	0.07
cg07967402	-468	upstream	0.17	0.19	-0.02	0.24	0.23	0.01
cg08324636	-442	upstream	-0.13	-0.20	0.07	-0.14	-0.21	0.07
cg06241380	-357	upstream	0.22	-0.17	0.40	0.28	-0.20	0.48
cg10218542	-326	upstream	-0.07	NA	NA	-0.06	NA	NA
cg04487155	-32	upstream	0.20	-0.11	0.31	0.25	-0.16	0.41
cg10481400	126	inside	0.20	-0.04	0.24	0.25	-0.06	0.31
cg01866162	203	inside	-0.05	-0.20	0.15	-0.05	-0.19	0.14
cg27250362	1005	inside	-0.06	-0.23	0.17	-0.10	-0.24	0.14
cg02215945	1063	inside	NA	-0.19	NA	NA	-0.22	NA
cg16517579	1832	inside	-0.25	NA	NA	-0.23	NA	NA
cg04545079	3613	inside	-0.01	-0.15	0.14	-0.22	-0.05	-0.17

Table 31: Correlation of probe regions in the CTCF gene of lung cancer samples.

When the Pearson and Spearman correlation differences were examined, it was seen that none of the probes exceeded the cut-off threshold.

3.2.2.13 Numerical summary of differential correlation results

The final step was to find out in which tissues differential correlation pattern between methylation and gene expression of CTCF can be detected. For this, the difference of the correlation coefficient in healthy and cancerous samples in the same probe regions was calculated. Probes exceeding the 0.5 cut-off threshold were shown in the table 32.

Tissue	CG probe ID	Distance to TSS	insideFeatur	Pearson correlation difference (r1-r2)	Spearman correlation difference (rho1-rho2)
bladder	**cg04487155**	**-32**	**upstream**	**0.63**	**0.59**
bladder	**cg06241380**	**-357**	**upstream**	**1.04**	**0.71**
bladder	**cg07967402**	**-468**	**upstream**	**0.64**	**0.81**
bladder	cg10481400	126	inside	0.45	0.53
bone	**cg01866162**	**203**	**inside**	**0.75**	**1**
bone	cg02215945	1063	inside	0.38	0.56
bone	**cg04487155**	**-32**	**upstream**	**0.77**	**0.49**
bone	**cg06241380**	**-357**	**upstream**	**0.73**	**0.54**
bone	**cg08324636**	**-442**	**upstream**	**0.52**	**0.66**
bone	**cg10218542**	**-326**	**upstream**	**0.75**	**0.79**
bone	cg10481400	126	inside	0.16	0.53
bone	**cg23858565**	**-608**	**upstream**	**0.62**	**0.7**
bone	cg27250362	1005	inside	0.52	0.38
brain	cg10218542	-326	upstream	0.46	0.52
breast	**cg04545079**	**3613**	**inside**	**0.83**	**0.98**
breast	cg07967402	-468	upstream	0.60	0.48
breast	**cg08324636**	**-442**	**upstream**	**1.04**	**1.06**
breast	cg10481400	126	inside	0.57	0.43
breast	**cg23858565**	**-608**	**upstream**	**0.74**	**0.78**
breast	**cg27250362**	**1005**	**inside**	**0.81**	**0.88**
colon	cg01866162	203	inside	0.79	0.2
colon	cg04487155	-32	upstream	0.72	0.2
colon	**cg06241380**	**-357**	**upstream**	**0.68**	**0.59**
colon	**cg08324636**	**-442**	**upstream**	**0.91**	**0.58**
colon	cg10218542	-326	upstream	0.55	0.48
colon	**cg10481400**	**126**	**inside**	**0.91**	**0.6**
colon	cg23858565	-608	upstream	0.67	0.48
colon	**cg27250362**	**1005**	**inside**	**0.88**	**0.65**
gastric	cg01866162	203	inside	0.29	0.51
gastric	cg04487155	-32	upstream	0.39	0.61
gastric	cg06241380	-357	upstream	0.49	0.7
gastric	**cg07967402**	**-468**	**upstream**	**0.57**	**0.59**
gastric	cg10218542	-326	upstream	0.30	0.57
pancreas	cg02215945	1063	inside	1.00	0.35
pancreas	**cg08324636**	**-442**	**upstream**	**0.94**	**0.6**
pancreas	cg10218542	-326	upstream	0.47	0.77
pancreas	**cg10481400**	**126**	**inside**	**0.94**	**0.67**
pancreas	cg23858565	-608	upstream	0.62	0.31
pancreas	**cg27250362**	**1005**	**inside**	**0.87**	**0.72**
prostate	cg01866162	203	inside	0.47	1.03
prostate	**cg04487155**	**-32**	**upstream**	**0.53**	**1**
prostate	**cg04545079**	**3613**	**inside**	**1.02**	**0.55**
prostate	cg10218542	-326	upstream	0.63	0.46
prostate	cg10481400	126	inside	0.88	0.4
prostate	cg16517579	1832	inside	0.90	0.45
prostate	cg23858565	-608	upstream	0.21	0.54
prostate	cg27250362	1005	inside	0.08	0.95
small_intestine	cg10218542	-326	upstream	0.01	0.5
small_intestine	cg10481400	126	inside	0.49	0.8
small_intestine	cg23858565	-608	upstream	0.71	0.25
				Count	**Percantage**
TOTAL	Probes exceeding 0.5 cut off threshold in at least one of the correlation types		**upstream:**	29	%58
			inside:	21	%42
			Sum	50	100%
	Probes exceeding 0.5 cut off threshold in both correlation types		**upstream:**	15	%65
			inside:	8	%35
			Sum	23	100%

Table 32: Probes exceeding 0.5 cut-off threshold for each cancer types. Probes that exceed the cut-off threshold in both correlations are shown in bold.

As a summary of this table, the table (Table 33) showing how many probe regions of each tissue type exceed the cut-off threshold is shown below:

Tissue	>0.5_threshold_difference
bone	5
colon	4
breast	4
pancreas	3
bladder	3
prostate	2
gastric	1
small_intestine	0
brain	0
kidney	0
liver	0
lung	0

Table 33: Probe region number that exceed the cut-off threshold in both types of correlation

CHAPTER 4

DISCUSSION

4.1 Mining CTCF Interactome

In this thesis, the CTCF gene, which is a candidate tumor suppressor gene, was investigated using multi-omic data mining. The analysis was performed on three levels, of molecular complexity, namely, protein-protein interaction (PPI) interactome, DNA methylome and transcriptome (whole genome expression). Conceptually, this work by design consists of 2 main components. The first part is about CTCF PPI network analysis and the second part is about correlation analysis of methylation and gene expression.

In the first part, the network of the CTCF gene was examined, with the ultimate goal of predicting potential biological role of this gene using the guilt by association paradigm

(GBA). The functional role of CTCF was predicted as a result of this conducted network-based enrichment analysis (overrepresentation of functional annotation terms among the genes comprising the network).

According to the literature screen, which was performed by the author, this thesis is the first such effort to examine CTCF's PPI network in the context of cancer. More specifically, the interactome level data mining addressed the long standing question about the candidate tumor suppressor CTCF from the perspective of its interaction partners based on the employed GBA approach. By and large, this is the most comprehensive scientific inquiry on the CTCF interactome until now. Particularly, the novelty is also underlined by the employed GBA approach.

Accordingly, to determine the potential roles of CTCF and to create a roadmap for the data to be selected in the next step, Cytoscape which is the most frequently used tool for network analysis, was used.

The searched and retrieved nodes (protein interaction partners) and edges (interactions) of the CTCF protein were used to reconstruct the CTCF PPI network. Interactions among the interaction partners were also obtained and included in the network to enable a more complete and mechanistic overview of the network. Accordingly, a network comprising 21 nodes and 83 edges was reconstructed (Table 2). NetworkAnalyzer plugin was used to calculate the relationship of genes in this network with CTCF, and the results were sorted according to degree value with cytoHubba plugin and top10 related gene was determined: SMAD binding family (SMAD4, SMAD5, SMAD6, SMAD1), ZMYM2, NPM1, ADNP, YBX1, SET (Figure 11) Then, using the MCODE plugin, the overrepresented clusters in this network, which are closely related to each other, were determined according to their topology values. According to this result, this network basically consists of 2 clusters: the 1st cluster (Table 5) with SMAD binding family members and the 2nd cluster consisting of 8 nodes including CTCF (Table 6).

The functional annotation roles of the SMAD genes in the 1st cluster generally contain the receptors involved in signal transmission. Especially TGFβ (transforming growth factor beta) receptor has been observed as the pathway in which they play a primary role in all SMAD genes. TGFβ is a growth factor and cytokine and is involved in paracrine signaling. It appears to be found in many different types of tissue, especially in the brain, heart, kidney, liver, bone, and testicles. High expression of TGFβ was previously shown to be associated with kidney diseases [99]. Based on this information, according to the

GBA approach, it is thought that CTCF, which appears to be in close relationship with the genes involved in these pathways, can also take part in these pathways. For this, tissue types specified to contain TGFβ mentioned above, will be examined in the next correlation stage of this study.

In addition to this pathway, BMP signaling pathway (Bone morphogenetic proteins), which is a member of TGFβ, also appeared to be annotated. Based on this, bone samples will be examined in the next stage.

In the 2nd cluster, the pathways that generally concern the structural formation of chromatin are annotated. These genes, including CTCF, participate in the nucleosome organization and function in chromatin assembly or disassembly. Thus, it appeared to be critical for the gene expression control.

In addition to these, when another result found to be annotated, "post-transcriptional regulation", is examined, it is seen that the genes in this cluster also play a role in this mechanism in which RNA polymerase II regulates gene expression by binding to the promoter of the gene at the transcription stage. Based on this, we can say that the cluster mainly contains genes involved in the regulation of gene expression. In this case, according to the GBA approach, CTCF is also thought to play a role in these pathways, and appropriate data have been selected to test this in the next stage.

Using a Gene Ontology (GO) source, a Cytoscape plugin GOlorize that enables the study of genes and gene products functions, was used to perform functional annotation of this network. The GO source basically includes 3 main domains: Cellular component (CC), which examines the cell and its extracellular environment, the molecular function (MF), which examines the activities of gene products at the molecular level, and finally the Biological process (BP), which studies the functioning of molecular events that have a beginning and an end [100].

Among the overrepresented GO categories in this network, the first annotation was performed using the BP domain, and the results obtained were ranked as including the highest grades. As a result 5 GO categories were specified, these are respectively; “regulation of transforming growth factor beta receptor signaling pathway”, “chromatin assembly or disassembly”, “BMP signaling pathway”, “regulation of histone acetylation”, “regulation of cell death”. Looking at these results, it was seen that the first 3 annotations results were the same with the annotations in the clusters in the MCODE, which is of great importance for the consistency of the results. When the 4th highly annotated category was examined, it was seen that acetylation, which is an epigenetic regulation mechanism, is similarly involved in the regulation of chromatin and indirectly in the regulation of gene expression. This result supports the importance of using methylation which is another important epigenetic mechanism in this study.

The regulation of cell death, the last category of the 5 most annotated GO-BP domains, is a pathway that plays a very important role in cancer. Since we are working on CTCF, which is thought to be a tumor suppressor gene, finding that the genes with which it is closely related play such a vital biological role, supports the idea of CTCF's association with cancer.

A second annotation was performed using the GO-MF domain, and the results obtained were ranked as including the highest grades. As a result 5 GO categories were specified, these are respectively; “binding”, “nucleic acid binding”, “DNA binding”, “transcription regulator activity”, “protein binding”. As can be seen from the results, the most important annotation results are the binding functions. These DNA binders include transcription factors that modulate transcription, various polymerases, histones that participate in packaging the chromosome. Looking at the transcription regulatory activity, another result of these annotation results, we can say that the molecular functions of genes in this network are effective regulators at the transcription level in general.

