REPUBLIC OF TURKEY

ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY

INSTITUTE OF HEALTH SCIENCES

INTEGRATION of MULTI-OMICS DATA for PREDICTING INDIVIDUAL COLON CANCER AETHIOLOGY

BEGÜM ÖZEMEK

MASTER THESIS

DEPARTMENT of BIOSTATISTICS and HEALTH INFORMATICS

SUPERVISOR

Prof. Dr. Uğur Sezerman

ISTANBUL - 2018

Department: Biostatistics and Health Informatics

Program: Bioinformatics and Biostatistics

Thesis Title: Integration of Multi-omics Data for Predicting Individual Colon Cancer Aethiology

Student Name-Surname: Begüm Özemek

Date of dissertation: 11.09.2018

This thesis is approved as Master thesis by jury members.

Head of Jury	Prof. Dr. Uğur Sezerman	
	Acibadem Mehmet Ali Aydinlar University	
Supervisor	Prof. Dr. Uğur Sezerman	
	Acibadem Mehmet Ali Aydinlar University	
Jury Member	Assoc. Prof. Emel Timuçin	
	Acibadem Mehmet Ali Aydinlar University	
Jury Member	Prof. Dr. Semra Demokan	
	Istanbul University	

In accordance with the relevant provisions of the Acibadem University Mehmet Ali Aydinlar University Graduate Education and Examinations Regulations, this thesis is approved by the above jury members and Health Sciences Boards of Directors accepted by decision.

Prof. Dr. Uğur Özbek

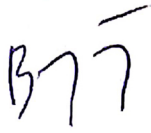
Head of the Inst

DECLERATION

I declare that this thesis study is my own work, I had no unethical behavior at any stage from planning to writing, I obtained all the information in this thesis following academic and ethical rules, I have shown all resources and shown these resources in the bibliography and there is no violation of patent and copyrights.

11.09.2018

Begüm Özemek



To my family…

ACKNOWLEDGEMENTS

First, I would like to thank my supervisor Prof. Dr. Uğur Sezerman, for his guidance that shapes my whole career, his endless patience when I feel broke and his support during the times when I lost my belief in myself. He has been a perfect role model for me, and it has been a great honor to be one of his students. I also would like to thank my jury members, Prof. Dr. Semra Demokan for her positive energy that fills me up whenever we meet and her support, and; Assoc. Prof. Dr. Emel Timuçin for her valuable and constructive comments. This project is financially supported by TUBITAK – 1001 114E623.

Dear Sezerman Lab members, especially, Ph.D. Ahmet Sinan Yavuz, for his guidance and supervision, without him this work would not be possible; Ph.D. Aslı Yenenler, for her power and support starting from my undergraduate years; Ph.D. Ceren Saygı, for her advices during hard and good times for both scientific and non-scientific topics; MD. Ege Ülgen for his unending wish for helping; Nogay Seymen for being a perfect friend and little-brother time to time; Okan Soykam for nice talks and jokes; Rüçhan Ekren for his kindness and friendship; Zeynep Özkeserli for her ability to feel what I feel and always being there whenever I need. Also, thanks to Ph.D. Melis Durası and Ph.D. Ozan Özışık and all those I could not mention. Additionally, I would like to thank my irreplaceable friends, Bensu Uçar, Serçin Çabuk and Basak Özçelik for all the things we shared, for always being beside me regardless of the kilometers in between. Also, thank you, Didem Koçhan, for all the laughs, support and sharing.

Lastly, the greatest thanks to my family; I know I will always owe them, will try to become better and better in whatever I do and be worthy of their support, their belief and their love. Dear Mom, Dad, I would not find the courage to shift my career from engineering to science if I was not sure you would support me. Lovely Canberk, being your sister and feeling that you are proud of me is the best feeling I had, including all the diplomas and titles, and probably will have. Finally, my dear husband, my greatest power-supplier, Alper, seeing pride and love in your eyes kept me motivated all the time. I feel fortunate to have such a great family, and I am sure that I will feel this way whole my life.

ABSTRACT

Integration of Multi-omics Data for Predicting Individual Colon Cancer Aethiology

Epigenetic changes are hereditary, dynamic and reversible changes that play a role in the formation of tumorigenesis. DNA methylation is one of the well-studied epigenetic mechanisms due to gene regulation, genomic imprinting and its importance on cell differentiation. miRNAs are small, non-coding, endogenous RNAs and play an important role as regulators in signal pathways in various cell mechanisms.

Cancer is affected by both genetic and epigenetic changes. Therefore, in order to understand the mechanisms leading to the formation of cancer, genetic and epigenetic changes should be studied together with a comprehensive approach. On the other hand, complex diseases follow different pathways for each patient, even if they show a similar phenotype. Personalized mechanisms should be identified in order to provide better health care and treatment. However, it is difficult to find the affected paths for a patient because there are many mechanisms that are affected by cancer.

In this study, we propose a holistic approach, along with the mentioned changes, to predict each change in gene expression changes, methylation amount-varying regions and miRNA transcription levels. In this study, DNA methylation, miRNA transcription and the omic data of gene expression were analyzed separately. Then, for each of these, pathways were analyzed, and pathways affected by separate mechanisms were found. Subsequently, candidate pathways were identified by combining these different data sources, and finally, pathways that were individually affected by DNA methylation levels or gene expression levels for colorectal adenocarcinoma patients were identified. This study presents the necessary bioinformatics background to be applied to all kinds of complex diseases, including different types of cancer or Alzheimer's disease.

Keywords:

Epigenetics, multi-omics, pathway analysis, personalized medicine

ÖZET

Kişisel Kolon Kanseri Etiyolojisini Tahmin Etmede Çoklu-Omik Verilerin Entegrasyonu

Epigenetik değişiklikler, tümörijenez oluşumunda rol oynayan kalıtsal, dinamik ve geri dönüşümlü değişimlerdir. DNA metilasyonu, gen regülasyonu, genomik imprinting ve hücre farklılaşması üzerindeki önemi nedeniyle iyi çalışılmış epigenetik mekanizmalardan biridir. miRNA'lar küçük, kodlayıcı olmayan, endojen RNA'lardır ve çeşitli hücre mekanizmalarında, esas olarak sinyal yollarındaki düzenleyiciler olarak önemli rol oynarlar.

Kanser hem genetik hem de epigenetik değişikliklerden etkilenir. Bu nedenle, kanser oluşumuna yol açan mekanizmaları anlamak için, genetik ve epigenetik değişimleri birlikte kapsamlı bir yaklaşımla birlikte incelemek gerekir. Öte yandan, karmaşık hastalıklar, benzer bir fenotip gösterseler de her hasta için farklı yollar izler. Daha iyi bir sağlık bakımı ve tedavi sağlamak için kişiselleştirilmiş mekanizmalar belirlenmelidir. Bununla birlikte, bir hasta için etkilenen yolları bulmak zordur çünkü kanserden etkilenen birçok mekanizma vardır.

Bu çalışmada, gen ekspresyonu değişiklikleri, metilasyon miktarı değişen bölgeler ve miRNA transkripsiyon düzeylerinde, etkilenen her bir yolu öngörmek için bu değinilen değişiklikler ile birlikte, bütüncül bir yaklaşım önermekteyiz. Bu tez çalışmasında, DNA metilasyonu, miRNA transkripsiyonu ve gen ekspresyonunun omik verileri ayrı ayrı analiz edilmiştir. Daha sonra bunların her biri için, yolak analizi uygulanmış ve ayrı ayrı mekanizmalardan etkilenmiş yolaklar bulunmuştur. Ardından, bu farklı veri kaynaklarının birleştirilmesiyle aday yolaklar belirlenmiş ve son olarak, kolorektal adenokarsinom hastaları için ya DNA metilasyon seviyelerinden ya da gen ekspresyon seviyelerinden tek tek etkilenen yollar tespit edilmiştir. Bu çalışma, farklı kanser türleri veya Alzheimer hastalığı dahil olmak üzere, her tür karmaşık hastalığa uygulanacak gerekli biyoinformatik arka planı sunmaktadır.

Anahtar kelimeler:

Epigenetik, çoklu-omik veri, kişisel tıp, yolak analizi

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVATIONS

$-CH_3$	Methyl Group
ASRi	Age-Standardized Incidence Rate
BCV	Biological Variation
BMIQ	Beta-Mixture Quantile Normalization
CGI	CpG Island
CR	Cox-Reid
CRAN	The Comprehansive R Archive Network
CRC	Colorectal Adenocarcinoma
CTC	Computed Tomography Colonography
DEGs	Differentially Expressed Genes
DE-miRNAs	Differentially Expressed miRNAs
DMP	Differentially Methylated Probes
DMR	Differentially Methylated Regions
DNA	Deoxiribonucleicacid
DNMT	DNA Methyltransferase
EWAS	Epigenome-Wide Association Study
FIT	Fecal Immunechemical Test
GEO	Gene Expression Omnibus
gFOBT	Guaiac-Based Fecal Occult Blood
GWAS	Genome-Wide Association Study

HCC	Hepatocellular carcinoma
HGP	Human genome project
logFC	log_2Fold Change
KEGG	Kyoto Encylopedia of Genes and Genomes
MBD	Methyl-Binding Domain
MID	Middle domain
miRISC	microRNA Induced Silencing Complex
miRNA	micro-RNA
mRNA	Messenger RNA
NGS	Next-Generation Sequencing
ORF	Open Reading Frames
PAZ	Piwi-Argonuate-Zwille Domain
PCR	Polymerase Chain Reaction
PIN	Protein Interaction Network
pre-miRNA	Preprocessed micro-RNA
piRNA	Piwi-Interacting RNA
RISC	RNA Induced Silencing Complex
RNA	Ribonucleicacid
SAM	S-Adenosyl Methionine
siRNA	Small Interfering RNA
SLIM	Sliding Linear Model

SNP	Single Nucleotide Polymorphism
SVD	Singular Value Decomposition
TCGA	The Cancer Genome Atlas
TMM	Trimmed Mean of M-Values
TNF	Tumor Necrosis Factor
TRBP	TAR-RNA binding protein
TUSEB	Health Institutes of Turkey
UTR	Untranslated Region
WGBS	Whole-Genome Bisulfite Sequencing

1 INTRODUCTION

Genetics and epigenetics are the two main categories that explain how information in the cells flows. The term “genetic” stands for the information there is encoded in DNA, and responsible for synthesis of cellular building blocks, such as proteins, and regulators, such as non-coding small RNAs. Apart from genetic mechanisms, the term “epigenetic” involves the additional guidance on how, when and where the genetic information should be used. These two fundamental mechanisms regulate each other on and on for the lifetime of the organisms.

Epigenetics, as it’s the first definition by C. H. Waddington, is the “the branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being.” [1]. Epigenetic regulation is responsible for changes in gene expression without causing any change in DNA sequence. Epigenetic control mechanisms play a key role in various aspects of living systems, such as X-chromosome inactivation, genomic imprinting, differentiation, transcriptional regulation and disease aethiology as in the case of cancer, which is the focus of this thesis [2]. The most studied and known epigenetic mechanisms involve DNA methylation, chromatin organization and histone modifications [3]. For this thesis, DNA methylation, its mechanism and its role in cancer aethiology was explained in Section 2.1.

Complex or multifactorial diseases are affected by genetic and environmental factors, which includes the lifestyle, as well. It is known that, complex diseases do not follow traditional Mendelian inheritance pattern but instead, genetic factor only represent the “risk” of the disease phenotype [4]. For a disease to be classified as “complex” there can be three possible ways it influences the phenotype: Firstly, if hereditary affects are clear but also phenotype depends on environmental effects, as well, as in the case of diabetes. Secondly, if genetic inheritance has a risk for obtaining the disease however the phenotype is not only decided by the heritage, as in the case of cancer. Finally, if the disorder aggregates in families without being consistent with

Mendelian Inheritance but, in the most cases, consistent with polygenic inheritance [5].

Due to complexity of each type of multifactorial diseases mentioned above, studying with individual aethiologies of multifactorial diseases is a challenging area. Genome-wide association studies (GWAS) aims to reveal unyielding mechanisms behind the complex diseases by identifying common variants that may increase the risk of the disease [6,7]. Over the years passed after the first successful clinical application of GWAS in 2005, GWA approach enlightened lots of complex diseases' causative genes [7–10]. However, GWAS was not sufficient to resolve all complex diseases' aethiology. Therefore, to reveal epigenetic variants that affect disease formation epigenome-wide association studies (EWAS) took to the stage. With the knowledge of epigenetic variations' effect on genome function, it is trivial to guess that these variations could affect the disease aethiology, as well. EWAS provided scientists a better understanding of complex diseases and their aethiology [11].

Even though the great impact GWAS and EWAS made in researches focusing in complex diseases, there are still some challenges left to overcome with EWAS design and execution. For instance, limited sample size or wrong tissue/blood selection with the specific phenotype could result in change in differences observed [12]. Moreover, both GWAS and EWAS aim to reveal common variations in genomic and epigenetic levels. However, these approaches are not sufficient for investigation of personal drivers of the disease, which may lead to a better understanding of the disease mechanism.

Personalized medicine (or *precision medicine*) aims to provide optimized diagnostic and treatment strategies for each individual patient by analyzing and modeling their specific characteristics [13]. Cancer is one of the most studied disease group in terms of precision medicine approaches. By analyzing an individual's tumor, determining the therapy or combination of therapies will work best and this approach would decrease the costs of the therapy when increasing the chance of cure [14]. Ever since Human Genome Project (HGP) was completed in 2001 and next-generation sequencing (NGS) techniques were developed and became accessible, lots of application of personalized medicine were seen in cancer treatment. Several large-

scale genome sequencing projects are being conducted across the globe, such as 1000 Genome Project at the UK and Turkish Genome Project led by the Health Institutes of Turkey (TÜSEB) One of the key aims of these projects is to develop new methods and technologies that enables studying of personalized mechanisms of disease development and personalized therapy target which are instrumental parts of personalized medicine.

One of the first applications of personalized medicine is targeted therapy which started in 1970s with tamoxifen which binds to estrogen receptor (ER) and preventing transcription factors from binding to the receptor, hence regulating its activity to simulate cancer cell growth in ER+ breast cancer patients [15]. This approach improved over time and was followed by various drugs designed for different target genes, such as *HER2, VEGF, mTOR* and even for targeting micro-environment as well [16]. Several studies showed that targeted therapies are resulting better compared to non-targeted therapies, as in the case of dabrafenib versus dacarbazine comparison. Dabrafenib targets MAP kinase signaling pathway and has a median progression-free survival rate as 5.1, whereas it is 2.7 months for non-targeted dacarbazine [17]. Hence, currently, many institutions and hospitals perform DNA sequencing for cancer patients to determine the optimal treatment option.

Unlike mutations and changes in the sequence, epigenetic changes are reversible. Therefore, they are promising ways for cancer treatments. It has been shown that, even simple methods such as green tea usage and physical activity have a role in inversing DNA methylation patterns in cancer and can influence the diseases formation [18,19]. On the other hand, targeting DNA methylation enzymes could be another option for treatment of cancer and there are many candidate agents are in trial phase [20].

Colorectal cancer is one of the most common cancers in the world-wide, second-most in women and the third-most in men. Colorectal cancer is the forth-most deadly cancer type across all cancer-related deaths by causing 771,000 deaths in 2013 [21]. Epigenetic regulators, such as DNA methylation and miRNAs play a key role in colorectal cancer formation and progression, additionally to the genetic regulations. For example, the genes whose promoter regions are hypermethylated, are

epigenetically silenced and they may take part several processes in colorectal cancer tissues, such as cell cycle regulation, DNA repair, apoptosis, and so on [22,23].

In this thesis, I aimed to identify personal and regional pathways that are regulated by DNA methylation and have a significant role in formation and progression of colorectal adenocarcinoma. For the future work, the pathways suggested, can be promising targets for epigenetic therapies. As a side-work, I also aim to provide candidate epigenetic biomarkers for the diagnosis, as well.

The dissertation is organized as 6 sections: In Section 1 (this section) a general introduction was provided, in Section 2 necessary biological background on DNA methylation, colorectal adenocarcinoma as well as theorical background on integrative analysis of multi-omics data with the support from previous works were explained. Section 3 has descriptions of method that were used in the predictive analysis. In Section 4 all the results obtained from the analysis were demonstrated. Section 5 has the discussions about the implications of the results in the context of the dissertation aims and finally, in Section 6 a general conclusion with directions for future work was provided.

2 BACKGROUND AND RELATED WORK

2.1 DNA Methylation

DNA methylation is an epigenetic process in which a methyl (-CH_3) group binds into the C5 position of the cytosine and forms 5-methylcytosine, in mammals. This process takes place under the catalytic activity of DNA methyltransferase (DNMTs) family. DNMTs transfer a methyl group from S-adenosyl methionine (SAM) to the fifth carbon of a cytosine residue and causes the formation of 5mC (Figure 3.1). There are three DNMTs which regulates DNA methylation patterns in different ways.

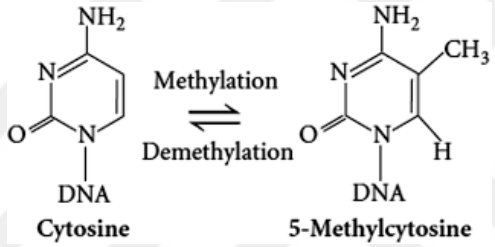


Figure 2.1: Methylation / demethylation of a cytosine residue (retrieved from [24])

DNMT1 plays a key role during DNA replication and copies the DNA methylation pattern from the parental DNA strand onto the newly synthesized strand (Figure 3.2a). However, DNMT3A and DNMT3B can establish a new methylation pattern to unmodified DNA and that is why they are also known as *de novo* DNMTs. All three DNMTs function extensively during development of an embryo and as the cells reach their terminal differentiation, DNMT expression is reduced [25,26] (Figure 3.2b).

There are two main categories for demethylation process, namely active and passive. These two categories differentiated by their need for DNA replication: In passive demethylation, DNMT1 losses its activity which was regulated by PCNA and Uhr1. Whereas in the case of active demethylation, 5mC is removed by “reverse” enzymatic reaction in the absence of SAM. TET family is known regulator of active

demethylation process by catalyzing the oxidation of 5-methylcytosine to 5-hydroxymethycytosine (Figure 3.1) [27].

CpGs or CG sites are DNA regions where a cytosine is followed by a guanine in the linear sequence, in 5' to 3' direction. In human genome, the percentage of CpG dinucleotides are very low, 1% of the genome is made up of CpG dinucleotides whereas the expected ratio is 21%). Conversely with the representation rate, 60% to 90% of CpG cytosines are methylated in the genome [28,29]. According to the usual definition, if a region with at least 200 base-pairs GC percentage greater than 50% and an observed-to-expected CpG ratio greater than 60%, it is called as CpG islands or CG islands (CGI) [30].

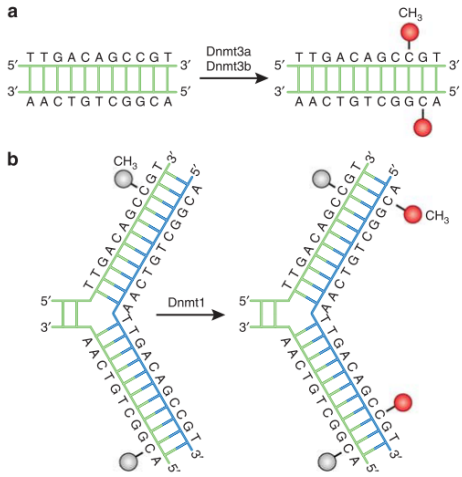


Figure 2.2: a. Role of DNMT3A and DNMT3B in DNA methylation (introducing new DNA methylation patterns). **b**. Role of DNMT1 in DNA Methylation (copying parental DNA methylation patterns) (Retrieved from [25]).

In addition to CpG islands, CpG sites in ±2000 base pairs away from CpG islands are named as "CpG island shores", ± 2000 base pairs – 4000 base pairs away CpG islands are called as "CpG island shelves", and lastly other CpG sites are as called as "open sea". Methylation of these different regions have different functions. CpG islands are commonly found in the promoter regions of the majority of annotated genes and methylation of CpG islands are observed in various diseases, as in the case of

cancer [19,30,31]. However, methylation in CpG island shores is associated with tissue-specific gene expression and its regulation [32].

DNA methylation is also observed in gene bodies and it is suggested that it is associated with a higher level of gene expression in dividing cells. By the term “gene body”, refers to the region of the gene past the first exon because methylation of the first exon, similarly to the promoter methylation, causes to the silencing of the corresponding gene [33]. The relationship between DNA methylation and transcriptional regulation might not be so straightforward but depends to a complex mechanism, which is not well-known for now.

Normally, DNA methylation patterns are established during development and maintained for the lifetime. Nevertheless, DNA methylation status of a CpG or a whole CpG island can be changed as a result of various factors, including environmental exposures, lifestyle, age and disease [34]. These differences in DNA methylation patters is called as “differential methylation” and divided into two categories: *hypermethylation*, when a CpG or CpG region becomes methylation and *hypomethylation*, vice-versa. Hypermethylation and hypomethylation have distinct roles in during embryonic development, chromatin structure, genomic imprinting, chromosome stability, cell differentiation and disease [35].

Few of these major roles summarized below in the scope of the dissertation.

2.1.1 Major Roles of DNA Methylation

DNA methylation is known by its role in embryonic development, genomic imprinting, X-chromosome inactivation and alterations in DNA methylation is associated with many diseases, including cancer.

Even though the direct mechanism is not discovered yet, DNA methylation on the CpG-dense promoter sites of the genes is associated with down-regulation of the gene expression [36]. Low-level methylation or even no methylation is observed on the highly transcribed genes’ CpG-dense promoter sites whereas genes with highly methylated promoters are lowly expressed. Therefore, generally, low-level

methylation is associated with high expression, on the other hand, high-level methylation is linked with low expression. The alteration of DNA methylation in the promoter site of a gene regulates the gene's expression over and over (Figure 3.3) [36].

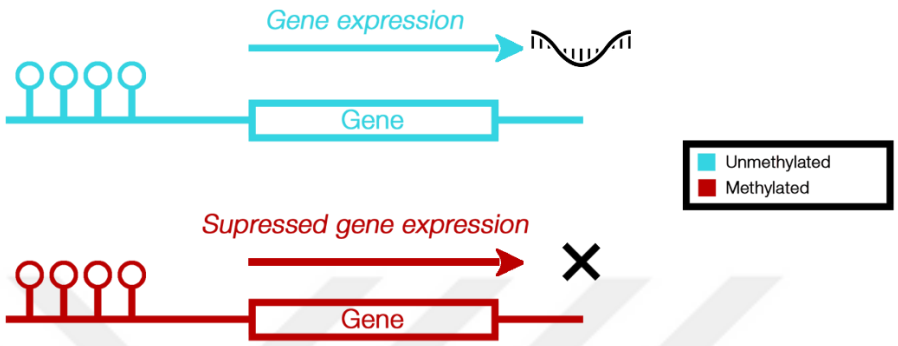


Figure 2.3: Representation of role of DNA methylation on the promoter region of a gene

DNA methylation on the promotor site can affect gene expression level by two ways: Firstly, binding methyl group can physically prevent binding of transcriptional proteins to the gene and secondly, a protein family with methyl-CpG-binding domain (MBDs). MBD proteins initially themselves bind to DNA and then recruit additional proteins to the region, such as chromatin remodeling proteins, and hence forms in a compact, inactive chromatin, called as *heterochromatin* [37]. Loss of some MBDs are linked with diseases, as in the case of Methyl-CpG-binding domain protein with Rett [38].

