



UCAM

UNIVERSIDAD CATÓLICA
DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO
Programa de Doctorado en Ciencias de la Salud

Multi-omics approaches for diagnosis, prognosis and
response to treatment of colorectal cancer

Autor:

Dr. Gaetano Gallo

Directores:

Dr. D. Alessio Gordon Naccarati

Dr. D. Pablo Conesa Zamora

Murcia, 24 de febrero de 2023



UCAM

UNIVERSIDAD CATÓLICA
DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO
Programa de Doctorado en Ciencias de la Salud

Multi-omics approaches for diagnosis, prognosis and
response to treatment of colorectal cancer

Autor:

Dr. Gaetano Gallo

Directores:

Dr. D. Alessio Gordon Naccarati

Dr. D. Pablo Conesa Zamora

Murcia, 24 de febrero de 2023



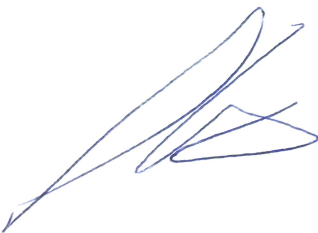
UCAM

UNIVERSIDAD CATÓLICA
DE MURCIA

AUTORIZACIÓN DE LO/S DIRECTOR/ES DE LA TESIS PARA SU PRESENTACIÓN

El Dr. D. Pablo Conesa-Zamora y el Dr. D. Alessio Gordon Naccarati como Directores⁽¹⁾ de la Tesis Doctoral titulada “Multi-omics approaches for diagnosis, prognosis and response to treatment of colorectal cancer” realizada por D. Gaetano Gallo en el Programa de Doctorado “Ciencias de la Salud”, **autoriza su presentación a trámite** dado que reúne las condiciones necesarias para su defensa.

Lo que firmamos, para dar cumplimiento al real decreto 99/2011 de 28 de enero, en murcia a 14 de septiembre de 2022.



Alessio G. Naccarati

RESUMEN

Palabras clave: Cirugía Abdominal, Gastroenterología, Patología, Biología Molecular

En los últimos años se ha hecho evidente que la gran heterogeneidad del cáncer colorrectal (CCR) influye en la biología de la enfermedad, su pronóstico y la respuesta al tratamiento. Por esta razón, se requiere una estratificación exhaustiva de la enfermedad. El advenimiento del análisis de Big Data en la investigación médica ha revolucionado el enfoque tradicional basado en hipótesis. El análisis de Big Data brinda una oportunidad única para mejorar la salud individual y la salud pública. De hecho, la disponibilidad de grandes bases de datos para capturar y almacenar el panorama genómico de los pacientes con CCR proporciona información muy valiosa sobre los genes que con frecuencia están alterados en el CCR. Además, la posibilidad de utilizar perfiles de expresión génica y análisis de secuenciación de próxima generación altamente sensibles para diferenciar los subtipos de CCR en grupos pronósticos también puede conducir a una mejor comprensión del tratamiento adecuado del CCR, mejorando así el pronóstico y la calidad de vida de los pacientes. El objetivo de este estudio, realizado en colaboración con el Instituto Italiano de Medicina Genómica (Turín, Italia) y el Departamento de Ciencias de la Computación de la Universidad de Turín (Turín, Italia), fue dilucidar las relaciones entre múltiples marcadores relevantes identificados por un multi- enfoque ómico (transcriptoma codificante y no codificante, metagenómica y estado mutacional) investigado, en diferentes tipos de muestras biológicas de los mismos sujetos (heces, plasma, tejido primario), firmas moleculares específicas para relacionarlas con el diagnóstico y pronóstico de pacientes con CCR. Se analizaron muestras recogidas de pacientes con diagnóstico de CCR, reclutados en nuestra consulta externa y sometidos a cirugía durante el periodo de estudio. Además, las características demográficas, así como la información clínico-patológica obtenida tanto en el diagnóstico previo (tomografía computarizada, resonancia magnética y colonoscopia) como

postoperatorio (examen histopatológico) se relacionaron con los resultados clínicos de los pacientes. Los pacientes fueron seguidos regularmente de acuerdo con las guías vigentes. Por último, estos datos se compararon con los obtenidos en nuestros estudios transversales previos que incluyeron pacientes con CCR, controles sanos y sujetos con diferentes tipos de pólipos. Las técnicas Omicas utilizadas fueron secuenciación de ARN pequeño (sRNA seq), secuenciación de ARN (RNA-seq), secuenciación metagenómica "shotgun" y perfil genómico basado en el ensayo de secuenciación de próxima generación (NGS) (estado mutacional e inestabilidad de microsatélites). Entre los resultados más relevantes demostramos, en primer lugar, que el análisis de miRNoma fecal identificó una firma predictiva que discrimina con precisión el CCR y las lesiones precancerosas que es de utilidad para un diagnóstico no invasivo destinado a mejorar la eficacia de los programas de detección actuales. En segundo lugar, señalamos el papel crucial de la expresión alterada de microARN (miARN) relacionados con el segmento cromosómico 8q24 para el inicio y/o progresión del cáncer, así como la correlación con el sistema de clasificación de subtipos moleculares de consenso (CMS). Por último, los perfiles de miARN en las heces permitieron reflejar rasgos comunes y hábitos de estilo de vida y deben considerarse en relación con estudios de enfermedades y asociaciones basados en la expresión de miARN en heces.

The high heterogeneity of colorectal cancer (CRC) in the disease biology, therapy response, and prognosis has become evident in the recent few years. For this reason, an extensive disease stratification is required. The advent of Big Data analysis in medical research has revolutionized the traditional hypothesis-driven approach. Big Data analysis provides an invaluable opportunity to improve individual and public health. In fact, the availability of large databases to capture and store the genomic landscape of patients with CRC provides information on the genes that are frequently deregulated in CRC. Moreover, the possibility of using gene-expression profiling and highly sensitive Next-Generation Sequencing (NGS) analyses to differentiate the subtypes of CRC into prognostic groups can also lead to a better understanding of adequate CRC treatment, improving prognosis and patients' quality of life. The aim of this study, conducted in collaboration with the Italian Institute for Genomic Medicine (Turin, Italy) and the Department of Computer Science, University of Turin (Turin, Italy), was to elucidate relationships between multiple relevant markers identified by a multi-

omics approach (coding and non-coding transcriptome, metagenomics, and mutational status) investigated in different types of biospecimens from the same subjects (stool, plasma, primary tissue) for the diagnosis and prognosis of patients with CRC, identifying specific signatures. Samples collected from patients with diagnosis of CRC, recruited in our outpatient clinic and undergoing surgery during the study period were analyzed. Furthermore, the demographic characteristics as well as the clinical-pathological information obtained at both pre- (imaging, i.e. TC-scan, MRI, and colonoscopy) and post-operative (histopathological examination) diagnosis were related to the clinical outcomes of patients. Patients were regularly followed-up in accordance with current guidelines. Lastly, these data were compared with those obtained in our previous cross-sectional studies including CRC patients, healthy controls and subjects with different types of polyps. The Omics techniques used were small RNA-sequencing (sRNA seq), RNA-sequencing (RNA-seq), shotgun metagenomics sequencing and genomic profile based on target sequencing (for the cancer mutational and MSI status). In this PhD project, we have obtained several results. Firstly, we demonstrated that fecal miRNome analysis identified a predictive signature accurately discriminating CRC and precancerous lesions for a non-invasive diagnosis aimed at improving the effectiveness of current screening programs. Secondly, we pointed out the crucial role of the altered expression of 8q24-related microRNAs (miRNAs) for the initiation and/or progression of cancer as well as the correlation with the consensus molecular subtypes (CMS) classification system. Lastly, miRNA profiles in stool may reflect common traits and lifestyle habits and should be considered in relation to disease and association studies based on faecal miRNA expression.

ACKNOWLEDGEMENTS

I am deeply indebted to Dr. Mario Trompetto, Dr Giuseppe Clerico and Dr. Alberto Realis Luc. You have strongly inspired my attitudes regarding colorectal surgery. I hope to be able to learn much more from you. Thanks for your friendship and mentorship.

This endeavor would not have been possible without the Genetic and Molecular Epidemiology Unit of the Italian Institute for Genomic Medicine (IIGM). My collaboration with Dr. Alessio Naccarati and his team started in 2015 and after 7 years, almost 500 samples collected and several publications and meeting presentations, it is still going on with Alessio as one of the two supervisors of this research project.

I would like to express my deepest gratitude to Prof. Pablo Conesa-Zamora who in this very difficult period, during the COVID-19 pandemic, transmitted to me his knowledge, professionalism, and outstanding expertise in the field.

I am a colorectal surgeon, and the world of translational research can be extremely difficult due to the huge number of skills required. For this reason, I would like to thank my two supervisors even more.

I'm also grateful to my tutor Silvia Montoro García who has constantly monitored my academic career during these PhD years

Words cannot express my gratitude to my father and my mother for their love and unconditional lifelong support. There hasn't been a day they haven't been on my side. Thanks for always believing in me, more than I did, and for representing my strength.

I could not have undertaken this journey without my beloved partner and future wife, Veronica. You are a special gift that surgery has given me. I feel so lucky to have met you. This alone would be enough to pay off my choice to be a surgeon.

Thanks for the patience and understanding that you have had and that you will have in the near future in supporting my academic ambitions.

"He who loves practice without theory is like the sailor who boards ship without a rudder and compass and never knows where he may cast".

- Leonardo Da Vinci -

TABLE OF CONTENTS

AUTORIZACIÓN DE LOS DIRECTORES	
ACKNOWLEDGEMENT	
SUMMARY	
LIST OF ABBREVIATIONS.....	17
LIST OF FIGURES.....	19
LIST OF PUBLICATIONS	21
CHAPTER I: INTRODUCTION	23
1.1. MicroRNAs	
1.2. The 8q24 locus	
1.3. Big Data	
CHAPTER II: RATIONALE.....	29
CHAPTER III: OBJECTIVES.....	31
CHAPTER IV: MATERIALS and METHODS.....	35
CHAPTER V: RESULTS	51
CHAPTER VI: DISCUSSION	71
CHAPTER VII: CONCLUSIONS	77
CHAPTER VIII: LIMITATIONS and FUTURE PERSPECTIVES.....	81
CHAPTER IX: REFERENCES	85

LIST OF ABBREVIATIONS

Cell migration-inducing and hyaluronan-binding protein = CEMIP

Colorectal cancer = CRC

Computed Tomography = CT

Consensus molecular subtypes = CMS

Czech = CZ

Deregulated miRNA = DEmiRNA

European Prospective Investigation into Cancer and Nutrition = EPIC

Extracellular vesicles = EVs

Faecal immunochemical test = FIT

Gastrointestinal = GI

Genome-wide association studies = GWAS

Inflammatory bowel disease = IBD

Italian = IT

Kirsten rat sarcoma virus = KRAS

MicroRNAs = miRNAs

Microsatellite instability = MSI

Next Generation Sequencing = NGS

Receiver Operating Characteristic = ROC

RNA-sequencing = RNA-seq

Small-RNA-sequencing = sRNA-seq

The Cancer Genome Atlas = TCGA

TruSight Oncology 500 = TSO

LIST OF FIGURES

LIST OF FIGURES

Figure 1. Representation of the study design.

Figure 2. Workflow of the study.

Figure 3. Scatterplot reporting the correlations of log₂FC of the 20 stool DEmiRNAs from the comparison between CRC and healthy subjects in common between the *IT-cohort* (x-axis) and the *CZ-cohort* (y-axis). In red are represented the up-regulated miRNAs, in blue the down-regulated ones.

Figure 4. DEmiRNA levels comparing CRC patients stratified for clinical data. The color of the dot is related to the log₂FC while the -size is proportional to the statistical significance.

Figure 5. Characterization of the 20 fecal DEmiRNAs in different sample types. Bar plot reporting the median expression levels in tumor, advanced (AA) and non-advanced adenoma (nAA) tissues. The color code represents the log₂FC from the paired differential expression analysis between CRC/adenoma tissues and matched adjacent mucosa *******, adj. p<0.001; ******adj. p<0.01; *****adj. p <0.05.

Figure 6. Box plots representing expression levels respectively of miR-151a-3p and miR-30d-5p in **(A, B)** CRC tissues versus normal adjacent mucosa from Study 1; **(C, D)** in the CRC-TCGA database (both for paired and not paired tissues); **(D, E)** in stool samples from CRC patients and healthy controls.

Figure 7. Box plots showing the expression levels of selected DEmiRNAs among individuals stratified according to the investigated common traits features. P values were computed using DESeq2 and adjusted using the FDR method. *******adj. p value 0.001, ******adj. p value 0.01, *****adj. p value 0.05.

Figure 8. Box plots showing the expression levels of selected DEmiRNAs among individuals stratified according to the investigated lifestyle features. P values

were computed using DESeq2 and adjusted using the FDR method. ***adj P value 0.001, **adj P value 0.01, *adj P value 0.05

Figure 9. Heatmap of the hierarchical clustering of significant associations between miRNAs and the investigated common traits and lifestyle variables (p-values adjusted for multiple testing). For each DEmiRNA, the log₂FCs of all the comparisons are reported.

Figure 10. Volcano Plot showing 3.849 differentially expressed genes between tissues that were observed after RNA-sequencing. On these genes we can build Consensus Molecular Subtypes.

Figure 11. Illumina TruSight Oncology 500 (TSO500 HT) was used as Next Generation Sequencing assay targeting 523 genes, assessing somatic mutations, SNV, indels, TMB and MSI. Eleven subjects of the cohort with MSI high were further investigated in blood samples where they resulted all MSI stable

Figure 12. CMS distribution in the study population

Figure 13. Top 5 enriched terms from Gene Ontology connected to the most overrepresented genes in each CMS subtype. No terms were identified for CMS-3

Figure 14. CMS subtypes stratified according to MSI and TMB

Figure 15. miRNA-Target interactions between the differentially expressed miRNAs and genes stratified by CMS subtypes. No interactions were identified for CMS-3

LIST OF PUBLICATIONS

Overall, this project included 457 patients who were enrolled at the Operative Unit of Colorectal Surgery of the S. Rita Clinic in Vercelli (Vercelli, Italy). Not all patients (and samples) have yet been included in the studies that will be discussed in this thesis. Moreover, to complete the analyzes it was necessary to recruit a Czech cohort, from Prague and Plzen, and a validation cohort from Brno, within Dr Naccarati's collaboration with the Academy of Sciences of the Czech Republic (**Study I**).

Concerning **Study III**, patients recruited have been divided as follows: 132 volunteers from a previous study on dietary habits (**Tarallo S, et al. Gut. 2022 Jul;71(7):1302-1314. doi: 10.1136/gutjnl-2021-325168**), 76 individuals as controls in a study on colorectal cancer (**Ferrero G, et al. Oncotarget. 2017 Dec 14;9(3):3097-3111. doi: 10.18632/oncotarget.23203**) and 127 healthy individuals on gluten-free diet with no dietary restrictions (no published data).

The main results from the following manuscripts are included in the present thesis:

Study I.

Pardini B, Ferrero G, Tarallo S, **Gallo G**, Francavilla A, Licheri N, Trompetto M, Clerico G, Senore C, Peyre S, Vymetalkova V, Vodickova L, Liska V, Vycital O, Levy M, Macinga P, Hucl T, Budinska E, Vodicka P, Cordero F, Naccarati A (2022) A fecal miRNA signature accurately distinguishes colorectal cancer and adenomas: evidence from a large-scale small RNA sequencing investigation on different populations and multiple biospecimens [In review]

Study II.

Gagliardi A, Francescato G, Ferrero G, Birolo G, Tarallo S, Francavilla A, Piaggeschi G, Di Battista C, **Gallo G**, Realis Luc A, Sacerdote C, Matullo G, Vineis P, Naccarati A, Pardini B (2022). The 8q24 region hosts miRNAs altered in biospecimens of colorectal and bladder cancer patients. *Cancer Med.* 2022 Nov 10. doi: 10.1002/cam4.5375. Epub ahead of print

Study III.

Francavilla A, Gagliardi A, Piaggieschi G, Tarallo S, Cordero F, Pensa RG, Impeduglia A, Caviglia GP, Ribaldone DG, **Gallo G**, Grioni S, Ferrero G, Pardini B, Naccarati A. Faecal miRNA profiles associated with age, sex, BMI, and lifestyle habits in healthy individuals. *Sci Rep.* 2021 Oct 19;11(1):20645. doi: 10.1038/s41598-021-00014-1

Study IV.

