André Filipe Afonso de Sousa Fonseca

Epigenetic of colorectal cancer: a focus on microRNAs

UNIVERSIDADE DO ALGARVE

Departamento de Ciências Biomédicas e Medicina

2018

André Filipe Afonso de Sousa Fonseca

Epigenetic of colorectal cancer: a focus on microRNAs

Master in Oncobiology - Molecular Mechanisms of Cancer

This work was done under the supervision of:

Pedro Castelo-Branco, Ph.D

Vânia Palma Roberto, Ph.D

UNIVERSIDADE DO ALGARVE

Departamento de Ciências Biomédicas e Medicina

2018

Epigenetic of colorectal cancer:

a focus on microRNAs

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"I declare that I am the author of this work, that is original and unpublished.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."

_____

(André Fonseca)

Copyright © 2018 André Filipe Afonso de Sousa Fonseca

A Universidade do Algarve reserva para si o direito, em conformidade com o disposto no Código do Direito de Autor e dos Direitos Conexos, de arquivar, reproduzir e publicar a obra, independentemente do meio utilizado, bem como de a divulgar através de repositórios científicos e de admitir a sua cópia e distribuição para fins meramente educacionais ou de investigação e não comerciais, conquanto seja dado o devido crédito ao autor e editor respetivos.

“If you want to succeed as bad as you want to breathe

then you will be successful!”

- Eric Thomas

Agradecimentos

Em primeiro lugar gostaria de agradecer ao meu orientador, o professor doutor Pedro Castelo-Branco por me ter acolhido no seu laboratório, por depositar em mim a sua confiança e acreditar na minha capacidade para realizar o seguinte trabalho. Gostaria também de lhe agradecer por partilhar comigo toda a sua sabedoria e todas as suas ideias que me permitiram adquirir um conhecimento inimaginável ao longo da realização desta tese.

Agradeço à minha co-orientadora, à doutora Vânia Palma Roberto por me ter acompanhado ao longo deste percurso, por toda a incansável ajuda que me forneceu e pela disponibilidade e paciência que teve para comigo. Estou extremamente agradecido por todo os conceitos que partilhou comigo, imprescindíveis para a concretização deste trabalho.

Gostaria também de agradecer à professora doutora Ana Marreiros por todo o conhecimento que me dispôs, pela disponibilidade que apesar de escassa era sempre valiosa. Acima de tudo, gostaria de agradecer pela boa disposição que apresentou diariamente, o que me proporcionou sempre imensas alegrias e sorrisos, melhorando substancialmente a experiência de concretização desta tese.

Um imenso obrigado à minha colega Sara Ramalhete e ao meu colega André Mestre, sem os quais esta tese não teria o mesmo sabor. Estou profundamente agradecido por toda a ajuda que me forneceram e por todo o tempo que passámos juntos.

Um obrigado muito especial à minha colega Joana Dias Apolónio que é umas das pessoas mais simpática, divertida e humilde que tive o prazer de conhecer, que se encontrou sempre disposta a ajudar-me ao longo jornada.

Por último gostaria de agradecer aos meus pais por me incentivarem a persistir e a dar o meu melhor ao longo desta etapa e pela paciência que só uns pais têm para aturar a impaciência de um filho.

Abstract

Colorectal cancer (CRC) is one of the most commonly diagnosed cancers and a frequent cause of cancer related deaths worldwide. Despite recent advances, CRC characterization still exhibits a lot to unveil, especially regarding the epigenetic contribution of miRNAs on disease initiation and progression.

MicroRNAs (miRNAs) are a class of small (21-23 nucleotides) endogenous non-coding RNAs that regulate gene expression at a post-transcriptional level. MiRNAs are known to regulate about 2/3 of our genes and since their discovery they have been implicated in almost every physiological process. Thus, it is not a surprise to find alterations of miRNA expression linked to several pathological conditions, including CRC. However, the role of miRNAs in CRC carcinogenesis is far from being completely understood. A clearer comprehension of these players in CRC would contribute to better understand this pathology and unveil new biomarkers for CRC detection and/or clinical management.

Here we propose to study the patterns of miRNAs behaviour throughout disease progression. Our main goal was to identify stage specific miRNAs alterations that could act as novel biomarkers. For that, we aimed to investigate (i) miRNAs expression and methylation patterns across CRC development; (ii) the targets of the differentially expressed miRNAs in order to better understand their role in CRC progression; (iii) stage specific alterations that could characterize patients within each stage of disease and provide a more accurate patient subclassification, and (iv) miRNA expression and methylation as prognostic value for CRC patients.

We found that the major miRNA deregulation events occur during Normal to Stage I transition, and are generally maintained throughout disease progression. Importantly, we show that alterations in both miRNAs expression and methylation were able to distinguish normal from malignant tissue and to predict patient's outcome, which evidences their potential as CRC diagnostic and prognostic biomarkers.

Keywords: Colorectal cancer, epigenetic, miRNAs, DNA methylation, biomarkers.

Resumo

O cancro pode ser definido como um conjunto de doenças extremamente complexas cuja heterogeneidade dificulta o seu combate. Caracterizado por uma proliferação celular descontrolada, subjacente a uma invasão de tecidos adjacentes ou órgãos distantes (metastização), o cancro é hoje em dia um problema mundial de saúde.

Dos diversos tipos de cancros identificados atualmente, o cancro colorretal (CCR) é um dos cancros mais proeminentes a nível mundial, afectando cerca de 1.23 milhões de indivíduos e contabilizando a morte de 600 mil pessoas anualmente em todo o mundo. Apesar da extensa investigação realizada no âmbito do combate a esta patologia, o CCR é o terceiro cancro mais frequente e a quarta principal causa de morte por cancro a nível mundial. Uma das principais dificuldades registadas no combate a esta doença deve-se ao facto dos métodos de rastreio atualmente empregues na prática clínica revelarem-se pouco sensíveis/específicos ou altamente invasivos. Assim sendo, torna-se evidente a necessidade de uma ferramenta de rastreio que não seja invasiva mas que permita simultaneamente identificar a doença num estadio inicial com elevada especificidade/sensibilidade. Adicionalmente, após o diagnóstico é fundamental determinar como o paciente irá progredir e ultimamente definir o seu prognóstico. Com o intuito de contornar estes obstáculos, têm aumentado os estudos que focam a descoberta de biomarcadores que permitam identificar precocemente esta doença e/ou "prever" o prognóstico dos pacientes.

pacientes.

Biomarcadores são componentes celulares ou alterações moleculares que refletem modificações a nível celular/tecidual sugestivas de um estado patológico. Ao longo dos últimos anos, diversos biomarcadores de CCR têm sido descortinados com o intuito de identificar e tratar esta doença. Os biomarcadores atualmente em uso têm a vantagem de poderem ser identificados na corrente sanguínea ou nas fezes, consistindo por isso num processo não invasivo (ou menos invasivo) e ultrapassando assim um dos principais obstáculos das técnicas standard utilizadas hoje em dia. Alguns biomarcadores atualmente utilizados na prática clínica são o *Carcinoembryonic antigen* (CEA), *Carbohydrate antigen 19-9* (CA19-9), *Tissue polypeptide specific antigen* (TPS), *Tumor-associated glycoprotein 72* (TAG72) e o *Tissue inhibitor of metalloproteinases-1* (TIMp-1)). Contudo, a maioria dos biomarcadores utilizados atualmente não são exclusivos para o CCR e pecam também pela sua baixa sensibilidade e/ou

especificidade. Assim, é urgente novos estudos que visem a descoberta e identificação de biomarcadores mais precisos e capazes de dar resposta às necessidades clínicas atuais.

Nesse sentido, mais recentemente verificou-se um aumento do número de estudos que pretendem identificar marcadores moleculares alternativos, como biomarcadores epigenéticos. Alterações epigenéticas são modificações que regulam a expressão genética sem alterar a sequência do DNA. Três grandes eventos epigenéticos são conhecidos atualmente: modificação de histonas, metilação do DNA e regulação por RNAs não codificantes, entre os quais se encontram os microRNAs (miRNAs), piwi-interacting RNAs (piRNAs), entre outros. Contudo, no CCR, as alterações epigenéticas mais extensivamente descritas englobam a metilação do DNA e os RNAs não codificantes, mais propriamente os miRNAs. A metilação do DNA consiste na adição de um grupo metilo (CH3) no carbono 5 de um nucleótido de citosina, que consequentemente tem impacto na regulação génica. Relativamente aos miRNAs, estes são pequenos RNAs não codificantes que regulam a expressão genética ao nível pós-transcricional por ligação ao 3'UTR dos seus genes alvo.

Ao longo dos últimos anos o número de miRNAs identificados que se encontram alterados no CCR tem crescido de forma exponencial, sugerindo um forte envolvimento dos mesmos nesta patologia. Para além disso, os miRNAs são extremamente estáveis quando excretados pelas células que os expressam, permitindo a sua deteção de forma menos invasiva, nomeadamente, em fluídos biológicos como sangue ou soro, ou até mesmo em fezes. Esta característica torna assim os miRNAs biomarcadores ideais para a identificação desta doença. Complementarmente, alterações na expressão dos miRNAs têm sido demonstradas estar não só envolvidas na génese mas também na progressão do CCR. MiRNAs encontrados desregulados em CCR compreendem os miRNAs: miR-21and miR-29b, miR-20a, miR-92a, miR-203, miR-145, miR-17-3p, entre outros.

Desta forma, a caracterização e a compreensão dos padrões de expressão dos miRNAs e mecanismos subjacentes ao desenvolvimento do CCR parece essencial para entender o papel dos mesmo no contexto patológico. Seguindo esta linha de pensamento, neste trabalho propomo-nos então a caracterizar os padrões de expressão e de metilação dos miRNAs na iniciação e progressão do CCR. O nosso principal objetivo é desvendar se estas alterações poderiam servir como uma ferramenta não só de diagnóstico mas também de prognóstico de pacientes com CCR. Para além disso, procurámos caracterizar os genes/vias de sinalização regulados pelos miRNAs

encontrados diferencialmente expressos a fim de perceber o seu modo de ação no contexto do CCR.

Os nossos resultados demonstram que tanto alterações na expressão como na metilação dos miRNAs ocorrem numa fase muito precoce do desenvolvimento do CCR, nomeadamente durante a transição de tecido normal para o estadio I. Para além disso, ambos os valores de expressão e de metilação são mantidos constantes ao longo da progressão da doença. Adicionalmente, os nossos resultados evidenciam explicitamente que miRNAs encontrados diferencialmente expressos ao longo da progressão de CRC se encontram maioritariamente sub-expressos.

No entanto, as nossas análises sugerem que tanto os miRNAs sub-expressos como os sobre expressos interagem com genes frequentemente alterados durante a progressão de CRC tais como *p53*, *APC*, *WNT3A* e *KRAS*. Estas interações podem então sugerir um possível envolvimento destes miRNAs no controlo da expressão destes genes durante o processo de carcinogénese. Adicionalmente, verificámos que a vasta maioria das vias de sinalização reguladas tanto pelos miRNAs sobre expressos como pelos miRNAs sub-expressos são equivalentes.

Relativamente ao potencial clinico, os nossos resultados demonstram que tanto alterações na expressão como na metilação dos miRNAs têm bons valores de diagnóstico num estadio precoce da doença (estadio I). De facto, certas CpGs e diversos miRNAs conseguiram distinguir pacientes normais de tumorais em estadio I com sensibilidades e especificidades de 100 %.

Por fim as nossas análises demonstram que quer as alterações de expressão quer as alterações de metilação de miRNAs são possíveis biomarcadores de prognóstico. De facto alguns dos painéis desenvolvidos neste trabalho conseguiram distinguir de forma bastante inequívoca pacientes com melhor prognostico daqueles com pior pronósticos tanto a nível de sobrevivência como de recorrência de doença.

Embora sejam necessários mais estudos, este trabalho evidencia claramente que os padrões de expressão e metilação dos miRNAs no CCR podem constituir importantes ferramentas no âmbito da clinica num futuro próximo.

Palavras-chave: Cancro colorectal, alterções epigenéticas, metilação do DNA, microRNAs, biomarcadores, diagnóstico, prognóstico.

Index of Contents

Index of figures

Index of tables

Index of Annexes

Abbreviations

A - Adenine

AD - Anderson-Darling

Ago2 -Argonaute 2

AJCC - American Joint Committee on Cancer

APC - Aadenomatous polyposis coli

AUC – Area under the curve

BP- Base pairs

C - Cytosine

CA19-9 - Carbohydrate antigen 19-9

CEA - Carcinoembryonic antigen

CRC – Colorectal cancer

DAVID - Database for Annotation, Visualization and Integrated Discovery

DNA - Deoxyribonucleic Acid

DNMTs - DNA methyltransferases

EMT - Epithelial-to-mesenchymal transition

FAP - Familial adenomatous polyposis

FOBT – Fecal Occult Blood Test

FPR – False positive rate

G - Guanine

GA - Genome analyzer

GI – Gastrointestinal

H – Histone

HATs - Histone acetyltransferases

HDACs - Histone deacetylases

HDMs -Histone demethylases

HMTases- Histone methyltransferases

HNPCC - Hereditary non-polyposis colon cancer

HR –Hazard ratio

IARC - International Agency for Research on Cancer

KEGG - Kyoto Encyclopedia of Genes and Genomes

KM – Kaplan-Meier

KRAS - Kirsten rat sarcoma viral oncogene homolog

KS - Kolmogorov-Smirnov

LF - Lilliefors

LncRNA - long non-coding RNAS

MiRNAs - MicroRNAs

MTIs - microRNA-target interactions

NCI - National Cancer Institute

NcRNAs - Non-coding RNAs

NGS - Next-generation sequencing

NHGRI - National Human Genome Research Institute

OS - Overall survival

PiRNAs - Piwi-interacting RNAs

Pre-miRNAs - precursor miRNAs

Pri-miRNAs - Primary miRNAs

PTEN - Phosphatase and tensin homolog

PTM - Post-translational Modifications

Q - Quartile

RISC - RNA induced silencing complex

RLC - RISC loading complex

RFS - Recurrence free survival

RNA - Ribonucleic acid

RNAseq – RNA sequencing

ROC - Receiver Operating Characteristic

RUNX3 - Runt-related transcription factor 3

SEPT9 – Septin 9

SBS – Sequencing by synthesis

SnoRNAs - Small ncRNAs

SnoRNAs - Small non-coding RNAs

SW - Shapiro-Wilk

T - Timin

TAG72 - Tumor-associated glycoprotein 72

TCGA - The Cancer Genome Atlas

TGFβ - Transforming growth factor beta

TGFBR - Transforming growth factor beta receptor

TIMp-1 - Tissue inhibitor of metalloproteinases-1

TNM - Tumour-node-metastasis

TPR – True Positive rate

TPS - Tissue polypeptide specific antigen

TRBP - RNA (tar)-binding protein

TSG - Tumor suppressor gene

TSP1 - Thrombospondin 1

TSS – Transcription start site

U - Uracil

UCSC - University of California, Santa Cruz cancer

UTR - Untranslated region

VIM –Vimentin

WGA - whole-genome amplification

Wnt - Wingless/Integrated

XPO5 - Exportin-5

CHAPTER 1 - INTRODUCTION:

1.1 Cancer

Cancer is a class of complex and heterogeneous diseases that share common features and are characterized by an abnormal and uncontrolled proliferation of cells that ultimately invade surrounding tissues or spread to distant organs (metastasize)[1,2]. In 2000, Douglas Hanahan and Robert Weinberg proposed the existence of six fundamental properties transversal to all cancers essential for the carcinogenic process[2]. These properties designated "hallmarks of cancer" included: sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis and activating invasion and metastasis (Figure 1.1)[2]. Eleven years later Hanahan and Weinberg revisited the hallmarks of cancer in an updated publication named "Hallmarks of cancer: the next generation" and four new hallmarks were proposed: deregulating cellular energetic, avoiding immune destruction, tumour-promoting inflammation and genome instability and mutations[3]. These publications revolutionized the comprehension of cancer and potentiated new perspectives in the approach to this condition.

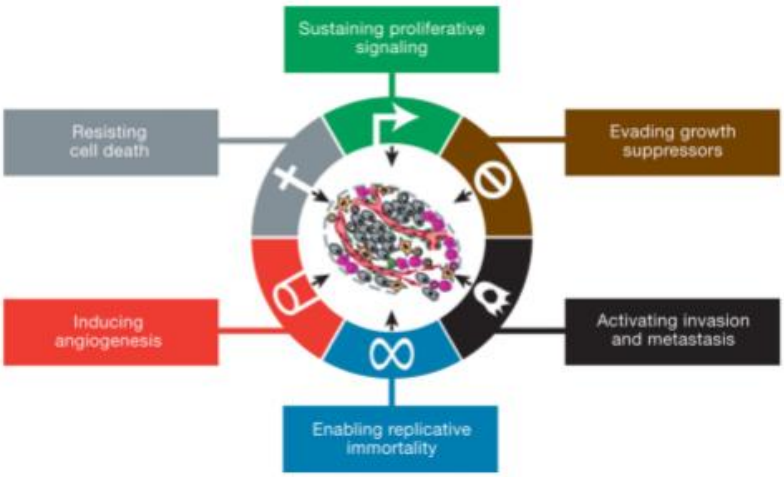

Figure 1. 1 Hallmarks of cancer. (A) Illustration of the six hallmarks of cancer first established by Douglas Hanahan and Robert Weinberg in 2000 in the publication "Hallmarks of Cancer". From *Hanahan and Weinberg* 2000.

Cancer develops when normal cells lose their ability to control proliferation due to the accumulation of genetic and epigenetic changes[4]. These alterations disrupt the normal function of cells leading to abnormal or damaged cells which start to grow out of control resulting in the formation of masses of tissue called tumor[1]. As tumors develop they can eventually progress into a cancerous state, however not all tumors do so. The last types of tumors are known as benign tumors, as they do not spread or invade nearby tissues and thus are considered non-cancerous[1,5]. On the other hand malignant tumors or cancerous tumors spread into nearby tissues, and, as they grow, can travel to distant places in the body through the blood or the lymphatic system and form new tumors far from the original tumor site[5].

There are numerous types of cancers reported in humans, however all begin with defective cells that are the result of either somatically acquired alterations due to environmental exposures and errors during DNA, or inherited (hereditary) alterations[6,7]. Among the different types of cancers, one of the most prominent is colorectal cancer (CRC). CRC is the third most commonly diagnosed cancer worldwide and the fourth most common cause of cancer related death globally[8]. Every year, over 1.23 million individuals are diagnosed with CRC and about 600 000 succumb to this disease. Incidence is approximately 30% higher in men than women, with approximately 814,000 cases reported in men and 664,000 cases in women every year. This makes CRC the third most commonly diagnosed cancer in man and the second in women[8]. Regardless, in both sexes CRC incidence is generally low for patients younger than 50 years old, however it strongly increases with age[9].

In Portugal CRC rises as the most frequent diagnosed cancer and the main cause of cancer related deaths. According to the last data estimated by the International Agency for Research on Cancer (IARC) in 2012, CRC was predicted to account for 14.5 % of all incident cancers in Portugal and responsible for 15.7% of all deaths by cancer (Figure 1.2).

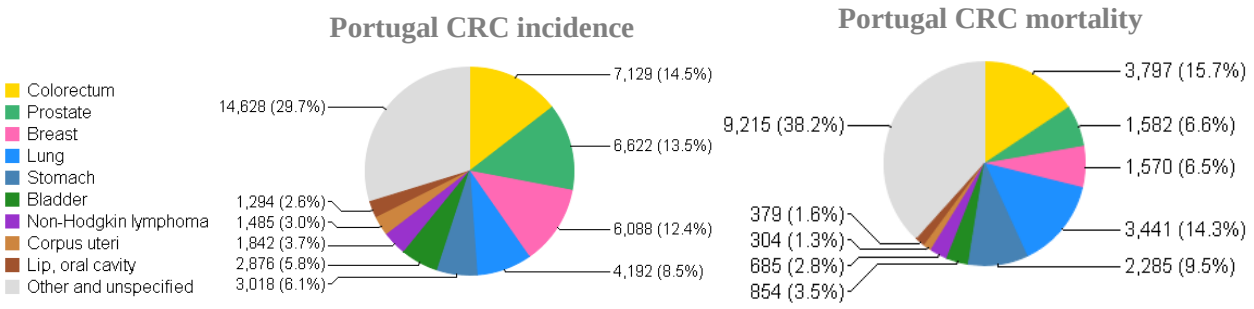

Figure 1. 2 Colorectal cancer estimated incidence and mortality rates for the year 2012 in Portugal. Pie charts illustrating the different cancer estimated incidence (left) and mortality (right) rates in the Portuguese population for the year 2012. From GLOBOCAN project data, available at International Agency for Research on Cancer (IARC) (http://globocan.iarc.fr/Pages/fact_sheets_population.aspx).

1.2 Colorectal cancer (CRC)

Colorectal cancer (CRC) is a malignant tumor that usually begins as an abnormal tissue growth in the cells of the colon or rectum which combined constitute a segment of the large intestine, the terminal portion of the gastrointestinal (GI) system[9]. These small growths designated as polyps, are projections of tissue that emerge from the innermost layer of the colon (the mucosa) into the lumen (hollow center) of the colon (Figure 1.3)[9,10] . Polyps are as non-cancerous (benign) growths, however as they slowly grow over time, normally over a period of 10 to 20 years, they can eventually become cancerous (malignant)[11,12]. Whether a polyp changes into a cancerous state or not depends on the type of polyp it is[11]. Hyperplastic and inflammatory polyps are the most common growths yet in general they are not pre-cancerous. On the other hand adenomatous polyps also known as adenomas, although less common, have the ability to transit into a malignant state, and thus are often called as pre-cancerous conditions[11]. Adenomas arise from glandular cells which produce mucus that lubricate the inside of the colon and rectum[9]. Although all adenomas have the potential to become cancerous, fewer than 10% are estimated to progress into a cancerous state[13]. As an adenoma becomes larger, the likelihood of becoming cancerous increases, and when that transition happens it is referred to as adenocarcinoma. Among the five subtypes of CRC (adenocarcinomas, carcinoid tumors, gastrointestinal stromal tumors, lymphomas and sarcomas), adenocarcinomas are the most predominant making up about 95% of all CRCs[14].

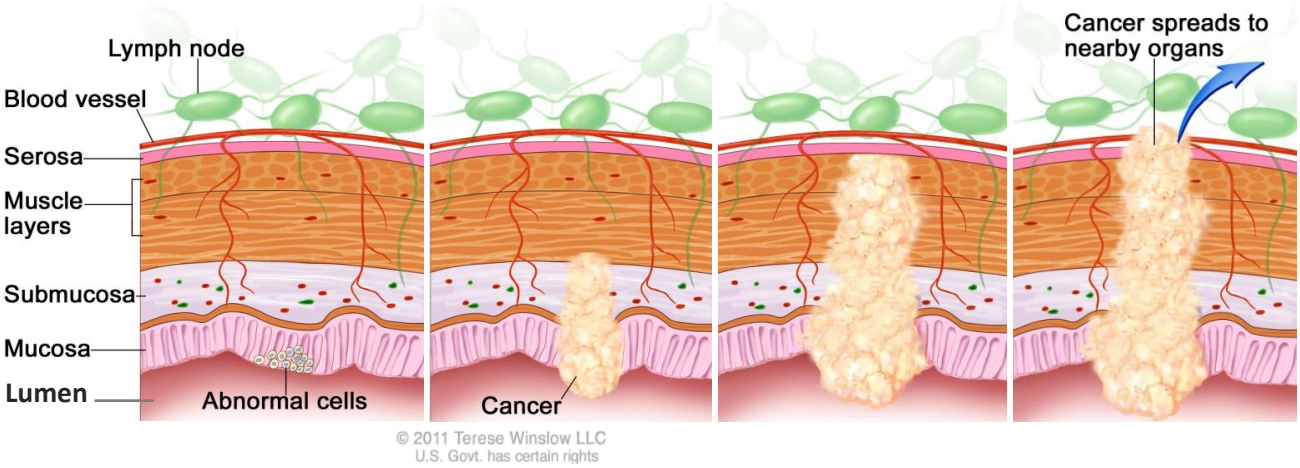

Figure 1. 3 Histological scheme of polyp formation within the large intestine walls. The four main layers that compose the large intestine from the outer portion to the inner portion (lumen): Serosa, Muscle layer, Submucosa and Mucosa. Polyps emerge in the mucosa layer and protrude to the lumen. However, as polyps grow they can invade the surrounding layers, spread into the blood vessels and metastasize. Adapted from Terese Winslow LCC, Medical And Scientific Illustration.

CRC is considered to be an environmental disease, were cultural, social, and lifestyle factors heavily influence the risk of disease[15,16]. In fact sporadically occurring CRC cases which results from complex interactions between gene susceptibility and environmental factors account for the vast majority (75% - 80%) of all CRCs in the population[17]. Nevertheless, like many other cancers CRC also exhibits a heritable component, as familial adenomatous polyposis (FAP) and hereditary non-polyposis colon cancer (HNPCC or Lynch syndrome) account for close to 20% of all colorectal cancer incidence[16,17].

1.2.1 Molecular Pathogenesis of Colorectal Cancer

CRC is a multistage process that results from the progression of a sequential accumulation of genetic mutations[18]. According to Fearon and Vogelstein, genetic events during the adenoma to carcinoma sequence lead to the development of CRC through specific genetic changes in oncogenes and tumor suppressor genes[19]. The accumulation of mutations in key tumor suppressor genes or oncogenes deregulates the cellular homeostatic functions affecting a wide range of cellular functions from proliferation, migration, differentiation, adhesion, cell death, to DNA stability and repair, causing the transformation of normal cells into cancer cells

(Figure 1.4)[18,20]. Mutations typically alter the gene product by changing the amino acid sequence of proteins which leads to truncated or dysfunctional proteins or by altering the quantity of protein produced. Common mutation in the context of CRC include inactivation of the tumor suppressor gene Adenomatous polyposis coli (APC) which leads to activation of the Wingless/Integrated (*Wnt*) pathway, a common mechanism (which occurs in approximately 70% of adenomas) for initiating the adenoma to carcinoma sequence[20,21]. Subsequently mutations in genes such as *Kirsten rat sarcoma viral oncogene homolog* (*KRAS)* or *TP53* (*p53*) also contribute to the progression of adenomas into carcinomas. CRC development can also involve mutations in the Transforming growth factor beta (*TGFβ*) signaling pathway as mutations in type II TGFβ receptor (*TGFBR2*) gene occur in approximately 30% of CRCs. Additionally, mutations affecting other TGF signaling pathway members, including *SMAD2*, *SMAD4*, Runt-related transcription factor 3 (*RUNX3*) and Thrombospondin 1 (*TSP1*) have been reported in colorectal cancers. Furthermore Serine/threonine-protein kinase B-raf (*BRAF)*, phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha (*PIK3CA)*, Phosphatase and tensin homolog (*PTEN*) and *β-catenin* alterations have been associated with CRC carcinogenesis[20].

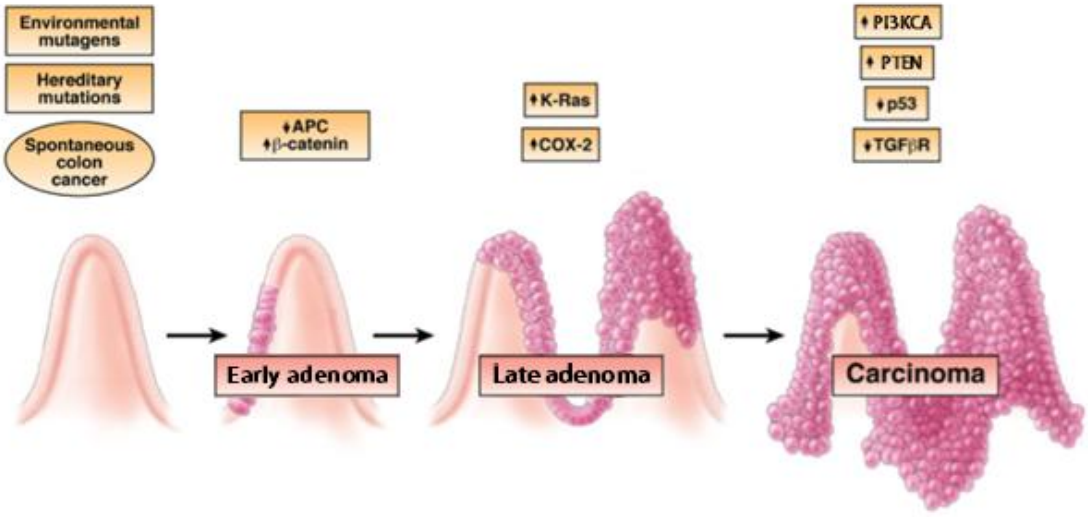

Figure 1. 4 Morphological and molecular changes implicated in colorectal cancer development. Accumulation of specific genetic mutations in oncogenes and tumor suppressor genes contribute to adenoma-carcinoma progression. Mutations in *APC*, *β-catenin* and other players of this pathway are postulated to occur early in the carcinogenic process mediating the transition of single neoplasm cells to early adenoma. Meanwhile mutations involving *KRAS*, *PTEN* and *p53* tend to occur later on leading to the emergence of carcinoma. Adapted from *Terzić, et al*., 2010.

Since Fearon and Vogelstein formulated the multi-step events of the molecular pathway of CRC formation involving oncogenes and tumour suppressor genes, there have been considerable advances in the understanding of colorectal carcinogenesis. Recent studies have shown that epigenetic alterations are an alternative mechanism in carcinogenesis[22]. Most of CRCs have epigenetic abnormalities such as over-expression of miR-21 , and hypermethylation of hsa-miR-129 and hsa-miR-137 that coexist with classical genetic changes such as *APC*, *p53*, *KRAS* and *β-catenin* mutations[22]

1.2.2 Histopathological classification of colorectal cancer

Colorectal cancers are classified according to a tumour-node-metastasis (TNM) staging system established by the American Joint Committee on Cancer (AJCC)[23,24]. This system describes the anatomical extent of the disease based on the assessment of three features: local invasion depth (T), lymph node involvement (N), and presence of distant metastases (M) (Table I)[25].

Table I Colorectal cancer classification according to local invasion depth (T), lymph node involvement (N), and presence of distant metastases (M). From *Brenner*, 2014.

	Definition
T stage	
Tx	No information about local tumour infiltration available
Tis	Tumour restricted to mucosa, no infiltration of lamina muscularis mucosae
T1	Infiltration through lamina muscularis mucosae into submucosa, no infiltration of lamina muscularis propria
T2	Infiltration into, but not beyond, lamina muscularis propria
T3	Infiltration into subserosa or non-peritonealised pericolic or perirectal tissue, or both; no infiltration of serosa or neighbouring organs
T4a	Infiltration of the serosa
T4b	Infiltration of neighbouring tissues or organs
N stage	
Nx	No information about lymph node involvement available
N0	No lymph node involvement
N1a	Cancer cells detectable in 1 regional lymph node
N1b	Cancer cells detectable in 2–3 regional lymph nodes
N1c	Tumour satellites in subserosa or pericolicor perirectal fat tissue, regional lymph nodes not involved
N2a	Cancer cells detectable in 4–6 regional lymph nodes
N2b	Cancer cells detectable in 7 or greater regional lymph nodes
M stage	
Mx	No information about distant metastases available
M0	No distant metastases detectable
M1a	Metastasis to 1 distant organ or distant lymph nodes
M1b	Metastasis to more than 1 distant organ or set of distant lymph nodes or peritoneal metastasis

Table II Union International Against Cancer stage classification of colorectal cancer. From *Brenner*, 2014.

	T	N	M
Stage 0	Tis	N0	M0
Stage I	T1/T2	N0	M0
Stage II	T3/T4	N0	M0
IIA	T3	N0	M0
IIB	T4a	N0	M0
IIC	T4b	N0	M0
Stage III	Any	N+	M0
IIIA	T1–T2	N1	M0
	T1	N2a	M0
IIIB	T3–T4a	N1	M0
	T2–T3	N2a	M0
	T1–T2	N2b	M0
IIIC	T4a	N2a	M0
	T3–T4a	N2b	M0
	T4b	N1–N2	M0
Stage IV	Any	Any	M+
IVA	Any	Any	M1a
IVB	Any	Any	M1b

Combined, these features produce an overall stage classification, the Union Internationale Contre le Cancer stage classification (UICC) that ranges from stage 0 to stage IV (Table II) and provides the basis for therapeutic decisions[25].

Although stage classification according to TNM and UICC provides valuable prognostic information and guides clinicians therapeutic decisions, the response and outcome of individual patients to therapy or treatment is not predicted[25].

1.2.3 Screening, diagnosis and prognosis

Patients with CRC are often asymptomatic in the early stages of the disease[13]. When symptoms appear, CRC has already grown or spread. Due to this particularity, early stage CRC detection is not always feasible. CRC patient survival is significantly affected by the stage of disease at diagnosis, as patients diagnosed in early stage have 5-year survival rates of 90%, which decrease to less than 10% when diagnosed at later stages (when distant metastasis develop)[15]. In fact if polyps are detected in an early stage they can be removed and cancer may be prevented[26]. This implies that an early detection of this disease is imperial for the success of the treatment, thus reducing both morbidity and mortality[17].

Early detection of CRC is manageable through screening techniques which involves the detection and removal of pre-cancerous growths or early stage cancer, before the manifestation of any symptoms (in patients who have no symptoms)[27].The idea is to detect the disease in a curable state, before it has a chance to grow or spread which makes treatment easier to manage, less expensive, and more likely to be successful[9,27]. Currently, the most predominantly used screening modalities for CRC, are Fecal Occult Blood tests (FOBTs) and colonoscopy[28].