Gene expression regulation by DNA methylation is one of the key mechanisms that controls cell differentiation and reprogramming. In the process of cell differentiation, tissue-specific genes are demethylated for expression and developmental genes are methylated for silencing. This process was explained by C. H. Waddington in 1957 with an "*epigenetic landscape*" metaphor (Figure 3.4). In this landscape, a not-differentiated cell starts (shown as the marble at the top of the hill) its

life with full of developmental potential and goes down the differentiation process, all potential paths (represented as whole landscape) competes to end up as a different cell or tissue type. However, changing the way during the journey is possible for the marble, and marks gained during the cell development process, can be reversed to regain pluripotency [39].

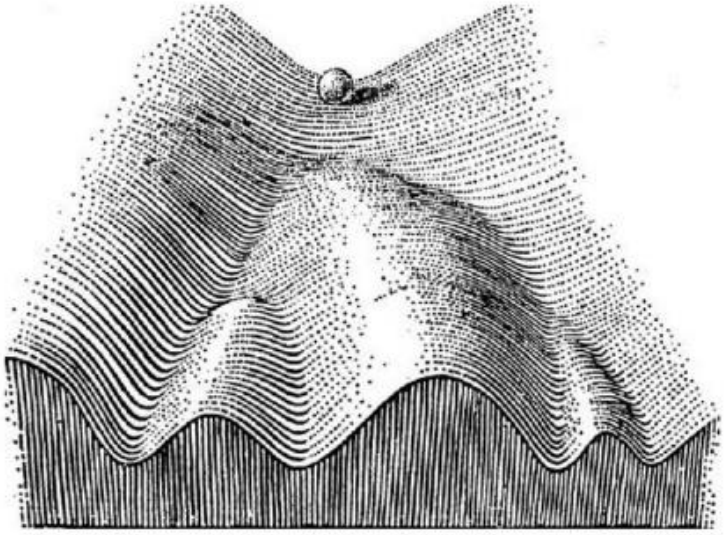


Figure 2.4: C. H. Waddington's "epigenetic landscape" which symbolizes a not-differentiated cell (the marble at the top) and possible paths (the landscape itself) that it can goes through to end up as a different cell or tissue type.

Genomic imprinting and dosage compensation are another important function of DNA methylation. Normally, both parental and maternal genes are expressed in the daughter cells. However, in the case of imprinting, only maternal or paternal genes are transcribed. The methylation status of the DNA at an imprinted locus shows a pattern which is unique for each allele. This unique methylation pattern causes the divergence of expression patterns from expected Mendelian inheritance. Not only in genomic imprinting but also in X chromosome inactivation, DNA methylation plays a key role, as well. During embryogenesis, one of two X chromosomes of a female is silenced, randomly, causing of expression of X linked genes as equal in both sexes. Silenced X

chromosome then becomes much dense and compact. This dense formation is named as "*Barr body*" [40]. Loss of imprinting is linked with approximately 30 known diseases or disorders, including Prader-Willi and the Angelman syndrome [41].

2.1.2 DNA Methylation in Cancer

Alterations in DNA methylation has effect in progression of various cancers, including gastric, colorectal, hepatocellular and breast cancers [42]. Historically, DNA *hypomethylation* in human colon and lung cancer was reported in 1983 and it was the first time the relationship between DNA methylation level changes is linked with cancer progression [43]. Even though the first studies focus on hypo-methylation in cancer, recent studies revealed that hyper-methylation is also a key factor in cancer initiation and progression [44].

DNA methylation regulates gene expression by providing a stable gene silencing mechanism, in healthy tissues. It is responsible from the cell differentiation and X-chromosome silencing during developmental process, as well. However, in cancer tissues, this mechanism loses its stability and either silences tumor-suppressor genes or activates oncogenes [35].

Together with the other epigenetic changes, such as chromatin modifications, DNA methylation takes part in all six hallmarks of cancer, declared by Hanahan and Weinberg [45] (Figure 2.5). These hallmarks explain the features that cancer cells gain during the formation of tumorigenesis. In other words, they are the characteristics which differentiate cancer cells than the normal cells. In the epigenetically updated version of "Hallmarks of cancer" shown in Figure 2.6, the possible epigenetic mechanisms are mentioned which may cause of acquisition these hallmarks.

One of the key properties that cancer cells have is the capability of sustaining proliferation. Normally, mitogenic growth signals are must for a cell to change its current growth state from an inactive state to a proliferative status. These signals are received via transmembrane receptors and in the literature, there are studies show that, many of the oncogenes are behaving as growth factor signals. Normal cells depend on

external signals for proliferation, however in the case of tumor cells, they can produce their own growth signals so that they do not depend transmissible signals. Furthermore, most of growth factors are responsive, which results in producing an endless positive feedback in cancer cells [46,47]. The studies done in the past ten years showed the importance of epigenetic mechanisms, miRNAs, long non-coding RNAs and DNA methylation, in controlling growth factor signals [48–50].

For a cell to become cancerous, growth suppressor mechanisms must be suppressed. In the literature, there are many studies show that, mutations on tumor suppressor genes are frequently observed in cancer patients, for instance *TP53* (Tumor Protein 53, encodes a tumor suppressor protein) is found as mutated in 45% of cancer patients [51]. However, there are various recent studies revealed that, evading growth suppressor mechanisms can be achieved by epigenetic mechanisms, as well. As an example, a study showed that growth suppressor genes such as *SMPD3* are silenced by DNA methylation in HCC patients [52].

Cancerous cells increase their numbers not only by growing and diving fast and uncontrollably but also by resisting to apoptosis. In 1992, up-regulation of *BCL-2* oncogene's activity on escaping apoptosis was discovered, and apoptosis-resistance was linked with cancer formation [53]. Recent studies showed epigenetic regulators, such as histone modifications by acetylation or methylation, take place during regulation of apoptosis process [54].

With the support of previously mentioned abilities of cancer cells (resistance to apoptosis, self-control on growth signal, capability of inactivating tumor-suppressor mechanisms), they also gain another related ability: unlimited replicative potential. In normal tissues, after reaching a particular number of cells, the process of dividing stops and tissue stays in a steady state. However, in the case of cancer, cells can replicate themselves without any boundaries [55]. Studies done in the past five years revealed that epigenetic mechanisms are associated with changes in cell replication timing and frequency. Acetylation, DNA methylation, histone modifications are some of these reported epigenetic regulators [55].

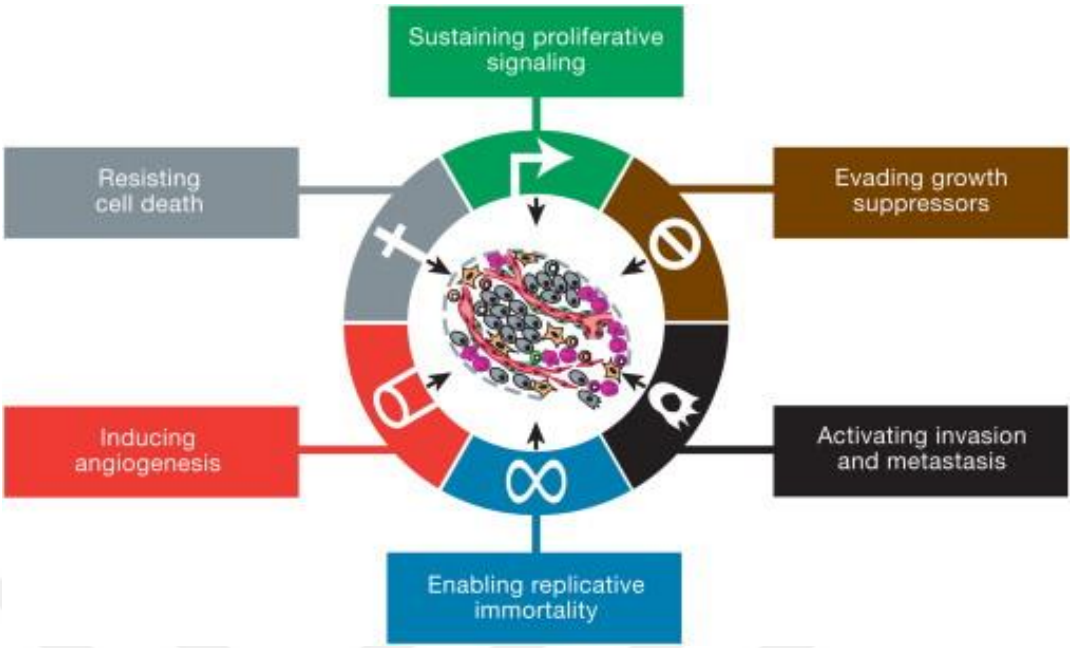


Figure 2.5: Hallmarks of cancer. Retrieved from [56].

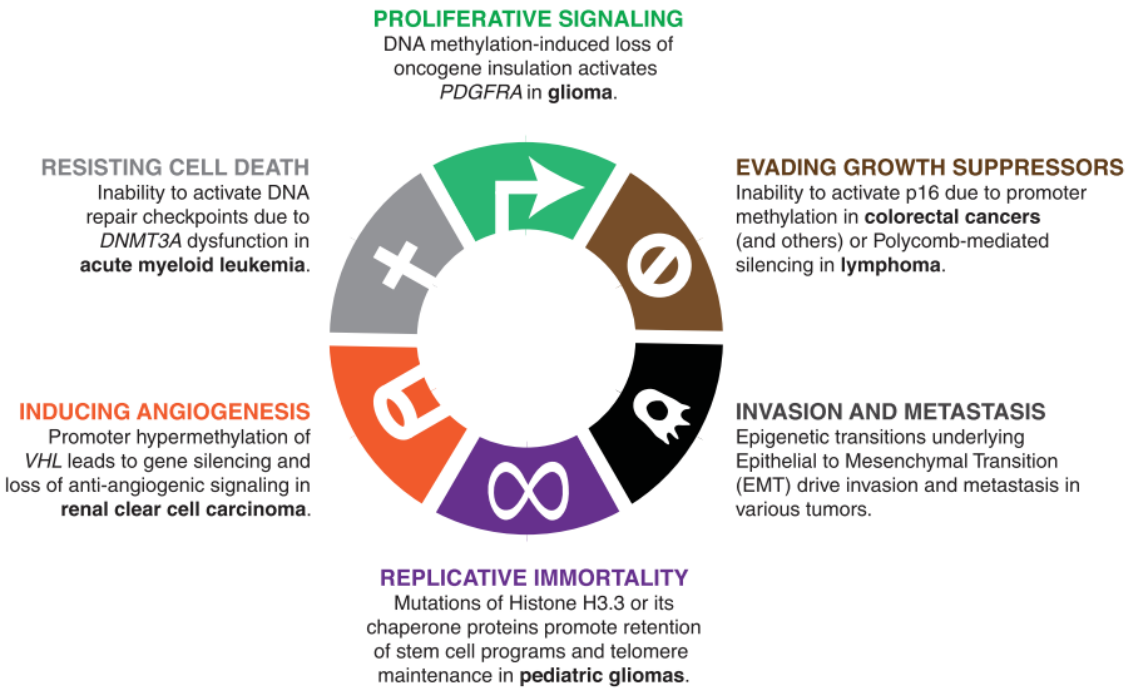


Figure 2.6: Hallmarks of cancer with their possible epigenetic causes. Retrieved from [51].

All cells need to oxygen and nutrients in order to keep living and functioning and these requirements are provided by capillary blood vessels which are nearby the tissue. Proliferative cells do not have the ability of encouraging growth of capillary vessels in their nature, however they gain this ability during tumorigenesis. Experimental studies showed that angiogenic regulation is achieved via switching angiogenesis activators and countervailing inhibitors. By way of change on the responsible genes' expression levels, the switching mechanism works. According to the a recognizable study, *VEGF* and *FGF*s are the two of these regulators and they are up-regulated in the tumor cells [57]. Almost 20 years after this study, various researches were focused on epigenetic regulators, mainly DNA methylation, that affect *VEGF* activity and epigenetic mechanisms were suggested as possible targets for targeted therapy [58–60].

Most of the types of cancer, a sub-clone of the primary tumor migrates through either adjacent tissues or even more wide tissues and settled, if they can find an appropriate environment. Sooner or later, the primary tumor becomes a mass and therefore space and nutrients become limiting for the growth of cancer. The ability of metastasis gives the tumor cells a chance to escape these limiting factors and grow up in a more suitable environment, at least for some time [56]. Metastases is a complex mechanism and the mechanism behind it still not understood completely. However, several mechanisms are reported which includes the change on links between the cells and their microenvironment and enabling of extracellular proteases. During the past five years, several studies focused on epigenetic regulators of metastases mechanism. There are many searches show regulatory role of epigenetic mechanisms, not only DNA methylation but also regulation via miRNAs as well, on metastasis mechanisms in several cancer types including skin cancer or colorectal carcinoma [61–63].

The "*hallmarks of cancer*" proposed in 2000 is explained above and their relationship with epigenetic regulators, particularly with DNA methylation, were detailed. Even though in 2011 four new hallmarks are added to these well-known six hallmarks, they cannot be thought off independent from their epigenetic regulators [45]. With the knowledge of epigenetic mechanisms' role in all these hallmarks, they can also explain observing cancer in patients with no family-history.

2.2 MicroRNAs (miRNAs)

Small RNAs are short (approximately from 18 to 30 nucleotides), non-coding RNA molecules. They are responsible of post-transcriptional regulation of gene expression and RNA silencing. Small RNAs are broadly defined in three main classes as microRNAs (miRNAs), siRNAs and Piwi-interacting RNAs (piRNAs). Small-RNAs have linked with various significant biological functions [64]. In the scope of this dissertation, miRNAs and their functions are summarized in this Section.

microRNAs are approximately 22 nucleotide length non-coding RNAs which take place in gene regulation at post-transcriptional level. miRNA genes are highly conserved parts of the genome. As an example, majority of known human miRNAs have their homologs in zebrafish [65].

Historically, the first miRNA was discovered in the year 1993 in *C. elegans* [66]. However, for the following, almost, 10 years miRNAs were not identified as a distinct class of regulatory mechanisms. After the 2000s, studies focusing on miRNAs and their roles in the cell rapidly increased [67,68]. During the time passed from 2000s until know, the knowledge about miRNAs were dramatically increased. It is known that; distinct sets of miRNAs are transcribed in the different cell-types and tissues. Additionally, it is shown that, miRNAs take part in various biological processes including the developmental processes, as well [69–72].

According to miRbase, the genome has over 1000 miRNAs [73]. Some of those miRNAs are found as clusters that are co-expressed and some of them are found individually. They are generally located in the intergenic regions of the genome; however, some miRNAs are found in intronic regions, oriented as reverse or sense. That location regularization show that, expression of miRNAs can be triggered by their own promoters or they can be expressed as a transcript product of other genes [65].

2.2.1 miRNA Biogenesis

miRNA biogenesis can be examined in two categories: canonical and non-canonical. These two categories will be explained separately.

In the case of canonical miRNA biogenesis, the miRNAs are found in the genome as either separately or as groups containing a small number of to several hundred different miRNAs. Additionally, many miRNAs are located in introns of protein-coding genes, and these genes are called as “host gene”. Primary miRNA transcripts (pri-miRNAs) are produced by RNA Polymerase II (Pol II) or RNA Polymerase III (Pol III). Pri-miRNAs are longer than mature, functional miRNAs since they contain local loop structures, caps, near to the 5’ and 3’ ends and polyadenylated [74]. Firstly, pri-miRNAs are converted into pre-miRNAs by microprocessors, a nuclear protein complex. Then, the pre-miRNA is transferred to the cytoplasm by Exp5, the export receptor exportin 5. Following the transfer of pre-miRNA to cytoplasm, a dsRNA which is 20-25 nucleotides long from the stem of the pre-miRNA is subjected to the process regulated by Dicer. This part of the preprocess during miRNA maturation is regulated by different enzymes in different species. Eventually, the miRNA is gathered with a set of proteins to form the RNA-induced silencing complex (RISC). RISC searches for its target mRNAs by seeking for complementary nucleotide sequences. Argonaute (AGO) protein, which takes part in RISC complex, plays the key role for the search of mRNA by representing 5’ of the miRNA to create the optimal position for it to bind with its target mRNA molecule [75,76]. A detailed explanation of miRNA biogenesis is shown in Figure 2.7.

2.2.2 Major Roles of miRNAs

The main function of miRNAs is gene regulation by silencing target genes. RISC is guided by its component miRNA to specifically recognize its target mRNA to suppress its expression level. Silencing is achieved by two main mechanisms: Translational repression and mRNA cleavage [74].

The mature miRNAs function with a complex which includes AGO proteins. The complex has a single polypeptide chain with four domains: the amino-terminal (N) domain, the Piwi-Argonaute-Zwille (PAZ) domain, the middle (MID) domain and P-element induced wimpy testes (PIWI) domain. L1 and L2, two linker domains, are responsible of connection between these four domains. L1 links N-terminal and PAZ

domains whereas L2 links the PAZ and MID domains. MID and PIWI domains binds 5’ of miRNA and PAZ domains holds its 3’ end. AGO1, AGO2, AGO3 and AGO4 are four AGO proteins which are found as encoded in human’s genome. However, only AGO2 protein has the ability of cleavage a target which has a high complementary to the guide strand of the miRNA. AGO2 protein is the most expressed one in humans, as well [77].

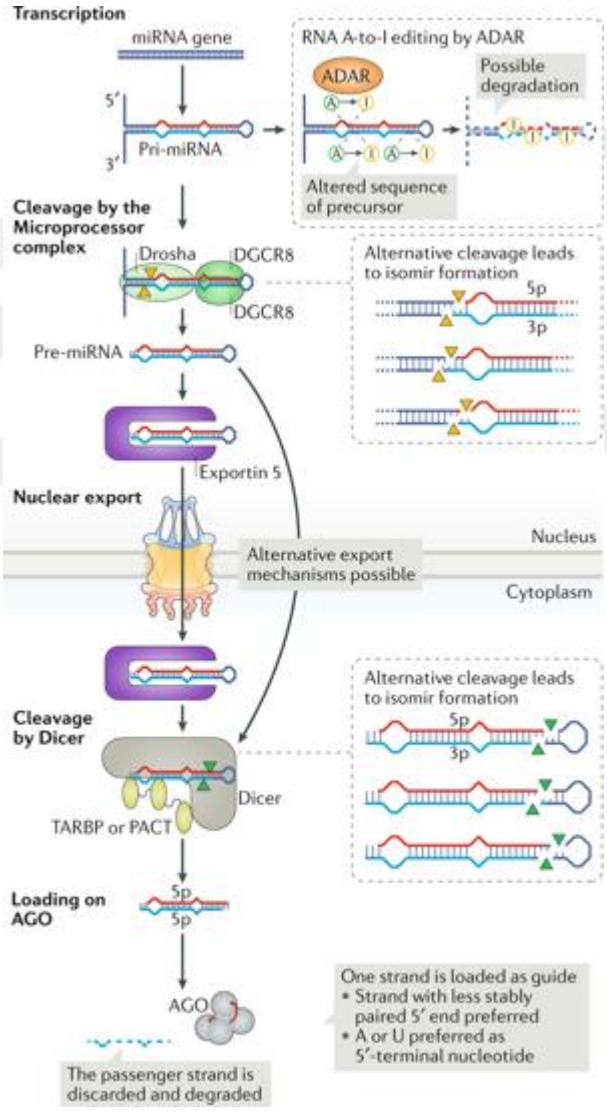


Figure 2.6: miRNA Biogenesis. Republished from the original publication [77].

.

miRNA target sites are mostly found in 3’ UTR of mRNAs. The affinity of miRNA-mRNA complex is the key factor on performance of translational repression mechanism. If miRNA and RNA have high complementarity, then silencing mechanism functions properly. However, if they have a mismatch in their binding sites, then degradation is omitted and repression mechanism does not work [78]. Considering this information, miRNA-mRNA binding behavior is similar to RNA-binding proteins as well.

Non-canonical binding is reported thanks to various large microarray-based studies and it is linked with AGO-target mRNA binding sites. The biological meaning of non-canonical binding is discussed since they do not provide a significant repression when meta-analyses are considered [77].

Even though most of the miRNA-based silencing mechanisms are reported in cytoplasm, there are some recent discoveries of nuclear mRNAs are targeted by miRISC complex, as well. RISC complex found as assembled or imported into nuclease, which explains of miRNA regulated silencing in nuclei [79].

2.2.3 miRNAs in Cancer

Tumorigenesis is a result of various regulatory mechanisms, both heritable (mutations and modifications) and inheritable, dynamic changes (epigenetic) in the cells. Non-coding RNAs, in the scope of this thesis, microRNAs are known as one of the main regulators of gene expression. Considering their roles in transcription, it is clear that they have also important effects on cancer progression and formation.

The miRNA processing is a complex mechanism in which various proteins are involved and regulated by several enzymatic reactions. Deficiencies in some of those components or mutations on responsible genes may result in tumorigenesis. Additionally, epigenetic changes, such as DNA methylation, is more frequent in miRNAs compared to mRNAs. Currently, there are 122 known miRNAs which are subjects of epigenetic alteration. 55 of those miRNAs are identified to specific to malignancies and 67 of them are thought to be cancer-related [80].

In 2002, the first time a miRNA complex is linked with a cancer, Chronic Lymphocytic Leukemia [81]. miRNAs are linked with tumorigenesis not only by expression changes on themselves but also mutations they carry and results of these genetic alterations [80]. For instance, a single nucleotide polymorphism on pri-miR-196a2 is reported as a leader mutation to chronic lymphoid leukemia, breast, gastric and lung solid tumors, with experiments *in vivo* [80].

Currently, we have knowledge about various changes occurring during miRNA biogenesis and leading to or play a role on cancer formation or progression. Examples of dysregulations caused by miRNAs can be grouped as, changes in nucleus and changes in cytoplasm. In nucleus, miR-15/16 loss related to 13q14.3 deletion in CLL is one of the earliest reported pre-translational changes. Also, some mutations on Drosha, such as R414, E993K and D1151, or DGCR8, such as E518K, A558T, L694S and Y721H, or Exportin 5, such as R1167, F1179 and K1181 are miRNA related mutations that takes part in tumorigenesis [82].

Mutations on either Dicer or TAR RNA binding protein (TRBP), transcriptional or translational changes, changes on the target mRNA and preferential processing are the other factors that are related with cancer occurring in cytoplasm. Examples of mutations on Dicer are S839, D1705, G1809, D1810 and E1813 whereas TRBP mutations include M145, P151, D221G, R296H and R353. Transcriptional changes on Dicer, such as being affected by KDM6A, KDM6B, miR-630, let-7, miR-103/107 or effect of EGFR-mediated phosphorylation on AGO2 protein also play role on oncogenesis. Mutations located in binding site (near to the 3' UTR) of mRNA also can be another player on tumorigenesis. Lastly, KSRP-mediated miRNA loading to RISC is a change occurring during preferential processing [82].