Gallo G, Gagliardi A, Ferrero G, Francescato G, Di Battista C, Tarallo S, Francavilla A, Pardini B, Naccarati A and Conesa Zamora P (2022) Combining RNA-seq and small RNA-seq analyses in colorectal cancer subtypes and their reflection in surrogate biospecimens. [Submitted]. Results were presented as oral presentation at the “VIII Jornadas de Investigación y Doctorado UCAM: Ética en la Investigación Científica” June 24, 2022 (Sala 1 | Pabellón 9 Aula 6 Ciencias Médicas | Odontología)

The following studies, concerning an update of the literature on the main topic, represented the *primum movens* and the rationale of the thesis

Study V.

Gallo G, Vescio G, De Paola G, Sammarco G. Therapeutic Targets and Tumor Microenvironment in Colorectal Cancer. *J Clin Med.* 2021 May 25;10(11):2295. doi: 10.3390/jcm10112295

Study VI.

Gallo G, Naccarati A, Conesa-Zamora P (2022). The role of Integrated Multi-omics Data Analyses for the management and treatment of Colorectal Cancer. [Submitted]

I - INTRODUCTION

I - INTRODUCTION

CRC is a heterogeneous disease, molecularly and anatomically, that develops in a multistep-process requiring the accumulation of several genetic and epigenetic mutations that lead to the gradual transformation of normal mucosa into cancer.

The worldwide burden of CRC is predicted to increase to more than 2.2 million new cases and 1.1 million fatalities by 2030, while there were 861700 CRC-related deaths in 2018 [1].

Over 70% of CRC cases are sporadic, 20% of cases have an associated hereditary component, and less than 5% of cases are inherited (Lynch Syndrome, 2-5%).

There are currently 3 main routes of CRC carcinogenesis: chromosomal instability, DNA replication errors and epigenetic regulation, which includes aberrant hypermethylation and gene silencing [2]. In fact, recent genome-targeting investigations confirmed that each patient is genetically and epigenetically unique.

At the transcriptional level, several classification schemes have identified different biologically subtypes of CRCs. The recent identification of four CMS has provided evidence that the expression subtypes have clinical relevance independent of cancer stage [3]

Often CRC becomes symptomatic in the more advanced stage of the disease and for this reason the patient's probability of survival increases only as the diagnosis is made in the early stages.

In fact, the mortality rate varies greatly depending on the stage of the disease and only 80% of patients can be potentially cured.

However, after thirty years from the first description of a non-invasive screening CRC method, the current focus concerns the use of molecular biomarkers, e.g. KRAS mutation [4].

Currently, colonoscopy and faecal immunochemical test (FIT) represents the most frequently used combination worldwide for CRC screening. Unfortunately, colonoscopy is uncomfortable and patient compliance is not particularly favorable, with complications occurring in between 3% and 16% of cases.

In this context, Computed Tomography (CT) colonography has a higher acceptability in comparison with colonoscopy, but its effectiveness is still debated. Furthermore, FIT has a relatively low sensitivity, and colonoscopy may fail in diagnosing lesions below 6 mm diameter.

The identification of the “ideal biomarkers”, for CRC screening, diagnosis and treatment remains a high priority. Due to the rapid increase in the availability of patient data, there is significant interest in precision medicine that could facilitate the development of a personalized treatment plan for each patient on an individual basis. In this context, gut microbiota which has emerged as a central player mechanistically linking various risk factors to CRC pathogenesis, have added even more complexity to the study of CRC.

Gastrointestinal disorders are often heterogeneous [e.g., malignancy, inflammatory bowel disease (IBD)] with a wide range of clinical phenotypes depending on age of onset, severity, natural course of disease, association with other diseases and treatment response. Big Data analysis allows for the subclassification of a disease entity into distinct subgroups (i.e., phenomapping), which enhances understanding of disease pathogenesis, as well as the development of more precise predictive models of disease outcomes [5]. The use of only clinical and laboratory data (as in traditional clinical research) in predicting disease course, outcome and treatment response may not achieve a high accuracy

In fact, personalized cancer treatment requires comprehensive genetic information of individual cancers.

While isolated analysis of genomic data types is of clinical value, an integrated and comprehensive analysis of multiple genomic data types from individual cancers leverages the predictive power of each data type and allows a better understanding of the complex molecular networks that drive tumor behavior at systemic level. Such information is extremely valuable in not only developing therapeutic strategies, but also predicting tumor response to specific treatment modalities for individual cancers.

Consequently, the concomitant analysis of DNA (genomics), RNA (transcriptomics), proteins (proteomics), metabolites (metabolomics) and images (radiomics) may provide a more representative evaluation of tumor heterogeneity improving CRC diagnosis and treatment response.

1.1 MicroRNAs

Recently, the field of epigenetics has significantly grown due to the discovery of new high-throughput miRNA profiling platforms such as NGS allowing genome-wide miRNA and mRNA expression analyses.

miRNAs are a class of short endogenous single-stranded non-coding RNAs that are 18-25 nucleotides in length and are able to post-transcriptionally repress gene expression by binding to the 3' untranslated region (UTR) of their target mRNAs. Interestingly, microRNAs can be found in several biospecimens including stools and extracellular vesicles (EVs) [6, 7]. In fact, miRNA in EVs already represent a promising biomarker in liquid biopsy. In particular, EVs are nanosized, membrane-bound vesicles released from almost all type of cells and contain proteins, lipids, nucleic acids, and membrane receptors of the cells from which they originate.

Faecal miRNAs have been shown to correlate with tumour stage, due to both their continued release into the intestinal lumen by CRC cells and their detection in stool samples. In fact, the rationale for using microRNAs in CRC is precisely the direct contact of the stools with the intestinal wall [8].

Koga et al [9] studied exfoliated colonocytes by comparing miRNAs expression in 197 patients with CRC and 119 healthy controls. The authors demonstrated that the miRNA-17-92 cluster and miRNA-135 were most highly expressed in patients with CRC ($p < 0.0001$). Several other studies reported the discriminatory power of microRNAs in stool samples, although the majority of studies have investigated a small fraction of the whole human miRNome [10].

1.2 The 8q24 locus

The 8q24 locus has been described for years as a “gene desert” due to the relatively low number of protein-coding genes mapped in the region [11]. Recently, several genome-wide association studies (GWAS) have identified in this

region a considerable number of genetic variants linked to susceptibility to different cancers, including prostate, breast, colon and many others [12, 13].

In particular, some genetic elements, which resides in 8q24, such as *MYC* oncogene but also genes (*FAM84B*, *GSDMC*, *FAM49B*, and *ASAP1*) and pseudogenes connected to the tumorigenesis have captured the attention of researchers [14-18]. Moreover, this region hosts a large number of other small non-coding RNAs even if a a deep characterization of miRNA profiles is still missing.

1.3 Big Data

Intratumoral heterogeneity is an important obstacle for effective diagnosis and treatment [19]. In this context, the use of omics technologies (epigenomics, transcriptomics, proteomics, metabolomics, pharmacogenomics, radiomics) is becoming increasingly popular with the potential to contribute in a different way in advancing to our understanding of the molecular basis and cellular changes occurring in CRC.

The concept of Big Data was introduced in late 1990s by Michael Cox and David Ellsworth [20] but Francis X. Diebold [21], in 2000, was the first to give an appropriate definition of Big Data, i.e. “explosion in the quantity (and sometimes, quality) of available and potentially relevant data”.

There is currently no consensus on the core characteristics of Big Data. Over the years there have been several changes and from the initial 3v, volume, velocity and variety [22], we have moved on to the 5v by adding veracity and value. In healthcare, volume and variety refer to the huge amount of different and heterogeneous medical data that are often stored in different data formats. The velocity concerns the speed of data generation. Veracity refers to the quality of data that is to be analyzed and can be influenced by inconsistencies, missing data, ambiguities, deception, fraud, duplication, or latency [23]. Last but not least, the value represents the benefit that the use of Big Data should bring in terms of medical decision by the clinician.

II - RATIONALE

II - RATIONALE

Specific microRNAs detected in surrogate tissues, such as stool samples, are shown to be promising diagnostic biomarkers in patients with CRC. However, new studies are necessary to establish the sensibility and specificity of the individual microRNAs in order to use them in clinical practice.

Large sample size is key to success in genome wide approaches. The application of Big Data analysis in healthcare research has revolutionized clinical study approaches. Recent developments in computational biology have driven the integration of big data and molecular networks using the principles of systems biology and machine learning. Machine learning algorithms provide the means and opportunity to investigate large amounts of data and thus help identify patterns behind complex medical conditions. These analytical approaches allow categorization of patients based on their specific differences through screening a patient's genome, transcriptome, proteome, epigenome, immunome and microbiome.

III - OBJECTIVES

III - OBJECTIVES

There is increasing evidence to support the use of molecular biomarkers in CRC for the diagnosis as well as tailoring of adjuvant and neoadjuvant treatment to individual patients with both economic and clinical benefits.

Assessing variation only of a single omic data type can miss complex models that require variation across multiple levels of biological regulation. Data integration approaches can provide a key to making sense of greater complexity by identifying the important genomic factors and their interactions. It also enables the study of rare exposures, rare events and long-term effects within a relatively short period of time. The huge sample size of Big Data permits subgroup analysis to investigate interactions between different variables with the outcome of interest without sacrificing statistical power.

A better understanding of diagnostic value, and of the relationships between multiple relevant markers identified by a multi-omics approach investigated in different types of biospecimens (stool, plasma, primary tissue), are the objectives of our study in order to help the clinical decision-making as well as the development of personalized-target therapeutic treatments.

In particular, the main objectives are the following:

- To test the predictivity of stool miRNA profiles and their potential use to improve screening strategies
- To explore all miRNAs profiles residing in 8q24 evaluating their expression in both tumor tissue and non-malignant adjacent mucosa of CRC patients from the whole miRNome análisis as well as validating the miRNA expression dysregulation in 8q24 in The Cancer Genome Atlas (TCGA)
- To evaluate the relationship between the stool miRNA levels and common traits (sex, age, BMI, and menopausal status) or lifestyle

habits (physical activity, smoking status, coffee, and alcohol consumption)

- To integrate data from CMS classification with small RNA sequencing and pan-cancer target assay to better describe CRC heterogeneity. The outcomes will be explored also in stool miRNome of patients in order to identify novel biomarkers in this surrogate tissue mirroring tissue alterations

IV - MATERIALS and METHODS

IV – MATERIALS AND METHODS

My PhD work, from October 2020 until September 2022, was focused on the following activities:

- Patients' clinical evaluation and enrollment (including collection of demographic characteristics such as lifestyle, dietary and anthropometric parameters);

The demographic characteristics, including past medical history, were collected and a questionnaire on lifestyle and anthropometric parameters was administered in order to better understand the complex interactions between the gut microbiome, metabolite composition, host condition and diet validated in the European Prospective Investigation into Cancer and Nutrition (EPIC) study [24]. In patients undergoing surgical treatment the pathological features were analyzed (tumor budding, lymph node involvement, circumferential resection margin, grading, type of tumor).

Patients were regularly followed-up (in accordance with recent guidelines [25, 26]).

Clinical information on surgery and/or treatments (including response to therapies, toxicity, living status) were collected for all CRC patients included.

Inclusion Criteria

- Age > 18 and < 90 years old
- Diagnosis of CRC
- Patients undergoing colorectal surgical resection with curative intent (any type of surgery)
- All patients should provide Written Informed Consent

Exclusion Criteria

- Current or previous diagnosis of other solid or hematological tumors
- Inability or refusal to give informed consent
- Inability or refusal to be regularly followed up

- Samples Collection (Plasma and Stool samples plus Pathological tissue and corresponding adjacent healthy mucosa, least 20cm proximally from the cancer, only in patients undergoing colorectal surgical resection with curative intent) from patients recruited in our outpatient clinic;

- Clinical interpretation of the data to be analyzed

During the study period several multidisciplinary meetings were held, which included colorectal surgeons, pathologists, biologists, to discuss the histopathological classification of the samples and the clinical characteristics useful for the subdivision of the samples into the various categories of analysis as described in methods section.

All the experimental and bioinformatics/statistical analyzes were performed in collaboration with the IIGM, Torino.

Study I.

Naturally evacuated fecal samples were obtained from all subjects previously instructed to self-collect the specimen at home. For all the cohorts, stool samples were collected in nucleic acid collection and transport tubes with RNA stabilizing solution (Norgen Biotek Corp.) and returned to the endoscopy unit. Stool aliquots (200µl) were stored at -80°C until RNA extraction [27].

Plasma samples were obtained from 8ml of blood centrifuged for 10min at 1000rpm, and aliquots were stored at -80°C until use. Plasma EVs were isolated from 200µl of plasma using the ExoQuick exosome precipitation solution (System Biosciences, Mountain View, CA, USA) according to the manufacturer's instructions. Paired tumor/adenoma tissue and adjacent non-malignant mucosa

(at least 20cm distant) were obtained from CRC and adenoma patients during surgical resection and immediately immersed in RNA later solution (Ambion). All samples were stored at -80°C until use.

The Study Design is shown in **Figure 1**.

Italian (IT) cohort – Stool specimens, clinical and demographic data were collected from 219 subjects recruited in a hospital-based study at one hospital in Vercelli, Italy. Based on colonoscopy results, participants were classified into: (i) 62 CRC patients (individuals with newly diagnosed sporadic CRC); (ii) 40 polyp patients, stratified in hyperplastic polyps (n=6), non-advanced adenomas (nAA, n=14), or advanced adenomas (AA, n=20); (iii) 36 subjects with a Gastrointestinal (GI) disease, including inflammatory bowel disease (IBD, including Crohn's disease, or ulcerative colitis), or diverticular disease; and (iv) 81 healthy subjects with negative colonoscopy.

Czech (CZ) cohort – Stool specimens, clinical and demographic data were collected from 162 Czech individuals recruited in two hospitals in Prague and one in Plzen, Czech Republic. Based on colonoscopy results, subjects were divided in: (i) 66 CRC patients, (ii) 28 individuals with colorectal polyps grouped in hyperplastic polyps (n=9), nAA (n=13), and AA (n=6); (iii) 32 patients with other GI diseases; and (iv) 36 healthy subjects.

Validation cohort – Stool specimens from 141 CRC patients recruited in the hospital in Brno, Czech Republic [28] and 50 stool samples of healthy subjects [29, 30] were included.

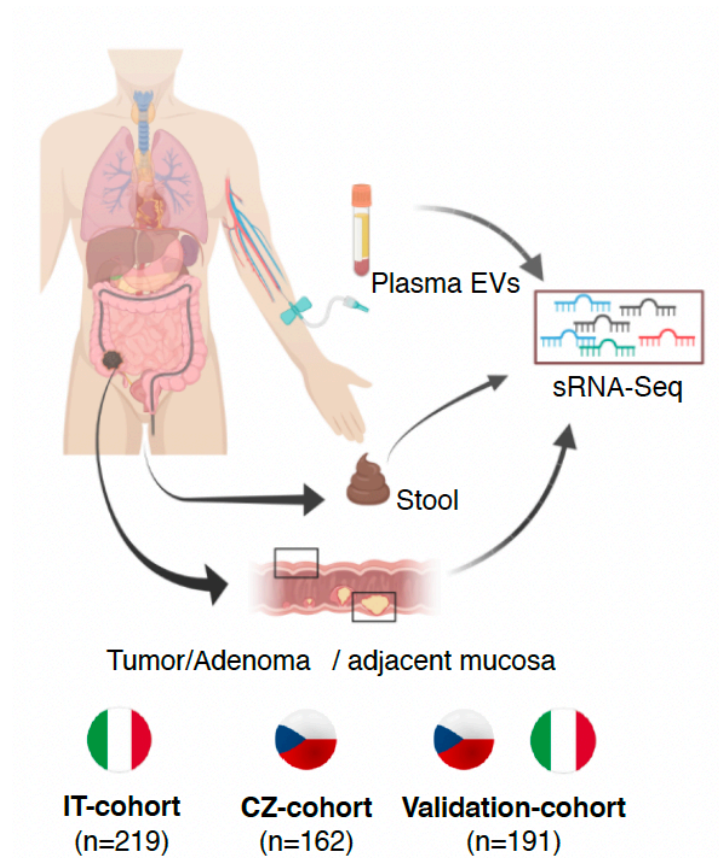


Figure 1. Representation of the study design.