FOBTs are non-invasive screenings methods that can detect microscopic amounts of blood in feces indicating bleeding from the gastrointestinal tract[29]. FOBTs are the commonly used methods for CRC screening worldwide and the primary choice in most screening programs in Europe[17]. Given the non-invasive nature and low cost, it is one of the most accepted techniques by the population[17]. However, despite being widely used due to their simplicity, low cost and non-invasive nature, FOBTs have a suboptimal diagnostic accuracy suffering from low sensitivities and low specificities[30].

Colonoscopy on the other hand is the most reliable method for CRC screening. It offers the opportunity to visualize the entire colonic mucosa and provides the ability to remove colon polyps and potentially prevent CRC[31]. For this reason it is considered to be the gold-standard screening test for CRC[32]. Nevertheless this technique is highly invasive, expensive and presents a high risk of complications such as perforation, bowel tears and bleeding when compared to other screening tests[26]. Moreover, the quality of the colonoscopy depends on the bowel preparation, which many patients find unpleasant[33]. All of this factors lead to a poor patient compliance[34].

There is undeniable evidence that individuals who do not comply with the current screening programs have higher risk of developing cancer[34]. Therefore simpler, more efficient and less invasive screening methods that would improve compliance and ultimately decrease the incidence and mortality of this disease are needed. Thus, a drive to identify new screening methods that can overcome the limitations of the current techniques has stimulated a considerable interest in researching for potential molecular markers (biomarkers)[35].

Moreover, specific histological characteristics and pathological staging of the tumors are necessary to accurately assess CRC patient prognosis[36]. However, individual to therapy and survival times of patients within the same stage of CRC are very heterogeneous, highlighting the necessity for a more precise system to assess patient prognosis[36].

1.2.4 Biomarkers in CRC

Currently, biomarkers play an important role in the detection and treatment of patients with CRC[37]. A biomarker is a cellular biochemical or molecular alteration found in body fluids or in tissues that serves as an indicator of a biological process, pathogenic process, or pharmacological responses to a therapeutic intervention[38]. In cancer, a biomarker can be either a molecule secreted by the tumor or a specific response of the body to the presence of the tumor such as an antibody[38]. Biomarkers can act as a precious tool for cancer detection, diagnosis, and patient prognosis and can influence treatment choice[39,40].

Although several molecular biomarkers (Carcinoembryonic antigen (CEA), Carbohydrate antigen 19-9 (CA19-9), Tissue polypeptide specific antigen (TPS), Tumor-associated glycoprotein 72 (TAG72), Tissue inhibitor of metalloproteinases-1 (TIMp-1)) are able to detect CRC and determinate the progression of the disease there is still a great way to go to implement these biomarkers as first line of screening, prevention and treatment of CRC patients as they usually lack sensitivity/specificity or are unable to detect cancer at early stages of disease[41].

Recently it has been demonstrated that epigenetic changes contribute to tumor progression, by influencing key transformation steps in CRC development (disrupting pivotal signaling pathways or affecting genes that regulate DNA repair and cellular proliferation)[42]. Epigenetic alterations appear to occur very early in the adenoma to carcinoma sequence, making them ideal screening biomarkers[43]. Recently, several studies have provided the identification of a variety of specific epigenetic alterations as potential clinical biomarkers for CRC patients[30,44].

1.3 Epigenetics

Epigenetics are environmentally-mediated, frequent, powerful and widespread changes that affect gene expression without modifying the DNA sequence[45]. The epigenetic regulation of gene expression occurs in normal tissues and plays an important role in embryonic development, imprinting, and tissue differentiation. However, when disturbed, epigenetic mechanisms may lead to the development of several pathologies such as cancer[46,47]. Contrarily to gene mutations, epigenetic alterations can be reverted. This feature provides an opportunity to correct these epigenetic changes, and possibly to help combat disease progression or development[48]. Epigenetic modifications include histone post-translational modifications (PTMs), DNA

methylation, and gene regulation through non-coding RNAs, especially post-transcriptional regulation by microRNAs (miRNAs)[45].

1.3.1 Histone Modifications

Post-translational histone modifications constitute an epigenetic mechanism that affects the compaction state of chromatin which influences the structure and folding (packaging) of the DNA, thereby affecting gene expression[42,49]. These modifications orchestrate the unraveling of the chromatin into a "open" or active chromatin state accessible for transcription (euchromatin), or into, a "closed" or inactive chromatin inaccessible for transcription (heterochromatin) modifying DNA accessibility leading to transcription regulation[30]. Histones are structures that package and order the DNA into structural units called nucleosomes. The nucleosomes are the fundamental units of the DNA and are composed of octamers of four core histones (H3, H4, H2A, H2B) around which 147 base pairs of DNA are wrapped[50](Figure 1.5). The core histones are characterized by the presence of N-terminal tails that are subjected to extensive post-translational modifications such as acetylation, methylation, phosphorylation, ubiquitination, sumoylation, citrullination and ADP-ribosylation[30]. Among the various histone modifications the most extensively characterized in CRC are histone acetylation and methylation[51].

Histone modifications, recently recognized as "histone code", have been proposed to play an important role in the establishment of gene silencing during tumorigenesis[52]. However much is still to uncover regarding the contribution of post-translational histone modifications to the CRC carcinogenic process[45].

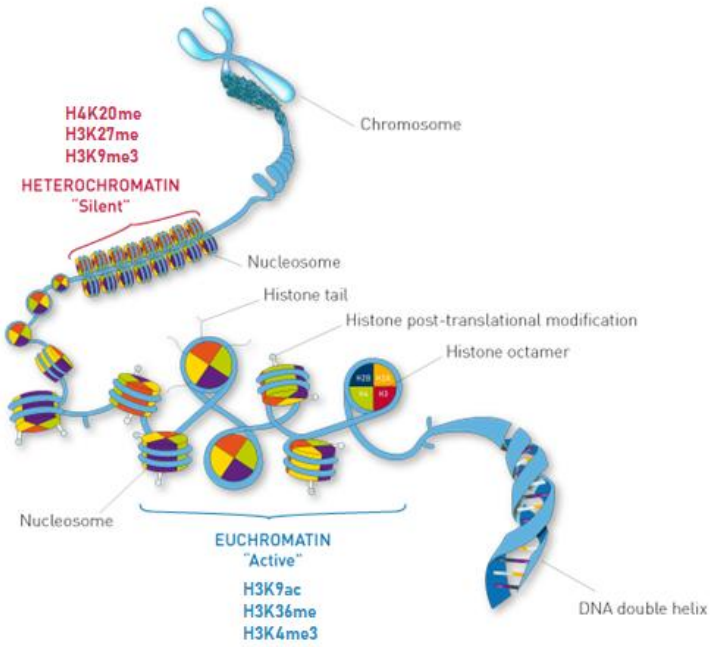

Figure 1. 5 Post-translational histone modifications regulate chromatin compaction. Histone`s N-terminal tails are susceptible to PTMs that can lead either to an active (euchromatin) or silent (heterochromatin) conformation. These modifications can directly mediate the DNA binding of proteins necessary for transcription and thus impact gene expression. Well characterized histone modifications and the respective effect on chromatin conformation are shown. Adapted from *http://www.biosense.it/prodotto/chip-seq-service/.*

1.3.2 DNA Methylation

DNA methylation occurs when a methyl group is covalently added to the 5-position of the pyrimidine ring of Cytosines (C) bases that are preceded by Guanine (G) nucleotides, typically designated CpG dinucleotides[20]. Approximately 80 % of all CpG sites are found dispersed genome-wide, while a small portion of CpGs can be found concentrated in specific regions called CpG islands[50]. CpG islands are genomic region of approximately 1–2 kilobases (kb) located in approximately 70% of all promoter regions of human genes, playing an important role in transcriptional regulation[51]. In healthy cells, the genome-wide scattered distributed CpG sites are heavily methylated, whereas the CpGs located in the promoter regions are unmethylated. However, CpG islands often become aberrantly methylated in cancer cells affecting not only the expression of protein coding genes but also the expression of various non-coding RNAs, some of which play a role in malignant transformation[20,53].

Generally, methylation of CpG islands in promoter regions is associated with transcriptional repression. In these regions, DNA methylation contributes to chromatin conformation changes influencing gene expression by affecting DNA exposure to transcription factors binding (Figure 1.6)[20]. In fact, epigenetic silencing of various tumor suppressor genes by hypermethylation of their promoters has been observed in a diversity of cancers, including CRC[22]. One of the best-characterized epigenetic events in tumor progression is the sporadic hypermethylation of the promoter of the mismatch repair gene *MLH1*, associated with approximately 12% of all CRCs cases[30]. Yet, recent evidences have demonstrated some exceptions to this classical view of DNA methylation repressor effect, as promoter DNA hypermethylation has been associated with increased gene expression[54]. Interestingly, global DNA hypomethylation may also play an important role in CRC development, possibly through genomic instability, however this process is far from being well understood[35].

The addition of a methyl group to C5 position of cytosine is catalyzed by a family of enzymes designated DNA methyltransferases (DNMTs)[50]. Eukaryotes have three different DNMTs, DNMT1 is considered the maintenance DNA methyltransferase being responsible for mimicking the methylation pattern of the unreplicated strand of DNA onto the newly generated DNA strand[55]. Conversely DNMT3A and DNMT3B are responsible for *de novo* methylation which refers to the methylation of DNA without the use of a DNA template that carries an existing methylation pattern[51].

DNA methylation plays a significant role in normal cells as it is involved in securing DNA stability through transcriptional silencing of genetic elements such as repetitive nucleotide sequences and endogenous transposons. Furthermore, DNA methylation contributes to gene imprinting, X-chromosome inactivation, homeostasis maintenance and genomic adaption in response to environmental stimuli besides other biological activities[46,55]. However, changes in DNA methylation were shown to promote progression of adenomatous precursor lesions into malignant tumors[56]. In fact thousands of genes are thought to be aberrantly methylated in the average colorectal cancer genome.

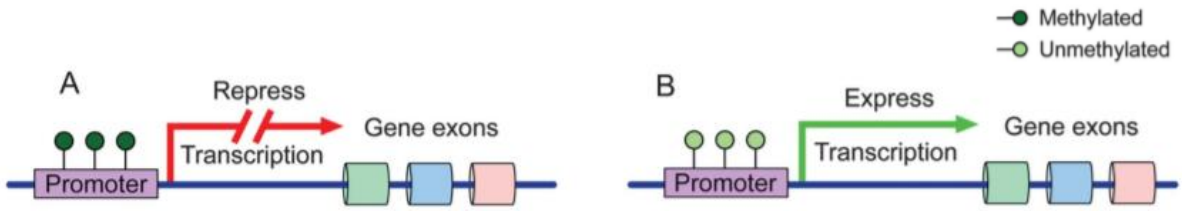

Figure 1. 6 Schematic representation of DNA methylation. Gene expression is heavily regulated by CpG islands methylation status. A) Methylated CpG islands in promoter regions usually lead to gene repression whereas B) unmethylated CpG islands are associated with gene expression. In healthy cells the methylation pattern is characterized by methylated CpGs spread throughout the genome and unmethylated CpG islands. However, during carcinogenesis CpG islands often become hypermethylated and the global genome hypomethylated. These methylation changes have profound consequences in gene expression regulation. From *Klein et al*., 2014.

1.3.3 Non-coding RNAs

Non-coding RNAs (ncRNAs) represent a class of ribonucleotide acids (RNAs) that do not code for proteins[57]. Although it was initially believed that ncRNAs had no biological function this idea was soon dismissed as it has they play a significant role in many biological and pathological processes, ranging from metabolic disorders to diseases of various organ systems such as cancer[58]. NcRNAs can be divided into small ncRNAs (snoRNAs) and long ncRNAs (lncRNAs) based upon their size[30]. Small ncRNAs comprise microRNAs (miRNAs), piwi-interacting RNAs (piRNAs), and small nucleolar RNAs (snoRNAs) and are usually shorter than 200 nucleotides. Conversely, long ncRNAs are longer and are often larger than 200 nucleotides[59].

The most widely studied class of ncRNAs are miRNAs, which play an important role in cancer initiation and progression. These small RNAs have thus revolutionized our understanding of cancer pathogenesis and also provided important insights into the feasibility of their use as clinically relevant biomarkers in cancer[30,58]

1.3.3.1 MicroRNAs

MicroRNAs (miRNAs) are endogenous small non-coding RNAs, of approximately 20 ~ 22 nucleotides, that mediate gene expression by binding to the 3' untranslated region (UTR) of their target mRNA[42]. This base pairing of miRNAs with their targets consequently drives either to translational repression or messenger RNAS (mRNA), cleavage and consequent decay[60]. As a

consequence the interaction of the miRNAs with their target mRNAs results in downregulation of gene expression of protein levels.

MiRNAs biogenesis starts with the transcription of primary miRNAs (pri-miRNAs) by RNA polymerase II in the nucleus which are then capped, spliced and polyadenylated (Figure 1.7)[61]. Pri-miRNAs are then folded into a base-paired stem-loop and further processed by the nuclear RNase III enzyme Drosha and co-factor DGCR8, which cleave these hairpin structures generating precursor miRNAs (pre-miRNAs)[62]. Pre-miRNAs are then translocated into the cytoplasm through the nuclear export factor, Exportin-5, where they are processed by the RNase III enzyme Dicer[63]. Dicer cleaves the pre-miRNAs at the terminal loop, liberating a ~22 nucleotide long RNA duplex[64,65]. This miRNA duplex contains two mature strands named 5p and 3p (the small RNA in the opposite side of the pre-miRNA stem loop), which are partially paired due to the 5' and 3' overhangs resulting from both Drosha and Dicer cleavages[65]. The 5p strand is present in the forward (5'-3') position while the 3p strand is located in the reverse position. The miRNA duplex is then assembled into the RNA induced silencing complex (RISC). This process starts with the binding of the duplex to the trans-activator RNA (tar)-binding protein (TRBP) which leads to the recruitment of Argonaute 2 (Ago2). Ago2 along with Dicer assembles the RISC[66,67]. Once the miRNA duplex is loaded into the RISC complex, the strand with the lowest thermodynamic stability remains bounded to this complex while the opposite strand (called passenger strand) is degraded[68]. Next, the remaining miRNA guides the RISC complex to the target mRNA[69]. MiRNAs bind to their mRNA targets 3'UTR by imperfect complementarily, mainly through the binding of nucleotides 2-8 in the 5' region of the miRNA named the "seed region". Although, the 3' region of the miRNA can also participate in the interaction miRNA:mRNA(Figure 1.7)[70,71]. Depending on the degree of complementarity different post-translational processes can occur, as partial complementarity is associated with translation repression, while near perfect complementarity can lead to mRNA degradation[72,73].

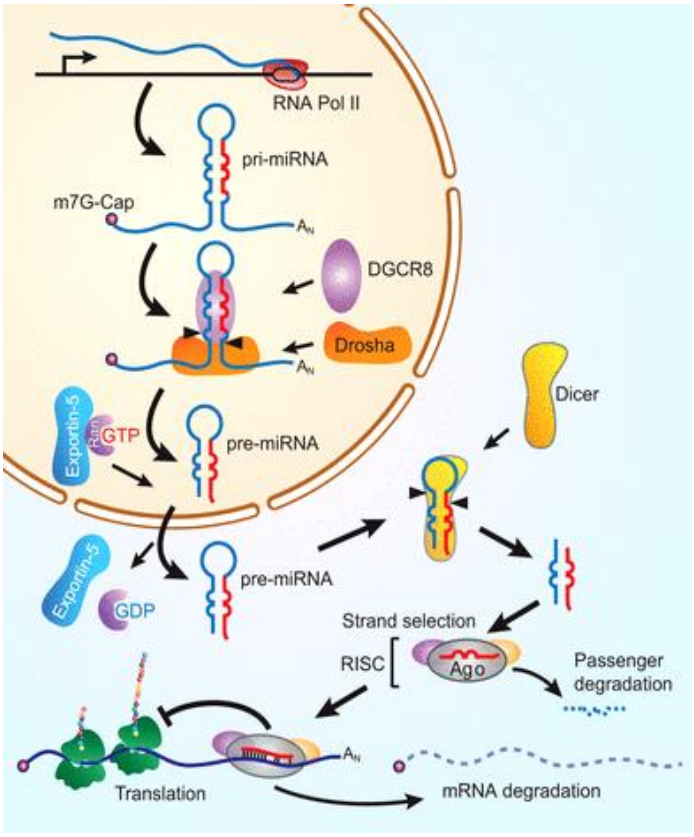

Figure 1. 7 MiRNA biogenesis. Transcription of miRNA genes by RNA PolII originates capped and polyadenylated transcripts designated primary miRNAs (pri-miRNAs). Pri-miRNAs then undergo cleavage by the microprocessor complex (consisting of the RNase III nuclease Drosha and RNA-binding protein DGCR8) to generate short hairpin-shaped structures designated pre-miRNAs. Pre-miRNAs are exported from the nucleus to the cytoplasm by exportin-5 and are further processed by Dicer, an RNase III nuclease, which generates 21-22 nucleotide double-stranded miRNAs. This duplex is assembled into RISC, with the assistance of TRBP and Ago2. During this assemblage, one of the mature strands called the passenger strand (in blue) is degraded while the remaining mature miRNA guides the RISC complex to the target mRNA. MiRNA binding to the 3'UTR of target mRNA triggers translation inhibition or mRNA degradation. From *Strubberg. et all*., 2017.

MiRNAs are transcribed from diverse regions scattered along the genome, however the vast majority of mammalian miRNAs are located within the intronic region of either protein-coding genes or non-coding transcripts[64,74]. Intronic miRNAs usually have the same orientation (are sense orientated) as their host gene and the expression of both miRNA and host gene largely coincides, which suggests a co-regulation and generation from common precursor transcripts[75]. However, miRNAs can also be found in intergenic regions (between genes) and a small subset of miRNAs has even been identified within exons of non-coding genes[76]. Moreover pri-miRNAs transcripts can comprise more than one miRNA. In fact, about 45% of known miRNAs are found in clusters and might be transcribed as a single polycistronic primary transcripts[74].Also, the same mature miRNA can have different genomic locations that will give rise to different pri- and pre-miRNAs, which upon processing generate the same mature miRNA (miRNA isoforms). All the biogenesis steps are thus critical do designate the mature miRNA being formed and expressed, impacting on the miRNA target genes[77,78].

Due to the limited complementarity between miRNAs and their mRNA targets, each miRNA can interact with several different mRNAs and a single mRNA can be suppressed by

various miRNAs[50]. Moreover, miRNAs from the same family are known to regulate the same target genes and/or different genes from the same signaling pathway to achieve their function[79,80].

MicroRNAs, therefore, control a diversity of cellular processes ranging from developmental transitions and organ morphology to cell proliferation and apoptosis[81]. Consequently, abnormal expression of miRNAs can affect the normal expression of numerous genes and ultimately deregulate several biological processes, resulting in development of certain diseases such as cancer[50]**.**

Since the discovery of miRNAs in patients with chronic lymphocytic leukemia in 2002, the role of miRNAs in regulating post-translational gene expression in cancer development, growth, and metastasis have been well-established in a diversity of publications[82,83]. Almost all cancer types exhibit their unique profile of upregulated and downregulated miRNAs[84]. This striking feature potentiates the use of miRNAs as useful cancer biomarkers. Moreover miRNAs are very stable outside cells which allow them to be safely extracted, stored, and studied in feces or various body fluids enabling a less invasive prognostic and diagnostic tool[85]. In CRC, aberrations of miRNA expression seem to play a significant role in tumor development and progression, and several miRNAs have been identified as potential biomarkers[30]. In fact several miRNAs have been described to function as tumor suppressors, oncogenes or both[86].

As mentioned above miRNAs are transcribed from genes. As such miRNAs gene promoters display all features of a normal gene commonly associated with Pol II-mediated transcription, including CpG islands[87,88]. Therefore abnormal CpG islands methylation in miRNAs promoter regions can affect their expression[89].

1.4 Epigenetic biomarkers

1.4.1 DNA methylation as potential biomarkers in CRC

Advances in understanding the molecular pathology of CRC, has led to the identification of promising early detection molecular markers with potential to be used in non-invasive CRC screening assays[90].Since DNA methylation appears to be an early event in tumorigenesis and particularly stable in blood and stool it has been proposed as a non-invasive diagnostic tool[30].

Moreover, aberrant DNA methylation also contributes to later stages of colon cancer formation and progression potentiating the ability to be a therapeutic or prognostic marker for CRC[20].

Recently, novel DNA methylation biomarker assays (Colovantage and ColoGuard) have gone through clinical trials and are commercially available[49]. Colovantage is a blood–based assay that detects the methylation of SEPT9 (*septin 9*) which is associated with impaired cytokinesis and loss of cell cycle control[91]. Colovantage has shown to provide an overall sensitivity of 90% and specificity of 88% and is being marketed in multiple countries, as a colon cancer screening assay[92,93]. ColoGuard is as stool based methylation assay for early detection of CRC. This test exploits the fact that the vimentin gene (*VIM*) is aberrantly methylated in the majority of colorectal cancers (53–84%) and has a reported a sensitivity and specificity of 83% and 82% in CRC[94]. *VIM* has been described to play a significant role in the epithelial-to-mesenchymal transition (EMT), disassembly of cell adherent junctions, reorganization of the actin cytoskeleton and acquisition of motility[95,96].

However, the majority of DNA methylation biomarkers are not clinically available. Nevertheless, blood-based methylation of Homeobox protein aristaless-like 4 *(ALX4)*, Nerve growth factor receptor *(NGFR*), Tomoregulin-2 precursor (*TMEFF2)*, Neurogenin 1 (*NEUROG1)* and *RUNX3* have been proposed as in CRC[30,93]. Furthermore hypermethylation of genes such as *APC*, Bone Morphogenetic Protein 3 *(BMP3)*, Ataxia telangiectasia mutated *(ATM)*, Secreted frizzled related protein 2 (*SFRP2)*, Cyclin-dependent kinase Inhibitor 2A *(CDKN2A/p16)*, *GATA4*, Glutathione s-transferase p1 (*GSTP1)*, Helicase like transcription factor (*HLTF)*, Human mutL homolog 1(*MLH1)*, O-6-methylguanine-DNA methyltransferase (*MGMT)*, N-myc downstream regulated gene 4 (*NDRG4)*, Ras association domain family member 2 *(RASSF2A)*, Tissue Factor Pathway Inhibitor 2 *(TFPI2)* and WNT Inhibitory factor 1 *(WIF1)* has been suggested as stool-based methylation biomarkers for early detection of CRC[30].

When considering the prognostic value of DNA methylation changes, hypermethylation of *CDKN2A/p16*, Checkpoint with FHA and RING finger domains (*CHFR)*, Enah/Vasp-like (*EVL)*, Insulin like growth factor binding protein 3 (*IGFBP3), KISS1*, *RET*, *HTLF*, and Hyperpigmentation Progressive 1 (*HPP1)* genes and hypomethylation of Inactivation escape 1 *(INE-1)*, MGMT, Transcription factor AP-2-alpha (*TFAP2A)* and Insulin-like growth factor 2 *(IGF2)* have been associated with poor prognosis of CRC[36]. Several other studies have also provided evidence that aberrantly methylated DNA has the potential to be used as prognostic

biomarkers in CRC. Nonetheless, further investigation is required to develop clinically robust assays in order to allow these biomarkers to be used in a clinical setting[30].

1.4.2 MiRNAs as potential biomarkers in CRC

The discovery of miRNAs in extracellular body fluids strongly increased the number of studies showing deregulated expression of circulating miRNAs in cancer diseases[30]. The first comprehensive miRNA expression profiling study was conducted by Ng and colleagues[97]. In this study the authors evaluated miRNA expression alterations in tissue and plasma samples from CRC patients and healthy subjects and demonstrated that a high expression of miR-92a and miR-17-3p, could discriminate CRC patients from healthy individuals[97]. Since this landmark study numerous miRNAs found in both feces and plasma/serum have been identified as potential biomarkers for an early CRC diagnosis. Basati et al. reported that two miRNAs: miR-194 and miR-29b were down-regulated in CRC patients when compared to control subjects, suggesting these miRNAs as powerful CRC serum biomarkers for early CRC[97]. Juan et al. also revealed miR-145 and miR-378 as potential biomarkers, which could help in early CRC diagnosing, with sensitivities reaching 100% and specificities of 60% and 98% respectively[97]. Expression of the miR-17-92a cluster and miR-135b in feces has also been found to discriminate patients with CRC from healthy subjects[98]. Additionally miR-21, one of the most promising biomarkers for early diagnosis of CRC since it is frequently deregulated in early stages of the adenoma-carcinoma sequence, has also been exhaustively reported as a potential serum or plasma diagnostic biomarker[30,99]. Despite the number of miRNAs identified as potential biomarkers in CRC the lack of consistency between biomarker panels in independent studies represents a major obstacle for the development of robust miRNA biomarkers[30].

MiRNAs also have been demonstrated to carry useful clinical information, as over-expression of miR-21 has been associated with metastasis and poor survival[100]. Furthermore over-expression of miR-372 and miR-15b has also been associated with metastasis and poor overall survival[36].

Finally, some miRNAs have been demonstrated to be methylation-sensitive, meaning that they can become epigenetically silenced due to CpG island promoter hypermethylation[101]. Bandres et al. identified a link between hypermethylation and downregulated expression of three

miRNAs, hsa-miR-9-1, hsa-miR-129, and hsa-miR-137 in primary CRC samples when compared to normal mucosa[89]. Moreover it was suggested that methylation of hsa-miR-9-1 could potentially be used as a poor CRC prognosis biomarker, and it could also be involved in metastatic events[89]. MiR-149 whose lower expression in CRC is a consequence of a neighbouring CpG island hypermethylation has also been referred as a promising prognostic marker in CRC. Lower expression of miR-149 was associated with lower 5-year survival rate and a higher tumor invasion in CRC[102]. In this sense miRNA methylation analysis can provide several insights on CRC carcinogenesis and work as potential biomarkers for this condition. Nevertheless, additional studies are still required in order to explore the potential of miRNA silencing by DNA methylation as a CRC biomarker for both diagnosis and prognosis[103].

1.5 Investigating miRNA expression and DNA methylation of miRNA genes

Since their discovery, miRNA research and understanding has grown exponentially[104]. For that to happen, new methods and technologies had to be developed and optimized in the last 20 years[105]. Nowadays we not only can "easily" study miRNA transcription, expression, methylation and function, as this information is usually available in open source databases. The big-data era has also arrived to the miRNA world, although all the information available is based on important research methods performed before, and enclosed behind the numbers we analyse[106]. Currently, high-throughput methods facilitate large-scale miRNA profiling and are the most widely used due to their accuracy, sensitivity and amount of data produced[107,108]. Those methods are used for miRNA expression profiling, as well as to investigate mechanisms behind miRNA patterns (e.g. miRNA interactions; DNA methylation; histone modifications), both essential to understand miRNAs role in pathological contexts[107–109].

In this work, we have used miRNA expression and DNA methylation data generated by next-generation sequencing (NGS, RNA-seq) and Infinium HumanMethylation450K array respectively, which information is publicly available. Thus, in this section, the basics of these technologies are briefly explained.

1.5.1 RNA-sequencing

The development of high-throughput next-generation sequencing (NGS) has led to the emergence of several tools such as RNA sequencing (RNA-seq)[108]. RNA-Seq is a sequencing approach that utilizes the capabilities of high-throughput sequencing methods to provide an understanding of the cellular transcriptome, which is defined as the complete set of transcripts (RNA molecules) in a cell that results from the transcription of a subset of genes into complementary RNA molecules[110].

Currently, the Illumina HiSeq platform is the most commonly applied next-generation sequencing technology for RNA-Seq[110]. Illumina Hiseq starts with library preparation in which RNA is converted to cDNA, which is then randomly shredded into small DNA fragments. Afterwards adaptors are ligated to both ends of the DNA-fragments, amplified by PCR and purified in a gel (Figure 1.8 A)[111]. The library is loaded into a flow cell and the fragments hybridize with the oligos on the flow cell surface which work as primers for amplification[112]. Each bound fragment is amplified into a cluster through bridge amplification (Figure 1.8 B). Using a method designated as sequencing by synthesis (SBS), fluorescently labeled reversible terminators deoxyribonucleotidetriphosphates (dNTPs) are incorporated into the DNA template strand in each cycle[109]. Each of the four dNTPs (A, C, T, and G) has a single different fluorescent label, which serves to identify the base. After each synthesis cycle, the clusters are excited by a laser which causes the last incorporated bases to emit fluorescence which is captured through a camera system (Figure 1.8 C)[109,113]. Subsequently, after imaging the reversible terminator fluorescent label nucleotides are removed and the template strands become ready for the next incorporation cycle[109]. Finally during data analysis, the newly identified sequence reads are aligned to the reference genome (Figure 1.8 D). After alignment, the mapped reads can be assembled into transcripts and the expression levels estimated[110].

Normalization of high-throughput sequencing of small RNA such as miRNAs is necessary in order to compare their levels across different samples[114]. Therefore data from RNA-seq experiments are typically normalized and perceived in reads per million genome-matching reads (RPMs)[115].

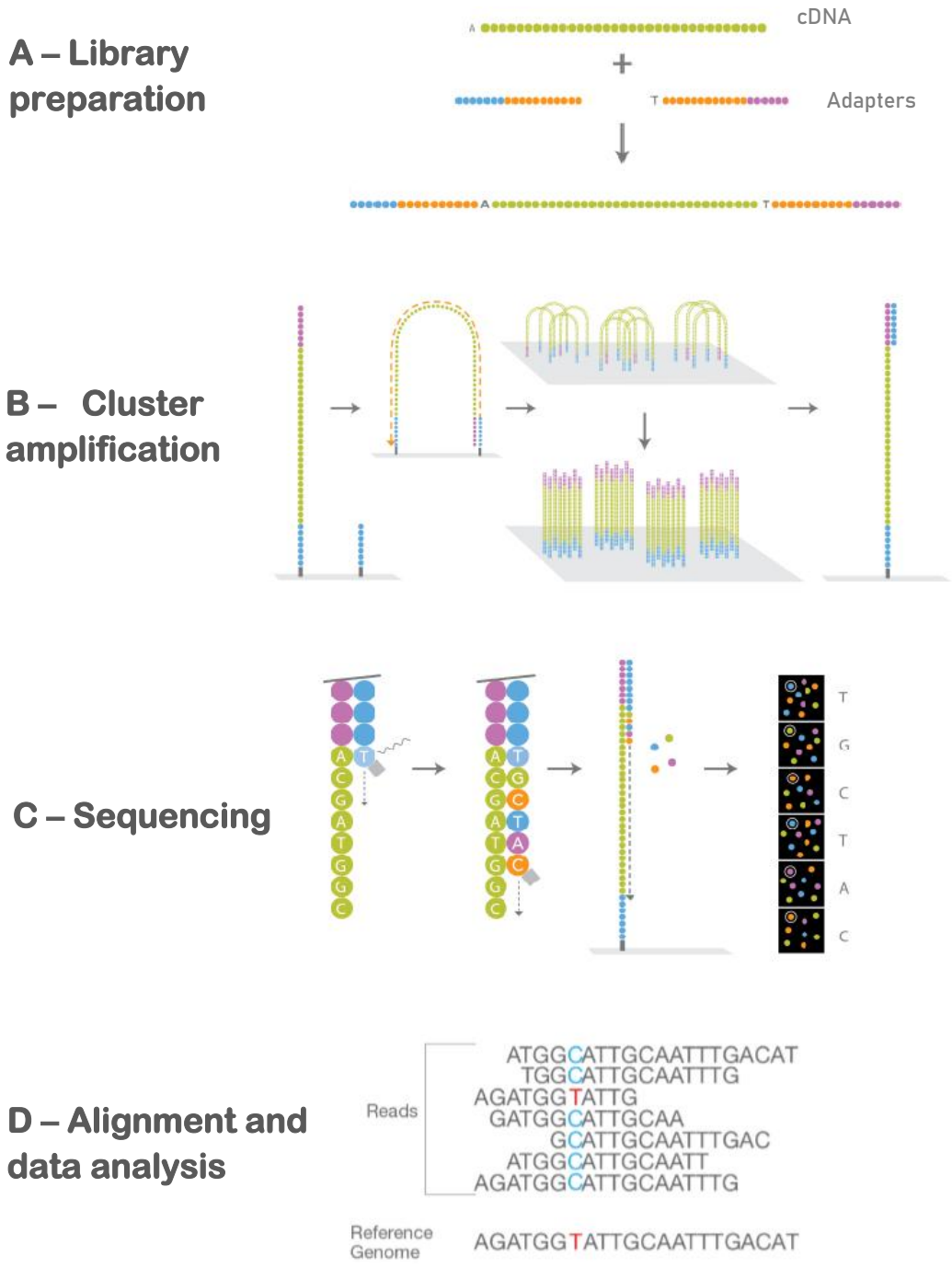

Figure 1. 8 Outline of the Illumina workflow. Representation of the main steps comprised in an Illumina workflow. Adapted from the Genome Analyzer brochure, http://www.solexa.com & An introduction to Next-Generation Sequencing Technology.

1.5.2 Illumina Infinium HumanMethylation450K array

The illumina Infinium HumanMethylation450K array is an assay that allows genome-wide DNA methylation analysis[116]. This tool makes it possible to assess the methylation status of more than 450000 CpGs located throughout the genome by providing quantitative methylation measurements at a single CpG site level[117].

Illumina Infinium HumanMethylation450K array makes use of two different types of chemical assays (Infinitium I and Infinitium II), which increase depth of coverage and allow quantitative genotyping of C/T polymorphisms generated by DNA bisulfite treatment[116,117]. For each CpG locus, Infinitum I uses 2 probes (one for methylated and other for unmethylated) placed on two bead types, which generate the same color channel. Infinitum II uses only one bead type for CpG locus but unmethylated and methylated signals are generated in different color channels[118].

Briefly, a small amount of genomic DNA is used for bisulfite treatment, which converts unmethylated cytosines to uracils, while keeping methylated cytosines unaltered[119]. During whole-genome amplification (WGA), uracils are converted into thymines generating C/T polymorphism at CpG locus. The products are then fragmented, purified and denature before being applied into the chip[119].

