University of Turin

PhD Program in Complex Systems for Life Sciences

XXXIII Cycle

TITLE

Small non-coding RNAs in Gastrointestinal Diseases: an investigation on Celiac Disease and Colorectal Cancer

Candidate
Antonio Francavilla

Tutor
Prof. Raffaele A. Calogero

Co-tutor
Dr. Alessio G. Naccarati

Contents

Acknowledgements

I would like to express my sincere gratitude and appreciation to people who have helped me during the whole PhD period.

First of all, I would like to thank my family who followed and supported me in this way.

I would like to express my gratitude to my supervisor Dr. Alessio Naccarati, for his encouragement, scientific contribution as well as all the members of the Molecular Epidemiology and Exposomics Unit of the Italian Institute for Genomic Medicine (IIGM): Dr. Barbara Pardini, Dr. Sonia Tarallo, Dr. Beatrice Piaggeschi, Dr. Amedeo Gagliardi, Dr. Giulia Francescato for the given support and friendship in all these years.

I would to thank Dr. Giulio Ferrero and Prof. Francesca Cordero for their computational support and Prof. Raffaele Calogero for having been my PhD Tutor.

Finally a great thank belongs to the Associazione Italiana Celiachia (AIC) which funded my PhD research for these three years.

Abstract

Over the past decade, small non-coding RNAs (sncRNAs), including microRNAs (miRNAs) have been recognized as important modulators of various cellular processes. Numerous studies showed the potentiality of miRNAs as powerful biomarkers in different human conditions, including gastrointestinal (GI) diseases. The human gut microbiome, referring to the microbial communities populating our intestinal tract, is also emerging as a relevant factor in human diseases. A fascinating aspect is the interplay of host fecal miRNAs with the gut microbiome which has been highlighted and that could be the starting point for novel suitable biomarkers for GI diseases detection, monitoring or innovative therapeutic approaches development.

In the present study, a comprehensive experimental and computational analysis was proposed to explore small RNA and metagenomic sequencing data obtained from surrogate tissue of patients with different GI disorders.

We investigated fecal sncRNA profiles and gut microbiome composition on samples from Celiac Disease (CD) (Study 1) and from Colorectal Cancer (CRC), colorectal polyps and Inflammatory Bowel Diseases (IBD) individuals (Study 2) and compared to those of healthy subjects, performing small RNA-sequencing and shotgun metagenomics sequencing, respectively.

The cohort of Study 1 was composed of 51 CD treated patients with low levels (CD-ltTG) and 11 CD treated patients with high levels of transglutaminase 2 (TG2) antibodies levels on a gluten-free diet (GFD), 3 CD untreated patients (recruited before and after GFD adherence), 2 non-celiac gluten sensitivity individuals (NCGS), and 65 healthy controls. Several miRNAs and other sncRNAs were differentially expressed (DE) in CD-ltTG and CD-htTG compared to controls. The majority of the observed sncRNAs were specific for each group while a small set was common. For a group of DEmiRNAs, a significant correlation with GFD duration was noticed.

The metagenomic analysis highlighted significant differences in bacterial abundances among the CD categories compared to controls. Interestingly, the abundance of some of the microbial species significantly correlated with the expression levels of several DEmiRNAs.

The cohort of Study 2 included subjects recruited in a gastroenterological department of a hospital in Vercelli: 58 patients with CRC, 43 with polyps, 41 with gut inflammations, and 79 controls with negative colonoscopy. For this cohort, stool and primary tissue samples were also collected. Additional stool samples were collected in an independent cohort from the Czech Republic, including 67 CRC, 27 polyps, 32 gut inflammations, and 36 controls with negative colonoscopy.

In stool of both CRC cohorts, several DEmiRNAs and DEsncRNAs were found in cancer and precancerous lesions compared to controls. A set of 19 miRNAs were differentially expressed in CRC patients of both cohorts and their pattern of expression mirrored those of tumor/polyp tissues. A predictive model including 11 miRNAs accurately classified CRC from healthy subjects, clearly distinguishing also CRC from adenomas and both CRC and adenomas from controls.

Microbiome composition analysis on a subset of 80 individuals (29 CRC, 27 Polyp, and 24 healthy controls subjects) of Cohort-IT was explored highlighting differences in fecal microbiome composition among healthy, polyp, and CRC patients. A combined predictive signature composed of 32 features from human and microbial small RNAs and DNA-based microbiome accurately classified CRC samples separately from healthy and adenoma samples.

Overall, the results presented in this thesis demonstrated the associations between the expression of specific sncRNAs and microbiome profiles measured in stool samples and different GI disorders, suggesting that their profiling in surrogate tissue could become a powerful tool for the diagnosis and monitoring of these pathologies.

1 INTRODUCTION

Gastrointestinal (GI) diseases refer to any condition or disorder occurring in the tract including the oesophagus, stomach, small intestine, large intestine and rectum, and the accessory organs of digestion, liver, gallbladder, and pancreas.

Many factors are involved in the development of GI diseases, including genetic, epigenetic, and environmental factors. Obesity, inactive lifestyle, diabetes, meat intake, fat-rich and fiber deficient diet, smoking habit, and alcohol consumption are among the principal risk factors characterizing the "Westernized lifestyle", which many studies have related to an increased incidence of GI diseases over the years [1].

Several conditions or disorders can affect the GI tract, leading to disability and poor quality of life for the patients and high healthcare costs [2].

Among the most frequent GI diseases, there are inflammatory bowel diseases (IBDs, which mainly include Chron's disease and ulcerative colitis), celiac disease (CD), colorectal polyps and colorectal cancer (CRC) (https://www.drugs.com/article/gastrointestinal-disorders.html). One of the main common features of these disorders is the presence of a chronic inflammatory condition leading to a cascade of events causing the alteration of the intestinal barrier and, as a consequence, an increased permeability [3]. Although all the molecules and mechanisms regulating the chronic inflammation processes are not fully understood, a number of evidence indicate a putative role of non-coding RNAs, especially microRNAs (miRNAs), in the stability and maintenance of gene expression patterns that characterize such inflammatory pathways [4]

Also "dysbiosis", defined as the alteration of the gut microbiome composition, has been observed in GI diseases [5] compromising the gut microbiome contribution to the maintenance of intestinal barrier function [6]. More recently, the interplay between miRNAs and gut microbiota has been reported highlighting miRNAs potentiality to regulate bacterial gene transcripts [7]. Unrevealing their relationship could bring new insights into the mechanisms at the bases of GI disorders and, hopefully, provide new potential biomarkers for the diagnosis and prognosis of these disorders (currently based on invasive, expensive and sometimes not effective procedures). Indeed, thanks to the recent advancements of Next Generation Sequencing (NGS) technologies, to date, the analysis of miRNome and microbiota are more affordable, accessible, and doable in different biospecimen types, including blood, saliva, urine, and stool [8, 9]. The latter, in particular, is the optimal biospecimen for investigations concerning GI disorders, considering that besides its minimal invasiveness, miRNAs from exfoliated fecal colonocytes are directly and continuously released into the intestinal lumen where also hundreds of bacterial species live in symbiosis with the host [10].

1.1 Celiac Disease

Celiac disease (CD) is a complex autoimmune disorder characterized by different forms and symptoms. The immune reaction is triggered by gluten ingestion which induces a succession of events leading to duodenal damage characterized by villous atrophy, intraepithelial lymphocytosis, infiltration of inflammatory cells in the lamina propria and crypt hyperplasia. Despite remaining asymptomatic in many cases, CD patients can exhibit intestinal and extraintestinal manifestations related to the GI tract and malabsorption. Classic symptoms include weight loss, chronic or recurrent diarrhea, abdominal pain and anorexia [11, 12]. Moreover, extraintestinal symptoms, such as arthritis, aphthous stomatitis, dental enamel defects, iron-deficiency anemia, osteoporosis, neurological and psychological problems may be present [13, 14]

To date, the only treatment for this pathology is the adherence to a lifelong GFD, strictly avoiding cereals containing gluten such as wheat, barley, and rye. This diet enables the disappearance of symptoms in symptomatic patients within few weeks from the beginning of GFD while serological and histological normalization may require from few months to one year [15].

1.1.1 Epidemiology

About three decades ago, CD was considered a disease mainly affecting children of Western Europe. Over the time, with the improvement of diagnostics including the application of CD specific serological tests transglutaminase 2 antibodies (TG2-Abs) and endomysial antibodies (EmAs), coupled with histological tests, a more reliable evaluation of its prevalence in the general population has become feasible [16]. However, epidemiological data on CD prevalence based on serological data are in general more reliable than the histological ones, considering that the small intestinal mucosal biopsy is not performed in all the seropositive patients [17].

It is estimated that CD affects about 1% of people worldwide (**Figure 2**) [18]. People of all ages can be affected, with a slight predisposition for women, with a ratio between 2:1 and 4:1 according to the countries [19].

In Europe, the overall prevalence of CD is 1%, on average, varying among different countries. The group of European countries with a high CD prevalence includes Germany (0.3%), Northern Ireland (0.9%) Italy (1.2%), Finland (2.0%) and Sweden (2,4%), whereas in Switzerland, Estonia, and Poland the disease is less common [11]. In the US population, rates similar to the European countries have been reported (0.7%)[20], as well as for other developed countries populated by individuals of European origin, like Australia [21] and New Zealand (0.4-1.3%) [22]. The presence of CD has been established also in many South American countries that are mostly populated by individuals of European origin with a mean prevalence of 0,6-1% [23]. Data about African prevalence are less abundant. However, the Sharawi, a black-haired African population originally living in Western Sahara, has been described as the population with the highest CD prevalence in the world (5.6%)[24]. The reasons for this high frequency are unclear but probably explainable with the highest frequencies of HLA-DQ2 and-DQ8 genotypes and the high gluten consumption of this population [25]. CD frequency is also relevant in the middle East with a prevalence of 0.8%, 1.5% and 0.8% in Iran, Turkey, and Israel, respectively [26-28]. In India, CD prevalence is estimated to range between 0.6% and 2.2% based on data coming from serological test on healthy blood donors [29]. In particular, the prevalence is higher in people of the Northern part of India which is in line with the wheat-rice consumption shift from the north to the south [30]. In China, both HLA predisposing genotypes and gluten consumption are largely diffused. However, very few data on Chinese CD prevalence are available to date. CD prevalence in Japan, Korea, Philippines, and other Pacific islands is very low probably due to low gluten consumption and low HLA genotypes diffusion (**Figure 2)**.

Based on serological data, CD prevalence worldwide is increasing over time even if the disorder still remains largely unrecognized [31, 32]. Indeed, it is estimated that for each clinically diagnosed CD patient, an average of five to ten seropositive individuals remain undiagnosed, usually because of atypical, minimal or often absent symptoms [33, 34].

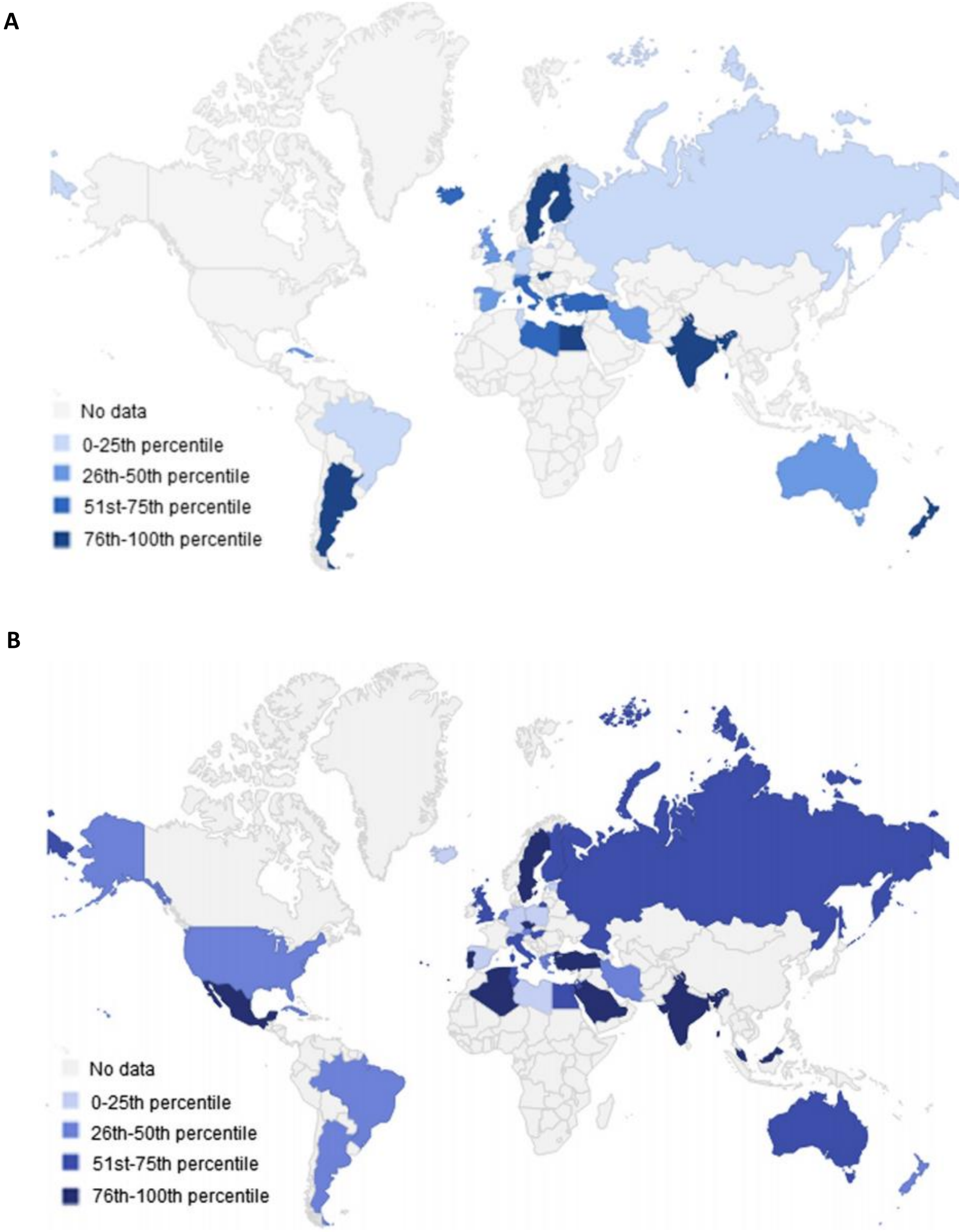


Figure 2. A. Worldwide CD seroprevalence rates (for only countries reporting data). The lowest and highest percentiles include countries with pooled national prevalence ranging from 0.2% to 0.8% and 2.1% to 8.5%, respectively [17]). **B**. Worldwide CD prevalence rates (based on biopsy) for the countries reporting data. Prevalence values were stratified into 4 groups of percentiles representing the 0 to 25th percentile (light blue) to the 76th to 100th percentile (dark blue). The lowest and highest percentiles include countries with a pooled national prevalence ranging from 0.2% to 0.4% and 0.9% to 2.4%, respectively [17].

1.1.2 Risk factors

The rise in the prevalence of CD is only partially explainable with the improvement of diagnostic criteria. This suggests that other factors might be involved behind these increasing rates [35].

Genetic factors are certainly among the most influential. Indeed, more than 90 % of CD patients carry one or two copies of the HLADQ2.5 which is encoded by the DQA1∗05 (alpha chain) and the DQB1∗02 (beta chain) genes. Interestingly, DQ2-negative CD patients are almost invariably HLA-DQ8 positive (DQA1∗0301/DQB1∗0302). The association of CD with HLA class II genes is explained by DQ molecule binding a peptide fragment of an antigen involved in the pathogenesis of CD to present it to T cells. Also, the major histocompatibility complex (MHC) class I region is associated with CD risk. Different studies mapping MHC association signal highlighted several new loci associations as risk factors independent of the HLA-DQ accounting for an additional 18% of CD heritability [36]. It is estimated that only 87% of the total CD heritability can be explained [37]. The MHC-HLA region accounts only for 41% of this percentage, while the remaining is shared by other non-HLA genes whose contribution is estimated to be about 6% globally, meaning that there is still a 40% of "missing heritability" [38]. This is also supported by data on monozygotic twins studies. A study by Nisticò et al. [37] showed that monozygotic twins of CD subjects have the 70% of probability to develop CD within 5 years from the diagnosis of the first twin; in dizygotic twins, this probability is reduced to 7%, indicating that other genes, in addition to HLA, increase the susceptibility to CD.

The CD prevalence also varies in populations with a similar genetic background [35]. Such variance may be explained by environmental factors rather than genetics. For instance, while infant feeding pattern association has been almost entirely denied, viral infections seem linked to CD. In particular, exposure to Adenovirus and Rotavirus gastrointestinal infections during early life and adulthood are linked to CD development even if the results still need to be cautiously considered [39, 40].

Other CD risk factors are organ-specific autoimmune disorders, even if the causal relationship still needs to be clarified[41]. Type 1 diabetes mellitus (T1DM) is the most severe form of autoimmunity associated with CD, with ∼5% of patients with CD having T1DM and vice versa [42]. Autoimmune thyroid disorders, including Hashimoto thyroiditis and Graves disease, are also very frequent autoimmune diseases associated with CD [43, 44].

1.1.3 Pathogenesis

The pathogenesis of CD is the consequence of gluten ingestion which can trigger an adaptive and/or innate immune response (**Figure 3)**[45]. Gluten is a protein mixture of gliadins and glutenins mainly contained in cereal grains wheat, rye, and barley. After its ingestion, gliadins and glutamins are partially hydrolysed by proteases of the gastrointestinal tract. The resulting fragments pass through the epithelial barrier of the small intestines entering the lamina propria where they are deamidated by tissue transglutaminase 2 (TG2)[46]. This modification increases the avidity of CD-associated gluten peptides to specific HLA variants expressed on antigen-presenting cells (APCs). Consequently, CD4+ gluten-specific T helper (Th) cells recognize the deamidated gluten-derived peptides presented by HLA-DQ2 or HLA-DQ8 and respond by expressing high levels of cytokines interleukin 21 (IL21) and interferon ɣ (IFN ɣ) [47]. These cytokines affect the epithelial cells and activate intraepithelial lymphocytes (IELs), licensing them to ‘kill’ the epithelial cells, ultimately leading to villous atrophy. Moreover, IL21 is critical for Th-cell-driven antibody responses and thus provides a link to B cells, a cell type that is now attracting more attention as an important player in CD.

Innate immunity plays a critical role in CD initiation. Cytokines such as interleukin 15 (IL15) and interferon α can prime the innate immune response by polarizing dendritic cells and intraepithelial lymphocyte function [48, 49]. These events occurring in the mucosa, together with the inhibition of the epithelial barrier function mediated by the gliadin-mediated zonulin release [50], enable the passage of undigested peptides from the gut lumen to the lamina propria. Once gliadin peptides cross the epithelial barrier, neutrophil recruitment through IL8 production [51] or a direct neutrophil chemoattractant effect leads to gluten intolerance in genetically susceptible individuals.

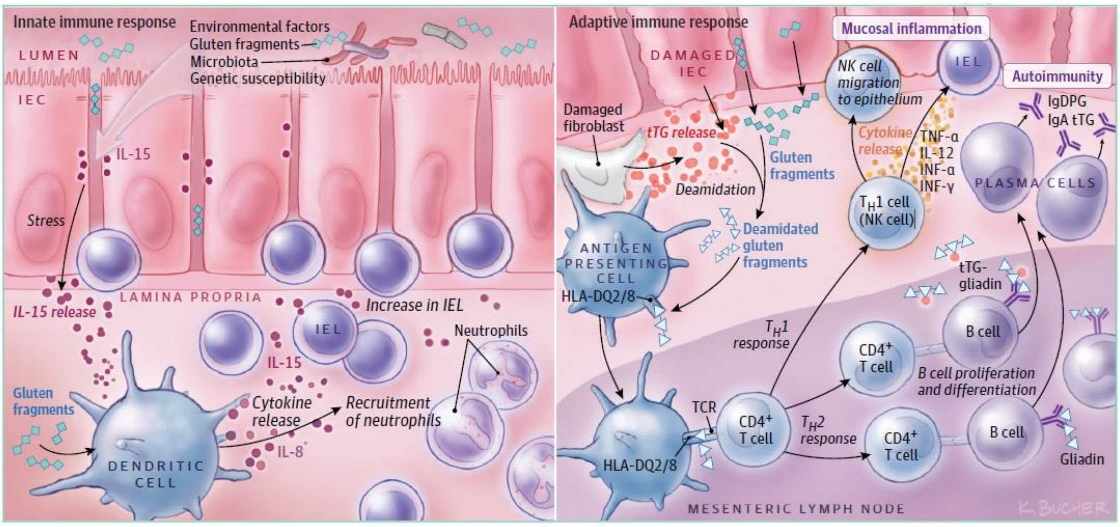


Figure 3. Immune response in CD: Innate immune response (left) and Adaptive immune response (right) triggered by gluten presence in the lamina propria [52].

1.1.4 CD diagnosis and classification

The diagnosis of CD, apart from typical clinical symptoms, relies on serological tests and subsequent confirmation by characteristic biopsy findings. The most reliable and used serological test include transglutaminase 2 (TG2) antibodies which have an excellent sensitivity (90–100%) and almost 100% specificity for CD [53]. For those subjects resulting positive to serological tests, the final diagnosis is generally based on the observation of small bowel mucosal villous atrophy, intraepithelial lymphocytosis and crypt hyperplasia through biopsy samples obtained upon gastroscopy [54](**Figure 4)**.

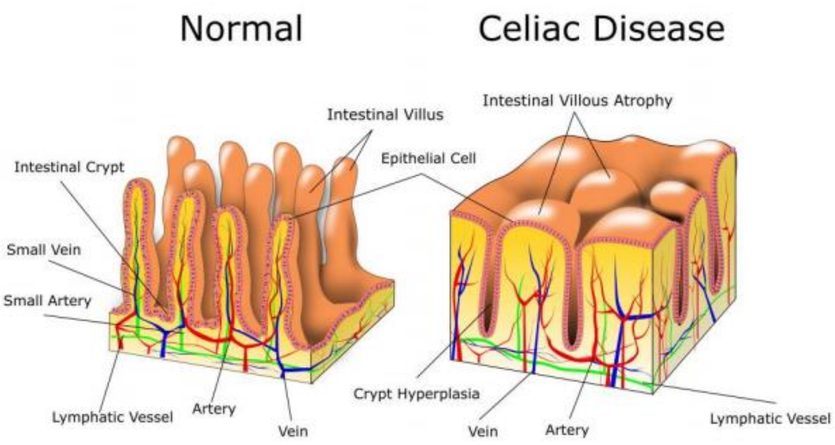


Figure 4: A representation of normal and CD intestinal villi.
Adapted from: https://www.dreamstime.com/stoc

CD is a disease with typical gastrointestinal symptoms and several, highly variable, non-gastrointestinal symptoms representing different types of CD [55](**Table 1).** Classical Celiac Disease is characterized by malabsorption symptoms such as diarrhea, failure to thrive, and weight loss and may occur both in adults and children. Non-classical Celiac Disease type has no important gastrointestinal symptoms or malabsorption but reflux, abdominal pain, bloating, vomiting, constipation, and dyspepsia can be present in some cases. It occurs in late childhood or adulthood and it is more common than the classic CD. About 70% of the subjects are diagnosed on the basis of extraintestinal symptoms associated with CD. About the 1-1,5% of CD patients can present a refractory celiac disease (RCD) [56]. This CD form is characterized by persistent or recurrent malabsorptive symptoms and villous atrophy despite strict GFD adherence. RCD is distinguished according to the normal (type I RCD) or abnormal (type II RCD) phenotypes of intraepithelial lymphocytes (IELs) and it is associated with serious complications, such as ulcerative jejunitis and enteropathy-associated T-cell lymphoma (EATL) [57]. Strict GFD is indispensable in RCD together with complementary treatments. In both types of RCD and RCD, the standard option consists of administration of open-capsule Budesonide which allows clinical remission and villous recovery in around 90% of both types of RCD [58]. Another type of CD is the Potential or latent celiac disease which occurs when an individual with a positive CD serology presents a normal small-bowel biopsy, with no characteristic villous atrophy [59]. Other non-celiac gluten-related disorders are Wheat Allergy, with an adverse immunologic reaction to wheat proteins and anti-wheat IgE antibodies production and Non-celiac gluten or wheat sensitivity (NCGS), typical of individuals with symptoms that respond to a gluten-free diet but without any CD histologic findings (e.g., characteristic findings on intestinal biopsy) or specific antibodies (e.g., tTG or EMA IgA;**Table 1)**.

Once CD diagnosis has been confirmed, the patient is usually followed up until its serological tests turn negative, which generally occurs within six to 12 months from the GFD starting. In parallel, a periodic examination of growth, nutritional status and the termination of disease manifestations should be evaluated. Moreover, since some subjects can heal gradually even if on a strict GFD, a follow-up biopsy to assess the intestinal villi healing can also be taken into consideration

Table 1: Different forms of gluten-related disorders.

Disorder	**Characteristics**
Classical Celiac Disease	Malabsorption symptoms such as diarrhoea, failure to thrive, and weight loss may occur both in adults and children.
Non-classical Celiac Disease	Absent or not important gastrointestinal symptoms, occurring in late childhood or adulthood. It is more common than the classic CD. 70% of diagnoses are made on the basis of extraintestinal symptoms.
Refractory Celiac Disease	Persistent or recurrent malabsorptive symptoms and villous atrophy despite a strict GFD adherence. The first-line drug treatment is typically a form of steroid medication with steroids.
Potential or latent Celiac Disease	Positive CD serology but normal small-bowel biopsy not presenting the characteristic villous atrophy
Non-celiac gluten or wheat sensitivity	Symptoms that respond to a GFD. No CD histologic findings or specific antibodies
Wheat Allergy	Immunologic reaction to wheat proteins and anti-wheat IgE antibodies production

1.1.5 Weaknesses and limitations in CD diagnosis and monitoring

Although in the last years the improvement in the CD diagnostic criteria has enabled an easier and more reliable detection of this disease, still nowadays there are some weaknesses and limitations in both CD diagnosis and management. First of all, about 10% of the affected patient presents a seronegative CD (SNCD) form. This happens because the antibodies remain in the intestinal mucosa forming immune-complexes unable to cross the lamina propria and enter the blood vessels [60]. A fraction of SNCD patients (about 2%) has a selective IgA deficiency/partial deficiency, which from a diagnostic point of view can be overcome with the IgG serological test [61]. Unfortunately, IgG test is not always reliable because of its lower sensitivity and specificity with respect to the IgA one, highlighting the need for a villous atrophy investigation by gastroscopy [62].

However, villous atrophy can also be the consequence of a medical treatment to cure viral or bacterial infections or a result of other non-celiac autoimmune enteropathies [63]. In addition, villous atrophy could appear as the final step of intestinal villi damage and thus take years to develop while other symptoms could turn up before the development of small intestinal lesion [64]. Another limitation, which still is a matter of discussion among the gastroenterologists, is the fact that the mucosal histology interpretation needs to be done on correctly cut, well oriented, and high quality samples in order to avoid erroneous diagnosis and misclassification [65].

CD patients monitoring after the diagnosis also needs to be improved. Indeed, excluding the symptoms disappearance and the serological tests normalization, to date, there is no definite indicator(s) for intestinal villi healing (which can improve gradually even if the patient is on a strict GFD), thus requiring a follow-up biopsy [66]. In this respect, molecular markers based on newly discovered sncRNA species, including microRNAs, or the gut microbiome composition, both detectable in surrogate tissues, may represent an interesting field of research.

1.2 Colorectal Cancer

Colorectal cancer (CRC) is an adenocarcinoma occurring in the colon or rectum, both of which are parts of the large intestine [67]. CRC usually develops from the epithelial cells of the large intestine: the accumulation of genetic mutations and epigenetic modifications in these cells arise into benign neoplasms (adenomas) and subsequently into invasive carcinomas. The result is a type of cancer not representing a single pathological entity, but rather a heterogeneous group of tumors arising through various molecular pathways [68]. Indeed, caecum or ascending colon cancers occurring in the right colon are biologically different from that in the left colon (from the splenic flexure down), both in terms of molecular characteristics and response to treatment.

CRC formation can be sporadic, being linked to predisposing lifestyle factors and ageing, or due to familial syndrome or even because of the presence of inflammatory bowel disease (IBD) including Crohn's disease and ulcerative colitis [69].

1.2.1 Epidemiology

The incidence and mortality of CRC consistently vary around the world. Globally, in males it is the third most commonly diagnosed cancer following lung and prostate tumors while the second in females behind breast cancer. According to the World Health Organization GLOBOCAN data of 2018, CRC encloses 11% of all cancer diagnoses with 1.8 million of new cases and almost 861,000 deaths, with rates substantially higher in males than in females and between the age of 50 and 65 [70].

CRC incidence varies a lot also by region, with up to eight-fold variations between countries. In countries undergoing a major developmental transition, incidence rates tend to rise uniformly with the increase of the Human Development Index (HDI), suggesting a causal relationship. Indeed, the highest incidence rates are recorded in countries with a recent economic development such as the Czech Republic and Slovakia, while remains high in Australia, New Zealand, Europe, and North America. The lowest rates are found in Africa and South-Central Asia (**Figure 5).**

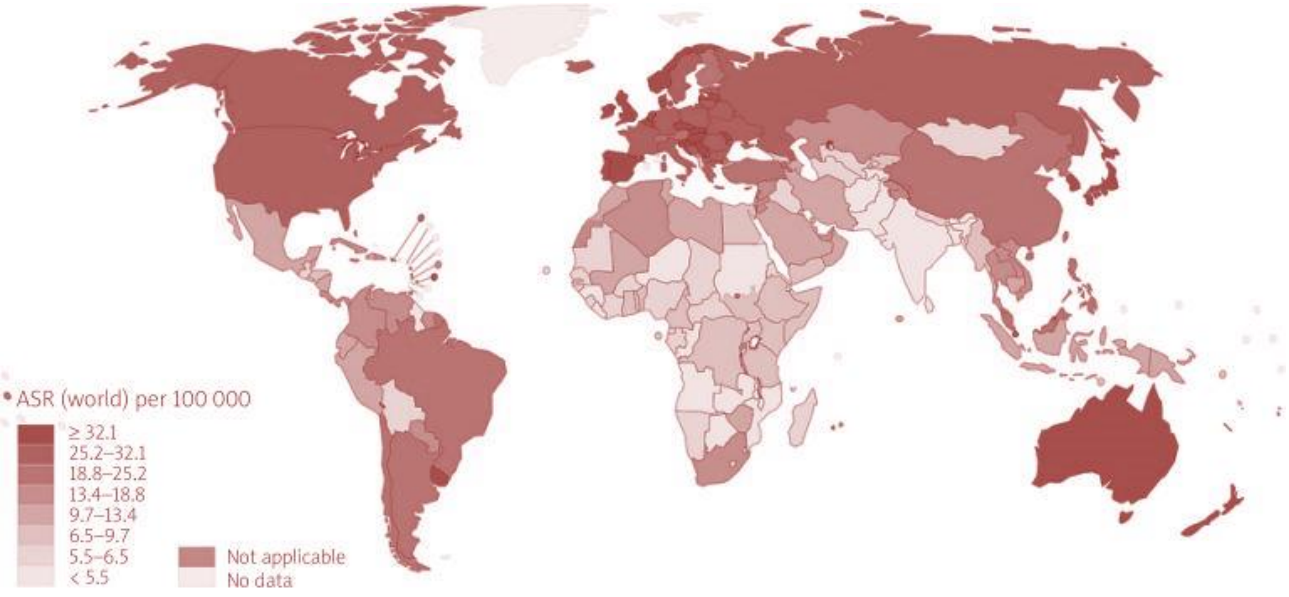


Figure 5 Colorectal cancer incidence rates worldwide in 2018. It includes all age and gender (age-standardised rates per 100 000; GLOBOCAN 2018).

These geographic differences could be mostly attributable to dietary and environmental exposure differences imposed upon a background of genetically determined susceptibility. Also, CRC mortality varies with the developmental condition of the nation, even if with a lesser degree than incidence, with a 2–3-fold difference between low and high HDI. In males, the age-standardised mortality is 12.8/100000 in high HDI nations and 5.7/100,000 in those with a low HDI, on the other hand for females the same rates are 8.5/100,000 and 3.8/100,000, respectively. However, in many Western countries, CRC mortality in the last years has progressively declined [71, 72]. This amelioration is partially explainable with an improvement of the screening strategies, enabling the detection of the tumor at an earlier stage and the preventive removal of colonic polyps, but also to a more effective primary and adjuvant treatments.

