Sara Maria Ventura Ramalhete

Epigenetic biomarkers as predictors of clinical outcomes in colorectal cancer

UNIVERSIDADE DO ALGARVE

Departamento de Ciências Biomédicas e Medicina

2019

Sara Maria Ventura Ramalhete

Epigenetic biomarkers as predictors of clinical outcomes in colorectal cancer

Mestrado em Oncobiologia:

Mecanismos Moleculares do Cancro

Trabalho efetuado sob a orientação de:

Professor Doutor Pedro Castelo-Branco

UNIVERSIDADE DO ALGARVE

Departamento de Ciências Biomédicas e Medicina

2019

Epigenetic biomarkers as predictors of clinical outcomes in colorectal cancer

Declaração de autoria do trabalho

Declaro ser a autora deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

I declare that I am the author of this work that is original and unpublished. Authors and works consulted are properly cited in the text and included in the list of references.”

(Sara Maria Ventura Ramalhete)

"If you want to make the world a better place,

Take a look at yourself and make a change."

Michael Jackson

Agradecimentos

Em primeiro lugar, agradeço ao meu orientador, Professor Doutor Pedro Castelo-Branco, por me ter aceite no seu grupo, ter confiado em mim, e por todo o apoio que me deu ao longo da realização desta dissertação.

Agradeço à Professora Doutora Ana Marreiros por toda motivação, disponibilidade e ajuda em tudo o que precisei para a realização deste projeto.

Aos meus colegas André Mestre e André Fonseca, um muito obrigada por todo o apoio, motivação e amizade que demonstraram ao longo deste último ano.

Quero também agradecer a todos os restantes membros desta equipa por me terem apoiado e acolhido, e por todas as discussões científicas que proporcionaram ao longo da realização deste projeto.

Um agradecimento especial ao Luís Carlos por todo o apoio e motivação para que pudesse chegar até aqui.

Por último, mas não menos importante, agradeço à minha família por todo o carinho, confiança e por me terem ajudado a chegar aqui.

Abstract

Colorectal Cancer is the third most common cancer and the second leading cause of death by cancer worldwide with about 1.3 million new cancer cases and 693,933 deaths reported in 2012.

Here, we intend to determine an epigenetic roadmap of Colorectal Cancer to predict tumor progression and patient outcome.

We analyzed whole-genome DNA methylation (*Illumina Infinium HumanMethylation 450K array*) and gene expression (*Illumina HiSeq*) in multiple stages of CRC (21 normal, 54 stage I, 131 stage II, 111 stage III, and 51 stage IV). The data is available in TCGA database, and was downloaded, processed and analyzed through R programming.

Results show that, in stages I, II, III, and IV, 307, 400, 305 and 233 genes are differentially expressed (fold-change absolute value > 1.5, *p-value* adjusted<0.05) and 924, 1814, 1169, and 618 CpG sites are differentially methylated (Δβ absolute value > 0.2, *p-value* adjusted<0.05), respectively. In addition, all these CpG sites are correlated with the respective gene. When the KEGG and Gene Ontology analysis was performed, we found that the enriched functions are related to nervous system, one of the processes deregulated in cancer progression. Moreover, we also identified 66, 85, 41, and 40 specific genes for stages I, II, III, and IV, respectively.

Regarding the diagnosis, were found 238 genes and 835 CpG sites as good diagnosis tool for stage I (AUC>0.8). Furthermore, 6, 1, and 5 genes and 87, 7, and 3 CpG sites were classified as good biomarkers for overall survival for stages I-IV, respectively. In addition, 3, 3, and 2 genes and 30, 12, 9 CpG sites were identified as good biomarkers for recurrence free survival for stages I-IV, respectively.

These results suggest that different methylation events are associated to specific stages of CRC which can predict patient outcome and might improve colorectal cancer diagnosis and prognosis.

Keywords: Colorectal Cancer, Epigenetics, DNA methylation, Biomarkers, Diagnosis, Prognosis.

Resumo

Introdução: O cancro colorretal é um evento biológico que compreende múltiplos passos, decorrendo de diversas alterações genéticas e epigenéticas.

Apesar das melhorias no rastreio, diagnóstico e prognóstico de cancro, incluindo de cancro colorretal, este continua a ser o terceiro tipo de cancro mais comum em homens e segundo em mulheres, com mais de 1,3 milhões de novos casos diagnosticados, e 693.933 mortes reportados em todo o mundo no ano de 2012. Em parte, a incidência e mortalidade continuam elevadas devido à baixa sensibilidade e especificidade na deteção de cancro colorretal nos estádios iniciais da doença.

Atualmente, entre os diversos meios de diagnóstico, a técnica mais eficiente é a colonoscopia, contudo apresenta baixa especificidade e sensibilidade. Estudos mais recentes têm apontado outros biomarcadores como forma de diagnóstico e prognóstico para o cancro colorretal, incluindo a septina 9. Este último é um biomarcador epigenético atualmente comercializado.

Este projeto teve como objetivos realizar uma análise global do genoma em termos de metilação do ADN e expressão genética através de um código em R, identificar mutações epigenéticas que ocorram ao longo da progressão do cancro colorretal, e, por último, relacionar estas alterações com o efeito causado nos doentes.

Métodos: Neste projeto, foi efetuada uma análise global do genoma de um cohort de cancro colorretal, em termos de metilação do ADN (*Illumina Infinium HumanMethylation 450K array*) e expressão genética (*Illumina HiSeq*). Neste projeto, foram analisadas 21 amostras de tecido normal adjacente ao tumor e 347 amostras tumorais divididas de acordo com a classificação TNM (54 estadio I, 131 estadio II, 111 estadio III e 51 estadio IV). Estes dados estão publicamente disponíveis, sendo que foram descarregados da base de dados do *The Cancer Genome Atlas* (TCGA) e analisados através de programação em R.

Resultados: Os resultados sugerem que nos estádios I, II, III e IV, estão diferencialmente expressos 307, 400, 305 e 233 genes (valor absoluto de *fold-change* > 1,5 e *p-value* ajustado (FDR) < 0.05) e diferencialmente metilados 924, 1.814, 1.169 e 618 locais de metilação (valor absoluto de $\Delta\beta > 0,2$ e *p-value* ajustado (FDR) < 0.05), respetivamente.

Em adição, cada um destes locais de metilação encontra-se correlacionado com os respetivos genes encontrados diferencialmente expressos no mesmo estadio (*p-value* < 0.05). De seguida, efetuou-se uma análise nas bases de dados KEGG e *Gene Ontology* (GO). A utilização destas ferramentas revelou que as funções mais enriquecidas estão relacionadas com o sistema nervoso. Estudos anteriores já tinham descrito alterações em genes envolvidos no desenvolvimento e regulação do sistema nervoso como desreguladas em diversos tipos de cancro. Em adição, foi ainda realizada uma análise com o objetivo de encontrar quais dos genes encontrados diferencialmente expressos e que continham locais de metilação diferencialmente metilados ainda não tinham sido reportados em associação com cancro colorretal e cancro em geral. Esta análise sugere que 87 genes nunca foram associados nem com cancro colorretal nem com cancro no geral. Em oposição, 511 já forma reportados em algum tipo de cancro. Destes últimos, 278 já foram também reportados em cancro colorretal enquanto 233 nunca foram descritos neste tipo de cancro.

Como forma de validação, realizou-se, ainda, uma técnica multivariada de representação gráfica, a qual demonstrou que tanto os genes como os locais de metilação selecionados conseguem distinguir amostras tumorais de amostras normais. Esta técnica permitiu-nos ainda diferenciar amostras tumorais em dois grupos principais distintos.

Ainda neste estudo, foram identificados 66, 85, 41 e 40 genes que estão somente diferencialmente expressos nos estádios I, II, III e IV. Curiosamente, apenas 85 genes são comuns aos 4 estadios de desenvolvimento de cancro colorretal

O potencial dos genes e locais de metilação, encontrados como diferencialmente expressos e metilados, respetivamente, para distinguir tecido tumoral do tecido normal também foi avaliado através da análise de curvas de receiver operating characteristic (ROC). Como resultado, obteve-se que 238 genes e 835 locais de metilação são bons marcadores de tecido tumoral em estadio I, quando comparado com tecido normal adjacente (AUC > 0,8, sendo que apenas foram selecionados os pontos ótimos com especificidade e sensibilidade > 60%). *ASTN1*, por exemplo, foi um dos genes classificados como um excelente marcador de diagnóstico (AUC =0,989). Este gene contém ainda o local de metilação cg08104310, o qual foi considerado um excelente marcador de diagnóstico (AUC=1,000).

De seguida, a capacidade de prever o *outcome* do paciente em termos de sobrevida em geral e sobrevida livre de progressão, através dos valores de metilação e expressão dos genes e locais de metilação específicos para cada um dos estádios, foi avaliada. Em relação à sobrevivência em geral, para os estádios II, III e IV, foram identificados 6, 1 e 5 genes e 87, 7 e 3 locais de metilação, respetivamente, como possíveis biomarcadores de prognóstico (*p-value* < 0.05). Especificamente, genes como o *ZNF536* (*p-value*=0,018; HR=3,133), *SOX1* (*p-value*=0,041; HR=0.459) e *BFSP2* (*p-value*=0,027; HR=2.828), por exemplo, foram identificados como bons preditores de sobrevivência em geral dos estádios II, III e IV, respetivamente. Relativamente aos locais de metilação, as cg02430935 localizada no gene *HMX* (*p-value*=0,013; HR=3,139), cg26489108 localizada no gene *DMRT3* (*p-value*=0,027; HR=0,407) e a cg01847754 localizada no gene *CXorf1* (*p-value*=0,019; HR=3,155), por exemplo, foram identificadas como bons marcadores para a sobrevivência em geral dos estádios II, III e IV, respetivamente.

Quanto à sobrevivência livre de recorrência, para os estádios II, III e IV, foram identificados 3, 3 e 2 genes e 30, 12 e 9 locais de metilação, respetivamente, capazes de prever se o doente para recorrer ou não. Mais concretamente, genes como o *CNTD2* (*p-value*=0,00033; HR=0,196), *SOX1* (*p-value*=0.01; HR=0,359) e *HTR2C* (*p-value*=0,0064; HR=0,285) foram identificados como bons preditores de prognóstico para a sobrevivência livre de progressão nos estádios II, III e IV, respetivamente. Relativamente aos locais de metilação, as cg06162589 localizada no gene *SLC5A8* (*p-value*=0.0066; HR=0,2924), cg03700449 localizada no gene *ASCL1* (*p-value*=0.0055; HR=0,3114) e cg14772660 localizada no gene *SLC5A7* (*p-value*=0.0047; HR=4,3174) são exemplos de bons preditores de sobrevivência livre de progressão para os estádios II, III e IV, respetivamente.

Conclusão: Este estudo sugere que as alterações epigenéticas são dinâmicas ao longo da progressão de cancro colorretal, demonstrando que há alterações que são características de estádios específicos, enquanto outras se mantêm alteradas desde o primeiro estadio. Notavelmente, algumas das alterações conseguem distinguir doentes com um prognóstico mais severo de doentes com um prognóstico mais indolente.

Assim sendo, este estudo mostrou que existem possíveis biomarcadores para cancro colorretal que devem ser melhor estudados no futuro. Este estudo pode ainda demarcar o início da melhoria das técnicas de diagnóstico e prognóstico.

Palavras-chave: Cancro Colorretal, Epigenética, Metilação do ADN, Biomarcadores, Diagnóstico, Prognóstico.

INDEX OF CONTENTS

INDEX OF FIGURES

INDEX OF TABLES

ABBREVIATIONS LIST

3’ UTR- Three Prime Untranslated Region

5’ UTR- Five Prime Untranslated Region

ALS- Amyotrophic Lateral Sclerosis

APC- Adenomatous Polyposis Coli

ATP- Adenosine Triphosphate

AUC- Area Under the Curve

BAX- BCL2 Associated X

BRAF- B-Raf Proto-Oncogene

CA- Cancer Antigen

CDC4- Cell Division Cycle 4

CEA- CarcinoEmbryonic Antigen

CIMP- CpG Island Methylator Phenotype

CIN- Chromosome Instability

CpG- Cytosine phosphate Guanine

CRC- Colorectal cancer

DNA- Deoxyribonucleic acid

DNMT- DNA methyltransferase

EGFR- EGF receptor

FDR- False Discovery Rate

FIT- Fecal Immunochemical Test

FOBT- Fecal Occult Blood Testing

FPF- False positive fraction

FS- Flexible Sigmoidoscopy

GO- Gene Ontology

H3K27ac- Acetylation of histone 3 at lysine 27

H3K27me3- Tri methylation of histone 3 at lysine 27

H3K36me3- Tri methylation of histone 3 at lysine 36

H3K4me1- Mono methylation of histone 3 at lysine 4

H3K4me3- Tri methylation of histone 3 at lysine 4

H3K9me3- Tri methylation of histone 3 at lysine 9

HAT- Histone acetyltransferases

HDAC- Histone deacetylases

HDM- Histone demethylases

HMT- Histone methyltransferases

HR- Hazard ratio

ID4- Inhibitor of DNA Binding 4

IGF2R- Insulin-like Growth Factor 2 Receptor

Indel- Insertion and Deletion

IRF8- Interferon Regulatory Factor 8

ITGA4- Integrin Subunit Alpha 4

KM- Kaplan-Meier

Last1- Large Tumor Suppressor 1

Last2- Large Tumor Suppressor 2

MBD- Methyl-CpG-binding domain

MGMT- O-6-Methylguanine-DNA Methyltransferase

MiRNA- MicroRNA

MLH1- MutL Homolog 1

mRNA- Messenger RNA

MSH3- MutS Homolog 3

MSH6- MutS Homolog 6

MSI- Microsatellite instability

Mst1- Mammalian STE20-like Protein Kinase 1

Mst2- Mammalian STE20-like Protein Kinase 2

NCI- National Cancer Institute

NHGRI- National Human Genome Research Institute

NIH- National Institute of Health

PCA- Principal Component Analysis

PIP_2- phosphatidylinositol (4,5)-bisphosphate

PIP_3- phosphatidylinositol (3,4,5)-triphosphate

Pol II- Polymerase II

Pre-miRNA- Precursor miRNA

Pri-miRNA- Primary microRNA

PTEN- Phosphatase and Tensin Homolog

Rb- Retinoblastoma

RISC- RNA-induced silencing complex

RNA- Ribonucleic acid

RNase- Ribonuclease

ROC - Receiver operating characteristic

RTKs- Receptor tyrosine kinases

SDC2- Syndecan 2

SEPT9- Septin 9

SFRP2- Secreted Frizzled Related Protein 2

SLC5A8- Solute Carrier Family 5 Member 8

TAZ- Transcriptional Co-Activator with PDZ-Binding Motif

TCGA- The Cancer Genome Atlas

TEAD- Transcriptional Enhanced Associate Domain

TEAP2E- Transcription factor AP-2 epsilon

TGFBR2- TGFβ receptor 2

THBD- Thrombomodulin

TIMP3- TIMP Metallopeptidase Inhibitor 3

TP53- Tumor Protein 53

TPF- True positive fraction

TSP-1- Thrombospondin-1

TSS- Transcription Start Site

VEGF-A- Vascular Endothelial Growth Factor-A

VIM- Vimentin

YAP- Yes-associated Protein

1. CHAPTER I- INTRODUCTION

1.1. Cancer

Historically, the humoral theory, proposed by Hippocrates, was the first trying to explain what cancer is. He believed that humans contained 4 body fluids, named humor fluids: blood, phlegm, yellow bile, and black bile, which could be the cause of cancer. Specifically, Hippocrates proposed that alterations on these substances including an abnormal increase of black bile led cancer to arise[1].

It was only in 1838 that it was demonstrated that cancer is formed by cells which are derived from other cells- the blastema theory[23]. After, other theories arose, including the chronic irritation theory, which suggested that cancer was caused by chronic irritation; the trauma theory, which asserted that trauma led to cancer, and the parasite theory, which characterized cancer as a contagious disease that could be transmitted among humans through parasites[2,4,5].

Despite multiple attempts to understand the cause of cancer, it was in the 20th century that the mystery started to be solved. Firstly, both Watson and Crick uncovered the structure of deoxyribonucleic acid (DNA). Then, it was revealed how genes work and that genes can be affected by mutations. Later, it was also discovered that DNA can be altered and cause cancer through the exposure to chemicals, radiation, viruses and other carcinogens. It was also in the same century, that oncogenes and tumor suppressor genes were identified[2].

Nowadays, it is known that cancer is a group of diseases characterized by uncontrolled cell division that ultimately can spread to other tissues and metastasize. Although proliferation and cellular growth being normal and essential processes for development of organisms, cell division can become out of control, resulting in the accumulation of both mutations and epimutations[6]. This condition may lead to an uncontrolled cellular growth and, ultimately, in the invasion of distant tissues[7].

1.1.1. Epidemiology of Cancer

Despite the significant improvement in treatment and screening and the search for tumor biomarkers, cancer is the second leading cause of death in the world with more than 14 million new cancer cases reported and 8.2 million deaths worldwide in 2012. In 2018 it was estimated about 18.1 million new cases and 9.6 million cancer-related deaths. Additionally, assenting in statistical predictions, it is expected that over 23 million new cancer cases are diagnosed and 14 million deaths by cancer are reported in 2035.

Among all cancer types, the most frequents are lung, breast, colorectal and prostate cancers[8–10].

1.1.2. Mutation in Tumorigenesis

Mutations and epimutations might have an impact in gene expression by modifying DNA sequence or chromatin structure, respectively[11]. Those changes can occur under many circumstances such as exposure to tobacco, chemicals, radiation or infectious organisms- external factors- and inherited mutations, hormones, immune conditions and random mutations- internal factors[6]. Additionally, other events can also arise during cancer development, such as genomic rearrangements, amplification, insertion and deletion (indel)[12,13].

Importantly, neither the total number of mutations nor epimutations are directly related to the outcome. These events can be assembled in two main groups: driver and passenger mutations. Driver mutations provide selective advantage to tumor cell growth, contributing to the tumor initiation and progression. In contrast, passenger mutations do not provide selective growth advantage, meaning that they do not contribute to tumor initiation and progression. Driver mutations happen in small scale in cancer, whereas passenger mutations are the most common alterations found in cancer cells. Additionally, there is another type of mutations, named gatekeeping mutations, which provide advantages to the growth of normal cells[13–17].

Among all cellular processes, cell fate determination, cell survival, and genome maintenance are the three main processes related to cancer driver genes[13]. These processes are regulated by several oncogenes and tumor suppressor genes, which are often activated or inactivated, respectively, across the tumor development[16].

1.1.3. Hallmarks of Cancer

Hanahan and Weinberg have originally proposed six hallmarks that normal cells acquire during the malignant transformation, which promote tumor growth and progression, revolutionizing the knowledge of tumorigenesis (***Figure 1.1***)[18,19]:

a. **Sustaining proliferative signaling**.

Cancer cells can affect the production and release of growth-promoting signals, such as growth factors that bind to cell-surface receptors. This control affects the cell cycle and cell growth, leading to an uncontrolled proliferation. Specifically, there are different known ways to take control of proliferation such as an autocrine proliferative stimulation, meaning that cancer cells produce growth factors themselves; or stimulating normal cells to produce growth factors. Moreover, mechanisms as somatic mutations that activate additional downstream pathways, or the disruption of negative feedback mechanisms that inhibits proliferative signaling are also commonly observed[18–21].

b. **Evading growth suppressors**.

Besides cancer cells constitutively activate proliferative signals, they inhibit growth suppressors (tumor suppressor genes). Among all known tumor suppressor genes, the most studied are Retinoblastoma (*Rb*) and Tumor Protein 53 (*TP53*). Both are involved in the control of cell cycle, being responsible to decide if the cell proliferates or enters in senesce or apoptosis. Moreover, the cell-cell contact is also lost in several types of cancer, in order to maintain the uncontrolled cell growth[18,19,22,23]. Indeed, this fact contributes to cancer development and metastization as well[24].

c. **Resisting cell death**.

Cancer cells avoid apoptosis, a programed mechanism of cellular death. Indeed, there are regulators that receive and process the extracellular death-inducing signals, as well as regulators that sense and integrate signals of intracellular origin. As a consequence of the activation of any of these regulators, the apoptotic effectors are also activated and the cell suffers apoptosis, being digested by both its neighbors and phagocytic cells[25]. In cancer, this mechanism is abnormally altered, leading to the proliferation of damaged cells. For example, cancer cells lose the tumor suppressor gene *TP53*, which is responsible for inducing apoptosis. Other strategies, such as to increased expression (upregulation) of antiapoptotic regulators and survival signals, or to downregulate proapoptotic factors are also commonly observed in several types of cancer as a way to avoid apoptosis[18,19,26,27].

d. **Enabling replicative immortality**.

Although normal cells have a limited number of cell divisions, cancer cells acquire the capability of dividing indefinitely. Specifically, cancer cells evade both senescence and crisis/apoptosis, being able to proliferate indeterminately. There are evidences that this feature is, in part, due to the activation of telomerase, a DNA polymerase that is responsible by the maintenance of the repetitive sequences located at the ends of chromosomes, named telomeres, which ultimately leads to cell immortalization. Remarkably, studies demonstrated that most non-immortalized cells do not express the gene that encodes for telomerase whereas about 90% of spontaneous immortalized cells do. Moreover, there are evidences that this alteration is correlated to resistance to senescence and crisis/apoptosis, and is associated to poor prognosis[18,19,28,29].

e. **Inducing angiogenesis**.

During malignant transformation, cancer cells are able to induce angiogenesis with the purpose of obtaining nutrients and oxygen as well as remove metabolic wastes and carbon dioxide. This process is mediated by vascular endothelial growth factor-A (*VEGF-A*), a promotor of angiogenesis, and thrombospondin-1 (*TSP-1*), an inhibitor of angiogenesis. Moreover, the production of new blood vessels due to a chronic activation of angiogenesis has liabilities, resulting in precocious capillary sprouting, convoluted and excessive vessel branching, distorted and enlarged vessels, erratic blood flow, micro hemorrhaging, leakiness, and abnormal levels of endothelial cell proliferation and apoptosis.

Studies have revealed that angiogenesis is important in microscopic premalignant stages as well as in later cancer stages as it promotes tumor mass growth[18,19,30,31].

f. Activating invasion and metastasis.

In advanced stages of the disease, a tumor mass with epithelial origin can spread to other tissues through the epithelial-mesenchymal transition mechanism. During this process, cancer cells must be altered in order to efficiently invade and metastasize. This process is characterized by shape alterations, as well as the loss of adhesion properties to neighboring cells and to the extracellular matrix. In detail, loss of proteins such as E-cadherin, cytokeratin, or laminin-1, involved in the cell adhesion, is often observed in tumors of epithelial origin. Additionally, studies have also demonstrated alteration of these class of proteins in other types of cancer, including breast cancer[32], lung cancer[33], and colorectal cancer[34]. Not only that, molecules associated to cell migration during embryogenesis and the inflammation processes were found deregulated[18,19,35,36].

More recently, four additional tumor characteristics were added to the "hallmarks of cancer": genome instability and mutation, tumor-promoting inflammation, deregulating cellular energetics and avoiding immune destruction (***Figure 1.1***)[19].

g. Genome instability and mutation

Throughout tumorigenesis, cancer cells acquire mutations and genomic instability, due to aberrant alterations in multiple genes including oncogenes and tumor suppressor genes. In this sense, cancer cell ability to detect and resolve DNA errors is reduced, and therefore there is increased mutation burden. Thus, cancer cells can acquire alterations that confer selective advantage, promoting cancer progression. Remarkably, these alterations are transmitted to daughter-cells during the cell cycle, leading to a mass constituted by clones of those cells. Moreover, there are evidences that genes involved in the detection and repair of DNA damage, or cell growth and proliferation, as *TP53*, *ATM,* and *BRCA1* are frequently altered in order to promote tumorigenesis[17,19,37–39].

h. Tumor-promoting inflammation

Tumor-promoting inflammation is also considered a cancer characteristic, since it has been found infiltrated innate and adaptative immune cells in tumors. Specifically, inflammatory cells, which are present in the tumor microenvironment, play a key role in tumor progression by facilitating the availability of molecules that promote tumorigenesis, such as growth factors, survival factors, proangiogenic factors, extracellular matrix-modifying enzymes, and inductive signals to induce invasion and metastasis. Importantly, those inflammatory cells can also release chemicals that act as mutagenic factors to cancer cells, promoting cancer development[19,40,41].

i. Deregulating cellular energetics

Cancer cells need to change their metabolic program in order to facilitate cancer progression. Therefore, in both absence and presence of oxygen, cancer cells metabolize glucose through anaerobic glycolysis, a process commonly used by normal cells only in the absence of oxygen. Although, glycolysis is a faster process when compared to mitochondrial phosphorylation, it is a less efficient way of adenosine triphosphate (ATP) production. Thus, in a process of aerobic glycolysis, cancer cells increase glucose transporters (GLUTs) as well as the uptake and utilization of glucose. Additionally, glycolysis is associated with cell proliferation, due to the facilitation of macromolecules and organelles biosynthesis achieved from glycolytic intermediates[19,42,43].

j. Avoiding immune destruction

Although avoiding immune destruction is an emerging hallmark of cancer, this process is yet to be fully understood. The immune system cannot eliminate cancer cells neither in early/later stages nor in micro metastases. Studies have suggested that in order for cancer cells to escape from immune destruction, they block the function of components from the immune system as well as secretions that can eliminate them. For example, cancer cells alter their cell surface antigens in order to avoid recognition by the immune system cells. In this sense, cancer cells develop strategies to evade immune destruction, leading to the down-regulation of the immune system and consequently increasing cancer cells proliferation[19,44].

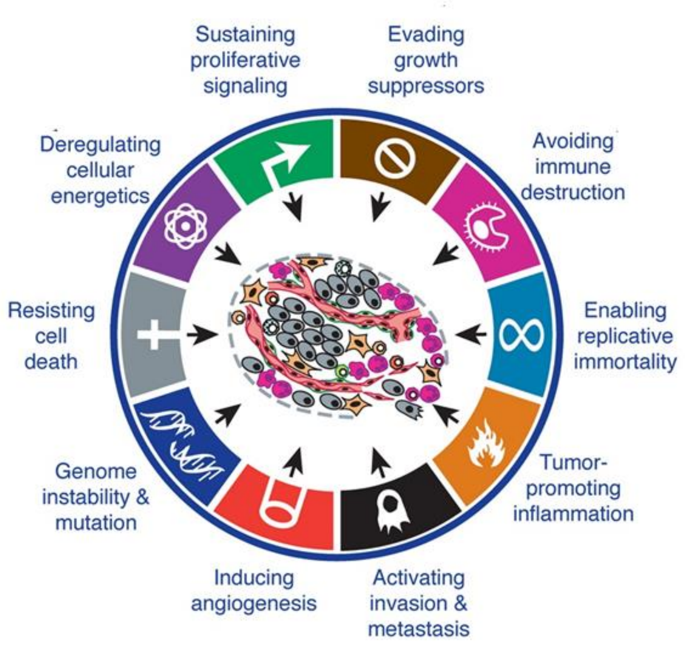

Sustaining proliferative signaling
Evading growth suppressors
Avoiding immune destruction
Enabling replicative immortality
Tumor-promoting inflammation
Activating invasion & metastasis
Inducing angiogenesis
Genome instability & mutation
Resisting cell death
Deregulating cellular energetics

Figure 1.1 Hallmarks of cancer. Capabilities of tumor cells acquired during tumorigenesis (adapted from Hanahan and Weinberg, 2011).

1.2. Main Pathways Altered in Cancer

For the past years, several pathways have been reported to be aberrantly regulated during cancer development and progression, including the following[45]:

1.2.1. TGFβ Pathway

TGFβ pathway is frequently affected in cancer, since it regulates processes such as cell proliferation, apoptosis, and immortalization which are often altered in this disease. When TGFβ activates its receptor (TGFβ receptor), both Smad2 and Smad3 are phosphorylated. and associated with Smad4, constituting a complex that migrates to the nucleus. As a result, proteins that inhibit the cell cycle, as Smad7 and Skil, are produced, leading to cell cycle blockade (***Figure 1.2***)[46–48].

In cancer, mutations/deletions in Smad2, can inactivate the TGFβ pathway, leading to cell cycle progression even in the presence of cell damage[7,49]. Moreover, the TGFβ receptor can also loose it functions due to mutations or DNA methylation of its promoter, leading to inactivation of the pathway.

However, several studies have also demonstrated that TGFβ can be up-regulated in metastatic cancer cells when compared to normal cells. Specifically, TGFβ can induce the remodulation of the extracellular matrix, leading to immunosuppression, angiogenesis and activation of myofibroblast differentiation[50–52].

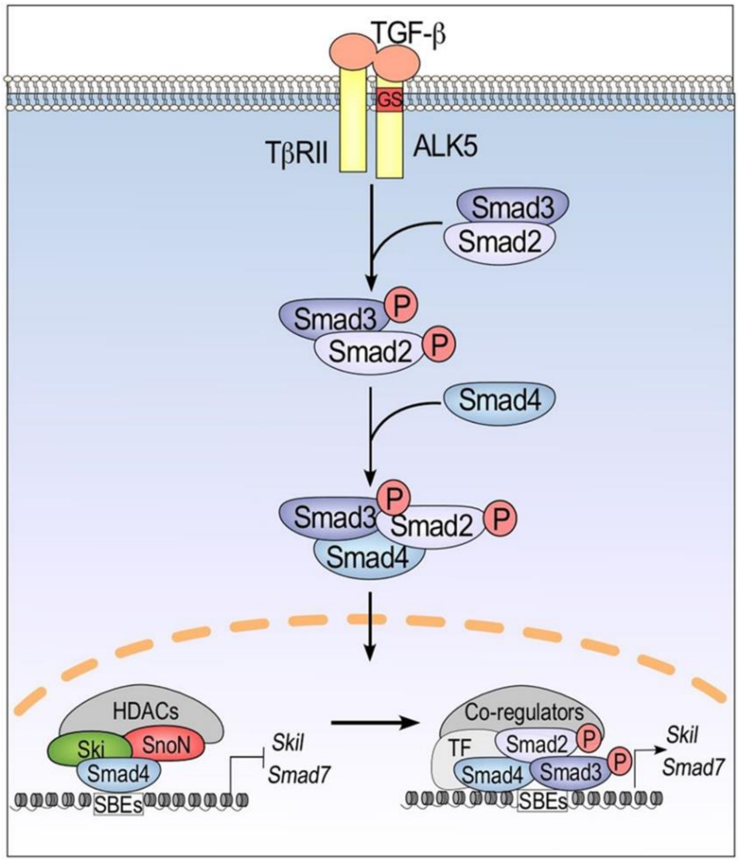


Figure 1.2 Schematic figure representing TGFβ signaling pathway. The activation of TGFβ receptor induces proteins that inhibits cell cycle progression (from Tecalco-Cruz et al. 2018).

1.2.2. Myc Pathway

Myc is considered to have oncogenic properties due to its ability to promote cell cycle progression. In fact, in order for the cell to divide it needs to fulfill multiple requisites which are verified in a checkpoint (R point). If all is correct, Myc forms a heterodimer with Max, inducing the expression of proteins that promote the cell cycle. Simultaneously, Myc can initiate the S phase through the activation of transcription factors. To note there are other pathways that can trigger Myc activation, such as Wnt, Notch, which are approached below. Contrarily, TGFβ signaling can block it (***Figure 1.3***).

Hence, genetic or epigenetic alterations that induce aberrant expression of Myc in cancer promotes cell growth and proliferation[7,53,54].

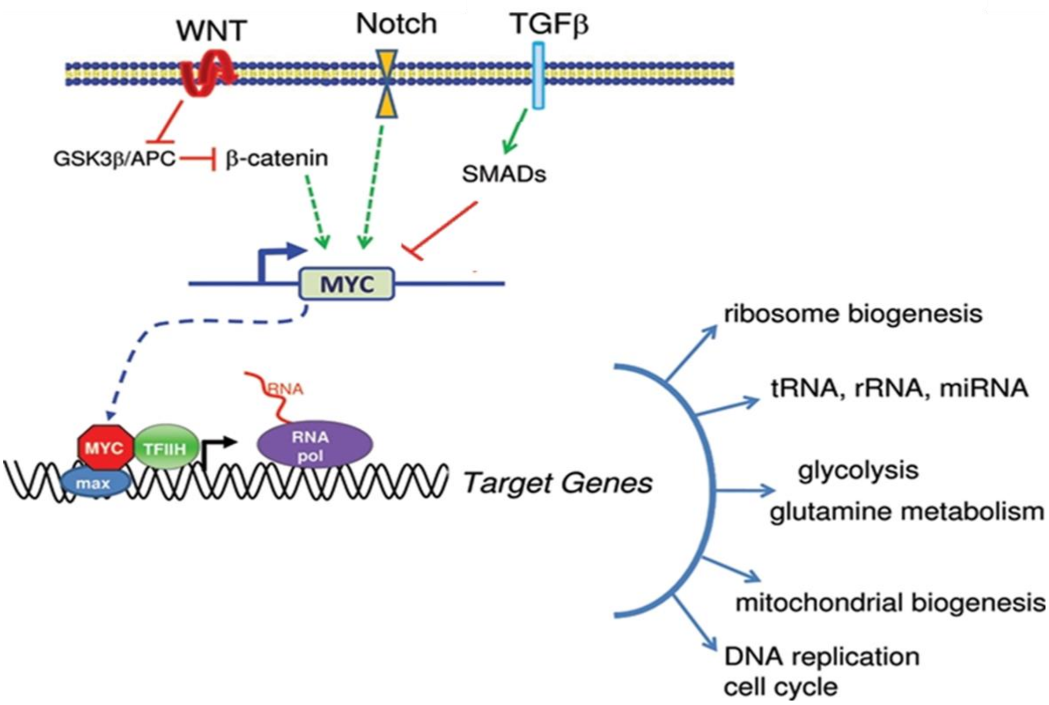


Figure 1.3 Schematic figure representing the Myc pathway. Myc protein can induce processes as ribosome biogenesis, glycolysis, and DNA replication cell cycle (adapted from Dang 2010).

1.2.3. PI3K Pathway

PI3K is an intracellular lipid kinase that, when activated, leads to the conversion of phosphatidylinositol (4,5)-bisphosphate (PIP_2) into phosphatidylinositol (3,4,5)-triphosphate (PIP_3), by phosphorylate PIP_2. As a result, cytoplasmic proteins, including AKT, can bind to PIP_3. Then, two kinases, PDK1 and PDK2, phosphorylate AKT in two sites, leading to its activation. Consequently, AKT kinase phosphorylates other substrates that regulate cell proliferation, survival, and size. Recently, there are evidences that PDK1, when activated, can also induce the expression of Myc through phosphorylation of PLK1. Moreover, PTEN can dephosphorylate PIP_3, converting it to PIP_2, leading to the block of the activity of AKT (***Figure 1.4***)[7,55–57].

Since PI3K signaling regulates several mechanisms, including cell motility, growth, proliferation, and metabolism, it can play a key role in carcinogenesis. Therefore, this pathway is commonly activated in cancer through several mechanisms, including genomic alterations involving PIK3CA, PIK3R1, PTEN, AKT, TSC1, MTOR, and TSC2, [58].