In the hybridization step, the Infinium I assay uses two types of beads per locus, one for the methylated allele (M) and one for the unmethylated allele (U) (Figure 1.9)[118]. Unmethylated or methylated CpG sites are thus recognized by bead-bound probes, which detect the presence of Ts or Cs by hybridization. Unmethylated and methylated signal of the same locus are generated in the same color channel[118]. Oppositely in the Infinitum II assay one bead type is utilized with two read out colors. If bound to Ts, the bead issues a certain color while if bound to Cs emanates another. In both scenarios, percentage of methylation of each CpG site corresponds to the ratio of the methylated signal over the cumulative methylated and unmethylated signals (M/M+U) and is expressed in β-values (beta-values)[118]. These DNA methylation values are continuous variables that range between 0 and 1 and higher beta values correspond to higher level of DNA methylation while lower beta values represent lower levels of DNA methylation[117].

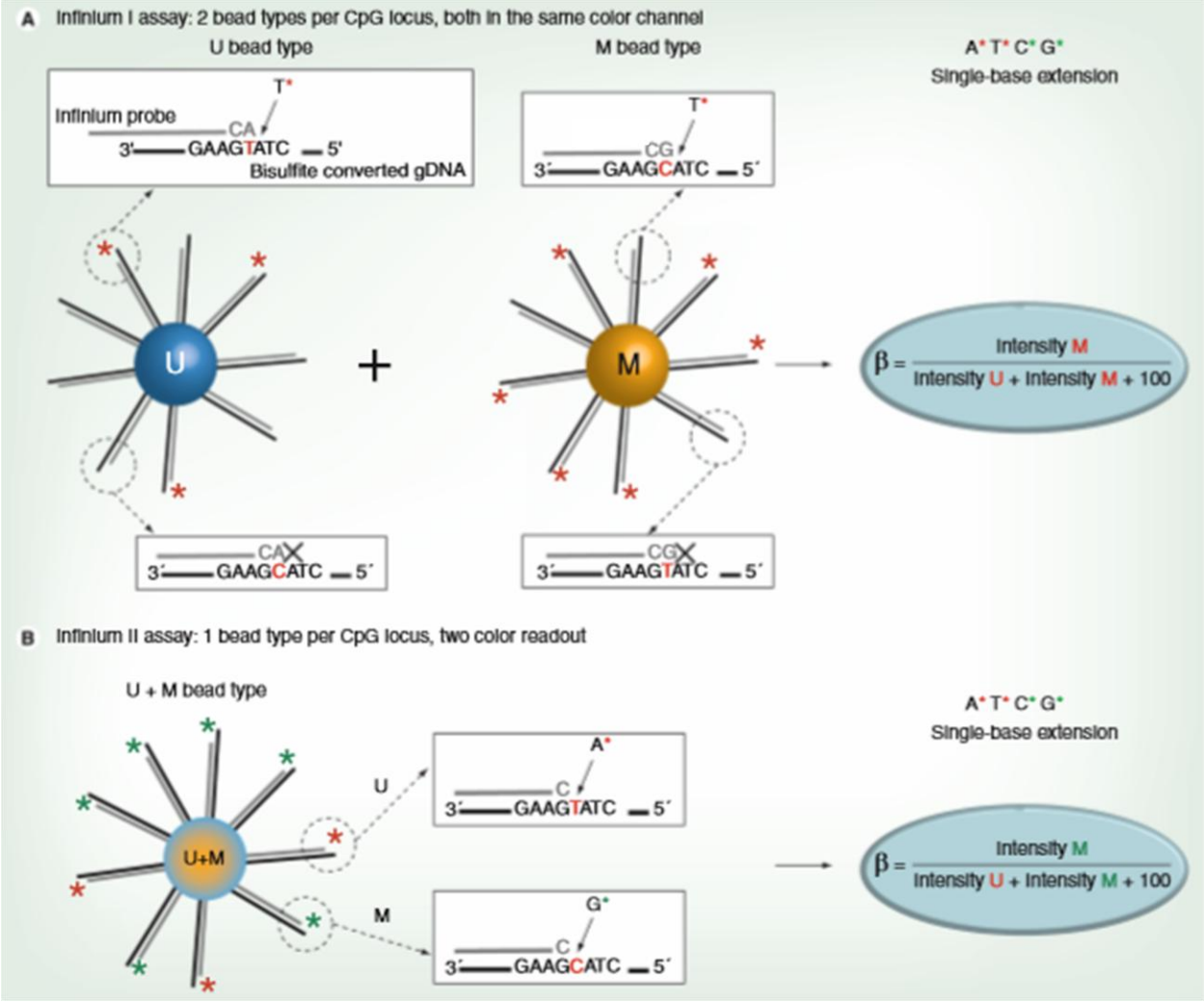

Figure 1. 9 The Infinium Assay for Methylation. The Illumina HumanMethylation450K array employs both Infinium I and II Infinium assays, enhancing its coverage depth. Infinium I assay utilizes two bead types per CpG locus, one for the methylated and one for the unmethylated state. Contrarily the Infinium II uses only one bead type, with the methylation state determined at the single base extension step after hybridization. From *Brouwer. et all.,2007*

CHAPTER 2 - AIMS:

Characterization and understanding of miRNA expression patterns and underlying mechanisms are essential to understand miRNAs role in pathologic contexts[120]. Nowadays this information is publicly available and when combined with clinical data of patients represents a powerful tool to give answer to current challenges. One of those challenges is to detect cancer at a stage that we can still effectively defeat it[121]. For that, we should not only develop effective diagnostic screenings based on accurate biomarkers, but also be able to discriminate which patients are at higher risk and require immediate treatment. Here, we have combined different approaches to meet these needs.

In order to achieve our main goal, we set up to:

1) Characterize miRNAs expression and methylation patterns in CRC initiation and progression;

2) Identify CRC stage specific miRNAs and differentially expressed miRNAs during disease progression. Investigate if they can be good predictors of disease outcome and characterize their target genes/pathways;

3) Study DNA methylation of miRNAs genes as epigenetic biomarkers in CRC;

CHAPTER 3 – Methodology:

3.1 Data collection

We analyzed mature miRNAs expression (RNAseq – IlluminaHiseq) and whole-genome DNA methylation (*Illumina Infinium HumanMethylation 450K array*) of Colon and Rectal cancer patients ("TCGA-COAD" and "TCGA-READ") publicly available at The Cancer Genome Atlas (TCGA) database (http://cancergenome.nih.gov/) through the use of University of California, Santa Cruz cancer (UCSC) Xena Public Data Hubs (https://xena.ucsc.edu/public-hubs/).

3.1.1 The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA) project, which began in 2006, is a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI)[122,123]. Created to discover and catalogue cancer-related alterations in order to improve diagnostic methods and therapeutic strategies of cancer, this public funded project has to this date generated the most comprehensive repository of human cancer molecular and clinical data with data on more than 11000 patients across 33 distinct cohorts of human cancers[122].

3.1.2 MiRNA expression and patient data collection

Level 3 miRNA expression data was obtained from the miRNA mature strand expression RNAseq – IlluminaHiseq dataset for both TCGA Colon Cancer (COAD; n=261) and Rectal Cancer (READ; n= 92) cohorts (https://tcga.xenahubs.net). All isoform expression for the same miRNA mature strand were added together, and $\log_2$ (total_RPM+1) transformation for miRNA expression values (normalization) was performed for 2164 miRNAs.

Clinical data for patients was then extracted from Xena Functional Genomics Explorer and added together to the Data obtained for mature miRNA expression, through sample intersection. Tissue samples obtained from both cohorts comprised a total of 340 primary tumor samples, 1 metastatic sample 1 recurrent tumor sample, and 11 normal samples. Further detailed patient information is provided in Anex I.

3.1.3 DNA Methylation and patient data collection

Level 3 DNA methylation data was collected from the DNA Methylation – Methylation 450k dataset for both TCGA CCOAD (n=334) and READ (n=105) cohorts. DNA methylation values were obtained for 485578 CpGs (probes) from a total of 463 samples. Both cohorts comprised a total of 414 primary tumor samples and 45 normal samples. It is important to mention however that some samples derived from to the same patients. In these cases (designated duplicated cases) the DNA methylation measurements for each probe were substituted by the median value. Further detailed patient information is in Anex II.

DNA methylation values and respective patient clinical data were extracted to R software using the TCGAbiolinks package.

3.1.4 TCGAbiolinks package

TCGAbiolinks is a Bioconductor package for R programming that allows users to query, download and perform integrative analyses of TCGA data. This package was developed in order to facilitate TCGA data retrieval as there are several major challenges for anyone interested in harnessing data from TCGA[123]. Among several tools developed to access TCGA data, TCGAbiolinks is currently the most versatile one[123].

3.2 TCGA data processing - Preparing data for analysis

All analyzes regarding data processing were performed using the software R studio. The several R packages used in order to develop this work will be mentioned when appropriated.

3.2.1 - Missing data treatment

In order to remove information that could cause noise in our results for differential expression and differential methylation analyzes we treated variables with missing data. Missing data treatment was performed using listwise case deletion. This technique consists on excluding those cases/variables with missing data and analyze only the remaining variables[124]. However, in opposition to listwise deletion where the absence of a single value in a variable results in its

exclusion, in our work we consider variable elimination only when more than half of the information was missing. This technique represented a more conservative approach to listwise case deletion, as the last would result in the omission of an extensive number of cases which could lead to having insufficient data to perform the analysis. Missing data treatment was the first step in our data processing analysis.

3.2.2 Sample selection and stratification

Following data collection, only primary tumor samples with patient TNM stage attribution and normal tissue samples and were analysed in our work. Concerning the miRNA expression dataset a total of 321 tumor samples (n=321, 49 stage I, 122 stage II, 106 stage III and 44 stage IV) and 11 normal tissue samples (n=11) were analysed (Table III).

Regarding the DNA methylation dataset, 373 tumor samples (n=373, 55 stage I, 144 stage II, 120 stage III and 54 stage IV) and normal tissue samples (n=45) were used (Table III).

After sample selection, for each cohort primary tumor samples were grouped according to the stage of disease of the respective patient provided by TCGA database. Thus samples were separated into five groups containing only stage I, stage II, stage III, stage IV patients samples and normal patients sample. This stratification process allowed us to analyse data for each stage of CRC independently, granting the opportunity to understand how both miRNA expression and DNA methylation data behaved throughout disease progression.

Table III Distribution of Colorectal Cancer patients samples by groups. After patient selection, 5 groups were made depending on CRC staging provided by TCGA database for both miRNA expression and DNA methylation cohorts.

	MiRNA expression dataset	**DNA methylation dataset**
Groups	**Samples** (n=321)	**Samples** (n=373)
Normal Solid Tissue	11	45
Primary Solid Tissue		
Stage I	49	55
Stage II	122	144
Stage III	106	120
Stage IV	51	54

3.2.3 Outlier detection and removal

Outliers refer to extreme values that abnormally lie outside the overall pattern of the variable distribution [125]. Among the several techniques to detect outlier values we selected the Box plot technique, which is often used to pinpoint possible outliers (Figure 3.1)[126,127]. In this plot, a box is drawn from the first quartile Q1 = $x_{[n/4]}$ (25^{th} percentile) to the third quartile Q3 = $x_{[3n/4]}$ (75^{th} percentile) of the data. The distance between Q1 and Q3 is denominated the interquartile range (IQR) and result from the difference between Q3 - Q1. Any points that lie outside Q1 − 1.5 IQR and Q3 +1.5 IQR, called upper and lower fences is considered an outlier[126]. In both miRNA expression and DNA methylation datasets, outliers were detected for each group of samples mentioned in section 3.2.1 were eliminated from our analyses [128].

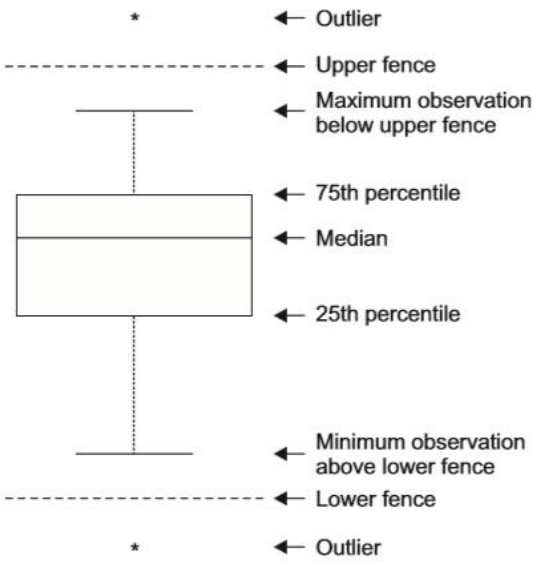

Figure 3. 1 Boxplot with outliers. The upper and lower fences represent values above the 75th and 25th percentiles (3rd and 1st quartiles), respectively, by 1.5 times the difference between the 3rd and 1st quartiles. Outliers are the value above or below the upper or lower fences. From *Kwak*, 2017

3.3 MiRNA Expression and DNA methylation Analysis

In order to determine if miRNAs and CpGs were differentially expressed and methylated respectively, a series of statistical inferences should be tested before, to establish if the differences between the groups being compared are large enough so that they are considered statistically significant. Statistical inferences rely on hypotheses testing, and our statement (H1)

is tested against the null hypothesis (H0) which can be either result in its acceptance or rejection. If there is significant evidence against H0 then it is rejected and H1 is accepted[129]. However claiming that there is significant evidence depends on the level of significance (α) established which corresponds to the probability of rejecting the null hypothesis when in fact it is true[129,130].

After selecting α, the value of the test statistic from the data (*p-value*) can be determined, and if is lower than (or equal to) α, there is enough significance to rejected the null hypothesis in favour of the alternative hypothesis (H_1)[129,131]. In this work, decision-making of the hypothesis tested, a level of significance of 0.05 was fixed (meaning we accepted a 5% error in our analysis), and the results derived from the inequality *p-value* ≤ 0.05 were considered significant.

3.3.1 Normal distribution assessment - Shapiro-Wilk test

The first step when performing statistical analysis is to assess normal distribution. The most commonly used tests to assess normal distribution are Shapiro-Wilk (SW) test, Kolmogorov-Smirnov (KS) test, Anderson-Darling (AD) test and Lilliefors (LF) test[132]. In the present study we applied the Shapiro-Wilk test, since in a 2014 published article Shapiro-Wilk demonstrated to be the most powerful normality test for all types of distribution and sample sizes[132].

SW test is a test of departure from normality that compares the scores in the sample to a normally distributed set of scores with the same mean and standard deviation[133,134]. The null hypothesis of the SW test is that the sample comes from a normally distributed population[135]. If the null hypothesis is rejected there is evidence that the data tested doesn't originate from a normally distributed population[135]. In this sense the SW test was used to assess if miRNAs expression and DNA methylation values across all samples (tumor and normal combined) followed a normal distribution. MiRNAs/CpGs whit *p-value*s > 0.05 were elected to a two sample t-test, whereas miRNAs/CpGs with *p-values* ≤ 0.05 were selected to a Wilcoxon-Mann-Whitney test.

Shapiro-Wilk test was performed using the *shapiro.test* function available in R studio.

3.3.2 Two sample t-test

A two-sample t-test is a parametric test that is used to compare whether the mean value between two groups of individuals in a normally distributed variable are significantly different[136]. This test assesses whether two groups mean values are large enough that they do not derive from the same population and that this difference is not due either chance or sampling variation[137]. The null hypothesis in this test infers that there is no difference between the two group's means. Two sample t-tests were therefore used to determine if the mean expression/methylation values of miRNAs/CpGS in each of the four groups of tumor samples was different from the normal patients samples mean, and if these differences were large enough so that they could be considered statistically significant. Only miRNAs/CpGs considered statistically different were further used in our study.

Two sample t-tests were implemented using the *t.test* function available in R studio.

3.3.3 Levene's test

Two sample t-tests can only be performed under the assumption that the two samples display a normal distribution and an equal variance (σ^2)[136]. Therefore accessing if the samples have the same variance must be performed before proceeding to the t-test. The condition of equal variance can be verified using Levene's or Bartlett's test[138].

The Levene's test explores whether the variances of two samples are equal[139]. The null hypothesis is that the variances of two samples are equal. If the null hypotheses is rejected we acknowledge that the variances between the two samples are different, and therefore this must be stipulated when performing the t-test[137]. In this sense Levene's test was used to assess if the variances of the miRNAs/CpGS elected for T-Test were different enough between each of the four groups of tumor samples and normal patient samples, so that they could be perceived as statistically significant.

Levene's test was executed using the *leveneTest* function provided by *car* R package.

3.3.4 Wilcoxon-Mann-Whitney test

Wilcoxon–Mann–Whitney, is a non-parametric statistical technique often presented as an alternative to a t-test when the data doesn't follow a normal distribution[140]. However, instead of differences between means, this test compares the differences between the medians of two data sets[140]. The null hypothesis of this test infers that the two groups being compared have the same median value or derive from the same population[137]. Therefore rejecting the null hypothesises leads to the acknowledgment that the medians between the two are large enough that they do not emerge from the same population[133].

Wilcoxon-Mann-Whitney test was performed in order to determine if the median expression/methylation values of miRNAs/CpGS in each of the four groups of tumor samples were different from the normal patient samples median, and if these differences were large enough so that they could be considered statistically significant. Once more, only miRNAs/CpGs considered statistically different were further used in our study.

In a parallel analysis we also used a two-sample t-test to determine if the mean expression/methylation values of miRNAs/CpGS between four groups of tumor samples were different within them, and if these differences were large enough so that they could be considered statistically significant

Wilcoxon-Mann-Whitney test was performed using the *wilcox.test* function available in R studio.

3.3.5 Multiple testing correction

When performing multiple tests of hypothesis the probability of obtaining *p-values* lower than the critical value by chance, which results in the rejection of null hypothesis (type I error), strongly increases with the number of hypotheses[141]. In order to avoid type I errors in these situations, *p-values* must be corrected. One approach used to perform *p-values* correction is through the use of the false discovery rate (FDR) method[142]. Considering simultaneously m (null) hypotheses H_1, H_2 ..., H_m and assuming the respective *p-values* P_1,P_2 ..., P_m, in the FDR method, for a desired FDR level q (our α selected), the ordered *p-value* $P_{(i)}$ is compared to the critical value $q \times i / m$. Let k be the largest i for which:

$$P_{(i)} \leq \frac{i}{m} \times q$$

the we reject all $H_{(i)}$ i = 1, 2, ..., k[141]. Multiple testing correction was performed for results obtained in both Two sample t-test and Wilcoxon-Mann-Whitney test using the function *p.adjust* function available in the stats package and only miRNAs and CpGs with a corrected *p-value* ≤ 0.05 were selected.

3.4 Biomarker Analysis

3.4.1 Receiver Operating Characteristic (ROC) curves

Receiver operating characteristic (ROC) curve is a statistical tool to assess the ability of a quantitative test or biomarker to discriminate between two mutually exclusive states (*e.g.* healthy and disease)[143,144]. ROC curves result from the several combinations of sensitivity (true-positive rate - TPR) and the 1-specificity (false-positive rate - FPR) that the whole range of values across both patients and controls would have, if we selected each value as a disease marker to discriminate the two exclusive states[144].

The TPR is obtained by the ratio between the number of true decisions obtained when the cases where positive (true positives - TP) and the actually number of positive cases. The FPR results from the ratio between the number of false decisions obtained when the cases where actual positive (False positives -FP) and the number of actually positive cases[145].

All possible combinations of sensitivity and 1-specificity that can be summarised into a single parameter, the area under the ROC curve (AUC)[145]. The AUC provides the global estimate of the diagnostic accuracy that ranges from 0.5 to 1[146].

Roc curves were performed to assess the diagnostic ability of the miRNAs/ CpGs found differentially expressed and methylated respectively in stage I patients were able to accurately discriminate normal from stage I patients. The AUC values for MiRNAs and CpGs were considered perfect (AUC = 1) excellent (0.9 ≤ AUC < 1), good (0.8 ≤ AUC 0.9), fair (0.7 ≤ AUC 0.8) poor (0.6 ≤ AUC 0.7) and failed for (0.5 ≤ AUC 0.6) according to the classification provided in Khouli in 2009 "*Relationship of temporal resolution to diagnostic performance for dynamic contrast enhanced MRI of the breast"*[147].

ROC curves were performed using the *roc* function available in the *pROC* package.

3.4.2 Kaplan-Meier

Kaplan–Meier (KM) method makes it possible to calculate the incidence rate of events such as death or relapse-free survival time by using information from all subjects at risk for those specific events. [148]. The Kaplan–Meier method can be graphically represented through a curve. The KM survival curve is a graphical representation of the KM survival probability against time, were the cumulative survival probability is on the Y axis and the time passed after entering the study on X axis[148,149]. KM survival plots display the proportion of patients free of the event declining over time[150]**.** In this graphical representation the proportion surviving remains unchanged between the events, even if there are some intermediate censored observations, only dropping when an event occurs[149].

Kaplan–Meier curves were utilized to determine if miRNAs/CpGs found differentially expressed an methylated could provide good prognostic values regarding both overall survival (OS) or recurrence free survival (RFS). In these analyzes only tumor samples were used and the thresholds to divide patients into two distinct groups (Group 1 and 2 patients with miRNA/CpGs expression/methylation above or below the median, respectively).

When comparing the survival between two different groups, the KM method provides a mean of visually assessing whether survival was different for the two groups[150]. However, it does not provide a comparison of the total survival experience between the two groups[150]. In order to overcome this limitation we can use the nonparametric test logrank test[151].

The Kaplan–Meier method was performed using the *coxph* function available in the *survival* R package and the curves were plotted using the *gsurvplot* function accessible in the *survminer* package.

3.4.2.1 Logrank test

The logrank test is used to formally test whether the difference between two or more curves is statistically significant[150]. It is considered to be the most robust test for this purpose and tests the null hypothesis where there is no difference between the populations in the probability of an event at any time point[148,152]. This method calculates, for each time event, the expected number of events for each group if there were in reality no difference between the groups and the total number of observed events in each group[148]. These values are then summed every time an

event occurs to give the total expected number of events in each group. In our analysis only KM curves with logrank test *p-value* ≤ 0.05 with at least 20 patients in both groups were considered.

However, the log-rank test is incapable of estimating the size of the difference between groups. For these purposes, a regression technique like the Cox proportional hazards model can be used[150]

The logrank test was provided by the *coxph* function available in the *survival* R package.

3.4.2.2 Cox proportional hazards model

The hazard ratio (HR) represents a measure of the relative survival experience between two groups[148]. The HR indicates the risk of an event at any point in time among patients in one group compared with those in the other group[148].

The Hazard ratio was obtained through the *coxph* function available in the *survival* R package.

3.5 MiRNA functional analysis

3.5.1 MiRNAs target genes analysis

MirTarBase was used to retrieve the target genes of the differentially expressed miRNAs. MirTarBase is a manually collected dataset of microRNA-target interactions (MTIs) experimentally validated by reporter assay, western blot, microarray and next generation sequencing experiments that holds more than three hundred and sixty thousand MTIs. Only functional MTIs were considered in our analysis. MirTarBase 7.0, the latest version, released in September 15th 2017 was the one used in this work (http://mirtarbase.mbc.nctu.edu.tw).

3.5.2 Function and pathway enrichment analysis

Functional analysis using The Database for Annotation, Visualization and Integrated Discovery (DAVID) DAVID Bioinformatics Resources 6.8 (https://david.ncifcrf.gov/home.jsp) was performed in order to obtain (functional annotation analysis) gene-annotation enrichment analysis. DAVID integrates functional annotations from other datasets providing investigators a comprehensive set of functional annotation tools to understand the biological meaning behind lists of genes. Among the several datasets available in DAVID, Kyoto Encyclopedia of Genes

and Genomes (KEEG) (www.genome.jp/kegg/) was the one selected to accomplish our analysis.

A *p-value* <0.05 was considered to indicate statistical significance.

3.6 Bibliographic research analysis

Bibliography research was performed through the use of two R packages: OncoScore and RISmed. OncoScore is a tool that allows the user to measure the association of a term to cancer, based on citation frequency in biomedical literature found in PubMed. This tool not only provides the number of citations of a gene in cancer related articles but also provides the number of citations in any Pubmed article and thus produces a score (oncoscore) that represents the prevalence of the gene in cancer[153].

RISmed package allows the user to extract bibliographic content from the National Center for Biotechnology Information (NCBI) databases, including PubMed (https://cran.r-project.org/web/packages/RISmed/index.html). However, in comparison to OncoScore, this tool enables the user to perform a more narrow research, as the user can obtain the number of citations of a term in a specific query (*e.g.* colorectal cancer) while simultaneously selecting the time period he wants the research to be performed (*e.g.* Articles published from 2010 to 2018). In our analysis we performed a research in Pubmed articles published from 1787 to June 2018 for four queries: “cancer”, “colon cancer”, “rectal cancer” and “colorectal cancer”. Despite the similarity between the two packages the simultaneous use of both packages further strengthens the accuracy of our search analysis.

3.7 Study pipeline

Initially miRNA mature expression (RNAseq – IlluminaHiseq) and whole-genome DNA methylation (Illumina Infinium HumanMethylation 450K array) data was collected for both Colon and Rectal cancer patients (“TCGA-COAD” and “TCGA-READ”). After patient selection and stratification both miRNA mature expression and DNA methylation data were analyzed in order to determine which miRNAs and CpGs were differentially expressed and methylated respectively in each stage (I, II, III and IV) when compared to normal patients (Figure 3.2 – Step 1). To perform this analysis we first handled missing values and outliers in both datasets.

MiRNAs and CpGs with more than fifty percent of information missing were excluded and outliers for each miRNA and CpG in both solid tissue normal and primary tumor tissue for each group were eliminated. After this initial data processing we proceeded to perform statistical analysis in order to identify if there was statistical evidences for differences in miRNA expression and CpG methylation between each group and normal patients. Normal distribution was assessed through the use of Shapiro-Will test, and Wilcoxon-Mann-Whitney test together with Levene's test followed by a two-sample t-test used to determine if there were statistical differences. False discovery rate (FDR) was performed for both T-Test and Mann-Whitney and only miRNAs and CpGs with a FDR lower than 5% were considered (Figure 3.2 – Step 2). The logarithm base 2 of miRNAs expression value rates between each stage and normal patients ($\log_2$ (fold-change)) and CpG difference between the methylation mean values in each stage and normal patients ($\Delta\beta$) were obtained. Only miRNAs with absolute values higher than 1 and CpGs with absolute values higher than 0.2 were considered differentially expressed and methylated, respectively (Figure 3.2 – Step 3)[116]. MRNAs and CpGs diagnostic values were assessed through the use of ROC curves an their ability to behave as potential diagnostic biomarkers was stratified as suggested by Swets in 1988[154]. Prognosis value was determined by the log-rank test *p-value* when performing Kaplan-Meier^{o}s for both overall survival and recurrence free-survival, with a *p-value* ≤ 0.05 being statistically significant.

Moreover, a bibliographic research of miRNAs in the PubMmed repository was conducted. At last, miRNA functional analyzes were accomplished, where miRNA targets were determined and enriched pathways identified (Figure 3.2 – Step 5).

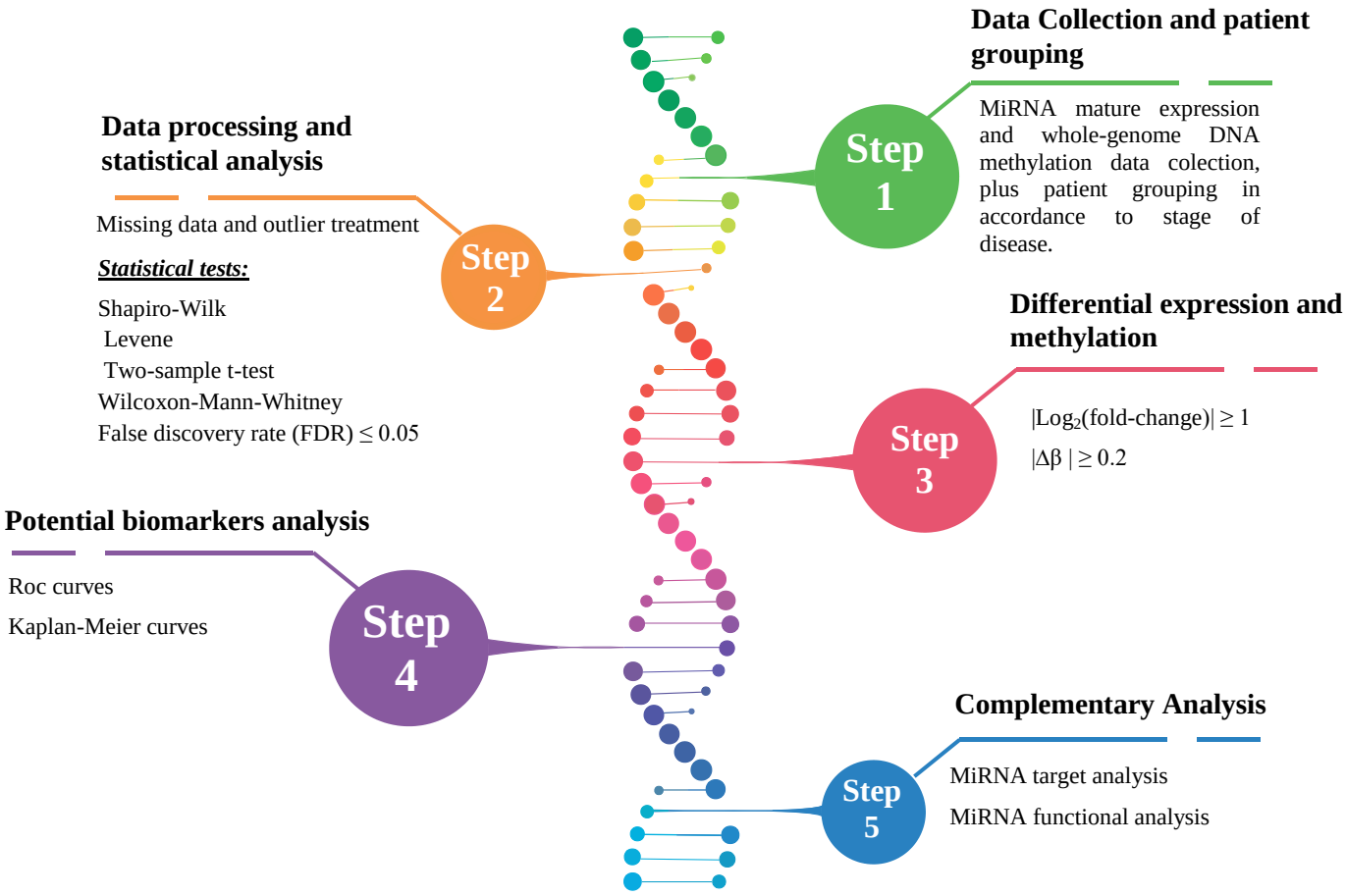

Figure 3. 2 Study Pipeline. (1) Mature expression and whole-genome DNA methylation data was obtained for both colon and rectal cancer patients from TCGA, and patients were grouped according to their stage of disease (2) Data processing and statistical analysis were the following steps. (3) Differentially expressed miRNAs and differentially methylated CpGs were determined (4) Diagnostic and prognostic value were evaluated through the use of Roc curves and Kaplan-Meier curves (5) Finally, complementary statistical analyzes such as miRNA bibliographic research and miRNA functional analyzes (target genes and pathways enrichment) were conducted.

CHAPTER 4 – Results:

4.1 - Differentially expressed miRNAs as potential CRC biomarkers

In recent years the role of miRNAs in the regulation of gene expression in cancer development, growth and metastasis has been intensely described[155,156]. These small non-coding RNAs participate both in early and later stages of disease, since tumor initiation to metastasis formation[155,157]. A better understanding of miRNAs behaviour throughout disease progression should provide crucial insights to better comprehend the carcinogenic process of CRC and how these small ncRNAs intervene in this process. Here we proposed to identify miRNAs differentially expressed during CRC progression. Also, we investigated if the identified miRNAs could be new potential diagnostic and/or prognostic biomarkers.

4.1.1 MiRNA deregulation is an early event in tumorigenesis

In order to identify differentially expressed miRNAs in CRC we started by collecting all mature miRNA expression data from the TCGA colon and rectal cohorts. From the 2116 initial miRNAs we were left with 491 miRNAs after applying listwise deletion. Then, outlier removal and statistical analysis was performed for stage I, stage II, stage III, stage IV and normal patient samples. Finally the $\log_2$ (fold-change) between miRNA expression values in each of the four groups of tumor samples and normal patients samples was calculated. MiRNAs with a $\log_2$ (fold-change) absolute value higher than 1 were considered differentially expressed in regard to normal patient samples expression values.

Our results showed that a total of 230 miRNAs were found differentially expressed between tumor and normal tissues (Supplementary Table I). Moreover, when comparing the groups containing samples from stage I, II, III or IV patients to the normal patient samples, a total of 213, 204, 213 and 216 miRNAs were found differentially expressed (Figure 4.1).