In Italy, CRC has a higher occurrence in the population, with nearly 52,000 new diagnoses in 2016 (13% of all new cancers) (http://www.registri-tumori.it). It is still the second most diagnosed cancer following the mammalian and the second deadliest cancer (19,407 total deaths in 2017, 10.8% of the totality of cancer deaths) after lung cancer. However, in the last decade, it has been observed a reduction in CRC Italian incidence (-4.1% and -3.0% annual mean for man and women, respectively). Notably, incidence and mortality estimates are heterogeneous across the country. In males, the incidence has slowly tended to stabilize, after a period of growth in the Centre-North while it continues to rise in the South. On the contrary, in the female population, CRC incidence is more stable across the regions with the lowest levels in the South.

1.2.2 Risk Factors

CRC is a complex disease with several well-established risk factors. Among those consistently increasing CRC risk, there are certain hereditary forms (including Familial adenomatous polyposis (FAP) and Lynch syndrome, which together account for the 4-5% of CRC cases), a personal or family history of sporadic CRC, inflammatory bowel disease, and a history of abdominal irradiation [73]. Other renowned factors are advanced age, male sex and those typical habits of “Westernized lifestyle” which includes smoking, excessive alcohol drinking, high consumption of red and processed meat (also including certain methods of cooking meat [74], overweight, and lack of physical activity [75]. On the other hand, some protective factors have also been identified. Decades of research on dietary factors suggested a protective effect against CRC onset of diets rich in fruits, vegetables, fish, fibers and whole grains, calcium and dairy products[76](**Table 2**). Epidemiological studies have also highlighted an association between circulating vitamin D concentrations and CRC risk [77]. Similarly, a link between high rates of coffee consumption and a reduced risk of this disease has been largely observed [78, 79]. Other important preventing factors are menopausal hormone therapy [80] and, interestingly, the use of aspirin and other nonsteroidal anti-inflammatory drugs (NSAIDs). Indeed, many observational and intervention trials evidence confirmed that regular assumption of aspirin and other NSAIDs reduced CRC risk of 20 to 40 % in average risk individuals [81, 82]. Since the elevated use of these drugs worldwide, their assumption may characterize a preventive strategy for this cancer. For this reason, several trials are examining the effect of aspirin administration on CRC development both in average-risk subjects and individuals with hereditary colorectal cancer (e.g. Lynch syndrome) and such a strategy could be extended to individuals at higher risk because of other risk factors. All the discussed risk and protective factors for CRC are reported in **Table 2**.

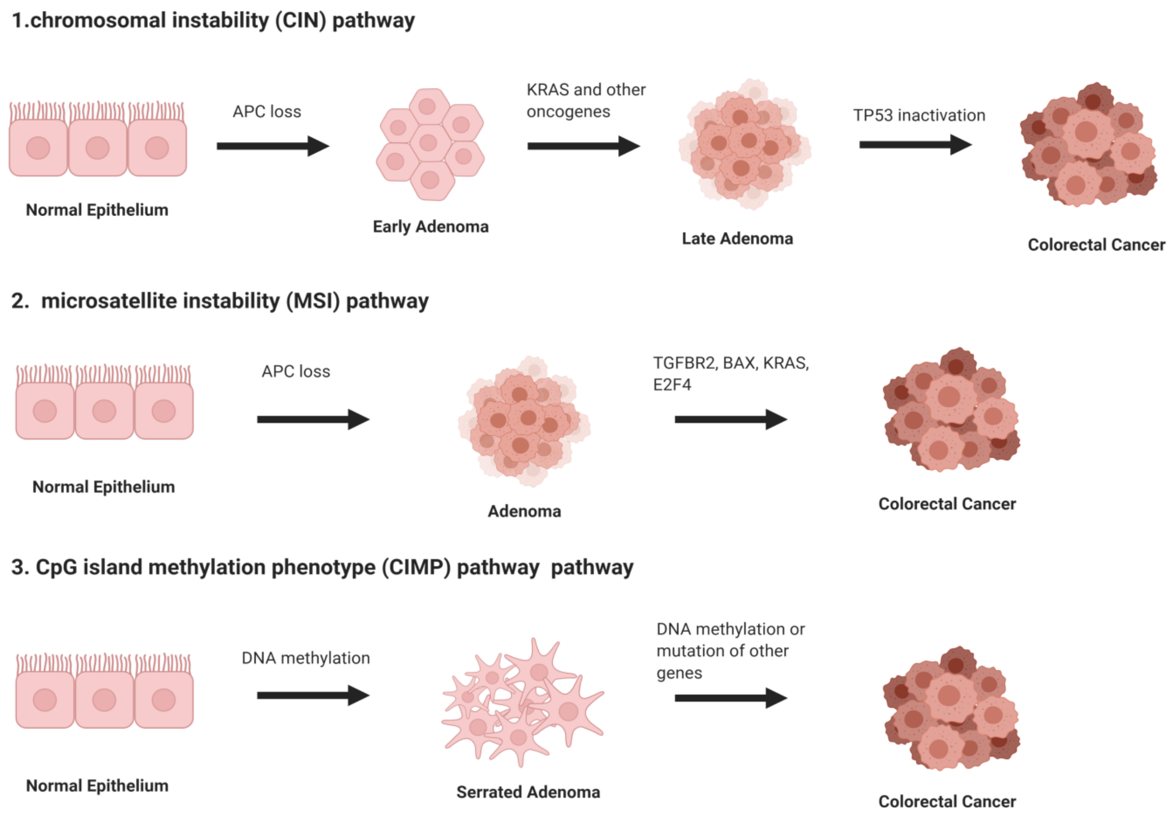
Table 2: Overview of risk factors of CRC [83].

Risk factors	
Sociodemographic factors	
Older age	+++
Male sex	++
Medical factors	
Family history	++
Inflammatory bowel disease	++
Diabetes	+
Helycobacter pylori infection	(+)
Other infections	(+)
Large bowel endoscopy	--
Hormone replacement therapy	-
Aspirin	-
Statin	(-)
Lifestyle factors	
Smoking	+
Excessive alcohol consumption	+
Obesity	+
Physical activity	-
Diet factors	
High consumption red and processed meat	+
Fruit and vegetables	(-)
Cereal fiber and whole grains	(-)
Fish	(-)
Dairy products	(-)
+++= very strong risk increase, ++=strong increase risk, +=moderate risk increase, --=strong decrease risk, -=moderate risk decreased. Parentheses show probable established but not fully association	

1.2.3 Pathogenesis

Several genomic alterations are linked to the development of CRC, mostly attributable to three different genetic pathways, namely 1) the chromosomal instability (CIN) pathway; 2) the microsatellite instability (MSI); and 3) the CpG island methylation phenotype (CIMP) pathway [84], all simplified in **Figure 6**. The CIN pathway includes mutation of APC and/or loss of chromosome 5q (i.e., a region including the APC gene, linked to the formation of the dysplastic aberrant crypt focus), mutation of the KRAS oncogene, loss of chromosome 18q and deletion of chromosome 17p, which contains the critical tumor suppressor gene TP53 [85]. The MSI pathway is another type of genomic instability accounting for about 10–15% of sporadic CRC. Microsatellites refer to sequences of repeated nucleotides scattered throughout the genome and MSI stands for the alteration in the number of nucleotide repeats found within these microsatellite regions in tumor versus germline DNA. MSI event leads to a dramatic increase in genetic errors and many microsatellites are present in genes implicated in colorectal carcinogenesis, such as *MSH3, TGFBR2, BAX, CASP5, MSH6, CTNNB1, APC, IGF2, KRAS* and *E2F4.* The CIMP pathway is the second most common pathway related to sporadic CRC. It provides the epigenetic instability necessary in sporadic cancers to methylate the promoter regions of key tumor suppressor genes such as *MLH1*, and thus epigenetically inactivate their expression. CIMP-positive CRCs are currently defined by a panel of CpG island methylation markers, which are classified as having or not DNA methylation within certain thresholds [86].

In CRC development, all these alterations can occur either separately or in combination, leading to the formation of tumors biologically different and enabling patients classification by function on their prognosis and management [87]. More recently, transcriptomic analyses have enabled a CRC classification into four consensus molecular subtypes (CMS) with distinct features: CMS1, accounting for 14% of CRC, shows hypermutated status, MSI and strong immune activation; CMS2 (37%), shows marked Wnt and MYC signaling activation; CMS3 (13%) is characterized by substantial metabolic dysregulation; and CMS4 (23%) is a mesenchymal subtype that exhibits prominent transforming growth factor-b activation, stromal invasion and angiogenesis [88].



Created with BioRender.com

Figure 6: Three genetic pathways involved in CRC pathogenesis: the chromosomal instability (CIN), the microsatellite instability (MSI) and the CpG island methylation phenotype (CIMP) pathways. The sequential genetic and epigenetic changes occurring in each pathway are simplified.

1.2.4 CRC classification

CRC is mainly classified in three different forms related to its origin and expression profiles.

Sporadic CRC is the most common form represented by 60 to 80 % of patients that do not carry any germline mutation known to be associated with this cancer [89]. From a histological point of view, these are adenocarcinomas that, as a rule, develop from a benign adenomatous polyp, which can be tubular, villous or tubulovillous in architecture. Only a limited proportion (estimated at 10%) of benign adenomas progresses to carcinoma; large adenomas with villous architecture have a high risk of progression [84]. More recently, a different multistep mechanism of carcinogenesis, namely the “serrated pathway”, has been described [90]. This pathway is responsible for about 10% of all CRCs and is characterized by serrated polyps replacing the traditional adenomas as the precursor lesion to CRC. Serrated polyps refer to a group of colorectal lesions that includes hyperplastic polyps, sessile serrated adenoma, traditional serrated adenoma and mixed polyps. From a genetic point of view, serrated polyps also exhibit a peculiar pattern, with *KRAS* and *BRAF* mutation having an important contribution to their development. MSI and the CIMP pathways, previously described, are also implicated in the serrated pathway [91].

Familial forms of CRC constitute 20–40% of the cases and occur in individuals where family members of primary consanguinity have suffered from sporadic colon cancer. So far, no genes have been associated to this form but probably a combination of environmental and inherited genetic factors plays a role in CRC development in these families. Colorectal adenoma (>10 mm) is the precursor lesions of CRC and it is present with high-grade dysplasia and/or a villous component, termed as advanced adenoma (ADA). A high prevalence of ADA has been described among young first-degree relatives aged 40-45 years as well as in older subjects. Additional risk factors are male sex and family history which increase the risk of developing CRC or ADA by 1.5-3.0-fold compared to the normal population [92, 93].

The third form of CRC is the hereditary one which can be distinguished in FAP and Lynch forms. FAP is the most common polyposis syndrome, accounting for approximately 1% of all CRC cases. It is classically characterized by the development of hundreds to thousands of adenomatous polyps (polyposis i.e. a malignant tumor with a high risk for developing non-digestive cancer) in both rectum and colon that, in general, begin to develop during the second decade of life, and nearly 100% of untreated patients will have malignancy by the age of 40-50 years.

Lynch syndrome, arising from a germline mutation in either the *hMSH2* or *hMLH1* mismatch repair gene, accounts for at least 3% of all the CRC cases [94]. This syndrome equally affects men and women in the same family, with a genetic alteration transmitted from parents without skipping any generation. The cancers associated with Lynch syndrome most likely affect the cecum or the right colon initially appearing as large and flat polyps or adenomas, with a high degree of dysplasia, and can or cannot be fluffy [95].

1.2.5 Limitations in CRC screening

In the last decade, as previously mentioned, a decline in CRC incidence and mortality has been observed, thanks to the adoption of effective screening programs. The methods traditionally adopted for screening include a spectrum of invasive and non-invasive tests [96]. Among invasive tests there are the endoscopic methods which have the advantage of being able to detect cancers (including non-bleeding lesions) and confirming the diagnosis histologically. These tests include flexible sigmoidoscopy, capsule endoscopy, and colonoscopy, the most reliable, which allows direct mucosal inspection of the entire colon and same-session biopsy sampling or definitive treatment by polypectomy in case of precancerous polyps or early-stage cancers. Non-invasive tests include guaiac-based fecal occult blood testing (FOBT) and fecal immunochemical occult blood test (FIT), both checking for hidden occult blood in fecal samples. However, both invasive and non-invasive tests present some limitations. Indeed, the endoscopic investigations are invasive and require a bowel preparation in the days before the inspection which is unpleasant for patients; moreover, they are costly for the national health system to be performed on a large scale [97]. In this respect, there is a large discussion about the proper age for starting the CRC screening: in the US healthcare system, this has been lowered from 50 to 45 years [98]. As a consequence, the management of this cancer will imply a further huge economic burden but a consistent change on the CRC rates and survival [99]. Also, FIT and FOBT tests, even if associated with a reduction of CRC incidence thanks to the effectiveness in detecting asymptomatic patients, are both not ideal for CRC detection since they are less sensitive to detect proximal compared to advanced distal neoplasia over multiple rounds of screening [100, 101].

For all these reasons, new biomarkers are constantly looked for with particular attention given to potential markers detectable in body fluids (including plasma, stool, and urine) by non-invasive (or minimally invasive) approaches. Among these, there are proteins, metabolites, and nucleic acids-based markers such as microRNAs and other sncRNAs or gut microbiome composition, which may all represent potential candidates for this research field.

1.3 Inflammatory Bowel Diseases

IBD refers to a group of chronic GI disorders, occurring in the colon and small intestine mainly including Crohn’s Disease (CrD) and Ulcerative Colitis (UC)[102]. Both these disorders are characterized by a chronic relapsing intestinal inflammation. Specifically, CrD induce a transmural inflammation affecting any part of the gastrointestinal tract (most commonly, the terminal ileum or the perianal region), while UC is characterized by a mucosal inflammation localized in the colon. Similar to CD and CRC, IBD arises as a result of the mixture of environmental and genetic factors leading to immunological responses and inflammation [102]. Indeed, the industrialized environment, the genetic make-up, the gut microbiota dysbiosis and the dysregulated immune response that cause chronic inflammation, are recurrent elements characterizing IBD [103].

1.3.1 Epidemiology.

IBD affects over 2 million individuals in North America, 3.2 million in Europe, and its incidence is increasing worldwide (**Figure 7**)[104]. This lifelong disorder occurs early in life, at 15-25 years old for CrD and 23-35 for UC, in both males and females. Historically, these diseases most commonly affected white people, particularly those of Ashkenazi Jewish heritage. However, over the last decade, an improved incidence was also observed in both Asian and Hispanic populations. In general, IBD incidence and prevalence markedly increased over the second half of the 20th century in industrialized countries. Therefore, IBD are considered the most prevalent GI diseases with accelerated incidence in newly industrialized countries [105]. Among various components of modern lifestyle, several of them have emerged as modifiers of systemic and intestinal immunity, such as antibiotics, diet, smoking and vitamin D intake. In addition, alterations of gut microbiota, derived from antibiotics exposure in children, gastroenteritis or lower level of lipopolysaccharide are associated with IBD onset [105].

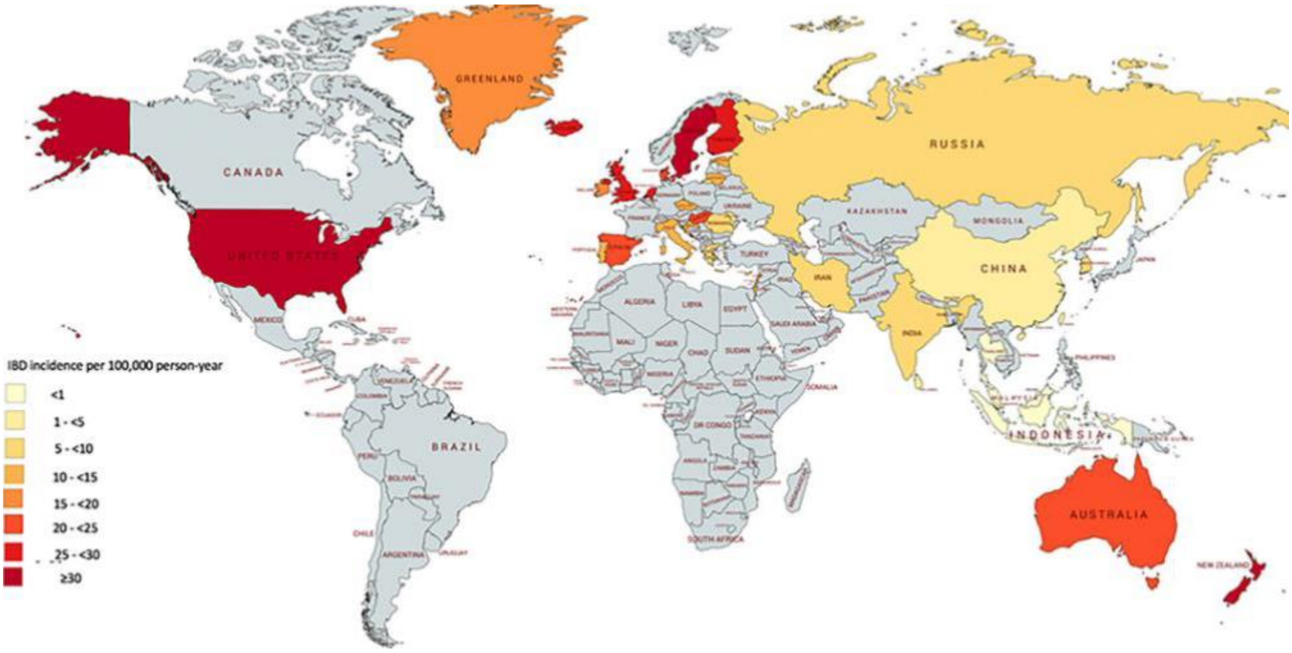


Figure 7. The latest reported worldwide IBD incidence according to population-based studies from 2010 to 2019. Adapted from Mak et al. [104].

1.3.2 Risk factors and pathogenesis.

IBD pathogenesis involves a complex interaction between distinct elements, such as genetic component, environmental factors, alteration of the intestinal microbiota and dysregulation of the innate and adaptive immune response at the intestinal mucosal level. The high risk of developing IBD in first-degree relatives of patients and data from twin studies have clearly demonstrated the role of the genetic component [105]: more than 201 polymorphisms have been implicated in the development of IBD, which is likely a polygenic process. Among them, 41 CrD-specific and 30 UC-specific genetic polymorphisms were identified, and 137 loci were associated with both CrD and UC. From 80% to 90% of the identified loci associated with IBD encodes for ncRNAs, such as miRNAs, revealing their involvement in the development of these diseases [106]. The contributions of *IL23R*, *NOD2* and HLA are very well-established but different genetic variants may have divergent effects: *NOD2* and *PTPN22*, for example, are risk factors for CrD but protective for UC. However, all these variants have low penetrance underlining the important role of environmental factors in IBD occurrence [106].

A dysbiotic gut microbiome has been extensively characterized in IBD. Some microbial populations have been found completely altered with an effect on the epithelial barrier permeability and immune response [105]. An altered immune response is a hallmark of IBD. The first indicator of intestinal inflammation is the infiltration of neutrophils in the gut mucosa and epithelium: this event affects epithelial barrier function and it leads to tissue damage and the perpetuation of inflammation through the release of multiple inflammatory mediators [107]. The adaptive immune response is also altered in UC and in CrD; various alterations in immunoglobulin subclasses and T cell populations are reported in both these pathologies. Moreover, also other factors are involved in IBD pathogenesis including damage-associated molecular patterns, regulatory RNAs, as well as epithelial, endothelial and mesenchymal cells [103]. Concerning intestinal barrier dysfunction, the increased intestinal permeability in patients with CrD is well-recognized and has also been associated with symptomatic status [108]. Conversely, patients with UC generally do not always have notable permeability defects [109].

Chronic inflammation in the colon is a key hallmark for CRC development: indeed, it is associated with an increased level of pro-inflammatory cytokines, damage of epithelial barriers, cell death, mutations in epithelial cells that, in turn, can initiate neoplastic growth. Patients with extensive IBD or diagnosed with IBD in childhood have a shorter life expectancy that may be related to an increased risk of CRC [110]. The risk of CRC in CrD is less known than in UC, where a direct relationship between UC inflammation and CRC occurrence was defined through the involvement of *TP53* mutations, altered miRNA expression and dysbiotic conditions [111].

1.3.3 IBD Diagnosis

Endoscopy remains the primary diagnostic method in IBD. However, more recently, other potential biomarkers have been assessed. An example is the C-reactive protein, resulting one of the most sensitive blood markers for inflammation although it is not highly specific. Indeed, its expression increases in presence of different tissue alterations, smoking, obesity, and drug therapies[112]. Also fecal calprotectin, an indicator of neutrophils migration to the intestinal mucosa, has been reported as reliable marker of intestinal mucosal inflammation [113]. In addition, the levels of this protein can also increase in the setting of diverticulitis, infectious colitis, intestinal neoplasms, cirrhosis, and in association with the use of nonsteroidal anti-inflammatory and proton pump inhibitors [105]. Other proposed biomarkers for IBD includes some serum antibodies, including perinuclear anti-neutrophil cytoplasmic antibodies (pANCA), which are antibodies that react with lysosomal enzymes in the cytoplasm of neutrophils and monocytes, anti-*Saccharomyces cerevisiae* antibodies (ASCA), antibodies of the mannan protein of *S. cerevisiae*, anti-granulocyte macrophage colony-stimulating factor (anti-GM-CSF) antibodies and other anti-microbial antibodies. However, all these mentioned antibodies have shown a low sensitivity, proving to be still far from representing ideal IBD biomarkers [114].

In the last years, numerous studies have evaluated the expression of miRNAs in both primary tissues and body fluid specimens from patients with IBD to define unique miRNA expression patterns that may distinguish IBD subtypes [115]. These molecules, together with other sncRNAs and gut microbiome profiles detectable in fecal samples, could constitute potential non-invasive biomarkers to improve the IBD diagnosis, monitoring as well as be a possible target for the treatment of this group of disorders.

1.4 SncRNAs and Microbiota as potential biomarkers of GI diseases

1.4.1 SncRNAs

Until a couple of decades ago, about 98.5% of the genome was thought to be inactive and considered by scientists as “junk” because of its non-coding nature. Thanks to the development and advancement of sequencing techniques, this “belief” has been overcome and we have realized that many areas of the genome had a biological functionality too. Among these genomic areas, there are non-coding genes which include introns, pseudogenes, repeated sequences, and cis/trans-regulatory elements that are transcribed in RNA without translation. It is estimated that 99% of the total RNA is constituted by ncRNA, with the number of validated ncRNAs increasing every year [116].

ncRNAs are currently classified by length in “long ncRNAs”, if longer than 200 nucleotides, and “small ncRNAs” ranging between 18 and 200 nucleotides. SncRNAs include various species: miRNAs (the most famous and extensively investigated), P-element-induced wimpy testis (PIWI) interacting RNAs (piRNAs), small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs) and transfer RNAs (tRNAs) (**Figure 8)**.

These molecules have many roles involving gene regulation through RNA interference, RNA modification or spliceosomal involvement [117]. Their expression is altered in correspondence of physiological events such as aging but also during progression of diseases such as cancer, cardiovascular, neurodegenerative and gastrointestinal diseases [118]. This aspect, together with the fact that they are easily detectable and remain stable in several body fluids, are the reasons why in the last years, these molecules gained a lot of interest for their potential application in diagnosis, prognosis, and therapeutics.

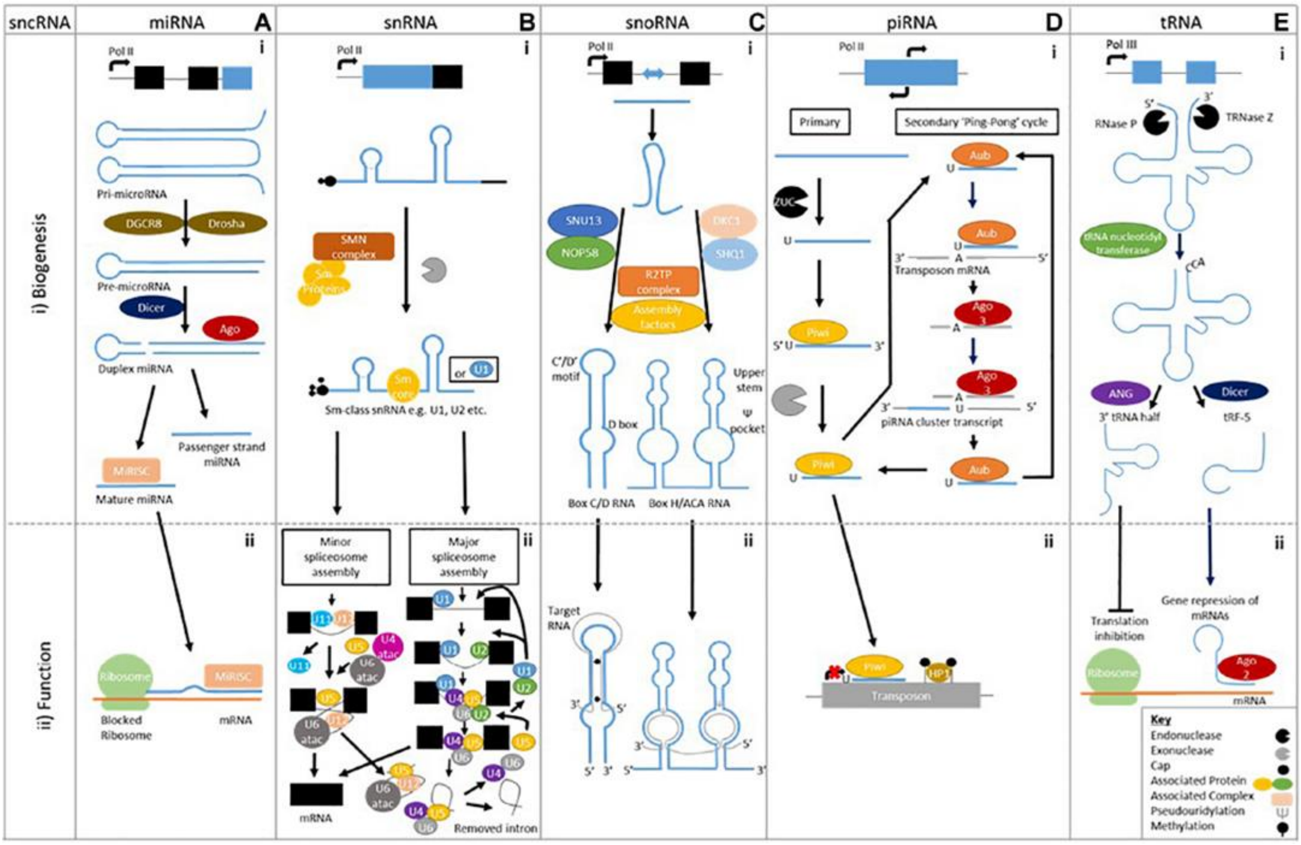


Figure 8. The main studied classes of sncRNAs. For each category, biogenesis and examples of the main function (s) are illustrated. **A** (i) MicroRNAs are single stranded ~22 bp sequences formed from double stranded precursors (ii) that prevent target mRNA translation. **B** (i) Small nuclear RNAs biogenesis is made up of two classes Sm class snRNA and Lsm-class snRNA (Not shown), (ii) which form the major and minor spliceosome. **C** (i) Small nucleolar RNAs have two different classes formed using different machinery; Box C/D RNA and Box H/ACA RNA, (ii) which cause methylation and pseudouridylation respectively. **D** (i) Piwi interacting RNAs are formed by either primary alone or by both primary and secondary biogenesis (ii) that prevent transposon translation through methylation. **E** (i) Transfer RNA cleavage forms transfer RNA derived fragments to be formed, (ii) which can prevent translation or cause gene repression (from [119]).

1.4.2 miRNA potentialities as biomarkers of human diseases

miRNAs are 22-nt-long sncRNAs whose interest is exploded since their discovery, as testified by the 119,416 and still growing PUBMED entries for these molecules. Their presence has been assessed in plant, animals and virus with the main role of RNA silencing and post transcriptional regulation of gene expression. This function is explicated with the binding of the “miRNA seed”, 5’ end of miRNA spanning from nucleotide position 2 to 7, and the 3’untranslated region (UTR) of the mRNA target. This binding leads to one of these consequences: i) the cleavage of the mRNA strand in two strands, ii) the destabilization of the mRNA through shortening of its poly (A) tail, and iii) a less efficient translation of the mRNA into proteins by ribosomes [120, 121].

In addition to this classical paradigm of miRNA functioning, other unconventional roles have also been described further improving the interest and potentiality in miRNA-based research field: Among the main promising there are pri-miRNA (the hairpin containing the primary transcripts) coding for proteins, miRNAs activating Toll-like Receptors, miRNA targeting nuclear ncRNAs etc.[122].

However, one of the main interests of miRNA research field is their potential role as human disease biomarkers. In particular, miRNAs raised to the attention in the field of molecular biomarkers because their dysregulation is associated to the initiation and progression of human tumors. The first evidence of this type came from the finding that miR-15 and miR-16 are down-regulated or deleted in most patients with chronic lymphocytic leukemia [123]. Starting from this discovery, in the past few years, a myriad of candidate or genome-wide miRNA expression profiling analyses have shown a general dysregulation of miRNAs in all tumors [124, 125], as well as in many other human diseases, including neurodegenerative and cardiovascular ones, viral infections, diabetes etc [126]. In addition, researchers have established peculiar characteristics of miRNAs (resistance to degradation by RNaseA, stability at high temperature, extreme pH, and freeze-and-thaw cycles) which make them suitable for their use both as diagnostic and as prognostic biomarkers of diseases [127, 128].

1.4.3 miRNAs in relation to CD

Despite research on miRNA alterations has been done already for more than a decade, only few studies explored their role in relation to CD (**Table 3)**. The first studies on this topic mainly focused on miRNA analyses on duodenal tissue while subsequently also miRNA circulating levels were investigated [129, 130]. One of the first work on this field has been performed by Capuano and colleagues which compared miRNA profiles extracted from bioptic tissues of CD and healthy control children. The expression of about 20% of the tested miRNAs resulted different between the two groups of patients investigated. Authors found that high miR-449a levels reduced both NOTCH1 and KLF4 in HEK-293 cells. NOTCH1 and KLF4 levels and the number of goblet cells were lower in small intestine of children with untreated CD but also in those on a GFD compared to controls, whereas more nuclear beta-catenin staining, as a sign of the WNT pathway activation, and more Ki67 staining, as sign of proliferation, were present in crypts from CD patients than in controls [131]. A similar study measured miRNA expression levels isolated from duodenal mucosa from adult CD patients and controls. Also in this case, several miRNAs resulted dysregulated, including miR-31-5p, miR-192-3p, miR-194-5p, miR-551a, miR-551b-5p, miR-638 and miR-1290 [129]. In addition, authors noticed that miR-192-3p levels were subjected to a specific modulation by gliadin peptides and that the miRNA cluster miR-192/194 was involved in matrix remodelling, possibly leading to cell apoptosis which, in turn, promotes the proliferative state of intestinal crypts.

Similar results were achieved by Magni et al in another study performed in the duodenum of adult CD patients and controls. Four miRNAs were validated as significantly down-regulated in CD and *in-silico* analysis revealed possible gene targets involved in innate and adaptive immunity [132]. Moreover, as already shown also by Vaira and colleagues, miR-192-5p and miR-31-5p expression were triggered by gliadin exposure in CD patients [129].

After the initial works performed on duodenal tissues, also the potentiality of circulating miRNAs has been evaluated [133, 134]. As an example, Buoli Comani et al. reported a lower expression for miR-192-5p and miR-31-5p in plasma samples of untreated CD individuals when compared with healthy controls. miR-192-5p was also decreased in treated CD patients compared to controls, whereas miR-31-5p and miR-21-5p returned to normal levels after at least one year of GFD [135]. Similar results were obtained by Amr et al. demonstrating an overexpression of miR-21 and a down-regulation of miR-31 in serum of untreated paediatric CD patients in comparison with healthy controls, without any significant difference between treated paediatric CD patients and controls [134]. Bascunan et al., besides duodenal biopsies, also measured miRNA levels in plasma, monocytes, and peripheral blood mononuclear cells (PBMCs). They found miR-146a, miR-155, and miR-21 in PBMCs, miR-155 in monocytes, and miR-155, miR-21, and miR-125b in plasma up-regulated in both untreated and treated CD subjects, while their expression in the intestinal mucosa did not change among all groups [133].

Altogether, these summarized findings represent a solid starting point for future investigations in the field. However, these results need to be weighed with caution until confirmatory studies are conducted. Moreover, studies analysing miRNA expression in stool samples and by more advanced technics based on NGS are still missing and this could represent the next step to achieve in the near future to assess miRNA potential usage as biomarkers of CD.