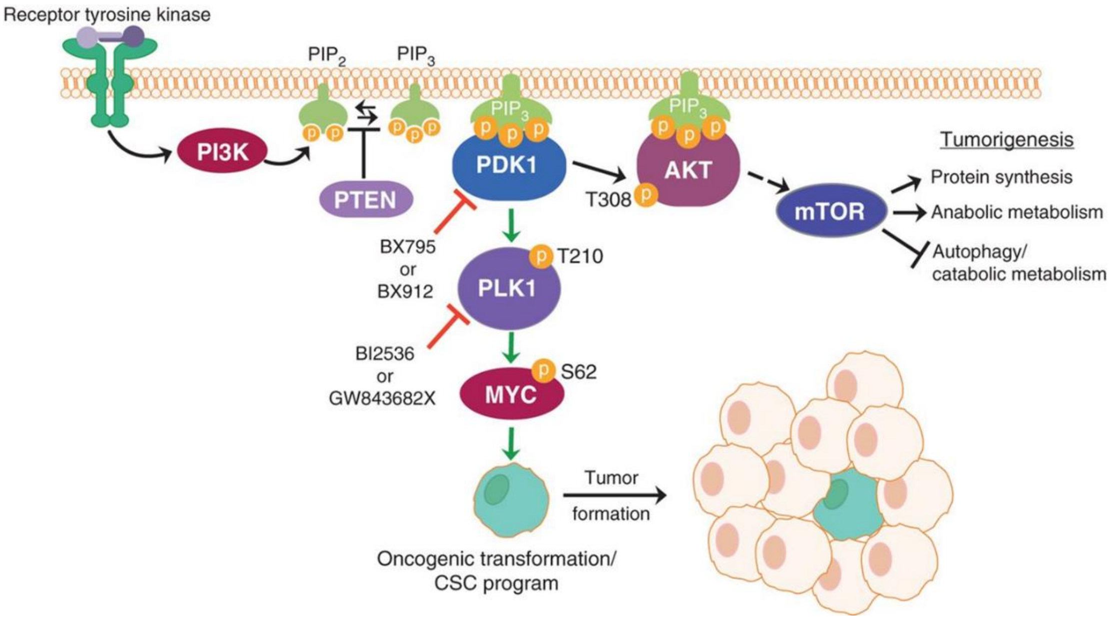


Figure 1.4 Schematic figure representing the PI3K pathway. PI3K converts PIP2 into PIP3, leading to the activation of AKT and Myc. As a result, genes involved in cell proliferation and survival are activated (from Cunningham et al. 2013).

1.2.4. RTK/RAS Pathway

Receptor tyrosine kinases (RTKs) are receptors located in the cell surface and constituted by an extracellular (N-terminal), a transmembrane and a cytoplasmic kinase domain. This type of receptors, when activated by growth factors, hormones, cytokines, neurotrophic factors and other extracellular signaling molecules, stimulate cell proliferation, differentiation, survival and cell migration.

RTKs are monomers, which, when activated by an extracellular stimulus of its N-terminal region, forms a dimer. This dimerization leads to the auto phosphorylation of the receptor, creating a dock site to a complex that can activate Ras, a GTPase protein, that hydrolysis GTP into GDP. Consequently, when the RTK is phosphorylated, Ras is activated, inducing pathways as MAPK and PI3K. Thus, genes involved in cell proliferation and survival are activated (***Figure 1.5***)[59,60].

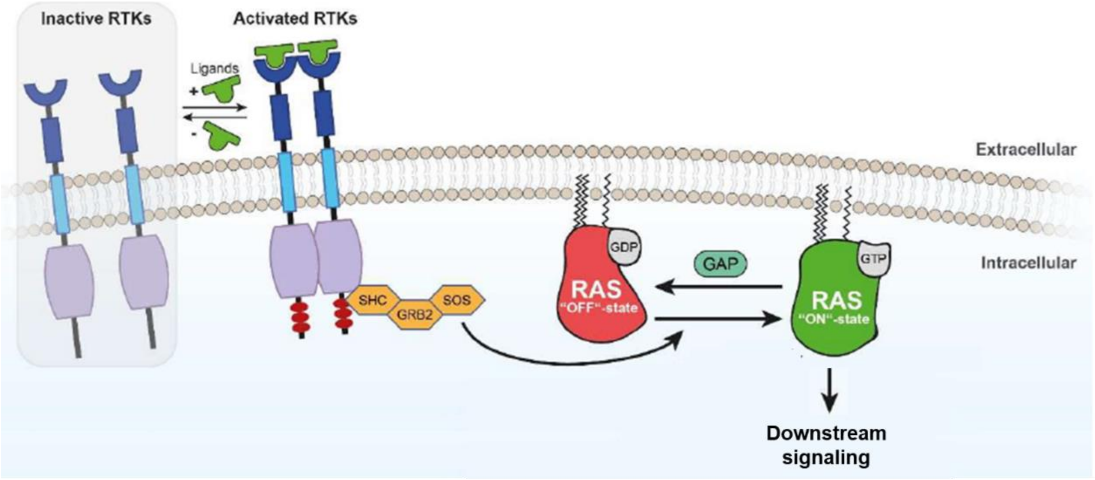


Figure 1.5 Schematic figure representing Ras activation. The phosphorylation of RTK leads to the activation of Ras (adapted from Schöneborn et al. 2018)

In cancer, Ras is found frequently mutated, leading to its constitutive activation. Once, permanently activated Ras is incapable of releasing GTP, and therefore the hydrolysis of GTP into GDP is blocked, leading to a constitutive activation of downstream signaling[7].

1.2.5. NRF2 Pathway

Generally, the transcription factor Nrf2 is considered a tumor suppressor gene, since its activation leads to the stimulation of genes involved in the defense of the cell against metabolic, xenobiotic, and oxidative stress. In fact, when the cell experiences endogenous or exogenous stress, there is an increase in Nrf2 levels, due to the non-ubiquitination of it by KEAP1. Thus, Nrf2 is translocated to the nucleus where it forms a heterodimer with MAF and binds to the antioxidant response element (***Figure 1.6***). As a result, genes involved in metabolism, intracellular redox-balancing, apoptosis, and autophagy are transcribed[61–63].

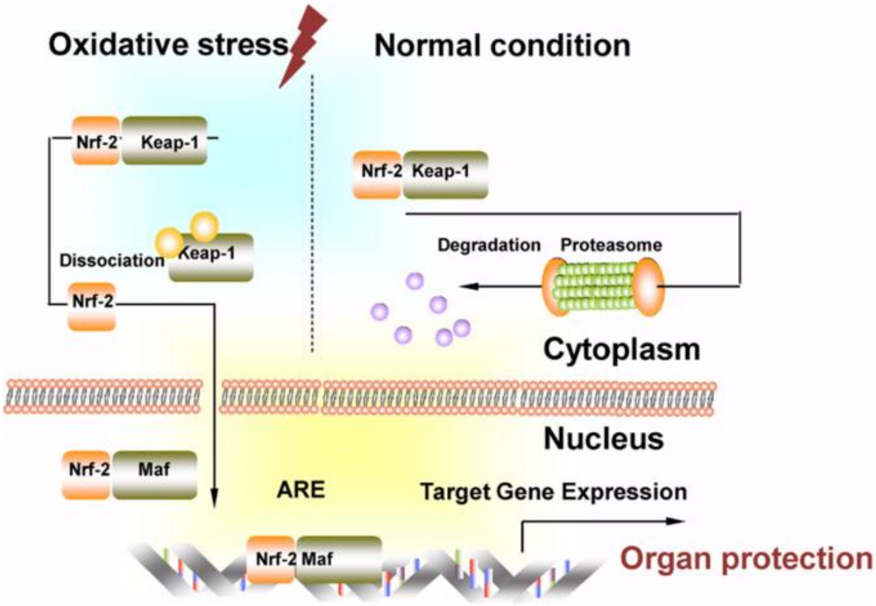


Figure 1.6 Schematic figure representing NRF2 pathway. NRF2 pathway induces the transcription of genes involved in the protection of oxidative stress (adapted from Zhao et al. 2017).

Also, it is believed that Nrf2 can also act as an oncogene, by promoting the survival of cancer cells. Specifically, studies suggested that due to the anti-oxidant effect of Nrf2, cancer cells can be protected from excessive oxidative stress, chemotherapeutic agents, or radiotherapy. However, this oncogenic role in carcinogenesis is yet to be fully understood[61–63].

1.2.6. Wnt Pathway

When the Wnt protein binds to its receptor it leads to the inactivation of glycogen synthase kinase-3β (GSK-3β) preventing the phosphorylation of β-catenin and blocking its degradation. Therefore, β-catenin migrates to the nucleus, where it associates with transcription factors leading to the expression of genes involved in cell proliferation (***Figure 1.7***).

In cancer, the aberrant activation of Wnt pathway can lead to increased translocation of β-catenin into the nucleus, and, consequently, promote the transcription of genes that promote cell survival and proliferation. Moreover, alterations of Apc, a protein that participates in the complex that promotes the degradation of β-catenin, are also frequent in cancer[7,64,65].

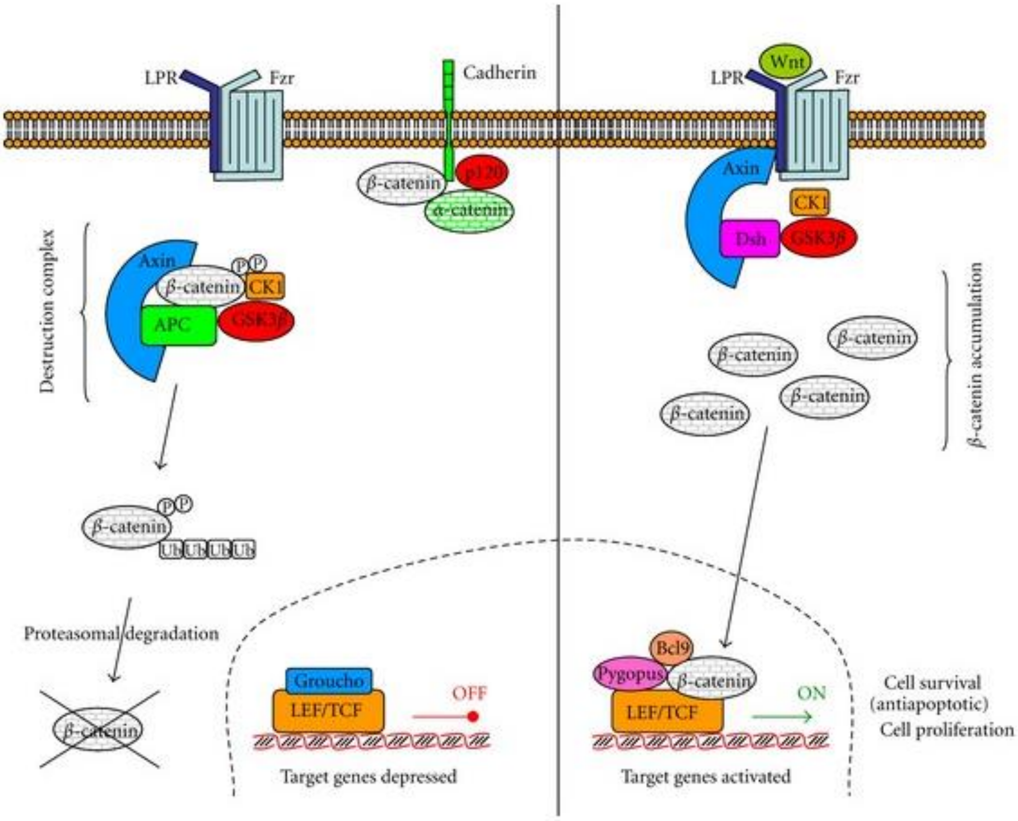


Figure 1.7 Schematic figure representing the Wnt pathway. The expression of Wnt protein blocks the degradation of β-catenin, leading to the transcription of genes involved in cell proliferation (from Centelles 2012).

1.2.7. p53 Pathway

p53, considered the master guardian of the genome, plays a key role in apoptosis control, cell cycle arrest and DNA damaged repair (***Figure 1.8***). Cell stress events, including DNA damage, oncogenic stress, hypoxia, and telomerase erosion, activate the p53 pathway. Specifically, the kinase ATM can block Mdm2, a p53 inhibitor, by phosphorylate it. This event leads to p53 activation, which in turn induces the expression of genes that block cell division and DNA repair, or trigger programmed cell death (***Figure 1.8***)[66–69].

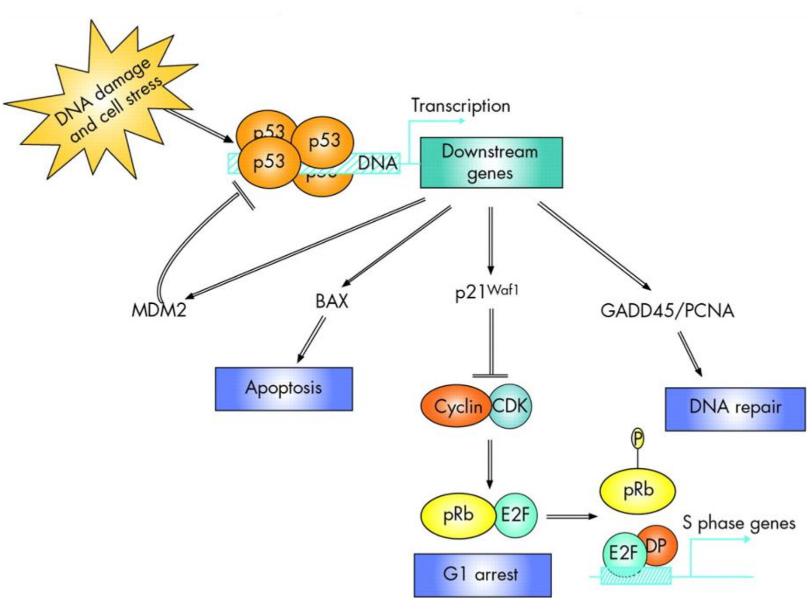


Figure 1.8 Schematic figure representing p53 pathway. When p53 is activated, genes responsible by apoptosis, cell cycle arrest and DNA repair are transcribed (adapted from Boland et al. 2005).

In cancer, levels of p53 can be reduced, or the protein can be sequestered in the nucleus, inactivating its function[7]. Furthermore, mutations in *TP53* can affect its folding resulting in the proliferation of cells with DNA damage and therefore promoting cancer[69].

1.2.8. Notch Pathway

When the Notch receptor is activated by its Delta or Jagged ligands, suffers a proteolytic cut. As a consequence, a cytoplasmatic fragment is translocated into the nucleus, where it activates the expression of genes involved in cell proliferation, by participating in a transcription factor complex (***Figure 1.9***).

In cancer, there are reports that an increased expression or truncated forms of the Notch receptor are common ways to induce cell proliferation. Moreover, an increased expression of Notch ligands is also observed in several types of cancer. Also constitutive expression of Notch, due to deletions in the gene that encodes the extracellular domain of the protein, is also reported in cancer[7,70–72].

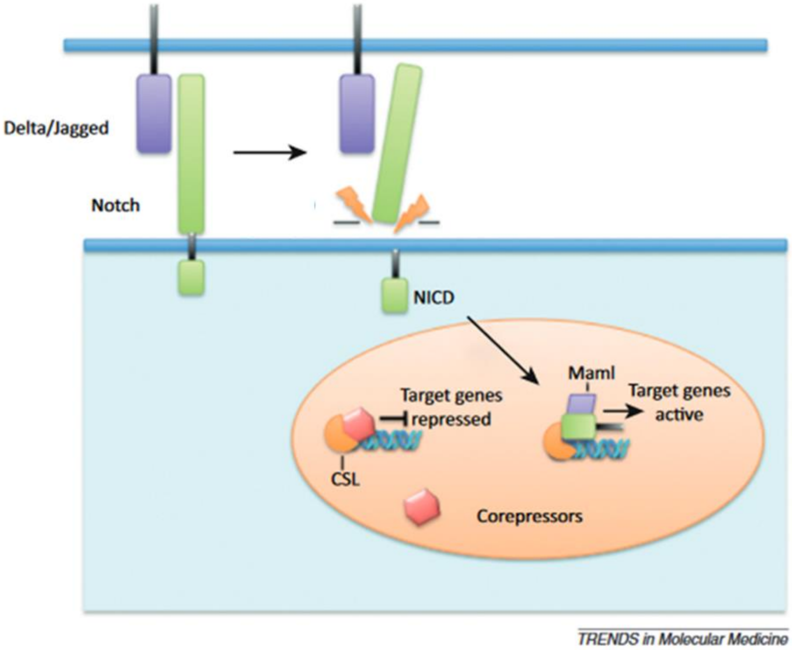


Figure 1.9 Schematic figure representing the Notch pathway. The expression of Notch ligands leads to the translocation of a cytoplasmatic fragment of Notch receptor. As a result, genes involved in cell proliferation are transcbribed (adapted from Avila et al. 2013).

1.3. Colorectal Cancer

Colorectal cancer (CRC) consists in a multistep process which occurs due to both genetic and epigenetic alteration leading to silencing of tumor suppressors genes and increased expression of oncogenes, ultimately promoting cellular growth[73]. This process evolves from a hyperplasia into a adenocarcinoma which ultimately becomes able to metastasize to organs such as liver, lung, peritoneum, bone or brain[74].

1.3.1. Epidemiology of CRC

According to statistical data available in GLOBOCAN about the year 2012, CRC is the third most common cancer in men and the second in women with 1, 3 million new cases diagnosed, and 693,933 deaths in the world. In Portugal, during 2012, 7,129 new cases of CRC, and 3,797 deaths due to this disease were reported (***Figure 1.10***).

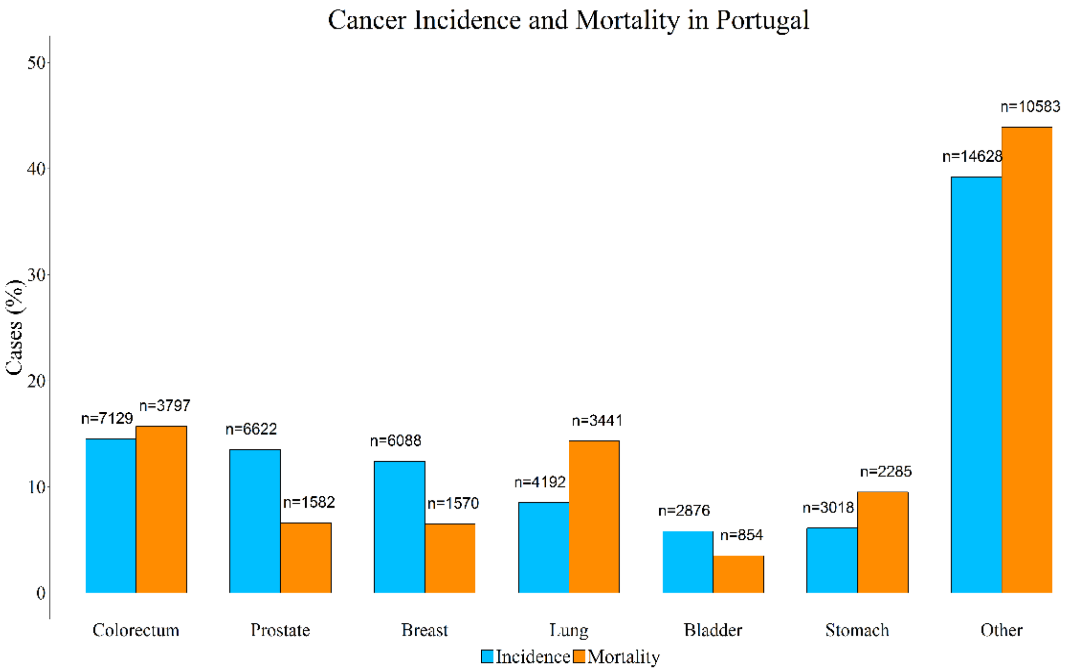


Figure 1.10 Cancer incidence and mortality in Portugal, 2012. Colorectal Cancer is the most common cancer in Portugal, representing 14.5% of all cancer cases, accounting with 7129 new cases in 2012. Moreover, Colorectal Cancer is also the deadliest cancer, being associated to over 15% of mortality by cancer (data source: GLOBOCAN)

The incidence is higher after 50 years of age, being the median age of diagnosis around 70 years[8,75,76]. In developed regions as Australia/New Zealand, Europe and Northern America the incidence of CRC is higher due to risk factors as diet and lifestyle[77]. In contrast, Western Africa, Middle Africa and South-Central Asia are the regions where incidence rates are lower. Despite of this, mortality rates are higher in less developed regions due to a lack of healthcare resources. With regard to 5-year survival rates, these can vary greatly, ranging from around 90% in early stages of the disease to less than 10% when the disease has metastasized[78].

Sporadic CRC is the most frequent form of CRC representing about 75% of all CRC cases[79]. Many risk factors may contribute to cancer initiation and progression including family history[80], age[81], smoking habits[82], alcohol[83], and diet, including both red and processed meat[84,85].

1.3.2. Disease subtypes

CRC can arise sporadically or affect patients who have a genetic predisposition, with family history, including genetic syndromes as Lynch Syndrome and familial adenomatous polyposis. Several genes altered in familial CRC have been identified, including DNA mismatch repair genes, the Adenomatous Polyposis Coli (*APC*), MutL Homolog 1 (*MLH1*), and Phosphatase and Tensin Homolog (*PTEN*)[86].

Sporadic CRC are divided into three main subtypes, depending on the molecular alteration in its origin: microsatellite instability (MSI), chromosomal instability (CIN), and CpG island methylator phenotype (CIMP). However, the tumor can be characterized by features of these different subtypes[86].

1.3.3. Colorectal Cancer Model

The model for colon and rectum tumorigenesis was initially suggested by Fearon and Vogelstein. According to that model, CRC is a multistep process that arises from benign lesions into a malignant tumor. Across the malignant transformation, somatic alterations occur, including alterations in oncogenes and tumor suppressor genes[87].

It is believed that those alterations are generated and propagated through clonal evolution, meaning that mutations/epimutations occur in a cell, and are inherited by daughter cells during mitosis. When the mutations are acquired, the cell has two ways to go: either undergo senescence before entering in the cell cycle or avoid apoptosis and to entry in cell cycle. In the second case, that cell might accumulate mutations and epimutations, originating clones which altogether can be able to cause a heterogeneous tumor mass (***Figure 1.11***)[17,88,89].

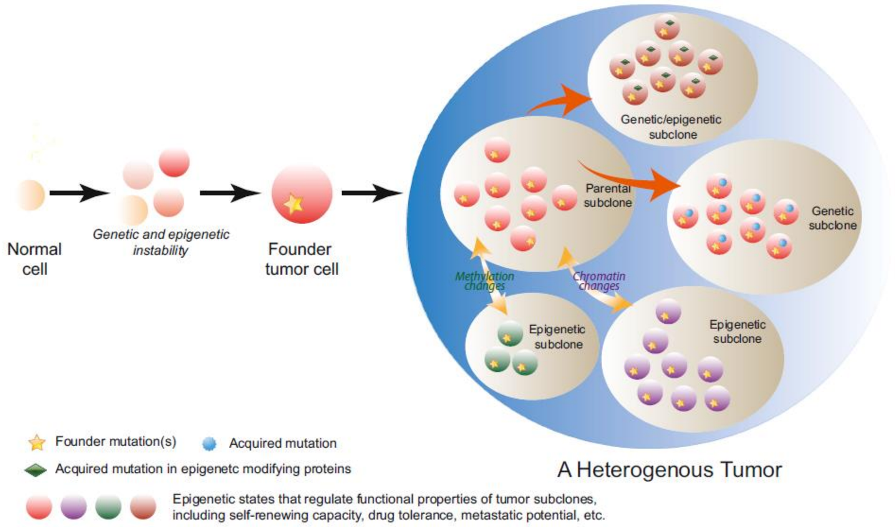

Normal cell
Genetic and epigenetic instability
Founder tumor cell
Genetic/epigenetic subclone
Parental subclone
Genetic subclone
Methylation changes
Chromatin changes
Epigenetic subclone
Epigenetic subclone
Founder mutation(s)
Acquired mutation
Acquired mutation in epigenetc modifying proteins
A Heterogenous Tumor
Epigenetic states that regulate functional properties of tumor subclones, including self-renewing capacity, drug tolerance, metastatic potential, etc.

Figure 1.11 A heterogenous tumor. The accumulation of mutations and epimutations leads to a heterogeneous tumor mass (from Easwaran et al. 2014).

In CRC (***Figure 1.12***), chromosomal instability drives tumorigenesis, initiated by the inactivation of *APC* gene, and followed by mutations in *KRAS*. The increasing chromosomal instability leads to other successive alterations, including loss of heterozygosity (loss of 18q-long arm) and mutations of *SMAD4*, and Cell Division Cycle 4 (*CDC4*). Ultimately, mutations in *TP53* allow the transition from late adenomas to cancer[90].

Another less common way to develop CRC is through microsatellite instability which can facilitate tumor initiation and progression, due to lacking mismatch repair mechanisms. This pathway is often initiated by abnormal alterations in the Wnt signaling and followed by activating mutations in B-Raf Proto Oncogene (*BRAF*) and *KRAS* genes. Importantly, the inefficiency of mismatch repair genes, caused due to hypermethylation of *MLH1* promoter, is increased throughout tumorigenesis. Therefore, tumor cells with mutations in genes as MutS Homolog 3 (*MSH3*), MutS Homolog 6 (*MSH6*), TGFβ receptor 2 (*TGFBR2*), Insulin-like Growth Factor 2 Receptor (*IGF2R*), and *BCL2* Associated X (*BAX*) are positively selected. Altogether, these events lead to the activation of a mechanism responsible for tumor progression independent of *TP53*[91].

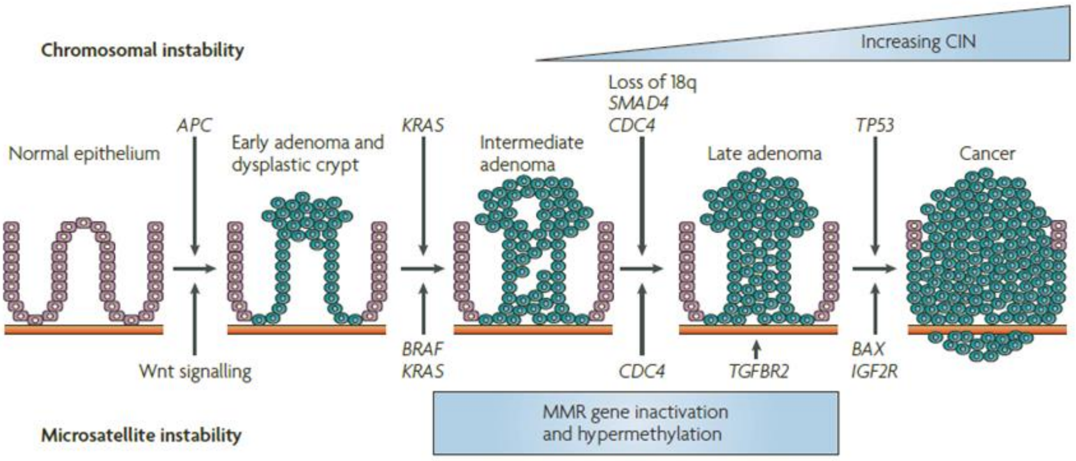


Figure 1.12 Adenoma–carcinoma sequence model schematic representation of genomic events that occur in colon and rectum tumorigenesis (from Walther et al. 2009)

1.3.4. Staging Systems

CRC can be classified according to molecular and histological features- histological staging- or physical exams, biopsies, and imaging tests- clinical staging. These classifications allow to differentiate the state of cancer evolution and decide the best treatment option to the patient.

The most common method of classification used is the TNM system (***Table 1.1***) which distinguishes the cancer stages based on:

a. **Tumor size** (T): size of primary tumor (range from T0-T4),
b. **Lymph nodes** (N): whether cancer has spread to lymph nodes (range from N0-N3),
c. **Metastasis** (M): whether cancer has metastasized (M0 or M1).

Higher numbers of T, N, and M are associated to most advanced disease, and, consequently, to worst prognosis[92]. Importantly, when the category cannot be determined, it is classified by X (TX or NX).

The overall stage is obtained by the combination of these three characteristics[93].

Table 1.1 Colorectal Cancer staging according to the most recent AJCC system effective on January 2018 (adapted from American Cancer Society®)

Overall Stage	T	N	M
Stage I	T1	N0	M0
	T2	N0	M0
Stage IIA	T3	N0	M0
Stage IIB	T4a	N0	M0
Stage IIC	T4b	N0	M0
Stage IIIA	T1-T2	N1/N1c	M0
	T1	N2a	M0
Stage IIIB	T3-T4a	N1/N1c	M0
	T2-T3	N2a	M0
	T1-T2	N2b	M0
Stage IIIC	T4a	N2a	M0
	T3-T4a	N2b	M0
	T4b	N1-N2	M0
Stage IVA	Any T	Any N	M1a
Stage IVB	Any T	Any N	M1b
Stage IVC	Any T	Any N	M1c

1.3.5. Screening, Diagnosis and Prognosis

The detection of CRC in early stages of the disease- screening- is based on colonoscopy, flexible sigmoidoscopy (FS), fecal occult blood testing (FOBT), and fecal immunochemical test (FIT).

Currently, colonoscopy remains the most accurate test for CRC screening and diagnosis. This technique can detect 88-98% of advanced neoplasia. Importantly, several studies have reported a decrease in mortality due to colonoscopy[94]. FS is also used to diagnose CRC, with a sensitivity of 90% to detect advanced neoplasia. However, both colonoscopy and FS are invasive and expensive techniques.

As an alternative to colonoscopy, FOBT and FIT can also be used to screen CRC at lower costs and in a simpler way. Nevertheless, these tests exhibit low sensitivities and specificities. FOBT only detects 13-50% of CRC cases, and 9-24% of advanced neoplasia. On the other hand, the sensitivity of FIT to detect CRC, and advanced neoplasia is 79%, and 32-53%, respectively[81,95–97].

To predict CRC outcome, blood tests targeting tumor markers might be performed. Common CRC marker are the carcinoembryonic antigen (CEA), and cancer antigen 19-9 (CA 19-9). These markers have poor sensitivity and specificity in early stages of the disease. Nonetheless, over the disease progression, both specificity and sensitivity increase[98,99].

The success of treatment and survival depends on the efficiency of screening/detection of cancer. In case of local CRC, the success rate is 70-90% however, in advanced CRC, the mortality is high[88]. In fact, the statistics presented by National Cancer Institute indicate that 92% of stage I, 63-87% stage II, 53-89% stage III, and 11% stage IV colon cancer patients survive at least 5 years. Similarly, the rectum cancer patients in stage I-IV have a 5-years survival rate about 87%, 49-80%, 58-84%, and 12%, respectively[100].

1.4. Epigenetics

Epigenetics, firstly introduced by Conrad Waddington in 1940s, is defined by reversible alterations that affect gene expression without altering DNA sequence[88,101–103]. Regulation of gene expression mediated by epigenetic alterations, including DNA methylation at cytosine residues in CpG dinucleotides, posttranslational modifications of amino acids on the amino-terminal tail of histones, and post-transcriptional regulation by small non coding RNAS, including microRNAs, is frequent in normal cells during embryonic development, imprinting or tissue differentiation[104–106] (***Figure 1.13***). Moreover, these epigenetic changes contribute to the different gene expression profiles of distinct cell types[107]. For example, in humans there are several cell types that are originated from the same fertilized egg cell, presenting the same DNA. However, each one of these cell types have distinct function, due to the inactivation and activation of different sets of genes through epigenetic mechanisms[108].

Remarkably, this process can become abnormal, resulting in aberrant changes of gene expression, and consequently in several diseases, including cancer[107,109].

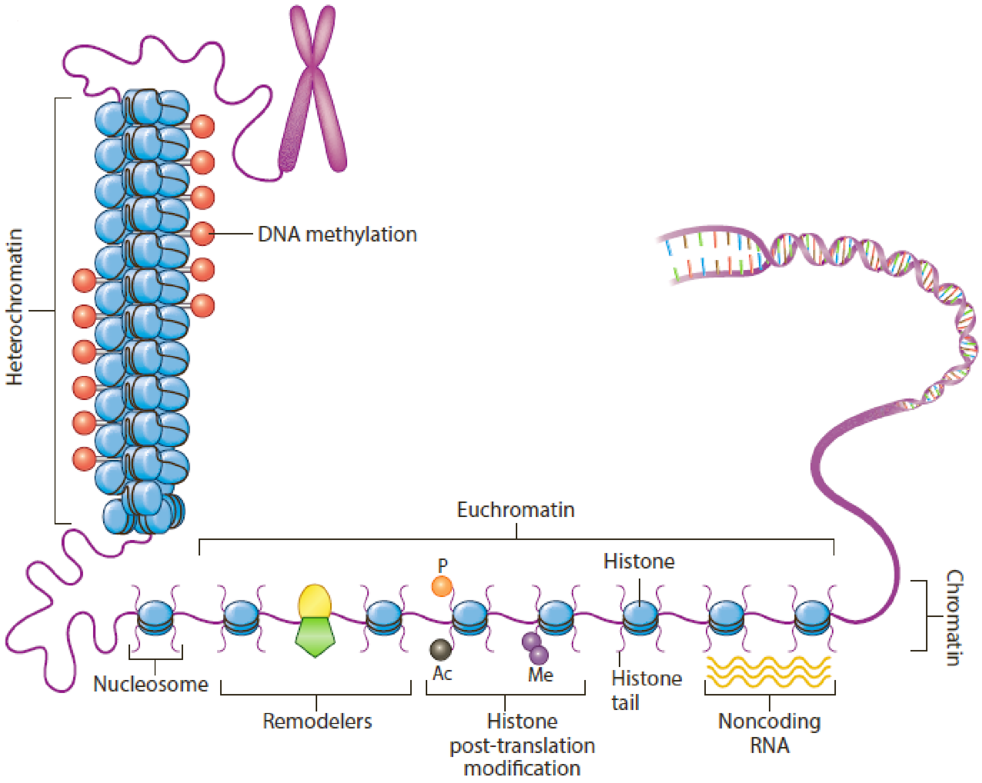


Figure 1.13 Schematic representation of epigenetic modifications. DNA methylation, histone modification, and post-transcriptional regulation by noncoding RNA are reversible alterations which affect gene expression (from Ahuja et al. 2016).

Besides that, epigenetic alterations are also determinant to tumor heterogeneity and different treatment responses. An example is the chemoresistance due to hypermethylation of the Transcription Factor AP-2 epsilon (*TEAP2E*) gene, that occurs in 51% of CRC[110,111].

1.4.1. microRNAs

MicroRNAs (miRNAs) were discovered in *Caenorhabditis elegans* in 1993[112] and are small non-coding ribonucleic acids (RNA) about 21-25 nucleotides in length, which are related to regulation of gene expression through complementary binding to 3'untranslated region (UTR) of its messenger RNA (mRNA) target molecules. The consequence of this binding depends on the complementarity between miRNA and its target. In case of complete complementarity, the most probable effect is mRNA degradation. In contrast, incomplete complementarity leads to translation inhibition[113–115]. Therefore, any alteration in the regulation of these non-coding RNAs may drive changes in gene expression which may lead to silencing or overexpression of many genes.

miRNAs are encoded either in intronic regions or in intergenic regions and are usually transcribed by polymerase II (Pol II), producing primary miRNAs (pri-miRNAs). The pri-miRNA is cleaved by DROSHA, which is constituted by two ribonuclease (RNase) III domains, generating a precursor miRNA (pre-miRNA) which is exported to the cytoplasm, where is recognized by DICER1. This RNase III enzyme cleaves the pre-miRNA, producing an RNA duplex which later associates with RNA-induced silencing complex (RISC). Importantly, this complex will be guided by the guide strand of the mature miRNA incorporated in RISC[116,117] (***Figure 1.14***).

miRNAs have also revealed important in cancer biology, since miRNAs are able to control several targets implicated in tumor growth, invasion, angiogenesis, and immune invasion. Therefore, the function of miRNAs could be considered as tumor suppressor genes or oncogenes, depending on their target. Additionally, recent studies have demonstrated different miRNA patterns between normal and tumor tissue, and that these patterns are also able to distinct tumor types and their subtypes[118].