Other analyzed bio-specimens

For 135 subjects submitted to a surgical procedure at Vercelli hospital, primary tumor (n=105) or adenoma (n=30) tissues paired with adjacent colonic mucosa were collected. Among them, 69 (51 CRC and 18 adenomas) provided stool and blood samples and were included in the IT-cohort.

Blood samples were collected from 209 subjects of IT-cohort stratified in patients with CRC (n=52), AAs (n=19), nAAs (n=14), hyperplastic polyps (n=6), other GI diseases (n=34), and healthy subjects (n=79).

The rationale for the study was based on the following phases:

- stool miRNA biomarkers discovery

- stool deregulated miRNA (DEmiRNA) analyses in different contexts

The following analyzes were performed:

- **Total RNA extraction**

Total RNA was extracted from stool samples and leftovers from FIT tubes using the Stool Total RNA Purification Kit (Norgen Biotek Corp) as previously described [27]. Total RNA from plasma EVs was extracted as described in [31]. For tissue samples, total RNA was extracted using Trizol reagent (Thermofisher), according to the manufacturer's instructions.

- **Library preparation for sRNA-Seq**

Small RNA transcripts were converted into barcoded cDNA libraries for Illumina sequencing protocol as previously described [29]. The obtained libraries were subjected to 75 cycles of single-end Illumina sequencing pipeline on HiSeq 2000 sequencer, (Illumina Inc., USA).

- **sRNA-Seq bioinformatics and statistical analysis**

sRNA-Seq pipeline analyses were performed using a previously published Docker-embedded software to guarantee the computational reproducibility of the analysis [27]. Trimmed reads were mapped against an in-house curated reference of human miRNAs based on miRbase v22. The age- and sex-adjusted differential expression analysis was performed using DESeq2 R package v1.22.2 [32]. For tissue samples, to test the significance of miRNA differential expression levels between CRC/adenoma tissue and matched adjacent non-malignant colonic mucosa, a paired DESeq2 analysis was applied. A miRNA was considered differentially expressed if associated with an Benjamini Hochberg (BH) adjusted p-value < 0.05 and a median number of reads > 20 in at least one study group.

Functional enrichment analysis was performed with RBiomirGS v0.2.12 [33] in default settings and considering the validated miRNA-target interactions. A term was considered enriched if associated with a BH adj. p<0.05 and at least two target genes.

Starting from DESeq2-normalized miRNA levels, a feature selection was performed to define the minimal miRNA predictive set by considering an

increasing number of features prioritized from ANOVA F-Test or trained logistic regression coefficients applied to the training set (70% of IT and CZ-cohorts). Each miRNA set was tested concomitantly using a stratified 10-Fold Cross-Validation procedure with four methods (Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machines). The smallest miRNA set providing the highest AUC was isolated and its predictive power (average AUC) was tested by Stratified 10-Fold Cross-Validation repeated 500 times. The identified signature was tested on 30% of IT- and CZ-cohorts excluded from the training and on the Validation cohort.

The statistical analyses between continuous variables were performed using Wilcoxon Rank Sum test or Kruskal-Wallis's test while chi-squared test was used for categorical variables.

Study II.

Tissues. Primary tumor and adjacent normal mucosa (at least 20cm distant) tissues from CRC subjects (Study I) were collected during surgical resection and immediately transferred in cryogenic vials with RNAlater™ Solution (Invitrogen, Milan) and stored at -80°C until use.

Stool. Naturally evacuated faecal samples were obtained from all patients of Study I, previously instructed to self-collect the specimen at home before the colonoscopy. Stool samples were collected in stool nucleic acid collection and transport tubes with RNA stabilizing solution (Norgen Biotek Corp.). Aliquots (200µl) were stored at -80°C until RNA extraction.

Urine. Urine samples were collected in the morning from all the participants in the Study II and stored at 4°C until they were centrifuged at 3000 g for 10 min. The urine supernatant aliquots were then transferred in tubes and stored at -80°C until use.

Plasma. For both study cohorts, plasma samples were obtained from 5-8mL of blood, centrifuged at 1000 rpm for 10 minutes. From each tube of blood, about 1-2ml of plasma was obtained: 200µl aliquots were stored at -80°C until use. EVs were isolated from 200µL of plasma using the ExoQuick™ Exosome Precipitation Solution (System Biosciences, Mountain View, CA, USA) according to manufacturer's instructions. Briefly, plasma was mixed with 50.4 µL of

ExoQuick™ Solution and refrigerated at 4°C overnight (at least 12h). The mixture was then centrifuged at 1500g for 30min. The EVs pellet was dissolved in 200µL of nuclease-free water and RNA was extracted immediately from this solution.

The study I population consisted of 200 subjects (89 women and 111 men) recruited at the Clinica S. Rita in Vercelli, Italy. Based on colonoscopy results, participants were classified as healthy controls (80 subjects with negative colonoscopy for tumor or other GI disorders) or CRC patients (120 subjects)[27]. CRC patients were also stratified according to the localization of the tumor (colon/sigma-region or rectum), the stage of the cancer (stage 0-IV) and its grade (G1-G3)[34] Stool and blood samples were collected at the time of recruitment and, for CRC cases only, tissue pairs of primary tumor and adjacent normal mucosa were also collected at the surgery.

The study II consisted of a set of subjects recruited in the context of previous research [35] nested in the Turin Bladder Cancer Study [36, 37] and consisting of 116 men. For all the subjects, urine and plasma samples were collected. The results of the latter population were not reported in the present project.

The following analyzes were performed:

- RNA extraction

Extraction of total RNA from stool, urine, plasma EVs and tissues was performed using appropriate kits/methodologies for total RNA purification according to the specimen to be analysed.

RNA from tissues was isolated using QIAzol (QIAGEN, Hilden, Germany) after tissue homogenization performed with ULTRA-TURRAX® Homogenizer [37], followed by phenol/chloroform extraction according to the manufacturer's standard protocol.

Total RNA from stool samples was extracted with the Stool Total RNA Purification Kit (Norgen Biotek Corp., Canada) following the manufacturer's standard protocol. Total RNA from plasma EVs was extracted with the miRNeasy Plasma/Serum Mini-kit (Qiagen, Hilden, Germany) using the QIAcube extractor (QIAGEN, Hilden, Germany). Total RNA from urine samples was extracted with

the Urine microRNA Purification Kit (Norgen Biotek Corp., Canada), following the manufacturer's standard protocol. The RNA concentration was quantified by Qubit™ 4 fluorometer with Qubit™ microRNA or RNA Broad range Assay kits (Invitrogen, Monza, Italy).

- Library preparation for sRNA-seq [30, 38]

sRNA-seq libraries were prepared from RNA extracted from tissues, stool, plasma EVs, and urine. Small RNA transcripts were converted into barcoded cDNA libraries using NEBNext® Multiplex Small RNA Library Prep for Illumina® (New England Biolabs, Inc., Ipswich, MA, USA). For each library, 6µL of RNA (35ng for EVs RNA, and 250ng for tissue/stool/urine RNA) were used in all the experimental procedures as starting material. Each library was prepared with a unique indexed primer. Multiplex adaptor ligations, reverse transcription primer hybridization, reverse transcription reaction and PCR amplification were performed according to the manufacturer's protocol. Further details concerning cDNA constructs and final libraries preparation are described in [30].

The obtained libraries were subjected to the Illumina® sequencing pipeline, passing through clonal cluster generation on a single-read flow cell (Illumina Inc., San Diego, CA, USA) by bridge amplification on the cBot (TruSeq SR Cluster Kit v3-cBOT-HS, Illumina, Inc., San Diego, CA, USA) and 50 cycles sequencing-by-synthesis on the HiSeq™ 2000 Sequencing System (Illumina, Inc., San Diego, CA, USA) (in collaboration with EMBL, Gene core facility, Heidelberg, Germany).

- Library preparation for total RNA-seq

Before RNA-seq library preparation, total RNA from tissue samples was cleaned up and DNase-treated with the RNA Clean & Concentrator™-5 kit, following manufacturer's protocol (Zymo Research, USA) to remove all traces of DNA. Next, the quality of the input RNA was determined by RNA Integrity Number (RIN) measurement obtained by running the samples on an Agilent Bioanalyzer® RNA 6000 Nano chip (Agilent Technologies, Milan, Italy). For each sample, 500ng of RNA was used as starting material to libraries preparation. RNA-seq libraries were prepared with the NEBNext® Ultra II Directional RNA Library Prep for

Illumina® kits (New England Biolabs, Ipswich, MA, USA) after ribosomal RNA depletion, following manufacturer's instructions. The generated barcoded libraries of about 300bp fragments were run on an Illumina® NovaSeq™ 6000 platform (Illumina, Inc., San Diego, CA, USA).

- DNA extraction

For DNA extraction, tissues were initially homogenized in a homogenization solution (Promega, Milan) and then processed with Maxwell® RSC Tissue DNA Kit (Promega, Milan). Before loading samples onto Maxwell® RSC Cartridges, 300µL of Lysis Buffer and 30µL of Proteinase K were added to the homogenized samples, and further incubated for 20min at 56°C. DNA was quantified with Qubit™ 4 fluorometer using Qubit™ dsDNA High Sensitivity Assay Kit (Invitrogen, Carlsbad, CA, USA).

- TruSight™ Oncology 500 High-Throughput (TSO500-HT).

DNA libraries were prepared using the hybrid capture based TruSight™ Oncology 500 High Throughput (TSO500-HT) Library Preparation Kit (Illumina, San Diego, CA, USA) following Illumina® TruSight™ Oncology 500 Reference Guide (document # 100000094853 v02). In brief, the genomic DNA was fragmented using the Covaris® ME220 focused-ultrasonicator (Covaris, Woburn, MA) for 10 seconds at 50 watts. After end repair, A-tailing, and adapter ligation, the adapter-ligated fragments were amplified using primers to add index sequences for sample multiplexing. Libraries were enriched through two hybridization/capture steps using specific probes: a pool of oligos specific to 523 genes targeted by TSO500-HT was hybridized to the DNA libraries overnight. Next, streptavidin magnetic beads were used to capture probes hybridized to the targeted regions. PCR amplification, cleanup, and quantification of the enriched DNA using Qubit™ dsDNA HS Assay Kit (Invitrogen, Carlsbad, CA, USA) were the final steps. Following pooling and denaturation, libraries were diluted to the appropriate loading concentration and finally sequenced on Illumina® NovaSeq™ 6000 Sequencer (Illumina, Inc., San Diego, CA, USA) (read length of 200 bp paired end).

- 8q24 region miRNA profiles from TCGA data

The locus 8q24 is a genomic region enriched in cancer-associated polymorphisms, it has been described as a “gene desert” due to sparse presence of protein-coding genes [39]. Nevertheless, this locus hosts the MYC oncogene and genetic elements connected to tumorigenesis, such as pseudogenes and long non-coding RNAs. This region hosts also many genes coding for 30 miRNAs genes, scarcely investigated so far. Research on such miRNAs may provide new insights to characterize the multiple-cancer associated variants annotated in this genomic region.

Data of the TCGA related to CRC (COAD, colon adenocarcinoma and READ–rectum adenocarcinoma) and BC (BLCA–Bladder carcinoma) were retrieved from TCGA Data portal (v.29.0). For each project, the isoform_expression_quantification.txt file, containing information on miRNA expression levels as raw counts, was downloaded along with information on the coordinates for each miRNA and their accession number on miRBase (v.22). Raw counts were then normalized using the DESeq2 package (v.1.28.1) [32] for the statistical software R (v. 4.0.2). TCGA-COAD and TCGA-READ data were merged in the same dataset of CRC patients. Files containing clinical data, information on patient biospecimen, and metadata were downloaded from the TCGA Data portal (v.29.0). To prepare the count matrix, each mature miRNA name was retrieved with the use of the ‘miRBaseConverter’ [40] R package. Data were filtered to keep only white-Caucasian individuals. Differential expression analyses were performed initially on tumor tissue samples and paired adjacent mucosa and then considering also non-matched tumor samples.

- miRNA targets functional enrichment analysis

To perform a functional enrichment analysis, miRNA target genes were retrieved using miRWalk database (v 2.0) as previously described in Sabo et al. [31]. Only miRTarBase validated interactions involving miRNAs targeting the gene 3’UTR and associated with a score greater or equal than 0.95, were retained. Target genes were subset based on miRNAs log₂ fold-change (log₂FC, up- or down-regulated) and separately analysed with the Metascape web tool [41] to retrieve the enriched functional terms.

Study III.

Sample collection was performed as described in Studies I-II.

We collected stool samples from healthy donors participating in different studies (**Figure 2**). Briefly, 132 volunteers were recruited from a study investigating the role of different dietary habits described in Tarallo et al. (Study 1) [29], 76 individuals were recruited as controls in a study on colorectal cancer (i.e., negative at colonoscopy for any other gastrointestinal diseases) (Study 2) [38], and 127 individuals from a comparative study (healthy individuals either on gluten-free diet or with no dietary restrictions).

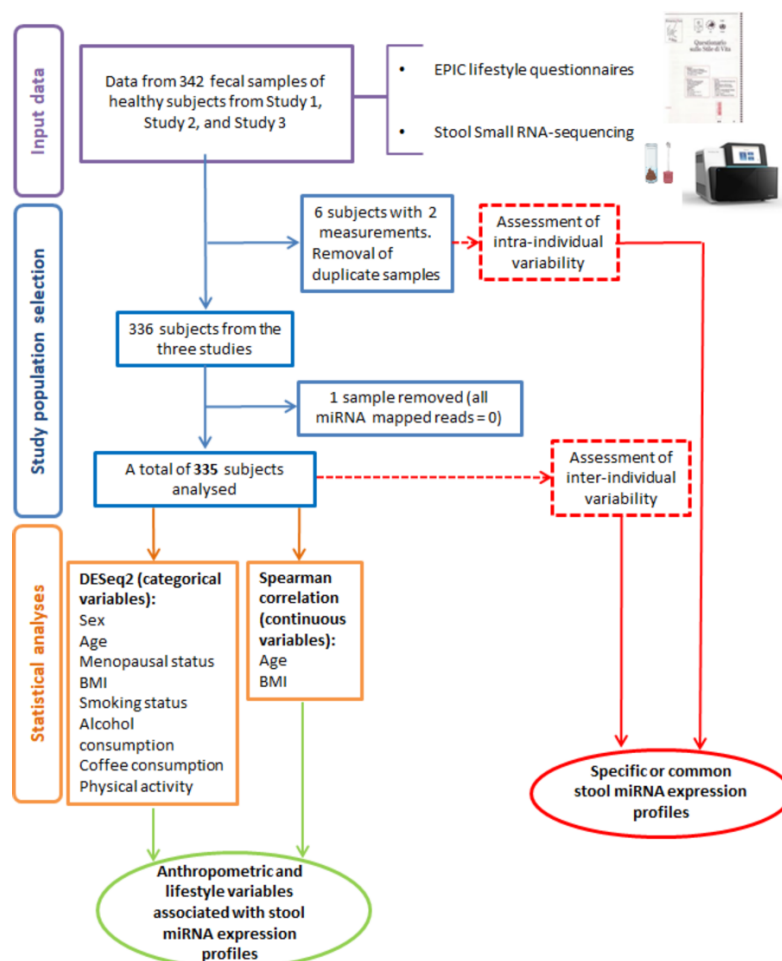


Figure 2. Workflow of the study

The rationale for the study was based on the need to correlate miRNA expression levels in stool with sex, age, BMI, smoking, alcohol and coffee consumption, and physical activity in order to understand their possible modulatory effects on the human faecal miRNome.

The following analyzes were performed:

- Total RNA extraction from stool

Total RNA was extracted from 200 µl faecal aliquots with the Stool total RNA purification kit (Norgen Biotek Corp) using the protocol recommended by the manufacturer. RNA quality and quantity were verified according to the MIQE guidelines ([http:// miqe. gene- quant ifica tion. info/](http://miqe.gene-quantification.info/)). For all samples, RNA concentration was quantified by Qubit fluorometer with a Qubit microRNA assay kit (Invitrogen).

- Library preparation for sRNA-Seq

The present step was performed according to the previous described studies.

- Analysis of miRNAs from sRNA-seq data

A full description of miRNA data analysis is detailed elsewhere [30, 40, 42].

- Classification criteria for common traits and lifestyle habits

Information on smoking status, alcohol and coffee consumption, and physical activity were collected from the quantitative and qualitative EPIC dietary and lifestyle questionnaires [24] whereas individual characteristics (i.e. age, height and weight) were reported in a baseline questionnaire. For women, menopausal status information (premenopausal and postmenopausal) was also included in EPIC lifestyle questionnaire.