Similarly, annotation analysis is performed using the GO source using the licensed tool ClueGO, which is another Cytoscape plugin. However, similar results are fused and displayed in a single category in this plug-in in order to prevent similar results from being displayed more than once, as in the GO-MF analysis conducted in GOlorize, and thus to give the most important biological roles in the network as a comprehensive summary.

First, an annotation analysis by selecting GO BP and MF domain together was performed by using ClueGO plugin. A network of 2 different clusters was obtained and the most significant results were determined as group names: regulation of protein acetylation and SMAD binding. Similar to the result of GO-BP analysis obtained using GOlorize before, the first group consists of protein acetylation, histone acetylation, peptidyl-lysine acetylation, which are an epigenetic regulation mechanism. When looking at the barchart showing the detailed annotation values of the genes involved in these functions and which genes are involved in the specified pathways, it was seen that the CTCF, SET and SMAD4 genes were involved in this acetylation mechanism.

The second group name, SMAD binding, is the only result generated term in these results using the MF domain. The reason for this is that the significant results obtained with MF are very similar to each other, so they may be fused. When the genes involved in this group are examined, as the name suggests, there are SMAD 1, 4 and 6 from the SMAD binding family. Looking at another annotated term in which these genes are involved, it was seen that the entire SMAD binding family was involved in the embryonic pattern specification. Based on the GBA approach, in order to examine whether CTCF, which works with genes that play an effective role in the embryonic development process, has an effect on this developmental process, a detailed examination of the change of CTCF in the developmental process has been carried out in the next step of this thesis.

As other annotation results, it has been observed that there are pathways that function with RNA polymerase II and thus are effective in the regulation of transcription. One of these is transcribing miRNAs that play a role in regulating gene expression by processing pri mRNAs with RNA polymerase II. The other is the positive regulation of transcription from the RNA polymerase II promoter involved in the cellular response to chemical stimulus. Genes involved in these functional pathways: SMAD family and POU5F1 gene. The fact that genes in this network are generally effective in regulating transcription suggests that CTCF may also have this biological role.

In addition to the GO source, annotation analysis was performed with the Reactome pathway option using the ClueGO plugin. 4 different clusters were obtained and the most significant terms were determined as the group name.

Looking at the group names and the genes that play a role in this pathway, it seems that there are members of the SMAD binding family. In particular, when the signaling by BMP and RUNX2 regulates bone development pathways were examined, it was seen that SMAD1 and SMAD4 genes formed a complex with the RUNX2 gene in the nucleus as response to the BMP signalling and induced SMAD6 transcription. In this way, the development of intramembranous and endochondral bones is provided [101]. Based on the idea that CTCF, which is closely related to SMAD binding genes, may play an important role in bone development, bone tissue was selected as one of the primary investigations in this study.

When looking at the other Reactome results, it was seen that the members of the SMAD binding family play a role in the BMP signal transmission pathway and form the necessary complexes to ensure (Phospho-R-Smad1/5/8 complex [102]) or prevent (SKI complex, Ubiquitin-dependent [103]) this transmission.

ReactomeFI plugin was used to increase the accuracy of these Reactome pathway annotations and to examine the main pathways to which they are connected. Among the obtained results, those with FDR value less than 0.05 were selected as cut-off threshold. Accordingly 5 results were obtained respectively; Transcriptional regulation of pluripotent stem cells, Transcriptional regulation by RUNX2, Signaling by TGF-beta family members, Signaling by BMP, RUNX2 regulates bone development.

Considering these results, it was seen that the results were consistent and the same categories were annotated with the results obtained using ClueGO and MCODE. To summarize all the results, this network consists of 3 main domains; developmental biology, gene expression regulation, and signal transduction. It has been observed that they are enriched especially in signal transduction pathway necessary for bone development. Through this plugin, it is also possible to visualize the gene complexes involved in these pathways.

In order to increase the accuracy of the results and highlight the missing points, online annotation tools have been used as a complementary second step of network analysis. Webgestalt was used first for this. Unlike other Cytoscape plugins, this tool automatically colors the results according to FDR values and uses high visualization techniques such as Volcano plot. First, overrepresentation analysis was performed using the GO source as previous. When the GO-BP result is examined, the results are very similar to the results obtained with other Cytoscape plugins. In addition as stated above, the results are shown on the barchart to be significant according to the FDR values. Accordingly, the most significant result is “positive regulation of transcription from RNA polymerase II promoter involved in cellular response to chemical stimulus”. In addition to the previous results, the result of urogenital system development was also obtained, but when the FDR value is considered, it can be said that it is less reliable than other terms because it is greater than 0.05.

When the GO-MF results obtained using Webgestalt were examined, it was shown that the FDR values of all results were less than 0.05. This explains why ClueGO shows only one MF result which is SMAD binding. When these results were examined, like the previous results it was seen that binding and acetylation regulation were obtained in general, however their reliability was not as high as those in BP.

When Cellular Component was used at the end of the GO categories, it was seen that the FDR value of only one term was significant which is transcription factor complex. It can be said that this result is consistent as it contains the transcription regulatory complex similar to the results obtained with BP. The methyltransferase complex, which is observed as another remarkable result, seems to be an important inference since methylation will be used as a basis in this study. However, it cannot be said to be precise since the FDR value is greater than cut-off threshold.

In addition to GO source overrepresentation analysis was performed using KEGG pathway by using Webgestalt. Looking at the results, it was observed that the signaling pathways were generally enriched. However, only 2 results were obtained significant which are TGFβ signaling pathway and signaling pathways regulating pluripotency of stem cells. Similar to other results, it has been observed that pathyways that are involved in developmental biology are annotated. In addition, although the FDR value was greater than the cut-off threshold, pancreatic cancer, colorectal cancer and gastric cancer were also annotated as a remarkable result. Since the tumor suppressor gene characteristic of CTCF will be tested in this study, these tissues are included among the primary tissues to be examined, just like bone.

According to the result obtained using the Reactome pathway by using Webgestalt, 3 pathways appeared to be significant which are “Signaling by BMP, RUNX2 regulates bone development, Signaling by TGFβ family members”. It can be said that the reliability of these results is very high because they both merge with the results obtained before and FDR values are also tested with this tool.

According to the annotation results created using the Wikipathway database, which is a database that is not available in other tools, only one significant result was obtained which is TGFβ receptor signaling. Looking at other results, an interesting result was the pancreatic adenocarcinoma pathway, however the significance of this result is low because the FDR value cannot exceed the cut-off threshold. Nevertheless, as in the other results, pancreatic tissue was used as the primary research tissue in this study.

A similar study was conducted using the Babelomics online tool. As a result, annotation of 5 different categories was obtained using GO-BP: “protein complex subunit organization”, “cellular component assembly”, “macromolecular complex assembly”, “cellular component biogenesis”, and “macromolecular complex subunit organization”. In general, when the results were examined, it was seen that similarly, the regulation of binding and the formation of the structure of chromatin and thus, they took part in the complex structure organizations that regulate the control of gene expression.

As a final online tool DAVID was used. Three categories were obtained when significance highest was selected. When the results were examined, it was seen that similar results were obtained such as “BMP signaling pathway”, “growth factor beta signaling pathway”, “cellular response via RNA polymerase II against chemical stimulus”. Therefore, in the second stage of this study, the change of CTCF in the developmental process, and in the formation of cancer and some diseases, which was determined by the GBA approach, was examined.

The CellWhere online tool was used to find out where genes in this network play a role in metabolism. As a result, it has been observed that the CTCF is involved in the nucleolus, and the genes with which it is closely related are in the nucleus.

4.2 Methylome and Transcriptome-driven Data Analysis

Being a multi-omic effort, this thesis also deals with the epigenomic and transcriptomic aspects of the CTCF biology. The relationship between methylation and expression of genes is a long studied topic, which is usually researched only using methylation and gene expression levels, without considering other remaining possible factors. In contrast to previous approaches, this is the first study to take importantly into accounts location of methylation sites (distance from TSS) and compare various biological contexts, while examining relationship between CTCF methylation and its expression. In another words, the originality derives from including specifics of methylation sites and tissue specific differences in the analysis. Originality also comes from each methylation sites independently, instead of putting all of them into one bin. Consequently, instead of the averages, the site-level resolution was achieved. All together, this unique level of elaboration and scrutiny distinguishes this work from previous studies.

The second stage of this thesis; deals with combined analysis of methylome and transcriptome. Here, the question of candidate tumor suppressor protein candidacy of CTCF is approached by adopting an integrated perspective, which attempts to link methylation with gene expression. To this end, correlation between methylation level and gene expression data (mRNA level) of CTCF was assessed in multiple biological contexts. With the purpose of this integrative analysis, the methylation and gene expression datasets, matching same samples, were selected as input. The data was divided into the following categorizes: developmental process (spanning 7 different age ranges) and cancer (consisting of 12 different types). Each category was analyzed separately, in order to gain a detailed insight into the underlying molecular biology.

4.2.1 Development-related methylation-gene expression correlation analysis

Firstly the dependence of CTCF methylation to gene expression in the developmental process was measured by Pearson and Spearman correlation method. For this purpose, 7 different age ranges have been studied separately in order to examine the change of CTCF methylation at each development stage. In this regard, the lifespan was divided into the following categories: fetal, newborn, infancy, childhood (5-17), early adulthood (18-40), late adulthood (41-80) and senescence (80+).

In this study, unlike previous studies, the methylation level of CTCF was performed represented not by a single value, but by all probe regions located 5 kb upstream and downstream of the CTCF's TSS point. As a result, a total of 12 methylation-specific probe sites were determined and demarcated. Half of them were located in the upstream region of the CTCF gene and the remaining half were located inside of the gene. When the fetal result was examined in general, it was seen that the correlation level was obtained as ~0 values. Based on this, it can be inferred that methylation of CTCF at the embryonic stage has no significant effect on gene expression. However, this initial observation should be viewed as rather preliminary result and interpreted with caution due to the following limitations in the input data. First of all, when during the inspection of “fetal samples” result, which comprised the correlation level of methylation and gene expression based on the 3 matched biological samples (corresponding to 3 data point), it was noticed that correlation could not be calculated for 5 probe regions. This is because the records corresponding to some probes in methylation input data contained “NA” (abbreviation for Not Applicable) instead of values. Furthermore, the number of analyzed samples was too limited, as result of the scarcity of fetal samples in the data sets. Thus, in order to reach a clear conclusion, a follow-up study using a larger sample size is necessary to confirm the accuracy of the result and arrive to a solid conclusions. When focusing on the newborns-related, correlation appears to be remarkably highly positive. To be more specific, the correlation was found to be close to 0.8 in almost all probe regions. In another words, gene expression tends to increase with the rising methylation. Based on this, it can concluded that methylation importantly CTCF gene, is thought to play role in developmental biology.

In the infant-related results, the correlation level was obtained as approximately 0. This findings suggests similarity to the fetal-related correlation pattern. Thus, for the inspected probes, methylation and gene expression appear to be unrelated in these two stages.

Inspection of correlation scatter plot, which was generated using childhood data, revealed resemblance with the infant-related results. The correlation were weak (0-0.5), both negative and positive correlations were observed (with coefficients around 0.25).