2.3 Colorectal Adenocarcinoma

Colorectal adenocarcinoma (CRC or colon cancer) is one of the most frequently observed and the second deadliest cancer types worldwide [83]. CRC is seen in colon or rectal parts of the digestive system and its formation starts as a polyp inside the colon or rectum which keeps on growing [84].

2.3.1 Epidemiology

Colorectal adenocarcinoma is the third-most common cancer in worldwide, which makes up 9.7% of all cancers (except non-melanoma skin cancer). It is observed more frequently in women (the second-most common cancer type) compared to men (the third most-common type of cancer) in worldwide. The age-standardized incidence rate (ASRi) of colon cancer is 20.6 per 100,000 people in men and 14,3 per 100,000 people in women. Colon cancer is mostly observed in individuals greater than 50 years old and 80% of colon cancer patients are diagnosed at >60 years old age.

Colorectal cancer is mostly observed in Australia and New Zeland (ASRi: 44.8 and 32.2 per 100,000 men and women, respectively) and it is rarely observed in Western Africa (ASRi: 4.5 and 3.8 per 100,000 men and women, respectively). In a general manner, colon cancer is mostly observed in more-developed regions, such as Europe, Northern America and Australia compared to less-developed regions as in the case of whole Africa. The ASRi variations between different geographical regions are linked with socioeconomic levels. The distribution of colon cancer in world-wide is shown in Figure 2.5 in detail with incidence and mortality rates [21].

2.3.2 Risk Factors

One of the main risk factors of CRC is age. CRC is mostly observed people older than 50 years old (90% of all CRCs are diagnosed after that age) and clearly the risk factor rises intensely after the age of 50. A study showed that, African Americans have a greater incidence and mortality rates, with respect to others [85]. Additionally, genetic background has a great impact on the formation of the CRC. Individuals who have a first-degree relative diagnosed with CRC, notedly if they were diagnosed at an early age (<55), have the risk around twice than the others. Also, having ovarian cancer, high-risk adenoma or being previously had CRC growths the risk [86].

2.3.3 Symptoms and Diagnosis

The symptoms of colon cancer are observed both in local and systemic manners. Some of the main local symptoms are alterations in defecation routine, constipation, diarrhea or change in between them [87]. These symptoms can be seen as unremarkable since they are basically symptoms of several common diseases, either. For example, diarrhea can be a symptom of alcohol abuse, a reaction to a specific type of food or drug, a bacterial infection [88]. Some of the symptoms of CRC affect the whole body, including feeling tired or exhausted, losing weight and appetite and vomiting [87]. Similarly, these symptoms are shared with some frequently observed diseases such as vomiting is observed in gastritis or poisoning, as well [89].

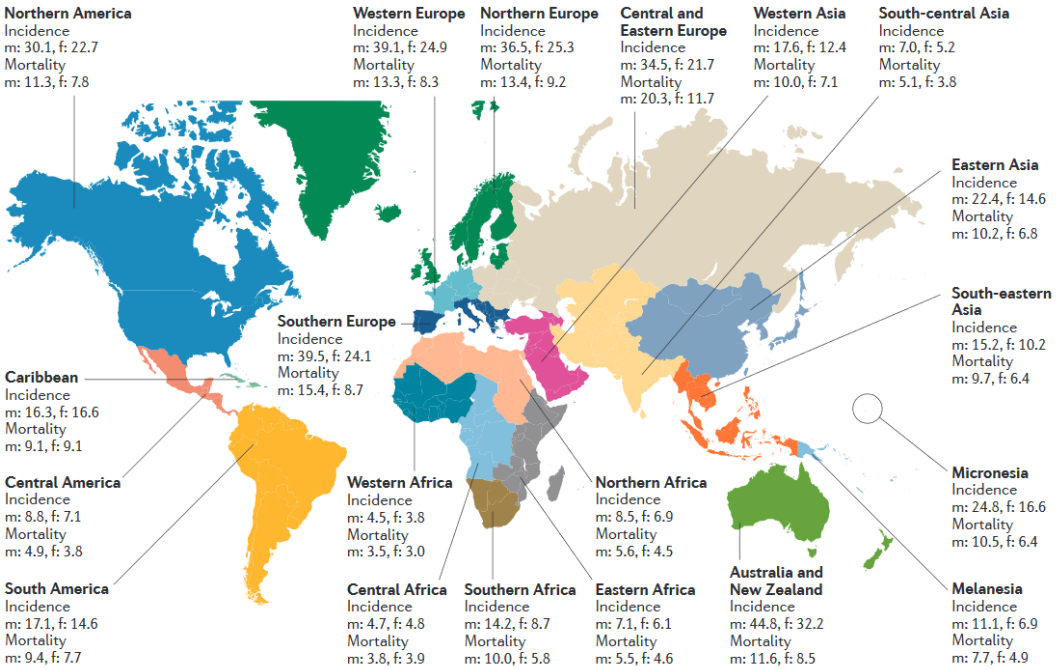


Figure 2.7: The distribution of colon cancer in world-wide with incidence and mortality rates.

Colon cancer is generally diagnosed at its later stages; therefore, screening is strongly recommended for early detection and treatment, especially for the individuals in the risk groups. There are four main screening technologies available for now, which are stool tests, flexible sigmoidoscopy, colonoscopy and virtual colonoscopy (computed tomography colonography).

Stool tests can be applied via three different methods: guaiac-based fecal occult blood test (gFOBT), fecal immunochemical test (FIT) and FIT-DNA test. gFOBT and FIT should be done once in a year, whereas FIT-DNA test should be done in once every one or three years. gFOBT can be done at home, with a kit obtained from health-care provider, by brushing a slight piece of stool, then the samples are checked in a laboratory to check the existence of blood. FIT is done similarly with the gFOBT, however it uses antibodies for detection of blood. Lastly, FIT-DNA test detect changed DNA in the stool addition to the FIT.

Flexible Sigmoidoscopy is done with a flexible, short, thin, lighted tube putted into the rectum and doctor checks if any polyps of cancer formation's existence in the rectum or the lower-third of the colon. It should be applied once in every five years, or once in every ten years if FIT is applied annually. In the case of colonoscopy, a longer stick is used, and it can detect cancer formation in the whole colon. Additionally, during colonoscopy most of the polyps can be removed, as well, and it should be applied in every ten years. Finally, in CT colonography, images of the whole colon are produced via X-rays and computers, then the doctor analyses these images. Virtual Colonoscopy should be applied once in every five years [90,91].

2.3.4 Molecular Pathology

Colorectal cancer is a result of a combination of hereditary and environmental factors which are resulting in gaining of "hallmarks of cancer" in colon epithelial cells. These behaviors are acquired by accumulation of mutations and epigenetic aberrations that trigger oncogenes and down-regulate tumor-suppressor genes. Almost all types of tumor formations, abnormal genomic or epigenomic stability is observed. In the specific manner of colon, a molecular event starts the formation of polyps, then they

grow up and divide uncontrollably. There are two distinct mechanisms discovered in which colon cancer forms from a polyp and they are both described in Figure 2.6.

In summary, aberrant crypt loci is formed from normal colon epithelial cells and then early polyps were formed in both cases. Then these polyps turn into either early or advanced carcinomas. During the "traditional" pathway (which is shown at the top in Figure 2.6), tubular adenomas grow and turn into adenocarcinomas. The alternative pathway (which is shown at the bottom in Figure 2.6), serrated polyps come to the stage and they progress into tumor tissue. During all of the stages of tumor formation, several genes are affected both in genetically or epigenetically, they are noted in the Figure 2.6, as well. Some of these altered genes shared by both mechanisms whereas some of them are unique for each mechanism [21].

2.3.5 Treatment

Treatment options differ the stage in which the CRC is diagnosed and can include surgery, chemotherapy and targeted therapy.

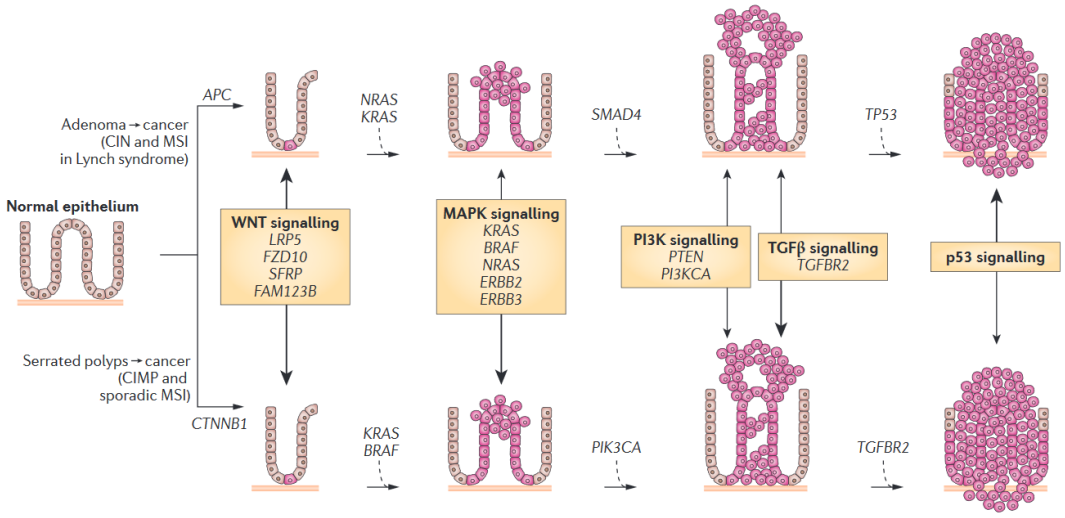


Figure 2.8: Formation of colon cancer from a healthy colon epithelial tissue. Retrieved from [21].

If the cancer is diagnosed before it spreads away from the wall of the colon, they are generally treated with surgery. However, if they have already metastasized, treatments focus to improve life-quality instead of curing the cancer [92]. In the case of rectal cancer, the standard protocol is to remove the rectum and mesorectum with the neighboring envelope. Mesorectum should be cleaned during the surgery, because it contains most of the lymph nodes affected by the tumor and tumor deposits. If it is a colon cancer surgery, similarly, the tumor and the associated lymph vessels are removed. The colorectal cancer surgery can be applied as open surgery or a laparoscopic surgery [93].

Chemotherapy can be used as an addition to the surgery for some cases, depending on the stage of the tumor. If the tumor is in its early onset stages (Stage I or II) chemotherapy is either not offered or it is debatable. However, if the cancer is diagnosed at later stages (Stage III or IV), especially if the cancer metastasized to the lymph nodes or different organs. Addition to the chemotherapy, radiation therapy can be used in rectal cancer, as well. Due to the sensitive nature of the colon, radiotherapy is not used in colon cancers [94].

3 METHODS

In this dissertation, a predictive statistical analysis was performed to provide insights on prediction of risk score in colon adenocarcinoma cases with epigenetic biomarkers. The method can be summarized in 4 main steps:

1. Dataset selection
2. Separate analysis of the data obtained from different sources
3. Selection of candidate genes that are regulated by DNA methylation
4. Selection of candidate genes that are regulated by miRNAs
5. Integration of omics data
6. Network analysis to find out candidate mechanisms
7. Identification of individualized affected pathways

Details of each step will be explained in their corresponding subsections. Overview of the applied method is provided in Figure 3.1.

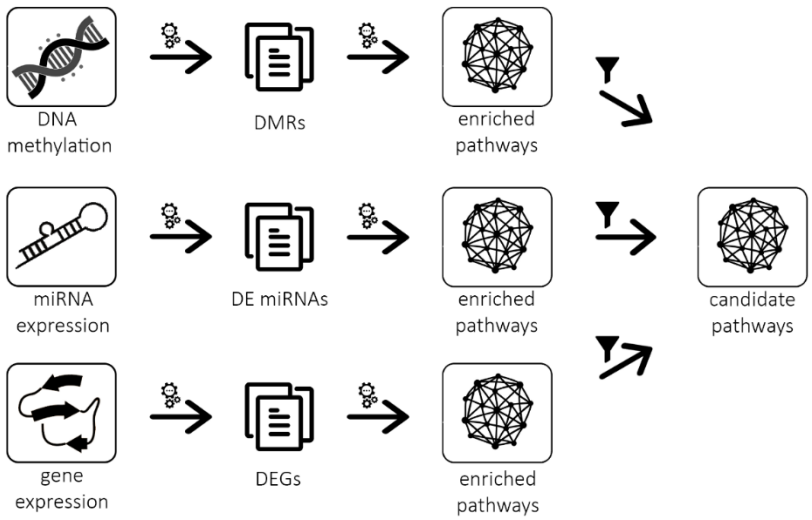


Figure 3.1: Overview of the method

3.1 Dataset Selection

For this dissertation, several DNA methylation, gene expression and miRNA expression datasets were obtained from past studies.

ß-values (level 3) of DNA methylation data of 37 patients was obtained from Illumina Human Methylation 450k platform The Cancer Genome Atlas (TCGA) Data Portal (http://cancergenome.nih.gov/). Bisulfide raw sequencing data of two patients' colon adenocarcinoma and adjacent normal tissue was obtained from Gene Expression Omnibus (GEO) to be used as verification dataset. [95] [23]

Gene expression levels of 132 colorectal tumors and 9 normal colon epitheliums produced with Agilent-014850 Whole Human Genome Microarray 4x44K G4112F were obtained from GEO portal [96]. As verification dataset, mRNA-Seq data of 9 colon adenocarcinoma and adjacent normal tissue obtained by using Illumina HiSeq were acquired from TCGA portal, mentioned above.

miRNA expression levels of 8 tumor and adjacent tissue data produced by Illumina Genome Analyzer IIx downloaded from a previous study done by Hamfjord J [97].

3.2 Separate Analysis of Data Obtained from Different Sources

Before integrating DNA methylation, mRNA and miRNA transcription data, all different kinds of data were analyzed separately. All of these analyses were performed in R (version 3.4.1, "Supposedly Educational") unless otherwise reported [98].

3.2.1 DNA Methylation

The DNA Methylation dataset obtained from TCGA was produced with Illumina HumanMethylation450k BeadChip arrays and the other two datasets downloaded from GEO portal were produced with Bisulfite-Seq method. Datasets produced with these

two different techniques were analyzed using two different methods. Therefore, the analyses were explained separately in the following two subsections.

3.2.1.1 DNA Methylation - Array Based

Illumina HumanMethylation 450k BeadChip Array is made up of genome-wide more than 450k probes at a single nucleotide resolution, and it covers 99% of RefSeq genes with an average 17 CpG sites per gene, 96% of CpG islands and additional probes for CpG shores and flanking regions. Gene associated probes locate among promoter, 5' UTR, first exon, gene body and 3'UTR regions. Moreover, the array also includes probes for miRNA promoters and cancer specific differential methylation sites [99].

ß-values that were converted from raw optical files were analyzed using a pipeline developed in R with functions from packages available in Bioconductor or CRAN.

ß-values represent the fraction of methylation observed in that probe and calculated by

$$ß = \frac{\max(Methylated, 0) + \alpha}{\max(Methylated\ + 0) + \max(Unmethylated, 0) + \ \alpha}$$

where *Methylated* stands for the signal intensity of methylated probes, *Unmethylated* stands for the signal intensity of unmethylated probe and α is an offset which is used to regularize the ß-value when both methylated and unmethylated intensities are low. Most commonly, α is considered as 100, and has no detrimental effects since most of the intensities are greater than 1000 [100].

ß-values take values in between 0 and 1 or 0% and 100%, and it defines methylation status perceptively. ß-value = 0.6 for a probe tells that, 60% of CpGs at that location is methylated. ß-values obtained from raw methylation data was publicly available in the TCGA portal. Before applying any quality control test, 65 probes which are interrogating SNPs were discarded. Additionally, SNP-overlapping (29481) or cross-reactive probes (9341), probes in X and Y chromosomes (11524) and probes

that are missing more than 20% of the samples (67496) were discarded and further analysis were completed with remaining 368570 (76%) probes.

Missing ß-values were imputed with *k*-nearest-neighbors method for *k=15*. For imputing, firstly, *K*-nearest neighbors was found for each probe with missing values, by using a Euclidean metric limited with the columns for which that probe is not missing. The distance is calculated by using each candidate neighbor might be missing some of the coordinates. After having the known the k-nearest-neighbors for a probe, the missing element is imputed by taking the average of non-missing values from its neighbors. If finding k-nearest-neighbors step is failed, then column average was imputed for missing value [101].

Quality control was performed with functions from ChAMP package and R's base functions. Raw values were examined with ß-value distribution, multidimensional scaling and hierarchical clustering plots.

After quality control and exploratory analysis, Beta mixture quantile dilation (BMIQ) normalization was performed using appropriate function from ChAMP package. Since Illumina HumanMethylation 450k arrays contain two different designs (*type-I* and *type-II* probes) in the same array, the signals coming from different probes contain different distributions. Therefore, when applying normalization, it is required to account these two different designs as different sources of variation and normalize each one.

Singular value decomposition analysis is one of the most powerful methods for revealing the number and the nature of the important components of variation. Ideally, the significant components are expected to be biological factors, however this is not the case all the time. Technical sources of variation, such as batch effects, can be leading components of the variation. To identify source of the variation, SVD analysis is applied with normalized data. For identification of variation from categorical variables, Kruskal test is used whereas linear regression is used for numeric covariates [102].

As technical covariates batch, portion, plate and center where samples collected are considered. As it can be seen below, "*Array*" is one of the important components on the phenotype with a $p<0.01$ as Principal Component #1, therefore ComBat correction was applied for "*Array*" covariate to remove technical variation [103].

ComBat algorithm combines systematic batch biases common across genes in making adjustments and assumes that the batch effect affects many genes in a similar way. The algorithm can be explained in three steps: Firstly, standardizing the gene expression levels across genes so that all genes will have similar overall mean and variance so that empirical Bayesian method will not be biased by magnitude changes caused by different gene expression levels or probe intensities. Afterwards, by assuming the data will satisfy the distribution form $\mathbf{Z_{ijg} \sim N\,(\gamma_{ig},\, \delta^2_{ig})}$, hyperparameters are estimated empirically. By the distribution assumptions made, empirical Bayesian estimates of batch effect parameters calculated. Finally, after calculating the adjusted batch effect estimators, the data is adjusted [104].

Differentially methylated regions (DMRs) are genomic regions that have distinct methylation patterns across multiple CpG sites. Reducing differentially methylated probes (DMP) into DMRs are preferable since it is required to find regions of interests and their position according to genes. DMRs were identified using Probe Lasso approach implemented in ChAMP Bioconductor package [105]. Probe Lasso algorithm uses a flexible window-based approach, which considers uneven probe spacing, for gathering neighboring significant methylation signals to define a clear DMR. Once derived, probe-lassos can be thrown around a probe and its radius extents upstream and downstream, centered on the targeted CpG. For this study in order to call a probe-lasso as DMR, minimum probe threshold was set as 3 [106].

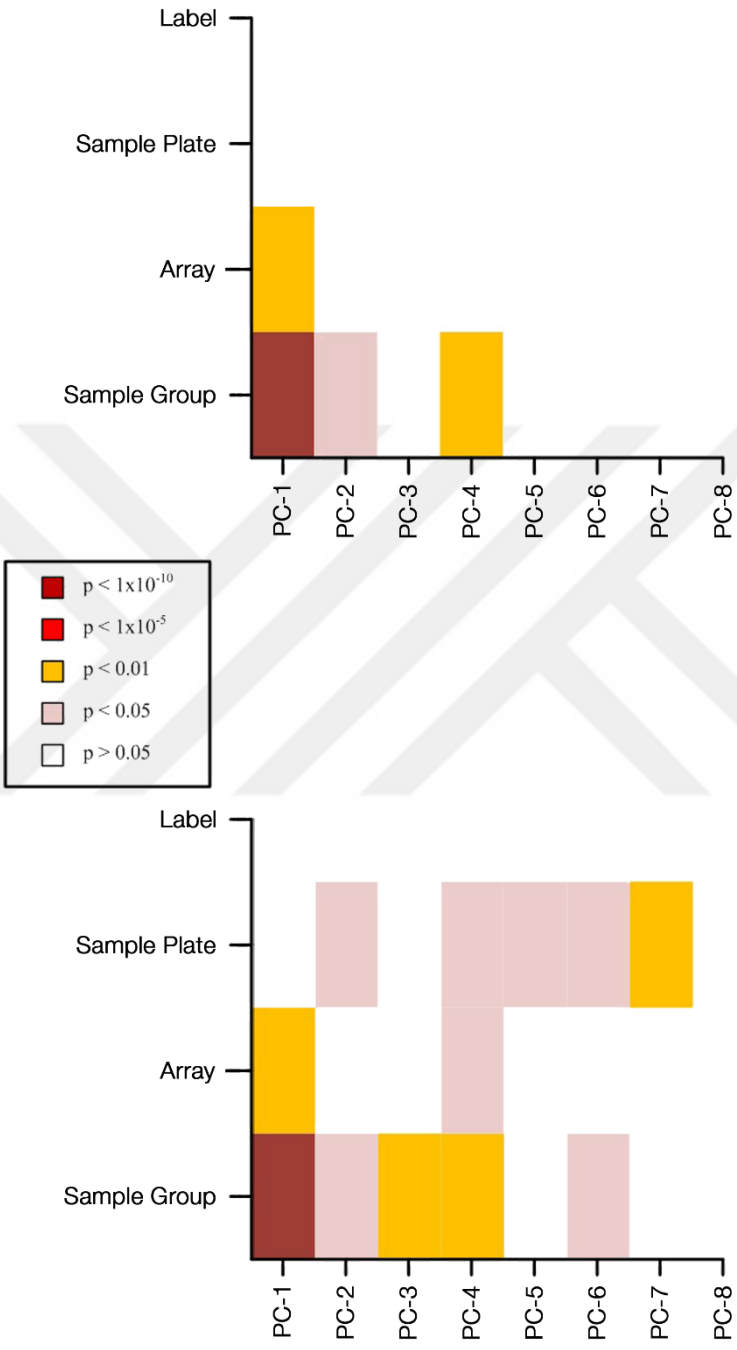


Figure 3.2: Singular value decomposition analysis, above after ComBat correction is applied and below, before ComBat correction was applied.

3.2.1.2 DNA methylation - bisulfite-sequencing based

Recall that, in DNA methylation process a methyl group (-CH_3) covalently binds to cytosines in the DNA strand. When DNA is treated with bisulfide, unmethylated cytosines in the DNA turns into uracil in a sulfonation-deamination-desulfonation reaction whereas methylated cytosines remain unchanged. After bisulfite treatment, PCR amplification is performed with appropriate primers, then PCR product is purified and finally prepared for sequencing [107].