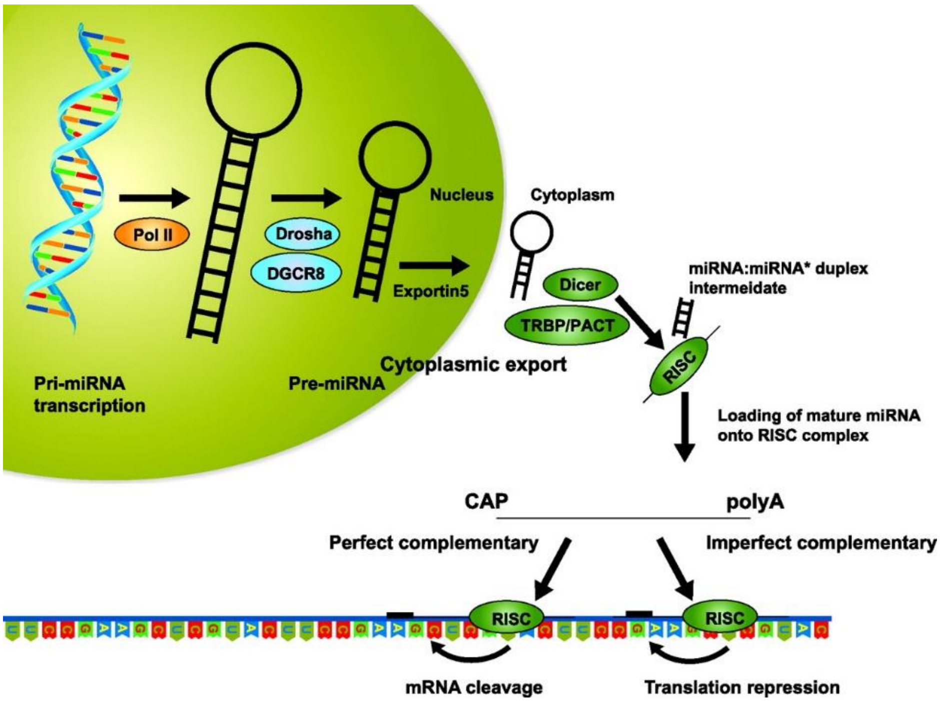


Figure 1.14 miRNA processing. The gene that codifies the miRNA is transcribed originating the pri-miRNA. This is processed by DROSHA in the nucleus and, originating the pre-miRNA.It is exported to the cytoplasm where it is cleaved by DICER and associated to the RISC complex. Lastly, the mature miRNA guides the RISC complex to the target mRNA (from Nelson et al. 2008).

1.4.2. Histone Modifications

Cells do not express all genes at the same time as gene expression depends on the needs of the cell. This is possible due to proteins associated with chromatin called histones, which stabilizes the negative charge of DNA and provides stability to the chromatin.

Histones regulate gene expression through alterations in the chromatin structure, either by condensing the chromatin, which leads to gene inactivation, or by stretching the chromatin, which results in gene activation[104]. Therefore, protein binding sites may be exposed or masked, and consequently gene expression is altered.

A group of 8 histones (an octamer) forms the nucleosome, which comprises two of each H2A, H2B, H3 and H4 histones. Moreover, there is an additional histone, H1, that works as a linker (***Figure 1.15A***). Each one of these histones is susceptible to suffer posttranslational modifications, especially in the N-terminal tails. The impact of these modifications, caused by histone methyltransferases (HMT), histone acetyltransferases

(HAT), histone deacetylases (HDAC), and histone demethylases (HDM), depends on the modification- generally acetylation, methylation, phosphorylation, and ubiquitination- and residue where it takes place- commonly lysine or arginine residues[88,119–121] (***Figure 1.15B***).

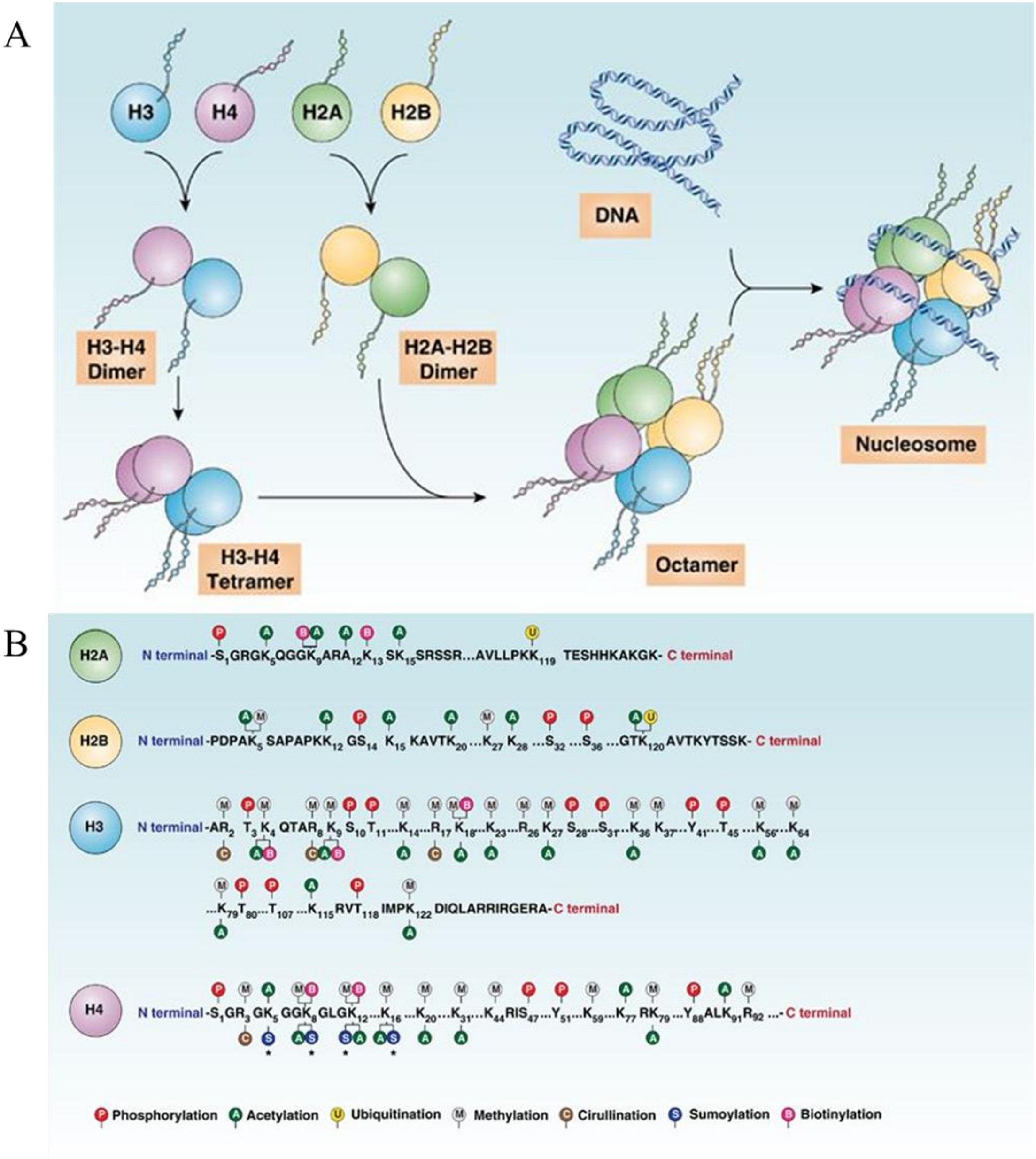

A
H3
H4
H2A
H2B
DNA
H3-H4 Dimer
H2A-H2B Dimer
H3-H4 Tetramer
Octamer
Nucleosome
B
H2A
N terminal $-S_1GRGK_5QGGK_9ARA_{12}K_{13}SK_{15}SRSSR...AVLLPKK_{119}$ TESHHKAKGK- C terminal
H2B
N terminal $-PDPAK_5$ $SAPAPKK_{12}$ GS_{14} K_{15} $KAVTK_{20}...K_{27}$ $K_{28}...S_{32}...S_{36}...GTK_{120}AVTKYTSSK$- C terminal
H3
N terminal $-AR_2$ T_3K_4 $QTAR_8K_9S_{10}T_{11}...K_{14}...R_{17}K_{18}...K_{23}...R_{26}K_{27}S_{28}...S_{31}...K_{36}K_{37}...Y_{41}...T_{45}...K_{56}...K_{64}$
$...K_{79}T_{80}...T_{107}...K_{115}RVT_{118}IMPK_{122}$ DIQLARRIRGERA- C terminal
H4
N terminal $-S_1GR_3GK_5GGK_8GLGK_{12}...K_{16}...K_{20}...K_{31}...K_{44}RIS_{47}...Y_{51}...K_{59}...K_{77}RK_{79}...Y_{88}ALK_{91}R_{92}$ ...- C terminal
P Phosphorylation
A Acetylation
U Ubiquitination
M Methylation
C Cirullination
S Sumoylation
B Biotinylation

Figure 1.15 Nucleosome assembly and post-translational modification of histone tails. (**A**) A nucleosome is an octamer of histones. (**B**) Each histone can suffer post-translational modifications in its tails (from Chen et al. 2014).

Artem Barski has identified histone modifications patterns associated to promoters, insulators, enhancers, and transcribed regions. Modifications such as mono methylation of histone 3 at lysine 4 (H3K4me1), tri methylation of histone 3 at lysine 4 (H3K4me3), and acetylation of histone 3 at lysine 27 (H3K27ac) have been associated to active enhancers, active promoters, and active enhancers and promoters, respectively. In contrast, tri methylation of histone 3 at lysine 27 (H3K27me3), tri methylation of histone 3 at lysine 9 (H3K9me3), and tri methylation of histone 3 at lysine 36 (H3K36me3) have been associated to repressive chromatin[122].

1.4.3. DNA Methylation

DNA methylation consists in the covalent addition of a methyl group to the 5-carbon of a cytosine residue by DNA methyltransferases (DNMT)[123]. That reaction often takes place in CG dinucleotides- CpG sites. These dinucleotides can be located in CpG Islands, which are DNA regions constituted by more than 50% of CG dinucleotides in a minimum length of 200-500 bases[124,125]. CpGs are usually methylated in human normal cells and located outside of the promoter. Paradoxically, CpG Islands are usually unmethylated and overlapping promoter regions (***Figure 1.16A***)[81,108].

The DNMTs enzyme family includes DNMT1, DNMT3a, and DNMT3B, where DNMT1 is responsible for maintaining methylation patterns during replication, and DNMT3a and DNMT3b are responsible for *de novo* methylation[81,108].

This epigenetic mechanism is essential during the embryonic development, imprinting, X-chromosome inactivation, and suppression of repetitive element transcription. Importantly, there are evidences that DNA methylation plays a key role in cancer development[86,126].

DNA methylation is often associated with gene inactivation, particularly when it takes place at the gene promoter (***Figure 1.16B***). Nevertheless, there are evidences that promoter hypermethylation can also lead to gene activation. The result of DNA methylation seems to be dependent on the region where it happens. This means that DNA methylation may affect regulatory regions, blocking protein binding sites due to the recruitment of methyl-CpG-binding domain (MBD) proteins. If the region affected is an activator binding site, the gene will be not expressed. In contrast, hypermethylation of gene promoters on repressor binding site, prevents the DNA access, leading to gene expression. Another explanation of gene activation associated with promoter hypermethylation, suggested by Bert et. al, is that regional hypermethylation forces the activation of alternative transcription start sites (TSS-***Figure 1.17***)[108,125,127].

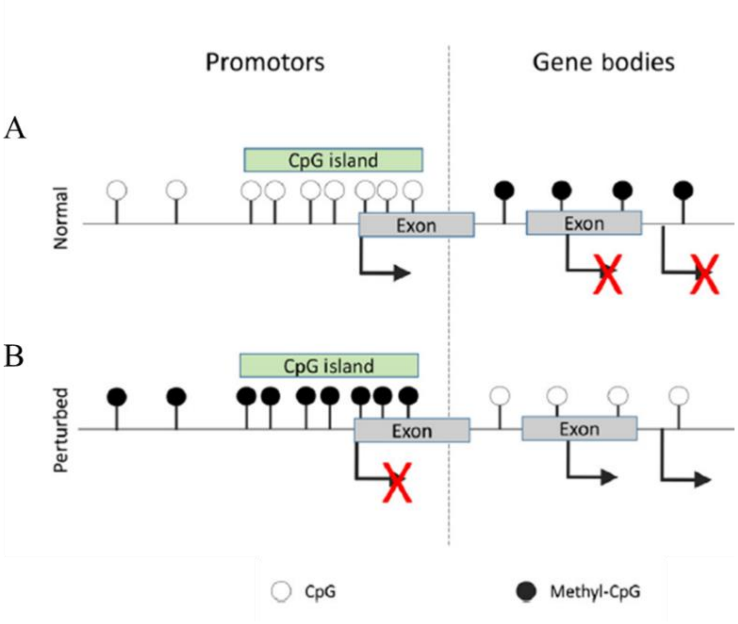


Figure 1.16 Hypermethylation can lead to gene inactivation. (A) In normal cells, there is a generalized methylation of the gene body, in contrast to a promotor region that is un-methylated. **(B)** In the case of diseases like cancer, promotor region can be aberrantly methylated and the region of gene body un-methylated. Consequently, that gene can be silenced. (Adapted from McBryan et al. 2014)

In cancer, this epigenetic mechanism is frequently deregulated, leading to an unbalance of gene expression. When aberrant DNA methylation changes- either hypermethylation or hypomethylation- occurs in driver genes, the normal cellular function is altered, and tumorigenesis may arise.

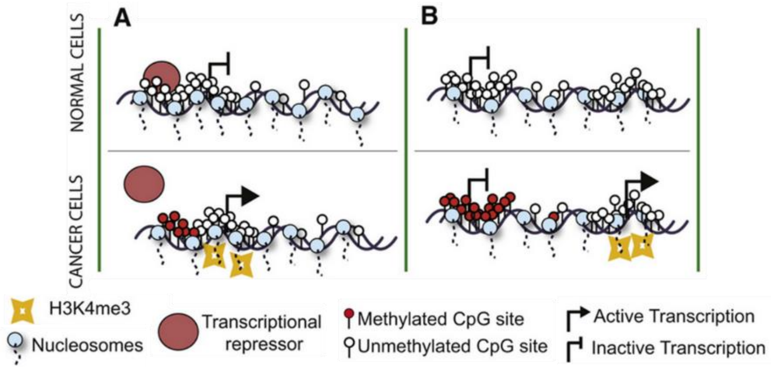


Figure 1.17 Promoter hypermethylation can be associated with gene activation. (A) Promoter hypermethylation can happen in a region of repressors binding. Hence, when hypermethylation occurs, the transcriptional repressor is blocked, leading to abnormal gene activation. (B) Promoter hypermethylation can lead to the gene activation through alternative TSS. (Adapted from Bert et al. 2013)

1.4.4. Epigenetic Alterations and Colorectal Cancer

Epigenetic events, which may also occur during normal ageing, have been associated with higher risk of cancer. In the early 80's hypomethylation was associated to cancer. Moreover, in 1986, hypermethylation of calcitonin was associated with tissue-specific gene silencing. Nevertheless, hypermethylation was also associated with inactivation of tumor suppressor genes, through observations based on the *Rb* promoter[101].

Several studies have shown that DNA methylation patterns of many genes become aberrant during carcinogenesis, including genes belonging to the Wnt and Ras signaling pathways, DNA repair genes, and cell cycle-related genes[73]. Specifically, in CRC, aberrantly methylated genes as Integrin Subunit Alpha 4 (*ITGA4*), O-6-Methylguanine-DNA Methyltransferase (*MGMT*), Solute Carrier Family 5 Member 8 (*SLC5A8*), and Secreted Frizzled Related Protein 2 (*SFRP2*) have been reported since early stages. Therefore, it is evident that methylation is involved in the initiation and progression of CRC. However, among all abnormally methylated genes, there is no evidence that a specific functional class of genes is more affected during specific steps of CRC initiation or progression[81].

In addition, studies have suggested that DNA methylation as well as genetic alterations play a role in cancer progression and metastasis. Methylated genes as *TIMP* Metallopeptidase Inhibitor 3 (*TIMP3*), Inhibitor of DNA Binding 4 (*ID4*) and Interferon Regulatory Factor 8 (*IRF8*) are more frequent in advanced CRC than in adenomas, providing clonal growth advantage. Despite association to advanced stages, DNA methylation seems to be most prevalent in CRC initiation rather than in its progression[81].

1.4.5. Epigenetic biomarkers as predictors of clinical outcome

A biomarker is any substance, structure, or process that can be estimated, and used in order to identify normal biological processes, pathogenic processes, treatment responses, or evolution of the disease[128]. Hence, epigenetic biomarkers could be useful in the clinic for diagnosis, prognosis or prediction of responsiveness to therapy.

Specifically, in CRC, in spite of efforts to identify new biomarkers capable to detect or predict progression and therapy response, there is a lack of accurate biomarkers. Moreover, as reported before, the detection of CRC in early stages is crucial to the efficiency of the treatment.

Therefore, measuring DNA methylation levels of specific sites can be a potential biomarker, since DNA methylation patterns are found usually altered in CRC. Not only that, DNA methylation levels can be detected through non-invasive methods, such as evaluation of tumor-derived cell-free from blood or feces, making it a good biomarker[129,130].

Until now, few epigenetic biomarkers have been reported in CRC, including aberrant methylation of Septin 9 (*SEPT9*) detected in plasma (sensitivity and specificity of almost 90%), methylation of *SFRP2* detected in serum and fecal DNA (sensitivity of almost 67%)[131], methylation of Thrombomodulin (*THBD*) detected in blood (sensitivity of 74% to stage I/II CRC at a specificity of 80%)[132] and methylation of Syndecan 2 (*SDC2*) also detected in blood (sensitivity of 92% for stage I)[133]. An epigenetic biomarker based on aberrant methylation of Vimentin (*VIM*) is currently commercialized in the United States for early detection of CRC with 83% of sensitivity and 82% of specificity[81]. Among all existent epigenetic biomarkers, two meta-analysis estimate that the sensitivity to diagnose CRC and adenomas is about 62%-75%[134,135].

1.5. Databases Analysis and Statistic Methodologies

In the last years, the amount of data available in public repositories has increased enormously. The Cancer Genome Atlas (TCGA), The National Center for Biotechnology

Information (NCBI), Ensembl, Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), and Kyoto Encyclopedia of Genes and Genomes (KEGG) are good examples of these repositories where vast amounts of clinical and biological information is present with relatively easy accessibility[136,137].

Indeed, before the new computational era and the web repositories, the costs and time spent collecting data could restrict scientific production. Nowadays, the improvement of online platforms not only allows for its access anywhere in the world but also significantly reduced the time spent in the data processing (***Figure 1.18***)[138,139].

However, to handle the amount of available data can be a challenge. Therefore, researchers from different fields of knowledge have been developing bioinformatic approaches combined with statistical analysis[139]. This promotes the creation of methodologies that can test hypothesis to be validated by inferential statistics with the help of computational tools[136]. Some of these tools comprise different programming languages (including R language used in the present work) since statistical software has its limitations regarding the management and processing of big data[140,141].

Moreover, TCGA, the principal repository of data collection used in these studies has become an important repository for cancer research, since it stores more than 2.5 petabytes of information, including genetic, epigenetic and clinical data, allowing for the analysis of more than 440 thousand variables (e.g. in case of DNA methylation).

Also, other methodological strategies were implemented in order to validate our analysis that included:

1. **Univariate** approaches to analyze the population of the study, including the socio-demographic and clinical characterization;
2. **Bivariate** approaches to analyze the linear association of gene expression and DNA methylation in order to select which genes and CpG sites are differentially expressed and methylated, respectively;
3. **Multivariate** approaches to observe the relationship between genes and CpG sites selected, and the sample distribution.

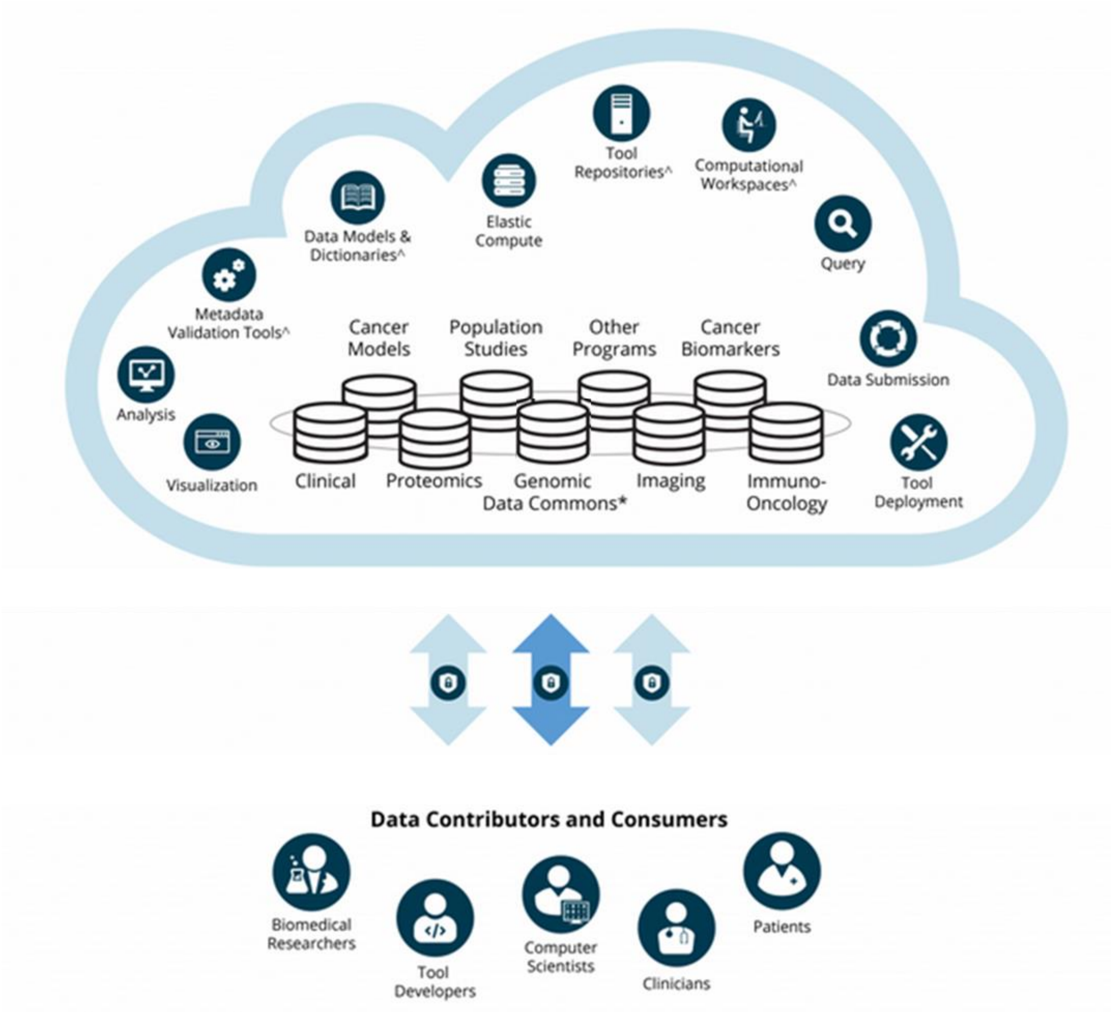


Figure 1.18 A network of data availability. Data can be upload and download from different databases to be used in different research projects (adapted from https://cbiit.cancer.gov/ncip/cancer-research-data-commons).

This characterization allowed to identify genes and CpG sites that can discriminate tumor samples from normal samples, and, subsequently, among different stages, to identify all significant associations between gene expression and DNA methylation. Finally, using multivariate approaches, we aimed to prove that the distribution of samples is able to reflect the intrinsic distancing of gene expression of DNA methylation of that samples, which were previously selected.

2. CHAPTER II- OBJECTIVES

Even though several studies have identified aberrant expression of several genes in CRC to be associated to epigenetic events, including aberrant DNA methylation, there is still a lot to know about how DNA methylation impacts gene expression during CRC carcinogenesis. Moreover, there is a lack of biomarkers that can accurately identify patients with early stages of CRC or predict patient outcome. Therefore, the identification of new biomarkers that play a key role in the initiation and progression of CRC might improve personalized treatments.

Here, we hypothesize that there is an epigenetic roadmap in CRC progression. Therefore, we:

1. **Performed a genome-wide analysis** of both DNA methylation and gene expression, contributing to the knowledge of epigenetic dynamics on CRC;
2. **DNA methylation and gene expression of CRC patients in different stages of its progression** through developing a Bioinformatics based tool (script);
3. **Identified epigenetic mutations** responsible for CRC initiation and progression;
4. **Identified potential epigenetic biomarkers** to help in the diagnosis and prediction of CRC progression.

3. CHAPTER III- MATERIALS AND METHODS

3.1. Data Collection

We analyzed whole-genome DNA methylation (*Illumina Infinium HumanMethylation 450K array*) and gene expression (*Illumina HiSeq*) of CRC patients ("TCGA-COAD" and "TCGA-READ") publicly available in The Cancer Genome Atlas database (TCGA; http://cancergenome.nih.gov/). The data was obtained through *TCGAbiolinks* package using the functions: *GDCquery*, *GDCdownload*, and *GDCprepare*[142].

3.1.1. The Cancer Genome Atlas

TCGA is an American database funded by National Cancer Institute (NCI) from the National Institute of Health (NIH) and the National Human Genome Research Institute (NHGRI) with the main aim to understand the genetics of cancer. TCGA has over 2.5 petabytes of data, aggregating 33 different tumor types, including 10 rare cancers, based on paired tumor and normal tissue sets collected from 11,000 patients[143]. Importantly, it means that normal tissue samples were obtained from cancer patients (***Figure 3.1***).

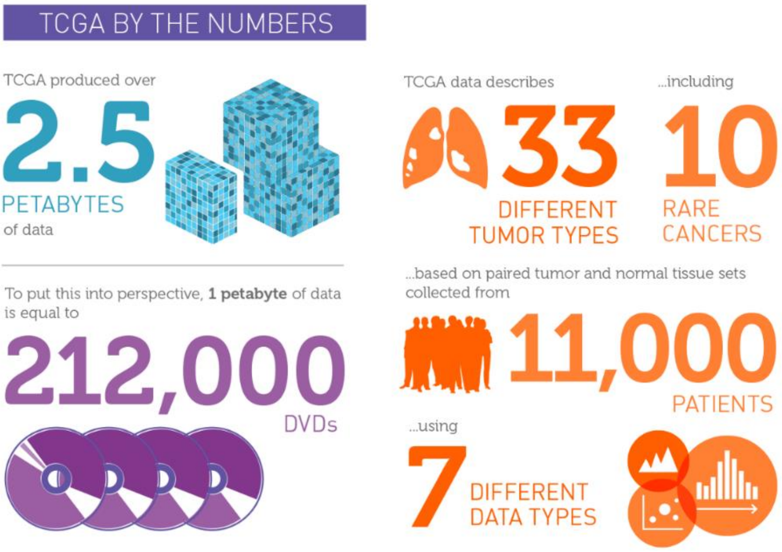


Figure 3.1 TCGA by numbers. TCGA provides data for different tumor types regarding a significant amount of patients (adapted from https://cancergenome.nih.gov/abouttcga).

a. Gene Expression

Gene expression values were generated through *Illumina HiSeq 2000 RNA Sequencing* platform by the University of North Carolina TCGA genome characterization center. These values were normalized (RSEM normalized count) and transformed by the application of base 2 logarithm to the expression measures plus one. The purpose of this transformation is to linearize the relationship between gene expression and DNA methylation[144]. Furthermore, *UCSC Xena HUGO probeMap* was used to map genes.

b. DNA Methylation

DNA methylation values (beta values) were measured experimentally through *Illumina Infinium HumanMethylation450K* array and were noted using *BeadStudio software*. The beta values range from 0 to 1, depending on the intensity ratio between the methylated bead type to the combined locus intensity, meaning that higher beta values correspond to higher methylation levels whereas lower beta values correspond to lower methylation levels[145].

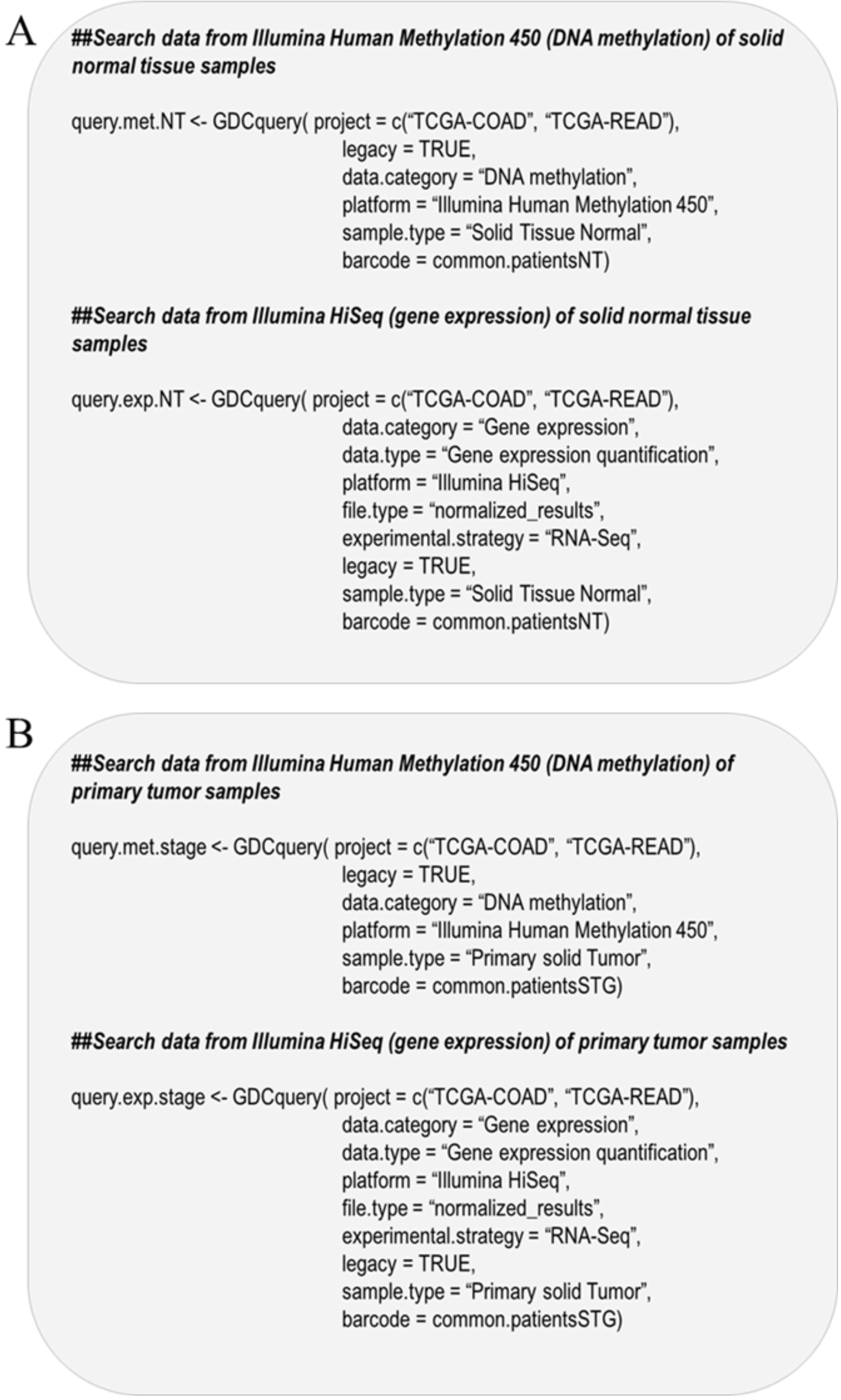
c. TCGAbiolinks Package

To download clinical data, from both DNA methylation and gene expression datasets for both colon (TCGA-COAD) and rectum (TCGA-READ) cohorts, we used the Bioconductor package *TCGAbiolinks* version 2.7.2 available for R programming[140]. The package is a software tool developed to query, download and analyze genomic and epigenomic data, once TCGA is a challenge for bioinformaticians, clinicians and molecular biologists. When compared to other tools developed to analyze TCGA open access data, *TCGAbiolinks* is the most complete package[142].

3.2. Patient Selection

In order to group CRC patients, the clinical information was imported to R using *GDCquery_clinic* function provided by *TCGAbiolinks* package or were obtained online (https://xena.ucsc.edu/).

The patients were separated into 5 groups according to CRC staging: Solid Tissue Normal, and stage I-IV (Primary Solid Tumor). To perform this analysis, only patients with tumor stage information, DNA methylation and gene expression data were included (***Figure 3.2A, Figure 3.2B***).

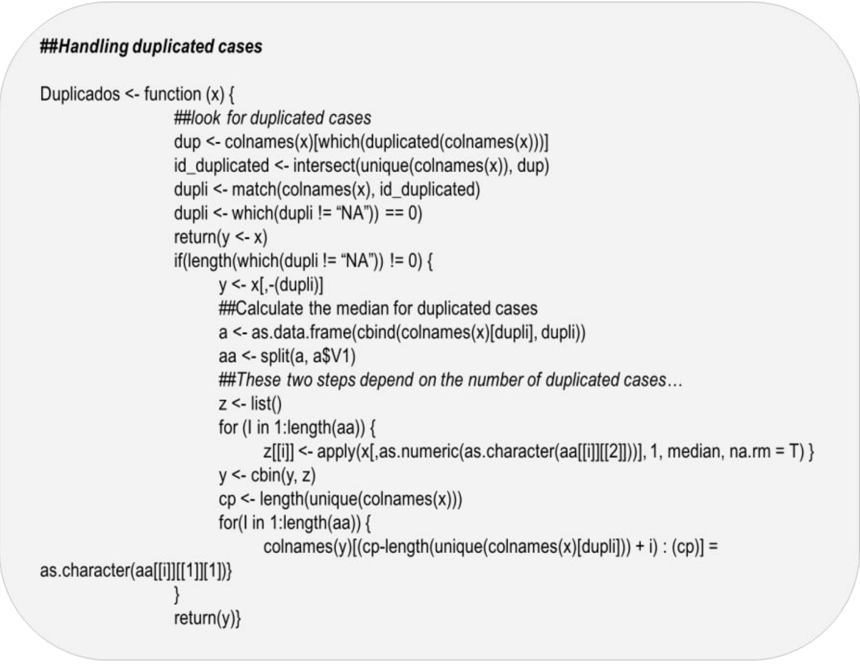


```
##Search data from Illumina Human Methylation 450 (DNA methylation) of solid
normal tissue samples

query.met.NT <- GDCquery( project = c("TCGA-COAD", "TCGA-READ"),
                          legacy = TRUE,
                          data.category = "DNA methylation",
                          platform = "Illumina Human Methylation 450",
                          sample.type = "Solid Tissue Normal",
                          barcode = common.patientsNT)

##Search data from Illumina HiSeq (gene expression) of solid normal tissue
samples

query.exp.NT <- GDCquery( project = c("TCGA-COAD", "TCGA-READ"),
                          data.category = "Gene expression",
                          data.type = "Gene expression quantification",
                          platform = "Illumina HiSeq",
                          file.type = "normalized_results",
                          experimental.strategy = "RNA-Seq",
                          legacy = TRUE,
                          sample.type = "Solid Tissue Normal",
                          barcode = common.patientsNT)
```

```
##Search data from Illumina Human Methylation 450 (DNA methylation) of
primary tumor samples

query.met.stage <- GDCquery( project = c("TCGA-COAD", "TCGA-READ"),
                          legacy = TRUE,
                          data.category = "DNA methylation",
                          platform = "Illumina Human Methylation 450",
                          sample.type = "Primary solid Tumor",
                          barcode = common.patientsSTG)

##Search data from Illumina HiSeq (gene expression) of primary tumor samples

query.exp.stage <- GDCquery( project = c("TCGA-COAD", "TCGA-READ"),
                          data.category = "Gene expression",
                          data.type = "Gene expression quantification",
                          platform = "Illumina HiSeq",
                          file.type = "normalized_results",
                          experimental.strategy = "RNA-Seq",
                          legacy = TRUE,
                          sample.type = "Primary solid Tumor",
                          barcode = common.patientsSTG)
```

Figure 3.2 Sample selection using R programming. DNA methylation and gene expression data were obtained for the groups of patients: (A) normal samples, and (B) primary tumor samples.