In sharp contrast to newborn, negative correlation between methylation and gene expression was observed in early adulthood. In another words, this observation (negative correlation) was quite opposite of to the mentioned finding (positive correlation) in newborn. Thus, apparently depending on the tissue type, two opposite effects of methylation on gene expression is possible. Moreover, methylations of the analyzed probe sites seems to have different effects on CTCF expression in across the studied age ranges of development. Remarkably, in early adulthood, negative correlation was detected for all investigated probe regions. Therefore, increase in the methylation of the investigated probes seems to consistently cause a decrease in CTCF gene expression at this developmental stage. Importantly, strong negative correlation was observed especially in the probe regions within the CTCF gene.

Interestingly, a positive correlation was observed in late adulthood with a correlation profile similar to that in newborn. High positive correlation appears to be particularly in probes located in the upstream region. Based on this, it can be interpreted that the methylation profile of CTCF in the early stages of aging is very similar to the methylation profile in neonates.

When the scatter plot of senescence samples was examined, it was observed that the correlation was low. However, although the values were low, there was a positive correlation. This methylation profile is also similar to that of the infant. Based on this, it can be assumed that CTCF plays an active role in the period when development or aging is active, metabolism is in high change, and it is more passive in other developmental stages.

In the scatter plot created to better compare all results, especially when the values above 0.4 were examined, it was seen that the CTCF profile was very similar in newborns and in late adulthood. In the early adulthood period, on the contrary, a high negative correlation was observed.

4.2.2 Cancer-related methylation-gene expression correlation analysis

Secondly a similar comparison of methylation and gene expression correlations was carried out using cancer and healthy tissues. Based on the integrated analysis approach, 12 different cancer tissues have been used to test the hypothesis that CTCF can be effective role in cancer pathways.

Bone, pancreas, gastric, colorectal which are determined by network analysis as the main research subject; In addition to these, the data of bladder, brain, breast, colon, kidney, liver, lung, prostate, and small intestine tissues were examined.

Correlations of methylation and gene expression of colon samples were calculated for each probe in both cancerous samples and healthy samples, and red ellipses were shown as cancer and blue triangles as correlation values of healthy samples on the scatter plot. When this scatter plot was examined, a high positive correlation was observed in almost all probes in healthy samples, and a negative correlation was observed in cancerous samples. Based on this, it can be said that the CTCF methylation profile in colon cancer has completely changed and this caused the tissue to become cancerous.

Similar results were obtained in the literature review of colon cancer, which was found as a result of annotation in the network analysis. In a study conducted in August 2020 [104], it was emphasized that overexpression of CTCF causes colorectal cancer. When we look at our results, the negative correlation in cancerous samples can be explained by the increase in gene expression due to low methylation, and consequently, overexpressed CTCF causes cancer. Similarly, in a study conducted in 2017 ([105], 5 regions in CTCF for colorectal cancer were identified as methylation specific biomarkers. In addition to this study, 4 biomarker regions can be used for colorectal cancer with the results obtained from this thesis.

When correlation analysis was performed with bone data, another primary examination tissue, the results appeared to be completely different in cancerous tissues as expected. Although it cannot be said that there is a positive or negative correlation in the cancer patient, it is clear that a correlation appears to be the opposite of the healthy one. These results have proven that CTCF, which is a part of the network that is effective in bone development, also plays an active role in this pathway.

The literature review, performed by the author of this thesis in order to interpret the results, highlights a demonstrated relationship between CTCF and osteosarcoma based on previous studies. Interestingly, CTCF was shown to function together with genes that play a role in osteosarcoma formation, especially insulin-like growth factor 2 (IGF2) and H19 imprinted maternally expressed transcript (H19) [106]. Similarly, a study conducted in 2018 found that the change in the expression of the Glutamate metabotropic receptor 4 (GRM4) gene is associated with osteosarcoma and that CTCF is influences regulate of this gene's transcription [107]. However, previous studies have not specifically studied CTCF cooperates with. Accordingly, the following novel insight from this thesis was gained into this cancer type: a total of 5 methylation probe sites in the CTCF gene can be used as biomarkers for osteosarcoma.

Similar to colon and bone, another tissue type that manifests high alteration in terms of correlation between cancer and healthy tissue is breast. Combined breast tissue correlation scatter plot, which contains levels for both normal and cancer tissues, distinguishes cancerous samples from healthy samples. In another words, samples are to certain extend visually grouped together based on their cancer status (normal vs cancer). Such separation is accompanied by an overall difference in the correlation direction: cancerous tissues display positive correlation and healthy samples negative correlation. Thus, the relationship between methylation and gene expressions seems to differs between the normal breast tissue and cancer breast tissue. This is a demonstration for the phenomenon, which is referred to as "differential correlation pattern".

The literature was reviewed in order to interpret this result, CTCF has been shown to play a role in breast cancer based on its association with the Bax gene [108]. In a study conducted in 2017, it was found that activation of p53 in CTCF knockdown mice increased p21 and Bax expressions and this was found to be effective in breast cancer [109]. In the light of the results found in this thesis, it has been shown that 4 probe sites on CTCF can be used as breast cancer biomarkers.

When the results of pancreas, which is one of the primary research tissues as a result of network analysis, were examined, it was observed that healthy and cancerous samples were clustered together, just like breast cancer. There was a high positive correlation in healthy tissues and a negative correlation in cancer patients. When a detailed literature review was performed, it was shown that CTCF was not directly associated with pancreatic cancer. Nonetheless, genes thought to be working together with CTCF were associated with pancreatic cancer [110].

Based on these results, it can be said that CTCF acts as a tumor suppressor gene in pancreatic tissue, where its gene expression appears to be negatively correlated with methylation. As such, this thesis may be a pioneering study in inking CTCF to pancreatic cancer. A total of 3 methyl specific probe regions on CTCF can be used as a pancreatic cancer biomarker.

When the prostate data were examined, it could not be said that there was a positive or negative correlation in cancerous cells, but when compared with healthy results, it appeared to be an opposite correlation value. Generally, there was a negative correlation in healthy tissues. When a detailed literature review was performed, it was found that there is hypermethylation in the cancerous prostate tissue and therefore there is low expression of CTCF gene [111]. On the contrary in this study, it was found that CTCF was already low-expressed in normal tissues, and the opposite it was overexpressed in cancerous tissues. As a feature that distinguishes this thesis from previous studies, since its effect on gene expression can be examined more easily due to the fact that methylation alone was not examined, it was possible to examine genes that are normally upregulated or downregulated.

Similar to the results in this study, the result that CTCF is upregulated in prostate cancer was shown in a study conducted in January 2020 [112]. However, although it was stated in this study that it could not be used as a powerful biomarker, in our study, 2 methylation-specific prostate cancer biomarkers on CTCF were found that could properly discriminate the cancerous tissue, since probe regions was examined separately in our study.

Similar to the prostate cancer in the bladder results, it was observed that there was a negative correlation in healthy tissues, and a positive correlation in cancer patients.

When the literature searched, it was seen that there were very few studies on this subject. Only a study conducted in 2001 showed that there is an upregulation of CTCF in bladder cancer due to hypomethylation on CTCF [113]. There are not as many Noticeably, less candidate biomarker probe sites in were identified in the bladder when compared to the prostate. To be precise, only 3 methylation-specific probe sites have been identified as bladder cancer biomarkers.

When the gastric cancer result was examined, it was observed that there was a negative correlation in healthy tissues (similar to bladder and prostate) in the scatter plot and a positive correlation in cancers. However, when the correlation values were examined, it was thought that it would not be appropriate to use it to make a strong distinction because they were very low. A strong cancerous tissue separation can be made with more than 0.5 in only 1 probe area. Therefore only one probe site can be used as a biomarker. In the literature review, only one related article could be found, and it was stated without any detail that dysregulation of CTCF may cause gastrointestinal tumor [114].

When the small intestine results were examined, it was seen that there were similar correlation values in generally cancerous and healthy tissues. Only one probe located in the upstream region appears to possess a discriminatory feature.

Similarly, correlation values in brain, kidney, liver and lung were very close to each other in healthy and cancerous samples, so no distinction could be made. Although there are studies in the literature showing that there is cancer in brain tumors due to downregulation of CTCF [115], such a result was not obtained in our study. However, it should not be forgotten that although different studies are collected to increase the accuracy and a large number of samples are studied, since these results are obtained based on the data used, there may be a possibility of deviation in the results when using different samples.

CHAPTER 5

CONCLUSION AND FUTURE WORK

CTCF is widely known as “master weaver of the genome” due to its role in chromatin organization as an architectural protein. The involvement of this protein has remained as an interesting open question. The current literature and genetic databases only loosely links CTCF gene to cancer as a candidate tumor suppressor gene. This inconclusive situation is mainly due to the lack of experimental findings with conclusive evidence. The complex nature of this multi functional protein and myriad of DNA bindings sites makes the issue even more complicated. This work aim to approach the issue of CTCF tumor suppressor candidacy. One of the benefits of data mining techniques is their power in extracting interesting patterns from the data, which otherwise remain elusive. To meet that objective in this effort, two approaches were employed with this purpose: the GBA approach for in-depth interactomics data mining (based mainly on topology based clustering and functional annotation-based over representation analysis) and the integrative approach for combined methylome and transcriptome data-mining (based mainly on correlation analysis and taking into account distances from the transcription start site). While the former aimed to detect functionally-relevant patterns in the interaction network, the latter concentrated on identifying methylation-gene expression correlation patterns in diverse tissues/conditions. To the best of the authors knowledge, the accomplished network analysis is the most comprehensive study on the CTCF interactome up to date. Furthermore, it is the first cancer-focused interactomics study on CTCF. The overrepresentation analysis, based on functional enrichment tests, put the reconstructed network and the unraveled two functional modules into a mechanistic context. Remarkably cancer-related function annotation terms, such as “Signaling by TGF-beta family members pathway (Reactome Pathway)”, “SMAD Binding (GO-MF)” and “TGF-beta Receptor Signaling (Wiki cancer pathway)” were enriched in one for the modules. This finding highlight the potential cancer context and relevance of the reconstructed network and implicates CTCF in cancer by means of the GBA approach. Remarkably, this result also suggest that CTCF’s involvement in cancer may be mediated by the SMAD proteins. This observation was obtained by multiple bioinformatics tools, which increases the confidence on the obtained results. The described network is an important step towards more sophisticated interactomics studies. Especially the validation and in-depth elucidation of the SMADS-related functional module, using wet lab characterization studies, is of great importance.

The link between CTCF methylation and CTCF gene expression was studied separately in two biological contexts, namely, development and cancer. Two correlation methods, namely Pearson correlation and Spearman correlation used to compute the correlation coefficients. Only consensus results (when both coefficients are higher than 0.5) were considered as substantial and regarded as influential (impacting mRNA level). This level of scrutiny adds novelty to this work and enhances the outcome. The first step in integrated methylation and expression data analysis was to investigate the possible effect of CTCF in developmental biology, by focusing on the specific age ranges. As result of the carried out research, similar probe-level (methylation site-related) CTCF correlation patterns was observed especially in newborns and late adulthood. Such similar patterns apparently reflect commonalities between the CTCF activity in the two developmental stages, when compared to other investigated stages. Concordantly, these results suggest that this gene could be functioning especially in the early stages of development and aging. Subsequently, comparison of the methylation and expression CTCF correlation patterns between normal and cancer status of the same tissues types, highlight considerable probe-level differential correlation pattern (which indicates differential regulation of CTCF activity) in the following cancer types: osteosarcoma, colorectal, breast, pancreatic, bladder, and prostate. While some CTCF-associated methylated sites were observed to be positively correlated in colorectal and pancreatic tissues (with normal status), the same sites appears to be negatively correlated in the cancer counterparts, meaning that gene expression decreases with methylation of this sites in these cancerous tissues. In the latter, methylation supposedly results into reduction of gene expression, which is evident from the lower mRNA levels. Based on this, CTCF's regulation mechanism could be different between some cancer and normal condition at least for some distinct tissue types. On the contrary, certain CTCF-related methylation site were observed to be negatively correlated in bone, breast, prostate and bladder tissues (all with normal status), but positively correlated in the respective cancer tissues counterparts. Based on this, it can be concluded that methylation in these cancer tissues results into CTCF over-expression (Table 32). In sum such differential correlation patterns could be underlying separate regulatory mechanisms for regulating CTCF activity between normal and cancer tissues. Such a regulatory remodeling, in its turn, hints fort CTCF's supposed tumor suppressor role. Since probe-level analysis was preferred over averaging of probe values (as a summary gene methylation level), the overall differential correlation pattern (for all probes at once) remains unknown. Consequently, more work is needed to confidently conclude whether CTCF is a differentially correlated gene (DCG) or not. Thus, this report can be regarded as a initial effort in this new exciting direction.