Table 3: Studies investigating miRNA expression levels in relation to CD

References	**Year**	**Country**	**N. of patients and specimens**	**N. of healthy controls and specimens**	**Technique**	**n. of miRNAs analysed**	**Main outcomes**		**Other findings**	**Sensitivity/ Specificity**
							Up-regulated miRNAs	***Down-regulated miRNAs***		
Bascunan K.A., et al.	2019	Chile	10 untreated and 20 treated CD patients PBMCs, monocytes, intestinal mucosa, and plasma	10 PBMCs, monocytes, intestinal mucosa, and plasma	qRT_PCR	4	miR-146a, miR-155, and miR-21 in PBMCs miR-155 in monocytes miR-155, miR-21, and miR-125b in plasma of both CD groups vs control subjects		Treatment with gliadin peptides, increased miR-146a and miR-155 expression in PBMCs and monocytes	miR-146a (AUC=0.91, 95% CI 0.83–0.99) miR-155 (AUC=0.92, 95% CI 0.86–0.99)
Amr K.S., et al.	2019	Egypt	25 untreated and 25 treated CD patients (on a GFD) serum	20 serum	qRT_PCR	2	miR-21 in untreated vs treated CD and control subjects	miR-31 in untreated vs treated CD and control subjects	miR-21 expression level positively correlates with the tTG IgA auto-antibodies	miR-21 (AUC=0.85, 95% CI 0.70 - 0.99) miR-31 (AUC=0.801, 95% CI 0.65 - 0.94)
Comincini S., et al.	2017	Italy	23 untreated CD patients blood 25 untreated CD patients duodenal biopsies	33 blood 24 duodenal biopsies	qRT_PCR	2		miR-17 and miR-30a in blood and duodenal biopsies of untreated CD vs controls subjects	ROC analyses did not identified any miRNAs as able to distinguish among CD patients and controls	
Comani G.B., et al.	2015	Italy	17 untreated and 7 treated CD patients plasma 20 untreated CD patients duodenal	12 plasma 8	qRT_PCR	6	miR-21-5p in plasma of untreated CD vs control subjects miR-21-5p, miR-	miR-192-5p in plasma of treated CD vs controls subjects miR-192-5p	plasma miR-31-5p and miR-21-5p returned to normal levels after at least 1 year of a GFD.	

			biopsies	duodenal biopsies			21-3p, miR-486-5p in duodenal biopsies of untreated CD vs control subjects	and miR-31-5p in duodenal biopsies of untreated CD vs control subjects		
Vaira V., et al.	2014	Italy	15 CD patients (5 untreated with anemia (CA), 5 untreated with classical symptoms (CC) and 5 treated) duodenal biopsies 22 untreated CD patients (10 CA and 12 CC) duodenal biopsies (*validation*)	5 duodenal biopsies 12 duodenal biopsies	Array (*discovery*) qRT_PCR (*validation*)	377 36	miR-638 in CA vs CC and control subjects; miR-551b-5p in CC vs CA and control subjects miR-1290 in CC and CA vs control subjects	miR-194-5p in CA vs CC and control subjects miR-31-5p, miR-551a and 192-3p in CD vs control subjects	miR-192-3p levels were subjected to a specific modulation by gliadin peptides	
Magni S., et al.	2014	Italy	6 untreated CD patients duodenal biopsies 21 untreated CD patients duodenal biopsies	5 duodenal biopsies 10 duodenal biopsies	Array (*discovery*) qRT_PCR (*validation*)	377 7		miR-192-5p, miR-31-5p, miR-338-3p, and miR-197 in CD vs control subjects		
Capuano M., et al.	2011	Italy	20 untreated and 9 treated CD patients duodenal biopsies	11 duodenal biopsies	Array (*discovery*) qRT_PCR (technical *validation*)	365 3	miR-449a in untreated and treated CD vs control subjects			

CD "treated" refers to those CD subjects on a GFD at least from one year
In studies with both discovery and validation phases only validated results were reported

1.4.4 miRNAs in relation to IBD and CRC

A role in pathogenesis and as diagnostic and prognostic biomarkers has been attributed to miRNAs in IBD [136]. Most of the studies on this field measured levels of miRNAs in blood, stool and tissue by microarray profiling, RT-qPCR, and NGS. In detail, miR-21, miR-155, and miR-31 have been repeatedly related to IBD. miR-21, become over the years one of the most studied, was reported in many studies, including those analysing mouse models, demonstrating its involvement in the loss of tight junctions in intestinal epithelial cells [137]. Some studies showed that miR-31 is a regulator of the inflammatory response while miR-155 contributes to the pathogenesis of IBD directly binding to SHIP-1 mRNA, which regulates cell membrane trafficking [138]. Altered miRNA levels associated with IBD were also found in fecal samples, such as up-regulation of miR-223, miR-155, miR-21 and miR-1246 [139, 140].

A larger number of evidence are available on miRNA expression in the context of CRC. Indeed, in the last decade, several studies illustrated the role of miRNAs in CRC onset, progression treatment and diagnosis/prognosis. These investigations were conducted on primary tissues as well as surrogate biospecimens including whole blood, plasma, serum, extracellular vescicles (EVs) and stool [8, 9].

A consistent number of studies performed miRNA analysis in primary tissue showing miRNAs dysregulation in CRC compared to adjacent non-malignant tissue [141]. Among these, the up-regulation of miR-21 together with that of miR-31, miR-92a, miR-135b, and miR-200 family has been extensively reported [142]. Similarly, miR-34a, miR-145, miR-143, miR-195 and miR-378 are among those repeatedly observed down-regulated in CRC vs adjacent non-malignant tissue [143]. Some of these miRNAs were also reported dysregulated in colorectal adenomas, particularly advanced colorectal adenomas (colorectal polyps greater than 1 cm in diameter and/or with villous component and/or severe dysplasia), which are recognized as critical premalignant lesions for CRC development and are the primary target lesions for CRC screening. Indeed, miR-21 is recognized as oncogenic miRNA in CRC, and it is frequently overexpressed also in colorectal adenoma tissues compared with normal colonic mucosa [144].

In addition to studies focusing on primary tissues, several other researchers have demonstrated that colorectal adenoma and carcinoma have unique expression profiles of various classes of RNAs in serum, plasma and EVs suggesting that their determination as circulating markers could provide a novel and promising early diagnostic option for CRC screening [142, 145]. In particular, miR-92a-3p, miR-17-5p, miR-29a and miR-196b-5p showed the best performances in this context and, together with miR-21, were repeatedly observed dysregulated in CRC [8].

Also, fecal miRNA analyses in the context of CRC have been extensively explored [9]. This could be due to the strong rationale for determination of sncRNA expression levels in stool which includes the following observations: i) colonocytes are continuously shed into the fecal stream, with a periodicity of replacement roughly every 3–4 days, and neoplastic cells exfoliate at even a higher rate; ii) tumor secreted non-coding RNAs (mainly miRNAs) are directly and continuously released from the tumors into intestinal lumen; iii) alterations in the expression of oncogenic or tumor suppressive non-coding RNAs are very specific for pre-cancer or cancer; iv) sncRNAs are extremely stable, enabling accurate and reproducible detection in the stool without need of special stabilization or logistical requirements [146].

The first study reporting stool miRNAs as potential biomarkers in CRC was conducted by Ahmed et al in 2009 [147]: miRNA expression was determined in colonocytes extracted from stool specimens of CRC and ulcerative colitis patients as well as healthy controls. Authors identified seven up-regulated miRNAs (miR-20a, miR-21, miR-92, miR-96, miR-106a, miR-203, and miR-326), and seven down-regulated (miR-16, miR-125b, miR-126, miR-143, miR-145, miR-320, and miR-484-5p) in CRC vs controls individuals. In another study, Link et al. have compared stool samples of patients suffering from CRC and adenoma with healthy individuals and have found a higher expression of fecal miR-21 and miR-106 in CRC patients [148]. Wu et al., examined the fecal expression of miR-21 and miR-92a not only in CRC patients but also in individuals with adenomatous polyps. Both miR-21 and miR-92a were overexpressed in tumors. Regarding the localization of CRC, they found that miR-92a showed a higher sensitivity in distal tumors. This phenomenon can be explained by the denser consistency of the stool in the distal parts of the large bowel, so a greater amount of tumor cells shed into the feces. The fecal expression of miR-21and miR-92a was found to be lower postoperatively or after the endoscopic removal of the adenomatous polyps [149].

Two miRNAs, miR-221 and miR-18a, known to be up-regulated in CRC tumor tissue, showed an increasing expression levels also in stool samples of stages I-IV CRC patients, independently on the location of the tumor, or previous antibiotic intake [150].

Recently, the expression of five fecal miRNAs (miR-19-b-3p, miR-20a-5p, miR-21- 3p, miR-92a-3p, miR-141) was found to be significantly higher in CRC patients compared to healthy subjects, and their expression significantly decreased after curative surgery [151].

In a recent study, Chang PY et al. evaluated miR-223 and miR-92a expression levels in stool and blood plasma in CRC. This combined approach yielded the highest sensitivity of 96.8% and specificity of 75% for CRC (AUC = 0.907). These results established a two-miRNA signature in two types of CRC clinical specimens with a high sensitivity for CRC detection [152].

The detection of miRNAs in stool may be another non-invasive screening method for CRC [153]. However, there are some limitations due to the complexity and density of the stool but also to the higher intrinsic variability to daily changes compared to blood serum/plasma. For this reason, further investigations and validation using standardized protocols on large cohorts are necessary before such markers can be seriously considered for adaptation in the clinic for non-invasive CRC screening [142].

1.4.5 Evidence on other sncRNAs

miRNAs are the best known and promising sncRNAs for their potential role as biomarkers. However, other sncRNA species such as piRNAs, snoRNAs, and tRNAs are also gaining attention as key component of cellular regulation and thus might be potentially assessed as biomarker of diseases [80, 142]. These molecules, together with miRNAs, are also among the most numerous sncRNAs human biotypes, accounting for 32,007 piRNAs, 231 snoRNAs and 2,723 tRNAs according to the last updated version of RNACentral (https://rnacentral.org/) and DASHR (https://dashr2.lisanwanglab.org/) databases.

piRNAs are sncRNAs of 24–31 nucleotides in length mainly involved in the epigenetic and post-transcriptional silencing of transposable elements and other genetic elements in germ line cells [154]. These molecules, similarly to miRNA precursors, are characterized by the absence of specific sequence motifs or secondary structures. Despite their large diversity, most piRNAs can be mapped to a relatively small number of genomic regions called piRNA clusters. In contrast to miRNAs, these are Dicer-independent and interact with the PIWI subfamily of Argonaute proteins involved in the regulation of genome stability. PIWI proteins are involved in gene regulation through RNA degradation and have been linked to DNA methylation [155]. snoRNAs are non-coding RNAs with a length of 60-nt conserved from archaebacteria to mammals [156]. They are present in the nucleolus in associations with proteins to form small nucleolar ribonucleoproteins (snoRNPs) and are responsible for sequence-specific 2′-O-Methylation of ribosomal RNA (rRNA). In addition to this main role, recent reports highlighted tumor-suppressive or oncogenic functions in various cancer types including inactivation of growth suppressors and cell death, activation of invasion and metastasis, and sustained proliferative signalling [157]. Therefore, in this context, snoRNAs could have potential applications in cancer diagnosis and therapy.

tRNAs originate when nucleases (such as Dicer and ANG) cut the tRNA ring in a given cell/tissue and under specific conditions such as cell stress [158]. tRNAs are divided into two main types based on the cleavage sites: tRNA-derived fragments (tRFs) and tRNA-derived stress-induced RNAs (tiRNAs, also known as tRNA halves). tRFs originate precisely from the extreme 5' (tRF-5) or 3' ends (tRF-3) of mature tRNAs or from the 3' trailer sequence of precursor tRNA transcripts (tRF-1); tiRNAs are produced by specific cleavage at the mature tRNA anticodon loop of over 31 nt (including 5'- and 3'- fragments, named 5'-tiRNAs and 3'-tiRNAs, respectively) induced by situations of stress or starvation [159]. Despite their role is still poorly understood, tiRNAs are commonly known to regulate gene expression and epigenetic inheritance [158]. This is in line with their altered expression levels observed in several diseases which poses these molecules as candidates for biomarker research.

Even if the studies on other sncRNAs than miRNAs are only at an initial stage, it seems really interesting to investigate their role as molecular biomarkers. Indeed, piRNAs, snoRNAs and tRNAs, measured in different biospecimens, have all been proposed as potential biomarkers in several human diseases [160]. In a study conducted by Cheng and colleagues [161], it has been reported that piR-651 expression in gastric, colon, lung, and breast cancer tissues was higher than that in paired non-cancerous tissues. This finding was also confirmed in different cancer cell lines including gastric, lung, mesothelium, breast, liver, and cervical ones observing that the growth of gastric cancer cells was inhibited by a piR-651 inhibitor and arrested at the G2/M phase. Thus, piR-651 might be involved in the development of gastric and other cancers and for this reason usable as potential biomarker. The up-regulation of a group of piRNAs was also seen by Herrera et al. in relation to CRC. Specifically, authors found a panel of 50 ncRNAs dysregulated in cancer-associated fibroblasts (CAFs) and non-malignant mucosa derived fibroblasts (NFs). Among the observed signals, besides lncRNAs, snRNAs and miRNAs, also 7 piRNAs were up-regulated in CAFs vs NFs cells [162]. On the other hand, one of the first implications of snoRNAs in carcinogenesis comes from the study conducted by Mei YP and colleagues [163] on the putative oncogene snoRNA42 (SNORA42) associated with carcinogenesis. snoRNA42 is located on chromosome 1q22, a genomic region frequently subjected to amplification in lung carcinomas and the over-expression of snoRNA42 is observed in non-small-cell-lung carcinoma (NSCLC). Repression of SNORA42 in NSCLC cells caused a marked decrease in lung cancer growth *in vitro* and *in vivo*; enforced SNORA42 expression in bronchial epithelium increased cell growth and colony formation. Thus, this study on SNORA42 provided evidence for the functional importance of snoRNAs in cancer showing its involvement in tumor development.

A number of evidence is also available on tiRNAs and tRFs in relation to various disorders, including CRC. Indeed, Xiong et al. found 16 tRFs and 5 key miRNAs significantly altered in CRC compared to matched non-malignant tissues [164]. Target mRNAs of the 16 tRFs and 5 miRNAs were primarily involved in vitamin metabolic pathways and the cyclic guanine monophosphate-protein kinase G signaling pathway, indicating their potential roles in CRC development. Similarly, Li et al. showed that 5'-tiRNA-Val was up-regulated in CRC patients vs controls and in highly metastatic cells being involved in angiogenin (ANG)-mediated tumor metastasis [165].

Taken together, all these results suggest a promising role for other sncRNAs in addition to well-known miRNAs as potential target in several disease diagnosis, monitoring and prognosis. Even if their functions and mechanisms need to be clarified, their regulation in physiological and pathological processes seems to be similar to that of miRNAs, increasing researchers' expectations for their clinical usage. However, the knowledge of sncRNAs related to GI diseases and their expression and detectability in different biospecimens is very limited. Besides miRNAs, very few studies investigated sncRNAs expression in stool and plasma samples in relation to CRC, while no studies at all are available on CD, which is one of the main issue of the present work.

1.4.6 Microbiome analysis: a complex research field with promising applications

Human microbiota refers to the variety of species inhabiting our body. These are microbes, bacteria, archaea, fungi and viruses present in a larger number of our own cells in the surfaces and specific niches of our organism such as gut, skin, mouth, etc. [13]. Decades ago, a limited knowledge was available about the human microbiota and this was mostly based on culture methods (it has been estimated that 20%–60% of human microbes is uncultivable). More recently, with the development of culture-independent methods to study microbiome composition such as NGS methodologies [166] and the parallel implementation of powerful computational tools for "omics" technologies, microbes and their genes have been efficiently catalogued, and the impact of the host-microbe interaction on human metabolism has started to be widely investigated and elucidated [167, 168]. The two most commonly used methodologies for microbial identification and genotyping are based on gene amplicon/marker genes (e.g. 16S rRNA) and shotgun metagenomic sequencing [169]. The most widely used is the 16S rRNA (or 16S rDNA) sequencing, the gold standard in microbial typing. However, in recent years, it has been overcome by metagenomics approaches. Unlike 16S sequencing, which only targets 16S rRNA genes, shotgun metagenomic sequencing technique sequences all given genomic DNA from a sample hence providing a better taxonomic resolution and genomic information [170].

To date, several functions have been attributed to the gut microbiome testifying its importance. These include the fermentation of indigestible food components into absorbable metabolites, the synthesis of essential vitamins, the removal of toxic compounds, the strengthening of the intestinal barrier, and the stimulation and regulation of the immune system. Indeed, the gut microbiota has systemic effects via secretion of anti-inflammatory chemokines, metabolites, antimicrobial and neuropeptides and the induction of immune activation (activity on dendritic cell and macrophage subsets, T cell priming and polarization in the mesenteric lymph nodes) [171, 172]. The relevance of human microbiome has also been testified by evidence reporting a dysbiosis associated with a variety of diseases such as cancer, metabolic and neurodegenerative disorders, chronic fatigue syndrome and GI diseases [173-176]. In particular, the analysis of human gut microbiome composition in relation to GI have been the focus of several studies highlighting gut dysbiosis in IBD, CD and CRC opening a new field of research with potential implications in the context of GI diagnosis, monitoring and therapeutics.

1.4.7 Microbiome in CD

Many studies have been performed to explore microbiome composition in both adult and pediatric forms of CD. These studies include experimentation performed on different biospecimens including duodenal biopsies, saliva, and feces, mainly conducted by 16S rRNA gene sequencing. In general, changes in the number of bacterial species, their diversity, and proportions have been described in CD as well as in and other GI. A reduced number of bacteria with anti-inflammatory capacity and increased bacteria with inflammatory capacity were reported in patients with CD diseases compared to healthy individuals [177]. The most consistent changes observed are a reduction in the diversity of gut microbiota and changes in the abundance of Firmicutes/Bacteroidetes. One of the first studies on this field noticed an increased association of rod-shaped bacteria in small bowel biopsies of both treated and untreated CD patients with respect to controls [178]. Subsequent studies performed in both feces and duodenal biopsies reported an increased abundance of gram-negative bacteria such as *Bacteroides*, *Clostridium*, *E.Coli* in CD patients compared to healthy adults [179, 180]. Since these initial findings, other studies on fecal samples and duodenal mucosa have reported similar results [177, 181, 182]. Oral microbiota in CD was also investigated analyzing saliva and oropharyngeal swabs [183, 184]. As a result, in contrast with the findings reported for stool and duodenal biopsy biospecimens, a reduction of *Bacteroidetes* and *Fusobacteria* and an increase of *Actinobacteria* was reported in CD patients compared to controls. A summary of the available studies is reported in **Table 4**.

Overall, the analyses of different biospecimens from CD patients compared to healthy subjects have shown dysbiosis in relation to the disease. Although not being always concordant, a large part of them has revealed an increased number of gram-negative bacteria, *Firmicutes*, *E. Coli, Enterobacteriaceae*, *Staphylococcus*, and a decrease in *Bifidobacterium*, *Streptococcus, Provetella* and *Lactobacillus spp*. However, there are also contradictory findings among the studies, mainly about the microbiome composition in treated CD patients compared to controls, where the GFD seems to be not always able to restore a normal microbiota in CD patients [185]. In addition, as for other diseases, from the available studies it is difficult to determine whether an altered gut microbiota is the cause or the consequence of CD, considering that diet can also modulate the gut microbiota over time.

Table 4. Gut dysbiotic features observed in CD**.** Table adapted from [185].

Biological sample	**CD associated dysbiosis**	**References**
Feces	↑ Gram (−)/Gram (+) bacteria ratio ↓ Firmicutes (Lactobacillus spp., Fecalibacterium prausnitzii, Clostridium spp.) ↓ Actinobacteria (Bifidobacterium spp.) ↑ Bacteroidetes (Bacteroides spp.) ↑ Proteobacteria (E. coli) ↑ Firmicutes (Staphylococcus spp.)	Sanz et al. Di Cagno et al. Collado et al. De Palma et al. Quagliariello et al. Olivares et al.
Duodenal Mucosa	↑ Gram (−) bacteria ↓ Firmicutes (Lactobacillus spp., Streptococcus spp.) ↓ Bacteroidetes (Prevotella spp.) ↑ Proteobacteria (Neisseria spp., E. coli)	Collado et al. Schippa et al. Di Cagno et al. Nistal et al. Wacklin et al. Sánchez et al. D’Argenio et al. Iaffaldano et al.
Saliva	↓ Bacteroidetes ↓ Fusobacteria ↑ Actinobacteria ↓ Bacteroidetes	Tian et al.
Oropharingeal swab	↓ Bacteroidetes ↓ Fusobacteria ↑ Actinobacteria (Actimomyces spp.) ↑ Proteobacteria (Nf)	Iaffaldano et al.

1.4.8 Microbiome in IBD and CRC

Evidence on gut microbiome composition in relation to IBD are less abundant compared to those of CD [186]. In general, metagenomic exploration in this regard demonstrated an increase of *Proteobacteria* (mainly *E. coli* species), *Pasteurellaceae, Veillonellaceae, Fusobacterium* species, and *Ruminococcus gnavus* in IBD patients compared to controls. On the contrary, a decrease in abundance and overall diversity of anti-inflammatory taxa was associated with IBD, among which *Bacteroides Bifidobacterium, Clostridium, Roseburia* and *Surella* species together with *Fecalibacterium prausnitzii* were among those mainly reported [187]. However, it remains difficult to understand whether these microbiota changes in IBD patients are causative or consequence of the inflammatory process, treatment, or both.

A much larger number of studies has focused on microbiome analysis in relation to CRC, mainly analyzing intestinal mucosa and fecal samples. A number of evidence proposed a link between intestinal dysbiosis and CRC also reporting specific microbiome signatures for different subtypes of CRC as well as the enrichment of some oncogenic microbes in CRC cases in comparison with healthy controls [188, 189].

So far, several bacterial species have been linked to CRC. Among these, *Streptococcus bovis* (S. bovis), a gram-positive cocci, has been proposed as a CRC risk factor [190, 191]; *Enterococcus fecalis* (*E. fecalis*) was observed enriched in CRC vs healthy controls individuals [192, 193] and its infection induces superoxide production, thus damaging DNA in epithelial cells [194, 195]. Similarly, Yu et al. found an higher amount of *Peptostreptococcus anaerobius* (*P. anaerobius*) in fecal and mucosal microbiota from patients with CRC compared to those of controls [196, 197]. In addition to the aforementioned bacterial species, *Fusobacterium nucleatum* has been repeatedly associated to CRC. This bacteria was found enriched in human colorectal adenomas and carcinomas, probably contributing to disease progression from adenoma to cancer [174, 198, 199].

Recent evidence have also hypothesized *F. nucleatum*'s effects on the host miRNome as one potential major contributor of CRC onset. In this respect, a study by Yang, Y. et al. showed the ability of *F. nucleatum* to up-regulate miR-21, activating the TLR4–MyD88 signaling cascade and thus leading to an increase of CRC cell proliferation and tumor growth in mice [200]. Accordingly, CRC patients with high levels in tissue of both *F. nucleatum* DNA and miR-21 showed an increased risk for poor outcomes. Interestingly, in addition to the miR-21 up-regulation, *F. nucleatum* infection led also to miR-18a and miR-4802 down-regulation, thus promoting the TLR4–MyD88 signaling pathway activation [201]. Consequently, the key components of the autophagy pathway ULK1 and ATG7 resulted up-regulated, being targeted by miR-18a and miR-4802, respectively, thus preventing CRC cells enriched in *F. nucleatum* from the apoptosis induced by chemotherapy. Also, *E. coli* although being a gut commensal bacterium, has been observed at higher levels of colonization in colonic mucosa of human CRC xenograft mice vs non infected controls [202]. In addition, also for some strains of this bacteria it has been reported possible roles in CRC in a miRNA-dependent manner. Indeed, the majority of *E. coli* isolated from CRC cases produced colibactin, a genotoxic compound inducing c-Myc expression. High c-Myc expression led to miR-20a-5p up-regulation thus targeting *SENP1*, a key negative regulator of p53 small ubiquitin-like modification (SUMOylation). As a consequence, p53 SUMOylation, drove a senescence-associated secretory phenotype (SASP) and the release of carcinogenic growth factors promoting colon tumor growth [203].

1.4.9 miRNAs may shape Gut Microbiota

The studies previously discussed show an effective impact of the gut microbiome on miRNAs and their target gene expression. Conversely, host-derived miRNAs can also influence gut microbiota (**Figure 9)** [204]. One of the first evidence on this mechanism came from Liu et al. demonstrating how host-fecal miRNAs can shape gut microbiota composition by modifying the relative bacteria abundance. Indeed, mice with specific miRNAs devoid in intestinal epithelial cell (IEC) showed an increased diversity of bacterial genera compared to wild-type mice. In addition, using knock-out mouse model, they observed a marked exacerbation of dextran sulfate sodium (DSS)-induced colitis in IEC-specific miRNA-deficient mice compared to wild type that was ameliorated by transplanting wild type fecal miRNAs [7].

As another example, it has been reported that CRC cells harbouring mutant p53 selectively shed exosomes enriched with miR-1246. Uptake of exosomes enriched in this miRNA triggers macrophages reprogramming into an anti-inflammatory state supporting tumor survival [205].

More recently, also the role of diet on miRNA and microbiome interactions has gained a growing interest. Indeed, dietary habits have been shown to influence host-derived miRNAs expression but also many food-derived exogenous miRNAs can have an active role in gut homeostasis regulation.

For example, high-fat diets trigger the release of hepatic bile acids which is metabolized by the gut microbiota into genotoxic secondary bile acids, such as deoxycholic acid (DCA) [206]. DCA was shown to promote CRC through down-regulation of miR-199a-5p expression. The up-regulation of this miRNA in CRC cells causes the suppression of tumor cell growth and the restoration of the tumor cell drug sensitivity, likely through inhibiting the expression of *CAC1*, a direct miR-199a-5p target and a driver of the cell cycle usually found to be highly expressed in CRC [207]. For example, it has been established that miRNAs derived from edible plants can affect microbiome composition and host-gut barrier function [208]. Specifically, miR-7267-3p contained in ginger exosome-like nanoparticles (GELN) was shown to modulate the expression of Lactobacillus rhamnosus (LGG) monooxygenase ycnE. This event leads to increased production of indole-3-carboxaldehyde (I3A) which is an aryl hydrocarbon receptor (AhR) ligand. Thus, GELN-derived miRNAs enhanced the IL-22 production ameliorating mouse colitis through improving barrier function [208].

All the discussed evidence suggest miRNAs as an important link in host–microbiota-diet interactions in regulating gut health and GI disease. However, the causal relationships remain to be established, and whether the mechanisms identified in the previous investigation could have a significant impact on human GI diseases development require further examination.

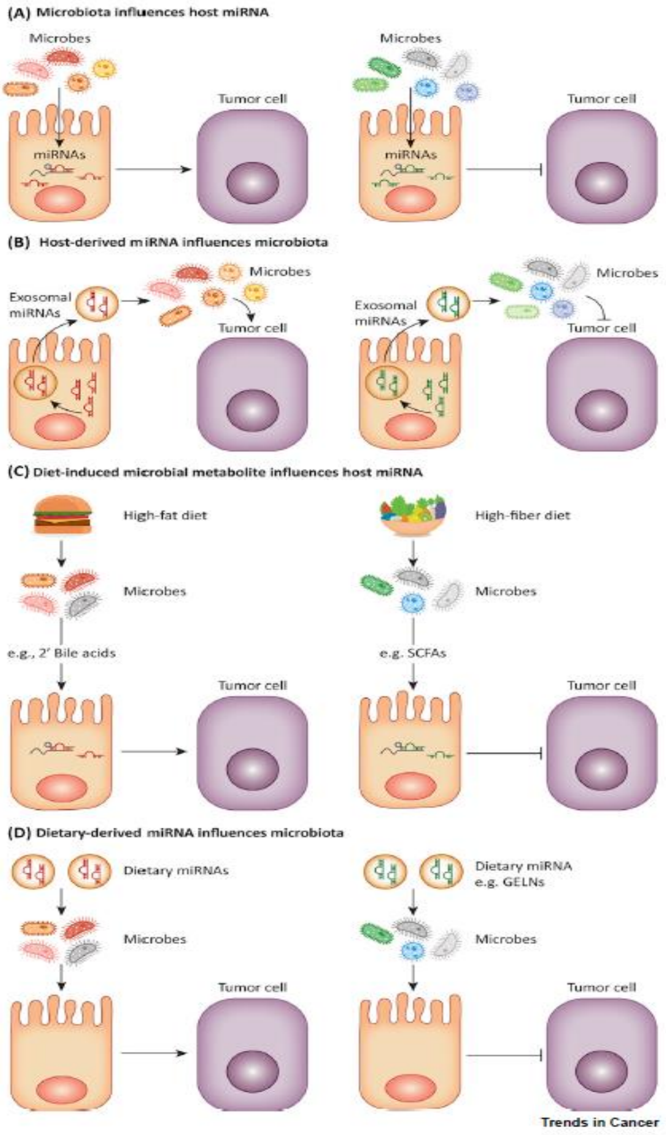


Figure 9 Reciprocal regulation/interaction of host miRNAs and microbiota [204]. **A.** Microbioma influencing host miRNA expression. **B.** Release of miRNAs from host cells via extracellular vesicles to control intestinal homeostasis and gut microbiota. **C.** Diet-induced microbial metabolites, such as secondary bile acids, can affect miRNA expression in host cells and promote or inhibit tumor. **D.** Dietary miRNAs, such as miR-7267-3p from ginger exosome-like nanoparticles (GELNs) can regulate gut microbiota and intestinal homeostasis.

1.4.10 Fecal miRNAs and gut microbiome interactions as a source of potential biomarkers of CD

To the best of our knowledge, there are no studies available both on miRNAs deregulated by microbiota and, vice versa, about miRNAs targeting bacteria in CD patients. Thus, the miRNA-mediated crosstalk between the host and the microbiota is a totally unexplored research field in the context of CD. To date, the only available investigation to fill this gap is the study of Mohan et al. that combined the microbiota dysbiosis with the expression levels of selected miRNAs [209].

Firstly, authors showed that in a model of gluten-sensitive (GS) macaques under gluten-containing diet (GD) the diversity of gut microbiota was significantly reduced compared to that of the healthy age-matched peers. A phenotype comparable with that of healthy controls was restored by the GFD indicating a direct relationship between microbiota composition and tissue damages mediated by inflammatory response to gluten [209]. miRNA expression analyses showed a group of up-regulated miRNAs, such as miR-203, miR-204, miR-23b, and miR-29b, in macaques under GD. Interestingly, the analysis of putative miRNA targets highlighted their complementarity on 16S ribosomal RNA of bacterial species such as *Lactobacillus reuteri*, *Prevotella stercorea* and *Streptococcus luteciae* that were found more abundant in fecal samples of GS macaques under GD. This finding is the first one highlighting the role of miRNAs potentially targeting bacteria and then contributing to dysbiosis in CD. Further investigation on this field could help in clarifying the dynamics of miRNA-microbiome relationship and provide new potential biomarkers for the diagnosis and monitoring of CD.

In this context, fecal miRNAs represent ideal candidates to study intestinal diseases and gut microbiota shaping. However, as already mentioned, very few data are available so far on fecal miRNAs of CD patients, revealing a hole in the CD scientific literature. Fecal samples are highly informative for intestinal-related diseases and extremely easy to collect without any invasive procedures. Together with those previously discussed, these characteristics make fecal miRNAs and gut microbiome ideal biomarkers for several gastrointestinal disorders and, therefore, should be further investigated.

2 AIMS OF THE STUDY

The present PhD Thesis reflects the growing interest in the analysis of sncRNA spectra as potential source of non-invasive biomarkers for GI disorders. In this respect, the main objectives of the study were: i) to evaluate miRNA and other sncRNA profiles in stool samples to identify potential biomarkers for CD and CRC; ii) to perform a comparison among the results obtained from CD and CRC to highlight similar and/or different features in terms of sncRNA expression; and iii) to explore gut microbiota and its relation with miRNA expression levels in stool.

Specifically, the work has been focused in:

> **Study 1**: We aimed at assessing whether the different dietary regimes and the CD condition may induce differences in sncRNA expression and gut microbiome composition. To achieve this, sncRNA expression profiles in stool specimens were obtained by small RNA-seq and gut microbiome composition by shotgun metagenomic sequencing in relation to CD. sncRNA expression levels and gut microbiome composition were assessed analysing different types of CD patients (individuals with high or normal levels of TG2-Abs) vs healthy controls.
>
> **Study 2**: We aimed at identifying sncRNA and microbial signatures in surrogate tissue able to accurately discriminate CRC cases at diagnosis from healthy controls, "precancerous lesions" or gut inflammatory disease (such as adenomas and IBDs), the latters being at higher risk of developing CRC. In parallel, we aimed at establishing if there is a reflection of the found fecal profiles on primary tumor tissue. This was achieved by investigating sncRNA expression profiles and shotgun metagenomics in stool in relation to CRC, precancerous lesions (polyps) and IBD in two independent cohorts. In one cohort, sncRNA profiles were also investigated in tumor/polyp and adjacent normal mucosa by the same approach (small RNA-seq).