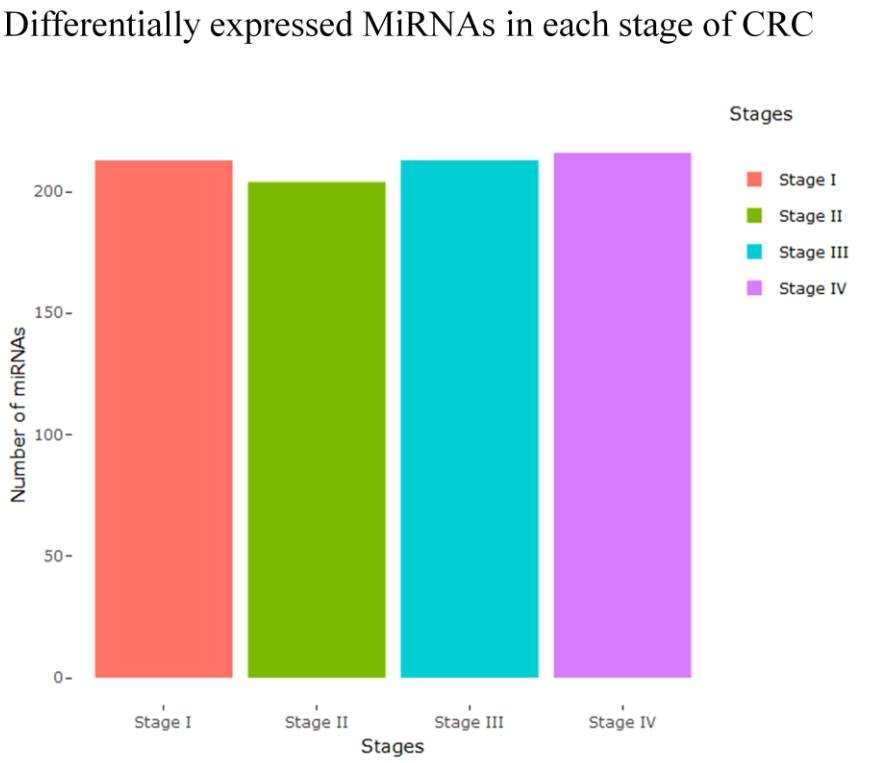

Figure 4.1 MiRNAs differentially expressed in each stage of disease. MiRNA expression values in each of the four groups of tumor samples with a $\log_2$ (fold-change) absolute value higher than 1 in regard to normal patient samples were considered differentially expressed. 213, 204, 213 and 216 miRNAs were found differentially expressed in stages I, II, III and IV respectively.

This result shows that throughout CRC progression, the number of miRNAs found differentially expressed is kept very similar since the initial phase of disease (stage I) until the latest phase (stage 4). Simultaneously Figure 4.2 clearly shows that the major alteration in miRNAs expression values occur in the Normal to stage I transition defined by the accentuated color changes, which further support our previous analysis. Not only that, once there is a change of colour in the transition from normal to stage I this colour is kept in the same gradient through the remaining stages of disease further emphasizing that during disease progression miRNA expression is kept very similar.

Also, the colours of the heatmap are representative of miRNA expression in disease progression and evidence that the vast majority of miRNAs transit from higher expression values in normal tissue into lower values in CRC. In fact, 83% of the miRNAs (191 out of the 230) are downregulated and only 17% (39 out of the 230 miRNAs) are upregulated from normal to malignancy (Figure 4.3), implying that the vast majority of miRNAs transit into a downregulated

state during the carcinogenic process. These results therefore may suggest that the vast majority of miRNAs under normal physiological conditions may act as tumor suppressors.

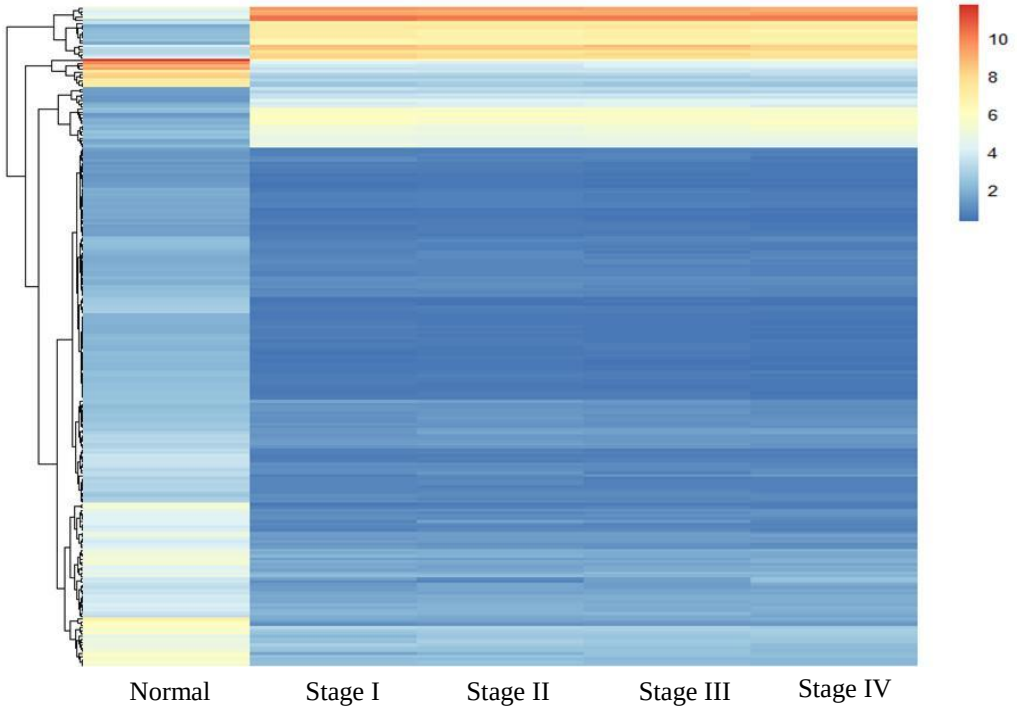

Figure 4. 2 Non-hierarchical heatmap of 11 Normal Tissue samples and 321 Primary Tumor samples across the four stages of CRC based on the total 230 miRNAs found differentially expressed between the four stages of disease. MiRNAs expression values are displayed in a gradient of colors that vary from dark blue, representative of lower expression values, to dark red representative of higher expression values.

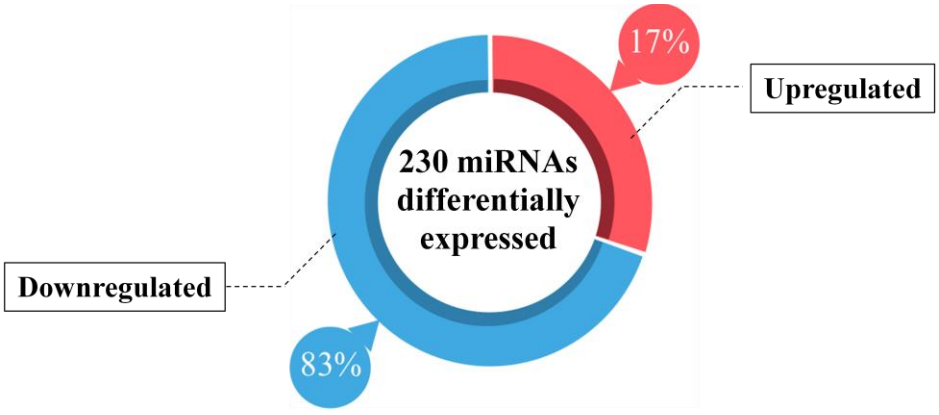

Figure 4. 3 Pie chart of Tumor (T) vs. Normal (N) miRNAs expression status. 191 out of the total 230 miRNAs found differentially expressed throughout disease progression, equivalent to 83%, were found downregulated (blue). Oppositely 39 miRNAs equivalent to 17%, were found upregulated (red).

Our analysis also demonstrated that the dimension of the differences between the expression values in tumor and normal patients is greater for downregulated miRNAs, regardless of the stage of disease (Figure 4.4). Log_2 (fold-change) values for the upregulated miRNAs barely surpass an absolute value of 2, while downregulated miRNAs exceed an absolute value of 3 (Figure 4.4).

We next asked if the majority of the miRNAs deregulated through all stages of disease were the same. As a mean to determine if differentially expressed miRNAs were simultaneously allocated in multiple stages of disease, a Venn diagram was plotted (Figure 4.5). This tool allowed us to determine how the differentially expressed miRNAs previously identified were distributed across the four stages of disease. We strengthened that the vast majority of miRNAs (213 out of 230, corresponding to 92 % of all miRNAs found altered during CRC) become deregulated in an initial phase of disease. Moreover we confirmed that the vast majority of miRNAs deregulated in the initial transition phase kept their deregulation through all stages of disease, as 191 out of the 213 (90%) miRNAs found differentially expressed in stage I were deregulated in all stages of CRC (Figure 4.5). Among the 230 deregulated miRNAs, we found 13 miRNAs exclusively deregulated in specific stages. 6 miRNAs (hsa-miR-141-5p, hsa-miR-3651, hsa-miR-3653-3p, hsa-miR-432-5p, hsa-miR-6087 and hsa-miR-887-3p) were specifically altered in stage I while 7 miRNAs(hsa-miR-128-1-5p, hsa-miR-1291, hsa-miR-30b-3p, hsa-miR-4664-3p, hsa-miR-501-5p, hsa-miR-659-5p, hsa-miR-939-5p) were differentially expressed only in stage IV (Figure 4.5). This result highlights a potential panel of miRNAs to help diagnose these specific stages.

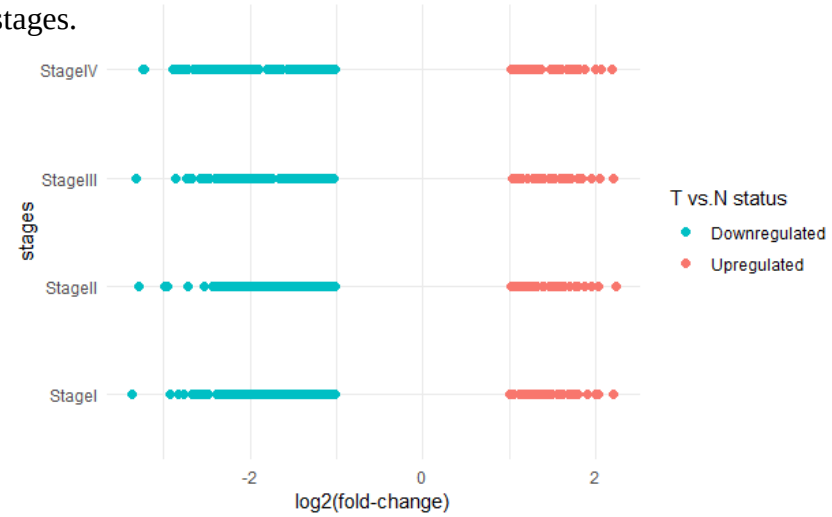

Figure 4. 4 Log_2 (fold-change) values for the miRNAs found differentially expressed in each stage. Range of log_2 (fold-change) values of differentially expressed miRNAs in each stage of disease. Values of upregulated miRNAs are represented in red while the ones of downregulated are represented in blue.

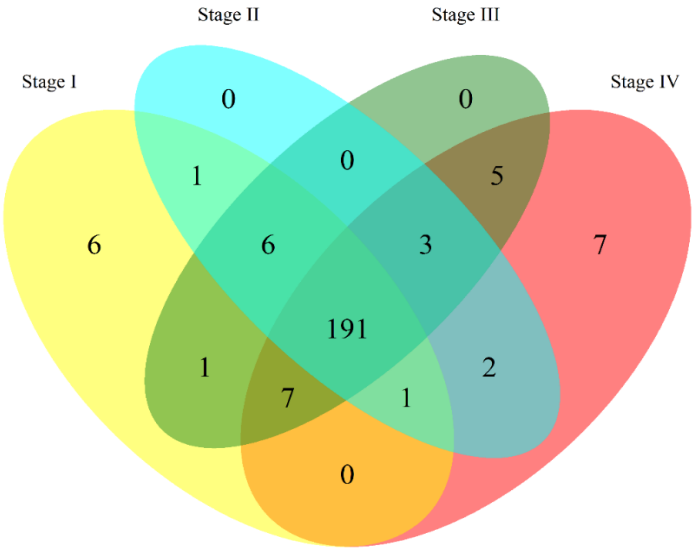

Figure 4. 5 Venn diagram of differentially expressed miRNAs correlated to the stages of disease throughout colorectal cancer progression. Allocation of the 230 differentially expressed miRNAs throughout disease progression across the four stages of disease.

Furthermore, we also found that stage III shared the highest number of differentially expressed miRNAs with the other stages: 206 miRNAs with stage IV (191+5+3+7), 200 with stage II (191+6+3+0) and 205 with stage I (191+6+7+1). Thus, stage III might be the most heterogeneous stage regarding miRNAs expression.

4.1.2 *Deregulated miRNAs target genes that are often altered in CRC*

Despite the ambiguous function of miRNAs, upregulated miRNAs are expected to be targeting Tumor Suppressor Genes (TSG), while downregulated miRNA are supposed to interact with oncogenes. Thus it is expected that upregulated and downregulated miRNAs target different genes involved in the carcinogenic process.

In order to get more insights into the function of the deregulated miRNAs we analysed the target genes (and associated metabolic pathways) of the 191 miRNAs constantly deregulated throughout tumor progression. Experimentally validated functional miRNA-target interactions available in the mirTarbase database were retrieved for the 155 downregulated and for the 36 upregulated miRNAs separately, and the target genes were obtained. A total of 360 target genes

were collected for the set of downregulated miRNAs, while 497 target genes were obtained for the panel of upregulated miRNAs.

Curiously, despite the inferior number of upregulated miRNAs (almost ¼ of the number of downregulated miRNAs) they showed a higher number of target genes. Interestingly when comparing the target genes of the upregulated and downregulated miRNAs 102 genes were found simultaneously targeted by both sets of miRNAs (Figure 4.6).

Figure 4. 6 Venn diagram of the target genes of down- and upregulated miRNAs. Functional validated target genes for upregulated and downregulated miRNAs were found and compared. The numbers of genes found for both upregulated (yellow) and downregulated miRNAs (blue) are shown. The section in green represents simultaneously targeted genes by both upregulated and downregulated miRNAs.

However, the validated functional miRNA-target interactions available in the mirTarbase database derive from a large diversity of experiments in different settings, and thus don`t necessarily represent what occurs in a CRC environment.

Nevertheless, genes heavily associated with the development of CRC such as *p53*, *KRAS*, APC, *PTEN*, *TGFβ2*, *SMAD4*, *Wnt3A* and *PIK3CA* were some of the genes found commonly targeted by both upregulated and downregulated miRNA (Table IV). Our analysis also demonstrates that among these genes *PTEN* was the most targeted one by the upregulated panel of miRNAs, while being one of the less targeted genes in the downregulated panel promoting a decreased expression of this gene in tumor tissue. Conversely, p53 showed the highest number of interactions in the downregulated miRNA panel, with fewer interactions in the upregulated miRNAs panel (Table IV) being more expressed in tumor tissue. For the remaining genes no great differences regarding the number of interaction in both panels were seen.

Table IV Genes involved in the CRC carcinogenic process targeted by both upregulated and downregulated miRNAs. Eight genes involved in the CRC carcinogenic process and the respective Tumor (T) vs. Normal (N) expression status. Upregulated and downregulated miRNAs found to target each one of these genes is shown.

Genes	Gene expression T. vs. N. status	Upregulated miRNAs	Downregulated miRNAs
p53	Up	miR-15a-5p , miR-19b-3p	miR-605-5p , miR-504-5p , miR-491-5p , miR-1228-3p , miR-125b-1-3p , miR-150-3p
KRAS	Down	miR-217 , miR-452-5p , miR-1-3p	miR-433-3p , miR-543
APC	Down	miR-106b-5p , miR-142-3p	miR-129-5p
PTEN	Down	miR-217 , miR-20a-5p , miR-19b-3p , miR-106b-5p , miR-29b-3p miR-429, miR-124-5p, miR-106b-5p	miR-205-p
TGFβ2	Up	miR-29b-3p , miR-142-5p	miR-490-3p , miR-23a-5p , miR-23b-5p
WNT3A	Up	miR-15-5p	miR-491-5p
SMAD4	Down	miR-20a-5p , miR-19b-3p , miR-144-5p	miR-483-3p , miR-205-5p
PIK3CA	Down	miR-1-3p	miR-139-5p

Hence miR-605-5p, miR 504-5p, miR-491-5p, miR-1228-3p, miR-125b-1-3p , miR-150-3p under normal physiologic conditions might be involved in suppressing *p53* expression, and their deregulation, in CRC, can contribute to higher expression of *p53.* Similarly, downregulation of miR-490-3p, miR-23a-5p, and miR-23b-5p can perhaps promote an increase expression of TGFβ2 in CRC. Moreover upregulation of WNT3A on the other hand might be associated with alteration of miR-491-5p.

Contrarily upregulation of these two panels of miRNAs: miR-217, miR-452-5p, miR-1-3p and miR-106b-5p, miR-142-3p might be in some way associated with the downregulation of KRAS and APC respectively. Furthermore upregulation of miR-217, miR-20a-5p, miR-19b-3p, miR-106b-5p, miR-29b-3p, miR-429, miR-124-5p, and miR-106b-5p may play a role in the diminished expression of PTEN. Finally our results suggest that upregulation of miR-20a-5p,

miR-19b-3p, miR-144-5p and miR-1-3p can also be associated with lower expressions of SMAD4 and PI3KCA respectively.

Nonetheless, a considerable amount of genes were exclusively targeted by either the upregulated or the downregulated set of miRNAs (supplementary Table II).

4.1.3 *Upregulated and Downregulated miRNAs target similar pathways*

After identifying the target genes for the both sets of miRNAs we proceeded to identify in which metabolic these genes were participating. Pathways were obtained from KEEG dataset, and for the upregulated panel of miRNAs a total of 103 pathways out of the 530 Pathways available in KEGG were found with a p-value ≤ 0.05 (Supplementary Table III). Meanwhile for the downregulated panel of miRNAs a total of 110 the 530 Pathways exhibited a *p-value* ≤ 0.05(Supplementary Table IV). Our results show that 97 pathways were found in common, portraying that the two sets of miRNAs affect genes that intervene mostly in the same pathways. These pathways were grouped into several subgroups according to their description: "Cancer related alterations", "Cellular pathways", "Cellular functions", "Non-cancerous pathologies" and "Viral/bacterial infection related pathologies". The number of pathways in each subgroup (in percentage) for both the upregulated and downregulated groups is shown in Figure 4.7 A and B, respectively. Comparing the two pie charts, the resemblances of the pathways regulated by the upregulated and downregulated set of miRNAs are notorious.

Moreover, our sub grouping system evidenced that a significant percentage (20% in the upregulated set and 17% in downregulated set) of the pathways are associated with cancer related alterations (Figure 4.7). These results further support the association between miRNAs deregulations and several alterations seen in cancer.

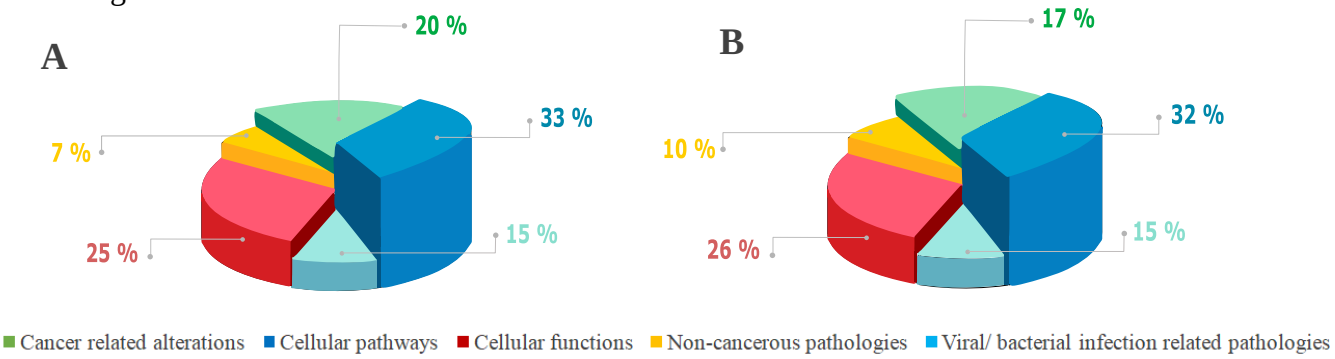

Figure 4. 7 Pie charts depicting pathway subgrouping for both upregulated (A) and downregulated (B) miRNAs. The number of pathways in each of the five subgroups is represented as percentage.

Additionally, some pathways in the cancer related alterations subgroup, such as: "Pathways in cancer", "MicroRNAs in cancer", "Transcriptional misregulation in cancer" and most importantly "Colorectal cancer" strongly support that association (Table V). Together these results solidify the idea that alterations in miRNA expression might be associated with the development of cancer, and more specifically of CRC.

Furthermore, various pathways usually affected in CRC such as: "PI3K-Akt signaling pathway", "RAS signaling pathway"," TGF-beta signaling pathway", "TGF-beta signaling pathway" and "p53 signaling pathway" were present in the subgroup Cellular pathways (Table V).

Table V List of pathways found in the subgroups "Cancer related alterations", "Cellular pathways" and "Cellular functions". The number of total genes found associated with each pathway (Total Hits), the number of genes belonging to the Total Hits found in our analysis (Nº of hits), the *p-value* and the false discovery rate (FDR) are provided.

Group	Term	Nº of Hits	Total Hits	*P-Value*	FDR
Cancer related alterations	Pathways in cancer	79	394	2.62E-39	3.30E-36
	MicroRNAs in cancer	50	262	4.52E-23	5.71E-20
	Prostate cancer	31	89	3.39E-22	4.29E-19
	Colorectal cancer	26	66	4.32E-20	5.45E-17
	Non-small cell lung cancer	20	56	1.78E-14	2.24E-11
	Renal cell carcinoma	19	66	6.14E-12	7.75E-09
	Small cell lung cancer	20	91	2.51E-10	3.16E-07
	Transcriptional misregulation in cancer	20	171	8.47E-06	0.010692
Cellular pathways	PI3K-Akt signaling pathway	52	341	1.72E-19	2.17E-16
	p53 signaling pathway	19	71	2.38E-11	3.00E-08
	Rap1 signaling pathway	31	214	6.06E-11	7.64E-08
	MAPK signaling pathway	34	259	8.57E-11	1.08E-07
	Ras signaling pathway	31	231	4.16E-10	5.25E-07
	HIF-1 signaling pathway	20	95	5.48E-10	6.91E-07
	TNF signaling pathway	21	107	6.74E-10	8.51E-07
	Wnt signaling pathway	20	139	3.55E-07	4.49E-04
	TGF-beta signaling pathway	14	84	6.63E-06	0.008368
	mTOR signaling pathway	11	58	3.00E-05	0.037807
Cellular functions	Adherens junction	24	73	1.67E-16	2.78E-13
	Apoptosis	16	70	1.42E-08	1.79E-05
	Regulation of actin cytoskeleton	26	217	1.48E-07	1.87E-04
	Cell cycle	17	126	8.15E-06	0.01029

Finally, for the subgroup "Cellular functions", functions such as "Apoptosis", "Regulation of actin cytoskeleton" and "Cell cycle" came up in our results. These mechanisms are highly important to maintain a correct cellular function under normal physiological conditions and become impaired during the carcinogenic process (Table V).

For the "Non-cancerous pathologies" and "Viral/ bacterial infection related pathologies" no relevant results were found in a cancer context.

4.1.4 Identification of novel miRNAs as diagnostic biomarkers for early stage CRC

Patients with CRC are often asymptomatic in the earlier stages of the disease. A sooner diagnosis correlates with patient survival evidencing the need to develop accurate diagnostic tools. Hence we looked at the 213 differentially expressed miRNAs in the early stage of disease (Stage I) as aiming to identify potential biomarkers in CRC.ROC curve analysis showed that all the 213 differentially expressed miRNAs could distinguish normal and stage I CRC (AUC >0.8). In fact, 107 miRNAs (50%) could be considered perfect biomarkers with AUC equal to 1 (Figure 4.7). Moreover, 97 miRNAs (46%) were excellent biomarkers ($0.9 \leq AUC < 1$) while only 9 miRNAs (4%) were good biomarkers ($0.8 < AUC < 0.9$) (Figure 4.8 and Supplementary Table V)[147]. These results thus suggest that that differently expressed miRNAs could be considered good diagnostic biomarker in an early stage of disease (Stage I), with almost half of the 213 miRNAs found to be have good prognostic values being considered perfect biomarkers. As examples of miRNAs for the different classifications according to *Khouli in 2009* classification, Roc curves and respective box plots are present in Figure 4.8.

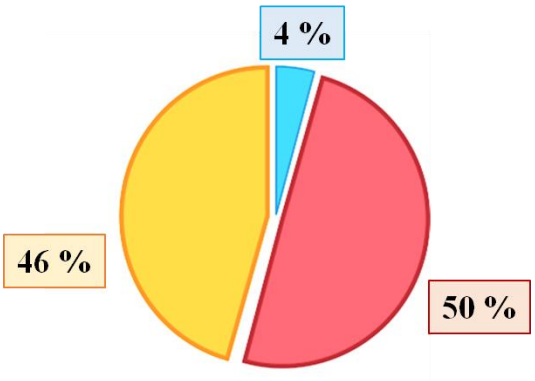

Figure 4. 8 Pie chart of the Stage I differentially expressed miRNAs distributed in accordance to the stratification suggested by Khouli in 2009. Blue represents good, orange excellent and red perfect miRNAs as biomarkers for diagnosis.

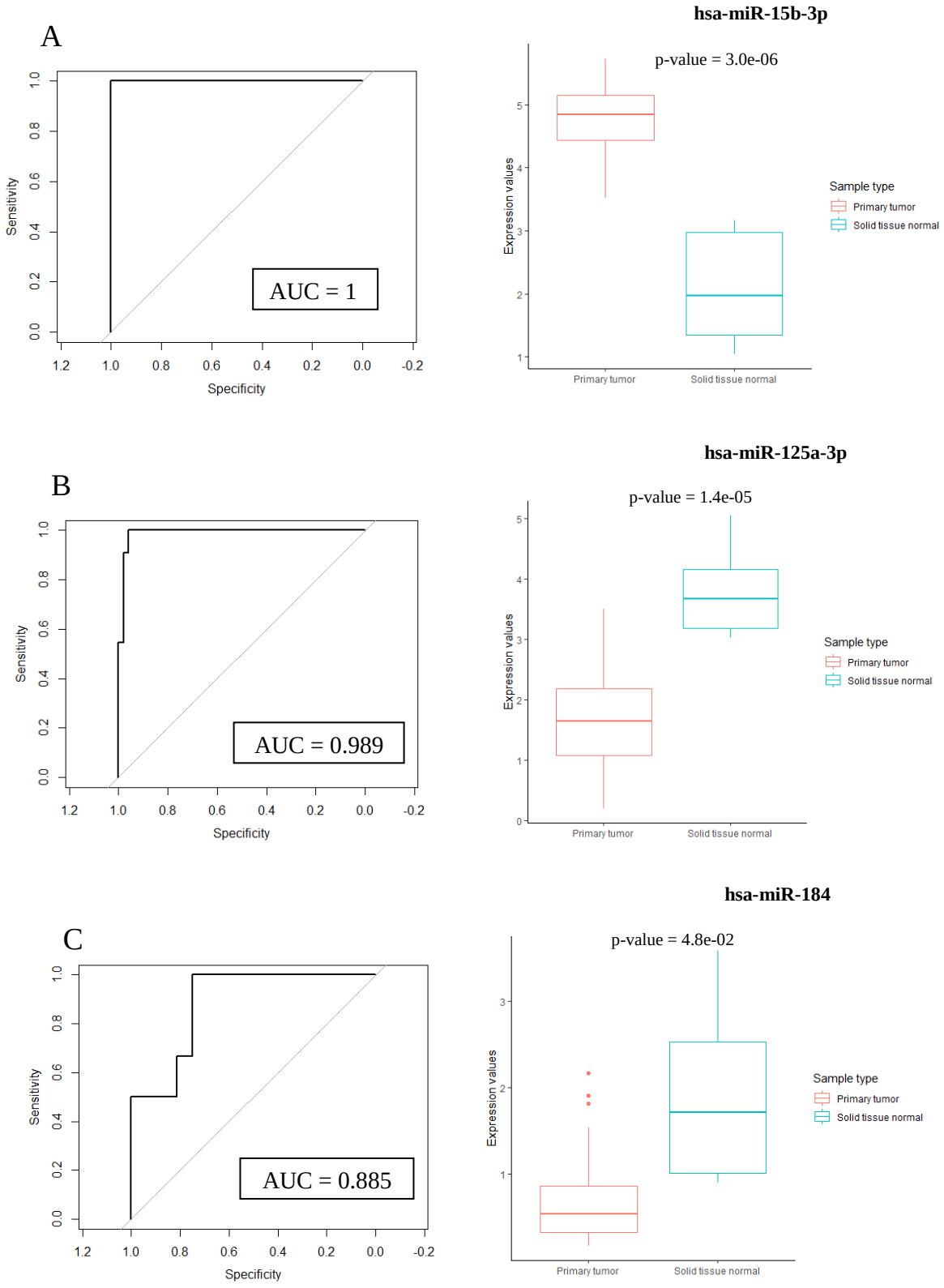

Figure 4. 9 MiRNA expression profiling and diagnostic accuracy for stage I differentially expressed miRNAs. Roc curve plots and box plots of the respective expression level in both Normal and Stage I sample are shown for A) hsa-miR-15b-3p, B) hsa-miR-125a-3p and C) hsa-miR-184.

In line with these results we then performed bibliographical analyzes to interrogate if the 213 miRNAs identified as good diagnostic biomarkers had been described in the literature. Out of the 213 miRNAs, 143 (dark blue -Figure 4.8) had already been described either in colorectal, colon or rectal cancers related articles. In contrast, the remaining 70 miRNAs had never been associated to any of these pathologies. Within these 48 (green) had already been reported in other cancers, 6 (orange) had never been associated to cancer but were described in non-cancer related articles and 16 (red) had never been cited in any PubMed article until the date of our analysis (Figure 4.10).

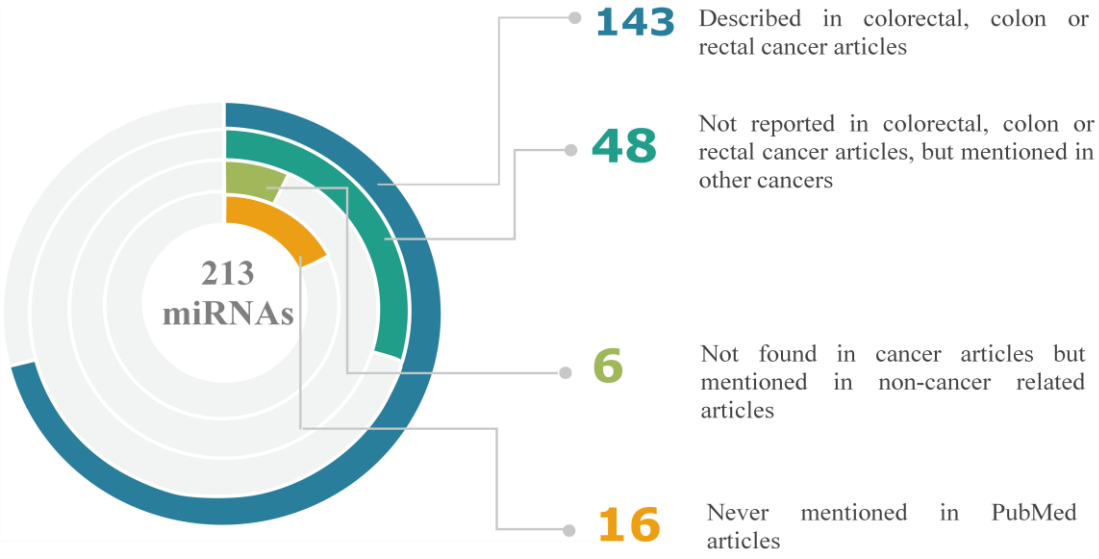

Figure 4. 10 Bibliographic search for the 213 miRNAs as potential diagnostic biomarkers. Rismed and oncoScore packages were first performed for the 213 miRNAs in order to uncover if they had ever been mention in either: colorectal cancer, Colon cancer or Rectal cancer articles found in PubMed. The set of miRNAs found in at least one of these cancers (dark blue) was dissociated from the group of not reported miRNAs. This sec**o**nd group was discriminated into three categories: "Not reported in colorectal, colon or rectal cancer articles, but mentioned in cancer" (dark green), "Not found in cancer articles but mentioned in non-cancer related articles" (light green) and "never mentioned in any PubMed article"(orange).

Our results thus suggest that to the best of our knowledge we have identified 70 novel good diagnostic biomarkers in CRC. Importantly, from the 70 miRNAs that have not been associated to colorectal, colon or rectal cancer, 4 miRNAs could be classified as good biomarkers according to *Khouli in 2009* classification, 33 highly excellent and 33 perfect biomarkers as they provided the ability to distinguish normal from stage I patients with 100% certainty (Table VI)[147]. Interestingly all 6 miRNAs found differently expressed only in stage I

have already been reported in either colorectal, colon or rectal cancers related articles. Curiously the 6 miRNAs found exclusively deregulated in stage I, despite providing good to perfect diagnostic values, had already been described in CRC (Supplementary Table VI).