Importantly, the results demonstrated the of the employed multi-omics data mining approach. To this end, potential novel DNA methylation biomarkers for cancer have been identified in each mentioned step of the accomplished integrated analysis. Relatedly, the differential correlation patterns can be explored further as DNA methylation signatures. In future studies, wet lab experiments, focusing on the highlighted methylation specific probe regions, will be helpful in confirming the accuracy of obtained data mining results and for validating the made predictions. Overall, the undertaken pioneering line of action, demonstrated the applicability of the employed approach, complemented previous findings, and yielded novel insights into the tumor suppressor candidacy of CTCF. The methodology can be extended into a broader workflows and frameworks. Additionally, it can be applied to other candidate tumor suppressor genes and other cancer types. As such, this work is a step towards more advanced data mining approaches, which are required for a broader elucidation of the putative CTCF tumor activity function. Future progress in the analytical tools for multi-omics, combined with the growing amount of biomedical data, is expected to pave the way for unraveling the molecular mechanisms, which underlie activity of the multifunctional protein CTCF in health and disease.

REFERENCES

[1] A. R. Sonawane, S. T. Weiss, K. Glass, and A. Sharma, "Network medicine in the age of biomedical big data," *Front. Genet.*, vol. 10, no. APR, pp. 1–16, 2019.

[2] Y. Ke-Jun, D. Jie, L. Ling-Yun, and P. Meng-Jia, "Network-based gene function inference method to predict optimal gene functions associated with fetal growth restriction," *Mol. Med. Rep.*, vol. 18, no. 3, pp. 3003–3010, 2018.

[3] C. J. Wolfe, I. S. Kohane, and A. J. Butte, "Systematic survey reveals general applicability of 'guilt-by-association' within gene coexpression networks," *BMC Bioinformatics*, vol. 6, pp. 1–10, 2005.

[4] S. S. Li, J. M. Tian, T. H. Wei, and H. R. Wang, "Identification of key genes for type 1 diabetes mellitus by network-based guilt by association," *Rev. Assoc. Med. Bras.*, vol. 66, no. 6, pp. 778–783, 2020.

[5] C. J. Gloeckner and P. Porras, "Guilt-by-Association – Functional Insights Gained From Studying the LRRK2 Interactome," *Front. Neurosci.*, vol. 14, no. May, pp. 1–14, 2020.

[6] L. D. Moore, T. Le, and G. Fan, "DNA methylation and its basic function," *Neuropsychopharmacology*, vol. 38, no. 1, pp. 23–38, 2013.

[7] M. Kulis and M. Esteller, "DNA Methylation and Cancer," *Adv. Genet.*, vol. 70, no. C, pp. 27–56, 2010.

[8] M. Esteller, "Cancer epigenomics: DNA methylomes and histone-modification maps," *Nat. Rev. Genet.*, vol. 8, no. 4, pp. 286–298, 2007.

[9] O. T. Avery, C. M. Macleod, and M. McCarty, "Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type iii," *J. Exp. Med.*, 1944.

[10] R. Holliday and J. Pugh, "DNA modification mechanisms and gene activity during development," *Science (80-.).*, 1975.

[11] A. M. Orgueira, “Hidden among the crowd: Differential DNA methylationexpression correlations in cancer occur at important oncogenic pathways,” *Front. Genet.*, 2015.

[12] M. Gardiner-Garden and M. Frommer, “CpG Islands in vertebrate genomes,” *J. Mol. Biol.*, 1987.

[13] V. K. Rakyan *et al.*, “DNA methylation profiling of the human major histocompatibility complex: A pilot study for the Human Epigenome Project,” *PLoS Biol.*, 2004.

[14] R. A. Irizarry *et al.*, “Genome-wide methylation analysis of human colon cancer reveals similar hypo-and hypermethylation at conserved tissue-specific CpG island shores,” *Nat. Genet.*, 2009.

[15] A. Hellman and A. Chess, “Gene body-specific methylation on the active X chromosome,” *Science (80-.).*, 2007.

[16] D. Aran, G. Toperoff, M. Rosenberg, and A. Hellman, “Replication timing-related and gene body-specific methylation of active human genes,” *Hum. Mol. Genet.*, 2011.

[17] W. Xie *et al.*, “Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome,” *Cell*, 2012.

[18] M. F. Paz *et al.*, “A systematic profile of DNA methylation in human cancer cell lines,” *Cancer Res.*, 2003.

[19] O. L. Caballero and Y. T. Chen, “Cancer/testis (CT) antigens: Potential targets for immunotherapy,” *Cancer Science*. 2009.

[20] V. Greger, E. Passarge, W. Höpping, E. Messmer, and B. Horsthemke, “Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma,” *Hum. Genet.*, vol. 83, no. 2, pp. 155–158, 1989.

[21] A. Catteau and J. R. Morris, “BRCA1 methylation: A significant role in tumour development?,” *Semin. Cancer Biol.*, 2002.

[22] E. Gordian, K. Ramachandran, and R. Singal, “Methylation mediated silencing of TMS1 in breast cancer and its potential contribution to docetaxel cytotoxicity,” *Anticancer Res.*, 2009.

[23] R. Hrdlickova, M. Toloue, B. Tian, M. Genetics, and R. New, “p-xylene, 26.07.2017, 11:17AM,” vol. 8, no. 1, pp. 1–24, 2017.

[24] C. Manzoni *et al.*, “Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences,” *Brief. Bioinform.*, vol. 19, no. 2, pp. 286–302, 2018.

[25] S. Picelli, “Single-cell RNA-sequencing : The future of genome biology is now,” *RNA Biol.*, vol. 14, no. 5, pp. 637–650, 2017.

[26] M. J. Hangauer, I. W. Vaughn, and M. T. McManus, “Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs,” *PLoS Genet.*, 2013.

[27] R. Andersson, A. Sandelin, and C. G. Danko, “A unified architecture of transcriptional regulatory elements,” *Trends in Genetics*. 2015.

[28] J. Shendure, “The beginning of the end for microarrays?,” *Nat. Methods*, 2008.

[29] A. Fukushima, “DiffCorr: An R package to analyze and visualize differential correlations in biological networks,” *Gene*, 2013.

[30] P. Heger, B. Marin, M. Bartkuhn, E. Schierenberg, and T. Wiehe, “The chromatin insulator CTCF and the emergence of metazoan diversity,” *Proc. Natl. Acad. Sci. U. S. A.*, 2012.

[31] J. Ziatanova and P. Caiafa, “CTCF and its protein partners: Divide and rule?,” *J. Cell Sci.*, vol. 122, no. 9, pp. 1275–1284, 2009.

[32] R. Ohlsson, R. Renkawitz, and V. Lobanenkov, “CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease,” *Trends in Genetics*. 2001.

[33] J. A. Wallace and G. Felsenfeld, “We gather together: insulators and genome organization,” *Current Opinion in Genetics and Development*. 2007.

[34] M. Furlan-Magaril *et al.*, “An insulator embedded in the chicken α-globin locus regulates chromatin domain configuration and differential gene expression,” *Nucleic Acids Res.*, 2011.

[35] M. Merkenschlager and E. P. Nora, “CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation,” *Annu. Rev. Genomics Hum. Genet.*, 2016.

[36] H. Hashimoto, D. Wang, J. R. Horton, X. Zhang, V. G. Corces, and X. Cheng, “Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA,” *Mol. Cell*, 2017.

[37] M. J. MacPherson, L. G. Beatty, W. Zhou, M. Du, and P. D. Sadowski, “The CTCF Insulator Protein Is Posttranslationally Modified by SUMO,” *Mol. Cell. Biol.*, 2009.

[38] C. T. Ong, K. Van Bortle, E. Ramos, and V. G. Corces, “XPoly(ADPribosyl)ation regulates insulator function and intrachromosomal interactions in drosophila,” *Cell*, 2013.

[39] G. A. Busslinger *et al.*, “Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl,” *Nature,* 2017.

[40] T. Hirayama, E. Tarusawa, Y. Yoshimura, N. Galjart, and T. Yagi, “CTCF Is Required for Neural Development and Stochastic Expression of Clustered Pcdh Genes in Neurons,” *Cell Rep.*, 2012.

[41] I. V. Chernukhin *et al.*, “Physical and functional interaction between two pluripotent proteins, the Y-box DNA/RNA-binding factor, YB-1, and the multivalent zinc finger factor, CTCF,” *J. Biol. Chem.,* 2000.

[42] M. E. Donohoe, L. F. Zhang, N. Xu, Y. Shi, and J. T. Lee, “Identification of a Ctcf Cofactor, Yy1, for the X Chromosome Binary Switch,” *Mol. Cell,* 2007.

[43] P. A. Defossez *et al.*, “The human enhancer blocker CTC-binding factor interacts with the transcription factor Kaiso,” *J. Biol. Chem.,* 2005.

[44] P. Majumder, J. A. Gomez, B. P. Chadwick, and J. M. Boss, “The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance chromatin interactions,” *J. Exp. Med.,* 2008.

[45] T. M. Yusufzai and G. Felsenfeld, “The 5′-HS4 chicken β-globin insulator is a CTCF-dependent nuclear matrix-associated element,” *Proc. Natl. Acad. Sci. U. S. A.,* 2004.

[46] T. Li *et al.*, “CTCF Regulates Allelic Expression of Igf2 by Orchestrating a Promoter-Polycomb Repressive Complex 2 Intrachromosomal Loop,” *Mol. Cell. Biol.,* 2008.

[47] M. Lutz, “Transcriptional repression by the insulator protein CTCF involves histone deacetylases,” *Nucleic Acids Res.,* 2000.

[48] K. Ishihara, M. Oshimura, and M. Nakao, “CTCF-Dependent Chromatin Insulator Is Linked to Epigenetic Remodeling,” *Mol. Cell,* 2006.

[49] W. Stedman, H. Kang, S. Lin, J. L. Kissil, M. S. Bartolomei, and P. M. Lieberman, “Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators,” *EMBO J.,* 2008.

[50] K. S. Wendt *et al.*, “Cohesin mediates transcriptional insulation by CCCTCbinding factor,” *Nature,* 2008.

[51] E. D. Rubio *et al.*, “CTCF physically links cohesin to chromatin,” *Proc. Natl. Acad. Sci. U. S. A.,* 2008.

[52] M. Mohan *et al.*, “The Drosophila insulator proteins CTCF and CP190 link enhancer blocking to body patterning,” *EMBO J.,* 2007.

[53] S. D. Pope and R. Medzhitov, “Emerging Principles of Gene Expression Programs and Their Regulation,” *Mol. Cell,* vol. 71, no. 3, pp. 389–397, 2018.

[54] F. P. Fiorentino and A. Giordano, “The tumor suppressor role of CTCF,” *J. Cell. Physiol.,* vol. 227, no. 2, pp. 479–492, 2012.