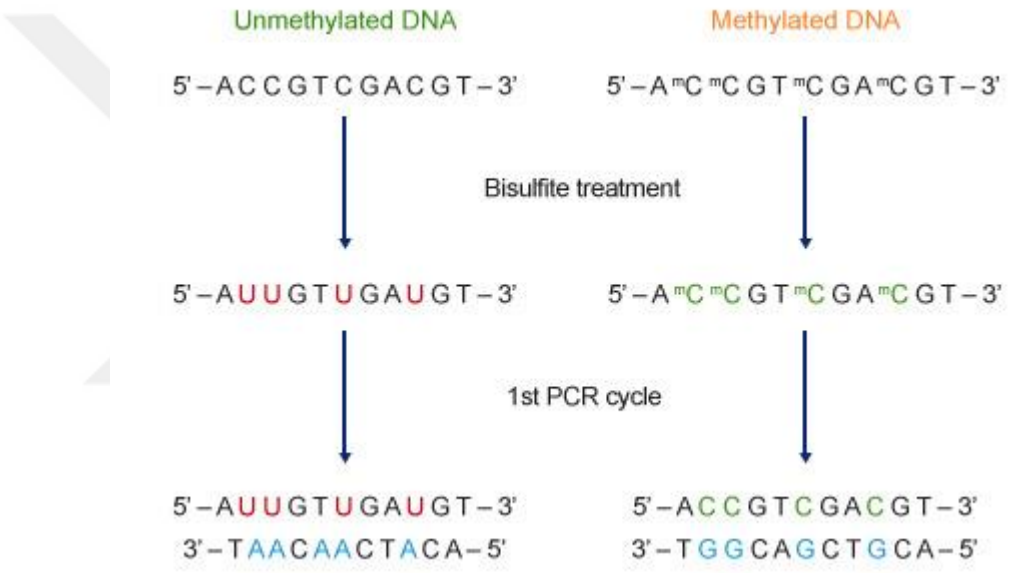


Figure 3.3: Bisulfite treatment and products of PCR. Republished from [108]

The data used for this study was obtained as bed files which contains DNA methylation levels. To obtain mentioned bed files, Bi-Seq raw sequencing reads were aligned against human reference genome version hg19 and duplicated reads were discarded. Then DNA methylation calling was performed as explained in [95,109].

Firstly, the bed files containing DNA methylation signals, were read to workspace with `methRead` function written in MethylKit package in R. Then, regions having read count less than 10 are discarded and united. As explanatory analysis, Pearson

correlation coefficient matrix was generated, and hierarchical clustering applied between samples to reveal the relationship between samples. Additionally, principle component analysis was performed to reveal association between the phenotype and source of data (details and background of the method was explained in the section 3.2.1.1). Since the first principal component was the source of data, the first principal component was removed from the data by using functions from MethylKit package [110].

After the steps explained above, differentially methylated regions were detected, and p-values were calculated by logistic regression. Then, q-values were obtained from p-values by using SLIM method proposed by Wang, Tuominen and Tsai [111].

For the sake of simplicity, let's consider the simplest form where there are not any covariates: The logistic regression will model the log-odds ratio which can be written as following where π_i's are treatment vectors that denote the sample group for the CpGs in the model.

$$log(\frac{\pi_i}{1-\pi_i}) = ß_0 + ß_1 Treatment_i$$

After calling out differentially methylated regions, the genes and gene parts were annotated against human reference genome version Hg19 using GRanges package in R [112].

3.2.2 Transcription

3.2.2.1 mRNA – miRNA Transcription – Array Based

mRNA transcription data of 132 micro dissected colorectal tumor and 9 micro dissected normal colon epithelium samples obtained from GEO Database (GSE21815) produced with Agilent DNA Microarray Scanner.

Raw signal densities were imported to R environment using *limma* Bioconductor package [113]. Background correction was applied with *normexp* method as it is suggested that the method is superior other correction methods [103]. After that,

quantile normalization between two-color arrays was done. Quantile normalization ensures that the intensities coming from two different channels share the same empirical distribution across array [114].

After background correction and between array normalization was done, control probes were removed out for further analysis. Probes, that are not corresponding any gene (e.g. locating at non-coding regions) were checked, however there were no such probe.

To decide on differentially expressed genes (mRNAs or miRNAs) multiple linear models were fitted by weighted or generalized least squares. A linear model was fitted to the expression data of each gene. Then, the genes were ranked in order to evidence for differential expression by using an empirical Bayes method to reduce gene-wise sample variances towards a common value [115]. After assigning p-values and log-FC rates were assigned, an adjusted p-value threshold ($p<0.05$) was set to decide significantly differential expression.

3.2.2.2 mRNA – miRNA Transcription – Sequencing Based

Differentially expression analysis was performed using edgeR Bioconductor package [116]. Raw read counts were extracted Level 3 data obtained from TCGA (for the case of mRNA) and the available dataset in GEO (for the case of miRNA), and then count matrices were created [97].

Firstly, genes which have average log2 counts per million lower than zero were filtered out, that means genes with read count of ~5 or more per sample were remained for the analysis. Then, normalization was done with `calcNormFunctions` from edgeR package, which normalizes RNA composition by finding a set of scaling factors that minimize the log-FC between the samples for most genes. For computing these scaling factors trimmed mean of M-values (TMM) between each pair of samples was used. Later, dispersions were estimated with the Cox-Reid profile adjusted likelihood method. The CR method relies on the idea of approximate conditional likelihood which reduces the residual maximum likelihood. As explanatory analysis,

multidimensional scaling, mean-variance relationship and biological coefficient of variation (BCV) plots were drawn.

To call differentially expressed RNAs, a negative binomial generalized log-linear model was fitted for each gene (mRNA or miRNA). Then likelihood ratio test was performed to assign p-values and test differential expression. The genes with significance indicator p-value < 0.05 and change level, |logFC| > 1 are considered as above threshold and kept for the further analysis.

3.3 Selection of Candidate Genes That are Regulated by DNA Methylation

Differentially methylated regions were annotated with their corresponding genes via probe information database, available in `ChAMPdata` package in R environment [117]. For the genes with several DMRs detected, the most-significant change (in other words, DMR with the smallest p-value) was selected and rest of them were discarded.

Candidate genes were decided in two ways, firstly the ones which are down-regulated, secondly the ones which are up-regulated.

To select *up-regulated genes by methylation*, the genes which were hypo-methylated and expression level is increased were chosen. The genes which were hyper-methylated and transcription levels are down-regulated were chosen as *down-regulated genes by methylation.*

3.4 Selection of Candidate Genes That are Regulated by miRNAs

Differentially expressed microRNAs were annotated using miRTarBase database available in http://mirtarbase.mbc.nctu.edu.tw [118]. Only experimentally validated relationships with miRNAs and their target genes were extracted. Then, these targeted genes were matched with their miRNAs.

Candidate genes were decided in two ways, firstly the ones which are down-regulated, secondly the ones which are up-regulated, as in the case of selection of genes regulated by methylation.

An *up-regulated gene by miRNA* selected if the gene is up-regulated and at least one of its regulatory miRNA's expression level is lowered. Similarly, a down-regulated gene by miRNA selected if the gene itself is down-regulated and at least one of the miRNAs which targets that particular gene is *up-regulated*.

3.5 Pathway Analysis to Find out Candidate Mechanisms

Pathway analysis were completed in R environment, using the `PathfindR` package developed by our group [119]. Generally, pathway analysis algorithms use a set of genes and identify the pathways that genes function commonly. However, genes do not independently perform their functions in the pathways, in other words, they are interacting with each other, not only physically but also functionally. The method pathfindR practices firstly finds out active sub-networks in which differentially expressed genes are interacted with each other. For finding active-subnetworks, greedy algorithm was used in this dissertation [120]. After calling out active sub-networks, then pathway enrichment analysis was applied on the orientation of active sub-networks (Figure). The pathway enrichment analysis uses one-sided hypergeometric test. Instead of using whole protein-protein interaction network, this method uses only proteins and their interactions take place in statistically significant active sub-networks. All of these searches repeated for 10 times (since greedy algorithm does not promise to provide the best solution, but only the local-best solution) and resulting pathways are reported. Resulting p-values were adjusted using Bonferroni method.

For the active-subnetwork analysis, genes in Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were used as protein interaction network (PIN) [121,122]. Threshold p-value or adjusted p-value is taken as $p < 0.05$ unless otherwise noted.

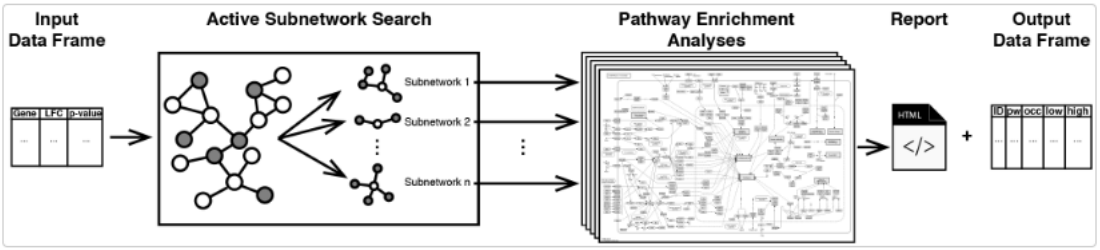


Figure 3.4: Workflow of active-subnetwork-oriented pathway enrichment analysis. Retrieved from [123].

After finding DMRs, DEGs, DE-miRNAs separately, pathway analysis was performed for each type of dataset separately. As an example, a pathway analysis was performed with gene list obtained from array-based DNA methylation dataset with corresponding p-values and Δß values instead of log-FC values.

All of these affected pathways obtained from different data sources were combined if they are observed in more than one type of dataset and considered as "*candidate mechanisms for colon cancer progression*".

3.6 Integration of omics data

Epigenetic factors, such as DNA methylation or small non-coding RNAs effect the expression level of a gene by either inactivating or activating their transcription. In the case of miRNAs, they are mostly silence the genes that they targeted. Likely, methylation on a gene body or a gene's promoter site is linked with down-regulation, as well.

To integrate these different levels of regulation, firstly enriched pathways were revealed by gene expression, DNA methylation and miRNA expression-based regulation levels. The pathways in which genes are silenced through miRNAs or hyper-methylation mechanisms were found out by intersecting the pathways obtained by hyper-methylation, up-regulated miRNAs and down-regulated mRNAs.

3.7 Identification of Individual Affected Pathways

To investigation of individual pathways leading to colorectal adenocarcinoma, pathways are affected by aberrations of DNA methylation, differentially expressed miRNAs and differentially expressed mRNAs were detected separately.

Array-based datasets have some disadvantages for personalized investigation. Two main disadvantages are their ability to relative measurement and dependence on PCR. Since array-based technologies measure the signals coming from the probes of interest, differentiating the reason of variance, if it is biological or technical, is not possible without replicates. However, in the case of sequencing, read counts of a gene can be obtained and they can be compared as tumor and control cases. Additionally, PCR step required in array technologies can provide synthetic duplicates and consequently create a bias, as well. [124]. Therefore, array-based datasets are used for creating a database of mechanisms shared by all the people in the dataset. Sequencing-based datasets are used for prediction of individual mechanisms leading to colorectal cancer.

3.7.1 Detection of DMRs for Each Individual in the Dataset

In order to find out individual mechanisms regulated by DNA methylation leading to colorectal adenocarcinoma, firstly DMRs were detected for the person of interest. Different approaches were used for different technologies used for detection of methylation levels. These approaches were explained in the following subsections, Section 3.7.1.1 and Section 3.7.1.2

For investigating differentially methylated regions for each person, only that particular person's DNA methylation data was analyzed, and corresponding genes were annotated as explained in the Section 3.2.1.2. The significance threshold $p<0.05$ and differentially methylation threshold $|\Delta ß| >= 20$ were kept. The pipeline is performed for each individual's methylation levels gathered from tumor and adjacent normal tissue.

3.7.2 Detection of DEGs for Each Individual in the Dataset

In order to find out individual mechanisms in gene expression level which lead to colorectal adenocarcinoma, firstly DEGs were detected for the person of interest. Different approaches were used for different technologies used for detection of gene expression levels. These approaches were explained in the following subsections, Section 3.7.2.1 and Section 3.7.2.2.

To reveal personalized pathways causing to colon cancer formation, differentially expressed genes were detected for each individual in the dataset.

Analysis of differentially expression in transcription level is completed with a different method than the method which was explained in Section 3.2.2.2. This pipeline is mainly based on an R package DESeq2's negative binomial distribution approach. For differentially expression call, a generalized linear model was used:

$$K_{ij} \sim NB\left(\mu_{ij}, \alpha_i\right)$$

$$\mu_{ij} = s_i * q_{ij}$$

$$\log_2(q_{ij}) = X_j * ß_j$$

where K_{ij} stands for counts of gene i coming from sample j and modeled with a Negative Binomial distribution with a mean μ_{ij} and gene-specific dispersion parameter α_i. Since the mean μ_{ij} is fitted, it is composed of a sample-specific size factor s_j and a parameter q_{ij}. The parameter q_{ij} is proportional to the expected true concentration of fragments for sample j. The coefficients ßi provide the $\log_2$FC values for gene i, for each column of model matrix X. Since in this case, we do not have any replicates but only one patient's data, estimation of the dispersion of counts around the expected value is not applicable [125]. However, differentially expressed genes are decided by comparatively.

3.7.3 Detection of DE miRNAs for Each Individual in the Dataset

For identification of personalized pathways causing to colorectal carcinoma regulated by miRNAs, differentially expressed miRNAs were detected for each patient in the dataset. In order to analyze differentially expressed miRNAs, the same pipeline explained in Section 3.7.2.2. was modified and used. After detecting differentially expressed miRNAs, their target genes were annotated as explained in Section 3.2.2.2.

3.7.4 Identification of Personalized Pathways Affected at DNA Methylation, Gene Expression or miRNA Expression Levels

After calling out differentially methylated genes, differentially expressed genes and the genes whose regulatory miRNAs differentially expressed, pathway analysis was done for each person and each set of gene-list as it is detailed in Section 3.6. In summary, for each patient, differentially methylated genes, their differentiation levels (Δß) are used for pathway enrichment analysis. Similarly, for each individual, differentially expressed genes and their logFC levels were used in order to reveal affected pathways as gene expression level. Finally, differentially expressed miRNAs' target genes were used for pathway analysis, separately for each individual.

Since in cancer, lots of mechanism are corrupted, pathway analysis ends up with lots of affected pathways for each type of data source, as expected. To determine which pathways are targetable and causative, each pathway was stored to create a database. If a pathway is observed more frequently than others, with other words, as more people is affected by that particular pathway, then it is decided to be more relevant with the colorectal cancer.

4 RESULTS

4.1 Multi-Omics Analysis Results

4.1.1 DNA Methylation

4.1.1.1 Array-Based Dataset

Before calculating differentially methylated regions, differentially methylated probes were called out. 207.384 differentially methylated probes were detected with an adjusted p-value cut-off <0.05. 125.142 of those are hyper-methylated whereas remaining 82.242 of them are hypomethylated. These probes correspond to 5.812 distinct genes or open reading frames (ORFs).

To determine differentially methylated regions, probes were gathered with a Probe-Lasso approach, which was explained in Section 4.1, and distribution of the Lassos shown in Figure 5.2. In the figure, radius of the circles represents the size of the Lassos. Greater radiuses correspond larger Lassos, in other words, larger area which has more differentially methylated probes comparing to the smaller radiuses. The numbers on the circles show number of Lassos at that specific sub-group of DNAs. For a region to be defined as “differentially methylated” there should be at least 3 differentially methylated probes at that particular region. Sliding window size used in Probe Lasso is set as 7 for this dissertation.

As it is seen in Figure 5.1, differentially methylation is mostly observed in CpG shelves of 3’UTRs and 5’UTRs. Also CpG shelves and open seas of intergenic regions are one of the mostly aberrant methylated regions. Regions away from at least 200 base-pairs and 1500 base-pairs away from the transcription start sites are the least differentially methylated regions, as they are represented with smaller circles in the figure.

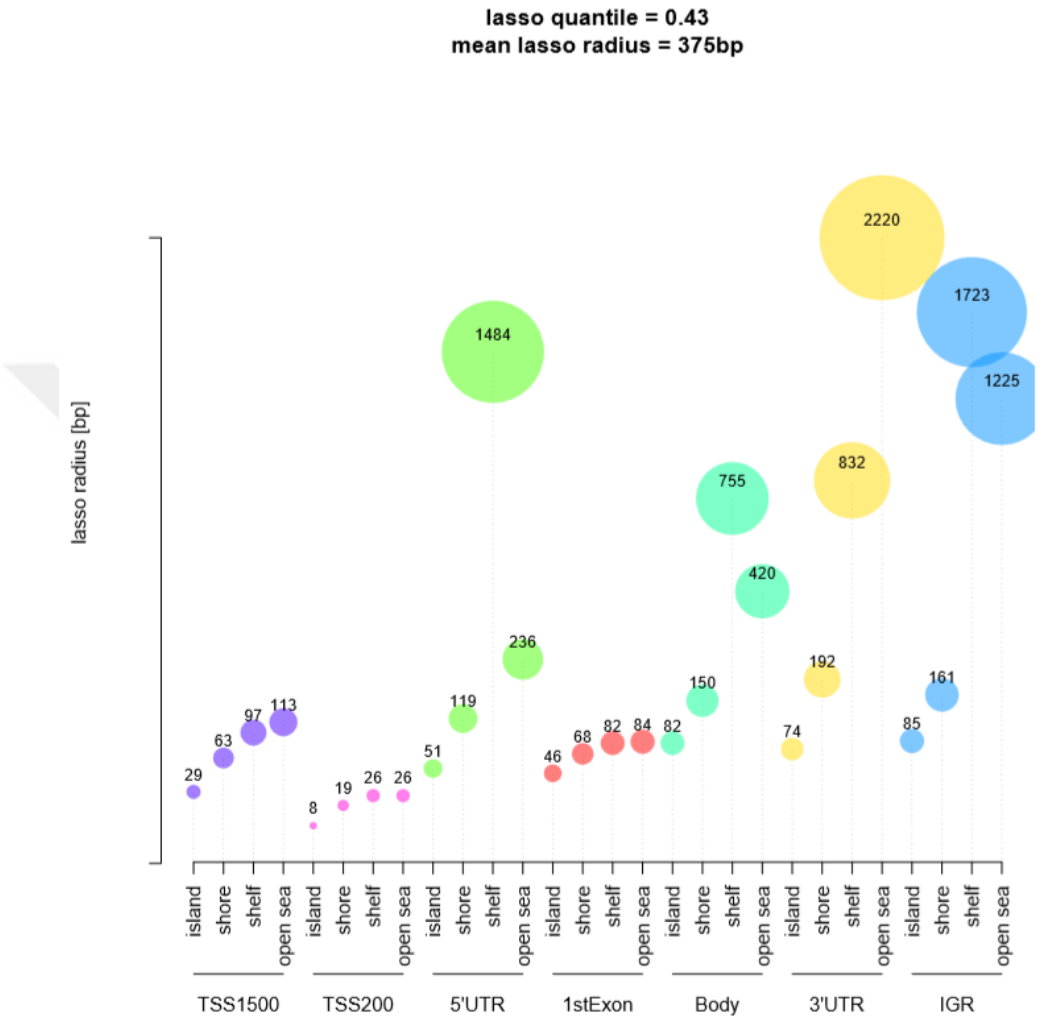


Figure 4.1: Distribution of DMRs gathered by DMPs using Probe-Lasso approach

4.1.1.2 Bisulfite Conversion-Based Dataset

Differentially methylated regions called out from bisulfite sequencing data was filtered with q-value <0.05 and |Δß| >= 0.20. After applying those filters, 5.679 DMRs were detected as hypo-methylated whereas 342 of them were hyper-methylated. When the DMRs mapped to their corresponding genes, 121 gene detected as hyper-methylated and 566 genes are hypo-methylated.

The distribution of aberrant methylation both in hyper and hypo methylation cases is provided in Figure 5.3. In the figure, parts colored with green represents CpG islands, purple color shows CpG shores, blue colored part represents CpG shelves and lastly pink part shows intergenic regions. As it can be seen, intergenic regions are the most effected regions by aberrant DNA methylation. They made up 82.63% of hyper-methylation and 80.65% of hypo-methylation. With 6.95% hyper-methylation and 9.86% hypo-methylation percentages, CpG shores are the second most affected regions. After those, CpG islands have 5.35% contribution to the hyper-methylation patterns, whereas they did 3.41% to the hypo-methylation patterns. Finally, least hyper-methylated region is CpG shelves with a 5.07% and they made up 6.1% of the hypo-methylation patterns.

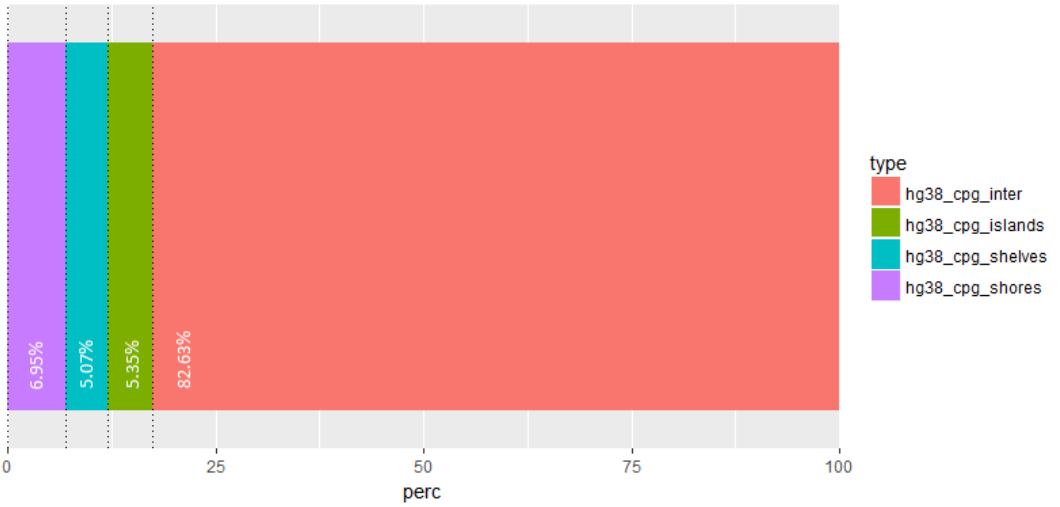


Figure 4.2a: Distribution of differentially hyper- methylation across CpGs

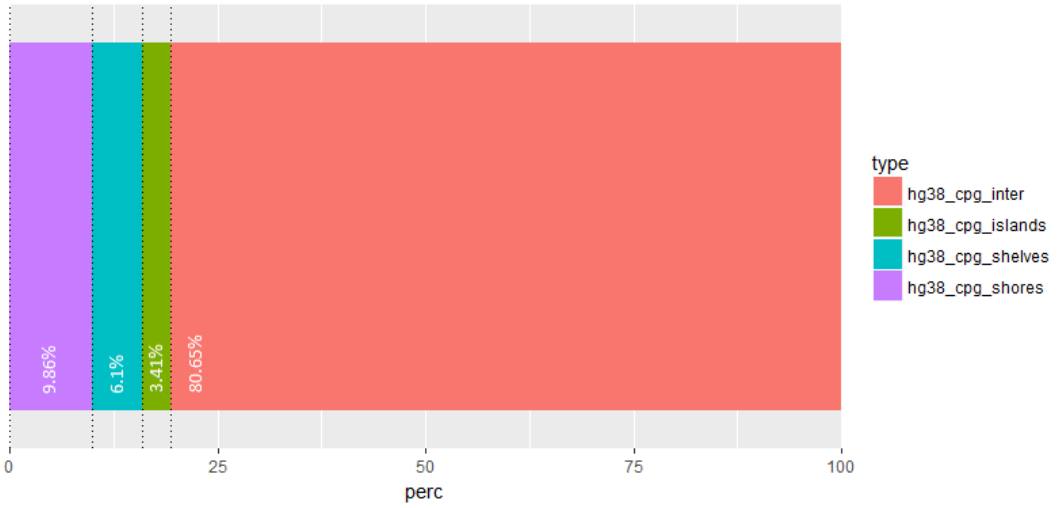


Figure 4.2b: Distribution of differentially hypo-methylation across CpGs

In order to understand the role of aberrantly methylated genes, pathway analysis was applied, which is explained in Section 5.2.

