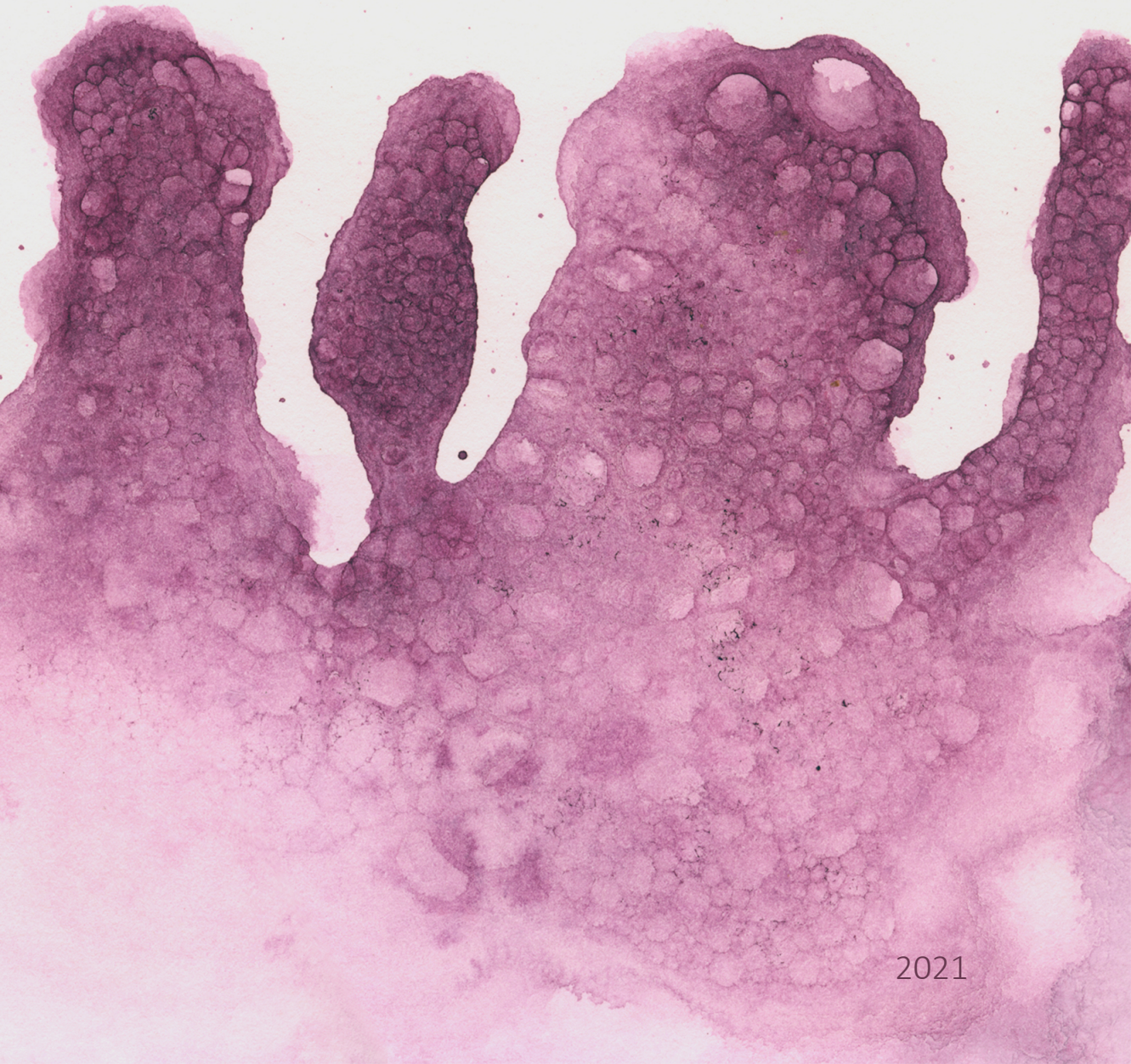
UNIVERSITAT DE BARCELONA

Genetic and epigenetic insights into colorectal tumorigenesis

Júlia Matas Gironella

GENETIC AND EPIGENETIC INSIGHTS INTO COLORECTAL TUMORIGENESIS

Júlia Matas Gironella



2021

Doctoral Programme in Biomedicine

Genetic and epigenetic insights into colorectal tumorigenesis

PhD thesis by:

Júlia Matas Gironella

to qualify for the degree of Doctor

by the University of Barcelona

This thesis was performed at the Institut Germans Trias i Pujol (IGTP) under the supervision of Dr. Miguel Ángel Peinado.

Supervisor	Tutor	Doctoral student
Dr. Miguel Á. Peinado Morales	Dr. Montserrat Corominas Guiu	Júlia Matas Gironella

Barcelona, September 2021

ABSTRACT

Colorectal cancer (CRC) is a major health burden with large numbers of new cases worldwide and high disease-specific mortality, despite great advances made towards improving patient clinical outcomes. It arises through the gradual acquisition of particular genetic and epigenetic alterations within normal cells, giving them selective advantage in driving malignant transformation. As such process takes over a decade, early cancer detection actions should strongly impact reducing morbidity. Identification of reliable CRC biomarkers is a permanent challenge for improving CRC management. Thanks to the emergence of new powerful technologies and the advances in the knowledge of the mechanistic bases of the disease, recent genetic and epigenetic markers are becoming promising candidates for early detection, risk stratification, prognosis, and prediction of treatment response. In this doctoral thesis, we have addressed mechanistic and clinical aspects of colorectal tumorigenesis: the deregulation and function of FOXD2 and FOXD2-AS1 in CRC tissues and cell lines (**study I**) and the role of precancerous mutations in normal colorectal mucosa (**study II**).

In **study I**, we characterized the transcriptomic and epigenetic profiles of FOXD2 and FOXD2-AS1 genes in normal and tumor colorectal samples. As bidirectional genes in head-to-head disposition, they showed a strong correlation at the transcriptomic level. However, in tumors they displayed an unbalanced bidirectional expression, whereas FOXD2 was strongly downregulated in association with higher methylation levels outside the promoter region. Interestingly, when we induced overexpression of such genes in CRC cell lines, FOXD2 behaved as a tumor suppressor by reducing migration and colony formation, while FOXD2-AS1 increased migration rates. Overall, our findings suggest the involvement of major mechanisms rewiring cancer, responsible for an altered bidirectional transcription of FOXD2 and FOXD2-AS1.

In **study II**, we focused on the characterization of somatic mutation in normal colorectal mucosa of individuals with and without CRC, using an ultra-deep sequencing technology, CRISPR-Duplex Sequencing. We identified coding mutations in normal colon of most individuals on the 4 cancer genes included in the panel: *BRAF*, *KRAS*, *PIK3CA*, and *TP53*. However, *TP53* and *KRAS* driver mutations were commonly found in normal colon of CRC patients, often displaying clonal expansions in early onset CRC. Additionally, we developed a primary and integrative mutational model based on the mutational analysis of normal biopsies with potential for CRC risk prediction. Overall, our results support a model where somatic evolution contributes to the expansion of mutated clones in the normal colon tissue, but this process is enhanced in young individuals with cancer.

CONTENTS

FIGURE INDEX

Introduction

Methods Study I

Methods Study II

Results Study I

Results Study II

Appendix

TABLE INDEX

ABBREVIATIONS

5-mC	5-methyl cytosine
APC	Adenomatous Polyposis Coli gene
bp	Base pair
cDNA	Complementary DNA
COAD	Colon adenocarcinoma
CTNNB1	Catenin Beta 1
CRC	Colorectal cancer
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
crRNA	CRISPR RNA
DAC	5-aza-2'-deoxycytidine
DCS	Duplex consensus sequence
DNMT	DNA methyltransferase
DS	Duplex Sequencing
EGFR	Epidermal growth factor receptor
EPCAM	Epithelial Cell Adhesion Molecule
FAP	Familial Adenomatous Polyposis
FOXD2	Forkhead Box D2
FOXD2-AS1	FOXD2 adjacent opposite strand RNA 1
CpGi	CpG islands
CIMP	CpG island methylator phenotype
CIN	Chromosomal instability
gRNA	guide RNA
GO	Gene ontology
GTEX	Genotype-Tissue Expression
H2K4me1	Histone 2 lysine 4 monomethylation
H3K27ac	Histone 3 lysine 27 acetylation
H3K27me3	Histone 3 lysine 27 trimethylation
H3K36me3	Histone 3 lysine 36 trimethylation
H3K4me3	Histone 3 lysine 4 trimethylation
H3K9ac	Histone 3 lysine 9 acetylation
H3K9me3	Histone 3 lysine 9 trimethylation
HNPCC	Hereditary Non-Polyposis Colorectal cancer
HRAS	Harvey Rat Sarcoma Viral Oncogene Homolog
HUB	Hospital Universitari de Bellvitge
KRAS	Kirsten Rat Sarcoma Viral Oncogene Homolog
lncRNA	Long non-coding RNAs
MAF	Mutant Allele Frequency
MALAT1	Metastasis associated lung adenocarcinoma transcript 1
MBP	Methyl-cytosine-binding protein
miRNA	Micro RNA
MLH1	MutL homolog 1
MMR	Mismatch repair
MSH2	MutS homolog 2

MSH6	MutS homolog 6
MSI	Microsatellite instability
ncRNA	non-coding RNA
NGS	Next-generation sequencing
NRAS	Neuroblastoma Rat Sarcoma Viral Oncogene Homolog
ORF	Open reading frame
PMS2	PMS1 Homolog 2, Mismatch Repair System Component
PCR	Polymerase chain reaction
READ	Rectal adenocarcinoma
RT-qPCR	Real-time quantitative PCR
SSCS	Single Strand Consensus Sequence
TA	Transit amplifying cells
TAD	Transactivating domain
TCGA	The Cancer Genome Atlas
TD	Tetramerization domain
TET	Ten eleven translocation
TNM	TNM classification of malignant tumors
tracrRNA	Trans-activating crispr RNA
TSS	Transcriptional Start Site
LOH	Loss of heterogenicity
TF	Transcription factor
UV	Ultraviolet light

INTRODUCTION

1. Colorectal Cancer

1.1. The large intestine, a short introduction

The large intestine is the final section of the gastrointestinal tract extending from the terminal ileum to the anal canal, comprising the cecum, colon, and rectum. Its principal function is to absorb water and electrolytes, converting the digestive residues into feces and transporting them to the anus.

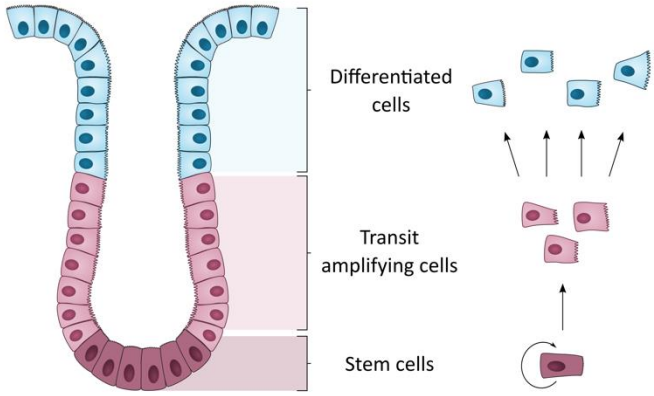


Figure 1. **Colonic crypt.** Representation of a normal colonic crypt. Stem cells have self renewal capacity and give rise to transit-amplifying (TA) cells with limited proliferation capacity. After a few rounds of duplication TA cells migrate to the top of the crypt becoming differentiated cells.

Histologically, four main layers make up the wall of the colon: mucosa, submucosa, muscularis mucosa, and serosa. Focusing on the mucosa, it is comprised of a single layer of epithelial cells followed by connective tissue and muscle. The epithelial cells constitute a barrier between the lumen gut and the host tissue and are folded into millions of invaginations, named crypts (**Figure 1**). The adult epithelium has about 15 million crypts (Boman and Huang 2008), each composed of a clonal population of 2000 cells (C. S. Potten et al. 1992). Towards the base of each crypt, there are about four to six intestinal stem cells –ISCs– with self-renewal capacity that give rise to progenitor cells, also known as transit-amplifying (TA) cells. TA cells occupy the middle compartment of the crypt and have a limited proliferation capacity, undergoing up to five rounds of cell division (Barker et al. 2009) and being able to migrate up to the top of the crypt, finally becoming fully differentiated intestinal cells (Gehart and Clevers 2019; Christopher S. Potten 1998) (**Figure 1**). Differentiated cells, which are colonocytes (absorptive cells), goblet cells

(mucus-secreting cells), entero-endocrine cells, and Paneth cells (present only in the ascending colon) (Humphries and Wright 2008), are continually extruded into the lumen. Therefore, the crypt is a dynamic structure that is constantly self-renewed and replaced every five days (Radtke and Clevers 2005).

The intestinal epithelium gives rise to colorectal cancers, starting with an aberrant crypt and progressing to a precursor lesion (polyp) that can eventually evolve into cancer. Given the high self-renewal capacity of adult stem cells, ISCs have been proposed as the main cell of origin for tumorigenesis initiation and development (Barker et al. 2009; Nassar and Blanpain 2016). However, it is still under debate which cell type sustains the cancer-initiation mutation.

1.2. Colorectal cancer incidence

Colorectal cancer (CRC) is the second and third most common cancer in women and men, respectively, with an estimated 1.9 million new diagnosed cases worldwide in 2020 (GLOBOCAN 2020). Although most CRCs occur in individuals aged 50 and older, 10% of new cases diagnosed in 2020 were in individuals younger than 50 years old, pointing out the increased incidence in young individuals reported in recent years (Siegel, Miller, and Jemal 2017). In 2020, as a result of CRC, about 1 million patients died, making it the fifth leading cause of cancer-related deaths (GLOBOCAN 2020).

1.3. Etiology and risk factors

The etiology and pathogenicity underlying the development of CRC are complex and heterogeneous. Most cases of CRC are sporadic (75-80%) and result from the progressive accumulation of both genetic and epigenetic alterations that contribute to the transformation from normal epithelial cells to neoplastic cells. Only a small proportion of CRC are related to heritable factors or family history. One of the most common hereditary syndromes is the Lynch Syndrome, also known as Hereditary Non-Polyposis Colorectal cancer (HNPCC), accounting for 2-4% of all colorectal cancer cases (Heather et al. 2015; Samowitz et al. 2001). It is an autosomal dominant syndrome characterized by early onset colorectal cancer, as it has an earlier diagnosis age than the rest of CRC. HNPCC tumors are caused by mutations in one of the DNA mismatch-repair genes *MLH1*, *MSH2*, and more rarely *MSH6*, *PMS2,* and E*PCAM*. Another hereditary CRC is the Familial Adenomatous Polyposis (FAP) that only accounts for 1% of all colorectal cancers and is caused by mutations in the Adenomatous Polyposis Coli gene (*APC*), a tumor suppressor gene involved in the β-Catenin/Wnt signaling pathway (see section 1.5.4).

Apart from genetic predisposition, and given that the majority of CRC are sporadic, environmental agents and lifestyle, as well as other biological characteristics (e.g., age), are important risk factors associated with tumor development and progression (**Figure 2**) (Dekker et al. 2019). Aging is one of the most well-known risk factors in cancer, as cancer incidence steadily increases with age. Also, male sex and polyps have shown strong associations with CRC incidence (Siegel et al. 2020; Click et al. 2018; Song et al. 2020). A range of environmental lifestyle factors increase the risk of developing the disease, such as smoking (Erhardt et al. 2002; Kikendall et al. 1989; Liang, Chen, and Giovannucci 2009), high alcohol intake (Park et al. 2019; Zisman et al. 2006), overweight, obesity (Johnson and Lund 2007), sedentary lifestyle (Cong et al. 2014; Namasivayam and Lim 2017) and high-fat or low-fiber diet habits (Johnson and Lund 2007). To date, many other exposures have been studied for their associations with the risk of developing CRC, but some have yielded ambiguous results.

As the relevance of environmental risk factors to develop CRC has been proven by many studies, it is not surprising that CRC is more predominant in developed countries as the lifestyles harbor many of the risk factors mentioned above.

Figure 2. **CRC risk factors.** Increased risk of developing colorectal cancer has been associated with genetic, environmental and other risk factors.

1.4. Screening and staging

Most patients developing colorectal cancer will eventually present a series of symptoms, being rectal bleeding, change in bowel habits, and abdominal discomfort/pain, the most specific ones. Other general symptoms may appear, like bodyweight loss, fatigue and fever. However, the presentation of symptoms in colorectal cancer often depends on the tumor site and extent of disease, being patients in an early CRC stage often asymptomatic.

The diagnosis of colorectal cancer can be assessed as a result of a symptomatic patient or a screening test. In apparently healthy people, screening programs for CRC aim to detect tumors at early stages, enabling successful treatment and improving survival rates. There are different tests for CRC detection: stool-based tests such as fecal occult blood test (FOBT) and fecal immunochemical test (FIT), which are the non-invasive techniques frequently used for first-line screening. Other structure-based screening tests are applied, being colonoscopy and flexible sigmoidoscopy the most common in clinical practice (Davidson et al. 2021). Recently, epigenetic-based tests have been introduced as non-invasive CRC screening (Okugawa, Grady, and Goel 2017; Yunfeng Zhang et al. 2021).

Colonoscopy is considered the current gold standard test to detect both adenomatous polyps and cancer, being the most complete screening procedure. Although invasive, in addition to asses a high accurate diagnostic, it allows biopsy sampling to confirm the diagnosis histologically and for molecular profiling (Kuipers et al. 2015), as well as it offers the potential for direct removal of precursor lesions. In clinical practice, colonoscopy is a method of detection usually applied when a patient is symptomatic and when the FOBT was positive. Also, it is the recommended procedure for people with a history of previous polyps or colorectal cancer and people with elevated risk (e.g., positive family history). The colonoscopy approach has demonstrated a significant impact on decreasing CRC incidence and mortality (Doubeni et al. 2018).

After a colorectal tumor is diagnosed, it is crucial to determine the stage of cancer in order to design an appropriate treatment plan. The TNM system of the American Joint Committee on Cancer (Frederick L et al. 2002) is the current strategy to classify tumors and determine the prognosis of the disease and the treatment management. It is assessed according to the extent of the tumor (T), the spread to nearby lymph nodes (N), and the distant metastasis (M) (**Table 1A** and **1B**).

Table 1A. TNM staging according to AJCC.

AJCC stage	TNM stage
Stage 0	Tis, N0, M0
Stage I	T1 o T2, N0, M0
Stage II-A	T3, N0, M0
Stage II-B	T4a, N0, M0
Stage II-C	T4b, N0, M0
Stage III-A	T1 o T2, N1, M0
Stage III-B	T3 o T4, N1, M0
Stage III-C	any T, N2, M0
Stage IV	any T, any N, M1

Table 1B. TNM staging according to AJCC.

TNM classification					
Primary tumor		Regional lymph Nodes		Metastasis	
Tis	Tumor confined to mucosa	N0	No regional lymp nodes affected	M0	Distant metastases not present
T1	Tumor invades submucosa	N1	Methastasis to 1 to 3 regional lymph nodes	M1	Distant metastases present
T2	Tumor invades muscularis propria	N2	Methastasis to 4 or more regional lymph nodes		
T3	Tumor invades subserosa or beyond				
T4	Tumor invades adjacent organs or perforates the visceral peritoneum				

Different types of exams and tests can be done to determine the tumor stage, like physical exams, biopsies, and imaging procedures. Sometimes after surgery, definitive cancer staging can be better defined. Furthermore, other molecular analyses are being introduced into the classification of CRC to better stratify patients, such as immunohistochemistry, PCRs, and microarrays (Shia et al. 2012).

1.5. CRC pathogenicity

1.5.1. Cancer evolution

The traditional model of how cancer emerges is based on a Darwinian evolutionary process that typically occurs over decades, whereas cells gradually accumulate somatic mutations and evolve by selection and clonal expansion (Nowell, 1976). DNA sequencing efforts have revealed high variability rates of somatic mutations in adult cancers, from hundreds to thousands of single nucleotide substitutions to smaller numbers of insertions, deletions, and chromosomal alterations (Alexandrov and Stratton 2014). While most mutations are believed to be passenger (neutral), a modest but undefined number can affect key genes and confer a selective advantage compared to their neighboring cells, giving rise to the well-known driver mutations (**Figure 3**). The constant cell exposure to environmental factors and cell-intrinsic damage (e.g., replication, deficient DNA repair) can result in DNA alterations already present in healthy tissues that generate sub-clones, leading to tissue heterogeneity (Kennedy, Zhang, and Risques 2019).

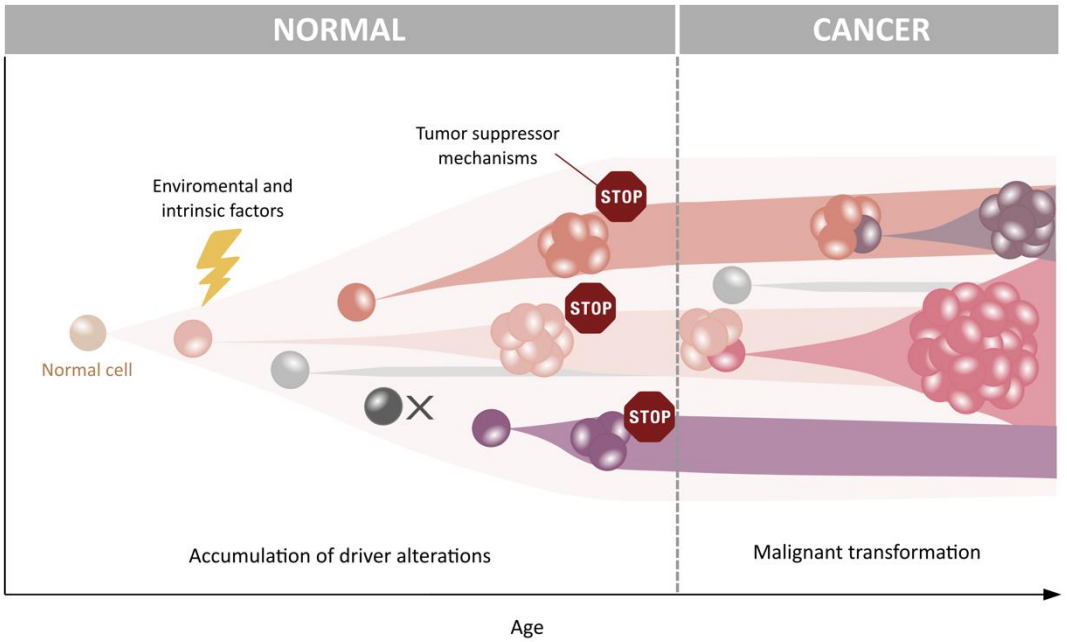


Figure 3. **Somatic evolution in cancer.** DNA mutations accumulate through life due to several intrinsic and environmental factors. Acquired mutations might be deleterious (X), neutral (light grey) or might confer a selective advantage to clonally expand (colored cells). Tumor suppressor mechanisms are critical to restrain potential premalignant clones to further expand. However, a minority will bypass these mechanisms by the accumulation of additional mutations that will acquire malignant properties leading to cancer. *Adapted from Kennedy et al 2019.*

Passenger changes have a weak or neutral effect on the neoplastic process, being the majority of mutations accumulated in cancer cells. They provide information on mutagenic events that cells have experienced in the past, as well as information about the specific patterns of mutations across the genome. A clear example is the case of melanomas or lung cancers in smokers. As the skin is exposed to UV-light and smokers to tobacco carcinogens (Nik-Zainal et al. 2015), tumors arising from these tissues can accumulate large numbers of mutations, which are considered passengers as they do not confer a selective advantage (Bert Vogelstein et al. 2013). Interestingly, *TP53* mutations in skin cancers carry a large number of C>T substitutions, while in cancers arising in the lung of smokers, *TP53* are often C>A (Nik-Zainal et al. 2015). Therefore, studying the type and distribution of passenger mutations can help us understand the underlying processes involved in mutagenesis.

Driver mutations are defined as mutations under positive selection within a population of cells that promote the cell growth to the next step of cancer disease. Cancer genes carrying drivers are classified as tumor suppressors and oncogenes, which by inactivation or activation respectively, have the ability to gain malignancy. Accumulation of driver

mutations is observed in a range of healthy tissues, as well as in cancer; however, they require the bypass of tumor suppressor mechanisms such as senescence and immune surveillance to further progress (Kennedy, Zhang, and Risques 2019) (**Figure 3**). It is believed that multiple driver mutations affecting key genes are needed for a cell to gain malignancy.

1.5.2. The adenoma to carcinoma sequence

Colorectal cancer is a complex multistep disease driven by genetic and epigenetic alterations that lead to the activation of key genes that provide a selective advantage to neoplastic cells. As previously mentioned, most colorectal carcinomas are believed to develop from polyp lesions through a long process, often taking ten or more years to develop. Fearon and Vogelstein first described the classic model of colorectal carcinogenesis in the 1990s as a transition from normal colonic epithelium to adenomas (benign neoplasms), followed by invasive carcinomas that can lead to the formation of metastases (B Vogelstein et al. 1988). This classic "adenoma to carcinoma" sequence has served as a paradigm for solid tumor progression. Since then, the molecular classification of CRC has evolved with the purpose of understanding the underlying mechanisms in the multistep carcinogenic process and predict the biological behavior of a particular tumor.

1.5.3. Genomic and epigenomic instability

The multiple mutations found in human cancer exceed the baseline mutation rate of normal cells. Almost 40 years ago, a "mutator phenotype" was proposed stating that the acquisition of multiple mutations observed in cancer cells is driven by genomic instability (Loeb, Loeb, and Anderson 2003). To date, three different mechanisms have been proposed for colorectal carcinogenesis according to the global genomic and epigenomic status: chromosomal instability (CIN), microsatellite instability (MSI), and CpG island methylator phenotype (CIMP) (**Figure 4**). CIN, MSI, and CIMP molecular subtypes may overlap by sharing similar molecular events. Thus, single events such as driver mutations in key cancer genes and aberrant signaling pathways are also useful classifiers for colorectal cancer and are reviewed in the next section (section 1.5.4).

CIN

Chromosomal instability is recognized by the presence of aneuploidy, a hallmark of cancer. Aneuploidy is a state with an abnormal number of chromosomes or multiple structural aberrations that promotes carcinogenesis by copy number gains of common oncogenes and loss of tumor suppressors. The mechanisms driving CIN are still poorly understood,

but defects in chromosomal segregation, DNA damage repair, telomere function, and specific mutations in key cancer genes have been discussed (Pino and Chung 2010; Tariq and Ghias 2016).

CIN is the most common form of genomic instability in colorectal carcinomas, accounting for ~85% of CRC. CIN tumor progression, also known as the adenoma-carcinoma sequence (B Vogelstein et al. 1988), arises via progressive accumulation of mutations in several genes. Inactivation of the *APC* gene is common during the early stages of progression, from normal epithelium to early adenoma, followed by *KRAS* mutations from early to late adenoma transition. Other mutations in genes such as *TP53* and *PIK3CA* take place during later stages to carcinoma progression (B Vogelstein et al. 1988), generating a malignant tumor that can lead to local and distant invasion (**Figure 4A**).

From a clinical point of view, CIN tumors typically arise in the left colon, with a well-differentiated phenotype. They usually display poor prognosis as there is a high tendency to invade local lymph nodes and produce distant metastases.

MSI

Microsatellite instability refers to the altered lengths of short tandem repeats in the DNA, named microsatellites. Microsatellites are composed of 2-6 nucleotide repeats found throughout the human genome, and due to their repetitive structure, they are prone to accumulate errors during cell replication. DNA polymerase is not perfect; thus, some base mismatches or indel loops can arise as a consequence of DNA polymerase slippage, and the mismatch repair system acts to preserve genomic integrity by eliminating these mistakes. Due to this, MSI mechanisms are driven by the functional loss of MMR genes, increasing the rate of polymerase-generated errors and degrading the fidelity of DNA replication.

MSI tumors account for approximately 15% of CRC and have been suggested as an alternative mechanism to the CIN pathway, being MSI and CIN mutually exclusive. As mentioned before, in the case of Lynch Syndrome, MSI is driven by germline mutations in MMR genes. However, in sporadic CRC, MMR inactivation is caused either by promoter hypermethylation or somatic mutations in MMR genes (Grady and Pritchard 2014) (**Figure 4B**), being *MLH1* promoter DNA hypermethylation the most common cause of sporadic MSI.

The accumulation of different driver mutations in MSI tumors has been reported, being *BRAF* V600E the most frequently acquired mutation, which is also commonly mutated in the serrated neoplasia pathway (**Figure 4**). In such cases, MSI tumors with *BRAF* V600E

effectively exclude the possibility of Lynch syndrome (Grady and Pritchard 2014; Tariq and Ghias 2016).

Unlike CIN, MSI tumors are usually located in the right colon, poorly differentiated, with high levels of lymphocyte infiltration, but with a better prognosis.

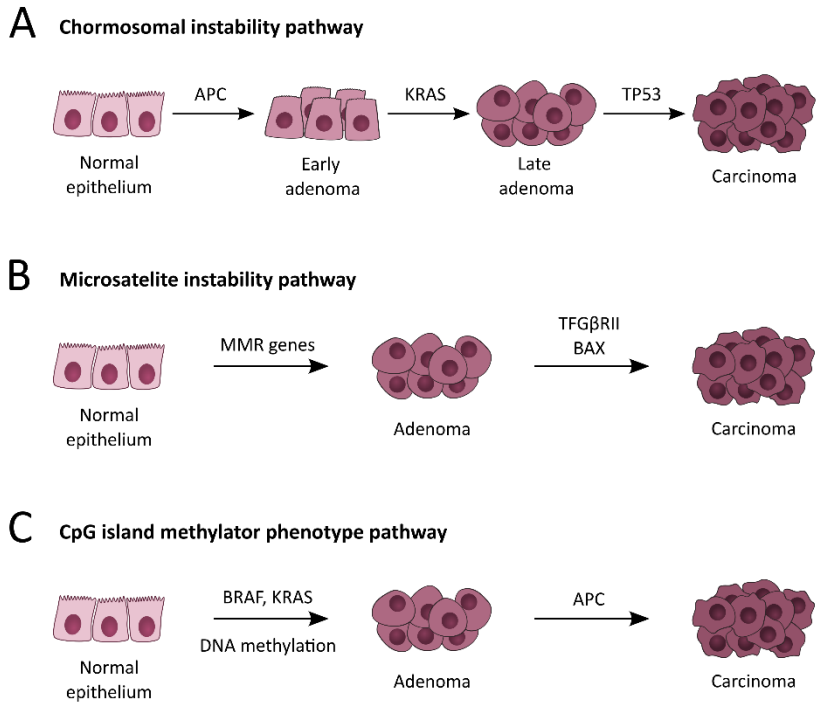


Figure 4. **Adenoma to carcinoma sequence.** Progression pathways from normal colon epithelium to carcinoma: Chromosomal instability pathway, CIN (**A**); Microsatellite instability pathway, MSI (**B**) and CpG island methylator phenotype pathway, CIMP (**C**).

CIMP

Epigenetic instability in CRC is characterized by a global DNA hypomethylation and localized hypermethylation of gene promoters that contain CpG islands. Aberrant methylation is present in most CRCs, but 20% of them display recurrent hypermethylation of several CpG loci, a class named CpG Island Methylator Phenotype (CIMP). This hypermethylation in promoters results in the transcriptional inactivation of key genes for normal cell functioning.

The mechanisms underlying CIMP are still unclear, but a strong association between CIMP and *BRAF* V600E mutation has been established, as well as the correlation between *BRAF* V600E and hypermethylation of *MLH1*, suggesting a link between sporadic MSI and CIMP (Grady and Pritchard 2014) (**Figure 4C**). In addition, age-related promoter hypermethylation of genes related to key pathways, such as WNT signaling and PI3 kinase, has been shown to sensitize cells to *BRAF* V600E induced transformation in CIMP tumors (Tao et al. 2019). Unlike CIN, *APC* alterations leading to WNT activations are typically seen at later stages of CIMP tumor progression (Tariq and Ghias 2016).

1.5.4. Driver genes in CRC

Besides genomic and epigenomic instability, several mutations in key genes are relevant in CRC pathogenesis. The most commonly mutated genes in CRC are involved in key signaling pathways, including the WNT signaling pathway, RAS/RAF/MAPK pathway, the PI3K pathway, and the TGFβl/SMAD pathway (**Figure 5**). Deregulation of the mentioned signaling pathway leads to alterations in cell proliferation, differentiation, apoptosis, angiogenesis, and invasion, common hallmarks in cancer (Grady and Pritchard 2014). Below I summarize the most relevant aspects of key genes in CRC studied in this thesis.

Adenomatous Polyposis Coli (*APC*)

Mutations in *APC* inactivation occur in 70% of sporadic CRC. *APC* function is lost by point somatic mutations, promoter hypermethylation, or loss of heterogenicity (LOH). Importantly, most cases of FAP are associated with *APC* germline mutations.

APC is a large protein with multiple functional domains that acts as tumor suppressor gene by negative regulating the canonical WNT signaling pathway, which controls the coordinated cell proliferation and differentiation. *APC* protein disruption results in pathway activation, specifically by increasing β-catenin levels in the nucleus that promote cell proliferation, apoptosis, and cell-cycle progression(L. Zhang and Shay 2017). Mutations in other genes of the pathway, particularly in β-catenin (*CTNNB1*), can also lead to WNT signaling activation. Importantly, *APC* mutations are often found at the earliest stages of neoplasia in CIN tumors (B Vogelstein et al. 1988); however, due to the large size of the gene and the high diversity of inactivating mutations, analysis of *APC* mutations for early cancer detection is not performed, except in families with FAP.

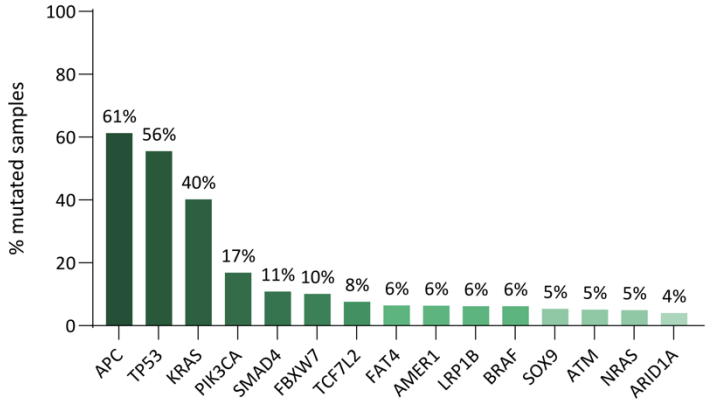


Figure 5. Landscape of mutations in CRC. Top 15 mutated genes in CRC. Data source: www.intogen.org

Tumor suppressor transcription factor 53 (*TP53*)

TP53 is one of the most mutated genes in human cancers. Approximately 50% of all CRC harbor mutations in *TP53* and is generally believed to occur in the transition from late adenoma to carcinoma, but it has also been reported to be an early event on CRC development (Gerstung et al. 2020). They are more commonly found in cancers that arise from the left colon and is frequently associated with the CIN subtype.

TP53 protein acts as a transcription factor by controlling the expression of a large number of target genes. Its activation usually occurs upon stress signals and results in different cellular outcomes, like proliferation prevention by cell cycle arrest, DNA repair, or apoptosis (Liebl and Hofmann 2021). *TP53* is located in chromosome 17 and the coding region contains 11 exons. Its protein is structured in a transactivation domain (TAD), a central DNA binding domain (DBD), and a tetramerization domain (TD). Most *TP53* mutations in CRC are missense mutations that occur in 8 hotspot codons (175, 273, 248, 282, 245, 213, 196 and 306) (Tate et al. 2019), mainly clustering at the DBD, which is responsible for recognizing specific p53-binding elements in the promoters of its target genes. Of note, 17p LOH frequently occurs in CRC (70%) (Fearon 2011), usually associated with somatic mutations in the *TP53* remaining allele.

Most somatic mutations observed in *TP53* lead to loss of function, frequently having a dominant effect over the wild-type *TP53*. Nonetheless, there is increasing evidence that gain of function *TP53* mutations can also act as key driver mutations for cancer progression (Pitolli et al. 2019).

Regarding *TP53* clinical use, the prognostic and predictive value of mutated *TP53* has been investigated, and it is still under debate. Some studies have associated *TP53* mutations with poor prognosis (Iacopetta et al. 2006; Lattery 2002), but others have failed to demonstrate it. Currently, *TP53* mutations have no clinical applications in CRC.

Kirsten rat sarcoma viral oncogene (*KRAS*)

KRAS is the most frequently mutated oncogene in human cancers, being approximately 40% of CRC *KRAS*-mutated. Although mutations in *KRAS* usually occur after *APC* mutations in the adenoma to carcinoma sequence, they are still considered an early event in tumor progression (Gerstung et al. 2020; B Vogelstein et al. 1988).

The RAS proto-oncogene family includes *HRAS*, *NRAS*, and *KRAS* genes, being *KRAS* the most commonly mutated. *KRAS* acts downstream the EGFR pathway, being a member of the MAP kinase pathway (RAS/RAF/MAPK) that regulates cell proliferation, differentiation, development, and apoptosis. It is a small protein activated by somatic mutations mainly in codons 12 and 13, being codon 12 the most affected. *KRAS* mutations lead to activation of the phosphoinositol kinase (PI3K) pathway, which inhibits apoptosis and activation of RAF, which stimulates cellular proliferation(Armaghany et al. 2012).

Mutations in *KRAS* are established as predictive biomarkers for treatment with EGFR inhibitors (such as cetuximab and panitumumab), being *KRAS*-mutated metastatic tumors non-responders to the therapy. Therefore, *KRAS* is tested before starting anti-EGFR treatment in patients with metastatic CRC.

V-RAF murine sarcoma viral oncogene homolog B (*BRAF*)

BRAF gene is mutated in approximately 5-10% of CRC and it usually occurs at early stages of colorectal tumorigenesis. It is commonly mutated in tumors arising from the right colon and associated with MSI and CIMP alterations.

BRAF is an oncogene, member of the RAS family that acts as a downstream effector of *KRAS* in the MAPK signaling pathway, thus downstream EGFR pathway. *BRAF* and *KRAS* mutations are mutually exclusive in CRC, suggesting that the activation of either one of them can promote tumorigenesis affecting the same signaling pathways. The main activation of the *BRAF* gene is V600E substitution which maintains the protein active, independently of RAS activity, that results in MEK-ERK activation, accelerating tumor cell proliferation, survival, and migration (X. Li et al. 2020).

BRAF mutations tend to be associated with poor clinical outcomes, and even though some studies have shown that *BRAF*^{V600E} tumors have a worse response to anti EGFR therapies, there is no sufficient evidence to use it as a prognostic biomarker in the clinical setting.

Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha (*PIK3CA*)

Somatic mutations in the *PIK3CA* gene are found in approximately 15–25% of CRCs (Fearon 2011), promoting the progression from late adenoma to carcinoma in CIN tumors.

PIK3CA oncogene encodes for p110α protein, a subunit of PI3K. In CRC, *PIK3CA* mutations mainly happen in codons 545 and 1047, leading to activation of the kinase activity. PI3K pathway, which regulates several cellular functions, such as migration and proliferation, and it is regulated by EGFR signaling in part via *KRAS* activation, is activated as a result of both *PIK3CA* mutations or loss of PTEN tumor suppressor gene (Grady and Pritchard 2014),.

Some studies have reported evidence of *PIK3CA* mutations as predictive biomarkers for some therapies. Nevertheless, as they are often found along with *KRAS* and *BRAF* mutation and their incidence is low, it has been challenging to establish *PIK3CA*'s role as a biomarker in CRC.

2. Epigenetics

Identical genetic information is shared within all somatic cells from the same organism; however, genetics is not sufficient to explain the diversity of observed phenotypes. The concept of epigenetics refers to the layer of information beyond the DNA sequence responsible for each cell-type unique gene expression pattern and biological function (Berger et al. 2009). Therefore, epigenetics is considered to provide a link between the genotype and the phenotype (Bell and Beck 2010).

The main mechanisms involved in establishing the epigenome include DNA methylation, histone post-transcriptional modifications, and non-coding RNAs regulation (**Figure 6**). Given the major role of epigenetics in gene expression regulation and chromatin remodeling, epigenetic changes are gaining importance to explain the underlying mechanisms in both normal physiological functions and disease conditions. Indeed, it has become clear that epigenetics plays a significant role in tumorigenesis in addition to genetic alterations.

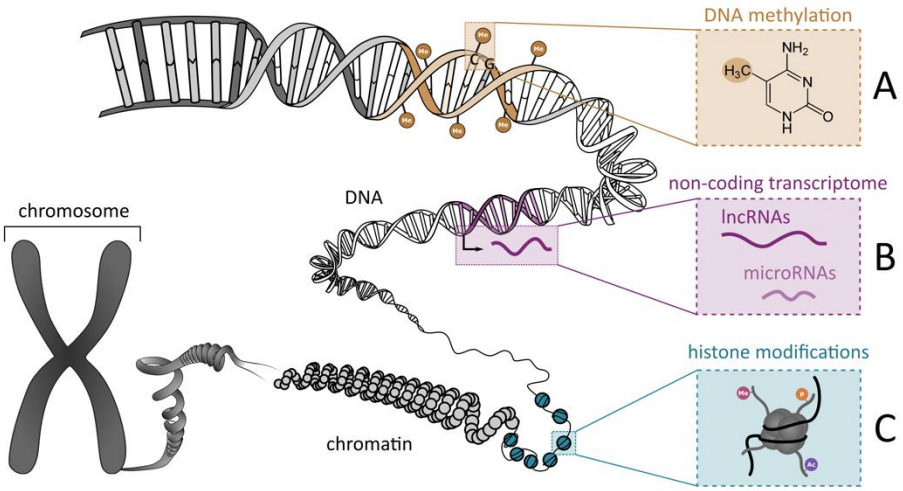


Figure 6. **Overview of DNA packaging and epigenetic mechanisms.** Chromosome, chromatin and DNA structure representation. DNA methylation (**A**), non-coding RNAs (**B**) and histone modifications (**C**) play an important role in gene regulation and chromatin conformation.

2.1. DNA methylation

So far, DNA methylation is the most well-known epigenetic mark. It consists of the covalent addition of a methyl group to the C-5 position of cytosine, resulting in a 5-methylcytosine (5mC), also known as the "fifth base" (**Figure 7**). In mammals, this phenomenon is almost exclusive of the CpG dinucleotide, but a minority has been reported to be outside the CpG context in specific cell types (Lister et al. 2009; Ziller et al. 2011).

The 5mC base is highly mutagenic by deamination, resulting in C>T transitions and representing a high proportion of the point mutations detected in many cancer types (Gehrke 1986; Poulos, Olivier, and Wong 2017). Thus, it is not surprising that CpG dinucleotides show a fivefold underrepresentation in the human genome, considering that there is a 42% GC content. It is worth noting that CpG dinucleotide distribution across the genome is not random, being most CpG concentrated in genomic regions denominated CpG islands (CpGi) and in repetitive sequences (e.g., Alus). CpGi are usually defined as regions larger than 200bp with a GC content higher than 50% and with a ratio of observed/expected CpGs equal or greater than 0.6 (M.Gardiner-Garden and M.Frommer 1987).

CpGi are generally found at the 5' regions of the genes, representing 60-70% of promoters, often associated with housekeeping, developmental and tissue-specific genes (Deaton and Bird 2011). Interestingly, promoter CpG islands are usually unmethylated during development and in most cell types allowing gene transcription. However, some CpGi become methylated in a tissue-specific manner leading to gene silencing. Indeed, this major role of DNA methylation has been well-described in developmental processes, such as X-chromosome inactivation and gene imprinting (Reik and Lewis 2005). The association between promoter methylation and gene silencing is broadly accepted, and it has been proposed that DNA methylation interferes with the binding of transcription factors (TF), causing transcriptional repression. However, it is thought that DNA methylation acts in maintaining gene repression rather than being an initiating event of repression itself (Jones 2012).

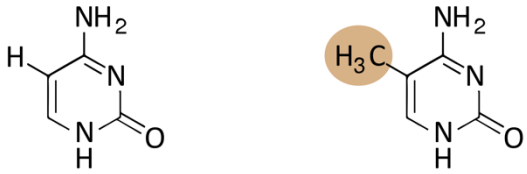


Figure 7. **Cytosine and 5-methylcytosine structures.** Two forms of cytosine bases in mammalian DNA. The 5-position of cytosine is covalently methylated, resulting in 5-mC.

Beyond promoter CpGi methylation as a silencing mechanism, other genomic contexts also display variable DNA methylation. For instance, methylation in gene bodies can also lead to transcriptional activation, playing an opposite role to promoters and leading to transcript elongation (Jones 2012; Moore, Le, and Fan 2012). CpGs in repetitive regions in the genome are heavily methylated, acting as a mechanism of protection from chromosomal instability, translocations, and gene disruption prevention (Esteller 2007). Also, it has been suggested the repressive role of DNA methylation on enhancers, which are regulatory regions situated at variable distances from promoters that regulate gene expression (Jones 2012).

2.1.1. DNA methylation regulators

DNA methylation patterns are established and maintained by a family of conserved enzymes, named DNA methyltransferases (DNMTs), which catalyze the transfer of the methyl group to DNA. There are five different types of DNMTs in mammals; however, only

three have been determined to be active in generating DNA 5mC so far: DNMT1, DNMT3a, and DNMT3b. While DNMT1 is mainly involved in maintaining DNA methylation patterns during replication, DNMT3a and DNMT3b are responsible for de *novo* methylation patterns.

Even though DNA methylation was initially considered an irreversible event, some mechanisms have been recently reported mediating DNA demethylation. The ten eleven translocation (TET) family of enzymes can reduce DNA methylation by oxidation from 5mC to hydroxymethylcytosine (5hmC) and then to other oxidized variants of methylated cytosine (Auclair and Weber 2012; Rasmussen and Helin 2016). DNMT1 might not recognize these variants during cell division resulting in a passive loss of methylation in the newly synthesized strand (Auclair and Weber 2012).

Other regulatory DNA methylation mechanisms include the role of methyl-cytosine-binding proteins (MBPs), central readers of methylated CpG that can recruit chromatin remodelers, histone deacetylases and methylases, typically causing transcriptional repression (Du et al. 2015).

2.1.2. DNA methylation in cancer cells

With all the evidence of epigenetics' fundamental role in proper cell functions, it is not surprising that epigenetics takes part in tumorigenic development in intimate cooperation with genetic events. Regarding DNA methylation alterations, a global loss of DNA methylation, as well as a substantial gain of DNA methylation in specific regions, have been reported in cancerous cells.

Global DNA hypomethylation was the first epigenetic alteration reported in cancer (Andrew P and Bert 1983; Gama-Sosal et al. 1983), almost forty years ago. Cancer cells have an estimated 20-60% less methylation levels than normal tissue (Esteller 2007). Indeed, it has been suggested global hypomethylation is an early event in cancer. For instance, in CRC, aberrant DNA methylation can already be detected in colon adenomas and adenocarcinomas (Feinberg et al. 1988; Goelz et al. 1985; Qasim, Al-Wasiti, and Azzal 2016).

The pattern of loss of methylation was thought to occur mainly in repetitive regions, but recent genome-wide methylation studies determined that it also affects gene bodies and intergenic regions through methylation changes in large domains, named cancer-specific differentially DNA-methylated regions, that cover up to half of the genome (K. D. Hansen

et al. 2011). This loss of global methylation levels is associated with the reactivation of oncogenes and retrotransposable elements.

Hypermethylation events mainly occur at specific genome sites, especially at CpGi promoters of tumor suppressor genes and genes involved in DNA repair. This hypermethylation acts as a mechanism of transcriptional repression, contributing to cancer formation. Many genes have been reported to be silenced in several cancers through promoter hypermethylation. In CRC, methylation levels of *MLH1*, VIM, and SEPT9 genes have been reported to occur at early stages and are used as early detection markers (Lao and Grady 2011).

2.2. Histone modifications

The human DNA is a macromolecule of about 2 meters long, and, in order to fit inside the nucleus, it is compacted into chromatin. Chromatin is a very dynamic structure consisting of DNA wrapped around nucleosomes, an octamer core of histones, two of each: H2A, H2B, H3, and H4 (**Figure 6.C**). Histone proteins are key regulatory players in coordinating the heterochromatin (compacted, inaccessible, and transcriptionally inactive) and the euchromatin states (uncondensed, accessible, and transcriptionally active) that lead to transcription regulation (Felsenfeld and Groudine 2003). The particular disposition of histones in the nucleosome leaves their N-terminal tail accessible for a wide variety of post-translational modifications, such as acetylation, methylation, phosphorylation, ubiquitination, sumoylation, isomerization, and ADP ribosylation, among others (Bannister and Kouzarides 2011).

Methylation and acetylation of lysine residues are the most common histone modifications with a clear role in gene regulation. Methylation marks are associated with both gene activation and silencing according to the residue that undergoes methylated (**Figure 8**)**.** For instance, H3K4me3 (histone 3 lysine 4 trimethylation) is an active mark in promoter genes (Schneider et al. 2004), while H3K9me3 (histone 3 lysine 9 trimethylation) and H3K27me3 (histone 3 lysine 27 trimethylation) are usually associated with repressed states (Barski et al. 2007). Also, H2K4me1 (histone 2 lysine 4 monomethylation) is a mark for active enhancers and H3K36me3 (histone 3 lysine 36 trimethylation) is enriched in transcribed regions (Barski et al. 2007). Unlike methylation, histone acetylation is generally associated with gene activation. For example, H3K9ac (histone 3 lysine 9 acetylation) is located in promoter regions, together with H3K4me3, associated with positive marks of transcription, and H3K27ac (histone 3 lysine 27 acetylation), which is

found at both enhancers and promoters, being associated with active transcription of neighboring genes.

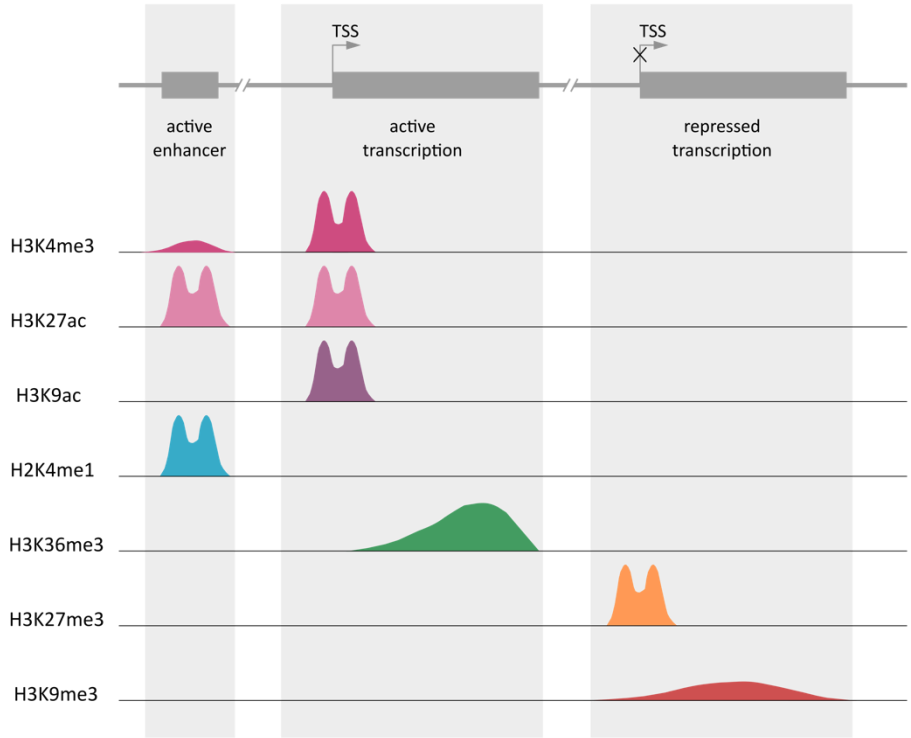


Figure 8. Distribution of histone modifications involved in transcription. Chromatin states are defined by combinations of different histone modifications, among other events. H3K27ac and H2K4me1 occupancy usually indicates enhancer state. Active transcription is usually characterized by H3K4me3, H3K27ac and H39ac active promoter marks and H3K36me3 in actively transcribed sites. H3K27me3 and H3K9me3 are usually found in repressed states.

Histone modifications tend to co-localize rather than being placed alone. As mentioned before, H3K9ac and H3K4me3 are usually found together in active promoters, while H3K9me3 and H3K27me3 co-localize in repressed regions (**Figure 8**). Considering the variety of amino acid residues in histone tails and all the possible modifications, the potential combinations are countless. Therefore, the reported co-localization of histone marks leads to the "histone code hypothesis", which states that the combination of histone marks at certain genomic loci constitutes an informative code associated with the activity of the underlying gene (Jenuwein and Allis 2001). However, this hypothesis is still in debate, as there is no clear picture of how all these modifications and their combinations can predict the transcriptional state of genes.

2.3. Non-coding RNAs

For decades, cancer research was mainly focused on protein-coding genes that were thought to be the main macromolecules of the cell with important functional roles (except infrastructural RNAs, like ribosomal RNA). At the same time, the non-coding portion of the DNA was considered "junk DNA". However, the rapid development of high-throughput sequencing technologies and computational platforms revealed transcriptional events in regions that did not appear to code for proteins (Djebali et al. 2012), suggesting that the transcriptional landscape is way more complex than initially thought. Surprisingly, studies have determined that, although about 80% of the human genome is transcribed, less than 2% encode for proteins, suggesting that the vast majority are non-coding RNAs (ncRNAs) (Djebali et al. 2012; Dunham et al. 2012). The recent discovery of a large number of ncRNAs has revolutionized the field by raising doubts about the biological relevance and functional role of these non-coding transcripts.

There are several types of ncRNAs that can be divide into housekeeping and regulatory ncRNAs (**Figure 9**). Housekeeping ncRNAs are, in general, abundant and regulate general cellular functions and cell viability, such as the well-known rRNAs and tRNAs. In contrast, regulatory ncRNAs can be further classified into short non-coding RNAs (sncRNAs) or long non-coding RNAs (lncRNAs) according to their size being smaller or larger than 200nt, respectively. They play key roles in gene regulation at the epigenetic, transcriptional, and post-transcriptional levels (P. Zhang et al. 2019). The most well-known subtype of short RNAs are microRNA, mature transcripts of 18-22nt, that have been extensively investigated in multiple human diseases, mainly in cancer. The lncRNAs are a less-understood subtype but are getting increasing attention, as described below.