Therefore, the analysis of TCGA-COAD and TCGA-READ cohorts was based on 21 Solid Tissue Normal patients, and 347 Primary Solid Tumor (54 stage I, 131 stage II, 111 stage III, and 51 stage IV). Importantly, few patients had DNA methylation measures from more than one sample. In these cases (duplicated cases), the DNA methylation measure was substituted by the median value of duplicated cases for each methylation probe (***Figure 3.3***).



```
##Handling duplicated cases

Duplicados <- function (x) {
                        ##look for duplicated cases
                        dup <- colnames(x)[which(duplicated(colnames(x)))]
                        id_duplicated <- intersect(unique(colnames(x)), dup)
                        dupli <- match(colnames(x), id_duplicated)
                        dupli <- which(dupli != “NA”)) == 0)
                        return(y <- x)
                        if(length(which(dupli != “NA”)) != 0) {
                                y <- x[,-(dupli)]
                                ##Calculate the median for duplicated cases
                                a <- as.data.frame(cbind(colnames(x)[dupli], dupli))
                                aa <- split(a, a$V1)
                                ##These two steps depend on the number of duplicated cases…
                                z <- list()
                                for (I in 1:length(aa)) {
                                        z[[i]] <- apply(x[,as.numeric(as.character(aa[[i]][[2]]))], 1, median, na.rm = T) }
                                y <- cbin(y, z)
                                cp <- length(unique(colnames(x)))
                                for(I in 1:length(aa)) {
                                        colnames(y)[(cp-length(unique(colnames(x)[dupli])) + i) : (cp)] =
as.character(aa[[i]][[1]][1])}
                        }
                        return(y)}
```

Figure 3.3 The approach to remove duplicated cases. DNA methylation measurements of different samples from the same patient was substituted by the median value of duplicated cases for each methylation CpG site.

3.3. Study Pipeline

After patient selection, we analyzed both DNA methylation and gene expression data to identify which CpG sites and genes were differentially methylated and expressed in Primary Solid Tumor comparatively to Solid Tissue Normal.

Initially, absence of gene location, or probe identification were exclusion criteria for CpG sites selection whereas absence of gene name was exclusion criteria for gene selection. Importantly, before any statistical analysis, outlier values were removed from both expression and methylation databases. Then the set of selected genes and CpG sites were analyzed (***Figure 3.4***).

Secondly, a statistical analysis was performed to identify if there is statistical evidence for differences on gene expression/ DNA methylation between normal and tumor tissue samples of each stage of the CRC. At that point, only genes/CpG sites with false discovery rate (FDR) lower than 5% were considered (***Figure 3.4***).

Then, the mean for each gene/CpG site was calculated for normal and tumor samples (calculated separately depending on disease stage), aiming to measure the base 2 logarithm of fold-change (referred only as fold-change) and Δβ values, respectively. Thus, genes with fold-change absolute value higher than 1.5 and CpG sites with Δβ absolute value higher than 0.2 were considered as differentially expressed and methylated between normal and tumor samples, respectively. Importantly, only differentially expressed genes which contained CpG sites differentially methylated were considered. Similarly, only CpG sites located in genes differentially expressed were considered (***Figure 3.4***).

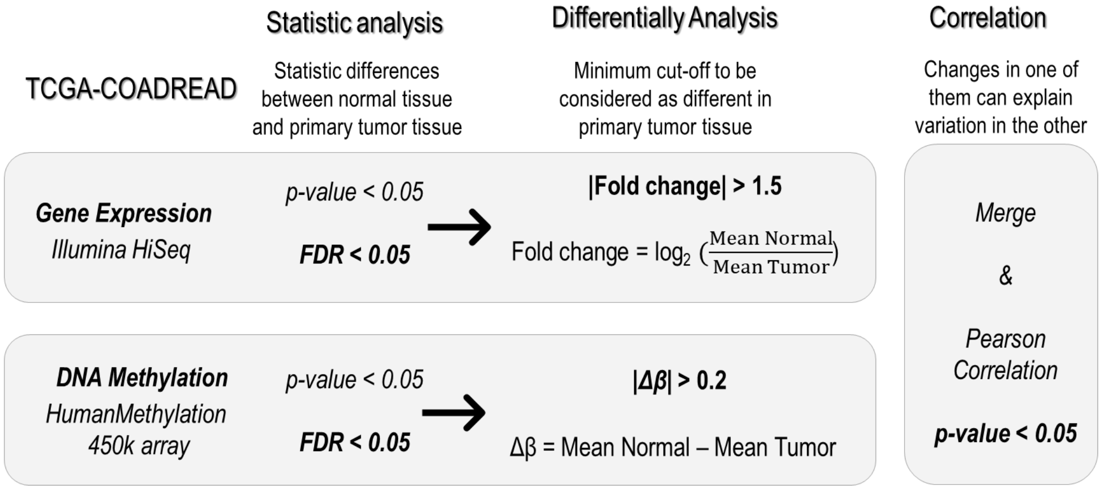

Statistic analysis
Differentially Analysis
Correlation
TCGA-COADREAD
Statistic differences between normal tissue and primary tumor tissue
Minimum cut-off to be considered as different in primary tumor tissue
Changes in one of them can explain variation in the other
Gene Expression
Illumina HiSeq
p-value < 0.05
FDR < 0.05
|Fold change| > 1.5
Fold change = $\log_2(\frac{\text{Mean Normal}}{\text{Mean Tumor}})$
DNA Methylation
HumanMethylation 450k array
p-value < 0.05
FDR < 0.05
|Δβ| > 0.2
Δβ = Mean Normal – Mean Tumor
Merge
&
Pearson Correlation
p-value < 0.05

Figure 3.4 Study Pipeline. Both whole-genome *Illumina HiSeq* and *Illumina Infinium HumanMethylation 450K* array data were analyzed. Firstly, were selected genes and CpG sites with statistic differences (FDR cut-off < 0.05). Then, an additional cut-off was applied for gene expression values- fold-change absolute value higher than 1.5- and for DNA methylation- Δβ absolute value higher than 0.2. After, only genes with CpG sites differentially methylated as well as CpG sites located in differentially expressed genes were admitted. At last, Pearson coefficient was measured, and a *p-value* cut-off was applied (*p-value* <0.05).

Lastly, a Pearson correlation test was performed as criteria of both CpG sites and genes selection. This analysis was executed aiming to identify a relationship between methylation and expression levels in tumor tissue. At this step, a *p-value* cut-off lower than 0.05 was established (***Figure 3.4***).

Moreover, *enrichR* R package was used to clarify in which pathways (KEGG 2016) selected genes were involved. The same package was used to identify enriched biological processes (GO Biological Process 2018)[146].

Another package available for R software, *RISmed*, was used to investigate which selected genes had not yet been reported in PubMed database as associated with CRC or with cancer in general[147].

Thereafter, in order to identify potential good diagnostic biomarkers that could discriminate tumor from normal tissue, a receiver operating characteristic (ROC) curve analysis was performed using *pROC* R package[148]. Here, it was applied a cut-off, in which it was considered as potential good diagnostic biomarkers genes/CpG sites with an area under the curve (AUC) higher than 0.8[149]. Then, *survival* R packages was used to perform Cox regression analysis in order to identify prognostic biomarkers (overall survival and recurrence free survival)[150,151]. In this analysis, the threshold to divide patients into two groups was based on the median, and a *p-value* cut-off at 0.05 was considered.

3.4. Statistical Testing

In order to identify if there were any evidences of statistically significance differences between Solid Tissue Normal and Primary Solid Tumor, test hypothesis were formulated, and statistical hypothesis tests were performed[152]. The main aim of statistical hypothesis tests is to achieve characteristics of a certain population by statistical inference[153]. This means that statistical hypothesis tests are performed based on a sample and extrapolated to a population.

Therefore, statistical testing implies the formulation of a null hypothesis, the selection of the most appropriate statistical test, and the *p-value* estimation to asses if the null hypothesis is true. The null hypothesis generally asserts that there are no differences between our groups. In opposition, the alternative hypothesis sustains that there are differences between our groups, and that those differences did not arise due to chance. The most common way to decide if the null hypothesis is rejected is based on *p-value*[154–156].

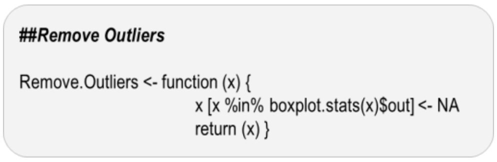
The choice of statistical test depends on the data distribution. If the data are normally distributed, parametric tests are most adequate. On the other hand, if the data are "free distributed", nonparametric tests should be used. Parametric methods are based on mean estimator, being more powerful. In contrast, nonparametric methods use the median, being more robust, especially in case of existence of outliers[157].

All statistical analysis were performed using functions provided by available packages for R programming[140]. Importantly, in all statistical analysis, it was used a *p-value* <0.05 as the significance level.

3.4.1. Handling Outliers

Experimental data can present observations with values deviated from the other ones- outliers. Indeed, these cases may influence our analysis. For example, in the concrete case of the current study, the presence of outliers can influence the fold-change measured. Hence, the decision if a specific gene is or not differentially expressed can be biased. Therefore, to deal with this issue, outliers should be identified and properly handled[158–160].

Here, outliers were assessed using *boxplot.stats* function provided by R base[140]. Moreover, these values were not replaced, meaning that missing values were introduced in our data (***Figure 3.5***).



```
##Remove Outliers

Remove.Outliers <- function (x) {
                        x [x %in% boxplot.stats(x)$out] <- NA
                        return (x) }
```

Figure 3.5 Function to remove outliers. Outlier values were substituted by missing data.

3.4.2. Shapiro-Wilk Normality Test

Shapiro test was used to assess whether the sample is normally distributed. The null hypothesis assumes that the population is normally distributed, meaning that lower *p-values* suggest that the sample is not normally distributed whereas higher *p-values* suggest that the sample is normally distributed. Importantly, compared to other tests to assess normality, Shapiro-Wilk test is considered the be the most powerful test, independently of either the size or the type of distribution of the sample[161].

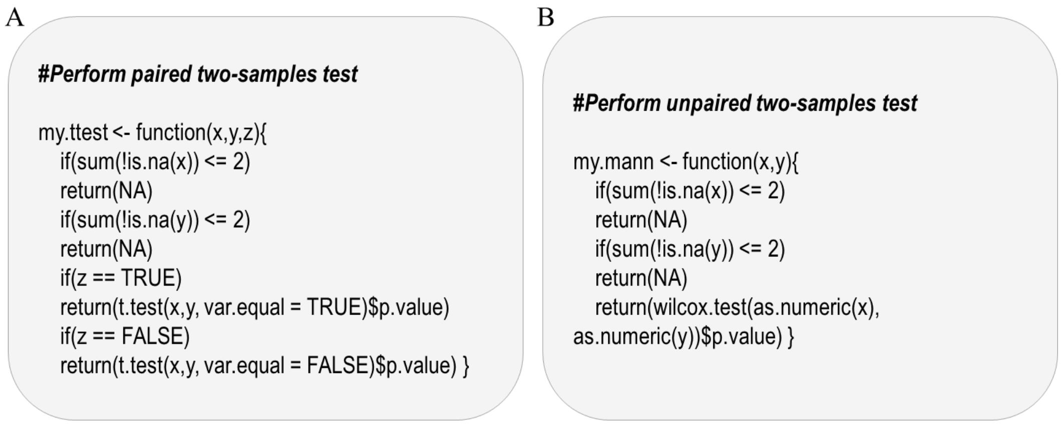
Here, the function *shapiro.test* provided by *stats* package was used[140]. Importantly, few variables were excluded from this analysis due to:

a. missing data- only variables with at least 3 non-missing values were considered[161];
b. all non-missing data to be equal 0.

3.4.3. Paired and Unpaired Two-Sample Tests

The Wilcoxon rank sum test, also known as Mann-Whitney U test, is a non-parametric test which was used with the aim to compare two samples (e.g. to compare DNA methylation measures of both normal and tumor tissue) that are not normally distributed. The null hypothesis is that the distribution of both groups is the same[162]. By contrast, when we intend to compare two normally distributed samples, a parametric test was applied. If both samples are normally distributed and they had equal variances, the unpaired Student's t test was applied. Otherwise, when samples were normally distributed, but they had unequal variances, the Welch test was applied[163]. The null hypothesis in t test is that the mean of both samples is equal[164].

In order to perform both Wilcoxon signed-rank and t test, *wilcox.test* (***Figure 3.6A***) and *t.test* (***Figure 3.6B***) functions was used[140]. Here, it was only considered variables with two or more observations in both samples (normal and tumor tissue).



```
#Perform paired two-samples test

my.ttest <- function(x,y,z){
   if(sum(!is.na(x)) <= 2)
   return(NA)
   if(sum(!is.na(y)) <= 2)
   return(NA)
   if(z == TRUE)
   return(t.test(x,y, var.equal = TRUE)$p.value)
   if(z == FALSE)
   return(t.test(x,y, var.equal = FALSE)$p.value) }
```

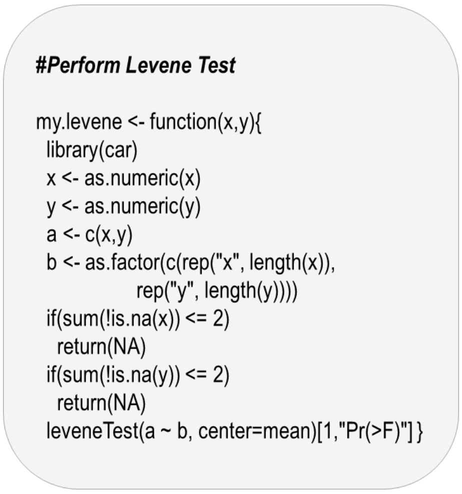
```
#Perform unpaired two-samples test

my.mann <- function(x,y){
   if(sum(!is.na(x)) <= 2)
   return(NA)
   if(sum(!is.na(y)) <= 2)
   return(NA)
   return(wilcox.test(as.numeric(x),
as.numeric(y))$p.value) }
```

Figure 3.6 Adapted functions to compare two-samples. (A) Function to perform a t-test considering the variance homogeneity. (B) Function to perform Wilcoxon signed-rank test.

3.4.4. Levene Test

The Levene test was used to assess the equality of variances of two samples. This test was performed before the implementation of t test to decide if the Welch test should be applied. The null hypothesis is that the variances of two samples are equal[165].

Levene test was implemented using the *leveneTest* function provided by *car* R package (***Figure 3.7***)[166].



```
#Perform Levene Test

my.levene <- function(x,y){
  library(car)
  x <- as.numeric(x)
  y <- as.numeric(y)
  a <- c(x,y)
  b <- as.factor(c(rep("x", length(x)),
                  rep("y", length(y))))
  if(sum(!is.na(x)) <= 2)
    return(NA)
  if(sum(!is.na(y)) <= 2)
    return(NA)
  leveneTest(a ~ b, center=mean)[1,"Pr(>F)"] }
```

Figure 3.7 Adapted function to perform the Levene test.

3.4.5. Correction for Multiple Testing

Many simultaneous statistical tests might originate *p-values* less than the critical value by chance, hence resulting in the rejection of null hypothesis even if it is true (false positive or type I error). In case of a *p-value* of 0.05, when 100 null hypotheses are simultaneous tested, the chance of commit a type I error is 5%. Due to this fact, each *p-value* must be corrected. One approach to adjust the *p-value* is using the false discovery rate (FDR) method, which corrects the falsely rejected hypothesis[155,167].

In order to correct multiple testing effect, *p-values* obtained were corrected by FDR method using the function *p.adjust* provided by *stats* package[140].

3.4.6. Pearson Correlation

Pearson correlation test is a parametric statistic test which aims to detect if two pared continuous variables are linearly associated. The relationship between the variables are measured by correlation coefficient that range from -1 (perfect negative correlation) to 1 (perfect positive correlation). A correlation coefficient of 0 means that there is not a linear correlation between the variables (***Table 3.1***). Importantly, correlation analysis do not indicate which variable vary in response to the other one[168–170].

Table 3.1 Correlation coefficient interpretation. The relationship between two pared continuous variables can be measured by correlation coefficient. This coefficient ranges from -1 to 1, being dependent on the degree of association between these variables (adapted from Mukaka et al. 2012).

|Correlation Coefficient|[1]	Meaning
0.00 – 0.30	Negligible correlation
0.30 – 0.50	Low correlation
0.50 – 0.70	Moderate correlation
0.70 – 0.90	High correlation
0.90 – 1.00	Very high correlation

[1] *Absolute value*

In this specific case, the Pearson correlation was performed to identify which DNA methylation variations are associated to unbalance in gene expression in Primary Solid

Tumor. The null hypothesis is that correlation coefficient is 0[155]. To perform Pearson correlation, the function *cor.test* provided by *stats* R package was used[140].

3.4.7. Receiver operating characteristic (ROC) curve analysis

ROC curve analysis was used as a technique to measure the quality of a diagnostic biomarker (e.g. identify CpG sites as a potential diagnostic tool). This curve is represented based on the true positive fraction (TPF), the same of sensitivity, and false positive fraction (FPF), given by 1 – specificity. The TPF is the ratio between the number of true positive decision and the number of actually positive cases. Otherwise, FPF is the ratio/division between the number of false positive decisions (considered as positive, but actually are negative) and the number of actually positive cases. Therefore, each coordinate (x,y) in a plot of the ROC curve corresponds to a pair of TPF and FPF[171].

Moreover, the relationship between these two measures- accuracy- can be determined through the area under the ROC curve (AUC). These accuracy measure range from 0.5 to 1. As described in the ***Table 3.2***, lower AUC values represent an inaccurate test -bad diagnostic tool- whereas higher AUC values represent an accurate test- good diagnostic tool[149,172].

Table 3.2 Classification of the diagnostic accuracy. The area under the ROC curve, an accuracy measure, is used to classify the diagnosis potential of a specific tool (Khouli et al. 2009).

Area Under the Curve (AUC)	Meaning
0.5 - 0.6	Failed
0.6 – 0.7	Poor
0.7 – 0.8	Fair
0.8 – 0.9	Good
0.9 - 1	Excellent

The ROC curve analysis was implemented using the *roc* function provided by *pROC* R package[148]. Importantly, an AUC cut-off of 0.8 was established to consider good biomarkers[149].

3.4.8. Overall Survival and Recurrence Free Survival Analysis

The prognostic ability of selected genes and CpG sites was tested fitting a Kaplan-Meier (KM) analysis, a logrank test, and a Cox proportional hazards regression model. This approach compares the survival time, with possible censoring, of two groups divided according to a specific threshold of the predictor variable. Firstly, the KM, an estimator that was used to determine the survival function for our two groups (better prognosis and worst prognosis), was drawn[173,174]. Then, it was used the logrank test in order to compare the survival functions of both groups. This test is a non-parametric test where the null hypothesis is that the distribution of both functions is the same[174,175].

Finally, the effect of the selected factor (methylation values of a CpG site or expression values of a gene) on the survival was measured through hazard ratio (HR). If a hazard ratio is equal 1, that predictor variable has no effect on survival. A hazard ratio lower than 1 is a bad prognostic factor for group 1 when compared to group 2 whereas a hazard ratio higher than 1 is a good prognostic factor for group 1 when compared to group 2[176–178].

This analysis was performed using *coxph* function provided by *survival* R package. Then, to estimate the survival/recurrence proportion, it was used the *surfit* function provided by *survival* R package. Kaplan-Meier curves were done using the *ggsurvplot* function provided by *survminer* R package[150,151,179]. Importantly, the threshold was based on the median value of the predictor variable.

3.4.9. HJ-Biplot and Hierarchical Clusters

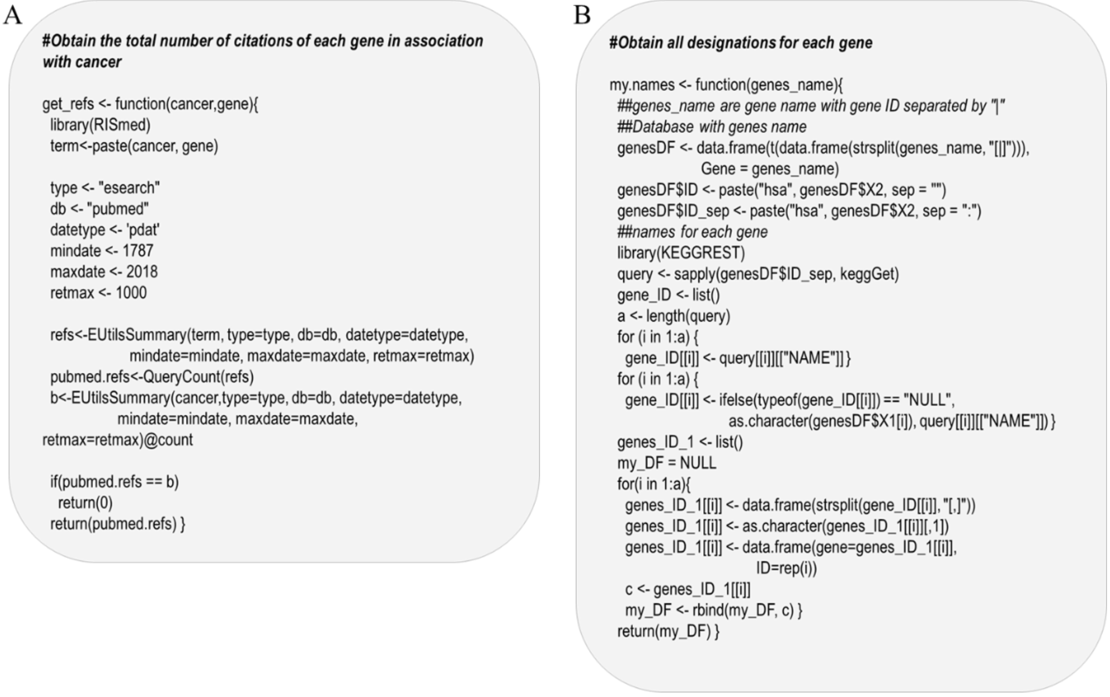
HJ-Biplot is a data reduction technique to analyze, in a multivariate perspective, the samples distribution considering all relationships of variables. A hierarchical cluster analysis[156] was performed considering the patient coordinates[180]. Thus, we used the Ward method to aggregate samples in clusters considering the square Euclidean distance. Moreover, in this approach was fixed the contributions, of factor to the element, over than 0.7. This contribution allows to know which variables are more directly related to each axis, and, consequently, it also allows to identify which variables are the most responsible by distributing the individuals on a reduced space, for posterior orthogonal projections in each variable. Importantly, it was only selected genes which have CpG sites with contributions of factor to the element over 0.7, as well as CpG sites which are located in genes with contributions of factor to the element over 0.7.

HJ-Biplot was implemented using the function *HJ.Biplot* provided by *MultBiplotR* R package. The Analysis of Hierarchical Clusters was executed using the function AddCluster2Biplot provided by the same R package[181]. Importantly, this technique does not deal with missing data. Hence, missing observations were replaced by the median value of the respective variable.

3.5. Citation Tool

RISmed is a text data mining package able to interact with Pubmed, a public query database of scientific literature. This package allows to count how many times a term was referred in Pubmed abstracts and, when possible, in PubMed articles[147].

EUtilsSummary and *QueryCount* functions were used to investigate which genes were mentioned in both CRC and cancer in general on Pubmed published articles from 1787 to June 2018, and how many times have been mentioned (***Figure 3.8*****A**). The keywords used include "cancer", "colorectal cancer", "rectum cancer", and "colon cancer". Importantly, the function *keggGet*, provided by *KEGGREST* package available for R software, was also used to identify all names given to each gene (***Figure 3.8*****B**)[182]. All these names were used in our analysis.



```
#Obtain the total number of citations of each gene in association
with cancer

get_refs <- function(cancer,gene){
  library(RISmed)
  term<-paste(cancer, gene)

  type <- "esearch"
  db <- "pubmed"
  datetype <- 'pdat'
  mindate <- 1787
  maxdate <- 2018
  retmax <- 1000

  refs<-EUtilsSummary(term, type=type, db=db, datetype=datetype,
                    mindate=mindate, maxdate=maxdate, retmax=retmax)
  pubmed.refs<-QueryCount(refs)
  b<-EUtilsSummary(cancer,type=type, db=db, datetype=datetype,
                 mindate=mindate, maxdate=maxdate,
retmax=retmax)@count

  if(pubmed.refs == b)
    return(0)
  return(pubmed.refs) }
```

```
#Obtain all designations for each gene

my.names <- function(genes_name){
  ##genes_name are gene name with gene ID separated by "|"
  ##Database with genes name
  genesDF <- data.frame(t(data.frame(strsplit(genes_name, "[|]"))),
                      Gene = genes_name)
  genesDF$ID <- paste("hsa", genesDF$X2, sep = "")
  genesDF$ID_sep <- paste("hsa", genesDF$X2, sep = ":")
  ##names for each gene
  library(KEGGREST)
  query <- sapply(genesDF$ID_sep, keggGet)
  gene_ID <- list()
  a <- length(query)
  for (i in 1:a) {
    gene_ID[[i]] <- query[[i]][["NAME"]] }
  for (i in 1:a) {
    gene_ID[[i]] <- ifelse(typeof(gene_ID[[i]]) == "NULL",
                         as.character(genesDF$X1[i]), query[[i]][["NAME"]]) }
  genes_ID_1 <- list()
  my_DF = NULL
  for(i in 1:a){
    genes_ID_1[[i]] <- data.frame(strsplit(gene_ID[[i]], "[,]"))
    genes_ID_1[[i]] <- as.character(genes_ID_1[[i]][,1])
    genes_ID_1[[i]] <- data.frame(gene=genes_ID_1[[i]],
                                ID=rep(i))
    c <- genes_ID_1[[i]]
    my_DF <- rbind(my_DF, c) }
  return(my_DF) }
```

Figure 3.8 Functions to obtain the number of citations in PubMed for each gene. (A) Function to obtain all designations for each gene. (B) Function of obtain the total number of citations of each gene in association with a cancer term.

3.6. Enrichment Analysis

To better understand the functional profile of a set of genes differentially expressed and methylated in our CRC cohort, it was performed an enrichment analysis. In this step, the function *enrichr*, provided by *enrichR* package, was used to obtain and display the most statistically significant enriched pathways (KEGG_2016) and biological processes (GO_Biological_Process_2018)[146].

4. CHAPTER IV- RESULTS

4.1. Clinical Features

CRC patients analyzed in this study were characterized according to the parameters showed in ***Table 4.1***. The clinical data was exported from TCGA website, being processed and analyzed through R programming[140].

Stage I patients are a mean age of 66 years old. Approximately 60% of our group is constituted by male, and 40% of female patients. Almost all patients are white- 76%- against 15% of black or African American. Additionally, colon is the more affected region comprising 80% of all cases. The anatomic subdivision more common is the cecum (35%), being splenic flexure the less common (2%-***Table 4.1***).

Moreover, the mean age of stage II CRC is about 66 years old. In this stage, the genders are more balanced, with 51% of males and 49% of females. Regarding to the race, it is observed that white patients constitute the great majority (68%). In addition, colon is the most frequently affected region (81%), being sigmoid colon the most predominant subdivision (18%-***Table 4.1***).

Furthermore, in stage III, was observed a mean age of 63 years old. Additionally, 55% of all patients are male against 45% of females. In terms of race, 76% are white patients, keeping the patterns observed in the other stages. Moreover, it is observed that in 70% of all patients, the tumor site is colon against 30% of rectum. Regarding to the anatomic subdivision, as in the stage I, cecum is the region most afflicted (***Table 4.1***).

Regarding to stage IV, the mean age of these patients is 61 years old. In addition, 57% are male patients whereas 43% are females. The predominant race continues to be white patients (71%). Moreover, 75% of all cases are colon tumor, being sigmoid colon, the most anatomic subdivision affected (27%-***Table 4.1***).

At last, the patients with normal samples are a mean age of 68 years old, and 24% are male patients. Furthermore, 48% of these patients are white, 14% are black or african American, and 38% have no information available about their race. Moreover, as normal samples are collected from patients with the disease, there are tumor related information about these patients. Specifically, about 90% of them are colon cancer, and sigmoid colon is the anatomic subdivision most affected (38%-***Table 4.1***).

Table 4.1 Descriptive statistics about patients, as well as the distribution of patients by stage.

Groups n (%)	**Normal** 21 (5.7)	**Stage I** 54 (14.7)	**Stage II** 131 (35.6)	**Stage III** 111 (30.1)	**Stage IV** 51 (13.9)
Age (mean ± sd[1] years)	68 ± 13	66 ± 13	66 ± 13	63 ± 13	61 ± 13
< 65 years old	7 (33%)	24 (44%)	57 (44%)	61 (55%)	31 (61%)
> 65 years old	14 (67%)	30 (56%)	74 (56%)	50 (45%)	20 (39%)
Gender					
Male	12 (57%)	32 (59%)	67 (51%)	61 (55%)	29 (57%)
Female	9 (43%)	22 (41%)	64 (49%)	50 (45%)	22 (43%)
Race					
Black or African American	3 (14%)	8 (15%)	19 (15%)	22 (20%)	11 (22%)
American Indian or Alaska Native	0 (0%)	0 (0%)	0 (0%)	1 (1%)	0 (0%)
White	10 (48%)	41 (76%)	89 (68%)	85 (76%)	36 (70%)
Asian	0 (0%)	0 (0%)	11 (8%)	1 (1%)	0 (0%)
Not Reported	8 (38%)	5 (9%)	12 (9%)	2 (2%)	4 (8%)
Anatomic Subdivision					
Ascending Colon	2 (10%)	6 (11%)	24 (18%)	10 (9%)	6 (12%)
Descending Colon	1 (5%)	2 (4%)	5 (4%)	5 (4%)	2 (4%)
Cecum	4 (18%)	19 (35%)	21 (16%)	23 (21%)	9 (17%)
Hepatic Flexure	2 (10%)	2 (4%)	7 (5%)	7 (6%)	1 (2%)
Rectosigmoid Junction	0 (0%)	6 (11%)	10 (8%)	17 (15%)	7(14%)
Rectum	2 (10%)	4 (7%)	14 (11%)	15 (14%)	6 (12%)
Sigmoid Colon	8 (37%)	10 (18%)	33 (25%)	22 (20%)	14 (27%)
Splenic Flexure	0 (0%)	1 (2%)	3 (2%)	0 (0%)	1 (2%)
Transverse Colon	0 (0%)	3 (6%)	9 (7%)	9 (8%)	1 (2%)
Not Reported	2 (10%)	1 (2%)	5 (4%)	3 (3%)	4 (8%)
Tumor Site					
Colon	19 (90%)	43 (79%)	106 (81%)	78 (70%)	38 (75%)
Rectum	2 (10%)	10 (19%)	25 (19%)	33 (30%)	12 (25%)

[1]standard deviation

4.2. Epigenetic Roadmap in Colorectal Cancer

Aiming to investigate which CpG sites are differentially methylated in tumor tissue when compared to normal tissue, we performed a comparative analysis between normal tissue and each stage of CRC.

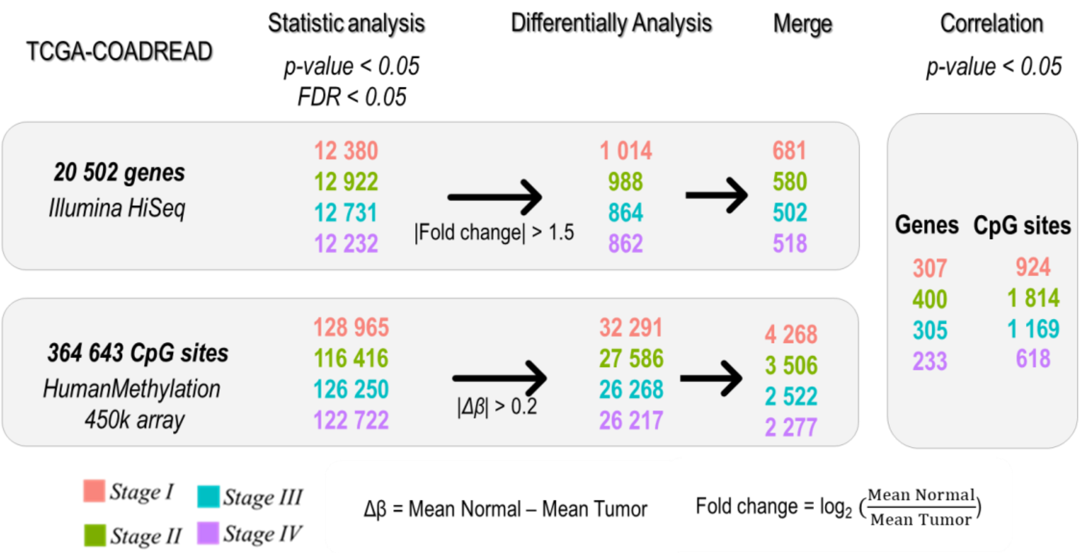


Figure 4.1 Study pipeline indicating CpG sites and genes selected for each stage. The pipeline presented at Figure 3.2 was applied to CpG sites and genes available in the TCGA database. Here, it is shown how many CpG sites and genes were selected in each step of that pipeline.

We analyzed 364,643 probes and 20,502 genes, from CRC patients in different stages of the disease. To consider genes as differentially expressed and CpG probes as differentially methylated a *p-value* cut-off lower than 0.05, a fold-change absolute value cut-off higher than 1.5, and a Δβ absolute value cut-off higher than 0.2 were defined (***Figure 3.4***, ***Figure 4.1***).

The comparative analysis with normal tissue suggested that 4,268 CpG sites corresponding to 681 genes are differentially methylated and expressed in stage I of the disease (***Figure 4.1***). Additionally, the correlation analysis shows that DNA methylation alterations of 924 CpG sites are correlated to gene expression changes of 307 genes (***Figure 4.1***, ***Appendix 1***, and ***Appendix 2***). The majority of these CpG sites are hypermethylated in tumor tissue (597 hypermethylated and 327 hypomethylated- ***Figure 4.2A1***, ***Appendix 1***) and associated to down-regulated genes (226 down-regulated and 81 up-regulated, ***Appendix 2***). Interestingly, when we determine the position of those CpG sites across the 307 genes, we found that the majority of them are in the gene body (43.8%), followed by regions near the TSS (34.6%). The five prime untranslated region (5’UTR), three prime untranslated region (3’UTR), and first exon are regions with less differentially methylated CpG sites (10.4%, 4.7%, and 6.5%, respectively; ***Figure 4.3A***).

Additionally, from the 924 CpG sites, 543 are negatively correlated with gene expression, while 381 are positively correlated (***Figure 4.2A2***, ***Appendix 1***).

Moreover, when we investigate the top 15most differentially expressed genes (***Table 4.2***), we found that most genes are up-regulated in tumor tissue, excepting *CACNG5, GABRG1* and *HTR3B,* with an absolute fold-change value ranging from 4.47 to 6.62. Interestingly, altogether these genes have 27 CpG sites differentially methylated associated with them ($0.20 < |\Delta\beta| < 0.41$), which are linearly correlated with gene expression levels ($0.27 < |\rho| < 0.64$). Moreover, our analysis shows that both *SOX14* and *HTR3B* are strongly regulated by DNA methylation. Specifically, *SOX14* is negatively correlated to the methylation levels of 6 CpG sites whilst *HTR3B* is positively correlated with 4 CpG sites in CRC tissue (***Appendix 1***, ***Appendix 2***).