Table VI List of the 70 miRNAs not previously associated with colorectal/colon/rectal cancers. The name, fold change, regulation status in Tumor tissue (T) versus normal tissue (N), Area under the curve (AUC), Optimal cut point, Sensibility and Specificity of each miRNA is presented.

miRNAs	Log_2(fc)	Status	AUC	Optimal cut point	sensibility	specificity
hsa-miR-105-5p	-1.3530	Down	0.8889	3.0489	0.8857	1.0000
hsa-miR-187-3p	-2.6630	Down	0.9802	0.9607	0.8696	1.0000
hsa-miR-383-5p	-1.6573	Down	0.9902	2.6303	1.0000	0.8333
hsa-miR-191-3p	-1.1382	Down	1.0000	3.6982	0.9783	1.0000
hsa-miR-616-5p	-1.2253	Down	0.9146	2.3356	0.9792	0.7000
hsa-miR-629-3p	-1.2566	Down	1.0000	4.2407	0.9796	1.0000
hsa-miR-767-5p	-2.0021	Down	0.9667	2.9985	0.9333	0.9091
hsa-let-7e-3p	-1.2559	Down	1.0000	3.4003	0.9792	1.0000
hsa-miR-19b-1-5p	1.5569	Up	1.0000	1.4324	1.0000	0.8333
hsa-miR-22-5p	1.0420	Up	0.9458	2.3911	0.9583	0.7000
hsa-miR-23a-5p	-1.3189	Down	0.9777	3.5441	0.9184	1.0000
hsa-miR-7-1-3p	1.4696	Up	1.0000	2.3594	1.0000	0.8889
hsa-miR-10a-3p	1.1406	Up	0.9271	1.5653	0.8980	0.7143
hsa-miR-218-1-3p	-2.1416	Down	1.0000	1.0238	0.9615	1.0000
hsa-miR-23b-5p	-1.7132	Down	0.9977	1.6147	0.9767	1.0000
hsa-miR-125b-1-3p	-1.3026	Down	0.9907	1.6370	0.9444	1.0000
hsa-miR-144-5p	1.9980	Up	1.0000	2.5619	1.0000	0.8750
hsa-miR-150-3p	-1.8673	Down	1.0000	2.8604	0.9792	1.0000
hsa-miR-296-3p	-2.0021	Down	0.9903	1.0450	0.9355	1.0000
hsa-miR-20b-3p	-1.8480	Down	0.9427	0.7299	0.7500	1.0000
hsa-miR-193b-5p	-1.6019	Down	1.0000	3.2347	0.9796	1.0000
hsa-miR-92b-5p	-1.6764	Down	1.0000	1.1365	0.9667	1.0000
hsa-miR-877-3p	-2.3394	Down	1.0000	0.9629	1.0000	1.0000
hsa-miR-887-3p	-1.0167	Down	0.8913	2.2805	0.9783	0.8000
hsa-miR-1228-3p	-2.0232	Down	1.0000	1.2470	1.0000	1.0000
hsa-miR-1181	-2.0091	Down	1.0000	0.9629	1.0000	1.0000
hsa-miR-1295a	-1.4937	Down	1.0000	1.0680	0.9600	1.0000
hsa-miR-1270	-1.6092	Down	0.9686	1.1847	0.9143	0.9000
hsa-miR-1306-3p	-1.4924	Down	0.9941	2.2634	0.9783	1.0000
hsa-miR-365a-5p	-2.5640	Down	1.0000	1.1807	0.9737	1.0000
hsa-miR-1976	-1.1859	Down	1.0000	4.3381	0.9783	1.0000
hsa-miR-2116-3p	-2.1244	Down	1.0000	1.1123	0.9767	1.0000
hsa-miR-3130-3p	-1.7169	Down	0.9711	1.2169	1.0000	0.8000

miRNAs	Log_2(fc)	Status	AUC	Optimal cut point	sensibility	specificity
hsa-miR-323b-3p	-1.1666	Down	0.9735	2.7894	0.8776	1.0000
hsa-miR-3193	-1.0442	Down	0.9167	1.4688	0.9737	0.5000
hsa-miR-3199	-1.9640	Down	1.0000	0.9407	0.9737	1.0000
hsa-miR-2277-5p	-1.4043	Down	0.9634	0.9761	0.8718	1.0000
hsa-miR-3614-5p	-1.0506	Down	0.9958	3.1928	0.9583	1.0000
hsa-miR-3680-3p	-1.3504	Down	0.9432	0.9787	0.8409	1.0000
hsa-miR-3690	-1.5471	Down	0.9361	0.9263	0.8333	0.9000
hsa-miR-3909	-1.2023	Down	0.9628	0.8962	0.8378	1.0000
hsa-miR-3928-3p	-1.7876	Down	1.0000	2.2035	0.9778	1.0000
hsa-miR-3940-3p	-2.6464	Down	1.0000	1.3584	0.9756	1.0000
hsa-miR-3944-3p	-2.3274	Down	1.0000	1.0347	1.0000	1.0000
hsa-miR-3158-5p	-1.4839	Down	0.9926	0.8879	0.9630	1.0000
hsa-miR-3173-5p	-2.2078	Down	1.0000	1.6138	0.9756	1.0000
hsa-miR-3940-5p	-1.9551	Down	0.9960	2.2664	0.9778	0.9091
hsa-miR-1343-3p	-1.8975	Down	1.0000	1.9092	1.0000	1.0000
hsa-miR-203b-3p	1.6845	Up	1.0000	2.8524	1.0000	0.9091
hsa-miR-4742-3p	-1.4952	Down	0.9887	0.8603	0.9211	1.0000
hsa-miR-5090	-2.2603	Down	1.0000	0.9154	0.9630	1.0000
hsa-miR-5091	-1.3201	Down	0.8930	1.2451	0.8837	0.8000
hsa-miR-5698	-2.0150	Down	0.9937	1.2375	1.0000	0.8889
hsa-miR-365b-5p	-1.6125	Down	0.9365	1.4744	0.9714	0.6667
hsa-miR-381-5p	-1.0436	Down	0.8905	1.0473	0.8667	0.8571
hsa-miR-503-3p	-1.2981	Down	0.9277	1.1713	0.7436	1.0000
hsa-miR-6509-5p	-1.5113	Down	0.9742	1.1113	0.8837	1.0000
hsa-miR-6715b-3p	-1.3732	Down	0.9246	1.8185	0.9024	0.8182
hsa-miR-6720-3p	-2.3817	Down	1.0000	1.3576	0.9756	1.0000
hsa-miR-552-5p	1.7517	Up	1.0000	3.0397	1.0000	0.8750
hsa-miR-874-5p	-1.5078	Down	0.9365	1.1997	0.9167	0.8571
hsa-miR-6793-5p	-2.8333	Down	1.0000	0.5260	0.9600	1.0000
hsa-miR-6798-3p	-1.9971	Down	0.9764	0.8812	0.9091	1.0000
hsa-miR-6802-3p	-2.5549	Down	1.0000	1.0182	0.9643	1.0000
hsa-miR-6803-3p	-2.6437	Down	1.0000	1.0228	0.9697	1.0000
hsa-miR-6808-3p	-2.1357	Down	1.0000	0.9911	0.9615	1.0000
hsa-miR-6877-5p	-1.9550	Down	0.9913	1.2455	0.9697	0.8571
hsa-miR-6892-5p	-1.6355	Down	1.0000	2.4277	0.9787	1.0000
hsa-miR-7704	-1.1441	Down	0.9222	2.2288	0.9500	0.6667
hsa-miR-7706	-1.3752	Down	1.0000	2.8750	0.9792	1.0000

These analyses further demonstrate that differently expressed miRNAs are able to distinguish tumor from normal tissues, even at early stages of disease (Stage I) with high values of sensitivity and specificity.

4.1.5 Identification of miRNAs with potential prognostic value in CRC

Clinical stratification of patients according to TNM stage is well established in the clinic, and is the most common classification system to stratify patients according to disease progression[24]. Despite being one of the most important clinical features to stratify patient according to the extent of disease, this classification method isn`t able to predict patient outcome[158]. New biomarkers that are able to discriminate patients according to their outcome within each specific stage are thus in need and would help clinical decisions regarding patient surveillance and/or treatment.

In this sense we used Kaplan-Meier graphical representations to perform Overall Survival (OS) and Recurrence Free Survival (RFS) analyzes for the miRNAs differentially expressed in each stage of disease. Also, in order to increase the prognostic potential of the biomarkers here identified we combined the miRNAs with good individual prognostic values. This approach enabled us to create a panel of miRNAs that could distinguish patients with different outcomes with higher precision and accuracy.

Due to the lack of patients with reported death or recurrence status in stage I, prognostic analyses in this group of patients were not performed. However stage II analysis revealed that from a total of 204 differentially expressed miRNAs, 10 could be considered good prognostic biomarkers for OS while 9 miRNAs could be considered good prognostic biomarkers for RFS (Table VII). The hazard ratio (HR) indicates the probability of survival at any time point and thus helps assess how good the prognosis of one group compared to the other is. In our analyzes the HR was obtained comparing group 1 to group 2, meaning that HR superior to one indicates a better prognosis for group 1 patients while HR below one implies a better prognosis for group 2 patients. The bigger (or smaller) the HR value the better is the biomarker.

Our stage II OS analysis evidenced that higher expression values of hsa-miR-5187-5p, hsa-miR-105-5p and hsa-miR-10a-3p in tumor patients were associated with better OS. In contrast, higher expression of hsa-miR-7704, hsa-miR-5091, hsa-miR-1271-5p, hsa-miR-142-5p, hsa-miR-6734-5p, hsa-miR-142-3p and hsa-miR-6509-5p were associated with worst outcomes. Moreover, hsa-miR-5187-5p was identified as the best individual OS prognostic biomarker. Interestingly, hsa-miR-10a-3p, which is upregulated in tumor samples, provided a better OS value when more expressed whereas hsa-miR-7704, hsa-miR-5091, hsa-miR-1271-5p, hsa-miR-

6734-5p and hsa-miR-6509-5p are all downregulated in tumor samples andshowed better OS when less expressed. Furthermore hsa-miR-142-5p beholds both good OS and RFS values.

Table VIII Stage II differentially expressed miRNAs with good prognostic value for both OS and RFS. The median expression value, number of patients in each group (group 1 – patients above miRNA median, group 2- patients below the miRNA median), p-value, Hazard Ratio (HR) and regulation status of Tumor (T) vs. Normal (N) are presented.

Overall survival analysis

miRNAs	median	group 1	group 2	*p-value*	HR	T vs. N status
hsa-miR-5187-5p	0.4306	46	45	0.0095	4.1568	Down
hsa-miR-105-5p	0.5434	33	33	0.0311	3.8461	Down
hsa-miR-10a-3p	2.9958	58	59	0.0239	2.9355	Up
hsa-miR-7704	1.1364	51	53	0.0470	0.3977	Down
hsa-miR-5091	0.7839	55	54	0.0452	0.3793	Down
hsa-miR-1271-5p	1.2297	54	53	0.0049	0.2648	Down
hsa-miR-142-5p	5.9155	57	60	0.0029	0.2550	Up
hsa-miR-6734-5p	0.4659	35	35	0.0129	0.2215	Down
hsa-miR-142-3p	10.3024	57	60	0.0003	0.1593	Up
hsa-miR-6509-5p	0.6067	49	49	0.0020	0.1313	Down

Recurrence free survival analysis

miRNAs	median	group 1	group 2	*p-value*	HR	T vs. N status
hsa-miR-2277-3p	0.3573	25	26	0.0269	5.0999	Down
hsa-miR-6877-5p	0.4156	36	38	0.0315	4.5765	Down
hsa-miR-4521	0.3655	30	32	0.0489	3.3995	Down
hsa-miR-491-5p	1.1629	52	53	0.0341	2.5642	Down
hsa-miR-642a-5p	1.4251	51	56	0.0319	0.3853	Down
hsa-miR-142-5p	5.9155	53	59	0.0164	0.3432	Up
hsa-miR-142-3p	10.3024	52	60	0.0018	0.2412	Up

Concerning RFS, higher expression levels of hsa-miR-2277-3p, hsa-miR-6877-5p, hsa-miR-4521 and hsa-miR-491-5p were associated with better RFS prognosis. In contrast higher expression values of hsa-miR-642a-5p, hsa-miR-142-5p and hsa-miR-142-3p were associated with lower RFS prognostic values in tumor samples. In stage II patients, hsa-miR-2277-3p was thus the best RFS prognostic biomarker.

In order to identify a reliable panel of miRNAs as potential CRC biomarkers, we combined the prognostic value of the miRNAs mentioned above. For that, we assembled the miRNAs with good prognostic value found in each analysis (10 in the OS analysis and 7 in the

RFS) in combinations of 2, 3 and 4. When combining 2 and 3 miRNAs we found 15 good combinations for OS and 4 for RFS, while combinations of 4 miRNAs were not statistically significant.

The panel of miRNAs that provided the best discriminatory value identified in the OS analysis was the combination of hsa-miR-142-3p and hsa-miR-5091 (*p-value* = 0.00022). This combination had a HR of 0.0541, meaning that the survival probability at any time point was approximately 18.5 (1/0.0541) times higher in patients with lower miRNA expression (group 2) than the ones with high miRNA expression (group 1) (Figure 4.11 A).

Regarding RFS analysis, the panel that could better discriminate both groups was the combination of hsa-miR-142-5p, hsa-miR-142-3p and hsa-miR-642a-5p (*p-value* = 0.0034). The HR for this panel was 0.1913, meaning that the probability of patients with lower miRNA expression (group 2) being free of disease at any point in time is approximately 5.2 (1/0.1913) times higher than patients in group 1 (Figure 4.11 B).

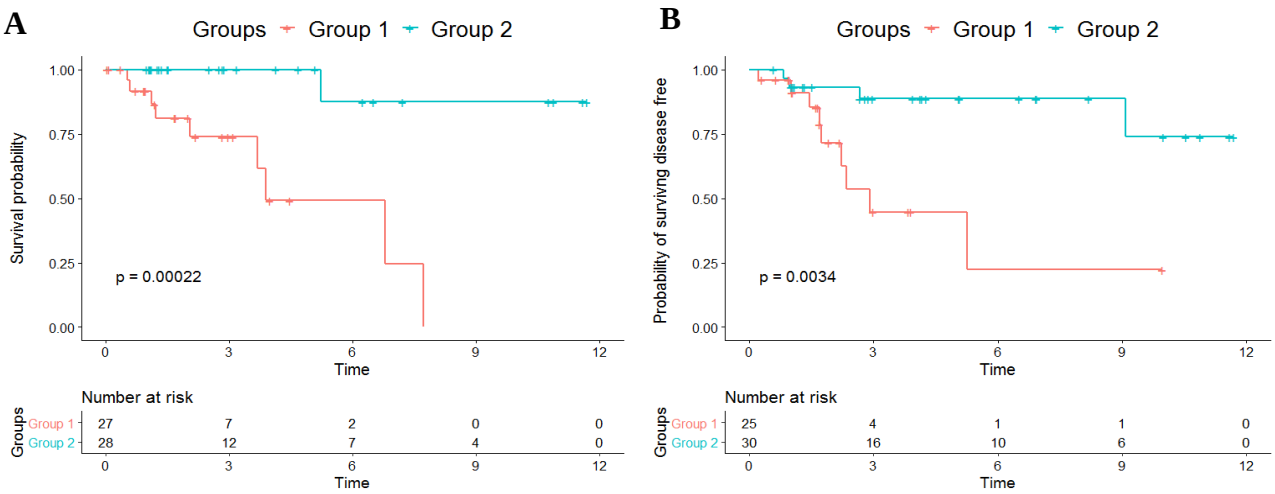

Figure 4. 11 Best miRNA panel for prognosis of Stage II patients. (A) Kaplan-Meier overall survival curve based on hsa-miR-142-3p and hsa-miR-5091 combined expression (p = 0.00022, Logrank test) and (B) Kaplan-Meier recurrence free survival curve based on hsa-miR-142-5p, hsa-miR-142-3p and hsa-miR-642a-5p conjoint expression (p = 0.0034, Logrank test). The respective number of patients at risk for each group at several time points is shown in the table below each graph.

The same analyses were then performed for stages III and IV. Regarding stage III we found that from a total of 213 miRNAs differentially expressed in this stage, 14 exhibited good prognostic values for OS while 13 could b be considered good prognostic biomarkers for RFS (Table VIII). In respect to OS analysis, miRNAs such as hsa-miR-543, hsa-miR-3150b-3p, hsa-miR-1180-3p, hsa-miR-4521, hsa-miR-433-3p, hsa-miR-1228-5p, hsa-miR-1254 and hsa-miR-

3614-5p were found downregulated in cancer and coincidentally showed better OS values when more expressed. On the other hand hsa-miR-133b, hsa-miR-23b-5p, hsa-miR-187-3p, hsa-miR-20b-3p, hsa-miR-34c-3p despite being downregulated in cancer, provided better prognostic values in tumor samples with higher expression values. Hsa-miR-144-5p, which was found upregulated in CRC, paradoxally exhibited better prognostic values in tumor samples with higher expression values. Interestingly both hsa-miR-543 and hsa-miR-1228-5p concomitantly provided good OS and RFS values.

Table VIII Stage III differentially expressed miRNAs with good prognostic value for both OS and RFS. The median expression value, number of patients in each group (group 1 – patients above miRNA median, group 2- patients below the miRNA median), p-value, and HR and regulation status of Tumor (T) vs. Normal (N) are presented.

Overall survival analysis

miRNAs	median	group 1	group 2	*p-value*	HR	T vs. N status
hsa-miR-543	0.8043	47	50	0.0018	4.1959	Down
hsa-miR-3150b-3p	1.4239	46	49	0.0119	2.9249	Down
hsa-miR-1180-3p	3.6438	51	52	0.0106	2.8295	Down
hsa-miR-4521	0.3738	33	35	0.0335	2.6908	Down
hsa-miR-433-3p	1.0820	47	49	0.0240	2.6186	Down
hsa-miR-1228-5p	0.5338	35	38	0.0387	2.5667	Down
hsa-miR-144-5p	5.4766	51	52	0.0498	2.3192	Up
hsa-miR-1254	0.7454	46	49	0.0424	2.2720	Down
hsa-miR-3614-5p	1.7178	50	53	0.0430	2.2257	Down
hsa-miR-133b	1.7656	48	47	0.0474	0.4383	Down
hsa-miR-23b-5p	0.7120	47	47	0.0495	0.4228	Down
hsa-miR-187-3p	0.8760	36	38	0.0351	0.3893	Down
hsa-miR-20b-3p	0.5462	25	25	0.0428	0.3413	Down
hsa-miR-34c-3p	1.1101	49	49	0.0048	0.3047	Down

Recurrence free survival

miRNAs	median	group 1	group 2	*p-value*	HR	T vs. N status
hsa-miR-639	0.5236	38	39	0.0157	3.3224	Down
hsa-miR-125b-1-3p	0.7831	39	39	0.0229	3.0651	Down
hsa-miR-1228-3p	0.5245	33	34	0.0458	2.7812	Down
hsa-miR-543	0.8043	43	47	0.0291	2.5977	Down
hsa-miR-3199	0.5418	40	43	0.0458	2.5729	Down
hsa-miR-7-1-3p	4.4708	47	49	0.0444	2.3481	Up
hsa-miR-6837-3p	0.8738	43	46	0.0439	0.3917	Down
hsa-miR-126-5p	6.5366	48	48	0.0139	0.3627	Up

miRNAs	median	group 1	group 2	*p-value*	HR	T vs. N status
hsa-miR-6511b-3p	1.1908	49	45	0.0195	0.3458	Down
hsa-miR-381-5p	0.3830	34	34	0.0378	0.3188	Down
hsa-miR-6820-3p	0.3527	30	29	0.0240	0.2870	Down
hsa-miR-365b-5p	0.4360	32	29	0.0014	0.1278	Down
hsa-miR-6802-3p	0.3567	28	28	0.0077	0.1012	Down

Regarding RFS analysis, hsa-miR-639, hsa-miR-125b-1-3p, hsa-miR-1228-3p, hsa-miR-543, hsa-miR-3199 downregulation was associated with better diagnostic.

When combining the prognostic ability of 2 or 3 miRNAs it was found 27 interactions for OS. Meanwhile combinations of 4 miRNAs were not statistically significant in either analysis. Regarding RFS analysis, only combinations of 2 were found statistically significant with a total of 9 combinations statistically significant.

MiRNAs hsa-miR-187-3p and hsa-miR-34c-3p provided the best discriminatory value in the OS analysis (*p-value* = 0.00097). This combination had a HR of 0.0721, implying that the survival probability at any time point was approximately 13.9 (1/0.0721) times higher in patients with lower miRNA expression than the ones with high miRNAs expression (Figure 4.12 A).

Regarding RFS analysis, the preferable panel to better discriminate both groups was the combination of hsa-miR-7-1-3p and hsa-miR-543 (*p-value* = 0.0010). The HR for this panel was 14.4650, revealing that, at any point in time, the probability of patients in group 1 being free of disease is approximately 14.45 times higher than patients in group 2 (Figure 4.12 B).

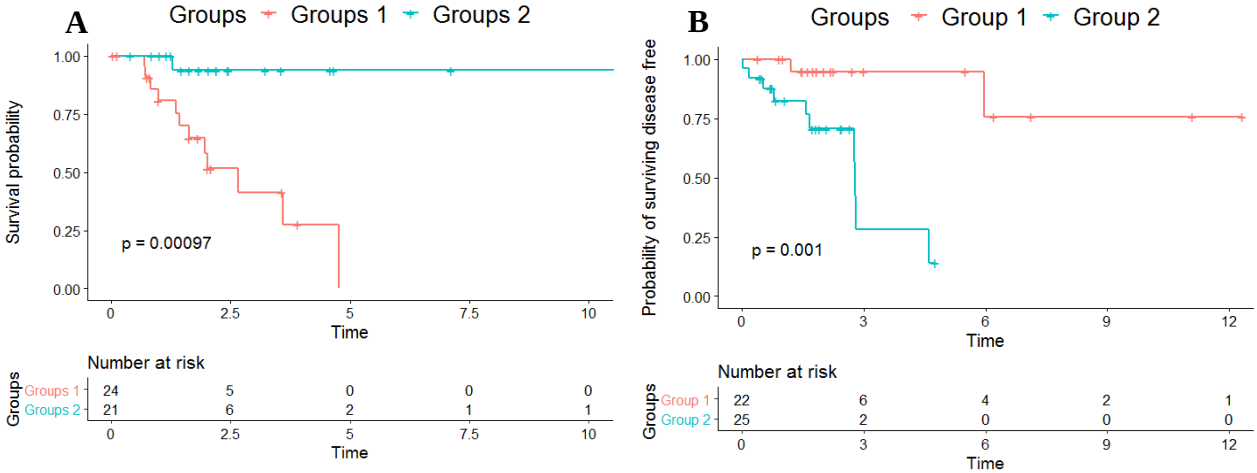

Figure 4. 12 Best miRNA panel for prognosis of Stage III patients. (A) Kaplan-Meier overall survival curve based on hsa-miR-187-3p and hsa-miR-34c-3p combined expression (p = 0.00097, Logrank test) and (B) Kaplan-Meier recurrence free survival curve based on hsa-miR-7-1-3p and hsa-miR-543 3p conjoint expression, (p = 0.0010, Logrank test). The respective number of patients at risk for each group at several time points is shown in the table below each graph.

Finally from the 216 differentially expressed miRNAs in stage IV, 5 miRNAs could be considered good prognostic biomarker for OS (Table IX). Our results demonstrate that higher expression values for all 5 miRNAs are associated with better OS prognosis. Nonetheless, we cannot fail to mention that better OS values resultant of higher of hsa-miR-215-5p and hsa-miR-144-5p are incongruent with their upregulated status in tumor patients. Interestingly, hsa-miR-144-5p was found as a potential OS biomarker in stage II and stage IV. Interestingly no panel of miRNAs was found statistically significant for stage IV and thus hsa-miR-3609 could be perceived the best OS biomarker for stage IV.

Table IX Stage IV differentially expressed miRNAs with good prognostic value for OS. The 5 differentially expressed miRNAs in stage III with good OS values are represented together with the median expression value, number of patients in each group (group 1 – patients above miRNA median, group 2- patients below the miRNA median), p-value, HR and regulation status of Tumor (T) vs. Normal (N).

Overall survival analysis						
miRNAs	**median**	**Group 1**	**Group2**	***p-value***	**HR**	**T vs. N status**
hsa-miR-3609	0.8722	21	21	0.0014	6.2810	Down
hsa-miR-129-5p	1.8047	20	21	0.0209	4.2295	Down
hsa-miR-215-5p	6.9004	20	21	0.0153	3.5821	Up
hsa-miR-144-5p	8.0587	21	22	0.0203	3.5598	Up
hsa-miR-320c	5.5391	22	21	0.0407	2.7285	Down

Regarding RFS, no miRNA appeared as a good potential prognostic biomarker. Moreover no combination of 2, 3 or 4 miRNAs was found statistically significant for both OS and RFS analysis.

4. 2 - DNA methylation of miRNAs as potential CRC biomarkers.

4.2.1 MiRNA methylation is an early event in tumorigenesis

Alterations in the DNA methylation status are known to play a significant role in several cancers including CRC [90]. Also, previous reports evidenced the potential of DNA methylation events as epigenetic biomarkers[92,93].

MiRNAs are transcribed as any other gene, and thus miRNA genes are also targets of DNA methylation[89]. Here, we investigated if DNA methylation of miRNA genes could act as diagnostic and/or prognostic epigenetic biomarkers in CRC.

First, we identified the 3439 CpGs sites associated with miRNAs genes from the 485578 CpG that are covered by 450k arrays available at TCGA. From those, after listwise deletion we continued our analysis with 2916 CpGs. Then, outlier removal and statistical analysis were performed and the difference between the mean of CpGs methylation values in each of the four groups and normal patients was obtained and expressed as $\Delta\beta$. Finally CpGs with an absolute $\Delta\beta$ value higher than 0.2 were considered differentially methylated in regard to normal patient methylation values.

When comparing each group to normal patients we obtained 413, 310, 344 and 349 CpGs differentially methylated in the stages I, II, III and IV, respectively (Figure 4.13). These results show that contrarily to what we observed in our miRNA expression analysis, the number of differentially methylated CpGs is not as constant throughout the four stages of disease, reaching its maximum in the initial phase of disease (Stage I) and significantly dropping in stage II, only to start increasing in the subsequent stages.

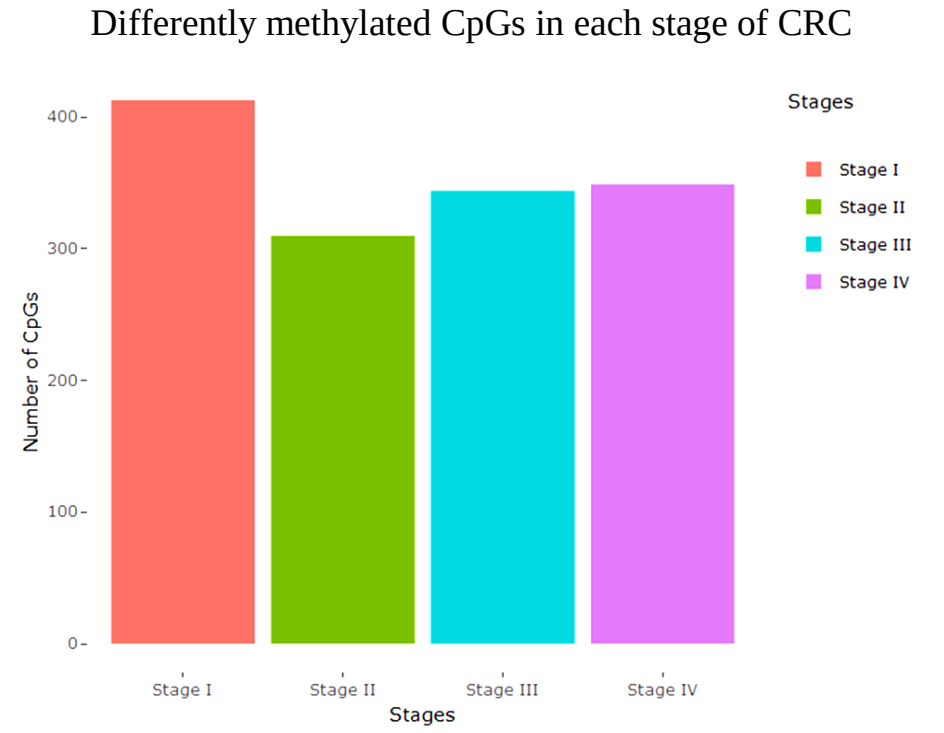

Figure 4. 13 CpGs differentially methylated in each stage of disease. CpGs methylation values in each of the four groups of tumor samples with a Δβ absolute value higher than 0.2 when compared to normal patient samples were considered differentially methylated. 413, 310, 344 and 349 CpGs were found differentially methylated in stages I, II, III and IV respectively.

We next asked how the differentially methylated CpGs were distributed through all the four stages of disease. To answer this we used a Venn diagram and found that a total of 439 CpG sites were differentially methylated throughout colorectal cancer development (Figure 4.14). We observed that only about 60% of CpGs (264 out of the total 439) were differentially methylated across all stages, keeping their altered methylation status from stage I until stage IV (Figure 4.14). Also, we could identify alterations exclusively found to specific stages of CRC. For each stage I, II, III and IV, we identified 45, 1, 2 and 17 CpGs specifically differentially methylated, respectively (Figure 4.14).

Our results show that similarly to what previously observed in our miRNA expression analysis, an extensive deregulation event is seen in the early transition from Normal tissue to stage I, as from the total of 439 distinct CpGs found differentially methylated throughout all four stages of disease, 413 became altered during this initial transition (Figure 4.14). These evidences that such as miRNA expression alterations, modifications in the miRNAs methylation are early

events in tumorigenesis further sustaining the idea that this changes may contribute to CRC initiation. Furthermore, the heatmap of the CpG methylation values in normal samples and in the different stages of CRC (Figure 4.15) evidenced that the most accentuated color changes occur in the Normal to stage I transition, supporting the Venn diagram results that the main methylation alteration occurs in CRC initial phase (Figure 4.14).

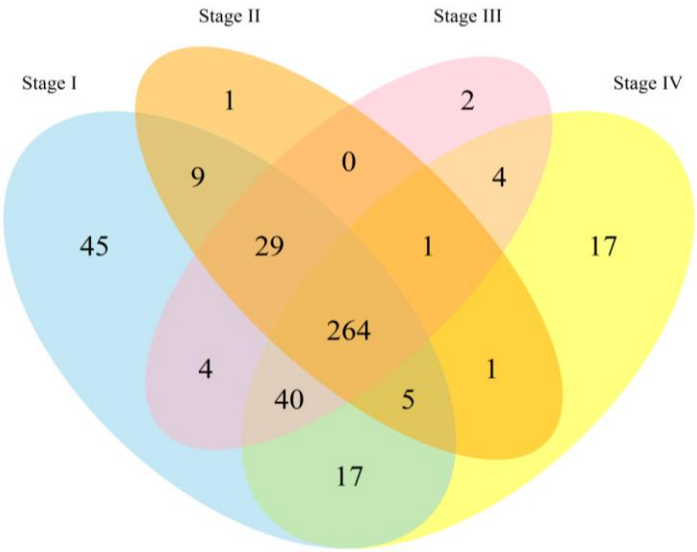

Figure 4. 14 Venn diagram of differentially methylated CpGs correlated to the stages of disease throughout colorectal cancer progression. Allocation of the 439 differentially methylated CpGs across the four stages of disease.

These results propose that a dominant deregulation in miRNA methylation occurs in the initial phase of disease coincidentally to the considerable miRNA expression deregulation observed in the same phase. Moreover, our results suggest that alterations in the miRNA methylation pattern might play, somehow, a more significant role in an early phase of CRC than in later stages of disease.

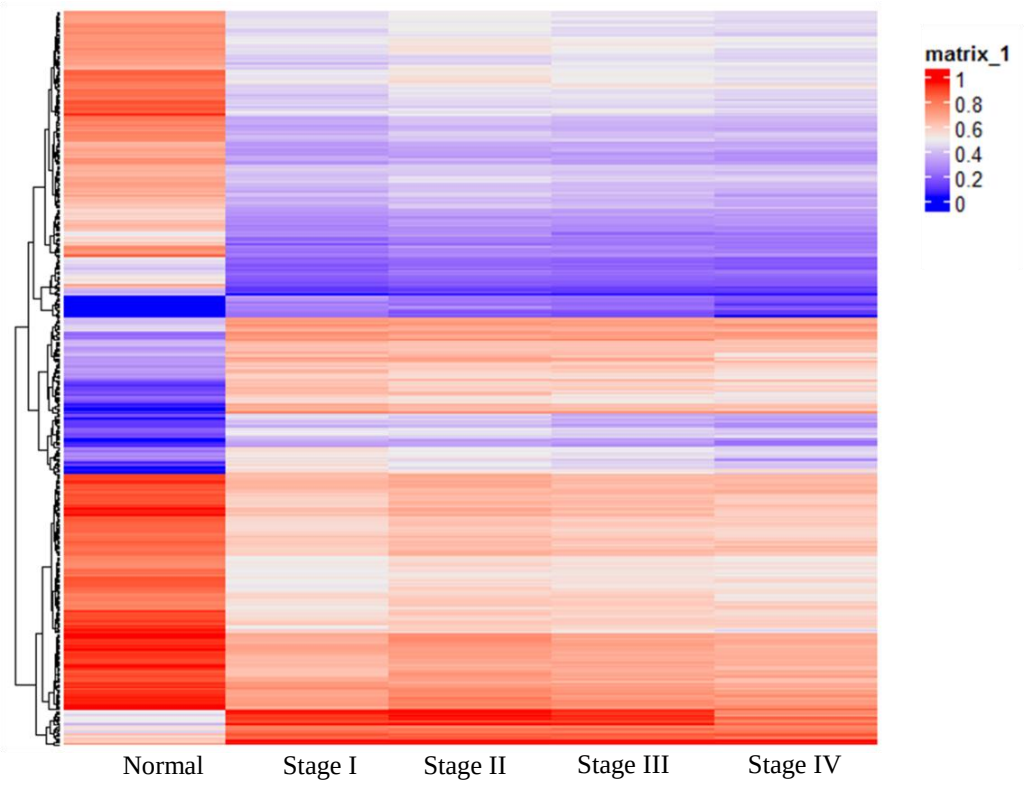

Figure 4. 15 Non-hierarchical heatmap of 45 Normal Tissue samples and 373 Primary Tumor samples into the four stages of CRC based on the total 439 CpGs found differentially methylated between the four stages of disease. CpGs methylations values are displayed in a gradient of colors that vary from dark blue, representative of lower expression values, to dark red representative of higher expression values.