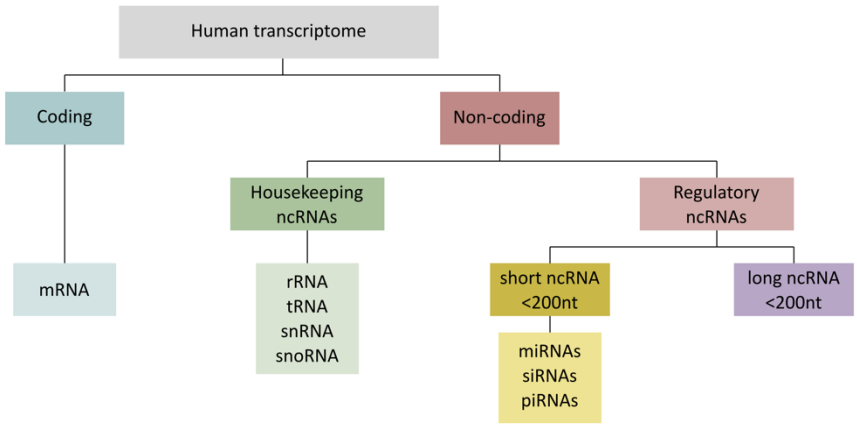


Figure 9. **Non-coding RNAs classification.** Schematic diagramillustrating the classification of the human transcriptome. Non-coding RNAs are divided according to their biological role and length.

2.3.1. LncRNAs

LncRNAs are RNA transcripts longer than 200 nucleotides with little or non-coding potential. According to the LNCipedia current release (version 5.2), 56,946 lncRNA genes have been identified in the human genome, yielding 127,802 lncRNA transcripts (Volders et al. 2019).

LncRNAs are very diverse in function, expression, localization, and size. A wide variety of classifications have been proposed to date according to several criteria, such as transcript length, genomic location, or functionality (Laurent, Wahlestedt, and Kapranov 2016). One of the most prevalent classifications is based on the genomic location relative to protein-coding genes, that divides lncRNAs into five classes: sense-overlapping, antisense, intronic, intergenic and bidirectional (Balas and Johnson 2018; Hermans-Beijnsberger, van Bilsen, and Schroen 2018; Rinn and Chang 2012; Volders et al. 2019) (**Figure 10**). Sense lncRNAs are transcribed in the same direction as a protein-coding gene overlapping one or more exons, thus they can be considered transcript variants of protein-coding mRNAs. Most of them lack substantial ORFs. On the contrary, antisense lncRNAs are transcribed from the opposite strand of a protein-coding gene by at least overlapping one exon. The expression levels of these antisense lncRNAs and the corresponding sense coding mRNA have been reported to be both positive and negatively correlated (Katayama et al. 2005). Intronic lncRNAs initiate and finish their transcription inside introns of protein-coding genes in either sense or antisense orientation, but without overlapping exons. In general, they are regulated by the same transcriptional mechanism as the protein-coding gene. Intergenic lncRNA (lincRNAs) are transcribed in between protein-coding genes, and many of them have been described as cis-acting chromatin regulators (Ulitsky and Bartel 2013). Bidirectional lncRNAs, also known as divergent lncRNAs, are transcribed in opposite directions of a protein-coding gene, usually in head-to-head disposition. Interestingly, bidirectional lncRNAs usually display a high coordinated expression with their paired protein-coding gene (Sigova et al. 2013), as they share the same promoter.

Most annotated lncRNAs resemble mRNA as they are often Pol II transcribed, 5’ capped, 3’ polyadenylated, subjected to splicing mechanisms, and transported to the cytoplasm (Dunham et al. 2012; Quinn and Chang 2016). In contrast, other described lncRNAs remain in the nucleus, lack post-transcriptional modifications, and undergo rapid degradation. These differences in lncRNA biogenesis suggest that they have a wide variety of functional roles in human cells. Unfortunately, to date, only a tiny portion of lncRNA has been well-characterized. While some lncRNAs have been hypothesized to be by-products of the transcription machinery or transcriptional noise, many others have been reported to act as functional regulators in diverse biological processes, including cancer disease (Chen,

Fan, and Song 2016; X. Hu et al. 2018; Slack and Chinnaiyan 2019). Furthermore, lncRNAs display lower expression levels than mRNAs, but they exhibit more specific expression patterns at subcellular, cellular, and tissue levels (Cabili et al. 2011; Djebali et al. 2012).

Based on the location where lncRNAs act, they can be further classified into cis- and trans-acting transcripts if they act locally or away from their site of transcription, respectively. Cis-regulation by lncRNAs can influence the expression and chromatin state of neighboring genes. A well-known cis-acting lncRNA is Xist, which is involved in the X chromosome inactivation in mammals (Pontier and Gribnau 2011). In addition, other lncRNAs can leave their site of transcription and regulate gene expression in distant locations of the genome or in the cell cytoplasm, such as HOTAIR, which is able to recruit chromatin-modifying complexes at a distant locus (Rinn et al. 2007). Other examples of trans-acting lncRNAs that play structural roles in the nuclear architecture are NEAT1, which interacts with proteins associated with transcription and RNA processing in the nucleus (Clemson et al. 2009), and *MALAT1*, which can recruit proteins involved in splicing processes (Bernard et al. 2010). Also, lncRNAs can regulate other RNAs and proteins, such as ciRS-7 that functions as a molecular sponge of microRNAs (T. B. Hansen et al. 2013), or NORAD that acts as an inhibitor of PUM1 and PUM2 proteins (Tichon et al. 2016).

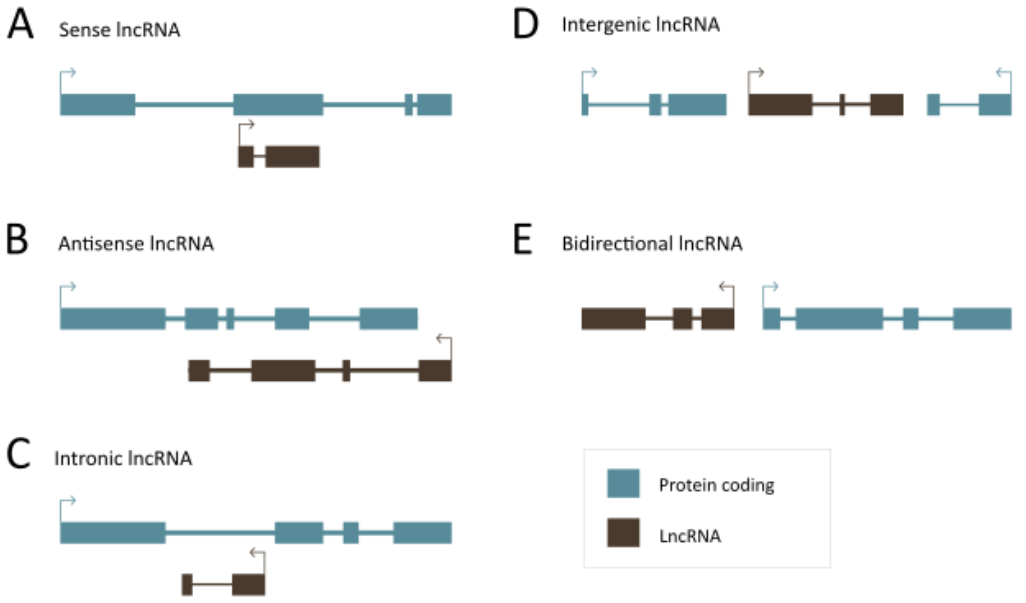


Figure 10. **LncRNAs classification according to their genomic context.** On the basis of genomic location and orientation relative to protein coding genes, lncRNAs can be classified into sense (**A**), antisense (**B**), intronic (**C**), intergenic (**D**) and bidirectional (**E**).

2.3.2. lncRNAs in cancer

With the increasing studies on lncRNA, dysregulated lncRNAs in several cancers have been reported to play critical roles in the regulation of the malignant progression, either by acting as tumor suppressors or as oncogenes. Specifically, data from several research studies has confirmed the involvement of some lncRNAs in CRC development. A classic example is the oncogene role of H19 that acts by several mechanisms, like targeting tumor suppressor retinoblastoma protein (RB) in CRC (Tsang et al. 2010). Many other lncRNAs have been demonstrated to be involved in colorectal carcinogenesis, such as HOTAIR, *MALAT1*, MEG3, and CCAT1 (reviewed in Siddiqui et al. 2019; Ye et al. 2015), among many others.

The plasticity of lncRNAs to bind RNA, DNA and/or proteins reveals their functional diversity in gene regulation at transcriptional, post-transcriptional, and epigenetic levels. In consequence, the clinical applications of lncRNAs are wide open. As they typically show expression-specific patterns in cells and tissues, they can be used as potential biomarkers in cancer or as targets for treating the disease. However, lncRNA research is a relatively new area of study. There are critical gaps in our knowledge trying to understand their mechanisms of action and their specific patterns of expression and regulation. Consequently, further studies need to be performed to better optimize their use as potential biomarkers or drug targets.

3. FOXD2 and FOXD2-AS1

3.1. Forkhead-box (FOX) Transcription Factor Family

The Forkhead-box (FOX) proteins are a family of evolutionarily conserved transcription factors characterized by a common DNA-binding domain of ~100 amino-acids, termed “forkhead box” or “winged helix” domain. They were first discovered in *Drosophila melanogaster* (Weigel et al. 1989) and since then, the family has expanded including large numbers of proteins in mammals that have been further classified into nineteen subclasses, from FOXA to FOXS.

Although high analogous DNA-binding domain and similar recognition motifs are shared between forkhead proteins, they display a wide variety of functions that could be explained in part by the distinct sequences surrounding the common binding domain that lead to different patterns of expression and post-translational modifications (reviewed in Golson and Kaestner 2016). They are involved in development and tissue-specific

functions, and many studies have associated their abnormal regulation with several human diseases, including cancer. In 2007, a snapshot of human FOX TF was published, including their regulatory roles, involvement in cellular and developmental processes, in human disease, and their known mouse phenotypes (Tuteja and Kaestner 2007).

As mentioned before, FOX protein deregulation is often associated with cancer development and progression. Among the most studied members, FOXA1 has been associated with breast cancer development acting as a co-factor for ER (Carroll et al. 2005), FOXA2 tumor suppressor role is involved in lung (Tang et al. 2011), breast (Z. Zhang et al. 2015) and liver cancers (Jian Wang et al. 2014) and FOXM1, a master regulator of cell cycle, has been described to play a role as an oncogene (Laoukili, Stahl, and Medema 2007). FOXO proteins have been described to act as tumor suppressors, whereas FOXO1A loss is observed in prostate cancers (X. Y. Dong et al. 2006), and FOXO3 has been linked to lung cancer (Cheng et al. 2015). Finally, FOXD1 has been associated with many types of cancer, including gastric (Feng et al. 2015), colorectal (Pan, Li, and Chen 2018), glioma (Gao et al. 2017), and lung, among many others.

3.2. FOXD2

FOXD2 TF belongs to the FOXD family subclass, which also includes FOXD1, FOXD3, FOXD4, FOXD4L1, FOXD4L2, FOXD4L3, FOXD4L5, and FOXD4L6 proteins. So far, FOXD1 is the most well-characterized member of the FOXD subclass, followed by FOXD3. On the contrary, very little is known about FOXD2 and the rest of FOXD members.

FOXD2, also known as FREAC-9 or FKHL17, was cloned and characterized for the first time in the human kidney in 1997 (Ernstsson et al. 1997). They reported its involvement in kidney development; however, a FOXD2 deficient mice model revealed mild abnormalities in the organ development (Kume, Deng, and Hogan 2000). Another study described FOXD2 role in leukocytes, whereas was determined to regulate RIα expression in T cells (Johansson et al. 2003). Regarding cancer disease, FOXD2, together with FOXD1, was shown up-regulated in prostate cancer samples and in lymph node metastases compared to normal prostate tissues (Van Der Heul-Nieuwenhuijsen, Dits, and Jenster 2009), and it was found deleted in chromosome 1 locus in meningioma (Sulman, White, and Brodeur 2004). Interestingly, FOXD2 3'UTR was reported differentially methylated in colorectal serrated adenocarcinomas compared to conventional adenocarcinomas and was further correlated with expression changes (Conesa-Zamora et al. 2015). The role of FOXD2 in cancer needs further investigation, as no recent studies aim to better characterize the precise functions and mechanisms of FOXD2.

3.3. FOXD2-AS1

FOXD2-AS1 (FOXD2 adjacent opposite strand RNA1) is a lncRNA in bidirectional disposition towards FOXD2. Unlike FOXD2, FOXD2-AS1 has been widely characterized due to its expression deregulation in several human cancers. Interestingly, most studies have been performed in the last five years, and the first report about FOXD2-AS1 was published in 2016, describing its upregulation in human gastric cancer (C. Y. Li et al. 2016). Since then, many other studies have revealed its upregulation in human cancer, including colorectal cancer (Yang, Duan, and Zhou 2017; Zhu et al. 2018), non-small cell lung cancer (Rong, Zhao, and Lu 2017), breast cancer (Jiang et al. 2019), glioma (H. Dong, Cao, and Xue 2019; Jin Wang et al. 2019), thyroid cancer (Yayuan Zhang et al. 2019), among many others (Bao et al. 2018; Q. Hu, Tai, and Wang 2019; Ren et al. 2019; Yang, Duan, and Zhou 2017). According to these studies, high expression levels of FOXD2-AS1 lead to increased proliferation, migration, and invasion, suggesting it has an oncogenic role. Furthermore, it acts by sponging tumor suppressor microRNAs, being miR-185-5p the most well studied (H. Dong, Cao, and Xue 2019; Zhu et al. 2018), but many other have been reported, such as miR-27a-3p and miR-31 (Jin Wang et al. 2019; Y. Wang et al. 2019). In addition, FOXD2-AS1 expression may have clinical significance, as high expression levels are associated with poorer prognosis in several cancer types (H. Dong, Cao, and Xue 2019; Jiang et al. 2019; Mao et al. 2020; Su et al. 2018).

Four studies have revealed FOXD2-AS1 upregulation in CRC compared to normal tissues, demonstrating its role in proliferation migration and invasion through *in vitro* assays. As well, it has been associated withmiR-185-5p (Zhu et al. 2018), Sema4C and miR-25-3p (M. Zhang et al. 2019), miR‑4306 (J. Ye et al. 2021) and its role as regulator of the EMT and Notch signaling pathway has been described (Yang, Duan, and Zhou 2017). Despite FOXD2-AS1 being well characterized in several cancer types, its mechanism of action and clinical significance are unknown and need to be further investigated.

OBJECTIVES

Colorectal cancer emerges through the progressive accumulation of genetic and epigenetic alterations that evolve to malignancy. CRC is one of the cancers with the highest incidence and a leading cause of cancer-related deaths worldwide. Early cancer detection and the reliable identification of CRC biomarkers are permanent challenges aiming to increase patient survival rates. Motivated by these premises, we addressed mechanistic and clinical aspects of colorectal tumorigenesis with two global aims:

- To investigate the deregulation and function of FOXD2 and its antisense long non-coding transcript FOXD2-AS1 in CRC tissues and cell lines (**study I**).

- To bring light into the contribution of precancerous somatic mutations in CRC tumorigenesis by ultra-deep sequencing of healthy colon mucosa from patients with and without CRC (**study II**).

Specifically, for each study we propose the following specific aims:

Study I:

- To characterize the transcriptomic and epigenetic profiles of FOXD2 and FOXD2-AS1 in CRC.
- To gain insights into the prognostic value of FOXD2 and FOXD2-AS1 expression and methylation profiles.
- To explore the functional involvement of FOXD2 and FOXD2-AS1 in cancer cell biology.

Study II:

- To determine the presence of somatic mutations in common CRC genes (*BRAF*, *KRAS*, *PIK3CA*, and *TP53*) in normal colorectal tissue.
- To explore differences in mutations in the normal colon of patients with and without CRC.
- To evaluate the potential of the mutational profiling in non-tumor colon biopsies for CRC prediction.

MATERIALS AND METHODS

Methods Study I

FOXD2 and FOXD2-AS1 in colorectal cancer

All experimental procedures from **study I** have been performed at the "Epigenetic Mechanisms of Cancer and Cell Differentiation" department led by Miguel Ángel Peinado, at the Institut Germans Trias i Pujol (IGTP), Badalona.

1. Samples

1.1. Patients

In this study, 110 CRC tissues and matched adjacent non-tumor epithelial tissues were provided from Biobanc HUB-ICO-IDIBELL (Hospital Universitari de Bellvitge, HUB) following the legislation guidelines and with the approval of the local ethics committees. Clinico-pathological characteristics of colorectal tumors are summarized in **Table 2**. Patients had an average age of 67 years old (from 33 to 88). All patients were on cancer stages II and III, with no distant metastasis at the time of collection and with high survival rates. Fresh frozen tissues were stored at -80ºC after collection up until DNA and RNA extractions (see methods 2.1 and 3.1, respectively).

Table 2. Clinico-pathological characteristics of patients (HUB cohort). NA, not available.

Variables	Cathegories	Patients
Sex	Females	50 (46.3%)
	Males	58 (53.7%)
Cancer Stage	II	60 (55.6%)
	III	48 (44.4%)
Tumor size and invasion	T1+T2	5 (4.6%)
	T3	78 (72.2%)
	T4	25 (23.1%)
Lymph node involvement	N0	60 (55.5%)
	N1	29 (26.9%)
	N2	19 (17.6%)
Distant methastasis	M0	107 (99.1%)
	NA	1 (0.9%)
Relapse	yes	18 (16.7%)
	no	90 (83.3%)
Colon location	distal	61 (56.5%)
	proximal	47 (43.5%)
Survival	yes	97 (89.8%)
	no	11 (10.2%)

1.2. Human cell lines

The three human colorectal cell lines HCT116, LoVo, and SW480 used in this study were obtained from the American Type Culture Collection (ATCC, Manassas, Virginia) (**Table 3**). Cells were grown at 37ºC and 5% of CO_2 in DMEM/F12 (Dulbecco's Modified Eagle Medium: F12) supplemented with 10% of inactivated fetal bovine serum (FBS) (Ref.10270106), 20µM of L-glutamine (Ref.25030024) and 10µM of pyruvate (Ref. 11360039). Packaging 293T cells, obtained from ATCC, were maintained in DMEM (Dulbecco's Modified Eagle Medium) supplemented with 10% FBS, 20µM of L-glutamine, and 10µM of L-glutamine pyruvate (all products from GBICO® Invitrogen, Carlsbad, CA).

Cell lines were kept in culture without antibiotics and checked periodically to ensure *Mycoplasma*-free conditions.

Table 3. List of human colorectal cancer cell lines used.

Cell line	Cancer type	Karyotype	Genomic instability	Mutations	ATCC code
HCT116	Human colorectal carcinoma	Near diploid	MSI	*KRAS*	CCL-247
LoVo	Human colorectal adenocarcinoma, Dukes' type C, grade IV	Hyperdiploid	MSI	*KRAS*, MYC, *TP53*	CCL-229
SW480	Human colorectal adenocarcinoma, Dukes' type B	Hypotriploid	CIN	*KRAS*, MYC, *TP53*	CCL-228

2. RNA analysis

2.1. RNA extraction

Tissue RNA extractions were performed using the Pure Link™ RNA Mini Kit (Ref. 12183018A, Ambion, Life Technologies), including On-Column PureLink® DNase treatment. Cell lines RNA were extracted with Pure Link™ RNA Mini Kit or Maxwell® 16 LEV simplyRNA Cells Kit (Ref.AS1270, Promega). All procedures were done according to manufacturer instructions.

RNA concentration and purity were assessed using NanoDrop™ spectrophotometer (ND-1000, Thermo Fisher Scientific, Massachusetts, USA) and RNA integrity by 18S and 28S ribosomic observation on 1% agarose gel electrophoresis.

2.2. RNA Fractionation

RNA subcellular fractionation protocol was adapted from Dumbović *et al*. Cell line pellets (~2M cells) were lysed with 175 µL/10^6 cells of cold home-made buffer (50 mM Tris–HCl pH 8.0, 140 mM NaCl, 1.5 mM MgCl2, 0.5% NP-40, 2 mM Vanadyl Ribonucleoside Complex (NEB: S1402S)) and incubated 5 min on ice. After 2 min of 300g centrifugation at 4ºC, supernatant (cytoplasmic fraction) and pellet (nuclear fraction) were used for RNA extraction with Pure Link™ RNA Mini Kit, according to manufacturer's protocol. On-Column PureLink® DNase treatment was included for each extraction. RNA concentration and integrity were assessed as previously described (see methods 2.1).

2.3. Reverse Transcription

Synthesis of cDNA was carried out using SuperScript® IV First-Strand Synthesis System (Ref. 18091050, Invitrogen, Massachusetts, USA) following the manufacturer's protocol. Briefly, 500ng of RNA were brought up to 10uL with water, mixed with 150ng of random hexamers (or 50µM oligo d(T) primers when indicated), 10mM of dNTPs and incubated 5 min at 65ºC and 1 min on ice. Then, 4uL of 5x SSIV Buffer, 1uL of 100mM DTT, 40 units of RNase inhibitor (Promega, Wisconsin, USA), and 200 units of SSIV reverse transcriptase were added before incubation. Each experiment included a negative control without RNA and a negative control without reverse transcriptase enzyme. Reactions were incubated in a PCR thermocycler for 10 min at 23ºC, 10 min at 55ºC, and 10 min at 80ºC. Obtained cDNA products were stored at -20ºC.

2.4. Real-Time Quantitative PCR (qPCR)

Transcript expression levels were quantified by real-time qPCR. Retrotranscribed cDNA was diluted with water 5 to 10 fold before amplification. Each sample was analyzed in triplicates in a LightCycler® 480 platform with LightCycler® 480 SYBR® Green I Master in a final volume of 10uL (Roche Life Science, Penzberg, Germany).

Chainy web tool designed in our lab (maplab.imppc.org/chainy/) was used to calculate the efficiency of individual reactions (cqD2 method) and to select the best reference genes for each experiment. The average efficiency for each gene was calculated prior to data analysis. Normalization was performed using at least two reference genes on the LightCycler® software (Roche Life Science).

We used Primer-BLAST tool (www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi) for primer design. When possible (not for FOXD2 and FOXD2-AS1), primers were designed within exons flanking one intron to avoid gDNA amplification products. Primer properties

were evaluated with IDT Oligo Analyzer tool (https://eu.idtdna.com/calc/analyzer), such as CG content, melting Tº, and hetero-dimer formation. In addition, primer specificity was assessed *in silico* by running the UCSC In-Silico PCR tool (https://genome.ucsc.edu/cgi-bin/hgPcr). Primer sequences used for expression analysis are listed in **Table 4**. All primer annealing temperatures were at 62ºC.

Table 4. List of primer sequences used for PCR analysis. *primers used for conventional PCR. The rest for qPCR.

Gene name	Forward (5' to 3')	Reverse (5' to 3')	Size (bp)
Targets			
FOXD2	CTGACGTTGAGCGAGATCTG	GGGATCTTGACGAAGCAGTC	125
FOXD2-AS1_1	CGTGTAACCCTTCTGAGTCC	CCCTGGCTTTGCTTCTATGAG	137
FOXD2-AS1_2	GAGAAATCTGCGGGCGTAGT	GATGCCTGTTGGGCTTTTCC	335
FOXD2-AS1_3	CTGTAACCAAGACCCGCAGAG	ACCGCGGGATTCGGAATTTAT	234
FOXD2-AS1_4	GTTCTGGGCTAGGAACCCG	ACTTGCTGCCCAAATTTCCTG	296
FOXD2-AS1_5	GGTCCATGGTGTGGGGTATC	CTGTCCGGGGAAAAAGGTCT	193
FOXD2-AS1_A*	CCAGCGATTATGCGGATCTAA	CCCTGGCTTTGCTTCTATGA	1607
FOXD2-AS1_B*	AAATCCCTGCTCCAGTCCT	CTCTCAGTTTCCTCCTGCATTC	503
References and controls			
GAPDH	ACATCGCTCAGACACCATG	ATGACAAGCTTCCCGTTCTC	222
MALAT1	AAGGTCAAGAGAAGTGTCAGC	AATGTTAAGAGAAGCCCAGGG	125
MRPL9	CAGTTTCTGGGGATTTGCAT	TATTCAGGAGGGCATCTCG	197
PSMC4	TGTTGGCAAAGGCGGTGGCA	TCTCTTGGTGGCGATGGCAT	182
PUM1	CGGTCGTCCTGAGGATAAAA	CGTACGTGAGGCGTGAGTAA	121

2.5. Conventional PCR

Conventional PCRs were performed using Phusion High-Fidelity DNA polymerase (Thermo Fisher Scientific). PCRs were done following the manufacturer's instructions using the 5x Phusion GC Buffer, without DMSO, and with 1:5 diluted cDNA in a final volume of 20µL. Amplified PCR products were checked on 2% agarose gel electrophoresis, purified with ExoSAP-IT™ PCR Product Cleanup Reagent (Thermo Fisher Scientific), and analyzed by Sanger sequencing (GATC Biotech Service, Germany). Annealing temperatures were 60ºC and 60-62ºC for PCRs A and B, respectively. Primers are included in **Table 4**.

2.6. RNA FISH

The subcellular localization of FOXD2-AS1 transcript in CRC cells was identified using RNA FISH (fluorescence *in situ* hybridization) according to the provider instructions with minor

changes (Stellaris®, Biosearch Technologies, Inc., Petaluma, CA). Briefly, cells were grown on 12 mm Ø round coverslips for 24h. After fixation for 10 minutes, cells were immersed in ice-cold 70% ethanol for up to 2 hours at 4ºC. Then, cells were washed and hybridized inside a humidified chamber overnight (<16h) at 37ºC onto 50-100uL of hybridization buffer containing the probe. Cells were washed and incubated with DAPI nuclear stain for 30 min at 37ºC in the dark. Finally, coverslips were mounted with cells side down onto a microscope slide with Vectashield Mounting Medium and sealed with clear nail polish. Images were obtained and processed with Zeiss Axio AxioObserver Z1 wide-field fluorescence microscope (63x objective).

Custom Stellaris® FISH Probes were designed against FOXD2-AS1 (NR_026878.1) by using the Stellaris® RNA FISH Probe Designer (Biosearch Technologies, Inc., Petaluma, CA) (**Supplementary Table S1**). Stellaris® FISH Probes recognizing *MALAT1* and GAPDH were kindly provided by Dr. Sonia Forcales (Universitat de Barcelona). All probes were labeled with Quasar 570 and stored at -20ºC upon arrival.

3. DNA methylation analysis

3.1. DNA extraction

Genomic DNA extractions were done using PureLink® Genomic DNA Mini Kit (Invitrogen) following manufacturer instructions. DNA quantification and purity were evaluated by NanoDrop™ and integrity was checked by 1% agarose gel electrophoresis.

3.2. Bisulfite conversion

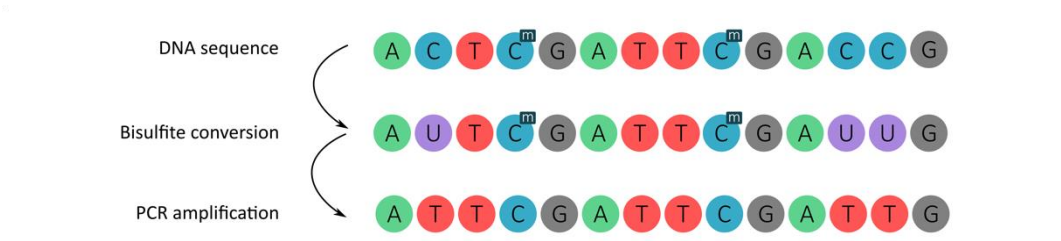


Figure 11. Bisulfite conversion scheme. Deamination of unmethylated cytosine (C) by sodium bisulfite treamtent leads to uracil (U), while methylated cytosines remain unaffected. Uracil is further converted to thymine (T) during PCR amplification.

Bisulfite conversion is the current gold standard method for DNA methylation analysis, providing high coverage at single-base resolution. Sodium bisulfite treatment results in the deamination of unmethylated cytosines to uracil (**Figure 11**), while methylated cytosines remain unmodified (Clark et al. 2006). EZ DNA Methylation Gold™ Kit (ZymoResearch) was used for bisulfite conversion of ~300ng of genomic DNAs following the manufacturer's protocol. Bisulfite-converted DNA was eluted in 40-50µL of water.

3.3. Direct bisulfite sequencing

Bisulfite-converted DNA was used as a template to amplify our CpG of interest. First, amplification with 1µL of converted DNA was performed by conventional PCR method using Taq DNA Polymerase (Roche Life Science, Penzberg, Germany) (external PCR). PCR products (1/10 diluted if necessary) were used for a subsequent nested PCR (internal PCR). All reactions were performed in duplicate. After amplicons were checked on 2% agarose gel electrophoresis, they were pooled, purified with ExoSAP-IT™ PCR Product Cleanup Reagent (Thermo Fisher Scientific), and analyzed by Sanger sequencing (GATC Biotech Service or Macrogen Service). The methylation degree was calculated by comparing the peak height of the cytosine residues with the peak of the thymine residues [C/(C+T)*100] in the Sanger sequencing chromatogram. Results were represented using the Methylation plotter tool (http://maplab.imppc.org/methylation_plotter/). Bisulfite primers, designed with MethPrimer (https://www.urogene.org/methprimer/), annealing temperatures, and PCR conditions are listed in **Table 5**.

Table 5. List of primers used to study DNA methylation at the FOXD2 and FOXD2-AS1 locus on chr 1. Ext., external; Int., internal.

Region	PCR	Forward (5' to 3')	Reverse (5' to 3')	Size (bp)	Annealing Tº
1	Ext.	TTTTAAGGTTTTGGGTTAGTTT	CAACTCATTTATAAAAACCAAAA	525	54/56
	Int.	GGGATTGGGAGAAGGGTTAT	ATAAAAAAACCCAACAAACATCC	379	60
2	Ext.	AGATAGTTATAGAGATTGAG	CACCCTATACTCCCTAAA	495	50/52
	Int.	ATTTTTTTTAGGTTTAAGGTTG	CCCTAAATATTAAAACTCACT	264	54
3	Ext.	GTTTTGGTTATGTTGATTGT	TAAAACCTAACCAAACATCT	479	50/52
	Int.	GGTTTTAGTGGTTGTTATTTT	CCCCTACTTTTATTTCTCAA	288	54

4. Protein analysis

4.1. Protein extraction

Protein extraction was performed by resuspending cell pellets RIPA buffer, which contained 150mM NaCl, 1% Triton X-100, 0.5% Deoxycholate, 0.1% SDS and 50mM Tris pH 8 and was supplemented with fresh protease inhibitors (1M DTT, 5mg/mL Aprotinin, 0.2M PMSF, 100mM Sodium Orthovanadate and 500mM Sodium fluoride). Lysis was allowed 30 min on ice, and then protein extracts were obtained by collecting the supernatant after a centrifugue of 30 min at 4ºC at 13.000rpm. Proteins were quantified using the Pierce™ BCA Protein Assay Kit (Thermo Fisher Scientific), following the manufacturer's instructions.

4.2. Western blot

Total protein extracts were separated by running 10-60ug in 10% SDS-polyacrylamide electrophoresis and electrotransferred to polyvinylidene difluoride membranes (Millipore, Massachusetts, USA). Total levels of transferred protein were detected using Sypro Rubi protein blot stained following the manufacturer's protocol (ThermoFisher Scientific). The membranes were blocked in Intercept™ Blocking buffer blocking (TBS) (Li-COR Biosciences, NE, USA.). and probed with the primary antibody overnight at 4ºC (FOXD2 #ab49128 or #ab104411, Abcam, UK). Then, membranes were incubated with fluorescent secondary antibody (#926-32211, Li-COR Biosciences) and visualized on an Odyssey CLx Imaging System (Li-COR Biosciences).

5. Cell culture experimental procedures

5.1. Maintenance and collection

Cell lines were divided three times per week by trypsinization (dilutions 1:4 to 1:10). Briefly, cells were washed with 1x PBS, trypsinized for 2-5 minutes at 37ºC, centrifuged for 5 min at 300g, and resuspended with fresh medium. When collected, pellet cells were resuspended in cold 1x PBS after centrifugation, centrifugated an additional 5 min at 4ºC and 300g, and stored at -80ºC.

5.2. 5-aza-2'-deoxycytidine (DAC) treatment

Cell lines were treated with demethylation drug DAC (A3656 Sigma-Aldrich, Missouri, USA) to investigate the role of DNA methylation in gene expression regulation. DAC is a pyrimidine analog of cytidine that is incorporated into DNA causing DNA damage and

depletion of DNA methyltransferases (DNMTs) (Sheikhnejad et al. 1999), resulting in global hypomethylation in a replication-dependent manner.

Depending on the cell type, 1-1.5 x 10^6 cells were seeded into 100mm plates to reach a confluence of 50-70% at 24h. Then, cells were treated with 0.5µM (HCT116 and SW480) or 1µM (LoVo) DAC for 48h by renewing the medium with drug every 24h. After 48h, fresh medium without DAC was added to let cells rest for 24h before cell collection.

5.3. CRISPR SAM genome editing

FOXD2 and FOXD2-AS1 overexpression was first performed using the CRISPR/Cas9 SAM method based on the use of a single guided RNA (sgRNA) designed according to the gene of interest, which targets the catalytically inactive Cas9-VP64 bounded to MS2-P65-HSF1 activation helper protein to increase gene expression near the TSS (Konermann et al. 2014).

First, cell lines were infected with lentiviruses containing the catalytically inactive Cas9-VP64 (addgene #61425) and the MS2-P65-HSF1 activation helper protein (addgene #61426). Once cells with stable integration of those vectors were generated, we proceeded to infect cells with lenti sgRNA(MS2)_zeo backbone (addgene #61427) vector containing the corresponding sgRNA of interest.

5.3.1. sgRNA cloning

The sgRNA sequences were designed using the Broad Institute GPP sgRNA Design tool (https://portals.broadinstitute.org/gpp/public/) and off-target prediction was assessed with Cas-OFFinder tool (www.rgenome.net/cas-offinder/). Six sgRNA were designed for FOXD2 and six for FOXD2-AS1, respectively (**Table 6**).

One microgram of lenti sgRNA(MS2)_zeo backbone (addgene #61427) was digested with Fast Digest Esp3I enzyme (Fermentas #FD0454) in a final volume of 20 µl 2 hours at 37ºC. Meanwhile, each pair of sgRNA were hybridized and phosphorylated with T4 PNK (NEB #M02015) using T4 Buffer 10X (Fermentas #EL0011) by 30 min incubation at 37ºC, 5 min at 95ºC and then ramped down to 25ºC at -5ºC/min. Ligation of sgRNAs (diluted 1/100) and 60ng of the digested vector was performed using T4 ligase as described by the manufacturer (Fermentas #EL0011). Finally, ligated products were transformed in 40 µl of Stbl3 E. coli strain by heat shock method and 200 µl of transformed bacteria from each sgRNA were seeded on a selective LB/agar plate with ampicillin and incubated overnight at 37°C. Colonies were picked and sgRNA insertion was verified by PCR amplification (Fw: 5'-GAGGGCCTATTTCCCATGAT-3' and Rv: 5'-CGGTGCCACTTTTTCAAGTT-3') and Sanger

sequencing. Positive colonies were grown overnight on 4mL of LB supplemented and miniprep was performed following the manufacturer's instructions (NucleoSping® Plasmid, Machery-Nagel). The presence of plasmid was determined by electrophoresis in a 2% agarose gel and Sanger sequencing was performed (GATC Biotech) to check the minipreps.

Table 6. sgRNA sequenced used for CRISPR SAM. Sequences and overhangs used for CRISPR SAM experiments in SW480 cell line.

Target gene	sgRNA ID	TSS (hg18)	TSS distance (bp)	Forward (5' to 3') & Reverse (3' to 5')	PAM	strand
FOXD2	a	47436017	-126	Fw caccgCTGTCCGGGGAAAAAGGTCT Rv aacAGACCTTTTTCCCCGGACAGc	TGG	+
	b	47436017	-79	Fw caccgTGGACAGGGACTAGTAGCCC Rv aacGGGCTACTAGTCCCTGTCCAc	TGG	+
	c	47436017	-100	Fw caccgACCCAGGAAATTCCAATTCC Rv aacGGAATTGGAATTTCCTGGGTc	AGG	-
	g	47438044	-85	Fw caccgGGGGACTGAGGCAGGCAGGG Rv aacCCCTGCCTGCCTCAGTCCCCc	AGG	-
	h	47438044	-130	Fw caccgAAGGTGAGCGCGGCCGAGCT Rv aacAGCTCGGCCGCGCTCACCTTc	GGG	+
	i	47438044	-149	Fw caccgCCGGCGGGTCGTGCCCTGGA Rv aacTCCAGGGCACGACCCGCCGGc	AGG	+
FOXD2-AS1	d	47434641	-142	Fw caccgAACTGGCCCAGAACCTTGGA Rv aacTCCAAGGTTCTGGGCCAGTTc	AGG	+
	e	47434641	-146	Fw caccgGGCCCAGAACCTTGGAAGGG Rv aacCCCTTCCAAGGTTCTGGGCCc	AGG	+
	f	47434641	-96	Fw caccgTGTCAGGAGTAAGTGCACTG Rv aacCCCTTCCAAGGTTCTGGGCCc	GGG	-
	j	47437695	-109	Fw caccgACCAACCACTGCCACCCCGA Rv aacCCCTTCCAAGGTTCTGGGCCc	GGG	+
	k	47437695	-118	Fw caccgTGCCTAAGGCGGAAGAGCTG Rv aacCCCTTCCAAGGTTCTGGGCCc	GGG	+
	l	47437695	-137	Fw caccgGAGGCTTTCCCTTACCTGGG Rv aacCCCTTCCAAGGTTCTGGGCCc	CGG	-

5.3.2. Lentivirus generation

293T cell line was seeded into 10mm plates to be at a confluence of 50% at 24h for calcium phosphate transfection. The next day, 5µg of the corresponding vector were combined with 3µg of CMV-VSV-G vector (addgene #8454), 10µg of psPAX2 (addgene #12260) and 50µL of 2M $CaCl_2$ in a final volume of 350µL with water. Then, 400µL of 2M HBSS were added dropwise with brief vortexing and 5 min of resting at RT. Mix was added dropwise

to 293T cells and left for 6-16h before fresh medium replacement. Each transfection included a GFP (addgene #12247) positive transfection to ensure transfection at 48-72h. Virus particles were collected from medium 72h post-transfection and filtered with 0.45µm polysulfone membrane to remove cell debris.

5.3.3. Target cells infection and selection

Target cells were plated into 10mm plates to reach 60-70% confluence on the day of infection. Cell medium was removed and replaced with 5mL of 293T filtered media containing the viruses mixed with 5mL of fresh DMEM-F12 and 8µg/mL of polybrene (H9268, Sigma-Aldrich, Missouri, USA). Cells were incubated o/n and then replaced with free-virus fresh medium. Finally, cells were selected by adding antibiotic-containing medium (concentrations indicated in **Table 7**) to select successfully infected cells and guarantee the stable integration of vectors. Each infection included a GFP (addgene #12247) positive control to ensure infection at 48-72h. Treatment efficiency was assessed by checking gene expression by qPCR.

Table 7. Antibiotic treatment in SW480 cell line

Antibiotic	#Reference	Concentration	Treatment length
Hygromycin	10687-010 (Invitrogen)	500 µg/ml	6-8 days
Blasticidin	203408 (Merck)	8 µg/ml	6 days
Bleomycin	A1113903 (Thermo Fisher Scientific)	40 µg/ml	10-12 days
Puromycin	P8833 (Sigma-aldrich)	1.5 µg/ml	7 days
G418	4727878001 (Merck)	600 µg/ml	5-7 days

5.4. FOXD2 overexpression

FOXD2 overexpression was performed with pLenti-C-mGFP-P2A-Puro cloned with the ORF of FOXD2 gene (#RC222086L4) (**Figure 12.A**), purchased from OriGene (OriGene, Rockville, Maryland). Lentiviruses containing pLenti-C-mGFP-P2A-Puro Empty vector (PS100093) and cloned FOXD2 were generated as described above (methods 5.3.2), and cells were infected and selected as indicated (methods 5.3.3).

5.5. FOXD2-AS1 overexpression

FOXD2-AS1 overexpression was carried out with pCMV6-Neo vector (PCMV6NEO, OriGene, Rockville, Maryland) (**Figure 12.B**). OriGene performed FOXD2-AS1 gene synthesis (NR_026878.1) and cloning into pCMV6-Neo vector. Transfection was done with JetOPTIMUS® DNA Transfection Reagent (Polyplus-transfection SA, Illkirch-Graffenstaden, Francia), following the manufacturer's instructions. Briefly, target cells were plated 24h

before transfection to a confluence of 70% into 12-well plates. For transfection, 1µg of plasmid was mixed with 100µL of JetOPTIMUS® Buffer by quick vortexing. After adding 1µL of JetOPTIMUS reagent, the mix was incubated for 20 min at room temperature. Target cells were then replaced with fresh medium containing the transfection mix. GFP control was included in all experiments and fluorescence was checked at 48h post-transfection by microscope. Successfully transfected clones were selected with G418 (**Table 7**).

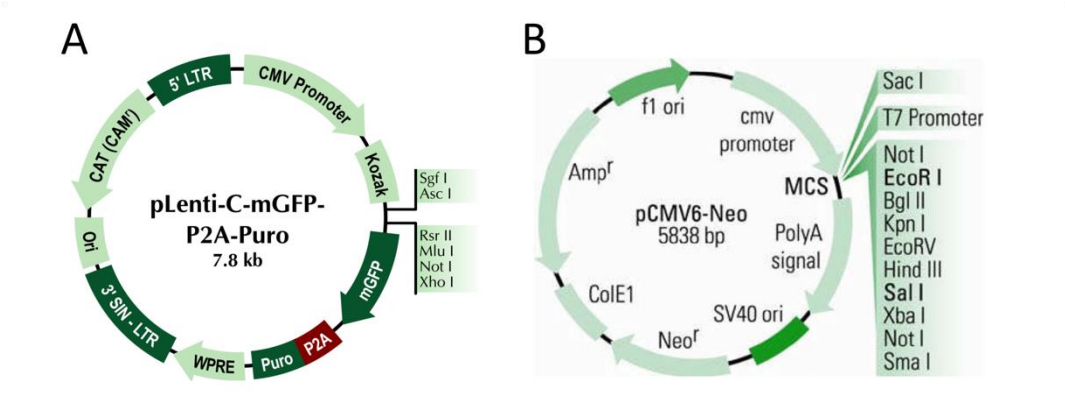


Figure 12. Plasmids vector map. Schematic representation of **A.** plenty-C-mGFP-P2A-Puro and **B.** pCMV6-Neo. Abbreviations: 5'LTR, 5' long terminal repeat; CMV, cytomegalovirus; Ori, origin of replication; P2A, 2A self-cleaving peptide; Puro^r, puromycin resistance for eukaryotic selection; WPRE, Woodchuck Hepatitis Virus (WHV) Posttranscriptional Regulatory Element; 3' SIN-LTR, 3' self inactivating long terminal repeat; CAM', chloramphenicol resistance for bacterial selection; Neo^r, neomycin resistance for eukaryotic selection; Amp^r, ampicillin resistance for bacterial selection.

6. Functional assays

Different functional assays were performed to explore the role of FOXD2 and FOXD2-AS1 on several cellular properties (viability/proliferation, migration, and colony formation).

6.1. Cell proliferation

In order to perform cell proliferation assay, 5x10^3 cells per well were seeded in sextupled in 96 well-plates. Proliferation was monitored at 24, 48, and 72 hours by colorimetric assay method using Cell Counting Kit-8 (Dojindo Molecular Technologies, Rockville, MD) or XTT Cell proliferation kit (Merk, Darmstadt, Germany) according to the manufacturer's instructions. The absorbance was measured at 450 nm in a Spectrophotometer (SpectraMax 340PC384 Microplate Reader, Thermo Fisher Scientific) and at 690nm reference wavelength.

6.2. Migration

We adopted the wound healing assay to assess cell migration ability. Cells were seeded in a two-well culture-insert (ref 80209, ibidi GmbH, Germany) as described by the manufacturer. After cell attachment and confluent monolayer formation, we removed the culture-insert and washed non-adherent cells with 1x PBS (time 0h). We then added fresh medium and monitored wound closure every 24h by microscopy photography (Leica DMI6000B). Wound area at different time points was measured with MRI Wound Healing tool (https://dev.mri.cnrs.fr/projects/imagej-macros/wiki/Wound_Healing_Tool) on ImageJ 1.52a (Schneider, Rasband, and Eliceiri 2012) and % of the cell-free area was calculated relative to time 0h.

6.3. Colony formation

A total of 100 or 200 cells were plated on 10 mm dishes and incubated for 10-15 until visible colonies were formed. Then, cells were washed twice with cold 1x PBS and fixed with ice-cold 100% methanol for 10 minutes. Cells were stained with a 0.05% crystal violet solution in 25% methanol for 10-15 minutes. Exceeding staining solution was removed with water and dishes were air-dried o/n. Plates were photographed and analyzed with OpenCFU software (Geissmann 2013).

7. Computational analysis

7.1. TCGA public data

The Cancer Genome Atlas (TCGA) project aims to accelerate our understanding of the molecular basis of cancer by providing large-scale genome sequencing data of multiple cancer types. We used the colon adenocarcinoma dataset (COAD) to investigate gene expression and DNA methylation profiles of our regions of interest.

Gene expression analysis was performed with RNA-seq data downloaded from the GDC Data Portal (https://portal.gdc.cancer.gov/repository). The TCGA COAD cohort with available RNA-seq data comprised 451 tumor tissues and 41 paired normal colonic mucosae. Data was downloaded as FPKM (Fragments Per Kilobase Million), normalized by transcript length and library size. Methylation analysis was carried out with Illumina 450K Infinium array, and data accounting for 30 CpGs located in region Chr1:47,897,000-47,907,000 were downloaded using the Wanderer tool (http://maplab.imppc.org/wanderer/). In total, 302 tumors and 38 paired normal tissues were included from the COAD dataset and methylation levels were assessed as Beta-

values. Clinico-pathological characteristics of the TCGA COAD cohort are listed in **Table 8**.

Table 8. Clinico-pathological characteristics of patients (TCGA-COAD cohort). A total of 451 and 302 tumors were included for RNA and DNA analysis, respectively. Patients with NA values were not included in the table.

		n	
Variables	**Categories**	**RNA-seq**	**Meth. Array**
Cancer Stage	Stage I	78	48
	Stage II	180	116
	Stage III	127	85
	Stage IV	64	36
Tumor size and invasion	T1	11	6
	T2	80	48
	T3	311	200
	T4	60	37
Lymph node involvement	N0	273	176
	N1	104	72
	N2	83	44
Distant metastasis	M0	337	204
	M1	64	35
Colon location	Distal	140	95
	Proximal	219	194
Survival	Surviver	351	253
	Non-surviver	91	40

7.2. ENCODE datasets

We downloaded RNA-seq and Chip-seq data from normal colonic tissues and CRC cell lines from ENCODE portal (www.encodeproject.org). Data sets were visualized using UCSC Genome Browser (http://genome.ucsc.edu/). The experiment's reference of the used data are listed in **Supplementary Table S2**.

7.3. Online tools

A set of bioinformatic tools have been used in the course of my thesis to visualize, download and process both public and in-house generated data. The most important tools are listed in **Table 9**.

7.4. Statistical analysis

Statistical analyses were performed with GraphPad Prism version 9.0.0 for Windows (GraphPad Software, San Diego, California USA). Comparisons between normal and tumor tissues were performed with the Wilcoxon rank test (if paired) or Mann-Whitney test (if

unpaired). Correlations were tested with Spearman's rank test. Analysis to explore the prognostic value of FOXD2 and FOXD2-AS1 in CRC patients were done with Mann-Whitney tests or non-parametric ANOVA tests. Survival curves were tested with the log-rank test. Cell line proliferation assays were tested with two-way ANOVA tests, and one-way ANVOA tests were applied to analyze migration and colony formation abilities. Each statistical test is indicated in the corresponding experiment.

For DNA methylation analysis, Beta-values were converted to M-values prior analysis (M-Value=log2(Beta-value/(1-Beta-value)). Differentially Expressed Gene analysis of RNA-seq was conducted with R using DESeq2 (Love, Huber, and Anders 2014).

Table 9. Bioinformatic tools.

Tool	Link
Benchling	https://www.benchling.com/
BLAST	https://blast.ncbi.nlm.nih.gov/Blast.cgi
Cas-OFFinder	http://www.rgenome.net/cas-offinder/
Chainy	http://maplab.imppc.org/chainy/
Corre	http://maplab.imppc.org/corre/
CPAT	http://lilab.research.bcm.edu/
ENCODE	https://www.encodeproject.org/
EnrichR	https://maayanlab.cloud/Enrichr/
FANTOM CAT	https://fantom.gsc.riken.jp/
GEPIA	http://gepia.cancer-pku.cn/
Gorilla	http://cbl-gorilla.cs.technion.ac.il/
GPP sgRNA Designer	https://portals.broadinstitute.org/gpp/public/
GTEX	https://gtexportal.org/home/
IDT Oligo Analyzer	https://eu.idtdna.com/calc/analyzer
In-Silico PCR tool	http://genome.ucsc.edu/cgi-bin/hgPcr
Methylation plotter	http://maplab.imppc.org/methylation_plotter/
muTarget	http://mutarget.com
ORFfinder	https://www.ncbi.nlm.nih.gov/orffinder
Primer-BLAST	https://www.ncbi.nlm.nih.gov/tools/primer-blast/
UCSC	http://genome.ucsc.edu/
Wanderer	http://maplab.imppc.org/wanderer/

Methods Study II

Colorectal cancer is associated with the presence of cancer driver mutations in normal colon

Data analysis and most experimental procedures from **study II** were performed at Rosana Risques research lab, at the University of Washington (UW), Seattle, USA. Other experimental procedures, such as sample collection and DNA extractions, were done at William M. Grady's research group at the Fred Hutchinson Cancer Research Center, Seattle, USA and at Miguel Ángel Peinado's research group at the IGTP, Badalona, Spain.

1. Samples

1.1. Normal colon mucosa

This study included normal colon mucosa samples (n=47) collected at the University of Washington Medical Center and affiliated practice sites (Seattle, WA, USA) from 24 patients without colorectal adenocarcinoma (CRC) undergoing colonoscopic screening or surveillance and from 23 patients with a newly diagnosed primary invasive colorectal adenocarcinoma undergoing surgical resection. Clinico-pathological characteristics of patients are listed in **Supplementary Table S3**. None of the patients had hereditary cancer syndrome. The groups of patients were matched by age and history of polyp(s) and were enriched with young individuals to explore differences in somatic mutations in early vs late onset CRC. All normal samples from individuals with CRC were located 10 to 15 cm from the tumor except for two samples collected between 3 to 5 cm from the tumor. Only one patient had neoadjuvant therapy (P40). Immediately after collection, samples were frozen in liquid nitrogen and stored at -80ºC until DNA extraction. Patients consented to sample collection, and the study was conducted following protocols approved by the appropriate Institutional Review Board committees.

1.2. CRC tumors

Formalin-Fixed Paraffin-Embedded (FFPE) tumor blocks from patients with CRC were histologically examined with hematoxylin and eosin staining followed by microdissection and DNA extraction in 19 cases with sufficient tumor content. DNA extraction and library preparation from tumor DNA (see methods 2) was performed after all normal tissues were analyzed to avoid any chances of cross-contamination. In all but one case, microsatellite instability was determined by mismatch repair defect based on routine clinical

immunohistochemistry of proteins MLH1, MSH2, MSH6, and PMS2 in tumor FFPE tissue sections. Two cases were MSI positive (**Supplementary Table S3**).

1.3. Cell lines

DNA from colorectal cancer cell lines HCT116, HT29, Lovo, and SW480, was used for method validation. Cell lines were obtained from the ATCC (Virginia, USA). Somatic mutations of each cell line are listed in **Table 10**.

Table 10. List of human colorectal cancer cell lines used.

Cell line	Gene	cDNA mutation	Protein mutation	Genotype
HCT116	*KRAS*	c.38G>A	p.G13A	Heterozygous
	PIK3CA	c.3140A>G	p.H1047A	Heterozygous
HT29	*BRAF*	c.1799T>A	p.V600E	Heterozygous
	TP53	c.818G>A	p.R273H	Homozygous
LoVo	*KRAS*	c.38G>A	p.G13A	Heterozygous
SW480	*KRAS*	c.35G>T	p.G12V	Homozygous
	TP53	c.818G>A	p.R273H	Heterozygous
	TP53	c.925C>T	Pro309Ser	Heterozygous

2. DNA extraction

Genomic DNA was extracted from frozen normal colon tissue samples using the DNeasy Blood & Tissue Kit (Qiagen, Germany), from FFPE tumor samples using the QIAamp DNA FFPE Tissue Kit (Qiagen) and from CRC cell lines DNA using PureLinkTM Genomic DNA Mini Kit (Invitrogen, MA, USA). Forceps mucosal biopsies procured at endoscopy were approximately 6mm x 4mm x 3mm in size, and the whole biopsy was used for DNA extraction. In normal colon samples from surgical resections, the mucosal epithelium was selected to match the size of the endoscopic biopsies. DNA was quantified by Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies). For cell lines and a subset of normal colon samples, DNA quality was assessed with Genomic TapeStation (Agilent Technologies, CA, USA). It demonstrated high quality in all samples (DNA integrity number (DIN) ≥7) (**Figure 13**).

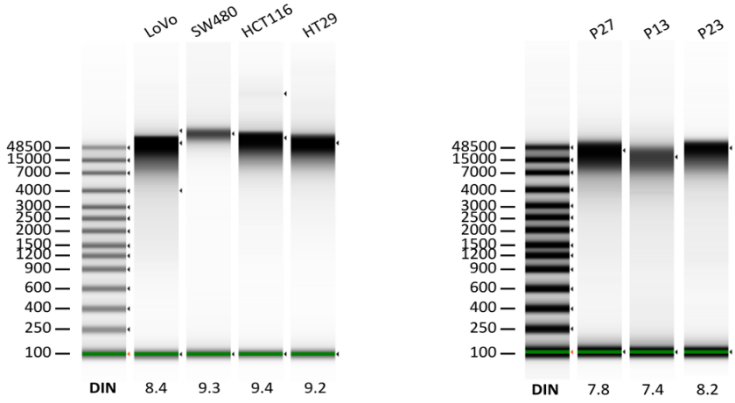


Figure 13. **Genomic Tape Station visualization.** Tape Station gels show genomic DNA of four CRC cell lines (left) and of normal colon of three patients included in the study (right).

3. CRISPR-DS

3.1. CRISPR guide design and annealing

CRISPR-DS employs CRISPR-Cas9 digestion of target regions followed by size selection of excised fragments as a method for efficient target enrichment prior to library preparation (Nachmanson et al. 2018). We used Benchling [Biology Software, 2020] (CA, USA) to design guide RNAs (gRNAs) to excise the coding regions of the *TP53* gene and the hotspot mutation codons of *BRAF*, *KRAS,* and *PIK3CA* genes into fragments of ~250-280bp (**Figure 14**). Then we used the CRISPOR web tool (Concordet and Haeussler 2018) to select the best candidates, which included 24 gRNAs (**Table 11**) that excised the target region into 13 fragments with a total panel size of 3461bp. The panel comprised 1953 coding bp and 1508 non-coding bp from intronic regions flanking the excised exons.