3 MATERIALS AND METHODS

3.1. Study populations

3.1.1 Study 1

A total of 130 subjects were enrolled in this study by the Molecular Epidemiology and Exposomics Unit at the Italian Institute for Genomic Medicine (IIGM) of Turin (n=88), the Gastroenterology Unit of Ospedale Mauriziano Umberto I (n=11) and the Gastroenterology outpatient clinic of San Giovanni Antica Sede (SGAS, n=31). The cohort included 60 treated CD, 3 untreated CD, 2 non-celiac gluten sensitivity (NCGS) patients and 65 healthy volunteers. For CD patients inclusion criteria were: Caucasian ethnicity, ages> 15 and <80 years, negative medical history of concomitant or previous gastrointestinal diseases (including tumors) and CD diagnosis histologically approved. Exclusion criteria included: self-diagnosis, use of antibiotics and other drugs at sampling, age <15 and other concomitant diseases. CD treated group of patients was further stratified in subjects with normal value of TG2-Ab serum levels (levels of TG2-Ab <3.0 UA are considered normal) (CD-ltTG, n=51) and those with a value above the 3.0 UA threshold (CD-htTG, n=11). This latter category also included a follow-up collection of 2 CD untreated individuals recruited after three months from the GDF diet adherence. Healthy volunteers were enrolled among the general population and were age and sex-matched to recruited CD. They followed an omnivore diet without any dietary restrictions, food intolerance or allergies.

All participants received a kit containing detailed information about the study, two questionnaires from the European Prospective Investigation into Cancer and Nutrition (EPIC) study (one about the dietary habits and another sex-specific one regarding lifestyle habits) [210], an additional short questionnaire about changes in their dietary habits, medical history, additional questions not present in the EPIC questionnaires, and finally a disposable 30mL polystyrene screw cap container with spoon for stool collection. All volunteers signed a written informed consent form to participate in the study. The study was conducted according to the guidelines in the Declaration of Helsinki. The protocol of the study was approved by the Azienda Ospedaliera-Universitaria, Città della Salute e della Scienza di Torino (Protocol n 0030717, 23th March, 2018).

3.1.2 Study 2

Biological specimens, clinical and demographic data were collected from patients recruited in a hospital-based study at the Clinica S. Rita in Vercelli, Italy and two hospitals in Prague and one in Plzen, Czech Republic. On the basis of colonoscopy results, participants were classified into four categories: (i) healthy subjects (individuals with colonoscopy results negative for tumor, polyps, and other GI; (ii) subjects with gut inflammatory pathologies such as IBD and diverticular disease; (iii) polyps patients (individuals with any type of colorectal polyps); and (iv) colorectal cancer patients (individuals with newly diagnosed CRC).

Stool and blood were collected for a total of 221 subjects (80 Healthy, 41 Inflammation, 43 Polyps, 57 CRC) in the Cohort-IT. When available, for CRC and polyps cases tumor/malignant tissue and adjacent normal mucosa were collected and stored in RNA later.

Stool and plasma samples were also collected from an independent cohort of 162 subjects (36 Healthy, 32 Inflammation, 28 Polyps, 66 CRC) recruited in the Czech Republic (Cohort-CZ).

All CRC patients were recruited at the first CRC diagnosis and had not received any treatment before the fecal sample collection.

The study was approved by the local ethics committee (Ethics Committee of Azienda Ospedaliera SS.Antonio e Biagio e C. Arrigo of Alessandria, Italy; protocol N. Colorectal miRNA CEC2014 and Institute of experimental medicine CE, Prague, Czech Republic), and informed consent was obtained from all participants.

3.2. Sample collection

3.2.1 Study 1

Naturally evacuated stool samples were collected at home in stool nucleic acid collection and transport tubes with RNA stabilizing solution (Norgen Biotek Corp.). The participants brought them to the IIGM laboratory or at the gastroenterology units of the two relative Hospitals. The fecal samples were aliquoted in 200 µl into Eppendorf LoBind tubes prepared for the next step and stored at -80°C.

Plasma and serum samples were collected according to standard phlebotomy procedures at the moment when volunteers brought the stool samples to the laboratory or, in case of those patients hospitally recruited, the day of their gastroenterological visit. Five ml of blood was collected into both Ethylenediaminetetraacetic acid (EDTA) tubes for plasma isolation and in BD Vacutainer SST II Advance tubes for the serum, and immediately placed on ice. Tubes were centrifuged at 1000 and 4000 rpm for 10 minutes at room temperature for plasma and serum isolation, respectively, and then aliquoted in 250µl in Eppendorf LoBind tubes and stored at -80ºC. The time from sample procurement to storage at -80°C was less than 3 hours.

3.2.2 Study 2

Stool samples were collected at home in the same tubes described in Study1, before any bowel preparation for colonoscopy and returned at the time of performing a colonoscopy in the endoscopy unit or at the time of blood sampling. Aliquots (200 µl) of the stool samples were stored at –80°C until RNA/DNA extraction.

Plasma samples were collected and processed as in Study 1.

For CRC and Polyp patients fresh colorectal tissue samples were prospectively collected at the Clinica S. Rita in Vercelli, Italy CRC. Paired primary tumor/polyp tissues and adjacent normal mucosa were obtained from the CRC patients of the Italian cohorts undergoing surgery. All tissues samples were immediately preserved in RNA later and stored at –80°C until use.

Sampling for all biospecimens was performed in the same time period (years 2016-2021) for both cohorts.

3.3 Nucleic acid isolations

3.3.1 Extraction of total RNA and DNA from stool

In both studies, a 200 µl fecal aliquot was used for RNA extraction with stool total RNA purification kit (Norgen Biotek Corp.) using the protocol recommended by the manufacturer. The RNA quality and quantity were verified according to the MIQE guidelines (http://miqe.gene-quantification.info/). The RNA concentration was quantified by Qubit with a Qubit microRNA assay kit (Invitrogen). In study 1, DNA extraction was performed using the DNeasy PowerSoil Pro Kit (Qiagen Cat No./ID: 47014, Cat No./ID: 47016). A smaller volume of the final elution buffer to increase DNA concentration was adopted as a unique modification of the kit protocol.. In Study 2, DNA extraction was performed using the Qiamp DNA stool kit (Qiagen) following the manufacturer's instructions. The DNA quantification was performed in both studies with a Qubit DNA high-sensitivity (HS) assay kit (Invitrogen).

3.3.2 Extraction of total RNA from plasma

For Study 1, 200 µl of plasma volume was used for RNA extraction. Total RNA was extracted with the Qiagen miRNeasy Serum/Plasma Kit using the protocol recommended by the manufacturer. The RNA concentration was quantified by Qubit using the Qubit microRNA assay kit (Invitrogen). For Study 2, plasma exosomes/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. Briefly, the plasma was mixed with 50.4µl of ExoQuick solution and refrigerated at 4°C overnight (at least 12 h). The mixture was then further centrifuged at 1500 g for 30 min. The EVs pellet was dissolved in 200 µl of nuclease-free water and RNA was extracted immediately from the solution with the same kit of Study 1 but using the QiaCube extractor (Qiagen) as described in [211] and RNA concentration quantified as in Study 1.

3.3.3 Extraction of total RNA from tissues

In Study 2, total RNA from tissues samples was extracted using Trizol reagent (Thermofisher) according to the manufacturer's instructions.

The RNA quality and quantity were verified according to the MIQE guidelines (http://miqe.gene-quantification.info). The RNA concentration was quantified by Qubit with a Qubit microRNA assay kit (Invitrogen).

3.4 Library preparation for small RNA sequencing

Small RNA transcripts were converted into barcoded cDNA libraries. Library preparation was performed with a NEBNext multiplex small RNA library prep set for Illumina (protocol E7330; New England BioLabs Inc., USA). For each sample, 250 ng of RNA were used as the starting material to prepare libraries. Each library was prepared with a unique indexed primer so that the libraries could all be pooled into one sequencing lane. Multiplex adapter ligations, reverse transcription primer hybridization, reverse transcription reactions, and the PCR amplification were performed as described in the protocol provided by the manufacturer. After PCR preamplification, the cDNA constructs were purified with a QIAQuick PCR purification kit (Qiagen, Germany) following the modifications suggested in the NEBNext multiplex small RNA library prep protocol. Further quality control checks and size selections were performed following the NEBNext multiplex small RNA library prep protocol (protocol E7330; New England BioLabs Inc., USA). Size selection of the amplified cDNA constructs was performed using Novex Tris-borate-EDTA (TBE) gels (Invitrogen) (6%) and following the procedure of gel electrophoresis running and purification of the construct described in the Illumina TruSeq small RNA library prep protocol. The 140-nt and 150-nt bands correspond to adapter-ligated constructs derived from RNA fragments of 21 to 30 nt. A concluding Bioanalyzer 2100 run performed with a high-sensitivity DNA kit (Agilent Technologies, Germany) permitted checking final size, purity, and concentration for the sequences in the DNA libraries. The obtained libraries (24 samples were multiplexed) were subjected to the Illumina sequencing pipeline, passing through clonal cluster generation on a single-read flow cell

(Illumina Inc., USA) by bridge amplification on a cBot (TruSeq SR cluster kit, v3-cBOT-HS; Illumina Inc., USA) and 50 cycles of sequencing by synthesis using a HiSeq 2000 (Illumina Inc., USA) at the Gene Core Facility of the European Molecular Biology Laboratory (EMBL), Heidelberg (Germany).

3.5 Analysis of sncRNAs from small RNA-seq data

Small RNA-Seq pipeline analyses were performed using a previously published Docker-embedded software to guarantee the computational reproducibility of the analysis [212]. Briefly, trimmed reads were mapped against an in-house reference of human small RNA sequences. The alignment was performed using BWA algorithm v0.7.12 [213]. Human miRNAs were annotated and quantified using two methods called the “knowledge-based” and “position-based” methods as described in [212]. miRNAs whose assigned arms were derived from the “position-based” methodology were indicated in italics. The sequences of the mature miRNAs were aligned between each other, and in case of mature miRNAs characterized by identical sequences, the associated read counts were summed. A detected miRNA was considered as expressed if supported by at least 15 (Study 1) and 20 (Study 2) normalized reads.

3.6 miRNA targets functional enrichment analysis

Functional enrichment analysis was performed using RBiomirGS v0.2.12 [214] considering validated targets of DEmiRNAs among the categories. The log2FC and adjusted p-value computed between the categories were used as inputs for the tool. The gene sets characterized by an adjusted p-value lower than 0.001 were considered as significantly enriched.

3.7 Model-based learning for disease classes classification

The predictive model to distinguish among all disease classes involved in the study 2 was defined through a supervised strategy based on three steps. Step1 consisted of the selection of differentially expressed sncRNAs between two disease classes, independently in Cohort-IT and Cohort-CZ. In Step2, for each comparison considered (i.e., CRC versus healthy, and adenoma versus healthy) the disease feature model was defined based on differentially expressed sncRNAs associated with the highest AUC value. In Step3, merging the disease feature models previously selected, the sncRNAs predictive model able to distinguish among all disease classes was identified.

The classification capability of the predictive model of sncRNAs was assessed through a stratified ten-fold cross-validation to address the imbalance in the number of patients belonging to the sample groups. In the cross-validation, a training set from the pooled samples from both Cohort-IT and Cohort-CZ was defined using the remaining samples for testing the model. The evaluation of the model was performed using three classification algorithms (i.e., random forest, logistic regression, and gradient boosting) and repeated 100 times. Each final accuracy value was therefore derived by an average value of over 1,000 validation folds.

3.8 Library preparation for shotgun metagenomic sequencing and bioinformatics analyses

The same experimental procedure was adopted for both studies. Library preparation of Study 2 was performed in collaboration with the Computational Metagenomics Laboratory (UNITN). Nextera DNA Flex Library Prep kit was adopted for the Illumina NovaSeq metagenomic sequencing of DNA from stool samples. The experimental procedure follows the Illumina reference guide. Two adjustments were performed:

- at the “clean up library step” stage, 0.6x AMPure XP beads were used
- re-suspension of the library pool occurred with ¼ of the initial pool volume

These modifications allow the achievement of a higher quality of resulting reads and a lower content of adapters. Sequencing was performed on the NovaSeq 6000 Sequencing System (average of 6.5GB/sample).

Pre-processing steps of metagenomics data analysis, including read quality control, trimming and adapter removal, and removal of reads aligned on phiX or human genome were performed in collaboration with the computer Science Department of the University of Torino, using the procedure implemented in collaboration with CIBIO (Trento, as provided in https://github.com/SegataLab/preprocessing). Taxonomic profiling was carried out by using

MetaPhlAn3 in default settings and using mpa_v30_CHOCOPhlAn_201901 as microbial markers database. Alpha and beta diversity metrics were computed using the *vegan* R package Functional profiles were obtained by HUMAnN 3.0 [215]. Difference in microbial relative abundances among the sample groups was evaluated using SIAMCAT v1.11.1 package in default settings [216]. Species with abundance lower than 0.001 were filtered using the *filter.features* function while association testing was performed using the *check.associations* function.

3.9. Data integration and feature selection

The R package *mixOmics* was used for the integration of the three datasets (taxonomic profiles, miRNA expression profiles, and dietary information), using the Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO) [217]. Datasets were integrated after normalization (*scale* R function) and removing of near-zero variables (*nearZeroVar* function of R package *caret*). The function *block.splsda* was used to compute the DIABLO model while the function *plotLoadings* was used to extract the 25 microbial species, miRNAs, and nutrients associated with the highest loading on the first and the second variate of the model. These attributes were further used as input for a classification and a feature selection analysis using Weka v.3.8.5 [218]. The classification analysis was performed using the Random Forest classifier with default settings and a 10-Fold cross-validation control to evaluate the predictive model. Feature selection analysis was performed using seven methods (*ClassifierAttributeEval*, *CorrelationAttributeEval*, *GainRatioAttributeEval, InfoGainAttributeEval, OneRAttributeEval, ReliefFAttributeEval*, *SymmetricalUncertAttributeEval*) providing a ranking statistic to score the contribution of each attribute to the classification.

3.10 Statistical Analysis

Differential expression analysis was performed with DESeq2 R package v1.22.2 using the likelihood ratio test (LRT) function. This function was selected in order to correct the analysis including age, gender, and BMI covariates [219].

Correlation analysis calculated using the Spearman's rank correlation coefficient (SCC) using age, gender, and BMI covariates to correct the analysis.

Multiple testing correction was performed using the BH method and correlations associated with an adjusted p-value <0.05 were considered as statistically significant. For microbiome composition analysis the differential abundances both at phylum and species levels were considered significant if associated with a Wilcoxon adjusted p-value lower than 0.05.

4 RESULTS

4.1 Study 1

4.1.1 Population Characteristics

A total of 130 participants were recruited for the present study categorised as: i) untreated CD, recruited at diagnosis before starting the GFD (n=3), ii) CD treated, already on a GFD (n=60), iii) NCGS individuals (n=2), and iv) healthy sex-/age-matched controls, not following any specific diet (n=65). CD treated group were further grouped according to serological levels of tTG2-Ab in those under (CD-ltTG, n=51) and over (CD-htTG, n=11) the <3.0 UA threshold (see Material & Methods section). The mean age of the healthy subjects was 40.8±14.3 years. Treated CD subjects had a similar mean age of 42±14.5 years while for CD untreated and NCGS categories was 46.5±13.9 and 32±0.3 years, respectively (**Table 5)**. CD treated group included 75% of females and 25% males while 77% of females and 23% of males characterized the healthy group. No significant differences were observed considering other variables, except for the Vitamin D3 levels (p<0.0001.) consistently higher in treated CD individuals (average of 92.6±134.1 ng/ul) compared to healthy controls (average of 32.5±2.8). This is probably due to the use of vitamin supplements usage in the former group.

Table 5: Demographic and clinical characteristics of the CD cohort analysed in the study

Covariates		CD Untreated	CD Treated	CD ltTG	CD htTG	NCGS	Healthy controls	P-value
		(n=3)	(n=60)	(n=51)	(n=11)	(n=2)	(n=65)	(CD treated vs Healthy)
Age	Average±SD	46.5±13.9	42±14.5	43±15.1	44.4±14.7	32.0±0.3	40.8±14.3	0.775
(years)	Range	31-58	19-75.6	19-74	24-76		20.6-77.5	
Sex	Female	3	45	38	9	2	50	0.801
	Male		15	13	2		15	
BMI (Kg/m^2)	Average±SD	NA	22.3±3.4	21.9±3.9	22.3±3.1	18.4±1.9	22.5±3.1	0.206
Length of Gluten free diet	Average±SD	_	8.9±7.9	10.2±7.9	9.0±6.2	1.2	_	
(years)	Range		0.04-29.8	0.2-29.9	0.1-16.4	(partially)		
Smoking status	current	1	8	7	1	1	8	0.989
	former	1	14	12	3	-	15	
	never	1	37	31	6	1	40	
	NA	-	1	1	1	-	-	
Marsh grade	l	-	5	4	1			
(at CD diagnosis)	ll	-	2	2	0			
	llla	1	14	13	2			
	lllb	2	11	9	3			
	lllc	-	16	13	3			
	NA	-	12	10	2			
*s-Ab anti transglutaminase lgA	≤3.0	0	37	37	0	2	50	
(AU/ml)	>3.0	3	9	-	11	0	0	
	NA		14	14	-		15	
**s-FERRITIN	Average±SD	6±2	54.4±54.6	60.0±59.9	22.1±29	34.0±21.2	73.0±80.6	0.46
(ng/ml)	Range	4-8	8-246	11-46	4-105	19-49	3-429	
	<11	3	2	-	3	0	4	
	≥11	0	44	37	7	2	46	
	NA		14	14	1			
***s-VITAMIN B12	Average±SD	369	286.9±122.2	295.6±123.4	271.5±101.4	177.5±87.0	270.1	0.395
(pg/ml)	Range	369	111-708	111-708	135-440	116-239	101-1152	
	<180	0	7	4	3	1	11	
	≥180	1	39	33	7	1	39	
	NA	2	14	14	1			
****s-VITAMIN D3	Average±SD	17.8	92.6±134.1	91±133.3	82.7±127.2	32.5±2.8	22.6±12.9	1.14E-04
(ng/ml)	Range	17.8	10-428	10.6-428	10.5-394	30.6-34.5	3.8-92.7	
	<30	1	24	19	4	0	44	
	≥30	0	22		6	2	6	
	NA	2	14		1			

* Ab anti transglutaminase lgA serum test is negative when ≤3 AU/ml, dubious when between 3 AU/ml - 12.0 and positive if >12 AU/ml.
** FERRITIN serum level is low when <11 ng/ml, normal when ranging from 11 to 307 ng/ml and high if >307 ng/ml.
***VITAMINB12 serum level is low when <180 ng/ml, normal when ranging from 180 to 914 ng/ml and high if >914 ng/ml.
****VITAMIN D3 serum level is low when <30 ng/ml, normal when ranging from 30 to 100 ng/ml and high if >100 ng/ml.

4.1.2 miRNA profiles from small-RNA sequencing data

Small-RNA sequencing of fecal samples was performed for all the 130 individuals recruited in this study. On average, 12.2 million single-end reads per sample were obtained from sequencing output, with a median of 101,671 reads (1.07%) assigned to human miRNA annotations (**Figure 10 A**). The distribution of miRNAs detected in all groups is reported in **Figure 10 B**. In total, 2,830 miRNAs were detected in at least one sample. Considering miRNAs with a median of reads > 15 at least in one of the investigate groups, an average of 757 detected miRNAs was observed. miR-3125 was characterized by the highest median expression levels in all groups.

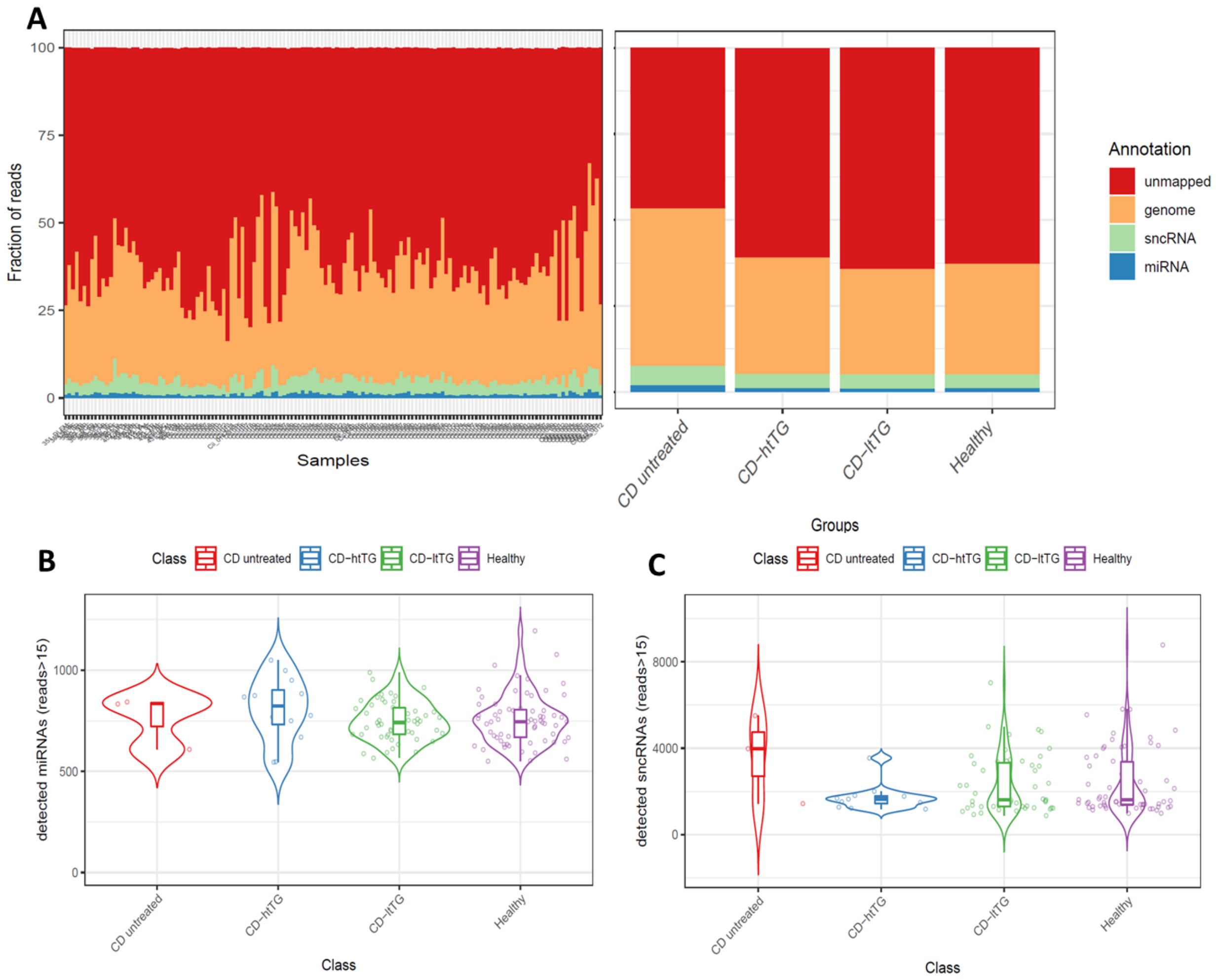


Figure 10. Read assignments for stool samples of each subject (left) and for each of the investigated categories (right) (**A)**; Stool miRNAs (**B)** and other sncRNAs (**C**) detected in the four groups.

Overall, one hundred DEmiRNAs were identified by the differential expression analysis by comparing the CD categories with the healthy controls (**Figure 11 A**). Forty-nine DEmiRNAs were detected comparing CD-ltTG with healthy individuals, among which 22 were up-regulated and 27 down-regulated. In the CD-htTG vs controls comparison, 54 miRNAs showed an altered expression, with 16 of them down-regulated and 38 up-regulated in the CD group. Interestingly, six DEmiRNAs overlapped between the comparisons of CD-ltTG and CD-htTG vs healthy controls (**Figure 11 B**). All of them shared the same trend of expression in both comparisons, with slightly higher FCs in the CD-htTG group. A comparison between these two groups highlighted 20 DEmiRNAs, 18 up-regulated and two down-regulated (namely, *miR-3137-3p* and miR-4259). Interestingly, for some of the DEmiRNAs found among the comparisons we noticed a trend of expression going from CD untreated to CD-htTG, CD-ltTG and healthy individuals**.** An example for three dysregulated miRNAs is reported in **Figure 11 C**. Due to the limited number of individuals (n=3), CD untreated group was not individually considered for DE analysis.

To study the relationship between fecal DEmiRNA expression profiles and the duration of the GFD a correlation analysis was run. The average GFD duration in the whole group of CD patients was 8.9±7.9 years, ranging from a couple of weeks to 29.8 years.

By analysing the whole group of CD treated patients (i.e., CD-htTG and CD-ltTG), a significant correlation with GFD duration was observed for six DEmiRNAs. *miR-3652-3p* and miR-6505-3p resulted negatively correlated (SCC = -0.29, p=0.02 and SCC = -0.25, p=0.04 respectively) while *miR-4771-5p*, miR-4684-3p, *miR-662-5p* and miR-320d resulted positively correlated (SCC ranging from 0.24 to 0.26, p<0.05). The same analysis focusing only on the CD-ltTG group highlighted miR-6505-3p, *miR-4771-5p*, *miR-3652-3p*, miR-3619-3p and miR-4684-3p significantly correlated with GFD duration (p<0.05). The first three miRNAs, already found in the previous analysis based on the whole CD group, showed the same correlation trend with stronger correlation for miR-6505-3p and *miR-4771-5p* (SCC= -0.34, p=0.01 and SCC= 0.33, p=0.02 respectively). None of the DEmiRNAs correlated to the duration of the GFD was also correlated to subjects' age

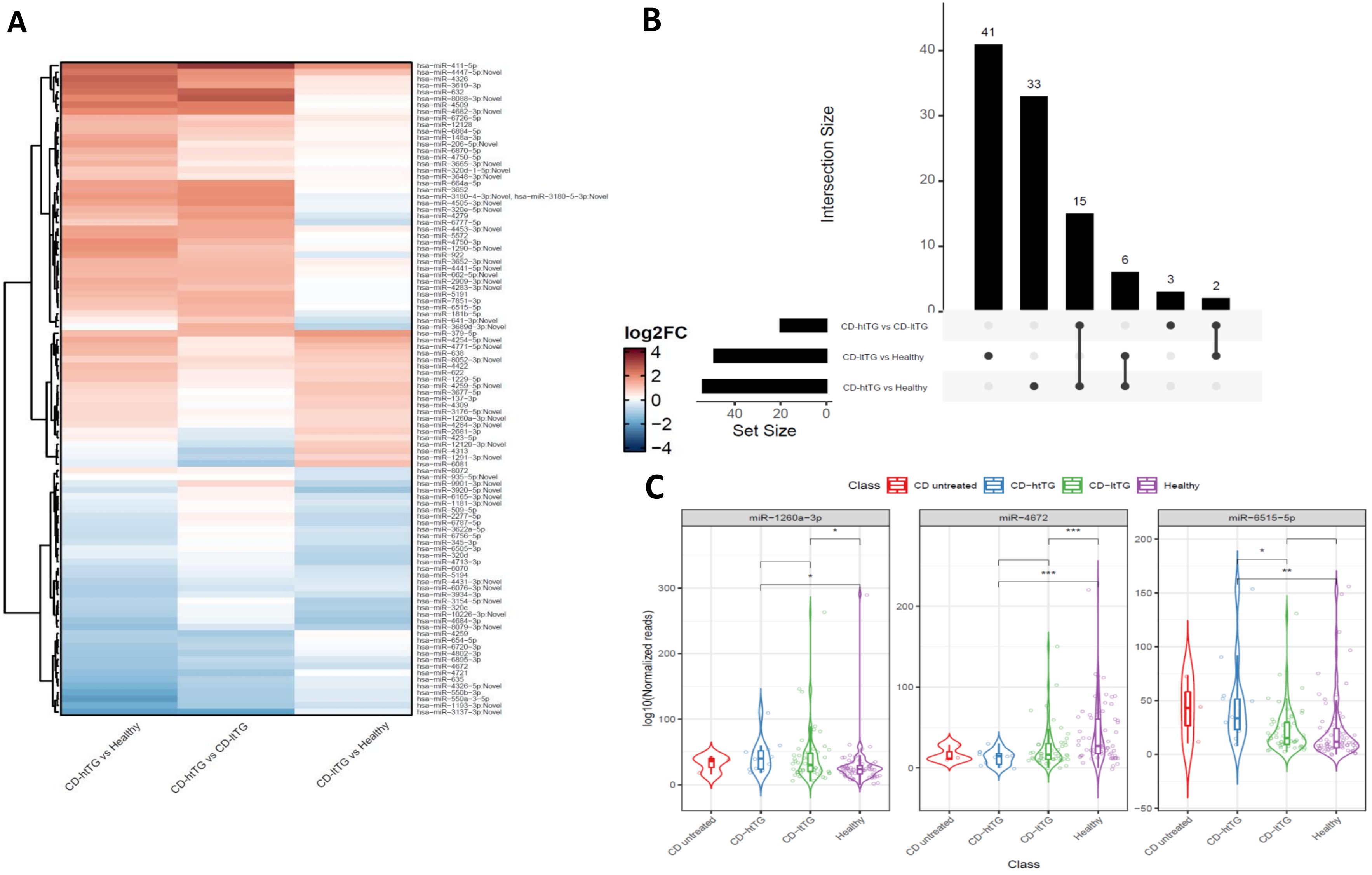


Figure 11. Results of miRNA differential expression analysis among the investigated categories. **A.** Heatmap representing FCs of the 100 significant DEmiRNAs among the previously mentioned comparisons. **B.** Upset plot representing specific and common DEmiRNAs among the investigated categories. **C.** Violin plots reporting expression levels of selected DEmiRNAs showing a trend of expression (down-regulation for miR-1260a-3p and miR-6515-5p and up-regulation for miR-4672) going from CD untreated to CD-htTG, CD-ltTG and the healthy controls categories**.**

4.1.3 DEmiRNAs enrichment analysis

A functional analysis was performed for the 3 groups of identified DEmiRNAs (i.e., CD-htTG or CD-ltTG specific miRNAs and the 6 DEmiRNAs overlapping among the two comparisons). A total of 104 significant enriched terms were observed (p<0.001): the most relevant are reported in **Table 6**. DEmiRNAs characterizing CD-htTG resulted mainly associated with the inflammatory (*Nlrp3 Inflammasome*, *Inflammasomes)* and metabolic pathways (*Intestinal Lipid Absorption, Cellular Carbohydrate Biosynthetic Process,* etc.). Similarly, enrichment analysis on CD-ltTG specific DEmiRNAs highlighted a role in the immunity response (*Positive Regulation Of Macrophage Migration*, and *T Cell Migration*) and metabolic process (*Cellular Modified Amino Acid Catabolic Process*) related terms. The enrichment analysis performed considering the 6 DEmiRNAs common among both CD groups showed apoptosis as the main process in which the DEmiRNAs target genes are involved (*Execution Phase Of Apoptosis, Regulation Of Execution Phase Of Apoptosis).*

Table 6: Enriched terms for the CD-htTG and CD-ltTG specific DEmiRNAs and those of the 6 shared DEmiRNAs. Three libraries were considered for the functional analysis, KEGG, Reactome and GO-Biological Process. Only terms enriched with a p-adj<0.05 were reported.