4.2.2 MiRNA methylation status decrease in CRC

Furthermore, the colours of the heatmap in Figure 4.15, illustrative of CpG methylation in disease progression, with dark blue being representative of lower methylation values and red representative of higher methylation values, evidences that the vast majority of CpGs shifts from higher methylation values in normal tissue into methylation values in CRC. In fact, our results show that close to 71% of all altered CpGs (311 out of the 439 CpGs) become hypomethylated during tumorigenesis (Figure 4.16). Complementarily, about 29% (128 out of the 439 CpGs) become hipermethylated. The list of CpGs differentially methylated and respective methylation status of Tumor (T) vs. Normal (N) is presented in Supplementary Table VII.

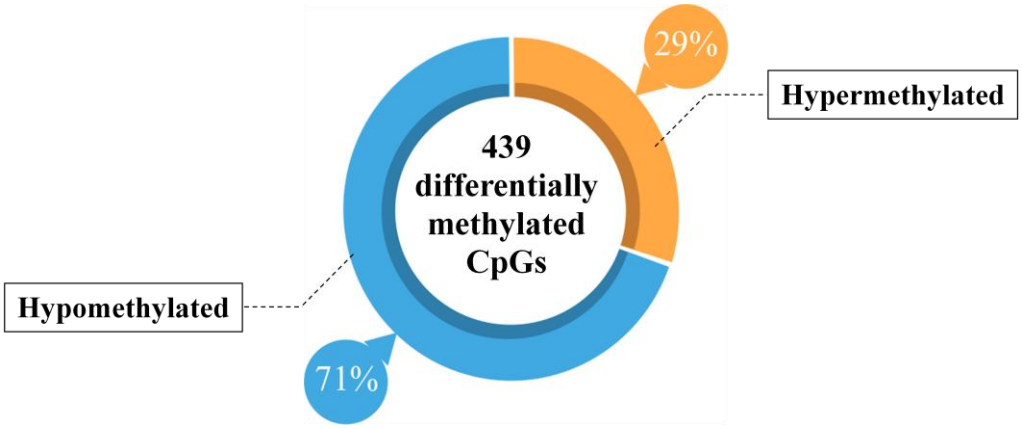

Figure 4. 16 Pie chart of Tumor (T) vs. Normal (N) CpGs methylation status. 311 out of the total 439 CpGs found differentially methylated throughout disease progression, equivalent to 71%, were found hypomethylated (blue). In contrast, 128 CpGs equivalent to 29%, were found hypermethylated (orange).

After evaluating the methylation status of the differently methylateed CpGs found in CRC patients, we proposed that this predominant downregulation event might be occurring due to the extensive downethylation events observed in CpGs scattered throughout the genome during the carcinogenic process[20,53]. In order to assess if this hypothesis was true we assess location of the 439 CpGs found differentially methylated in CRC. In order to do so, we determined what miRNAs these CpGs were affecting and the location of the CpGs for the respective miRNAs. Our results suggest that results about 80 % of all of all CpGs were located in promoter regions (comprising the portions up to 200 base pairs (bp) and 1500 bp from the transcription start site) (Figure 4.17). In cancer, promoter regions are exhaustively described to become aberrantly methylated (hypermethylated) affecting not only the expression of protein coding genes but also the expression of various non-coding RNAs[20,53]. Therefore out hypothesis could not be corroborated by these results, and the explanation to why this predominantly downmethylation even is observed is still unknown.

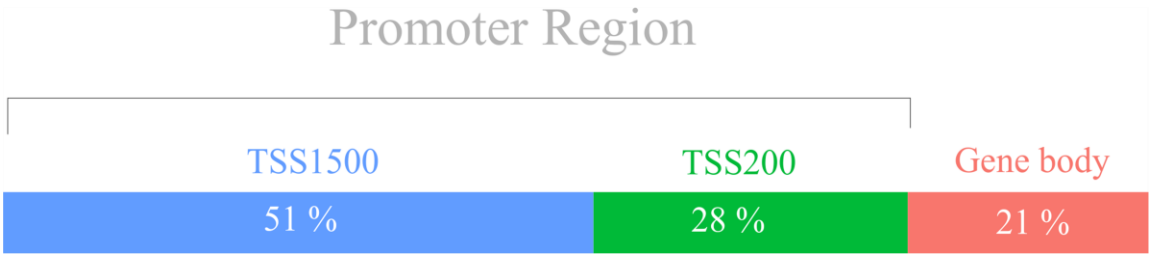

Figure 4. 17 CpGs sites location for the 439 differentially methylated CpGs. The percentage of CpGs located in miRNAs gene body (orange), up to 200bp from the TSS (TSS200, green) and up to 1500bp from the TSS (TSS1500, blue) is represented.

4.2.3 MiRNA methylation alterations as early diagnostic tools in CRC

Although altered miRNA expression could provide good diagnostic values, we were further interested in assessing their methylation value as a diagnostic tool for early stage of disease. In order to assess the diagnostic potential of the observed methylation changes, we performed Roc curves on CpGs found differentially expressed in stage I. From 413 miRNAs found differentially expressed in stage I patients, 3 CpGs (1%) could be considered poor ($0.6 \leq AUC < 0.7$), 21 (5%) fair ($0.7 \leq AUC < 0.8$), 127 (31%) good ($0.8 \leq AUC < 0.9$) , 228 (55%) excellent ($0.9 < AUC < 1$) and 34 (8%) perfect tests ($AUC = 1$) (Figure 4.18)[154]. A table with the 413 CpGs, the miRNAs they affect and respective AUC is provided in Supplementary Table VIII. Our results thus demonstrate CpGs found differently expressed can in stage I, close to 96 % (389 CpGs) could be considered good diagnostic biomarkers.

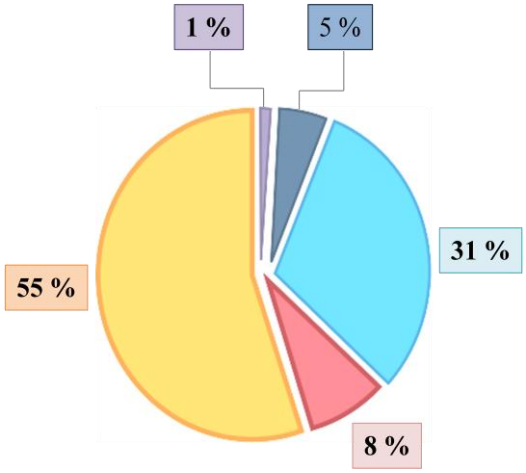

Figure 4. 18 Stage I differentially methylated CpGs distributed in accordance to the stratification suggested by Khouli in 2009. Purple represents poor, dark blue fair, light blue good, orange excellent and red perfect CpGs as biomarkers for diagnosis.

In accordance to our results, miRNA methylation, similarly to miRNA expression can provide good diagnostic values for early disease detection. Our result support that the vast majority of CpGs showed to be highly accurate to discriminate tumor form normal patients with a total of 34 CpGs (8%) being able to do so with 100% certainty.

4.2.4 MiRNA methylation alterations as prognostic tools in CRC

Finally, we interrogated if miRNA methylation could also predict CRC patients' outcome. We thus performed Kaplan-Meier curves for OS and RFS, and once again, due to the scarcity of patients with reported death and recurrence status in stage I, this analysis was only performed for stages II, III and IV.

Stage II analysis revealed that from a total of 310 differentially methylated CpGs, 68 could be consider good prognostic biomarkers for OS while 14 CpGs were good prognostic biomarkers for RFS (Table X). In regards to OS analysis, patients with higher methylation values of cg02558026 showed better OS, inconsistent with the hypermethylation observed in tumor patients. For the remaining 67 CpGs, lower methylation values were synonymous of better prognosis.

Concerning RFS analysis, higher methylation values of cg05509179 encompassed a better RFS value, congruent with its hypomethylation status in tumor patients. As for the remaining CpGs, found hypomethylated in tumor patients, reduced methylation values were paradoxically associated with better RFS prognosis. Interestingly cg01963147, cg02145866, cg02336334, cg05827233, cg09852439, cg1186929, cg16908824 and cg20587874 were simultaneously considered good potential OS and RFS biomarkers.

Table X Stage II differentially methylated CpGs with good prognostic value for both OS and RFS. The respective genes/miRNAs each CpG is affecting, median, β-value , number of patients in each group (group 1 – patients above miRNA median, group 2- patients below the miRNA median), p-value, Hazard ratio (HR) and methylation status of Tumor (T) vs. Normal (N) are presented.

Overall survival analysis

CpGs	genes/miRNAs	median	group 1	group 2	*p-value*	HR	T vs. N Status
cg02558026	MIR762;BCL7C	0.5927	67	68	0.0196	2.6777	Hyper
cg14148088	MIR494	0.5087	67	71	0.0485	0.4463	Hypo
cg12386297	LOC642587;MIR205	0.2731	70	68	0.0481	0.4403	Hypo
cg16908824	MIR1283-2	0.3423	68	68	0.0496	0.4397	Hypo
cg01882870	MIR9-3	0.4912	70	68	0.0461	0.4310	Hyper
cg12176191	MIR1295;FMO3	0.5327	68	70	0.0452	0.4300	Hypo
cg02632490	MIR346;GRID1	0.4883	67	70	0.0466	0.4290	Hypo
cg00505001	MAB21L1;MIR548F5;NBEA	0.4993	70	68	0.0441	0.4269	Hyper
cg18412777	MGC16121;MIR424;MIR503	0.6660	68	70	0.0380	0.4264	Hyper
cg00571033	MIR124-1;LOC157627	0.5464	70	68	0.0411	0.4224	Hyper
cg03019112	MIR487A	0.4184	67	71	0.0425	0.4188	Hypo

CpGs	genes/miRNAs	median	group 1	group 2	*p-value*	HR	T vs. N Status
cg13888600	MIR9-3	0.4344	69	69	0.0457	0.4179	Hyper
cg12038641	C20orf166;MIR133A2	0.4088	67	70	0.0466	0.4179	Hypo
cg17848546	LOC642587;MIR205	0.3399	69	68	0.0380	0.4170	Hypo
cg05896714	MIR1180;B9D1	0.7369	71	67	0.0449	0.4166	Hyper
cg23230910	ATP2B2;MIR885	0.7737	68	70	0.0401	0.4140	Hypo
cg26238975	MIR134;MIR668;MIR382;MIR485	0.7375	69	69	0.0404	0.4125	Hypo
cg18614984	MIR1253	0.3899	69	69	0.0358	0.4104	Hyper
cg02145866	MIR548G	0.5074	70	68	0.0315	0.4080	Hypo
cg23073467	MIR516A2;MIR519A2	0.3652	68	69	0.0382	0.4040	Hypo
cg22492271	MIR1200;ELMO1	0.5289	72	66	0.0384	0.4032	Hypo
cg25696807	MIR891A	0.1837	68	66	0.0304	0.4029	Hypo
cg09887953	SLIT3;MIR218-2	0.6117	69	69	0.0352	0.3976	Hypo
cg01192900	MIR34B;BTG4;MIR34C	0.4456	70	68	0.0320	0.3934	Hyper
cg03799024	MIR515-2;MIR515-1	0.3702	69	69	0.0308	0.3908	Hypo
cg19400113	TP63;MIR944	0.6233	69	69	0.0389	0.3903	Hypo
cg21592803	MIR525	0.7123	70	68	0.0285	0.3876	Hypo
cg15003468	MIR519B;MIR526B	0.4277	66	69	0.0252	0.3827	Hypo
cg25562958	MIR1197;MIR380;MIR323	0.4810	69	69	0.0250	0.3822	Hypo
cg08519216	MIR548N;PLEKHA3	0.2492	69	69	0.0229	0.3805	Hyper
cg11074814	MIR129-2	0.5320	69	69	0.0195	0.3788	Hyper
cg24428232	MIR921;FAM78B	0.3501	67	68	0.0366	0.3763	Hypo
cg26548251	MIR548F3;CNTNAP2	0.2789	67	71	0.0196	0.3747	Hypo
cg24554839	MIR921;FAM78B	0.4934	69	69	0.0240	0.3744	Hypo
cg12033297	MIR488;ASTN1	0.4089	68	70	0.0185	0.3740	Hypo
cg15282281	MIR346;GRID1	0.3630	67	71	0.0203	0.3722	Hypo
cg11638181	MIR129-2	0.5301	70	68	0.0208	0.3676	Hyper
cg06660530	MIR124-3	0.5202	69	69	0.0171	0.3675	Hyper
cg06961429	MIR329-2	0.2966	67	70	0.0168	0.3635	Hypo
cg23059797	VWA5B2;MIR1224	0.5470	69	69	0.0184	0.3605	Hypo
cg16506910	MIR30A	0.2301	68	67	0.0162	0.3604	Hypo
cg19157647	MIR1180;B9D1	0.9486	64	62	0.0173	0.3581	Hyper
cg24185864	MIR124-1;LOC157627	0.5325	71	67	0.0174	0.3579	Hyper
cg09177473	MIR1256	0.3992	67	68	0.0244	0.3560	Hypo
cg11869269	MIR377	0.4850	68	70	0.0128	0.3547	Hypo
cg22952287	H2BFWT;MIR1256	0.4210	69	67	0.0229	0.3537	Hypo
cg20587874	MIR548N	0.2178	67	68	0.0179	0.3516	Hypo
cg21881253	MIR34B;BTG4;MIR34C	0.5212	69	69	0.0127	0.3407	Hyper
cg10647025	MIR892A;MIR892B	0.2082	68	67	0.0109	0.3390	Hypo
cg11125104	NCRNA00164;MIR663B	0.5301	68	70	0.0100	0.3310	Hyper
cg01807688	MGC16121;MIR424;MIR503	0.7614	70	68	0.0089	0.3297	Hyper
cg05827233	MIR187	0.5754	69	69	0.0125	0.3207	Hypo
cg01963147	MIR320D2	0.5845	69	69	0.0094	0.3199	Hypo
cg20272287	MIR1180;B9D1	0.8692	71	66	0.0100	0.3157	Hyper
cg23322812	MIR346;GRID1	0.4869	68	70	0.0086	0.3149	Hypo

CpGs	genes/miRNAs	median	group 1	group 2	*p-value*	HR	T vs. N Status
cg18246262	MIR124-1;LOC157627	0.5202	70	68	0.0091	0.3113	Hyper
cg02975060	MIR1185-2	0.4985	69	69	0.0065	0.3074	Hypo
cg25972714	MIR346;GRID1	0.6456	69	69	0.0064	0.3033	Hypo
cg24553547	MIR1180;B9D1	0.9236	62	60	0.0065	0.2977	Hyper
cg02336334	MIR518A2	0.3643	69	69	0.0050	0.2958	Hypo
cg09852439	MIR516A1	0.3535	70	68	0.0050	0.2936	Hypo
cg13446906	MIR548F5;NBEA;MAB21L1	0.2315	66	68	0.0065	0.2919	Hypo
cg20474788	MIR377	0.5075	67	71	0.0027	0.2860	Hypo
cg08145178	MIR1256;H2BFM	0.4454	68	68	0.0030	0.2551	Hypo
cg24041078	BTG4;MIR34C	0.4766	70	68	0.0017	0.2518	Hyper
cg04864152	MIR1180;B9D1	0.9321	69	63	0.0025	0.2452	Hyper
cg17857791	MIR520B	0.4353	69	69	0.0026	0.2451	Hypo
cg00739582	MIR888;MIR890	0.5518	70	68	0.0027	0.2448	Hypo

Recurrence free survival analysis

CpGs	genes/miRNAs	median	group 1	group 2	*p-value*	HR	T vs. N Status
cg05509179	MIR548A2	0.4075	67	62	0.0266	2.4307	Hypo
cg11869269	MIR377	0.4850	65	65	0.0456	0.4530	Hypo
cg16908824	MIR1283-2	0.3423	65	63	0.0366	0.4288	Hypo
cg05827233	MIR187	0.5754	66	64	0.0426	0.4255	Hypo
cg17918158	COL8A1;MIR548G	0.5424	64	66	0.0348	0.4250	Hypo
cg12823329	MIR494	0.6448	54	56	0.0456	0.4148	Hypo
cg11201447	MIR1204;PVT1	0.1006	61	59	0.0372	0.4112	Hypo
cg02145866	MIR548G	0.5074	66	64	0.0202	0.3865	Hypo
cg26754262	MIR2052	0.2285	63	66	0.0162	0.3765	Hypo
cg20587874	MIR548N	0.2178	64	63	0.0179	0.3740	Hypo
cg01963147	MIR320D2	0.5845	66	64	0.0171	0.3692	Hypo
cg26024682	MIR670	0.3359	63	65	0.0143	0.3619	Hypo
cg02336334	MIR518A2	0.3643	66	64	0.0153	0.3589	Hypo
cg09852439	MIR516A1	0.3535	67	63	0.0147	0.3581	Hypo

Then we combined the prognostic ability of 2 or 3 CpGs sites and a total of 25648 combinations were found statistically significant for OS. However, no combination for 4 CpGS was significant for OS. Meanwhile with regards to RFS, when combining 2, 3 and 4 CpGs a total of 852 combinations were established as statistically significant.

The panel of CpGs that provided the best discriminatory value identified in the OS analysis was the combination of cg20474788 and cg24041078 (*p-value* = 0.000011). This combination had a HR of 0.0345, meaning that the survival probability at any time point was approximately 29 (1/0.0345) times higher in patients with lower methylation (group 2) than the ones with high methylation expression (group 1) (Figure 4.19A).

Regarding RFS analysis, the panel that could better discriminate both groups was the combination of cg01963147, cg02336334 and cg26024682 (*p-value* = 0.0002). The HR for this panel was 0.0529, implying that patients with lower miRNA expression (group 2) have a probability of being free of disease, at any point in time, approximately 19 (1/0.1913) times higher than patients in group 1 (Figure 4.19B).

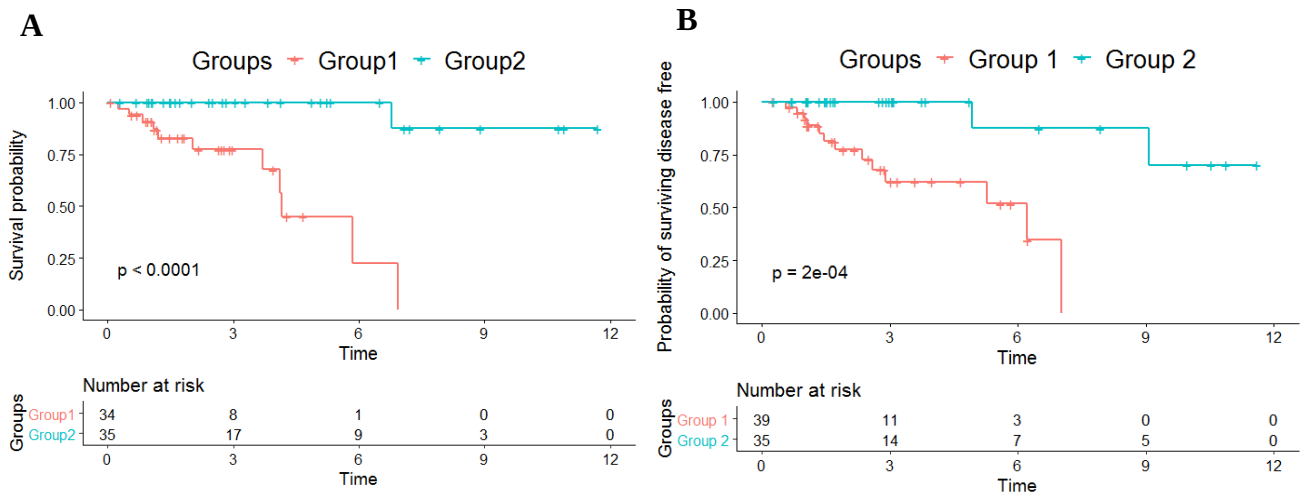

Figure 4. 19 Best CpGs panel for prognosis of Stage II patients. (A) Kaplan-Meier overall survival curve based on cg20474788 and cg24041078 collective methylation (p = 0.000011, Logrank test) and (B) Kaplan-Meier recurrence free survival curve based on cg01963147, cg02336334 and cg26024682 conjoint methylation (p = 0.0002, Logrank test). The respective number of patients at risk for each group at several time points is shown in the table below each graph.

Regarding stage III we found that from a total of 344 CpGs differentially methylated in this stage, 12 exhibited good prognostic values for OS while 16 could be considered good prognostic biomarkers for RFS (Table XI). Higher methylation values for cg1431159, cg06749053 and cg02226645 were associated with better OS prognostic values, conflictingly with the superior methylation value perceived in tumor samples. Meanwhile higher methylation values for all the other 9 CpGs were associated to a worst OS prognosis.

Regarding RFS, higher methylation values for 9 CpGs encompassed better prognostic values, whereas lower methylation values hold better prognosis for the remaining 7 CpGs. Curiously cg16407471, cg202653076 and cg21492137 were simultaneously linked to better OS and RFS biomarkers.

Table XI Stage III differentially methylated CpGs with good prognostic value for both OS and RFS. The respective genes/miRNAs each CpG is affecting, median, β-value, number of patients in each group (group 1 – patients above miRNA median, group 2- patients below the miRNA median), *p-value*, Hazard ratio (HR) and methylation status of Tumor (T) vs. Normal (N) are presented.

Overall survival analysis

CpGs	genes/miRNAs	median	group 1	group 2	*p-value*	HR	T vs. N Status
cg14311597	MIR1253	0.4000	60	58	0.0060	2.9701	Hyper
cg06749053	MIR125B1;LOC399959	0.5893	56	57	0.0344	2.2257	Hyper
cg02226645	MIR124-2	0.6712	59	57	0.0484	2.1436	Hyper
cg20653075	MIR124-2	0.6325	58	59	0.0434	0.4593	Hyper
cg07036112	PKD1;MIR1225	0.9148	54	55	0.0478	0.4563	Hyper
cg21492137	MIR496;MIR154	0.6885	59	59	0.0309	0.4408	Hypo
cg02278768	MIR9-3	0.5536	57	57	0.0217	0.4096	Hyper
cg16407471	MIR129-2	0.6306	57	57	0.0240	0.3983	Hyper
cg01939477	MIR129-2	0.6090	57	57	0.0161	0.3763	Hyper
cg14944647	MIR129-2	0.5325	56	56	0.0146	0.3685	Hyper
cg08613350	MIR548F5	0.7158	57	60	0.0080	0.3459	Hypo
cg17796010	MIR663	0.5603	57	58	0.0042	0.3227	Hyper

Recurrence free survival analysis

CpGs	genes/miRNAs	median	group 1	group 2	*p-value*	HR	T vs. N Status
cg02999711	MIR135B	0.6879	52	57	0.0087	3.1971	Hypo
cg25571269	MIR193A	0.5208	52	57	0.0108	2.8009	Hyper
cg02520707	MIR135B	0.4689	52	57	0.0276	2.4768	Hypo
cg16908824	MIR1283-2	0.3322	54	51	0.0467	2.3962	Hypo
cg12386297	LOC642587;MIR205	0.2612	54	54	0.0308	2.3764	Hypo
cg06660530	MIR124-3	0.4792	57	52	0.0340	2.3466	Hyper
cg02336334	MIR518A2	0.3801	53	56	0.0476	2.2699	Hypo
cg13888600	MIR9-3	0.3910	57	52	0.0427	2.2623	Hyper
cg00395657	MIR487A	0.5091	56	53	0.0498	2.2006	Hypo
cg12675571	LOC728264;MIR145	0.4141	51	58	0.0467	0.4550	Hypo
cg05376374	MIR129-2	0.6005	53	51	0.0449	0.4326	Hyper
cg21492137	MIR496;MIR154	0.6885	56	53	0.0313	0.4190	Hypo
cg20653075	MIR124-2	0.6325	53	55	0.0314	0.4183	Hyper
cg05455720	MIR124-2	0.6697	53	55	0.0339	0.4141	Hyper
cg16407471	MIR129-2	0.6306	52	53	0.0229	0.3898	Hyper
cg18515591	BTG4;MIR34C;MIR34B	0.5959	50	53	0.0246	0.3857	Hyper

When combining the prognostic ability of 2, 3 or 4 CpGs a total of 119 combinations were found statistically significant for OS and 112 were established as statistically significant for RFS

The combination of cg01939477, cg08613350 and cg14944647 (*p-value* = 0.00016) came up as the best panel in the OS analysis. This combination showed a HR of 0.0567, denoting that at any time point the survival probability was approximately 17.6 (1/0.0567) times higher in group 2 than in group 1 (Figure 4.20A).

Meanwhile, taking into account the RFS analysis, the panel that could better discriminate both groups was the combination of cg00395657 and cg25571269 (*p-value* = 0.000063). The HR for this panel was 22.0173, and thus patients in group 2 had a probability of being free of disease, roughly 22 times higher than patients in group 1 (Figure 4.20B).

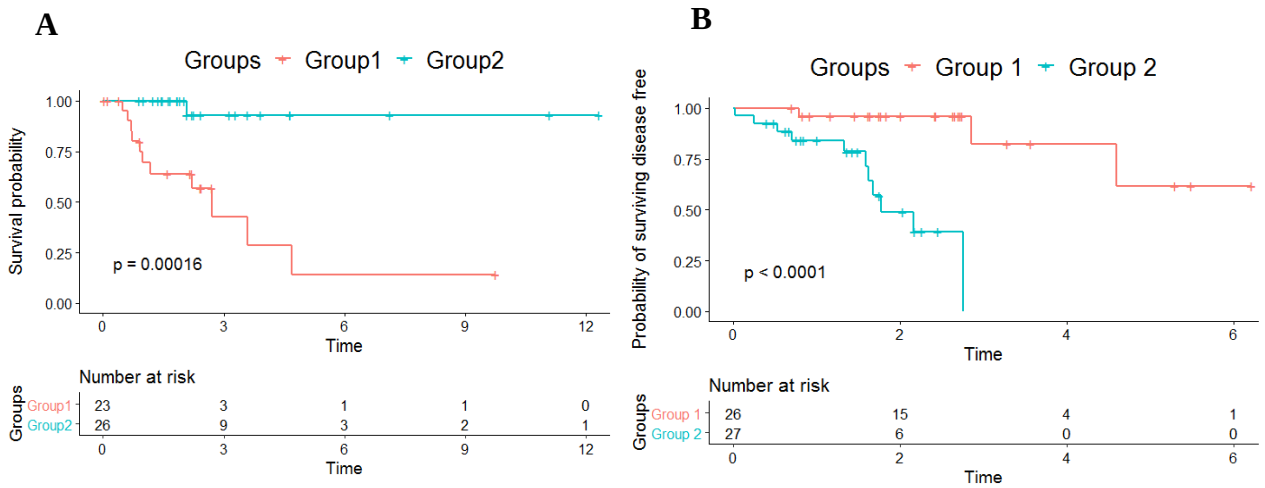

Figure 4. 20 Best CpG panel for prognosis of Stage III patients. (A) Kaplan-Meier overall survival curve based on cg01939477, cg08613350 and cg14944647 combined methylation (p = 0.00016, Logrank test) and (B) Kaplan-Meier recurrence free survival curve based on cg00395657 and cg25571269 collective methylation, (p =0.000063, Logrank test). The respective number of patients at risk for each group at several time points is shown in the table below each graph.

Finally, in stage IV, from a total of 349 differentially methylated CpGs found in this stage, 24 and 31 CpGs exhibited good prognostic values for OS and RFS respectively (Table XII). Patients with higher methylation values of cg05878887, cg10285618 and cg12675571 were tie to better OS values. Better prognostic values associated with higher expression of cg10285618 were however in incongruent with the superior methylation value perceived in tumor patients. Oppositely for the subsisting CpGs higher methylation values for all were accompanied with worst OS prognosis.

Concerning RFS analysis, all CpGs provided a prognosis values for patients with higher methylation. Nonetheless both cg21881253 and cg14901205 were hypermethylation in tumor

patients, and therefore a better prognosis associated with superior methylation values seems incoherent.

Table XII Stage IV differentially methylated CpGs with good prognostic value for both OS and RFS. The respective genes/miRNAs each CpG is affecting, median, β-value, number of patients in each group (group 1 – patients above miRNA median, group 2- patients below the miRNA median), p-value, Hazard ratio (HR) and methylation status of Tumor (T) vs. Normal (N) are presented.

Overall survival analysis

CpGs	genes/miRNAs	median	group 1	group 2	*p-value*	HR	T vs. N Status
cg05878887	C3orf26;FILIP1L;MIR548G	0.2789	25	27	0.0264	2.8991	Hypo
cg10285618	MIR559;EPCAM	0.7434	26	26	0.0302	2.8764	Hyper
cg12675571	LOC728264;MIR145	0.3966	25	27	0.0403	2.5129	Hypo
cg03000711	MIR548A1	0.5738	26	27	0.0484	0.4154	Hypo
cg25512249	MIR518F;MIR520B	0.5664	26	27	0.0383	0.4076	Hypo
cg26024682	MIR670	0.3499	26	27	0.0476	0.4070	Hypo
cg05283542	COL8A1;MIR548G	0.4938	26	27	0.0373	0.3934	Hyper
cg14278808	LOC157627;MIR124-1	0.6700	24	23	0.0415	0.3884	Hyper
cg11638181	MIR129-2	0.4325	26	27	0.0308	0.3825	Hyper
cg17033471	MIR1257	0.6628	26	27	0.0337	0.3774	Hypo
cg13767940	MIR34B;BTG4;MIR34C	0.6042	24	25	0.0325	0.3767	Hyper
cg24480735	MIR521-2	0.4591	24	25	0.0320	0.3743	Hypo
cg01411759	MIR498	0.6085	26	27	0.0331	0.3709	Hypo
cg15248835	LOC157627;MIR124-1	0.4977	26	27	0.0227	0.3657	Hyper
cg24198004	COL8A1;MIR548G	0.4456	25	26	0.0407	0.3634	Hypo
cg11251554	MIR518B	0.4515	26	27	0.0222	0.3505	Hypo
cg14944647	MIR129-2	0.5553	25	24	0.0179	0.3388	Hyper
cg16407471	MIR129-2	0.6250	24	26	0.0196	0.3361	Hyper
cg19267861	MIR124-3	0.6361	25	24	0.0179	0.3264	Hyper
cg09532726	MIR548F5;DCLK1	0.6559	26	27	0.0117	0.3168	Hypo
cg01939477	MIR129-2	0.6315	25	24	0.0196	0.3163	Hyper
cg01273384	MIR1257	0.3347	26	26	0.0110	0.3041	Hypo
cg19973758	C20orf166;MIR133A2	0.4523	26	27	0.0043	0.2635	Hypo
cg22879515	MIR34B;BTG4;MIR34C	0.6273	23	22	0.0055	0.2516	Hyper

Recurrence free survival analysis

CpGs	genes/miRNAs	median	group 1	group 2	*p-value*	HR	T vs. N Status
cg14361526	MIR487B;MIR539	0.4252	21	24	0.0002	8.3642	Hypo
cg16595458	MIR548F3;CNTNAP2	0.5231	24	21	0.0001	7.2126	Hypo
cg03816029	MIR489;CALCR;MIR653	0.5917	23	22	0.0004	6.5106	Hypo
cg15319027	MIR30C2	0.5499	23	22	0.0005	6.2521	Hypo

CpGs	genes/miRNAs	median	group 1	group 2	*p-value*	HR	T vs. N Status
cg01615815	MIR548I4;CNTNAP2	0.5229	23	22	0.0024	5.1648	Hypo
cg22887663	MIR206	0.6948	21	24	0.0023	5.0704	Hypo
cg25952096	MIR1260;NGB	0.6494	23	22	0.0014	5.0672	Hypo
cg12942606	MIR548F5;DCLK1	0.2020	20	24	0.0070	4.1592	Hypo
cg00989505	MIR299;MIR411	0.6542	24	21	0.0049	3.8480	Hypo
cg07191657	MIR1265	0.4841	23	22	0.0064	3.8419	Hypo
cg24428232	MIR921;FAM78B	0.3179	22	21	0.0138	3.7904	Hypo
cg05509179	MIR548A2	0.3707	21	22	0.0108	3.7244	Hypo
cg19098437	MIR30A	0.4698	22	23	0.0139	3.4840	Hypo
cg26995690	DCLK1;MIR548F5	0.3120	24	21	0.0093	3.4758	Hypo
cg21492137	MIR496;MIR154	0.6431	23	22	0.0133	3.4226	Hypo
cg02691745	MIR411	0.6322	23	22	0.0201	3.3541	Hypo
cg14440205	MIR379	0.4689	20	22	0.0322	3.3137	Hypo
cg21182526	MIR548N;OSBPL6	0.5601	21	24	0.0267	3.1169	Hypo
cg07018107	MIR1208	0.5664	21	24	0.0264	3.0452	Hypo
cg21881253	MIR34B;BTG4;MIR34C	0.4340	23	22	0.0322	3.0399	Hyper
cg13600227	MIR521-1	0.5444	21	24	0.0299	3.0162	Hypo
cg09404289	MIR526A1	0.7075	21	24	0.0342	2.9474	Hypo
cg10199730	HTR2C;MIR448	0.7077	24	21	0.0297	2.8587	Hypo
cg25562958	MIR1197;MIR380;MIR323	0.4175	21	20	0.0408	2.8291	Hypo
cg26551791	MIR888;MIR890	0.4172	24	21	0.0475	2.8289	Hypo
cg14901205	MIR129-2	0.6321	24	21	0.0286	2.8001	Hyper
cg00739582	MIR888;MIR890	0.4876	23	21	0.0416	2.7979	Hypo
cg18384960	MIR299;MIR411	0.5454	24	21	0.0360	2.7434	Hypo
cg01823120	MIR548H3	0.7050	23	20	0.0382	2.7156	Hypo
cg04274236	MIR512-1;MIR512-2	0.6559	22	23	0.0452	2.6588	Hypo
cg06159667	H2BFWT;MIR1256	0.5819	22	23	0.0475	2.6211	Hypo

When combining the prognostic capacity of the CpGs at this stage, only 6 combinations of 2 CpGs for OS analysis were identified as statistically significant. The panel of CpGs that provided the best discriminatory value was the combination of cg01273384 and cg26024682 (p-*value = 0.0036*). This combination displayed a HR of 0.1426, meaning that the survival probability at any time point was approximately 7.1 (1/0. 1426) times higher in patients with lower methylation (group 2) than the ones with high methylation levels (group 1) (Figure 4.21).