The gRNAs (guide RNAs) are composed of a complex of CRISPR RNA (crRNA), which contains the ~20bp unique sequence responsible for target recognition, and the trans-activating crRNA (tracrRNA), which has a universal sequence (Nachmanson et al. 2018). Each designed crRNA (30nM) was incubated with tracrRNA (30nM) (IDT, IA, USA) and with Nuclease-free Duplex Buffer (IDT) for 5 min at 95ºC in a total volume of 100uL. The obtained gRNAs were pooled, aliquoted, and stored at -80ºC.

Table 11. crRNA sequences for CRISPR-Cas9 digestion.

Gene	Exon	crRNAs sequence plus ***pam site***		Cut position	Fragment Length (bp)
BRAF	15	Up	ACACTGATTTTTGTGAATACTGG	Chr7:+140753164	277
		Down	TTCATAATGCTTGCTCTGATAGG	Chr7:-140753440	
KRAS	2	Up	CGAATATGATCCAACAATAGAGG	Chr12:-25245279	278
		Down	GATACACGTCTGCAGTCAACTGG	Chr12:-25245556	
PIK3CA	10	Up	AATCATCTGTGAATCCAGAGGGG	Chr3:+179218157	270
		Down	CATGTTTTTACCATACCTATTGG	Chr3:+179218426	
PIK3CA	21	Up	TCATCAAAAGATTGTAGTTCTGG	Chr3:-179234194	251
		Down	CAGGCAAAGACCGATTGCATAGG	Chr3:+179234444	
TP53	11	Up	ACGCACACCTATTGCAAGCAAGG	Chr17:+7669519	264
		Down	TGCTTTGAAGGGCCTAAGGCTGG	Chr17:+7669782	
TP53	10	Up	ATGTGGTTATAGGATTCAACCGG	Chr17:+7670450	278
		Down	CGGATCTGCAGCAACAGAGGCGG	Chr17:+7670729	
TP53	9	Up	CAATTGGGGCATTTGCCATCAGG	Chr17:-7673433	261
		Down	ACTAAGCGAGGTAAGCAAGCAGG	Chr17:-7673693	
TP53	8	Up	ACTAAGCGAGGTAAGCAAGCAGG	Chr17:-7673694	276
		Down	TGGCTTCTCCTCCACCTACCTGG	Chr17:-7673969	
TP53	7	Up	CCCGCCGGGGATGTGATGAGAGG	Chr17:+7674064	250
		Down	GATAACACAGGCCCAAGATGAGG	Chr17:+7674313	
TP53	6	Up	CATTTACTTTGCACATCTCATGG	Chr17:+7674739	246
		Down	AGACCTAAGAGCAATCAGTGAGG	Chr17:+7674984	
TP53	5	Up	AGACCTAAGAGCAATCAGTGAGG	Chr17:+7674985	253
		Down	TGAGGGCAGGGGAGTACTGTAGG	Chr17:+7675237	
TP53	4	Up	TTGCACGGTCAGTTGCCCTGAGG	Chr17:-7675984	282
		Down	CATTGCTTGGGACGGCAAGGGGG	Chr17:+7676265	
TP53	2&3	Up	ACAACGTTCTGGTAAGGACATGG	Chr17:-7676376	275
		Down	GGGTTGGAAGTGTCTCATGCAGG	Chr17:-7676650	

3.2. CRISPR-DS Library Preparation

Genomic DNA from normal colon tissues and CRC cell lines was processed for CRISPR-DS as previously described in Nachmanson *et al* . 300nM of pooled gRNAs were incubated with Cas9 nuclease (NEB) at ~30 nM, 1× NEB Cas9 reaction buffer, and water in a volume of 27 µL for 10 min at 25°C. Then, 100ng of DNA were added for digestion in a final volume of 30 µL, and the reaction was incubated at 37ºC o/n. The next day, the Cas9 enzyme was inactivated by heat shock for 10 min at 70ºC, followed by a double size selection with 0.5x and 1.8x ratio AMPure XP Beads (Beckman Coulter, CA, USA), to remove off-target, undigested high molecular weight DNA (**Figure 14.B**). DNA fragments were end-repaired, A-tailed, and ligated with duplex adapter containing 8 bp random double-stranded molecular tags (TwinStrand Biosciences, WA, USA) using the NEBNext Ultra II DNA Library Prep Kit (NEB, MA, USA) (**Figure 14.C**). Ligated DNA was amplified using the KAPA Real-

Time Amplification kit with fluorescent standards (KAPA Biosystems, MA, USA). Samples were amplified until they reached Fluorescent Standard 3, which typically takes 8 to 9 cycles. After, a 0.8x ratio AMPure Bead was was performed to purify the amplified fragments in 40uL of elution volume.

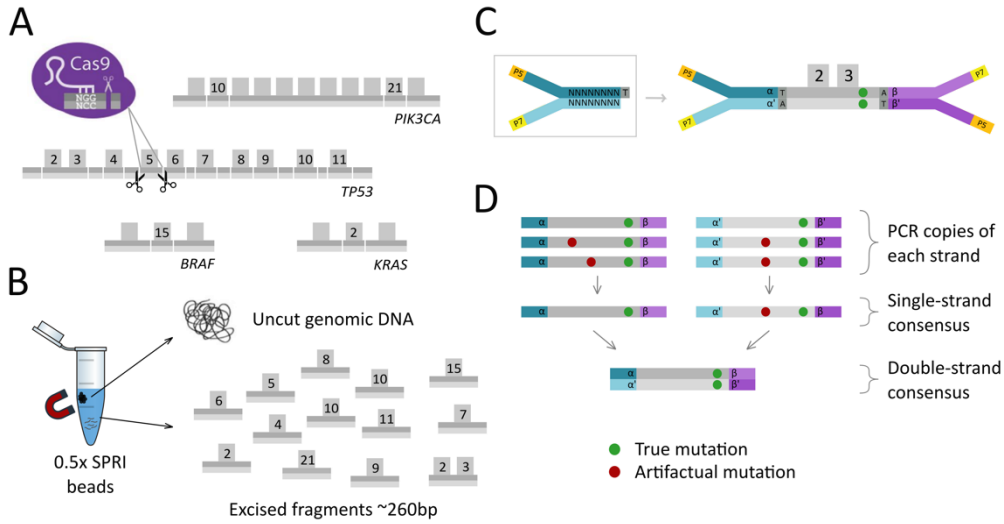


Figure 14. **Ultra-deep sequencing CRISPR-DS. A.** The coding region of *TP53* and hotspot mutations of *BRAF, KRAS* and *PIK3CA* are excised in fragments of ~260bp by targeted Cas9 digestion. **B.** Fragments containing the regions of interest are enriched by size selection. **C.** Ligation to duplex adapters, which contain standard Illumina P5 and P7 sites (orange and yellow boxes) and double-stranded molecular tags (green and purple boxes). **D.** PCR produces multiple copies of each strand, which can be distinguished by their tags after sequencing and grouped to generate highly accurate duplex reads based on their consensus sequence. Duplex reads only contain mutations identified in both DNA strands. *Adapted from* (Nachmanson et al., 2018).

3.3. Hybridization capture and post-capture PCR

Regions of interest were captured by hybridization with 120bp biotinylated xGen Lockdown probes (Integrated DNA Technology, IA, USA) (**Supplementary Table S4**). Probes were mixed in equimolar amounts and the final capture pool was diluted to 0.75pmol/µL. Hybridization capture was performed according to the IDT protocol, except for three modifications. First, we used blockers MWS60, 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTIIIIIIIIIIIITG ACT-3' and MSW61, 5' -GTCAIIIIIIIIIIIIAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3', which are specific to DS adapters. Second, we used 50 µL of Dynabeads M270 Streptavidin beads instead of 100 µL. Third, the post-capture PCR was performed with the KAPA HiFi HotStart PCR kit (KAPA Biosystems) using MWS13 (5'-AATGATACGGCGACCACCGAG-3')

and indexed primer MWS21 (5'-CAAGCAGAAGACGGCATACGAGATXXXXXXGTGACTGGAGT TCAGACGTGTGC-3'). The PCR product was purified with a 0.8× AMPure Bead wash.

Libraries were visualized on the Agilent 4200 TapeStation to confirm the expected peak size (**Figure 15**). If peaks were not present, a second round of hybridization capture was performed.

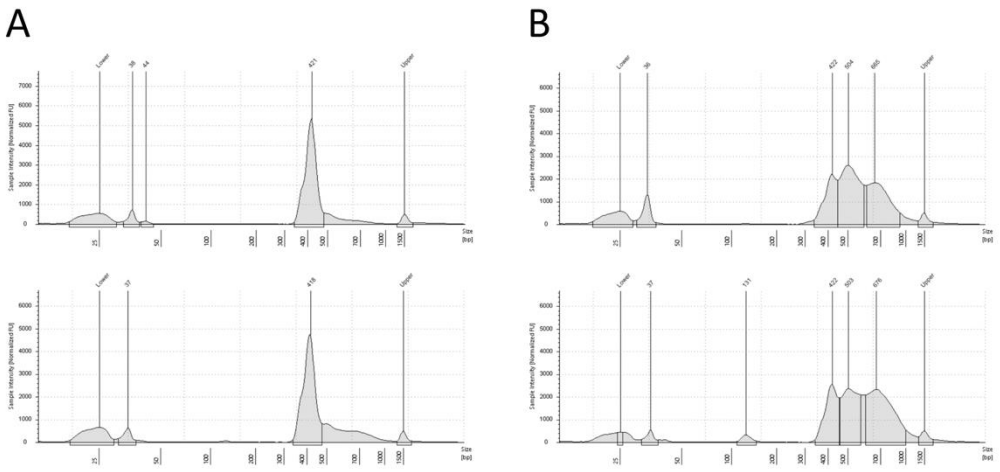


Figure 15. **Visualization of sequencing libraries prepared with CRISPR-DS.** Library electropherogram visualization for quality control of libraries prior to sequencing. **A.** Expected peak size (~420bp). **B.** Peaks of low quality libraries (>400bp).

3.4. CRISPR-DS sequencing and data processing

Libraries were quantified using the Qubit dsDNA HS Assay kit, diluted, and pooled for sequencing on a MiSeq Illumina platform on-site or a HiSeq 3000 (Genewiz, NJ, USA), allocating ~2 million reads per sample. CRISPR-DS sequencing data was analyzed using the Duplex Sequencing pipeline v1.1.4 available at https://github.com/Kennedy-Lab-UW/Duplex-Seq-Pipeline. First, raw reads were grouped using the double-stranded molecular tag included in the duplex adapters, and a Single-Strand Consensus Sequence read was built from reads sharing the same tag. Then Single-Strand Consensus Sequence reads with complementary tags were compared to produce a single, highly accurate duplex read (**Figure 14.D**). Duplex reads were aligned to the human genome reference hg38 (GRCH38), end-trimmed, locally realigned, and overlap-trimmed. Variants were called using a samtools mpileup-based variant caller and output VCF files were converted to MAF files using the Vcf2Maf script (https://github.com/mskcc/vcf2maf) with VEP version 99. All variants identified in SNP positions were discarded for subsequent analysis.

Three samples with potential cross-contamination based on SNP frequency were removed from the study. Mutant Allele Frequency (MAF) was calculated for each mutation as the number of mutated duplex reads divided by the total duplex depth at the given position.

3.5. CRISPR-DS validation using CRC cell lines

CRISPR-DS reproducibility, sensitivity, and efficiency were evaluated using DNA extracted from four common human colorectal cancer cell lines: HCT116, HT29, LoVo, and SW480. 100 ng of DNA from each cell line were processed for CRISPR-DS in two independent replicate libraries to test reproducibility. To test sensitivity, HT29 DNA was spiked in HCT116 DNA at three different ratios (1:10, 1:20, and 1:100). Library preparation and data analysis were performed using the same methods employed for tissue samples. As each duplex read corresponds to an original DNA molecule, duplex depth indicates the number of haploid genomes analyzed in each position. Thus, the efficiency (also called recovery rate) was calculated as the average duplex depth divided by the number of input genomes corresponding to 100ng and 500ng of DNA.

4. Data analysis of normal colon

4.1. Calculation of mutation frequency

For each sample, the overall duplex depth was calculated as the total number of duplex nucleotides sequenced divided by the size of the panel. On average, we sequenced 8.6 M duplex nucleotides per sample, corresponding to a duplex depth of 2,484x (minimum 1,268x; maximum 4,306x) (**Supplementary Table S5**). To correct for the variability in sequencing depth across samples (**Figure 16**), sample comparisons were made based on mutation frequencies, which were calculated as the number of mutations in a given region (e.g., coding, non-coding, *TP53* coding) divided by the total number of duplex nucleotides sequenced in that region. Similarly, mutation frequencies were calculated for specific types of mutations (e.g., drivers) by dividing the number of mutations in the category of interest by the total number of duplex nucleotides sequenced in the target region. Mutation counts and corresponding mutation frequencies for each sample are indicated in **Supplementary Table S5**.

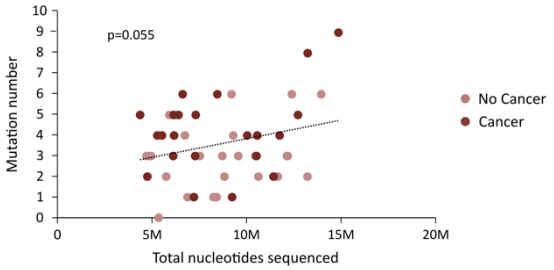


Figure 16. Number of mutations tends to increase with number of total nucleotides sequenced. Correlation between the number of mutations and the total nucleotides sequenced by patient. P-value corresponds to Pearson correlation.

4.2. Mutational analysis

Coding mutations were extracted from MAF files and were further annotated by mutation type (missense, nonsense, splice, indel, and synonymous), mutation spectrum (C>A, C>G, C>T, T>A, T>C, and T>G), localization in CpG dinucleotides, and driver mutations. Mutations in *BRAF*, *KRAS*, and *PIK3CA* were considered driver mutations if they corresponded to common oncogenic hotspot mutations in these genes according to large intestine carcinoma data from the COSMIC database (https://cancer.sanger.ac.uk/cosmic) (Tate et al. 2019). The following oncogenic driver mutations were considered: *BRAF* V600E, which accounts for >90% of *BRAF* mutations in CRC; *KRAS* hotspot mutations in codons 12 and 13, which account for >90% of *KRAS* mutations in CRC; and *PIK3CA* hotspot mutations E545K, H1047R, and E542K, which account for >50% of *PIK3CA* mutations in CRC. Driver mutations in *TP53* included the 10 most common substitutions according to the COSMIC database, which represent >50% of all mutations reported in large intestine carcinomas (p.R175H, p.R273H, p.R248Q, p.R282W, p.R273C, p.R248W, p.G245S, p.R213*, p.R196* and p.R306*), and all splice, indels and nonsense mutations. The list of annotated coding mutations for oncogenes and *TP53* are presented in **Supplementary Tables S6** and **S7**, respectively. Large intestine carcinoma variants from COSMIC were also used to determine the mutation spectrum (6 possible nucleotide substitutions) of CRC (n=70,525) as well as the distribution of CRC mutations within the protein domains of the genes of interest.

4.3. *TP53* mutation characterization with Seshat

All *TP53* mutations were further characterized using the Seshat web service tool (https://p53.fr/*TP53*-database/seshat) (Tikkanen et al. 2018). A MAF file containing all the *TP53* mutations observed (n=118) was submitted to Seshat to accurately annotate, validate and analyze *TP53* variants using data derived from the UMD *TP53* database (Leroy

et al. 2017) . From the Seshat output, coding and splice mutations were extracted (n=85) (**Supplementary Table S7**) along with the information about their frequency in the UMD cancer database and predicted pathogenicity. Frequency in the cancer database was categorized as "Common in cancer" (including very frequent and frequent mutations) and "Not common in cancer" (including not frequent, rare, unique, and not seen before mutations). Predicted pathogenicity was categorized as "Pathogenic" (including pathogenic and likely pathogenic mutations) and "Benign or unknown" (including likely benign, benign and variants of unknown significance). In addition, the distribution of *TP53* mutations in normal colon based on mutation type, cancer frequency, and pathogenicity was compared to the expected distribution for random *TP53* mutations as well as CRC *TP53* mutations based on the UMD database.

4.4. *TP53* mutations without selection

To compare *TP53* mutations in normal colon with the theoretical make up of mutations if they were to occur completely at random and without selection, we generated a list of all possible mutations in the *TP53* coding region (n=3,546) and submitted it to Seshat to determine their frequency in cancer and predicted pathogenicity.

4.5. UMD *TP53* cancer database mutational analysis

To compare *TP53* mutations observed in normal colon with the mutations present in CRC, we used the most recent UMD *TP53* cancer database (2021) kindly provided by Dr. Thierry Soussi (Sorbonne Université, Paris, France). We selected all mutations corresponding to colorectal carcinoma samples (n=17,681) and determined the distribution of mutations according to mutation type, frequency in CRC, and predicted pathogenicity. The distribution of *TP53* CRC mutations across these variables was compared with the distribution of normal colon mutations for the same variables, divided by patients younger and older than 55 years old and patients with and without CRC.

5. Tumor Sequencing and data processing

50-100ng of tumor DNA were sonicated, end-repaired, A-tailed, and ligated to DS adapters using commercial kits (TwinStrand Biosciences, Seattle, WA). Hybridization capture was performed with the same probe pool used for CRISPR-DS of the normal colon but using two rounds of hybridization capture as previously recommended (Schmitt et al. 2015) (**Supplementary Table S4**). Enriched libraries were amplified, quantified using the Qubit dsDNA HS Assay kit, diluted, and pooled for sequencing on a MiSeq Illumina platform on-site, allocating ~0.8 million reads per sample. Raw reads were processed with the DS

pipeline v1.1.4 available at https://github.com/Kennedy-Lab-UW/Duplex-Seq-Pipeline. Data analysis was performed on Single-Strand Consensus Sequence (SSCS) reads instead of duplex reads to provide higher depth (mean 284x) and because the goal was to identify tumor driver mutations, which are expected to be clonal and harbor large MAF. For each tumor, we cataloged *BRAF, KRAS, PIK3CA,* and *TP53* non-synonymous or indel mutations with MAF ≥0.1 and determined whether these mutations coincided with mutations identified in normal colon from the same individual (**Supplementary Table S8**).

6. Bi-Sulfite Conversion and methylation assessment

DNA (500 ng) from each sample was bisulfite converted using the EZ DNA Methylation Kit (Zymo Research, Irvine, USA). The DNA samples were submitted to the Genomics Core at the Fred Hutchinson Cancer Research Center, where they were processed and run on MethylationEPIC arrays following the manufacturer's instructions (Illumina, Inc., CA, USA). As previously described (Wang et al. 2020), the raw intensity files (IDAT) were preprocessed, normalized, and the array results were assessed using four epigenetic age clocks (Hannum, Horvath, PhenoAge, and EpiTOC).

7. Statistical analysis

Statistical analyses were performed with IBM SPSS Statistics version 25 and R version 3.6.3. Correlations were tested with Spearman's rank test. Comparison of mutation frequency means across groups of individuals was performed by t-test. Comparison of the distribution of mutational features across groups of individuals, with the mutational distributions in CRC and in the absence of selection was performed with Chi-Square. The predictive model was estimated with the glmnet R package (Friedman, Hastie, and Tibshirani 2010), with parameters for Lasso logistic regression. The penalization parameter was selected to restrict the model to 5 covariates. Predictive accuracy was calculated with the area under the ROC curve and its 95% confidence intervals as implemented in the pROC R package (Turck et al. 2011).

8. Data access

Sequencing data from this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject) under accession number PRJNA767868.

RESULTS Study I

The role of FOXD2 and FOXD2-AS1 in CRC

Background

Epigenetic variability constitutes a signature of human cancer underlying chromatin architecture and genomic regulation changes with direct implications in the cell's functional reprogramming. DNA methylation dynamics are intrinsically interconnected with the multiple layers of factors that drive genome remodeling and cell programs alterations in cancers.

Following our previous findings, a novel epigenetic network based on DNA methylation co-variability between pairs of CpGs was assessed in primary colon carcinomas and their normal tissue counterparts in two independent cohorts (TCGA-COAD and Colonomics) (Mallona, Aussó, Díez-Villanueva, Moreno, & Peinado, 2018). The co-methylation network was modular in both normal and tumor with partly shared structure. We focused on trans-modules composed of at least ten pairs of CpGs placed >1Mb apart or in different chromosomes. As the co-methylation trans network properties were evaluated, two noticeable giant components, named modules 1 and 2, displayed opposite methylation behaviors in tumors but lacked such negative correlation in normal samples (**Figure 17**). Out of all the pairs of CpGs with this flipping trend in tumors compared to normal tissue networks, we focused on identifying CpGs near genes involved in cancer progression. We further explored such CpGs using *in silico* tools and basic experimental approaches, and cg08638320 located on the FOXD2-AS1 body gene rapidly caught our attention. We observed interesting DNA methylation changes related to gene expression, and the literature regarding those genes at the time was rather poor. Therefore, we decided to better characterize methylation and expression associations of the two neighboring genes FOXD2 and FOXD2-AS1 near the cg08638320 DNA locus.

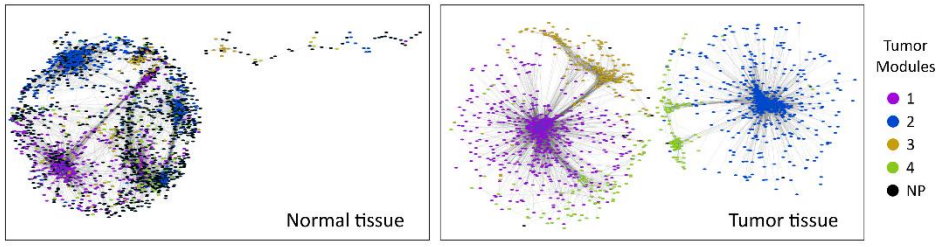


Figure 17. Co-methylation network module comparison. Networks highlight the nodes (CpGs) color coded by the module they belong in the tumor dataset. Data from the *Colonomics* cohort. NP, not present. *Adapted from* (Mallona et al., 2018).

Results Study I

1. *In silico* characterization of FOXD2 and FOXD2-AS1 genomic locus

To begin studying the role of FOXD2 and FOXD2-AS1 genes in CRC, we closely examined their DNA locus using in silico available tools.

Accurate gene annotation is challenging and requires a combination of experimental techniques, computational predictions, and validations. As a result, annotations are still inaccurate and incomplete, especially for specific genes. Non-coding RNAs are particularly affected, often automatically annotated and based in predictive models lacking detailed curation and validation.

FOXD2 and FOXD2-AS1 adjacent genes are located on chromosome 1 on a head-to-head opposite orientation, being FOXD2 transcribed in the + strand and lncRNA FOXD2-AS1 on the – strand of the DNA. Since gene annotations are constantly updated and differ between datasets, we characterized the TSS and gene boundaries among different sources since we started this project in 2018 (**Figure 18**). GENCODE (v33, released in 2019) and NCBI RefSeq initially had the TSS of FOXD2 and FOXD2-AS1 annotated at chr1:47,436,017 and chr1:47,434,641, respectively. Instead, FANTOM CAT (v1.0.0, published in 2017) defined TSS as chr1:47,438,044 for FOXD2 and chr1:47,437,695 for FOXD2-AS1, being their 5' in close proximity (<0.5Kb). However, at the end of 2020, both GENCODE (v36) and RefSeq were updated and agreed with FANTOM CAT's TSS for FOXD2 (chr1:47,438,044), but not for FOXD2-AS1.

Considering these differences between genome consortiums and lncRNA poorly annotations, we further explored the genomic landscape of FOXD2 and FOXD2-AS1 by adding transcriptomic and epigenomic layers of information. Massive parallel sequencing of histone modifications provides deeper insights into the genome architecture, equally RNA sequencing data helps in the understanding and profiling of the human transcriptome. Thus, we analyzed RNA-seq and Chip-seq data from large intestine samples and CRC cell lines available at ENCODE (see methods 7.2). Our results support FANTOM CAT TSS annotations, being FOXD2 and FOXD2-AS1 transcribed in opposite directions, with their TSS close together, whereas H3K4Me3 (associated to initiation of transcription) and H3K27Ac (enriched at promoters) histone marks co-localize (**Figure 18**).

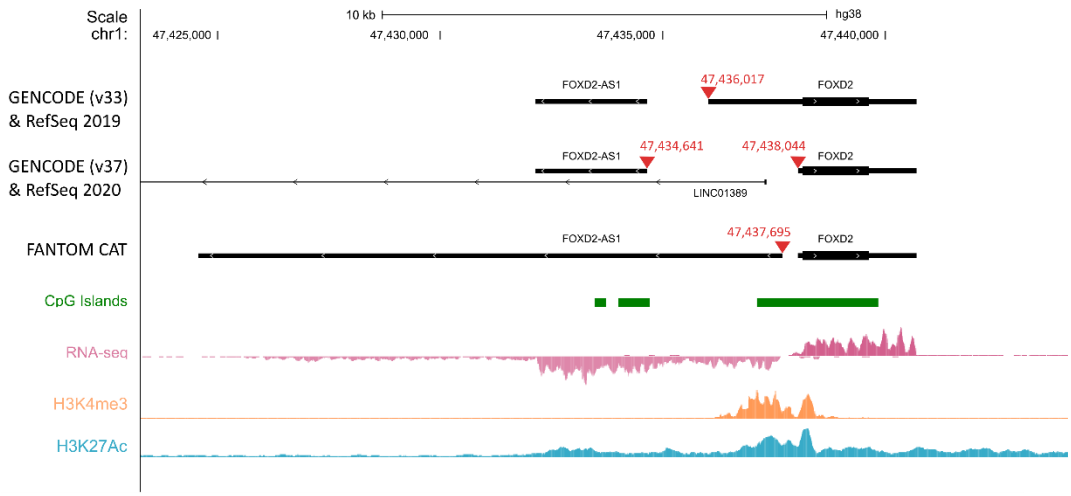


Figure 18. **FOXD2 and FOXD2-AS1 loci on chromosome 1.** FOXD2 and FOXD2-AS1 genes head-to-head disposition in chromosome 1. Figure describes gene annotations from GENCODE, RefSeq and FANTOM CAT and CpGi disposition. TSS coordinates (hg38) are indicated in red. RNA-seq, H3K4me3 and H3K27Ac Chip-Seq data obtained from colon samples from ENCODE.

2. FOXD2-AS1 transcript characterization

LncRNAs are emerging components of the genome and remain relatively unexplored. To widen our knowledge about FOXD2-AS1, we aimed to characterize its transcript using *in silico* and experimental approaches.

2.1 FOXD2-AS1 has no predicted coding potential

To validate the non-coding probability already described for FOXD2-AS1, we first used the NCBI ORFfinder tool (https://www.ncbi.nlm.nih.gov/orffinder). This tool searches open reading frames based on Met start codon within an in-frame stop codon. Although FOXD2-AS1 transcript is not well defined, we established our region of study based on RNA-seq data in colon samples (**Figure 18**), exploring the – strand of chromosome 1: 47,432,133-47,437,695 (5563bp) and using a cutoff of 100aa (300nt), usually applied to establish a protein-coding definition (Dinger, Pang, Mercer, & Mattick, n.d.). The tool identified only two ORF of at least 100 aa, potentially giving rise to 158 and 116aa proteins. We then performed a protein alignment search of the two potential coding regions using the online BLAST tool (www.blast.ncbi.nlm.nih.gov) and found no significant similarities with previously described proteins in humans.

By chance, non-coding sequences might contain ORF sequences, being the ORF presence and size insufficient parameters to predict RNAs' coding potential accurately. Therefore, we further explored FOXD2-AS1 coding probability by using CPAT: Coding-Potential Assessment Tool (Wang et al., n.d.), which searches for ORF also considering other features comparable to sets of protein-coding genes and sets of non-coding genes. CPAT determined a coding probability of only 0.082, which translates into low protein-coding potential values even compared to other lncRNAs (values >0.364 indicate coding sequence). Gathering all this data, we assumed FOXD2-AS1 has no protein-coding potential, as previously reported.

2.2 FOXD2-AS1 is a polyadenylated cytoplasmic lncRNA

As reviewed in the introduction (section 2.3.1), most lncRNAs are polyadenylated, and we wondered if FOXD2-AS1 transcript has a poly-A tail. To address this question, we analyzed the abundance of the transcript when preparing the cDNA with random hexamers or oligo-dT primers in three different CRC cell lines. We amplified the generated cDNAs by qPCR and calculated the expression ratio as oligo(dT) versus random primed retrotranscription. Results showed that FOXD2-AS1 has a higher expression when cDNA is amplified using oligo(dT) primers, behaving as an mRNA (e.g., PUM1) rather than a non-polyadenylated RNA (e.g., *MALAT1*) and thus indicating the presence of a polyA tail (**Figure 19**).

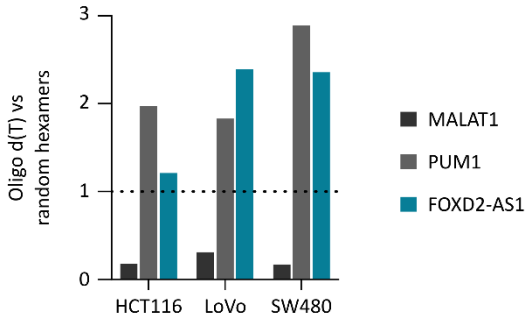


Figure 19. **FOXD2-AS1 is a polyadenilated lncRNA.** Expression ratio of amplified transcripts by qPCR using oligo-d(T) versus random hexamers. PUM1 and *MALAT1* were used as poly A + and poly A- controls, respectively.

While most mRNAs are stable transcripts located in the cytoplasm, lncRNAs are in general less stable and can localize to a specific cell compartment putatively linked to the biological function. We performed subcellular fractionation of the cytoplasmic and nuclear RNA content in HCT116, LoVo, and SW480 (methods section 2.2). Transcript enrichment

analysis by RT qPCR determined that FOXD2-AS1 is localized in both cytoplasmic and nuclear fractions but slightly enriched in the cytoplasmic fraction in HCT116 and SW480 (**Figure** 20.A). We used GAPDH and *MALAT1* as fraction-specific controls for cytoplasmic and nuclear RNAs, respectively.

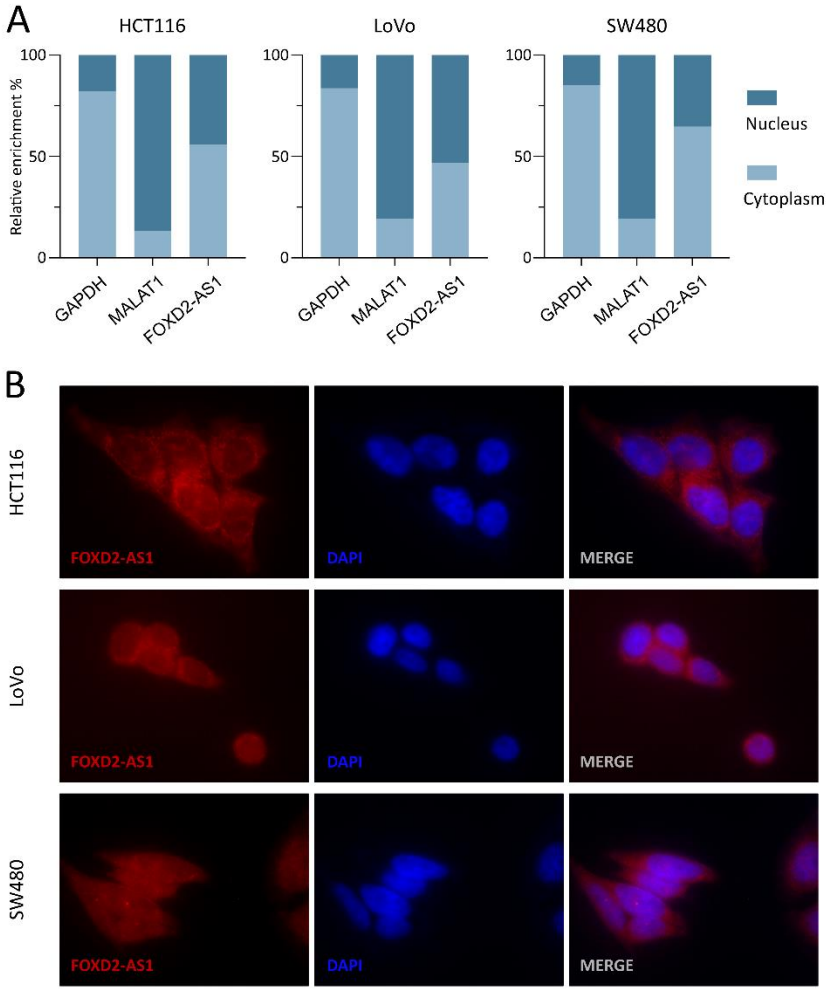


Figure 20. **FOXD2-AS1 is slightly enriched in the cytoplasm. A.** Relative enrichment (%) of transcripts in nuclear and cytoplasmic subcellular fractions. GAPDH and *MALAT1* are used as cytoplasmic and nuclear controls, respectively. **B.** RNA FISH signal location using probes against FOXD2-AS1 (red). The nucleus was stained with DAPI (blue).

We performed single-molecule RNA-FISH (Stellaris®) to further validate RNA fractionation results. FOXD2-AS1 signal was found in both cytoplasmic and nuclear fractions but enriched in the cytoplasm (**Figure** 20.B), confirming the subcellular fractionation results.

We also used GAPDH and *MALAT1* as fraction-specific controls of the experiment (data not shown).

Next, regarding the difference of gene annotations between public data sources, we tried to gain insights on gene boundaries by performing several PCRs across the possible transcribed regions. We attempted to perform long conventional PCR amplification of the region to fully characterize the transcript by designing several primers (**Figure 21.A**). After several attempts, we could only set up the PCR conditions for PCRs A (1607bp) and B (503bp) that successfully amplified in cDNA from CRC cell lines HCT116, SW480 and LoVo.

Alternatively, we designed primers to amplify short regions close to the expected 5' by qPCR, aiming to determine if these regions were expressed in CRC cell lines (**Figure 21.B**). All PCRs (1-5) amplified in cDNA from CRC cell lines, suggesting FOXD2-AS1 transcript is longer than the one annotated by GENCODE and RefSeq (**Figure 18**), being FANTOM CAT a more reliable source in this case. Although these data give us insights into the FOXD2-AS1 gene boundary, it is not enough to define the full-length transcript precisely.

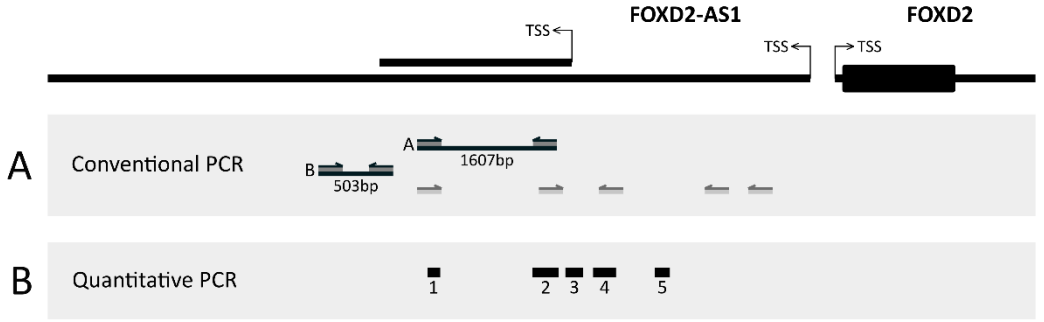


Figure 21. **Alternative FOXD2-AS1 transcripts by PCR amplification.** Genomic disposition and transcripts of FOXD2 and FOXD2-AS1. **A.** Long conventional PCR and **B.** qPCRs designs to amplify FOXD2-AS1 cDNA. In grey PCR primers we attempted to amplify.

2.3 Attempts on FOXD2 protein detection by Western blot

As FOXD2 is a protein-coding gene, we were interested in detecting the endogenous protein in CRC samples and cell lines. We attempted to quantify its protein with available antibodies; however, either the antibody had specificity problems, or the detected band did not match the expected theoretical weight of FOXD2. In consequence, we were not able to characterize FOXD2 protein due to a lack of valid antibodies.

3. Epigenetic and transcriptomic profiles of FOXD2 and FOXD2-AS1 in COAD-TCGA data

3.1 FOXD2 and FOXD2-AS1 display a coordinated expression; however, only FOXD2 is downregulated in CRC tumors

We explored the transcriptomic profiles of FOXD2 and FOXD2-AS1 using public RNA-seq data from the TCGA-COAD project that accounts for 461 colon adenocarcinoma samples and 41 matched normal colon mucosae.

We first investigated changes in expression levels between normal and tumor paired tissues (n=41). FOXD2 expression was remarkably downregulated in tumor samples compared to normal ($p<0.0001$, Wilcoxon test), while FOXD2-AS1 expression did not significantly differ between tissues ($p=0.1555$, Wilcoxon test) (**Figure 22.A**, **Supplementary Table S9**). Being FOXD2 and FOXD2-AS1 bidirectionally divergent, we wondered if they have a coordinated expression. As expected, they were significantly co-expressed in both normal ($r=0.77$, $p<0.0001$, Spearman) and tumor tissues ($r=0.48$, $p=0.0017$, Spearman)(**Figure 22.B**). In addition, we calculated the ratio of expression FOXD2-AS1/FOXD2 and found that tumors have a significantly higher ratio than normals ($p<0.0001$, Wilcoxon test) (**Figure 22.C**, **Supplementary Table S9**).

Overall, we analyzed the co-expression of FOXD2 and FOXD2-AS1 in other tissues through GEPIA webserver (Tang et al., 2017). We found a robust positive co-expression in TCGA tumors ($r=0.72$, $p<0.0001$, Spearman), in TCGA normal tissues ($r=0.91$, $p<0.0001$, Spearman), and in multiple healthy tissues from the GTEX database ($r=0.86$, $p<0.0001$, Spearman) (**Supplementary Figure S1**). These results indicate a consistent and coordinated expression of FOXD2 and FOXD2-AS1 among all cell types, suggesting a shared transcriptional mechanism.

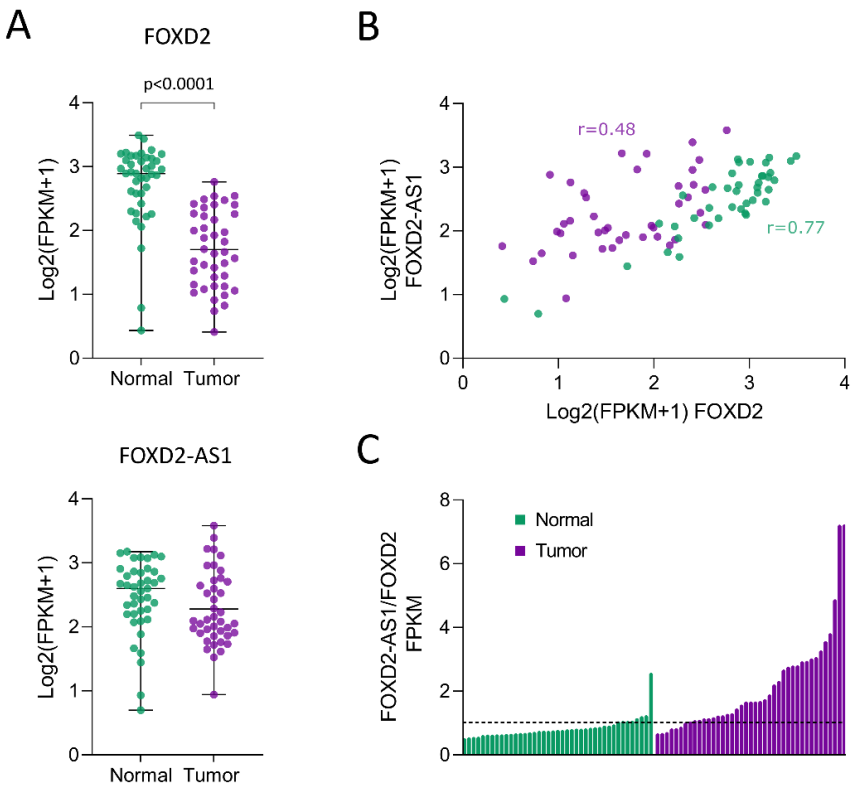


Figure 22. **FOXD2 and FOXD2-AS1 expression profiles in TCGA-COAD cohort. A.** FOXD2 and FOXD2-AS1 expression in CRC tissues compared to normal paired tissues (n=41). Only significant p values are displayed. P values were obtained using Wilcoxon test. **B.** Correlation of expression analysis between FOXD2 and FOXD2-AS1 in paired normal-tumor CRC tissues. Pearson correlation. **C.** Ratio of expression between FOXD2-AS1 and FOXD2 in normal and tumor samples in ascending disposition.

Table 12. ROC curves between normal and tumor in TCGA COAD cohort.

	FOXD2	FOXD2-AS1	FOXD2-AS1/FOXD2
AUC	0.896	0.624	0.897
95% CI	0.822 to 0.971	0.498 to 0.750	0.829 to 0.966
P value	<0.0001	0.0532	<0.0001
cuttof	< 4.464	< 3.591	> 1.080
Sensitivity%	87.8	56.1	80.49
Specificity%	75.61	75.61	90.24

Aiming to characterize the differential expression of FOXD2 and FOXD2-AS1 in normal and tumor tissue we performed ROC curve analysis to determine the cutoff values of FOXD2, FOXD2-AS1, and FOXD2-AS1/FOXD2 expression discriminating normal and tumor samples. As expected, FOXD2 and FOXD2-AS1/FOXD2 ratio had the best AUC, displaying high sensitivity and specificity (**Table 12**). These results suggest that modulation of the co-expression between FOXD2-AS1 and FOXD2 is a landmark of CRC.

3.2 Differentially expressed genes (DEGs) in high and low FOXD2 or FOXD2-AS1 expression tumors

To explore the potential mechanisms underlying the dysregulation of these genes, we compared the transcriptional profiles of tumors with low (25^{th} quartile) versus high (75^{th} quartile) expression of FOXD2 and FOXD2-AS1 in COAD tumoral samples (n=461).

The overall results are displayed using a Volcano plot (**Figure 23.A-B**). Interestingly, we identified 3805 and 3495 differentially expressed genes ($p<0.001$ and fold change >2) between tumors expressing low and high FOXD2 or FOXD2-AS1, respectively. Indeed, Gene Ontology (GO) analysis (Enrichr: https://maayanlab.cloud/Enrichr/) of DEGs revealed an enrichment of immune system processes for both comparisons based on FOXD2 and FOXD2-AS1 expression, as well as enrichment of extracellular organization regarding FOXD2 expression (**Figure 23.C-D**).

3.3 FOXD2 and FOXD2-AS1 lower expression is associated with higher mutational landscape in CRC

Next, we used the muTarget platform (www.mutarget.com) to explore mutational profiles associated with FOXD2 and FOXD2-AS1 expression changes in the TCGA-COAD cohort. We identified 1139 and 885 genes that show a statistically significant association between mutation and expression of FOXD2 or FOXD2-AS1, respectively ($p<0.01$, Mann-Whitney *U* test). The three most significant mutated genes with differential FOXD2 and FOXD2-AS1 expression are represented in **Figure 24.A-B**, being *BRAF* the most significant one. Interestingly, lower FOXD2 and FOXD2-AS1 expression levels were associated with mutant genes in CRC, except for APC, which was the only mutated gene associated with higher FOXD2-AS1 expression levels.

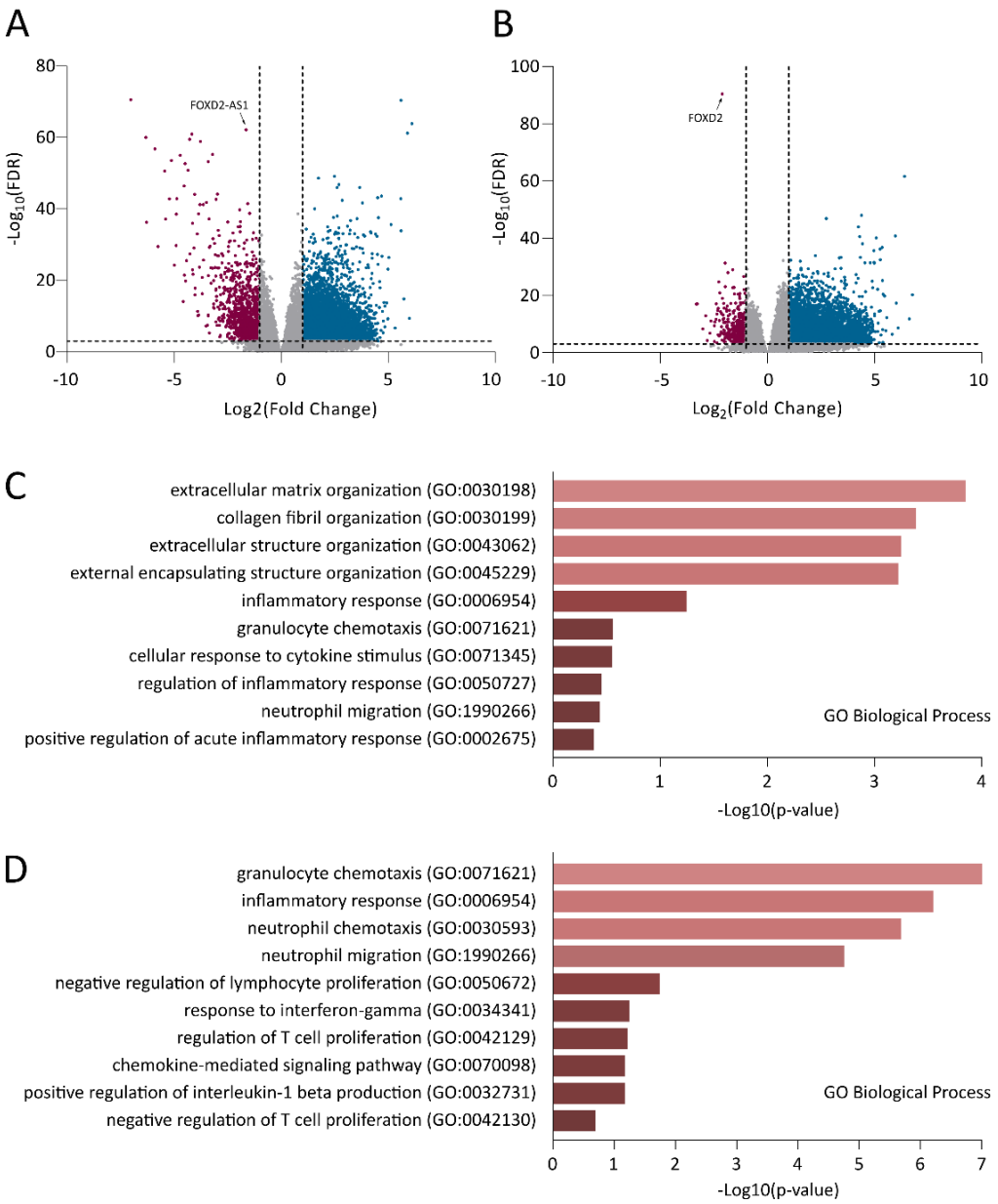


Figure 23. **Differential gene expression analysis.** Volcano plot of differentially expressed genes in tumors expressing low versus high FOXD2 (A) and FOXD2-AS1 (B), respectively. Upregulated genes are marked in blue; downregulated genes in red. DEG were selected with thresholds of fold change >2 and adjusted p<0.01. **C**. and **D.** Gene Ontology Biological Process of DEG from plots A and B, respectively. Enrichment analysis performed with Enricr (https://maayanlab.cloud/Enrichr/).

Then we wondered if the observed association could reflect a differential mutational load. Consequently, we explored tumors' overall mutation rates in coding regions and their

relationship with FOXD2 and FOXD2-AS1 expression. We observed a significant negative correlation between mutation rates and expression levels of FOXD2 (r= -0.25, p=0.002, Spearman) and FOXD2-AS1 (r=-0.17, p=0.03, Spearman) (**Figure 24.C**), suggesting that a subgroup of tumors with high mutation rates display low levels of FOXD2 and FOXD2-AS1.

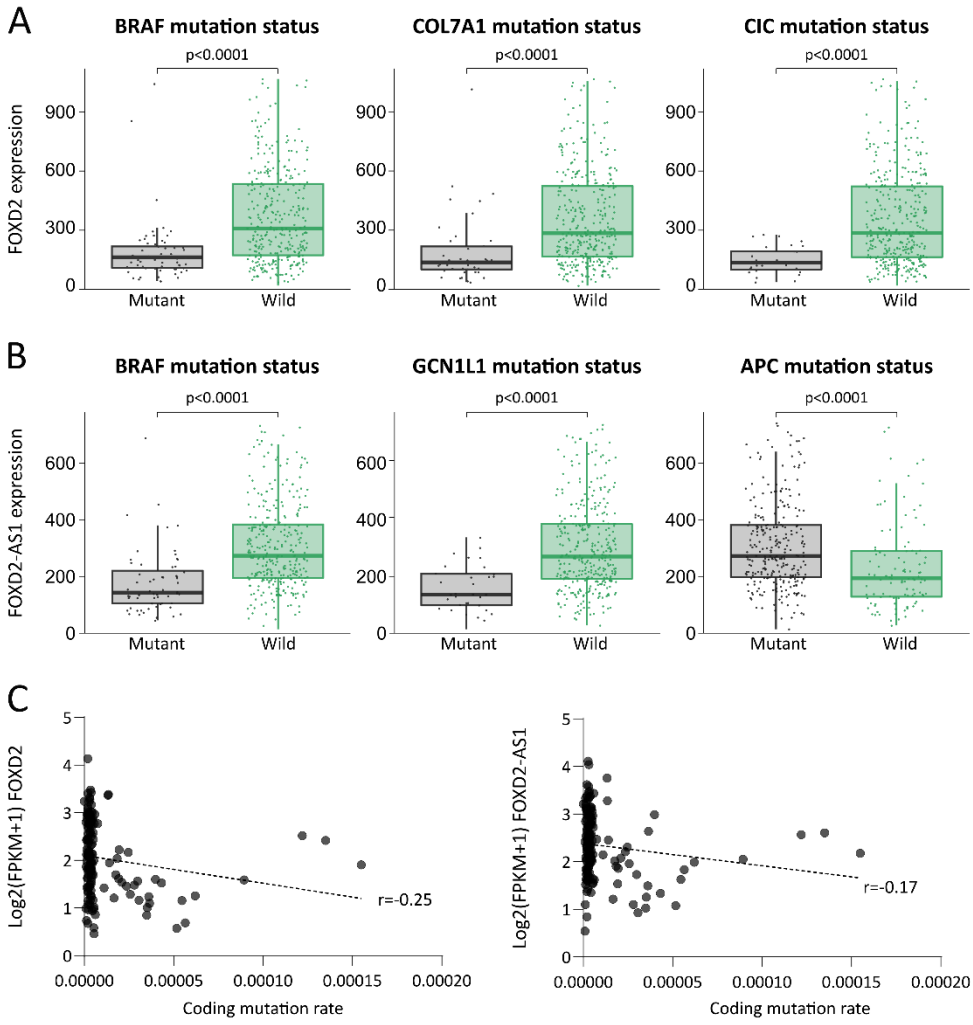


Figure 24. **Linking FOXD2 and FOXD2-AS1 expression changes to CRC genotype.** Boxplots show the three most significant mutated genes associated with changes of expression of FOXD2 (**A**) and FOXD2-AS1 (**B**). Analysis was performed with MuTarget (Nagy & Győrffy, 2021) and TCGA-COAD data (n=396). P values correspond to t-test. C. FOXD2 and FOXD2-AS1 correlations with overall coding mutation rate. Correlation coefficients r correspond to Spearman.

As the analysis generated a large number of mutated genes, we then performed a Gene Ontology enrichment analysis (GOrilla: http://cbl-gorilla.cs.technion.ac.il/) of all the genes that showed a statistically significant association between mutation states and FOXD2 and FOXD2-AS1 expression level changes. While various processes were associated with mutated genes regarding FOXD2 expression, FOXD2-AS1 deregulation was correlated with genes in cancer-associated processes (e.g., WNT-signaling pathway) (**Supplementary Table S10**). Although GO analyses are not conclusive, they can help to gain insights into the underlying biological significance of the alterations of FOXD2 and FOXD2-AS1 expression in colorectal cancers.

3.4 Tumors display coordinated hypermethylation outside the CpGi promoter

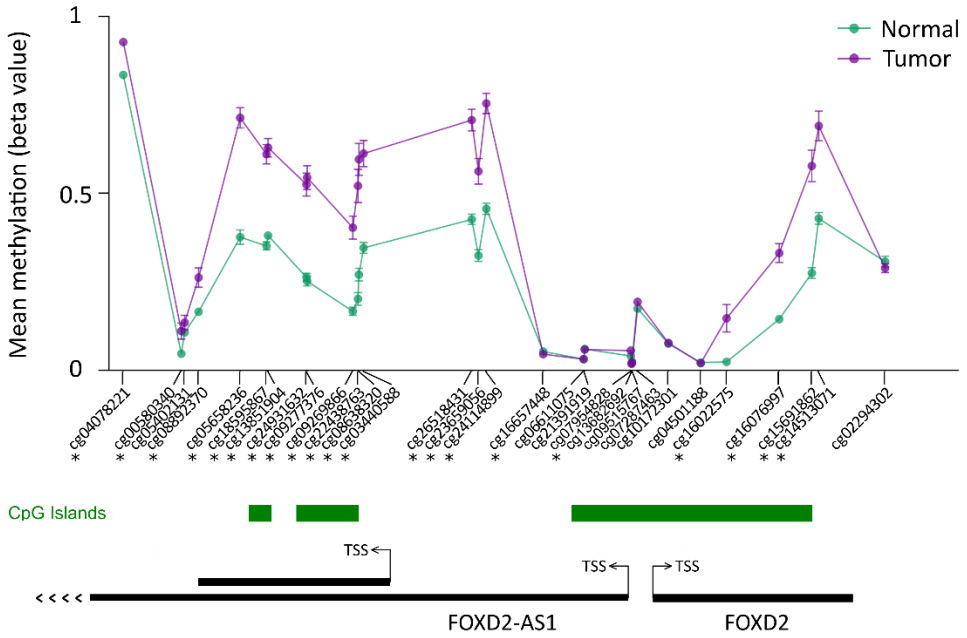


Figure 25. **DNA methylation changes between normal and tumor TCGA COAD samples.** Graph represents CpGs (x-axis) covering the FOXD2 and FOXD2-AS1 locus and DNA methylation level (y-axis) in colon cancer and normal samples. * Differences are statistically significant between normal and tumor according to Wilcoxon test. CpG islands are indicated in green and genes and their TSS in black.