Regarding to stage II of the disease, our analysis suggested that 3,506 CpG sites located throughout 580 differentially expressed genes are differentially methylated compared to normal tissue (***Figure 4.1***). However, only 1,814 CpG sites (1,366 hypermethylated, and 448 hypomethylated- ***Figure 4.2B1***, ***Appendix 3***) are significantly correlated to 400 genes (294 down-regulated, and 106 up-regulated- ***Figure 4.1***, ***Appendix 4***), being 1,305 negatively correlated and 509 positively correlated (***Figure 4.2B2***, ***Appendix 3***). Furthermore, the majority of CpG sites selected are located in the gene body (36.5%), TSS1500 (23.6%), and TSS200 (19.5%). Only 10.1% are in the 5'UTR, 6.4% in the first exon, and 3.9% in the 3'UTR region (***Figure 4.3B***).

Additionally, our analysis identified genes represented in ***Table 4.2*** as the top 15 genes more differentially expressed in tumor tissue when compared to normal tissue ($4.47 < |\text{fold-change}| < 6.57$). These top 15 genes are linearly correlated to methylation levels of 33 CpG sites ($0.21 < |\Delta\beta| < 0.40$; $0.18 < |\rho| < 0.66$). Surprisingly, *PTF1A* is negatively correlated with methylation levels of 13 CpG sites (***Appendix 3***, ***Appendix 4***).

Additionally, from normal tissue to stage III, 2,522 CpG sites located on 502 genes are statistically differentially methylated and expressed (***Figure 4.1***). Moreover, 1,169 CpG sites (758 hypermethylated and 412 hypomethylated- ***Appendix 5***) are also correlated to gene expression changes of 305 genes (205 down-regulated and 102 up-regulated- ***Figure 4.1***, ***Figure 4.2C1***, and ***Appendix 6***), being 690 negatively correlated and 479 positively correlated (***Figure 4.2C2***, ***Appendix 5***). Additionally, these CpG sites distributed along the gene, being the gene body (39.9%) the most enriched region, followed by TSS1500

(21%), TSS200 (20.8%), 5’UTR (9.1%), first exon (4.7%), and 3’UTR (4.5%, ***Figure 4.3C***).

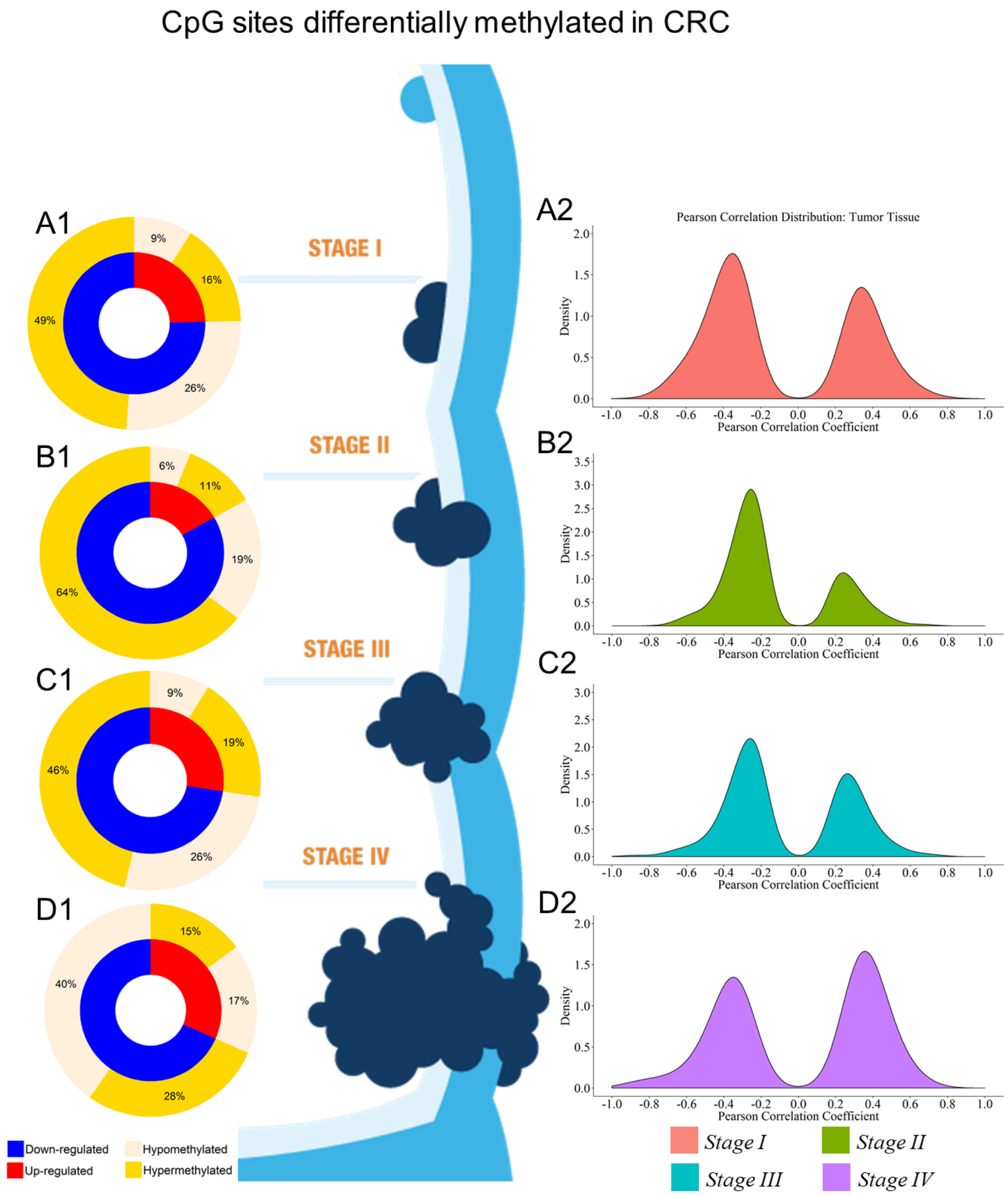


Figure 4.2 Characterization of CpG sites differentially methylated throughout CRC development. Status of both CpG sites and the respective gene in (A1) stage I, (B1) stage II, (C1) stage III, and (D1) stage IV of CRC development. Moreover, it is also represented the Pearson correlation- between gene expression and methylation values of each CpG site-distribution in (A2) stage I, (B2) stage II, (C2) stage III, and (D2) stage IV.

Furthermore, this analysis reveals that the top 15 genes more differentially expressed on stage III, represented in the ***Table 4.2***, have absolute fold change values ranging from 4.34 to 6.79. These genes are correlated with methylation levels of 34 CpG sites (0.20 < |Δβ| < 0.41; 0.19 < |ρ| < 0.90). Among all these genes, *PTF1A* is the gene correlated with more CpG sites (11 CpG sites- ***Appendix 5***, ***Appendix 6***).

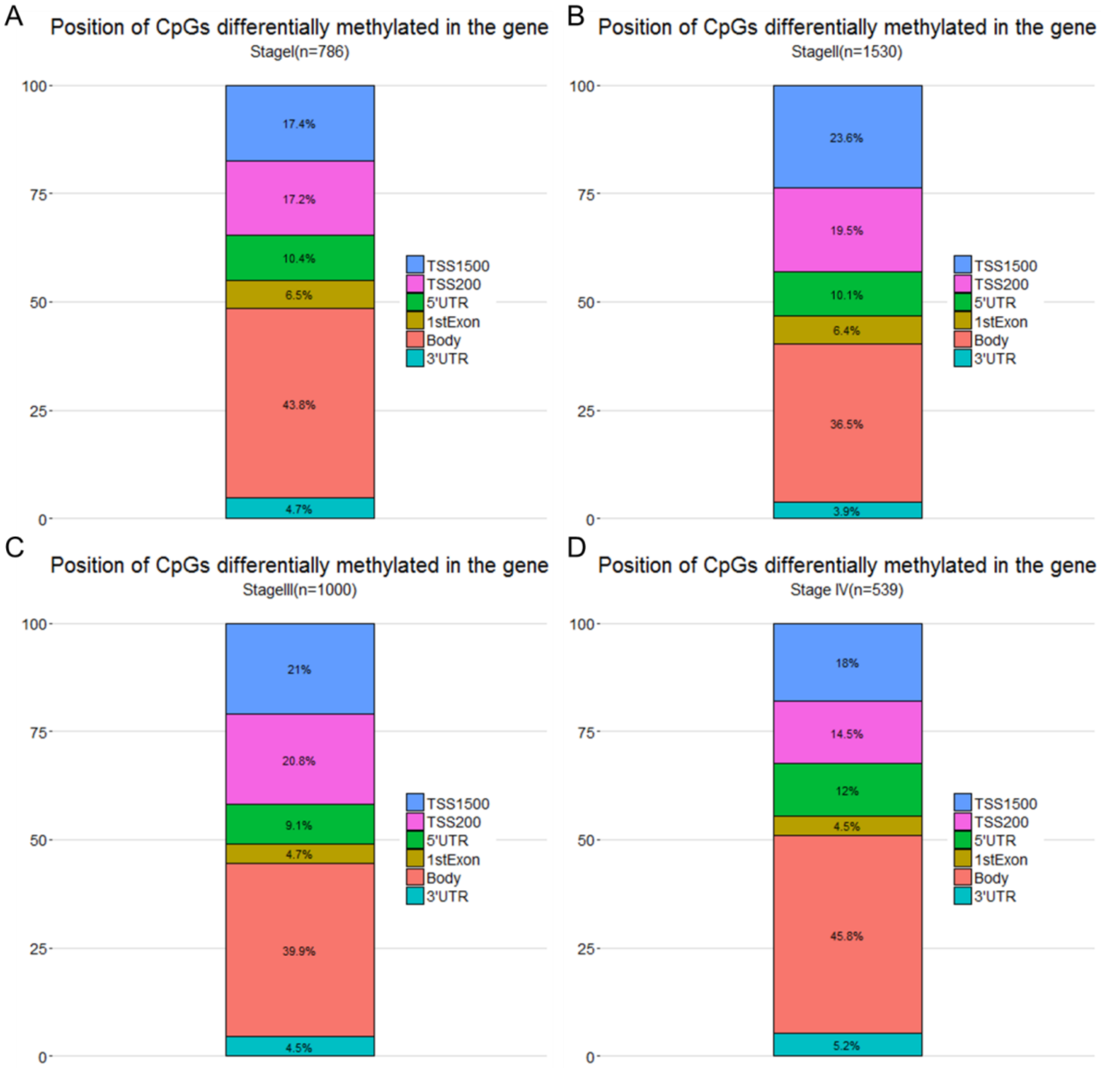


Figure 4.3 Localization of CpG sites differentially methylated in the gene. Distribution of CpG sites differentially methylated (A) in stage I, (B) stage II, (C) stage III, and (C) stage IV. Only CpG sites with one gene location were taken into account. **TSS1500** and **TSS200**: probes located within 1500 and 200 base pairs from the transcription start site, respectively; **5'UTR** and **3'UTR**: five and three prime untranslated regions, respectively.

Ultimately, it was identified 2,277 CpG sites related to 518 genes differentially expressed in stage IV when compared to normal tissue (***Figure 4.1***). From these set of genes and

CpG sites, were identified 618 CpG sites (266 hypermethylated, and 352 hypomethylated- ***Appendix 7***) which are statistically correlated to gene expression of 233 genes (146 down-regulated, and 87 up-regulated-***Figure 4.1***, ***Figure 4.2D1***, and ***Appendix 8***), being 288 negatively correlated and 330 positively correlated (***Figure 4.2D2***, ***Appendix 7***). Similarly, to previously stages, the gene body (45.8%) is where most events occur, followed by TSS1500 (18%), TSS200 (14.5%), 5'UTR (12%), 3'UTR (5.2%), and first exon (4.5%; ***Figure 4.3D***).

Table 4.2 The top 15 most differentially expressed genes for each stage of the disease

Stage I	**Stage II**	**Stage III**	**Stage IV**
FEZF1	*FEZF1*	*FEZF1*	*KRTAP3-1*
SOX14	*SOX14*	*LOC84931*	*ZIC5*
LOC84931	*SPRR1A*	*SOX14*	*PTF1A*
SPRR1A	*LOC84931*	*SPRR1B*	*DKK4*
GBX2	*GBX2*	*C14orf105*	*NKX2-8*
ZIC5	*ZIC5*	*ZIC5*	*DIRC1*
CACNG5	*C14orf105*	*KRTAP3-1*	*SPRR3*
PGLYRP3	*SPRR3*	*DIRC1*	*HTR2C*
GABRG1	*DKK4*	*PTF1A*	*LY6G6E*
SEMG2	*DIRC1*	*ONECUT3*	*HOXC13*
ONECUT3	*OTOP3*	*ELF5*	*TBX5*
HTR3B	*PTF1A*	*PGLYRP3*	*ELF5*
DIRC1	*SEMG2*	*NXPH1*	*SPERT*
SPRR3	*KRTAP3-1*	*LY6G6E*	*ONECUT3*
KRTAP3-1	*PGLYRP3*	*NKX2-8*	*FGF3*

In the ***Table 4.2***, the top 15 genes, which have absolute fold change values ranging from 4.51 to 5.75, were identified. Additionally, 40 CpG sites are correlated either positively or negatively with these top 15 genes ($0.20 < |\Delta\beta| < 0.43$; $0.28 < |\rho| < 0.90$). Specifically, *ELF5* is strongly negatively correlated with methylation levels of 9 CpG sites as well as *FGF3* (***Appendix 7***, ***Appendix 8***).

4.3. Nervous System Related Functions are Enriched in CRC

Aiming to clarify the biological relevance of genes involved in cancer progression, a functional enrichment analysis was performed using *enrichR* package available for R software[146].

This analysis showed that the great majority of enriched pathways are common across CRC progression. In detail, "Neuroactive ligand-receptor interaction" is significantly enriched in all stages, being the most significantly enriched pathway (adjusted *p-value* < 0.05) in all stages (***Figure 4.4***, ***Appendix 9***, ***Appendix 10***, ***Appendix 11***, and ***Appendix 12***). The nicotine addiction pathway is also present among all stages of the disease. Salivary secretion is significantly enriched in both stages II and IV, cAMP signaling pathway is also significantly enriched in stage II, and amyotrophic lateral sclerosis (ALS) is enriched in stage I (***Figure 4.4***, ***Appendix 9***, ***Appendix 10***, ***Appendix 11***, and ***Appendix 12***).

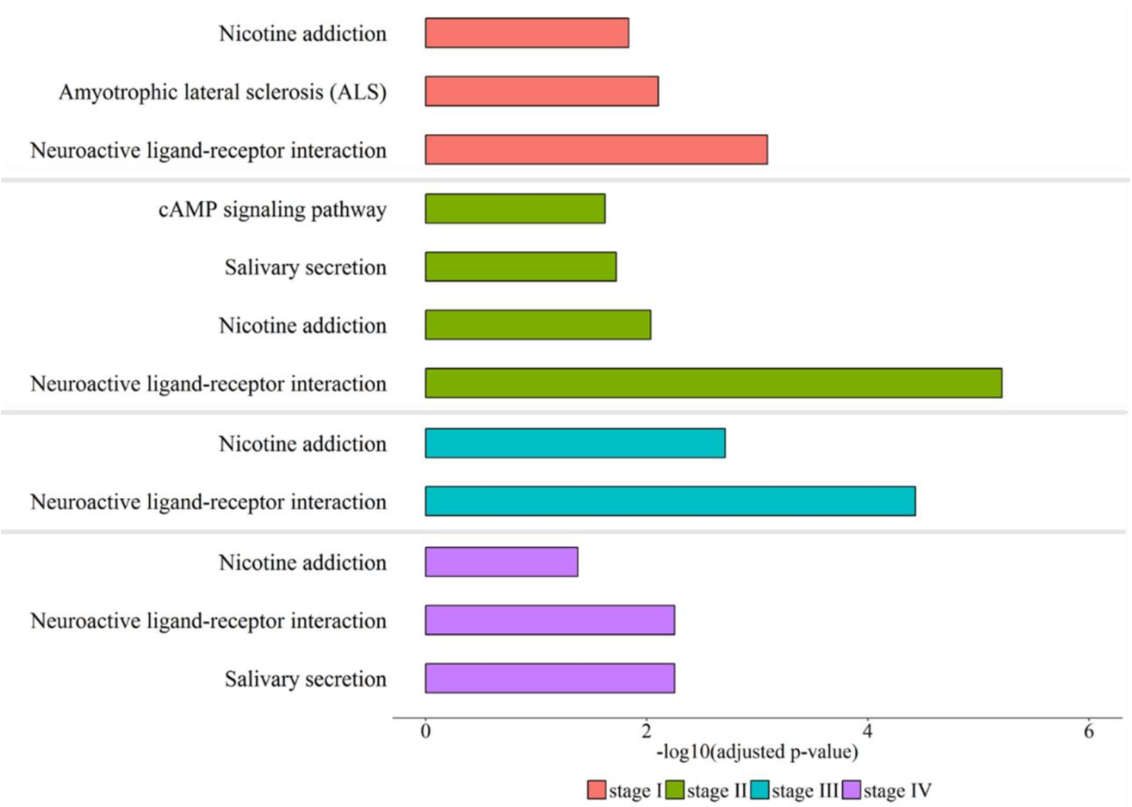


Figure 4.4 Enriched pathways across colorectal cancer development. The most significant enriched pathways (adjusted p-value < 0.05) in colorectal cancer in each stage.

Moreover, when the focus is on the top 9 of the most significantly enriched biological process of each stage (adjusted *p-value* lower than 0.05) gene ontology (GO), we found "dopaminergic neuron differentiation" as the most enriched biological process, considering the ratio of overlapped genes. Interestingly, this is significantly enriched in all stages of CRC development, although stage II, and stage III are stages with more genes that play a role in these functions (both overlap 7/21, adjusted *p-value* 4.94e-05, 6.77e-06, respectively- ***Figure 4.5***, ***Appendix 13***, ***Appendix 14***, ***Appendix 15***, and ***Appendix 16***).

Curiously, other functions are also significantly over-represented in all stages of CRC development, meaning that there are biological processes altered since early stages which are maintained across CRC progression. Specifically, anterograde trans-synaptic signaling, chemical synaptic transmission, and neuron differentiation are altered processes common to all stages (***Figure 4.5***, ***Appendix 13***, ***Appendix 14***, ***Appendix 15***, and ***Appendix 16***).

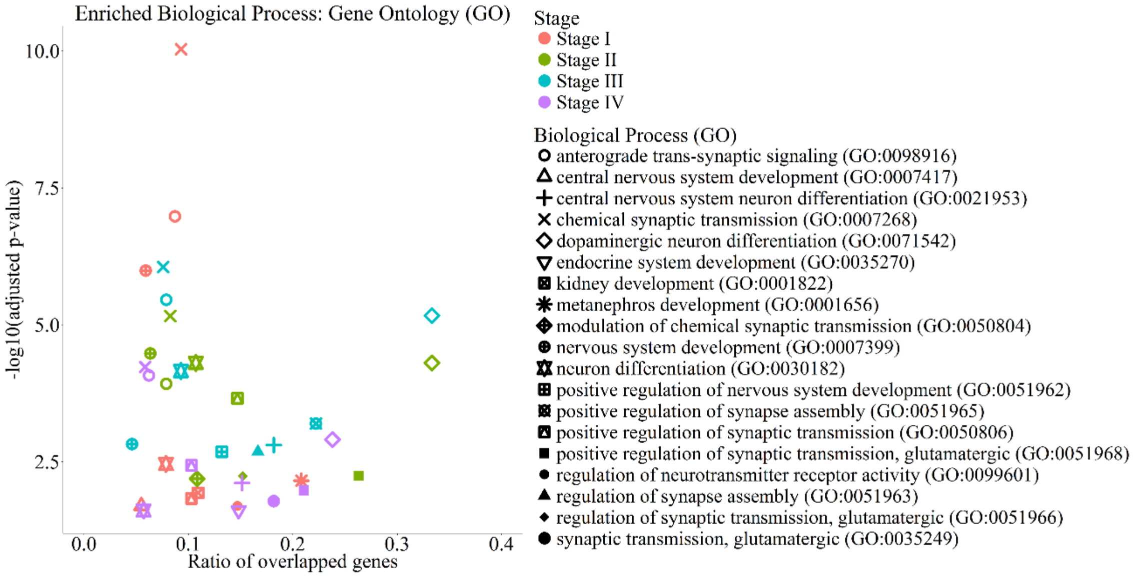


Figure 4.5 Enriched biological process: gene ontology (GO). The top 9 more enriched biological processes for each stage of colorectal cancer development. Colors represent each stage of the disease whereas shape represents the enriched biological process. Ratio of overlapped genes (overlapped genes/total genes).

4.4. Identification of potential New Biomarkers for CRC

After identifying the genes differentially expressed and methylated in each stage of CRC, we were interested in assessing if they had already been associated with CRC or any other cancer in general. Our results showed that out of the 598 genes, 87 of them had not been associated neither CRC nor other cancers. On the other hand, 511 genes were mentioned in cancer related articles. From this set of genes, 278 also appears related with CRC whereas 233 have not yet been associated to these terms (***Figure 4.6A***, ***Appendix 17***, and ***Appendix 18***).

Then the top 20 genes most strongly associated with CRC (*COMP*, *CD5L*, *CALCA*, *ACTBL2*, *SLC6A2*, *F2*, *KRT17*, *TMEFF2*, *ELANE*, *AICDA*, *C10orf90*, *PRB2*, *CRP*, *MMP7*, *KRT24*, *SCN11A*, *FOXD3*, *GPR149*, *CASR*, *HCRT*) and the bottom 20 genes, that had not yet been previously associated with CRC (*ADRB3*, *AKR1CL1*, *ALOXE3*, *ASPG*, *ASTN1*, *ATCAY*, *B3GALT1*, *BAI3*, *C13orf36*, *CACNG7*, *CADM2*, *CALY*, *CAMKV*, *CHRM2*, *CHST4*, *CHST8*, *CIDEA*, *CNGA3*, *CNTNAP4*, and *COL29A1*) are shown in ***Figure 4.6B***.

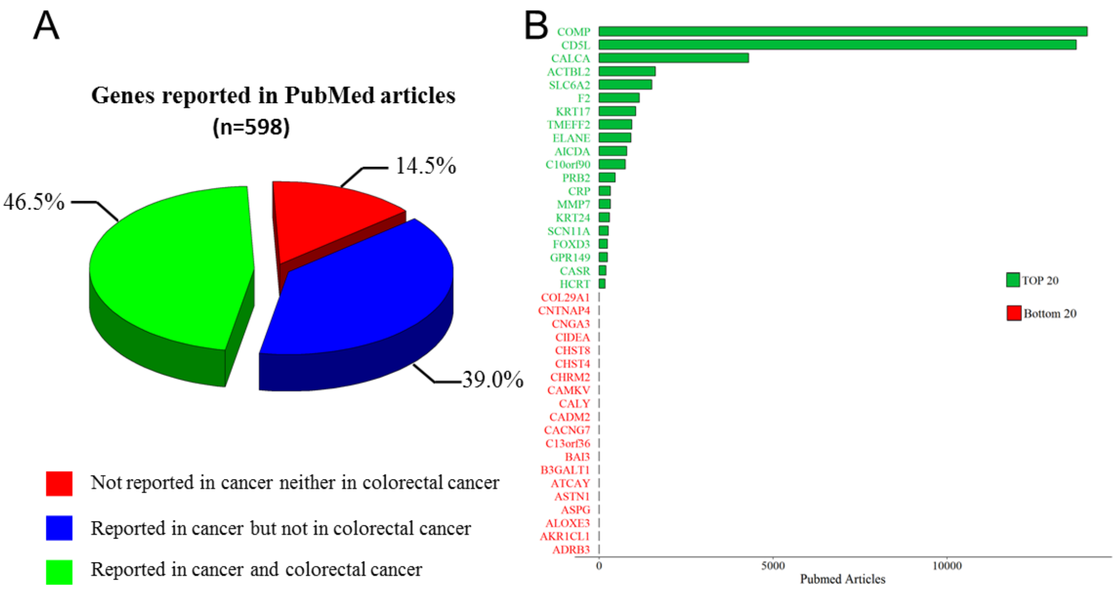


Figure 4.6 Potential new epigenetic biomarkers for colorectal cancer. (A) Only 14.5% of all genes considered as differentially expressed and methylated had not yet been previously associated with cancer. In contrast, 85.5% have already been reported in cancer PubMed articles. From these, 39.0% had not yet been reported in colorectal cancer articles whereas 46.5% were mentioned in “colorectal cancer” associated articles. (B) From the set of genes reported in cancer, we identify the top 20 genes reported in CRC (green), and 20 genes that had not been previously reported in CRC (red).

4.5. Identification of genes epigenetically regulated which characterize CRC progression

We then performed intersections between all genes identified as differentially expressed and methylated in each stage of the disease, using *VennDiagram* package available for R software[183]. The analysis showed that 85 genes are differentially methylated and expressed across all stages of the disease when compared to normal tissue whereas 66, 85, 41, and 40 genes are altered only on a specific stage (stage I, stage II, stage III, and stage IV, respectively) of CRC development (***Figure 4.7***, ***Figure 4.8***, ***Appendix 19***, ***Appendix 20***, ***Appendix 21***, ***Appendix 22***, and ***Appendix 23***).

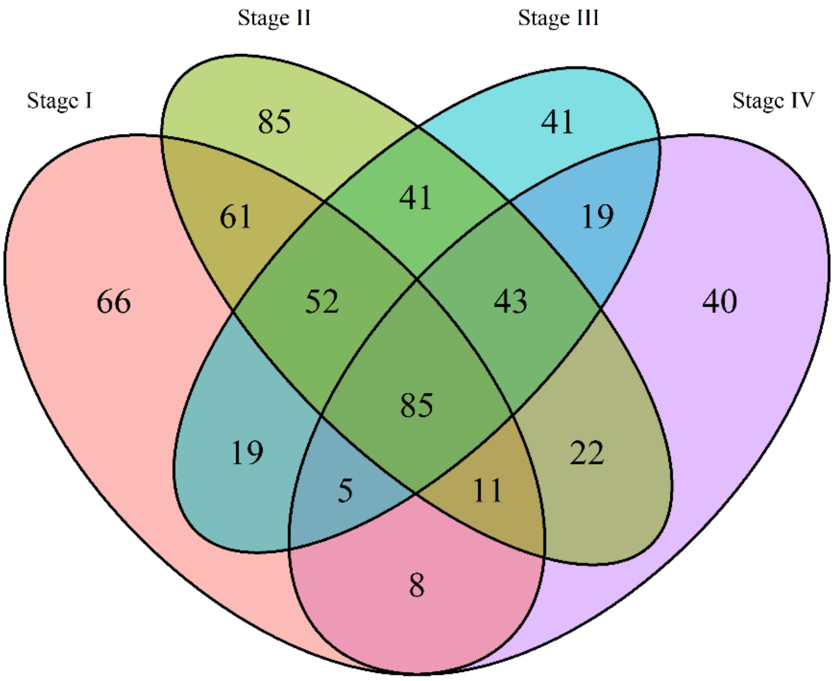


Figure 4.7 Genes epigenetically regulated across colorectal cancer development. From all genes found as differentially expressed and correlated to DNA methylation changes, there are specific set of specific genes which are differentially expressed and methylated only on stage I, stage II, stage III, and stage IV. Moreover, there are common set of genes that are differentially expressed and methylated throughout all stages.

Additionally, this approach also showed that few genes are epigenetically deregulated in 2 or more stages. In more detail, 61, 41, and 19 genes are common to stages I and II, stages II and III, and stages III and IV, respectively (***Figure 4.7***, ***Figure 4.8***).

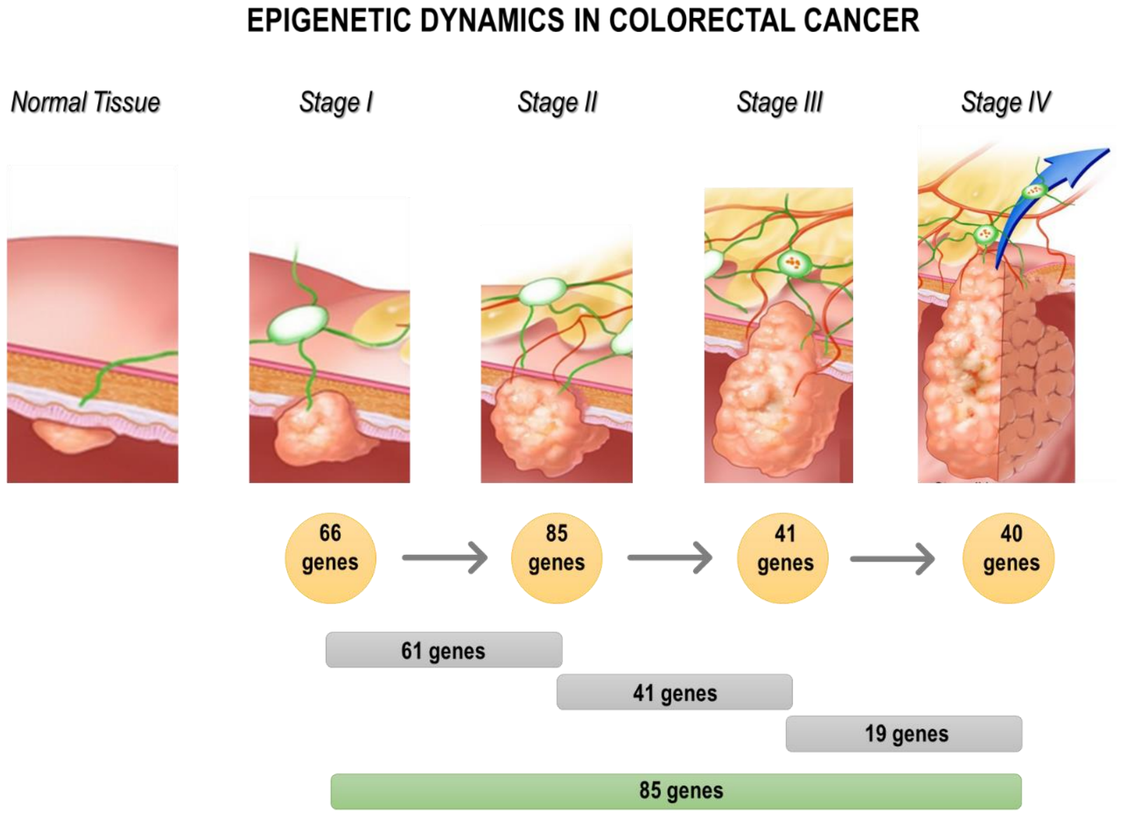


Figure 4.8 Epigenetic dynamic in CRC. There are genes differentially expressed which present CpG sites differentially methylated altered in only one stage of CRC. Specifically, 66, 85, 41, and 40 genes revealed as differentially expressed only on stages I, II, III, and IV, respectively. Moreover, 61, 41, and 19 genes are common to stages I and II, stages II and III, and stages III and IV, respectively. Additionally, 85 genes are common to all stages of CRC (adapted from Chen et al. 2015).

Similarly, the intersection of CpG sites differentially methylated throughout CRC development was performed. This revealed that stage II is the stage with more CpG sites differentially methylated, followed by stages I, III, and IV. Specifically, 121 CpG sites are differentially methylated across all stages of the disease when compared to normal tissue (***Appendix 24***). 342, 815, 298, and 178 CpG sites are differentially methylated only on a specific stage- stage I, stage II, stage III, and stage IV, respectively (***Figure 4.9***, ***Appendix 25***, ***Appendix 26***, ***Appendix 27***, , and ***Appendix 28***). Moreover, ***Table 4.3*** shows the top 10 most differentially methylated CpG sites specific for each tumor stage when compared to normal tissue.

Table 4.3 Top 10 specific CpG sites most differentially methylated in each stage. Positive values for Δβ mean that CpG site is hypermethylated in tumor tissue when compared with normal tissue.

CpG	Δβ	Stage	Gene	CpG	Δβ	Stage	Gene
cg01566592	0.6182190	Stage I	*RIMS2*	cg05937969	0.4898993	Stage III	*CASR*
cg21769093	-0.5632798	Stage I	*VWC2*	cg19868631	0.4786458	Stage III	*VSTM2A*
cg05135828	0.5567621	Stage I	*SLITRK4*	cg16732616	0.4692574	Stage III	*DMRTA2*
cg00662647	0.5393673	Stage I	*SLC35F3*	cg01163842	0.4580375	Stage III	*GSC*
cg09813525	0.5331854	Stage I	*PCDH8*	cg27341472	0.4567671	Stage III	*TMEM196*
cg14170313	0.5202518	Stage I	*RORB*	cg01134282	0.4539574	Stage III	*RALYL*
cg20129213	0.5114721	Stage I	*RIMS2*	cg04347874	0.4488147	Stage III	*NKX2-1*
cg27486637	0.4967454	Stage I	*WDR17*	cg23097402	0.4455062	Stage III	*DMRTA2*
cg17535595	0.4963827	Stage I	*PCDH8*	cg07961994	0.4371072	Stage III	*GRID2*
cg18443378	0.4843755	Stage I	*WDR17*	cg24882673	-0.4334559	Stage III	*ZMAT4*
cg24403845	0.5834750	Stage II	*SORCS1*	cg00241002	-0.5653321	Stage IV	*LOC731789*
cg16437728	0.5450563	Stage II	*SYT9*	cg25884711	0.5186585	Stage IV	*NPY*
cg25146017	0.5413849	Stage II	*C6orf186*	cg07369569	-0.4910343	Stage IV	*HMGCLL1*
cg25729826	0.5324301	Stage II	*CASR*	cg02468050	0.4885486	Stage IV	*GRID2*
cg04842146	0.5197947	Stage II	*RALYL*	cg01395254	-0.4680523	Stage IV	*MYT1L*
cg27361134	0.5170881	Stage II	*RIMS4*	cg24977670	-0.4572511	Stage IV	*CTNND2*
cg20872937	0.5140720	Stage II	*GALR1*	cg20388823	0.4566364	Stage IV	*GRID2*
cg24190603	0.5032448	Stage II	*SNAP91*	cg24543552	-0.4483070	Stage IV	*KHDC1L*
cg26296488	0.5027418	Stage II	*DRD5*	cg01847754	-0.4460867	Stage IV	*CXorf1*
cg09258813	0.5003622	Stage II	*ADRB3*	cg02608452	-0.4386950	Stage IV	*KCNA4*

Interestingly, when we investigated which specific CpG sites were associated to specific genes, we found 200, 223, 89, and 55 CpG sites differentially methylated located within 66, 85, 41, 40 specific genes for stages I, stage II, stage III, and stage IV, respectively. In regard to alterations maintained throughout CRC development, we found 121 CpG sites located in 55 differentially expressed genes.