[55] T. Xiao, J. Wallace, and G. Felsenfeld, “Specific Sites in the C Terminus of CTCF Interact with the SA2 Subunit of the Cohesin Complex and Are Required for Cohesin-Dependent Insulation Activity,” *Mol. Cell. Biol.,* 2011.

[56] V. Parelho *et al.*, “Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms,” *Cell*, 2008.

[57] R. Nativio *et al.*, “Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus,” *PLoS Genet.*, 2009.

[58] J. G. Kirkland, J. R. Raab, and R. T. Kamakaka, “TFIIIC bound DNA elements in nuclear organization and insulation,” *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*. 2013.

[59] L. Carrire *et al.*, “Genomic binding of Pol III transcription machinery and relationship with TFIIS transcription factor distribution in mouse embryonic stem cells,” *Nucleic Acids Res.*, 2012.

[60] A. Henderson, A. Ptito, J. Doyon, and M. Levin, “Imaging of post-stroke arm motor recovery and compensation,” *Exp. Brain Res.*, vol. 15, no. 4, pp. 234–246, 2015.

[61] H. Heath *et al.*, “CTCF regulates cell cycle progression of αβ T cells in the thymus,” *EMBO J.*, 2008.

[62] P. Delgado-Olguín *et al.*, “Epigenetic repression of cardiac progenitor gene expression by Ezh2 is required for postnatal cardiac homeostasis,” *Nature Genetics*. 2012.

[63] N. P. Gomes and J. M. Espinosa, “Gene-specific repression of the p53 target gene PUMA via intragenic CTCF-Cohesin binding,” *Genes Dev.*, 2010.

[64] J. A. Beagan *et al.*, “YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment,” *Genome Res.*, 2017.

[65] M. Franke *et al.*, “Formation of new chromatin domains determines pathogenicity of genomic duplications,” *Nature*, 2016.

[66] M. Mallo, D. M. Wellik, and J. Deschamps, “Hox genes and regional patterning of the vertebrate body plan,” *Developmental Biology*. 2010.

[67] R. G. Arzate-Mejía, F. Recillas-Targa, and V. G. Corces, “Developing in 3D: the role of CTCF in cell differentiation,” *Development*, vol. 145, no. 6, 2018.

[68] E. K. Allen *et al.*, “SNP-mediated disruption of CTCF binding at the IFITM3 promoter is associated with risk of severe influenza in humans,” *Nat. Med.*, 2017.

[69] M. Lazniewski, W. K. Dawson, A. M. Rusek, and D. Plewczynski, “One protein to rule them all: The role of CCCTC-binding factor in shaping human genome in health and disease,” *Semin. Cell Dev. Biol.*, vol. 90, no. August, pp. 114–127, 2019.

[70] A. Gregor *et al.*, “De novo mutations in the genome organizer CTCF cause intellectual disability,” *Am. J. Hum. Genet.*, 2013.

[71] Y. Shi, X. Bin Su, K. Y. He, B. H. Wu, B. Y. Zhang, and Z. G. Han, “Chromatin accessibility contributes to simultaneous mutations of cancer genes,” *Sci. Rep.*, 2016.

[72] D. Chitayat, J. M. Friedman, L. Anderson, and J. E. Dimmick, “Hepatocellular carcinoma in a child with familial Russell-Silver syndrome,” *Am. J. Med. Genet.*, 1988.

[73] E. Bruckheimer and A. Abrahamov, “Russell-Silver syndrome and Wilms tumor,” *The Journal of Pediatrics*. 1993.

[74] G. R. Weiss and M. B. Garnick, “Testicular cancer in a Russell-Silver dwarf,” *J. Urol.*, 1981.

[75] M. B. Draznin, M. W. Stelling, and A. J. Johanson, “Silver-Russell syndrome and craniopharyngioma,” *J. Pediatr.*, 1980.

[76] S. Oh, C. Oh, and K. H. Yoo, “Functional roles of CTCF in breast cancer,” *BMB Rep.*, vol. 50, no. 9, pp. 445–453, 2017.

[77] J. E. J. Rasko *et al.*, “Cell growth inhibition by the multifunctional multivalent zinc-finger factor CTCF,” *Cancer Res.*, vol. 61, no. 16, pp. 6002–6007, 2001.

[78] P. Shannon *et al.*, “Cytoscape: A software Environment for integrated models of biomolecular interaction networks,” *Genome Res.*, 2003.

[79] D. Warde-Farley *et al.*, “The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function,” *Nucleic Acids Res.*, 2010.

[80] J. Montojo *et al.*, “GeneMANIA cytoscape plugin: Fast gene function predictions on the desktop,” *Bioinformatics,* 2010.

[81] Y. Assenov, F. Ramírez, S. E. S. E. Schelhorn, T. Lengauer, and M. Albrecht, “Computing topological parameters of biological networks,” *Bioinformatics*, 2008.

[82] G. D. Bader and C. W. V. Hogue, “An automated method for finding molecular complexes in large protein interaction networks,” *BMC Bioinformatics,* 2003.

[83] O. Garcia *et al.*, “GOlorize: A Cytoscape plug-in for network visualization with Gene Ontology-based layout and coloring,” *Bioinformatics,* 2007.

[84] G. Bindea *et al.*, “ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks,” *Bioinformatics*, 2009.

[85] G. Wu, X. Feng, and L. Stein, “A human functional protein interaction network and its application to cancer data analysis,” *Genome Biol.*, 2010.

[86] Y. Liao, J. Wang, E. J. Jaehnig, Z. Shi, and B. Zhang, “WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs,” *Nucleic Acids Res.*, 2019.

[87] R. Alonso *et al.*, “Babelomics 5.0: Functional interpretation for new generations of genomic data,” *Nucleic Acids Res.*, 2015.

[88] D. W. Huang, B. T. Sherman, and R. A. Lempicki, “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources,” *Nat. Protoc.*, 2009.

[89] L. Zhu *et al.*, “CellWhere: Graphical display of interaction networks organized on subcellular localizations,” *Nucleic Acids Res.*, 2015.

[90] A. D. Moher D, Liberati A, Tetzlaff J, “PRISMA 2009 Flow Diagram,” *The PRISMA statement*. 2009.

[91] X. Ma, Y. W. Wang, M. Q. Zhang, and A. F. Gazdar, “DNA methylation data analysis and its application to cancer research,” *Epigenomics*. 2013.

[92] K. Nakabayashi, “Illumina humanmethylation beadchip for genome-wide DNA methylation profiling: Advantages and limitations,” in *Handbook of Nutrition, Diet, and Epigenetics*, 2019.

[93] Hansen. K. D., “IlluminaHumanMethylation450kmanifest: Annotation for Illumina’s 450k methylation arrays,” *R. Package Version 0.4.0*, 2012. .

[94] L. J. Zhu *et al.*, “ChIPpeakAnno: A Bioconductor package to annotate ChIP-seq and ChIP-chip data,” *BMC Bioinformatics*, 2010.

[95] M. Carlson, “org.Mm.eg.db: Genome wide annotation for Mouse. R package version 3.8.2.,” *Bioconductor*. 2019.

[96] T. Metsalu and J. Vilo, “ClustVis: A web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap,” *Nucleic Acids Res.*, 2015.

[97] M. D. Robinson and T. P. Speed, “A comparison of affymetrix gene expression arrays,” *BMC Bioinformatics*, 2007.

[98] M. M. Mukaka, “Statistics corner: A guide to appropriate use of correlation coefficient in medical research,” *Malawi Med. J.*, 2012.

[99] B. Sutariya, D. Jhonsa, and M. N. Saraf, “TGF-β: The connecting link between nephropathy and fibrosis,” *Immunopharmacology and Immunotoxicology*. 2016.

[100] M. Ashburner *et al.*, “Gene ontology: Tool for the unification of biology,” *Nature Genetics*. 2000.

[101] G. Chen, C. Deng, and Y. P. Li, “TGF-β and BMP signaling in osteoblast differentiation and bone formation,” *International Journal of Biological Sciences*. 2012.

[102] S. Faure, M. A. Lee, T. Keller, P. Ten Dijke, and M. Whitman, “Endogenous patterns of TGFβ superfamily signaling during early Xenopus development,” *Development*, 2000.

[103] W. Wang, F. V. Mariani, R. M. Harland, and K. Luo, “Ski represses bone morphogenic protein signaling in Xenopus and mammalian cells,” *Proc. Natl. Acad. Sci. U. S. A.*, 2000.

[104] Q. Lai *et al.*, “CTCF promotes colorectal cancer cell proliferation and chemotherapy resistance to 5-FU via the P53-hedgehog axis,” *Aging (Albany. NY).*, 2020.

[105] F. Model *et al.*, “Identification and validation of colorectal neoplasia-specific methylation markers for accurate classification of disease,” *Mol. Cancer Res.*, 2007.

[106] G. A. Ulaner *et al.*, “Loss of imprinting of IGF2 and H19 in osteosarcoma is accompanied by reciprocal methylation changes of a CTCF-binding site,” *Hum. Mol. Genet.*, 2003.

[107] Y. Pang, J. Zhao, M. Fowdur, Y. Liu, H. Wu, and M. He, “To Explore the Mechanism of the GRM4 Gene in Osteosarcoma by RNA Sequencing and Bioinformatics Approach,” *Med. Sci. Monit. Basic Res.*, 2018.

[108] C. F. Méndez-Catalá *et al.*, “A novel mechanism for CTCF in the epigenetic regulation of bax in breast cancer cells,” *Neoplasia (United States)*, 2013.

[109] J. Y. Lee, M. Mustafa, C. Y. Kim, and M. H. Kim, “Depletion of CTCF in breast cancer cells selectively induces cancer cell death via p53,” *J. Cancer*, 2017.

[110] D. Wang, C. Li, and X. Zhang, “The promoter methylation status and mRNA expression levels of CTCF and SIRT6 in sporadic breast cancer.,” *DNA Cell Biol.*, 2014.

[111] N. A. Damaschke *et al.*, “CTCF loss mediates unique DNA hypermethylation landscapes in human cancers,” *Clin. Epigenetics*, 2020.

[112] D. Höflmayer *et al.*, “Expression of CCCTC-binding factor (CTCF) is linked to poor prognosis in prostate cancer,” *Mol. Oncol.*, vol. 14, no. 1, pp. 129–138, 2020.

[113] D. Takai, F. A. Gonzales, Y. C. Tsai, M. J. Thayer, and P. A. Jones, “Large scale mapping of methylcytosines in CTCF-binding sites in the human H19 promoter and aberrant hypomethylation in human bladder cancer,” *Hum. Mol. Genet.*, 2001.

[114] C. Fang *et al.*, “Cancer-specific CTCF binding facilitates oncogenic transcriptional dysregulation,” *bioRxiv*. 2020.

[115] W. A. Flavahan *et al.*, “Insulator dysfunction and oncogene activation in IDH mutant gliomas,” *Nature*, 2016.

Appendix

Appendix A

Numerical Summary of the Literature Related with the Candidate Tumor Suppressor Gene CTCF

Year	Specific year	Average	Running average	Cumulative sum
1998	2	5	#N/A	2
1999	1	5	#N/A	3
2000	3	5	2	6
2001	2	5	2	8
2002	1	5	2	9
2003	3	5	2	12
2004	2	5	2	14
2005	5	5	3	19
2006	4	5	4	23
2007	5	5	5	28
2008	5	5	5	33
2009	2	5	4	35
2010	9	5	5	44
2011	7	5	6	51
2012	9	5	8	60
2013	3	5	6	63
2014	12	5	8	75
2015	8	5	8	83
2016	9	5	10	92
2017	11	5	9	103
2018	4	5	8	107
2019	6	5	7	113
MEAN	5			

Table A.1: Numerical Summary of the Literature Related with the Candidate Tumor Suppressor Gene CTCF. Table shows the yearly number, average (arithmetic mean), running average and cumulative sum of the number of publications, related to the CTCF tumor suppressor gene, in the NCBI Pubmed database (https://pubmed.ncbi.nlm.nih.gov/). Numbers were obtained by a simple search (using default options) using the query term “tumor suppressor CTCF”. The yearly chronological record covers the timeline 1998 and 2019.