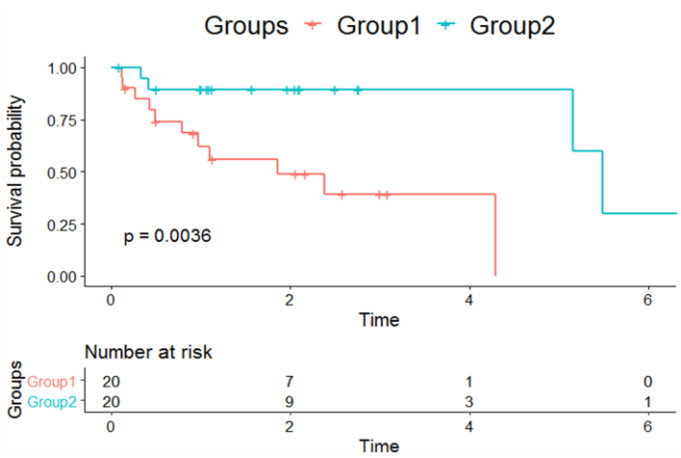

Figure 4. 21 Best CpG panel for prognosis of Stage IV patients. Kaplan-Meier overall survival curve based on cg01273384 and cg26024682 collective methylation (p = 0.0036, Logrank test). The respective number of patients at risk for each group at several time points is shown in the table below each graph.

CHAPTER 5 – Discussion:

In this study, we performed analysis on miRNA expression and methylation on samples of Colon and Rectal cancer patients, available at TCGA dataset, in order to identify novel differentially expressed and methylated miRNAs with potential prognostic and/or diagnostic power.

5. 1 Expression of miRNAs is a valuable tool for diagnosis and prognosis of CRC

In order to understand the pattern of miRNAs deregulation throughout CRC progression we compared miRNAs expression between normal tissue and each of the four different stages of CRC. Our data suggested that similar numbers of miRNAs are being deregulated (~200) at each stage of the disease (Figure 4.1).Interestingly, the vast majority of miRNAs (92%) were found deregulated during the initial Normal to Stage I transformation. In fact, upon transformation, expression level of these deregulated miRNAs was maintained throughout CRC progression and is similar even at the later stage of disease (stage IV). These results are in agreement with previous data from *Pizzini et al* where it was shown that most miRNA expression changes occur during normal mucosa to tumor transformation and are maintained throughout metastatic transition[159]. Interestingly, the pattern of miRNA expression was also observed in the initiation of other tumor types including chronic lymphocytic leukemia[82,160].Nonetheless our approach allowed us to reinforce this alteration in miRNA expression during the initial transition from normal to Stage I tissue in CRC, in an extent never reported before.

Moreover we observed that the vast majority of miRNAs become downregulated between normal to CRC transition, which is in line with previous expression analyzes performed in multiple human cancers including pancreatic, prostate and breast cancers[161]. These results therefore may suggest that the vast majority of miRNAs under normal physiological conditions may act as tumor suppressors.

Although the mechanisms driving global miRNA downregulation in cancer are still unclear, mechanisms such as amplifications, deletions or mutations, transcriptional and epigenetic regulation, seem to contribute to this phenomenon[162,163]. *Sun et al.* recently revealed that activation of the extracellular signal–regulated kinases (ERK) can cause wide spread of miRNA repression by suppressing the major steps of miRNA biogenesis[164]. In this article, the

authors show that in a carcinogenic context, ERK suppresses pre-miRNA export from the nucleus into the cytoplasm through phosphorylation of exportin-5 (EXPO5). This phosphorylation event leads to an alteration in the exportin-5 conformation resulting in the reduction of pre-miRNA loading and consequently to the depletion of miRNA in the cytoplasm[164]. Interestingly, analysis of miRNAs levels in cancer cell lines showed that many pre-miRNAs are retained in the nucleus implying that the function of the nuclear-cytoplasmatic export machinery might be compromised in tumor cells. [165].

Furthermore, we found that the dimension of the differences between the expression values in tumor patients and normal patients are greater for downregulated miRNAs. These results may suggest that genes targeted by the downregulated miRNA are less indulgent/permissive to miRNAs expression alterations, while target genes of upregulated miRNAs might be more sensible to less pronounced alterations. However, one cannot exclude that the higher number of downregulated miRNAs might also contribute to the differences observed. Also, the exclusion criteria (due to the scarcity of information) might have biased the results as we could have excluded upregulated miRNAs with greater differences. Nevertheless, this evidence has been reported in other contexts, thus supporting our finding. For instance, Cheng and colleagues showed significant differences between the range of fold change values for upregulated and downregulated miRNAs in membranous nephropathy[166]. Thus, to the best of our knowledge, this is the first report presenting distinct extends of deregulation regarding downregulated and upregulated miRNAs in CRC.

We then proceeded to profile which specific miRNAs were deregulated at each stage of the disease. Our analysis revealed that 90% of the identified miRNAs were the same in all stages of CRC. Theses result might be suggestive that the miRNAs involved the initial transition from normal to stage I can also be participating in the progression of CRC in later stages of the disease. This analyzes also revealed stage specific miRNAs that could be used as markers for patient stratification. Stage I and IV presented 4 and 7 miRNAs respectively that were exclusive to each stage (Figure 4.5). If on one hand stage I specific miRNAs are able to identify the presence of the disease, on the other stage IV specific miRNAs identify patients with a very poor prognosis. In fact, the transition from stage III to stage IV is defined by the dissemination of disease to other organs (metastasis)[25]. Therefore in stage IV, patients with metastatic CRC are usually conducted to palliative treatment rather than a curative one[167,168]. In this sense this 7

miRNAs may help to better stratify patients and thus facilitate the selection of the most appropriate treatment that will ultimate benefit the patient.

After the initial miRNAs characterization, we next aimed to unveil the regulatory pathways regulated by the downregulated and upregulated miRNAs in order to better understand their role in CRC. In this sense we first obtained their target genes and, interestingly, we found that the total of genes modulated by the upregulated and downregulated panels of miRNAs was of 497 and 360 genes respectively. Also, about 100 genes were simultaneously targeted by down- and upregulated miRNAs. Of notice, in this group were genes known to be associated with the development of CRC such as *TP53, KRAS, APC, PTEN, TGFβ2, SMAD4, Wnt3A and PIK3CA*. We found that miR-605-5p, miR-504-5p, miR-491-5p, miR-1228-3p, miR-125b-1-3p and miR-150-3p seem to play a role in suppressing *p53* expression under normal physiologic conditions and, as they become downregulated in CRC, they might be contributing to the observed higher expression of *p53* in this cancer. Similarly, downregulation of miR-490-3p, miR-23a-5p, and miR-23b-5p can perhaps promote an increase expression of *TGFβ2* in CRC. Moreover upregulation of *WNT3A* on the other hand might be associated with alteration of miR-491-5p. Contrarily upregulation of miR-217, miR-452-5p, miR-1-3p and miR-106b-5p, miR-142-3p might be in some way associated with the downregulation of *KRAS* and *APC*, respectively. Furthermore, upregulation of miR-217, miR-20a-5p, miR-19b-3p, miR-106b-5p, miR-29b-3p, miR-429, miR-124-5p and miR-106b-5p may play a role in the diminished expression of *PTEN*. Finally our results suggest that upregulation of miR-20a-5p, miR-19b-3p, miR-144-5p and miR-1-3p can also be associated with lower expressions of *SMAD4* and *PI3KCA*, respectively.

Nonetheless we cannot fail to mention that the validated functional miRNA-target available in the mirTarbase database derive from a large diversity of experiments in different settings, and thus don`t necessarily represent what occurs in the CRC environment. Therefore, in order to strengthen these associations, validation of these interactions in a CRC context and correlation analysis (e.g. Pearson correlation coefficient) should be performed.

In addition, we show that the deregulated miRNAs interact with a diversity of pathways, in agreement with the distinct target genes we also found. From a total of 530 metabolic pathways provide by KEGG, upregulated and downregulated miRNAs panels interacted with 103 and 110 pathways, respectively. Due to their promiscuous binding to their targets miRNAs

are known to interact with several genes, and thus are able to control a multitude of different pathways. This reality can thus be explaining the large amount of distinctive pathways found in our analysis. In this sense these results may suggest that upregulated and downregulated miRNAs work in the same direction by targeting genes with opposing functions in the same pathway, thus co-contributing to its activation or repression.

Interestingly, many of the pathways found in our analysis were associated with cancer related alterations, such as: "Pathways in cancer", "MicroRNAs in cancer", Transcriptional misregulation in cancer" and most importantly "Colorectal cancer" strongly supporting the idea that alterations in miRNA expression are associated with the development of cancer, including CRC. Furthermore, various pathways usually affected in CRC such as: "PI3K-Akt signaling pathway", "RAS signaling pathway"," TGF-beta signaling pathway", and "p53 signaling pathway" were also found (Figure 4.7 and Table V). However, the involvement of miRNAs in pathways related to CRC progression is not a novelty. In the article by Reid and his colleagues reveal that many of miRNAs deregulated in CRC were computationally mapped to targets involved in pathways related to progression[169].

Moreover, various studies have demonstrated that miRNAs play an important role in negatively regulating the *p53* function through binding to 3′-UTR of the *p53* mRNA[170,171]. In addition a group of several miRNAs such as miR-192, miR-194, miR-215, and miR-605 has oppositely been identified to activate p53 by directly repressing *MDM2*[170,172,173]. Furthermore, upregulation (overexpression) of miR-31 has also been associate with the repression of oncogene RAS by repressing RASA1 in CRC[174]. MiR-21, miR-106a have also been reported to inhibit the expression of TGFBR2 and loss of miR-101 expression as been associated with promotion the activation of the Wnt/β-catenin signalling pathway and malignancy in colon cancer cells[175]. Additionally miR-574-5p has been shown to impact β-catenin/Wnt signalling and the development of colorectal cancer through negative regulation of Qki6/7[176,177]. All together, these findings further support the substantial involvement of miRNAs in the correct regulation of pathways often seen altered in CRC, and that impaired miRNA expression can lead to notorious alterations of such pathways.

Finally, we aimed at scrutinizing the potential use of miRNAs expression in the clinics, and thus we evaluated the diagnostic and prognostic value of the deregulated miRNAs.

We started by investigating putative early stage of disease biomarkers. In this sense we assessed the diagnostic value of the differentially expressed miRNAs in stage I. Interestingly, all the 213 miRNAs differentially expressed in stage I could be classified as good diagnostic biomarkers or better evidencing their capacity of early stage disease biomarkers. Importantly, bibliographic analysis indicated that we have identified 70 novel miRNAs in CRC and from those, 33 miRNAs could discern normal from tumor patients with an accuracy of 100% (sensibility=1 and specificity =1) and thus could be considered perfect biomarkers for an early disease (Stage I) diagnosis[147]. Curiously, supporting our findings, the miRNAs found exclusively deregulated in stage I had already been reported in either colorectal/colon/rectal cancer[178–180].

Although other miRNAs were reported to accurately distinguish early stage of disease, we provided novel biomarkers that surpass by far the sensibilities and specificities of biomarkers used nowadays in a clinical setting such as CEA and CA-19[41,181–183].

CEA is an oncofetal protein found significantly increased in the serum of some patients with adenocarcinoma of the colon[184,185]. This biomarker is widely used for CRC detection, nonetheless serum CEA can be detected in patients with alternative types of carcinoma, severely hindering it's specificity and sensibility for colorectal/colon detection[186]. Similarly CA19-9 is a found in high concentration in colorectal patients' serum[184]. However CA19-9 has been demonstrated to less sensitive than CEA, and has been reported in other GI malignancies as well leading to low values of specificity for colorectal/colon cancer identification[187]. Future studies should test the expression levels of miRNAs here described in the serum/blood of the, in order to sustain our findings. These studies will contribute to develop a potential new panel of early diagnostic CRC biomarkers, which might be used alone or in combination with the current biomarkers employed in the clinic.

Lastly, we evaluated the individual and combinatorial prognostic value of all deregulated miRNAs in each stage of disease for both OS and RFS. We discovered that combinatory panels of 2 to 3 miRNAs could better predict patients' outcome than individual miRNAs, which was clearly evidenced by the hazard ratios found in the Kaplan-Meier curves. Our approach resulted in several panels of miRNAs that were able to distinguish the outcome of patients in stages II, III and IV. Low expression values of hsa-miR-142-3p and hsa-miR-5091 combined were associated with better OS for stage II patients. Similarly lower expression values of hsa-miR-187-3p and hsa-miR-34c-3p were associated with better OS for stage III. No panel of miRNAs was found

statistically significant for stage IV, and thus hsa-miR-3609 may be considered the best OS biomarker.

On the other hand, lower expression values of hsa-miR-142-5p, hsa-miR-142-3p and hsa-miR-642a-5p were related do better RFS in stage II patients while higher expression of hsa-miR-7-1-3p and hsa-miR-543 implied better RFS values for stage III patients. None of the miRNAs was able to accurately discriminate patients in stage IV.

Some incongruence's were found in our analysis, as upregulated miRNAs were found to provide better OS and RFS prognostic values when more expressed in CRC patient samples. Similarly, downregulated miRNAs showed better prognostic in patients with lower expression values. One hypothesis for this might be related to feedback loops that alter the expression of some miRNAs to counteract the biological effects of others miRNAs[188]. Other possibility can be that deregulation of some miRNAs in cancer may function as a compensatory process to oppose the cellular alterations resultant from malignant transformation (in order to restore the normal cell homeostasis), and therefore despite being observed in cancer, those regulations may imply a better prognosis.

Nevertheless, here we were able to identify different panels of miRNAs that are able to help improve CRC screenings as well as to potentially impact on clinical decisions in a near future.

5.2 Methylation of miRNA genes are potential epigenetic biomarkers for CRC management

DNA methylation alterations are well described as important events for CRC progression[189,190]. Here, we explored methylation of miRNAs patterns during CRC initiation and progression and investigated their value as epigenetic biomarkers in this cancer.

We found that miRNA methylation, similarly to mRNA expression, suffers a significant alteration during the initial transition from normal to stage I tumorigenesis in CRC. However, after this initial transition (were alterations in miRNA methylation reached their peak) alterations in miRNA methylation severely decreased in stage two. Nonetheless, we also observed an increase of miRNA methylation in later stages of disease (through stages III and IV) the alterations in miRNA methylation. Together these results propose that alterations in the miRNA methylation pattern might play a more significant role in an early phase of CRC than in later

stages of disease. Although little is known regarding genome-wide miRNA methylation in CRC, involvement of DNA methylation in miRNA expression alterations in early stages of CRC is described. Balaguer and colleagues demonstrated that miR-137 silencing in colorectal adenomas by promoter hypermethylation is an early event in CRC carcinogenesis[191]. Additionally *Suzuki et al.* reported the methylation of miR1-1 in approximately 80% of primary CRC tissue samples and in 70% of colorectal adenoma tissue samples, suggesting that methylation of miR1-1 is an early event in colorectal tumorignesis[192]. Moreover miRNAs silencing as an early event in tumorigenesis through promoter hypermethylation has been reported in several other cancers[192]. MiRNA-148a silencing by hypermethylaton has been described in pancreatic cancer and preneoplastic pancreatic lesions, suggesting it is an early event in pancreatic carcinogenesis[193].

Contrarily to those studies, we found that about 70% of the CpGs affecting miRNAs were lowly methylated in tumor tissue. Under normal physiologic conditions the genome is characterized by a general methylation, of the CpGs scatter throughout the genome while the Cpgs found in promoter regions are unmethylated[20,53]. However during carcinogenesis a major event of hypomethylation of the CpGs dispersed along the genome is observed[20,53]. In this sense we hypothesized was that that the downmethylation event observed in our analysis could be due to this phenomenon. Strikingly, when analysing the location of the CpGs we observed that the vast majority (85%) of all CpGs were located in the promoter region of miRNAs genes. Although this methylation pattern is not well reported, a recent study described similar findings. When comparing methylation of human femoral atherosclerotic plaques cases with healthy mammary arteries, the authors found an hypomethylation of the promoter region in almost 84% of the cases[194].

DNA methylation of promoter regions is commonly linked to gene silencing[192]. However, we now know that it can also lead to gene activation. This can be explained if the CpGs lay in or near repressors binding sites, where DNA methylation would create a physical barrier thus blocking repressors biding and enabling transcription. However, this blocking by selective methylation seems to be limited to some genes and not a general regulatory mechanism189. Thus, the hypothesis that the observed downregulation in miRNA expression is mainly linked to the lower methylation event found in this study is unlikely, and one must consider other mechanisms underlying the silencing of miRNAs in CRC. Nonetheless, further studies should accurately understand the impact of miRNA methylation on miRNA expression.

Nonetheless our results suggest that, despite mainly located in promoter regions, the vast majority of CpGs become downmethylated during tumorigenesis which seems to contradict the literature.

Nevertheless a recent study has highlighted new findings on DNA methylation patterns[194]. While studying human femoral atherosclerotic plaques compared with healthy mammary arteries, Aavik and colleagues reported that changes in methylation status are a frequent phenomenon in atherosclerotic plaques. However, most significantly when analysing several genes differentially methylated, they found that hypomethylation of the promoter region was found in almost 84% of the cases[194]. Despite this study being performed outside the cancer filed, and in genes, instead of miRNAs the results obtained here are cohesive with rour results.

Nevertheless in this article Aavik and colleagues further mention that an incrise in mRNA transcription was observed as a result of this hpomethyaltion event. In this sense this results cannot explain why the vast majority of miRNAs is found downmethylated in cancer. Thus, we provide here some hypothesis that can explain this phenomenon.

Our first hypothesis is that miRNA methylation despite influencing miRNA expression, account only for a small amount of miRNA expression changes, being miRNAs expression higly submissive to other regulation events. However this is just a speculation, as we could not find any article to corroborate this hypothesis.

Another hypothesis is that promoter hypermethylation of miRNAs can be potentiating in some way the transcription of miRNAs in normal cells and therefore a decrease in promoter methylation can lead to a decrease miRNA expression. In fact in a study by Medvedev and her colleagues in which the effects of cytosine methylation on transcription binding sites were assessed, the authors affirm: “These observations allow us to suggest that blocking of TFBSs by selective methylation is unlikely to be a general mechanism of methylation-dependent transcription regulation and that such a mechanism is limited to special cases”[195]. These evidences thus stress the possibility of promoter methylation in the enhancment of miRNA expression, probably by preventing the binding of transcription repressor factors.

Nonetheless further studies need to be performed in order to accurately understand the process of miRNA on miRNA expression.

Moreover, information about the diagnostic and prognostic value of miRNA methylation is still lacking for CRC. In order to get some insights into these clinical features we first assessed the value of miRNA methylation as early disease (Stage I) biomarkers in CRC. We found that 34 CpGs (8%) were able to accurately discriminate stage I samples from normal tissue with 100% certainty indicating that miRNA methylation can be good diagnostic tools.

Also, when evaluating the individual and combinatorial prognostic value of all deregulated CpGs we discovered that combinatory panels of 2 to 3 CpGs could better predict patients' outcome than individual CpG sites. In fact, stages II, III and IV the hazard ratios of the Kaplan-Meier curves were consistently higher for combinatory CpGs panels than for individual CpGs, for both OS and RFS. Our approach resulted in several panels of CpGs that were able to distinguish the outcome of patients in stages II, III and IV. Lower methylation of cg20474788 and cg24041078 combined were associated with better OS for stage II patients. Similarly lower methylation values of cg01939477, cg08613350 and cg1494464 and lower methylation values of cg01273384 and cg26024682 were considered the best panel for OS in stages III and IV, respectively.

On the other hand lower methylation values of cg01963147, cg02336334 and cg2602468 and were related do better RFS in stage II. Oppositely, higher methylation values of cg00395657 and cg25571269 were related do better RFS in stage III patients. No panel of CpGs was found statistically significant for stage IV, and thus cg09852439 was considered the best RFS biomarker. Some incongruence's were once again found in our analysis, as hypoermethylated CpGs were found to provide better OS and RFS prognostic values when the methylation values were higher in CRC patient samples and hypomethylated CpGs showed better prognosis in patients with lower methylation values. Future studies might help understand the significance of these contradictions.

Our results evidenced that not only miRNA expression is massively repressed in CRC initiation but also that this is accompanied by a lower methylation status of miRNA genes. More importantly, these alterations can distinguish normal from malignant tissue in early stage disease and further predict patients' survival and risk of recurrence. These biomarkers should increase the power of current biomarkers and help clinicians to detect and treat CRC patients.

5.3 Limitations of our study

Nonetheless, our study presents some limitations including:

- Normal patients samples used in this analysis weren't obtained from normal patients but rather from adjacent tissues of CRC patients which might already comprise abnormalities and thus affect the viability of our results.
- The clinical information available for patients in TCGA could be more complete as some relevant information is missing.
- The high percentage of information missing especially in the miRNA mature strand expression RNAseq – IlluminaHiseq dataset
- The lack of normal patients in the in the miRNA mature strand expression RNAseq – IlluminaHiseq dataset which is the major limitation in our miRNA mature expression analyzes
- The use of tissues from the colon and rectum instead of body fluids or feces in our analyzes. The ideal frameworks would be to have perform the analyzes here described in blood, serum or feces samples which can be collected in a non-invasive manner

CHAPTER 6 – Conclusion:

This study provides an overview on miRNAs expression and methylation behaviour throughout CRC initiation and progression. Here we present a description of miRNA alterations during the carcinogenesis of CRC in an extent never described before and provide novel biomarkers for early detection and patient outcome.

In this work we have identified miRNAs differently expressed in CRC tumor patient samples when compared to normal tissue samples. Our results suggest that miRNA alterations are an early event in tumorigenesis and therefore may play a significant role in disease initiation. Furthermore our biological analysis suggested that miRNAs might control several pathways usually altered in CRC.

Also, we showed that miRNAs are able to accurately distinguish normal from early stage CRC patients proving to be very good diagnostic biomarkers. Moreover to the best of our knowledge, we have identified 70 novel biomarkers for early detection of CRC never previously referenced in any PubMed article as associated with CRC. Moreover differentially expressed miRNAs could predict patients' outcome evidencing their prognostic values for both OS and RFS.

Simultaneously, a parallel analysis of large-scale miRNA methylation was performed in order to understand the alteration patterns perceived during CRC initiation and progression. We demonstrate that miRNAs methylation is also an early event in CRC and may have a significant impact on disease initiation. Moreover we revealed that miRNA methylation could not only distinguish normal from stage I CRC patients but also disassociate CRC patients based on their outcome.

In conclusion, our study evidences the potential of miRNA alterations as CRC diagnostic and prognostic biomarkers. Therefore, this work may provide a basis for further analyzes on miRNA potential as epigenetic biomarkers of CRC. Future studies should verify if the miRNAs alterations here described (or at least some) are maintained in feces or body fluids, which would allow establishing novel non–invasive biomarkers for CRC management.

Bibliography

1. Sudhakar A. History of Cancer, Ancient and Modern Treatment Methods. *J Cancer Sci Ther*. 2009;1(2):1-4. doi:10.4172/1948-5956.100000e2
2. Hanahan D, Weinberg R a. The hallmarks of cancer. *Cell*. 2000;100(1):57-70. http://www.ncbi.nlm.nih.gov/pubmed/10647931.
3. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. *Cell*. 2011;144(5):646-674. doi:10.1016/j.cell.2011.02.013
4. Tomasetti C, Vogelstein B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*. 2015;347(6217):78-81. doi:10.1126/science.1260825
5. Cooper GM. The Development and Causes of Cancer. 2000. https://www.ncbi.nlm.nih.gov/books/NBK9963/. Accessed May 21, 2018.
6. Garraway LA, Lander ES. Lessons from the cancer genome. *Cell*. 2013;153(1):17-37. doi:10.1016/j.cell.2013.03.002
7. Rahman N. Realizing the promise of cancer predisposition genes. *Nature*. 2014;505(7483):302-308. doi:10.1038/nature12981
8. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*. 2015;136(5):E359-E386. doi:10.1002/ijc.29210
9. American Cancer Society. Colorectal Cancer Facts & Figures 2017 - 2019. *Atlanta*. 2017:1-40. doi:http://dx.doi.org/10.1016/S0140-6736(13)61649-9
10. Das V, Kalita J, Pal M. Predictive and prognostic biomarkers in colorectal cancer: A systematic review of recent advances and challenges. *Biomed Pharmacother*. 2017;87:8-19. doi:10.1016/j.biopha.2016.12.064
11. Tariq K, Ghias K. Colorectal cancer carcinogenesis: a review of mechanisms. *Cancer Biol Med*. 2016;13(1):120-135. doi:10.28092/j.issn.2095-3941.2015.0103
12. Winawer SJ, Zauber AG. The advanced adenoma as the primary target of screening. *Gastrointest Endosc Clin N Am*. 2002;12(1):1-9. doi:10.1016/S1052-5157(03)00053-9
13. Levine JS, Ahnen DJ. Clinical practice. Adenomatous polyps of the colon. *N Engl J Med*. 2006;355(24):2551-2557. doi:10.1056/NEJMcp063038
14. Huo T, Canepa R, Sura A, Modave F, Gong Y. Colorectal cancer stages transcriptome analysis. *PLoS One*. 2017;12(11):1-11. doi:10.1371/journal.pone.0188697
15. De La Chapelle A. Genetic predisposition to colorectal cancer. *Nat Rev Cancer*. 2004;4(10):769-780. doi:10.1038/nrc1453
16. Haggar F a, Boushey RP, Ph D. Colorectal Cancer Epidemiology : Incidence , Mortality , Survival , and Risk Factors. *Clin Colon Rectal Surg*. 2009;22(4):191-197. doi:10.1055/s-0029-1242458.
17. Binefa G, Rodríguez-Moranta F, Teule À, Medina-Hayas M. Colorectal cancer: From prevention to personalized medicine. *World J Gastroenterol*. 2014;20(22):6786-6808. doi:10.3748/wjg.v20.i22.6786
18. Bardhan K, Liu K. Epigenetics and colorectal cancer pathogenesis. *Cancers (Basel)*. 2013;5(2):676-713. doi:10.3390/cancers5020676
19. Fearon EF, Vogelstein B. for Colorectal Tumorigenesis. *Cell*. 1989;61(1):759-767.
20. Khare S, Verma M. Epigenetics of colon cancer. *Methods Mol Biol*. 2012;863(12):177-185. doi:10.1007/978-1-61779-612-8_10

21. Kinzler KW, Vogelstein B. Lessons from hereditary colorectal cancer. *Cell*. 1996;87(2):159-170. doi:10.1016/S0092-8674(00)81333-1
22. Kondo Y, Issa J-PJ. Epigenetic changes in colorectal cancer. *Cancer Metastasis Rev*. 2004;23(1/2):29-39. doi:10.1023/A:1025806911782
23. Li J, Yi CH, Hu YT, et al. TNM Staging of Colorectal Cancer Should be Reconsidered According to Weighting of the T Stage. *Med (United States)*. 2016;95(6):1-8. doi:10.1097/MD.0000000000002711
24. Brierley JD, Gospodarowicz MK, Wittekind C. TNM classification of malignant tumours - 8th edition. *Union Int Cancer Control*. 2017:241. doi:10.1002/ejoc.201200111
25. Brenner H, Kloor M, Pox CP. Colorectal cancer. *Lancet*. 2014;383(9927):1490-1502. doi:10.1016/S0140-6736(13)61649-9
26. Levin B, Lieberman DA, McFarland B, et al. Screening and Surveillance for the Early Detection of Colorectal Cancer and Adenomatous Polyps, 2008: A Joint Guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *CA Cancer J Clin*. 2008;58(3):130-160. doi:10.3322/CA.2007.0018
27. Andrew R Marley HN, Hansen RD. Molecular epidemiology of colorectal cancer. *Mol ecular Epidemiol ogy Genet*. 2016;7(3):105-114. doi:10.1093/bmb/64.1.1
28. Jha P, Wang X, Auwerx J. Analysis of Mitochondrial Respiratory Chain Supercomplexes Using Blue Native Polyacrylamide Gel Electrophoresis (BN-PAGE). *Curr Protoc Mouse Biol*. 2016;6(1):1-14. doi:10.1002/9780470942390.mo150182
29. Elsafi SH, Alqahtani NI, Zakary NY, Al Zahrani EM. The sensitivity, specificity, predictive values, and likelihood ratios of fecal occult blood test for the detection of colorectal cancer in hospital settings. *Clin Exp Gastroenterol*. 2015;8:279-284. doi:10.2147/CEG.S86419
30. Okugawa Y, Grady WM, Goel A. Epigenetic Alterations in Colorectal Cancer: Emerging Biomarkers. *Gastroenterology*. 2015;149(5):1204-1225.e12. doi:10.1053/j.gastro.2015.07.011
31. Young PE, Womeldorph CM. Colonoscopy for Colorectal Cancer Screening. *J Cancer*. 2013;4(3):217-226. doi:10.7150/jca.5829
32. Strul H, Arber N. Screening techniques for prevention and early detection of colorectal cancer in the average-risk population. *Gastrointest Cancer Res*. 2007;1(3):98-106. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2632817&tool=pmcentrez&rendertype=abstract.
33. U.S. Preventive Services Task Force. Screening for colorectal cancer: recommendation and rationale. *Ann Intern Med*. 2002;137(2):129-131. http://www.ncbi.nlm.nih.gov/pubmed/12118971.
34. Simon K. Colorectal cancer development and advances in screening. *Clin Interv Aging*. 2016;11:967-976. doi:10.2147/CIA.S109285
35. Qasim BJ, Al-Wasiti EA, Azzal HS. Association of global DNA hypomethylation with clinicopathological variables in colonic tumors of Iraqi patients. *Saudi J Gastroenterol*. 2016;22(2):139-147. doi:10.4103/1319-3767.178525
36. Zamani M, Hosseini SV, Mokarram P. Epigenetic biomarkers in colorectal cancer: premises and prospects. *Biomarkers*. 2018;23(2):105-114. doi:10.1080/1354750X.2016.1252961
37. Lech G, Słotwiński R, Słodkowski M, Krasnodębski IW. Colorectal cancer tumour

markers and biomarkers: Recent therapeutic advances. *World J Gastroenterol*. 2016;22(5):1745-1755. doi:10.3748/wjg.v22.i5.1745
38. Mishra A, Verma M. Cancer biomarkers: Are we ready for the prime time? *Cancers (Basel)*. 2010;2(1):190-208. doi:10.3390/cancers2010190
39. Henry NL, Hayes DF. Cancer biomarkers. *Mol Oncol*. 2012;6(2):140-146. doi:10.1016/j.molonc.2012.01.010
40. Bhatt AN, Mathur R, Farooque A, Verma A, Dwarakanath BS. Cancer biomarkers - current perspectives. *Indian J Med Res*. 2010;132(August):129-149. http://www.ncbi.nlm.nih.gov/pubmed/20716813.
41. Article R, Access O. Biomarkers in Colorectal Cancer Screening. *J Gastrointest Dig Syst*. 2016;6(1):2-9. doi:10.4172/2161-069X.Page
42. Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. *Carcinogenesis*. 2009;31(1):27-36. doi:10.1093/carcin/bgp220
43. Shen L, Toyota M, Kondo Y, et al. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proc Natl Acad Sci*. 2007;104(47):18654-18659. doi:10.1073/pnas.0704652104
44. Castells A. Choosing the optimal method in programmatic colorectal cancer screening: Current evidence and controversies. *Therap Adv Gastroenterol*. 2015;8(4):221-233. doi:10.1177/1756283X15578610
45. Bishnupuri KS, Mishra MK. Epigenetics of colorectal cancer. *Epigenetic Adv Cancer*. 2016:97-121. doi:10.1007/978-3-319-24951-3_5
46. Fedoriw A, Mugford J, Magnuson T. Genomic imprinting and epigenetic control of development. *Cold Spring Harb Perspect Biol*. 2012;4(7):1-15. doi:10.1101/cshperspect.a008136
47. GRØNBAEK K, HOTHER C, JONES PA. Epigenetic changes in cancer. *Apmis*. 2007;115(10):1039-1059. doi:10.1111/j.1600-0463.2007.apm_636.xml.x
48. Choong MK, Tsafnat G. Genetic and epigenetic biomarkers of colorectal cancer. *Clin Gastroenterol Hepatol*. 2012;10(1):9-15. doi:10.1016/j.cgh.2011.04.020
49. Vaiopoulos AG, Athanasoula KC, Papavassiliou AG. Epigenetic modifications in colorectal cancer: Molecular insights and therapeutic challenges. *Biochim Biophys Acta - Mol Basis Dis*. 2014;1842(7):971-980. doi:10.1016/j.bbadis.2014.02.006
50. Babashah S. MicroRNAs: Key regulators of oncogenesis. *MicroRNAs Key Regul Oncog*. 2013:1-433. doi:10.1007/978-3-319-03725-7
51. Dawson MA, Kouzarides T. Cancer epigenetics: From mechanism to therapy. *Cell*. 2012;150(1):12-27. doi:10.1016/j.cell.2012.06.013
52. Cohen I, Poręba E, Kamieniarz K, Schneider R. Histone modifiers in cancer: Friends or foes? *Genes and Cancer*. 2011;2(6):631-647. doi:10.1177/1947601911417176
53. Baylin SB, Jones P a. A decade of exploring the cancer epigenome - biological and translational implications. *Nat Rev Cancer*. 2011;11(10):726-734. doi:10.1038/nrc3130
54. Wajed SA, Laird PW, DeMeester TR. DNA methylation: An alternative pathway to cancer. *Ann Surg*. 2001;234(1):10-20. doi:10.1097/00000658-200107000-00003
55. Bird A. DNA methylation patterns and epigenetic memory. *Genes Dev*. 2002;16(1):6-21. doi:10.1101/gad.947102
56. Ashktorab H, Brim H. DNA Methylation and Colorectal Cancer. *Curr Colorectal Cancer Rep*. 2014;10(4):425-430. doi:10.1007/s11888-014-0245-2
57. Shahrouki P, Larsson E. The non-coding oncogene: A case of missing DNA evidence?