4.1.2 mRNA Transcription

4.1.2.1 Array-Based Dataset

Differentially expression call resulted with 496 differentially expressed genes, with a p-value threshold ($p<0.05$), at this point no fold change level threshold was applied. 336 of these genes are detected as up-regulated whereas 160 genes were found as down-regulated.

These down-regulated genes include well-known tumor-suppressor genes such as *WT1*, *BCL2L1* and *NPAS2*. WT1 encodes a transcription factor which plays a role in cell development and cell survival [126]. *BCL2L1* is a protein coding gene and in literature, it is associated with regulation of cell death and regulator of G2 checkpoint during the mitosis [127]. Suppression of *BCL2L1* is linked with cancer in the previous studies [128]. *NPAS2* is one of the key CLOCK genes which is a transcription activator and forms a core component of circadian clock. Down-regulation of *NPAS2* gene is associated with DNA damage repair mechanism as well and trivially cancer in the

literature [129,130]. The top-10 down-regulated genes can be found in table (sorted by increasing p-value) 5.1.

Gene Symbol	p-value	Adj p-value	log-FC
AK001903	8,8177E-07	0,03824764	-5,19154E+13
A_24_P50139	2185,960429	4799170,844	-2,75701E+14
THC2341518	39460,58348	65832394,96	-7,55313E+14
DCC1	59376,30327	95389130,76	-9,72423E+14
DDX31	281489,5577	368997297,2	-2,84378E+14
ZMYM3	330883,1531	410068218,6	-1,20501E+14
BCL2L1	976756,762	103336100,8	-1,30219E+14
BF688944	16150557,51	1,07776E+14	-7,64718E+14
FZD6	61154604,65	3,49325E+14	-8,73093E+14
SALL4	102129560,9	5,33732E+14	-7,42116E+14

Table 4.1: The most significantly down-regulated 10 genes based on array-based transcription dataset's analysis results

In the up-regulated genes, there are several known oncogenes as well. For instance, *BRCA2* is detected as up-regulated and it is associated with several cancer types, especially with breast and ovarian cancer [131]. *BRCA2* plays a key role in DNA repair mechanism, comes to the stage in a double stranded break repair mechanism: homologous recombination [132]. Regulation by DNA methylation of *BRCA1* and *BRCA2* genes in breast cancer cell lines are shown in a recent study, as well [133]. Another up-regulated gene is *EIF5A2* (Eukaryotic Translation Initiation Factor) and it takes place in several pathways including metabolism of proteins, stress response and also has a regulatory role in apoptosis [134]. Previous studies showed that, up-regulation of *EIF5A2* is linked with metastasis in colon cancer [135]. Additionally, there are several studies focus on *EIF5A2*'s role in human cancer, too [reviewed in 59]. The most significantly up-regulated 10 genes can be found in Table 5.2.

4.1.2.2 Sequencing Based Dataset

At the end of differentially expression analysis, 2.447 genes were found as down-regulated and 2.033 genes were found as up-regulated, in total 4.480 genes were differentially expressed. During the differentially expression call, a p-value threshold (p-value < 0.05) and absolute log-fold change (|logFC] > 1.00) threshold were applied.

Gene Symbol	p-value	Adj p-value	log-FC
EIF5A2	2,33112E-05	0,050557307	2,2189E+14
BC035417	0,000950531	0,137434044	9,48776E+13
NKD1	0,00252716	27,40452079	3,94993E+14
ABHD7	0,023565503	204,4354476	9,65187E+14
ETV4	0,129259284	903,3720989	2,57031E+14
TMPRSS3	0,145785796	903,3720989	4,10198E+14
SLC39A6	16,11336453	87366,66248	3,80643E+14
LOC196394	20,64684903	99508,63594	4,28047E+14
TBC1D16	24,26777477	105263,8999	2,03612E+14
KDELC1	38,10271313	15024,93895	1,42203E+14

Table 4.2: The most significantly up-regulated 10 genes based on array-based transcription dataset's analysis results

In these down-regulated 3.507 genes, there are several known tumor suppressor genes as well. Some of them are, *FAS*, *CASP5* and *EGR1*. *FAS* is a protein encoding gene which belongs to TNF-receptor subfamily. TNF-receptors has a key role by inducing subsequent cascade of caspases and mediating in programmed cell death mechanisms and they are linked with several cancer types and malignancies [137,138]. *CASP5* gene encodes a member from cysteine-aspartic acid protease (caspase) family, which takes part in cell apoptosis's execution-phase and has parts in TRAF Pathway and Presenilin-Mediated Signaling. In the literature, *CASP5* gene' relation with several

cancers, including prostate and colon was shown [139,140]. EGR1, Early Growth Response-1, is a protein coding gene and acts as transcriptional regulator of differentiation and mitogenesis. It is also known as a tumor suppressor gene, meaning that down-regulation of it is highly correlated with cancer formation [141,142]. The most significantly down-regulated 10 genes can be found in Table 5.3.

Gene Symbol	**p-value**	**Adj p-value**	**log-FC**
CLEC3B	2,17615E-28	6,10628E-25	-4,1425E+14
LYVE1	8,65022E-20	1,45635E-16	-4,3155E+14
CLDN5	2,3865E-17	3,09071E-15	-2,8198E+14
FGL2	8,67201E-17	1,04287E-13	-2,4978E+14
UGP2	1,47227E-16	1,65248E-13	-2,0319E+14
CFD	3,96958E-16	3,93129E-14	-4,0923E+14
GLP2R	1,42756E-13	1,04497E-11	-3,3549E+14
PLPP1	2,07061E-13	1,45253E-11	-1,7868E+13
NR5A2	5,87876E-12	3,53481E-09	-2,3269E+14
ITM2C	1,71945E-11	9,04647E-09	-2,3725E+14

Table 4.3: The most significantly down-regulated 10 genes based on sequencing-based transcription dataset's analysis results

In the case of up-regulated genes, there are several oncogenes observed, such as *MYC*, *TP53*, *BRAF*. *MYC* is known as a proto-oncogene, so its role in various cancer is well-studied. *MYC* encodes a nuclear phosphoprotein which has several metabolic functions, such as cell cycle progression, programmed cell death and cellular transformation. It plays role in several pathways including ERK Signaling and p38 MAPK Signaling Pathway [143,144]. *TP53* is the one of the most well-studied genes whose mutated versions are frequently observed in tumor formation and progression. It encodes a tumor-suppressor protein which can respond to several cellular stress by several mechanisms such as cell cycle arrest, apoptosis, DNA repair, or other changes [145–148]. *BRAF* is another proto-oncogene which encodes a protein belonging RAF

family of serine/threonine protein kinases and takes part in regulation of MAP kinase / ERK signaling pathway. These pathways are associated with cell division, differentiation and secretion. Previous studies showed the link between *BRAF* gene and various cancers, such as melanoma, colorectal cancer and so on [17,149,150]. The most significantly 10 up-regulated genes are shown in Table 5.4.

Gene Symbol	**p-value**	**Adj p-value**	**log-FC**
TRIB3	6,35136E-42	1,06932E-37	3,65184E+14
GRIN2D	1,73283E-39	1,45869E-36	5,1832E+13
KRT80	3,80079E-37	2,133E-33	7,14617E+14
ETV4	2,70593E-36	1,13893E-34	5,88366E+14
CLDN1	9,45728E-33	3,18446E-29	5,12444E+14
AJUBA	5,18516E-27	1,2471E-23	3,20435E+14
NFE2L3	7,41583E-21	1,56066E-17	2,99433E+14
OSBPL3	2,61364E-20	4,88925E-17	2,24577E+14
CDH3	1,2256E-19	1,87583E-16	6,28058E+13
KIAA1257	2,07923E-17	2,91716E-14	3,58927E+14

Table 4.4: The most significantly up-regulated 10 genes based on sequencing-based transcription dataset's analysis results

4.1.3 miRNA Transcription

For miRNA transcription, only a sequencing-based dataset is used, as mentioned in Section 3.1. Result of differentially expression call ended up with 77 miRNAs, 29 of them are down-regulated whereas 48 of them are up-regulated. Since 18 of these differentially expressed miRNAs' target genes were not shown experimentally yet, they were discarded for the further analysis, leaving 59 miRNAs. 21 of them were down-regulated whereas remaining 38 miRNAs are up-regulated. These miRNAs target 1986 distinct genes. Down-regulated miRNAs control 841 distinct genes and up-regulated miRNAs control 1258 distinct genes. 113 of these genes are controlled by both up-regulated and down-regulated miRNAs.

One of the most significantly down regulated miRNAs, hsa-miR-422a targets TGFB1 (Transforming Growth Factor Beta 1) gene, which is a protein coding gene. As its regulatory miRNA's expression level goes down, TGFB1 is expressed with greater amounts. Besides its regulatory effect on cell cycle, TGFB1 is also a regulator of other growth factors and it is frequently observed up-regulated in cancer [151].

miRNA	**p-value**	**Adj p-value**	**log-FC**
hsa-miR-422a	288341486,2	2,85562E+14	-3,0442E+14
hsa-miR-195	329524639,8	2,85562E+14	-2,1757E+14
hsa-let-7d	1090508070	6,36584E+14	-1,5483E+14
hsa-miR-378	1844160156	6,90595E+14	-2,2066E+14
hsa-miR-378c	1922427450	6,90595E+14	-2,6686E+14
hsa-miR-145	4240894577	1,23781E+14	-2,2832E+14
hsa-miR-3656	1,1149E+14	2,63434E+14	-3,4347E+14
hsa-miR-3151	1,18461E+14	2,63434E+14	-2,8802E+14
hsa-miR-497	2,31341E+14	4,00134E+13	-1,5717E+14
hsa-miR-140-3p	4,82439E+14	7,04059E+14	-1,2451E+14

Table 4.5: The most significantly down-regulated 10 miRNAs

hsa-miR-7 is one of the most-significantly up-regulated miRNAs and it controls the activity of *BCL2* gene. *BCL2* encodes a mitochondrial membrane protein and regulates cell death. As has-miR-7 expressed highly, *BCL2* is being transcribed lower levels, therefore failures on apoptosis is observed. The relationship between *BCL2* and several cancer types, such as lymphoma and colorectal adenocarcinoma, was revealed with previous studies [53,128,152].

miRNA	p-value	Adj p-val	log-FC
hsa-miR-7	6291047	2937919047	6,20808E+14
hsa-miR-96	12864786	3003927441	3,14114E+14
hsa-miR-135b	2,02E+08	2,85562E+14	3,61814E+14
hsa-miR-584	3,67E+08	2,85562E+14	2,5516E+14
hsa-miR-542-3p	9,99E+08	6,36584E+14	1,6448E+14
hsa-miR-1269	1,38E+09	6,90595E+14	5,16645E+14
hsa-miR-549	1,56E+09	6,90595E+14	4,73137E+14
hsa-miR-493	1,63E+08	6,90595E+14	2,42582E+14
hsa-miR-889	3,41E+09	1,13639E+14	1,96302E+14
hsa-let-7i	4,22E+09	1,23781E+14	1,35821E+14

Table 4.6: The most significantly up-regulated 10 miRNAs

4.2 Candidate Pathways

In order to define possible mechanisms leading colorectal adenocarcinoma, pathways affected at DNA methylation, gene expression and miRNA regulation levels were called out separately.

For separate analysis of pathways, differentially methylated DNAs were divided into two categories as hyper-methylated gene set and hypo-methylated gene set, for both array and sequencing-based datasets. Then pathway analysis was performed, separately.

4.2.1 Candidate Pathways Obtained from Gene Transcription Datasets

Gene based pathways were discovered with four different gene lists, two of them obtained from array-based dataset and remaining two is coming from sequencing-based dataset. Down-regulated genes in the array-based dataset affect 19 pathways whereas up-regulated genes in the array-based dataset affect 188 pathways. Down-regulated genes in sequencing-based dataset affect 193 pathways and up-regulated genes in sequencing-based dataset affect 188 pathways. Top three most significant pathways for each of these mentioned cases is shown in Table 4.7a, b, c, d. The most

affected pathways are, almost always, directly linked with cancer formation, such as P53 signaling pathway, cell cycle, ErbB signaling pathway.

P53 Signaling Pathway is one of the pathways responsible from cellular processes, cell growth and dead. Including the tumorigenesis process, several stress signals activates P53. The protein encoded by P53 is a transcriptional regulator whose activation yields with three main outputs: arresting cell cycle, cellular senescence or programmed cell death. The genes regulated by P53 have roles several major functions in the cell, including communication with neighbor cells, DNA damage repair mechanisms and feedbacks to P53 for inactivate or trigger its activity [153–155].

Pathway ID	Pathway	p-value
hsa04115	p53 signaling pathway	0.0065
hsa05212	Pancreatic cancer	0.0088
hsa04110	Cell cycle	0.0153

Table 4.7a: Top three most-significantly affected pathways by down-regulation on mRNA transcription, obtained from the gene transcription array-based dataset.

Pathway ID	Pathway	p-value
hsa04215	Apoptosis - multiple species	0. 0094
hsa05014	Amyotrophic lateral sclerosis (ALS)	0. 0227
hsa04621	NOD-like receptor signaling pathway	0.0019

Table 4.7b: Top three most-significantly affected pathways by up-regulation on mRNA transcription, obtained from the gene transcription array-based dataset.

Pathway ID	Pathway	p-value
hsa00190	Oxidative phosphorylation	3.6e-17
hsa05012	Parkinson's disease	9.9e-15
hsa04714	Thermogenesis	2.9e-14

Table 4.7c: Top three most-significantly affected pathways by down-regulation on mRNA transcription, obtained from the gene transcription sequencing-based dataset.

Pathway ID	Pathway	p-value
hsa04110	Cell cycle	5.6e-29
hsa03040	Spliceosome	5.5e-18
hsa03020	RNA polymerase	2.2e-03

Table 4.8d: Top three most-significantly affected pathways by up-regulation on mRNA transcription, obtained from the gene transcription sequencing-based dataset.

Cell cycle is another highly affected pathway both by up-regulated and down-regulated genes. Similar to the P53 Signaling Pathway, Cell Cycle pathway also belong to the class cellular processes and cell growth and death. The pathway is responsible for cell division by regulating transition between phases. The Cell Cycle pathway is also in control for response of DNA damage by promoting cell cycle arrest by up-regulating *P53* gene [143,153,156,157].

Apoptosis, in other words programmed cell death, is one of the fundamental mechanisms that cancer cells must control in order to keep growing and reproducing [56]. In healthy tissues, apoptosis oversees keeping cell numbers in balance and also eliminates the cells which can be potentially harmful to the organism, as well. The main group of actors of the apoptosis mechanism are caspases and their inhibitors, members of BCL-2 family of pro-apoptotic and anti-apoptotic proteins [158–160].

4.2.2 Candidate Pathways Obtained from DNA Methylation Datasets

Pathways affected by aberrant DNA methylation were discovered with four different gene lists, similar to the gene transcription-based pathways. Genes corresponding to the differentially methylated regions were annotated and used for pathway analysis. For pathway analysis, four different gene set were used. Two of them obtained from array-based dataset and remaining two is coming from bisulfite conversion sequencing-based dataset. Hyper-methylated genes in the array-based dataset affect 83 pathways whereas hypo-methylated genes in the array-based dataset affect 50 pathways. Hyper-methylated genes in the bisulfite conversion sequencing-based dataset affect 19 pathways and hypo-methylated genes in the bisulfite conversion sequencing-based dataset affect 91 pathways. Top three most significant pathways for each of these mentioned cases is shown in Table 4.8a, b, c, d. The most affected pathways are strongly associated with not only colorectal adenocarcinoma but also many other cancer types in the literature. In addition to Cell Cycle pathway, TNF signaling pathway and Central Carbon Mechanism in Cancer are observed as highly affected pathways by aberrant DNA methylation.

Tumor necrosis factor (TNF) signaling pathways is one of the most affected mechanisms by changes in DNA methylation patterns. TNF signaling pathway directly or indirectly touch various other significant signaling pathways. Some of these regulated pathways are apoptosis and cell survival, inflammation and immunity. TNF is a cytokine and in its activated form, it is found as a homotrimer bound with TNFR1 or TNFR2 [161]. TNFR1 signaling is responsible of activation of genes associated with apoptosis and necroptosis by interacting with NF-kappa ß pathway and MAPK cascade. Conversely, TNFR2 signaling is charged on cell survival by activating NF-kappa ß pathway and JNK pathway. Malfunction on a regulatory mechanism of cell survival and cell death is trivially linked with cancer formation [162,163].

Pathway ID	Pathway	p-value
hsa04668	TNF signaling pathway	2.6e-02
hsa04110	Cell cycle	2.8e-09
hsa05218	Melanoma	0.1e-09

Table 4.8a: Top three most-significantly affected pathways by hyper-methylation, obtained from the DNA methylation array-based dataset

Pathway ID	Pathway	p-value
hsa00190	Oxidative phosphorylation	7.3e-11
hsa05012	Parkinsons disease	1.4e-08
hsa04392	Hippo signaling pathway - multiple species	1.8e-06

Table 4.8b: Top three most-significantly affected pathways by hypo-methylation, obtained from the DNA methylation array-based dataset

Pathway ID	Pathway	p-value
hsa05230	Central carbon metabolism in cancer	0.0043
hsa00030	Pentose phosphate pathway	0.0193
hsa00052	Galactose metabolism	0.0207

Table 4.8c: Top three most-significantly affected pathways by hyper-methylation, obtained from the bisulfite conversion sequencing-based dataset

Pathway ID	Pathway	p-value
hsa05230	Central carbon metabolism in cancer	1.3e-03
hsa04730	Long-term depression	0.2e-6
hsa05219	Bladder cancer	2.6e-5

Table 4.8d: Top three most-significantly affected pathways by hypo-methylation, obtained from the bisulfite conversion sequencing-based dataset

4.2.3 Candidate Pathways Obtained from miRNA Transcription Dataset

Pathways affected by alterations on miRNA transcription were identified with twp different gene lists, similar to the DNA methylation-based or gene transcription-based pathways. Genes targeted by differentially expressed miRNAs were annotated and used for pathway analysis. For pathway analysis, two different gene set were used since there is only one sequencing-based miRNA transcription dataset was used.

One of them includes down-regulated miRNAs and their target genes whereas the other includes up-regulated miRNAs with their target genes. Up-regulated miRNAs fount to be enriched in 105 pathways where down-regulated miRNAs are associated with 193 pathways. 77 of these pathways are affected by both up-regulated and down-regulated miRNAs. Top three most significant pathways obtained for down-regulated miRNAs' target genes and up-regulated miRNAs' target genes are shown in Table 4.9a, b. The most affected pathways, similarly to the pathways obtained from gene expression and DNA methylation datasets, include fundamental mechanisms for cancer formation. Additional to the pathways that are responsible of fundamental mechanisms of cellular function, the main cancer pathways are observed, as well. Pancreatic cancer, viral carcinogenesis and chronic myeloid leukemia are the most affected cancerous pathways observed in the datasets.

Pathway ID	Pathway	p-value
hsa04110	Cell cycle	9.9e-17
hsa05212	Pancreatic cancer	7.7e-16
hsa05220	Chronic myeloid leukemia	7.4e-14

Table 4.9a: Top three most-significantly affected pathways by down-regulation, obtained from the miRNA transcription sequencing-based dataset

Pathway ID	Pathway	p-value
hsa05203	Viral carcinogenesis	0.1e-08
hsa04064	NF-kappa B signaling pathway	1.7e-03
hsa05322	Systemic lupus erythematosus	0.6e-04

Table 4.9b: Top three most-significantly affected pathways by up-regulation, obtained from the miRNA transcription sequencing-based dataset

4.2.4 Candidate Pathways Obtained with an Integrative Approach

As candidate gene list, only array-based DNA methylation, array-based expression and miRNA dataset were used, since sequencing-based dataset is used for predicting individual pathways.

There were 19 pathways found to be enriched by down-regulated mRNAs using the gene list obtained from annotated differentially methylation analysis of array-based dataset. 153 pathways are enriched by hyper-methylation and 105 pathways are affected by up-regulated miRNAs. 16 of these pathways were found to be enriched commonly in three cases (Table 4.10).

The pathways found to be affected by all three regulatory mechanisms are decided as "*candidate pathways*" and they are used for prediction of individualized pathways.

The p-values are combined with Fisher's combined probability test. Basically, it is used for combining probabilities obtained from independent tests answering the same hypothesis. Combining p-values are done in R environment using the `metap` package [164].

Pathway ID	mRNA	miRNA	DNA Methylation	Combined
hsa04110	1,53E-02	8,50E-11	2,75E-09	4,14E-04
hsa04115	6,50E-03	4,30E-03	6,53E-03	2,50E+09
hsa04137	3,70E-02	4,40E-02	1,09E-04	2,45E+09

hsa04218	3,27E-02	1,20E-04	5,61E-05	5,95E+05
hsa04934	1,40E-03	5,20E-03	2,50E-03	3,23E+08
hsa05161	5,80E-03	1,20E-06	1,33E-03	3,24E+05
hsa05162	4,96E-02	1,60E-04	1,44E-03	2,13E+08
hsa05166	6,60E-03	1,80E-02	1,30E-05	3,52E+07
hsa05168	1,23E-02	2,30E-03	4,24E-05	2,79E+07
hsa05203	5,00E-03	1,00E-07	5,00E-03	9,62E+04
hsa05212	8,80E-03	1,30E-06	4,82E-03	1,67E+06
hsa05214	4,41E-02	1,10E-06	7,56E-03	9,49E+06
hsa05218	4,54E-02	6,60E-04	1,01E-02	3,88E+09
hsa05220	9,10E-03	4,00E-04	8,63E-04	6,66E+07
hsa05222	1,67E-02	1,10E-04	1,67E-02	5,16E+08
hsa05223	3,81E-02	4,30E-03	3,28E-03	6,43E+09

Table 4.10: Common pathways across hyper-methylation, highly expressed miRNA and down-regulated mRNAs, with corresponding significance indicators (p-values)

The idea behind the Fisher's combined probability test is the fact that the probability of rejecting the null hypothesis is associated with probabilities of each individual test. Combined p-value is calculated as following, for the p-values obtained from Fisher's exact test:

$$X = -2\sum_{i=1}^{k} \ln(P_i)$$

The same idea applied for the up-regulated mRNAs, hypo-methylated gene regions and poorly expressed miRNAs, as well. 188 pathways are affected by up-regulated mRNAs, 160 pathways are affected by hypo-methylated regions and 193 pathways are affected by poorly transcribed miRNAs. 34 pathways are identified as affected by both up-regulation at gene level, down-regulation at miRNA level and hypo-methylation (Table 11). Similarly, p-values of these pathways are combined with Fisher's combined probability test.