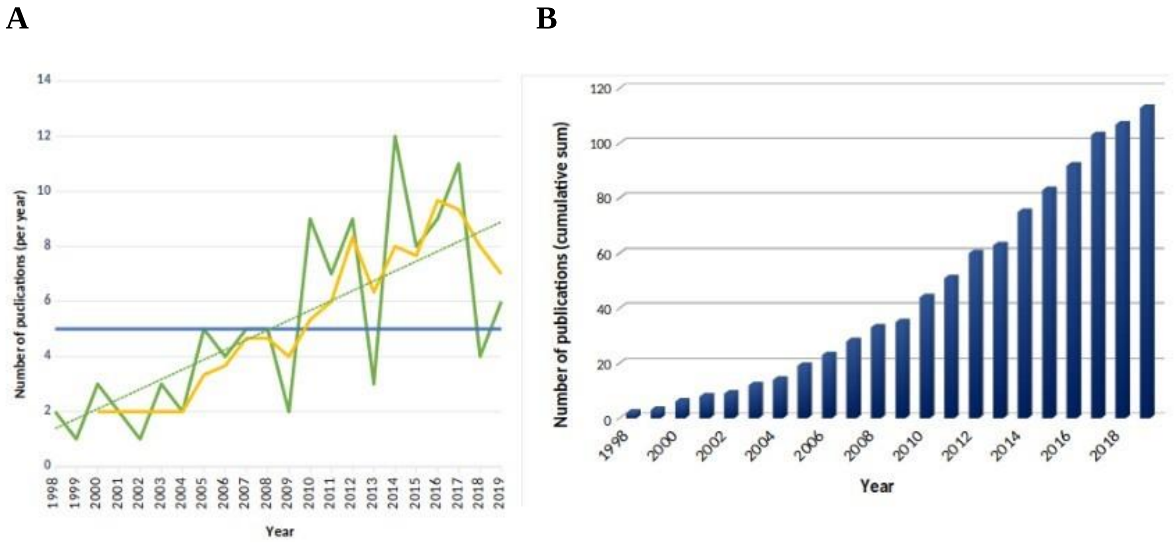


Figure A.1: Growing Body of the Literature Related with the Candidate Tumor Suppressor Gene CTCF. The figure is based on the data, presented in the Table A.1. A. Number of publication per year. In addition to the number of yearly publications, the average, the running average and the trend line (linear fit for the number of yearly publications) are shown. B. Cumulative sum of publication per year.

Appendix B

B.1 Annotation for the Selected Development-related CTCF Microarray Samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Age range	Cell type	Selected file type
2016	GSE79056	Methylation	36	36	fetal	cord blood	Series matix file (TXT)
2018	GSE100197	Methylation	41	102	fetal	cental Tissue Cod whole blood	Series matix file (TXT)
2015	GSE74738	Methylation	21	79	fetal	Placental chronic villi	Series matix file (TXT)
2008	GSE9984	Gene_expression	12	12	fetal	Placental Tissue	Series matix file (TXT)
2015	GSE60403	Gene_expression	16	16	fetal	cord blood	Series matix file (TXT)
2016	GSE83556	Gene_expression	18	40	fetal	amniotic fluid	Series matix file (TXT)
2013	GSE48521	Gene_expression	16	16	fetal	cell free mRNA	Series matix file (TXT)
2017	GSE101141	Gene_expression	20	20	fetal	cell free mRNA amniotic fluid	Series matix file (TXT)
2017	GSE86171	Gene_expression	16	16	fetal	villus cytotrophoblast	Series matix file (TXT)

Table B.1.1: Description of the selected microarray datasets of fetal samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Age range	Cell type	Selected file type
2019	GSE129841	Methylation	40	114	newborn	cord blood	Series matix file (TXT)
2014	GSE62924	Methylation	20	38	newborn	cord blood	Series matix file (TXT)
2012	GSE30870	Methylation	20	40	newborn	cord blood	Series matix file (TXT)
2017	GSE3240	Gene_expression	29	29	newborn	cord blood	Series matix file (TXT)
2012	GSE82155	Gene_expression	11	46	newborn	epicardial adipose tissue	Series matix file (TXT)
2012	GSE35683	Gene_expression	30	30	newborn	umbilical cord blood	Series matix file (TXT)
2012	GSE39840	Gene_expression	10	20	newborn	umbilical cord blood	Series matix file (TXT)

Table B.1.2: Description of the selected microarray datasets of newborn samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Age range	Cell type	Selected file type
2015	GSE64495	Methylation	6	113	infant	whole blood	Series matix file (TXT)
2014	GSE60598	Methylation	29	43	infant	whole blood	Series matix file (TXT)
2015	GSE67444	Methylation	25	70	infant	whole blood	Series matix file (TXT)
2011	GSE26378	Gene_expression	15	103	infant	whole blood	Series matix file (TXT)
2012	GSE32140	Gene_expression	22	147	infant	whole blood	Series matix file (TXT)
2011	GSE26440	Gene_expression	23	130	infant	whole blood	Series matix file (TXT)

Table B.1.3: Description of the selected microarray datasets of infant samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Age range	Cell type	Selected file type
2017	GSE104812	Methylation	32	48	5-17	whole blood	Series matix file (TXT)
2015	GSE64495	Methylation	13	113	5-17	whole blood	Series matix file (TXT)
2015	GSE73103	Methylation	106	355	5-17	whole blood	Series matix file (TXT)
2013	GSE35571	Gene_expression	64	131	5-17	peripheral blood	Series matix file (TXT)
2015	GSE72439	Gene_expression	51	60	infant	whole blood	Series matix file (TXT)
2009	GSE14844	Gene_expression	36	36	infant	whole blood	Series matix file (TXT)

Table B.1.4: Description of the selected microarray datasets of childhood (5-17) samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Age range	Cell type	Selected file type
2012	GSE41169	Methylation	25	95	18-40	whole blood	Series matix file (TXT)
2019	GSE77056	Methylation	27	47	18-40	whole blood	Series matix file (TXT)
2015	GSE59509	Methylation	6	42	18-40	whole blood	Series matix file (TXT)
2017	GSE107737	Methylation	12	24	18-40	whole blood	Series matix file (TXT)
2018	GSE93272	Gene_expression	13	275	18-40	whole blood	Series matix file (TXT)
2019	GSE110551	Gene_expression	42	156	18-40	whole blood	Series matix file (TXT)
2018	GSE93777	Gene_expression	15	448	18-40	whole blood	Series matix file (TXT)

Table B.1.5: Description of the selected microarray datasets of early adulthood (18-40) samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Age range	Cell type	Selected file type
2018	GSE99624	Methylation	42	48	41-80	whole peripheral blood	Series matix file (TXT)
2015	GSE59509	Methylation	3	42	41-80	whole blood	Series matix file (TXT)
2014	GSE52113	Methylation	12	24	41-80	whole blood	Series matix file (TXT)
2015	GSE62003	Methylation	25	70	41-80	whole blood	Series matix file (TXT)
2018	GSE93272	Gene_expression	28	275	41-80	whole blood	Series matix file (TXT)
2017	GSE95233	Gene_expression	22	124	41-80	whole blood	Series matix file (TXT)
2019	GSE110551	Gene_expression	32	156	41-80	whole blood	Series matix file (TXT)

Table B.1.6: Description of the selected microarray datasets of late adulthood (41-80) samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Age range	Cell type	Selected file type
2012	GSE30870	Methylation	11	40	80(+)	whole blood	Series matix file (TXT)
2018	GSE99624	Methylation	4	48	80(+)	whole peripheral blood	Series matix file (TXT)
2017	GSE67530	Methylation	11	144	80(+)	whole blood	Series matix file (TXT)
2017	GSE95233	Gene_expression	12	124	80(+)	whole blood	Series matix file (TXT)
2014	GSE57065	Gene_expression	11	107	80(+)	whole blood	Series matix file (TXT)
2018	GSE93272	Gene_expression	3	275	80(+)	whole blood	Series matix file (TXT)

Table B.1.7: Description of the selected microarray datasets of senescence (80+) samples

B.2 Annotation for the Selected Cancer Tissue-related CTCF Microarray Samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2017	GSE52955	Methylation	25	83	bladder	Series matix file (TXT)
2015	GSE69463	Methylation	4	8	bladder	Series matix file (TXT)
2013	GSE41525	Methylation	1	8	bladder	Series matix file (TXT)
2011	GSE30522	Gene_expression	2	17	bladder	Series matix file (TXT)
2007	GSE7476	Gene_expression	9	12	bladder	Series matix file (TXT)
2013	GSE31189	Gene_expression	19	92	bladder	Series matix file (TXT)

Table B.2.1: Description of the selected microarray datasets of bladder cancer samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2015	GSE58770	Methylation	16	24	bone	Series matix file (TXT)
2017	GSE97529	Methylation	20	36	bone	Series matix file (TXT)
2020	GSE125645	Methylation	13	66	bone	Series matix file (TXT)
2017	GSE87437	Gene_expression	21	21	bone	Series matix file (TXT)
2011	GSE33458	Gene_expression	16	18	bone	Series matix file (TXT)
2019	GSE129091	Gene_expression	12	12	bone	Series matix file (TXT)

Table B.2.2: Description of the selected microarray datasets of bone cancer samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2018	GSE103659	Methylation	148	181	brain	Series matix file (TXT)
2019	GSE128654	Methylation	35	74	brain	Series matix file (TXT)
2019	GSE123678	Methylation	70	78	brain	Series matix file (TXT)
2014	GSE50774	Gene_expression	21	66	brain	Series matix file (TXT)
2016	GSE43378	Gene_expression	50	50	brain	Series matix file (TXT)
2016	GSE73038	Gene_expression	182	182	brain	Series matix file (TXT)

Table B.2.3: Description of the selected microarray datasets of brain cancer samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2018	GSE39451	Methylation	10	38	breast	Series matix file (TXT)
2017	GSE78758	Methylation	93	116	breast	Series matix file (TXT)
2017	GSE72245	Methylation	34	118	breast	Series matix file (TXT)
2018	GSE103668	Gene_expression	21	21	breast	Series matix file (TXT)
2018	GSE120129	Gene_expression	55	108	breast	Series matix file (TXT)
2017	GSE102907	Gene_expression	61	61	breast	Series matix file (TXT)

Table B.2.4: Description of the selected microarray datasets of breast cancer samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2017	GSE91370	Methylation	2	8	colon	Series matix file (TXT)
2014	GSE59134	Methylation	12	21	colon	Series matix file (TXT)
2012	GSE42752	Methylation	22	63	colon	Series matix file (TXT)
2015	GSE62932	Gene_expression	15	68	colon	Series matix file (TXT)
2018	GSE92921	Gene_expression	6	59	colon	Series matix file (TXT)
2017	GSE85043	Gene_expression	15	29	colon	Series matix file (TXT)