To bring light to the putative mechanisms of FOXD2 and FOXD2-AS1 expression changes between normal and tumor tissues, we analyzed the DNA methylation profiles from the TCGA-COAD cohort. We downloaded the Illumina 450K methylation array data using our group's Wanderer web tool (Díez-Villanueva, Mallona, & Peinado, 2015). Data analyzed includes methylation from 302 colon tumors and 38 paired normal adjacent tissues. We

studied the methylation of 30 CpGs distributed along our region of interest of about 9 kb long.

The genomic locus of FOXD2 and FOXD2-AS1 is a CpG rich region that contains three CpG islands. We observed a general hypermethylation pattern in paired tumor samples compared to normal (n=38), having almost all individual CpGs significantly higher methylation levels in tumor samples (**Figure 25**). We first paid special attention to the promoter region defined by active histone marks at the TSS reported by FANTOM CAT (**Figure 18**). Interestingly, the promoter was completely unmethylated in all tissues analyzed independently if they were normal or tumors, and although some were significantly methylated in tumors, the differences were minimal. These results suggest that expression changes of these genes are not associated with promoter DNA hypermethylation. In contrast, CpGs displaying recurrent hypermethylation in the tumor compared with the normal were located at the 3' end of FOXD2 gene and along the gene body of FOXD2-AS1 (**Figure 25**).

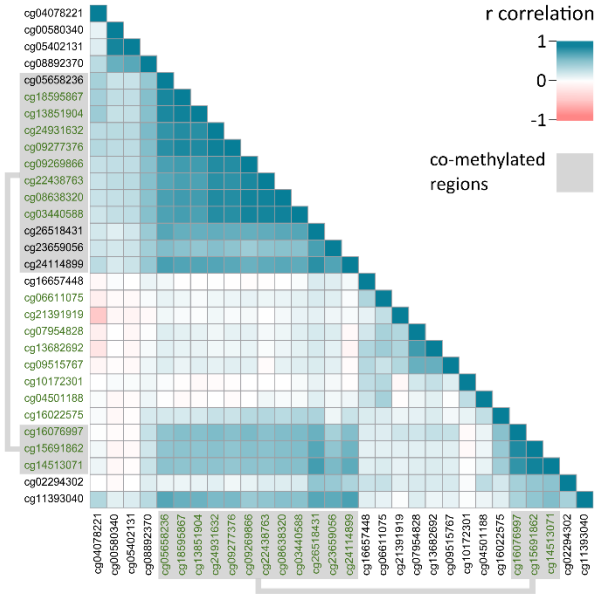


Figure 26. **Coordinated hypermethylation between CpGs in COAD tumors**. Correlation matrix describing pair wise Pearson correlation of methylation status of 30 CpGs sites in TCGA-COAD tumors (n=302). CpG sites are arranged according to their position within the FOXD2 and FOXD2-AS1 coding region. CpGs in green indicate CpG island. Co-methylated regions are highlighted with a grey box.

As the probes that had DNA hypermethylation in tumor samples are distributed in alarge region, we investigated the relationship between DNA methylation levels of individual CpGs within an individual. We performed Spearman correlations and plotted a correlation

matrix representing the methylation changes happening in all the tumor samples available from COAD-TCGA (n=302) (**Figure 26**).

CpGs with coordinated hypermethylation (r=0.36-0.76, p<0.0001) were clustered in two regions partially overlapping with two CpGi separated by about 6kb, one located at the FOXD2-AS1 gene body and at the 3' UTR of FOXD2 gene. These results indicate that regional DNA hypermethylation follows a specific pattern and that the FOXD2 promoter CpGi retains local protection from hypermethylation (**Figure 25** and **Figure 26**).

Upon seeing methylation differences between normal and tumor tissues, we plotted a heatmap of 38 paired normal and tumoral samples to study the potential value of the DNA methylation of this region to differentiate tissue types (**Figure 27**). Interestingly, the methylation levels were accurate enough to distinguish the tumor samples from the normal.

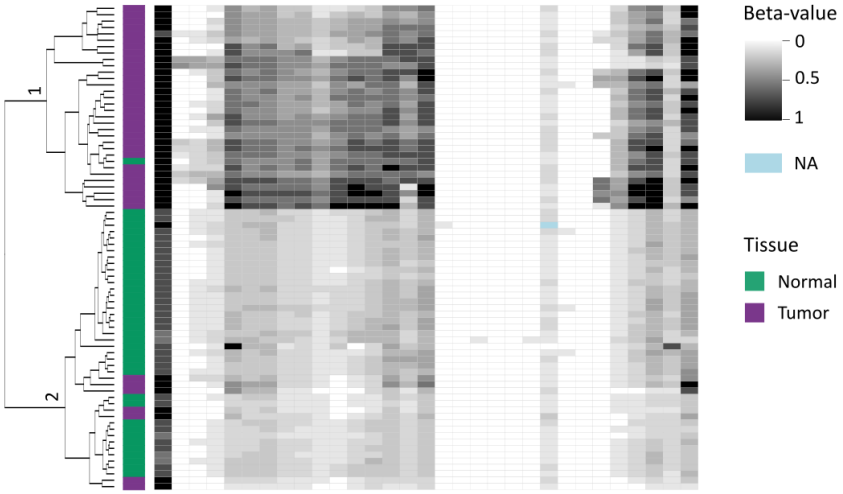


Figure 27. **DNA methylation segregates normal and tumor colon samples.** Heatmap representing DNA methylation levels between normal and tumor samples. Each column represents a CpG probes, each row corresponds to a tumor or normal colon sample as indicated on the right bar. Clustered samples are indicated on the left. Data from TCGA-COAD cohort.

3.5 DNA methylation is negatively correlated with FOXD2 and FOXD2-AS1 expression

After exploring methylation and expression changes separately, we wondered if the expression patterns observed for FOXD2 and FOXD2-AS1 genes were associated with

methylation alterations in the region of study. Consequently, we performed correlation analysis between each individual CpG and the expression levels of both genes in 300 patients with both RNA-seq and methylation array data available.

Table 13. DNA methylation correlation with gene expression in TCGA COAD cohort. Spearman correlation. Significant p values are displayed in bold.

	FOXD2 expression			FOXD2-AS1 expression		
CpG ID	n	r	p value	n	r	p value
cg04078221	319	-0.1532	**0.0061**	319	-0.04766	0.3963
cg00580340	319	-0.03307	0.5562	319	-0.05733	0.3073
cg05402131	319	0.0342	0.5428	319	-0.04593	0.4136
cg08892370	319	-0.1709	**0.0022**	319	-0.134	**0.0166**
cg05658236	319	-0.2938	**<0.0001**	319	-0.1266	**0.0238**
cg18595867	319	-0.286	**<0.0001**	319	-0.09985	0.0749
cg13851904	319	-0.2916	**<0.0001**	319	-0.1134	**0.0431**
cg24931632	319	-0.288	**<0.0001**	319	-0.08282	0.14
cg09277376	319	-0.2724	**<0.0001**	319	-0.09256	0.0989
cg09269866	319	-0.3325	**<0.0001**	319	-0.1502	**0.0072**
cg22438763	319	-0.3101	**<0.0001**	319	-0.1353	**0.0156**
cg08638320	319	-0.3138	**<0.0001**	319	-0.1326	**0.0178**
cg03440588	319	-0.3467	**<0.0001**	319	-0.1777	**0.0014**
cg26518431	319	-0.6744	**<0.0001**	319	-0.4908	**<0.0001**
cg23659056	319	-0.4496	**<0.0001**	319	-0.3334	**<0.0001**
cg24114899	319	-0.5202	**<0.0001**	319	-0.2828	**<0.0001**
cg16657448	319	-0.08538	0.1281	319	-0.1222	**0.0291**
cg06611075	319	-0.1208	**0.031**	319	-0.1524	**0.0064**
cg21391919	319	-0.1543	**0.0058**	319	-0.198	**0.0004**
cg07954828	319	-0.2174	**<0.0001**	319	-0.1502	**0.0072**
cg13682692	319	-0.1429	**0.0106**	319	-0.1447	**0.0097**
cg09515767	319	-0.1287	**0.0215**	319	-0.1488	**0.0078**
cg07287463	312	-0.0511	0.3683	312	-0.102	0.0719
cg10172301	319	-0.08924	0.1117	319	-0.1201	**0.032**
cg04501188	319	-0.124	**0.0267**	319	-0.1304	**0.0198**
cg16022575	319	-0.4027	**<0.0001**	319	-0.2862	**<0.0001**
cg16076997	319	-0.613	**<0.0001**	319	-0.3468	**<0.0001**
cg15691862	319	-0.6381	**<0.0001**	319	-0.3483	**<0.0001**
cg14513071	319	-0.6529	**<0.0001**	319	-0.3541	**<0.0001**
cg02294302	319	-0.4142	**<0.0001**	319	-0.2756	**<0.0001**

As expected, most CpGs displayed significant negative correlations between DNA methylation and gene expression, and in general, associations were more substantial with FOXD2 expression than with FOXD2-AS1. Interestingly, CpGs at the 3' of FOXD2 had the most prominent associations with expression levels, followed by CpGs through the gene body of FOXD2-AS1 (**Table 13**). Even though gene silencing by CpGi promoter hypermethylation is the most common association described in many studies, our results suggest a more complex view of the correlation between DNA hypermethylation and gene expression patterns.

4. Epigenetic and transcriptomic profiles of FOXD2/FOXD2-AS1 in normal-tumor paired colorectal tissues from HUB

4.1 FOXD2 and FOXD2-AS1 co-express and are downregulated in CRC

To further elucidate the involvement of FOXD2 and FOXD2-AS1 in CRC development and validate the results observed in COAD patients from the TCGA database, we analyzed another cohort of about 100 colorectal carcinoma tissues and adjacent normal mucosae collected at the Hospital de Bellvitge (HUB), Barcelona.

We first assessed the expression of FOXD2 and FOXD2-AS1 by RT and qPCR analysis (n=108). Results showed a significant downregulation of FOXD2 and FOXD2-AS1 in tumor tissues relative to normal adjacent tissues ($p<0.0001$, Wilcoxon test) as well as an increase of FOXD2-AS1/FOXD2 expression ratio ($p<0.0001$, Wilcoxon test) (**Figure 28**, **Supplementary Table S11**). As expected, FOXD2 and FOXD2-AS1 showed a significant positive correlated expression in both normal and tumor tissues ($r=0.73$ and 0.67, $p<0.0001$, Spearman).

In addition, ROC curves analysis showed the potential value of these expression parameters as biomarkers to distinguish normal and tumoral cells, whereas FOXD2 would be the most informative for discrimination between normal and tumor tissues (AUC=0.89), followed by the ratio FOXD2-AS1/FOXD2 (AUC=0.80) and FOXD2-AS1 (AUC=0.74) (**Table 14**).

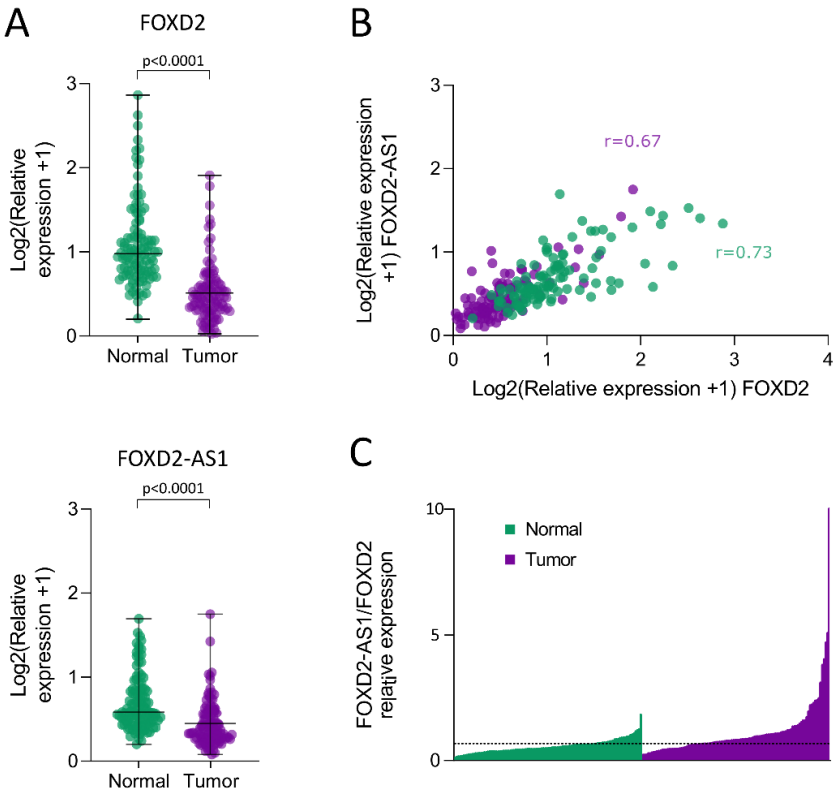


Figure 28. **FOXD2 and FOXD2-AS1 expression profiles in HUB cohort. A.** FOXD2 and FOXD2-AS1 expression in CRC tissues compared to normal paired tissues (n=108). Only significant p values are displayed. P values were obtained using Wilcoxon test. **B.** Correlation of expression analysis between FOXD2 and FOXD2-AS1 in paired normal-tumor CRC tissues. Pearson correlation. **C.** Ratio of expression between FOXD2-AS1 and FOXD2 in normal and tumor samples in ascending disposition. qPCR analysis were analyzed relative to PSMC4 and PUM1.

Table 14. ROC curves between normal and tumor in HUB cohort.

	FOXD2	FOXD2-AS1	FOXD2-AS1/FOXD2
AUC	0.8785	0.754	0.759
95% CI	0.831 to 0.926	0.690 to 0.819	0.695 to 0.823
P value	<0.0001	<0.0001	<0.0001
cuttof	< 0.5725	< 0.3737	> 0.6825
Sensitivity%	77.78	65.74	72.22
Specificity%	87.04	78.7	73.15

4.2 Methylation gain in tumors outside of the CpGi promoter

The DNA methylation at FOXD2 and FOXD2-AS1 loci was explored by Bisulfite sequencing in paired tissues. Due to the large size of our region of interest, three different regions of approximately 300bp were PCR amplified and Sequenced by Sanger (**Figure 29.A**). Region 1 is in FOXD2-AS1 coding gene and includes the cg08638320 from the Illumina 450K methylation array. Region 2 contains the promoter and is very close to cg21391919, and region 3 covers the 3' of FOXD2 including the CpG cg14513071 from the Illumina array.

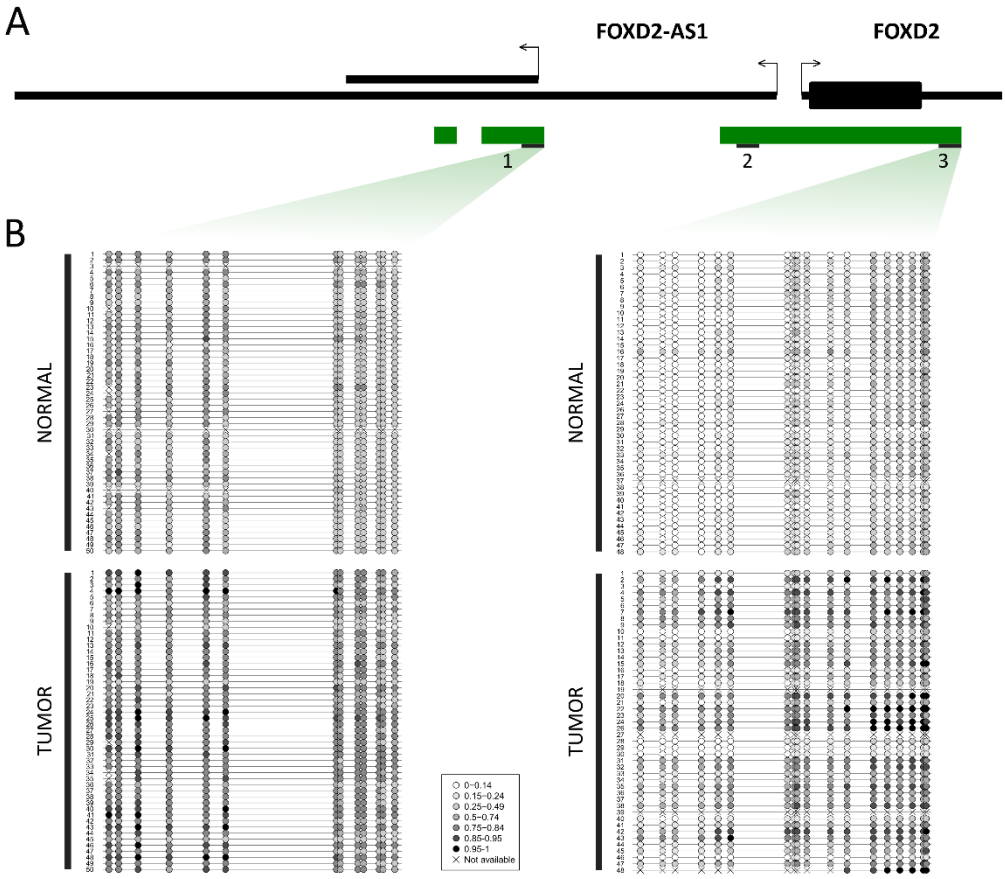


Figure 29. **FOXD2 and FOXD2-AS1 methylation profiles in the HUB cohort. A.** Gene (black) and CpG island (green) arrangement. **B.** DNA methylation values of paired normal-tumor tissues in regions 1 and 3. Each circle represents a CpG site. Data from HUB cohort.

As we previously observed in the TCGA-COAD cohort that the promoter region was completely unmethylated in all samples, we only analyzed methylation levels of region 2 in 25 patients. After observing the same pattern in HUB cohort (**Supplementary Figure S2**),

we assumed the promoter remains unmethylated in all patient samples. On the other hand, we analyzed regions 1 and 3 in paired samples of 50 patients, while the rest of the cohort is being analyzed. We found regions 1 and 3 significantly hypermethylated in tumor samples compared to normal ($p<0.0001$, Wilcoxon test) (**Figure 29**), in accordance with TCGA-COAD cohort. Furthermore, we observed positive co-methylation patterns between regions 1 and 3, as previously reported with TCGA-COAD cohort (**Supplementary Figure S3**).

To get further insights into the differential DNA methylation profiles between normal and tumor tissue, we created a heatmap of the methylation values of all the individual CpGs from Regions 1 and 3 (**Figure 30**). Overall, three main clusters classifying samples were generated. Most normal tissues were grouped in cluster 3, while cluster 1 was integrated by tumors with high hypermethylation levels in most CpGs of regions 1 and 3. Finally, cluster 2 consisted mainly of tumors with hypermethylation in region 1, but not in region 3, although six normal tissues also clustered with these tumors.

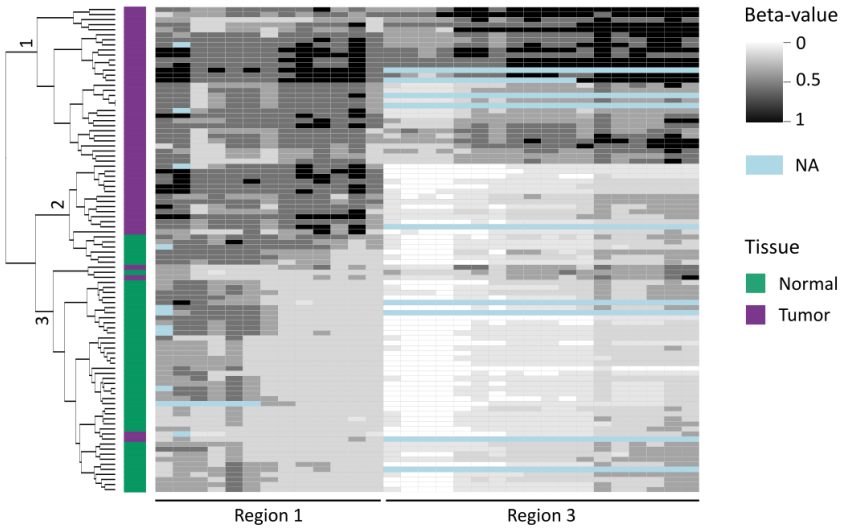


Figure 30. **DNA methylation segregates normal and tumor colon samples.** Heatmap representing DNA methylation levels in regions 1 and 3 between normal and tumor samples. Each column represents a CpG probe; each row corresponds to a tumor or normal colon sample (color coded on the left bar). Data from HUB cohort.

4.3 DNA methylation negatively correlates with FOXD2 and FOXD2-AS1 expression

Table 15. DNA methylation correlation with gene expression in HUB cohort. Spearman correlation. Significant p values are displayed in bold.

			FOXD2 expression			FOXD2-AS1 expression		
Region	**Gene body**	**CpG #**	n	r	p value	n	r	p value
1	FOXD2-AS1	CpG_1	98	-0.202	**0.0466**	98	-0.05	0.6239
		CpG_2	97	-0.279	**0.0056**	97	-0.061	0.5514
		CpG_3	98	-0.273	**0.0066**	98	-0.042	0.6791
		CpG_4	98	-0.195	0.0548	98	-0.077	0.4528
		CpG_5	98	-0.307	**0.0021**	98	-0.155	0.1273
		CpG_6	98	-0.406	**<0.0001**	98	-0.21	**0.0382**
		CpG_7	98	-0.321	**0.0013**	98	-0.187	0.0654
		CpG_8	98	-0.41	**<0.0001**	98	-0.241	**0.0169**
		CpG_9	98	-0.403	**<0.0001**	98	-0.22	**0.0298**
		CpG_10	98	-0.375	**0.0001**	98	-0.234	**0.0206**
		CpG_11	98	-0.226	**0.025**	98	-0.1	0.3296
		CpG_12	98	-0.343	**0.0005**	98	-0.239	**0.0179**
3	FOXD2	CpG_1	88	-0.548	**<0.0001**	88	-0.338	**0.0013**
		CpG_2	88	-0.548	**<0.0001**	88	-0.338	**0.0013**
		CpG_3	88	-0.552	**<0.0001**	88	-0.312	**0.0031**
		CpG_4	88	-0.575	**<0.0001**	88	-0.33	**0.0017**
		CpG_5	88	-0.595	**<0.0001**	88	-0.382	**0.0002**
		CpG_6	88	-0.539	**<0.0001**	88	-0.344	**0.001**
		CpG_7	88	-0.538	**<0.0001**	88	-0.286	**0.0069**
		CpG_8	88	-0.505	**<0.0001**	88	-0.214	**0.0456**
		CpG_9	88	-0.533	**<0.0001**	88	-0.322	**0.0022**
		CpG_10	88	-0.53	**<0.0001**	88	-0.294	**0.0055**
		CpG_11	88	-0.521	**<0.0001**	88	-0.306	**0.0038**
		CpG_12	88	-0.501	**<0.0001**	88	-0.352	**0.0008**
		CpG_13	89	-0.551	**<0.0001**	89	-0.349	**0.0008**
		CpG_14	89	-0.568	**<0.0001**	89	-0.313	**0.0028**
		CpG_15	89	-0.609	**<0.0001**	89	-0.325	**0.0019**
		CpG_16	89	-0.585	**<0.0001**	89	-0.365	**0.0004**
		CpG_17	89	-0.578	**<0.0001**	89	-0.353	**0.0007**
		CpG_18	89	-0.532	**<0.0001**	89	-0.351	**0.0007**

We previously observed a correlation pattern between DNA methylation outside the promoter and gene expression (see results **3.4**). By analyzing another cohort of patients using different experimental approaches, we confirmed that a gain of methylation in

regions 1 and 3 was associated with lower FOXD2 and FOXD2-AS1 expression levels (**Table 15**). More precisely, region 3 displayed the highest associations with FOXD2 expression.

5. Clinical associations of FOXD2 and FOXD2-AS1 expression and methylation profiles

After characterizing the expression and methylation profiles of FOXD2 and FOXD2-AS1, we wondered if those parameters were associated with clinicopathological features of the tumors. We considered both the TCGA-COAD and HUB series to perform the analysis, as they are very different cohorts regarding numbers and patients' characteristics. The clinicopathological characteristics of each cohort are summarized in **Table 2** (HUB) and **Table 8** (TCGA-COAD).

5.1 Clinicopathological features and overall survival (OS) associated with FOXD2 and FOXD2-AS1 expression

Table 16. Correlation between FOXD2 and FOXD2-AS1 expression and the clinical pathological parameters of TCGA-COAD pateints. P values correspond to Mann-Whitney test or Krustal-Wallis test. Significant p values are displayed in bold.

			FOXD2		FOXD2-AS1		FOXD2-AS1/FOXD2	
Variables		**n**	**mean ± SD**	**p value**	**mean ± SD**	**p value**	**mean ± SD**	**p value**
Cancer Stage	I	78	3.14 ± 2.48	0.9247	4.57 ± 2.58	0.6962	1.81 ± 0.98	0.1066
	II	180	2.94 ± 2.24		4.51 ± 2.93		1.93 ± 1.11	
	III	127	3.06 ± 2.52		5.19 ± 4.15		2.12 ± 1.27	
	IV	64	2.93 ± 2.06		4.91 ± 2.99		2.28 ± 1.51	
Tumor size and invasion	T1	11	3.39 ± 2.46	0.0842	4.74 ± 4.49	0.889	1.71 ± 0.89	**0.0027**
	T2	80	3.25 ± 2.49		4.57 ± 2.57		1.73 ± 0.97	
	T3	311	3.08 ± 2.36		4.95 ± 3.52		2.05 ± 1.26	
	T4	60	2.36 ± 1.76		4.63 ± 2.91		2.38 ± 1.28	
Lymph node involvement	N0	273	3.03 ± 3.03	0.1318	4.61 ± 2.92	0.58	1.90 ± 1.08	**0.0001**
	N1	104	3.20 ± 3.20		5.20 ± 4.00		1.95 ± 1.23	
	N2	83	2.76 ± 2.76		5.12 ± 3.61		2.53 ± 1.50	
Distant methastasis	M0	337	3.06 ± 2.41	0.9117	4.69 ± 3.31	0.2811	1.91 ± 1.05	**0.0364**
	M1	64	2.93 ± 2.06		4.91 ± 2.99		2.28 ± 1.51	

The expression levels of FOXD2 and FOXD2-AS1 were analyzed to investigate possible associations with clinicopathological parameters. We first included 451 tumoral samples from the TCGA-COAD cohort. **Table 16** shows that FOXD2-AS1/FOXD2 expression ratio was

remarkably higher in patients with tumor invasion (p=0.027, Kruskal-Wallis test), lymph node metastasis (p<0.0001, Kruskal-Wallis test), and distant metastasis (p=0.0364, Mann-Whitney test). However, neither FOXD2 nor FOXD2-AS1 alone were associated with the tumor's clinicopathological features, even though a tendency of a lower FOXD2 and a higher FOXD2-AS1 expression was associated with higher tumor stages, invasion, lymph node involvement, and distant metastasis. No significant associations were found with patients age, gender nor tumor site (data not shown).

In contrast, when we performed the same analysis for the 108 patients from HUB, neither FOXD2 nor FOXD2-AS1 nor their ratio of expression showed significant associations with clinicopathological features of the tumor (**Table 17**).

Table 17 . Correlation between FOXD2 and FOXD2-AS1 expression and the clinical pathological parameters of HUB pateints. P values correspond to Mann-Whitney test or Krustal-Wallis. Significant p values are displayed in bold.

			FOXD2		FOXD2-AS1		FOXD2-AS1/FOXD2	
Variables		**n**	**mean ± SD**	**p value**	**mean ± SD**	**p value**	**mean ± SD**	**p value**
Cancer Stage	II	60	0.53 ± 0.53	0.4815	0.44± 0.39	0.2798	1.10 ± 0.83	0.736
	III	48	0.41 ± 0.30		0.34 ± 0.20		1.28 ± 1.62	
Tumor size and invasion	T1 & T2	5	0.38 ± 0.11	0.3831	0.31 ± 0.16	0.911	0.81 ± 0.35	0.2466
	T3	78	0.52 ± 0.50		0.41 ± 0.36		1.16 ± 1.35	
	T4	25	0.35 ± 0.21		0.37 ± 0.23		1.31 ± 0.98	
Lymph node involvement	N0	60	0.53 ± 0.53	0.3925	0.44 ± 0.39	0.4852	1.10 ± 0.83	0.9109
	N1	29	0.43 ± 0.27		0.34 ± 0.17		1.25 ± 1.84	
	N2	19	0.37 ± 0.35		0.34 ± 0.25		1.32 ± 1.26	

Regarding the prognostic function of FOXD2 and FOXD2-AS1 expression in overall survival, we performed ROC curves to determine the best cutoff value for survival analysis (**Supplementary Table S12**). Kaplan-Meier analysis of the TCGA-COAD cohort revealed that expression of each individual gene does not result in useful prognostic factors in CRC. Nevertheless, Kaplan-Meier analysis taking into account the ratio of expression of FOXD2-AS1/FOXD2, exhibited a significant prognostic value in predicting the OS of patients with CRC (HR=2.5, p<0.0001 long rank test) (**Figure 31**, **Supplementary Table S12**).

Patients from the HUB cohort are only distributed into stages II and III, resulting in ~90% of the patients surviving the disease. Consequently, the OS analysis was of limited significance due to the uneven distribution of survivals and non-survivals. As expected,

when we performed ROC curves, results were very poor, displaying low AUC values (data not shown).

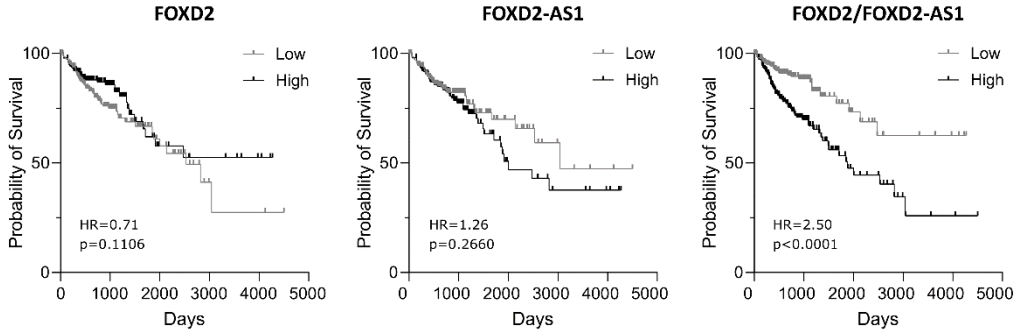


Figure 31. **Overall Survival analysis regarding FOXD2 and FOXD2-AS1 expression.** Klapan-meier overall survival (OS) curves according to FOXD2, FOXD2-AS1 and FOXD2/FOXD2-AS1 expression levels.

5.2 Clinicopathological features and overall survival (OS) associated to FOXD2 and FOXD2-AS1 methylation patterns

We next evaluated the correlations between methylation patterns and the clinicopathological features of the tumors. Neither analysis with TCGA-COAD cohort (~250) nor HUB cohort (n~50) showed relevant associations of DNA methylation with clinicopathological features of the tumors (data not shown).

Overall, we did not observe a potential value for gene expression or methylation levels to stratify CRC patients. However, in TCGA cohort, a higher ratio of expression FOXD2-AS1/FOXD2 was associated with tumor malignancy.

6. Analysis of FOXD2 and FOXD2-AS1 functions in CRC cell lines

In order to explore the potential implications of FOXD2 and FOXD2-AS1 deregulation in cancer, we started by comparing their levels of expression by RT-qPCR in three human colorectal adenocarcinoma cell lines: HCT116, LoVo, and SW480. As shown in **Figure 32.C**, LoVo cells exhibited the highest expression levels, while SW480 cells showed the lowest expression for both FOXD2 and FOXD2-AS1 genes. Indeed, FOXD2 displayed 1.8 to 3.5 fold

expression relative to FOXD2-AS1, which is expected in protein-coding genes in head-to-head orientation to lncRNAs (Derrien et al., 2012).

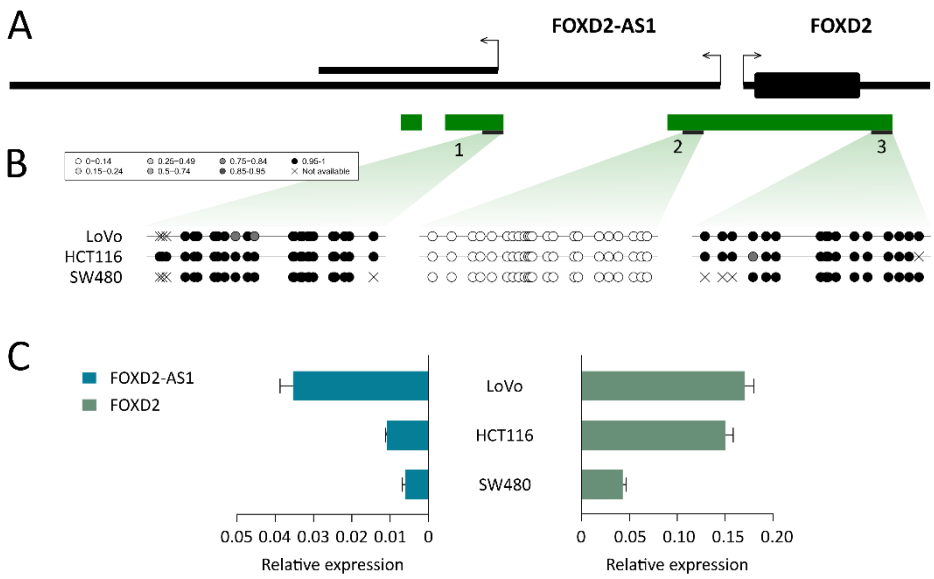


Figure 32. FOXD2 and FOXD2-AS1 expression and DNA methylation profiles in CRC cell lines. A. Gene locations and the corresponding CpG islands in green, divided in three regions of study. **B.** DNA methylation profiles of FOXD2 and FOX2-AS1 in the three flanking regions in LoVo, HCT116 and SW480 cell lines. Each circle indicates a CpG site, distributed proportionally to their position and color coded by beta-values. **C**. qPCR expression levels in LoVo, HCT116 and SW480 cell lines. Bar plots correspond to the average value of triplicates ± SD. qPCR analysis were normalized using to MRPL9, PSMC4 and PUM1 as reference genes.

In addition, we checked their methylation levels by bisulfite treatment, PCR amplification, and Sanger sequencing on three representative regions of the FOXD2 and FOXD2-AS1 locus, as previously described (**Figure 32.A**). Overall, all cell lines displayed the same methylation pattern across the region. Similar to primary tumor samples from HUB and TCGA-COAD cohorts, regions 1 and 3 were heavily methylated while region 2, which covers the promoter, was completely unmethylated in all cell lines (**Figure 32.B**).

6.1 Induction of DNA demethylation reactivates gene expression

We previously reported a negative association between gene expression and methylation levels of regions 1 and 3 (see results 3.5 and 4.3). To expand our observations on how DNA methylation regulates the expression of FOXD2 and FOXD2-AS1, we treated CRC cell lines with 5-aza-2'-deoxycytidine (DAC) DNA hypomethylating agent. After DAC treatment, we assessed the demethylating effect by measuring the methylation levels of regions 1 and 3

compared to the non-treated cells. DAC for 48h hours produced heterogeneous results in the three cell lines and within regions 1 and 3. Region 1 displayed complete demethylation in HCT116 and partial loss in SW480 and LoVo. However, region 3 demethylation was observed only for HCT116 and Lovo, while SW480 remained methylated (**Figure 33**).

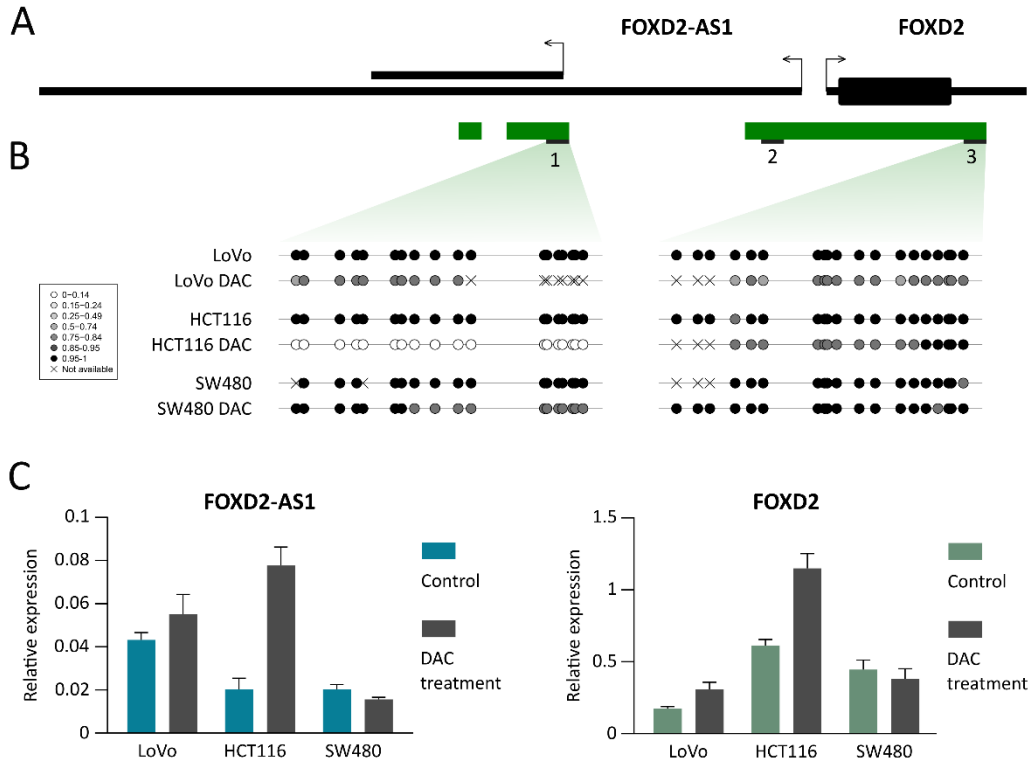


Figure 33. **DAC treatment effects on methylation and its association with expression. A.** Gene locations and the corresponding CpG islands (in green) are displayed indicating three regions of study **B**. DNA methylation profiles of untreated (control) and treated cell lines with DAC demethylating reagent. Each circle indicates a CpG site, distributed proportionally to their position and color coded by beta-values**. C.** qPCR expression levels of the corresponding DAC treated cell lines. Bar plots correspond to the average value of triplicates ± SD. qPCR analysis were analyzed relative to MRPL9, PSMC4 and PUM1**.**

Expression results revealed strong upregulation of both genes in HCT116, a tendency for higher expression levels in LoVo, and no changes in SW480 cell line. Notably, the relative increase in gene expression was consistent with methylation loss in regions 1 and 3, once more suggesting a negative association between methylation and expression changes.

6.2 CRISPR SAM assay induces coordinated overexpression of FOXD2 and FOXD2-AS1

As reviewed in the introduction, several published studies have reported a role for FOXD2-AS1 as tumor promoter by enhancing cell proliferation, migration, and invasion in several cancers, including colorectal cancer (Yang, Duan, & Zhou, 2017; Zhu et al., 2018). However, little is known about the functional implications of FOXD2 transcription factor or its putative relationship with lncRNA FOXD2-AS1 transcript. Consequently, we developed in vitro models with different genome editing tools to achieve different expression levels of FOXD2 and FOXD2-AS1 by themselves and in combination.

We first used the novel and powerful CRISPR/Cas9 synergistic activation mediator (SAM) technique to overexpress FOXD2 and FOXD2-AS1 in SW480, as it was the cell line with the lowest expression levels. The CRISPR SAM system is based on a single guide RNA (sgRNA) that directs the assembly of a multi-component transcriptional activation complex at a target site to induce endogenous transcriptional activation. The optimal transcriptional activation effect obtained has been reported when the sgRNA targets the first 200bp before the TSS (Konermann et al., 2015). According to this criterion, we first designed three different sgRNA for each gene (sgRNA a-f) according to the official TSS (RefSeq) using the GPP sgRNA Design tool from the Broad Institute (https://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design) (**Figure 34.A**). Unfortunately, none of the sgRNA achieved a significant upregulation, discarding them from the analysis. Considering that the distance to the TSS is a crucial factor in predicting the best target sites for sgRNAs, we designed other sgRNA (sgRNA g-l) according to the TSS reported by FANTOM CAT (**Figure 34.B**). Out of all designs, only "sgRNA j" provided suitable levels of upregulation. Such sgRNA was located ~100bp before FOXD2-AS1 TSS and ~225bp before FOXD2 TSS, resulting in increased expression of both genes. However, compared to the negative control transfected with an empty sgRNA vector, FOXD2-AS1 gene displayed a higher expression increase than FOXD2 (**Figure 34.C**).

Furthermore, we expanded clones from the overexpressing pool of cells with the transfected sgRNA "j" and assessed their expression levels. Most clones displayed a relevant upregulation of FOXD2 and FOXD2-AS1 (**Figure 34.D**), but more interestingly, they revealed a robust significant co-expression pattern ($r=0.98$, $p<0.0001$, Pearson) (**Figure 34.E**). These results suggest a strong co-regulation of gene expression through induction of transcription in both directions.

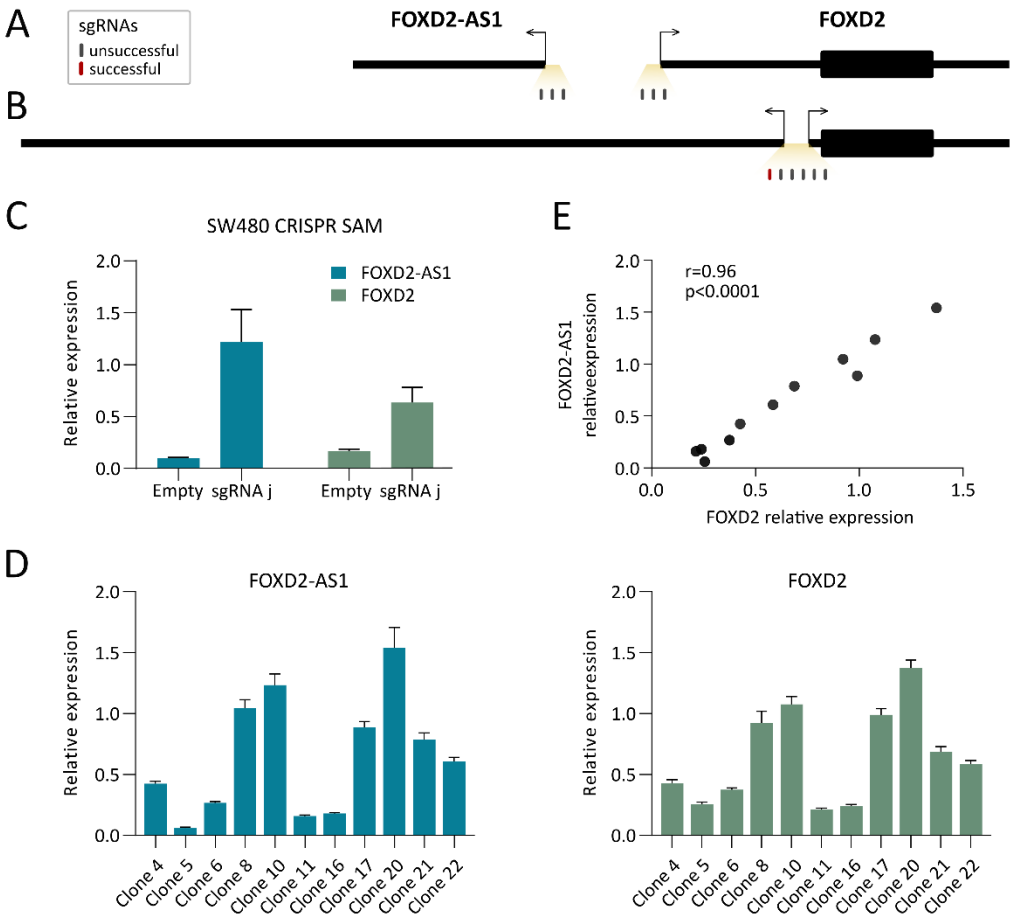


Figure 34. CRISPR SAM overexpression of FOXD2 and FOXD2-AS1 in SW480. A. and **B.** Representation of gene boundaries and sgRNA schematic desig. sgRNA j with successful overexpression is indicated in red, the rest of unsuccessful sgRNAs in grey. **C.** qPCR relative expression of FOXD2 and FOXD2-AS1 in CRISPR SAM empty sgRNA and sgRNA "j" in SW480 cell line. **D.** qPCR relative expression of sgRNA "j" clones overexpressing FOXD2 and FOXD2-AS1 in SW480 cell line. **E.** Correlation of expression between FOXD2 and FOXD2-AS1 in sgRNA "j" clones. P value and r coefficient correspond to Pearson correlation. Bar plot data correspond to the average value of triplicates ± SD. qPCR analysis were normalized using MRPL9 and PMSC4 as reference genes.

6.3 Coordinated FOXD2 and FOXD2-AS1 overexpression has no effect on cell proliferation, migration, nor colony formation

At the morphological level, cells overexpressing FOXD2 and FOXD2-AS1 induced by CRISPR SAM did not present any distinctive trait, looking identical to the empty vector cells and the untreated SW480. As FOXD2-AS1 has been reported to promote proliferation, migration, and colony formation, we wondered whether FOXD2-AS1 and FOXD2

simultaneous overexpression enhanced proliferation compared to negative controls. As observed in **Figure 35**, coordinated up-regulation of FOXD2 and FOXD2-AS1 did not affect proliferation rates, migration abilities, or the capacity to form colonies in SW480 cell line.

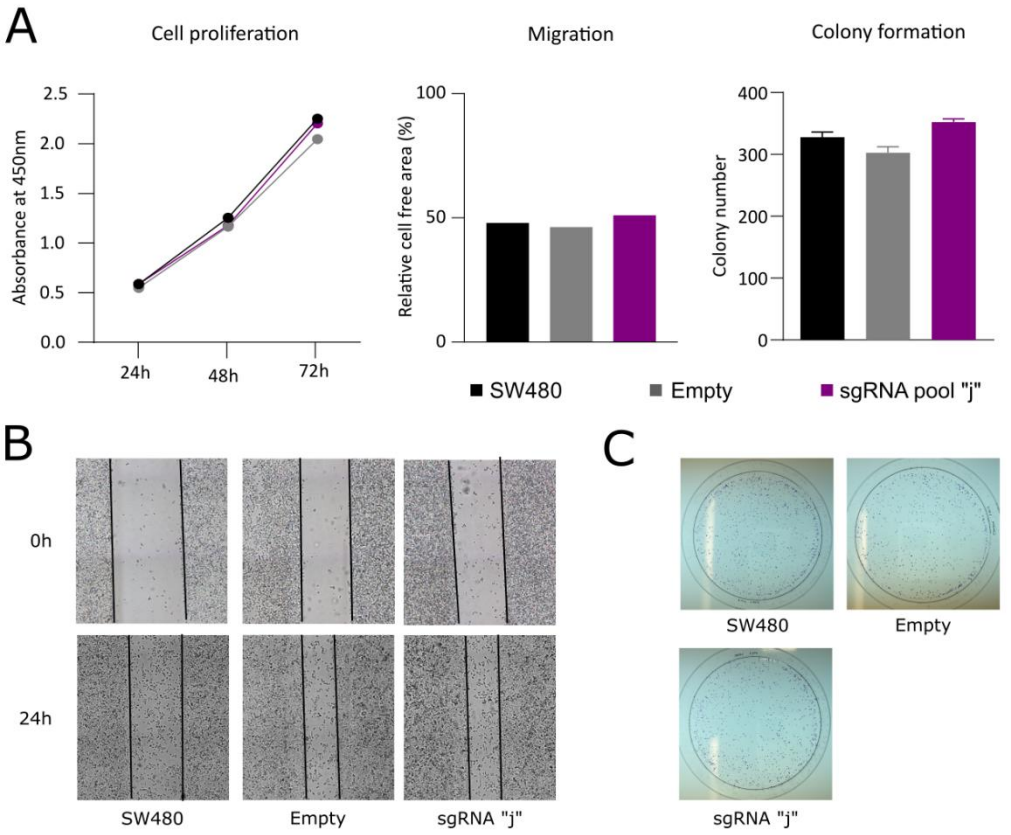


Figure 35. Overexpression of FOXD2 and FOXD2-AS1 has no effect on functional roles in SW480 cell line. A. Data representing cell proliferation, migration and colony formation assays. Each experiment was performed independently two times. **B.** Microscope images at 0h and 24h of wound healing assay to assess cell migration ability. **C.** Cell colony formation images. Negative controls: SW480, untreated SW480; Empty, SW480 infected with empty sgRNA vector; sgRNA "j", pool of cells overexpressing FOXD2 and FOXD2-AS1.

We also explored the proliferation and colony formation abilities of two clones expanded from the sgRNA "j" pool (named clones 10 and 20, **Figure 34.D**) and did not observe any differences compared to the negative controls (SW480 and Empty) (data not shown).

6.4 Ectopic overexpression of FOXD2 and FOXD2-AS1 independently

Given our problems to overexpress FOXD2 and FOXD2-AS1 independently with the CRISPR SAM system, we developed another *in vitro* model by using overexpression vectors (see

methods 5.4 and 5.5). Again, we used the SW480 cell line as it had the lowest expression levels of both FOXD2 and FOXD2-AS1 genes (**Figure 32.C**).

We overexpressed FOXD2 using a lentiviral vector containing the FOXD2 coding region. FOXD2-AS1 was overexpressed using a plasmid containing the officially reported transcript sequence according to RefSeq and GENCODE databases (2,527bp, NR_026878.1). Compared with the respective empty vectors, cells containing FOXD2 or FOXD2-AS1 exogenous plasmids produced a significantly increased expression of the respective gene as detected by qPCR (**Figure 36**). As some antisense transcripts regulate the expression of sense transcripts, we also investigated if FOXD2-AS1 could modulate FOXD2 expression. We did not observe changes in FOXD2 expression in cells overexpressing FOXD2-AS1 (and vice versa).

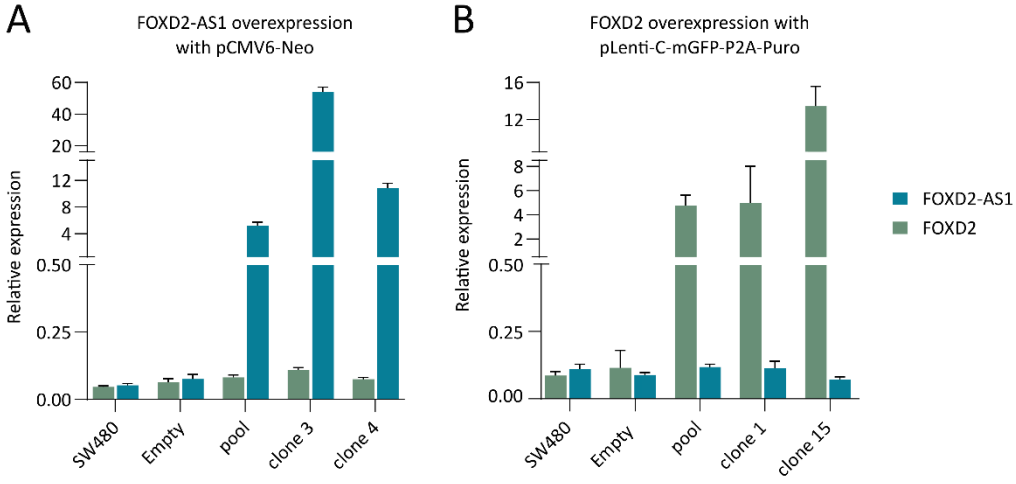


Figure 36. Overexpression of FOXD2 and FOXD2-AS1 in SW480 cell line. qPCR relative expression of SW480 cell line, cells transferred with the corresponding empty vector, pool of cells with overexpression of FOXD2-AS1 (**A**) or FOXD2 (**B**) and expanded clones from the pool of cells. Bars correspond to the average value of triplicates ± SD. qPCR analysis were normalized using MRPL9 and PMSC4 as reference genes.

We next expanded and selected two clones out of the cells overexpressing FOXD2 or FOXD2-AS1 to further investigate their functional implication. All the developed models displayed the same cell morphology as the non-treated SW480.

6.5 FOXD2 overexpression decreases cell migration and colony formation abilities, while FOXD2-AS1 promotes cell migration

We evaluated the functional effects of our genes of interest on cell proliferation, migration, and colony formation. Regarding proliferation, we did not observe any significant differences nor tendencies on treated cells overexpressing FOXD2 or FOXD2-AS1 compared to their negative controls (**Figure** 37.A and 38.A). By evaluating their migratory abilities, we observed opposite outcomes as FOXD2 expression was associated with decreased cell migration and FOXD2-AS1 promoted migration (**Figure** 37.A-B and 38.A-B). Finally, while FOXD2 ectopic expression displayed significant patterns in colon formation abilities, no associations were observed with FOXD2-AS1 expression (**Figure** 37.A-C and 38.A-C). Overall, clones, rather than the pool of cells, displayed the major changes in tumoral cells.

These results suggest FOXD2 behaves as a tumor suppressor gene while FOXD2-AS1 has no or mild functional properties as an oncogene in SW480 cell line.

6.6 FOXD2-AS1 did not confer any malignant properties in cells with high FOXD2 overexpression

Finally, we transfected two clones (1 and 15) exhibiting increased FOXD2 expression with FOXD2-AS1 plasmid aiming to perform functional assays and investigate if FOXD2-AS1 could reverse the effects promoted by FOXD2 (**Figure** 39.A). We observed no changes in cell proliferation, migration, or colony formation in clones 1 and 15, indicating that FOXD2-AS1 overexpression did not impact those clones that maintained a decreased migration capacity and a low ability to form colonies (**Figure** 39).