Pathway ID	mRNA	miRNA	DNA Methylation	Combined
hsa05212	1,20E-04	1,80E-08	7,70E-16	3,21E-10
hsa05220	3,18E-02	3,10E-08	7,40E-14	9,94E-06
hsa05210	4,76E-02	1,80E-07	1,90E-11	1,63E-03
hsa05223	1,04E-03	5,60E-06	5,50E-09	2,42E+00
hsa05214	5,01E-03	2,30E-04	1e-08	6,37E+02
hsa05218	9,41E-03	3,70E-05	1,10E-08	2,23E+02
hsa03460	1,28E-02	9,40E-05	6,50E-04	1,89E+07
hsa04012	1,47E-02	1,10E-04	2,10E-04	8,56E+06
hsa05222	1,69E-02	1,10E-04	5,90E-07	4,54E+04
hsa04115	1,85E-02	1,30E-04	4,90E-06	4,12E+05
hsa05205	1,91E-02	1,50E-04	4e-06	3,88E+05
hsa05215	4,05E-02	2,30E-04	4,50E-06	1,34E+06
hsa04350	2,03E-02	2,70E-04	4,20E-05	6,17E+06
hsa04933	2,09E-02	1,20E-02	5,20E-06	3,08E+07
hsa05213	2,15E-02	6,00E-06	6e-06	2,28E+06
hsa04211	2,21E-02	9,30E-04	3,00E-03	9,67E+08
hsa04213	2,28E-02	4,70E-02	1,40E-02	1,12E+14
hsa04664	2,41E-02	1,60E-03	1,10E-04	8,81E+07
hsa04066	2,47E-02	1,10E-02	1,40E-03	5,01E+09
hsa05231	2,54E-02	5,20E-03	7,10E-04	1,40E+09
hsa04620	3,10E-02	1,70E-02	4,10E-02	1,51E+14
hsa05230	3,25E-02	7,70E-03	1,30E-02	3,06E+14
hsa05120	3,49E-02	1,00E-04	1,50E-02	4,61E+14
hsa04210	3,63E-02	8,80E-03	8,90E-03	2,72E+14
hsa04215	3,81E-02	1,70E-02	1,10E-02	6,19E+14
hsa04917	3,81E-02	2,40E-02	1,80E-02	1,22E+14
hsa04750	3,99E-02	2,20E-02	1,80E-02	1,14E+14
hsa04660	4,14E-02	2,40E-02	2,10e-02	1,45E+14
hsa04625	4,14E-02	3,00E-02	2,40E-02	1,96E+14
hsa04912	4,32E-02	2,61E-02	2,40E-02	1,81E+14

hsa04130	4,32E-02	3,41E-02	2,50E-02	2,33E+14
hsa05211	4,40E-02	3,02E-02	3,80E-02	3,02E+14
hsa04015	4,49E-02	3,11E-02	4,20E-02	3,41E+14
hsa05100	4,76E-02	4,23E-02	4,70E-02	5,04E+14

Table 4.11: Common pathways across hypo-methylation, poorly expressed miRNA and up-regulated mRNAs, with corresponding significance indicators (p-values)

Interestingly, 7 pathways detected as both affected by up-regulation of gene expression with supporting methylation and miRNA regulators and down-regulation of gene transcription with supporting methylation and miRNA regulators, as well (Table 4.12)

Pathway ID	**Pathway**	**pVal-up Reg.**	**pVal-down Reg.**
hsa05222	Small Cell Lung Cancer	4,54E+04	5,16E+08
hsa05212	Pancreatic Cancer	3,21E-10	1,67E+06
hsa05214	Glioma	6,37E+02	9,49E+06
hsa04115	P53 Signaling Pathway	4,12E+05	2,50E+09
hsa05218	Melanoma	2,23E+02	3,88E+09
hsa05220	Chronic Myeloid Leukemia	9,94E-06	6,66E+07
hsa05223	Non-small cell lung cancer	2,42E+00	6,43E+09

Table 4.12: Pathways which are affected by both up-regulated and down-regulated genes and their epigenetic regulators and corresponding combined p-values.

The pathways which are affected by both up-regulation and down-regulation on gene expression levels and their supportive epigenetic regulators are mainly cancer pathways. The only exception is the P53 Signaling Pathway; however, it is one of the most related pathways with cancer formation and progression.

4.3 Individual Affected Pathways

To reveal individually affected pathways, each patient in the cohort is analyzed as explained. The DNA methylation bisulfite conversion sequencing-based dataset has 2 patients, mRNA transcription sequencing-based dataset contains 9 patients and lastly, miRNA transcription sequencing-based dataset is made up of 8 patients. Individualized analyses are completed and reported for these patients, separately.

4.3.1 Individual Affected Pathways by Aberrant DNA Methylation

The bisulfite conversion sequencing-based dataset contains two patients, for the sake of simplicity these patients are mentioned as Patient 1 and Patient 2, obtained from [109] and [95] respectively.

In Patient 1, there is one pathway found to be enriched by hypo-methylated genes, details can be seen in Table 4.10 and the Cytokine-Cytokine Receptor Interaction pathway with effected genes marked is shown in Figure 4.3.

Pathway ID	Pathway	p-value	Hypo-methylated Genes
hsa04060	Cytokine-cytokine receptor interaction	0.019	TNFRSF8, TNFRSF14, TNFRSF1B

Table 4.10: Personal affected pathways for Patient 1, obtained from hypo-methylated genes

According to KEGG BRITE hierarchy, Cytokine-Cytokine Receptor Interaction Pathway belong to the class of *"Environmental Information Processing; Signaling Molecules and Interaction"*[165]. Cytokines are defined as small proteins which functions in cell signaling and with their releasing, they effect the behavior of cells around them [166]. In addition to their extracellular function, cytokines also have critical intracellular regulatory role in mechanisms such as cell growth, differentiation, cell death, angiogenesis.

In Patient 1, hyper-methylated genes affected 253 distinct pathways. The most significant three pathways are shown in Table 4.11.

Pathway ID	Pathway	p-value	Hyper-methylated Genes
hsa04146	Peroxisome	7.3e-33	*PEX14, PEX26, PEX13, PEX5, PEX7...*
hsa03050	Proteasome	2.9e-31	*PSMD3, PSMD11, PSMD7, PSMD13, PSMD14…*
hsa03040	Spliceosome	1.8e-23	*DDX46, SNRNP70, SNRPA, DHX38, CDC40…*

Table 4.11: Top three most significant pathways affected by hyper-methylated genes, in Patient 1.

Peroxisomes are the small organelles with an envelope membrane and they contain enzymes which are used in various metabolic mechanisms. Peroxisomes' major role is to join fatty acid oxidation, resulting in conversion of long chain fatty acids into medium chain fatty acids and then they are sent to mitochondria for oxidation [167,168]. Insufficiencies of peroxisomes are linked with several diseases cause to severe and mostly lethal disorders [169].

The second and the third most significantly affected pathways of Patient 2 are Proteasome and Spliceosome Pathways. These two pathways are in the same class according to BRITE hierarchy: Genetic Information Processing. Proteasome taking part in *"Folding, Sorting and Degradation*" subclass whereas Spliceosome belongs to "*Transcription*" subclass [170,171]. Proteasome is a large protein composite which is mainly in charge of degradation of intracellular proteins. Additionally, it takes part in various significant cellular metabolism mechanisms including cell cycle, stress signaling, inflammatory responses and programmed cell death [172]. Spliceosome is a ribonucleoprotein (RNP) complex made up of five snRNPs and several proteins and

it is the catalyzer of pre-mRNA splicing. This pathway is mainly responsible of production of variable forms of mRNA from a single pre-mRNA [173].

When hyper-methylated and hypo-methylated genes are taken as input of pathway analysis, the top-three significantly affected pathways remain the same as in the case of only hyper-methylated genes are taken into consideration. However, their significance levels (p-values) are changed and hypo-methylated genes' affects are seen, as well (Table 4.12). When these affected genes are placed into the corresponding pathway's map, the failure is seen more clearly and comprehensively.

In Patient 2, 237 pathways are affected by hyper-methylation whereas 179 pathways are affected by hypo-methylation. 172 of these pathways are commonly affected by both hyper-methylation and hypo-methylation. The top three most affected pathways by hyper-methylation and hypo-methylation are shown in Table 4.13 and 4.14, respectively.

Pathway ID	**Pathway**	**p-value**	**Hyper-meth. Genes**	**Hypo-meth. Genes**
hsa04146	Peroxisome	9.1e-37	*PEX14, PEX26, PEX13…*	*PEX7, PEX11B, ABCD3…*
hsa03050	Proteasome	2.3e-37	*PSMD14, PSMD8, PSMD4…*	*PSMD3, PSMD11, PSMD7…*
hsa03040	Spliceosome	1.3e-25	*DDX46, SNRPA, DHX38…*	*SNRNP70, SNRPC, CDC40…*

Table 4.12: Top three most significantly affected pathways by aberrant DNA methylation in Patient 1.

Pathway ID	Pathway	p-value	Hyper-methylated Genes
hsa03020	RNA polymerase	0.5e-20	*PEX14, PEX26, PEX13, PEX5, PEX7...*
hsa04130	SNARE interactions in vesicular transport	4.3e-17	*STX18, BET1, GOSR1, SEC22B, GOSR2…*
hsa04120	Ubiquitin mediated proteolysis	1.5e-15	*SAE1, SKP2, FBXW7, RHOBTB1, RHOBTB2…*

Table 4.13: Top three most significantly affected pathways by hyper-methylation in Patient 2.

RNA Polymerase pathway is from *"Genetic Information Processing, Transcription"* class according to BRITE hierarchy [174]. RNA polymerases are responsible of transcription of all kinds of RNA by physically attaching promoter region of the DNA and moves from 3' end to 5' end of the template strand[175]. Dysregulation on RNA polymerase pathway is linked with various outcomes such as reduced rDNA transcription, reduced genomic stability and consequently abnormalities in growth, various diseases including cancer [176].

Pathway ID	Pathway	p-value	Hyper-methylated Genes
hsa04974	Protein digestion and absorption	6.5e-11	*SLC7A8, PRSS3, ATP1A1, ATP1B2, KCNQ1...*
hsa04350	TGF-beta signaling pathway	8.5e-11	*SMAD3, SMAD6, CDKN2B, MYC, INHBA…*
hsa05200	Pathways in cancer	1.4e-10	*DCC, CDH1, WNT1, WNT6, WNT16…*

Table 4.14: Top three most significantly affected pathways by hypo-methylation in Patient 2.

In order to detect which of these affected pathways are players in cancer formation, they are compared with the pathways obtained with pathways obtained from array-based datasets.

4.3.2 Individual Affected Pathways by Changes on mRNA Expression Levels

In the gene transcription sequencing-based dataset there are 9 patients available and they are used for personalized pathway prediction. Firstly, differentially expressed genes for each of these patients are called out as explained in Section 3.7.2. After differentially expressed genes and their transcription level change values are obtained, pathway analysis is applied with these values, for each patient. The pathways are identified as it is explained in Section 3.5. For each patient, affected pathways are collected and the ones included in previously created database are reported as players in cancer formation.

According to pathway analysis, top three most-effected pathways are shown in the Table 4.15 for each patient for up-regulated genes. Even though only the most affected pathways are represented in this Section, there are many more pathways are identified and a filtering should be applied in order to obtain more relevant information and possible treatment targets.

As individualized leading mechanisms to cancer, these affected pathways are filtered out if they are observed in array-based datasets or not. Applying such a filter provides a chance of classifying cancers by their affected pathways.

Patient	Pathway ID	Pathway	p-value	Up-regulated Genes
1	hsa05219	Bladder cancer	6,36E-05	*MMP1, CXCL8, CDKN2A, MYC*
1	hsa04512	ECM-receptor interaction	2,15E-04	*COL1A1, COL4A1, COL9A3, COMP*
1	hsa04974	Protein digestion and absorption	2,86E-04	*SLC7A9, PRSS2, SLC9A3, MME, COL1A1...*

2	hsa04610	Complement and coagulation cascades	4,83E-05	*SERPIND1, SERPINE1, PLAU*
2	hsa03440	Homologous recombination	7,31E-03	*XRCC2*
2	hsa05144	Malaria	1,05E-02	*MET, COMP, THBS2, CXCL8*
3	hsa05219	Bladder cancer	3,51E-06	*MMP1, CXCL8, CDKN2A, MYC*
3	hsa04218	Cellular senescence	1,13E-05	*MYC, CDKN2A, IL1A, CXCL8*
3	hsa00240	Pyrimidine metabolism	1,02E-04	*RRM2, POLR3G, NME1*
4	hsa04610	Complement and coagulation cascades	4,83E-05	*FGB, PLAU, C2, ITGAX, CFB*
4	hsa04621	NOD-like receptor signaling pathway	4,74E-04	*NOD2, IL1B, PLCB4, OAS2, CXCL8, CXCL1*
4	hsa05219	Bladder cancer	2,32E-03	*MMP1, MMP9, CXCL8, MYC*
5	hsa05219	Bladder cancer	6,11E-03	*VEGFA, MYC, CCND1*
5	hsa04657	IL-17 signaling pathway	3,81E-02	*FOSL1, TRAF5, MAPK15, CXCL5*
5	hsa04064	NF-kappa B signaling pathway	3,97E-02	*TRAF5, PLCG1*
6	hsa03440	Homologous recombination	7,31E-03	*RAD54B*
6	hsa05132	Salmonella infection	3,25E-02	*KLC3*
6	hsa04115	p53 signaling pathway	2,03E-02	*TP73, SERPINB5*
7	hsa04392	Hippo signaling pathway - multiple species	7,99E-06	*STK3, TEAD4, AJUBA*
7	hsa04512	ECM-receptor interaction	2,15E-04	*COL1A1, COL4A1, COL9A3, SPP1...*
7	hsa04974	Protein digestion and absorption	2,86E-04	*MME, COL1A1, COL4A1, COL7A1, COL9A3...*

8	hsa04062	Chemokine signaling pathway	6,34E-04	*CXCL1, CXCL9, CXCL11, CXCL5*
8	hsa04060	Cytokine-cytokine receptor interaction	1,98E-03	*IL1B, AMH, INHBA, EDAR, LIF, IL11...*
8	hsa05144	Malaria	1,05E-02	*ICAM1, HBB, IL1B*
9	hsa04218	Cellular senescence	4,39E-06	*MYC, CDKN2A, CCNE1, MYBL2, CDK1...*
9	hsa04062	Chemokine signaling pathway	7,88E-06	*CXCL1, CXCL2, CXCL8, CXCL9, PF4...*
9	hsa04060	Cytokine-cytokine receptor interaction	3,61E-05	*BMP7, INHBA, TGFB2, CSF2, OSM, IL11...*

Table 4.15: Top-three most affected pathways by up-regulated genes, for each patient in the dataset.

After filtering out the pathways which are common with the 34 selected candidate pathways, number of individualized pathways are drastically reduced. The pathways found for each patient in the gene expression sequencing-based dataset is shown in Table 4.16. Patient IDs are converted into integers from TCGA IDs (the conversion table can be found in Supplementary Materials). For gene expression level changes, 16 pathways in Patient 1, 6 pathways in Patient 2, 6 pathways in Patient 3, 10 pathways in Patient 4, 21 pathways are Patient 5, 5 pathways in Patient 6, 8 pathways in Patient 7, 8 pathways in Patient 8 and 15 pathways in Patient 9 are found to be enriched.

Pathway ID	**Pathways**	**Patient IDs**
hsa05210	Colorectal cancer	1, 2, 4, 5, 7, 9
hsa05213	Endometrial cancer	1, 4, 5, 6, 7, 9
hsa04012	ErbB signaling pathway	1, 2, 3, 4, 5, 7
hsa04350	TGF-beta signaling pathway	1, 3, 4, 6, 7, 9
hsa04933	AGE-RAGE signaling pathway in diabetic complications	1, 4, 5, 8, 9

hsa05214	Glioma	1, 5, 7, 8, 9
hsa04115	p53 signaling pathway	1, 3, 6, 8, 9
hsa05230	Central carbon metabolism in cancer	1, 2, 5, 9
hsa05120	Epithelial cell signaling in Helicobacter pylori infection	4, 5, 7, 9
hsa03460	Fanconi anemia pathway	1, 2, 6, 8
hsa05218	Melanoma	1, 5, 8, 9
hsa05223	Non-small cell lung cancer	1, 5, 8, 9
hsa05212	Pancreatic cancer	1, 5, 8, 9
hsa04215	Apoptosis - multiple species	3, 4, 9
hsa05220	Chronic myeloid leukemia	1, 3, 5
hsa04066	HIF-1 signaling pathway	2, 5, 9
hsa05205	Proteoglycans in cancer	1, 5, 9
hsa04620	Toll-like receptor signaling pathway	4, 6, 8
hsa05100	Bacterial invasion of epithelial cells	1, 7
hsa04750	Inflammatory mediator regulation of TRP channels	1, 5
hsa04917	Prolactin signaling pathway	5, 7
hsa05215	Prostate cancer	2, 4
hsa04130	SNARE interactions in vesicular transport	5, 9
hsa04210	Apoptosis	3
hsa04625	C-type lectin receptor signaling pathway	4
hsa05231	Choline metabolism in cancer	5
hsa04664	Fc epsilon RI signaling pathway	5
hsa04015	Rap1 signaling pathway	5
hsa05222	Small cell lung cancer	5
hsa04660	T cell receptor signaling pathway	5

Table 4.16: Pathways found to be affected by up-regulation in gene expression level individually for each person

Colorectal Cancer, Endometrial Cancer, ErbB Signaling and TGF-beta Signaling pathways are most commonly affected pathways across the cohort. These pathways

are observed in 6 out of 9 patients and they are strongly linked with colorectal adenocarcinoma formation, as it is mentioned in previous sections.

Similar approach is applied for down-regulated genes and their epigenetic supporters as well. Table 4.17 and 4.18 shows

Patient	Pathway ID	Pathway	P-Value	Down-regulated Genes
1	hsa04530	Tight junction	5,61E-06	*TIAM1, IGSF5, PRKACB, CLDN5, CLDN8...*
1	hsa00480	Glutathione metabolism	6,75E-05	*NAT8B, GSTA1, GSTM1, GSTM5, HPGDS*
1	hsa04721	Synaptic vesicle cycle	9,67E-05	*CACNA1A, SNAP25, SLC18A1*
2	hsa04514	Cell adhesion molecules (CAMs)	2,79E-06	*CD22, CNTN2, CNTN1, NLGN1, NRXN1...*
2	hsa04213	Longevity regulating pathway - multiple species	2,31E-05	*ADCY5, PRKAA2, CRYAB*
2	hsa03320	PPAR signaling pathway	1,45E-04	*PCK1, PLIN4, CD36, FABP4, FABP1...*
3	hsa00480	Glutathione metabolism	3,06E-07	*ANPEP, GGT7, GPX3, GSTM1, GSTM2...*
3	hsa00982	Drug metabolism - cytochrome P450	7,19E-07	*FMO2, FMO5, CYP2B6, GSTM1, GSTM2...*
3	hsa04662	B cell receptor signaling pathway	8,08E-07	*AKT3, PRKCB, VAV3, CR2, CD19...*
4	hsa04713	Circadian entrainment	8,72E-05	*NOS1, PRKCB, CACNA1H, RYR3...*
4	hsa04514	Cell adhesion molecules (CAMs)	2,91E-04	*MPZ, CNTN2, CNTN1, NLGN1, NRXN3...*

4	hsa04915	Estrogen signaling pathway	1,70E-03	*SP1, ADCY5, HSPA1A, HSPA2, GNAI1...*
5	hsa00730	Thiamine metabolism	2,78E-07	*ALPI, ALPPL2*
5	hsa00790	Folate biosynthesis	1,59E-06	*ALPI, ALPPL2*
5	hsa05130	Pathogenic Escherichia coli infection	6,39E-05	*TUBA1A, TUBAL3, TUBB4A, TUBB2B*
6	hsa04512	ECM-receptor interaction	5,41E-05	*LAMA1, COL4A5, COL4A6, COL9A2...*
6	hsa04640	Hematopoietic cell lineage	9,00E-05	*CD36, IL1R2, ANPEP, FCER2, CR2, MS4A1...*
6	hsa04662	B cell receptor signaling pathway	1,39E-04	*CARD11, CR2, CD19, FOS, PLCG2, CD79B...*
7	hsa04662	B cell receptor signaling pathway	3,49E-05	*CR2, CD79A*
7	hsa04612	Antigen processing and presentation	4,47E-05	*KIR2DL4, CD8A*
7	hsa04920	Adipocytokine signaling pathway	1,28E-04	*PCK1, PRKAA2, ADIPOQ, RXRG, CD36...*
8	hsa03320	PPAR signaling pathway	3,64E-05	*PLIN4, CD36, FABP4, FABP1, FABP3...*
8	hsa04371	Apelin signaling pathway	2,56E-04	*GNG4, GNG11, MYLK, MYL3, AGTR1*
8	hsa00790	Folate biosynthesis	2,90E-03	*AKR1B10, AKR1C3*
9	hsa04630	Jak-STAT signaling pathway	1,68E-03	*CNTFR, GHR, IL5RA, IL6R, IL10RA...*
9	hsa04662	B cell receptor signaling pathway	3,49E-05	*CR2, PLCG2, CD79A*
9	hsa04060	Cytokine-cytokine receptor interaction	7,82E-03	*LIFR, IL1R2, TNFRSF13B, TNFRSF17...*

Table 4.17: Top-three most affected pathways by down-regulated genes, for each patient in the dataset.

After calling out individually affected pathways, they are filtered out if they are intersecting with 16 candidate pathways. Similar to the up-regulated genes and their pathways, again the number of individually affected pathways are decreased. 4 pathways are found as enriched for Patients 1 and 8, 2 pathways are found to be affected for Patients 3, 5, 6, and 7, and 1 pathway is found for Patients 2, 4 and 9 (Table 4.18).

Pathway ID	**Pathways**	**Patient IDs**
hsa05222	Small cell lung cancer	6, 7, 8, 9
hsa05223	Non-small cell lung cancer	1, 7, 8
hsa05214	Glioma	3, 8
hsa05166	HTLV-I infection	3, 6
hsa05218	Melanoma	5, 8
hsa04137	Mitophagy- animal	4, 5
hsa05212	Pancreatic cancer	1, 2
hsa05161	Hepatitis B	1
hsa04115	P53 signaling pathway	1

Table 4.18: Pathways found to be affected by down-regulation in gene expression level individually for each person

5 DISCUSSION

In this dissertation, a predictive method was proposed for the detection of individual affected pathways in colorectal adenocarcinoma patients. To achieve this goal, an integrative approach was used for observing regulatory roles of epigenetic mechanisms as well as the gene level changes. Since there are many affected genes and mechanisms in cancerous cells, it fundamental to understand which pathways the leaders are. In this study, we used miRNA, DNA methylation and gene expression datasets all together and by merging results from these distinct levels of regulation, gained a comprehensive understanding.

In the Discussion Section, technical preferences with their reasonings were clarified, side-results were briefly explained, and their biological denotations were discussed.

5.1 Multi-Omics Data Analysis and Integration

For all steps of analysis and integration, several thresholds were set, and methods were chosen over another. This section clarifies these technical details briefly.