Enriched terms	Class	Targets
REACTOME_THE_NLRP3_INFLAMMASOME	**CD-htTG**	*HSP90AB1, PYCARD, SUGT1, RELA, TXNIP*
REACTOME_INFLAMMASOMES	**CD-htTG**	*HSP90AB1, PYCARD, SUGT1, BCL2L1, BCL2, RELA, TXNIP*
GO_NEGATIVE_REGULATION_OF_ANION_TRANSMEMBRANE_TRANSPORT	**CD-htTG**	*MTOR, THBS1*
GO_INTESTINAL_LIPID_ABSORPTION	**CD-htTG**	*LDLR, ABCG8*
GO_CELLULAR_CARBOHYDRATE_BIOSYNTHETIC_PROCESS	**CD-htTG**	*B4GALT1, PPP1CB*
GO_CELLULAR_MODIFIED_AMINO_ACID_CATABOLIC_PROCESS	**CD-ltTG**	*ABHD12, ALDH4A1*
GO_REGULATION_OF_MACROPHAGE_CHEMOTAXIS	**CD-ltTG**	*MDK, THBS1*
GO_REGULATION_OF_GRANULOCYTE_CHEMOTAXIS	**CD-ltTG**	*MDK, THBS1*
GO_POSITIVE_REGULATION_OF_MACROPHAGE_CHEMOTAXIS	**CD-ltTG**	*MDK, THBS1*
GO_REGULATION_OF_MACROPHAGE_MIGRATION	**CD-ltTG**	*MDK, THBS1*
GO_POSITIVE_REGULATION_OF_MACROPHAGE_MIGRATION	**CD-ltTG**	*MDK, THBS1*
GO_POSITIVE_REGULATION_OF_INSULIN_SECRETION	**CD-ltTG**	*GLUL, HIF1A, TCF7L2*
GO_T_CELL_MIGRATION	**CD-ltTG**	*CRK, RHOA, MSN*
GO_EXECUTION_PHASE_OF_APOPTOSIS	**Both**	*MTRNR2L7, MTRNR2L3, MTRNR2L10, MTRNR2L11, TAOK1*
GO_REGULATION_OF_EXECUTION_PHASE_OF_APOPTOSIS	**Both**	*MTRNR2L7, MTRNR2L3, MTRNR2L10, MTRNR2L11, TP53*
GO_NEGATIVE_REGULATION_OF_EXECUTION_PHASE_OF_APOPTOSIS	**Both**	*MTRNR2L7, MTRNR2L3, MTRNR2L10, MTRNR2L11*
GO_NEGATIVE_REGULATION_OF_SIGNALING_RECEPTOR_ACTIVITY	**Both**	*MTRNR2L7, MTRNR2L3, MTRNR2L10, MTRNR2L11*

4.1.4 Other sncRNA profiles from small RNA sequencing data

A median of 417,889 reads (4%) was assigned to other sncRNAs than miRNA annotations (**Figure 10 A**). The distribution of sncRNAs detected in all groups is reported in **Figure 10 C**. In total 25,849 sncRNAs were detected in each sample with an average of 2345 sncRNAs expressed in each sample. The differential analysis of other sncRNAs than miRNAs provided a total of 56 differentially expressed sncRNAs (DEsncRNAs) observed in at least one of the performed comparisons among the groups of interest of CD and healthy controls (**Figure 12 A**). Among the DEsncRNAs, 35 were piRNAs, 15 tRNAs, 3 snRNAs, 1 other snRNAs, 1 tRNA-derived fragments (tRFs) and 1 from the miscellaneous RNAs (miscRNA). No significant results were obtained comparing CD-ltTG vs healthy controls, while 42 DEsncRNAs (35 up and 7 down-regulated) were observed comparing CD-htTG vs healthy controls, **Figure 12 B**. The comparison between CD-htTG and CD-ltTG groups highlighted 22 DEsncRNAs, 14 up and 8 down-regulated in CD-htTG. Eight DEsncRNAs overlapped with consistent trends between CD-htTG vs CD-ltTG and CD-htTG vs healthy controls.

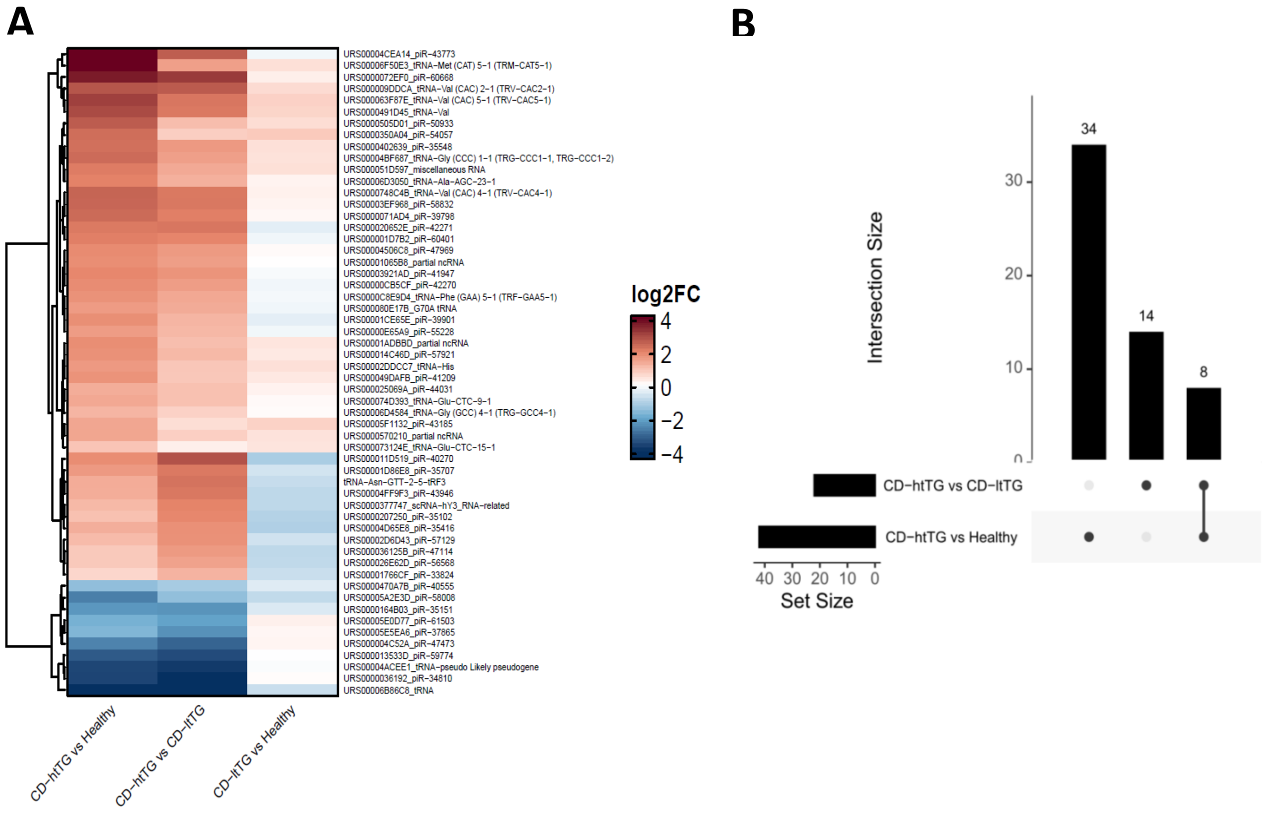


Figure 12. Results of sncRNA differential expression analysis among the investigated categories. **A.** Heatmap representing FCs of the 56 significant DEsncRNAs among the comparison previously mentioned. **B.** Upset plot representing specific and common DEmiRNAs among the investigated categories.

4.1.5 Gut microbiome composition

Metagenomic sequencing of DNA from fecal samples was performed for all the 130 individuals recruited in this study. On average, 44.382.527 raw reads were assigned to each sample of which on average 44.343.411 (99,91%) passed the trimming phase and were used for the following analysis. Initially, gut microbiome composition at phylum level was explored among the categories, (**Figure 13)**.

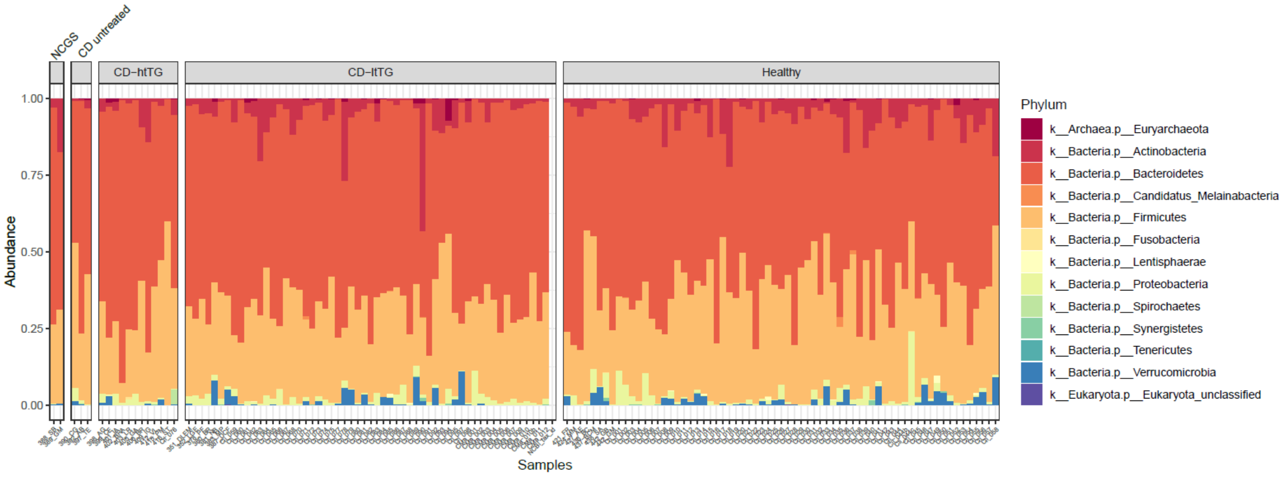


Figure 13. Relative abundances for each phylum in each sample of the investigated categories

Comparing CD-ltTG vs healthy controls a significant reduction of *Actinoabacteria* (p=0.00004) and *Verrucomicrobia* (p=0.004) abundance as well as an increase in *Bacteroidetes* (p=0.02) was noticed. CD-htTG group showed a reduction of *Euryarchaeota* (p=0.02) and *Fusobacteria* (p=0.04) abundance in comparison with the healthy controls while no significant differences were obtained comparing the CD groups. Subsequently, gut microbiome was explored at the species level. Considering the microbial richness, an average of 119, 104, 103 and 104 species were detected in CD untreated, CD-htTG, CD-ltTG, and healthy controls, respectively. Even if with a similar richness, a significantly different evenness was observed between CD-htTG and healthy controls (p<0.05). Comparing CD-ltTG vs healthy subjects five microbial species showed a significantly different abundance (p<0.05) in CD-ltTG. Specifically, *Bifidobacterium longum*, *Roseburia sp CAG 309*, *Ruminococcus bicirculans*, *Ruminococcus callidus* and *Eubacterium sp CAG 274* were less abundant while *Roseburia inulinivorans* more abundant in CD-ltTG (**Figure 14 A)**. When CD-htTG patients were compared to healthy controls, 7 microbial species resulted more abundant while *Ruminococcus bicirculans* was reduced in the CD group (p<0.05) (**Figure 14 B)**. There were no significant differences in the microbial abundances of the two largest CD groups.

A correlation analysis was run between gut microbiome profiles and the duration of the GFD. By analysing the whole group of CD treated patients (i.e., CD-htTG and CD-ltTG), a significant negative correlation with GFD duration was observed for *Aggregatibacter aphrophilus* (SCC=-0.28, p=0.02), *Haemophilus parainfluenzae* (SCC=-0.30, p=0.01) and *Streptococcus sanguinis* (SCC=-0.25, p=0.04) (**Supplementary Table 1A**). It is worth to notice that while *Streptococcus sanguinis* was also negatively correlated with age of investigated subjects (SCC=-0.26, p=0.03), no concomitant significant correlation was observed for *Aggregatibacter aphrophilus* and *Haemophilus parainfluenzae*. The same analysis focusing only on the CD-ltTG group confirmed the negative correlation for *Aggregatibacter aphrophilus* (SCC=-0.32, p=0.009) and *Haemophilus parainfluenzae* (SCC=-0.28, p=0.02) with GFD duration.

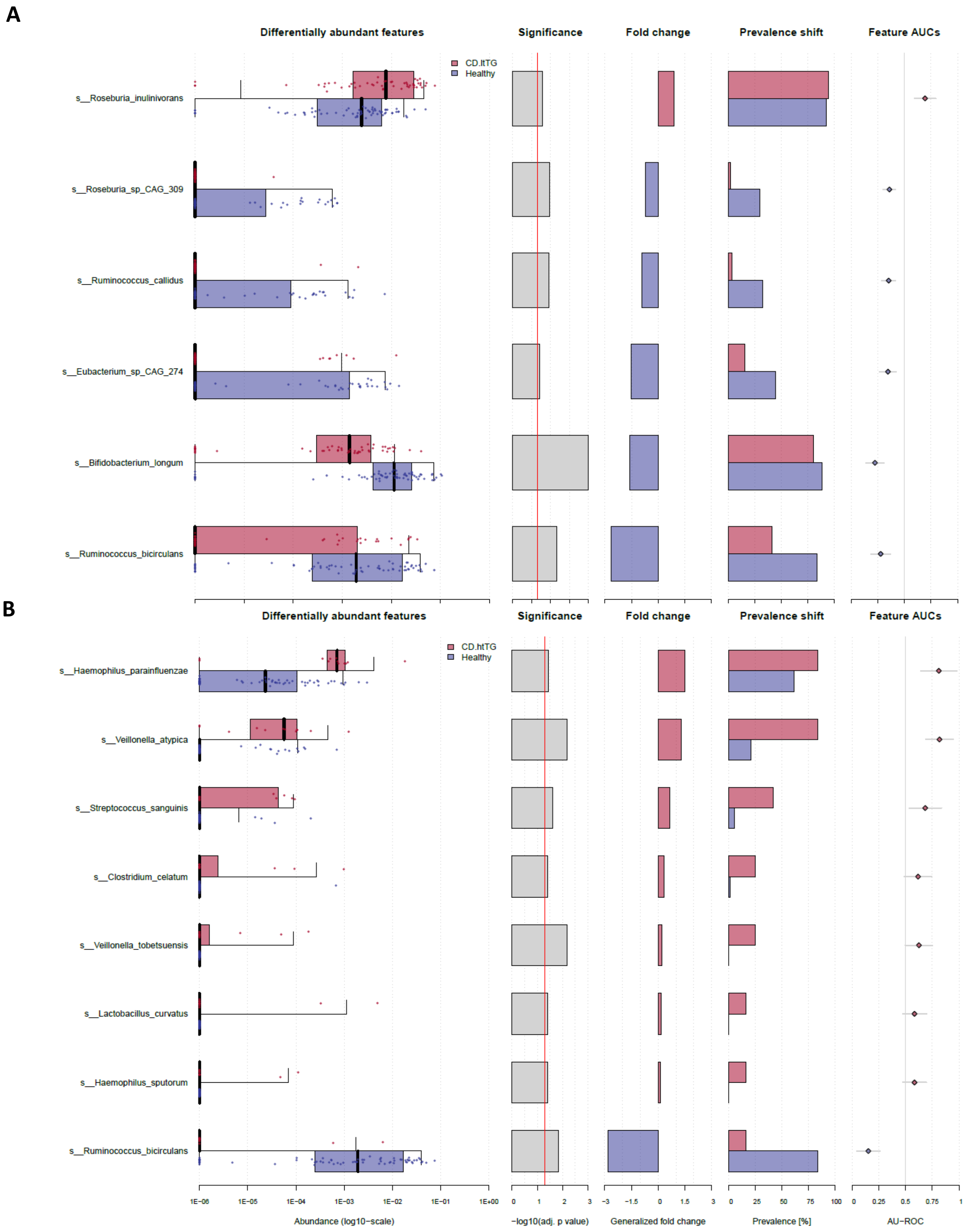


Figure 14. Relative abundances comparisons between CD-ltTG vs healthy (**A**) and CD-htTG vs healthy (**B**). For significantly associated microbial features (p<0.05), the plot shows: the abundances of the species across the two different classes (CD vs. controls), the significance of the enrichment calculated by a Wilcoxon test (after multiple hypothesis testing correction), the generalized fold change of each feature, the prevalence shift between the two classes, and - the Area Under the Receiver Operating Characteristics Curve (AU-ROC) as non-parametric effect size measure.

4.1.6 Gut microbiome functional profiles analysis

A functional profiling of the gut metagenomes was performed using HUMAnN3. Different gut microbiome composition is reflected in differential potential activities. A total of 24 pathways were identified with a significant different abundance among the analysed groups (**Figure 15**). Overall, in comparison with controls, the microbiome of both the CD groups showed a lower abundance in pathways related to the starch biosynthesis and degradation, amino acids biosynthesis (L-arginine biosynthesis I, L-methionine biosynthesis III), glucose (glycogen degradation I and II, gluconeogenesis lll) and an higher abundance in L-rhamnose degradation I, nitrate reduction and aerobactine biosynthesis pathways.

A correlation analysis was run between the pathways abundance and GFD duration. Considering the whole group of CD patients, PWY-5941: glycogen degradation II (eukaryotic) pathway resulted positive correlated (SCC=0.36, p=0.03) and DENITRIFICATION-PWY: nitrate reduction I (SCC= -0.26, p=0.03) negative correlated with GFD adherence (**Supplementary Table 1A**) while no significant results were observed considering only CD-ltTG group.

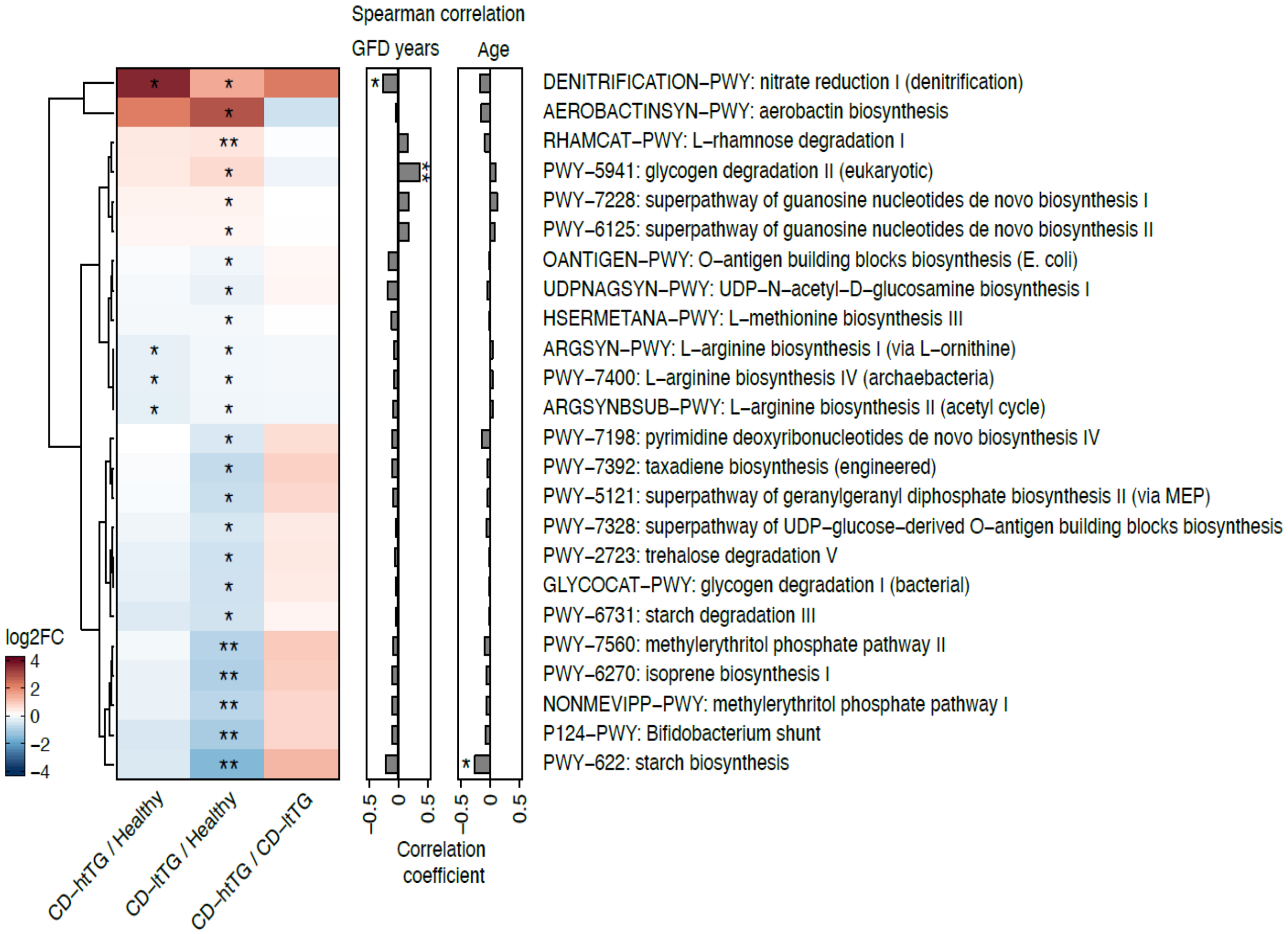


Figure 15: Heatmap representing log2FCs of the 24 pathways with a significant different enrichment among the comparisons performed. Correlation coefficient from spearman correlation analysis between the 24 pathways and GFD years duration and age are also reported.

4.1.7 miRNA-microbiome correlation analysis

To investigate the relationship between stool miRNA profiles and gut microbiome composition, the expression levels of all the observed DEmiRNAs were correlated to the abundances of the microbial species differently represented among the comparisons, **Figure 16**.

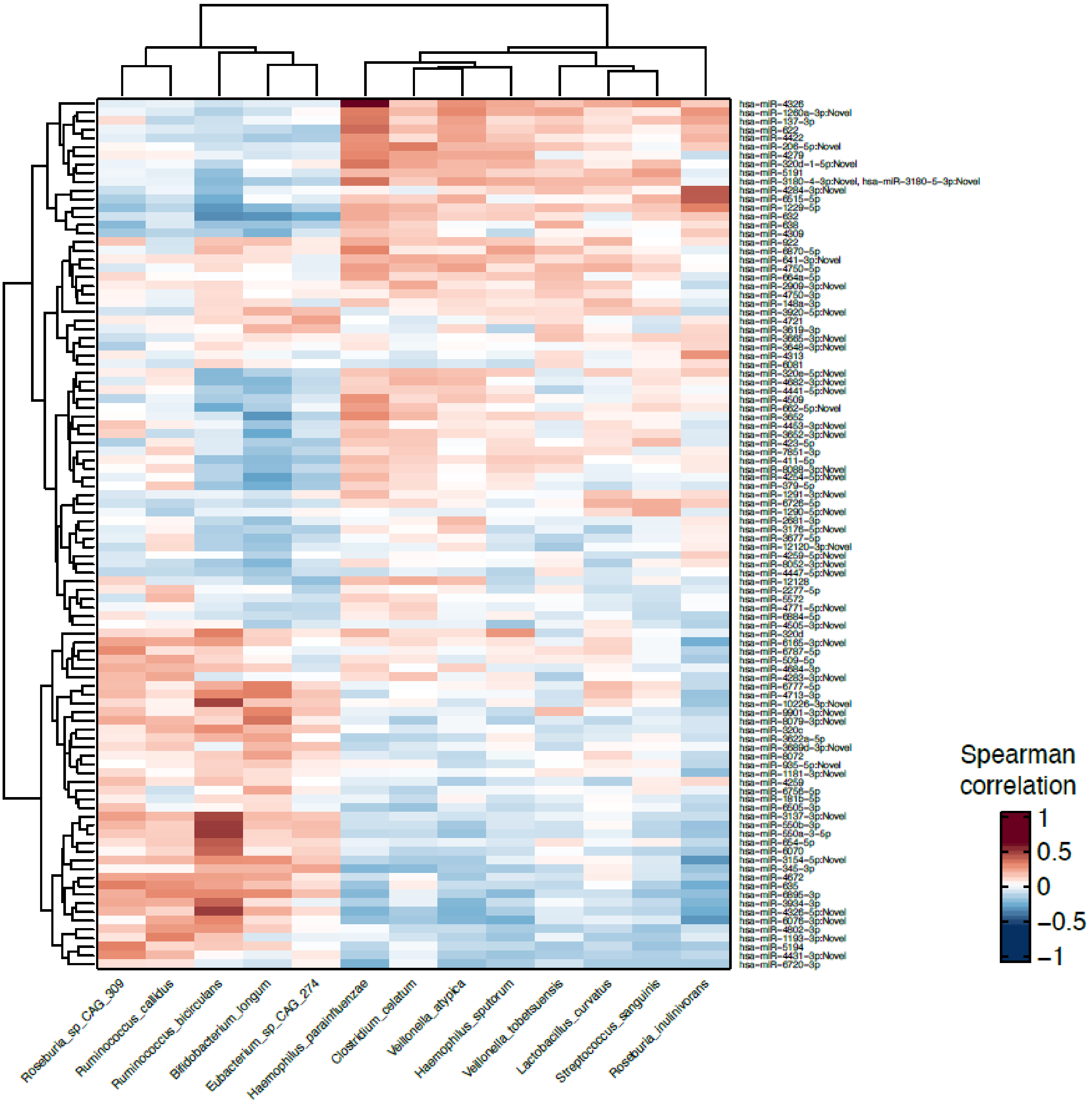


Figure 16: Heatmap representing SCCs computed between the 13 microbial species reporting a different abundance among the categories and the DEmiRNAs.

We chose to focus on *Ruminococcus bicirculans, Bifidobacterium longum,* and *Roseburia inulinivorans* since these microbial species showed a higher difference in relative abundances among the groups (**Supplementary Table 1B**). Significant correlations were observed between *Ruminococcus bicirculans* and 25 DEmiRNAs (p<0.05). Nine DEmiRNAs resulted inversely related with the abundance of this microbial species while 16 were positively correlated. Similarly, 24 DEmiRNAs significantly correlated with *Bifidobacterium longum* abundance (16 negatively correlated and 8 positively) (p<0.05). The analysis between DEmiRNA expression levels and the abundance of *Roseburia inulinivorans* highlighted 6 DEmiRNAs negatively correlated and 5 positively correlated (p<0.05). Noteworthy, miR-1229-5p and *miR-3154-5p* were associated with all the three bacterial species while 4 and 6 just with *R.bicirculans* and *B.longum* and with *R.bicirculans* and *R.inulinivorans*, respectively (**Figure 17**).

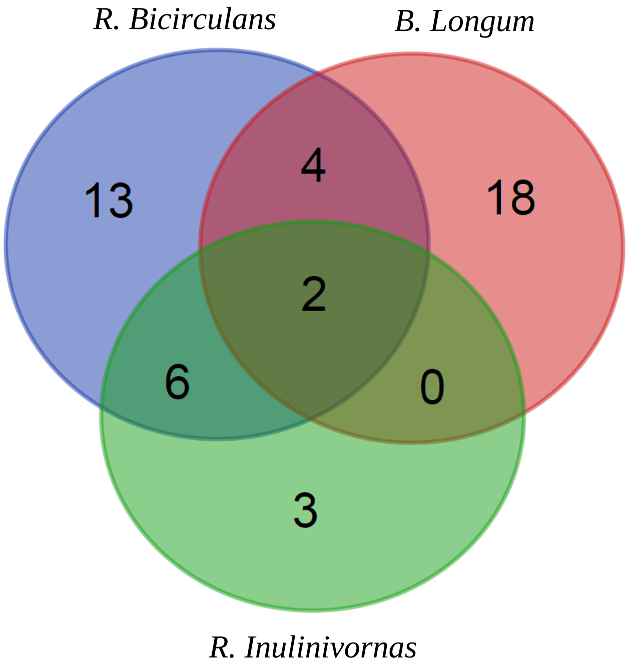


Figure 17. Venn diagrams showing the number of DEmiRNAs overlapping among the three correlation analyses performed with microbial species.

4.1.8. Integration of miRNAs, microbiome and diet

Data of the identified DEmiRNAs, microbial species abundance, and estimated daily nutrient intake from the dietary questionnaires were used to compute the DIABLO model. Results of the model are reported in **Figure 18A**. miRNAs and microbial species abundance showed a better efficiency compared to the nutrient intake for the discrimination of both CD categories from controls, with a major contribution of variate 1 compared to variate 2. **Figure 18B-C** shows the correlation analysis results among the three different types of data used for the DIABLO model for the variate 1 (**Figure 18B**) and variate 2 (**Figure 18C**).

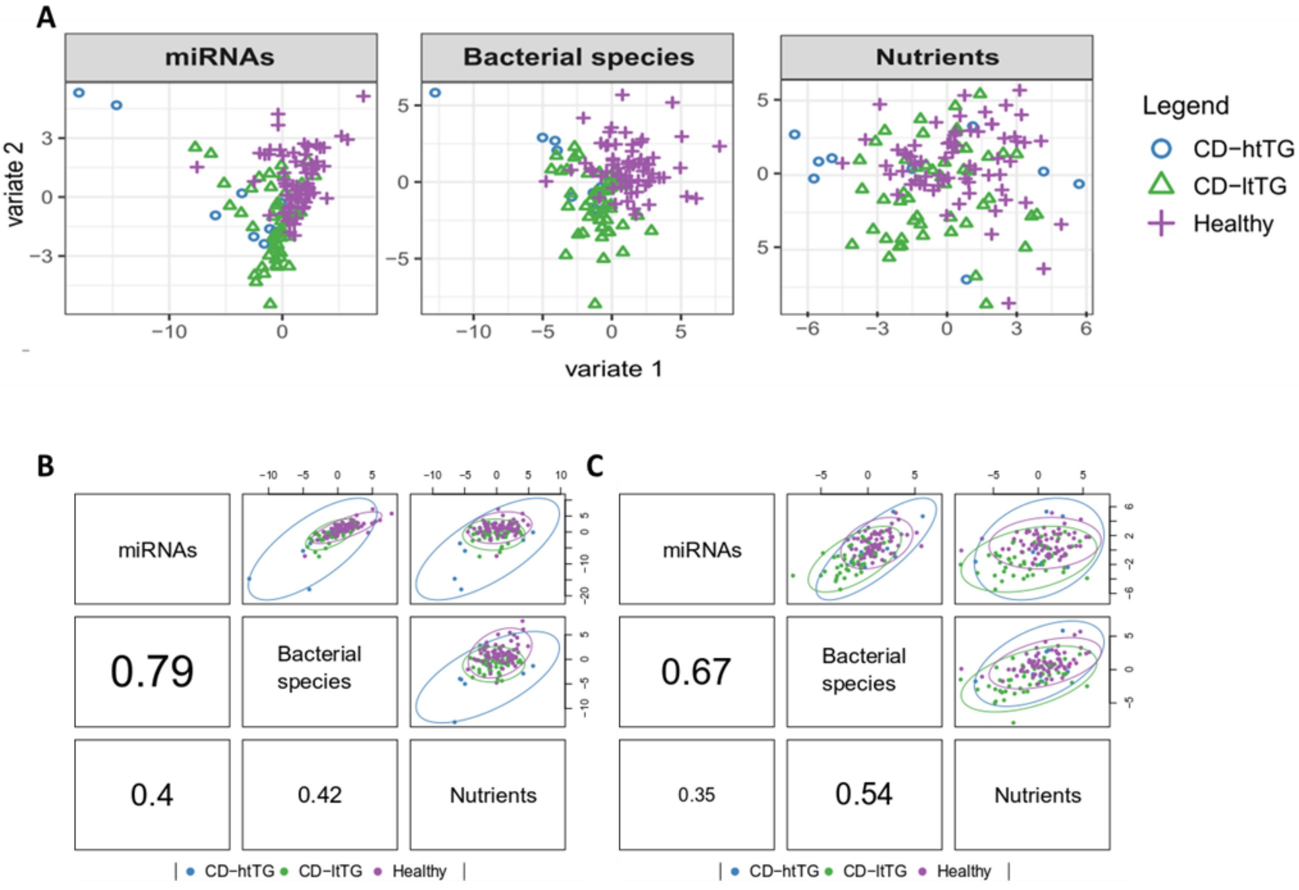
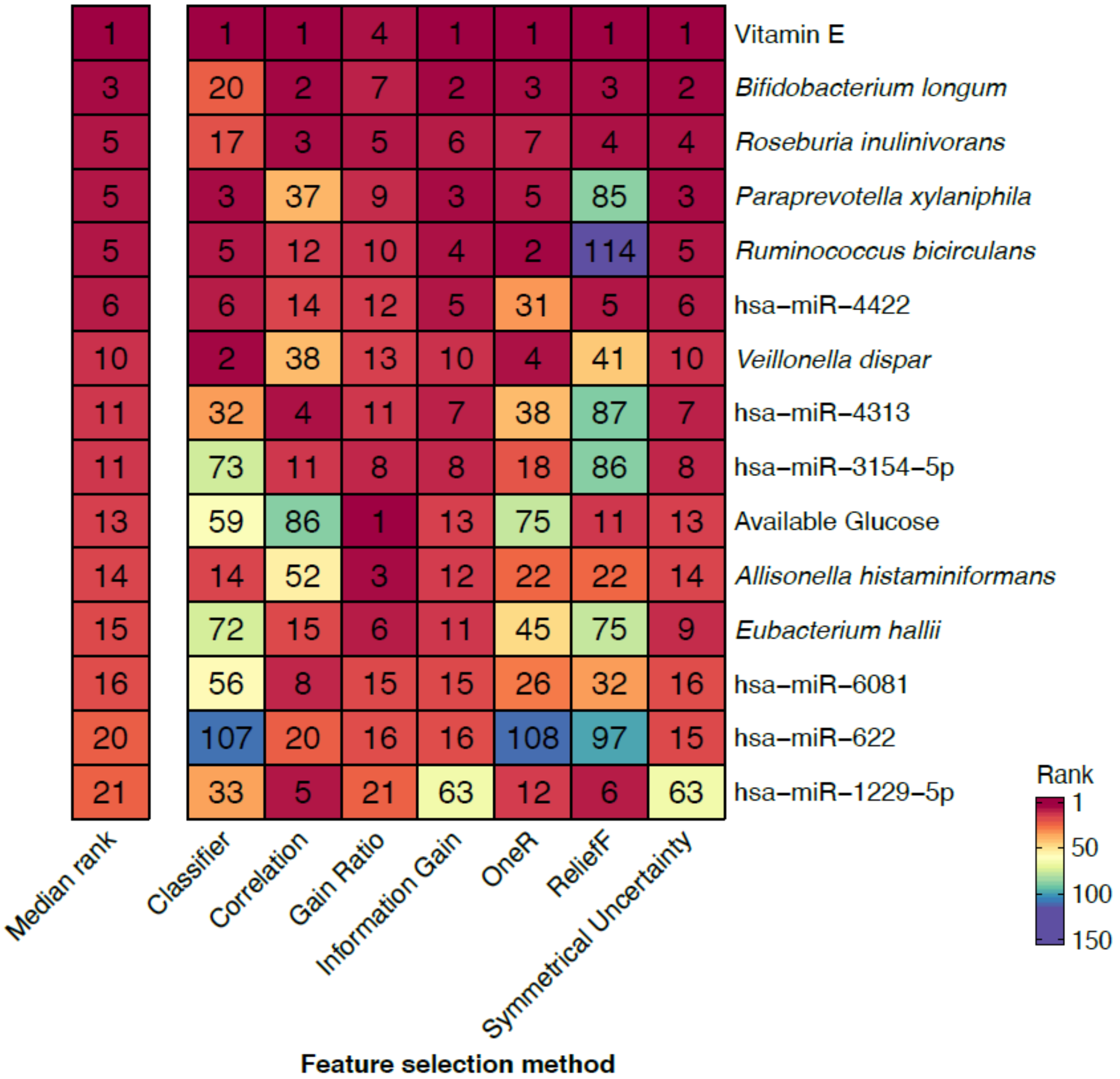


Figure 18: Results of the DIABLO model. **A.** Contribution of miRNAs, microbial species, and nutrients in the discrimination of CD categories and controls. **B-C** Results from correlation analysis of the three different types of data used for the variate 1 **(B)** and variate 2 **(C)**, respectively**.**

With the same input data used in the DIABLO model, a Weka model was also computed. Seven different feature selection methods were tested and the results ranked. A final median rank was calculated for each feature and the best 15 are reported in the heatmap of **Figure 19**. These 15 features includes 7 microbial species, 6 miRNAs and 2 food intake nutrients with Vitamin E and *B. longum* being those with the best performance in the tested models.