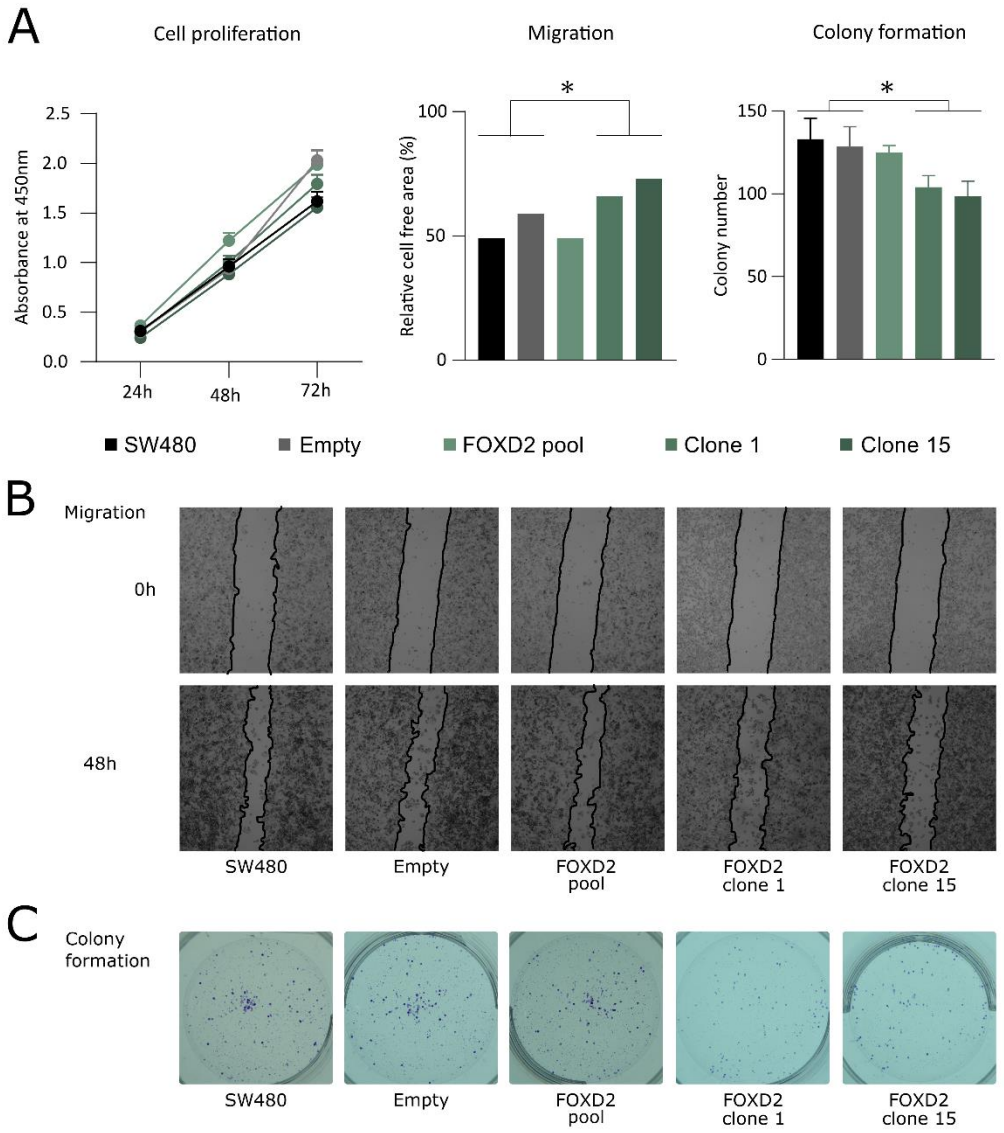


Figure 37. FOXD2 overexpression results in inhibition of migration and cell colony formation. A. Data representing cell proliferation, migration and colony formation assays. Each experiment was performed independently two times. P values correspond to t-test against negative controls. **B.** Microscope images at 0h and 48h of wound healing assay to assess cell migration ability. **C.** Cell colony formation images. Negative controls: SW480, untreated SW480; Empty, SW480 infected with empty pLenti-C-mGFP-P2A-Puro vector; FOXD2 pool, pool of cells overexpressing FOXD2 infected with pLenti-C-mGFP-P2A-Puro-FOXD2; Clones 1 and 15, expanded cells from FOXD2 pool.

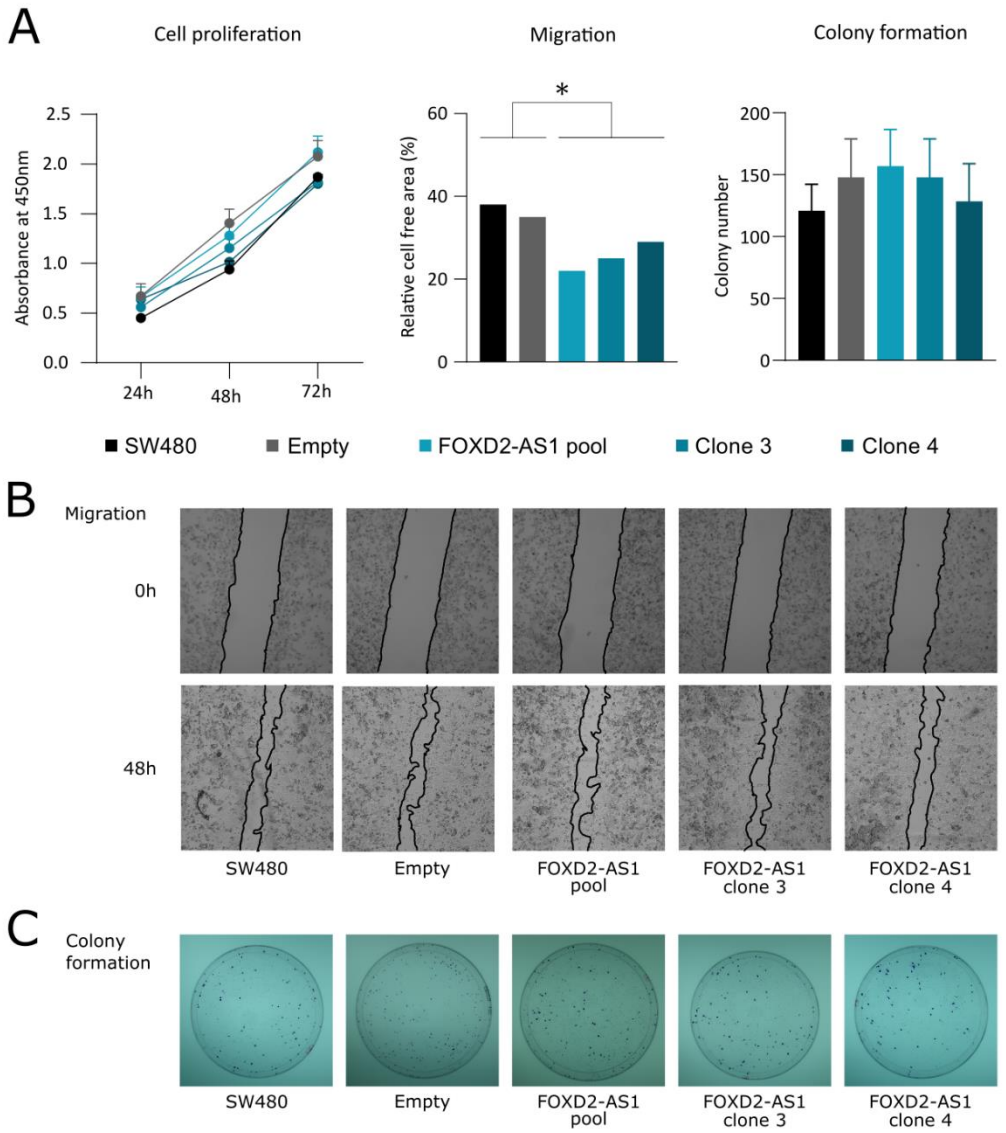


Figure 38. FOXD2-AS1 overexpression results in enhanced cell migration. A. Data representing cell proliferation, migration and colony formation assays. Each experiment was performed independently two times. P values correspond to t-test against negative controls. **B.** Microscope images at 0h and 48h of wound healing assay to assess cell migration ability. **C.** Cell colony formation images. Negative controls: SW480, untreated SW480; Empty, SW480 transfected with empty pCMV6-Neo; FOXD2-AS1 pool, pool of cells overexpressing FOXD2-AS1 transfected with pCMV6-Neo-FOXD2-AS1; Clones 3 and 4, expanded cells from FOXD2-AS1 pool.

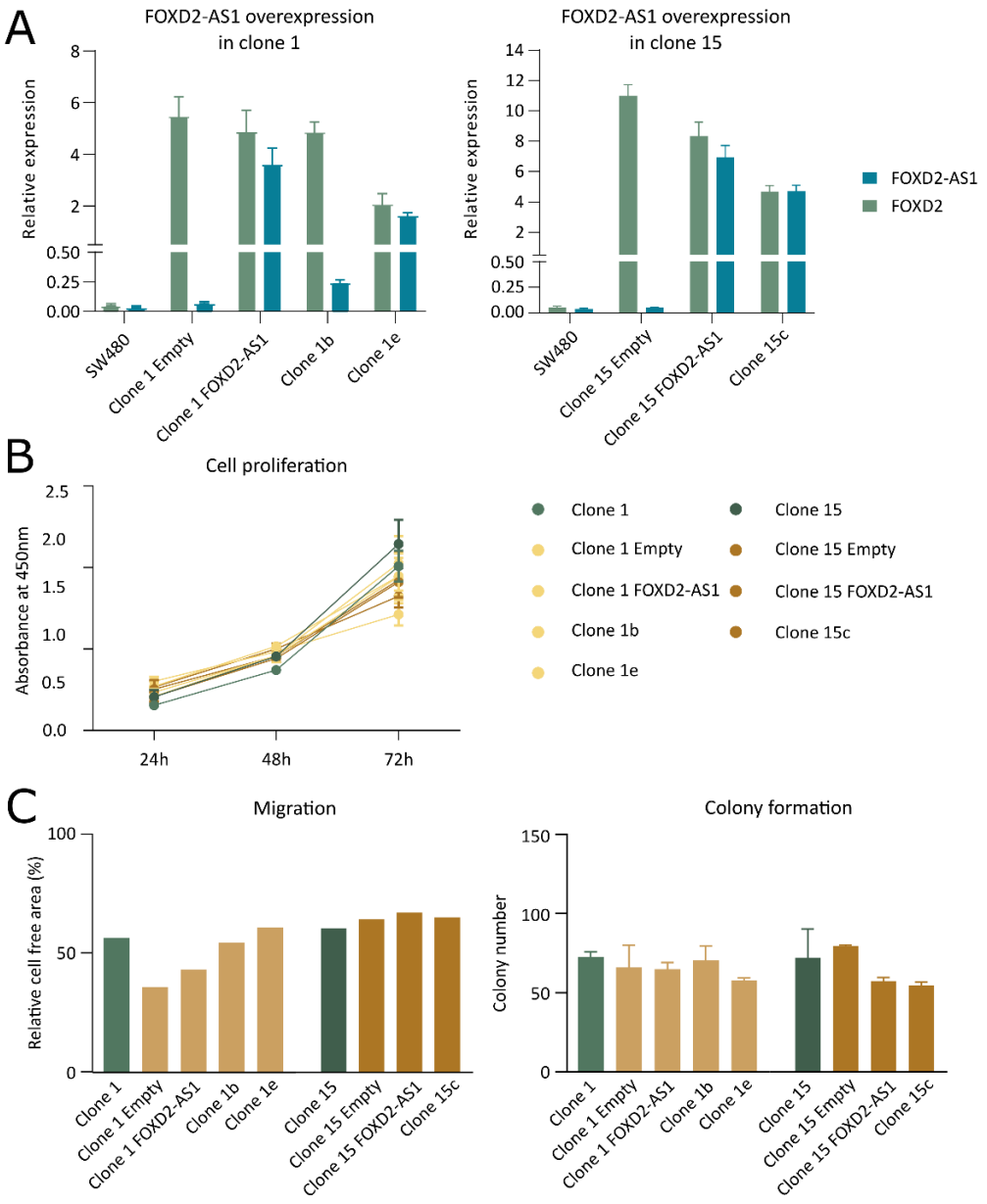


Figure 39. FOXD2-AS1 had no impact on cells already overexpressing FOXD2. A. qPCR relative expression of clones 1 and 15 overexpressing FOXD2, transfected with pCMV6-Neo to overexpress FOXD2-AS1. Bars correspond to the average value of triplicates ± SD. qPCR analysis were performed relative to MRPL9 and PMSC4. **B.** Cell proliferation, **C.** migration and colony formation analysis. Each experiment was performed independently two times. No statistically significant results were found. Negative controls: Clones 1 and 15, overexpressing FOXD2; Clone 1 Empty and 15 Empty, transfected with pCMV6-Neo empty vector; Clone 1 FOXD2-AS1 and 15 FOXD2-AS1, pool of cells transfected with pCMV6-Neo-FOXD2-AS1; Clones 1b and 1e, expanded cells from Clone 1 FOXD2-AS1; Clone 15c, expanded cells from Clone 15 FOXD2-AS1.

RESULTS Study II

Colorectal cancer is associated with the presence of cancer driver mutations in normal colon

Background

Normal tissue cells are very heterogeneous populations composed of myriads of small clones. The polyclonal nature of these tissues makes it challenging to detect somatic mutations, as the detection limit of conventional sequencing technologies is below the mutation frequency in normal cells (Risques and Kennedy 2018). With the rapid optimization of standard NGS technologies, error rates have been improved, allowing the detection of real somatic mutations in morphologically normal tissues. Indeed, the duplex-sequencing (DS) approach is able to detect low-frequency somatic mutations within normal tissue by using ultra-deep, high-accuracy sequencing. DS employs double-stranded molecular tags, which enable error correction by consensus sequence independently in each DNA strand, effectively decreasing the error rate of sequencing from 10^{-3} to $<10^{-7}$ (Schmitt et al. 2012). Because each duplex read corresponds to an original DNA molecule, this method detects single mutant DNA molecules among thousands of non-mutant genomes, thus providing the necessary resolution to identify mutant cells in normal tissue by analyzing a single biopsy.

Previously at Risques Lab (UW, Seattle, USA), they have used DS to perform ultra-deep sequencing of *TP53* in normal gynecological tissues across the human lifespan, revealing a progressive enrichment of *TP53* pathogenic mutations with older age (Salk et al. 2019). They have also demonstrated the presence of cancer driver *TP53* mutations in the peritoneal fluid (Krimmel et al. 2016), uterine lavage (Salk et al. 2019), and Pap test DNA (Krimmel-Morrison et al. 2019) of women with and without ovarian cancer. Interestingly, women with cancer tended to have higher *TP53* mutation burden (Krimmel-Morrison et al. 2019; Krimmel et al. 2016), suggesting increased *TP53* somatic evolution is associated with cancer progression.

As mentioned in the introduction (see section **1.3**), CRC is one of the most common cancers worldwide and its incidence is increasing in individuals aged <50 years old (R L Siegel et al. 2017). The easy access to normal tissue via colonoscopy makes this cancer type useful in examining the potential of clonal expansion detection in histologically normal tissue in order to identify early cancer progression. For that reason, the goal of this study was to investigate whether mutations in common CRC genes could be detected by ultra-deep sequencing (>1,000x) in histologically normal colon biopsies and to determine whether they were more frequent in individuals with CRC than in those who are cancer-free.

Results

1. CRISPR-DS enables ultra-sensitive detection of mutations

To investigate whether common cancer mutations are present at low frequency in normal colon in association with CRC development, we used ultra-accurate deep sequencing DS method combined with CRISPR digestion of the DNA (Nachmanson et al. 2018). Briefly, CRISPR-Cas9 digestion was employed to excise selected genomic regions of interest in fragments of a predetermined size. These fragments were then size-selected with SPRI beads for target enrichment before library preparation in order to enrich the target region (**Figure 14**). Then, the selected fragments were ligated with DS adapters of 8bp molecular tags composed by random nucleotides to uniquely label each DNA molecule and therefore, enable double-strand error correction leading to very low frequency mutations detection (Schmitt et al. 2012).

We designed a panel including the commonly mutated CRC genes *TP53*, *KRAS*, *PIK3CA,* and *BRAF*, and aimed for a minimum of 1000x depth. These genes were selected because, together with APC, they constitute the five most frequently mutated genes in CRC, based on the Catalogue of Somatic Mutations (COSMIC) database (Tate et al. 2019). In contrast to APC, which is a large gene, they tend to accumulate mutations in smaller hotspot regions, thus being excellent targets for developing ultra-sensitive sequencing tests for early cancer detection. The library size was 3461 bp, composed of 1953 coding bp and 1508 non coding bp from intronic regions flanking the excised exons.

1.1. Design of CRISPR-Cas9 guide RNA (gRNA)

First, we adapted the CRISPR-DS method to digest the DNA in small fragments of approximately ~260bp (between 246 and 282 bp) to maximize the read space of an Illumina MiSeq v2 300 cycle kit. To do so, we designed 24 guide RNAs (gRNAs) to specifically excise the DNA in 14 fragments containing the common oncogenic mutations of *BRAF, KRAS*, and *PIK3CA* oncogenes and the entire coding region of *TP53* (**Figure 14**, see methods 3.1). Initially, gRNAs were selected based on the highest specificity score that produced an appropriate fragment length according to CRISPOR online platform (Concordet and Haeussler 2018). In order to obtain a uniform sequencing coverage among the targeted fragments, gRNAs were experimentally tested. Unfortunately, the first set of gRNAs showed low coverage depths on some fragments, therefore 13 out of 24 gRNAs

were redesigned and tested again to ensure a proper excision and a homogeneous coverage. The best gRNA selected to perform the libraries are listed in **Table 11**.

1.2 CRISPR-DS proof of concept

Next, to demonstrate the efficiency of the assay, we used DNA from 4 common human colorectal cancer cell lines that carry driver mutations in our target genes (**Table 10**). First, 100 ng of DNA from three cell lines were processed for CRISPR-DS yielding average duplex coverage depths of ~2700x, 2500x, and 1800x for HCT116, SW480, and HT29, respectively.

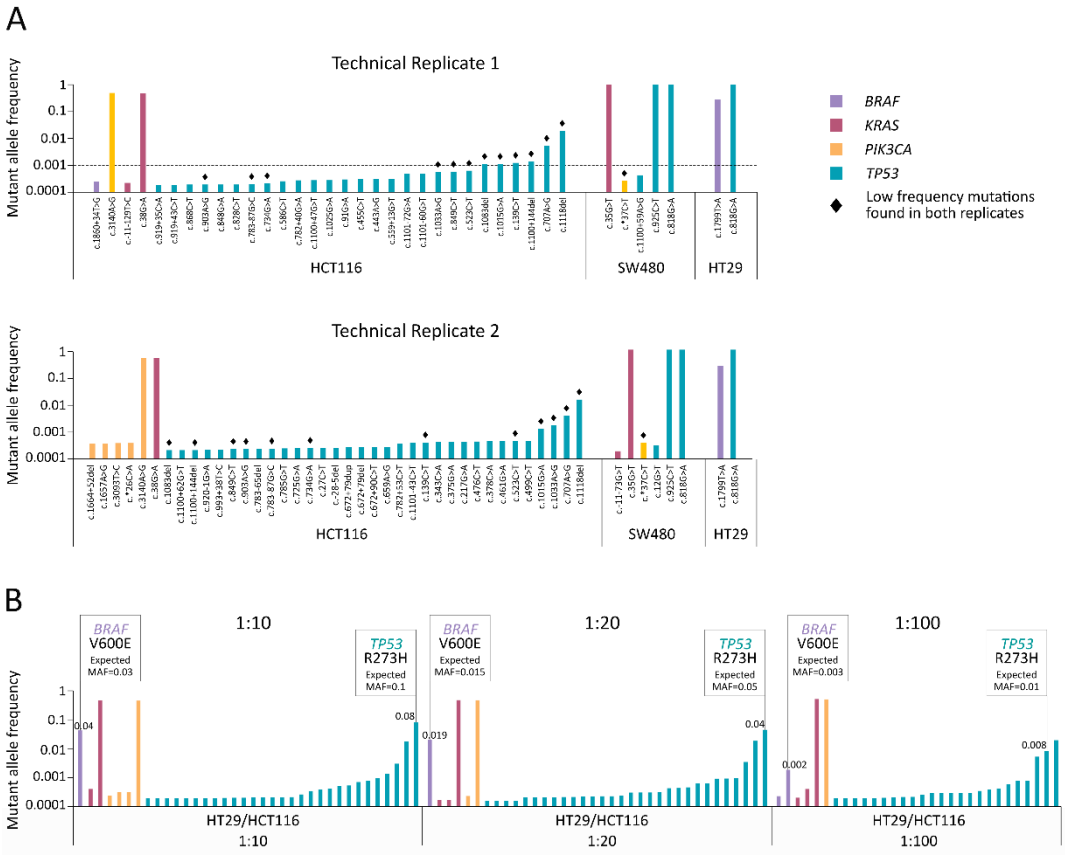


Figure 40. Ultra-deep sequencing of colorectal cancer cell lines with CRISPR-DS. A. Distribution of mutations in two technical replicates of HCT116, SW480 and HT29 cell lines. Each bar represents a mutation color coded by gene and the height of the bar indicates its Mutant Allele Frequency (MAF). MAF is calculated as the number of mutant alleles divided by the sequencing depth at a given position. Mutations are sorted by ascending MAF within each sample. The black diamonds highlight mutations found in both replicates. **B.** Spike in of HT29 in HCT116 at different concentrations: 1:10, 1:20 and 1:100. Expected HT29 mutations and their MAF for each dilution are indicated with boxes and the observed frequency is indicated above each of the corresponding bars.

Sequencing recovery was measured as the percentage of sequenced genomes equivalents compared to input genomes. Taking into account that 1 ng of DNA contains 300 genomes and each duplex read corresponds to one original DNA molecule, the efficiencies of the four cell lines were estimated as 9%, 8%, and 6%, respectively, in agreement with the previous study (Nachmanson *et al.*, 2018). In addition, 500ng of DNA were processed for SW480 and LoVo cell lines yielding average depths of 5800x and 4800x that correspond to sequencing efficiencies of 4% and 3%, respectively. These results suggest that low DNA inputs have higher enrichment rates.

To test the reproducibility, we performed an independent technical replicate experiment for HCT116, SW480, and HT29 that identified not only the expected driver mutations but also all the mutations with Mutant Allele Frequency (MAF) of as little as 0.001 as well as a subset of the very rare mutations below 0.001 despite the decreased likelihood of repeated sampling of rare events (**Figure 40.A**). These results demonstrate the high sensitivity and reproducibility of the assay even at very low MAF. To further demonstrate sensitivity in an independent experiment, we spiked DNA from HT29 into DNA from HCT116 at 3 different ratios (1:10, 1:20, 1:100). The MAF from the two driver HT29 mutations (*BRAF* V600E and *TP53* R273H) proportionally decreased in each dilution and the mutations could be identified even when present at very low frequencies (0.01 and 0.003) (**Figure 40.B**).

2. Normal colon tissue of CRC patients carries a higher frequency of coding mutations than individuals without cancer

Once we set up the technique with our panel of genes, we used CRISPR-DS to perform ultra-deep sequencing (mean depth ~2,500x) of DNA from the normal colonic epithelium of 47 individuals, 24 cancer-free and 23 with CRC. The groups were age-matched and enriched with individuals with CRC younger than 50 years of age to allow investigation of the role of somatic mutation load in with early age of onset CRC (**Figure 41.A** and **Supplementary Table S3**). We analyzed normal left colon epithelium in individuals without CRC and normal epithelium of at least 10 cm distant from the tumor (except for three patients at 3-5 cm from the tumor) in individuals with CRC.

While the mean duplex depth across samples was variable, all samples reached a minimum of 1,000x and the average depth for both groups of patients was similar (**Figure 41.B**). Overall, CRISPR-DS yielded a total of 404M duplex nucleotides, corresponding to 227M in coding regions and 177M in non-coding regions. On average, for each sample, we

sequenced 4.8M coding nucleotides and 3.7M non-coding nucleotides. A total of 168 mutations were identified: 117 coding and 51 non-coding (**Figure 41.B**) in *TP53* as well as in the targeted oncogenes in both groups. As expected, the number of mutations increased with the number of duplex nucleotides sequenced (**Figure 16**). To correct this effect, sample comparisons were made based on mutation frequencies, calculated as the number of mutations in a given region (e.g., coding) divided by the total number of duplex nucleotides sequenced in that region (see methods 4.1). Mutation counts and corresponding mutation frequencies for each sample are shown in **Supplementary Table S5**.

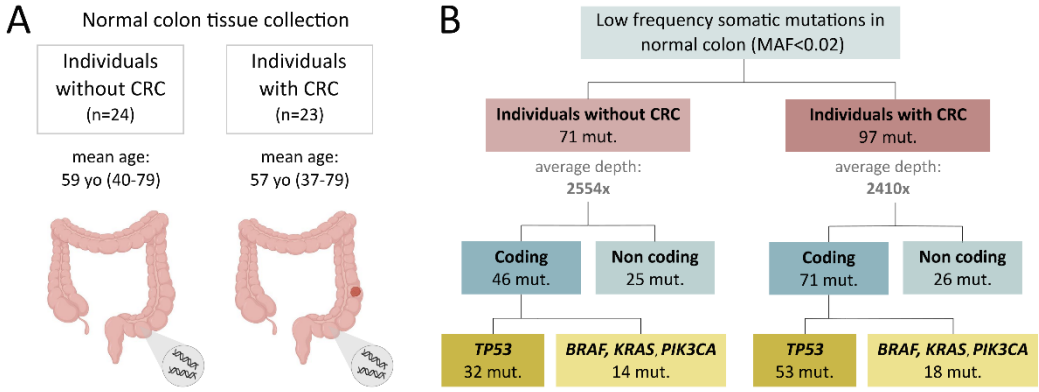


Figure 41. CRISPR-DS enables ultra-sensitive detection of cancer gene mutations in normal colon samples. **A**. Normal colon biopsies were procured from 47 individuals with and without CRC. **B.** Low frequency somatic mutations in cancer genes are identified in normal colon from patients with and without CRC.

Next, we compared coding and non-coding mutation frequencies in patients with and without CRC. Patients with CRC had a significantly higher coding mutation frequency than patients without cancer (p=0.005, t-test), while the non-coding mutation frequency was similar in both groups (**Figure 42.A**). Interestingly, the non-coding mutation frequency significantly increased with age (p=0.024, Spearman correlation), but this trend was not observed for the coding mutation frequency (**Figure 42.B**). In addition, the non-coding mutation frequency was significantly associated with higher epigenetic age of 3 out of the 4 epigenetic clocks previously measured using methylation arrays on the same samples (Wang et al. 2020). Specifically, higher non-coding mutation frequency correlated with advanced epigenetic age in the normal colon as measured by the Horvath clock, the PhenoAge clock, and the EpiTOC clock, which are well-established measurements of epigenetic aging (p=0.038, p=0.021, p=0.033, respectively, Spearman correlation) (**Figure**

42.C). Coding mutations did not associate with lower or higher epigenetic age determined by these clocks. These results suggest that while intronic, non-functional mutations accumulate with chronological and biological aging in the normal colon, coding mutations in target driver genes exceed the age-related background level, especially in patients that develop CRC.

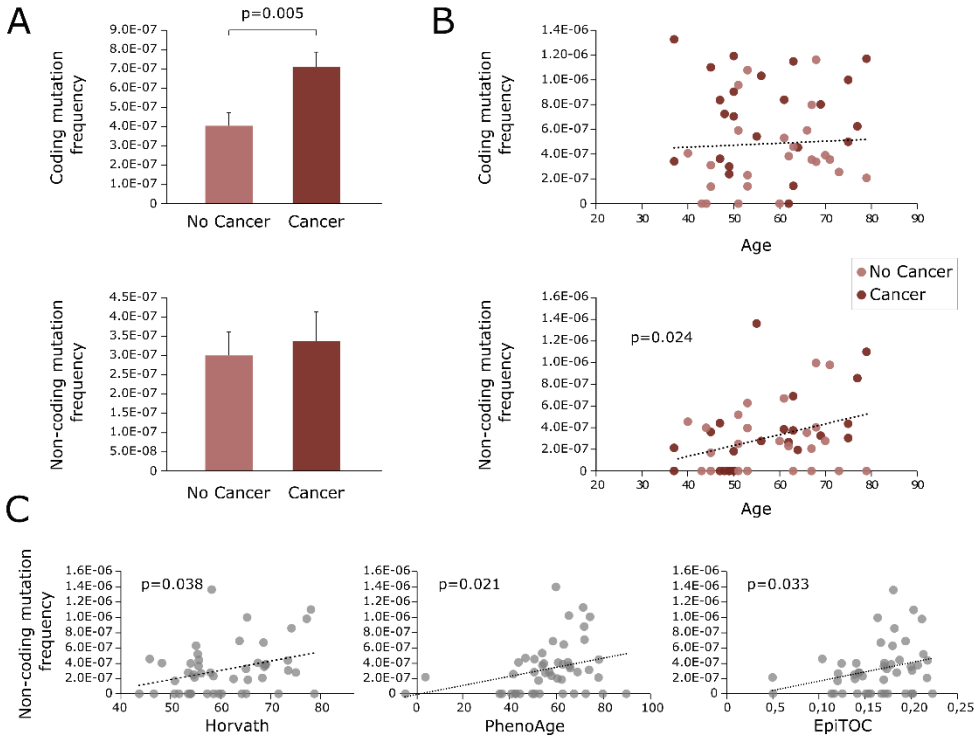


Figure 42. Normal colon of patients with CRC has higher, not age-related, coding mutation frequency. A. Coding and non-coding mutation frequency of normal colon mucosa of individuals with and without cancer. Mutation frequency is calculated as the number of mutations divided by the total number of duplex nucleotides sequenced in the coding or non-coding regions, respectively. P-value corresponds to t-test. Error bars represent standard error of the mean. **B.** Coding and non-coding mutation frequency and its correlation with age. P-value corresponds to Separman correlation. **C.** Non-coding mutation frequency correlation with Horvath, PhenoAge and EpiTOC epigenetic clocks. P-values correspond to Spearman correlation. Only significant p-values are displayed.

To further explore the nature of the coding mutations present in the normal colon of patients with and without cancer, we classified them by mutational spectrum (C>A, C>G, C>T, T>A, T>C, T>G) and compared the mutation frequency of each nucleotide change in patients with and without CRC. We observed that while all types of mutations were more

frequent in patients with CRC, two types were significantly overrepresented: C>A (p=0.007, t-test) and T>A (p=0.015, t-test) (**Figure 43.A**). Interestingly, these two types of mutations are also enriched in CRC according to COSMIC database (**Figure 43.B**), indicating that the mutational landscape of the normal colon of patients with CRC is more similar to the one observed in cancer, compared to the normal colon profile of patients without CRC. While C>T transitions are often caused by deamination of methylated cytosines, a common age-related mutation, C>A transversions are typically due to oxidative damage (Delaney et al. 2012) and T>A transversions are linked to environmental mutagens (Kucab et al. 2019). The relative increase in proportion of C>A and T>A mutations compared to C>T in the normal colon of patients with cancer suggests an increased mutational processes producing higher rates of coding mutations similar to the ones observed in CRC.

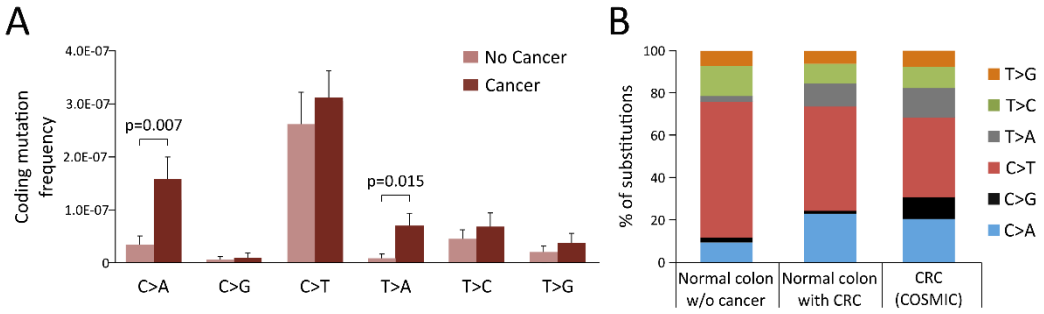


Figure 43. Mutation spectrum of patients with cancer resembles cancer mutation databases. A. Frequency of mutation by substitution type compared between normal colon of individuals with and without cancer. P-values correspond to t-tests. Only significant p-values are displayed. Error bars represent standard error of the mean. **B.** Mutation spectrum compared between coding mutations from normal colon of individuals without CRC (n=42), with CRC (n=65), and the CRC COSMIC database (n=70,525).

3. *KRAS* and *TP53* driver mutations are abundant in the colon of patients with CRC

To gain further insights into the mutational differences in the normal colon of patients with and without CRC, we plotted the mutant allele frequency (MAF) of each coding mutation identified (**Figure 44**). All patients except five (four without CRC and one with CRC) carried at least one coding mutation in their normal colon mucosa. All mutations had a very low MAF (100% <0.02 and 82% <0.001) which makes them unidentifiable by standard sequencing methods. Therefore, CRISPR-DS enabled a high resolution view of the landscape of common cancer gene mutations in the normal colon of individuals with (**Figure 44.A**) and without CRC (**Figure 44.B**).

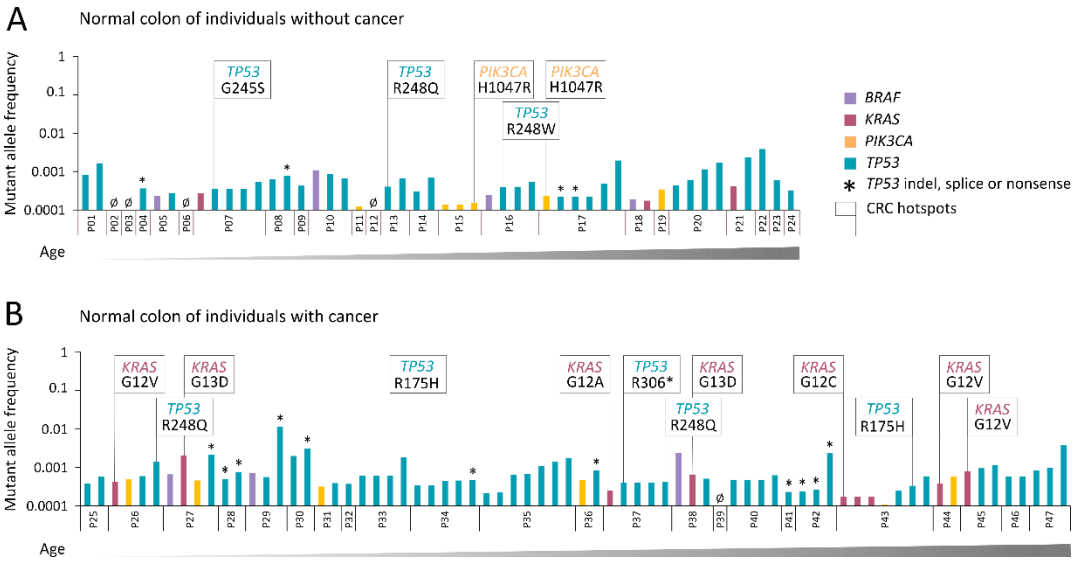


Figure 44. Landscape of coding mutations in normal colon of individuals with and without cancer. Distribution of mutations in normal colon from patients without CRC **(A)** and patients with CRC **(B).** Patient IDs are indicated in the x-axis and patients are sorted by ascending age. Each bar represents a mutation color coded by gene. The height of the bars indicate Mutant Allele Frequency (MAF), calculated as the number of mutant alleles divided by the sequencing depth at a given position. Mutations are sorted by ascending MAF within each patient. Hotspot codons and *TP53* indels, nonsense and splice mutations are highlighted.

We then explored the distribution of coding mutations by gene. While coding mutations in *BRAF*, *KRAS*, and *PIK3CA* were found in the normal colon of patients with and without CRC, cancer hotspot *KRAS* mutations were exclusively found in individuals with CRC (**Figure 44.B**). None of the mutations in *BRAF* (0/7) and only two of the mutations in *PIK3CA* (2/12) corresponded to the canonical cancer hotspots in these oncogenes (**Figure 45.A-B**) and their frequency was not statistically different between patients with and without CRC. In contrast, 7 out of the 13 *KRAS* mutations found in normal tissue corresponded to the canonical hotspots in codons 12 or 13 and all of them were present in the normal colon of individuals with CRC (**Figure 45.C**). Overall, 30% of patients with CRC carried a hotspot *KRAS* mutation in normal colon compared to none of the patients without CRC (p=0.003, Pearson Chi-Square) (**Figure 45.D**). All the mutations found in the oncogenes *BRAF*, *KRAS* and *PIK3CA* are listed in **Supplementary Table S6**.

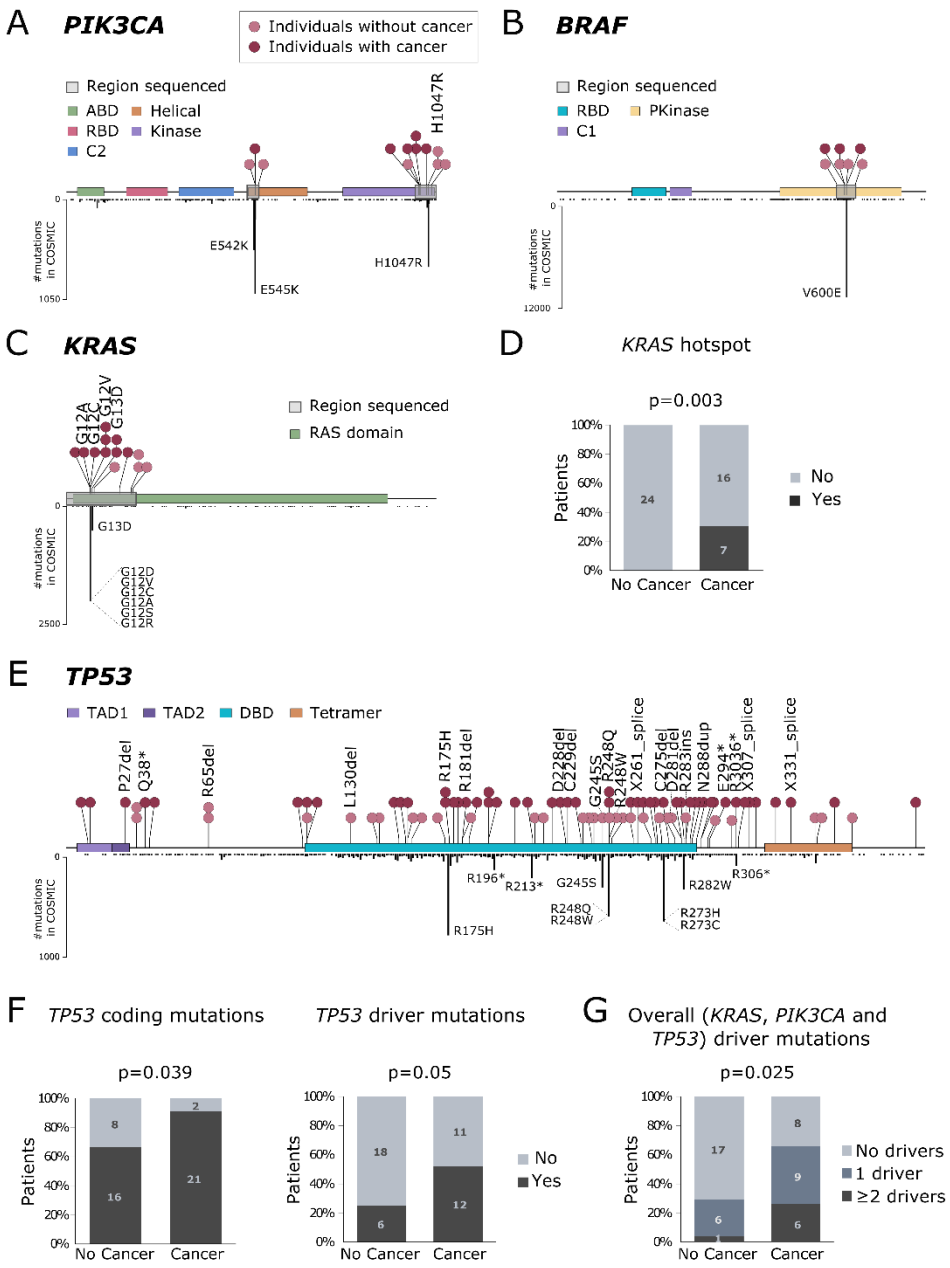


Figure 45. Normal colon carries mutations in common CRC genes, but these mutations are more abundant and pathogenic in patients with CRC. A-C. Distribution of mutations in *PIK3CA, BRAF* and *KRAS* in normal colon (above gene diagram) and in CRC samples from COSMIC database (below gene diagram)**.** Normal colon mutations are color-coded by individuals with or without CRC and mutations corresponding to cancer hotspots are indicated. **D.** Percentage of patients with and without CRC that carry *KRAS* hotspot mutations in normal colon. **E.** Distribution of mutations across *TP53* in normal colon (above gene diagram) and in CRC samples from COSMIC database (below gene diagram)**. F.** Percentage of patients with and without CRC that carry *TP53*

coding mutations and driver mutations in their normal colon. **G**. Percentage of patients with and without CRC that carry one or more different cancer driver mutations in PIK3CA, KRAS, or TP53 their normal colon. P-values correspond to Person Chi-Square. *ABD: adapter-binding domain; RBD: Ras-binding domain; Pkinase: protein tyrosine kinase domain; TAD: transactivation domain; DBD: DNA-binding domain; Tetramer: tetramerization domain.*

Given the tumor suppressor role of *TP53*, we deep sequenced all its coding exons. In total we identified 85 *TP53* coding mutations, which clustered around the DNA-binding domain similar to what is observed for COSMIC CRC *TP53* mutations (**Figure 45.E**) (**Supplementary Table S7**). These results suggest that *TP53* mutations identified in normal colon follow similar patterns than the ones that take place in CRC. In addition, 9.4% of the substitutions identified corresponded to the top ten most common *TP53* substitutions in large intestinal cancers reported in COSMIC (see methods **4.2**); and 18.8% of *TP53* mutations were high impact mutations (indels, nonsense or splice), which severely affect protein function. In total, more than a quarter of *TP53* mutations identified in normal colon (27.1%) were either hotspots or high impact mutations, which are likely to confer a selective advantage to the cells carrying them, and therefore were considered driver mutations. Patients with CRC more frequently carried *TP53* coding mutations in normal colon than patients without cancer (p=0.039, Pearson Chi-Square) and also more likely to carry *TP53* driver mutations (p=0.05, Person Chi-Square) (**Figure 45.F**). Additionally, we only identified one patient without cancer that carried ≥2 TP53 driver mutations, while 6 patients with CRC carried multiple TP53 driver mutations (p=0.025, Pearson Chi-Square) (**Figure 45.G**), suggesting the concurrence of driver mutations in these patients.

Overall, the higher prevalence of *KRAS* hotspot, *TP53* coding, and *TP53* driver mutations in the normal colon of individuals with CRC hints the presence of more extensive or advanced precancerous fields of somatic evolution.

4. The normal colon of patients with CRC displays a mutation profile different from the cancers of the same patients

We then investigated if the mutations observed in the normal colon of individuals with CRC coincided with those detected in the paired tumor, which might be indicative of a common clonal origin. We duplex sequenced the same 4 gene regions in tumor DNA from 19 patients with available cancer tissue and catalogued all non-synonymous mutations with VAF>0.1 (**Supplementary Table S8**). In 4 out of the 19 sequenced tumors, we did not identify mutations in any of the genes included in our panel, likely because they were driven by other non-sequenced cancer genes. From the remaining 15 tumors carrying

mutations in *TP53*, *KRAS* or *PIK3CA*, 7 of them had at least one non-synonymous mutation that was identified in the tumor as well as the corresponding normal tissue of the same patient (**Figure 46**) (**Supplementary Table S8**). However, in all cases, the normal tissue also carried additional cancer gene mutations that were not detected in the tumor. Also, in the remaining 8 cases, the tumor mutation in *TP53*, *KRAS,* or *PIK3CA* could not be identified in the normal tissue, which instead carried other cancer mutations. Overall, out of 44 non-synonymous mutations identified in the normal colon of these 15 individuals with CRC, only 9 (20.5%) coincided with the synchronous tumor mutation indicating that most clonal expansions observed in normal colon are not related to the expansion that eventually progressed to CRC. These results suggest that multiple independent clonal expansions might be common in normal colonic mucosa of individuals at risk of CRC, one of which can eventually give rise to a tumor.

While mutations in individuals without CRC were less abundant, in these patients we identified a higher frequency of *TP53* coding mutations in males (p=0.056, t-test) and patients carrying polyps (p=0.054, t-test) (**Figure 47**), both well known risk factors for CRC (Click et al. 2018; Rebecca L. Siegel et al. 2020). Smoking and high BMI are also CRC risk factors but were not associated with *TP53* mutation in patients without CRC. These results suggest a potential link between *TP53* clonal expansions in normal colon and some CRC risk factors.

5. Clones with cancer driver mutations are larger in patients with early CRC

Next, we wondered whether the clonal expansions identified in the normal colon of individuals with CRC were not only more abundant but also larger than in the normal colon of individuals without CRC. As previously mentioned, when using DS technology, each duplex read corresponds to an original DNA molecule. Therefore, the number of duplex reads that contain a given mutation is equivalent to the number of haploid genomes with that mutation, and thus, proportional to the size of the clone carrying the mutation. We observed that patients with CRC not only had more driver mutations, but these driver mutations were detected in multiple reads indicating their presence in larger clones (**Figure 46**). Large clones were identified in normal colon of individuals with cancer and these clones were especially enriched in individuals that developed CRC younger than age 50.

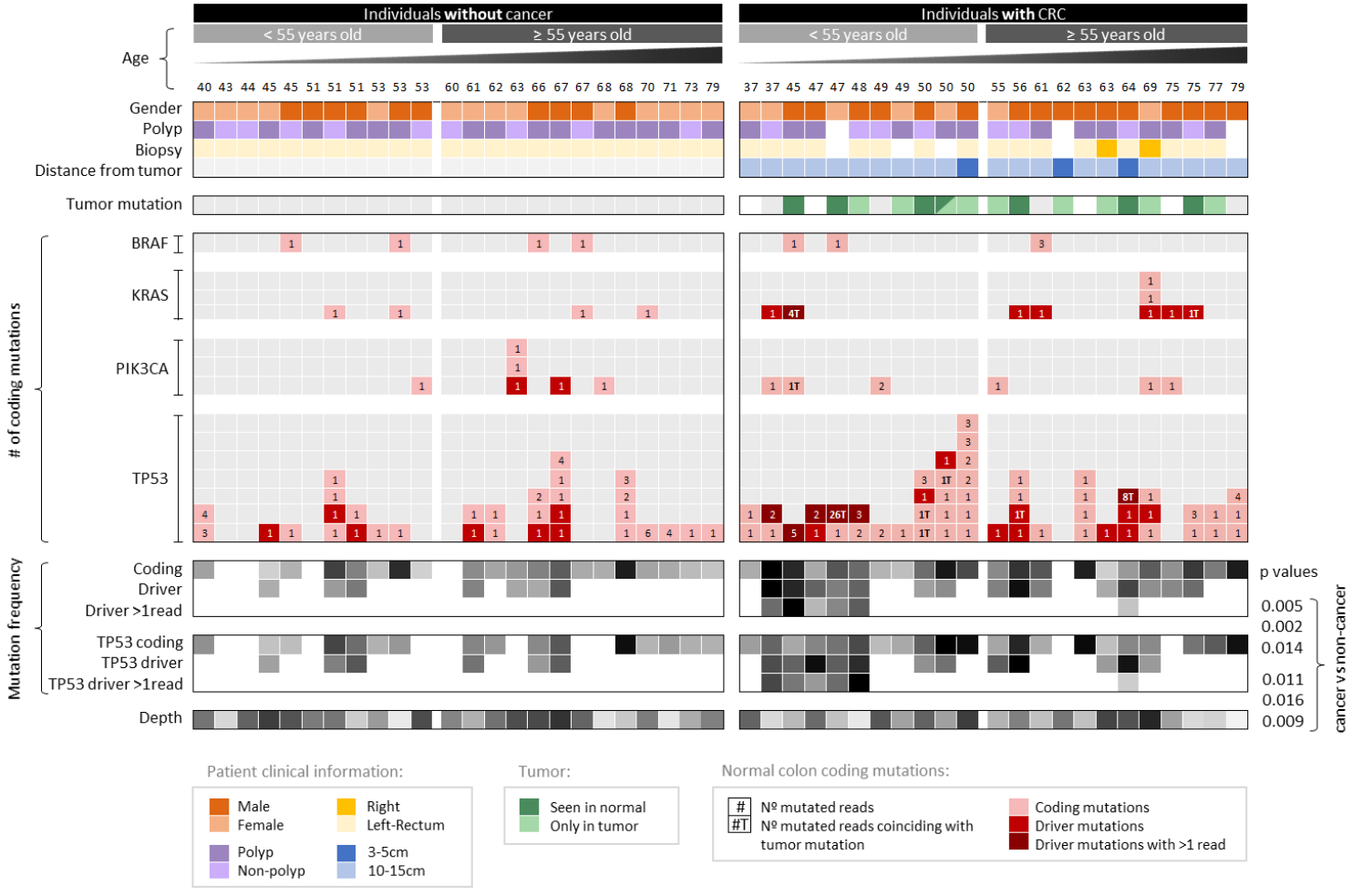

Individuals without cancer
< 55 years old
≥ 55 years old
Individuals with CRC
< 55 years old
≥ 55 years old
Age
40 43 44 45 45 51 51 51 53 53 53 60 61 62 63 66 67 67 68 68 70 71 73 79
37 37 45 47 47 48 49 49 50 50 50 55 56 61 62 63 63 64 69 75 75 77 79
Gender
Polyp
Biopsy
Distance from tumor
Tumor mutation
# of coding mutations
BRAF
KRAS
PIK3CA
TP53
Mutation frequency
Coding
Driver
Driver >1read
TP53 coding
TP53 driver
TP53 driver >1read
Depth
p values
0.005
0.002
0.014
0.011
0.016
0.009
cancer vs non-cancer
Patient clinical information:
Male
Female
Right
Left-Rectum
Polyp
Non-polyp
3-5cm
10-15cm
Tumor:
Seen in normal
Only in tumor
Normal colon coding mutations:
Nº mutated reads
#T Nº mutated reads coinciding with tumor mutation
Coding mutations
Driver mutations
Driver mutations with >1 read

Figure 46. Mutations in normal colon of patients with CRC are often different from mutations in synchronous tumors and, in early onset CRC patients, frequently include cancer driver mutations forming large clones. Each column corresponds to a patient. Patients are grouped by cancer status and sorted by age. Panels of data indicate clinical information, tumor information, normal colon mutation counts for each gene, normal colon mutation frequency, and depth. Clinical and tumor information is indicated with white squares if not available, and grey squares if negative. Tumor mutation was negative for four cases that did not show any mutation in the 4 tested genes. Mutations in normal colon identified in each gene are indicated with squares that contain the number of mutated reads color coded for mutations that are coding, drivers, and drivers with more than one (>1) mutated duplex read. 'T' next to the number indicates that the mutation was observed in the synchronous tumor (dark green cases). Driver mutations were conservatively defined as oncogenic hotspots and *TP53* hotspot, nonsense, splice and indel mutations. Bottom grey-scale heatmaps show mutation frequency values based on mutations that are coding, driver and driver with >1 duplex read for all genes and *TP53* only. P-values correspond to t-test comparison of the mean frequency between individuals with and without CRC. Depth indicates average duplex depth for all coding positions sequenced.

We compared the mutation frequency separately for coding mutations, driver mutations and large driver mutations (with >1 duplex read carrying the mutation) for the four genes included in the panel as well as only considering the *TP53* gene. We observed that all frequencies were significantly higher in the normal colon of individuals with CRC compared to those without cancer (**Figure 46**). However, the differences in mutation frequency between the two groups of patients (cancer and non cancer) were accentuated for younger (<55 yo) individuals (**Supplementary Figure S4**). In addition, younger individuals with CRC carried a higher level of large mutant clones compared to those with CRC but older (≥55 yo) suggesting different mechanisms of clonal expansion underlying CRC progression in young and old individuals.

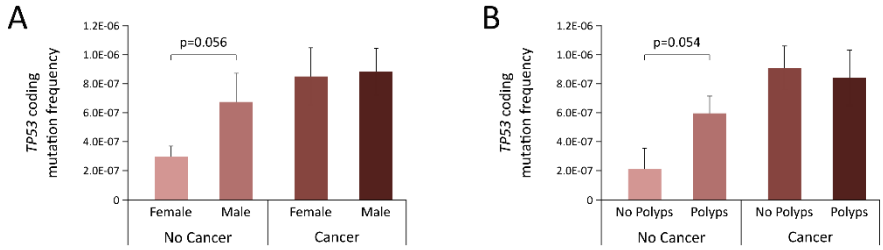


Figure 47. *TP53* coding mutations are more frequent in normal colon from individuals without CRC that are males or harbor polyps. Comparison of *TP53* mutation frequency in the normal colon of (**A**) females and males with and without CRC and (**B**) polyp and non-polyp formers with and without CRC. P-values correspond to t-test comparisons. Error bars represent standard error of the mean.

6. *TP53* mutations in normal colon are more pathogenic in individuals with CRC and resemble mutations reported in CRC

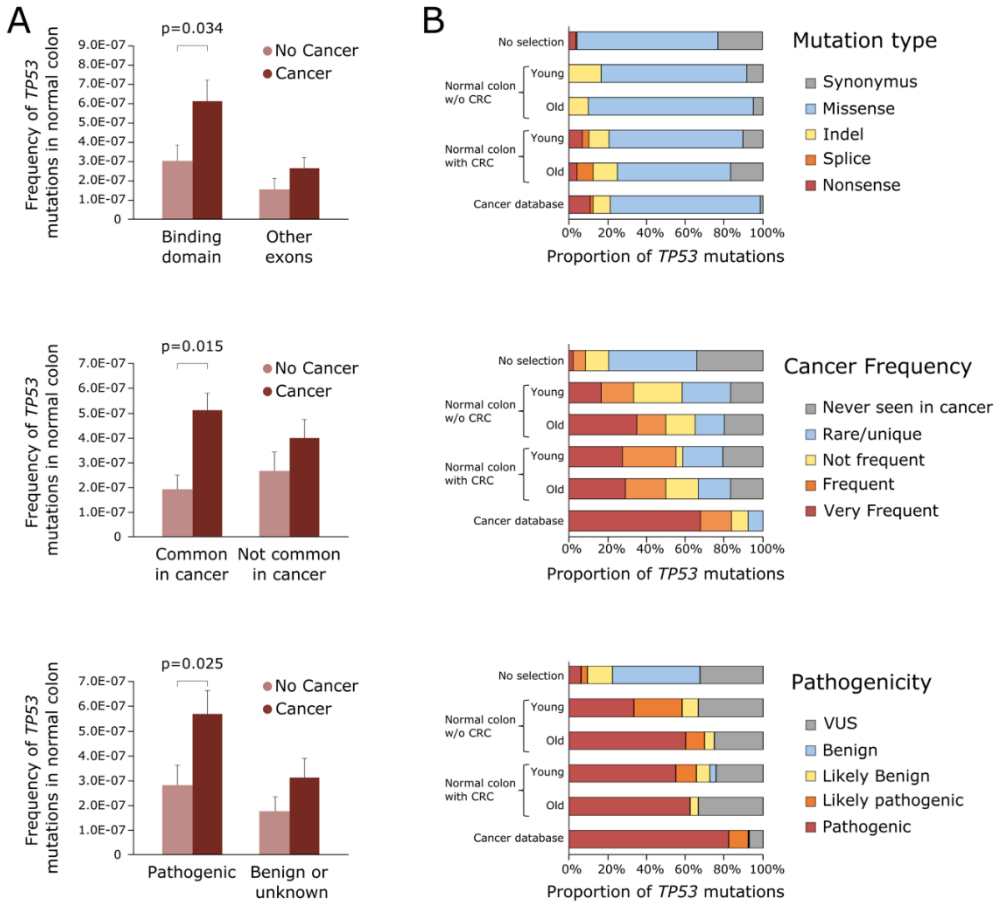


Figure 48. *TP53* mutations identified in normal colon are more pathogenic and more closely resemble *TP53* mutations identified in CRC in individuals with CRC than those cancer-free. A. *TP53* mutation frequency of individuals with and without CRC was compared based on mutations localized in the binding domain, mutations common in CRC, and mutations predicted to be pathogenic. Data was extracted from Seshat (Tikkanen et al. 2018). Only significant p-values of t-tests are displayed. Error bars represent standard error of the mean. **B.** Distribution of *TP53* mutations by mutation type, cancer frequency, and pathogenicity in normal colon of young (<55 years old) and old (≥55 years old) individuals without and with CRC compared to all possible *TP53* mutations in the coding region (no selection, n= 3,546) and *TP53* mutations reported in CRC in the UMD cancer database (n=17,681). Number of *TP53* mutations in each group: young without CRC n=12; old without CRC n=20; young with CRC n=29; old with CRC n=24.