Considering the age distribution of the study population, miRNA profiles were investigated comparing three categories: <37 (n=122), 37-53 (n=111) and >53 (n=102) years old.

Study IV.

Sample collection was performed as described in Studies I-II.

In this study, paired tissue specimens were collected from 115 CRC patients (tumor and adjacent normal mucosa). RNA-seq and sRNA-seq were performed on each sample. After the CMS classification, each subtype was genomically characterized with the TruSight Oncology 500 (TSO) cancer-panel. Specific genes and miRNA signatures of each CMS were identified using differential expression analysis.

The rationale for the study was based on the integration of data from CMS classification with sRNA sequencing and pan-cancer target assay to better describe CRC heterogeneity.

The following analyzes were performed (according to the previous described studies):

- **Library preparation for sRNA-seq**
- **Library preparation for total RNA-seq**
- **TruSight™ Oncology 500 High-Throughput**

The TSO 500 was employed as a comprehensive NGS assay targeting 523 full coding gene regions implicated in solid tumors pathogenesis [44]. Besides somatic mutations, single nucleotide variants, indels, amplifications (1.94 Mb genomic content) tumor mutation burden (TMB) and microsatellite instability (MSI) were assessed. Libraries from FFPE samples or tissues in RNA later (40ng of DNA in input) were prepared as manufacturer's protocol and sequenced on NovaSeq 6000 System.

- **miRNA targets functional enrichment analysis**

Ethical Approval

Local ethics committees (Universidad Católica San Antonio de Murcia (UCAM); Azienda Ospedaliera SS. Antonio e Biagio e Cesare Arrigo di Alessandria; Institute of Experimental Medicine in Prague; Masaryk Memorial Cancer Institute and Masaryk University in Brno) approved the study. All patients gave written informed consent following the Declaration of Helsinki prior to participating in the study.

V - RESULTS

V - RESULTS

Study I.

Study Population (**Figure 1**) included stool specimens, clinical and demographic data from an IT Cohort (219 subjects recruited in a hospital-based study at one hospital in Vercelli), a CZ cohort (162 CZ individuals recruited in two hospitals in Prague and one in Plzen) and a Validation cohort (141 CRC patients recruited in the hospital in Brno, CZ Republic and 50 stool samples of healthy subjects from Italy).

The mean age of cases in the IT Cohort was 58.5 (39-84) years, 55.9 (30-82) years, 65.4 (42-93) years and 71.1 (54-87) years, respectively, in Healthy, GI disease, Polyps and CRC patients ($p < 0.001$). Overall, 56.6% of patients were male (124/219) but without statistically significant differences ($p = 0.079$).

Moreover, the mean age of cases in the CZ Cohort was 57.8 (40-76) years, 58.7 (41-75) years, 63.1 (48-82) years and 68.0 (40-88) years, respectively, in Healthy, GI disease, Polyps and CRC patients ($p < 0.001$) with 55.5% of male patients (90/162) ($p = 0.17$).

In both groups, there were no statistically significant differences regarding BMI (IT $p = 0.096$; CZ $p = 0.16$). Conversely, while the smoking status was homogeneous in the IT Cohort ($p = 0.55$), there was a significant difference in the CZ Cohort ($p = 0.025$).

Stool miRNA profiles

An average of 267 ± 112 miRNAs were detected in each stool sample by sRNA-Seq. Differential expression analysis, adjusting for age and sex, was performed between CRC and healthy subjects independently in each of the two data sets, the IT-cohort and the CZ-cohort. A total of 116 and 30 significant DE miRNAs (median expression >20 reads and adj. $p < 0.05$) were detected in the IT-cohort and the CZ-cohort, respectively (Figures).

The two cohorts shared 20 common stool DEmiRNAs (**Figure 3**), all with a similar trend of expression (17 up-regulated and 3 down-regulated, $\rho=0.83$, $p<0.001$, **Figure 3**).

DEmiRNA profiles were further explored in relation to clinical data of cancer patients (**Figure 4**). The levels of three down-regulated miRNAs significantly decreased with increasing tumor grade and size (T). miR-607 also significantly decreased in patients with advanced stages of the disease or lymph node invasion (**Figure 4**). On the contrary, the levels of 10 CRC-up-regulated miRNAs significantly decreased in patients with metastatic disease.

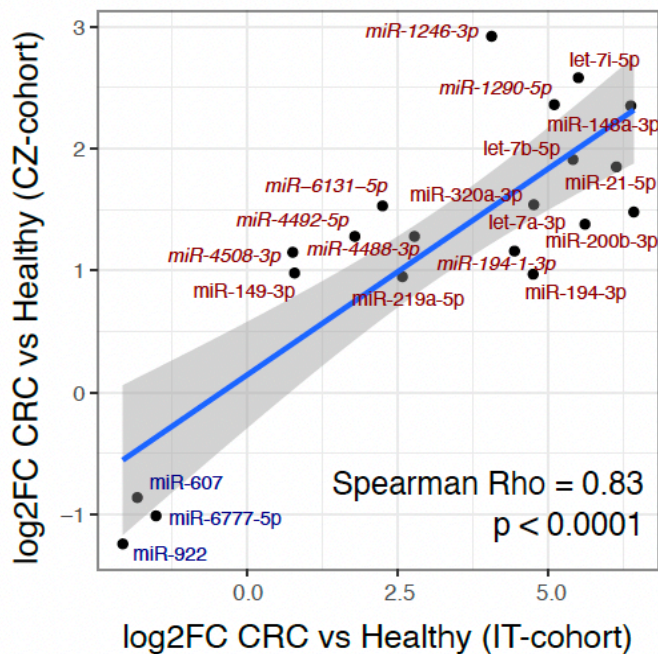


Figure 3. Scatterplot reporting the correlations of log2FC of the 20 stool DEmiRNAs from the comparison between CRC and healthy subjects in common between the IT-cohort (x-axis) and the CZ-cohort (y-axis). In red are represented the up-regulated miRNAs, in blue the down-regulated ones.

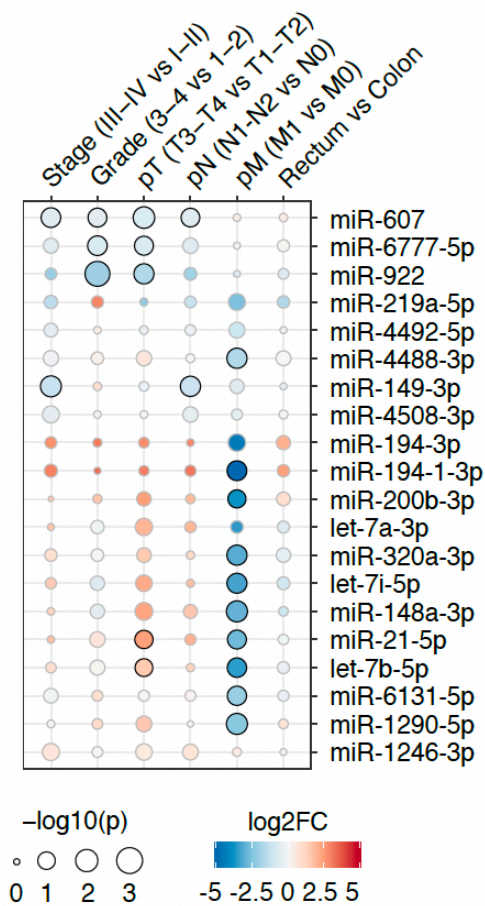


Figure 4. DEMiRNA levels comparing CRC patients stratified for clinical data. The color of the dot is related to the log₂FC while the -size is proportional to the statistical significance.

A fecal miRNA predictive signature distinguishes CRC patients from healthy individuals

From the integrated machine-learning pipeline, the best miRNA predictive signature which accurately distinguished CRC patients from healthy controls was composed of five miRNAs (miR-607, miR-6777-5p, miR-4488-3p, miR-149-3p, and miR-1246-3p, AUC=0.83).

By stratifying patients for CRC stage, the same five miRNA signature accurately distinguished both stages III-IV patients (Validation cohort, AUC=0.89) but also stages I-II patients from from healthy subjects (Validation cohort, AUC=0.85). Results remained similar even not including age and sex in the analysis.

DEmiRNA profiles in tumor tissue and adjacent mucosa

A differential expression analysis was performed between paired tumor tissues and matched adjacent mucosa collected of 105 CRC patients. Eleven miRNAs among the 20 DEmiRNAs were also significantly differentially expressed (adj. $p < 0.05$) in tumor tissue (Figure 5), with seven miRNAs (miR-21-5p, miR-1246-3p, miR-1290-5p, miR-148-3p, miR-4488-3p, miR-149-3p, miR-219-3p) up-regulated in tumor tissues coherently with their increase in stool of CRC patients.

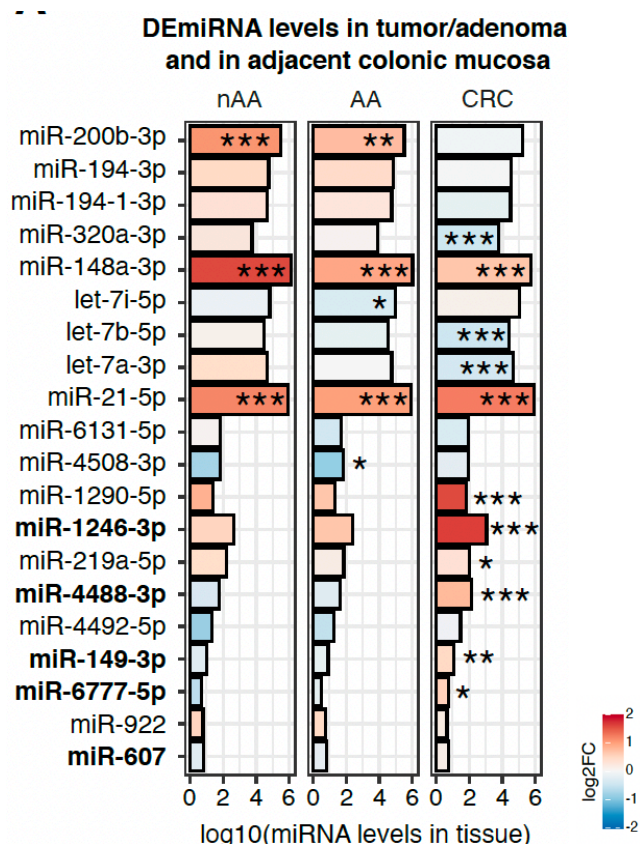


Figure 5. Characterization of the 20 fecal DEmiRNAs in different sample types. Bar plot reporting the median expression levels in tumor, advanced (AA) and non-advanced adenoma (nAA) tissues. The color code represents the log₂FC from the paired differential expression analysis between CRC/adenoma tissues and matched adjacent mucosa *** , adj. $p < 0.001$; **adj. $p < 0.01$; *adj. $p < 0.05$.

DEmiRNA profiles in circulating EVs

sRNA-Seq was performed on RNA isolated from EVs obtained from plasma samples collected from 209 subjects in the IT-Discovery cohort, detecting an average of 292 ± 57 miRNAs in these samples. Among the 20 DEmiRNAs identified in stool samples of CRC patients, only miR-4488-3p emerged as significantly dysregulated also in plasma EVs, albeit associated with low median levels (<10 normalized reads)

Study II

The Study I cohort consisted of 120 CRC patients and 80 controls. The mean age of cases was 70.5 ± 10.5 years, with 59.2% of males. The control group mean age was 57.9 ± 11.3 years, with 40 males and 40 females. Stool samples were available for 62 patients, while plasma EVs were available for 53 cases. For the control group, stool, and plasma EVs were available for all individuals. Tissue pairs were available for 108 patients. Forty-three CRC patients had all three biospecimens collected. In this population, 53 out of the 54 known miRNAs in the 8q24 region were detected in all specimens analysed (tissue, stool, and plasma EVs samples).

Small RNA sequencing results in CRC tissue samples

Twelve 8q-related miRNAs were differentially expressed (adj p-value <0.05) in cancer tissues when compared with the adjacent colonic mucosa. Among them, nine resulted up-regulated: miR-151a-3p (log₂FC=0.23; **Figure 6A**), miR-151a-5p (log₂FC=0.22), miR-548az-5p (log₂FC=0.37), miR-548d-3p (log₂FC=0.38), miR-937-3p (log₂FC=0.84), miR-939-5p (log₂FC=0.98), miR-1302 (log₂FC=1.16), miR-4472 (log₂FC=0.32), and miR-4664-3p (log₂FC=1.76). Three DE miRNAs were down-regulated: miR-30b-3p (log₂FC=-0.31), miR-30d-5p (log₂FC=-0.23; **Figure 6B**), and miR-4662a-5p (log₂FC=-0.46)

Seven miRNAs (namely miR-30b-3p, miR-30d-5p, miR-548d-3p, miR-937-3p, miR-939-5p, miR-1302, miR-4664-3p) resulted differentially expressed in both tumors with stages I-II and III-IV in comparison to their respective adjacent tissues.

When tumors were stratified for grade, six miRNAs (miR-151a-3p, miR-548d-3p, miR-937-3p, miR-939-5p, miR-1302, and miR-4664-3p) were up-regulated in G1-G2 cancers vs adjacent mucosa. Conversely, miR-30b-3p and miR-30d-5p were

down-regulated. In G3 CRC, miR-548az-5p, miR-548d-5p, miR-937-3p, miR-939-5p, miR-1302, miR-4664-3p, and miR-4472 were significantly up-regulated, while miR-30b-5p, miR-30d-5p, and miR-4662a-5p resulted down-regulated compared to adjacent tissues.

Validation with TCGA data

8q24-related miRNAs were investigated on the TCGA dataset. Out of 52 miRNAs, four were up-regulated in tumor tissues with respect to normal adjacent mucosa (22 matched samples): miR-151a-3p ($\log_2FC=1.82$ **Figure 6C**), miR-151a-5p ($\log_2FC=0.84$), miR-30b-5p ($\log_2FC=2.84$), and miR-4662a-5p ($\log_2FC=3.84$). Conversely, miR-30d-5p ($\log_2FC=-1.11$; **Figure 6D**) and miR-937-3p ($\log_2FC=-2.23$) resulted down-regulated. miR-151a-3p (**Figure 6C**) and miR-30d-5p (**Figure 6D**) showed similar significant expression levels also when also unmatched tumor tissue samples ($n=574$) were included in the analysis.

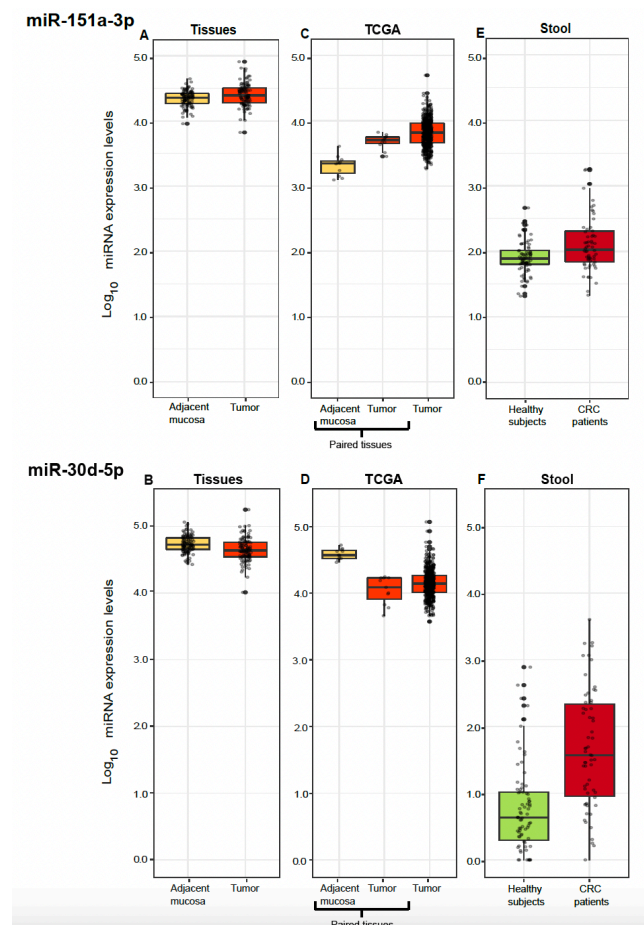


Figure 6. Box plots representing expression levels respectively of miR-151a-3p and miR-30d-5p in (A, B) CRC tissues versus normal adjacent mucosa from Study 1; (C, D) in the CRC-TCGA database (both for paired and not paired tissues); (D, E) in stool samples from CRC patients and healthy controls.