Table B.2.5: Description of the selected microarray datasets of colon cancer samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2018	GSE103186	Methylation	130	191	gastric	Series matix file (TXT)
2017	GSE89269	Methylation	11	22	gastric	Series matix file (TXT)
2019	GSE97686	Methylation	3	9	gastric	Series matix file (TXT)
2016	GSE79973	Gene_expression	10	20	gastric	Series matix file (TXT)
2015	GSE62254	Gene_expression	23	300	gastric	Series matix file (TXT)
2017	GSE54129	Gene_expression	111	132	gastric	Series matix file (TXT)

Table B.2.6: Description of the selected microarray datasets of gastric cancer samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2019	GSE92482	Methylation	24	24	kidney	Series matix file (TXT)
2018	GSE105288	Methylation	35	88	kidney	Series matix file (TXT)
2019	GSE113501	Methylation	144	144	kidney	Series matix file (TXT)
2010	GSE23629	Gene_expression	32	32	kidney	Series matix file (TXT)
2014	GSE46699	Gene_expression	67	130	kidney	Series matix file (TXT)
2017	GSE73731	Gene_expression	104	265	kidney	Series matix file (TXT)

Table B.2.7: Description of the selected microarray datasets of kidney cancer samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2019	GSE112791	Methylation	67	329	liver	Series matix file (TXT)
2018	GSE67170	Methylation	49	89	liver	Series matix file (TXT)
2019	GSE99036	Methylation	15	32	liver	Series matix file (TXT)
2019	GSE101685	Gene_expression	24	32	liver	Series matix file (TXT)
2019	GSE112791	Gene_expression	72	329	liver	Series matix file (TXT)
2019	GSE88839	Gene_expression	35	38	liver	Series matix file (TXT)

Table B.2.8: Description of the selected microarray datasets of liver cancer samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2019	GSE108124	Methylation	87	138	lung	Series matix file (TXT)
2015	GSE66836	Methylation	76	183	lung	Series matix file (TXT)
2017	GSE75008	Methylation	40	80	lung	Series matix file (TXT)
2019	GSE114761	Gene_expression	42	42	lung	Series matix file (TXT)
2017	GSE108492	Gene_expression	95	95	lung	Series matix file (TXT)
2017	GSE101929	Gene_expression	66	66	lung	Series matix file (TXT)

Table B.2.9: Description of the selected microarray datasets of lung cancer samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2018	GSE117852	Methylation	32	32	pancreas	Series matix file (TXT)
2014	GSE53051	Methylation	16	220	pancreas	Series matix file (TXT)
2017	GSE80241	Methylation	15	17	pancreas	Series matix file (TXT)
2017	GSE106189	Gene_expression	35	35	pancreas	Series matix file (TXT)
2018	GSE112282	Gene_expression	6	48	pancreas	Series matix file (TXT)
2014	GSE42404	Gene_expression	22	22	pancreas	Series matix file (TXT)

Table B.2.10: Description of the selected microarray datasets of pancreas cancer samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2017	GSE84043	Methylation	86	233	prostate	Series matix file (TXT)
2017	GSE76938	Methylation	73	136	prostate	Series matix file (TXT)
2018	GSE112047	Methylation	31	47	prostate	Series matix file (TXT)
2009	GSE17951	Gene_expression	141	154	prostate	Series matix file (TXT)
2015	GSE46602	Gene_expression	36	50	prostate	Series matix file (TXT)
2014	GSE55945	Gene_expression	13	21	prostate	Series matix file (TXT)

Table B.2.11: Description of the selected microarray datasets of prostate cancer samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2016	GSE73832	Methylation	22	133	small_intestine	Series matix file (TXT)
2013	GSE34387	Methylation	32	220	small_intestine	Series matix file (TXT)
2015	GSE61467	Methylation	10	56	small_intestine	Series matix file (TXT)
2008	GSE8167	Gene_expression	32	32	small_intestine	Series matix file (TXT)
2019	GSE132542	Gene_expression	29	29	small_intestine	Series matix file (TXT)
2008	GSE9576	Gene_expression	3	12	small_intestine	Series matix file (TXT)

Table B.2.12: Description of the selected microarray datasets of small intestine cancer samples

B.3 Annotation for the Selected Healthy Tissue-related CTCF Microarray Samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2017	GSE52955	Methylation	5	83	bladder	Series matix file (TXT)
2014	GSE50192	Methylation	4	70	bladder	Series matix file (TXT)
2012	GSE31848	Methylation	2	153	bladder	Series matix file (TXT)
2011	GSE30522	Gene_expression	3	17	bladder	Series matix file (TXT)
2007	GSE7476	Gene_expression	3	12	bladder	Series matix file (TXT)
2013	GSE31189	Gene_expression	5	92	bladder	Series matix file (TXT)

Table B.3.1: Description of the selected microarray datasets of bladder samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2013	GSE51759	Methylation	4	16	bone	Series matix file (TXT)
2015	GSE64490	Methylation	48	48	bone	Series matix file (TXT)
2014	GSE58477	Methylation	10	72	bone	Series matix file (TXT)
2010	GSE19429	Gene_expression	17	200	bone	Series matix file (TXT)
2018	GSE118985	Gene_expression	40	750	bone	Series matix file (TXT)
2014	GSE33075	Gene_expression	5	27	bone	Series matix file (TXT)

Table B.3.2: Description of the selected microarray datasets of bone samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2015	GSE64511	Methylation	49	372	brain	Series matix file (TXT)
2016	GSE80970	Methylation	138	286	brain	Series matix file (TXT)
2016	GSE79122	Methylation	9	78	brain	Series matix file (TXT)
2008	GSE11882	Gene_expression	109	173	brain	Series matix file (TXT)
2006	GSE5281	Gene_expression	74	161	brain	Series matix file (TXT)
2013	GSE50161	Gene_expression	13	130	brain	Series matix file (TXT)

Table B.3.3: Description of the selected microarray datasets of brain samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2019	GSE124367	Methylation	12	12	breast	Series matix file (TXT)
2015	GSE52865	Methylation	20	57	breast	Series matix file (TXT)
2011	GSE29290	Methylation	8	22	breast	Series matix file (TXT)
2011	GSE30010	Gene_expression	12	107	breast	Series matix file (TXT)
2015	GSE65194	Gene_expression	11	178	breast	Series matix file (TXT)
2013	GSE42568	Gene_expression	17	121	breast	Series matix file (TXT)

Table B.3.4: Description of the selected microarray datasets of breast samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2015	GSE66555	Methylation	6	65	colon	Series matix file (TXT)
2013	GSE32146	Methylation	10	25	colon	Series matix file (TXT)
2012	GSE42752	Methylation	19	63	colon	Series matix file (TXT)
2018	GSE92415	Gene_expression	6	183	colon	Series matix file (TXT)
2012	GSE38713	Gene_expression	13	43	colon	Series matix file (TXT)
2010	GSE20916	Gene_expression	16	145	colon	Series matix file (TXT)

Table B.3.5: Description of the selected microarray datasets of colon samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2016	GSE85464	Methylation	19	38	gastric	Series matix file (TXT)
2017	GSE99553	Methylation	32	84	gastric	Series matix file (TXT)
2013	GSE34387	Methylation	7	220	gastric	Series matix file (TXT)
2008	GSE13911	Gene_expression	31	69	gastric	Series matix file (TXT)
2010	GSE19826	Gene_expression	15	27	gastric	Series matix file (TXT)
2013	GSE44740	Gene_expression	12	26	gastric	Series matix file (TXT)

Table B.3.6: Description of the selected microarray datasets of gastric samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2014	GSE59157	Methylation	36	95	kidney	Series matix file (TXT)
2016	GSE70303	Methylation	12	24	kidney	Series matix file (TXT)
2010	GSE22459	Gene_expression	25	65	kidney	Series matix file (TXT)
2014	GSE53757	Gene_expression	23	144	kidney	Series matix file (TXT)

Table B.3.7: Description of the selected microarray datasets of kidney samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2019	GSE113017	Methylation	30	60	liver	Series matix file (TXT)
2019	GSE113019	Methylation	18	55	liver	Series matix file (TXT)
2016	GSE73832	Methylation	7	133	liver	Series matix file (TXT)
2019	GSE112790	Gene_expression	15	198	liver	Series matix file (TXT)
2018	GSE102079	Gene_expression	33	257	liver	Series matix file (TXT)
2014	GSE55092	Gene_expression	7	140	liver	Series matix file (TXT)

Table B.3.8: Description of the selected microarray datasets of liver samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2017	GSE75008	Methylation	28	80	lung	Series matix file (TXT)
2014	GSE52401	Methylation	100	244	lung	Series matix file (TXT)
2015	GSE68825	Methylation	5	144	lung	Series matix file (TXT)
2020	GSE51024	Gene_expression	41	96	lung	Series matix file (TXT)
2014	GSE33532	Gene_expression	20	100	lung	Series matix file (TXT)
2013	GSE40791	Gene_expression	72	194	lung	Series matix file (TXT)

Table B.3.9: Description of the selected microarray datasets of lung samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2013	GSE48472	Methylation	4	56	pancreas	Series matix file (TXT)
2015	GSE64491	Methylation	2	64	pancreas	Series matix file (TXT)
2013	GSE52578	Methylation	1	31	pancreas	Series matix file (TXT)
2017	GSE46234	Gene_expression	2	8	pancreas	Series matix file (TXT)
2011	GSE32688	Gene_expression	3	96	pancreas	Series matix file (TXT)
2013	GSE46385	Gene_expression	2	47	pancreas	Series matix file (TXT)

Table B.3.10: Description of the selected microarray datasets of pancreas samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2017	GSE52955	Methylation	5	83	prostate	Series matix file (TXT)
2013	GSE38240	Methylation	4	12	prostate	Series matix file (TXT)
2014	GSE47915	Methylation	4	8	prostate	Series matix file (TXT)
2014	GSE55945	Gene_expression	4	21	prostate	Series matix file (TXT)
2011	GSE32448	Gene_expression	4	80	prostate	Series matix file (TXT)
2011	GSE26910	Gene_expression	5	24	prostate	Series matix file (TXT)

Table B.3.11: Description of the selected microarray datasets of prostate samples

Year	GEO Accession ID	Source name	Analyzed samples	ll available samples	Cell type	Selected file type
2016	GSE73832	Methylation	4	133	small_intestine	Series matix file (TXT)
2015	GSE67485	Methylation	2	19	small_intestine	Series matix file (TXT)
2014	GSE50475	Methylation	3	45	small_intestine	Series matix file (TXT)
2015	GSE56525	Gene_expression	6	12	small_intestine	Series matix file (TXT)
2013	GSE18490	Gene_expression	1	360	small_intestine	Series matix file (TXT)
2013	GSE43346	Gene_expression	1	70	small_intestine	Series matix file (TXT)
2012	GSE33846	Gene_expression	1	32	small_intestine	Series matix file (TXT)

Table B.3.12: Description of the selected microarray datasets of small intestine samples