Focusing only on *TP53* tumor suppressor mutations, we analyzed its potential functional impact using the Seshat tool (Tikkanen et al. 2018). This web service provides functional data for specific *TP53* variants including the frequency of mutations in the UMD cancer database and the predicted pathogenicity. Out of all the variables given by Seshat, we classified them in three different functional categories: (1) "common in cancer" vs "not common in cancer"; (2) " pathogenic" vs "not pathogenic"; and (3) "located in DNA binding domain" vs "other codons" (see methods 4.3). Even though patients with CRC had higher levels of mutation frequency for all categories analyzed, they only displayed a significant higher frequency of *TP53* mutations that are common in cancer (p=0.015), that are predicted to be pathogenic (p= 0.025) and that are located in the DNA binding domain of the protein (p=0.034, t-test), compared to individuals without cancer (**Figure 48.A**).

In addition, we compared the type, frequency, and pathogenicity of *TP53* mutations observed in normal colon with mutations from colon carcinomas reported in the UMD database (2021, n=17,681), as well as with all the possible substitutions in the coding region of the gene in the theoretical absence of selection (n=3,546) (**Figure 48.B**). Normal colon mutations from individuals with and without CRC were predominantly missense, similar to mutations reported in CRC or random mutations in *TP53*. However, only the normal colon of patients with CRC carried nonsense and splicing mutations, which are considered highly damaging, in similar proportions to what it is observed in the cancer database. The distribution of pathogeneic and common cancer mutations reported in normal colon clearly differed from the expected pattern of random mutations in *TP53*, strongly resembling the pattern observed in the cancer database, especially in older individuals and those with CRC (**Figure 48.B**). These results suggest that there is a common process of positive selection of *TP53* mutant clones that is operative in normal colon as well as in CRC. However, this process appears to be enhanced with aging and in those patients that develop CRC.

7. Integrative mutational analysis proof-of-principle for the development of a CRC predictor

The ultimate goal of our research is to determine whether samples from histologically normal colonic mucosa could be informative of CRC risk. An essential step is to demonstrate that individuals with cancer can be identified based on the mutation profile of normal colonic biopsies, by constructing a predicting model that summarizes the

mutational analysis. We used regularized logistic regression with Lasso penalty estimated to determine the 5 variables that were the best predictors. All quantitative variables with prior demonstrated significance in univariate analyses were included in the model as well as their interaction with age, to determine potential differential effects between young and old individuals. As shown in **Table 18**, the variables with the largest effects were the frequency of driver mutations (OR=2.16) and the presence of hotspots in *KRAS* (OR=1.86). Additional information was gained when considering the frequency of *TP53* coding mutations, *TP53* mutations common in cancer, and the interaction between frequency of more than 1 mutant read and age (ORs of 1.26, 1.066 and 1.26, respectively). This later interaction indicates that the risk of CRC increases with increased frequency of larger clones (represented by mutations with more than 1 duplex read) but only in younger individuals. The predicted capability of the model was good, with AUC = 0.6866, 95% CI: 0.5277-0.8455 after 5 fold cross-validation. While this preliminary analysis included a small number of cases and requires validation in larger studies, it demonstrates the potential of this approach for the development of a CRC predictor based on the mutational analysis of biopsies collected from histologically normal mucosa.

Table 18. Logistic regression model for CRC prediction based on normal colon mutations. Abbreviations: OR, odds ratio; AUC, area under the curve; CI: confidence interval.

Variable	beta	OR
Mutation frequency of drivers	0.768	2.16
Hotspots in *KRAS*	0.621	1.86
Coding mutation frequency in *TP53*	0.229	1.26
TP53 mutations common in CRC	0.0638	1.066
Age * drivers>1 read	0.231	1.26

AUC	95% CI
0.8397	0.7239-0.9555

5 fold cross-validated AUC	95% CI
0.6866	0.5277-0.8455

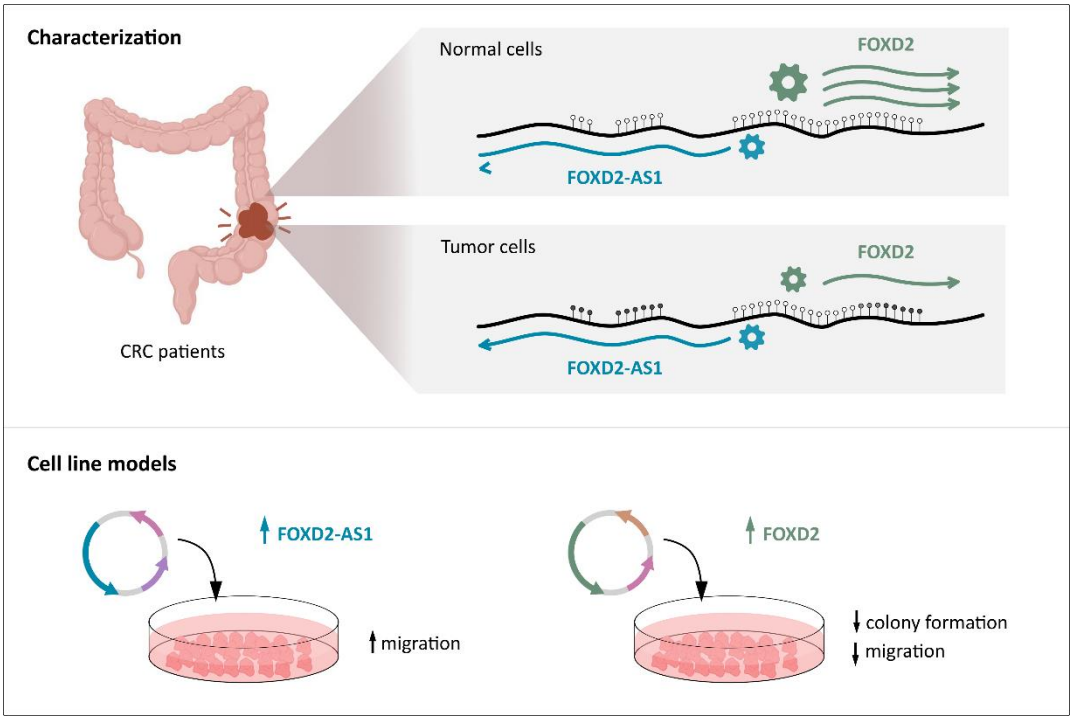
DISCUSSION

Despite the great advances made towards improving CRC patient clinical outcomes, it remains a major health burden with large numbers of new cases worldwide and high disease-specific mortality. Most CRCs are sporadic and emerge through the gradual accumulation of genetic and epigenetic alterations, able to drive malignant transformation after the bypass of tumor suppression mechanisms. As such transformation process may take over a decade, CRC is the perfect scenario for early cancer detection and prevention. The reliable identification of CRC biomarkers is a permanent challenge of growing interest for improving CRC management and reducing morbidity. Thanks to the emergence of new powerful technologies and the advances in the knowledge of the mechanistic bases of the disease, new markers are becoming promising candidates for early detection, risk stratification, prognosis and prediction of response to specific therapies. Overall, translating these advances into clinical practice should increase patient survival rates.

Motivated by these premises and previous work in our laboratory, my research during this thesis has been structured in two studies addressing both mechanistic and clinical aspects of colorectal tumorigenesis. In the first study I have investigated the deregulation and function of a transcription factor, FOXD2, and its antisense non-conding transcript, FOXD2-AS1, in CRC patients and cell lines. The second study aims to bring light into the contribution of precancerous mutations to CRC tumorigenesis by applying a high sensitive technique do detect mutations in healthy colorectal tissue.

Discussion Study I

FOXD2 and FOXD2-AS1 in colorectal cancer



Graphical abstract study I. Model summarizing FOXD2 and FOXD2-AS1 expression in CRC. In normal cells, FOXD2 and FOXD2-AS1 display a conventional profile for bidirectional protein-coding gene/lncRNA with full unmethylation of CpGi promoter and correlated expression profiles, although with a high FOXD2/FOXD2-AS1 ratio. In cancer cells, the promoter remains unmethylated, while hypermethylation of flanking CpGi strongly reduces the expression of FOXD2. According to functional experiments in CRC cell lines, FOXD2 may act as a tumor suppressor gene, decreasing migration and colony formation abilities, while FOXD2-AS1 increases cell migration.

1. FOXD2 and its natural antisense transcript FOXD2-AS1 are regulated by a bidirectional promoter

Massive sequencing technologies have led to a better understanding of the human genome organization and structure. Bidirectional initiation of transcription by RNA polymerase II has been a frequent observation, being over 10% of genes in the human genome arranged in a divergent or head-to-head configuration and regulated by a shared, bidirectional promoter (Trinklein et al. 2004). These promoters are reported to be short

inter-transcript regions of less than 1kb long, the majority being around 300bp or less, and often showing specific TF binding motifs and lack of a TATA box (Trinklein et al. 2004; M. Q. Yang and Elnitski 2008). Furthermore, most bidirectional promoters are characterized by the presence of a CpGi, but more interestingly, their GC content is increased on average (66% GC content) compared to regular promoters (M. Q. Yang and Elnitski 2008). Although not demonstrated, bidirectional promoters have been suggested to be more resistant to *de novo* methylation in cancer. Promoter DNA methylation is a known potent gene silencing mechanism observed in cancers, usually affecting tumor suppressors or DNA mismatch repair genes that, by getting silenced, contribute to tumorigenesis (Herman et al. 1995; Hibi et al. 2009; Liang et al. 2017; Maeda et al. 2003; Suzuki et al. 2004). Considering that bidirectional promoters co-regulate pairs of genes, their hypermethylation can potentially involve the simultaneous silencing of two genes, thus potentially having a higher impact on tumor development (S. Ahmad et al. 2021; Shu et al. 2006). A well-known example is the silencing of *MLH1* gene through methylation of its bidirectional promoter responsible for cases of MMR deficiency in CRC (Lin et al. 2007).

Our findings point out the co-regulation of FOXD2 and FOXD2-AS1 by a bidirectional promoter between their two 5' flanks. While several projects have helped to increase both the number and size of available gene annotations (Frankish et al. 2019; Hon et al. 2017; O'Leary et al. 2016), FANTOM resources based on Cap Analysis of Gene Expression (CAGE) provide annotations with highly accurate 5' positions (Hon et al. 2017). Our results support close proximity between FOXD2 and FOXD2-AS1 TSS in line with FANTOM 5' annotations rather than other genomic consortiums (**Figure 18**). Accordingly, their shared promoter of 350bp long is located at a CpGi, displaying 76% of CG content, surrounded by other CpGi, and lacking a TATA box, which are common characteristics of bidirectional promoters (S. S. Ahmad et al. 2021; Core, Waterfall, and Lis 2008; Trinklein et al. 2004). More evidence is reflected on the analysis of public available RNA-seq data from several colorectal tissues and cell lines showing opposite direction of transcribed regions in proximal 5' disposition. Chromatin Chip-seq data also revealed occupancy of H3K4me3 and H3K27ac marks, often enriched at bidirectional promoters compared to unidirectional promoters (Bornelöv, Komorowski, and Wadelius 2015), facilitating active co-expression of the associated genes (Chen et al. 2014). Given the positioning of bidirectional promoters, the expression of the gene pair is reported to be more positively correlated compared to other pairs of genes among the genome (Trinklein et al. 2004). In line with this scenario, FOXD2 and FOXD2-AS1 displayed consistent positive co-expression in all colorectal tissues analyzed, including normal and tumor samples, and in CRC cell lines and other normal and cancer tissue types analyzed from GTEX and TCGA databases, respectively. Indeed, by targeting their shared

promoter to induce expression using the CRISPR SAM system, we observed an exceptional pattern of coordinated upregulation of both genes in every single clone derived from the experiment.

Gathering all this data, we can affirm that FOXD2 and FOXD2-AS1 are regulated by a single promoter with bidirectional activity that drives strong positive co-expression among tissues. However, the mechanisms underlying the co-expression of bidirectional genes are still unknown. This is especially true in the case of FOXD2 and FOXD2-AS1. It remains to be resolved if the regulation is exclusively at the transcriptional level, if their expression can be altered in an independent manner and if their functional roles are related. All these questions will be further discussed in the appropriate following sections.

2. FOXD2 and FOXD2-AS1 expression dynamics in CRC

RNA-seq data from the TCGA-COAD cohort and RT-qPCR analysis from tissues collected at HUB allowed the comparison of expression profiles between paired normal-tumor colorectal samples (**Figure 22** and **23**). FOXD2 gene was statistically downregulated in tumors compared to their normal counterparts, as previously described by another study (Conesa-Zamora et al. 2015). However, FOXD2-AS1 displayed distinct expression patterns in our cohorts of study, showing no changes of expression (TCGA-COAD) or a slightly decreased expression (HUB) in tumors compared to normal. Moreover, these results differ from already published studies reporting an upregulation of FOXD2-AS1 in tumor tissues (X. Yang, Duan, and Zhou 2017; Ye et al. 2021; M. Zhang et al. 2019; Zhu et al. 2018). The mentioned studies also used RT-qPCR analysis to quantify FOXD2-AS1 expression levels. Usually, discrepancies can be partially explained because of the approach used to measure expression, the normalization methodology employed, or the natural heterogeneity of the cohorts of study.

As mentioned before, there it has been widely shown a co-expression pattern displayed in bidirectional gene pairs often due to the presence of common transcriptional regulatory elements (Trinklein et al. 2004). Of note, most lncRNAs originating from the opposite strand of a 5' protein-coding gene show considerably lower expression levels than the respective coding gene (Derrien et al. 2012). In consequence, we explored the ratio of expression, calculated as FOXD2-AS1 versus FOXD2 expression, which confirmed a tendency of lower FOXD2-AS1 expression relative to FOXD2 (ratio <1). However, we noticed a flipping trend in CRC tumors, whereas FOXD2-AS1/FOXD2 expression ratio was significantly increased compared to normal samples (**Supplementary Table S9**). Even though the mechanisms driving specific bidirectional transcription are far from clear, such

deregulation can potentially be explained by alterations during the transcriptional cycle. While RNA polymerase II recruitment and transcription initiation have been described to be very active at bidirectional promoters affecting both directions (Core, Waterfall, and Lis 2008), epigenetics and other molecular regulators interplay with RNA polymerase can result in altered elongation leading to asymmetric transcription (Hodges et al. 2009). Chromatin architecture, commonly altered in cancer cells, is typically accompanied by specific signatures in histone modifications, nucleosome positioning, and DNA methylation, and is an important regulator of gene expression by promoting or restricting the binding of complexes involved in transcription. In the particular case of bidirectional genes. Jangid *et al*. and colleagues reported that the histone modification landscape mirrors the transcriptional status of each independent gene of the pair, revealing a strong association between histone marks and transcription profiles. As cancer has atypical histone modifications patterns, asymmetric chromatin states upstream and downstream FOXD2 and FOXD2-AS1 shared promoter in cancer cells could explain altered patterns of expression and partial loss of co-expression. Also, chromatin remodeling involves changes in interactions with TF and enhancers, potentially driving the loss of bidirectionality (S. S. Ahmad et al. 2021; Hodges et al. 2009; Saunders, Core, and Lis 2006), which can be disrupted in cancer cells by aberrant DNA methylation profiles. DNA looping between active promoters and 3' ends with functional polyA signals enhances transcription (Perkins et al. 2008) and mediates transcription memory conservation (S. S. Ahmad et al. 2021). Indeed, gene loops in protein-coding genes maintain the direction of transcription in bidirectional promoters, restricting higher transcription levels of divergent ncRNAs (Tan-Wong et al. 2012).

Overall, I hypothesize that the impaired expression between FOXD2 and FOXD2-AS1 observed in colorectal cancer could be driven by asymmetric epigenetic states upstream and downstream their promoter, combined with transcription factors and associated machinery operating within cancer cells. However, our analysis has only addressed DNA methylation patterns (see results 3.4 and 4.2) but not a complete epigenetic characterization of the region. To fully understand the dynamics of FOXD2 and FOXD2-AS1 regulation in normal and cancer cells, we plan to explore other epigenetic factors, including nucleosome occupancy, histone modification marks, and binding of TF in both tissue types. Additionally, other mechanisms affecting RNA stability and posttranscriptional regulation cannot be excluded.

3. FOXD2/FOXD2-AS1 regulation beyond promoter

DNA methylation changes are one of the main molecular hallmarks associated with neoplastic cells. Hypermethylation of promoter CpGi affecting tumor suppressor genes is one of the most frequent epigenetic alterations reported in cancer. Referring back to the detected changes in FOXD2 and FOXD2-AS1 expression between normal and tumor samples, we wondered if DNA methylation could play a major role in their regulation. We thus, explored the DNA methylation patterns among the chromosome 1 locus containing FOXD2 and FOXD2-AS1 genes among TCGA-COAD and HUB cohorts. Contrary to what was expected, gain of DNA methylation outside the promoter was a predominant change in colorectal tumors. Indeed, the promoter remained unmethylated in all samples analyzed while gene bodies were hypermethylated, discarding gene silencing due to promoter hypermethylation and suggesting a differential methylation mechanism involved in gene expression. Methylation gain surrounding the promoter placed in gene bodies was significantly associated with a decreased expression of FOXD2 and FOXD2-AS1. Such negative association was stronger regarding FOXD2 expression and the CpGs located at the 3' UTR of FOXD2 displayed the highest associations with expression, as previously reported (Conesa-Zamora et al. 2015) (**Tables 13** and **15**). Furthermore, we treated CRC cell lines with DAC demethylating agent and observed an increase of expression only if there was a loss of gene body methylation (**Figure 33**). These results indicate the repressive role of DNA gene body methylation in FOXD2 and FOXD2-AS1 expression.

Despite the large number of recent studies on epigenetics, there is still a lack of understanding of how some genes behave in tumor cells with respect to their DNA methylation changes. Gene repression through promoter methylation is considered one of the most important epigenetic modifications in cancer. However, surprising numbers of methylated CpGs exist in non-promoter regions (Weber et al. 2005) and their methylation effects are less well understood. Our experiments show that gene body DNA methylation is associated with decreased expression of FOXD2 and FOXD2-AS1. There is growing evidence of the potential role of DNA methylation beyond promoters on gene expression (Peter A Jones 2012), but its functionality has been controversial. In agreement with our results, other studies have demonstrated a negative correlation between gene body methylation and gene expression (Xiaojing Yang et al. 2014), and several mechanisms have been hypothesized. According to *Lorincz et al.*, polymerase depletion in highly methylated intragenic regions reduces transcription elongation. Neri *et al.* found aberrant transcription events associated with Dnmt3b-dependent gene body DNA methylation. Also, Maunakea *et al.* reported that H3K4me3 within gene bodies, which is usually enriched at promoters, had an inversed correlation with methylated intergenic CpGi,

suggesting the role of alternative promoters. On the contrary, some studies have shown a positive association with gene expression levels (P. A. Jones 1999; Shann et al. 2008), claiming that methylation blocks transcription initiation but not elongation. For instance, H3K36me3 histone mark has been correlated with methylated and active transcribed regions (Ball et al. 2009) and reported to recruit DNMTs, thus inducing gene body methylation (Hahn et al. 2011). Overall, such studies indicate a more complex view of the functional implications of DNA methylation in different gene parts.

To conclude, our results indicate that DNA methylation surrounding the promoter and within body genes, and perhaps other epigenetic mechanisms, are the main factors responsible for the deregulation of FOXD2 and FOXD2-AS1 in CRC. However, further investigations are needed to elucidate possible mechanisms behind this correlation.

Of note, we observed that FOXD2 and FOXD2-AS1 were poorly expressed in most normal tissues, being the colon one of the tissues with the highest expression levels (data from GTEX). In addition, when exploring methylation changes within the regions spanning and flanking these genes, we realized that colorectal tissues (TCGA COAD and READ) had the lowest methylation levels among normal tissues and a major methylation gain in tumors, compared to the rest of tissue types (data from Wanderer tool (Díez-Villanueva, Mallona, and Peinado 2015)). Interestingly, tissue-specific methylation frequently occurs within gene bodies rather than in promoters (Maunakea et al. 2010). Together with these data, our results suggest tissue-specific expression associated with DNA methylation patterns of FOXD2 and FOXD2-AS1, which are critical for understanding their functions and implications in cancer.

4. Clinical correlates of FOXD2 and FOXD2-AS1 in CRC

We next explored FOXD2 and FOXD2-AS1 as potential prognostic biomarkers of CRC by studying associations between the tumor clinicopathological characteristics and methylation and expression levels reported in the corresponding tumor samples. The correlations detected were rather poor.

Before discussing the results, I would like to make some considerations about the limitations of our study. The TCGA project was initially designed for molecular analyses and clinical data collection was secondary, reflecting uneven disease stages and tumor characteristics distribution. However, considering the extensive collection of tumors in the TCGA-COAD cohort, tumor characteristics were heterogeneous enough to study associations with expression and methylation profiles. On the other hand, the HUB cohort

was much more limited as all patients were distributed on CRC stages II and III. Most tumors displayed low invasion rates, no patients had distant metastases and survival rates were ~90%. Therefore, we are aware that the HUB cohort is relatively small and homogeneous, and consequently, clinico-pathological potential associations might be short of statistical power.

To our knowledge, no published studies have evaluated FOXD2 expression as a prognostic biomarker in cancer. In contrast, many studies have predicted FOXD2-AS1 prognosis, typically associating its overexpression with poor survival rates in several cancer types such as osteosarcoma (Z. Ren et al. 2019; H. Zhang et al. 2019), breast (Jiang et al. 2019), glioma (Dong, Cao, and Xue 2019), thyroid (H. Li et al. 2019; Y. Zhang et al. 2019), gastric (Mao et al. 2020; Xu et al. 2018) and bladder (Su et al. 2018). Of note, one study reported the opposite, whereas FOXD2-AS1 expression levels were positively associated with higher survival rates in thyroid cancer (Lu et al. 2018). Two other studies showed no statistical associations between FOXD2-AS1 expression and patient's survival in melanoma (W. Ren, Zhu, and Wu 2019) and stomach cancer (Q. Li et al. 2020).

Regarding other clinicopathological characteristics of tumors, FOXD2-AS1 upregulation has been correlated with lymph node metastasis and advanced TNM stage in cholangiocarcinoma (Hu et al. 2021), cervical cancer (Dou et al. 2020), glioma (Gu et al. 2019), breast cancer (Jiang et al. 2019), gastric cancer (Xu et al. 2018) and thyroid cancer (Liu et al. 2019). However, some negative results have also been published, reporting no associations in breast cancer (Arabpour et al. 2021) and in esophageal cancer (Bao et al. 2018).

Specifically, FOXD2-AS1 clinical associations with CRC are relatively weak. Two studies reported an association of high expression of FOXD2-AS1 with lower survival rates in CRC (M. Zhang et al. 2019; Zhu et al. 2018), however, Zhu *et al*. only found such association for short-term survival rates, but not long-term. Yu *et al*. found no significant clinico-pathological characteristics associated with FOXD2-AS1 expression regarding CRC tumor size, position, metastasis, and TNM stage. These results are in agreement with our data, showing no relevance of FOXD2-AS1 expression on colorectal tumor characteristics. As most cancer biomarkers have been reported for specific cancer types or subtypes, it is not surprising that FOXD2-AS1 could serve as a prognostic biomarker in several cancer types, but it does not apply to CRC. Additionally, we should consider that most studies focused on FOXD2-AS1 are based only on Asian populations and sometimes have insufficient sample size. This, together with the bias of publishing positive results, leads us to believe that the potential prognostic value of FOXD2-AS1 could be overestimated.

Although we did not find significant clinical correlations with FOXD2 or FOXD2-AS1 expression, we observed a tendency of higher FOXD2-AS1 and lower FOXD2 associated with tumor malignancy. This is reflected in the significant associations observed between higher FOXD2-AS1/FOXD2 ratio of expression and bigger tumor size and invasion, lymph node metastasis, and distant metastasis (**Table 16**). Overall, our results indicate no prognostic value of FOXD2 and FOXD2-AS1 in CRC. Nevertheless, the disruption of the FOXD2/FOXD2-AS1 balance demonstrated association with enhanced tumor aggressiveness. Together with the fact that this region belongs to a large module of co-methylation CpGs (Mallona et al. 2018) (see background study I), the data suggests the involvement of major mechanisms rewiring cancer, that could be driving imbalanced bidirectional transcription in this locus and with a potential clinical impact.

5. Functional characterization of FOXD2 and FOXD2-AS1 in CRC cell line

Several studies have reported the oncogene role of FOXD2-AS1 in several cancer types, including CRC, but to date, none have studied FOXD2 functional role. In consequence, we modulated their expression in SW480 CRC cell line to explore their impact on cell biology.

We previously performed several attempts to silence FOXD2 and FOXD2-AS1 in LoVo and HCT116 cell lines, first by using CRISPR genome editing tool. However, CRISPR Cas9 system is less reliable to elucidate lncRNA function, as they lack ORF and small indels might not alter lncRNA function. Therefore, as expected, we did not observe changes of expression after disrupting FOXD2-AS1 using the CRISPR system. The second attempt to block FOXD2 and FOXD2-AS1 was with Gapmers, which are modified antisense nucleotides able to enter the nucleus that from a DNA-RNA complex that triggers RNase H1 degradation of the target RNA. Unfortunately, the downregulation levels achieved with this technique were mild and only lasted for 48 hours for both genes, difficulting the study with cell functional assays. We recently used shRNAs that successfully downregulated FOXD2 and FOXD2-AS1 levels in LoVo cell line, however we did not have time to characterize the functional implications associated with such silencing.

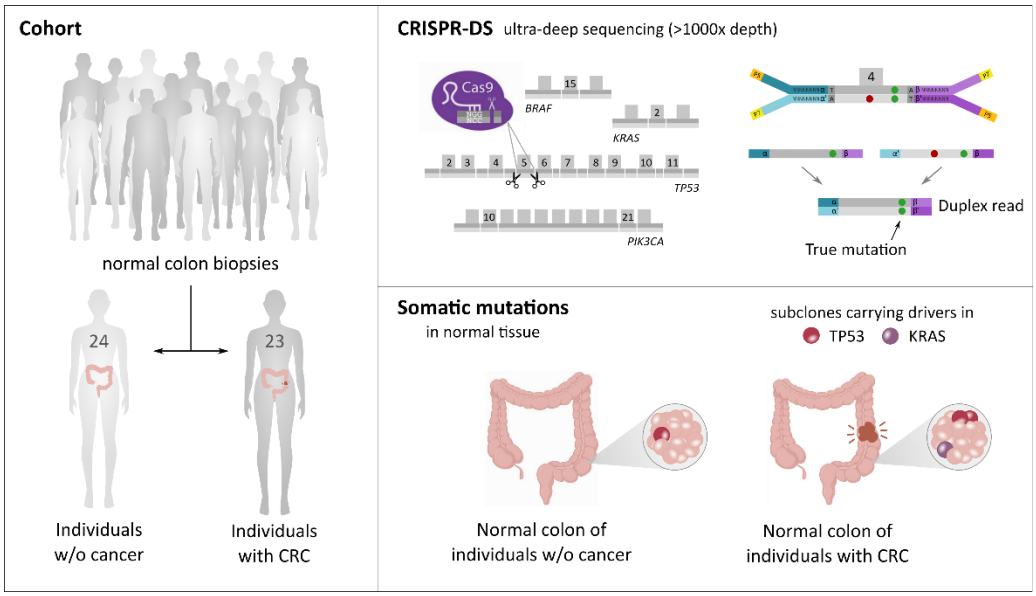
To overexpress FOXD2 and FOXD2-AS1 we first used the CRISPR activation system. Despite being an efficient and simple gene editing tool, the fact that both FOXD2 and FOXD2-AS1 were poorly annotated in two of the reference gene annotation databases (RefSeq and GENCODE), led to failed attempts on the their upregulation. Only one sgRNA successfully worked, but by upregulating both FOXD2 and FOXD2-AS1 genes, thus difficulting to study them independently. Interestingly, we observed no effects on cell proliferation, migration nor colony formation with the coordinated upregulation of FOXD2 and FOXD2-AS1.

We finally achieved FOXD2 and FOXD2-AS1 upregulation separately, by using overexpression vectors for each gene. FOXD2, which is strongly downregulated in CRC, behaved as a tumor suppressor gene by decreasing cell migration and colony formation abilities (**Figure 37**). FOXD2-AS1, previously described as an oncogene lncRNA, only displayed higher migration rates, but no effects on cell proliferation or colony formation (**Figure 38**). Three studies have *in vitro* modulated FOXD2-AS1 expression with siRNAs or overexpression vectors and performed functional assays in CRC cell lines. In addition to our observation of a higher increase of cell migration, they also reported higher proliferation rates, invasion and colony formation (X. Yang, Duan, and Zhou 2017; Ye et al. 2021; Zhu et al. 2018). However, none of them studied FOXD2-AS1 in SW480 cell line. Of note, our results are based only on one model, therefore we should extrapolate these results into other cell lines and by using another method to modulate their expression.

Overall, to our knowledge, this is the first time that FOXD2 is shown to have tumor suppressor functions as its upregulation enhanced migration and cell colony formation. Additional experiments need to be performed to further confirm these results and better explore FOXD2 functional role in cancer cells.

Study II

Colorectal cancer is associated with the presence of cancer driver mutations in normal colon



Graphical abstract study II. CRISPR-DS error-correction sequencing technology enabled the detection of low frequency mutations in the the normal colon mucosa of individuasl with and without CRC. *KRAS* and *TP53* mutations were more frequent in normal colon of patients with CRC, in larger clones specially in patients with early onset CRC.

1. Looking for a needle in the haystack: somatic mutations in non-cancer tissues

While somatic mutations in CRC are well characterized, little is known about the accumulation of cancer mutations in healthy colorectal tissue. Providing evidence of mutations acquired in early tumorigenesis is technically challenging due to the intrinsic technical limitations of currently available methods to reliably detect small subsets of mutant cells within morphologically normal tissue. To date, standard NGS technologies need to optimize the error rate to be useful to distinguish between a sequencing error and a real somatic mutation at low frequency. Different approaches have been developed to address this problem, such as *in vitro* clone expansion, microbiopsy sequencing, single-cell sequencing and error corrected deep-sequencing.

In vitro clone formation consists of the isolation of single cells and their expansion in cell culture to a large number of cells amenable for standard NGS methods (Jager et al. 2018). Several studies have used this methodology for different tissues, including skin (Abyzov et al. 2012; Tang et al. 2020), liver and intestine (Blokzijl et al. 2016), blood (Lee-Six et al. 2018; Ortmann et al. 2015), and lung (Yoshida et al. 2020). However, culture-related artifacts must be considered, such as clonal selection and acquired mutations during cell growth. Other disadvantages of this approach are that not all cell types can be successfully expanded in cell culture. On the other hand, tissue microbiopsies aim to reduce the sample size to very small or microscopic regions, in which mutant clones represent a sizable proportion of cells (Ellis et al. 2021). Thus, it takes advantage of the in vivo clonal expansions to identify mutations present in a limited number of adjacent cells. The collection can be done randomly or by targeting individual clonal structures, such as colonic crypts (Lee-Six et al. 2019). Multi-region sequencing will only be a valuable approach for tissues in which clonal areas are of a detectable size. It has been successfully applied to study esophagus (Yokoyama et al. 2019), liver (Brunner et al. 2019), skin (Martincorena et al. 2015), bladder (Lawson et al. 2020; Li et al. 2020), and colon (Lee-Six et al. 2019). The main inconvenience is that it requires the analysis of a large number of samples per individual. Overall, the two mentioned approaches are very labor-intensive, which challenges large cohort studies and transcriptional applications.

Single-cell sequencing methods aim to call mutations in genomes at the cellular level. However, the acquisition of high-quality data is technically challenging as it requires high amplification rates to obtain sufficient DNA for sequencing, introducing artifacts, such as amplification bias, genome loss, mutations, and chimeras (Gawad, Koh, and Quake 2016).

An alternative approach to detect low frequency somatic mutations within normal tissue consists of performing ultra-deep sequencing using high-accuracy NGS methods such as Duplex Sequencing (DS) (Schmitt et al., 2012). As previously mentioned, DS employs double-stranded molecular tags, allowing an additional level of correction by comparison of independent consensus derived from the two complementary strands of the original DNA molecule. This extra error correction enables the detection of a single mutation among $>10^7$ sequenced bases (Kennedy et al. 2014; Schmitt et al. 2012), providing extreme resolution to identify clonal expansions in a single normal biopsy. Several studies have used DS to study normal tissues and body fluids (Krimmel-Morrison et al. 2019; Krimmel et al. 2016; Jesse J. Salk et al. 2019; Short et al. 2020). The main disadvantage of this high-sensitive approach is that recapturing both strands of the DNA and achieving high depth requires large sequencing capabilities and, thus, is more suitable for small target regions.

In this study, we have used DS in combination with CRISPR/Cas9 digestion of the DNA (Nachmanson et al. 2018) (**Figure 14**). Using CRISPR we excised the DNA generating fragments of a similar size that cover regions of interest, which then were enriched by size selection and used for library preparation. Thereafter, only one round of hybridization capture of the panel of genes is required (instead of two), reducing costs and time. Compared to conventional DS, this method avoids sonication-related errors, improves fragment size homogeneity reducing PCR bias amplification of short fragments, and increases the recovery, allowing us to deep-sequence (~2,500x) a small panel using 100 ng or less of DNA. One inconvenience of this approach was the optimization process for successful DNA digestion. The CRISPR/Cas9 system (*Streptococcus pyogenes*) uses target DNA preceded by a PAM sequence (NGG) (Ran et al. 2013) which limits the sgRNA sequences surrounding the regions of interest. Additionally, the efficiency of the sgRNA, even though it can be predicted *in silico*, needs to be *experimentally* tested to achieve a homogeneous coverage of the targeted areas.

2. Normal colon of CRC patients show an increased mutation rate than individuals without cancer

In this study, we characterized the normal colon of 47 individuals, 23 with CRC and 24 cancer-free with a sensitivity of at least 1 mutation in 1000 genomes. Even though others have attempted to identify somatic mutations in normal cells (Abascal et al. 2021; Blokzijl et al. 2016; Lee-Six et al. 2019), all of them had worse sensitivity than CRISPR-DS to detect very low frequency cancer mutations. Such studies demonstrated an increase of age-related somatic mutations, which is in agreement with our findings regarding non-coding mutation frequency associated with normal colon (**Figure 42.B**). A remarkable observation from our data is that we observed key differences between the normal colon of patients with and without cancer. These include a higher coding mutation frequency and the presence of driver *KRAS* and *TP53* mutations in the normal colon of CRC patients, which suggests an enhanced process of somatic evolution compared to the normal colon of cancer-free individuals (**Figure 45**). Lee-Six *et al*. laser microdissected hundreds of normal colonic crypts to further sequence their genomes and demonstrated that around 1% of normal crypts carry a clonal driver mutation in middle-aged individuals. Most mutated genes identified in their study were not common CRC driver genes, e.g., ERBB2, ERBB3, and FBXW7 (Lee-Six et al. 2019). They also reported truncating mutations in *TP53*, but no *KRAS* mutations were identified. Interestingly, they reported a similar frequency of driver mutations in the normal colon of patients with and without CRC. The authors claimed that about 90% of the called mutations were fully clonal, suggesting that the observed driver

mutations likely contribute to crypt colonization by a mutant stem cell. Therefore, the hypothesis that *KRAS* and *TP53* mutations contribute to subsequent preneoplastic transformation is supported by our data that demonstrated a higher frequency of such mutations in the normal colon of individuals that progressed to CRC. Additionally, *KRAS* activating mutations were found to dramatically increase crypt fission in mice, driving field cancerization (Snippert et al. 2014), and were postulated to promote lateral expansions of mutant crypts in human colon epithelium (Nicholson et al. 2018), indicating its involvement in tumor development. Also, a recent evolutionary history model of cancers revealed that a set of driver genes characterizes early carcinogenesis stages, including *TP53* and *KRAS* (after APC) (Gerstung et al. 2020). Mutations in these two genes were identified as early drivers, participating in initiating and even precancerous events in colorectal cancer. Regarding *TP53*, the finding of prevalent *TP53* somatic evolution is not new. In accordance with previous observations at Risques lab, ultra-deep sequencing of *TP53* gene revealed that women with ovarian cancer had higher *TP53* mutation burden in the peritoneal fluid and pap tests (Krimmel-Morrison et al. 2019; J J Salk et al. 2019), suggesting an increased somatic evolution in the context of cancer development. Overall, our results are consistent with these data and highlight the role of *TP53* and *KRAS* mutations in histologically normal epithelium in the normal colon of individuals with CRC, suggesting that multiple mutant clones accumulate in these individuals through life, whith potential to evolve into cancer.

Of note, out of all driver mutations identified in the normal colon of patients with CRC, a minority matched mutations present in the paired tumors, possibly indicating potential dissemination of the cancer cells from the primary tumor site. However, most driver mutations identified in the normal colon (70%) did not coincide with the driver mutation in the tumor, revealing carcinogenic fields composed of multiple precancerous clonal expansions instead of a single large clonal patch (see results 4). Field cancerization (also known as field-effect) is the consequence of the evolution of somatic cells in the body, resulting in cells that carry some but not all phenotypes required for malignancy (Curtius, Wright, and Graham 2017). Prior studies exemplified this concept by reporting the presence of clonal expansions in the colon of patients with ulcerative colitis, who are prone to develop CRC (Baker et al. 2018). Overall, the data obtained demonstrate the applicability of using single-molecule ultra-deep sequencing to identify and characterize carcinogenic fields in CRC. However, more extensive studies in multiple normal tissue biopsies would be beneficial to deeper understand the field-effect phenomena.

Finally, it should be mentioned that another possibility is that normal samples were contaminated by tumor DNA. However, we minimized this risk by taking extreme care in

normal sample collection during surgery, as well as by performing DNA extractions and library preparation of normal colon tissues months before processing the tumors.

3. Early onset CRC and cancer prevention

We previously mentioned the increased incidence of CRC among adults younger than 50 years old (see introduction 1.2). In consequence, we designed the study to include individuals with early-onset CRC (≤50 years old) with the aim to investigate the nature of clonal expansions in the normal colon when cancer develops at a younger age. We demonstrated that two thirds of young individuals with CRC (7/11) (**Figure 46**) harbor *TP53* driver mutations in their normal colon, and out of these, in most of them (5/7) the mutated clones are large. Knowing that each duplex read corresponds to one original DNA molecule, large clones are represented by multiple duplex reads. In line with our findings, a study revealed an increased *TP53* functional loss in patients with early onset CRC (Kim et al. 2021). Therefore, the enrichment of large clones carrying driver mutations in young individuals with CRC suggests that their enhanced CRC development might be related to a high frequency of clones carrying *TP53* driver mutations or involving frequent *TP53* loss.

Cancer is a disease that evolves by the accumulation of somatic mutations, and understanding such phenomena provides clues on how we can manage the disease. As a multicellular organism with a long lifespan, imperfect DNA replication and repair generate mutations. In addition, many other extrinsic factors, sometimes related with lifestyle, can also contribute to tumorigenesis (see introduction 1.3). A study focused on esophageal epithelium revealed that alcohol consumption and smoking accelerate the emerge of clones carrying driver mutations (Yokoyama et al. 2019). Also, the skin exposed to UV radiation had a considerably higher mutation rate than non-exposed skin (Yizhak et al. 2019). In our study, patients without CRC displayed an association between higher mutation rates and male gender as well as with the presence of colorectal polyps (**Figure 47**), which are two CRC risk factors. Therefore, these data indicate that by studying somatic mutations in normal colon tissue, we could potentially investigate the mechanisms of action associated with extrinsic or intrinsic risk factors, overall serving as a potential biomarker for cancer risk prediction. Indeed, we developed a preliminary regression model that could predict the CRC risk by considering the age and frequency of driver mutations in the normal colon of individuals. Although the model only included a small number of individuals and needs further validation in other cohorts, it demonstrates the applicability of using the mutational study in normal colon mucosa to predict CRC risk.

4. Study limitations and future perspectives

Several limitations exist in our study. One is the small gene panel size analyzed by CRISPR-DS (~3.5Kb). Although the study aimed to determine if ultra-sensitive sequencing of normal colon mucosa could have clinical value to assess CRC risk, the limited span did not allow the assessment of mutational signature analysis to better understand the biological mechanisms responsible for mutations. However, even with this small sequenced region, we still observed a similar pattern of mutation signature distribution between the normal colon of patients with CRC and CRC cells. The advantage of our approach was that even in a reduced number of driver genes, we could provide accurate estimates of clone size and abundance in a single biopsy. Second is the nature of the samples, as we cannot assume that all the genomes analyzed were from epithelial cells, limiting estimations regarding percentages of mutated crypts as well as mutation rates associated with age. In addition, in patients with CRC, normal biopsies were collected at least 10 cm from the, whereas tissues from patients without cancer were collected during colonoscopies. This difference between groups of patients should be taken into account. Of note, there is debate about how non-malignant tissue near the primary tumor is truly “normal” (Yadav, Degregori, and De 2016). Also differences in the mutation rates in different locations of the colon cannot be excluded. The third limitation of the study is that normal colon was not tested prior CRC development, which would help us to better understand somatic evolution and tumor development.

Moving forward, the easy accessibility of the colon and the frequent procedure of colonoscopies in the clinical practice make possible the collection of normal colon biopsies in large cohorts of study, even in patients prior to CRC development and progression. The analysis of the normal clonal dynamics opens the potential to offer personalized cancer risk predictions of progression to malignancy that could detect cancer in early stages, improving survival and even prevent disease's progression. This risk assessment would be especially useful for young individuals, which demonstrated an excess of *TP53* driver mutations associated with CRC. Also, this study can be useful to catalog low frequency mutations that can help identify the cells and processes responsible for cancer origin, as well as to extrapolate the study of somatic evolution as a function of aging and exposure to cancer risk factors.

In summary, the work presented in this study contributes to a deeper understanding of the somatic mutational process that takes place in the normal colon epithelium and its potential impact on CRC initiation or progression. We have demonstrated the presence of somatic mutations in common CRC genes in normal colon of most individuals, however,

with a higher abundance in patients with CRC. We highlight the presence of driver *KRAS* and *TP53* in distant tissue from the primary tumor, with enrichment of larger clones carrying such mutations in early onset CRC. Overall, these results expand our knowledge of somatic evolution in the colon, offering insights about different mechanisms of carcinogenesis in early vs. late onset CRC. Moreover, they offer the possibility of using normal colon biopsies for CRC risk assessment based on ultra-deep sequencing analysis of a reduced panel of cancer genes.

CONCLUSIONS

The conclusions of this doctoral thesis are organized according to the experimental structure, which includes the two studies presented:

Study I – FOXD2 and FOXD2-AS1 in colorectal cancer

- FOXD2 and FOXD2-AS1, head-to-head genes regulated by a shared bidirectional promoter, are co-expressed in the human colon, but FOXD2 is remarkably downregulated in most CRC.
- Decreased expression of FOXD2 in CRC is associated with gain of DNA methylation in the gene body of FOXD2 and FOXD2-AS1.
- FOXD2 / FOXD2-AS1 co-expression disregulation is associated with poor prognosis in CRC.
- FOXD2 overexpression decreases cell migration and colony formation in CRC cell lines.
- FOXD2-AS1 overexpression promotes cell migration.

Study II - Colorectal cancer is associated with the presence of cancer driver mutations in normal colon

- High sensitive CRISPR-DS technique revealed somatic mutations in common CRC genes in the normal colon mucosa of patients with and without CRC.
- *KRAS* and *TP53* driver mutations are more commonly found in normal colon of CRC patients.
- Cancer driver mutations often display clonal expansion in early onset CRC, suggesting an enhanced cancer risk.
- Our integrative mutational analysis may have application for CRC risk prediction.

REFERENCES

Abascal, F., Harvey, L. M. R. R., Mitchell, E., Lawson, A. R. J. J., Lensing, S. V., Ellis, P., … Martincorena, I. (2021). Somatic mutation landscapes at single-molecule resolution. *Nature*, *593*(7859), 405–410. https://doi.org/10.1038/s41586-021-03477-4

Abyzov, A., Mariani, J., Palejev, D., Zhang, Y., Haney, M. S., Tomasini, L., … Vaccarino, F. M. (2012). Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. *Nature*, *492*(7429), 438–442. https://doi.org/10.1038/nature11629

Ahmad, S., Abbas, M., Ullah, M. F., Aziz, M. H., Beylerli, O., Alam, M. A., … Ahmad, A. (2021). Long non-coding RNAs regulated NF-κB signaling in cancer metastasis: Micromanaging by not so small non-coding RNAs. *Seminars in Cancer Biology*, (May). https://doi.org/10.1016/j.semcancer.2021.07.015

Ahmad, S. S., Samia, N. S. N., Khan, A. S., Turjya, R. R., & Khan, M. A.-A.-K. (2021). Bidirectional promoters: an enigmatic genome architecture and their roles in cancers. *Molecular Biology Reports*. https://doi.org/10.1007/s11033-021-06612-6

Alexandrov, L. B., & Stratton, M. R. (2014). Mutational signatures: The patterns of somatic mutations hidden in cancer genomes. *Current Opinion in Genetics and Development*, *24*(1), 52–60. https://doi.org/10.1016/j.gde.2013.11.014

Andrew P, F., & Bert, V. (1983). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature*, *301*(9), 89–92. https://doi.org/10.1038/301089a0

Arabpour, M., Layeghi, S. M., Bazzaz, J. T., Naghizadeh, M. M., Majidzadeh-A, K., & Shakoori, A. (2021). The potential roles of lncRNAs DUXAP8, LINC00963, and FOXD2-AS1 in luminal breast cancer based on expression analysis and bioinformatic approaches. *Human Cell*, *34*(4), 1227–1243. https://doi.org/10.1007/s13577-021-00539-7

Armaghany, T., Wilson, J. D., Chu, Q., & Mills, G. (2012). Genetic alterations in colorectal cancer. *Gastrointest Cancer Res*. https://doi.org/10.1007/978-0-85729-984-0_2

Auclair, G., & Weber, M. (2012). Mechanisms of DNA methylation and demethylation in mammals. *Biochimie*, *94*(11), 2202–2211. https://doi.org/10.1016/j.biochi.2012.05.016

Baker, K. T., Salk, J. J., Brentnall, T. A., & Risques, R. A. (2018). Precancer in ulcerative colitis: The role of the field effect and its clinical implications. *Carcinogenesis*, *39*(1), 11–20. https://doi.org/10.1093/carcin/bgx117

Balas, M. M., & Johnson, A. M. (2018). Exploring the mechanisms behind long noncoding RNAs and cancer. *Non-Coding RNA Research*, *3*(3), 108–117. https://doi.org/10.1016/j.ncrna.2018.03.001

Ball, M. P., Li, J. B., Gao, Y., Lee, J. H., Leproust, E. M., Park, I. H., … Church, G. M. (2009). Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature Biotechnology*, *27*(4), 361–368. https://doi.org/10.1038/nbt.1533

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, *21*(3), 381–395. https://doi.org/10.1038/cr.2011.22

Bao, J., Zhou, C., Zhang, J., Mo, J., Ye, Q., He, J., & Diao, J. (2018). Upregulation of the long noncoding RNA FOXD2-AS1 predicts poor prognosis in esophageal squamous cell carcinoma. *Cancer Biomarkers*, *21*(3), 527–533. https://doi.org/10.3233/CBM-170260

Barker, N., Ridgway, R. A., Van Es, J. H., Van De Wetering, M., Begthel, H., Van Den Born, M., … Clevers, H. (2009). Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature*, *457*(7229), 608–611. https://doi.org/10.1038/nature07602

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., … Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, *129*(4), 823–837. https://doi.org/10.1016/j.cell.2007.05.009

Bell, C. G., & Beck, S. (2010). The epigenomic interface between genome and environment in common complex diseases. *Briefings in Functional Genomics*, *9*(5–6), 477–485. https://doi.org/10.1093/bfgp/elq026

Berger, S. L., Kouzarides, T., Shiekhattar, R., & Shilatifard, A. (2009). An operational definition of epigenetics. *Genes and Development*, *23*(7), 781–783. https://doi.org/10.1101/gad.1787609

Bernard, D., Prasanth, K. V, Tripathi, V., Colasse, S., Nakamura, T., Xuan, Z., … Bessis, A. (2010). *A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression*. *29*(18), 3082–3093. https://doi.org/10.1038/emboj.2010.199

Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., … Van Boxtel, R. (2016). Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, *538*(7624), 260–264. https://doi.org/10.1038/nature19768

Boman, B. M., & Huang, E. (2008). Human colon cancer stem cells: A new paradigm in gastrointestinal oncology. *Journal of Clinical Oncology*, *26*(17), 2828–2838. https://doi.org/10.1200/JCO.2008.17.6941

Bornelöv, S., Komorowski, J., & Wadelius, C. (2015). Different distribution of histone modifications in genes with unidirectional and bidirectional transcription and a role of CTCF and cohesin in directing transcription. *BMC Genomics*, *16*(1), 1–13. https://doi.org/10.1186/s12864-015-1485-5

Brunner, S. F., Roberts, N. D., Wylie, L. A., Moore, L., Aitken, S. J., Davies, S. E., … Campbell, P. J. (2019). Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature*, *574*(7779), 538–542. https://doi.org/10.1038/s41586-019-1670-9

Cabili, M., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., & Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes and Development*, *25*(18), 1915–1927. https://doi.org/10.1101/gad.17446611

Carroll, J. S., Liu, X. S., Brodsky, A. S., Li, W., Meyer, C. A., Szary, A. J., … Brown, M. (2005). Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell*, *122*(1), 33–43. https://doi.org/10.1016/j.cell.2005.05.008

Chen, X., Fan, S., & Song, E. (2016). *The Long and Short Non-coding RNAs in Cancer Biology* (Vol. 927). https://doi.org/10.1007/978-981-10-1498-7

Chen, Y., Li, Y., Wei, J., & Li, Y. Y. (2014). Transcriptional regulation and spatial interactions of head-to-head genes. *BMC Genomics*, *15*(1), 1–9. https://doi.org/10.1186/1471-2164-15-519

Cheng, C. W., Chen, P. M., Hsieh, Y. H., Weng, C. C., Chang, C. W., Yao, C. C., … Shen, C. Y. (2015). Foxo3a-mediated overexpression of microRNA-622 suppresses tumor metastasis by repressing hypoxia-inducible factor-1a in erk-responsive of lung cancer. *Oncotarget*, *6*(42), 44222–44238. https://doi.org/10.18632/oncotarget.5826

Clark, S. J., Statham, A., Stirzaker, C., Molloy, P. L., & Frommer, M. (2006). DNA methylation: Bisulphite modification and analysis. *Nature Protocols*, *1*(5), 2353–2364. https://doi.org/10.1038/nprot.2006.324

Clemson, C. M., Hutchinson, J. N., Sara, S. A., Ensminger, A. W., Fox, A. H., Chess, A., & Lawrence, J. B. (2009). Article An Architectural Role for a Nuclear Noncoding RNA : NEAT1 RNA Is Essential for the Structure of Paraspeckles. *Molecular Cell*, *33*(6), 717–726. https://doi.org/10.1016/j.molcel.2009.01.026

Click, B., Pinsky, P. F., Hickey, T., Doroudi, M., & Schoen, R. E. (2018). Association of colonoscopy adenoma

findings with long-term colorectal cancer incidence. *JAMA - Journal of the American Medical Association*, *319*(19), 2021–2031. https://doi.org/10.1001/jama.2018.5809

Concordet, J. P., & Haeussler, M. (2018). CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res*, *46*(W1), W242–W245. https://doi.org/10.1093/nar/gky354

Conesa-Zamora, P., García-Solano, J., Turpin, M. del C., Sebastián-León, P., Torres-Moreno, D., Estrada, E., … Conesa, A. (2015). Methylome profiling reveals functions and genes which are differentially methylated in serrated compared to conventional colorectal carcinoma. *Clinical Epigenetics*, *7*(1), 1–14. https://doi.org/10.1186/s13148-015-0128-7

Cong, Y. J., Gan, Y., Sun, H. L., Deng, J., Cao, S. Y., Xu, X., & Lu, Z. X. (2014). *Association of sedentary behaviour with colon and rectal cancer : a meta-analysis of observational studies mode*. (October 2013), 817–826. https://doi.org/10.1038/bjc.2013.709

Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science*, *322*(DECEMBER), 1845–1849. https://doi.org/10.5040/9780755621101.0007

Curtius, K., Wright, N. A., & Graham, T. A. (2017). An evolutionary perspective on field cancerization. *Nature Reviews Cancer*, *18*(1), 19–32. https://doi.org/10.1038/nrc.2017.102

Davidson, K. W., Barry, M. J., Mangione, C. M., Cabana, M., Caughey, A. B., Davis, E. M., … Wong, J. B. (2021). Screening for Colorectal Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA - Journal of the American Medical Association*, *325*(19), 1965–1977. https://doi.org/10.1001/jama.2021.6238

Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes and Development*, *25*(10), 1010–1022. https://doi.org/10.1101/gad.2037511

Dekker, E., Tanis, P. J., Vleugels, J. L. A., Kasi, P. M., & Wallace, M. B. (2019). Colorectal cancer. *The Lancet*, *394*(10207), 1467–1480. https://doi.org/10.1016/S0140-6736(19)32319-0

Delaney, S., Jarem, D. A., Volle, C. B., & Yennie, C. J. (2012). Chemical and Biological Consequences of Oxidatively Damaged Guanine in DNA. *Free Radic Res*, *46*(4), 420–441. https://doi.org/10.3109/10715762.2011.653968

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., … Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Research*, *22*(9), 1775–1789. https://doi.org/10.1101/gr.132159.111

Díez-Villanueva, A., Mallona, I., & Peinado, M. A. (2015). Wanderer, an interactive viewer to explore DNA methylation and gene expression data in human cancer. *Epigenetics and Chromatin*, *8*(1), 1–8. https://doi.org/10.1186/s13072-015-0014-8

Dinger, M. E., Pang, K. C., Mercer, T. R., & Mattick, J. S. (n.d.). *Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities*. https://doi.org/10.1371/journal.pcbi.1000176

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., … Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, *489*(7414), 101–108. https://doi.org/10.1038/nature11233

Dong, H., Cao, W., & Xue, J. (2019). Long noncoding FOXD2-AS1 is activated by CREB1 and promotes cell proliferation and metastasis in glioma by sponging miR-185 through targeting AKT1. *Biochemical and Biophysical Research Communications*, *508*(4), 1074–1081. https://doi.org/10.1016/j.bbrc.2018.12.050

Dong, X. Y., Chen, C., Sun, X., Guo, P., Vessella, R. L., Wang, R. X., … Dong, J. T. (2006). FOXO1A is a candidate

for the 13q14 tumor suppressor gene inhibiting androgen receptor signaling in prostate cancer. *Cancer Research*, *66*(14), 6998–7006. https://doi.org/10.1158/0008-5472.CAN-06-0411

Dou, X., Zhou, Q., Wen, M., Xu, J., Zhu, Y., Zhang, S., & Xu, X. (2020). Long noncoding RNA FOXD2-AS1 promotes the malignancy of cervical cancer by sponging microRNA-760 and upregulating hepatoma-derived growth factor. *Frontiers in Pharmacology*, *10*(January), 1–13. https://doi.org/10.3389/fphar.2019.01700

Doubeni, C. A., Corley, D. A., Quinn, V. P., Jensen, C. D., Zauber, A. G., Goodman, M., … Fletcher, R. H. (2018). Effectiveness of screening colonoscopy in reducing the risk of death from right and left colon cancer: A large community-based study. *Gut*, *67*(2), 291–298. https://doi.org/10.1136/gutjnl-2016-312712

Du, Q., Luu, P. L., Stirzaker, C., & Clark, S. J. (2015). Methyl-CpG-binding domain proteins: Readers of the epigenome. *Epigenomics*, *7*(6), 1051–1073. https://doi.org/10.2217/epi.15.39

Dumbović, G., Biayna, J., Banús, J., Samuelsson, J., Roth, A., Diederichs, S., … Forcales, S. V. (2018). A novel long non-coding RNA from NBL2 pericentromeric macrosatellite forms a perinucleolar aggregate structure in colon cancer. *Nucleic Acids Research*, *46*(11), 5504–5524. https://doi.org/10.1093/nar/gky263

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Ellis, P., Moore, L., Sanders, M. A., Butler, T. M., Brunner, S. F., Lee-Six, H., … Campbell, P. J. (2021). Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat Protoc*, *16*(2), 841–871. https://doi.org/10.1038/s41596-020-00437-6

Erhardt, J. G., Kreichgauer, H. P., Meisner, C., Bode, J. C., Bode, C., & Contribution, O. (2002). Alcohol , cigarette smoking , dietary factors and the risk of colorectal adenomas and hyperplastic polyps – a case control study. *43*, 35–43. https://doi.org/10.1007/s003940200004

Ernstsson, S., Betz, R., Lagercrantz, S., Larsson, C., Ericksson, S., Cederberg, A., … Enerbäck, S. (1997). Cloning and characterization of freac-9 (FKHL17), a novel kidney- expressed human forkhead gene that maps to chromosome 1p32-p34. *Genomics*, *46*(1), 78–85. https://doi.org/10.1006/geno.1997.4986

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet*, *8*(4), 286–298. Retrieved from http://dx.doi.org/10.1038/nrg2005

Fearon, E. R. (2011). *Molecular Genetics of Colorectal Cancer*. https://doi.org/10.1146/annurev-pathol-011110-130235

Feinberg, A. P., Gehrke, C. W., Kuo, K. C., & Ehrlich, M. (1988). Reduced Genomic 5-Methylcytosine Content in Human Colonic Neoplasia. *Cancer Research*, *48*(5), 1159–1161.