Median rank	Classifier	Correlation	Gain Ratio	Information Gain	OneR	ReliefF	Symmetrical Uncertainty	
1	1	1	4	1	1	1	1	Vitamin E
3	20	2	7	2	3	3	2	*Bifidobacterium longum*
5	17	3	5	6	7	4	4	*Roseburia inulinivorans*
5	3	37	9	3	5	85	3	*Paraprevotella xylaniphila*
5	5	12	10	4	2	114	5	*Ruminococcus bicirculans*
6	6	14	12	5	31	5	6	hsa–miR–4422
10	2	38	13	10	4	41	10	*Veillonella dispar*
11	32	4	11	7	38	87	7	hsa–miR–4313
11	73	11	8	8	18	86	8	hsa–miR–3154–5p
13	59	86	1	13	75	11	13	Available Glucose
14	14	52	3	12	22	22	14	*Allisonella histaminiformans*
15	72	15	6	11	45	75	9	*Eubacterium hallii*
16	56	8	15	15	26	32	16	hsa–miR–6081
20	107	20	16	16	108	97	15	hsa–miR–622
21	33	5	21	63	12	6	63	hsa–miR–1229–5p

Figure 19: Results from the Weka model. Heatmap showing the best 15 features reporting the best rank among the feature selection methods tested.

4.2 Study 2

4.2.1 Population Characteristics

The study group consisted of two independent cohorts of individuals that after colonoscopy were diagnosed with the following status: i) subjects with CRC (70% colon and 30% rectum cancer), ii) subjects with precancerous lesions, iii) subjects with inflammatory diseases and iv) healthy subjects with negative colonoscopy. The Cohort-IT included 221 subjects (79 Healthy, 41 Inflammations, 43 colorectal Polyps, and 58 CRCs). The mean age of the healthy subjects was 57.9 years (range spanning from 39 to 81 years), subjects with inflammation had a similar mean age (55.1) while polyp and CRC subjects were older than other groups (mean age of 64.1 and 71.2, respectively) (**Table 7)**. The gender distribution was similar in the groups of healthy (39 males and 40 females) and inflammation (20 males and 21 females) subjects but different in polyp and CRC groups where the number of males individuals was higher compared to females (26 and 17, and 41 and 17, in polyp and CRC, respectively). No other significant differences were present for the other variables.

The Cohort-CZ consisted of 162 individuals: 36 healthy subjects, 32 inflammations, 27 polyps and 67 CRC. Those groups were significantly different for age, gender distribution, and smoking habits (all $p<0.03$). The mean age of the healthy subjects was 57.8 years (range spanning from 40 to 76 years), subjects with inflammation had a similar mean age (58.7) while polyp and CRC subjects were older compared to the other groups (mean age of 63.1 and 68.0, respectively) (**Table 7**). The gender distribution was significantly different in the groups of healthy (14 males and 22 females) and CRC (46 males and 20 females) individuals compared to that of polyp and inflammation groups represented by an equal gender distribution. Cohort-CZ differs from Cohort-IT since in the first colorectal polyps consisted only of adenomas with low-grade dysplasia and hyperplastic polyps, and CRC patients presented mainly low-grade tumors (**Table 7**).

Table 7: Demographic and clinical characteristics of the Italian and Czech cohorts analysed in the study

		Cohort-IT					**Cohort-CZ**				
Covariate		**Healthy (n=79)**	**Inflammation (n=41)**	**Polyp (n=43)**	**CRC (n=58)**	**p-value**	**Healthy (n=36)**	**Inflammation (n=32)**	**Polyp (n=28)**	**CRC (n=66)**	**p-value**
Age	Average	57.9	55.1	64.1	71.2	1.80E-10	57.8	58.7	63.1	68.0	8.34E-06
	Range	39-81	30-82	42-92	54-87		40-76	41-75	48-82	40-88	
Sex	Male	39	20	26	41	5.37E-02	14	16	14	46	1.74E-02
	Female	40	21	17	17		22	16	14	20	
BMI	Average	25.36±4.63	24.81±3.3	25.05±3.9	25.1±5.3	8.17E-01	28.2±6.1	28.8±7.0	29.0±3. 5	27.1±5.4	1.61E-01
	Range	15.4-36.9	19.5-33.7	18.9-36.0	16.0-44.1		20.9-43.8	21.97-60.8	22.6-34.7	16.9-47.6	
Smoking status	Non-smoker	29	17	19	20	5.68E-01	25	24	13	32	2.53E-02
	Ex-smoker	33	11	11	23		3	0	8	12	
	Smoker	13	8	10	9		8	8	6	18	
	na	4	5	3	6		0	0	1	4	
Localization*	Distal colon			10	26				16	16	
	Proximal colon			15	11				11	15	
	Rectum			18	19				6	34	
	na			3	2				0	1	
Polyp type	Hyperplastic polyp			5					9		
	Tubular adenoma			25					16		
	TubuloVillous adenoma			7					2		
	Serrated adenoma			2					1		
	na			4					0		
Adenoma Grade	Low			30					28		
	High			9					0		
	na			4					0		
Number of polyps	Single			32					20		

	Multiple		7			8	
	na		4			0	
pT (Combined)	T1-T2			19			20
	T3-T4			36			43
	Tis			0			1
	na			3			2
pN	0			35			38
	1			12			13
	2			7			10
	3			0			1
	na			4			4
Metastasis	No			40			52
	Yes			15			11
	na			3			3
Staging	I			17			16
	II			18			16
	III			16			15
	IV			4			14
	na			3			5
Grade	G1-G2			25			44
	G3			27			18
	na			6			4
Inflammation type	IBD	16			15		
	Diverticular disease	21			17		
	Phlogosis	3			0		
	na	1			0		

*Localization of polyps exceeds the number of subjects with polyps, due to polyp detection in multiple sites. BMI= Body Mass Index; IBD= Inflammatory Bowel disease

4.2.2 Small-RNA sequencing of fecal samples

Small-RNA sequencing was initially performed on stool samples of all the 221 subjects characterizing the Cohort-IT cohort. After sequencing, an average of 1.51% of the reads were aligned to miRNA sequences while 5.16% of the reads were aligned to other sncRNAs (**Figure 20 A**). The distribution of miRNAs detected in all groups is reported in **Figure 20 B**. In total, 2,531 miRNAs were detected, with an average of 230 miRNAs in each sample. Considering miRNAs with a median of reads greater than 20 within at least in one of the analysed group, a total of 152 miRNAs were detected. Among these, 56 were in all four groups, while 34, 9, 3, and 1 were specifically detected in CRC, Healthy, Polyps, and Inflammation, respectively (**Figure 20 C**).

Besides miRNAs, other classes of sncRNAs were detected, including piRNAs, tRNAs, and snoRNAs. In total, 317 sncRNAs were identified in all subjects whose distribution in each sample group is shown in **Figure 20 D-E**.

From the sequencing of the 162 stool samples of the Cohort-CZ, an average of 1.06% and 3.46% of the reads were respectively aligned to miRNAs and other sncRNAs (**Figure 20 F**). The number of miRNAs and sncRNAs detected in stool samples of all subjects is reported in **Figure 20 H** and **Figure 20 J**, while the distribution of miRNAs and sncRNAs detected in all groups is reported in **Figure 20 G-I**. On average, 310 miRNAs and 891 other sncRNAs were detected in each sample. A total of 240 miRNAs and 637 sncRNAs were detected in at least one group while 129 miRNAs and 420 sncRNAs were detected in all four subject classes.

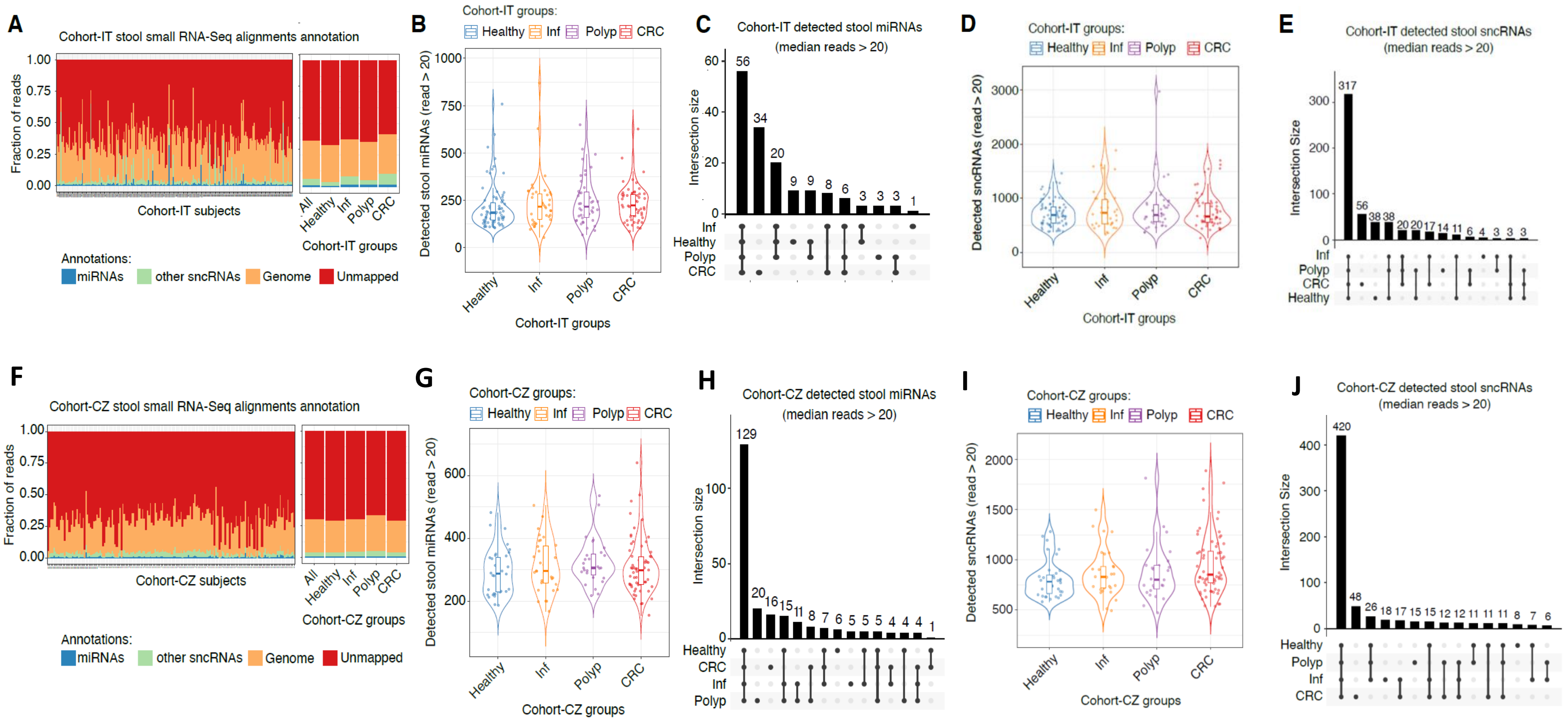


Figure 20. **A**) Read assignments for of each sample of Cohort-IT (left) and for each investigated disease category (right). **B**) miRNAs detected in the four main groups of Cohort-IT. **C**) Upset plots representing detected miRNAs in common among different groups of subjects in Cohort-IT. **D**) sncRNAs with a minimum of 20 reads cut-off detected in each Cohort-IT subject from the four main groups; **E**) Upset plots representing detected sncRNAs in common between the different groups of subjects considered in the Cohort-IT, **F**) Read assignments for each samples of Cohort-CZ (left) and for each investigated disease category (right). **G**) miRNAs detected in the four main groups of Cohort-CZ. **H**) Upset plots representing detected miRNAs in common among the different groups of subjects in Cohort-CZ. **I**) sncRNAs with a minimum of 20 reads cut-off detected in each Cohort-CZ subject from the four main groups. **J**) Upset plots representing detected sncRNAs in common between the different groups of subjects considered in the Cohort-CZ.

4.2.3 Differential expression analysis among the disease categories

Differential expression analysis performed on the Cohort-IT showed that 110, 69, and 64 miRNAs were altered in CRC cases, polyp patients, and colorectal inflammation, respectively, when compared with healthy subjects ($p<0.05$). Comparing CRC and healthy subjects, 71 miRNAs were up-regulated and 39 down-regulated. The expression levels of the most significant dysregulated miRNAs between these groups are reported in **Figure 21 A**. Fifty-one DEmiRNAs, all up-regulated, were in common among the comparisons performed against healthy individuals (**Figure 21 B-C**). Conversely, 38, 7, and 3 DEmiRNAs were specifically altered in CRC, Polyp, and Inflammation, respectively. The CRC-specific DEmiRNAs were mostly down-regulated (**Figure 21 C**). When compared with healthy subjects, all DEmiRNAs in CRC had consistent expression levels. This was also noted in the comparison between CRC and Polyps ($p<0.0001$), with 12 and 29 miRNAs significantly up- and down-regulated in both CRC vs. healthy and CRC vs. Polyp comparisons (**Figure 21 C**).

DEsncRNAs were 321, 89, and 155 in CRC, Polyps, or Inflammations vs. healthy subjects, respectively (for all $p<0.05$). In total, 78 DEsncRNAs were in common among all three comparisons but only 179 between CRC and healthy controls. All of the common DEsncRNAs were up-regulated while most of the DEsncRNAs altered only in CRC were down-regulated (n=137, 76.5%). The most frequent DEsncRNAs found in the comparison between CRC and healthy were piRNAs (n=147, 45.8%), followed by tRNAs (n=137, 42.7%), and misc_RNA (n=9, 2.8%).

In stool samples of the Cohort-CZ, 27 DEmiRNAs were identified between CRC and healthy controls and six between Polyp and healthy groups (**Figure 21 D**). No DEmiRNAs were detected in the comparison between Inflammations versus Healthy. Among the DEmiRNAs in CRC patients compared with healthy controls, 20 (74.1%) were up-regulated and seven down-regulated (25.1%) (**Figure 21 E**). Like the Cohort-IT, a subset of DEmiRNAs in CRC patients with respect to healthy subjects was also significantly differentially expressed in the comparison between CRC and Polyps (n=8) or between CRC and Inflammations (n=12). In all comparisons between CRC and the other categories, four common miRNAs were always significantly differentially expressed.

Among the other sncRNAs, 56 were differentially expressed in the comparison between CRC and healthy subjects, while only two between Polyps and healthy controls. Most of the stool DEsncRNAs between CRC and healthy were annotated as piRNAs (n=36, 64.3%), followed by tRNAs (n=9, 16.1%), and misc_RNAs (n=5, 8.9%) as already observed in the Cohort-IT.

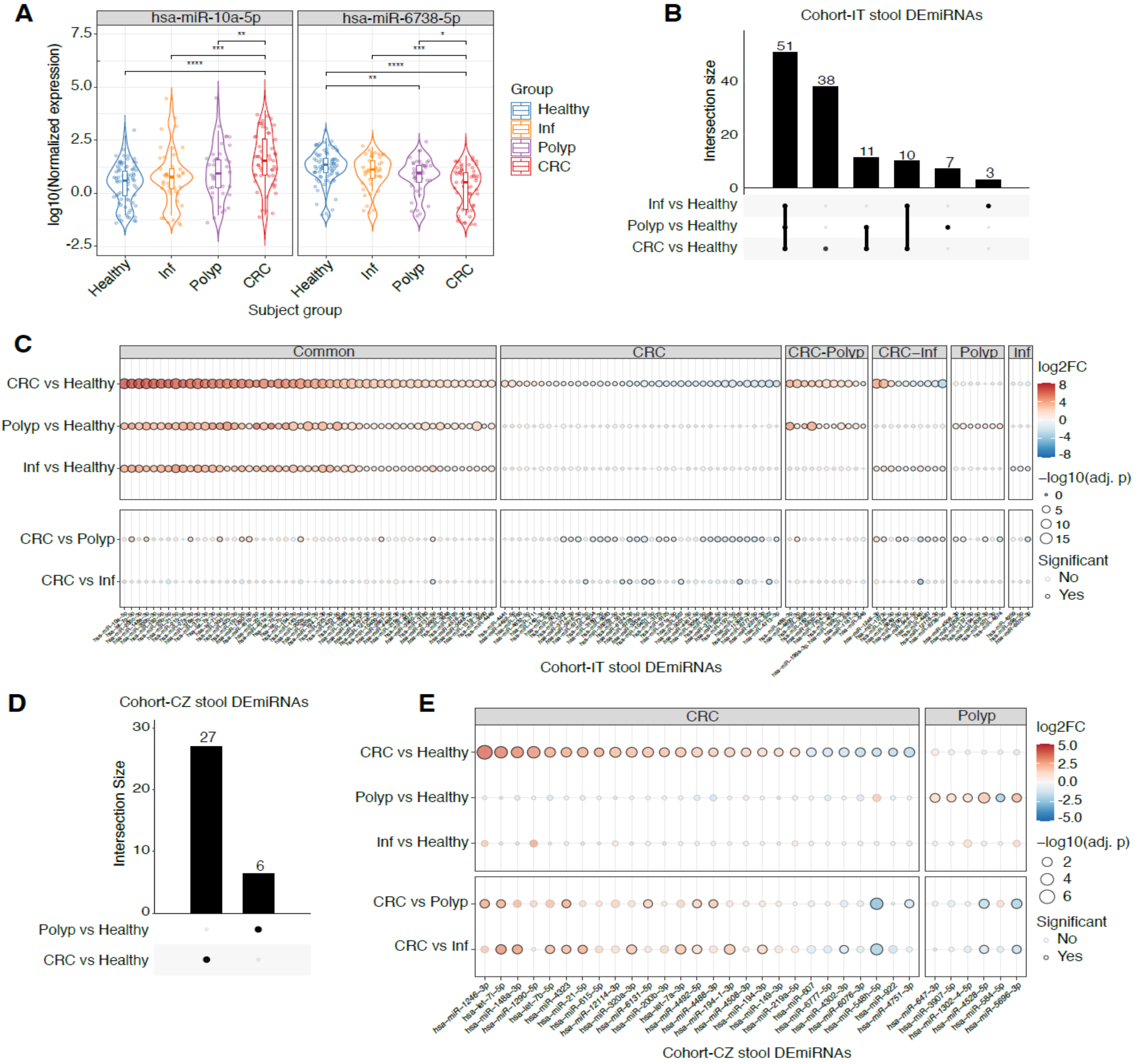


Figure 21 A) Examples of expression levels of two dysregulated miRNAs between CRC and Healthy in Cohort-IT. **B**) Upset plot showing overlapping/specific DEmiRNAs in different comparisons in Cohort-IT. **C**) Fold-changes and significance levels of the stool DEmiRNAs in all comparisons with the Healthy group in Cohort-IT. miRNAs are ordered by fold-change from the comparison CRC vs. Healthy and grouped in common or specific for disease categories. **D**) Upset plot showing overlapping/specific DEmiRNAs between different comparisons in Cohort-CZ. **E**) Fold changes and significance levels of stool DEmiRNAs in Cohort-CZ in the comparisons between CRC and Healthy (left) or Polyp and Healthy (right). miRNAs are ordered by fold change from the comparison CRC vs. Healthy.

4.2.4 Identification of common sncRNAs altered in both cohorts

Among the 27 DEmiRNAs in the Cohort-CZ, 19 were also dysregulated in the Cohort-IT and with the same expression trend (16 up- and 3 down-regulated in CRC versus Healthy). Overall, the fold changes of all the DEmiRNAs between CRC and healthy subjects in Cohort-IT and those in the Cohort-CZ positively correlated (r=0.73, p<0.0001). This correlation resulted even stronger considering only the 19 DEmiRNAs in common between both cohorts (r=0.86, p<0.0001) (**Figure 22 A**). All the 16 up-regulated overlapping DEmiRNAs were also significantly overexpressed in stool from the polyp and inflammation groups in comparison with healthy subjects (**Figure 22 B**). Conversely, the three miRNAs (namely, miR-607, miR-6777-5p, and miR-922) significantly down-regulated were also significantly underexpressed in CRC when compared to Polyps and Inflammations.

Clustering analysis based on the expression levels of the 19 DEmiRNAs defined two distinct groups of CRC patients (**Figure 22 C**). A further stratification according to clinical data showed that tumor grade was the major discriminant between these clusters (p<0.05).

Target enrichment analysis for the 19 identified DEmiRNAs highlighted several enriched terms. The top twenty hits included REACTOME *signaling by interleukin*, KEGG *Pathways in cancer*, REACTOME *transcriptional regulation by tp53* and others reported in **Supplementary Table 1C.**

Thirty-seven out of the 56 DEsncRNAs in Cohort-CZ were also significantly dysregulated in the Cohort-IT with a coherent expression trend.

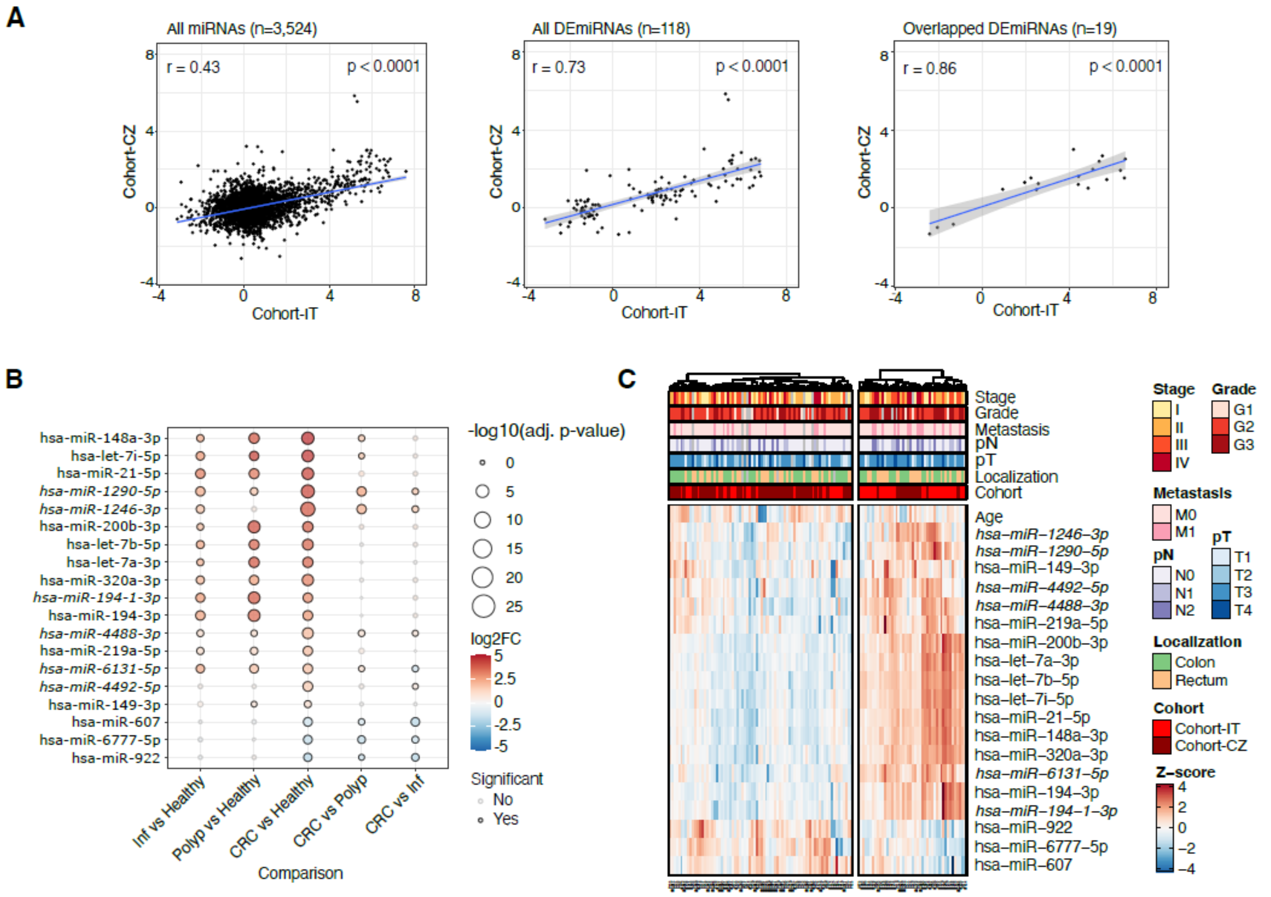


Figure 22 A. Scatterplots of stool miRNA log2FC in the Cohort-IT (x axis) and Cohort-CZ (y axis) resulting from the comparison between CRC and Healthy. Data are reported for all miRNAs (left panel), for DEmiRNAs in at least one cohort (middle) or in both (right). **B.** Log2FC and significance levels of 19 stool DEmiRNAs in both cohorts when comparing CRC and Healthy groups. miRNAs are ordered by decreasing log2FC computed in CRC vs. Healthy. **C.** Heat map reporting the normalized expression levels of stool DEmiRNAs in samples from CRC subjects of both cohorts together. The heatmap column annotations report the clinical data: tumor grade, presence of metastasis, lymph node invasion status (pN) and tumor size (pT) based on the TNM system.

4.2.5 fecal miRNA predictive model discriminating CRC and adenoma

After the application of the three-steps supervised strategy described in Materials and Methods, the CRCvsHealthy model was composed of three miRNAs: miR-607, miR-1246-3p, and miR-6777-5p. The overall classification performances of this model to distinguish CRC patients versus healthy subjects (given by the Area Under the Curve, AUC) were tested performing a 10-fold cross-validation without any overlap between training and test sets in two settings. First, both Cohort-IT (AUC values 0.89 and 0.81 adjusting for gender and age or not, respectively) and Cohort-CZ (AUC values 0.86 and 0.78 with or without adjustment for gender and age, respectively) were tested separately. Second, the two cohorts were pooled obtaining AUC values of 0.87 and 0.78 adjusting for gender and age or not, respectively (**Figure 23 A**).

Considering CRC versus adenoma (i.e., excluding hyperplastic polyps) patients, the three-steps supervised strategy defined the CRCvsAdenoma model composed of nine miRNAs: miR-1246-3p, miR-3180-3p, miR-300, miR-2110, miR-182-5p, miR-12114-3p, miR-92a-3p, miR-3169 and miR-6509-5p. The classification performances of CRC versus adenoma patients of the pooled cohort were AUC=0.73 and AUC=0.71 with or without adjustment for gender and age, respectively. Stratifying patients according to adenoma histology, the classification performances of CRC versus tubulovillous adenoma patients (AUC=0.68 and AUC=0.57, with or without adjustment for gender and age, respectively) and CRC versus tubular adenoma patients (AUC=0.73 and AUC=0.74 with or without adjustment for gender and age, respectively) were also tested (**Figure 23 A**).

Analyzing the pooled cohort, CRC patients were classified from healthy subjects with a similar AUC of 0.87 as previously (**Figure 23 B**). Only minor differences considering high-grade (AUC=0.86) or low-grade tumors (AUC=0.89) versus healthy individuals were observed. The union of CRCvsHealthy model and CRCvsAdenoma model generated the CRCprogressive predictive model composed of 11 miRNAs. This former model also discriminated the presence of CRC+adenoma with respect to healthy subjects (AUC=0.80) (**Figure 23 C**). Stratifying for the adenoma histology, the CRCprogressive predictive model also discriminated tubular adenoma (AUC=0.77 and AUC=0.63) and tubulovillous adenoma (AUC=0.67 and AUC=0.66) from CRC and healthy subjects, respectively.

Finally, the CRCprogressive predictive model was also challenged for its accuracy to discriminate the (pre)oncological conditions from subjects with inflammations. The AUC results (CRC versus inflammations: AUC=0.82 and CRC+Adenomas versus inflammations: AUC=0.75) together with the high sensitivity values in the identification of CRC subjects (CRC versus inflammations recall=0.88 and CRC+Adenomas versus inflammations recall=0.89) proved the classification accuracy of our CRCprogressive predictive model.