Functional enrichment analysis

In total, 213 genes are targeted by miRNAs up-regulated in tumor tissues, as retrieved from the miRWalk database.

sRNA sequencing in surrogate tissues

Stool samples. Out of the 54 miRNAs detected in the 8q24 region, four were differentially expressed in faecal samples of CRC cases in comparison with controls. Specifically, in patients, miR-6849-5p was down-regulated ($\log_2FC=-0.71$, adj p-value=0.05) while miR-151a-3p ($\log_2FC=1.40$, adj p-value<0.0001; **Figure 6E**), miR-30d-5p ($\log_2FC=4.10$, adj p-value<0.0001; **Figure 6F**), and miR-10400-5p ($\log_2FC=2.46$ adj p-value<0.0001) were up-regulated

miR-6849 was down-regulated in CRC stage I-II ($\log_2FC=-1.44$) and stage III-IV ($\log_2FC=-1.04$) when compared to healthy controls. miR-151a-3p ($\log_2FC=2.08$) and miR-1302 ($\log_2FC=-0.97$) were respectively up- and down-regulated only in stage III-IV.

After a stratification of CRC cases for tumor localization, miR-30d-5p was up-regulated both in colon/sigma ($\log_2FC=4.11$, adj p-value<0.0001) and rectal cancer ($\log_2FC=4.72$, adj p-value<0.0001) when compared to healthy controls while miR-6849-5p was down-regulated in colon/sigma patients only ($\log_2FC=-1.00$, adj p-value=0.01).

Plasma EVs samples. Only miR-30d-5p were significantly down-regulated in samples from patients with G1-G2 tumors ($\log_2FC=-0.31$, adj. p-value=0.02).

Study III.

Overall, 342 faecal samples were collected. Six patients provided a second stool sample one year after the first collection: for them, only data from the first sampling were considered in the analyses while the second sample was used to assess intra-individual variability. One patient was excluded because few sRNA-Seq reads were aligned on miRNA sequences. The final study population

consisted of 335 subjects [average age of 44.7 ± 14.7 years old (range: 18–81); 63.6% were females] provided by both stool sRNA-seq data and lifestyle questionnaires (**Figure 2**).

Stool miRNA profiles and analysis of the intra/inter-individual expression variability.

Four hundred and forty-nine miRNAs (13.8%) were detected in at least half of the analysed samples. Among them, nine miRNAs (miR-320e-5p, miR-607, miR-647-3p, miR-1246-3p, miR-1302, miR-3125, miR-5698, miR-6075, and miR-6777-5p) were detected in all the samples analysed. miR-3125 was characterised by the highest median expression levels (2,051 reads), followed by miR-6075 (921 reads), and miR-1246-3p (884 reads). Repeated samples collected from six subjects were used to assess the stability of stool miRNA expression levels over time. A second faecal sample was collected approximately one year after the first collection (min = 378 days, max = 560 days). No significant changes in lifestyle habits were reported from the questionnaires compiled by the participants on both occasions.

miRNA profiles in relation to common traits.

miRNA expression levels were analyzed in relation to sex, age, menopausal status, and BMI. Initially, miRNA profiles of 122 males and 213 females were compared and nine DE miRNAs were observed between sexes. Specifically, five were up-regulated (miR-324-3p, miR-324-5p, miR-1255b-5p, miR-3935, and miR-4675) and four down-regulated (miR-3615-5p, miR-4326, miR-4418, and miR-4632-5p) in males. The expression levels of miR-324-5p, miR-4326, and miR-4418 are reported as an example of DE miRNAs associated with sex. (**Figure 7a**).

In total, 19 DE miRNAs were identified in at least one comparison among these categories. miR-1231 and miR-4276-3p were down-regulated and miR-4487 was up-regulated in subjects of the age class 37–53 compared with those of the age class < 37. Comparing subjects of age class > 53 with those of the class < 37, 7 DE miRNAs (6 down- and miR-3169 up-regulated in the older group) were identified. Between the age classes 37–53 and > 53, 10 DE miRNAs (3 down and 7 up-regulated in the eldest group) were identified. Progressive reduced expression levels with increasing age were observed across categories for miR-4276-3p and increased for miR-3169 and miR-4505-3p (**Figure 7b**). For the 19 DE miRNAs, a similar expression pattern among age classes was observed when the analysis was performed considering males and females separately. However, only miR-550a-3-5p was significantly down-regulated in males of the age class 37–53 with

respect to the other two classes. Additionally, 13 and 6 DE miRNAs were specifically associated with age in males and females, respectively.

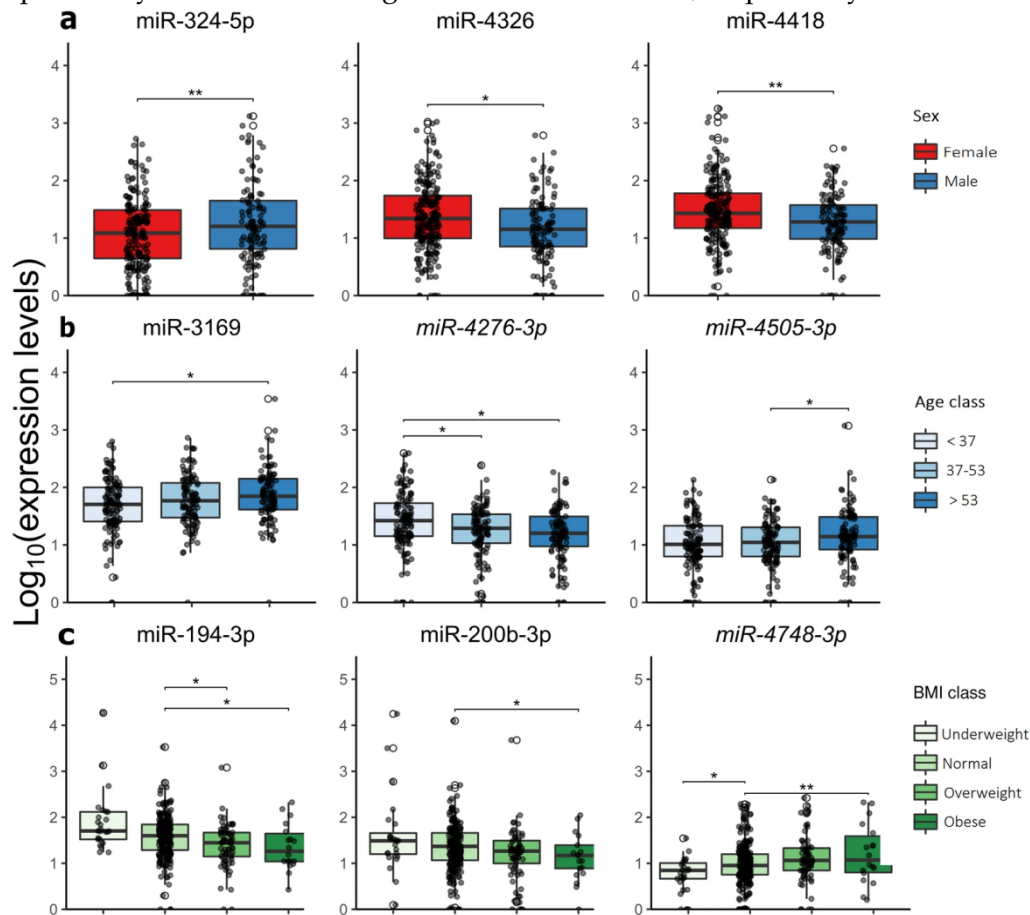


Figure 7. Box plots showing the expression levels of selected DE miRNAs among individuals stratified according to the investigated common traits features. P values were computed using DESeq2 and adjusted using the FDR method. ***adj. p value 0.001, **adj. p value 0.01, *adj. p value 0.05.

miRNA profiles according to lifestyle habits (Figures 7 and 8)

miRNA levels were further investigated in relation to smoking status, alcohol, and coffee consumption as well as physical activity.

For smoking habits, miRNA levels were analysed comparing subjects who smoked more than 16 cigs/day ($n = 16$), those smoking less than 16 cigs/day ($n = 41$) and former smokers ($n = 94$) with never smokers ($n = 181$). Overall, 84 DE miRNAs were identified from the three comparisons. Comparing individuals who smoke more than 16 cigs/day with non-smokers, 59 DE miRNAs were

identified (50 up- and nine down-regulated) while 22 miRNAs were differentially expressed in those who smoke less than 16 cigs/day compared to non-smokers (three up- and 19 down-regulated). Interestingly, mir-8075-5p and miR-12128 were down-regulated in both smoking categories compared to non-smokers. Finally, 13 miRNAs were differentially expressed in former smokers vs non-smokers, with miR-5090-3p up-regulated and the other 12 DE miRNAs down-regulated in the former group.

A similar expression pattern was observed for the 84 DE miRNAs when the population was stratified by sex, with 14 and 26 out of 84 DE miRNAs significantly dysregulated in men and women, respectively. Other additional sex specific DE miRNAs were uniquely identified in men (n = 17) and women (n = 14)

Participants were also categorized by their self-reported alcohol consumption (i.e., gr/day intake of alcohol) in non-drinkers (0 gr/day), low intake (0.1–24.0 gr/day for male and 0.1–12.0 gr/day for female), and high intake (> 24.0 gr/day for male and > 12.0 gr/day for female) according to WHO guidelines.

The relationship between miRNA expression levels and alcohol consumption was assessed comparing nondrinkers (n = 29) with low (n = 230) and high (n = 75) intake drinkers. Comparing high intake drinkers with non-drinkers, miR-3972 was down-regulated in the latter, whereas in low intake drinkers vs non-drinkers, miR-4254 and miR-4254-5p were down-regulated, and miR-6895-3p up-regulated in drinkers

miRNA profiles in relation to coffee consumption were studied comparing low (< 8.0 gr/day; n = 149) and high (> 8.0 gr/day; n = 50) intake coffee drinkers with non-drinkers (n = 135). In high intake coffee consumers vs non-drinkers, 44 miRNAs were differentially expressed, with three up- and 41 down-regulated, in drinkers

The stratification of the study population according to the physical activity identified the following categories: inactive (n = 90), moderately inactive (n = 114), moderately active (n = 108), and active (n = 21) subjects. Comparing each of the first three categories against active individuals, 11 DE miRNAs were identified

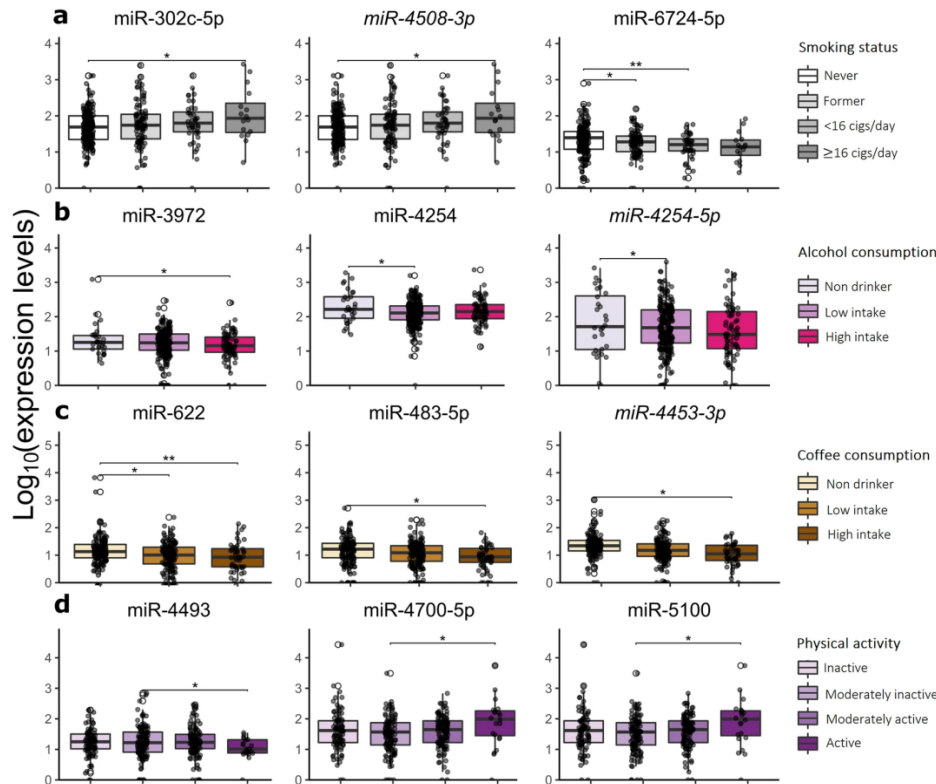


Figure 8. Box plots showing the expression levels of selected DE miRNAs among individuals stratified according to the investigated lifestyle features. P values were computed using DESeq2 and adjusted using the FDR method. ***adj P value 0.001, **adj P value 0.01, *adj P value 0.05

Overview of common miRNAs altered among investigated variables.

From a total of 3,041 miRNAs detected, 151 (5%) were associated with at least one of the analysed common traits or lifestyle habits while 52 DE miRNAs were significant in two or more comparisons. Considering separately for males and females the stool levels of the latter group of DE miRNAs, a subtle clustering of miRNAs emerged for both sexes, mainly related to smoking habit, BMI and coffee consumption, as reported in the heatmap in **Figure 9**.

To test the temporal stability of all the identified stool DE miRNAs expression levels in repeated samples collected from the same individuals, a Wilcoxon paired test was performed between the available samples collected at two time points. One hundred and seventy-three miRNAs (85.2%) showed no significant variation between the two measurements ($p \geq 0.05$, Supplementary Table 1D). The

remaining 30 DEMiRNAs (among which two variables shared 11 miRNAs), showing a significant variability between the two time points, were mostly related to BMI (n = 13), smoking habit (n = 14) or coffee consumption (n = 11).

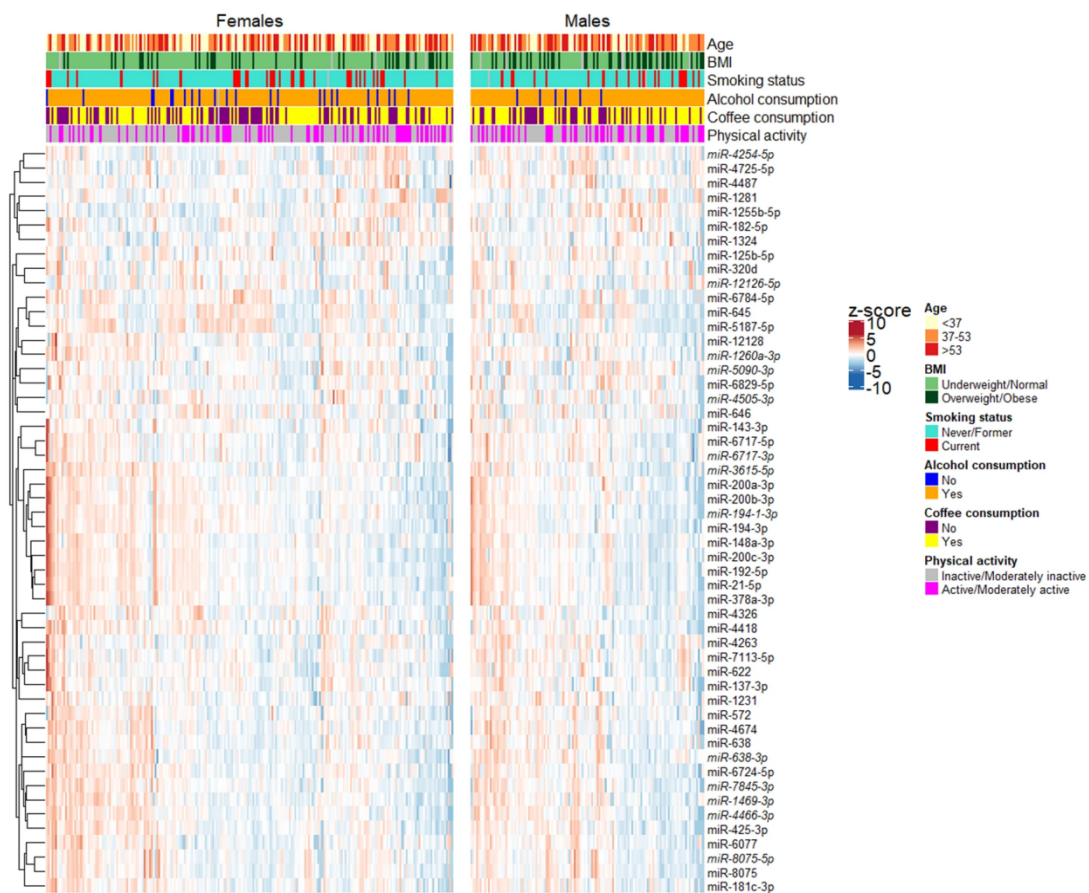


Figure 9. Heatmap of the hierarchical clustering of significant associations between miRNAs and the investigated common traits and lifestyle variables (p-values adjusted for multiple testing). For each DEMiRNA, the log₂FCs of all the comparisons are reported.