Felsenfeld, G., & Groudine, M. (2003). Controlling the double helix. *Nature*, *421*(6921), 444–448. https://doi.org/10.1038/nature01410

Feng, D., Ye, X., Zhu, Z., Wei, Z., Cai, Q., & Wang, Y. (2015). Comparative transcriptome analysis between metastatic and non-metastatic gastric cancer reveals potential biomarkers. *Molecular Medicine Reports*, *11*(1), 386–392. https://doi.org/10.3892/mmr.2014.2709

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., … Flicek, P. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, *47*(D1), D766–D773. https://doi.org/10.1093/nar/gky955

Frederick L, G., Page, D. L., Fleming, I. D., Fritz, A. G., Balch, C. M., Haller, D. G., & Morrow, M. (2002). *AJCC*

Cancer Staging Manual (6th ed.). https://doi.org/https://doi.org/10.1007/978-1-4757-3656-4

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Gama-Sosal, M. A., Slagell, V. A., Trewyn, R. W., Oxenhandler, R., Kuo, K. C., Gehrke, C. W., & Ehrlich, M. (1983). The 5-methylcytosine content of DNA from human tumors Miguel. *Nucleic Acids Res.*, *11*(9), 71–84.

Gao, Y. F., Zhu, T., Mao, X. Y., Mao, C. X., Li, L., Yin, J. Y., … Liu, Z. Q. (2017). Silencing of Forkhead box D1 inhibits proliferation and migration in glioma cells. *Oncology Reports*, *37*(2), 1196–1202. https://doi.org/10.3892/or.2017.5344

Gawad, C., Koh, W., & Quake, S. R. (2016). Single-cell genome sequencing: Current state of the science. *Nature Reviews Genetics*, *17*(3), 175–188. https://doi.org/10.1038/nrg.2015.16

Gehart, H., & Clevers, H. (2019). Tales from the crypt: new insights into intestinal stem cells. *Nature Reviews Gastroenterology and Hepatology*, *16*(1), 19–34. https://doi.org/10.1038/s41575-018-0081-y

Gehrke, M. E. K. F. N. R. Y. W. K. C. K. C. W. (1986). DNA cytosine methylation and heat-induced deamination. *Biosci Rep*, *6*(4), 387–393. Retrieved from https://doi-org.sire.ub.edu/10.1007/BF01116426

Geissmann, Q. (2013). OpenCFU, a New Free and Open-Source Software to Count Cell Colonies and Other Circular Objects. *PLoS ONE*, *8*(2), 1–10. https://doi.org/10.1371/journal.pone.0054072

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S. C., Gonzalez, S., Rosebrock, D., … Consortium, P. (2020). The evolutionary history of 2,658 cancers. *Nature*, *578*(7793), 122–128. https://doi.org/10.1038/s41586-019-1907-7

GLOBOCAN. (2020). GLOBOCAN 2020. Retrieved from gco.iarc.fr

Goelz, S. E., Vogelstein, B., Hamilton, S. R., & Feinberg, A. P. (1985). Hypomethylation of DNA from benign and malignant human colon neoplasms. *Science*, *228*(4696), 187–190. https://doi.org/10.1126/science.2579435

Golson, M. L., & Kaestner, K. H. (2016). Fox transcription factors: from development to disease. *Development*, *143*(24), 4558–4570. https://doi.org/10.1242/dev.112672

Grady, W. M., & Pritchard, C. C. (2014). Molecular alterations and biomarkers in colorectal cancer. *Toxicologic Pathology*, *42*(1), 124–139. https://doi.org/10.1177/0192623313505155

Gu, N., Wang, X., Di, Z., Xiong, J., Ma, Y., Yan, Y., … Yu, J. (2019). Silencing lncRNA FOXD2-AS1 inhibits proliferation, migration, invasion and drug resistance of drug-resistant glioma cells and promotes their apoptosis via microRNA-98-5p/CPEB4 axis. *Aging*, *11*(22), 10266–10283. https://doi.org/10.18632/aging.102455

Hahn, M. A., Wu, X., Li, A. X., Hahn, T., & Pfeifer, G. P. (2011). Relationship between gene body DNA methylation and intragenic H3K9ME3 and H3K36ME3 chromatin marks. *PLoS ONE*, *6*(4). https://doi.org/10.1371/journal.pone.0018844

Hansen, K. D., Timp, W., Corrada Bravo, H., Sabunciyan, S., Langmead, B., McDonald, O. G., … Feinberg, A. P. (2011). Increased methylation variation in epigenetic domains across cancer types HHS Public Access Author manuscript. *Nat Genet*, *43*(8), 768–775. https://doi.org/10.1038/ng.865.Increased

Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., & Kjems, J. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature*, *495*(7441), 384–388. https://doi.org/10.1038/nature11993

Heather, H., Frankel, W. L., Martin, E., Arnold, M., Khanduja, K., Kuebler, P., … Panescu, J. (2015). Screening for the Lynch Syndrome (Hereditary Nonpolyposis Colorectal Cancer) Heather. *New England Journal of Medicine*, 687–696.

Herman, J. G., Lapidus, R. G., Issa, J. P. J., Davidson, N. E., Sidransky, D., Baylin, S. B., … Mao, L. (1995). Inactivation of the CDKN2/pl6/MTS1 Gene Is Frequently Associated with Aberrant DNA Methylation in All Common Human Cancers. *Cancer Research*, *55*(20), 4525–4530.

Hermans-Beijnsberger, S., van Bilsen, M., & Schroen, B. (2018). Long non-coding RNAs in the failing heart and vasculature. *Non-Coding RNA Research*, *3*(3), 118–130. https://doi.org/10.1016/j.ncrna.2018.04.002

Hibi, K., Sakata, M., Yokomizo, K., Kitamura, Y. H., Sakuraba, K., Shirahata, A., … Sanada, Y. (2009). Methylation of the MGMT gene is frequently detected in advanced gastric carcinoma. *Anticancer Research*, *29*(12), 5053–5055.

Hodges, C., Bintu, L., Lubkowska, L., Kashlev, M., & Bustamante, C. (2009). Nucleosomal fluctuations govern the transcription dynamics of RNA polymerase II. *Science*, *325*(5940), 626–628. https://doi.org/10.1126/science.1172926

Hon, C. C., Ramilowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J. L., Gough, J., … Forrest, A. R. R. (2017). An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature*, *543*(7644), 199–204. https://doi.org/10.1038/nature21374

Hu, Q., Tai, S., & Wang, J. (2019). Oncogenicity of lncRNA FOXD2-AS1 and its molecular mechanisms in human cancers. *Pathology Research and Practice*, *215*(5), 843–848. https://doi.org/10.1016/j.prp.2019.01.033

Hu, X., Sood, A. K., Dang, C. V., & Zhang, L. (2018). The role of long noncoding RNAs in cancer: the dark matter matters. *Current Opinion in Genetics and Development*, *48*, 8–15. https://doi.org/10.1016/j.gde.2017.10.004

Hu, Z., Huang, L., Wang, W., Guan, C., Zhao, Y., Liu, L., & Jiang, X. (2021). Long Non-coding RNA FOXD2-AS1 Promotes Proliferation, Migration, and Invasion in Cholangiocarcinoma Through Regulating miR-760/E2F3 Axis. *Digestive Diseases and Sciences*. https://doi.org/10.1007/s10620-021-06876-9

Humphries, A., & Wright, N. A. (2008). Colonic crypt organization and tumorigenesis. *Nature Reviews Cancer*, *8*(6), 415–424. https://doi.org/10.1038/nrc2392

Iacopetta, B., Russo, A., Bazan, V., Dardanoni, G., Gebbia, N., Soussi, T., … Ishioka, C. (2006). *Functional categories of TP53 mutation in colorectal cancer : results of an International Collaborative Study*. (March), 842–847. https://doi.org/10.1093/annonc/mdl035

Jager, M., Blokzijl, F., Sasselli, V., Boymans, S., Janssen, R., Besselink, N., … Cuppen, E. (2018). Measuring mutation accumulation in single human adult stem cells by whole-genome sequencing of organoid cultures. *Nature Protocols*, *13*(1), 59–78. https://doi.org/10.1038/nprot.2017.111

Jangid, R. K., Kelkar, A., Muley, V. Y., & Galande, S. (2018). Bidirectional promoters exhibit characteristic chromatin modification signature associated with transcription elongation in both sense and antisense directions. *BMC Genomics*, *19*(1), 1–20. https://doi.org/10.1186/s12864-018-4697-7

Jenuwein, T., & Allis, C. D. (2001). Translating the histone code. *Science*, *293*(5532), 1074–1080. https://doi.org/10.1126/science.1063127

Jiang, M., Qiu, N., Xia, H., Liang, H., Li, H., & Ao, X. (2019). Long non-coding RNA FOXD2-AS1/miR-150-5p/PFN2 axis regulates breast cancer malignancy and tumorigenesis. *International Journal of Oncology*, *54*(3), 1043–1052. https://doi.org/10.3892/ijo.2019.4671

Johansson, C. C., Dahle, M. K., Blomqvist, S. R., Grønning, L. M., Aandahl, E. M., Enerbäck, S., & Taskén, K. (2003). A winged helix forkhead (FOXD2) tunes sensitivity to cAMP in T lymphocytes through regulation of cAMP-dependent protein kinase RIα. *Journal of Biological Chemistry*, *278*(19), 17573–17579. https://doi.org/10.1074/jbc.M300311200

Johnson, I. T., & Lund, E. K. (2007). Review article: Nutrition, obesity and colorectal cancer. *Alimentary Pharmacology and Therapeutics*, *26*(2), 161–181. https://doi.org/10.1111/j.1365-2036.2007.03371.x

Jones, P. A. (1999). The DNA methylation paradox. *Trends in Genetics*, *15*(1), 34–37. https://doi.org/10.1016/S0168-9525(98)01636-9

Jones, P. A. (2012). Functions of DNA methylation : islands , start sites , gene bodies and beyond. *Nature Reviews Genetics*, *13*(July), 484–492. https://doi.org/10.1038/nrg3230

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., … Wahlestedt, C. (2005). Antisense transcription in the mammalian transcriptome. *Science*, *309*(5740), 1564–1566. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16141073

Kennedy, S. R., Schmitt, M. W., Fox, E. J., Kohrn, B. F., Salk, J. J., Ahn, E. H., … Loeb, L. a. (2014). Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc*, *9*(11), 2586–2606. https://doi.org/10.1038/nprot.2014.170

Kennedy, S. R., Zhang, Y., & Risques, R. A. (2019). Cancer-Associated Mutations but No Cancer: Insights into the Early Steps of Carcinogenesis and Implications for Early Cancer Detection. *Trends in Cancer*, *5*(9), 531–540. https://doi.org/10.1016/j.trecan.2019.07.007

Kikendall, J. W., Bowen, P. E., Burgess, M. B., Magnetti, C., & Woodward, J. (1989). *Cigarettes and Alcohol as Independent Risk Factors for Colonic Adenomas*.

Kim, J. E., Choi, J., Sung, C. O., Hong, Y. S., Kim, S. Y., Lee, H., … Kim, J. I. (2021). High prevalence of *TP53* loss and whole-genome doubling in early-onset colorectal cancer. *Exp Mol Med*, *53*(3), 446–456. https://doi.org/10.1038/s12276-021-00583-1

Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., … Zhang, F. (2015). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, *517*(7536), 583–588. https://doi.org/10.1038/nature14136

Krimmel-Morrison, J. D., Ghezelayagh, T. S., Lian, S., Zhang, Y., Fredrickson, J., Nachmanson, D., … Risques, R. A. (2019). Characterization of *TP53* mutations in Pap test DNA of women with and without serous ovarian carcinoma. *Gynecol Oncol*. https://doi.org/10.1016/j.ygyno.2019.11.124

Krimmel, J. D., Schmitt, M. W., Harrell, M. I., Agnew, K. J., Kennedy, S. R., Emond, M. J., … Risques, R. A. (2016). Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic *TP53* mutations in noncancerous tissues. *Proc Natl Acad Sci U S A*, *113*(21), 6005–6010. https://doi.org/10.1073/pnas.1601311113

Kucab, J. E., Zou, X., Morganella, S., Joel, M., Nanda, A. S., Nagy, E., … Nik-Zainal, S. (2019). A Compendium of Mutational Signatures of Environmental Agents. *Cell*, *177*(4), 821–836 e16. https://doi.org/10.1016/j.cell.2019.03.001

Kuipers, E. J., Grady, W. M., Lieberman, D., Seufferlein, T., Sung, J. J., Boelens, P. G., … Watanabe, T. (2015). Colorectal cancer. *Nature Reviews Disease Primers*, *1*, 1–25. https://doi.org/10.1038/nrdp.2015.65

Kume, T., Deng, K., & Hogan, B. L. M. (2000). Minimal Phenotype of Mice Homozygous for a Null Mutation in the Forkhead/Winged Helix Gene, Mf2 . *Molecular and Cellular Biology*, *20*(4), 1419–1425. https://doi.org/10.1128/mcb.20.4.1419-1425.2000

Lao, V. V., & Grady, W. M. (2011). Epigenetics and colorectal cancer. *Nature Reviews Gastroenterology and Hepatology*, *8*(12), 686–700. https://doi.org/10.1038/nrgastro.2011.173

Laoukili, J., Stahl, M., & Medema, R. H. (2007). FoxM1: At the crossroads of ageing and cancer. *Biochimica et Biophysica Acta - Reviews on Cancer*, *1775*(1), 92–102. https://doi.org/10.1016/j.bbcan.2006.08.006

Lattery, M. L. S. (2002). *PROGNOSTIC SIGNIFICANCE OF P53 MUTATIONS IN COLON CANCER AT THE POPULATION LEVEL*. *602*, 597–602. https://doi.org/10.1002/ijc.10405

Laurent, G. S., Wahlestedt, C., & Kapranov, P. (2016). *The Landscape of long non-coding RNA classification The non-coding RNA universe*. *31*(5), 239–251. https://doi.org/10.1016/j.tig.2015.03.007.The

Lawson, A. R. J., Abascal, F., Coorens, T. H. H., Hooks, Y., O'Neill, L., Latimer, C., … Martincorena, I. (2020). Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science*, *370*(6512), 75–82. https://doi.org/10.1126/science.aba8347

Lee-Six, H., Øbro, N. F., Shepherd, M. S., Grossmann, S., Dawson, K., Belmonte, M., … Campbell, P. J. (2018). Population dynamics of normal human blood inferred from somatic mutations. *Nature*, *561*(7724), 473–478. https://doi.org/10.1038/s41586-018-0497-0

Lee-Six, H., Olafsson, S., Ellis, P., Osborne, R. J., Sanders, M. A., Moore, L., … Stratton, M. R. (2019). The landscape of somatic mutation in normal colorectal epithelial cells. *Nature*, *574*(7779), 532–537. https://doi.org/10.1038/s41586-019-1672-7

Leroy, B., Ballinger, M. L., Baran-Marszak, F., Bond, G. L., Braithwaite, A., Concin, N., … Soussi, T. (2017). Recommended guidelines for validation, quality control, and reporting of *TP53* variants in clinical practice. *Cancer Research*, *77*(6), 1250–1260. https://doi.org/10.1158/0008-5472.CAN-16-2179

Li, C. Y., Liang, G. Y., Yao, W. Z., Sui, J., Shen, X., Zhang, Y. Q., … Pu, Y. P. (2016). Integrated analysis of long non-coding RNA competing interactions reveals the potential role in progression of human gastric cancer. *International Journal of Oncology*, *48*(5), 1965–1976. https://doi.org/10.3892/ijo.2016.3407

Li, H., Han, Q., Chen, Y., Chen, X., Ma, R., Chang, Q., & Yin, D. (2019). *Upregulation of the long non-coding RNA* FOXD2-AS1 *is correlated with tumor progression and metastasis in papillary thyroid cancer*. *11*(9), 5457–5471.

Li, Q., Liu, X., Gu, J., Zhu, J., Wei, Z., & Huang, H. (2020). Screening lncRNAs with diagnostic and prognostic value for human stomach adenocarcinoma based on machine learning and mRNA-lncRNA co-expression network analysis. *Molecular Genetics and Genomic Medicine*, *8*(11), 1–14. https://doi.org/10.1002/mgg3.1512

Li, R., Du, Y., Chen, Z., Xu, D., Lin, T., Jin, S., … Bai, F. (2020). Macroscopic somatic clonal expansion in morphologically normal human urothelium. *Science*, *370*(6512), 82–89. https://doi.org/10.1126/science.aba7300

Li, X., Yang, L., Fan, Q., & Wang, Y. (2020). *Preliminary Study on the Identification of BRAF V600E Mutation in Colorectal Cancer by Near-Infrared Spectroscopy*. 13077–13085.

Liang, P. S., Chen, T., & Giovannucci, E. (2009). *Cigarette smoking and colorectal cancer incidence and mortality : Systematic review and meta-analysis*. *2415*, 2406–2415. https://doi.org/10.1002/ijc.24191

Liang, T. J., Wang, H. X., Zheng, Y. Y., Cao, Y. Q., Wu, X., Zhou, X., & Dong, S. X. (2017). APC hypermethylation

for early diagnosis of colorectal cancer: A meta-analysis and literature review. *Oncotarget*, *8*(28), 46468–46479. https://doi.org/10.18632/oncotarget.17576

Liebl, M. C., & Hofmann, T. G. (2021). The Role of p53 Signaling in Colorectal Cancer. *Cancers*, *13*(9), 2125. https://doi.org/10.3390/cancers13092125

Lin, J. C., Jeong, S., Liang, G., Takai, D., Fatemi, M., Tsai, Y. C., … Jones, P. A. (2007). Role of Nucleosomal Occupancy in the Epigenetic Silencing of the *MLH1* CpG Island. *Cancer Cell*, *12*(5), 432–444. https://doi.org/10.1016/j.ccr.2007.10.014

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., … Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, *462*(7271), 315–322. https://doi.org/10.1038/nature08514

Liu, X., Fu, Q., Li, S., Liang, N., Li, F., Li, C., … Sun, H. (2019). LncRNA FOXD2-AS1 functions as a competing endogenous RNA to regulate TERT expression by sponging miR-7-5p in thyroid cancer. *Frontiers in Endocrinology*, *10*(APR), 1–12. https://doi.org/10.3389/fendo.2019.00207

Loeb, L. A., Loeb, K. R., & Anderson, J. P. (2003). Multiple mutations and cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(3), 776–781. https://doi.org/10.1073/pnas.0334858100

Lorincz, M. C., Dickerson, D. R., Schmitt, M., & Groudine, M. (2004). Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nature Structural and Molecular Biology*, *11*(11), 1068–1075. https://doi.org/10.1038/nsmb840

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. https://doi.org/10.1186/s13059-014-0550-8

Lu, W., Xu, Y., Xu, J., Wang, Z., & Ye, G. (2018). Identification of differential expressed lncRNAs in human thyroid cancer by a genome-wide analyses. *Cancer Medicine*, *7*(8), 3935–3944. https://doi.org/10.1002/cam4.1627

M.Gardiner-Garden, & M.Frommer. (1987). CpG Islands in vertebrate genomes. *Journal of Molecular Biology*, *196*(2), 261–282. https://doi.org/https://doi.org/10.1016/0022-2836(87)90689-9

Maeda, K., Kawakami, K., Ishida, Y., Ishiguro, K., Omura, K., & Watanabe, G. (2003). Hypermethylation of the CDKN2A gene in colorectal cancer is associated with shorter survival. *Oncology Reports*, 935–938. https://doi.org/https://doi.org/10.3892/or.10.4.935

Mallona, I., Aussó, S., Díez-Villanueva, A., Moreno, V., & Peinado, M. (2018). Modular dynamics of DNA co-methylation networks exposes the functional organization of colon cancer cells' genome. *BioRxiv*, 428730. https://doi.org/10.1101/428730

Mao, R., Wang, Z., Zhang, Y., Chen, Y. Y., Liu, Q., Zhang, T., & Liu, Y. (2020). Development and validation of a novel prognostic signature in gastric adenocarcinoma. *Aging*, *12*(21), 22233–22252. https://doi.org/10.18632/aging.104161

Martincorena, I., Roshan, A., Gerstung, M., Ellis, P., Van Loo, P., McLaren, S., … Campbell, P. J. (2015). Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, *348*(6237), 880–886. https://doi.org/10.1126/science.aaa6806

Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., Dsouza, C., Fouse, S. D., … Costello, J. F. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, *466*(7303), 253–257. https://doi.org/10.1038/nature09165

Moore, L. D., Le, T., & Fan, G. (2012). DNA Methylation and Its Basic Function. *Neuropsychopharmacology*, *38*(1), 23–38. https://doi.org/10.1038/npp.2012.112

Nachmanson, D., Lian, S., Schmidt, E. K., Hipp, M. J., Baker, K. T., Zhang, Y., … Risques, R. A. (2018). Targeted genome fragmentation with CRISPR/Cas9 enables fast and efficient enrichment of small genomic regions and ultra-accurate sequencing with low DNA input (CRISPR-DS). *Genome Research*, *28*(10), 1589–1599. https://doi.org/10.1101/gr.235291.118

Nagy, Á., & Győrffy, B. (2021). muTarget: A platform linking gene expression changes and mutation status in solid tumors. *International Journal of Cancer*, *148*(2), 502–511. https://doi.org/10.1002/ijc.33283

Namasivayam, V., & Lim, S. (2017). *Recent advances in the link between physical activity , sedentary behavior , physical fitness , and colorectal cancer*. *6*(0), 1–8. https://doi.org/10.12688/f1000research.9795.1

Nassar, D., & Blanpain, C. (2016). Cancer Stem Cells: Basic Concepts and Therapeutic Implications. *Annual Review of Pathology: Mechanisms of Disease*, *11*, 47–76. https://doi.org/10.1146/annurev-pathol-012615-044438

Nicholson, A. M., Olpe, C., Hoyle, A., Thorsen, A. S., Rus, T., Colombe, M., … Winton, D. J. (2018). Fixation and Spread of Somatic Mutations in Adult Human Colonic Epithelium. *Cell Stem Cell*, *22*(6), 909–918 e8. https://doi.org/10.1016/j.stem.2018.04.020

Nik-Zainal, S., Kucab, J. E., Morganella, S., Glodzik, D., Alexandrov, L. B., Arlt, V. M., … Phillips, D. H. (2015). The genome as a record of environmental exposure. *Mutagenesis*, *30*(6), 763–770. https://doi.org/10.1093/mutage/gev073

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Okugawa, Y., Grady, W. M., & Goel, A. (2017). Gastroenterology. *Physiology & Behavior*, *176*(3), 139–148. https://doi.org/10.1053/j.gastro.2015.07.011.Epigenetic

Ortmann, C. A., Kent, D. G., Nangalia, J., Silber, Y., Wedge, D. C., Grinfeld, J., … Green, A. R. (2015). Effect of Mutation Order on Myeloproliferative Neoplasms. *New England Journal of Medicine*, *372*(7), 601–612. https://doi.org/10.1056/nejmoa1412098

Pan, F., Li, M., & Chen, W. (2018). FOXD1 predicts prognosis of colorectal cancer patients and promotes colorectal cancer progression via the ERK 1/2 pathway. *American Journal of Translational Research*, *10*(5), 1522–1530.

Park, S., Wilkens, L. R., Setiawan, V. W., Monroe, K. R., Haiman, A., & Marchand, L. Le. (2019). *Original Contribution Alcohol Intake and Colorectal Cancer Risk in the Multiethnic Cohort Study*. *188*(1), 67–76. https://doi.org/10.1093/aje/kwy208

Pino, M. S., & Chung, D. C. (2010). THE CHROMOSOMAL INSTABILITY PATHWAY IN COLON. *Gastroenterology; 138(6): 2059–2072.*, *138*(6), 2059–2072. https://doi.org/10.1053/j.gastro.2009.12.065.THE

Pitolli, C., Wang, Y., Mancini, M., Shi, Y., Melino, G., & Amelio, I. (2019). Do mutations turn p53 into an oncogene? *International Journal of Molecular Sciences*, *20*(24). https://doi.org/10.3390/ijms20246241

Pontier, D. B., & Gribnau, J. (2011). Xist regulation and function eXplored. *Human Genetics*, *130*(2), 223–236. https://doi.org/10.1007/s00439-011-1008-7

Potten, C. S. (1998). Stem cells in gastrointestinal epithelium: Numbers, characteristics and death. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *353*(1370), 821–830.

https://doi.org/10.1098/rstb.1998.0246

Potten, C. S., Kellett, M., Roberts, S. A., Rew, D. A., & Wilson, G. D. (1992). Measurement of in vivo proliferation in human colorectal mucosa using bromodeoxyuridine. *Gut*, *33*(1), 71–78. https://doi.org/10.1136/gut.33.1.71

Poulos, R. C., Olivier, J., & Wong, J. W. H. (2017). CpG methylation accounts for genome-wide C>T mutation variation and cancer driver formation across cancer types. *BioRxiv*, 1–49. https://doi.org/10.1101/106872

Qasim, B. J., Al-Wasiti, E. A., & Azzal, H. S. (2016). Association of Global DNA Hypomethylation with Clinicopathological Variables in Colonic Tumors of Iraqi Patients. *Saudi Journal of Gastroenterology*, *22*(2), 139–147. Retrieved from 10.4103/1319-3767.178525

Quinn, J. J., & Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. *Nature Reviews Genetics*, *17*(1), 47–62. https://doi.org/10.1038/nrg.2015.10

Radtke, F., & Clevers, H. (2005). Self-renewal and cancer of the gut: Two sides of a coin. *Science*, *307*(5717), 1904–1909. https://doi.org/10.1126/science.1104815

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., & Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. *Nature Protocols*, *8*(11), 2281–2308. https://doi.org/10.1038/nprot.2013.143

Rasmussen, K. D., & Helin, K. (2016). Role of TET enzymes in DNA methylation, development, and cancer. *Genes and Development*, *30*(7), 733–750. https://doi.org/10.1101/gad.276568.115

Reik, W., & Lewis, A. (2005). Co-evolution of X-chromosome inactivation and imprinting in mammals. *Nature Reviews Genetics*, *6*(5), 403–410. https://doi.org/10.1038/nrg1602

Ren, W., Zhu, Z., & Wu, L. (2019). FOXD2-AS1 correlates with the malignant status and regulates cell proliferation, migration, and invasion in cutaneous melanoma. *Journal of Cellular Biochemistry*, *120*(4), 5417–5423. https://doi.org/10.1002/jcb.27820

Ren, Z., Hu, Y., Li, G., Kang, Y., Liu, Y., & Zhao, H. (2019). HIF-1α induced long noncoding RNA FOXD2-AS1 promotes the osteosarcoma through repressing p21. *Biomedicine and Pharmacotherapy*, *117*(24), 109104. https://doi.org/10.1016/j.biopha.2019.109104

Rinn, J. L., & Chang, H. Y. (2012). Genome Regulation by Long Noncoding RNAs. *Annual Review of Biochemistry*, *81*(1), 145–166. https://doi.org/10.1146/annurev-biochem-051410-092902

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., … Chang, H. Y. (2007). *Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs*. 1311–1323. https://doi.org/10.1016/j.cell.2007.05.022

Risques, R. A., & Kennedy, S. R. (2018). Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLoS Genet*, *14*(1), e1007108. https://doi.org/10.1371/journal.pgen.1007108

Rong, L., Zhao, R., & Lu, J. (2017). Highly expressed long non-coding RNA FOXD2-AS1 promotes non-small cell lung cancer progression via Wnt/β-catenin signaling. *Biochemical and Biophysical Research Communications*, *484*(3), 586–591. https://doi.org/10.1016/j.bbrc.2017.01.141

Salk, J. J., Loubet-Senear, K., Maritschnegg, E., Valentine, C. C., Williams, L. N., Higgins, J. E., … Risques, R. A. (2019). Ultra-Sensitive *TP53* Sequencing for Cancer Detection Reveals Progressive Clonal Selection in Normal Tissue over a Century of Human Lifespan. *Cell Rep*, *28*(1), 132–144 e3. https://doi.org/10.1016/j.celrep.2019.05.109

Salk, J. J., Loubet-Senear, K., Maritschnegg, E., Valentine, C. C., Williams, L. N., Higgins, J. E., … Risques, R. A. (2019). Ultra-Sensitive *TP53* Sequencing for Cancer Detection Reveals Progressive Clonal Selection in Normal Tissue over a Century of Human Lifespan. *Cell Reports*, *28*(1), 132–144.e3. https://doi.org/10.1016/j.celrep.2019.05.109

Samowitz, W. S., Curtin, K., Lin, H. H., Robertson, M. A., Schaffer, D., Nichols, M., … Slattery, M. L. (2001). The colon cancer burden of genetically defined hereditary nonpolyposis colon cancer. *Gastroenterology*, *121*(4), 830–838. https://doi.org/10.1053/gast.2001.27996

Saunders, A., Core, L. J., & Lis, J. T. (2006). Breaking barriers to transcription elongation. *Nature Reviews Molecular Cell Biology*, *7*(8), 557–567. https://doi.org/10.1038/nrm1981

Schmitt, M. W., Fox, E. J., Prindle, M. J., Reid-Bayliss, K. S., True, L. D., Radich, J. P., & Loeb, L. A. (2015). Sequencing small genomic targets with high efficiency and extreme accuracy. *Nature Methods*, *12*(5), 423–425. https://doi.org/10.1038/nmeth.3351

Schmitt, M. W., Kennedy, S. R., Salk, J. J., Fox, E. J., Hiatt, J. B., & Loeb, L. A. (2012). Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A*, *109*(36), 14508–14513. https://doi.org/10.1073/pnas.1208715109

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, *9*(7), 671–675. https://doi.org/10.1038/nmeth.2089

Schneider, R., Bannister, A. J., Myers, F. A., Thorne, A. W., Crane-Robinson, C., & Kouzarides, T. (2004). Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nature Cell Biology*, *6*(1), 73–77. https://doi.org/10.1038/ncb1076

Shann, Y. J., Cheng, C., Chiao, C. H., Chen, D. T., Li, P. H., & Hsu, M. T. (2008). Genome-wide mapping and characterization of hypomethylated sites in human tissues and breast cancer cell lines. *Genome Research*, *18*(5), 791–801. https://doi.org/10.1101/gr.070961.107

Sheikhnejad, G., Brank, A., Christman, J. K., Goddard, A., Alvarez, E., Ford, H., … Cheng, X. (1999). Mechanism of inhibition of DNA (cytosine C5)-methyltransferases by oligodeoxyribonucleotides containing 5,6-dihydro-5-azacytosine. *Journal of Molecular Biology*, *285*(5), 2021–2034. https://doi.org/10.1006/jmbi.1998.2426

Shia, J., Klimstra, D. S., Bagci, P., Basturk, O., & Adsay, N. V. (2012). TNM staging of colorectal carinoma: issues and caveats. *Semin Diagn Pathology*, *29*(3), 142–153.

Short, N. J., Kantarjian, H., Kanagal-Shamanna, R., Sasaki, K., Ravandi, F., Cortes, J., … Jabbour, E. (2020). Ultra-accurate Duplex Sequencing for the assessment of pretreatment ABL1 kinase domain mutations in Ph+ ALL. *Blood Cancer Journal*, *10*(5). https://doi.org/10.1038/s41408-020-0329-y

Shu, J., Jelinek, J., Chang, H., Shen, L., Qin, T., Chung, W., … Issa, J. P. J. (2006). Silencing of bidirectional promoters by DNA methylation in tumorigenesis. *Cancer Research*, *66*(10), 5077–5084. https://doi.org/10.1158/0008-5472.CAN-05-2629

Siddiqui, H., Al-Ghafari, A., Choudhry, H., & Al Doghaither, H. (2019). Roles of long non-coding RNAs in colorectal cancer tumorigenesis: A review. *Molecular and Clinical Oncology*, *11*(2), 167–172. https://doi.org/10.3892/mco.2019.1872

Siegel, R. L., Fedewa, S. A., Anderson, W. F., Miller, K. D., Ma, J., Rosenberg, P. S., & Jemal, A. (2017). Colorectal Cancer Incidence Patterns in the United States, 1974-2013. *J Natl Cancer Inst*, *109*(8). https://doi.org/10.1093/jnci/djw322

Siegel, R. L., Miller, K. D., Goding Sauer, A., Fedewa, S. A., Butterly, L. F., Anderson, J. C., … Jemal, A. (2020). Colorectal cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, *70*(3), 145–164. https://doi.org/10.3322/caac.21601

Siegel, R. L., Miller, K. D., & Jemal, A. (2017). *Cancer Statistics , 2017*. *67*(1), 7–30. https://doi.org/10.3322/caac.21387.

Sigova, A. A., Mullen, A. C., Molinie, B., Gupta, S., Orlando, D. A., Guenther, M. G., … Young, R. A. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(8), 2876–2881. https://doi.org/10.1073/pnas.1221904110

Slack, F., & Chinnaiyan, A. (2019). The Role of Non-coding RNAs in Oncology. *Cell*, *179*(5). https://doi.org/10.1016/j.cell.2019.10.017

Song, M., Emilsson, L., Bozorg, S. R., Nguyen, L. H., Joshi, A. D., Staller, K., … Ludvigsson, J. F. (2020). Risk of colorectal cancer incidence and mortality after polypectomy: a Swedish record-linkage study. *The Lancet Gastroenterology and Hepatology*, *5*(6), 537–547. https://doi.org/10.1016/S2468-1253(20)30009-1

Su, F., He, W., Chen, C., Liu, M., Liu, H., Xue, F., … Jiang, C. (2018). The long non-coding RNA FOXD2-AS1 promotes bladder cancer progression and recurrence through a positive feedback loop with Akt and E2F1 article. *Cell Death and Disease*, *9*(2). https://doi.org/10.1038/s41419-018-0275-9

Sulman, E. P., White, P. S., & Brodeur, G. M. (2004). Genomic annotation of the meningioma tumor suppressor locus on chromosome 1p34. *Oncogene*, *23*(4), 1014–1020. https://doi.org/10.1038/sj.onc.1206623

Suzuki, H., Watkins, D. N., Jair, K. W., Schuebel, K. E., Markowitz, S. D., Chen, W. D., … Baylin, S. B. (2004). Epigenetic inactivation of SFRP genes allows constitutive WNT signaling in colorectal cancer. *Nature Genetics*, *36*(4), 417–422. https://doi.org/10.1038/ng1330

Tan-Wong, S. M., Zaugg, J. B., Camblong, J., Xu, Z., Zhang, D. W., Mischo, H. E., … Proudfoot, N. J. (2012). GENE LOOPS ENHANCE TRANSCRIPTIONAL DIRECTIONALITY. *Science*, *338*, 671–675. https://doi.org/10.1126/science.1224350

Tang, J., Fewings, E., Chang, D., Zeng, H., Liu, S., Jorapur, A., … Shain, A. H. (2020). The genomic landscapes of individual melanocytes from human skin. *Nature*, *586*(7830), 600–605. https://doi.org/10.1038/s41586-020-2785-8

Tang, Y., Shu, G., Yuan, X., Jing, N., & Song, J. (2011). FOXA2 functions as a suppressor of tumor metastasis by inhibition of epithelial-to-mesenchymal transition in human lung cancers. *Cell Research*, *21*(2), 316–326. https://doi.org/10.1038/cr.2010.126

Tang, Z., Li, C., Kang, B., Gao, G., Li, C., & Zhang, Z. (2017). GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Research*, *45*(W1), W98–W102. https://doi.org/10.1093/nar/gkx247

Tao, Y., Kang, B., Petkovich, D. A., Bhandari, Y. R., In, J., Stein-O'Brien, G., … Easwaran, H. (2019). Aging-like Spontaneous Epigenetic Silencing Facilitates Wnt Activation, Stemness, and *BRAF* V600E -Induced Tumorigenesis. *Cancer Cell*, *35*(2), 315–328.e6. https://doi.org/10.1016/j.ccell.2019.01.005

Tariq, K., & Ghias, K. (2016). Colorectal cancer carcinogenesis: a review of mechanisms. *Cancer Biology and Medicine*, *13*(1), 120–135. https://doi.org/10.28092/j.issn.2095-3941.2015.0103

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., … Forbes, S. A. (2019). COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(D1), D941–D947.

https://doi.org/10.1093/nar/gky1015

Tichon, A., Gil, N., Lubelsky, Y., Solomon, T. H., Lemze, D., Itzkovitz, S., … Ulitsky, I. (2016). A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells. *Nature Communications*, 1–10. https://doi.org/10.1038/ncomms12209

Tikkanen, T., Leroy, B., Fournier, J. L., Risques, R. A., Malcikova, J., & Soussi, T. (2018). Seshat: A Web service for accurate annotation, validation, and analysis of *TP53* variants generated by conventional and next-generation sequencing. *Hum Mutat*. https://doi.org/10.1002/humu.23543

Trinklein, N. D., Force Aldred, S., Hartman, S. J., Schroeder, D. I., Otillar, R. P., & Myers, R. M. (2004). An abundance of bidirectional promoters in the human genome. *Genome Research*, *14*(1), 62–66. https://doi.org/10.1101/gr.1982804

Tsang, W. P., Ng, E. K. O., Ng, S. S. M., Jin, H., Yu, J., Sung, J. J. Y., & Kwok, T. T. (2010). Oncofetal H19-derived miR-675 regulates tumor suppressor RB in human colorectal cancer. *Carcinogenesis*, *31*(3), 350–358. https://doi.org/10.1093/carcin/bgp181

Turck, N., Vutskits, L., Sanchez-Pena, P., Robin, X., Hainard, A., Gex-Fabry, M., … Sanchez, J.-C. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *8*, 12–77. Retrieved from http://link.springer.com/10.1007/s00134-009-1641-y

Tuteja, G., & Kaestner, K. H. (2007). SnapShot:Forkhead Transcription Factors I. *Cell*, *130*(6), 1160.e1-1160.e2. https://doi.org/10.1016/j.cell.2007.09.005

Ulitsky, I., & Bartel, D. P. (2013). Ulitsky, Igor, and David P Bartel. 2013. “lincRNAs: Genomics, Evolution, and Mechanisms.” Cell 154 (1): 26–46. doi:10.1016/j.cell.2013.06.020.lincRNAs: genomics, evolution, and mechanisms. *Cell*, *154*(1), 26–46. https://doi.org/10.1016/j.cell.2013.06.020.lincRNAs

Van Der Heul-Nieuwenhuijsen, L., Dits, N. F., & Jenster, G. (2009). Gene expression of forkhead transcription factors in the normal and diseased human prostate. *BJU International*, *103*(11), 1574–1580. https://doi.org/10.1111/j.1464-410X.2009.08351.x

Vogelstein, B., Fearon, E., Hamilton, S., Kern, S., Preisinger, A., Leppert, M., … JL, B. (1988). Genetic alterations during colorectal-tumor development. *The New England Journal of Medicine*, *319*(9), 525–532. https://doi.org/10.1056/NEJM198809013190901

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer Genome Landscapes. *Science*, *339*(6127), 1546–1558. https://doi.org/10.1126/SCIENCE.1235122

Volders, P. J., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdagh, P., & Vandesompele, J. (2019). Lncipedia 5: Towards a reference set of human long non-coding rnas. *Nucleic Acids Research*, *47*(D1), D135–D139. https://doi.org/10.1093/nar/gky1031

Wang, J., Li, B., Wang, C., Luo, Y., Zhao, M., & Chen, P. (2019). Long noncoding RNA FOXD2-AS1 promotes glioma cell cycle progression and proliferation through the FOXD2-AS1/miR-31/CDK1 pathway. *Journal of Cellular Biochemistry*, (October 2018), 1–12. https://doi.org/10.1002/jcb.29284

Wang, J., Zhu, C. P., Hu, P. F., Qian, H., Ning, B. F., Zhang, Q., … Xie, W. F. (2014). FOXA2 suppresses the metastasis of hepatocellular carcinoma partially through matrix metalloproteinase-9 inhibition. *Carcinogenesis*, *35*(11), 2576–2583. https://doi.org/10.1093/carcin/bgu180

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J.-P., & Li, W. (n.d.). *CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model*. https://doi.org/10.1093/nar/gkt006

Wang, T., Maden, S. K., Luebeck, G. E., Li, C. I., Newcomb, P. A., Ulrich, C. M., … Grady, W. M. (2020).

Dysfunctional epigenetic aging of the normal colon and colorectal cancer risk. *Clin Epigenetics*, *12*(1), 5. https://doi.org/10.1186/s13148-019-0801-3

Wang, Y., Cao, L., Wang, Q., Huang, J., & Xu, S. (2019). LncRNA FOXD2-AS1 induces chondrocyte proliferation through sponging miR-27a-3p in osteoarthritis. *Artificial Cells, Nanomedicine and Biotechnology*, *47*(1), 1241–1247. https://doi.org/10.1080/21691401.2019.1596940

Weber, M., Davies, J. J., Wittig, D., Oakele, E. J., Haase, M., Lam, W. L., & Schübeler, D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genet.*, *37*, 853–862. Retrieved from http://dx.doi.org/10.1038/ng1598

Weigel, D., Jürgens, G., Küttner, F., Seifert, E., & Jäckle, H. (1989). The homeotic gene fork head encodes a nuclear protein and is expressed in the terminal regions of the Drosophila embryo. *Cell*, *57*(4), 645–658. https://doi.org/10.1016/0092-8674(89)90133-5

Xu, T. peng, Wang, W. yu, Ma, P., Shuai, Y., Zhao, K., Wang, Y. fen, … Shu, Y. qian. (2018). Upregulation of the long noncoding RNA FOXD2-AS1 promotes carcinogenesis by epigenetically silencing EphB3 through EZH2 and LSD1, and predicts poor prognosis in gastric cancer. *Oncogene*, *37*(36), 5020–5036. https://doi.org/10.1038/s41388-018-0308-y

Yadav, V. K., Degregori, J., & De, S. (2016). The landscape of somatic mutations in protein coding genes in apparently benign human tissues carries signatures of relaxed purifying selection. *Nucleic Acids Research*, *44*(5), 2075–2084. https://doi.org/10.1093/nar/gkw086

Yang, M. Q., & Elnitski, L. L. (2008). Diversity of core promoter elements comprising human bidirectional promoters. *BMC Genomics*, *9*(SUPPL. 2), 1–8. https://doi.org/10.1186/1471-2164-9-S2-S3

Yang, X., Duan, B., & Zhou, X. (2017). Long non-coding RNA FOXD2-AS1 functions as a tumor promoter in colorectal cancer by regulating EMT and Notch signaling pathway. *European Review for Medical and Pharmacological Sciences*, *21*(16), 3586–3591.

Yang, X., Han, H., DeCarvalho, D. D., Lay, F. D., Jones, P. A., & Liang, G. (2014). Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell*, *26*(4), 577–590. https://doi.org/10.1016/j.ccr.2014.07.028

Ye, J., Liu, J., Tang, T., Xin, L., Bao, X., & Yan, Y. (2021). miR‑4306 inhibits the malignant behaviors of colorectal cancer by regulating lncRNA FOXD2‑AS1. *Molecular Medicine Reports*, *24*(4), 1–10. https://doi.org/10.3892/mmr.2021.12362

Ye, L. C., Zhu, D. X., Qiu, J. J., Xu, J., & Wei, Y. (2015). Involvement of long non-coding RNA in colorectal cancer: From benchtop to bedside (Review). *Oncology Letters*, *9*(3), 1039–1045. https://doi.org/10.3892/ol.2015.2846

Yizhak, K., Aguet, F., Kim, J., Hess, J. M., Kubler, K., Grimsby, J., … Getz, G. (2019). RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science*, *364*(6444). https://doi.org/10.1126/science.aaw0726

Yokoyama, A., Kakiuchi, N., Yoshizato, T., Nannya, Y., Suzuki, H., Takeuchi, Y., … Ogawa, S. (2019). Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature*, *565*(7739), 312–317. https://doi.org/10.1038/s41586-018-0811-x

Yoshida, K., Gowers, K. H. C., Lee-Six, H., Chandrasekharan, D. P., Coorens, T., Maughan, E. F., … Campbell, P. J. (2020). Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature*, *578*(7794), 266–272. https://doi.org/10.1038/s41586-020-1961-1

Yu, M., Song, X. G., Zhao, Y. J., Dong, X. H., Niu, L. M., Zhang, Z. J., … Xie, L. (2021). Circulating Serum Exosomal Long Non-Coding RNAs FOXD2-AS1, NRIR, and XLOC_009459 as Diagnostic Biomarkers for Colorectal Cancer. *Frontiers in Oncology*, *11*(March), 1–9. https://doi.org/10.3389/fonc.2021.618967

Zhang, H., Lu, Y., Wang, J., Zhang, T., Dong, C., Li, X., … Zhou, Y. (2019). Downregulation of the long non-coding RNA FOXD2-AS1 inhibits cell proliferation, migration and invasion in osteosarcoma. *Molecular Medicine Reports*, *20*(1), 292–302. https://doi.org/10.3892/mmr.2019.10254

Zhang, L., & Shay, J. W. (2017). Multiple Roles of APC and its Therapeutic Implications in Colorectal Cancer. *Journal of the National Cancer Institute*, *109*(8), 1–10. https://doi.org/10.1093/jnci/djw332

Zhang, M., Jiang, X., Jiang, S., Guo, Z., Zhou, Q., & He, J. (2019). LncRNA FOXD2-AS1 regulates miR-25-3p/sema4c axis to promote the invasion and migration of colorectal cancer cells. *Cancer Management and Research*, *11*, 10633–10639. https://doi.org/10.2147/CMAR.S228628

Zhang, P., Wu, W., Chen, Q., & Chen, M. (2019). Non-Coding RNAs and their Integrated Networks. *Journal of Integrative Bioinformatics*, *16*(3), 1–12. https://doi.org/10.1515/jib-2019-0027

Zhang, Y., Hu, J., Zhou, W., & Gao, H. (2019). LncRNA FOXD2-AS1 accelerates the papillary thyroid cancer progression through regulating the miR-485-5p/KLK7 axis. *Journal of Cellular Biochemistry*, *120*(5), 7952–7961. https://doi.org/10.1002/jcb.28072

Zhang, Y., Wu, Q., Xu, L., Wang, H., Liu, X., Li, S., … Fan, J. B. (2021). Sensitive detection of colorectal cancer in peripheral blood by a novel methylation assay. *Clinical Epigenetics*, *13*(1), 90. https://doi.org/10.1186/s13148-021-01076-8

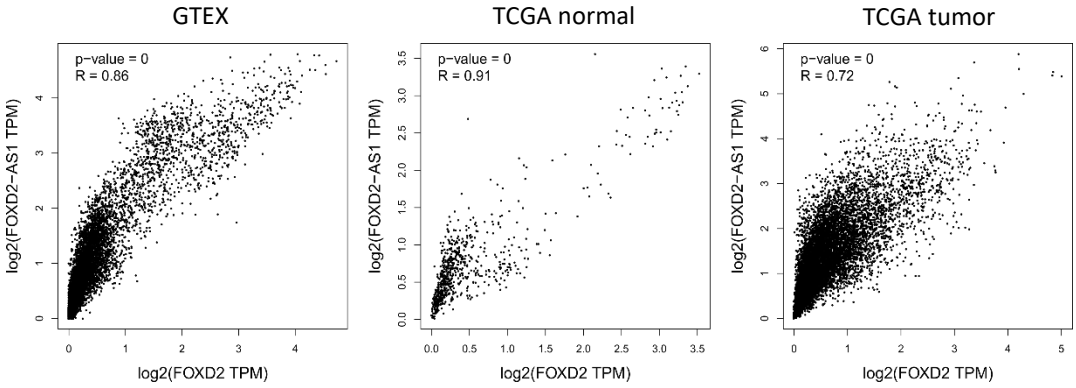
Zhang, Z., Yang, C., Gao, W., Chen, T., Qian, T., Hu, J., & Tan, Y. (2015). FOXA2 attenuates the epithelial to mesenchymal transition by regulating the transcription of E-cadherin and ZEB2 in human breast cancer. *Cancer Letters*, *361*(2), 240–250. https://doi.org/10.1016/j.canlet.2015.03.008

Zhu, Y., Qiao, L., Zhou, Y., Ma, N., Wang, C., & Zhou, J. (2018). Long non-coding RNA FOXD2-AS1 contributes to colorectal cancer proliferation through its interaction with microRNA-185-5p. *Cancer Science*, *109*(7), 2235–2242. https://doi.org/10.1111/cas.13632

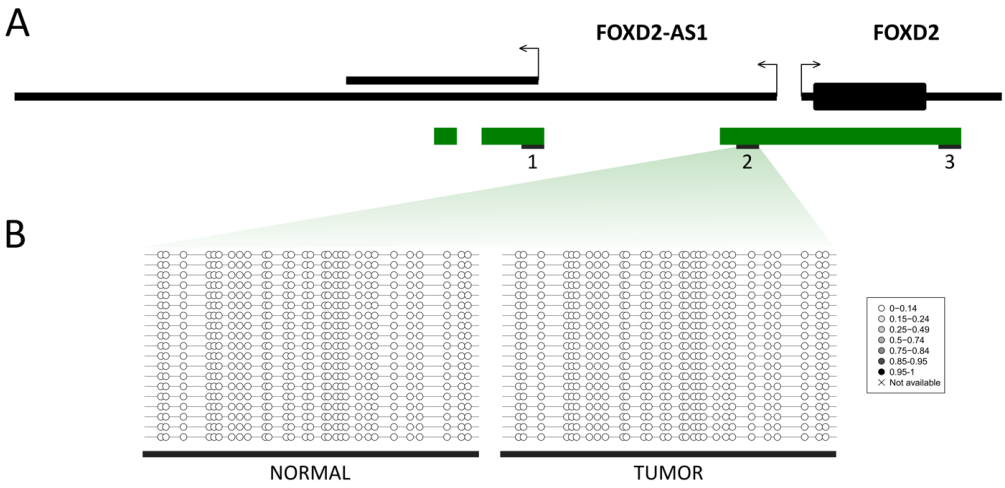
Ziller, M. J., Müller, F., Liao, J., Zhang, Y., Gu, H., Bock, C., … Meissner, A. (2011). Genomic distribution and Inter-Sample variation of Non-CpG methylation across human cell types. *PLoS Genetics*, *7*(12). https://doi.org/10.1371/journal.pgen.1002389

Zisman, A. L., Nickolov, A., Brand, R. E., Gorchow, A., & Roy, H. K. (2006). *Associations Between the Age at Diagnosis and Location of Colorectal Cancer and the Use of Alcohol and Tobacco*. *166*, 629–635.