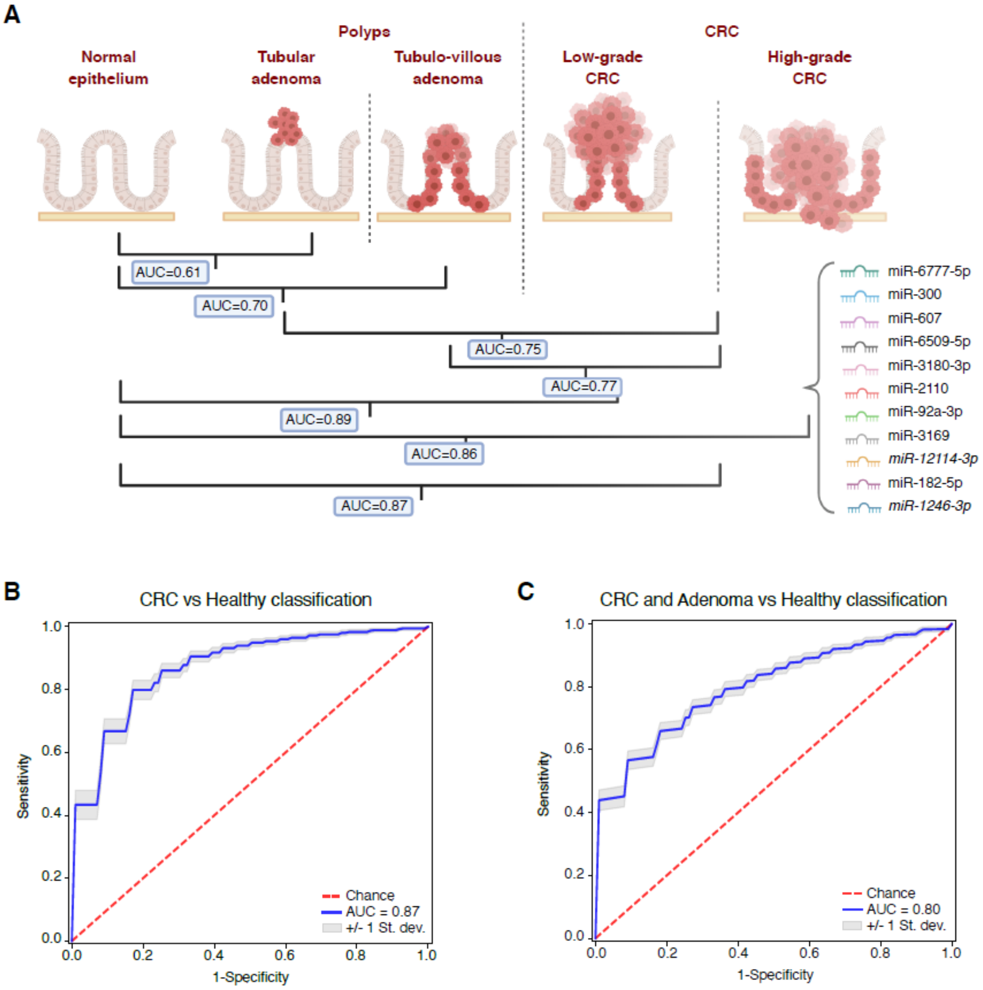


Figure 23. A. Discriminative capacity of the fecal miRNA predictive model. Receiver operating characteristic (ROC) curves and relative area under the curve (AUC) as a metric to assess the performance of 11 miRNAs in distinguishing **B.** CRC from healthy subjects and **C.** CRC and adenoma patients from healthy subjects.

4.2.6 Small RNA sequencing of tissue samples

Small RNA-seq was also performed on 24 CRC and 24 polyp tissue samples paired with their adjacent non-malignant colonic mucosa. An average of 71.6% and 8.9% of the reads were assigned to miRNAs and other sncRNAs, respectively, with an average of 543 miRNAs and 642 sncRNAs detected in each sample, **Figure 24**.

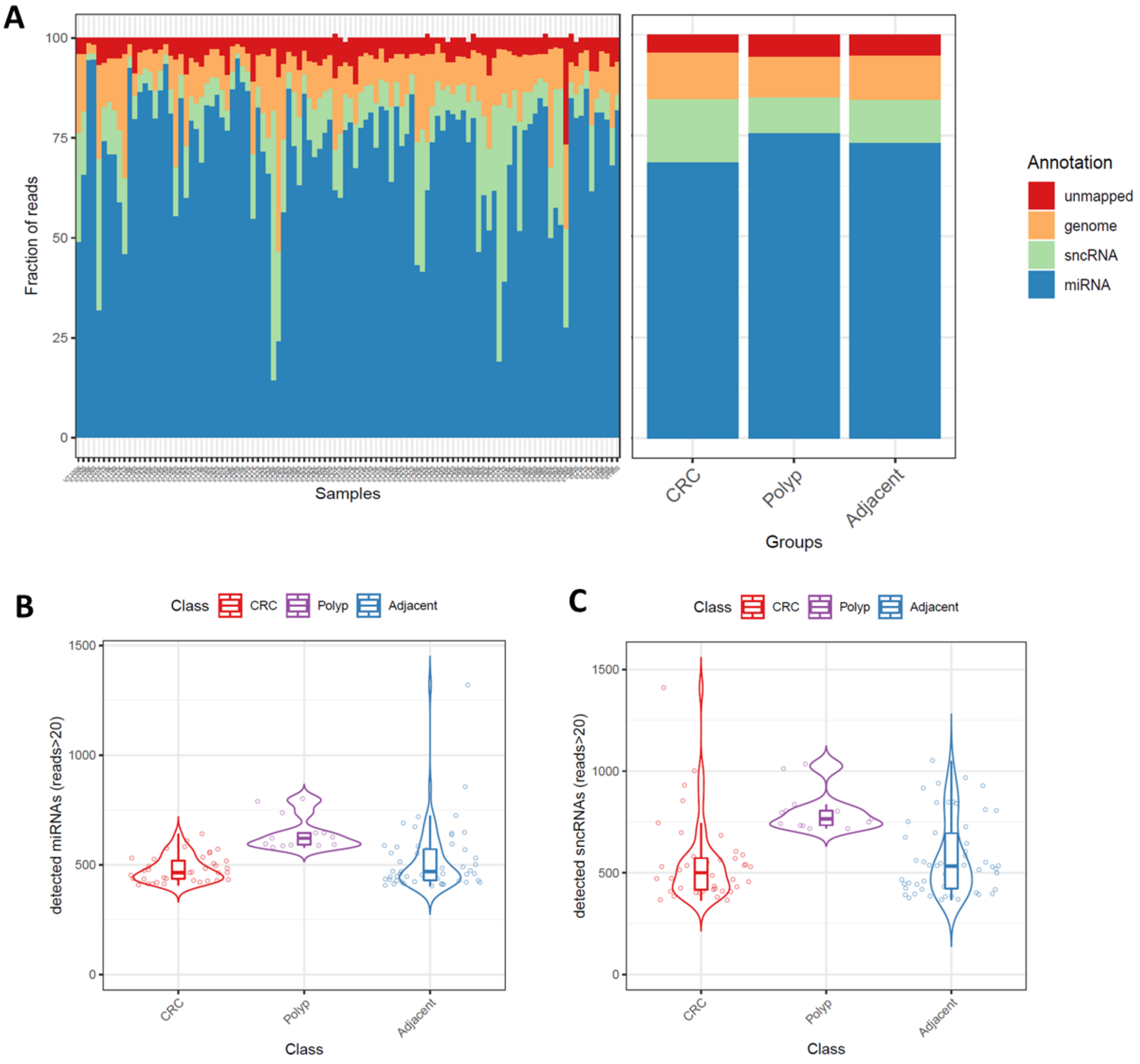


Figure 24. Read assignments for each tissue sample (left) and according to tissue types (right) (**A**). Stool miRNAs (**B**) and sncRNAs (**C**) detected in tissue types.

.

No significant difference was observed between the number of annotations detected in CRC tissues and adjacent mucosa (p=0.13 and p=0.33, for miRNAs and sncRNAs, respectively). Similarly, polyp and adjacent mucosa were also characterized by a non-significant difference in the total number of detected miRNAs (p=0.59) and other sncRNAs (p=0.93).

A paired differential expression analysis revealed 222 and 339 DEmiRNAs in tumor tissue vs. adjacent mucosa and polyp tissue vs. adjacent mucosa, respectively (p<0.05). The expression of the 19 stool DEmiRNAs in common between both cohorts was also investigated in tissue. All of the three down-regulated DEmiRNAs showed a low expression in tissue samples while 12 out of 16 up-regulated DEmiRNAs were highly expressed in tissues and up-regulated in the comparison between tumor and adjacent mucosa. Four (miR-21-5p, *miR-1290-5p*, miR-148a-3p, *miR-1246-3p*) and two (miR-320a-3p, let-7a-3p) miRNAs were significantly up- and down-regulated, respectively, in the comparison between tumor and adjacent mucosa. Four miRNAs (miR-21-5p, miR-194-3p, miR-148a-3p, and miR-200b-3p) were also significantly up-regulated in polyps with respect to adjacent mucosa while let-7i-5p was down-regulated, **Figure 25**.

In total, 199 and 269 DEsncRNAs were identified in the comparison between, respectively, tumor and polyp tissues with respect to the matched adjacent non-malignant mucosa. In the same samples, the expression levels of the 37 stool DEsncRNAs between CRC and healthy subjects in both cohorts were also investigated. Eleven and 17 DEsncRNAs resulted dysregulated between CRC vs. adjacent mucosa and polyps vs. non-malignant mucosa, respectively. Two DEsncRNAs, piR-33382 and the 28S rRNA were significantly down-regulated in both CRC and polyp tissues vs their matched adjacent mucosa.

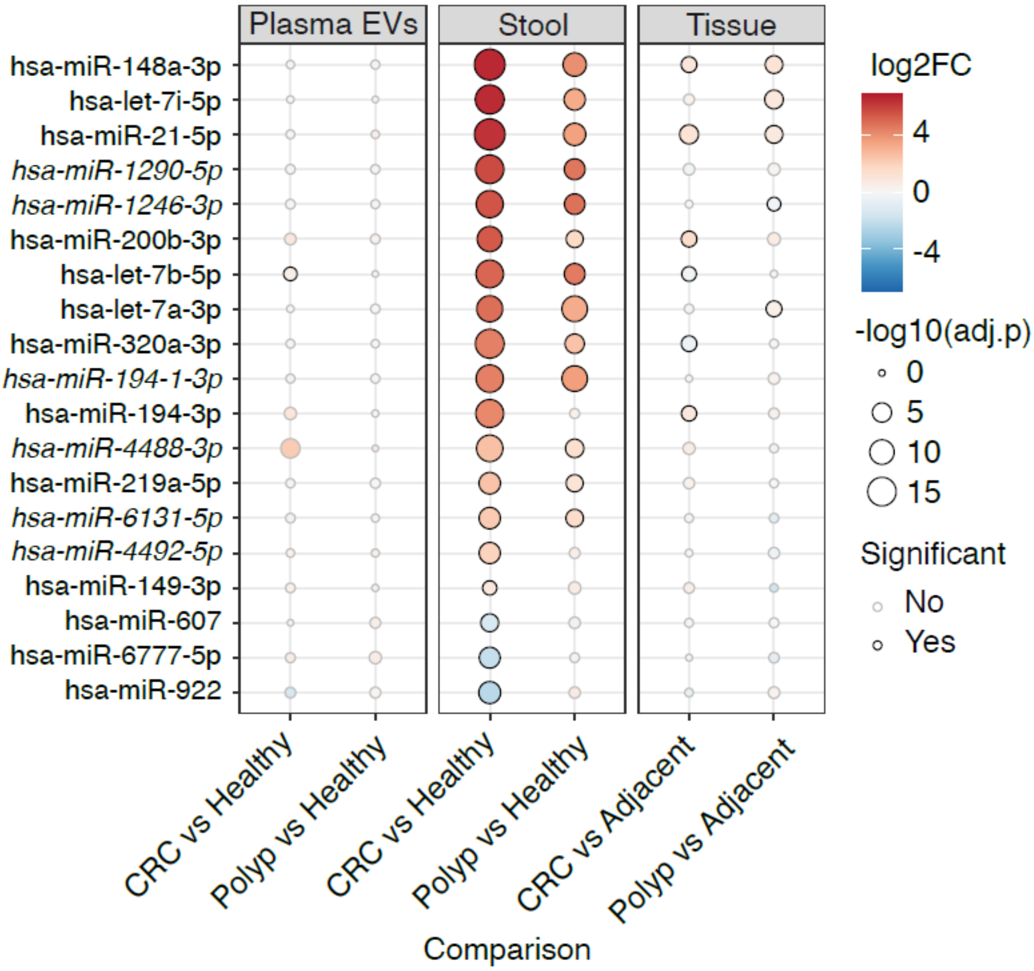


Figure 25. Dot plot reporting the log2FC and significance of 19 DEmiRNAs resulting when comparing CRC and healthy groups in stool of both cohorts (middle) and their levels as observed in plasma EVs (left), and in the paired analysis between CRC or polyps and associated adjacent mucosa (right).

4.2.7 Overview of DEmiRNA profiles among different investigated GI diseases

Finally, a comparison of the DEmiRNAs found among the GI diseases analysed in the two studies was performed to explore any similar or peculiar expression trend. For this purpose, Cohort-IT and Cohort-CZ groups from Study 2 were merged. First, the expression levels of fecal miRNAs commonly dysregulated in CRC, Adenoma and other inflammatory GI disorders vs controls were compared with those in stool and primary tissue (for CRC and Adenoma) of other GI diseases**.** No strong similarities with other examined GI disorders both in stool and tissue were observed for these group of DEmiRNAs **Figure 26** (left panel).

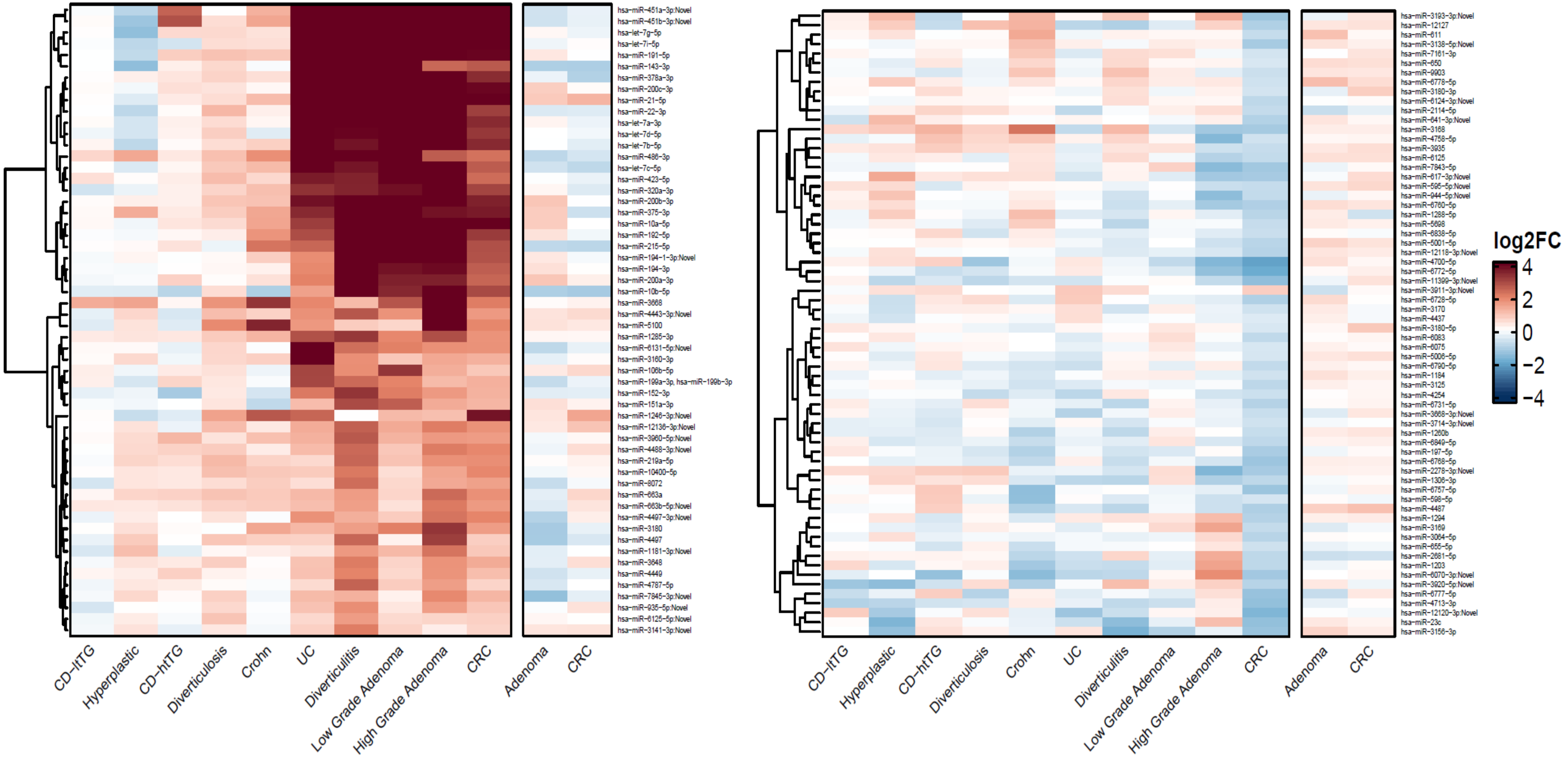


Figure 26. Heatmaps representing FCs fecal DEmiRNAs in common among CRC, Adenoma and other inflammatory GI disorders (left) and FCs of the fecal DEmiRNAs in CRC only (right), among the analysed GI diseases categories in stool. The additional columns represent expression of the same miRNA in investigated primary tissues (i.e., adenoma and tumor vs. normal mucosa).

A similar comparison among the categories was performed investigating the expression levels of those fecal miRNAs specifically dysregulated in CRC vs healthy controls (**Figure 26,** right panel). A group of DEmiRNAs in high grade Adenoma, Diverticulitis and Crohn's disease showed the same trend of down-regulation observed in CRC vs controls while the majority of miRNAs had not a particularly marked trend of expression (including those of CRC and Adenoma tissues).

Finally, the expression levels of fecal DEmiRNAs in CD-htTG vs controls were compared with those of other GI disease categories (**Figure 27**)**.** In general, miRNAs up-regulated in CD-htTG had similar expression trend in high grade Adenoma and Diverticulitis (i.e. miR-148a-3p, *miR-1290-5p*, miR-622*).*

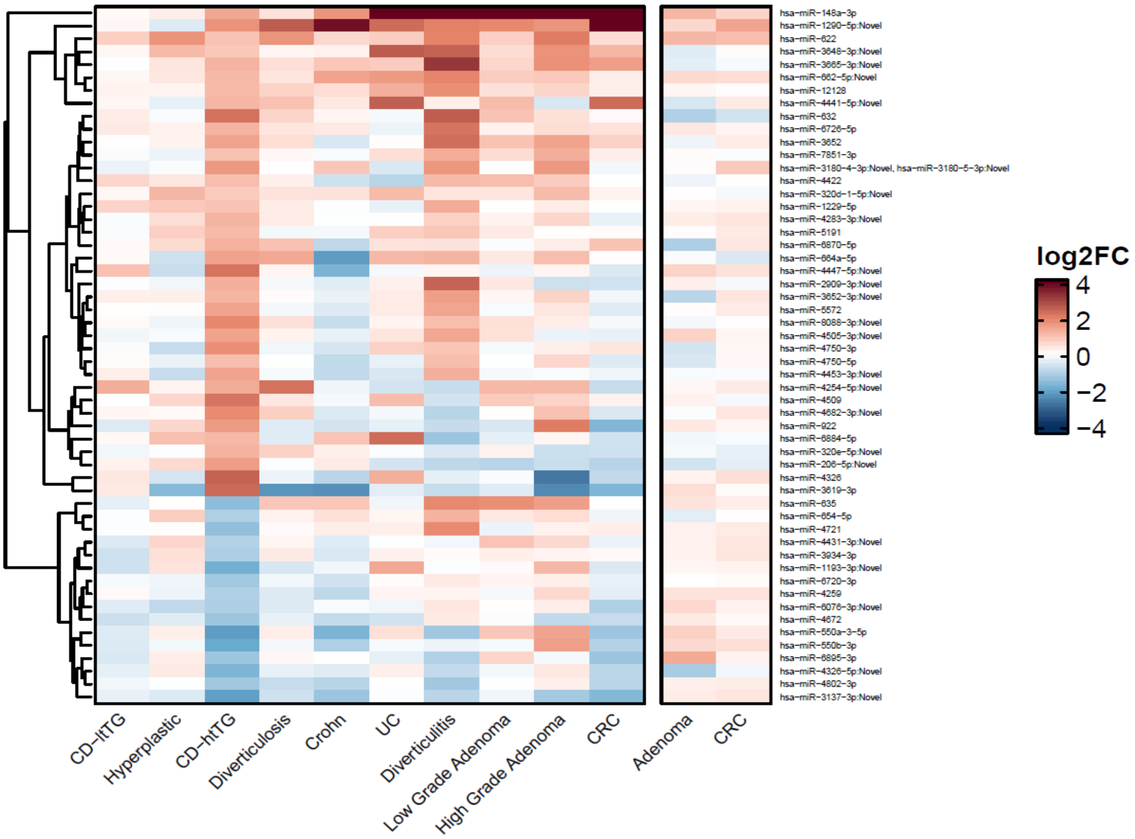


Figure 27. Heatmap representing FCs for the DEmiRNAs identified in CD-htTG vs controls across all the other GI disease categories investigated in stool and also in adenoma/tumor tissue pairs (last two columns on the right).

4.2.8 Microbial DNA and RNA detection in stool samples

For a subset of 80 individuals (29 CRC, 27 Polyp, and 24 healthy controls subjects) from the Cohort-IT, bacterial DNA from metagenomics analyses (published in Thomas et al. [189] and bacterial RNA from small RNA-seq analysis (not aligned on human genome) were further investigated and published in Tarallo et al. [212]. From metagenomic data resulted that the abundances of *Proteobacteria* and *Verrucomicrobia* in polyp patients were significantly different from those seen in the healthy group ($p<0.05$). In particular, *Verrucomicrobia* showed the lowest abundance, whereas *Proteobacteria* had intermediate abundance in the polyp group. *Proteobacteria* was the most significantly abundant phylum in the carcinoma group compared with both the healthy and polyp groups, while *Firmicutes* abundance significantly decreased from the healthy group to the carcinoma group.

Bacterial small RNAs (bsRNAs) analysis was performed using all the reads from the small RNA-seq previously not aligned to human miRNAs, human sncRNAs, human genome and miRNAs derived from animals or plants which can be commonly found in the Western diet. These surviving reads were mapped on bacteria annotations. Subsequently, data coming from this mapping step (18.9% of the non-mapped reads used as input) were integrated with the metagenomic data to improve the profiling of the bacteria (**Figure 28 A**). At taxonomic level, 50.2% of the annotations (n=130) detected from DNA and RNA were significantly correlated ($p<0.05$) and all of them were associated with a positive correlation (median r=0.64). The most correlated bacterial phyla were *Spirochaetes* ($r = 0.902$, $p<0.0001$), *Proteobacteria* ($r = 0.786$, $p<0.0001$), and *Fusobacteria* ($r = 0.646$, $p<0.0001$). At species level, *Porphyromonas asaccharolytica* was characterized by the highest correlation ($r = 0.999$; $p<0.0001$), while *F. nucleatum* ($r = 0.990$; $p<0.0001$) and *E. coli* ($r = 0.632$; $p<0.0001$) were among the 15 most highly correlated species. Differential expression analysis were performed based on the quantification of bsRNA annotations retrieved from RNA Central and Bacterial Small RNA Database tools. A total of 450 differentially expressed bsRNAs ($p<0.05$) were observed among the groups with the highest number (n=419) in CRC vs healthy controls comparison. Most of them (n=176) were annotated on *E. coli* and showed an increasing expression going from healthy, to polyp, and CRC subjects. Since the different abundance of *E. coli* DNA and bsRNA among the groups, a possible functional relationship with the DEmiRNAs was investigated. Thirteen miRNAs significantly correlated with the *E. coli* abundances estimated using the metagenomic data, with a trend of increased expression moving from the healthy to the carcinoma condition (**Figure 28 B**). Functional enrichment analysis of the human genes targeted by these hsa-miRNAs reported "*microRNA in cancer*" and "*Pathogenic Escherichia coli Infection*" as the most highly represented KEGG terms (p<0.001).

Finally, a combination of transcriptomic and genomic profiles to classify the recruited subjects according to disease status was tested. The Random Forest classification approach highlighted a signature providing the best accuracy in classifying CRC cases and controls when hsa-miRNAs, bsRNAs, and microbial DNA profiles were combined, with an AUC = 0.87 and AUC = 0.74 for CRC versus controls and CRC versus polyps, respectively. (**Figure 28 C-D**). The signature was composed of 57.7% human miRNAs and 42.4% microbial signals (**Figure 28 E**).

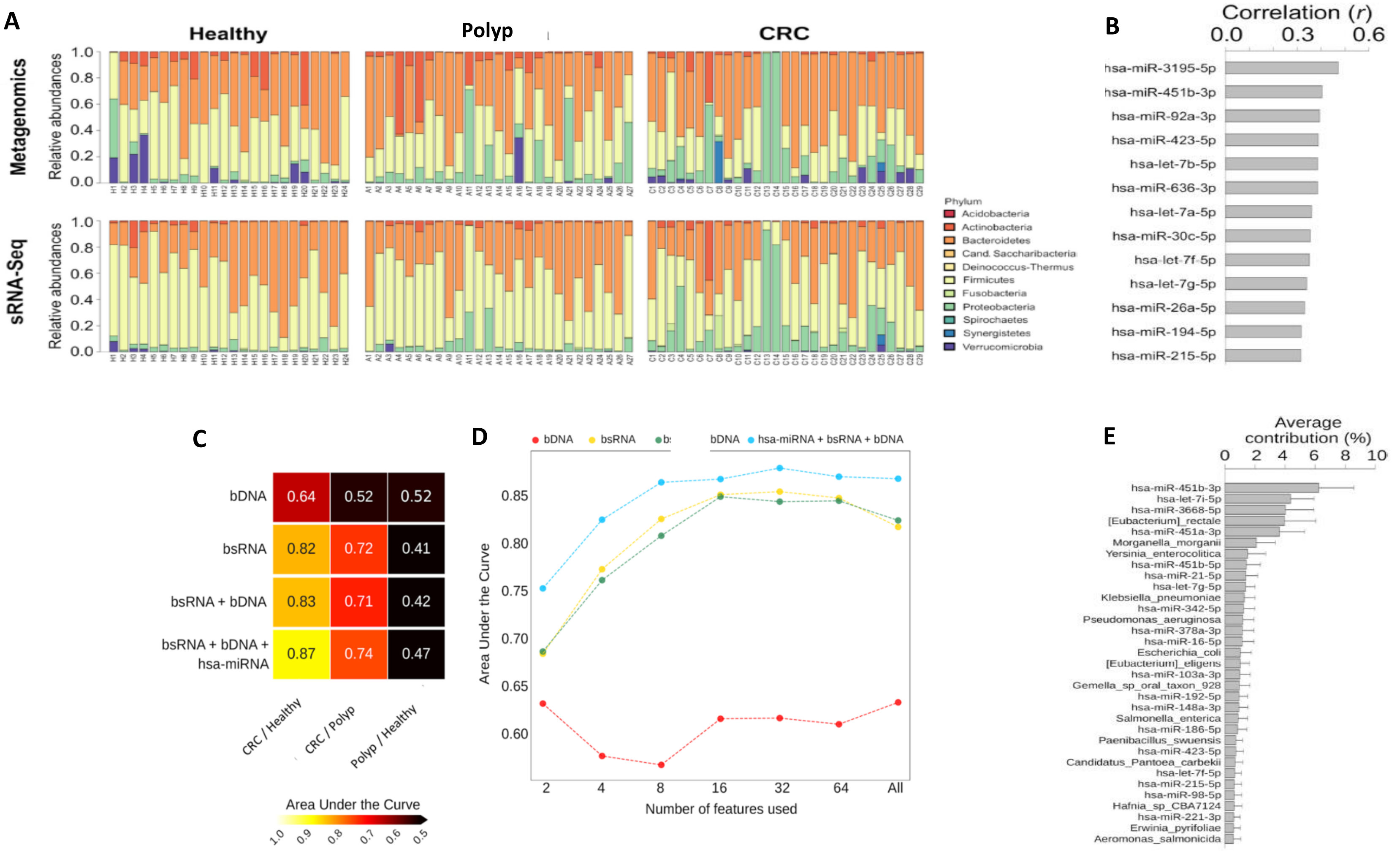


Figure 28. A. Stacked bar plots reporting bacterial phyla relative abundances by whole-metagenome sequencing (top) and small RNA-seq (bottom) data, respectively. **B.** Bar plot reporting hsa-miRNAs and hsa-sncRNAs correlated with the abundance of E. coli. **C.** Heat map reporting the area under the curve (AUC) computed by the Random Forest classifier using bacterial relative abundances provided by metagenomic data (bDNA), small RNA-Seq data (bsRNAs), and the combination of both bDNA and bsRNAs and combined with the expression levels of hsa-miRNAs (hsa-miRNAs + bDNA + bsRNAs). **D**. Line plot reporting the AUC computed by the Random Forest classifier. **E**. Bar plot reporting the average classification contribution of each of the 32 features providing the best classification accuracy of cancer and healthy samples

5 DISCUSSION

In the last decades, the RNA field has been consistently revolutionized by the discovery of additional players in the complex process of gene expression regulation, including miRNAs and other numerous sncRNAs. This has introduced a new era in which the central dogma of molecular biology was left behind, giving way to deciphering the complex interactions among molecular factors cooperating in the regulation of protein translation [116]. Concomitantly, the huge progress of NGS technologies, together with the development of new bioinformatics approaches, has provided the opportune tools for studying transcriptomics from another point of view. In this respect, due to the miRNome implication in several cellular processes, the study of miRNA deregulation has gained a great attention for broad research in several diseases, including those of the GI tract [220]. The high stability and easy detection of miRNAs, coupled with their direct correlation with individuals' physiological status, makes them excellent biomarkers [139]. Over the years, blood circulating miRNAs have been extensively investigated in various pathologies highlighting alterations in miRNA expression patterns that could be a consequence of heterogeneous physiological alterations since circulating miRNAs reflect a systemic condition of the patients [8]. Also, the characterization of miRNA profiles in primary tissues has revealed specific features of the disease under study; however, this investigation usually requires invasive procedures that are uncomfortable for the patients and unfeasible for comprehensive screening population programs. When studying GI diseases, these limitations can be overcome through the investigation of stool samples. Due to its strict relation with the altered site, this surrogate tissue contains factors released by intestinal cells, including tumor-derived ones [9]. As a confirmation, fecal miRNA dysregulation in affected individuals has been associated with inflammation, epithelial barrier dysfunction and altered apoptosis [221].

Fecal biospecimens also provide the possibility to study the human gut microbiome, since its essential functions related to metabolism, immune system education and regulation, and protection against pathogen invasion, the gut microbiome plays an essential role for the health of its host [171]. Indeed, the dysbiosis, similarly to fecal miRNA deregulation, has been associated with the alteration of intestinal barrier function and the inflammatory condition characterizing GI disorders [6]. In this respect, an emerging fascinating aspect is the interplay of fecal miRNAs with the gut microbiome, which could be the starting point for novel suitable biomarkers for GI diseases detection, monitoring or innovative therapeutic strategies development [7]. In this particular attempt, the sequencing of sncRNAs and the microbiota in stool samples could provide a non-invasive approach to detect common and conditions-specific molecular signatures of the different GI diseases.

This PhD work includes emerging findings connected with the above described aspects and reported in two main studies coauthored by the candidate (listed at the end of this thesis). The study by Thomas et al. [189] results from the metagenome analyses performed on the CRC study group. In Tarallo et al. [212] is reported the first description of the interplay of fecal miRNAs with the gut microbiome in relation to CRC. Moreover, three reviews have been useful exercises for collecting the available evidence from both *in vitro* and human studies on miRNAs and other ncRNA profiles as biomarkers of various types of cancer [9, 222] and more specifically as diagnostic, prognostic and predictive markers of CRC in EVs [8]. This last has been useful for enlarge the knowledge on this topic but also as a proper comparison of the results obtained from Study 2. All publications with co-authoships of the candidate and reporting findings from this PhD study, including those in preparation, are listed at the end of the Thesis.

In the present work, the expression profiles of miRNAs and other sncRNAs, as well as gut microbiome composition, were investigated in stool samples of patients with various GI disorders: celiac disease, inflammatory bowel diseases, diverticular diseases, colorectal polyps and colorectal cancer. The attention for CD has recently risen due to its increasing incidence in the industrialized countries, in part due to enhanced awareness of the disease [17]. Even though the specificity and sensitivity of the serological tests have improved over the last years, a duodenal biopsy is still required to confirm the diagnosis although its invasiveness. Moreover, patients ‘monitoring after the diagnosis remains a challenge. Indeed, excluding the disappearance of symptoms and the normalization of the serological test values, there is no strategy to evaluate healing of the intestinal villi without additional follow-up biopsies [66].

In this respect, the possibility to analyse stool miRNA profiles before and after CD diagnosis has never been explored and the present study, to the best of our knowledge, represents the first miRNome characterization in this biospecimen in relation to this pathology. For this purpose, 60 adult CD patients were recruited and matched for sex and age with healthy individuals without any allergies/intolerances or dietary restrictions. The individuals recruited were also screened for serum anti-transglutaminase antibodies levels. From the clinical point of view, this parameter allows to monitor the GFD adherence since the gluten-specific immune response causes their rapid increased levels. Moreover, in the present study, this analysis has also enabled to exclude the presence of any potential asymptomatic CD patient among the recruited healthy controls. Based on the anti-transglutaminase antibodies levels, the group of CD patients has been further stratified in those with low (CD-ltTG) and high (CD-htTG) anti-transglutaminase antibodies levels. The comparison of CD individuals with low and high TG levels with the healthy group revealed 44 and 49 DEmiRNAs, respectively. Interestingly, six DEmiRNAs (*miR-4447-5p*, *miR-4254-5p*, miR-622, miR-1229-5p, miR-3934-3p, and miR-4672) overlapped between these two comparisons with consistent trends (four up-regulated and two down-regulated). Finally, 20 DEmiRNAs were identified comparing the profiles of the two CD groups.