Study IV (Figures 10 to 15).

Starting from RNA-seq data, patients were assigned to CMS1 (n=9), CMS2 (n=24), CMS3 (n=26) and CMS4 (n=24). Differential expression analyses revealed 1096 differentially expressed genes in CMS1, 4218 in CMS2, 27 in CMS3 and 190 in CMS4. Notably, only cell migration-inducing and hyaluronan-binding protein

(CEMIP) was differentially expressed in all subtypes. CMS3, the most represented group, showed the highest number of coding variants (n=683) of which 384 were unique missense variants. CMS1 was the least represented subtype but it included almost all subjects with microsatellite instability. APC harboured the most coding variants in all CMS groups but CMS1.

In total, 380 miRNAs were differentially expressed in tumor tissue compared to adjacent mucosa. Based on these results CMS-specific miRNA-target interactions will be defined based on validated annotations and co-expression analysis. Putative upstream regulators will also be identified by functional enrichment analysis.

Our data are a reflect of the CMS classification system and ongoing investigations will aim to elucidate if the miRNome could implement RNA-seq-based CMS classification to facilitate patient prompt diagnosis and stratification and eventually further therapeutical approaches. Integration of TSO, RNA-seq and sRNA-seq data is a feasible approach to shed new light on CRC heterogeneity.

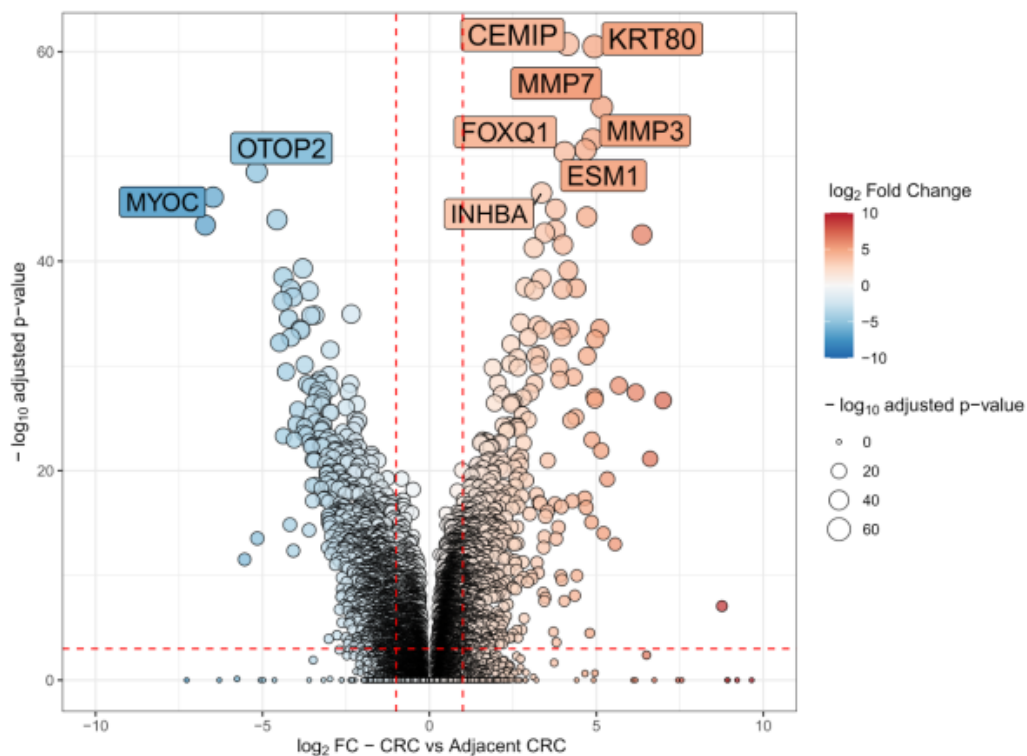


Figure 10. Volcano Plot showing 3.849 differentially expressed genes between tissues that were observed after RNA-sequencing. On these genes we can build Consensus Molecular Subtypes.

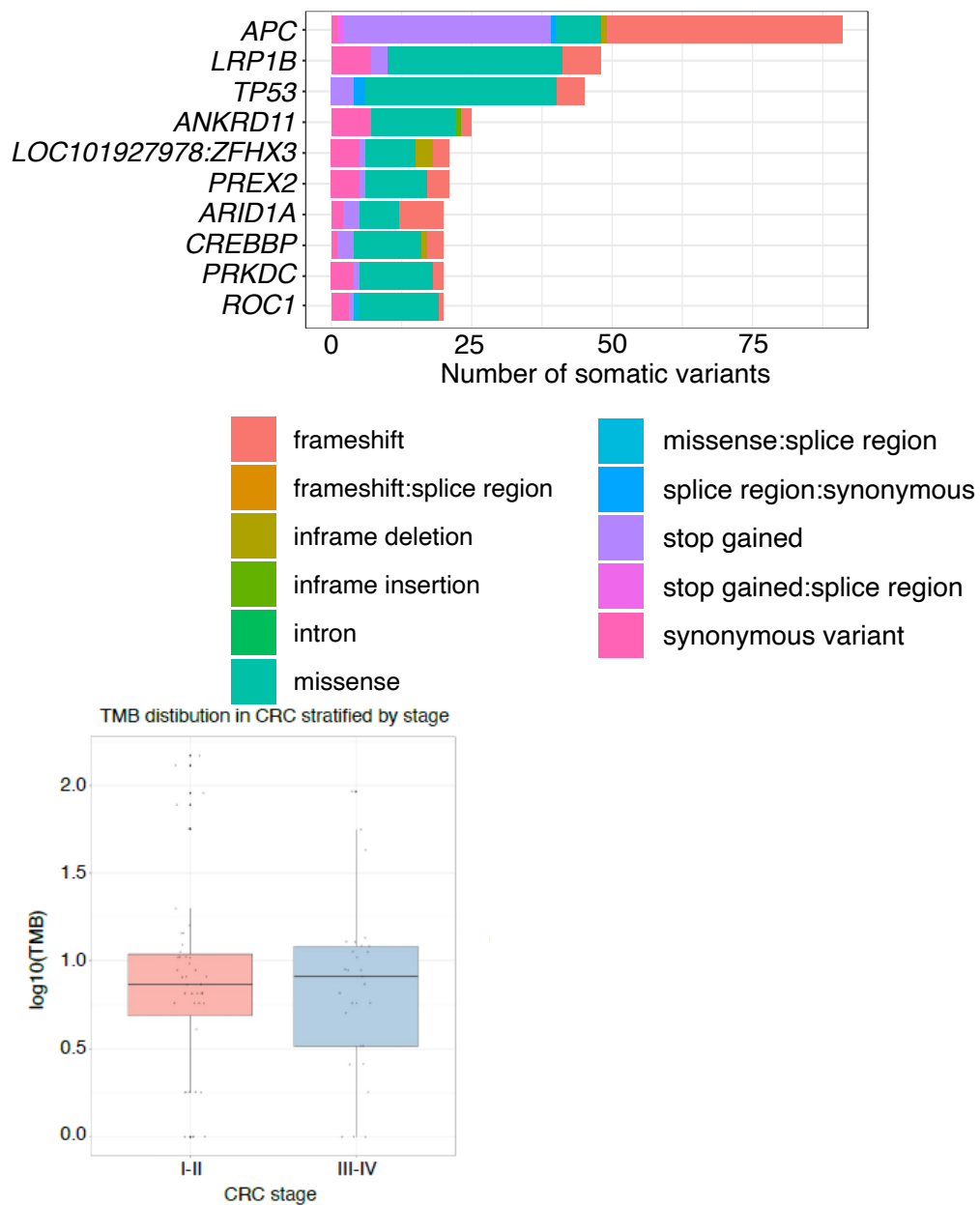


Figure 11. Illumina TruSight Oncology 500 (TSO500 HT) was used as Next Generation Sequencing assay targeting 523 genes, assessing somatic mutations, SNV, indels, TMB and MSI. Eleven subjects of the cohort with MSI high were further investigated in blood samples where they resulted all MSI stable

Eleven subjects of the cohort with MSI high were further investigated in blood samples where they resulted all MSI stable

According to RNA-seq data, each patient was assigned to a CMS subgroup using the CMSCaller package for R, with the use Nearest Template Prediction as a core algorithm.

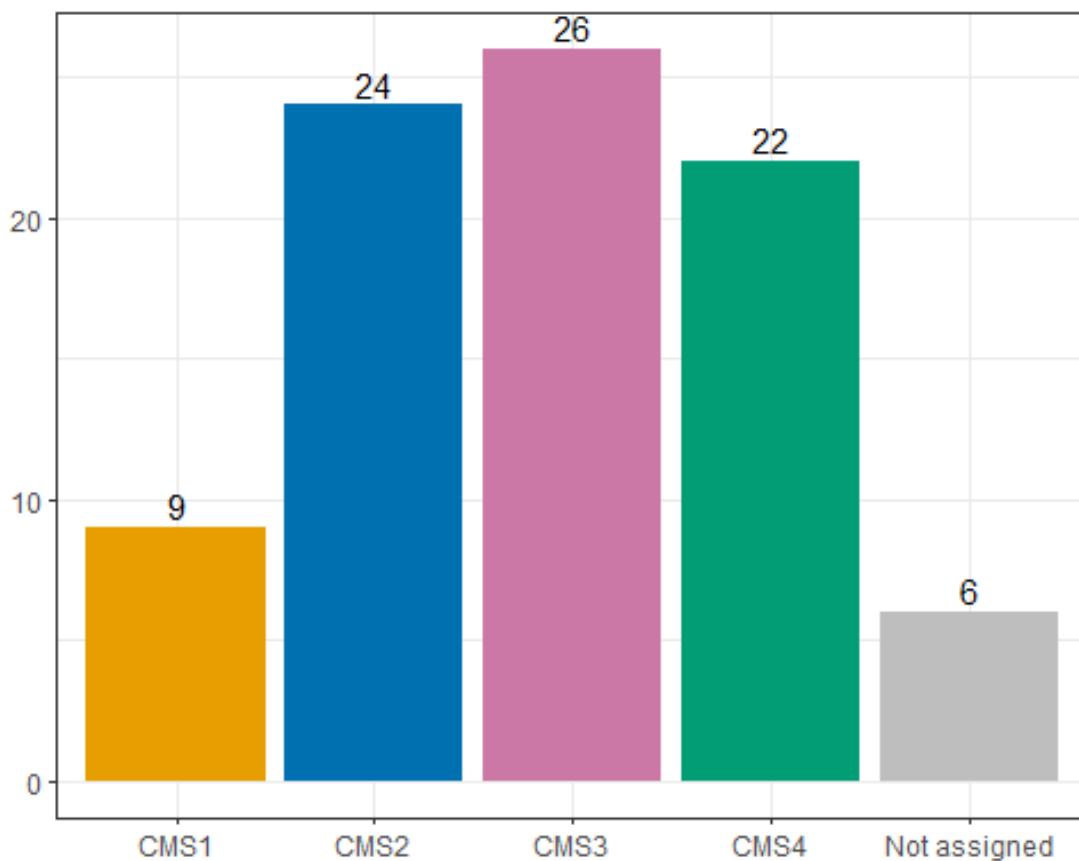


Figure 12. CMS distribution in the study population

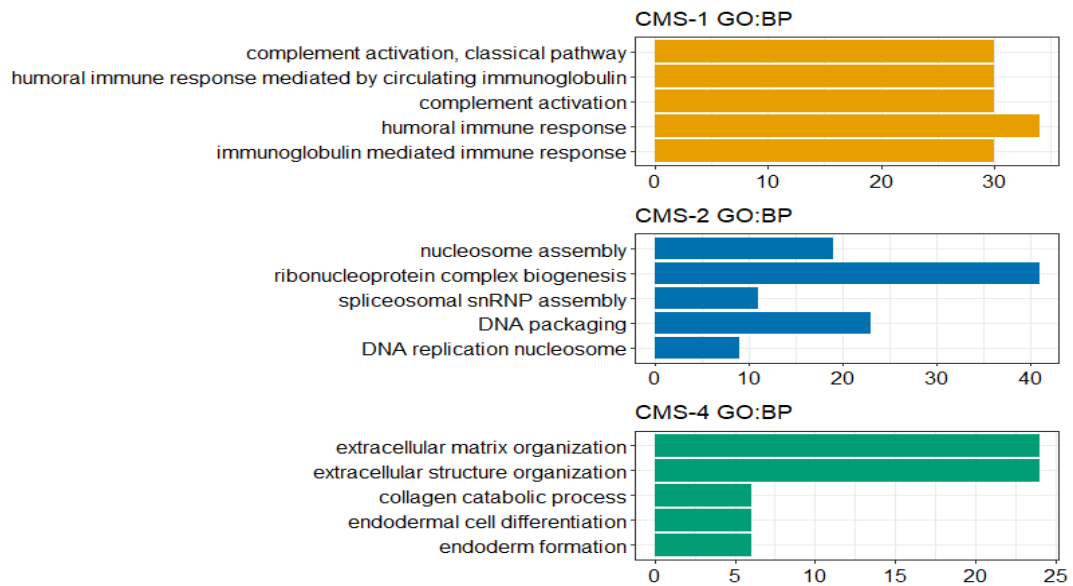


Figure 13. Top 5 enriched terms from Gene Ontology connected to the most overrepresented genes in each CMS subtype. No terms were identified for CMS-3

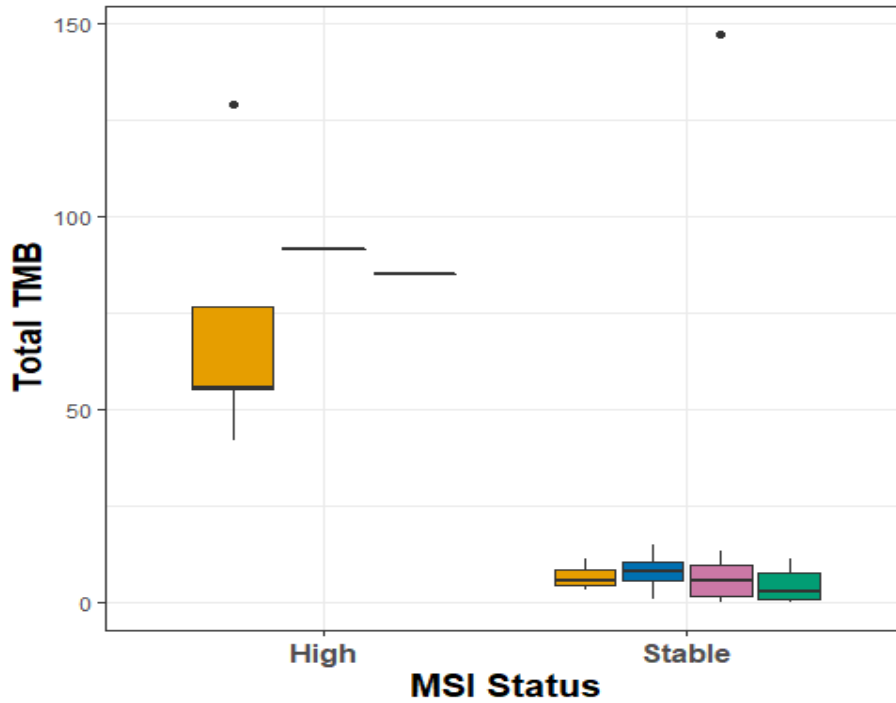


Figure 14. CMS subtypes stratified according to MSI and TMB

CMS 1				
miRNA	Target	# Evidence	miRNA Log2FC	Target Log2FC
miR-6777-5p	PER1	7	1.37	-1.51
miR-1976	PER1	6	1.51	-1.51
miR-30c-1-3p	RBM8A	6	-1.15	1.00
miR-30e-3p	KPNA2	6	-1.16	1.51
miR-495-3p	KPNA2	6	-1.41	1.51
miR-887-5p	RBM8A	6	-1.35	1.00
miR-4444	PER1	5	1.35	-1.51
miR-5094	MYLK	5	1.34	-1.55
let-7c-3p	CKS2	4	-1.61	1.70
miR-30e-3p	MYC	4	-1.16	1.56
miR-3183	VAV3	4	1.83	-2.31
miR-539-5p	SOD2	4	-1.53	1.29
miR-636	BTG2	4	1.28	-1.05
miR-6510-3p	UGDH	4	2.01	-1.09
miR-887-5p	H4C14	4	-1.35	1.18
miR-149-3p	BTG2	3	1.50	-1.05
miR-1910-3p	BTG2	3	1.39	-1.05
miR-1910-3p	ZBTB7B	3	1.39	-1.06
miR-193a-5p	PCSK9	3	-1.02	2.75
miR-224-3p	PSMA7	3	-1.00	1.03