As a significance threshold, p-value = 0.05 was defined for all types of analysis in the dissertation, unless otherwise noted. As DNA methylation difference threshold $|\Delta ß|=0.20$ is set, for array-based and bisulfite conversion sequencing-based dataset. Previous studies show that, all aberrations on DNA methylation does not affect the down-stream mechanisms [100]. For miRNA transcription and gene transcription datasets $|\log_2 FC| = 1.00$ is used as a differentially expression threshold, meaning that the genes (or miRNAs) whose expression level is doubled, at least, are considered as up-regulated and the genes whose expression level is halved at least are defined as down-regulated.

Additionally, through the analysis of DNA methylation microarray-based dataset, a minimum probes threshold was set 3 probes. That is, to define a region as "differentially methylated"; in that region at least 3 probes should be differentially methylated with the previously mentioned p-value and Δß thresholds. Default application for the probe number threshold is in-between 3 and 7 probes [106]. In the

dissertation, minimum number selected as 3, so that we can catch the changes occurred in small regions, since they can hit the promoter or enhancer binding regions and might affect the expression, as well. As a minimum size of a DMR, 50 base-pair long fragments were selected, meaning that if a region is differentially methylated however short than 50 base-pairs, then it will not be considered for further analysis.

Excluding the individualized analysis, adjusted p-value was used as the significance indicator for the thesis. Since for each gene (or each miRNA or each probe) statistical hypothesis testing was applied to detect if that gene is differentiated between two cases (tumor and normal in this study). Usage of multiple hypothesis testing causes a bias on Type I error rate, by causing an increase on change of a rare event. In this dissertation, Bonferroni method is used for correcting this bias. Bonferroni method simply adjusts the significance level by dividing number of tests applied. In other words, set $\alpha=0.05$ as the desired significance level. Bonferroni tests the hypothesis for a significance level $p = \frac{\alpha}{m}$ [177].

For gene and miRNA expression analysis, change level threshold is used with log-formation. For example, assume that gene X is read 15 times in normal tissue and 150 times in cancerous tissue, and gene Y is read 1000 times in normal tissue and 1500 times in cancerous tissue. Clearly, Y gene's expression level is 10-times higher than X gene. However, assuming that Y gene is 10-times more affective in cell function compared to X gene would not be correct. The biologically correct assumption here is, X gene needs to be 15 copies in the cell for its normal function, however it was 10-times more expressed in tumorous tissue. Y gene is required to be expressed 1000 copies in the healthy cells whereas it has 1500 copies in tumor cells, so it has 1,5 times more expressed in tumorous tissue. Expression level does not indicate the effect on regular cell function all the time [178]. Therefore, logFC formation is used for both gene and miRNA expression datasets.

5.2 Pathway Analysis and Integration

In this study, the aim was to propose a new method to predict individualized pathways in colorectal adenocarcinoma patients. 2 patients' whole-genome-bisulfite-sequencing data is used as test-set for predicting individual pathways from DNA

methylation information. 9 patients' gene expression dataset obtained with RNA-Seq is used as test-set for predicting personalized pathways affected in gene expression level.

Additionally, miRNA expression sequencing dataset of 8 patients, DNA methylation array-based dataset of 38 patients, gene expression array-based dataset of 132 tumor and 9 normal samples are used for creating a candidate pathways database.

In this study, we used the pathways which are affected by both gene expression, miRNA expression and DNA methylation levels. However, usage of pathways affected by aberrant DNA methylation as candidates for predicting DNA methylation driven mechanisms can be another approach. Similarly, pathways affected at gene-expression level can be used as predictors of personalized pathways at gene expression level and same approach can be used for miRNA expression levels and miRNA-level affected pathways, as well.

Pathway analysis are completed with a greedy-based algorithm. This algorithm is not promised to give the optimal solution, in other words, globally optimal solution. However, to overcome this issue, the active sub-network search step is repeated 10 times.

5.3 Candidate Pathways

Candidate pathways are selected from a set of pathways affected by aberrant DNA methylation, changes in mRNA transcription levels and differentially expressed miRNAs. In this study, an integrative approach is proposed at pathway level. The DNA methylation changes might not be directly affected on a gene, however it can affect the same pathway with that particular gene and with this approach it is aimed to catch such regulatory mechanisms, as well.

As a result of this approach, 16 pathways are identified as candidate pathways for down-regulated genes. These 16 pathways are also affected by hyper-methylation and up-regulation in regulatory miRNAs (Table 5.1).

The pathways selected are mostly related with Cell Cycle or they are directly Cancer pathways. The interesting thing is that, two lung cancer related (small cell lung

cancaer and non-small cell lung cancer) pathways are found as enriched in colorectal adenocarcinoma patients.

Pathway ID	Pathway
hsa04115	p53 signaling pathway
hsa05212	Pancreatic cancer
hsa04110	Cell cycle
hsa04934	Cushing's syndrome
hsa04218	Cellular senescence
hsa05203	Viral carcinogenesis
hsa05161	Hepatitis B
hsa05166	HTLV-I infection
hsa05220	Chronic myeloid leukemia
hsa05168	Herpes simplex infection
hsa05222	Small cell lung cancer
hsa04137	Mitophagy - animal
hsa05223	Non-small cell lung cancer
hsa05214	Glioma
hsa05218	Melanoma
hsa05162	Measles

Table 5.1: 16 selected candidate pathways from down-regulation in gene expression, supported with epigenetic mechanisms.

In literature, it is shown that lung cancer metastasis is commonly seen in primary colorectal cancer. Therefore, it can be assumed that the cohort includes patients with possible metastases in lung, as well [109]. Therefore, it can be concluded that our approach seems biologically meaningful, as well.

34 pathways are selected as candidate pathways via up-regulation of mRNAs, with support of pathways regulated by epigenetic regulators (Table 5.2). Most of these

pathways are directly linked with cancer generally or colorectal cancer more specifically and they are already explained previously.

Pathway ID	Pathway
hsa05120	Epithelial cell signaling in Helicobacter pylori infection
hsa04012	ErbB signaling pathway
hsa04620	Toll-like receptor signaling pathway
hsa04210	Apoptosis
hsa04130	SNARE interactions in vesicular transport
hsa04215	Apoptosis - multiple species
hsa03460	Fanconi anemia pathway
hsa05213	Endometrial cancer
hsa04213	Longevity regulating pathway - multiple species
hsa05230	Central carbon metabolism in cancer
hsa05223	Non-small cell lung cancer
hsa04664	Fc epsilon RI signaling pathway
hsa04115	p53 signaling pathway
hsa05211	Renal cell carcinoma
hsa05214	Glioma
hsa04917	Prolactin signaling pathway
hsa05218	Melanoma
hsa05100	Bacterial invasion of epithelial cells
hsa05212	Pancreatic cancer
hsa05220	Chronic myeloid leukemia
hsa04350	TGF-beta signaling pathway
hsa05210	Colorectal cancer
hsa04211	Longevity regulating pathway
hsa05205	Proteoglycans in cancer
hsa05222	Small cell lung cancer
hsa04912	GnRH signaling pathway
hsa04015	Rap1 signaling pathway

hsa04933	AGE-RAGE signaling pathway in diabetic complications
hsa04066	HIF-1 signaling pathway
hsa04660	T cell receptor signaling pathway
hsa04625	C-type lectin receptor signaling pathway
hsa05231	Choline metabolism in cancer

Table 5.2: 34 selected candidate pathways from up-regulation in gene expression, supported with epigenetic mechanisms.

6 CONCLUSION

In this study, the main hypothesis was cancer formation is an individual process with shared mechanisms. In order to prove this idea, an integrative approach was developed to understand the genetic and epigenetic mechanisms lead to cancer. Previously, several studies showed the systems biology approach is much more powerful the approaches focusing only gene-level changes. Additionally, there are various methods for integrating epigenetic changes with gene expression level changes into the whole pathway analysis. However, in this thesis we are suggesting a new method for predicting individual mechanisms causing to cancer with comparing pathway-based changes in cancer. Consequently, the two main contributions of this work are, investigation of common regulatory mechanisms to understand development of cancer, and to predict patient-specific mechanisms in cancer formation.

Pathway analysis showed that, pathways related to cell cycle are the most affected mechanisms both in epigenetic and genetic mechanisms. Even though it is expected by the nature of cancer itself, revealing each pathway in a personalized manner is important for understanding side-mechanisms and providing to patients an optimal treatment.

According to our investigations, 34 pathways are affected by up-regulation of genes and the epigenetic mechanisms leading to up-regulation. On the other hand, 16 pathways are affected by down-regulation at the gene level with supportive epigenetic mechanisms. These two cases share 7 pathways in common and 6 of them are cancer pathways.

The pathway-based approach used here, is more powerful compared to the traditional gene-based comparison approach, since it provides a more comprehensive understanding both in nature of cancer and predicting personalized mechanisms.

7 SUPPLEMENTARY MATERIAL

All supplementary material and codes are uploaded to figshare platform and can be reached via:

https://figshare.com/projects/Integration_of_Multi-Omics_Data_for_Predicting_Individual_Colon_Cancer_Aetiology/56948

8 BIBLIOGRAPHY

[1] Waddington C 1968 Towards a Theoretical Biology *Nature* **218** 525–7

[2] Felsenfeld G 2014 A brief history of epigenetics. *Cold Spring Harb. Perspect. Biol.* **6**

[3] Dupont C, Armant D R and Brenner C A 2009 Epigenetics: definition, mechanisms and clinical perspective. *Semin. Reprod. Med.* **27** 351–7

[4] Craig Johanna 2008 Complex Diseases: Research and Applications *Nat. Educ.* **1** 184

[5] Mitchell K J 2012 What is complex about complex disorders? *Genome Biol.* **13** 237

[6] Weissfeld J L, Lin Y, Lin H-M, Kurland B F, Wilson D O, Fuhrman C R, Pennathur A, Romkes M, Nukui T, Yuan J-M, Siegfried J M and Diergaarde B 2015 Lung Cancer Risk Prediction Using Common SNPs Located in GWAS-Identified Susceptibility Regions *J. Thorac. Oncol.* **10** 1538–45

[7] Low S-K, Takahashi A, Mushiroda T and Kubo M 2014 Genome-wide association study: a useful tool to identify common genetic variants associated with drug toxicity and efficacy in cancer pharmacogenomics. *Clin. Cancer Res.* **20** 2541–52

[8] Haines J L, Hauser M A, Schmidt S, Scott W K, Olson L M, Gallins P, Spencer K L, Kwan S Y, Noureddine M, Gilbert J R, Schnetz-Boutaud N, Agarwal A, Postel E A and Pericak-Vance M A 2005 Complement factor H variant increases the risk of age-related macular degeneration. *Science* **308** 419–21

[9] Han Z, Qu J, Zhao J and Zou X 2018 Analyzing 74,248 Samples Confirms the Association Between CLU rs11136000 Polymorphism and Alzheimer's Disease in Caucasian But Not Chinese population *Sci. Rep.* **8** 11062

[10] Xue A, Wu Y, Zhu Z, Zhang F, Kemper K E, Zheng Z, Yengo L, Lloyd-Jones

L R, Sidorenko J, Wu Y, McRae A F, Visscher P M, Zeng J and Yang J 2018 Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes *Nat. Commun.* **9** 2941

[11] Rakyan V K, Down T A, Balding D J and Beck S 2011 Epigenome-wide association studies for common human diseases *Nat. Rev. Genet.* **12** 529–41

[12] Chadwick L H, Sawa A, Yang I V., Baccarelli A, Breakefield X O, Deng H-W, Dolinoy D C, Fallin M D, Holland N T, Houseman E A, Lomvardas S, Rao M, Satterlee J S, Tyson F L, Vijayanand P and Greally J M 2015 New insights and updated guidelines for epigenome-wide association studies *Neuroepigenetics* **1** 14–9

[13] Carrasco-Ramiro F, Peiró-Pastor R and Aguado B 2017 Human genomics projects and precision medicine *Gene Ther.* **24** 551–61

[14] Anon Personalized Medicine: Redefining Cancer and Its Treatment

[15] Yan L, Rosen N and Arteaga C 2011 Targeted cancer therapies *Chin. J. Cancer*

[16] Røsland G V and Engelsen A S T 2015 Novel Points of Attack for Targeted Cancer Therapy *Basic Clin. Pharmacol. Toxicol.* **116** 9–18

[17] Hauschild A, Grob J-J, Demidov L V, Jouary T, Gutzmer R, Millward M, Rutkowski P, Blank C U, Miller W H, Kaempgen E, Martín-Algarra S, Karaszewska B, Mauch C, Chiarion-Sileni V, Martin A-M, Swann S, Haney P, Mirakhur B, Guckert M E, Goodman V and Chapman P B 2012 Dabrafenib in BRAF-mutated metastatic melanoma: a multicentre, open-label, phase 3 randomised controlled trial *Lancet* **380** 358–65

[18] Yuasa Y, Nagasaki H, Akiyama Y, Hashimoto Y, Takizawa T, Kojima K, Kawano T, Sugihara K, Imai K and Nakachi K 2009 DNA methylation status is inversely correlated with green tea intake and physical activity in gastric cancer patients *Int. J. Cancer* **124** 2677–82

[19] Kurkjian C, Kummar S and Murgo A J 2008 DNA methylation: its role in

cancer development and therapy. *Curr. Probl. Cancer* **32** 187–235

[20] Yoo C B and Jones P A 2006 Epigenetic therapy of cancer: past, present and future *Nat. Rev. Drug Discov.* **5** 37–50

[21] Kuipers E J, Grady W M, Lieberman D, Seufferlein T, Sung J J, Boelens P G, Van De Velde C J H and Watanabe T 2015 Colorectal cancer *Nat. Rev. Dis. Prim.* **1** 1–25

[22] Jia Y and Guo M 2013 Epigenetic changes in colorectal cancer *Chin. J. Cancer* **32** 21–30

[23] Heyn H, Vidal E, Ferreira H J, Vizoso M, Sayols S, Gomez A, Moran S, Boque-Sastre R, Guil S, Martinez-Cardus A, Lin C Y, Royo R, Sanchez-Mut J V., Martinez R, Gut M, Torrents D, Orozco M, Gut I, Young R A and Esteller M 2016 Epigenomic analysis detects aberrant super-enhancer DNA methylation in human cancer *Genome Biol.* **17** 11

[24] Saini A, Mastana S, Myers F and Lewis M P 2013 "From Death, Lead Me to Immortality" - Mantra of Ageing Skeletal Muscle *Curr. Genomics* **14** 256–7

[25] Moore L D, Le T and Fan G 2013 DNA methylation and its basic function. *Neuropsychopharmacology* **38** 23–38

[26] Jin B, Li Y and Robertson K D 2011 DNA methylation: superior or subordinate in the epigenetic hierarchy? *Genes Cancer* **2** 607–17

[27] Hill P W S, Amouroux R and Hajkova P 2014 DNA demethylation, Tet proteins and 5-hydroxymethylcytosine in epigenetic reprogramming: An emerging complex story *Genomics* **104** 324–33

[28] Jabbari K and Bernardi G 2004 Cytosine methylation and CpG, TpG (CpA) and TpA frequencies *Gene* **333** 143–9

[29] Lander E S, Linton L M, Birren B, Nusbaum C, Zody M C, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A,

Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov J P, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin J C, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston R H, Wilson R K, Hillier L W, McPherson J D, Marra M A, Mardis E R, Fulton L A, Chinwalla A T, Pepin K H, Gish W R, Chissoe S L, Wendl M C, Delehaunty K D, Miner T L, Delehaunty A, Kramer J B, Cook L L, Fulton R S, Johnson D L, Minx P J, Clifton S W, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng J-F, Olsen A, Lucas S, Elkin C, et al 2001 Initial sequencing and analysis of the human genome *Nature* **409** 860–921

[30] Gardiner-Garden M and Frommer M 1987 CpG Islands in vertebrate genomes *J. Mol. Biol.* **196** 261–82

[31] Asmar F, Søgaard A and Grønbæk K 2015 DNA Methylation and Hydroxymethylation in Cancer *Epigenetic Cancer Ther.* 9–30

[32] Deaton A M and Bird A 2011 CpG islands and the regulation of transcription. *Genes Dev.* **25** 1010–22

[33] Moore L D, Le T and Fan G 2012 DNA Methylation and Its Basic Function *Neuropsychopharmacology* **38**

[34] Mitchell C, Schneper L M and Notterman D A 2016 DNA methylation, early life environment and health outcomes *Pediatr. Res.* **79** 212–9

[35] Bergman Y and Cedar H 2013 DNA methylation dynamics in health and disease *Nat. Struct. Mol. Biol.* **20** 274–81

[36] Weber M, Hellmann I, Stadler M B, Ramos L, Pääbo S, Rebhan M and Schübeler D 2007 Distribution, silencing potential and evolutionary impact of

promoter DNA methylation in the human genome *Nat. Genet.* **39** 457–66

[37] Choy M-K, Movassagh M, Goh H-G, Bennett M R, Down T A and Foo R S-Y 2010 Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated *BMC Genomics* **11** 519

[38] Kriaucionis S and Bird A DNA methylation and Rett syndrome

[39] Waddington C H 1957 *The Strategy of The Genes* (Bristol: J W Arrowsmith Ltd.)

[40] Philips T and Lobo I 2008 Genetic Imprinting and X Inactivation *Nat. Educ.* **1** 117

[41] Biliya S and Bulla L A 2010 Experimental Biology and Medicine Minireview Genomic imprinting : the influence of differential methylation

[42] Tahara T and Arisawa T 2015 DNA methylation as a molecular biomarker in gastric cancer *Epigenomics* **7** 475–86

[43] Feinberg A P and Vogelstein B 1983 Hypomethylation distinguishes genes of some human cancers from their normal counterparts *Nature* **301** 89–92

[44] Sharma S, Kelly T K and Jones P A 2010 Epigenetics in cancer *Carcinogenesis* **31** 27–36

[45] Hanahan D and Weinberg R A 2011 Hallmarks of Cancer: The Next Generation *Cell* **144** 646–74

[46] Alimandi M, Wang L M, Bottaro D, Lee C C, Kuo A, Frankel M, Fedi P, Tang C, Lippman M and Pierce J H 1997 Epidermal growth factor and betacellulin mediate signal transduction through co-expressed ErbB2 and ErbB3 receptors *EMBO J.* **16** 5608–17

[47] Hanahan D and Weinberg R A 2000 The hallmarks of cancer. *Cell* **100** 57–70

[48] Cai H, An Y, Chen X, Sun D, Chen T, Peng Y, Zhu F, Jiang Y and He X 2016

Epigenetic inhibition of miR-663b by long non-coding RNA HOTAIR promotes pancreatic cancer cell proliferation via up-regulation of insulin-like growth factor 2 *Oncotarget* **7** 86857–70

[49] Perks C M and Holly J M 2015 Epigenetic regulation of insulin-like growth factor binding protein-3 (IGFBP-3) in cancer *J. Cell Commun. Signal.* **9** 159–66

[50] Abbastabar M, Kheyrollah M, Azizian K, Bagherlou N, Tehrani S S, Maniati M and Karimian A 2018 Multiple functions of p27 in cell cycle, apoptosis, epigenetic modification and transcriptional regulation for the control of cell growth: A double-edged sword protein *DNA Repair (Amst).* **69** 63–72

[51] Kandoth C, McLellan M D, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, McMichael J F, Wyczalkowski M A, Leiserson M D M, Miller C A, Welch J S, Walter M J, Wendl M C, Ley T J, Wilson R K, Raphael B J and Ding L 2013 Mutational landscape and significance across 12 major cancer types *Nature* **502** 333–9

[52] Revill K, Wang T, Lachenmayer A, Kojima K, Harrington A, Li J, Hoshida Y, Llovet J M and Powers S 2013 Genome-wide methylation analysis and epigenetic unmasking identify tumor suppressor genes in hepatocellular carcinoma. *Gastroenterology* **145** 1424-35.e1-25

[53] Korsmeyer S J 1992 Bcl-2 initiates a new category of oncogenes: regulators of cell death. *Blood* **80** 879–86

[54] Lapierre L R, Kumsta C, Sandri M, Ballabio A and Hansen M 2015 Transcriptional and epigenetic regulation of autophagy in aging. *Autophagy* **11** 867–80

[55] Wright W E, Pereira-Smith O M and Shay J W 1989 Reversible cellular senescence: implications for immortalization of normal human diploid fibroblasts. *Mol. Cell. Biol.* **9** 3088–92

[56] Hanahan D and Weinberg R A 2000 The hallmarks of cancer. *Cell* **100** 57–70

[57] Cheung C Y, Singh M, Ebaugh M J and Brace R A 1995 Vascular endothelial growth factor gene expression in ovine placenta and fetal membranes *Am. J. Obstet. Gynecol.* **173** 753–9

[58] Barr M P, O'Byrne K J, Al-Sarraf N and Gray S G 2015 VEGF-mediated cell survival in non-small-cell lung cancer: implications for epigenetic targeting of VEGF receptors as a therapeutic approach *Epigenomics* **7** 897–910

[59] Turunen M P, Husso T, Musthafa H, Laidinen S, Dragneva G, Laham-Karam N, Honkanen S, Paakinaho A, Laakkonen J P, Gao E, Vihinen-Ranta M, Liimatainen T and Ylä-Herttuala S 2014 Epigenetic Upregulation of Endogenous VEGF-A Reduces Myocardial Infarct Size in Mice ed R Morishita *PLoS One* **9** e89979

[60] Ping S-Y, Shen K-H and Yu D-S 2013 Epigenetic regulation of vascular endothelial growth factor a dynamic expression in transitional cell carcinoma *Mol. Carcinog.* **52** 568–79

[61] Tian Y, Wei W, Li L and Yang R 2015 Down-Regulation of miR-148a Promotes Metastasis by DNA Methylation and is Associated with Prognosis of Skin Cancer by Targeting TGIF2. *Med. Sci. Monit.* **21** 3798–805

[62] Marzese D M, Scolyer R A, Huynh J L, Huang S K, Hirose H, Chong K K, Kiyohara E, Wang J, Kawas N P, Donovan N C, Hata K, Wilmott J S, Murali R, Buckland M E, Shivalingam B, Thompson J F, Morton D L, Kelly D F and Hoon D S B 2014 Epigenome-wide DNA methylation landscape of melanoma progression to brain metastasis reveals aberrations on homeobox D cluster associated with prognosis *Hum. Mol. Genet.* **23** 226–38

[63] Alečković M and Kang Y 2015 Regulation of cancer metastasis by cell-free miRNAs *Biochim. Biophys. Acta - Rev. Cancer* **1855** 24–42

[64] Zhang C 2009 Novel functions for small RNA molecules. *Curr. Opin. Mol. Ther.* **11** 641–51

[65] Wang Y, Stircker H M, Gou D and Liu L 2007 MicroRNA: past and present

Front. Biosci. **12** 2316–29

[66] Lee R C, Feinbaum R L and Ambros V 1993 The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75** 843–54

[67] Lagos-Quintana M, Rauhut R, Lendeckel W and Tuschl T 2001 Identification of Novel Genes Coding for Small Expressed RNAs *Science (80-.).* **294** 853–8

[68] Lau N C, Lim L P, Weinstein E G and Bartel D P 2001 An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis elegans *Science (80-.).* **294** 858–62

[69] Poy M N, Eliasson L, Krutzfeldt J, Kuwajima S, Ma X, MacDonald P E, Pfeffer S, Tuschl T, Rajewsky N, Rorsman P and Stoffel M 2004 A pancreatic islet-specific microRNA regulates insulin secretion *Nature* **432** 226–30