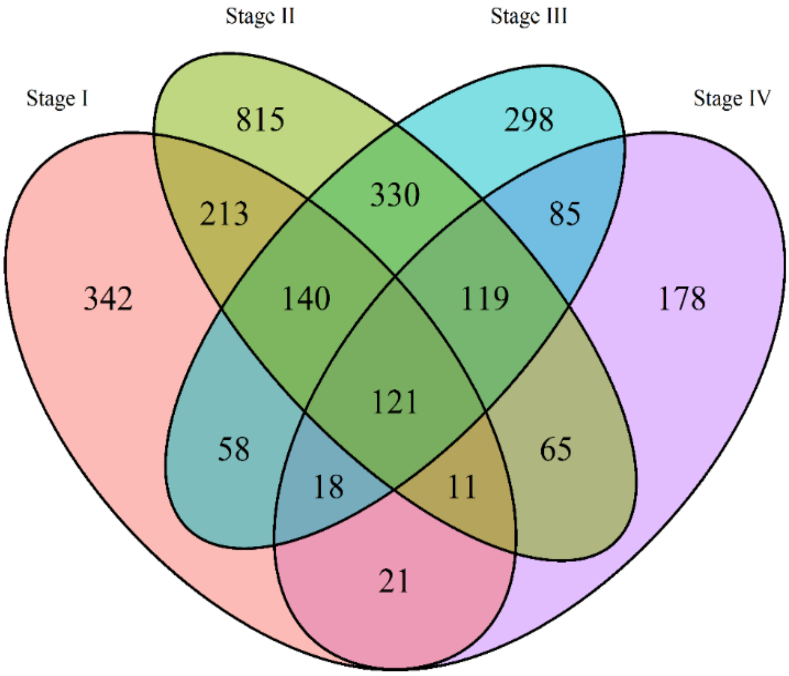


Figure 4.9 CpG sites differentially methylated across colorectal cancer development. A set of CpG sites, located in genes differentially expressed, was found differentially methylated in colorectal cancer. In more detail, when an overlap of these CpG sites is performed, 342, 814, 298, and 178 CpG sites were detected as differentially methylated only on stage I, stage II, stage III, and stage IV, respectively. 121 were differentially methylated across all stages in our analysis.

4.6. Potential New Biomarkers for Colorectal Cancer Diagnosis

ROC curve analyses were performed to establish which genes and CpG sites can distinguish tumor tissue from normal tissue with specific sensitivity and specificity. Therefore, aiming to discovery which genes can be potential biomarkers to CRC diagnosis (AUC > 0.8), we found that 238 differentially expressed and methylated genes from a set of 307 can differentiate stage I tumor tissue from normal tissue (***Appendix 29***). Surprisingly, 24 genes of these genes have not yet been associated with cancer in general, in contrast to 214 which have already been reported in cancer (***Appendix 18***, ***Appendix 29***). From these 214 genes, 93 have not yet been reported in CRC (***Appendix18***, ***Appendix 29***).

Additionally, we intended to identify which CpG sites could potentially differentiate tumor from normal tissue. In accordance to our results from a set of 924 CpG sites, 835 were able to differentiate stage I tumor tissue from normal tissue (***Appendix 30***).

As an example, we show that *ASTN1* is one gene which its expression values can differentiate stage I tumor tissue from normal tissue with an AUC of 0.989 (***Figure 4.10***A, ***Figure 4.10***B). Additionally, previous studies have already demonstrated that *ASTN1* function is related to the nervous system. Specifically, *ASTN1* is a receptor involved in the neuronal migration across glial fibers. A lack of these gene leads to a slow migration[184–186].

Interestingly, the methylation values of CpG site cg08104310 located in *ASTN1* gene can also differentiate stage I tumor tissue from normal tissue, with an AUC of 1.000 (***Figure 4.10***C, ***Figure 4.10***D).

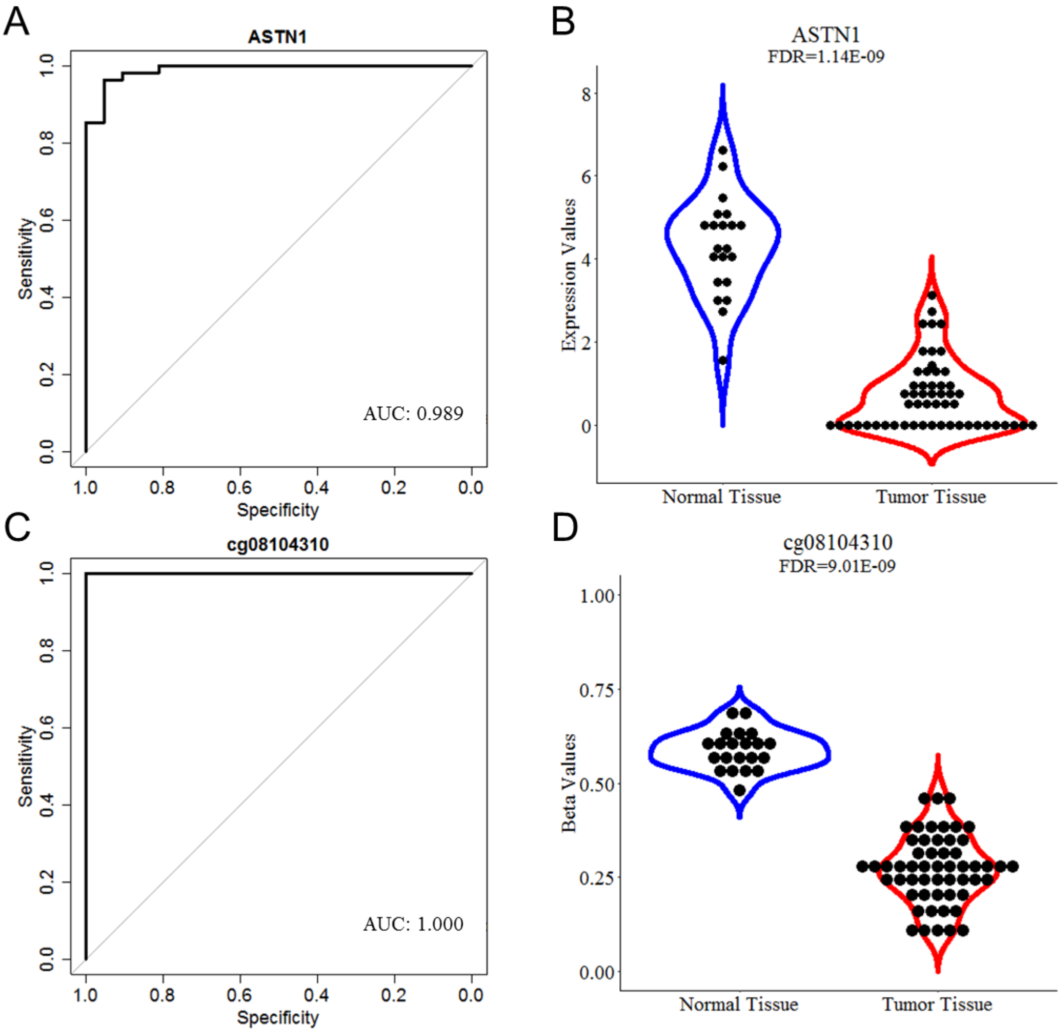


Figure 4.10 ASTN1 gene has the potential to distinguish stage I colorectal tumor tissue from normal tissue. (A, C) ROC curve analysis for ASTN1 expression values (AUC= 0.98), and cg08104310 which is located in the 3'UTR region (AUC=1.000). (B, D) Violin plots representing expression (FDR=1.14E-09) and methylation (FDR=9.01E-09) values for normal and tumor tissue. Each dot corresponds to a patient.

4.7. Identification of Epigenetic Biomarkers which predict patient outcome

The prognostic value of specific genes for each stage was investigated. Thus, a survival analysis was performed using the median as threshold. This analysis shows that from a set of 85 genes altered only on stage II (***Appendix 21***), 6 (*KRT83, LAIR2, SBSN, SNAP91, TMEM179, ZNF536*) have statistical significance to predict the outcome of stage II CRC patients (***Table 4.4***, and ***Figure 4.11***). Moreover, from a set of 41 specific genes altered only on stage III (***Appendix 22***), only *SOX1* can predict the outcome for stage III patients (***Table 4.4***). Lastly, from a set of 40 specific genes only altered in stage IV (***Appendix

23), only 5 genes (*BFSP2, F2, HOTAIR*, and *KHDC1L, SLC6A5*) have the capability to predict stage IV patient outcome (***Table 4.4***). However, using a median of *SLC6A5* expression value to divide our patients into two distinct groups, one of them is constituted by only one patient. Thus, we did not consider *SLC6A5* as a potential prognosis biomarker.

Table 4.4 Genes that better predict the outcome (overall survival). HR: hazard ratio.

Gene	Stage	HR	*p-value*	Cut-off	Gene Function
SNAP91	**Stage II**	0.3395086	0.01277283	0.5254676	Encodes a protein responsible for the transport of *EGF* receptor (*EGFR*)[187].
ZNF536	**Stage II**	3.1328533	0.01819290	1.7054029	Negatively regulates neuron differentiation[188,189].
SBSN	**Stage II**	0.3779900	0.02744733	1.2330904	Involved in epidermal differentiation and cornified envelope formation[190].
KRT83	**Stage II**	2.5024214	0.03404451	1.0393490	Keratin gene (RefSeq, Jul 2008).
TMEM179	**Stage II**	2.8799428	0.04846541	0.0000000	Not well established.
LAIR2	**Stage II**	0.4305100	0.04998472	2.4087934	Inhibits immune cell function upon collagen binding[191].
SOX1	**Stage III**	0.4594413	0.04070349	3.08714	Inhibition of cell growth, and promotion of apoptosis[192].
KHDC1L	**Stage IV**	0.2215415	7.522785e-04	0.0000000	Involved in Germ Cell and Early Development[193].
HOTAIR	**Stage IV**	0.3138915	1.457532e-02	2.5124539	Long noncoding RNA[194].
F2	**Stage IV**	3.332949	1.480315e-02	1.8856746	Coagulation factor[195,196].
BFSP2	**Stage IV**	2.828378	2.677998e-02	0.5357807	Codify for a filament protein[197].

As represented in the ***Table 4.4***, notwithstanding stage II having more genes that can predict the outcome, it is in stage IV that genes have the highest statistical significance for that prediction. Additionally, the set of initial genes under the analysis is higher in stage II (85 genes) than in stage IV (40 genes).

In ***Figure 4.11***, are shown survival curves for *ZNF536*, a negative regulator of neuron differentiation (stage II- ***Figure 4.11***A), *SOX1*, an inhibitor of cell growth and promoter of apoptosis (stage III- ***Figure 4.11***B), and *BFSP2*, which codifies for a filament protein (stage IV- ***Figure 4.11***C) as examples of outcome predictors from the three stages. Those curves show that lower expression levels of *ZNF536* (*p-value*=0.018, hazard ratio=3.133) and *BFSP2* (*p-value*=0.027, hazard ratio=2.828) is associated with a poor prognosis whereas in case of *SOX1* (*p-value*=0.041, hazard ratio=0.459) are associated with a better prognosis. Interestingly, neither *ZNF536* nor *BFSP2* have been previously reported to have a role in CRC.

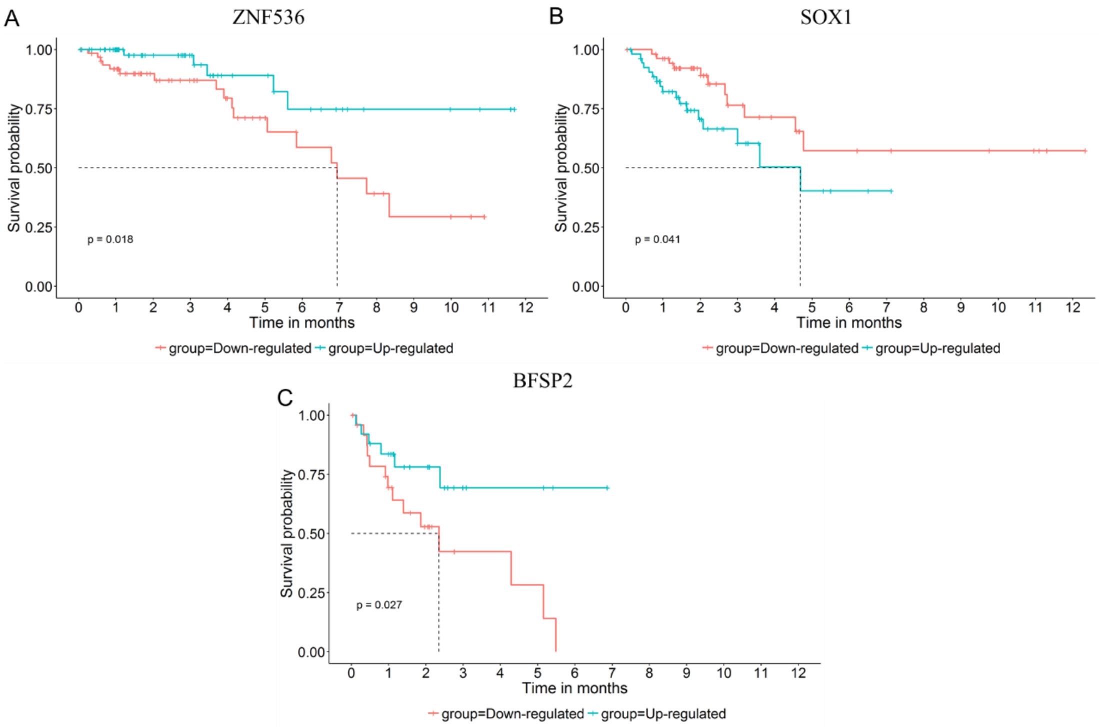


Figure 4.11 Epigenetically altered genes can predict patient outcome. (A) Kaplan-Meier curve representing the overall survival for stage II based on expression values of *ZNF536* gene. The median cut-off-1.71 - was used to divide patients into 2 groups. (B) Kaplan-Meier curve representing the overall survival for stage III based on expression values of *SOX1* gene. The median cut-off- 3.09- was used to divide patients into 2 groups. (C) Kaplan-Meier curve representing the overall survival for stage IV based on expression values of *BFSP2* gene. The median cut-off-0.54 - was used to divide patients into 2 groups. The group of patients with expression values lower than the cut-off was considered as down-regulated (red line) whereas patients with expression values higher than the cut-off was considered as up-regulated (blue line).

Furthermore, we also analyzed the potential of specific CpG sites to predict patient outcome and found that 88, 7, and 3 (***Appendix 31***, ***Appendix 32***, and ***Appendix 33***) from a set of 815, 298, and 178 CpG sites (***Appendix 26***, ***Appendix 27***, and ***Appendix 28***) are good predictors for stage II, III, and IV, respectively. Similar to the pattern observed on potential gene predictors, there are more CpG sites in stage II than in any other stage.

In addition, ***Figure 4.12*** shows one CpG sites per stage as examples. Specifically, hypomethylation of cg02430935- located in the body of the tumor suppressor gene *HMX2*[198] - in stage II (***Figure 4.12A***) and cg01847754- located in the first exon of *CXorf1* gene- in stage IV (***Figure 4.12C***) is associated with poor prognosis (*p-value*=0.013, and 0.019, respectively; hazard ratio=3.139, and 3.155, respectively). In contrast, hypomethylation of cg26489108- located in the region of TSS of a gene involved in the regulation of cell differentiation and survival, *DMRT3*[199]- in stage III (***Figure 4.12B***) is related to a better prognosis (*p-value*=0.027, hazard ratio=0.407).

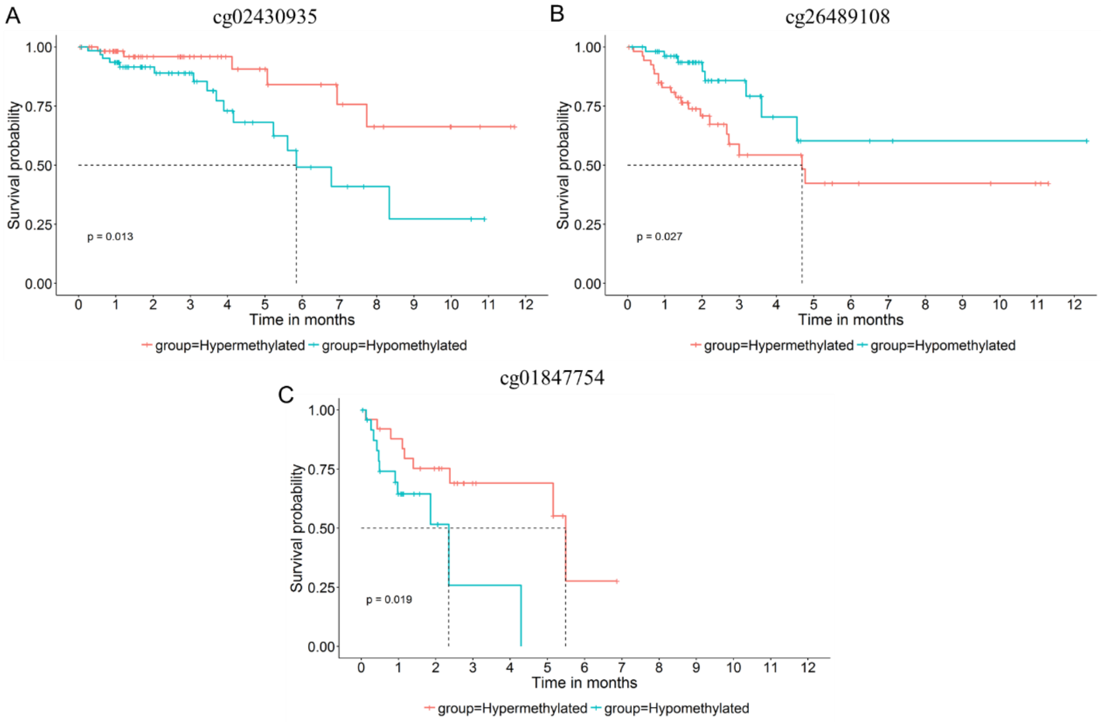


Figure 4.12 CpG sites predict patient outcome. (A) Kaplan-Meier curve representing the overall survival for stage II based on methylation values of cg02430935. The median cut-off- 0.63 - was used to divide patients into 2 groups. (B) Kaplan-Meier curve representing the overall survival for stage III based on methylation values of cg26489108. The median cut-off- 0.74- was used to divide patients into 2 groups. (C) Kaplan-Meier curve representing the overall survival for stage IV based on methylation values of cg01847754. The median cut-off-0.24 - was used to divide patients into 2 groups. The group of patients with methylation values lower than the cut-off was considered as hypomethylated (blue line) whereas patients with methylation values higher than the cut-off was considered as hypermethylated (red line).

Then, the potential of the same set of genes and CpG sites to differentiate patients that recurred from those who did not was also investigated. Regarding gene expression, Cox analysis showed that 3 genes can differentiate stage II CRC patients who recur from patients who do not recur (*CNTD2*, *SNAP91*, and *RPH3A*). In addition, 3 genes were identified as good predictors of recurrence for stage III (*SOX1*, *IZUMO1*, and *GNGT1*). Lastly, only 2 genes were capable to predict recurrence in stage IV (*KHDC1L*, and *HTR2C; **Table 4.5***).

Interestingly, SNAP91, SOX1, and KHDC1L genes can predict both overall survival and recurrence free survival for stage II, stage II, and stage IV, respectively (***Table 4.4***, ***Table 4.5***).

Table 4.5 Genes that better predict the outcome (recurrence free survival). HR: hazard ratio.

Gene	Stage	HR	*p-value*	Cut-off	Gene Function
CNTD2	**Stage II**	0.1962451	0.0003321157	5.8719668	Controls the cell cycle[200].
SNAP91	**Stage II**	0.3705250	0.0194745759	0.5066019	Encodes a protein responsible for the transport of *EGF* receptor (*EGFR*)[187].
RPH3A	**Stage II**	0.3897250	0.0299158467	0.6922486	Involved in the regulation of exocytosis and endocytosis processes at presynaptic sites[201].
SOX1	**Stage III**	0.3592034	0.01014127	3.0871403	Inhibition of cell growth, and promotion of apoptosis[192].
IZUMO1	**Stage III**	0.4474681	0.04563485	0.5046712	Involved in sperm–egg fusion[202].
GNGT1	**Stage III**	0.4415817	0.04985110	1.0921398	G-proteins of photoreceptors[203].
HTR2C	**Stage IV**	0.2846536	0.006409629	0	Receptor of serotonin which have function related to cell growth[204,205]
KHDC1L	**Stage IV**	0.2443717	0.007702573	0	Involved in Germ Cell and Early Development[193].

Figure 4.13 shows plotted curves representing the recurrence free survival for stages II-IV, based on gene expression levels of *CNTD2*, a controller of cell cycle, *SOX1*, and *HTR2C*, also an intermediate of cell cycle. Remarkably, high gene expression levels of these 3 genes are associated with a poor prognosis, meaning that patients with high levels of these genes have a higher probability for recurrence. Moreover, this analysis shows

that all patients in stage IV who have expression levels of the *HTR2C* gene above the cut-off recur by 3 months. In opposition, patients who have expression levels of *HTR2C* gene below the cut-off might have a delay on recurrence (***Figure 4.13***).

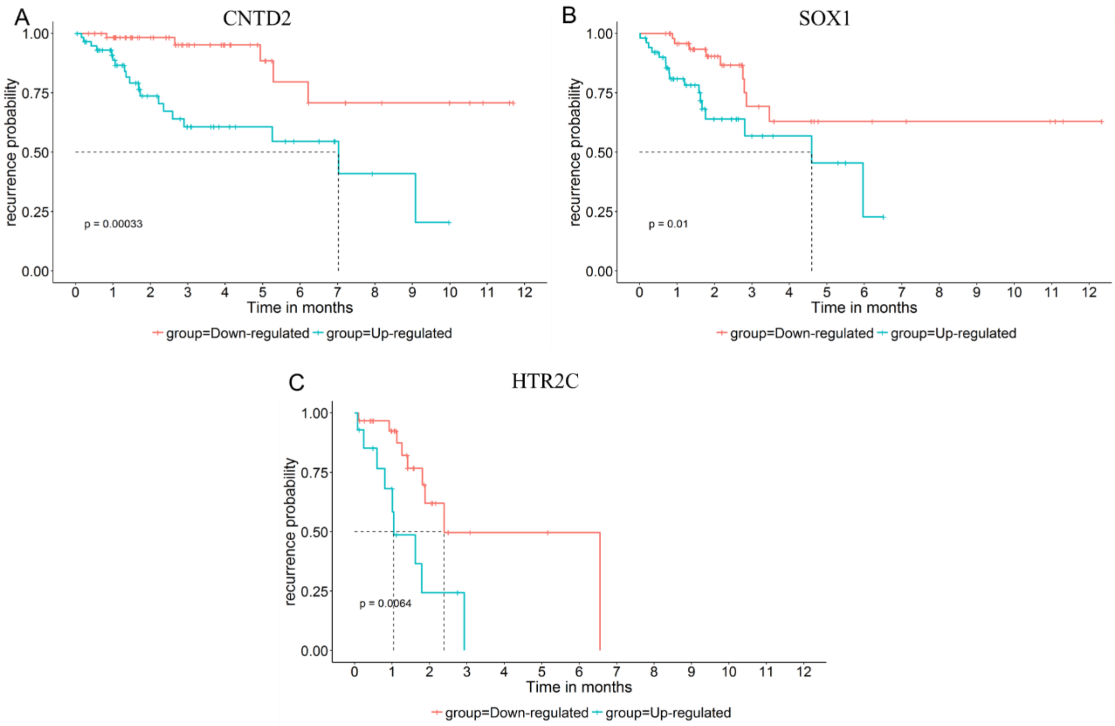


Figure 4.13 Recurrence Free Survival prediction through gene expression levels. (A) Kaplan-Meier curve representing the recurrence free survival for stage II based on expression values of *CNTD2* gene. The median cut-off-5.87- was used to divide patients into 2 groups. (B) Kaplan-Meier curve representing the recurrence free survival for stage III based on expression values of *SOX1* gene. The median cut-off-3.09 - was used to divide patients into 2 groups. (C) Kaplan-Meier curve representing the recurrence free survival for stage IV based on expression values of *HTRC2* gene. The median cut-off-0 - was used to divide patients into 2 groups. The group of patients with expression values lower than the cut-off was considered as down-regulated (red line) whereas patients with expression values higher than the cut-off was considered as up-regulated (blue line).

When looking at CpG sites, in order to identify potential biomarkers for recurrence free survival, the same set of CpG sites tested for overall survival were also analyzed here. For recurrence free survival in stage II patients, it was found 30 CpG sites that can distinguish two groups (those who recur and those who do not recur) based on a specific cut-off (***Appendix 34***). In stage III, and stage IV, it was found 12 and 9 genes as recurrence predictors, respectively (***Appendix 35***, ***Appendix 36***).

Figure 4.14 shows Kaplan-Meier plots representing recurrence free survival curves for the most significant CpG site from each stage. Specifically, high methylation levels of cg06162589- located in the 3'UTR region of a tumor suppressor gene *SLC5A8*[206] - in stage II (*p-value*=0.0066, hazard ratio= 0.2924- ***Figure 4.14***A) and cg03700449- located in the first exon of a gene involved in the neurogenesis, *ASCL1*[207]- in stage III (*p-value*=0.0055, hazard ratio= 0.3114- ***Figure 4.14***B) are associated with a poor prognosis whereas high methylation levels of cg14772660- located in body of a gene *SLC5A7* that encodes to a choline transporter[208]- in stage IV (*p-value*=0.0047, hazard ratio= 4.3174- ***Figure 4.14***C) are associated with a better prognosis.

Moreover, this analysis also showed that patients with high methylation levels of cg06162589, have a probability of 100% to recur at 7 months. In contrast, patients who present low methylation levels of that CpG site have a reduced probability to recur (***Figure 4.14***A). A similar scenario happens in stage IV, when methylation levels of cg14772660 are considered. Specifically, it is expected that all patients of both hypomethylated and hypermethylated groups recur. However, patients who present low methylation levels of cg14772660 recur before (***Figure 4.14***C).

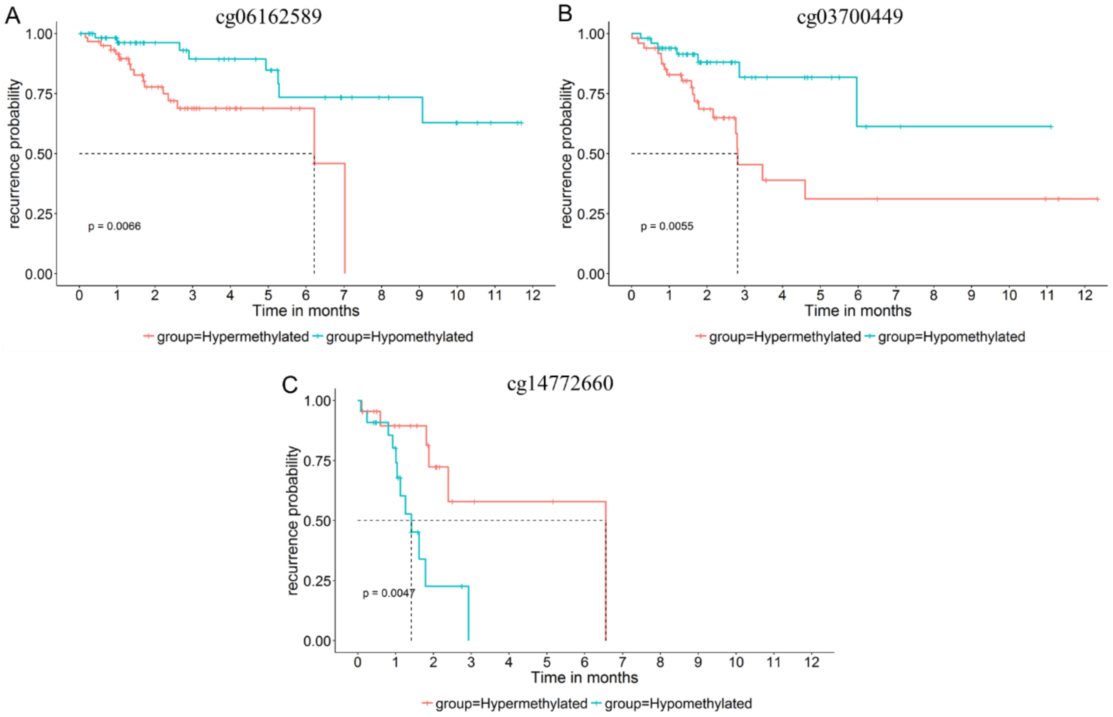


Figure 4.14 CpG sites can predict recurrence. (A) Kaplan-Meier curve representing the recurrence free survival for stage II based on methylation levels of cg06162589. The median cut-off-0.32- was used to divide patients into 2 groups. (B) Kaplan-Meier curve representing the recurrence free survival for stage III based on methylation levels of cg03700449. The median cut-off-0.52 - was used to divide patients into 2 groups. (C) Kaplan-Meier curve representing the recurrence free survival for stage IV based on methylation levels of cg14772660. The median cut-off-0.69 - was used to divide patients into 2 groups. The group of patients with methylation values lower than the cut-off was considered as hypermethylated (red line) whereas patients with methylation values higher than the cut-off was considered as hypomethylated (blue line).

4.8. CRC patients can be grouped according to their gene expression and DNA methylation patterns

Here, in order to validate the results obtained by our pipeline (**Erro! A origem da referência não foi encontrada.**), we used the HJ-biplot multivariate technique of graphical representation, followed by hierarchical clustering in the patient coordinates, considering the ward method and the Euclidean square distance. Furthermore, this approach was also used aiming to corroborate the pipeline and to distinguish groups of patients classified within the same stage. In order to facilitate the visualization of the results we established three groups per stage and selected genes and CpG sites with contributions over than 0.7 to increase powerful quality contributions to HJ-biplot (***Figure 4.15***, ***Figure 4.16***).

Firstly, we analyzed how samples from CRC patients are distributed based on genes differentially expressed in each stage when compared to normal tissue samples. Applying the contribution cut-off, we selected 43, 22, 12, and 12 genes (from a set of 307, 400, 305, and 233, respectively) from stages I, II, III, and IV. This suggests that each set of variables is the strongest when combining genes responsible for the patient distribution in a respective stage.

Since the HJ-biplot is a data reduction technique, the plans 1-2 (***Figure 4.15***) retains 87.3%, 85.4%, 87.4%, and 89.5% of the total absorbed inertia (***Figure 4.15***A, ***Appendix 37***, ***Figure 4.15***B, ***Appendix 38***, ***Figure 4.15***C, ***Appendix 39***, ***Figure 4.15***D, and ***Appendix 40***) from stages I, II, III, and IV, respectively. This fact shows that all genes in all stages are correlated and represent high levels of information (amount of accumulated variance). We also observed that no isolated gene (or a set of genes) exists with other behavior (correlated with axis 2). In fact, all of these set of variables are correlated to the axis 1, meaning that all variables are correlated with each other.

Moreover, this approach shows that we can distinguish primary tumor from normal samples (circles and triangles, respectively- ***Figure 4.15***). Thus, the pipeline applied was efficient. Also, it is also possible to identify two subgroups of primary tumor samples (red and green), suggesting that in the same stage, there are patients who have different patterns of gene expression, and might be responsible for different outcomes (***Figure 4.15***). Curiously, when we looked at the HJ-biplot representation of each stages, we observed that the distance between normal and primary tumor samples is bigger in stages I and IV. This fact shows that in these stages primary tumor samples are more distinct from normal samples than in other stages. Interestingly we observed two-subgroups of patients at stage II where the green subgroup is closer to normal tissue when compared to the red group. At a lesser extent the same was observed in stage III of the disease suggesting that different set of genes are characterizing distinct subgroups of patients.

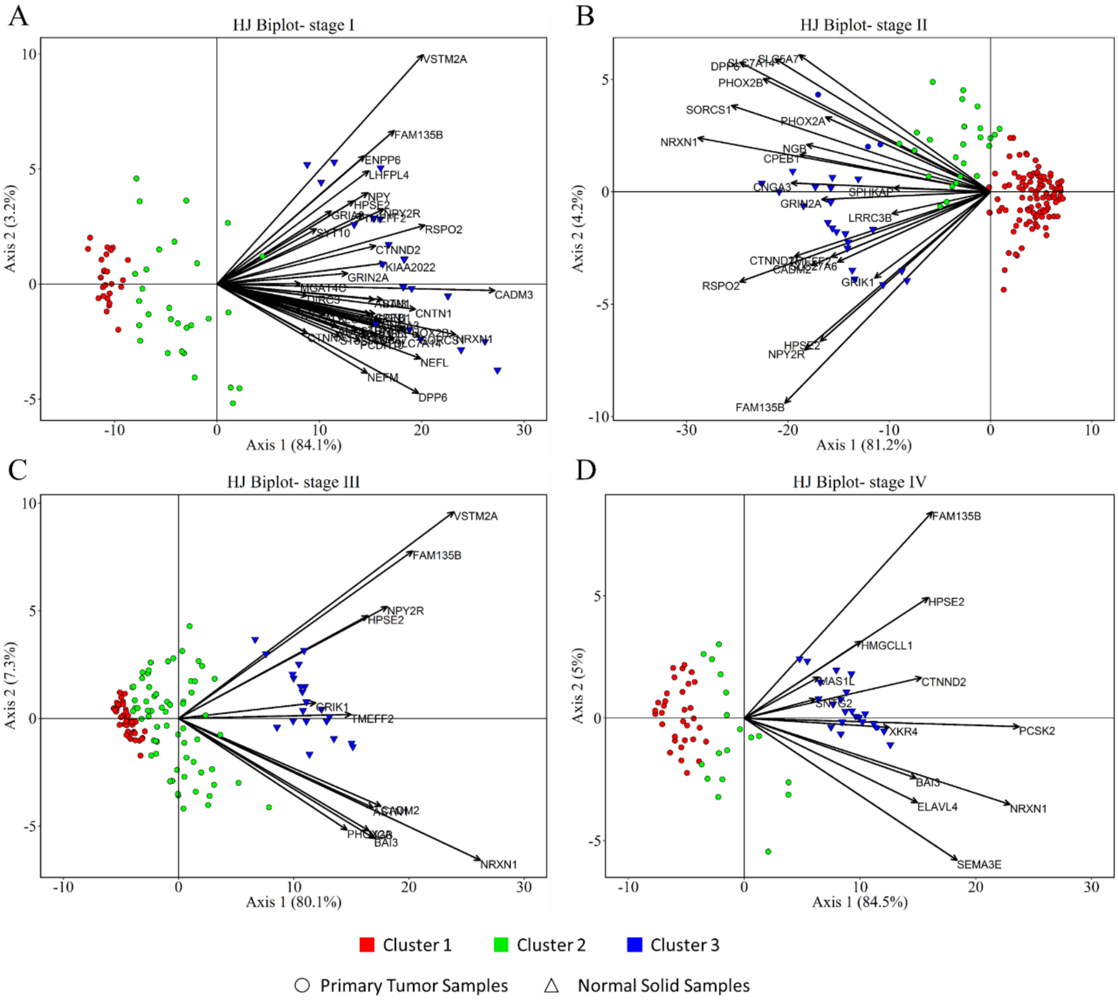


Figure 4.15 HJ-biplot representation of gene expression for (A) stage I, (B) stage II, (C) stage III, and (D) stage IV of CRC. In stage I, cluster 1 is composed by 25 patients, cluster 2 by 29 patients and cluster 3 by 21 patients. In stage II, cluster 1 is composed by 103 patients, cluster 2 by 25 patients and cluster 3 by 24 patients. In stage III, cluster 1 is composed by 46 patients, cluster 2 by 62 patients and cluster 3 by 21 patients. In stage IV, cluster 1 is composed by 32 patients, cluster 2 by 18 patients and cluster 3 by 22 patients.

The same approach was applied using CpG sites differentially methylated in CRC. For that we selected 116, 65, 16, and 30 from stages I, II, III, and IV, respectively. The total absorbed inertia of plans 1-2 were: 85.3%, 81.3%, 84.4%, and 85.2% for each stage (***Figure 4.16***A, ***Appendix 41***, ***Figure 4.16***B, ***Appendix 42***, ***Figure 4.16***C, ***Appendix 43***, ***Figure 4.16***D, and ***Appendix 44***). These results are similar to the ones obtained for gene expression since all differentially methylated CpG sites are correlated and also represent high levels of retained information. We also observed that no isolated CpG sites correlated with axis 2 exist, meaning that all of those CpG sites are correlated.