CMS2				
miRNA	Target	# Evidence	miRNA Log2FC	Target Log2FC
miR-6812-5p	PLCG2	8	1.01	-1.65
miR-29b-2-5p	H2AC12	6	-1.16	1.99
miR-135a-3p	KIFC1	5	-1.66	2.01
let-7e-5p	ATP6V1F	4	-1.03	1.08
let-7e-5p	CCND1	4	-1.03	1.55
miR-130b-5p	TNS1	4	1.19	-1.95
miR-22-3p	VSNL1	4	-1.10	1.18
miR-466	CDKL1	4	1.26	-1.04
miR-1288-5p	DEK	3	-1.78	1.04
miR-130b-5p	PDE3A	3	1.19	-1.66
miR-152-3p	CCNA2	3	-1.03	2.21
miR-29b-2-5p	INSIG1	3	-1.16	1.07
miR-33a-5p	ABCA1	3	1.24	-1.09
miR-466	GNAI1	3	1.26	-1.40
miR-511-5p	TFDP1	3	-1.10	1.09
miR-552-5p	PER1	3	2.76	-1.99
miR-628-3p	SINHCAF	3	-1.17	1.19
miR-93-5p	PHLPP2	3	1.13	-1.24
miR-93-5p	TXNIP	3	1.13	-2.01
let-7e-5p	AP1S1	2	-1.03	1.23

CMS4				
miRNA	Target	# Evidence	miRNA Log2FC	Target Log2FC
miR-200c-3p	FN1	3	-1.05	1.91
miR-223-5p	F2RL1	3	1.08	-1.13
miR-3180-3p	ENTPD5	3	1.02	-1.23
miR-200b-3p	FN1	2	-1.16	1.91
miR-708-5p	PAQR5	2	1.11	-1.23
miR-196a-3p	SLC11A1	1	-1.08	1.77
miR-200b-3p	ACSL4	1	-1.16	1.13
miR-200b-3p	IGF2	1	-1.16	1.50
miR-200b-3p	SERPINH1	1	-1.16	1.13
miR-200b-5p	PMEPA1	1	-1.09	1.06
miR-200b-5p	SOD2	1	-1.09	1.04
miR-200c-3p	ACSL4	1	-1.05	1.13
miR-200c-3p	CDH11	1	-1.05	1.29
miR-200c-3p	FLNA	1	-1.05	1.13
miR-200c-3p	IGF2	1	-1.05	1.50
miR-200c-3p	SERPINH1	1	-1.05	1.13
miR-215-5p	ELOVL5	1	-1.78	1.01
miR-215-5p	FBN1	1	-1.78	1.05
miR-215-5p	MCAM	1	-1.78	1.43
miR-215-5p	PLAU	1	-1.78	1.57

Figure 15. miRNA-Target interactions between the differentially expressed miRNAs and genes stratified by CMS subtypes. No interactions were identified for CMS-3

VI - DISCUSSION

VI - DISCUSSION

CRC is a heterogeneous disease that arises due to complex interactions of the transcriptome, metabolome, microbiome, immune system and many other leading actors.

The use of omics technologies is an increasingly popular and useful approach used by researchers to discover and identify the heterogeneity in cancer disease. So far, omics platforms have been applied for several biological and pathological processes [45-47]. Literature examining the multi-omics approach on CRC is still limited to large-scale data with multiple biospecimens but has started to increase in recent years. 80% of CRCs in Western countries are related to dietary factors [48]. In fact, diet has been demonstrated to have a crucial impact on the structure and composition of the gut microbiome and host metabolism [49, 50].

Interestingly, the microbiome has a well-defined role in the pathogenesis of CRC, as shown by the presence of potential pathogenic bacteria such as *Fusobacterium* as well as beneficial bacteria such as *Lactobacillales* [51].

Recently, a free online platform including multiomics and clinico-pathological features of CRC, the so called "ColPortal" has been introduced [52]. The platform included also detailed and specific information about demographics, location, histology, and staging of the tumor, molecular biomarker status and clinical outcomes. All these data allow a better knowledge of the pathogenesis, including not very common CRC subtypes, as well as of the prognosis, providing a valid support for personalized therapeutic strategies.

In **Study I** we performed the first large-scale profiling of stool miRNome by deep sequencing of samples from patients with CRC, colorectal polyps, or other GI diseases and healthy controls. We confirmed previous findings about their potential role as non-invasive molecular biomarkers but also reported novel evidence on specific markers across different disease conditions. Moreover, a fecal miRNA signature was able to accurately distinguish CRC or adenoma patients from controls.

20 fecal miRNAs emerged as coherently altered in two independent cohorts. The analysis showed that the fecal profile of some of these miRNAs reflected their altered expression in the tumor tissue or in adjacent colonic mucosa. These results were consistent with those described in the literature. In fact, more than half of those miRNAs have been already reported in other series [10, 53].

In the same study we evaluated the minimal set of stool miRNAs able to accurately discriminate CRC patients from healthy individuals reproducing a signature of five fecal miRNAs (miR-1246-3p, miR-607, miR-6777-5p, miR-4488-3p, miR-149-3p).

In this study, we sought to compare stool DE miRNA profiles of newly diagnosed CRC with those of subjects with other bowel precancerous lesions diagnosed at colonoscopy. Besides different polyp types, we also included samples from several GI diseases, like different types of IBDs and diverticulitis. Notably, we found that while CRC-specific miRNAs were down-regulated, most of miRNAs in common with adenomas and inflammatory diseases were up-regulated: miR-21-5p was the clearest example consistently with the literature [54]. As an exception, miR-607 was down-regulated in stool miRNA profiles of patients with AA and ulcerative colitis. Accordingly, recent studies showed altered miRNA profiles in fecal samples of patients with inflammations [55, 56], even in relation to microbiota [57]. We can therefore conclude that altered stool miRNA profiles reflect either the intestinal response to an inflammatory process or the transcriptional alteration related specifically to CRC development. Importantly, we clearly highlighted that the fecal levels of known CRC-related miRNAs are actually dysregulated in several disease contexts, suggesting that other miRNAs, such as miR-6777-3p and miR-149-3p, should be selected to design CRC-specific molecular signatures. This is the first evidence of CRC-specific fecal miRNAs from a large-scale analysis of subjects with different gastrointestinal diseases and it highlights an extensive influence of gut inflammation on the fecal miRNA levels.

In **Study II** we comprehensively analysed the expression profiles of all miRNA genes residing in the locus by next-generation sequencing in multiple biospecimens from a cohort of CRC patients and healthy controls (tumor tissue and paired adjacent mucosa for patients, stool, and plasma EVs for both). In CRC

tissues, out of the 54 identified mature miRNAs in the 8q24 region, twelve were differentially expressed between tumor and adjacent mucosa, with nine of them being up regulated. Some of the latter miRNAs have been previously reported as over-expressed also in other cancers tissues [58, 59].

Similar results were confirmed in the CRC-TCGA dataset with five out of twelve miRNAs (miR-151a-3p, miR-151a-5p, miR-30b-5p, miR-4662a-5p, miR-30d-5p) dysregulated in the same direction as in our dataset. Those analysis were necessary to assess if the dysregulation of these miRNAs was specific of CRC.

In summary, several dysregulated miRNAs mapping to chromosome 8q24 were found in CRC and BC, both in primary and surrogate tissues. The dysregulated miRNAs emerged from an investigation of the whole genome miRNome, highlighting the importance of the 8q24 locus also at the transcriptomic level. The strength of this study is that we took advantage of a large collection of several biospecimens from the same patients. Tissues, stool, and plasma samples were available for many CRC patients. In this respect, no previous study assessed 8q24 miRNA profiles in stool of CRC patients and healthy controls.

Study III aimed to provide the first evidence on how the faecal miRNome expression is affected by a set of common variables investigated in a population of 335 healthy subjects. A total of 203 miRNAs were significantly associated with at least one of the considered variables, with 52 associated with more than one. Several miRNAs showed a differential expression in agreement with other studies on blood or tissue samples [60-62]. Profiles of several miRNAs in stool reflect main common traits and lifestyle habits. Our findings were consistent with those already reported in the current literature. However, a definitive association between variables and miRNA profiles, cannot be established with certainty for the failure to rule out the potential for other confounding factors.

Lastly, in **Study IV** we demonstrated that the integration of TSO, RNA-seq and small-RNA-seq data is a feasible approach to shed new light on CRC heterogeneity and, potentially, improve the current CMS classification.

VII - CONCLUSIONS

VII CONCLUSIONS

The studies included in this PhD project give an overview of how a multi-omics approach, including different biospecimens, generating a large amount of data, can be useful for understanding the nature of CRC and applying correct diagnostic-therapeutic strategies.

- Fecal miRNome analysis identified a predictive signature accurately discriminating CRC and precancerous lesions, independently from age and sex, for a non-invasive diagnosis aimed at improving the effectiveness of current screening programs, potentially increasing sensitivity and maintaining high specificity and applicable on a large scale, with a reasonable cost/time required.
- Altered expression of 8q24-related miRNAs may be important for the initiation and/or progression of cancer.
- miRNA profiles in stool may reflect common traits and lifestyle habits and should be considered in relation to disease and association studies based on faecal miRNA expression.
- The Integration of TruSight Oncology 500 (TSO) cancer-panel, RNA-seq and sRNA-seq data is a feasible approach to shed new light on CRC heterogeneity and reveals a potential role for stool as less invasive biomarker of CRC.

VIII – LIMITATIONS AND FUTURE PERSPECTIVES

VIII – LIMITATIONS AND FUTURE PERSPECTIVES

Our studies have some limitations. In fact, in **Study I** although a similar approach was used for patients and controls' recruitment, the two cohorts were heterogeneous for individual categories. Despite the large number of samples, the variegated spectrum of CRC, adenomas and other precancerous lesions was not exhaustively represented and deserves further investigation. However, we should also consider the major strengths or rather the inclusion of independent cohorts from two countries who have different diet and lifestyle habits and CRC rates as well as the standardization of stool collection in both cohorts. Moreover, the miRNome-wide approach in different biospecimens and different GI diseases contexts has allowed us to discriminate miRNAs specifically dysregulated in stool of CRC patients.

In **Study II** we took advantage of a large collection of several biospecimens from the same patients firstly assessing 8q24 miRNA profiles in stool of CRC patients and healthy controls. Further studies are required to determine the possible functions of the altered miRNAs identified in this study and their role in tumorigenesis, as well as to expand the investigation also to other cancer types.

In **Study III** the self-reported data retrieved from the questionnaires may contain data entry errors, or individuals may have wrongly answered. Furthermore, the limited sample size for some of the categories investigated among the variables could have affected the results especially for the obese and underweight groups in the analyses on BMI or for the active category for the physical activity. Nonetheless, no studies have been performed on stool miRNA expression levels in relation to common traits and lifestyle habits in healthy individuals, so far.

In **Study IV** even though we reported integration of TSO, RNA-seq and small-RNA-seq, a correlation with other omics will be useful in the future to validate the results

Overall, in the next trials it will be useful to integrate other omics not investigated in the present study such as proteomics, methylation and radiomics. Diagnostic images can provide information on the entire tumor volume, reducing inaccuracy due to sampling errors in histopathological analyses.

The growing impact of non-invasive imaging techniques for disease diagnosis, in parallel with the evolution of NGS tools, provides powerful methods for investigating the phenotype. In fact, thanks to the correlation of radiomic features with genomic features, a new field of study called "radiogenomics" was born.

Radiomics has the potential to characterize the intratumoral CRC heterogeneity, with promising results in predicting treatment response and outcome as well as differentiating the nature of the tumor and assessing the relationship with genetics.

IX - REFERENCES

IX – REFERENCES

1. Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin DM, Piñeros M, Znaor A, Bray F. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int J Cancer*. 2019 Apr 15;144(8):1941-19531
2. Pellino G, Gallo G, Pallante P, Capasso R, De Stefano A, Maretto I, Malapelle U, Qiu S, Nikolaou S, Barina A, Clerico G, Reginelli A, Giuliani A, Sciaudone G, Kontovounisios C, Brunese L, Trompetto M, Selvaggi F. Noninvasive Biomarkers of Colorectal Cancer: Role in Diagnosis and Personalised Treatment Perspectives. *Gastroenterol Res Pract*. 2018 Jun 13;2018:2397863
3. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P, Bot BM, Morris JS, Simon IM, Gerster S, Fessler E, De Sousa E Melo F, Missiaglia E, Ramay H, Barras D, Homicsko K, Maru D, Manyam GC, Broom B, Boige V, Perez-Villamil B, Laderas T, Salazar R, Gray JW, Hanahan D, Tabernero J, Bernards R, Friend SH, Laurent-Puig P, Medema JP, Sadanandam A, Wessels L, Delorenzi M, Kopetz S, Vermeulen L, Tejpar S. The consensus molecular subtypes of colorectal cancer. *Nat Med*. 2015 Nov;21(11):1350-6. doi: 10.1038/nm.3967
4. Greigor DH. Diagnosis of large-bowel cancer in the asymptomatic patient. *JAMA*. 1967 Sep 18;201(12):943-53
5. Cheung KS, Leung WK, Seto WK. Application of Big Data analysis in gastrointestinal research. *World J Gastroenterol*. 2019 Jun 28;25(24):2990-3008. doi: 10.3748/wjg.v25.i24.2990
6. Wu CW, Ng SC, Dong Y, Tian L, Ng SS, Leung WW, Law WT, Yau TO, Chan FK, Sung JJ, Yu J. Identification of microRNA-135b in stool as a potential noninvasive biomarker for colorectal cancer and adenoma. *Clin Cancer Res*. 2014 Jun 1;20(11):2994-3002. doi: 10.1158/1078-0432.CCR-13-1750
7. Di K, Fan B, Gu X, Huang R, Khan A, Liu C, Shen H, Li Z. Highly efficient and automated isolation technology for extracellular vesicles microRNA. *Front Bioeng Biotechnol*. 2022 Aug 10;10:948757. doi: 10.3389/fbioe.2022.948757