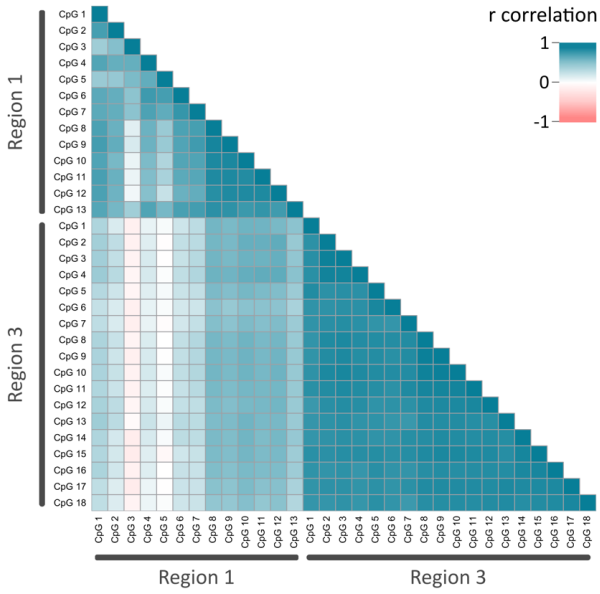
APPENDIX



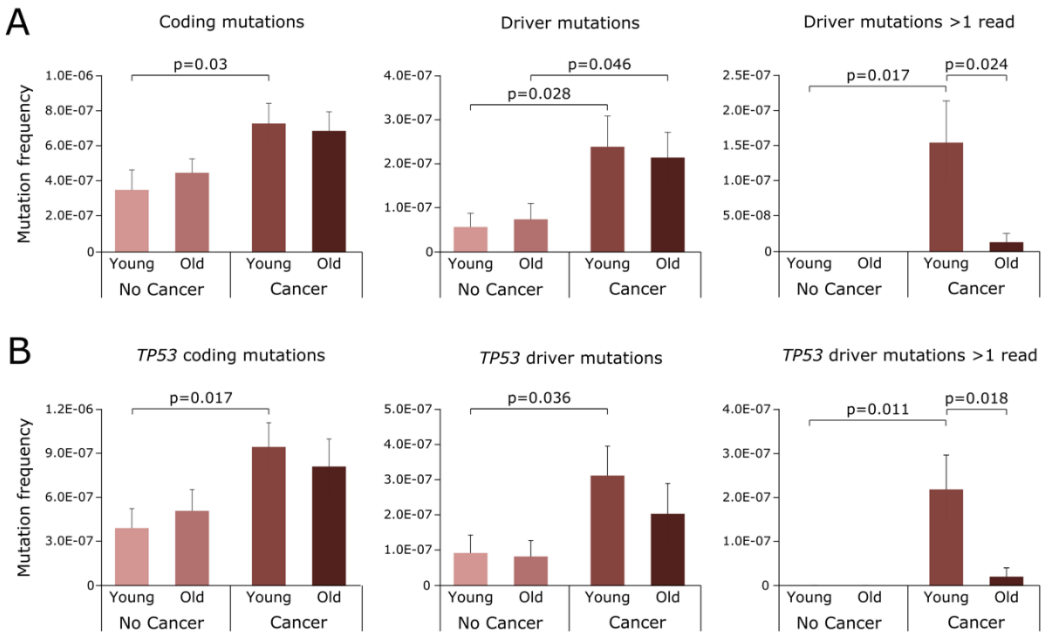
Supplementary Figure S1. Correlation of expression between FOXD2 and FOXD2-AS1. Correlation of expression analysis between FOXD2 and FOXD2-AS1 in normal colorectal tissues (GTEX and TCGA) and in colorectal tumors (TCGA). P values and r correspond to Pearson correlation. Results downloaded from GEPIA tool (Tang et al., 2017).



Supplementary Figure S2. Promoter methylation changes in CRC. A. Gene locations and the corresponding CpG islands in green. **B.** Methylation profiles of paired normal-tumor samples (n=20) at region 2, located around the promoter of FOXD2 and FOXD2-AS1. Data from HUB cohort.



Supplementary Figure S3. **Coordinated methylation between proximal and distant CpGs in HUB samples**. Correlation matrix describing pair wise Pearson correlation of methylation status among the amplified regions 1 and 3 within the FOXD2 and FOXD2-AS1 coding region from HUB samples. Region 1 (FOXD2-AS1 body gene) and region 3 (FOXD2 3‘ UTR) are indicated at the left and bottom bars.



Supplementary Figure S4. Driver mutations and larger clones are more abundant in the normal colon of young patients with CRC. Comparison of mutation frequency in the normal colon of younger (<55 year old) and older (≥55 year old) individuals with and without CRC. Mutation frequencies were calculated for overall (**A**) and *TP53*-only (**B**) mutations that are coding, drivers, and drivers with >1 mutated duplex read. P-values correspond to t-test comparisons.

Supplementary Table S1. List of 48 custom probes against FOXD2-AS1 used for RNA FISH.

Probe Sequence (5' to 3')	Probe name	Probe Sequence (5' to 3')	Probe name
TGCTGCAAGACGCCGAACAG	FOXD2-AS1_1	CTGCTCTCAAGAGCGTGGAG	FOXD2-AS1_25
GGAAAGAGTCCAGGTTTTCG	FOXD2-AS1_2	AGAAACCCACAAGAGCGCAC	FOXD2-AS1_26
CAGAGACGCTGTAACCAAGA	FOXD2-AS1_3	TCCAGAGGACAGACACATGA	FOXD2-AS1_27
AATTGTTCTGGGCTGCACGA	FOXD2-AS1_4	ACACAGGACACAGGATGCAA	FOXD2-AS1_28
TAGTGGAAAAGCCCAACAGG	FOXD2-AS1_5	GACACCAAGCATGAGGTCTG	FOXD2-AS1_29
TAGCAACGTACTCTTCGCAA	FOXD2-AS1_6	CTTCAGACTCTGGGGGGAAG	FOXD2-AS1_30
GAACAGCTCATTTATGGGGA	FOXD2-AS1_7	TACAGTCACAGACCCTCAAC	FOXD2-AS1_31
ATAATCGCTGGAGGGCTTTT	FOXD2-AS1_8	TGCATGAACTCCTTTTCCAT	FOXD2-AS1_32
AGGACAAACTCCGCTTCAAG	FOXD2-AS1_9	CTGGAGTATTCTTGGCTGTC	FOXD2-AS1_33
AGCTCGAACCGCTGAAAAGG	FOXD2-AS1_10	TGTAATTGGTAGGAGGGAGG	FOXD2-AS1_34
TAGGTCCAGAGTGGGAAGGA	FOXD2-AS1_11	TGTCTATGGTACACACAGGT	FOXD2-AS1_35
TAGACAGCTATCTCGCTTTG	FOXD2-AS1_12	CGTGAAGGTGAGCGCATGTG	FOXD2-AS1_36
CGAAGATCCGGGTGGAGAAA	FOXD2-AS1_13	TCTGGGACTCAGAAGGGTTA	FOXD2-AS1_37
CTAAGGGAGCTGATCGCTTC	FOXD2-AS1_14	CAGGGACGCGGCAATATTTC	FOXD2-AS1_38
GTCACAGGATCTGGAGTCTC	FOXD2-AS1_15	TTGCTTCTATGAGGCTTACG	FOXD2-AS1_39
CATGGGGAACATGTCTGAGG	FOXD2-AS1_16	AGGAATCCATTACTAGCGTC	FOXD2-AS1_40
CTTCAGAGTTGAAGGTGCAC	FOXD2-AS1_17	CTGACTCTGTGTGGATGAGA	FOXD2-AS1_41
AGACGCGTGGTGGTTATCTC	FOXD2-AS1_18	TCTCAGAACCAGTCCTTTAG	FOXD2-AS1_42
ATTCACTCTCGAACTTTGCC	FOXD2-AS1_19	AGTAGGGTGAGGAAAGGGTG	FOXD2-AS1_43
TACTTGGGTGCTTAAGCGAG	FOXD2-AS1_20	ATAACTTTTCCAAGCGGGTG	FOXD2-AS1_44
CTCGCCTTGGTCTCAACAAC	FOXD2-AS1_21	TCCGGAGGTTAAAAGTCTCT	FOXD2-AS1_45
CTCTTCCCACGAACAACAGC	FOXD2-AS1_22	GGAGAAGAGCAGGCAGGAAG	FOXD2-AS1_46
TTTCAAGTGGCGCTGTTTTC	FOXD2-AS1_23	TGGAGTGGATTCACAGTCTT	FOXD2-AS1_47
GGAACTCTGTGATCTTCAGG	FOXD2-AS1_24	AGCATGCAACACAAGGCGTG	FOXD2-AS1_48

Supplementary Table S2. ENSEMBL experiment ID of RNA-seq, H3K4me3 and H3K27ac data visualized on UCSC Genome Browser.

			Experiment ID		
Description	**Sex**	**Age**	**RNA-seq**	**H3K4me3**	**H3K27ac**
colonic mucosa	Female	16	ENCSR516BJM		
left colon	Female	46	ENCSR773COB		
mucosa of descending colon	Male	40	ENCSR674KHG		
left colon	Female	59	ENCSR759TPN		
colonic mucosa	Female	41	ENCSR202OWR	ENCSR574USP	
transverse colon	Female	51	ENCSR403SZN	ENCSR315EZG	ENCSR792VLP
transverse colon	Male	37		ENCSR813ZEY	ENCSR640XRV
transverse colon	Female	53		ENCSR933BVL	ENCSR208QRN
transverse colon	Male	54			ENCSR069EGE

Supplementary Table S3. Clinicopathological characteristics of individuals study II. Smoking: 0-Never; 1-Former; 2-Current. Abbreviations: M, male; F, female; n/a, not applicable; NA, not available; BMI, body mass index.

Patient ID #	Colorectal cancer	Age	Gender	Biopsy colon location	Distance from tumor	Polyp Former	BMI	Smoking	Tumor Stage	Tumor MSI	Chemo-therapy
P1	No	40	F	Left	n/a	Yes	36,7	2	n/a	n/a	No
P2	No	43	F	Left	n/a	No	NA	NA	n/a	n/a	No
P3	No	44	F	Left	n/a	No	30,4	1	n/a	n/a	No
P4	No	45	F	Left	n/a	Yes	21,8	0	n/a	n/a	No
P5	No	45	M	Left	n/a	No	30,8	0	n/a	n/a	No
P6	No	51	M	Left	n/a	Yes	25,6	2	n/a	n/a	No
P7	No	51	M	Left	n/a	No	32,0	2	n/a	n/a	No
P8	No	51	M	Left	n/a	Yes	26,7	2	n/a	n/a	No
P9	No	53	F	Left	n/a	Yes	22,0	0	n/a	n/a	No
P10	No	53	M	Left	n/a	Yes	25,3	2	n/a	n/a	No
P11	No	53	M	Left	n/a	No	29,6	2	n/a	n/a	No
P12	No	60	F	Left	n/a	No	26,8	0	n/a	n/a	No
P13	No	61	F	Left	n/a	Yes	26,9	0	n/a	n/a	No
P14	No	62	F	Left	n/a	Yes	21,6	0	n/a	n/a	No
P15	No	63	F	Left	n/a	No	32,8	0	n/a	n/a	No
P16	No	66	M	Left	n/a	Yes	34,8	0	n/a	n/a	No
P17	No	67	M	Left	n/a	Yes	27,8	0	n/a	n/a	No
P18	No	67	M	Left	n/a	No	28,1	1	n/a	n/a	No
P19	No	68	F	Left	n/a	Yes	27,0	1	n/a	n/a	No
P20	No	68	M	Left	n/a	Yes	27,8	2	n/a	n/a	No
P21	No	70	F	Left	n/a	Yes	37,7	1	n/a	n/a	No
P22	No	71	F	Left	n/a	Yes	22,4	0	n/a	n/a	No
P23	No	73	F	Left	n/a	No	21,9	0	n/a	n/a	No
P24	No	79	F	Left	n/a	Yes	34,2	0	n/a	n/a	No
P25	Yes	37	F	Left	10-15cm	Yes	25,6	0	IIIA	No	No
P26	Yes	37	F	Left	10-15cm	No	32,4	0	IV	No	No
P27	Yes	45	M	Left	10-15cm	Yes	31,1	0	IIA	No	No
P28	Yes	47	M	Left	10-15cm	Yes	27,7	0	NA	NA	No
P29	Yes	47	F	NA	10-15cm	NA	39,3	1	IIA	No	No
P30	Yes	48	M	Left	10-15cm	No	29,7	0	IIIA	No	No
P31	Yes	49	M	Left	10-15cm	No	29,8	0	IV	No	No
P32	Yes	49	F	NA	10-15cm	Yes	21,1	0	IIIA	No	No
P33	Yes	50	M	Left	10-15cm	No	29,9	0	IIIA	No	No
P34	Yes	50	F	NA	10-15cm	Yes	20,2	0	IIIC	Yes	No
P35	Yes	50	M	Left	3-5cm	Yes	26,3	0	IIIB	No	No
P36	Yes	55	F	Left	10-15cm	No	21,8	1	I	No	No
P37	Yes	56	M	Left	10-15cm	No	32,1	0	IV	No	No
P38	Yes	61	M	Left	10-15cm	Yes	20,3	2	IIA	No	No
P39	Yes	62	M	NA	3-5cm	NA	34,0	1	IIA	No	No
P40	Yes	63	M	Left	10-15cm	Yes	23,0	0	IV	No	Yes
P41	Yes	63	M	Right	10-15cm	Yes	34,6	1	IIIB	Yes	No
P42	Yes	64	M	Left	3-5cm	No	30,2	0	IIIB	No	No
P43	Yes	69	F	Right	10-15cm	Yes	27,7	0	I	No	No
P44	Yes	75	M	Left	10-15cm	Yes	28,6	1	I	No	No
P45	Yes	75	M	Left	10-15cm	No	30,0	1	I	No	No
P46	Yes	77	M	Left	10-15cm	Yes	33,2	1	I	No	No
P47	Yes	79	M	NA	10-15cm	NA	27,2	0	I	No	No

Supplementary Table S4. CRISPR-DS hybridization capture probes.

Probe Name	Probe Sequence
*BRAF*_e15_1	/5Biosg/ATATATTTCTTCATGAAGACCTCACAGTAAAAATAGGTGATTTTGGTCTAGCTACAGTGAAATCTCGATGGAGTGGGTCCCATCAGTTTGAACAGTTGTCTGGATCCATTTTGTGGATGG
*BRAF*_e15_2	/5Biosg/TAAGAATTGAGGCTATTTTTCCACTGATTAAATTTTTGGCCCTGAGATGCTGCTGAGTTACTAGAAAGTCATTGAAGGTCTCAACTATAGTATTTTCATAGTTCCCAGTATTCACAAAAA
*KRAS*_e2_1	/5Biosg/ATTAACCTTATGTGTGACATGTTCTAATATAGTCACATTTTCATTATTTTTATTATAAGGCCTGCTGAAAATGACTGAATATAAACTTGTGGTAGTTGGAGCTGGTGGCGTAGGCAAGAG
*KRAS*_e2_2	/5Biosg/TGCCTTGACGATACAGCTAATTCAGAATCATTTTGTGGACGAATATGATCCAACAATAGAGGTAAATCTTGTTTTAATATGCATATTACTGGTGCAGGACCATTCTTTGATACAGATAAA
*PIK3CA*_e10_1	/5Biosg/ATCCAGAGGGGAAAAATATGACAAAGAAAGCTATATAAGATATTATTTTATTTTACAGAGTAACAGACTAGCTAGAGACAATGAATTAAGGGAAAATGACAAAGAACAGCTCAAAGCAAT
*PIK3CA*_e10_2	/5Biosg/GGAAAATGACAAAGAACAGCTCAAAGCAATTTCTACACGAGATCCTCTCTCTGAAATCACTGAGCAGGAGAAAGATTTTCTATGGAGTCACAGGTAAGTGCTAAAATGGAGATTCTCTGT
*PIK3CA*_e21_1	/5Biosg/GATGCTTGGCTCTGGAATGCCAGAACTACAATCTTTTGATGACATTGCATACATTCGAAAGACCCTAGCCTTAGATAAAACTGAGCAAGAGGCTTTGGAGTATTTCATGAAACAAATGAA
*PIK3CA*_e21_2	/5Biosg/ATTGGATCTTCCACACAATTAAACAGCATGCATTGAACTGAAAAGATAACTGAGAAAATGAAAGCTCACTCTGGATTCCACACTGCACTGTTAATAACTCTCAGCAGGCAAAGACCGATT
*TP53*_e11_1	/5Biosg/CCCCGGGACAAAGCAAATGGAAGTCCTGGGTGCTTCTGACGCACACCTATTGCAAGCAAGGGTTCAAAGACCCAAAACCCAAAATGGCAGGGGAGGGAGAGATGGGGGTGGGAGGCTGTC
*TP53*_e11_2	/5Biosg/AGTGGGGAACAAGAAGTGGAGAATGTCAGTCTGAGTCAGGCCCTTCTGTCTTGAACATGAGTTTTTTATGGCGGGAGGTAGACTGACCCTTTTTGGACTTCAGGTGGCTGTAGGAGACAG
*TP53*_e10_1	/5Biosg/AATCCTATGGCTTTCCAACCTAGGAAGGCAGGGGAGTAGGGCCAGGAAGGGGCTGAGGTCACTCACCTGGAGTGAGCCCTGCTCCCCCCTGGCTCCTTCCCAGCCTGGGCATCCTTGAGT
*TP53*_e10_2	/5Biosg/TCCAAGGCCTCATTCAGCTCTCGGAACATCTCGAAGCGCTCACGCCCACGGATCTGCAGCAACAGAGGAGGGGGAGAAGTAAGTATATACACAGTACCTGAGTTAAAAGATGGTTCAAGT
*TP53*_e9_1	/5Biosg/AAGAGGTCCCAAGACTTAGTACCTGAAGGGTGAAATATTCTCCATCCAGTGGTTTCTTCTTTGGCTGGGGAGAGGAGCTGGTGTTGTTGGGCAGTGCTAGGAAAGAGGCAAGGAAAGGTG
*TP53*_e8_1	/5Biosg/GCATAACTGCACCCTTGGTCTCCTCCACCGCTTCTTGTCCTGCTTGCTTACCTCGCTTAGTGCTCCCTGGGGGCAGCTCGTGGTGAGGCTCCCCTTTCTTGCGGAGATTCTCTTCCTCTG
*TP53*_e8_2	/5Biosg/TGCGCCGGTCTCTCCCAGGACAGGCACAAACACGCACCTCAAAGCTGTTCCGTCCCAGTAGATTACCACTACTCAGGATAGGAAAAGAGAAGCAAGAGGCAGTAAGGAAATCAGGTCCTA
*TP53*_e7_1	/5Biosg/TGACCTGGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTTCATGCCGCCCATGCAGGAACTGTTACACATGTAGTTGTAGTGGATGGTGGTACAGTCAGAGCCAACCTAGGA
*TP53*_e7_2	/5Biosg/ATGTGATGAGAGGTGGATGGGTAGTAGTATGGAAGAAATCGGTAAGAGGTGGGCCCAGGGGTCAGAGGCAAGCAGAGGCTGGGGCACAGCAGGCCAGTGTGCAGGGTGGCAAGTGGCTCC
*TP53*_e6_1	/5Biosg/GACCTCAGGCGGCTCATAGGGCACCACCACACTATGTCGAAAAGTGTTTCTGTCATCCAAATACTCCACACGCAAATTTCCTTCCACTCGGATAAGATGCTGAGGAGGGGCCAGACCTAA
*TP53*_e6_2	/5Biosg/ATCTCATGGGGTTATAGGGAGGTCAAATAAGCAGCAGGAGAAAGCCCCCCTACTGCTCACCTGGAGGGCCACTGACAACCACCCTTAACCCCTCCTCCCAGAGACCCCAGTTGCAAACCA
*TP53*_e5_1	/5Biosg/CACCATCGCTATCTGAGCAGCGCTCATGGTGGGGGCAGCGCCTCACAACCTCCGTCATGTGCTGTGACTGCTTGTAGATGGCCATGGCGCGGACGCGGGTGCCGGGCGGGGGTGTGGAAT
*TP53*_e5_2	/5Biosg/CAACCCACAGCTGCACAGGGCAGGTCTTGGCCAGTTGGCAAAACATCTTGTTGAGGGCAGGGGAGTACTGTAGGAAGAGGAAGGAGACAGAGTTGAAAGTCAGGGCACAAGTGAACAGAT
*TP53*_e4_1	/5Biosg/AAGGGACAGAAGATGACAGGGGCCAGGAGGGGGCTGGTGCAGGGGCCGCCGGTGTAGGAGCTGCTGGTGCAGGGGCCACGGGGGGAGCAGCCTCTGGCATTCTGGGAGCTTCATCTGGAC
*TP53*_e4_2	/5Biosg/CTGGGTCTTCAGTGAACCATTGTTCAATATCGTCCGGGGACAGCATCAAATCATCCATTGCTTGGGACGGCAAGGGGGACTGTAGATGGGTGAAAAGAGCAGTCAGAGGACCAGGTCCTC
*TP53*_e2_1	/5Biosg/CGCTTCCCACAGGTCTCTGCTAGGGGGCTGGGGTTGGGGTGGGGGTGGTGGGCCTGCCCTTCCAATGGATCCACTCACAGTTTCCATAGGTCTGAAAATGTTTCCTGACTCAGAGGGGGC
*TP53*_e2_2	/5Biosg/TCGACGCTAGGATCTGACTGCGGCTCCTCCATGGCAGTGACCCGGAAGGCAGTCTGGCTGCTGCAAGAGGAAAAGTGGGGATCCAGCATGAGACACTTCCAACCCTGGGTCACCTGGGCC

Supplementary Table S5. Normal colon mucosa sequencing coverage and mutation frequency.

			Overall panel							TP53				BRAF, KRAS & PIK3CA			
			Duplex nucleotides sequenced			Nº mutations		Mutation frequency		Nº mutations		Mutation frequency		Nº mutations		Mutation frequency	
Patient # ID	Colorectal cancer	Mean depth	Total	Coding	Non-coding	Coding	Non-coding	Coding	Non-coding	Coding	Non-coding	Coding	Non-coding	Coding	Non-coding	Coding	Non-coding
P1	No	2,695	9,326,130	4,924,086	4,402,044	2	2	4.06E-07	4.54E-07	2	1	6.49E-07	3.48E-07	0	1	0.00E+00	6.55E-07
P2	No	1,551	5,367,067	3,047,020	2,320,047	0	0	0	0	0	0	0	0	0	0	0	0
P3	No	3,080	10,658,938	5,634,308	5,024,630	0	2	0	3.98E-07	0	1	0	4.02E-07	0	1	0	3.94E-07
P4	No	3,832	13,263,374	7,280,322	5,983,052	1	1	1.37E-07	1.67E-07	1	1	2.09E-07	2.58E-07	0	0	0.00E+00	0.00E+00
P5	No	3,378	11,690,912	6,452,655	5,238,257	2	0	3.10E-07	0	1	0	2.51E-07	0	1	0	4.05E-07	0
P6	No	2,558	8,854,304	4,980,823	3,873,481	0	2	0	5.16E-07	0	1	0	3.66E-07	0	1	0	8.79E-07
P7	No	2,669	9,236,235	5,221,476	4,014,759	5	1	9.58E-07	2.49E-07	4	1	1.31E-06	4.04E-07	1	0	4.61E-07	0.00E+00
P8	No	1,666	5,765,591	3,390,348	2,375,243	2	0	5.90E-07	0	2	0	9.57E-07	0	0	0	0.00E+00	0
P9	No	2,181	7,547,857	4,352,345	3,195,512	1	2	2.30E-07	6.26E-07	1	2	3.55E-07	8.36E-07	0	0	0.00E+00	0.00E+00
P10	No	1,363	4,718,201	2,786,196	1,932,005	3	0	1.08E-06	0	1	0	5.92E-07	0	2	0	1.82E-06	0
P11	No	3,530	12,216,241	7,156,605	5,059,636	1	2	1.40E-07	3.95E-07	0	2	0.00E+00	5.43E-07	1	0	3.18E-07	0.00E+00
P12	No	2,392	8,278,036	4,644,030	3,634,006	0	1	0	2.75E-07	0	1	0	3.88E-07	0	0	0	0.00E+00
P13	No	1,954	6,763,275	3,771,223	2,992,052	2	2	5.30E-07	6.68E-07	2	2	7.51E-07	9.48E-07	0	0	0.00E+00	0.00E+00
P14	No	2,771	9,589,139	5,223,745	4,365,394	2	1	3.8E-07	2.29E-07	2	0	6E-07	0	0	1	0	6.49E-07
P15	No	3,516	12,170,283	6,527,065	5,643,218	3	0	4.60E-07	0	0	0	0.00E+00	0	3	0	7.30E-07	0
P16	No	3,593	12,435,304	6,762,953	5,672,351	4	2	5.9E-07	3.53E-07	3	2	7.1E-07	5.43E-07	1	0	3.9E-07	0
P17	No	4,043	13,993,326	7,526,392	6,466,934	6	0	8E-07	0.00E+00	5	0	1.1E-06	0.00E+00	1	0	3.2E-07	0.00E+00
P18	No	3,030	10,488,236	5,631,727	4,856,509	2	1	3.55E-07	2.06E-07	0	0	0.00E+00	0.00E+00	2	1	6.00E-07	3.89E-07
P19	No	1,437	4,971,794	2,960,372	2,011,422	1	2	3.38E-07	9.94E-07	0	1	0.00E+00	7.02E-07	1	1	8.30E-07	1.7E-06
P20	No	1,715	5,934,067	3,444,544	2,489,523	4	1	1.2E-06	4.02E-07	4	1	1.8E-06	5.45E-07	0	0	0	0.00E+00
P21	No	2,524	8,737,008	5,118,042	3,618,966	2	1	3.91E-07	2.76E-07	1	0	3.35E-07	0.00E+00	1	1	4.69E-07	9.31E-07
P22	No	1,403	4,856,222	2,809,574	2,046,648	1	2	3.56E-07	9.77E-07	1	0	5.58E-07	0	0	2	0.00E+00	3.52E-06
P23	No	1,992	6,894,381	3,925,682	2,968,699	1	0	2.55E-07	0.00E+00	1	0	3.77E-07	0.00E+00	0	0	0.00E+00	0.00E+00
P24	No	2,439	8,440,231	4,839,530	3,600,701	1	0	2.07E-07	0	1	0	3.57E-07	0	0	0	0.00E+00	0
P25	Yes	3,050	10,557,305	5,844,131	4,713,174	2	1	3.42E-07	2.12E-07	2	1	5.24E-07	3.06E-07	0	0	0.00E+00	0.00E+00
P26	Yes	1,531	5,298,860	3,015,681	2,283,179	4	0	1.3E-06	0.00E+00	2	0	1E-06	0.00E+00	2	0	1.9E-06	0.00E+00
P27	Yes	1,855	6,418,565	3,638,045	2,780,520	4	1	1.10E-06	3.60E-07	1	0	4.34E-07	0.00E+00	3	1	2.25E-06	1.19E-06
P28	Yes	2,908	10,065,384	5,517,689	4,547,695	2	2	3.6E-07	4.4E-07	2	2	7.5E-07	8.32E-07	0	0	0	0
P29	Yes	1,772	6,132,436	3,590,012	2,542,424	3	0	8.4E-07	0	2	0	9.1E-07	0	1	0	7.2E-07	0
P30	Yes	1,379	4,774,170	2,761,965	2,012,205	2	0	7.2E-07	0.00E+00	2	0	1.3E-06	0.00E+00	0	0	0	0.00E+00
P31	Yes	3,311	11,461,064	6,683,886	4,777,178	2	0	2.99E-07	0.00E+00	1	0	2.73E-07	0.00E+00	1	0	3.31E-07	0.00E+00
P32	Yes	2,092	7,239,612	4,189,719	3,049,893	1	0	2.39E-07	0	1	0	4.27E-07	0	0	0	0.00E+00	0
P33	Yes	3,062	10,596,572	5,678,138	4,918,434	4	0	7E-07	0.00E+00	4	0	1.1E-06	0.00E+00	0	0	0	0.00E+00
P34	Yes	2,118	7,331,547	4,196,623	3,134,924	5	0	1.19E-06	0.00E+00	5	0	1.92E-06	0.00E+00	0	0	0.00E+00	0.00E+00
P35	Yes	3,835	13,272,777	7,748,274	5,524,503	7	1	9.03E-07	1.81E-07	7	0	1.80E-06	0	0	1	0.00E+00	4.73E-07
P36	Yes	1,917	6,636,208	3,690,634	2,945,574	2	4	5.42E-07	1.36E-06	1	3	4.58E-07	1.71E-06	1	1	6.64E-07	8.37E-07
P37	Yes	2,451	8,482,167	4,848,952	3,633,215	5	1	1.03E-06	2.75E-07	4	1	1.59E-06	4.84E-07	1	0	4.28E-07	0
P38	Yes	1,787	6,184,656	3,580,324	2,604,332	3	1	8.38E-07	3.84E-07	1	1	4.09E-07	5.02E-07	2	0	1.76E-06	0.00E+00
P39	Yes	2,678	9,268,032	5,493,337	3,774,695	0	1	0	2.65E-07	0	1	0	4.06E-07	0	0	0	0.00E+00
P40	Yes	1,781	6,163,835	3,485,544	2,678,291	4	1	1.15E-06	3.73E-07	4	0	1.84E-06	0.00E+00	0	1	0.00E+00	1.03E-06
P41	Yes	3,687	12,762,267	6,950,341	5,811,926	1	4	1.4E-07	6.88E-07	1	2	2.3E-07	5.45E-07	0	2	0	9.32E-07
P42	Yes	3,406	11,786,830	6,612,785	5,174,045	3	1	4.5E-07	1.93E-07	3	1	7.3E-07	3.09E-07	0	0	0	0
P43	Yes	4,306	14,904,427	8,742,469	6,161,958	7	2	8E-07	3.25E-07	3	1	7E-07	2.78E-07	4	1	8.9E-07	3.90E-07
P44	Yes	2,111	7,306,536	4,010,476	3,296,060	2	1	4.99E-07	3.03E-07	0	0	0.00E+00	0.00E+00	2	1	1.48E-06	8.81E-07
P45	Yes	1,531	5,300,225	3,001,331	2,298,894	3	1	1.00E-06	4.35E-07	2	1	9.00E-07	5.75E-07	1	0	1.28E-06	0
P46	Yes	1,602	5,545,003	3,203,376	2,341,627	2	2	6.2E-07	8.54E-07	2	2	1.1E-06	1.36E-06	0	0	0	0.00E+00
P47	Yes	1,268	4,387,202	2,565,183	1,822,019	3	2	1.17E-06	1.10E-06	3	1	1.77E-06	7.49E-07	0	1	0.00E+00	2.05E-06

Table S6. *BRAF*, *KRAS* and *PIK3CA* coding mutations detected by CRISPR-DS in normal colon tissue. Abbreviations: NA, not available.

Patient #ID	Colorectal Cancer	Age	Gene	Genomic location	Duplex Depth	Mutations	Mutant allele frequency	Mutation spectrum	cDNA variant	Protein Variant	Mutation type	Hotspot mutation	Seen in tumor
P5	No	45	*BRAF*	chr7:140753354	4335	1	0.000231	T>C	c.1781A>G	p.D594G	Missense		-
P7	No	51	*KRAS*	chr12:25245345	3672	1	0.000272	C>T	c.40G>A	p.V14I	Missense		-
P10	No	53	*BRAF*	chr7:140753274	923	1	0.00108	C>T	c.1860+1G>A	p.X620_splice	Splice		-
			KRAS	chr12:25245288	1183	1	0.000845	C>T	c.97G>A	p.D33N	Missense		-
P11	No	53	*PIK3CA*	chr3:179234217	8097	1	0.000124	A>G	c.3060A>G	p.A1020=	Silent		-
P15	No	63	*PIK3CA*	chr3:179218304	7345	1	0.000136	A>C	c.1634A>C	p.E545A	Missense		-
			PIK3CA	chr3:179218335	7347	1	0.000136	G>A	c.1664+1G>A	p.X555_splice	Splice		-
			PIK3CA	chr3:179234297	6632	1	0.000151	A>G	c.3140A>G	p.H1047R	Missense	x	-
P16	No	66	*BRAF*	chr7:140753345	4092	1	0.000244	A>C	c.1790T>G	p.L597R	Missense		-
P17	No	67	*PIK3CA*	chr3:179234297	4245	1	0.000236	A>G	c.3140A>G	p.H1047R	Missense	x	-
P18	No	67	*BRAF*	chr7:140753393	5251	1	0.00019	T>G	c.1742A>C	p.N581T	Missense		-
			KRAS	chr12:25245285	5747	1	0.000174	G>T	c.100C>A	p.P34T	Missense		-
P19	No	68	*PIK3CA*	chr3:179234328	2923	1	0.000342	G>A	c.3171G>A	p.W1057*	Nonsense		-
P21	No	70	*KRAS*	chr12:25245288	2430	1	0.000412	C>T	c.97G>A	p.D33N	Missense		-
P26	Yes	37	*KRAS*	chr12:25245350	2321	1	0.000431	C>A	c.35G>T	p.G12V	Missense	x	no
			PIK3CA	chr3:179234224	2041	1	0.00049	C>T	c.3067C>T	p.R1023*	Nonsense		no
P27	Yes	45	*BRAF*	chr7:140753332	1506	1	0.000664	T>A	c.1803A>T	p.K601N	Missense		no
			KRAS	chr12:25245347	1926	4	0.00208	C>T	c.38G>A	p.G13D	Missense	x	yes
			PIK3CA	chr3:179218307	2180	1	0.000459	A>G	c.1637A>G	p.Q546R	Missense		yes
P29	Yes	47	*BRAF*	chr7:140753354	1371	1	0.000729	T>C	c.1781A>G	p.D594G	Missense		no
P31	Yes	49	*PIK3CA*	chr3:179234280	6282	2	0.000318	A>T	c.3123A>T	p.K1041N	Missense		no
P36	Yes	55	*PIK3CA*	chr3:179234213	2162	1	0.000463	T>G	c.3056T>G	p.I1019S	Missense		no
P37	Yes	56	*KRAS*	chr12:25245350	3954	1	0.000253	C>G	c.35G>C	p.G12A	Missense	x	no
P38	Yes	61	*BRAF*	chr7:140753354	1264	3	0.00237	T>A	c.1781A>T	p.D594V	Missense		no
			KRAS	chr12:25245347	1521	1	0.000657	C>T	c.38G>A	p.G13D	Missense	x	no
P43	Yes	69	*KRAS*	chr12:25245306	5752	1	0.000174	G>T	c.79C>A	p.H27N	Missense		no
			KRAS	chr12:25245349	5767	1	0.000173	A>C	c.36T>G	p.G12=	Silent		no
			KRAS	chr12:25245351	5766	1	0.000173	C>A	c.34G>T	p.G12C	Missense	x	no
			PIK3CA	chr3:179234224	9605	1	0.000104	C>T	c.3067C>T	p.R1023*	Nonsense		no
P44	Yes	75	*KRAS*	chr12:25245350	2642	1	0.000379	C>A	c.35G>T	p.G12V	Missense	x	NA
			PIK3CA	chr3:179234218	1745	1	0.000573	T>C	c.3061T>C	p.Y1021H	Missense		NA
P45	Yes	75	*KRAS*	chr12:25245350	1282	1	0.00078	C>A	c.35G>T	p.G12V	Missense	x	yes

Supplementary Table S7. *TP53* coding mutations detected by CRISPR-DS in normal colon tissue. Abbreviations: MAF, mutant allele frequency; VUS, variant of unknown significance; NA, not available.

Patient #ID	Colorectal cancer	Age	Gene	Genomic location	Duplex Depth	Mutations	MAF	Exon	Mutation spectrum	cDNA variant	Protein variant	Mutation type	Hotspot mutation	Driver mutation	Frequency	Pathogenicity	Seen in tumor
P1	No	40	*TP53*	chr17:7673710	3652	3	0,00082	8	A>G	c.910A>G	p.T304A	Missense			Not frequent	VUS	-
			TP53	chr17:7676040	2446	4	0,00164	4	G>A	c.329G>A	p.R110H	Missense			Not frequent	VUS	-
P4	No	45	*TP53*	chr17:7675219	2740	1	0,00037	5	Deletion	c.388_393del	l	Indel		x	before	Pathogenic	-
P5	No	45	*TP53*	chr17:7670633	3649	1	0,00027	10	C>T	c.1076C>T	p.P359L	Missense			Rare/Unique	VUS	-
P7	No	51	*TP53*	chr17:7674217	2812	1	0,00036	7	G>A	c.746G>A	p.R249K	Missense			Frequent	Likely Pathogenic	-
			TP53	chr17:7674230	2812	1	0,00036	7	G>A	c.733G>A	p.G245S	Missense	x	x	Very frequent	Pathogenic	-
			TP53	chr17:7674245	2810	1	0,00036	7	A>T	c.718A>T	p.S240C	Missense			Rare/Unique	Likely Pathogenic	-
			TP53	chr17:7675070	1850	1	0,00054	5	G>A	c.542G>A	p.R181H	Missense			Frequent	Likely Pathogenic	-
P8	No	51	*TP53*	chr17:7676177	1292	1	0,00077	4	Deletion	c.192del	p.R65Efs*58	Indel		x	Not frequent	Pathogenic	-
			TP53	chr17:7676382	1607	1	0,00062	3	G>A	c.96G>A	p.L32=	Silent			before	Likely Benign	-
P9	No	53	*TP53*	chr17:7673787	2335	1	0,00043	8	C>T	c.833C>T	p.P278L	Missense			Very frequent	Pathogenic	-
P10	No	53	*TP53*	chr17:7673743	1491	1	0,00067	8	G>A	c.877G>A	p.G293R	Missense			Rare/Unique	VUS	-
P13	No	61	*TP53*	chr17:7674220	2496	1	0,0004	7	G>A	c.743G>A	p.R248Q	Missense	x	x	Very frequent	Pathogenic	-
			TP53	chr17:7675187	1523	1	0,00066	5	C>T	c.425C>T	p.P142L	Missense			Not frequent	VUS	-
P14	No	62	*TP53*	chr17:7674879	1422	1	0,0007	6	G>T	c.652G>T	p.V218L	Missense			Rare/Unique	VUS	-
			TP53	chr17:7676382	3258	1	0,00031	3	G>A	c.96G>A	p.L32=	Silent			before	Likely Benign	-
P16	No	66	*TP53*	chr17:7674221	2516	1	0,0004	7	C>T	c.742C>T	p.R248W	Missense	x	x	Very frequent	Pathogenic	-
			TP53	chr17:7674241	2515	1	0,0004	7	C>T	c.722C>T	p.S241F	Missense			Very frequent	Pathogenic	-
			TP53	chr17:7675139	3803	2	0,00053	5	G>A	c.473G>A	p.R158H	Missense			Very frequent	Pathogenic	-
P17	No	67	*TP53*	chr17:7673770	4462	1	0,00022	8	Deletion	c.841_850del	p.D281Qfs*61	Indel		x	before	Pathogenic	-
			TP53	chr17:7673772	4458	1	0,00022	8	Insertion	c.847_848insCTCTCCTC	p.R283Pfs*65	Indel		x	before	Pathogenic	-
			TP53	chr17:7673821	4457	1	0,00022	8	C>T	c.799C>T	p.R267W	Missense			Frequent	Pathogenic	-
			TP53	chr17:7675146	2058	1	0,00049	5	C>G	c.466C>G	p.R156G	Missense			Not frequent	Likely Pathogenic	-
			TP53	chr17:7675185	2063	4	0,00194	5	G>A	c.427G>A	p.V143M	Missense			Frequent	Pathogenic	-
P20	No	68	*TP53*	chr17:7670682	1756	2	0,00114	10	G>A	c.1027G>A	p.E343K	Missense			Rare/Unique	VUS	-
			TP53	chr17:7670684	1757	3	0,00171	10	G>A	c.1025G>A	p.R342Q	Missense			Rare/Unique	VUS	-
			TP53	chr17:7674256	1647	1	0,00061	7	A>G	c.707A>G	p.Y236C	Missense			Very frequent	Pathogenic	-
			TP53	chr17:7674891	2309	1	0,00043	6	C>T	c.640C>T	p.H214Y	Missense			Not frequent	Likely Pathogenic	-
P21	No	70	*TP53*	chr17:7674188	2528	6	0,00237	7	G>T	c.775G>T	p.D259Y	Missense			Frequent	Pathogenic	-
P22	No	71	*TP53*	chr17:7675095	1042	4	0,00384	5	G>A	c.517G>A	p.V173M	Missense			Very frequent	Pathogenic	-
P23	No	73	*TP53*	chr17:7676175	1668	1	0,0006	4	G>T	c.194G>T	p.R65I	Missense			before	VUS	-
P24	No	79	*TP53*	chr17:7673806	3047	1	0,00033	8	G>A	c.814G>A	p.V272M	Missense			Very frequent	Pathogenic	-
P25	Yes	37	*TP53*	chr17:7674953	1740	1	0,00058	6	A>T	c.578A>T	p.H193L	Missense			Very frequent	Pathogenic	NA
			TP53	chr17:7675148	2597	1	0,00039	5	C>T	c.464C>T	p.T155I	Missense			Frequent	Likely Pathogenic	NA
P26	Yes	37	*TP53*	chr17:7674220	1398	2	0,00143	7	G>A	c.743G>A	p.R248Q	Missense	x	x	Very frequent	Pathogenic	no
			TP53	chr17:7676261	1684	1	0,00059	4	G>T	c.108G>T	p.P36=	Silent			before	Likely Benign	no
P27	Yes	45	*TP53*	chr17:7673795	2289	5	0,00218	8	Deletion	c.822_825del	p.C275Pfs*69	Indel		x	before	Pathogenic	no
P28	Yes	47	*TP53*	chr17:7674281	2614	2	0,00077	7	Deletion	c.681_682del	p.D228Lfs*11	Indel		x	before	Pathogenic	NA
			TP53	chr17:7676257	2008	1	0,0005	4	C>T	c.112C>T	p.Q38*	Nonsense		x	Frequent	Pathogenic	NA
P29	Yes	47	*TP53*	chr17:7673558	1770	1	0,00057	9	G>A	c.970G>A	p.D324N	Missense			Rare/Unique	VUS	no
			TP53	chr17:7673740	2256	26	0,0115	8	G>T	c.880G>T	p.E294*	Nonsense		x	Very frequent	Pathogenic	yes
P30	Yes	48	*TP53*	chr17:7673838	981	3	0,00306	_07_S	G>T	c.783-1G>T	p.X261_splice	Splice		x	Frequent	Pathogenic	no
			TP53	chr17:7675157	1003	2	0,00199	5	C>T	c.455C>T	p.P152L	Missense			Very frequent	Pathogenic	no

Supplementary Table S7 (continuation). *TP53* coding mutations detected by CRISPR-DS in normal colon tissue. Abbreviations: MAF, mutant allele frequency; VUS, variant of unknown significance; NA, not available.

Patient #ID	Colorectal cancer	Age	Gene	Genomic location	Duplex Depth	Mutations	MAF	Exon	Mutation spectrum	cDNA variant	Protein variant	Mutation type	Hotspot mutation	Driver mutation	Frequency	Pathogenicity	Seen in tumor
P31	Yes	49	*TP53*	chr17:7673821	5091	2	0,00039	8	C>T	c.799C>T	p.R267W	Missense			Frequent	Pathogenic	no
P32	Yes	49	*TP53*	chr17:7670656	2679	1	0,00037	10	G>A	c.1053G>A	p.K351=	Silent			before	Likely Benign	no
P33	Yes	50	*TP53*	chr17:7675061	1658	1	0,0006	5	A>T	c.551A>T	p.D184V	Missense			Rare/Unique	VUS	yes
			TP53	chr17:7675075	1658	1	0,0006	5	T>G	c.537T>G	p.H179Q	Missense			Frequent	Pathogenic	yes
			TP53	chr17:7675083	1658	3	0,00181	5	C>A	c.529C>A	p.P177T	Missense			Rare/Unique	Likely Pathogenic	no
			TP53	chr17:7675088	1654	1	0,00061	5	G>A	c.524G>A	p.R175H	Missense	x	x	Very frequent	Pathogenic	no
P34	Yes	50	*TP53*	chr17:7673752	2957	1	0,00034	8	C>T	c.868C>T	p.R290C	Missense			Not frequent	Likely Pathogenic	no
			TP53	chr17:7673815	2959	1	0,00034	8	A>T	c.805A>T	p.S269C	Missense			Rare/Unique	VUS	no
			TP53	chr17:7674188	2205	1	0,00045	7	G>T	c.775G>T	p.D259Y	Missense			Frequent	Pathogenic	yes
			TP53	chr17:7674259	2181	1	0,00046	7	A>G	c.704A>G	p.N235S	Missense			Frequent	Benign	no
			TP53	chr17:7676398	2118	1	0,00047	3	Deletion	c.80del	p.P27Lfs*17	Indel		x	Very frequent	Pathogenic	no
P35	Yes	50	*TP53*	chr17:7669628	918	1	0,00109	11	A>G	c.1163A>G	p.E388G	Missense			before	VUS	no
			TP53	chr17:7673593	4451	1	0,00023	9	C>A	c.935C>A	p.T312N	Missense			before	VUS	no
			TP53	chr17:7673764	4591	3	0,00065	8	G>A	c.856G>A	p.E286K	Missense			Very frequent	Pathogenic	no
			TP53	chr17:7673772	4593	1	0,00022	8	G>A	c.848G>A	p.R283H	Missense			Frequent	Pathogenic	no
			TP53	chr17:7674893	1131	2	0,00177	6	G>A	c.638G>A	p.R213Q	Missense			Very frequent	Pathogenic	no
			TP53	chr17:7675057	2147	3	0,0014	5	C>T	c.555C>T	p.S185=	Silent			Rare/Unique	VUS	no
			TP53	chr17:7676564	2972	2	0,00067	2	G>A	c.31G>A	p.E11K	Missense			Rare/Unique	VUS	no
P36	Yes	55	*TP53*	chr17:7675067	1210	1	0,00083	5	Deletion	c.542_545del	p.R181Pfs*65	Indel		x	before	Pathogenic	no
P37	Yes	56	*TP53*	chr17:7673704	2517	1	0,0004	8	C>T	c.916C>T	p.R306*	Nonsense	x	x	Very frequent	Pathogenic	no
			TP53	chr17:7674220	2401	1	0,00042	7	G>A	c.743G>A	p.R248Q	Missense	x	x	Very frequent	Pathogenic	yes
			TP53	chr17:7676040	2505	1	0,0004	4	G>A	c.329G>A	p.R110H	Missense			Not frequent	VUS	no
			TP53	chr17:7676041	2505	1	0,0004	4	C>T	c.328C>T	p.R110C	Missense			Frequent	Pathogenic	no
P38	Yes	61	*TP53*	chr17:7676565	1962	1	0,00051	2	C>A	c.30C>A	p.V10=	Silent			before	Likely Benign	no
P40	Yes	63	*TP53*	chr17:7673584	1595	1	0,00063	9	C>A	c.944C>A	p.S315Y	Missense			before	VUS	NA
			TP53	chr17:7673751	2092	1	0,00048	8	G>T	c.869G>T	p.R290L	Missense			Not frequent	VUS	NA
			TP53	chr17:7673774	2092	1	0,00048	8	G>A	c.846G>A	p.R282=	Silent			Rare/Unique	VUS	NA
			TP53	chr17:7673811	2085	1	0,00048	8	T>G	c.809T>G	p.F270C	Missense			Frequent	Pathogenic	NA
P41	Yes	63	*TP53*	chr17:7673700	4322	1	0,00023	_08_S	G>A	c.919+1G>A	p.X307_splice	Splice		x	Frequent	Pathogenic	no
P42	Yes	64	*TP53*	chr17:7673534	3423	8	0,00234	_09_S	G>A	c.993+1G>A	p.X331_splice	Splice		x	Frequent	Pathogenic	yes
			TP53	chr17:7673759	4177	1	0,00024	8	Insertion	c.860dup	p.N288Efs*18	Indel		x	before	Pathogenic	no
			TP53	chr17:7674263	3817	1	0,00026	7	Deletion	c.686_700del	l	Indel		x	Not frequent	Pathogenic	no
P43	Yes	69	*TP53*	chr17:7674865	1694	1	0,00059	6	G>A	c.666G>A	p.P222=	Silent			Rare/Unique	VUS	no
			TP53	chr17:7675088	3003	1	0,00033	5	G>A	c.524G>A	p.R175H	Missense	x	x	Very frequent	Pathogenic	no
			TP53	chr17:7676046	4055	1	0,00025	4	G>T	c.323G>T	p.G108V	Missense			Rare/Unique	VUS	no
P45	Yes	75	*TP53*	chr17:7674954	2606	3	0,00115	6	C>T	c.577C>T	p.H193Y	Missense			Very frequent	Pathogenic	no
			TP53	chr17:7675146	1041	1	0,00096	5	C>T	c.466C>T	p.R156C	Missense			Not frequent	VUS	no
P46	Yes	77	*TP53*	chr17:7673726	1735	1	0,00058	8	G>A	c.894G>A	p.E298=	Silent			Rare/Unique	VUS	no
			TP53	chr17:7673778	1736	1	0,00058	8	A>T	c.842A>T	p.D281V	Missense			Frequent	Pathogenic	no
P47	Yes	79	*TP53*	chr17:7673767	1183	1	0,00085	8	G>A	c.853G>A	p.E285K	Missense			Very frequent	Pathogenic	no
			TP53	chr17:7674917	1028	1	0,00097	6	A>G	c.614A>G	p.Y205C	Missense			Very frequent	Pathogenic	no
			TP53	chr17:7674954	1027	4	0,00389	6	C>T	c.577C>T	p.H193Y	Missense			Very frequent	Pathogenic	no

Supplementary Table S8. Coding mutations with MAF>0.1 detected in paired tumors.

Patient #ID	Gene	Genomic location	Mutant allele frequency	Protein Variant	Mutation type	seen in normal
P27	*KRAS*	chr12:25245347	0,42	p.G13D	Missense	yes
	PIK3CA	chr3:179218307	0,22	p.Q546R	Missense	yes
P29	*TP53*	chr17:7673740	0,75	p.E294*	Missense	yes
P30	*KRAS*	chr12:25245350	0,38	p.G12D	Missense	no
P32	*TP53*	chr17:7675152	0,92	p.G154S	Missense	no
P33	*TP53*	chr17:7675061	0,92	p.D184V	Missense	yes
	TP53	chr17:7675075	0,92	p.H179Q	Missense	yes
P34	*PIK3CA*	chr3:179234296	0,30	p.H1047Y	Missense	no
	TP53	chr17:7673704	0,14	p.R306*	Missense	no
	TP53	chr17:7674188	0,17	p.D259Y	Missense	yes
P35	*TP53*	chr17:7675098	0,60	p.V172F	Missense	no
P36	*KRAS*	chr12:25245347	0,38	p.G13D	Missense	no
	TP53	chr17:7673803	0,62	p.R273C	Missense	no
P37	*TP53*	chr17:7674220	0,51	p.R248Q	Missense	yes
P39	*KRAS*	chr12:25245350	0,56	p.G12D	Missense	no
P41	*KRAS*	chr12:25245347	0,38	p.G13D	Missense	no
	PIK3CA	chr3:179218304	0,27	p.E545A	Missense	no
	TP53	chr17:7673751	0,49	p.R290H	Missense	no
P42	*TP53*	chr17:7673534	0,21	p.X331_splice	Splice	yes
P43	*KRAS*	chr12:25245349	0,39	p.G12W	Missense	no
P45	*KRAS*	chr12:25245350	0,56	p.G12V	Missense	yes
P46	*TP53*	chr17:7673820	0,16	p.L265_G266dup	Indel	no

Supplementary Table S9. FOXD2 and FOXD2-AS1 expression in TCGA-COAD patients. Wilcoxon matched-pairs signed rank test. Significant p values are highlighted in bold.

		FOXD2		FOXD2-AS1		FOXD2-AS1/FOXD2	
Tissue	**n**	**mean ± SEM**	**p value**	**mean ± SEM**	**p value**	**mean ± SEM**	**p value**
Normal	41	6.11 ± 0.35	**<0.0001**	4.87 ± 0.30	0,1555	0.85 ± 0.05	**<0.0001**
Tumor	41	2.54 ± 0.22		4.26 ± 0.36		2.16 ± 0.24	

Supplementary Table S10. GO terms for mutated genes associated with low FOXD2 and FOXD2-AS1 expression levels. N; total number of genes, B;total number of genes associated with a specific GO term, n;number of genes in the top of the user's input list or in the target wet when appropiate, b;number of genes in the intersection.

Gene	GO term	Description	P-value	FDR	Enrichment (N, B, n, b)
FOXD2	GO:0050772	positive regulation of axonogenesis	0,000191	1	8.36 (1104,12,66,6)
	GO:0010770	positive regulation of cell morphogenesis involved in differentiation	0,000601	1	5.33 (1104,23,72,8)
	GO:0071840	cellular component organization or biogenesis	0,000783	1	1.16 (1104,459,473,229)
	GO:0016043	cellular component organization	0,000783	1	1.16 (1104,459,473,229)
	GO:0070413	trehalose metabolism in response to stress	0,000906	1	1,104.00 (1104,1,1,1)
	GO:0005991	trehalose metabolic process	0,000906	1	1,104.00 (1104,1,1,1)
FOXD2-AS1	GO:1904837	beta-catenin-TCF complex assembly	0,000324	1	82.00 (861,3,7,2)
	GO:0016055	Wnt signaling pathway	0,00086	1	2.45 (861,26,216,16)
	GO:1905114	cell surface receptor signaling pathway involved in cell-cell signaling	0,000345	0,721	2.51 (861,27,216,17)
	GO:0032990	cell part morphogenesis	0,000636	0,997	3.02 (861,15,209,11)
	GO:0018205	peptidyl-lysine modification	0,000173	1	5.43 (861,28,51,9)
	GO:0032092	positive regulation of protein binding	0,000892	0,933	5.63 (861,9,102,6)
	GO:0044648	histone H3-K4 dimethylation	0,000927	0,831	47.83 (861,3,12,2)
	GO:0097692	histone H3-K4 monomethylation	0,000927	0,727	47.83 (861,3,12,2)

Supplementary Table S11. FOXD2 and FOXD2-AS1 expression in HUB patients. Wilcoxon matched-pairs signed rank test. Significant p values are highlighted in bold.

		FOXD2		FOXD2-AS1		FOXD2-AS1/FOXD2	
Tissue	**n**	**mean ± SEM**	**p value**	**mean ± SEM**	**p value**	**mean ± SEM**	**p value**
Normal	108	1.29 ± 0.1	**<0.0001**	0.67 ± 0.04	<0.0001	1.18 ± 0.12	**<0.0001**
Tumor	108	0.48 ± 0.4		0.39 ± 0.03		0.6 ± 0.02	

Supplementary Table S12. ROC curve and Overall Survival (OS) analysis. Significant p values are highlighted in bold.

	ROC curve			Long rank		
Variable	**AUC**	**95% CI**	**P value**	**HR**	**95% CI**	**P value**
FOXD2	0,57	0.51 to 0.64	**0,0369**	0,7	0.47 to 1.08	0,1106
FOXD2-AS1	0,54	0.47 to 0.61	0,054	1,3	0.84 to 1.90	0,266
FOXD2-AS1/FOXD2	0,65	0.59 to 0.72	**<0.0001**	2,5	1.65 to 3.77	**<0.0001**