Furthermore, similar to the observed in HJ-biplot applied to a set of genes, we found that the set of selected CpG sites can also differentiate primary tumor from normal samples (circles and triangles, respectively), and it can divide primary tumor samples into two distinct groups (red and green). Additionally, it is also possible to observe that the distance from the cluster constituted by normal samples (triangles) to the clusters constituted by primary tumor samples is higher in stages I, and IV when compared to stages II, and III (***Figure 4.16***). This suggests that the set of CpG sites selected can easily detect which are the tumor samples. Regarding to stage II, we detected a decrease of the proximity between clusters of normal and primary tumor samples (***Figure 4.16***B). As for stage III, we found that the cluster formed by normal samples are homogenous, in spite of the proximity to other clusters constituted by primary tumor samples, suggesting that there are tumor samples with more similarities to normal samples. (***Figure 4.16***C).

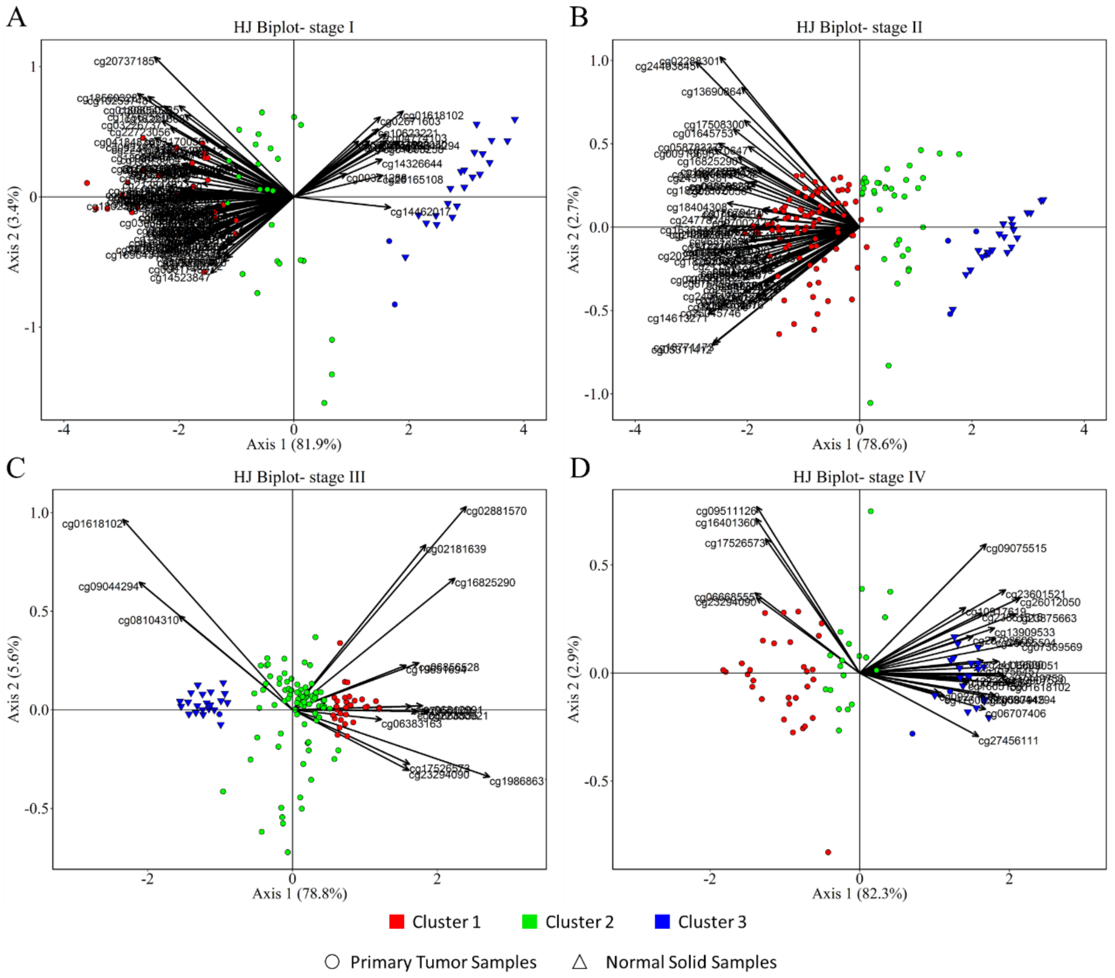


Figure 4.16 HJ biplot representation of DNA methylation for (A) stage I, (B) stage II, (C) stage III, and (D) stage IV. In stage I, cluster 1 is composed by 27 patients, cluster 2 by 25 patients and cluster 3 by 23 patients. In stage II, cluster 1 is composed by 91 patients, cluster 2 by 36 patients and cluster 3 by 25 patients. In stage III, cluster 1 is composed by 29 patients, cluster 2 by 81 patients and cluster 3 by 22 patients. In stage IV, cluster 1 is composed by 30 patients, cluster 2 by 19 patients and cluster 3 by 23 patients.

5. CHAPTER V- DISCUSSION

In this study, we performed a whole-genome analysis of the CRC patient cohort, available at the TCGA dataset, with the intention of identifying epigenetic signatures during CRC development. Contrarily to the majority of whole-genome analysis performed, we decided to analyze CRC initiation and progression, by distributing our patients through a correspondent TNM stage[209–211].

We found that epigenetic alterations are present throughout CRC development and that they can be associated either to the overexpression or silencing of genes. This was expected since DNA methylation alters the chromatin structure[212,213]. In addition, the dynamics of epigenetic alterations was also shown here, since there are different DNA methylation patterns in different stages of CRC development. Thus, as previous studies reported, understanding DNA methylation alterations may provide new insights on this disease, and hopefully help to improve both diagnostic and prognostic in CRC[214]. Indeed, it is known that epigenetic changes are dynamic during events as embryonic development and cell differentiation. However, until now there was a lack of studies that demonstrated these dynamics across CRC initiation and progression[215–217].

Moreover, we also found that there are groups of genes more affected by these changes than others, possibly due to the importance of specific classes of genes involved in cancer initiation and progression[7]. However, few genes are epigenetically altered in all stages of CRC suggesting that, depending on the CRC stage, different genes need to be overexpressed or down-regulated in order to favor the disease[91].

Furthermore, applying the developed pipeline, we found that stage IV contains less alterations, where DNA methylation is correlated with gene expression, than the other stages when compared to normal tissue. One possible explanation for this observation is that in the last stage, cells may not need to express genes responsible for cancer initiation. However, another reasonable hypothesis lies on the fact that during metastastatic formation, cancer cells might need to become more similar to normal cells for this process to be successful. Alternatively, sample size could have had an influence in our analysis. Moreover, we also found that stage I is not the stage with more alterations, in opposition to previous reports[81]. This difference can also be explained by sample size. Additionally, previous studies also shown that, although there are no significant differences in the proportion of early and advanced tumors which present aberrant DNA methylation patterns, the number of genes differentially methylated is different across cancer initiation and progression[81,218,219].

This study identified for each stage of CRC development, multiple CpG sites differentially methylated which are correlated to alterations in gene expression. In agreement with previous studies[220], when we investigated where these CpG sites are located, we found that promoters (TSS200, TSS1500) and gene body are the regions more affected by aberrant methylation. Indeed, the aberrant methylation in promoter region alters the accessibility of transcriptions factors to the DNA, leading to changes in gene expression[127]. Additionally, gene body methylation has already been associated to alteration on gene expression[221]. Recently, it was reported that intragenic DNA methylation can prevent the transcription initiation of aberrant transcripts in mouse embryonic stem cells[222]. Another study also suggested that gene body DNA methylation associated with H3K36me3 blocks aberrant transcription[223].

Moreover, when we compared methylation patterns across CRC progression, we found that the most common alteration is hypermethylation. Among all hypermethylated CpG sites, we observed that the great majority is located in down-regulated genes, meaning that high methylation levels are associated to low gene expression levels. This fact further strengthens the conventional idea that hypermethylation is often associated with down-regulation whereas hypomethylation leads to gene activation. However, there are cases where this is not verified. Indeed, the consequence of DNA methylation in gene expression depends on its location in the gene. Specifically, when CpG sites are located in repressor regions, and this region is hypermethylated, repressors are blocked. Consequently, hypermethylation will be associated with overexpression of the respective gene. Interestingly, this pattern was observed before in genes involved in cancer as *TERT* and *EGFR.* In both cases, promoter hypermethylation leads to gene activation[127,224].

Nevertheless, the pattern described before is less observed in stage IV. Specifically, in opposition to previous stages, hypomethylation is more common than hypermethylation and a positive correlation between DNA methylation and gene expression levels in tumor tissue is also more common than negative correlation. These findings suggest that the epigenetic pattern of stage IV is different from other stages. Importantly, across cancer progression, the heterogeneity increases, which can instigate this difference[225].

When the top 15 most differentially expressed genes with significant differences in DNA methylation were analyzed, we found that the vast majority were simultaneously present in the top 15 in all four stages. Specifically, genes as ZIC5, DIRC1, KRTAP3-1 are found in all stages, meaning that these genes are among the most differentially expressed genes in CRC progression. Curiously, ZIC5 is a zinc finger complex which has already been associated with cancer. Satow et al. also associated ZIC5 with higher survival of CRC cells by enhancing FAK and STAT3 activity[226]. DIRC1 is not well described, although it has been reported in familial clear cell renal cancer[227]. KRTAP3-1 is a keratin-related gene, and has not been previously reported in cancer PubMed articles[228]. Although these three genes are players in CRC their role has not yet been characterized.

Furthermore, *FEZF1*, *SOX14*, *LOC84931*, and *PGLYRP3* are among the top 15 most differentially expressed genes with significant differences in methylation in stages I, II, and III. All these genes have been associated with cancer, however only *FEZF1* was previously reported in CRC. Interestingly, both *FEZF1* and *SOX14* are reported as enhancers of tumor proliferation and metastasis[229,230]. *LOC84931* has also been associated with cancer, however its role in oncogenesis is poorly described[231]. At last, Jing et al. have associated *PGLYRP3* to inflammatory bowel disease (IBD)[232]. Importantly, higher risk of CRC is mostly connected to IBD[233,234]. This fact suggests that *PGLYRP3* can be effectively related to CRC directly or indirectly, although its association has not been yet well studied.

Biological process analysis was performed using KEGG and GO platforms. Results indicate that genes differentially expressed and methylated across CRC development have similar functions, being the majority related to the nervous system development. Specifically, KEGG pathway analysis revels “nicotine addiction”, “neuroactive ligand-receptor interaction”, “cAMP signaling pathway”, “salivary secretion” and “ALS” as the most enriched pathways in CRC. Previous studies have reported that smoking is strongly associated with higher incidence and mortality in CRC[235–238]. “Neuroactive ligand-receptor interaction” pathway also revealed to be important. This pathway was also found enriched in other types of cancer as well as in CRC[239,240]. Moreover, several studies have also described the relationship between cAMP and cancer. Specifically, studies referred that increasing cAMP levels leads to inhibition of cell growth by the induction of apoptosis and/or cell cycle arrest[241]. For example, Dong, H et al. have demonstrated that the activation of cAMP signaling leads to a decrease of cell migration in breast cancer[242].

Regarding the salivary secretion pathway, there are reports which suggest that cancer can interact with salivary glands, including pancreatic cancer, breast cancer, lung cancer, and ovarian cancer. However, these relationship is not yet well established[243]. Finally, the ALS pathway is also enriched in our analysis. Curiously, neurodegenerative diseases have been inversely associated with cancer[244–246]. Whereas cancer involve resistance to cell death, neurodegenerative diseases occurs due to premature cell death[247]. In case of ALS pathways, its relationship with cancer is not clear[248,249].

Interestingly, GO biological process analysis shows that the most enriched functions are related to the neural system. Several studies suggested that there is a crosstalk between neural system and CRC cells. Neurogenesis and axogenesis are important in cancer development, and nerves are constituents of the microenvironment[250–252]. In addition, studies suggested that neurotransmitters can stimulate migration, cell survival, and proliferation, immune suppression, angiogenesis, and provide mechanical support. Hence the nervous system pathways development is associated with poor prognosis, by inducing tumorigenesis and metastasis[253–255]. For example, in pancreatic cancer, cancer cells grow and invade nerves, leading to a poor prognosis and severe pain[256,257]. In a specific case of CRC progression, there are also evidences that the nervous system support cell migration through a processes called perineural invasion[258].

Unexpected, from all genes found as differentially expressed and methylated, 14.5% have not yet been previously reported in any cancer type, including CRC and might be critical to CRC initiation and progression. For example, *CIDEA* has already been associated with other cancers by promoting apoptosis in mammalian cells[259]. Interestingly, our analysis showed that this gene is differentially expressed, and it has CpG sites differentially methylated in all stages of CRC progression. Furthermore, *CIDEA* is down-regulated in all stages, which is corroborated by its biological function[259].

Then, we investigated the potential of genes differentially expressed and methylated between stage I and normal tissue as potential diagnostic biomarkers in CRC. This analysis revealed that 78% of differentially expressed genes and 90% of CpG sites differentially methylated are good candidates. This evidence suggests that DNA methylation levels of CpG sites can probably be more trustworthy than gene expression levels in relation to diagnosis. This because almost all CpG sites found differentially methylated in tumor tissue when compared to normal tissue could be used as potential biomarkers for diagnosis in contrast to the lower number of differently expressed genes with the same ability One possible explanation to this fact might be due to expression levels being more instable, meaning that they can vary easily due to different factors besides cancer[260]. Moreover, these are good early diagnostic candidates, since its methylation occurs in early stages of CRC. For example, *ASTN1* where both gene expression and DNA methylation values could differentiate stage I tumor tissue from normal tissue, had a higher AUC value for the methylation of one of its CpG cg08104310, rather than for its expression values. This result suggests that methylation of *ASTN1* gene can better differentiate tumor from normal tissue than its expression levels. *ASTN1* codifies for a membrane protein involved in the central nervous system development[261]. Interestingly, functions related to nervous system development are enriched in our analysis and have been associated with cancer, as reported before.

Additionally, this analysis showed that although enriched functions of genes differentially expressed with probes differentially methylated are similar across CRC, there are different genes associated to CRC initiation and progression. The same pattern was observed with specific CpG sites that were associated only to a specific stage, as well as CpG sites which characterize all stages of CRC development. Also, we found that there are genes differentially expressed and with probes differentially methylated in specific stages of CRC development, suggesting that these genes have an important function in a specific stage. In contrast, we also found genes differentially expressed and methylated in all stages, evidencing that there are genes that play a role across CRC development. It was also identified common genes between two stages suggesting a role in the transition between stages. Interestingly, there is a lack of studies that corroborate these findings, as the vast majority of whole-genome analysis performed until this date do not analyze DNA methylation and gene expression data across cancer initiation and progression taking into account the TNM staging.

Also, when we evaluated the potential of specific genes differentially expressed to predict the outcome of patients in stage II, III, and IV, we found that there are genes capable of distinguishing two groups depending on a gene expression cut-off value. For example, in stage II, *ZNF536* expression values can differentiate two distinct groups. In this case, the down-regulation of *ZNF536* is associated with poor prognosis. Interestingly, this gene encodes for a zinc finger protein which negatively regulates neuron differentiation[188,189] and was previously reported in association with cancer only once. Importantly, this type of proteins have been revealed as players in the progression of multiple cancer types[262].

Regarding stages III and IV, genes as *SOX1* and *BFSP2*, respectively, can also differentiate two groups with distinct outcome. Indeed, higher gene expression levels of *SOX1* were associated with a worst prognosis. Paradoxically, distinct studies reported that *SOX1* is associated with inhibition of cell growth as well as promoting apoptosis, by decreasing β-catenin levels[192]. Hence, it was expected that low methylation values of *SOX1* were associated with a worst prognosis. Interestingly, DNA methylation changes have also been reported in another study, where *SOX1* was hypermethylated in CRC. In the same study, *SOX1* was also found most significantly methylated in later stages of TNM classification[263].

Furthermore, *BFSP2* is a gene that codifies for a filament protein and lower *BFSP2* expression levels are associated with poor prognosis in this study. Interestingly, this gene has not been previously reported in CRC, perhaps due to their specificity to lens fiber cells, a structure in the eye[197].

Next, we evaluated the potential of CpG sites to predict patient outcome and found that there are CpG sites efficient to distinguish 2 groups. Specifically, CpG sites as cg02430935, cg26489108 and cg01847754 are able to predict the outcome in stage II, stage III, and stage IV, respectively. Furthermore, both cg0243093- located in the first intron of *HMX2* gene- and cg26489108- located in the region near to the TSS of *DMRT3* gene- are hypermethylated in tumor tissue (stages II and III, respectively). Moreover, *HMX2* gene was down-regulated, and *DMRT3* gene was up-regulated. In addition, previous studies have reported that *HMX2* inhibits the cellular growth, and are frequently silenced in colorectal cancer cell lines and primary tumors[123]. *DMRT3* was also found up-regulated in a recent study. The same study has reported an interaction between this gene and both *TP63* and *SOX2* in lung cancer.

Interestingly, hypermethylated of cg0243093 (*HMX2* gene) in tumor tissue is associated with a better prognosis in stage II patients. In fact, previous studies have suggested that *HMX2* has tumor suppressor activity in cancer[198], being down-regulated in other cancer types[123].

Regarding to cg26489108 (*DMRT3* gene), we found a positive correlation between methylation levels of this CpG site and *DMRT3* gene in tumor tissue which is associated with poor prognosis. However, in contrast with the standard epigenetic regulation mechanism, hypermethylation of this CpG site is associated with gene over-expression[264].

One possible explanation for this, is that this CpG site is located in a repressor binding region. When hypermethylation occurs, the repressors are blocked, and, consequently, the gene is constitutively expressed. The hypothesis that hypermethylation leads to the activation of an alternative promoter should be considered[127].

Considering cg01847754- located in the first exon of *CXorf1* gene is hypomethylated in stage IV. This is in accordance with survival analysis performed, which showed that low methylation levels are associated with a poor prognosis. Interestingly, one more time, the canonical pathway is not observed, meaning that hypomethylation of cg01847754 is associated with down-regulation of *CXorf1* gene. Indeed, this gene may be an interesting candidate biomarker for cancer prognosis although its role in cancer is not clear yet.

Then, were identified genes that could predict patient recurrence. *CNTD2*, *SOX1*, and *HTR2C* are some of the genes found as predictors of recurrence free survival in stages II, III, and IV, respectively. Specifically, *CNTD2* was found over-expressed in stage II CRC tissue which is in agreement with a poor prognostic associated with high gene expression levels. *CNTD2* is a member of the cyclin family which can control the cell cycle. When there are alterations on genes that regulate the cell cycle, the consequence may be uncontrolled cell growth. Recently, this gene was reported as a new oncogenic driver in lung cancer[200].

HTR2C is up-regulated in stage IV of our cohort, and, as expected, high expression levels of this gene are associated with a poor prognosis. In agreement with our finding, other studies have already reported an over-expression of *HTR2C* in tumor tissue[204]. Interestingly, *HTR2C* is a receptor of serotonin involved in cell growth[204,205]. Also, there are evidences that serotonin plays a key role in aggressive tumors, mainly when *HTR2C* is present[205].

We also observed that high *SOX1* expression levels in stage III is associated with a poor prognosis. Similar to *SOX1*, *SNAP91*, and *KHDC1L* were found to be predictors of both survival and recurrence for CRC patients in stages II, and IV, respectively. This suggest that the expression levels of these genes should be monitored even after cancer treatment. Moreover, *SOX1*, and *SNAP91* were also identified as correlated with survival in different types of cancers including glioblastomas[265–267]. Additionally, studies reported aberrant methylation of SOX1 in a CRC cohort, validating our results[263]. Regarding to *KHDC1L*, there are no reports associating it neither to CRC nor to cancer in general.

Furthermore, our findings revealed that there are CpG sites that allow to differentiate patient recurrence. For example, cg06162589 located in the 3'UTR of *SLC5A8*, cg03700449 located in the first exon of *ASCL1,* and cg14772660 located in the gene body of *SLC5A7* are CpG sites with potential for prognosis in stage II, III, and IV, respectively. Moreover, both cg06162589 and cg14772660 are hypomethylated, and positively correlated with gene expression levels of tumor tissue, whereas cg03700449 is hypermethylated, and negatively correlated with gene expression levels of tumor tissue. In addition, Cox analysis reveals that low methylation levels of cg14772660, and high levels of cg03700449 are associated with poor prognosis, regarding recurrence free survival. Although cg06162589 is hypomethylated in tumor tissue, low methylation levels are associated with a better prognosis of CRC patients. In addition, *SLC5A8* was also identified as down-regulated in colon and other tumors as lung cancer and acute myeloid leukemia[268–270]. In hepatocellular carcinoma, *SLC5A8* was also found to be down-regulated, which was associated with the inhibition of cancer progression by decreasing the expression of proteins such as β-catenin, and c-Myc[271]. Interestingly, in previous studies *SLC5A8* has already been reported as a tumor suppressor gene that can be silenced by epigenetic mechanisms[206,269,270,272–275]. In lung adenocarcinomas, *ASCL1*, a gene involved in cell proliferation, survival, and cell cycle control, was found overexpressed mainly in smokers, and associated with a poor prognosis. Additionally, in the same study it was observed a global hypomethylation of this gene[276]. Other studies have also reported that DNA methylation regulates expression levels of *ASCL1* in cancer. Interestingly, in medullary thyroid cancer, the expression levels of this gene are decreased by the action of NOTCH1[277]. Moreover, *SLC5A7* was also found to be down-regulated in different tumor types, being associated with poor prognosis. In contrast, high expression levels of this gene are associated with better prognosis in several cancer types. However, its role in cancer development remains unclear[278].

Finally, using HJ-biplot multivariate analysis we identified a set of gene expression and methylation profiles that can differentiate normal and tumor samples corroborating the pipeline previously applied. Also, this analysis allowed to differentiate for each tumor stage two different sub-groups suggesting that there are patients who are classified into the same stage but have distinct gene expression and methylation patterns. Indeed, this might explain why patients at the same stage differ in survival time. However, to validate these results, further studies need to be performed. Specifically, to evaluate whether, in fact, the outcome of these two groups are distinct, a Kaplan-Meier analysis followed by a logrank test can be performed.

In summary, epigenetic changes are dynamic across CRC initiation and progression, and can influence the expression of genes involved in this process, including genes with functions related to the nervous system regulation. Using specific cut-offs for certain differentially expressed genes or differentially methylated CpG sites, it was possible to distinguish tumor from normal tissue, in an early stages of the disease (stage I), and determine patient outcomes (overall survival and recurrence free survival).

5.1. Limitations

The current study presents some limitations including:

a. The sample size should be larger and similar for each stage.
b. Normal samples are from tumor adjacent tissues and, in some cases, they might not be normal anymore. It would be better to have normal tissue from individuals without disease.
c. The clinical information should be more robust as the current one presented some difficulties to standardize our groups.
d. In the HJ-biplot representation, it was only taken in account both first and second dimensions.

6. CHAPTER VI- CONCLUSION

This whole-genome analysis provides advances in the knowledge of epigenetic dynamics across CRC initiation and progression.

Here, we created an epigenetic roadmap, identifying new biomarkers for CRC that can be potentially used in the clinic for CRC detection and prediction of survival and recurrence. Firstly, we defined a pipeline based on a statistical analysis to identify which genes and CpG sites are differentially expressed and methylated in each TNM stage of tumor tissue, when compared to normal tissue.

Epigenetic alterations are present in CRC progression although only 85 genes with significant different expression profiles are common to all stages of CRC progression. Interestingly, we also found expression and methylation profiles specific to different stages of CRC.

Moreover, this specific expression and methylation profiles can distinguish two sub-groups of patients with distinct outcomes at the same cancer stage.

Additionally, we have identified genes differentially expressed with CpG sites differentially methylated in CRC that have never been reported to be associated to CRC or cancer in general.

Furthermore, our multivariate analysis showed that our pipeline was efficient in distinguishing tumor from normal samples and that a small set of genes with distinct expression and methylation profiles can distinguish different subgroups of patients in the same tumor stage.

In conclusion, our findings evidence that epigenetic alterations are dynamic across CRC initiation and progression and may have clinical applications.

However, further analyzes are needed. In the future, these findings should be validated in other cohorts, including cohorts with a larger sample size for each stage.

REFERENCES

1. Hajdu, S. I. Greco-Roman thought about cancer. *Cancer* **100**, 2048–2051 (2004).
2. Sudhakar, A. History of Cancer, Ancient and Modern Treatment Methods. *J. Cancer Sci. Ther.* **01**, i–iv (2009).
3. Hajdu, S. I. Thoughts about the cause of cancer. *Cancer* **106**, 1643–1649 (2006).
4. Alter, N. M. Mechanical Irritation as Etiologic Factor of Cancer: Clinical Observation. *Am. J. Pathol.* **1**, 511–518.3 (1925).
5. Plimmer, H. G. THE PARASITIC THEORY OF CANCER. *Br. Med. J.* **2**, 1511–5 (1903).
6. Mathur, G., Nain, S. & Sharma, P. K. Cancer: An Overview. *Acad. J. Cancer Res.* **8**, 1–9 (2015).
7. Weinberg, R. A. *The Biology of Cancer*. (Garland Science, Taylor & Francis Group, 1943).
8. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–E386 (2015).
9. Hatzimichael, E., Lagos, K., Sim, V. R., Briasoulis, E. & Crook, T. Epigenetics in diagnosis, prognostic assessment and treatment of cancer: an update. *EXCLI J.* **13**, 954–76 (2014).
10. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68**, 394–424 (2018).
11. Laurence, L. Genetic Mutation. *Nat. Educ.* **1**, 113 (2008).
12. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nature Reviews Genetics* **15**, 585–598 (2014).
13. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **340**, 1546–1558 (2013).
14. Merid, S. K., Goranskaya, D. & Alexeyenko, A. Distinguishing between driver and passenger mutations in individual cancer genomes by network enrichment analysis. *BMC Bioinformatics* **15**, (2014).
15. Pon, J. R. & Marra, M. A. Driver and Passenger Mutations in Cancer. *Annu. Rev. Pathol. Mech. Dis.* **10**, 25–50 (2015).
16. Macconaill, L. E. & Garraway, L. A. Clinical implications of the cancer genome. *J. Clin. Oncol.* **28**, 5219–28 (2010).
17. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
18. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
19. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–74 (2011).
20. Witsch, E., Sela, M. & Yarden, Y. Roles for Growth Factors in Cancer

Progression. *Physiology* **25**, 85–101 (2010).

21. Ruan, W.-J. & Lai, M.-D. Autocrine Stimulation in Colorectal Carcinoma (CRC): Positive Autocrine Loops in Human Colorectal Carcinoma and Applicable Significance of Blocking the Loops. *Med. Oncol.* **21**, 01–08 (2004).

22. Zilfou, J. T. & Lowe, S. W. Tumor Suppressive Functions of p53. *Cold Spring Harb. Perspect. Biol.* **1**, a001883–a001883 (2009).

23. Dyson, N. J. RB1 : a prototype tumor suppressor and an enigma. *Genes Dev.* **30**, 1492–1502 (2016).

24. Okegawa, T., Pong, R.-C., Li, Y. & Hsieh, J.-T. The role of cell adhesion molecule in cancer progression and its application in cancer therapy. *Acta Biochim. Pol.* **51**, 445–57 (2004).

25. Elmore, S. Apoptosis: A Review of Programmed Cell Death. *Toxicol. Pathol.* **35**, 495–516 (2007).

26. Bauer, J. H. & Helfand, S. L. New tricks of an old molecule: lifespan regulation by p53. *Aging Cell* **5**, 437–440 (2006).

27. Wong, R. S. Apoptosis in cancer: from pathogenesis to treatment. *J. Exp. Clin. Cancer Res.* **30**, 87 (2011).

28. Jafri, M. A., Ansari, S. A., Alqahtani, M. H. & Shay, J. W. Roles of telomeres and telomerase in cancer, and advances in telomerase-targeted therapies. *Genome Med.* **8**, 69 (2016).

29. Zvereva, M. I., Shcherbakova, D. M. & Dontsova, O. A. Telomerase: structure, functions, and activity regulation. *Biochemistry. (Mosc).* **75**, 1563–83 (2010).

30. Nishida, N., Yano, H., Nishida, T., Kamura, T. & Kojiro, M. Angiogenesis in cancer. *Vasc. Health Risk Manag.* **2**, 213–9 (2006).

31. Nagy, J. A., Chang, S.-H., Dvorak, A. M. & Dvorak, H. F. Why are tumour blood vessels abnormal and why is it important to know? *Br. J. Cancer* **100**, 865–869 (2009).

32. Wu, Y., Sarkissyan, M. & Vadgama, J. Epithelial-Mesenchymal Transition and Breast Cancer. *J. Clin. Med.* **5**, 13 (2016).

33. Xiao, D. & He, J. Epithelial mesenchymal transition and lung cancer. *J. Thorac. Dis.* **2**, 154–9 (2010).

34. Vu, T. & Datta, P. Regulation of EMT in Colorectal Cancer: A Culprit in Metastasis. *Cancers (Basel).* **9**, 171 (2017).

35. Kalluri, R. & Weinberg, R. A. The basics of epithelial-mesenchymal transition. *J. Clin. Invest.* **119**, 1420–1428 (2009).

36. Roche, J. The Epithelial-to-Mesenchymal Transition in Cancer. *Cancers (Basel).* **10**, 52 (2018).

37. Wei Dai, Y. Y. Genomic Instability and Cancer. *J. Carcinog. Mutagen.* **05**, (2014).

38. Iengar, P. Identifying pathways affected by cancer mutations. *Genomics* **110**,

318–328 (2018).

39. Broustas, C. G. & Lieberman, H. B. DNA Damage Response Genes and the Development of Cancer Metastasis. *Radiat. Res.* **181**, 111–130 (2014).

40. Bondar, T. & Medzhitov, R. The Origins of Tumor-Promoting Inflammation. *Cancer Cell* **24**, 143–144 (2013).

41. Colotta, F., Allavena, P., Sica, A., Garlanda, C. & Mantovani, A. Cancer-related inflammation, the seventh hallmark of cancer: links to genetic instability. *Carcinogenesis* **30**, 1073–1081 (2009).

42. Liberti, M. V. & Locasale, J. W. The Warburg Effect: How Does it Benefit Cancer Cells? *Trends Biochem. Sci.* **41**, 211–218 (2016).

43. Jiang, B. Aerobic glycolysis and high level of lactate in cancer metabolism and microenvironment. *Genes Dis.* **4**, 25–27 (2017).

44. Vinay, D. S. *et al.* Immune evasion in cancer: Mechanistic basis and therapeutic strategies. *Semin. Cancer Biol.* **35**, S185–S198 (2015).

45. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337.e10 (2018).

46. Tecalco-Cruz, A. C., Ríos-López, D. G., Vázquez-Victorio, G., Rosales-Alvarez, R. E. & Macías-Silva, M. Transcriptional cofactors Ski and SnoN are major regulators of the TGF-β/Smad signaling pathway in health and disease. *Signal Transduct. Target. Ther.* **3**, 15 (2018).

47. Kitamura, K., Aota, S., Sakamoto, R., Emori, T. & Okazaki, K. Smad7 induces G0/G1 cell cycle arrest in mesenchymal cells by inhibiting the expression of G1 cyclins. *Dev. Growth Differ.* **47**, 537–552 (2005).

48. Shinagawa, T., Dong, H.-D., Xu, M., Maekawa, T. & Ishii, S. The sno gene, which encodes a component of the histone deacetylase complex, acts as a tumor suppressor in mice. *EMBO J.* **19**, 2280–2291 (2000).

49. Colak, S. & ten Dijke, P. Targeting TGF-β Signaling in Cancer. *Trends in Cancer* **3**, 56–71 (2017).

50. Pardali, K. & Moustakas, A. Actions of TGF-β as tumor suppressor and pro-metastatic factor in human cancer. *Biochim. Biophys. Acta - Rev. Cancer* **1775**, 21–62 (2007).

51. Lebrun, J.-J. The Dual Role of TGF <math id="M1"> <mrow> <mi mathvariant="bold-italic">β</mi> </mrow> </math> in Human Cancer: From Tumor Suppression to Cancer Metastasis. *ISRN Mol. Biol.* **2012**, 1–28 (2012).

52. Seoane, J. & Gomis, R. R. TGF-β Family Signaling in Tumor Suppression and Cancer Progression. *Cold Spring Harb. Perspect. Biol.* **9**, a022277 (2017).

53. Dang, C. V. MYC on the Path to Cancer. *Cell* **149**, 22–35 (2012).

54. Dang, C. V. Enigmatic MYC Conducts an Unfolding Systems Biology Symphony. *Genes Cancer* **1**, 526–531 (2010).

55. Yu, J. S. L. & Cui, W. Proliferation, survival and metabolism: the role of

PI3K/AKT/mTOR signalling in pluripotency and cell fate determination. *Development* **143**, 3050–3060 (2016).

56. Cunningham, J. T. & Ruggero, D. New Connections between Old Pathways: PDK1 Signaling Promotes Cellular Transformation through PLK1-Dependent MYC Stabilization. *Cancer Discov.* **3**, 1099–1102 (2013).

57. Tan, J. *et al.* PDK1 Signaling Toward PLK1-MYC Activation Confers Oncogenic Transformation, Tumor-Initiating Cell Activation, and Resistance to mTOR-Targeted Therapy. *Cancer Discov.* **3**, 1156–1171 (2013).

58. Janku, F., Yap, T. A. & Meric-Bernstam, F. Targeting the PI3K pathway in cancer: are we making headway? *Nat. Rev. Clin. Oncol.* **15**, 273–291 (2018).

59. Regad, T. Targeting RTK Signaling Pathways in Cancer. *Cancers (Basel).* **7**, 1758–1784 (2015).

60. Schöneborn, H., Raudzus, F., Coppey, M., Neumann, S. & Heumann, R. Perspectives of RAS and RHEB GTPase Signaling Pathways in Regenerating Brain Neurons. *Int. J. Mol. Sci.* **19**, 4052 (2018).

61. Menegon, S., Columbano, A. & Giordano, S. The Dual Roles of NRF2 in Cancer. *Trends Mol. Med.* **22**, 578–593 (2016).

62. Leinonen, H. M., Kansanen, E., Pölönen, P., Heinäniemi, M. & Levonen, A.-L. Role of the Keap1–Nrf2 Pathway in Cancer. in *Advances in Cancer Research* 281–320 (Elsevier Inc., 2014). doi:10.1016/B978-0-12-420117-0.00008-6

63. Zhao, H., Eguchi, S., Alam, A. & Ma, D. The role of nuclear factor-erythroid 2 related factor 2 (Nrf-2) in the protection against lung injury. *Am. J. Physiol. Cell. Mol. Physiol.* **312**, L155–L162 (2017).

64. Zhan, T., Rindtorff, N. & Boutros, M. Wnt signaling in cancer. *Oncogene* **36**, 1461–1473 (2017).

65. Centelles, J. J. General Aspects of Colorectal Cancer. *ISRN Oncol.* **2012**, 1–19 (2012).

66. Boland, C. R. INFECTION, INFLAMMATION, AND GASTROINTESTINAL CANCER. *Gut* **54**, 1321–1331 (2005).

67. Hao, Q. & Cho, W. Battle Against Cancer: An Everlasting Saga of p53. *Int. J. Mol. Sci.* **15**, 22109–22127 (2014).

68. Brooks, C. L. & Gu, W. New insights into p53 activation. *Cell Res.* **20**, 614–621 (2010).

69. Toufektchan, E. & Toledo, F. The Guardian of the Genome Revisited: p53 Downregulates Genes Required for Telomere Maintenance, DNA Repair, and Centromere Structure. *Cancers (Basel).* **10**, 135 (2018).