8. Tarallo S, Ferrero G, Gallo G, Francavilla A, Clerico G, Realis Luc A, Manghi P, Thomas AM, Vineis P, Segata N, Pardini B, Naccarati A, Cordero F. Altered Fecal Small RNA Profiles in Colorectal Cancer Reflect Gut Microbiome Composition in Stool Samples. *mSystems*. 2019 Sep 17;4(5):e00289-19. doi: 10.1128/mSystems.00289-19
9. Koga Y, Yasunaga M, Takahashi A, Kuroda J, Moriya Y, Akasu T, Fujita S, Yamamoto S, Baba H, Matsumura Y. MicroRNA expression profiling of exfoliated colonocytes isolated from feces for colorectal cancer screening. *Cancer Prev Res (Phila)*. 2010 Nov;3(11):1435-42. doi: 10.1158/1940-6207.CAPR-10-0036
10. Francavilla A, Tarallo S, Pardini B, Naccarati A. Fecal microRNAs as non-invasive biomarkers for the detection of colorectal cancer: a systematic review. *Minerva Biotecnologica* 2019; 31(1):30-42
11. Huppi K, Pitt JJ, Wahlberg BM, Caplen NJ. The 8q24 gene desert: an oasis of non-coding transcriptional activity. *Front Genet*. 2012 Apr 30;3:69. doi: 10.3389/fgene.2012.00069
12. Grisanzio C, Freedman ML. Chromosome 8q24-Associated Cancers and MYC. *Genes Cancer*. 2010 Jun;1(6):555-9. doi: 10.1177/1947601910381380
13. Easton DF, Eeles RA. Genome-wide association studies in cancer. *Hum Mol Genet*. 2008 Oct 15;17(R2):R109-15. doi: 10.1093/hmg/ddn2874
14. Dang CV, O'Donnell KA, Zeller KI, Nguyen T, Osthus RC, Li F. The c-Myc target gene network. *Semin Cancer Biol*. 2006 Aug;16(4):253-64. doi: 10.1016/j.semcancer.2006.07.014
15. Gu Y, Lin X, Kapoor A, Chow MJ, Jiang Y, Zhao K, Tang D. The Oncogenic Potential of the Centromeric Border Protein FAM84B of the 8q24.21 Gene Desert. *Genes (Basel)*. 2020 Mar 15;11(3):312. doi: 10.3390/genes11030312
16. Wei J, Xu Z, Chen X, Wang X, Zeng S, Qian L, Yang X, Ou C, Lin W, Gong Z, Yan Y. Overexpression of GSDMC is a prognostic factor for predicting a poor outcome in lung adenocarcinoma. *Mol Med Rep*. 2020 Jan;21(1):360-370. doi: 10.3892/mmr.2019.10837
17. Müller T, Stein U, Poletti A, Garzia L, Rothley M, Plaumann D, Thiele W, Bauer M, Galasso A, Schlag P, Pankratz M, Zollo M, Sleeman JP. ASAP1 promotes tumor cell motility and invasiveness, stimulates metastasis formation in vivo, and correlates with poor survival in colorectal cancer patients. *Oncogene*. 2010 Apr 22;29(16):2393-403. doi: 10.1038/onc.2010.6
18. Chattaragada MS, Riganti C, Sassoe M, Principe M, Santamorenna MM, Roux C, Curcio C, Evangelista A, Allavena P, Salvia R, Rusev B, Scarpa A, Cappello P, Novelli F. FAM49B, a novel regulator of mitochondrial

- function and integrity that suppresses tumor metastasis. *Oncogene*. 2018 Feb 8;37(6):697-709. doi: 10.1038/onc.2017.358
19. Cyganek M, Graña B, Krawczyk A, Kasprzak P, Porwik K, Walkowiak M, Woźniak M. A survey of big data issues in electronic health record analysis. *Appl. Artif. Intell.* 2016 July 21; 30(6):497–520 doi.org/10.1080/08839514.2016.1193714
 20. Cox M, Ellsworth D. Application-controlled demand paging for out-of-core visualization. *Proceedings. Vis. 97 (Cat. No. 97CB36155) (1997)* 235–244
 21. Diebold FX. Big data dynamic factor models for macroeconomic measuring and forecasting. *Adv. Econ. Econom. Eighth World Congr. Econom. Soc. (2003)* 115–122
 22. Laney D. META delta, *Appl. Deliv. Strateg.* 949 (2001) 4, <http://dx.doi.org/10.1016/j.infsof.2008.09.005>
 23. Sukumar SR, Natarajan R, Ferrell RK. Quality of Big Data in health care. *Int J Health Care Qual Assur.* 2015;28(6):621-34. doi: 10.1108/IJHCQA-07-2014-0080
 24. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, Charrondière UR, Hémon B, Casagrande C, Vignat J, Overvad K, Tjønneland A, Clavel-Chapelon F, Thiébaud A, Wahrendorf J, Boeing H, Trichopoulos D, Trichopoulou A, Vineis P, Palli D, Bueno-De-Mesquita HB, Peeters PH, Lund E, Engeset D, González CA, Barricarte A, Berglund G, Hallmans G, Day NE, Key TJ, Kaaks R, Saracci R. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr.* 2002 Dec;5(6B):1113-24. doi: 10.1079/PHN2002394
 25. https://www.aiom.it/wp-content/uploads/2020/10/2020_LG_AIOM_Colon.pdf
 26. https://www.aiom.it/wp-content/uploads/2021/01/2020_LG_AIOM_Retto_e_Ano.pdf
 27. Tarallo S, Ferrero G, Gallo G, Francavilla A, Clerico G, Realis Luc A, Manghi P, Thomas AM, Vineis P, Segata N, Pardini B, Naccarati A, Cordero F. Altered Fecal Small RNA Profiles in Colorectal Cancer Reflect Gut Microbiome Composition in Stool Samples. *mSystems.* 2019 Sep 17;4(5):e00289-19. doi: 10.1128/mSystems.00289-19
 28. Zwinsová B, Petrov VA, Hrivňáková M, Smatana S, Micenková L, Kazdová N, Popovici V, Hrstka R, Šefr R, Bencsiková B, Zdražilová-Dubská L, Brychtová V, Nenutil R, Vídeňská P, Budinská E. Colorectal Tumour Mucosa Microbiome Is Enriched in Oral Pathogens and Defines

- Three Subtypes That Correlate with Markers of Tumour Progression. *Cancers* (Basel). 2021 Sep 25;13(19):4799. doi: 10.3390/cancers13194799
29. Tarallo S, Ferrero G, De Filippis F, Francavilla A, Pasolli E, Panero V, Cordero F, Segata N, Grioni S, Pensa RG, Pardini B, Ercolini D, Naccarati A. Stool microRNA profiles reflect different dietary and gut microbiome patterns in healthy individuals. *Gut*. 2022 Jul;71(7):1302-1314. doi: 10.1136/gutjnl-2021-325168
 30. Francavilla A, Gagliardi A, Piaggieschi G, Tarallo S, Cordero F, Pensa RG, Impeduglia A, Caviglia GP, Ribaldone DG, Gallo G, Grioni S, Ferrero G, Pardini B, Naccarati A. Faecal miRNA profiles associated with age, sex, BMI, and lifestyle habits in healthy individuals. *Sci Rep*. 2021 Oct 19;11(1):20645. doi: 10.1038/s41598-021-00014-1
 31. Sabo AA, Birolo G, Naccarati A, Dragomir MP, Aneli S, Allione A, Oderda M, Allasia M, Gontero P, Sacerdote C, Vineis P, Matullo G, Pardini B. Small Non-Coding RNA Profiling in Plasma Extracellular Vesicles of Bladder Cancer Patients by Next-Generation Sequencing: Expression Levels of miR-126-3p and piR-5936 Increase with Higher Histologic Grades. *Cancers* (Basel). 2020 Jun 9;12(6):1507. doi: 10.3390/cancers12061507
 32. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* 2014;15:550.25
 33. Zhang J, Storey KB. RBiomirGS: an all-in-one miRNA gene set analysis solution featuring target mRNA mapping and expression profile integration. *PeerJ*. 2018 Jan 12;6:e4262. doi: 10.7717/peerj.4262
 34. Weiser, M.R. (2018) *AJCC 8th Edition: Colorectal Cancer.*]
 35. Pardini B, Cordero F, Naccarati A, Viberti C, Birolo G, Oderda M, Di Gaetano C, Arigoni M, Martina F, Calogero RA, Sacerdote C, Gontero P, Vineis P, Matullo G. microRNA profiles in urine by next-generation sequencing can stratify bladder cancer subtypes. *Oncotarget*. 2018 Apr 17;9(29):20658-20669. doi: 10.18632/oncotarget.25057
 36. Sacerdote C, Guarrera S, Ricceri F, Pardini B, Polidoro S, Allione A, Critelli R, Russo A, Andrew AS, Ye Y, Wu X, Kiemeny LA, Bosio A, Casetta G, Cucchiara G, Destefanis P, Gontero P, Rolle L, Zitella A, Fontana D, Vineis P, Matullo G. Polymorphisms in the XRCC1 gene modify survival of bladder cancer patients treated with chemotherapy. *Int J Cancer*. 2013 Oct 15;133(8):2004-9. doi: 10.1002/ijc.28186
 37. Turinetto V, Pardini B, Allione A, Fiorito G, Viberti C, Guarrera S, Russo A, Anglesio S, Ruo Redda MG, Casetta G, Cucchiara G, Destefanis P, Oderda M, Gontero P, Rolle L, Frea B, Vineis P, Sacerdote C, Giachino C,

- Matullo G. H2AX phosphorylation level in peripheral blood mononuclear cells as an event-free survival predictor for bladder cancer. *Mol Carcinog.* 2016 Nov;55(11):1833-1842. doi: 10.1002/mc.22431
38. Ferrero G, Cordero F, Tarallo S, Arigoni M, Riccardo F, Gallo G, Ronco G, Allasia M, Kulkarni N, Matullo G, Vineis P, Calogero RA, Pardini B, Naccarati A. Small non-coding RNA profiling in human biofluids and surrogate tissues from healthy individuals: description of the diverse and most represented species. *Oncotarget.* 2017 Dec 14;9(3):3097-3111. doi: 10.18632/oncotarget.23203
 39. Ozawa T, Matsuyama T, Toiyama Y, Takahashi N, Ishikawa T, Uetake H, Yamada Y, Kusunoki M, Calin G, Goel A. CCAT1 and CCAT2 long noncoding RNAs, located within the 8q.24.21 'gene desert', serve as important prognostic biomarkers in colorectal cancer. *Ann Oncol.* 2017 Aug 1;28(8):1882-1888. doi: 10.1093/annonc/mdx248
 40. Xu T, Su N, Liu L, Zhang J, Wang H, Zhang W, Gui J, Yu K, Li J, Le TD. miRBaseConverter: an R/Bioconductor package for converting and retrieving miRNA name, accession, sequence and family information in different versions of miRBase. *BMC Bioinformatics.* 2018 Dec 31;19(Suppl 19):514. doi: 10.1186/s12859-018-2531-5
 41. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun.* 2019 Apr 3;10(1):1523. doi: 10.1038/s41467-019-09234-6
 42. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 2011 May; 17(1):10-12. doi.org/10.14806/ej.17.1. 200
 43. Kulkarni N, Alessandrì L, Panero R, Arigoni M, Olivero M, Ferrero G, Cordero F, Beccuti M, Calogero RA. Reproducible bioinformatics project: a community for reproducible bioinformatics analysis pipelines. *BMC Bioinformatics.* 2018 Oct 15;19(Suppl 10):349. doi: 10.1186/s12859-018-2296-x
 44. <https://emea.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/trusight-oncology-500-and-ht-data-sheet-1170-2018-010.pdf>
 45. Xu Y, Ku X, Wu C, Cai C, Tang J, Yan W. Exosomal proteome analysis of human plasma to monitor sepsis progression. *Biochem Biophys Res Commun.* 2018 May 23;499(4):856-861. doi: 10.1016/j.bbrc.2018.04.006

46. Dear JW, Street JM, Bailey MA. Urinary exosomes: a reservoir for biomarker discovery and potential mediators of intrarenal signalling. *Proteomics*. 2013 May;13(10-11):1572-80. doi: 10.1002/pmic.201200285
47. Cubedo J, Padró T, García-Moll X, Pintó X, Cinca J, Badimon L. Proteomic signature of Apolipoprotein J in the early phase of new-onset myocardial infarction. *J Proteome Res*. 2011 Jan 7;10(1):211-20. doi: 10.1021/pr100805h
48. Bingham SA. Diet and colorectal cancer prevention. *Biochem Soc Trans*. 2000 Feb;28(2):12-6. doi: 10.1042/bst0280012
49. De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, Collini S, Pieraccini G, Lionetti P. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc Natl Acad Sci U S A*. 2010 Aug 17;107(33):14691-6. doi: 10.1073/pnas.1005963107
50. Wang T, Cai G, Qiu Y, Fei N, Zhang M, Pang X, Jia W, Cai S, Zhao L. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *ISME J*. 2012 Feb;6(2):320-9. doi: 10.1038/ismej.2011.109
51. Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, Beghini F, Manara S, Karcher N, Pozzi C, Gandini S, Serrano D, Tarallo S, Francavilla A, Gallo G, Trompetto M, Ferrero G, Mizutani S, Shiroma H, Shiba S, Shibata T, Yachida S, Yamada T, Wirbel J, Schrotz-King P, Ulrich CM, Brenner H, Arumugam M, Bork P, Zeller G, Cordero F, Dias-Neto E, Setubal JC, Tett A, Pardini B, Rescigno M, Waldron L, Naccarati A, Segata N. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat Med*. 2019 Apr;25(4):667-678. doi: 10.1038/s41591-019-0405-7
52. Esteban-Gil A, Pérez-Sanz F, García-Solano J, Albuquerque-González B, Parreño-González MA, Legaz-García MDC, Fernández-Breis JT, Rodríguez-Braun E, Pimentel P, Tuomisto A, Mäkinen M, Slaby O, Conesa-Zamora P. ColPortal, an integrative multiomic platform for analysing epigenetic interactions in colorectal cancer. *Sci Data*. 2019 Oct 31;6(1):255. doi: 10.1038/s41597-019-0198-z
53. Slaby O. Non-coding RNAs as Biomarkers for Colorectal Cancer Screening and Early Detection. *Adv Exp Med Biol*. 2016;937:153-70. doi: 10.1007/978-3-319-42059-2_8
54. Jenike AE, Halushka MK. miR-21: a non-specific biomarker of all maladies. *Biomark Res*. 2021 Mar 12;9(1):18. doi: 10.1186/s40364-021-00272-1

55. Wohnhaas CT, Schmid R, Rolser M, Kaaru E, Langgartner D, Rieber K, Strobel B, Eisele C, Wiech F, Jakob I, Gantner F, Herichova I, Vinisko R, Böcher WO, Visvanathan S, Shen F, Panzenbeck M, Raymond E, Reber SO, Delić D, Baum P. Fecal MicroRNAs Show Promise as Noninvasive Crohn's Disease Biomarkers. *Crohns Colitis* 360. 2020 Jan;2(1):otaa003. doi: 10.1093/crocol/otaa003
56. Verdier J, Breunig IR, Ohse MC, Roubrocks S, Kleinfeld S, Roy S, Streetz K, Trautwein C, Roderburg C, Sellge G. Faecal Micro-RNAs in Inflammatory Bowel Diseases. *J Crohns Colitis*. 2020 Jan 1;14(1):110-117. doi: 10.1093/ecco-jcc/jjz120
57. Ambrozkiwicz F, Karczmarski J, Kulecka M, Paziewska A, Niemira M, Zeber-Lubecka N, Zagorowicz E, Kretowski A, Ostrowski J. In search for interplay between stool microRNAs, microbiota and short chain fatty acids in Crohn's disease - a preliminary study. *BMC Gastroenterol*. 2020 Sep 21;20(1):307. doi: 10.1186/s12876-020-01444-3
58. Li Y, Wang YW, Chen X, Ma RR, Guo XY, Liu HT, Jiang SJ, Wei JM, Gao P. MicroRNA-4472 Promotes Tumor Proliferation and Aggressiveness in Breast Cancer by Targeting RGMA and Inducing EMT. *Clin Breast Cancer*. 2020 Apr;20(2):e113-e126. doi: 10.1016/j.clbc.2019.08.010
59. Shen Y, Chen G, Gao H, Li Y, Zhuang L, Meng Z, Liu L. miR-939-5p Contributes to the Migration and Invasion of Pancreatic Cancer by Targeting ARHGAP4. *Onco Targets Ther*. 2020 Jan 14;13:389-399. doi: 10.2147/OTT.S227644
60. Fehlmann T, Lehallier B, Schaum N, Hahn O, Kahraman M, Li Y, Grammes N, Geffers L, Backes C, Balling R, Kern F, Krüger R, Lammert F, Ludwig N, Meder B, Fromm B, Maetzler W, Berg D, Brockmann K, Deuschle C, von Thaler AK, Eschweiler GW, Milman S, Barziliai N, Reichert M, Wyss-Coray T, Meese E, Keller A. Common diseases alter the physiological age-related blood microRNA profile. *Nat Commun*. 2020 Nov 24;11(1):5958. doi: 10.1038/s41467-020-19665-1
61. Cui C, Yang W, Shi J, Zhou Y, Yang J, Cui Q, Zhou Y. Identification and Analysis of Human Sex-biased MicroRNAs. *Genomics Proteomics Bioinformatics*. 2018 Jun;16(3):200-211. doi: 10.1016/j.gpb.2018.03.004
62. Georgiadis P, Hebels DG, Valavanis I, Liampa I, Bergdahl IA, Johansson A, Palli D, Chadeau-Hyam M, Chatziioannou A, Jennen DG, Krauskopf J, Jetten MJ, Kleinjans JC, Vineis P, Kyrtopoulos SA; EnviroGenomarkers consortium. Omics for prediction of environmental health effects: Blood leukocyte-based cross-omic profiling reliably predicts diseases associated with tobacco smoking. *Sci Rep*. 2016 Feb 3;6:20544. doi: 10.1038/srep20544