[70] Wilfred B R, Wang W-X and Nelson P T 2007 Energizing miRNA research: A review of the role of miRNAs in lipid metabolism, with a prediction that miR-103/107 regulates human metabolic pathways *Mol. Genet. Metab.* **91** 209–17

[71] Brennecke J, Hipfner D R, Stark A, Russell R B and Cohen S M 2003 bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell* **113** 25–36

[72] Wienholds E, Kloosterman W P, Miska E, Alvarez-Saavedra E, Berezikov E, de Bruijn E, Horvitz H R, Kauppinen S and Plasterk R H A 2005 MicroRNA Expression in Zebrafish Embryonic Development *Science (80-.).* **309** 310–1

[73] Kozomara A and Griffiths-Jones S 2014 miRBase: annotating high confidence microRNAs using deep sequencing data *Nucleic Acids Res.* **42** D68–73

[74] Wahid F, Shehzad A, Khan T and Kim Y Y 2010 MicroRNAs: Synthesis, mechanism, function, and recent clinical trials *Biochim. Biophys. Acta - Mol. Cell Res.* **1803** 1231–43

[75] Ha M and Kim V N 2014 Regulation of microRNA biogenesis *Nat. Rev. Mol. Cell Biol.* **15** 509–24

[76] Alberts B, Johnson A, Lewis J, Raff M, Roberts K and Walter P 2008 *Molecular Biology of the Cell* (Oxford: Garland Science, Taylor & Francis Group)

[77] Gebert L F R and MacRae I J 2018 Regulation of microRNA function in animals *Nat. Rev. Mol. Cell Biol.* 1

[78] Graves P and Zeng Y 2012 Biogenesis of Mammalian MicroRNAs: A Global View *Genomics, Proteomics Bioinforma.* **10** 239–45

[79] Macfarlane L-A and Murphy P R 2010 MicroRNA: Biogenesis, Function and Role in Cancer. *Curr. Genomics* **11** 537–61

[80] Rasool M, Malik A, Zahid S, Basit Ashraf M A, Qazi M H, Asif M, Zaheer A, Arshad M, Raza A and Jamal M S 2016 Non-coding RNAs in cancer diagnosis and therapy *Non-coding RNA Res.* **1** 69–76

[81] Calin G A, Dumitru C D, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K, Rassenti L, Kipps T, Negrini M, Bullrich F and Croce C M 2002 Nonlinear partial differential equations and applications: Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia *Proc. Natl. Acad. Sci.* **99** 15524–9

[82] Rupaimoole R and Slack F J 2017 MicroRNA therapeutics: Towards a new era for the management of cancer and other diseases *Nat. Rev. Drug Discov.* **16** 203–21

[83] Baretti M and Azad N S 2018 The role of epigenetic therapies in colorectal cancer *Curr. Probl. Cancer*

[84] Anon Colorectal Cancer—Patient Version - National Cancer Institute

[85] Lansdorp-Vogelaar I, Kuntz K M, Knudsen A B, van Ballegooijen M, Zauber A G and Jemal A 2012 Contribution of Screening and Survival Differences to Racial Disparities in Colorectal Cancer Rates *Cancer Epidemiol. Biomarkers Prev.* **21** 728–36

[86] Imperiale T F, Juluri R, Sherer E A, Glowinski E A, Johnson C S and Morelli M S 2014 A risk index for advanced neoplasia on the second surveillance colonoscopy in patients with previous adenomatous polyps *Gastrointest. Endosc.* **80** 471–8

[87] Anon Colorectal Cancer: What are the symptoms & signs? | CTCA

[88] Anon Diarrhea: What Causes It, How to Stop it & Home Remedies

[89] Anon Nausea and vomiting Causes - Mayo Clinic

[90] Anon CDC - Colorectal Cancer Screening Tests

[91] Bénard F, Barkun A N, Martel M and von Renteln D 2018 Systematic review of colorectal cancer screening guidelines for average-risk adults: Summarizing the current global recommendations. *World J. Gastroenterol.* **24** 124–38

[92] PDQ Adult Treatment Editorial Board 2002 *Colon Cancer Treatment (PDQ®): Patient Version*

[93] Brenner H, Kloor M and Pox C P 2014 Colorectal cancer *Lancet* **383** 1490–502

[94] DeVita V T, Lawrence T S and Rosenberg S A 2008 *DeVita, Hellman, and Rosenberg's cancer : principles & practice of oncology* (Wolters Kluwer/Lippincott Williams & Wilkins)

[95] Ziller M J, Gu H, Müller F, Donaghey J, Tsai L T-Y, Kohlbacher O, De Jager P L, Rosen E D, Bennett D A, Bernstein B E, Gnirke A and Meissner A 2013 Charting a dynamic DNA methylation landscape of the human genome *Nature* **500** 477–81

[96] Iwaya T, Yokobori T, Nishida N, Kogo R, Sudo T, Tanaka F, Shibata K, Sawada G, Takahashi Y, Ishibashi M, Wakabayashi G, Mori M and Mimori K 2012 Downregulation of miR-144 is associated with colorectal cancer progression via activation of mTOR signaling pathway *Carcinogenesis* **33** 2391–7

[97] Hamfjord J, Stangeland A M, Hughes T, Skrede M L, Tveit K M, Ikdahl T and Kure E H 2012 Differential Expression of miRNAs in Colorectal Cancer: Comparison of Paired Tumor Tissue and Adjacent Normal Mucosa Using High-Throughput Sequencing ed W C S Cho *PLoS One* **7** e34150

[98] R Development Core Team R 2016 R: A language and environment for statistical computing

[99] Anon 2011 *Epigenetics: Product Information Product Information*

[100] Ozer B and Sezerman U 2017 Analysis of the interplay between methylation and expression reveals its potential role in cancer aetiology *Funct. Integr. Genomics* **17** 53–68

[101] Beretta L and Santaniello A 2016 Nearest neighbor imputation algorithms: a critical evaluation. *BMC Med. Inform. Decis. Mak.* **16 Suppl 3** 74

[102] Anon *Singular Value Decomposition (SVD)*

[103] Ritchie M E, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A and Smyth G K 2007 A comparison of background correction methods for two-colour microarrays *Bioinformatics* **23** 2700–7

[104] Nygaard V, Rødland E A and Hovig E 2016 Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* **17** 29–39

[105] Morris T J, Butcher L M, Feber A, Teschendorff A E, Chakravarthy A R, Wojdacz T K and Beck S 2014 ChAMP: 450k Chip Analysis Methylation Pipeline *Bioinformatics* **30** 428–30

[106] Butcher L M and Beck S 2015 Probe Lasso: A novel method to rope in differentially methylated regions with 450K DNA methylation data *Methods* **72** 21–8

[107] Frommer M, McDonald L E, Millar D S, Collis C M, Watt F, Grigg G W, Molloy P L and Paul C L 1992 A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. U. S. A.* **89** 1827–31

[108] Anon DNA methylation guide | Abcam

[109] Simmer F, Brinkman A B, Assenov Y, Matarese F, Kaan A, Sabatino L, Villanueva A, Huertas D, Esteller M, Lengauer T, Bock C, Colantuoni V, Altucci L and Stunnenberg H G 2012 Comparative genome-wide DNA methylation analysis of colorectal tumor and matched normal tissues *Epigenetics* **7** 1355–67

[110] Akalin A, Kormaksson M, Li S, Garrett-Bakelman F E, Figueroa M E, Melnick A and Mason C E 2012 methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles *Genome Biol.* **13** R87

[111] Wang H-Q, Tuominen L K and Tsai C-J 2011 SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures *Bioinformatics* **27** 225–31

[112] Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan M T and Carey V J 2013 Software for Computing and Annotating Genomic Ranges ed A Prlic *PLoS Comput. Biol.* **9** e1003118

[113] Ritchie M E, Phipson B, Wu D, Hu Y, Law C W, Shi W and Smyth G K 2015 limma powers differential expression analyses for RNA-sequencing and microarray studies *Nucleic Acids Res.* **43** e47–e47

[114] Yang Y H and Thorne N P Normalization for Two-color cDNA Microarray Data

[115] Phipson B, Lee S, Majewski J, Alexander W S and Smyth G K 2016 Robust Hyperparameter Estimation Protects Against Hypervariable Genes and Improves Power to Detect Differential Expression *Ann. Appl. Stat.* **10** 946–63

[116] Robinson M D, McCarthy D J and Smyth G K 2009 edgeR: A Bioconductor package for differential expression analysis of digital gene expression data *Bioinformatics* **26** 139–40

[117] Tian Y, Morris T J, Stirling L and Teschendorff A E 2017 ChAMPdata: Data Packages for ChAMP package

[118] Chou C-H, Shrestha S, Yang C-D, Chang N-W, Lin Y-L, Liao K-W, Huang W-C, Sun T-H, Tu S-J, Lee W-H, Chiew M-Y, Tai C-S, Wei T-Y, Tsai T-R, Huang H-T, Wang C-Y, Wu H-Y, Ho S-Y, Chen P-R, Chuang C-H, Hsieh P-J, Wu Y-S, Chen W-L, Li M-J, Wu Y-C, Huang X-Y, Ng F L, Buddhakosai W, Huang P-C, Lan K-C, Huang C-Y, Weng S-L, Cheng Y-N, Liang C, Hsu W-L and Huang H-D 2018 miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions *Nucleic Acids Res.* **46** D296–302

[119] Ulgen E, Ozisik O and Sezerman O U 2018 pathfindR: An R Package for Pathway Enrichment Analysis Utilizing Active Subnetworks

[120] He H, Lin D, Zhang J, Wang Y and Deng H 2017 Comparison of statistical methods for subnetwork detection in the integration of gene expression and protein interaction network *BMC Bioinformatics* **18** 149

[121] Kanehisa M, Sato Y, Kawashima M, Furumichi M and Tanabe M 2016 KEGG as a reference resource for gene and protein annotation *Nucleic Acids Res.* **44** D457–62

[122] Kanehisa M, Furumichi M, Tanabe M, Sato Y and Morishima K 2017 KEGG: new perspectives on genomes, pathways, diseases and drugs *Nucleic Acids Res.* **45** D353–61

[123] Ulgen E, Ozisik O and Sezerman O U 2018 pathfindR - An R Package for Pathway Enrichment Analysis Utilizing Active Subnetworks

[124] Hurd P J and Nelson C J 2009 Advantages of next-generation sequencing versus the microarray in epigenetic research *Briefings Funct. Genomics Proteomics* **8** 174–83

[125] Love M I, Huber W and Anders S 2014 Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 *Genome Biol.* **15** 550

[126] Hamilton T B, Barilla K C and Romaniuk P J 1995 High affinity binding sites for the Wilms' tumour suppressor protein WT1. *Nucleic Acids Res.* **23** 277–84

[127] Bruey J-M, Bruey-Sedano N, Luciano F, Zhai D, Balpai R, Xu C, Kress C L, Bailly-Maitre B, Li X, Osterman A, Matsuzawa S, Terskikh A V, Faustin B and Reed J C 2007 Bcl-2 and Bcl-XL regulate proinflammatory caspase-1 activation by interaction with NALP1. *Cell* **129** 45–56

[128] Cai J, Liu T, Jiang X, Guo C, Liu A and Xiao X 2017 Downregulation of USP18 inhibits growth and induces apoptosis in hepatitis B virus-related hepatocellular carcinoma cells by suppressing BCL2L1 *Exp. Cell Res.* **358** 315–22

[129] Yi C, Mu L, de la Longrais I A R, Sochirca O, Arisio R, Yu H, Hoffman A E, Zhu Y and Katsaro D 2010 The circadian gene NPAS2 is a novel prognostic biomarker for breast cancer. *Breast Cancer Res. Treat.* **120** 663–9

[130] Hoffman A E, Zheng T, Ba Y and Zhu Y 2008 The circadian gene NPAS2, a putative tumor suppressor, is involved in DNA damage response. *Mol. Cancer Res.* **6** 1461–8

[131] Paul A and Paul S 2014 The breast cancer susceptibility genes (BRCA) in breast and ovarian cancers. *Front. Biosci. (Landmark Ed.* **19** 605–18

[132] Bhatia V, Barroso S I, García-Rubio M L, Tumini E, Herrera-Moyano E and Aguilera A 2014 BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2 *Nature* **511** 362–5

[133] Bosviel R, Durif J, Déchelotte P, Bignon Y-J and Bernard-Gallon D 2012

Epigenetic modulation of BRCA1 and BRCA2 gene expression by equol in breast cancer cell lines *Br. J. Nutr.* **108** 1187–93

[134] Jenkins Z A, Hååg P G and Johansson H E 2001 Human eIF5A2 on chromosome 3q25-q27 is a phylogenetically conserved vertebrate variant of eukaryotic translation initiation factor 5A with tissue-specific expression. *Genomics* **71** 101–9

[135] Tang D-J, Dong S-S, Ma N-F, Xie D, Chen L, Fu L, Lau S H, Li Y, Li Y and Guan X-Y 2010 Overexpression of eukaryotic initiation factor 5A2 enhances cell motility and promotes tumor metastasis in hepatocellular carcinoma. *Hepatology* **51** 1255–63

[136] Clement P M J, Johansson H E, Wolff E C and Park M H 2006 Differential expression of eIF5A-1 and eIF5A-2 in human cancer cells. *FEBS J.* **273** 1102–14

[137] Straus S E, Jaffe E S, Puck J M, Dale J K, Elkon K B, Rösen-Wolff A, Peters A M, Sneller M C, Hallahan C W, Wang J, Fischer R E, Jackson C E, Lin A Y, Bäumler C, Siegert E, Marx A, Vaishnaw A K, Grodzicky T, Fleisher T A and Lenardo M J 2001 The development of lymphomas in families with autoimmune lymphoproliferative syndrome with germline Fas mutations and defective lymphocyte apoptosis. *Blood* **98** 194–200

[138] Li Q, Ching A K-K, Chan B C-L, Chow S K-Y, Lim P-L, Ho T C-Y, Ip W-K, Wong C-K, Lam C W-K, Lee K K-H, Chan J Y-H and Chui Y-L 2004 A death receptor-associated anti-apoptotic protein, BRE, inhibits mitochondrial apoptotic pathway. *J. Biol. Chem.* **279** 52106–16

[139] Flood B, Oficjalska K, Laukens D, Fay J, O'Grady A, Caiazza F, Heetun Z, Mills K H G, Sheahan K, Ryan E J, Doherty G A, Kay E and Creagh E M 2015 Altered expression of caspases-4 and -5 during inflammatory bowel disease and colorectal cancer: Diagnostic and therapeutic potential *Clin. Exp. Immunol.* **181** 39–50

[140] Mittal R D, Mittal T, Singh A K and Mandal R K 2012 Association of Caspases with an Increased Prostate Cancer Risk in North Indian Population *DNA Cell Biol.* **31** 67–73

[141] Shimizu N, Ohta M, Fujiwara C, Sagara J, Mochizuki N, Oda T and Utiyama H 1992 A gene coding for a zinc finger protein is induced during 12-O-tetradecanoylphorbol-13-acetate-stimulated HL-60 cell differentiation. *J. Biochem.* **111** 272–7

[142] Zheng L, Pu J, Jiang G, Weng M, He J, Mei H, Hou X and Tong Q 2010 Abnormal expression of early growth response 1 in gastric cancer: association with tumor invasion, metastasis and heparanase transcription. *Pathol. Int.* **60** 268–77

[143] Taira N, Mimoto R, Kurata M, Yamaguchi T, Kitagawa M, Miki Y and Yoshida K 2012 DYRK2 priming phosphorylation of c-Jun and c-Myc modulates cell cycle progression in human cancer cells. *J. Clin. Invest.* **122** 859–72

[144] Scheller H, Tobollik S, Kutzera A, Eder M, Unterlehberg J, Pfeil I and Jungnickel B 2010 c-Myc overexpression promotes a germinal center-like program in Burkitt's lymphoma. *Oncogene* **29** 888–97

[145] Shajani-Yi Z, de Abreu F B, Peterson J D and Tsongalis G J 2018 Frequency of Somatic TP53 Mutations in Combination with Known Pathogenic Mutations in Colon Adenocarcinoma, Non–Small Cell Lung Carcinoma, and Gliomas as Identified by Next-Generation Sequencing *Neoplasia* **20** 256–62

[146] Klug S J, Ressing M, Koenig J, Abba M C, Agorastos T, Brenna S M F, Ciotti M, Das B R, Del Mistro A, Dybikowska A, Giuliano A R, Gudleviciene Z, Gyllensten U, Haws A L F, Helland A, Herrington C S, Hildesheim A, Humbey O, Jee S H, Kim J W, Madeleine M M, Menczer J, Ngan H Y S, Nishikawa A, Niwa Y, Pegoraro R, Pillai M R, Ranzani G, Rezza G, Rosenthal A N, Roychoudhury S, Saranath D, Schmitt V M, Sengupta S, Settheetham-Ishida W, Shirasawa H, Snijders P J F, Stoler M H, Suárez-Rincón A E, Szarka K,

Tachezy R, Ueda M, van der Zee A G J, von Knebel Doeberitz M, Wu M-T, Yamashita T, Zehbe I and Blettner M 2009 TP53 codon 72 polymorphism and cervical cancer: a pooled analysis of individual data from 49 studies. *Lancet. Oncol.* **10** 772–84

[147] Hu X, Zhang Z, Ma D, Huettner P C, Massad L S, Nguyen L, Borecki I and Rader J S 2010 TP53, MDM2, NQO1, and susceptibility to cervical cancer. *Cancer Epidemiol. Biomarkers Prev.* **19** 755–61

[148] Jiang P, Liu J, Zeng X, Li W and Tang J 2010 Association of TP53 codon 72 polymorphism with cervical cancer risk in Chinese women. *Cancer Genet. Cytogenet.* **197** 174–8

[149] Loupakis F, Ruzzo A, Cremolini C, Vincenzi B, Salvatore L, Santini D, Masi G, Stasi I, Canestrari E, Rulli E, Floriani I, Bencardino K, Galluccio N, Catalano V, Tonini G, Magnani M, Fontanini G, Basolo F, Falcone A and Graziano F 2009 KRAS codon 61, 146 and BRAF mutations predict resistance to cetuximab plus irinotecan in KRAS codon 12 and 13 wild-type metastatic colorectal cancer. *Br. J. Cancer* **101** 715–21

[150] Zlobec I, Molinari F, Kovac M, Bihl M P, Altermatt H J, Diebold J, Frick H, Germer M, Horcic M, Montani M, Singer G, Yurtsever H, Zettl A, Terracciano L, Mazzucchelli L, Saletti P, Frattini M, Heinimann K and Lugli A 2010 Prognostic and predictive value of TOPK stratified by KRAS and BRAF gene alterations in sporadic, hereditary and metastatic colorectal cancer patients. *Br. J. Cancer* **102** 151–61

[151] Ciftci R, Tas F, Yasasever C T, Aksit E, Karabulut S, Sen F, Keskin S, Kilic L, Yildiz İ, Bozbey H U, Duranyildiz D and Vatansever S 2014 High serum transforming growth factor beta 1 (TGFB1) level predicts better survival in breast cancer *Tumor Biol.* **35** 6941–8

[152] Chiorazzi M, Rui L, Yang Y, Ceribelli M, Tishbi N, Maurer C W, Ranuncolo S M, Zhao H, Xu W, Chan W-C C, Jaffe E S, Gascoyne R D, Campo E, Rosenwald A, Ott G, Delabie J, Rimsza L M, Shaham S and Staudt L M 2013

Related F-box proteins control cell death in Caenorhabditis elegans and human lymphoma. *Proc. Natl. Acad. Sci. U. S. A.* **110** 3943–8

[153] Harris S L and Levine A J 2005 The p53 pathway: positive and negative feedback loops. *Oncogene* **24** 2899–908

[154] Levine A J, Hu W and Feng Z 2006 The P53 pathway: what questions remain to be explored? *Cell Death Differ.* **13** 1027–36

[155] Bálint E E and Vousden K H 2001 Activation and activities of the p53 tumour suppressor protein. *Br. J. Cancer* **85** 1813–23

[156] Klein E A and Assoian R K 2008 Transcriptional regulation of the cyclin D1 gene at a glance. *J. Cell Sci.* **121** 3853–7

[157] Sherr C J and Roberts J M 1999 CDK inhibitors: positive and negative regulators of G1-phase progression *Genes Dev.* **13** 1501–12

[158] Twomey C and McCarthy J V Pathways of apoptosis and importance in development. *J. Cell. Mol. Med.* **9** 345–59

[159] Kumar S and Cakouros D 2004 Transcriptional control of the core cell-death machinery. *Trends Biochem. Sci.* **29** 193–9

[160] Fuchs Y and Steller H 2015 Live to die another way: modes of programmed cell death and the signals emanating from dying cells. *Nat. Rev. Mol. Cell Biol.* **16** 329–44

[161] Chu W-M 2013 Tumor necrosis factor. *Cancer Lett.* **328** 222–5

[162] MacEwan D J 2002 TNF receptor subtype signalling: differences and cellular consequences. *Cell. Signal.* **14** 477–92

[163] Pomerantz J L and Baltimore D 1999 NF-kappaB activation by a signaling complex containing TRAF2, TANK and TBK1, a novel IKK-related kinase. *EMBO J.* **18** 6694–704

[164] Dewey M 2018 metap: meta-analysis of significance values

[165] Anon KEGG BRITE: KEGG Orthology (KO) - Homo sapiens (human)

[166] Zhang J-M and An J 2007 Cytokines, inflammation, and pain. *Int. Anesthesiol. Clin.* **45** 27–37

[167] Waddington C H 2012 The Epigenotype *Int. J. Epidemiol.* **41** 10–3

[168] Cooper G 2000 *The Cell: A Molecular Approach* (Sunderland (MA): Sinauer Associates)

[169] Wanders R J A and Waterham H R 2006 Peroxisomal disorders: the single peroxisomal enzyme deficiencies. *Biochim. Biophys. Acta* **1763** 1707–20

[170] Anon KEGG BRITE: KEGG Orthology (KO) - Homo sapiens (human)

[171] Anon KEGG BRITE: KEGG Orthology (KO) - Homo sapiens (human)

[172] Tanaka K 2009 The proteasome: overview of structure and functions. *Proc. Jpn. Acad. Ser. B. Phys. Biol. Sci.* **85** 12–36

[173] Ritchie D B, Schellenberg M J and MacMillan A M 2009 Spliceosome structure: piece by piece. *Biochim. Biophys. Acta* **1789** 624–33

[174] Anon KEGG PATHWAY: RNA polymerase

[175] Griffiths A J, Miller J H, Suzuki D T, Lewontin R C and Gelbart W M 2000 Transcription and RNA polymerase

[176] Hannan K M, Sanij E, Rothblum L I, Hannan R D and Pearson R B 2013 Dysregulation of RNA polymerase I transcription during disease. *Biochim. Biophys. Acta* **1829** 342–60

[177] Dunn O J 1961 Multiple Comparisons among Means *J. Am. Stat. Assoc.* **56** 52–64

[178] Feng C, Wang H, Lu N, Chen T, He H, Lu Y and Tu X M 2014 Log-

transformation and its implications for data analysis. *Shanghai Arch. psychiatry* **26** 105–9