Appendix C

R Scripts

```
### Developmental CTCF ###
## Selection CTCF gene from methylation data for all developmental stage
colnames(age_range_methylation)[1] <- "cg_probe"
CTCF_age_range <- merge(reference_met, age_range, by = "cg_probe")
#reference_met contains all methylation probe on the CTCF gene
age_range_methylation <- CTCF_age_range[,6:ncol(CTCF_age_range)] #selection
numerical methylation data to make data ready to correlate. Other columns which
contains probe and other detailed information prepared as annotation file
CTCF_annotation <- age_range_methylation[,1:5]
```

```
## Selection CTCF gene from gene expression data for all cancer type
age_range_gene_expression <- age_range_gene_expression[,-1] #only the first column
which contains CTCF gene symbol was deleted to make data ready to correlate. Gene
expression data was already contains only CTCF gene.
```

```
## Correlations correlation_age_range <-
data.frame(diag(cor(t(age_range_methylation),t(age_range_gene_expression)))) #they
first returned to transpose to correlate probes rather than samples then only diagnol
samples were selected to avoid duplicate results.
!! cor function default method is Pearson, for Spearman correlation, following argument
was added: “method = “spearman”
```

```
colnames(correlation_age_range) <- "correlation_level"
CTCF_age_range <- cbind(CTCF_annotation, correlation_age_range) #probe
informations were added to the correlation results
```

```
# Scatter plot of CTCF gene for cancer type library(ggplot2)
ggplot(CTCF_age_range, aes(x=distance_to_TSS, y=correlation_level,
color=insideFeature, shape=insideFeature)) +   geom_point() +
  geom_smooth(method=lm, aes(fill=insideFeature)) +   ggtitle("CTCF_age_range")
```

```
# For 7 different developmental stages, this scripts were adjusted and run separately.
```

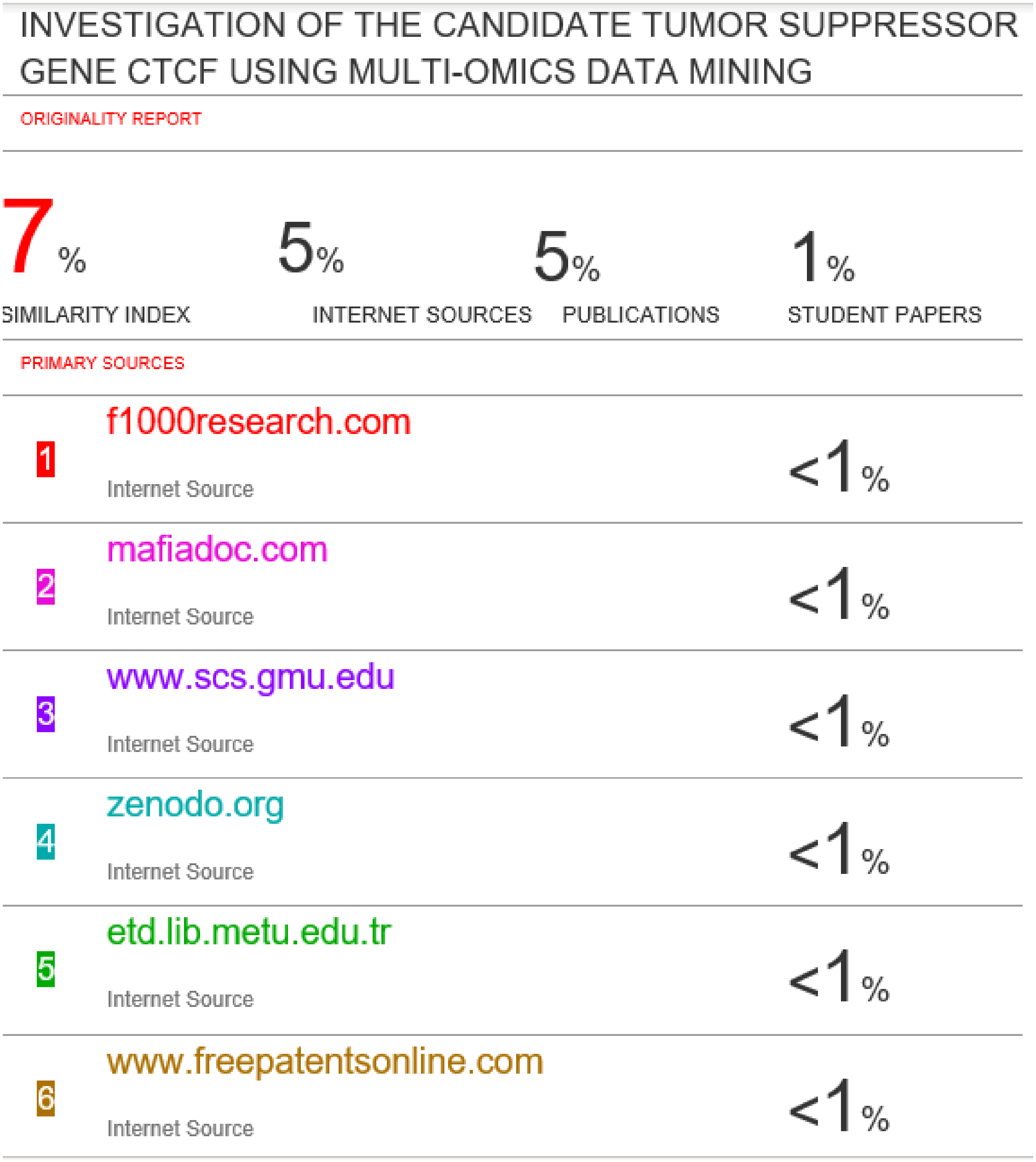
```
# Scatter plot of the all developmental stages
ggplot(developmental_CTCF, aes(x=distance_to_TSS, y=correlation_level,
color=cg_probe, shape=age_Range)) +
  geom_point() +   ggtitle("developmental_CTCF")+
scale_shape_manual(values=1:nlevels(developmental_CTCF$age_Range))
```

```
### PCA of the all developmental stages
```

```
rownames(pca_candidate_t) <- pca_candidate_t$age_Range
pca_candidate_t$age_Range <- NULL
PCA_raw <- prcomp(na.omit(pca_candidate_t), scale. = FALSE) percentVar <-
round(100*PCA_raw$sdev^2/sum(PCA_raw$sdev^2),1) sd_ratio <- sqrt(percentVar[2]
/ percentVar[1])
dataGG <- data.frame(PC1 = PCA_raw$x[,1], PC2 = PCA_raw$x[,2],
DevelopmentalStages = colnames(pca_candidate_t))

ggplot(dataGG, aes(PC1, PC2)) +
  geom_point(aes(colour = DevelopmentalStages)) +   ggtitle("PCA plot of the CTCF in
different developmental stages") +   xlab(paste0("PC1, VarExp: ", percentVar[1], "%"))
+   ylab(paste0("PC2, VarExp: ", percentVar[2], "%")) +   theme(plot.title =
element_text(hjust = 0.1))+   scale_color_brewer(palette = "Paired")

### heatmap annotation_for_heatmap <-
  data.frame(DevelopmentalStages = colnames(pca_candidate_t))
library(RColorBrewer) ann_colors <- list(group = brewer.pal(7, name = "Paired"))
names(ann_colors$group) <- unique(DevelopmentalStages)
rownames(annotation_for_heatmap) <- colnames(pca_candidate_t)
pheatmap((pca_candidate_t),
         annotation_col = annotation_for_heatmap,          annotation_colors = ann_colors,
scale = "column",          legend = TRUE,          cluster_cols = T,          cluster_rows = T,
show_rownames = T,          show_colnames = T,          clustering_distance_rows =
"manhattan",          clustering_method = "complete",          main = "", fontsize_col = 10)


### Tissue cancer related CTCF ###

## Selection CTCF gene from methylation data for all cancer type

colnames(cancer_type_methylation)[1] <- "cg_probe" CTCF_cancer_type <-

merge(reference_met, cancer_type_methylation, by = "cg_probe") #reference_met

contains all methylation probe on the CTCF gene cancer_type_methylation <-

CTCF_cancer_type[,6:ncol(CTCF_cancer_type)] #selection numerical methylation data

to make data ready to correlate. Other columns which contains probe and other detailed

information prepared as annotation file CTCF_annotation <-

cancer_type_methylation[,1:5]


## Selection CTCF gene from gene expression data for all cancer type
cancer_type_gene_expression <- cancer_type_gene_expression[,-1] #only the first
column which contains CTCF gene symbol was deleted to make data ready to correlate.
Gene expression data was already contains only CTCF gene.
```

```
## Correlation correlation_cancer_type <-
data.frame(diag(cor(t(cancer_type_methylation),t(cancer_type_gene_expression))))
#they first returned to transpose to correlate probes rather than samples then only
diagnol samples were selected to avoid duplicate results.

!! cor function default method is Pearson, for Spearman correlation, following argument
was added: “method = “spearman”
colnames(correlation_cancer_type) <- "correlation_level"

CTCF_cancer_type <- cbind(CTCF_annotation, correlation_cancer_type) #probe
informations were added to the correlation results


# Scatter plot of CTCF gene for cancer type library(ggplot2)

ggplot(CTCF_cancer_type, aes(x=distance_to_TSS, y=correlation_level,

color=insideFeature, shape=insideFeature)) +   geom_point() +

  geom_smooth(method=lm, aes(fill=insideFeature)) +   ggtitle("CTCF_cancer_type")


# For 12 different cancer type, this scripts were adjusted and run separately.
## Selection highly correlated probes by 0.4 cut-off threshold

cut-off threshold_all <- subset(cancer_tissue_related, correlation_level>=0.4 |
correlation_level<=-0.4) #cancer_tissue_related file contains all correlation results of
each cancer type in together.

cut-off threshold_all_positive <- subset(cut-off threshold_all, correlation_level>=0) #all
positive and negative correlation results were divided to examine all probes which
exceeding cut-off threshold according to cancer types

cut-off threshold_all_negative <- subset(cut-off threshold_all, correlation_level<=0)

number_table_p <- data.frame(table(cut-off threshold_all_positive$Tissue)) #to get

number of probes exceeding cut-off threshold table function used number_table_n <-

data.frame(table(cut-off threshold_all_negative$Tissue)) colnames(number_table_p) <-

c("Tissue", ">0.4_cut-off threshold") colnames(number_table_n)<- c("Tissue",

"<0.4_cut-off threshold")

final_table <- merge(number_table_p,number_table_n, by= "Tissue") #as a final table
the number of all positively and negatively correlated probes passing cut-off threshold
have been shown

overall_table <- data.frame(table(cut-off threshold_all$Tissue)) colnames(overall_table)

<- c("Tissue", "overall")

final_table_up <- merge(final_table, overall_table, by= "Tissue")
```

```
ordered_final_table <- final_table_up[order(final_table_up[,4], decreasing = TRUE),
drop = FALSE,] #results are ordered from greater to smaller, according to which cancer
type has more probe passing the cut-off threshold.


## Scatter plot for all cancer type together

ggplot(cancer_tissue_related, aes(x=distance_to_TSS, y=correlation_level,
color=cg_probe, shape=Tissue)) +   geom_point() +

  ggtitle("cancer_related_CTCF")+

  scale_shape_manual(values=1:nlevels(cancer_tissue_related$Tissue))
## The difference in correlation values between healthy and cancerous states for all
cancer types was computed and combined.

difference_cut-off threshold <- subset(cancer_tissue_related, Difference>=0.5) #the
difference greater than 0.5 cut-off threshold were selected

number_difference_cut-off threshold <- data.frame(table(difference_cut-off
threshold$Tissue)) #the number of probes exceeding cut-off threshold were calculated
for all cancer types

colnames(number_difference_cut-off threshold) <- c("Tissue", ">0.5_cut-off
threshold_difference")

ordered_number_difference_cut-off threshold <- number_difference_cut-off
threshold[order(number_difference_cut-off threshold[,2], decreasing = TRUE), drop =
FALSE,]
```



INVESTIGATION OF THE CANDIDATE TUMOR SUPPRESSOR GENE CTCF USING MULTI-OMICS DATA MINING

ORIGINALITY REPORT

7%	5%	5%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	f1000research.com Internet Source	<1%
2	mafiadoc.com Internet Source	<1%
3	www.scs.gmu.edu Internet Source	<1%
4	zenodo.org Internet Source	<1%
5	etd.lib.metu.edu.tr Internet Source	<1%
6	www.freepatentsonline.com Internet Source	<1%