Front Genet. 2012;3(SEP):1-8. doi:10.3389/fgene.2012.00170
58. Costa FF. Non-coding RNAs: New players in eukaryotic biology. *Gene*. 2005;357(2):83-94. doi:10.1016/j.gene.2005.06.019
59. Pauli A, Rinn JL, Schier AF. Non-coding RNAs as regulators of embryogenesis. *Nat Rev Genet*. 2011;12(2):136-149. doi:10.1038/nrg2904
60. He L, Hannon GJ. MicroRNAs: Small RNAs with a big role in gene regulation. *Nat Rev Genet*. 2004;5(7):522-531. doi:10.1038/nrg1379
61. Cai X, Hagedorn CH, Cullen BR. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *Rna*. 2004;10(12):1957-1966. doi:10.1261/rna.7135204
62. Han J, Lee Y, Yeom K, Kim Y, Jin H, Kim VN. The Drosha – DGCR8 complex in primary microRNA processing. *Genes Dev*. 2004:3016-3027. doi:10.1101/gad.1262504.mic
63. Lund E, Güttinger S, Calado A, Dahlberg JE, Kutay U. Nuclear export of microRNA precursors. *Science*. 2004;303(5654):95-98. doi:10.1126/science.1090599
64. Bartel DP. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*. 2004;116(2):281-297. doi:10.1016/S0092-8674(04)00045-5
65. MacRae IJ, Zhou K, Li F, et al. Structural basis for double-stranded RNA processing by Dicer. *Science (80-)*. 2006;311(5758):195-198. doi:10.1126/science.1121638
66. Chendrimada TP, Gregory RI, Kumaraswamy E, Cooch N, Nishikura K, Shiekhattar R. NIH Public Access. 2010;436(7051):740-744. doi:10.1038/nature03868.TRBP
67. Wahid F, Shehzad A, Khan T, Kim YY. MicroRNAs: Synthesis, mechanism, function, and recent clinical trials. *Biochim Biophys Acta - Mol Cell Res*. 2010;1803(11):1231-1243. doi:10.1016/j.bbamcr.2010.06.013
68. Schwarz DS, Hutvágner G, Du T, Xu Z, Aronin N, Zamore PD. Asymmetry in the assembly of the RNAi enzyme complex. *Cell*. 2003;115(2):199-208. doi:10.1016/S0092-8674(03)00759-1
69. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: Are the answers in sight? *Nat Rev Genet*. 2008;9(2):102-114. doi:10.1038/nrg2290
70. Béthune J, Artus-Revel CG, Filipowicz W. Kinetic analysis reveals successive steps leading to miRNA-mediated silencing in mammalian cells. *EMBO Rep*. 2012;13(8):716-723. doi:10.1038/embor.2012.82
71. Pillai RS. MicroRNA function : Multiple mechanisms for a tiny RNA ? *Rna*. 2005;11(Bartel 2004):1753-1761. doi:10.1261/rna.2248605.that
72. Eulalio A, Huntzinger E, Izaurralde E. Getting to the Root of miRNA-Mediated Gene Silencing. *Cell*. 2008;132(1):9-14. doi:10.1016/j.cell.2007.12.024
73. Wazen RM, Kuroda S, Nishio C, Sellin K, Brunski JB, Nanci A. NIH Public Access. 2014;8(9):1385-1395. doi:10.2217/nnm.12.167.Gene
74. Griffiths-Jones S, Saini HK, Van Dongen S, Enright AJ. miRBase: Tools for microRNA genomics. *Nucleic Acids Res*. 2008;36(SUPPL. 1):154-158. doi:10.1093/nar/gkm952
75. Baskerville S, Bartel DP. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *Rna*. 2005;11(3):241-247. doi:10.1261/rna.7240905
76. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. Identification of novel genes coding for small expressed RNAs. *Science*. 2001;294(5543):853-858.

doi:10.1126/science.1064921
77. Olena AF, Patton JG. NIH Public Access. 2014;222(3):540-545. doi:10.1002/jcp.21993.Genomic
78. Budak H, Bulut R, Kantar M, Alptekin B. MicroRNA nomenclature and the need for a revised naming prescription. *Brief Funct Genomics*. 2016;15(1):65-71. doi:10.1093/bfgp/elv026
79. Zhang NS, Dai GL, Liu SJ. MicroRNA-29 family functions as a tumor suppressor by targeting RPS15A and regulating cell cycle in hepatocellular carcinoma. *Int J Clin Exp Pathol*. 2017;10(7):8031-8042.
80. Jiang H, Zhang G, Wu JH, Jiang CP. Diverse roles of miR-29 in cancer (Review). *Oncol Rep*. 2014;31(4):1509-1516. doi:10.3892/or.2014.3036
81. Hummel R, Hussey DJ, Haier J. MicroRNAs: Predictors and modifiers of chemo- and radiotherapy in different tumour types. *Eur J Cancer*. 2010;46(2):298-311. doi:10.1016/j.ejca.2009.10.027
82. Calin GA, Dumitru CD, Shimizu M, et al. Nonlinear partial differential equations and applications: Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci*. 2002;99(24):15524-15529. doi:10.1073/pnas.242606799
83. Li M, Marin-Muller C, Bharadwaj U, Chow K-H, Yao Q, Chen C. MicroRNAs: control and loss of control in human physiology and disease. *World J Surg*. 2009;33(4):667-684. doi:10.1007/s00268-008-9836-x
84. Paranjape T, Slack FJ, Weidhaas JB. MicroRNAs: Tools for cancer diagnostics. *Gut*. 2009;58(11):1546-1554. doi:10.1136/gut.2009.179531
85. Cortez MA, Bueso-Ramos C, Ferdin J, Lopez-Berestein G, Sood AK, Calin GA. MicroRNAs in body fluids--the mix of hormones and biomarkers. *Nat Rev Clin Oncol*. 2011;8(8):467-477. doi:10.1038/nrclinonc.2011.76
86. Saito Y, Jones PM. Epigenetic Activation of Tumor Suppressor MicroRNAs in Human Cancer Cells. *Cell Cycle*. 2006;5(19):2220-2222. doi:10.4161/cc.5.19.3340
87. Corcoran DL, Pandit K V., Gordon B, Bhattacharjee A, Kaminski N, Benos P V. Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data. *PLoS One*. 2009;4(4):1-10. doi:10.1371/journal.pone.0005279
88. Ozsolak F, Poling LL, Wang Z, et al. Chromatin structure analyses identify miRNA promoters. *Genes Dev*. 2008;22(22):3172-3183. doi:10.1101/gad.1706508
89. Bandres E, Agirre X, Bitarte N, et al. Epigenetic regulation of microRNA expression in colorectal cancer. *Int J Cancer*. 2009;125(11):2737-2743. doi:10.1002/ijc.24638
90. Ahlquist DA, Sargent DJ, Loprinzi CL, et al. Stool DNA and occult blood testing for screen detection of colorectal neoplasia. *Ann Intern Med*. 2008;149(7):441-450. doi:10.7326/0003-4819-149-7-200810070-00004
91. Estey MP, Di Ciano-Oliveira C, Froese CD, et al. Mitotic regulation of SEPT9 protein by cyclin-dependent kinase 1 (Cdk1) and pin1 protein is important for the completion of cytokinesis. *J Biol Chem*. 2013;288(42):30075-30086. doi:10.1074/jbc.M113.474932
92. Warren JD, Xiong W, Bunker AM, et al. Septin 9 methylated DNA is a sensitive and specific blood test for colorectal cancer. *BMC Med*. 2011;9. doi:10.1186/1741-7015-9-133
93. Ng J, Yu J. Promoter Hypermethylation of Tumour Suppressor Genes as Potential Biomarkers in Colorectal Cancer. *Int J Mol Sci*. 2015;16(2):2472-2496. doi:10.3390/ijms16022472

94. Itzkowitz S, Brand R, Jandorf L, et al. A simplified, noninvasive Stool DNA test for colorectal cancer detection. *Am J Gastroenterol*. 2008;103(11):2862-2870. doi:10.1111/j.1572-0241.2008.02088.x
95. Toiyama Y, Yasuda H, Saigusa S, et al. Increased expression of slug and vimentin as novel predictive biomarkers for lymph node metastasis and poor prognosis in colorectal cancer. *Carcinogenesis*. 2013;34(11):2548-2557. doi:10.1093/carcin/bgt282
96. Pino MS, Kikuchi H, Zeng M, et al. Epithelial to Mesenchymal Transition Is Impaired in Colon Cancer Cells With Microsatellite Instability. *Gastroenterology*. 2010;138(4):1406-1417. doi:10.1053/j.gastro.2009.12.010
97. Ng EKO, Chong WWS, Jin H, et al. Differential expression of microRNAs in plasma of patients with colorectal cancer: a potential marker for colorectal cancer screening. *Gut*. 2009;58(10):1375-1381. doi:10.1136/gut.2008.167817
98. Koga Y, Yasunaga M, Takahashi A, et al. MicroRNA expression profiling of exfoliated colonocytes isolated from feces for colorectal cancer screening. *Cancer Prev Res*. 2010;3(11):1435-1442. doi:10.1158/1940-6207.CAPR-10-0036
99. Schetter AJ, Leung SY, Sohn JJ, et al. MicroRNA expression profiles associated with prognosis and therapeutic outcome in colon adenocarcinoma. *JAMA*. 2008;299(4):425-436. doi:10.1001/jama.299.4.425
100. Schnekenburger M, Diederich M. Epigenetics offer new horizons for colorectal cancer prevention. *Curr Colorectal Cancer Rep*. 2012;8(1):66-81. doi:10.1007/s11888-011-0116-z
101. Shirafkan N, Mansoori B, Mohammadi A, Shomali N, Ghasbi M, Baradaran B. MicroRNAs as novel biomarkers for colorectal cancer: New outlooks. *Biomed Pharmacother*. 2018;97(November 2017):1319-1330. doi:10.1016/j.biopha.2017.11.046
102. Wang F, Ma Y-L, Zhang P, et al. SP1 mediates the link between methylation of the tumour suppressor miR-149 and outcome in colorectal cancer. *J Pathol*. 2013;229(1):12-24. doi:10.1002/path.4078
103. Gyparaki MT, Basdra EK, Papavassiliou AG. DNA methylation biomarkers as diagnostic and prognostic tools in colorectal cancer. *J Mol Med*. 2013;91(11):1249-1256. doi:10.1007/s00109-013-1088-z
104. Almeida MI, Reis RM, Calin GA. MicroRNA history: Discovery, recent applications, and next frontiers. *Mutat Res - Fundam Mol Mech Mutagen*. 2011;717(1-2):1-8. doi:10.1016/j.mrfmmm.2011.03.009
105. Nair VS, Pritchard CC, Tewari M, Ioannidis JPA. Design and analysis for studying microRNAs in human disease: A primer on-omic technologies. *Am J Epidemiol*. 2014;180(2):140-152. doi:10.1093/aje/kwu135
106. Axtell MJ, Meyers BC. Revisiting criteria for plant miRNA annotation in the era of big data. *Plant Cell*. 2018;30(February):tpc.00851.2017. doi:10.1105/tpc.17.00851
107. Manuscript A. NIH Public Access. *J Genet*. 2011;38(3):95-109. doi:10.1016/j.jgg.2011.02.003.The
108. Cohen MM, Emanuel BS. Expressed sequence tags. *Science*. 1994;266(5192):1790-1791. doi:10.1038/nrg2484.RNA-Seq
109. Buermans HPJ, den Dunnen JT. Next generation sequencing technology: Advances and applications. *Biochim Biophys Acta - Mol Basis Dis*. 2014;1842(10):1932-1941. doi:10.1016/j.bbadis.2014.06.015
110. Montgomery SB. RNA Sequencing and Analysis. *Cold Spring Harb Protoc*.

2016;2015(11):951-969. doi:10.1101/pdb.top084970.RNA
111. Ansorge WJ. Next-generation DNA sequencing techniques. *N Biotechnol*. 2009;25(4):195-203. doi:10.1016/j.nbt.2008.12.009
112. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456(7218):53-59. doi:10.1038/nature07517
113. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol*. 2011;12(11):R112. doi:10.1186/gb-2011-12-11-r112
114. Lutzmayer S, Enugutti B, Nodine MD. Novel small RNA spike-in oligonucleotides enable absolute normalization of small RNA-Seq data. *Sci Rep*. 2017;7(1):1-6. doi:10.1038/s41598-017-06174-3
115. Meyer SU, Pfaffl MW, Ulbrich SE. Normalization strategies for microRNA profiling experiments: A “normal” way to a hidden layer of complexity? *Biotechnol Lett*. 2010;32(12):1777-1788. doi:10.1007/s10529-010-0380-z
116. Sugawara E, Nikaido H. Properties of AdeABC and AdeIJK efflux systems of Acinetobacter baumannii compared with those of the AcrAB-TolC system of Escherichia coli. *Antimicrob Agents Chemother*. 2014;58(12):7250-7257. doi:10.1128/AAC.03728-14
117. Dedeurwaerder S, Defrance M, Bizet M, Calonne E, Bontempi G, Fuks F. A comprehensive overview of Infinium Human Methylation450 data processing. *Brief Bioinform*. 2013;15(6):929-941. doi:10.1093/bib/bbt054
118. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F. Evaluation of the Infinium Methylation 450K technology. *Epigenomics*. 2011;3(6):771-784. doi:10.2217/epi.11.105
119. Weisenberger DJ, Berg D Van Den, Pan F, et al. Comprehensive DNA Methylation Assay Platform. *Appl Note Illumina Epigenetic Anal*. 2008.
120. Li Y, Kowdley K V. MicroRNAs in Common Human Diseases. *Genomics, Proteomics Bioinforma*. 2012;10(5):295-301. doi:10.1016/j.gpb.2012.07.005
121. Emery JD. The challenges of early diagnosis of cancer in general practice. *Med J Aust*. 2015;203(10):391-393.e1. doi:10.5694/mja15.00527
122. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Wspolczesna Onkol*. 2015;1A:A68-A77. doi:10.5114/wo.2014.47136
123. Colaprico A, Silva TC, Olsen C, et al. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*. 2016;44(8):e71. doi:10.1093/nar/gkv1507
124. Kang H. The prevention and handling of the missing data. *Korean J Anesthesiol*. 2013;64(5):402-406. doi:10.4097/kjae.2013.64.5.402
125. Kwak SK, Kim JH. Statistical data preparation: Management of missing values and outliers. *Korean J Anesthesiol*. 2017;70(4):407-411. doi:10.4097/kjae.2017.70.4.407
126. Aguinis H, Gottfredson RK, Joo H. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organ Res Methods*. 2013;16(2):270-301. doi:10.1177/1094428112470848
127. Rousseeuw PJ, Hubert M. Robust statistics for outlier detection. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2011;1(1):73-79. doi:10.1002/widm.2
128. Spector PE, Liu C, Sanchez JI. Methodological and Substantive Issues in Conducting

Multinational and Cross-Cultural Research. *Annu Rev Organ Psychol Organ Behav*. 2015;2(1):101-131. doi:10.1146/annurev-orgpsych-032414-111310
129. Massey A, Miller SJ. Tests of Hypotheses Using Statistics. 2006:1-32.
130. Suresh K, Chandrashekara S. Sample size estimation and power analysis for clinical research studies. *J Hum Reprod Sci*. 2012;5(1):7-13. doi:10.4103/0974-1208.97779
131. Silva-Aycaguer LC, Suarez-Gil P, Fernandez-Somoano A. The null hypothesis significance test in health sciences research (1995-2006): statistical analysis and interpretation. *BMC Med Res Methodol*. 2010;10:44. doi:10.1186/1471-2288-10-44
132. Yap BW. Power Comparisons of Shapiro-Wilk , Kolmogorov-Smirnov , Lilliefors and Anderson-Darling Tests. 2011;(January).
133. Road D. Approximating the Shapiro-Wilk W-test for non-normality. 2000;(1992):3-5.
134. Mendes M, Pala A. Type I Error Rate and Power of Three Normality Tests in Terms of Type I Error Rate and Power Under Different Distributions. *Turkish J Med Sci*. 2003;2(2):135-139.
135. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: A guide for non-statisticians. *Int J Endocrinol Metab*. 2012;10(2):486-489. doi:10.5812/ijem.3505
136. Davis RB, Mukamal KJ. Hypothesis testing: means. *Circulation*. 2006;114(10):1078-1082. doi:10.1161/CIRCULATIONAHA.105.586461
137. Lamb TJ, Graham AL, Petrie A. T testing the immune system. *Immunity*. 2008;28(3):288-292. doi:10.1016/j.immuni.2008.02.003
138. Kim TK. T test as a parametric statistic. 2015;(Table 2).
139. Gastwirth JL, Gel YR, Miao W. The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice. *Stat Sci*. 2009;24(3):343-360. doi:10.1214/09-STS301
140. Hart A. Mann-Whitney test is not just a test of medians: differences in spread can be important. *Bmj*. 2001;323(7309):391-393. doi:10.1136/bmj.323.7309.391
141. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*. 2003;19(3):368-375. doi:10.1093/bioinformatics/btf877
142. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Source J R Stat Soc Ser B J R Stat Soc Ser B J R Stat Soc B*. 1995;57(1):289-300. doi:10.2307/2346101
143. Gheva D. The biplot graphic technique for the representation of multivariate time series. *Eur J Oper Res*. 1986;27(1):95-103. doi:10.1016/S0377-2217(86)80011-X
144. Pandey M, Jain A. ROC Curve : Making way for correct diagnosis Manoj Pandey , Ephicacy Lifescience Analytics Pvt . Ltd ., Bangalore , India Abhinav Jain , Ephicacy Consulting Group Inc ., Iselin , NJ. 2016:1-12.
145. Greiner M, Pfeiffer D, Smith RD. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Prev Vet Med*. 2000;45(1-2):23-41. doi:10.1016/S0167-5877(00)00115-X
146. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Diagnostic methods 2: Receiver operating characteristic (ROC) curves. *Kidney Int*. 2009;76(3):252-256. doi:10.1038/ki.2009.171
147. El Khouli RH, Macura KJ, Barker PB, Habba MR, Jacobs MA, Bluemke DA. Relationship of temporal resolution to diagnostic performance for dynamic contrast enhanced MRI of the breast. *J Magn Reson Imaging*. 2009;30(5):999-1004. doi:10.1002/jmri.21947
148. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts

and first analyses. *Br J Cancer*. 2003;89(2):232-238. doi:10.1038/sj.bjc.6601118
149. Kishore J, Goel M, Khanna P. Understanding survival analysis: Kaplan-Meier estimate. *Int J Ayurveda Res*. 2010;1(4):274. doi:10.4103/0974-7788.76794
150. Jager KJ, Dijk PC Van, Zoccali C, Dekker FW. The analysis of survival data : the Kaplan – Meier method. *Kidney Int*. 2008;74(5):560-565. doi:10.1038/ki.2008.217
151. Peto R, Pike MC, Armitage P, et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *Br J Cancer*. 1977;35(1):1-39. doi:10.1038/bjc.1977.1
152. Bland JM, Altman DG, Bland JM, Altman DG. The logrank test service The logrank test. *Bmj*. 2004;(November 2008):10-12. doi:10.1136/bmj.328.7447.1073
153. Kristensen SL, Rørth R, Jhund PS, et al. Microvascular complications in diabetes patients with heart failure and reduced ejection fraction-insights from the Beta-blocker Evaluation of Survival Trial. *Eur J Heart Fail*. 2018. doi:10.1002/ejhf.1201
154. Swets JA. Measuring the accuracy of diagnostic systems. *Science*. 1988;240(4857):1285-1293. doi:10.1126/science.3287615
155. Jansson MD, Lund AH. MicroRNA and cancer. *Mol Oncol*. 2012;6(6):590-610. doi:10.1016/j.molonc.2012.09.006
156. Melo SA, Esteller M. Dysregulation of microRNAs in cancer: Playing with fire. *FEBS Lett*. 2011;585(13):2087-2099. doi:10.1016/j.febslet.2010.08.009
157. Sassen S, Miska EA, Caldas C. MicroRNA - Implications for cancer. *Virchows Arch*. 2008;452(1):1-10. doi:10.1007/s00428-007-0532-2
158. Marsh JW, Dvorchik I, Bonham CA, Iwatsuki S. Is the pathologic TNM staging system for patients with hepatoma predictive of outcome? *Cancer*. 2000;88(3):538-543. http://www.ncbi.nlm.nih.gov/pubmed/10649244.
159. Pizzini S, Bisognin A, Mandruzzato S, et al. Impact of microRNAs on regulatory networks and pathways in human colorectal carcinogenesis and development of metastasis. *BMC Genomics*. 2013;14(1):1. doi:10.1186/1471-2164-14-589
160. Fesler A, Jiang J, Zhai H, Ju J. Circulating microRNA testing for the early diagnosis and follow-up of colorectal cancer patients. *Mol Diagn Ther*. 2014;18(3):303-308. doi:10.1007/s40291-014-0089-0
161. Lu J, Getz G, Miska EA, et al. MicroRNA expression profiles classify human cancers. *Nature*. 2005;435(7043):834-838. doi:10.1038/nature03702
162. Qu Y, Shi B, Hou P. Activated ERK: An Emerging Player in miRNA Downregulation. *Trends in Cancer*. 2017;3(3):163-165. doi:10.1016/j.trecan.2017.01.002
163. Peng Y, Croce CM. The role of MicroRNAs in human cancer. *Signal Transduct Target Ther*. 2016;1(November 2015):15004. doi:10.1038/sigtrans.2015.4
164. Sun HL, Cui R, Zhou JK, et al. ERK Activation Globally Downregulates miRNAs through Phosphorylating Exportin-5. *Cancer Cell*. 2016;30(5):723-736. doi:10.1016/j.ccell.2016.10.001
165. Lee EJ, Baek M, Gusev Y, et al. Systematic evaluation of microRNA processing patterns in tissues , Systematic evaluation of microRNA processing patterns in tissues , cell lines , and tumors. *Rna*. 2008:35-42. doi:10.1261/rna.804508.miRNAs
166. Chen W, Lin X, Huang J, et al. Integrated profiling of microRNA expression in membranous nephropathy using high-throughput sequencing technology. *Int J Mol Med*. 2014;33(1):25-34. doi:10.3892/ijmm.2013.1554
167. Cook AD, Single R, McCahill LE. Surgical resection of primary tumors in patients who

present with stage IV colorectal cancer: An analysis of surveillance, epidemiology, and end results data, 1988 to 2000. *Ann Surg Oncol*. 2005;12(8):637-645. doi:10.1245/ASO.2005.06.012
168. K. C, S.M. S, L S. Palliative care of patients with colorectal cancer. *Libr Oncol*. 2013;41(1-3):93-98. http://hrcak.srce.hr/libri-oncologici?lang=en%0Ahttp://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emexa&NEWS=N&AN=616114476.
169. Reid JF, Sokolova V, Zoni E, et al. miRNA Profiling in Colorectal Cancer Highlights miR-1 Involvement in MET-Dependent Proliferation. *Mol Cancer Res*. 2012;10(4):504-515. doi:10.1158/1541-7786.MCR-11-0342
170. Hermeking H. MicroRNAs in the p53 network: Micromanagement of tumour suppression. *Nat Rev Cancer*. 2012;12(9):613-626. doi:10.1038/nrc3318
171. Zhang S, Zhang C, Li Y, Wang P, Yue Z, Xie S. MiR-98 regulates cisplatin-induced A549 cell death by inhibiting TP53 pathway. *Biomed Pharmacother*. 2011;65(6):436-442. doi:10.1016/j.biopha.2011.04.010
172. Rokavec M, Li H, Jiang L, Hermeking H. The p53/microRNA connection in gastrointestinal cancer. *Clin Exp Gastroenterol*. 2014;7:395-413. doi:10.2147/CEG.S43738
173. Hoffman Y, Pilpel Y, Oren M. MicroRNAs and Alu elements in the p53-Mdm2-Mdm4 regulatory network. *J Mol Cell Biol*. 2014;6(3):192-197. doi:10.1093/jmcb/mju020
174. Sun D, Yu F, Ma Y, et al. MicroRNA-31 activates the RAS pathway and functions as an oncogenic MicroRNA in human colorectal cancer by repressing RAS p21 GTPase activating protein 1 (RASA1). *J Biol Chem*. 2013;288(13):9508-9518. doi:10.1074/jbc.M112.367763
175. Ye J-J, Cao J. MicroRNAs in colorectal cancer as markers and targets: Recent advances. *World J Gastroenterol*. 2014;20(15):4288-4299. doi:10.3748/wjg.v20.i15.4288
176. Vinet L, Zhedanov A. A 'missing' family of classical orthogonal polynomials. *J Phys A Math Theor*. 2011;44(8):085201. doi:10.1088/1751-8113/44/8/085201
177. Ji S, Ye G, Zhang J, et al. MiR-574-5p negatively regulates Qki6/7 to impact β-catenin/Wnt signalling and the development of colorectal cancer. *Gut*. 2013;62(5):716-726. doi:10.1136/gutjnl-2011-301083
178. Wang Y-N, Chen Z-H, Chen W-C. Novel circulating microRNAs expression profile in colon cancer: a pilot study. *Eur J Med Res*. 2017;22(1):51. doi:10.1186/s40001-017-0294-5
179. Mullany LE, Herrick JS, Wolff RK, Stevens JR, Slattery ML. Association of cigarette smoking and microRNA expression in rectal cancer: Insight into tumor phenotype. *Cancer Epidemiol*. 2016;45:98-107. doi:10.1016/j.canep.2016.10.011
180. No Title.
181. Huang Z, Huang D, Ni S, Peng Z, Sheng W, Du X. Plasma microRNAs are promising novel biomarkers for early detection of colorectal cancer. *Int J Cancer*. 2010;127(1):118-126. doi:10.1002/ijc.25007
182. Wang Q, Huang Z, Ni S, et al. Plasma miR-601 and miR-760 are novel biomarkers for the early detection of colorectal cancer. *PLoS One*. 2012;7(9):e44398. doi:10.1371/journal.pone.0044398
183. Masuda T, Hayashi N, Kuroda Y, Ito S, Eguchi H, Mimori K. MicroRNAs as Biomarkers in Colorectal Cancer. *Cancers (Basel)*. 2017;9(9). doi:10.3390/cancers9090124

184. Vukobrat-Bijedic Z, Husic-Selimovic A, Sofic A, et al. Cancer Antigens (CEA and CA 19-9) as Markers of Advanced Stage of Colorectal Carcinoma. *Med Arch (Sarajevo, Bosnia Herzegovina)*. 2013;67(6):397-401. doi:10.5455/medarh.2013.67.397-401
185. Szajda SD, Snarska J, Jankowska A, Roszkowska-Jakimiec W, Puchalski Z, Zwierz K. Cathepsin D and carcino-embryonic antigen in serum, urine and tissues of colon adenocarcinoma patients. *Hepatogastroenterology*. 55(82-83):388-393. http://www.ncbi.nlm.nih.gov/pubmed/18613372.
186. Jin Z, Jiang W, Wang L. Biomarkers for gastric cancer: Progression in early diagnosis and prognosis (review). *Oncol Lett*. 2015;9(4):1502-1508. doi:10.3892/ol.2015.2959
187. Świderska M, Choromańska B, Dąbrowska E, et al. The diagnostics of colorectal cancer. *Wspolczesna Onkol*. 2014;18(1):1-6. doi:10.5114/wo.2013.39995
188. Zhou K, Liu M, Cao Y. New Insight into microRNA Functions in Cancer: Oncogene-microRNA-Tumor Suppressor Gene Network. *Front Mol Biosci*. 2017;4:46. doi:10.3389/fmolb.2017.00046
189. Hashimoto Y, Zumwalt TJ, Goel A. DNA methylation patterns as noninvasive biomarkers and targets of epigenetic therapies in colorectal cancer. *Epigenomics*. 2016;8(5):685-703. doi:10.2217/epi-2015-0013
190. Luo Y, Wong C-J, Kaz AM, et al. Differences in DNA methylation signatures reveal multiple pathways of progression from adenoma to colorectal cancer. *Gastroenterology*. 2014;147(2):418-29.e8. doi:10.1053/j.gastro.2014.04.039
191. Balaguer F, Link A, Lozano JJ, et al. Epigenetic silencing of miR-137 is an early event in colorectal carcinogenesis. *Cancer Res*. 2010;70(16):6609-6618. doi:10.1158/0008-5472.CAN-10-0622
192. Suzuki H, Maruyama R, Yamamoto E, Kai M. DNA methylation and microRNA dysregulation in cancer. *Mol Oncol*. 2012;6(6):567-578. doi:10.1016/j.molonc.2012.07.007
193. Hanoun N, Delpu Y, Suriawinata AA, et al. The silencing of microRNA 148a production by DNA hypermethylation is an early event in pancreatic carcinogenesis. *Clin Chem*. 2010;56(7):1107-1118. doi:10.1373/clinchem.2010.144709
194. Aavik E, Lumivuori H, Leppänen O, et al. Global DNA methylation analysis of human atherosclerotic plaques reveals extensive genomic hypomethylation and reactivation at imprinted locus 14q32 involving induction of a miRNA cluster. *Eur Heart J*. 2015;36(16):993-1000. doi:10.1093/eurheartj/ehu437
195. Medvedeva YA, Khamis AM, Kulakovskiy I V, et al. Effects of cytosine methylation on transcription factor binding sites. *BMC Genomics*. 2014;15:119. doi:10.1186/1471-2164-15-119

Anexes:

Annex 1 Detailed patient information for Colon and Rectal cancer patients used in miRNA expression analysis

Groups	Normal (n=11)	Stage I (n=49)	Stage II (n=122)	Stage III (n=106)	Stage IV (n=44)
Age (mean ± sd[1] years)	68 ± 18	64 ± 13	66 ± 13	63 ± 13	61 ± 13
< 65 years old	5 (45%)	23 (47%)	52 (43%)	55 (52%)	26 (59%)
> 65 years old	6 (55%)	26 (53%)	70(57%)	51 (48%)	18 (41%)
Gender					
Female	9 (82%)	21(43%)	57 (47%)	49 (46%)	19 (43%)
Male	2 (18%)	28 (57%)	65 (53%)	57 (54%)	25 (57%)
Anatomic Subdivision					
Ascending Colon	2 (18%)	6 (12%)	23 (19%)	10 (9%)	6 (14%)
Cecum	2 (18%)	17 (35%)	19 (16%)	22 (21%)	7 (16%)
Descending Colon	0 (0%)	1 (2%)	5 (4%)	5 (5%)	2 (5%)
Hepatic Flexure	2(18%)	0 (0%)	6 (5%)	7 (7%)	1 (2%)
Rectosigmoid Junction	1 (9%)	7 (15%)	10 (8%)	16 (15%)	7 (16%)
Rectum	1 (9%)	4 (8%)	14 (11%)	14 (13%)	6 (14%)
Sigmoid Colon	2 (18%)	10 (20%)	28 (23%)	21 (20%)	12 (27%)
Splenic Flexure	0 (0%)	1 (2%)	3 (3%)	0 (0%)	1 (2%)
Transverse Colon	0 (0%)	2 (4%)	9 (7%)	9 (8%)	1 (2%)
Not Reported	1 (9%)	1 (2%)	5 (4%)	2 (2%)	1 (2%)
Tumor Site					
Colon	8 (73%)	37 (76%)	97 (80%)	75 (70%)	31 (70%)
Rectum	3 (27%)	12 (24%)	25 (20%)	31 (30%)	13 (30%)

1 – standard deviation

Annex 2 Detailed patient information for colon and Rectal Cancer patients used in the DNA methylation analysis

Groups	Normal (n=45)	Stage I (n=55)	Stage II (n=144)	Stage III (n=120)	Stage IV (n=54)
Age (mean ± sd[1] years)	69 ± 13	66 ± 13	66 ± 13	63 ± 13	61 ± 13
< 65 years old	12 (27%)	25 (45%)	57 (40%)	61(50%)	23 (43%)
> 65 years old	33 (73%)	30 (55%)	85 (59%)	59 (50%)	31 (57%)
Not reported	0 (0%)	0 (0%)	2 (1%)	0 (0%)	0 (0%)
Gender					
Female	21 (47%)	22(40%)	71 (49%)	57 (48%)	22 (41%)
Male	24 (53%)	33 (60%)	73 (51%)	63 (52%)	32 (59%)
Anatomic Subdivision					
Ascending Colon	5 (11%)	6 (11%)	28 (19%)	11 (9%)	6 (11%)
Cecum	10 (23%)	19 (34%)	22 (15%)	24 (20%)	9 (17%)
Descending Colon	2 (4%)	2 (4%)	5 (3%)	5 (4%)	2 (4%)
Hepatic Flexure	5 (11%)	2 (4%)	8 (6%)	7 (6%)	1 (2%)
Rectosigmoid Junction	0 (0%)	6 (11%)	10 (7%)	17 (14%)	7 (13%)
Rectum	7 (16%)	4 (7%)	18 (13%)	17 (14%)	6 (11%)
Sigmoid Colon	12 (27%)	11 (20%)	34 (24%)	24 (20%)	17 (31%)
Splenic Flexure	0 (0%)	1 (2%)	3 (2%)	0 (0%)	1 (2%)
Transverse Colon	2 (4%)	3 (5%)	11 (8%)	10 (9%)	1 (2%)
Not Reported	2 (4%)	1 (2%)	5 (3%)	5 (4%)	4 (7%)
Tumor Site					
Colon	38 (84%)	44 (80%)	115 (80%)	85 (70%)	41 (76%)
Rectum	7 (16%)	11 (20%)	29 (20%)	35 (30%)	13 (24%)

1 – standard deviation