70. Avila, J. L. & Kissil, J. L. Notch signaling in pancreatic cancer: oncogene or tumor suppressor? *Trends Mol. Med.* **19**, 320–327 (2013).

71. Kopan, R. Notch Signaling. *Cold Spring Harb. Perspect. Biol.* **4**, a011213–a011213 (2012).

72. Lasky, J. L. & Wu, H. Notch Signaling, Brain Development, and Human Disease. *Pediatr. Res.* **57**, 104R–109R (2005).

73. Coppedè, F., Lopomo, A., Spisni, R. & Migliore, L. Genetic and epigenetic biomarkers for diagnosis, prognosis and treatment of colorectal cancer. *World J. Gastroenterol.* **20**, 943–956 (2014).

74. Vatandoust, S., Price, T. J. & Karapetis, C. S. Colorectal cancer: Metastases to a single organ. *World J. Gastroenterol.* **21**, 11767–11776 (2015).

75. Brenner, H., Kloor, M. & Pox, C. P. Colorectal cancer. in *The Lancet* **383**, 1490–1502 (2014).

76. Globocan, I. http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx. *World Health Organization* (2012). doi:10.1074/jbc.M111.260794

77. Center, M. M., Jemal, A., Smith, R. A. & Ward, E. Worldwide variations in colorectal cancer. *Diseases of the Colon and Rectum* **53**, 1099 (2010).

78. Ting, W.-C. *et al.* Common genetic variants in Wnt signaling pathway genes as potential prognostic biomarkers for colorectal cancer. *PLoS One* **8**, e56196 (2013).

79. Migliore, L., Migheli, F., Spisni, R. & Copped, F. Genetics, cytogenetics, and epigenetics of colorectal cancer. *Journal of Biomedicine and Biotechnology* **2011**, (2011).

80. Taylor, D. P., Burt, R. W., Williams, M. S., Haug, P. J. & Cannon-Albright, L. A. Population-Based Family History-Specific Risks for Colorectal Cancer: A Constellation Approach. *Gastroenterology* **138**, 877–885 (2010).

81. Lao, V. V. & Grady, W. M. Epigenetics and colorectal cancer. *Nat. Rev. Gastroenterol. Hepatol.* **8**, 686–700 (2011).

82. American Cancer Society. Colorectal Cancer Risk Factors. Available at: https://www.cancer.org/cancer/colon-rectal-cancer/causes-risks-prevention/risk-factors.html.

83. Klatsky, A. L. *et al.* Alcohol intake, beverage choice, and cancer: a cohort study in a large kaiser permanente population. *Perm. J.* **19**, 28–34 (2015).

84. Alexander, D. D., Weed, D. L., Miller, P. E. & Mohamed, M. A. Red Meat and Colorectal Cancer: A Quantitative Update on the State of the Epidemiologic Science. *Journal of the American College of Nutrition* **34**, 521–543 (2015).

85. Chan, D. S. M. *et al.* Red and processed meat and colorectal cancer incidence: Meta-analysis of prospective studies. *PLoS One* **6**, (2011).

86. Coppedè, F. Epigenetic biomarkers of colorectal cancer: Focus on DNA methylation. *Cancer Letters* **342**, 238–247 (2014).

87. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).

88. Verma, M. & Kumar, V. Epigenetic Biomarkers in Colorectal Cancer. *Molecular Diagnosis and Therapy* **21**, 153–165 (2017).

89. Easwaran, H., Tsai, H.-C. & Baylin, S. B. Cancer Epigenetics: Tumor Heterogeneity, Plasticity of Stem-like States, and Drug Resistance. *Mol. Cell* **54**, 716–727 (2014).

90. Armaghany, T., Wilson, J. D., Chu, Q. & Mills, G. Genetic alterations in colorectal cancer. *Gastrointest. Cancer Res.* **5**, 19–27 (2012).

91. Walther, A. *et al.* Genetic prognostic and predictive markers in colorectal cancer. *Nat. Rev. Cancer* **9**, 489–499 (2009).

92. American Cancer Society. Colorectal Cancer Stages. (2017).

93. Roadknight, C. *et al.* Biomarker Clustering of Colorectal Cancer Data to Complement Clinical Classification. *SSRN Electron. J.* (2012). doi:10.2139/ssrn.2828496

94. Zauber, A. G. *et al.* Colonoscopic Polypectomy and Long-Term Prevention of Colorectal-Cancer Deaths. *N. Engl. J. Med.* **366**, 687–696 (2012).

95. Rex, Douglas K; Boland, C Richard; Dominitz, Jason A; Giardiello, Francis M; Johnson, David A; Kaltenbach, Tonya; Levin, Theodore R; Lieberman, David; Robertson, D. J. Colorectal Cancer Screening: Recommendations for Physicians and Patients from the U.S. Multi-Society Task Force on Colorectal Cancer. *Am. J. Gastroenterol.* **112**, 1016–1030 (2017).

96. Center, M. M., Jemal, A., Smith, R. A. & Ward, E. Worldwide variations in colorectal cancer. *Dis. Colon Rectum* **53**, 1099 (2010).

97. Schreuders, E. H. *et al.* Colorectal cancer screening: A global overview of existing programmes. *Gut* **64**, 1637–1649 (2015).

98. Fakih, M. G. & Padmanabhan, A. CEA monitoring in colorectal cancer. What you should know. *Oncology (Williston Park).* **20**, 579–587; discussion 588, 594, 596 passim (2006).

99. Duffy, M. J. Carcinoembryonic antigen as a marker for colorectal cancer: Is it clinically useful? *Clinical Chemistry* **47**, 624–630 (2001).

100. Society, A. C. What Are the Survival Rates for Colorectal Cancer, by Stage? (2017). Available at: https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/survival-rates.html. (Accessed: 20th August 2003)

101. Feinberg, A. P. & Tycko, B. The history of cancer epigenetics. *Nat. Rev. Cancer* **4**, 143–153 (2004).

102. Moore, L. D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **38**, 23–38 (2013).

103. Waddington CH. The epigenotype. *Endeavour* (1942).

104. Sharma, S., Kelly, T. K. & Jones, P. A. Epigenetics in cancer. *Carcinogenesis* **31**, 27–36 (2009).

105. Ahuja, N., Sharma, A. R. & Baylin, S. B. Epigenetic Therapeutics: A New Weapon in the War Against Cancer. *Annu. Rev. Med.* **67**, 73–89 (2016).

106. Dupont, C., Armant, D. & Brenner, C. Epigenetics: Definition, Mechanisms and Clinical Perspective. *Semin. Reprod. Med.* **27**, 351–357 (2009).

107. Weksberg, R., Butcher, D. T., Grafodatskaya, D., Choufani, S. & Tycko, B. Epigenetics. in *Emery and Rimoin's Principles and Practice of Medical Genetics* 1–31 (Elsevier, 2013). doi:10.1016/B978-0-12-383834-6.00006-9

108. McBryan, T. & Adams, P. D. Epigenetics. in *Handbook of Pharmacogenomics and Stratified Medicine* 57–69 (Elsevier, 2014). doi:10.1016/B978-0-12-386882-4.00004-9

109. Renaud, F. *et al.* MUC5AC hypomethylation is a predictor of microsatellite instability independently of clinical factors associated with colorectal cancer. *Int. J. Cancer* **136**, 2811–2821 (2015).

110. Ebert, M. P. A. *et al.* TFAP2E-DKK4 and Chemoresistance in Colorectal Cancer. *N. Engl. J. Med.* **366**, 44–53 (2012).

111. Crea, F. *et al.* Epigenetics and chemoresistance in colorectal cancer: An opportunity for treatment tailoring and novel therapeutic strategies. *Drug Resist. Updat.* **14**, 280–296 (2011).

112. Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843–854 (1993).

113. Chuang, J. C. & Jones, P. A. Epigenetics and microRNAs. *Pediatr. Res.* **61**, 24–29 (2007).

114. Meltzer, P. S. NEWS & VIEWS Small RNAs with big impacts. *Nature* **435**, 0–1 (2005).

115. MacFarlane, L.-A. & R. Murphy, P. MicroRNA: Biogenesis, Function and Role in Cancer. *Curr. Genomics* **11**, 537–561 (2010).

116. Lin, S. & Gregory, R. I. MicroRNA biogenesis pathways in cancer. *Nat. Rev. Cancer* **15**, 321–333 (2015).

117. Nelson, K. M. & Weiss, G. J. MicroRNAs and cancer: past, present, and potential future. *Mol. Cancer Ther.* **7**, 3655–3660 (2008).

118. Hayes, J., Peruzzi, P. P. & Lawler, S. MicroRNAs in cancer: Biomarkers, functions and therapy. *Trends in Molecular Medicine* **20**, 460–469 (2014).

119. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **128**, 693–705 (2007).

120. Sawan, C. & Herceg, Z. *Histone Modifications and Cancer*. *Advances in Genetics* **70**, (2010).

121. Chen, R., Kang, R., Fan, X.-G. & Tang, D. Release and activity of histone in diseases. *Cell Death Dis.* **5**, e1370–e1370 (2014).

122. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**, 823–837 (2007).

123. Jin, B. *et al.* DNMT1 and DNMT3B Modulate Distinct Polycomb-Mediated

Histone Modifications in Colon Cancer. *Cancer Res.* **69**, 7412–7421 (2009).

124. Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).

125. Portela, A. & Esteller, M. Epigenetic modifications and human disease. *Nature Biotechnology* **28**, 1057–1068 (2010).

126. Jin, B., Li, Y. & Robertson, K. D. DNA methylation: superior or subordinate in the epigenetic hierarchy? *Genes Cancer* **2**, 607–17 (2011).

127. Bert, S. A. *et al.* Regional activation of the cancer genome by long-range epigenetic remodeling. *Cancer Cell* **23**, 9–22 (2013).

128. Strimbu, K. & Tavel, J. A. What are biomarkers? *Current Opinion in HIV and AIDS* **5**, 463–466 (2010).

129. Danese, E. *et al.* Epigenetic alteration: new insights moving from tissue to plasma – the example of PCDH10 promoter methylation in colorectal cancer. *Br. J. Cancer* **109**, 807–813 (2013).

130. Leygo, C. *et al.* DNA Methylation as a Noninvasive Epigenetic Biomarker for the Detection of Cancer. *Dis. Markers* **2017**, 1–13 (2017).

131. Tang, D. *et al.* Diagnostic and prognostic value of the methylation status of secreted frizzled-related protein 2 in colorectal cancer. *Clin. Invest. Med.* **34**, E88-95 (2011).

132. Lange, C. P. E. *et al.* Genome-Scale Discovery of DNA-Methylation Biomarkers for Blood-Based Detection of Colorectal Cancer. *PLoS One* **7**, (2012).

133. Oh, T. *et al.* Genome-wide identification and validation of a novel methylation biomarker, SDC2, for blood-based detection of colorectal cancer. *J. Mol. Diagnostics* **15**, 498–507 (2013).

134. Luo, Y. X., Chen, D. K., Song, S. X., Wang, L. & Wang, J. P. Aberrant methylation of genes in stool samples as diagnostic biomarkers for colorectal cancer or adenomas: A meta-analysis. *International Journal of Clinical Practice* **65**, 1313–1320 (2011).

135. Yang, H. *et al.* Diagnostic value of stool DNA testing for multiple markers of colorectal cancer and advanced adenoma: A meta-analysis. *Canadian Journal of Gastroenterology* **27**, 467–475 (2013).

136. Blei, D. M. & Smyth, P. Science and data science. *Proc. Natl. Acad. Sci.* **114**, 8689–8692 (2017).

137. Cao, L. Data science and analytics: a new era. *Int. J. Data Sci. Anal.* **1**, 1–2 (2016).

138. Marx, V. The big challenges of big data. *Nature* **498**, 255–260 (2013).

139. Dunn, M. C. & Bourne, P. E. Building the biomedical data science workforce. *PLOS Biol.* **15**, e2003082 (2017).

140. R Core Team. R Development Core Team. *R: A Language and Environment for Statistical Computing* **55**, 275–286 (2017).

141. Jonge, E. van der Loo, M. An introduction to data cleaningwith R. *Stat. Netherlands* (2013).

142. Colaprico, A. *et al.* TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71 (2016).

143. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. (Poznań, Poland)* **19**, A68-77 (2015).

144. Silva, T. C. *et al.* TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Research* **5**, 1542 (2016).

145. Dedeurwaerder, S. *et al.* A comprehensive overview of Infinium HumanMethylation450 data processing. *Brief. Bioinform.* **15**, 929–941 (2014).

146. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).

147. Kovalchik, S. RISmed: Download Content from NCBI Databases. R package version 2.1.7 (2017).

148. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77 (2011).

149. El Khouli, R. H. *et al.* Relationship of temporal resolution to diagnostic performance for dynamic contrast enhanced MRI of the breast. *J. Magn. Reson. Imaging* **30**, 999–1004 (2009).

150. Therneau, T. & Grambsch, P. Modeling Survival Data: Extending the Cox Model. *Springer-Verlag* (2000).

151. Therneau, T. A Package for Survival Analysis in S. version 2.38 (2015).

152. Greenland, S. *et al.* Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur. J. Epidemiol.* **31**, 337–350 (2016).

153. Bakan, D. The test of significance in psychological research. *Psychol. Bull.* **66**, 423–437 (1966).

154. Qu, H.-Q. H.-Q., Tien, M. & Polychronakos, C. Statistical significance in genetic association studies. *Clin. Investig. Med.* **33**, E266–E270 (2010).

155. MCDONALD, J. H. *HANDBOOK OF BIOLOLOGICAL STATISTICS*. (SPARKY HOUSE PUBLISHING, 2014).

156. Hair, J. F., Anderson, R. E., Tatham, R. E., Black, W. C. *Análisis Multivariante*. (1999).

157. Kitchen, C. M. R. Nonparametric vs Parametric Tests of Location in Biomedical Research. *American Journal of Ophthalmology* **147**, 571–572 (2009).

158. Aguinis, H., Gottfredson, R. K. & Joo, H. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organ. Res. Methods* **16**, 270–301 (2013).

159. Kwak, S. K. & Kim, J. H. Statistical data preparation: management of missing values and outliers. *Korean J. Anesthesiol.* **70**, 407 (2017).

160. Rousseeuw, P. J. & Hubert, M. Robust statistics for outlier detection. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **1**, 73–79 (2011).

161. Razali, N. M. & Wah, Y. B. Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J. Stat. Model. Anal.* **2**, 21–33 (2011).

162. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bull.* (1945). doi:10.2307/3001968

163. Welch, B. L. The Generalization of `Student's' Problem when Several Different Population Variances are Involved. *Biometrika* **34**, 28 (1947).

164. Kim, T. K. T test as a parametric statistic. *Korean J. Anesthesiol.* **68**, 540 (2015).

165. Levene, H. Robust tests for equality of variances. In Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. *Stanford Univ. Press* 278–292 (1960).

166. Weisberg, S. & Fox, J. *An R Companion to Applied Regression*. (Thousand Oaks: Sage, 2011).

167. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**, 289–300 (1995).

168. Mukaka, M. M. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* **24**, 69–71 (2012).

169. Spearman, C. Spearman ' s rank correlation coefficient. *Amer. J. Psychol.* **15**, 72–101 (1904).

170. Spearman, C. The Proof and Measurement of Association between Two Things. *Am. J. Psychol.* **15**, 72 (1904).

171. Metz, C. E. Basic principles of ROC analysis. *Semin. Nucl. Med.* **VIII**, 238–298 (1978).

172. Mandrekar, J. N. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology* (2010). doi:10.1097/JTO.0b013e3181ec173d

173. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *J. Am. Stat. Assoc.* **53**, 457 (1958).

174. Clark, T. G., Bradburn, M. J., Love, S. B. & Altman, D. G. Survival Analysis Part I: Basic concepts and first analyses. *Br. J. Cancer* **89**, 232–238 (2003).

175. Jager, K. J., van Dijk, P. C., Zoccali, C. & Dekker, F. W. The analysis of survival data: the Kaplan–Meier method. *Kidney Int.* **74**, 560–565 (2008).

176. Andersen, P. K. & Gill, R. D. Cox's Regression Model for Counting Processes: A Large Sample Study. *Ann. Stat.* (1982). doi:10.1214/aos/1176345976

177. Cox, D. R. Regression models and life tables. *J. R. Stat. Soc. Ser. B 34* 187–200 (1972).

178. Bradburn, M. J., Clark, T. G., Love, S. B. & Altman, D. G. Survival Analysis

Part II: Multivariate data analysis – an introduction to concepts and methods. *Br. J. Cancer* **89**, 431–436 (2003).

179. Kassambara, A. & Kosinski, M. survminer: Drawing Survival Curves using ‘ggplot2’. R package version 0.4.2 (2018).

180. Galindo, M. P. Una alternativa de representación simultánea: HJ-biplot. *Questíio* **10**, 13–23 (1986).

181. Vicente-Villardon, J. L. MultBiplotR: MULTivariate Analysis Using BIPLOTs. R package version 18.2.09 (2018).

182. Tenenbaum, D. KEGGREST: Client-side REST access to KEGG. R package version 1.20.0 (2018).

183. Chen, H. & Boutros, P. C. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* **12**, 35 (2011).

184. Fishell, G. & Hatten, M. E. Astrotactin provides a receptor system for CNS neuronal migration. *Development* **113**, 755–65 (1991).

185. Zheng, C., Heintz, N. & Hatten, M. E. CNS gene encoding astrotactin, which supports neuronal migration along glial fibers. *Science* **272**, 417–9 (1996).

186. Adams, N. C., Tomoda, T., Cooper, M., Dietz, G. & Hatten, M. E. Mice that lack astrotactin have slowed neuronal migration. *Development* **129**, 965–72 (2002).

187. Beggs, A. *et al.* Hypermethylation of SNAP91 as an alternative mechanism of epidermal growth factor signalling dysregulation: a genome-wide meta-analysis with validation of colorectal cancers. *Lancet* **383**, S25 (2014).

188. Somaiah, N. *et al.* Targeted next generation sequencing of well-differentiated/dedifferentiated liposarcoma reveals novel gene amplifications and mutations. *Oncotarget* **9**, 19891–19899 (2018).

189. Qin, Z. *et al.* ZNF536, a Novel Zinc Finger Protein Specifically Expressed in the Brain, Negatively Regulates Neuron Differentiation by Repressing Retinoic Acid-Induced Gene Transcription. *Mol. Cell. Biol.* **29**, 3633–3643 (2009).

190. Li, J. *et al.* A data mining paradigm for identifying key factors in biological processes using gene expression data. *Sci. Rep.* **8**, 9083 (2018).

191. Lebbink, R. J. *et al.* The soluble leukocyte-associated Ig-like receptor (LAIR)-2 antagonizes the collagen/LAIR-1 inhibitory immune interaction. *J. Immunol.* **180**, 1662–9 (2008).

192. Lin, Y.-W. *et al.* SOX1 suppresses cell growth and invasion in cervical cancer. *Gynecol. Oncol.* **131**, 174–181 (2013).

193. Geng, L. N. *et al.* DUX4 Activates Germline Genes, Retroelements, and Immune Mediators: Implications for Facioscapulohumeral Dystrophy. *Dev. Cell* **22**, 38–51 (2012).

194. Heubach, J. *et al.* The long noncoding RNA HOTAIR has tissue and cell type-dependent effects on HOX gene expression and phenotype of urothelial cancer cells. *Mol. Cancer* **14**, 108 (2015).

195. de Vetten, M., Ploos van Amstel, H. K. & Reitsma, P. H. RFLP for the human prothrombin (F2) gene. *Nucleic Acids Res.* **18**, 5917 (1990).

196. Pozzi, N. & Di Cera, E. Prothrombin structure: unanticipated features and opportunities. *Expert Rev. Proteomics* **11**, 653–655 (2014).

197. Jakobs, P. M. *et al.* Autosomal-Dominant Congenital Cataract Associated with a Deletion Mutation in the Human Beaded Filament Protein Gene BFSP2. *Am. J. Hum. Genet.* **66**, 1432–1436 (2000).

198. DEMOKAN, S. *et al.* Validation of nucleolar protein 4 as a novel methylated tumor suppressor gene in head and neck cancer. *Oncol. Rep.* **31**, 1014–1020 (2014).

199. Zhang, S., Li, M., Ji, H. & Fang, Z. Landscape of transcriptional deregulation in lung cancer. *BMC Genomics* **19**, 435 (2018).

200. Gasa, L. *et al.* A systematic analysis of orphan cyclins reveals CNTD2 as a new oncogenic driver in lung cancer. *Sci. Rep.* **7**, 10228 (2017).

201. Burns, M. E., Sasaki, T., Takai, Y. & Augustine, G. J. Rabphilin-3A: A Multifunctional Regulator of Synaptic Vesicle Traffic. *J. Gen. Physiol.* **111**, 243–255 (1998).

202. Kato, K. *et al.* Structural and functional insights into IZUMO1 recognition by JUNO in mammalian fertilization. *Nat. Commun.* **7**, 12198 (2016).

203. Lagman, D., Sundström, G., Ocampo Daza, D., Abalo, X. M. & Larhammar, D. Expansion of transducin subunit gene families in early vertebrate tetraploidizations. *Genomics* **100**, 203–211 (2012).

204. Soll, C. *et al.* Serotonin promotes tumor growth in human hepatocellular cancer. *Hepatology* **51**, 1244–54 (2010).

205. Sarrouilhe, D., Clarhaut, J., Defamie, N. & Mesnil, M. Serotonin and cancer: what is the link? *Curr. Mol. Med.* **15**, 62–77 (2015).

206. Babu, E. *et al.* Role of SLC5A8, a plasma membrane transporter and a tumor suppressor, in the antitumor activity of dichloroacetate. *Oncogene* **30**, 4026–4037 (2011).

207. Raposo, A. A. S. F. *et al.* Ascl1 Coordinately Regulates Gene Expression and the Chromatin Landscape during Neurogenesis. *Cell Rep.* **10**, 1544–1556 (2015).

208. Choudhary, P. *et al.* Discovery of Compounds that Positively Modulate the High Affinity Choline Transporter. *Front. Mol. Neurosci.* **10**, (2017).

209. Wei, J. *et al.* Integrated analysis of genome-wide DNA methylation and gene expression profiles identifies potential novel biomarkers of rectal cancer. *Oncotarget* **7**, (2016).

210. Naumov, V. A. *et al.* Genome-scale analysis of DNA methylation in colorectal cancer using Infinium HumanMethylation450 BeadChips. *Epigenetics* **8**, 921–934 (2013).

211. Vidal, E. *et al.* A DNA methylation map of human cancer at single base-pair resolution. *Oncogene* **36**, 5648–5657 (2017).

212. Lim, D. H. K. & Maher, E. R. DNA methylation: a form of epigenetic control of gene expression. *Obstet. Gynaecol.* **12**, 37–42 (2010).

213. Siegfried, Z. & Simon, I. DNA methylation and gene expression. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2**, 362–371 (2010).

214. Shigaki, H. *et al.* Epigenetic changes in gastrointestinal cancers. *J. Cancer Metastasis Treat.* **1**, 113 (2015).

215. Fardi, M., Solali, S. & Farshdousti Hagh, M. Epigenetic mechanisms as a new approach in cancer treatment: An updated review. *Genes Dis.* **5**, 304–311 (2018).

216. Burggren, W. W. Dynamics of epigenetic phenomena: intergenerational and intragenerational phenotype 'washout'. *J. Exp. Biol.* **218**, 80–87 (2015).

217. Kim, M. & Costello, J. DNA methylation: an epigenetic mark of cellular memory. *Exp. Mol. Med.* **49**, e322–e322 (2017).

218. Kim, Y.-H. *et al.* CpG island methylation of genes accumulates during the adenoma progression step of the multistep pathogenesis of colorectal cancer. *Genes, Chromosom. Cancer* **45**, 781–789 (2006).

219. Øster, B. *et al.* Identification and validation of highly frequent CpG island hypermethylation in colorectal adenomas and carcinomas. *Int. J. Cancer* **129**, 2855–2866 (2011).

220. Mishra, N. K. & Guda, C. Genome-wide DNA methylation analysis reveals molecular subtypes of pancreatic cancer. *Oncotarget* **8**, (2017).

221. Yang, X. *et al.* Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* **26**, 577–590 (2014).

222. Neri, F. *et al.* Intragenic DNA methylation prevents spurious transcription initiation. *Nature* **543**, 72–77 (2017).

223. Teissandier, A. & Bourc'his, D. Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. *EMBO J.* **36**, 1471–1473 (2017).

224. Castelo-Branco, P. *et al.* Methylation of the TERT promoter and risk stratification of childhood brain tumours: an integrative genomic and molecular study. *Lancet Oncol.* **14**, 534–542 (2013).

225. Dagogo-Jack, I. & Shaw, A. T. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **15**, 81–94 (2017).

226. Satow, R., Inagaki, S., Kato, C., Shimozawa, M. & Fukami, K. Identification of zinc finger protein of the cerebellum 5 as a survival factor of prostate and colorectal cancer cells. *Cancer Sci.* **108**, 2405–2412 (2017).

227. Druck, T. *et al.* The DIRC1 gene at chromosome 2q33 spans a familial RCC-associated t(2;3)(q33;q21) chromosome translocation. *J. Hum. Genet.* **46**, 583–589 (2001).

228. Kim, Y.-J., Yoon, B., Han, K. & Park, B. C. Comprehensive Transcriptome Profiling of Balding and Non-Balding Scalps in Trichorhinophalangeal Syndrome Type I Patient. *Ann. Dermatol.* **29**, 597 (2017).

229. Bian, Z. *et al.* LncRNA-FEZF1-AS1 promotes tumor proliferation and metastasis in colorectal cancer by regulating PKM2 signaling. *Clin. Cancer Res.* clincanres.2967.2017 (2018). doi:10.1158/1078-0432.CCR-17-2967

230. Li, F., Wang, T. & Tang, S. SOX14 promotes proliferation and invasion of cervical cancer cells through Wnt/β-catenin pathway. *Int. J. Clin. Exp. Pathol.* **8**, 1698–704 (2015).

231. Iancu, I. V. *et al.* LINC01101 and LINC00277 expression levels as novel factors in HPV-induced cervical neoplasia. *J. Cell. Mol. Med.* **21**, 3787–3794 (2017).

232. Jing, X. *et al.* Peptidoglycan Recognition Protein 3 and Nod2 Synergistically Protect Mice from Dextran Sodium Sulfate–Induced Colitis. *J. Immunol.* **193**, 3055–3069 (2014).

233. Kim, E. R. Colorectal cancer in inflammatory bowel disease: The risk, pathogenesis, prevention and diagnosis. *World J. Gastroenterol.* **20**, 9872–9881 (2014).

234. Testa, U., Pelosi, E. & Castelli, G. Colorectal Cancer: Genetic Abnormalities, Tumor Progression, Tumor Heterogeneity, Clonal Evolution and Tumor-Initiating Cells. *Med. Sci.* **6**, 31 (2018).

235. Luchtenborg, M., White, K. K. L., Wilkens, L., Kolonel, L. N. & Le Marchand, L. Smoking and Colorectal Cancer: Different Effects by Type of Cigarettes? *Cancer Epidemiol. Biomarkers &amp; Prev.* **16**, 1341–1347 (2007).

236. Anderson, J. C. *et al.* Prevalence of colorectal neoplasia in smokers. *Am. J. Gastroenterol.* **98**, 2777–2783 (2003).

237. Chao, A. *et al.* Cigarette smoking and colorectal cancer mortality in the cancer prevention study II. *J. Natl. Cancer Inst.* **92**, 1888–96 (2000).

238. Chen, K., Xia, G., Zhang, C. & Sun, Y. Correlation between smoking history and molecular pathways in sporadic colorectal cancer: a meta-analysis. *Int. J. Clin. Exp. Med.* **8**, 3241–57 (2015).

239. Fang, Z.-Q. *et al.* Gene expression profile and enrichment pathways in different stages of bladder cancer. *Genet. Mol. Res.* **12**, 1479–1489 (2013).

240. Liu, J. *et al.* Aberrantly methylated-differentially expressed genes and pathways in colorectal cancer. *Cancer Cell Int.* **17**, 75 (2017).

241. Vitale, G., Dicitore, A., Mari, D. & Cavagnini, F. A new therapeutic strategy against cancer: cAMP elevating drugs and leptin. *Cancer Biol. Ther.* **8**, 1191–1193 (2009).

242. Dong, H., Claffey, K. P., Brocke, S. & Epstein, P. M. Inhibition of breast cancer cell migration by activation of cAMP signaling. *Breast Cancer Res. Treat.* **152**, 17–28 (2015).

243. Wang, X., Kaczor-Urbanowicz, K. E. & Wong, D. T. W. Salivary biomarkers in cancer detection. *Med. Oncol.* **34**, 7 (2017).

244. Bajaj, A., Driver, J. A. & Schernhammer, E. S. Parkinson’s disease and cancer risk: a systematic review and meta-analysis. *Cancer Causes Control* **21**, 697–707 (2010).

245. Becker, C., Brobert, G. P., Johansson, S., Jick, S. S. & Meier, C. R. Cancer risk in association with Parkinson disease: A population-based study. *Parkinsonism Relat. Disord.* **16**, 186–190 (2010).

246. Roe, C. M. *et al.* Cancer linked to Alzheimer disease but not vascular dementia. *Neurology* **74**, 106–112 (2010).

247. Plun-Favreau, H., Lewis, P. A., Hardy, J., Martins, L. M. & Wood, N. W. Cancer and Neurodegeneration: Between the Devil and the Deep Blue Sea. *PLoS Genet.* **6**, e1001257 (2010).

248. Sato, Y. *et al.* Report of an autopsy case of colon cancer with amyotrophic lateral sclerosis. *Nihon Shokakibyo Gakkai Zasshi* **104**, 1365–70 (2007).

249. Taguchi, Y. & Wang, H. Genetic Association between Amyotrophic Lateral Sclerosis and Cancer. *Genes (Basel).* **8**, 243 (2017).

250. Lu, R. *et al.* Neurons generated from carcinoma stem cells support cancer progression. *Signal Transduct. Target. Ther.* **2**, 16036 (2017).

251. Jobling, P. *et al.* Nerve-Cancer Cell Cross-talk: A Novel Promoter of Tumor Progression. *Cancer Res.* **75**, 1777–1781 (2015).

252. Hanoun, M., Maryanovich, M., Arnal-Estapé, A. & Frenette, P. S. Neural Regulation of Hematopoiesis, Inflammation, and Cancer. *Neuron* **86**, 360–373 (2015).

253. Entschladen, F., Drell, T. L., Lang, K., Joseph, J. & Zaenker, K. S. Tumour-cell migration, invasion, and metastasis: navigation by neurotransmitters. *Lancet Oncol.* **5**, 254–258 (2004).

254. Mancino, M., Ametller, E., Gascón, P. & Almendro, V. The neuronal influence on tumor progression. *Biochim. Biophys. Acta - Rev. Cancer* **1816**, 105–118 (2011).

255. Grabowski, P. *et al.* Neuroendocrine differentiation is a relevant prognostic factor in stage III–IV colorectal cancer. *Eur. J. Gastroenterol. Hepatol.* **13**, 405–411 (2001).

256. Marchesi, F., Piemonti, L., Mantovani, A. & Allavena, P. Molecular mechanisms of perineural invasion, a forgotten pathway of dissemination and metastasis. *Cytokine Growth Factor Rev.* **21**, 77–82 (2010).

257. Bapat, A. A., Hostetter, G., Von Hoff, D. D. & Han, H. Perineural invasion and associated pain in pancreatic cancer. *Nat. Rev. Cancer* **11**, 695–707 (2011).

258. Duchalais, E. *et al.* Colorectal Cancer Cells Adhere to and Migrate Along the Neurons of the Enteric Nervous System. *Cell. Mol. Gastroenterol. Hepatol.* **5**, 31–49 (2018).

259. Inohara, N. CIDE, a novel family of cell death activators with homology to the 45kDa subunit of the DNA fragmentation factor. *EMBO J.* **17**, 2526–2533 (1998).

260. de Nadal, E., Ammerer, G. & Posas, F. Controlling gene expression in response to stress. *Nat. Rev. Genet.* **12**, 833–845 (2011).

261. Chang, H. Cleave but not leave: Astrotactin proteins in development and disease. *IUBMB Life* **69**, 572–577 (2017).

262. Jen, J. & Wang, Y.-C. Zinc finger proteins in cancer progression. *J. Biomed. Sci.* **23**, 53 (2016).

263. Huang, J. *et al.* DNA hypermethylated status and gene expression of PAX1/SOX1 in patients with colorectal carcinoma. *Onco. Targets. Ther.* **10**, 4739–4751 (2017).

264. Herman, J. G. & Baylin, S. B. Gene Silencing in Cancer in Association with Promoter Hypermethylation. *N. Engl. J. Med.* **349**, 2042–2054 (2003).

265. Gao, Y.-F. *et al.* COL3A1 and SNAP91: novel glioblastoma markers with diagnostic and prognostic value. *Oncotarget* **7**, (2016).

266. Lou, J. *et al.* Prognostic significance of SOX-1 expression in human hepatocelluar cancer. *Int. J. Clin. Exp. Pathol.* **8**, 5411–8 (2015).

267. Guan, Z. *et al.* SOX1 down-regulates β-catenin and reverses malignant phenotype in nasopharyngeal carcinoma. *Mol. Cancer* **13**, 257 (2014).

268. Miyauchi, S., Gopal, E., Fei, Y.-J. & Ganapathy, V. Functional Identification of SLC5A8, a Tumor Suppressor Down-regulated in Colon Cancer, as a Na + -coupled Transporter for Short-chain Fatty Acids. *J. Biol. Chem.* **279**, 13293–13296 (2004).

269. Park, J. Y. *et al.* Gene silencing of SLC5A8 identified by genome-wide methylation profiling in lung cancer. *Lung Cancer* **79**, 198–204 (2013).

270. Whitman, S. P. *et al.* DNA hypermethylation and epigenetic silencing of the tumor suppressor gene, SLC5A8, in acute myeloid leukemia with the MLL partial tandem duplication. *Blood* **112**, 2013–2016 (2008).

271. Hu, B.-S., Xiong, S.-M., Li, G. & Li, J.-P. Downregulation of SLC5A8 inhibits hepatocellular carcinoma progression through regulation of Wnt/β-catenin signaling. *Tumor Biol.* **37**, 13445–13453 (2016).

272. Ganapathy, V., Gopal, E., Miyauchi, S. & Prasad, P. D. Biological functions of SLC5A8, a candidate tumour suppressor. *Biochem. Soc. Trans.* **33**, 237–240 (2005).

273. Ueno, M. *et al.* Aberrant Methylation and Histone Deacetylation Associated with Silencing of SLC5A8 in Gastric Cancer. *Tumor Biol.* **25**, 134–140 (2004).

274. Park, J. Y. *et al.* Candidate tumor suppressor gene SLC5A8 is frequently down-regulated by promoter hypermethylation in prostate tumor. *Cancer Detect. Prev.* **31**, 359–365 (2007).

275. Bhutia, Y. D. *et al.* SLC transporters as a novel class of tumour suppressors: identity, function and molecular mechanisms. *Biochem. J.* **473**, 1113–1124 (2016).

276. Miyashita, N. *et al.* An Integrative Analysis of Transcriptome and Epigenome Features of ASCL1–Positive Lung Adenocarcinomas. *J. Thorac. Oncol.* **13**, 1676–1691 (2018).

277. Truong, N., Chun, S. M., Kim, T. I., Suh, Y. A. & Jang, S. J. Hypermethylation of adjacent CpG sites is negatively correlated with the expression of lineage oncogene ASCL1 in pulmonary neuroendocrine tumors. *Tumor Biol.* **39**, 101042831770622 (2017).

278. Li, M., Sun, Q. & Wang, X. Transcriptional landscape of human cancers. *Oncotarget* **8**, (2017).