DEmiRNAs associated with low levels of TG should mainly reflect the long-term effects of a GFD, as also supported by the observed correlation between expression levels of some of the DEmiRNAs (miR-6505-3p, *miR-4771-5p*, miR-4684-3p, miR-320d) and the GFD adherence duration, not related to the age of subjects. Conversely, miRNA signatures in patients with high TG could be explained by an inflammatory response triggered by gluten-ingestion. In this respect, *in silico* functional analysis on the validated target genes of such miRNAs revealed among significantly enriched terms both *inflammasome* and *NLRP3 inflammasome*. In agreement with this, the triggering of *NLRP3 inflammasome* pathways by gliadin in CD was already reported by Gòmez Castro and collaborators [223]. Recently, stool miRNA profiling has allowed discriminating healthy subjects who adhere to different diets [224] and Tarallo et al., in preparation. In this respect, to better understand how GFD affects stool miRNA expression, a comparison between a group of healthy subjects that adhere to a low-gluten or gluten-free diet would allow discriminating specific CD signatures.

The six DEmiRNAs overlapping between the two CD groups in comparison with healthy subjects may be related to peculiar molecular aspects associated with this pathology and not related to the GFD adherence. Interestingly, the enrichment analysis performed with their target genes revealed a link with apoptotic pathways, whose increased activity has been reported in small intestine of CD patients [225, 226].

By investigating the few available studies on dysregulated miRNAs in relation to CD, some of the fecal DEmiRNAs identified in the present study were already reported altered in biopsies. As an example, an investigation on primary tissue of untreated patients reported the up-regulation of miR-638 and miR-1290-5p, in agreement with results from the present work (miR-638 up-regulated in CD-ltTG patients and miR-1290-5p up-regulated in CD-htTG vs healthy controls) [129]. Another example is a study finding miR-379 down-regulated in duodenal biopsies of both untreated and treated CD pediatric patients [131] even if this is in contrast with the present work where this miRNA resulted up-regulated in stool of CD-ltTG patients. Notably, miR-638 and miR-379 have among their targets also *TG*, which plays a crucial role in CD pathogenesis. Overall, this feature of a group of miRNAs whose expression in stool mirrors that of primary tissue support fecal biospecimens to be suitable for searching CD miRNA biomarkers.

For a subgroup of 90 individuals from the recruited group of CD and control subjects, miRNA expression levels in plasma samples were also evaluated. No significant differences were noticed performing the same comparisons of interest as in stool; this most probably was due to the low quality of sample collection and processing performed by the collaborating hospitals. However, removing the plasma samples recruited by the Hospitals and focusing only on those recruited at IIGM, Spearman correlation analysis of stool miRNA levels with their expression levels in plasma showed a significant correlation (p<0.05) for a huge group of miRNAs, six of which were among those observed differentially expressed in the performed comparisons (data not shown). This further supports stool samples as suitable biospecimens in the context of CD biomarkers.

Besides miRNAs, many other sncRNA biotypes exist and carry the potential of being acceptable cancer biomarkers. The analysis of these molecules in this study highlighted a total of 56 DEsncRNAs in the performed comparisons. Forty-two and 22 were the DEsncRNAs in subjects with high and low TG levels versus controls, respectively, with 8 DEsncRNAs overlapped between the two comparisons. No significantly altered sncRNAs were obtained comparing the two CD groups. This is an interesting finding, showing the potentiality also of other sncRNAs than miRNAs for CD biomarker purposes [227].

Several evidence reported an impaired gut microbiome composition associated to CD [228]. In our cohort, metagenome analysis highlighted a reduction of *Actinobacteria* and *Verrucomicrobia* and an increase in *Bacteroidetes* abundance in CD patients with low TG levels vs controls, while in those patients with high TG levels *Euryarcheota* and *Fusobacteria* showed a reduced abundance compared to controls. These findings agree with those of other studies performing similar analysis in fecal samples reporting *Actinobacteria, Bacteroidetes* and *Fusobacteria* among those with an altered abundance in CD [185]. A further investigation at species levels evinced a different abundance of a group of bacteria in both the two CD groups compared to controls. Among these, a reduced abundance of *Bifidobacterium longum* that we noticed in both CD groups (even if not with significant results in CD-htTG vs controls) has been already reported in relation to the disease [229]. This bacterium has also been found to decrease the production of inflammatory cytokines, CD4+ T cells and peripheral CD3+ T lymphocytes [230] and its administration also ameliorates the enteropathy induced by gliadin ingestion [231]. *Streptococcus sanguinis* was also observed more abundant in the saliva of CD patients compared to controls similarly to our results in stool samples [232]. Indeed, we found a higher abundance of this microbial species in CD-htTG and CD-ltTG (even if not significantly) compared to controls, and together with *Aggregatibacter aphrophilus* and *Haemophilus parainfluenzae* to be negative correlated with GFD years duration, suggesting the role of GFD to restore over the time a comparable *Streptococcus sanguinis* abundance comparable with that of controls. Correlation analysis between miRNAs and microbial species altered among the groups provided a consistent list of significant correlations. Interestingly, a single miRNA resulted correlated with more than one bacterial species.

Finally, the integration of miRNome, microbiome and daily nutrient intake data showed the potential to discriminate the investigated CD categories from the healthy controls, with a major contribution of identified DEmiRNAs and differentially abundant microbial species. A subsequent features selection analysis performed with different methods, also attributed the best median rank to the same identified miRNA and microbial features, together with Vitamin E.

Taken together, our findings show that CD condition has specific fecal sncRNAs and microbiome profiles which could be due to the influence of the GFD (as for CD-ltTG group) or the result of an inflammatory condition triggered by the gluten ingestion (as in CD-htTG one). This encourages the hypothesis of a potential application of fecal miRNAs and microbiome in the diagnostic and monitoring strategies for CD.

The early detection of CRC is still the most efficient approach to enhance patient prognosis and survival. Although current existing screening programs have been proven efficient in the detection of early precancerous lesions and CRC in asymptomatic patients, a significant number of patients are still diagnosed in advanced stages of the disease [69]. The possibility to find reliable, non-invasive stool biomarkers able to discriminate earlier and premalignant CRC phases could considerably improve both diagnosis, prompt treatment and then prognosis. To approach this issue, in the second study of this PhD work, a whole miRNome profiling was performed on stool samples from two cohorts of CRC patients and individuals with polyps or other GI disorders, according to the diagnosis at colonoscopy. Besides different polyp types, which may evolve to CRC, we also included samples from several gastrointestinal chronic diseases, like different types of IBDs and diverticulitis. When comparing these categories of patients with healthy controls, several miRNAs and other sncRNAs showed a significant dysregulation. Some miRNAs were up- or down-regulated in all the analyzed gastrointestinal lesions and others only altered in specific categories. Notably, all miRNAs dysregulated in common among CRC, polyps and inflammations were up-regulated, while those only altered in CRC were mainly down-regulated. This is in line with recent studies where down-regulation of miRNAs seems to be a premature step in the development of several cancers [233, 234] and others showing altered miRNA profiles in fecal samples of patients with inflammations [235, 236].

A set of 19 miRNAs was differentially expressed between CRC and healthy subjects in both the Italian and the Czech cohorts. This is particularly relevant since the two cohorts were completely independent and analyzed using a hypothesis-free small RNA-sequencing approach. Among the 19 DEmiRNAs, miR-194-3p, miR-21-5p, and miR-320a-3p (reviewed in [8, 9]) have been repeatedly reported dysregulated both in CRC tissues and biofluids. Other fecal miRNAs were reported altered for the first time in this work, supporting the use of a miRNome-wide approach for the discovery of new biomarkers. miRNA target enrichment analysis performed for this set of 19 miRNAs highlighted several significant and inherent terms such as *Pathways in cancer, signaling by interleukin, transcriptional regulation by tp53,* whose role in CRC has been largely described [237, 238]. For the Italian cohort, the 19 DEmiRNAs were also tested in tumor and polyps tissues paired with non-malignant adjacent mucosa and in plasma EVs. In contrast with the mirroring observed between DEmiRNAs profiles in stool and primary tissue, only a subgroup of the fecal DEmiRNAs showed the same trend in plasma EVs where only let-7b-5p was significantly up-regulated in CRC patients compared to healthy subjects. Of note, in general, the observed expression differences in stool were stronger than those in EVs.

So far, researchers mainly focused on the analyses of miRNAs as novel CRC biomarkers in plasma and serum mainly selecting candidate miRNAs from the literature and focusing on free-circulating miRNAs in plasma and serum (and not on EVs) [142]. This last could be a limitation of the studies, especially for plasma in which the contribution of freely circulating miRNA population may also arise from blood cells [227]. The lack of correspondence between sncRNA profiles in plasma EVs and stool could be explainable to a general higher heterogeneity of feces and most probably on the fact that miRNA and other small RNA profiling in plasma EVs may not reflect the presence of CRC except than only in advanced/metastatic stages [239].

Subsequently, the fecal miRNA outcomes were implemented in a model-based learning approach that allowed us to design a predictive model composed of 11 fecal miRNAs (miR-1246-3p, miR-607, miR-6777-5p, miR-3180-3p, miR-300, miR-2110, miR-182-5p, miR-12114-3p, miR-92a-3p, miR-3169, and miR-6509-5p) able to distinguish the different classes of patients. This model showed a high sensitivity in the detection of both CRC patients alone (average recall value 82%) and CRC or adenomas patients (average recall value 70%) from healthy subjects. In addition, it was also able to discriminate CRC versus the adenoma class (AUC=0.79), as well as versus tubular adenomas (AUC=0.75) and tubulovillous adenomas (AUC=0.77) patients. The high discrimination power of our predictive model was also confirmed when CRC (AUC=0.86) and CRC together with adenoma subjects (AUC=0.78) were classified with respect to healthy subjects and patients with inflammation. One of the major strengths of our predictive model is the high discrimination capacity among the pathological classes, also without performing gender and age adjustment. This supports the fact that the set of miRNAs identified is characterized by an elevated stratification power. All these aspects encourage the future use of a miRNA signature measured in stool as a valuable tool to implement already existing non-invasive screening tests. In search of biomarkers for non-invasive diagnosis of CRC and adenomas, a hypothesis-free approach such as sequencing in stool could be more informative than starting the research of biomarkers directly from the tumor/adenoma tissues as for example performed by Duran-Sanchon et al. [153]. Moreover, the possibility to complement already existing non-invasive CRC screening tests with additional analyses relatively fast and inexpensive makes this approach particularly attractive. In the view to define a good non-invasive tumor biomarker signature, we attempted to reduce the number of miRNAs from 11 to three (miR-607, miR-1246-3p, and miR-6777-5p), reporting a fairly identical high performance in distinguishing the presence of CRC (AUC=0.87) and CRC + adenoma patients (AUC=0.8) from healthy subjects.

Some of the miRNAs identified in the predictive model have been previously associated with CRC. As an example, miR-300 has a role in promoting proliferation and EMT-mediated CRC migration and invasion by targeting p53 [142]. Conversely, miR-182-5p, down-regulated in CRC tissues [240], targets the well-known oncogene c-Myc [88] and is an important modulator of Metadherin (MTDH), a promoter of cell proliferation, invasion, and migration ability. Also, miR-2110 has been found up-regulated in rectal but not in colon cancer [241] while miR-3180-3p was described as up-regulated in sporadic CRC tissues, particularly in CRC-derived liver metastasis from the same patients [242]. Altered miR-1246 levels have been mainly found in circulating exosomes in relation to metastasis and prognosis of CRC [239]. Interestingly, two down-regulated miRNAs in CRC, miR-607 and miR-6777-5p, were also included in our predictive model. For these last miRNAs there is almost no evidence in the literature of a relationship with CRC. In The Cancer Genome Atlas both miRNAs are frequently deleted in CRC (data not shown), supporting their down-regulation in stool and tumor tissues observed by us.

Besides miRNAs expression analyses, in both the studies of the present thesis we attempted to investigate in stool, plasma and tissues also other sncRNAs that can be assessed by small RNA-seq [211]. Even though miRNAs are the most extensively investigated, other sncRNAs carry the potential of being acceptable biomarkers. Crucial issues for their analyses is that their number/nomenclature is currently not definitive and there is still limited evidence of their expression in different biospecimens [243]. Hereby, we proved that several sncRNAs could be detected in stool and their altered profiles could be associated with CD, CRC or other GI diseases. In particular, in the second study, the model-based learning approach was also applied to the union of DEsncRNAs obtained from the Cohort-IT and Cohort-CZ. The sncRNA-based predictive model included piR-35467, piR-38736, piR-35468 and piR-35469, and it was able to classify CRC versus healthy subjects with an accuracy of 0.60 (data not shown). This is an interesting finding because until now the role of sncRNAs other than miRNAs have scarcely been investigated [227]. However, these results should also be taken with caution since three piRNAs of this signature (namely, piR-35467, piR-35468 and piR-35469) have very similar sequences that partially overlap. More extensive analyses should be performed in this field.

We also investigated the composition of microbiota and microbial small RNAs in fecal samples from healthy subjects and patients with polyp or carcinoma. Metagenome analysis revealed that the *Firmicutes* abundance in the carcinoma group was significantly different from those characterizing either the healthy or the polyp group. Interestingly, the *Verrucomicrobia* phylum, characterized by a significant peak of expression in the polyp samples, may represent a potential candidate biomarker for pre-cancer lesions. A significant increase in the abundance of the *Fusobacteria* phylum was also noticed in the carcinoma group compared to the healthy and polyp groups, as previously reported [189, 244, 245]. The abundance of *F. nucleatum*, a well-known CRC-related bacterium, increased from the healthy to the CRC group, albeit the variation was not statistically significant. Conversely, a significant increase in the abundance of *E. coli* was observed at both the DNA and bsRNA levels in our analysis, consistent with previous studies reporting an increase in the abundance of *E. coli* in the gut of CRC subjects [246]. Furthermore, in the multicohort analysis performed by Thomas et al. [189], *E. coli* was the second-ranked bacterial species not only in our cohort (named Cohort1 in the paper) but also in another cohort from the United States which included metagenomic data of stool samples from 52 CRC patients and 52 healthy controls [247].

In addition, correlation analysis suggested an interaction between miRNAs and *E.coli* species via target genes involved in the bacterium adhesion and phagocytosis pathways. Indeed, target functional enrichment analysis highlighted among the enriched terms "*Fc-gamma receptor signaling pathway involved in phagocytosis*," a pathway involving bacterial phagocytosis by immune cells as well as enterocytes [248]. Furthermore, two Toll-like receptor-coding genes (*TLR5* and *TLR4*) were also targeted by the miRNAs identified in our study. TLR4 is involved in the innate immune response to bacterium recognition [249], but it is also necessary and sufficient for bacterium phagocytosis by IEC-6 intestinal epithelial cells [250].

From the clinical point of view, the results obtained by applying a machine learning approach provided a strong support for the idea of using a combination of fecal microbial and human RNA biomarkers to better distinguish subjects with colonic adenoma or carcinoma from healthy individuals. However, a validation on larger independent cohorts of patients and healthy controls is mandatory to assess the accuracy of these potential biomarkers.

Overall, the work of this thesis has various strengths. These include, the adoption for all the study populations of the same protocol for the collection of stool, using the same tubes that contain a preservative buffer that maintains the stability of nucleic acids as well as the used miRNome-wide approach that has allowed to search all potentially altered sncRNAs. Similarly, the adoption of the shotgun metagenomic sequencing approach provides a better taxonomic resolution and genomic information than 16S sequencing, which is still the gold standard of microbiome typing research [169]. A remarkable strength of Study 1 was the possibility to screen the whole cohort for the serum levels of TG2-Ab, enabling the stratification of the CD patient groups and, importantly, to check for any asymptomatic CD patients in the control group (all of them presented normal levels of TG2-Ab) which is a recurrent limitation presented by numerous previous studies on CD [251]. In addition, the healthy control group was composed by individuals that matched those of CD group for age and gender avoiding any biases linked to the influence of these anthropometric characteristics on miRNAs and microbiome profiles (Francavilla et al, *in preparation*). The study population also represents a strength in Study 2 with the inclusion of two different and independent cohorts from countries with different diet and lifestyle habits and CRC rates. Finally, it is worth putting in evidence the robust prediction model that discriminated not only cancer cases but also adenoma both using miRNAs only or combined with bsRNAs and bDNA.

However, there are also some limitations. In Study 1, a limited number of untreated CD and NCGS individuals was recruited, which is why both these categories were only partially included in the analyses. Future investigations of sncRNAs and microbiome profiles in CD untreated and NCGS groups could highlight new biomarkers for CD diagnosis and add knowledge to the tricky clinical interpretation of NCGS condition [133, 252].

Similarly, in Study 2 the two cohorts were heterogeneous for individual categories: despite a large number of samples, the variegated spectrum of CRC, adenomas and other precancerous lesions was not exhaustively represented and deserves further investigations.

To conclude, the extensive sequencing adopted allowed to detect several miRNAs and other sncRNAs differentially expressed in stool as well as different microbial abundances across the GI diseases investigated. Interestingly, the hypothesis of host miRNA-microbiome interaction was supported by the integrative analysis showing numerous correlations between miRNA expression and microbial species abundances. Additionally, the machine learning approach identified a signature of miRNAs or miRNAs combined with bacterial RNA and DNA with a good discriminating power for the presence of a tumor or an adenoma. This is an excellent starting point supporting further research on this field which in the near future could hopefully improve diagnosis and prognosis of GI diseases.

6 REFERENCES

1. Bishehsari, F., et al., *Epidemiological transition of colorectal cancer in developing countries: environmental factors, molecular pathways, and opportunities for prevention.* World J Gastroenterol, 2014. **20**(20): p. 6055-72.
2. Keefer, L., *Behavioural medicine and gastrointestinal disorders: the promise of positive psychology.* Nat Rev Gastroenterol Hepatol, 2018. **15**(6): p. 378-386.
3. Bjarnason, I., A. Macpherson, and D. Hollander, *Intestinal Permeability - an Overview.* Gastroenterology, 1995. **108**(5): p. 1566-1581.
4. Davidson-Moncada, J., F.N. Papavasiliou, and W. Tam, *MicroRNAs of the immune system: roles in inflammation and cancer.* Ann N Y Acad Sci, 2010. **1183**: p. 183-94.
5. Siegel, J., V. Andresen, and P. Layer, *[Microbiome And Gastrointestinal Diseases].* Dtsch Med Wochenschr, 2019. **144**(14): p. 949-956.
6. Quigley, E.M., *Leaky gut - concept or clinical entity?* Curr Opin Gastroenterol, 2016. **32**(2): p. 74-9.
7. Liu, S., et al., *The Host Shapes the Gut Microbiota via Fecal MicroRNA.* Cell Host Microbe, 2016. **19**(1): p. 32-43.
8. Francavilla, A., et al., *Exosomal microRNAs and other non-coding RNAs as colorectal cancer biomarkers: a review.* Mutagenesis, 2020. **35**(3): p. 243-260.
9. Francavilla, A., et al., *Fecal microRNAs as non-invasive biomarkers for the detection of colorectal cancer: a systematic review.* Minerva Biotecnologica, 2019. **31**(1): p. 30-42.
10. Guarner, F. and J.R. Malagelada, *Gut flora in health and disease.* Lancet, 2003. **361**(9356): p. 512-519.
11. Mustalahti, K., et al., *The prevalence of celiac disease in Europe: results of a centralized, international mass screening project.* Ann Med, 2010. **42**(8): p. 587-95.
12. Reilly, N.R., et al., *Celiac disease in children: an old disease with new features.* Minerva Pediatrica, 2012. **64**(1): p. 71-81.
13. Sanders, D.S., et al., *Association of adult celiac disease with surgical abdominal pain: a case-control study in patients referred to secondary care.* Ann Surg, 2005. **242**(2): p. 201-7.
14. Hadjivassiliou, M., A.P. Duker, and D.S. Sanders, *Gluten-related neurologic dysfunction.* Handb Clin Neurol, 2014. **120**: p. 607-19.
15. Murray, J.A., et al., *Effect of a gluten-free diet on gastrointestinal symptoms in celiac disease.* Am J Clin Nutr, 2004. **79**(4): p. 669-73.
16. Singh, A., et al., *Non-Invasive Biomarkers for Celiac Disease.* J Clin Med, 2019. **8**(6).
17. Singh, P., et al., *Global Prevalence of Celiac Disease: Systematic Review and Meta-analysis.* Clin Gastroenterol Hepatol, 2018. **16**(6): p. 823-836 e2.
18. Lebwohl, B., D.S. Sanders, and P.H.R. Green, *Coeliac disease.* Lancet, 2018. **391**(10115): p. 70-81.
19. Llorente-Alonso, M.J., M.J. Fernandez-Acenero, and M. Sebastian, *Gluten intolerance: sex and age-related features.* Can J Gastroenterol, 2006. **20**(11): p. 719-22.
20. Fasano, A., et al., *Prevalence of celiac disease in at-risk and not-at-risk groups in the United States: a large multicenter study.* Arch Intern Med, 2003. **163**(3): p. 286-92.
21. Hovell, C.J., et al., *High prevalence of coeliac disease in a population-based study from Western Australia: a case for screening?* Med J Aust, 2001. **175**(5): p. 247-50.
22. Cook, H.B., et al., *Adult coeliac disease: prevalence and clinical significance.* J Gastroenterol Hepatol, 2000. **15**(9): p. 1032-6.
23. Pratesi, R., et al., *Prevalence of coeliac disease: unexplained age-related variation in the same population.* Scand J Gastroenterol, 2003. **38**(7): p. 747-50.
24. Cataldo, F. and G. Montalto, *Celiac disease in the developing countries: a new and*

challenging public health problem. World J Gastroenterol, 2007. **13**(15): p. 2153-9.

25. Catassi, C., et al., *The distribution of DQ genes in the Saharawi population provides only a partial explanation for the high celiac disease prevalence.* Tissue Antigens, 2001. **58**(6): p. 402-6.
26. Akbari, M.R., et al., *Screening of the adult population in Iran for coeliac disease: comparison of the tissue-transglutaminase antibody and anti-endomysial antibody tests.* Eur J Gastroenterol Hepatol, 2006. **18**(11): p. 1181-6.
27. Israeli, E., et al., *Prevalence of celiac disease in an adult Jewish population in Israel.* Isr Med Assoc J, 2010. **12**(5): p. 266-9.
28. Ertekin, V., et al., *Prevalence of celiac disease in Turkish children.* J Clin Gastroenterol, 2005. **39**(8): p. 689-91.
29. Kochhar, R., et al., *Prevalence of coeliac disease in healthy blood donors: a study from north India.* Dig Liver Dis, 2012. **44**(6): p. 530-2.
30. Makharia, G.K., et al., *Prevalence of celiac disease in the northern part of India: a community based study.* J Gastroenterol Hepatol, 2011. **26**(5): p. 894-900.
31. Lohi, S., et al., *Increasing prevalence of coeliac disease over time.* Aliment Pharmacol Ther, 2007. **26**(9): p. 1217-25.
32. Rubio-Tapia, A., et al., *Increased prevalence and mortality in undiagnosed celiac disease.* Gastroenterology, 2009. **137**(1): p. 88-93.
33. West, J., et al., *Seroprevalence, correlates, and characteristics of undetected coeliac disease in England.* Gut, 2003. **52**(7): p. 960-5.
34. Korponay-Szabo, I.R., et al., *Population screening for coeliac disease in primary care by district nurses using a rapid antibody test: diagnostic accuracy and feasibility study.* BMJ, 2007. **335**(7632): p. 1244-7.
35. Caio, G., et al., *Celiac disease: a comprehensive current review.* BMC Med, 2019. **17**(1): p. 142.
36. Gutierrez-Achury, J., et al., *Fine mapping in the MHC region accounts for 18% additional genetic risk for celiac disease.* Nat Genet, 2015. **47**(6): p. 577-8.
37. Nistico, L., et al., *Concordance, disease progression, and heritability of coeliac disease in Italian twins.* Gut, 2006. **55**(6): p. 803-8.
38. Ricano-Ponce, I., C. Wijmenga, and J. Gutierrez-Achury, *Genetics of celiac disease.* Best Pract Res Clin Gastroenterol, 2015. **29**(3): p. 399-412.
39. Welander, A., et al., *Infectious disease and risk of later celiac disease in childhood.* Pediatrics, 2010. **125**(3): p. e530-6.
40. Stene, L.C., et al., *Rotavirus infection frequency and risk of celiac disease autoimmunity in early childhood: a longitudinal study.* Am J Gastroenterol, 2006. **101**(10): p. 2333-40.
41. Sarno, M., et al., *Risk factors for celiac disease.* Ital J Pediatr, 2015. **41**: p. 57.
42. Elfstrom, P., J. Sundstrom, and J.F. Ludvigsson, *Systematic review with meta-analysis: associations between coeliac disease and type 1 diabetes.* Aliment Pharmacol Ther, 2014. **40**(10): p. 1123-32.
43. Collin, P., et al., *Endocrinological disorders and celiac disease.* Endocr Rev, 2002. **23**(4): p. 464-83.
44. Elfstrom, P., et al., *Risk of thyroid disease in individuals with celiac disease.* J Clin Endocrinol Metab, 2008. **93**(10): p. 3915-21.
45. De Re, V., R. Magris, and R. Cannizzaro, *New Insights into the Pathogenesis of Celiac Disease.* Front Med (Lausanne), 2017. **4**: p. 137.
46. Sharma, N., et al., *Pathogenesis of Celiac Disease and Other Gluten Related Disorders in Wheat and Strategies for Mitigating Them.* Front Nutr, 2020. **7**: p. 6.
47. Parzanese, I., et al., *Celiac disease: From pathophysiology to treatment.* World J Gastrointest Pathophysiol, 2017. **8**(2): p. 27-38.
48. Matysiak-Budnik, T., et al., *Secretory IgA mediates retrotranscytosis of intact gliadin*

peptides via the transferrin receptor in celiac disease. J Exp Med, 2008. **205**(1): p. 143-54.

49. Kim, S.M., T. Mayassi, and B. Jabri, *Innate immunity: actuating the gears of celiac disease pathogenesis.* Best Pract Res Clin Gastroenterol, 2015. **29**(3): p. 425-35.
50. Sturgeon, C. and A. Fasano, *Zonulin, a regulator of epithelial and endothelial barrier functions, and its involvement in chronic inflammatory diseases.* Tissue Barriers, 2016. **4**(4): p. e1251384.
51. Jelinkova, L., et al., *Gliadin stimulates human monocytes to production of IL-8 and TNF-alpha through a mechanism involving NF-kappaB.* FEBS Lett, 2004. **571**(1-3): p. 81-5.
52. Leonard, M.M., et al., *Celiac Disease and Nonceliac Gluten Sensitivity: A Review.* JAMA, 2017. **318**(7): p. 647-656.
53. Giersiepen, K., et al., *Accuracy of diagnostic antibody tests for coeliac disease in children: summary of an evidence report.* J Pediatr Gastroenterol Nutr, 2012. **54**(2): p. 229-41.
54. Lindfors, K., et al., *Coeliac disease.* Nat Rev Dis Primers, 2019. **5**(1): p. 3.
55. Ludvigsson, J.F., et al., *The Oslo definitions for coeliac disease and related terms.* Gut, 2013. **62**(1): p. 43-52.
56. Hujoel, I.A. and J.A. Murray, *Refractory Celiac Disease.* Curr Gastroenterol Rep, 2020. **22**(4): p. 18.
57. Rishi, A.R., A. Rubio-Tapia, and J.A. Murray, *Refractory celiac disease.* Expert Rev Gastroenterol Hepatol, 2016. **10**(4): p. 537-46.
58. Mukewar, S.S., et al., *Open-Capsule Budesonide for Refractory Celiac Disease.* Am J Gastroenterol, 2017. **112**(6): p. 959-967.
59. Volta, U., et al., *Features and Progression of Potential Celiac Disease in Adults.* Clin Gastroenterol Hepatol, 2016. **14**(5): p. 686-93 e1.
60. Korponay-Szabo, I.R., et al., *In vivo targeting of intestinal and extraintestinal transglutaminase 2 by coeliac autoantibodies.* Gut, 2004. **53**(5): p. 641-8.
61. Chow, M.A., et al., *Immunoglobulin A deficiency in celiac disease.* J Clin Gastroenterol, 2012. **46**(10): p. 850-4.
62. Ierardi, E., et al., *Seronegative celiac disease: where is the specific setting?* Gastroenterol Hepatol Bed Bench, 2015. **8**(2): p. 110-6.
63. Aziz, I., et al., *The clinical and phenotypical assessment of seronegative villous atrophy; a prospective UK centre experience evaluating 200 adult cases over a 15-year period (2000-2015).* Gut, 2017. **66**(9): p. 1563-1572.
64. Kamboj, A.K. and A.S. Oxentenko, *Clinical and Histologic Mimickers of Celiac Disease.* Clin Transl Gastroenterol, 2017. **8**(8): p. e114.
65. Ensari, A. and M.N. Marsh, *Diagnosing celiac disease: A critical overview.* Turk J Gastroenterol, 2019. **30**(5): p. 389-397.
66. Kelly, C.P., et al., *Advances in diagnosis and management of celiac disease.* Gastroenterology, 2015. **148**(6): p. 1175-86.
67. Cunningham, D., et al., *Colorectal cancer.* Lancet, 2010. **375**(9719): p. 1030-47.
68. Molinari, C., et al., *Heterogeneity in Colorectal Cancer: A Challenge for Personalized Medicine?* Int J Mol Sci, 2018. **19**(12).
69. Dekker, E., et al., *Colorectal cancer.* Lancet, 2019. **394**(10207): p. 1467-1480.
70. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA Cancer J Clin, 2018. **68**(6): p. 394-424.
71. Jemal, A., et al., *Annual Report to the Nation on the Status of Cancer, 1975-2014, Featuring Survival.* J Natl Cancer Inst, 2017. **109**(9).
72. Center, M.M., et al., *Worldwide variations in colorectal cancer.* CA Cancer J Clin, 2009. **59**(6): p. 366-78.
73. Stigliano, V., et al., *Early-onset colorectal cancer: a sporadic or inherited disease?* World J Gastroenterol, 2014. **20**(35): p. 12420-30.

74. Cross, A.J., et al., *A large prospective study of meat consumption and colorectal cancer risk: an investigation of potential mechanisms underlying this association.* Cancer Res, 2010. **70**(6): p. 2406-14.
75. Yurgelun, M.B., et al., *Cancer Susceptibility Gene Mutations in Individuals With Colorectal Cancer.* J Clin Oncol, 2017. **35**(10): p. 1086-1095.
76. Johnson, C.M., et al., *Meta-analyses of colorectal cancer risk factors.* Cancer Causes Control, 2013. **24**(6): p. 1207-22.
77. Gorham, E.D., et al., *Vitamin D and prevention of colorectal cancer.* J Steroid Biochem Mol Biol, 2005. **97**(1-2): p. 179-94.
78. Giovannucci, E., *Meta-analysis of coffee consumption and risk of colorectal cancer.* Am J Epidemiol, 1998. **147**(11): p. 1043-52.
79. Sinha, R., et al., *Caffeinated and decaffeinated coffee and tea intakes and risk of colorectal cancer in a large prospective study.* Am J Clin Nutr, 2012. **96**(2): p. 374-81.
80. Green, J., et al., *Menopausal hormone therapy and risk of gastrointestinal cancer: nested case-control study within a prospective cohort, and meta-analysis.* Int J Cancer, 2012. **130**(10): p. 2387-96.
81. Rothwell, P.M., et al., *Effect of daily aspirin on long-term risk of death due to cancer: analysis of individual patient data from randomised trials.* Lancet, 2011. **377**(9759): p. 31-41.
82. Rothwell, P.M., et al., *Long-term effect of aspirin on colorectal cancer incidence and mortality: 20-year follow-up of five randomised trials.* Lancet, 2010. **376**(9754): p. 1741-50.
83. Brenner, H., et al., *Prevention, early detection, and overdiagnosis of colorectal cancer within 10 years of screening colonoscopy in Germany.* Clin Gastroenterol Hepatol, 2015. **13**(4): p. 717-23.
84. Vogelstein, B., et al., *Genetic alterations during colorectal-tumor development.* N Engl J Med, 1988. **319**(9): p. 525-32.
85. Takayama, T., et al., *Analysis of K-ras, APC, and beta-catenin in aberrant crypt foci in sporadic adenoma, cancer, and familial adenomatous polyposis.* Gastroenterology, 2001. **121**(3): p. 599-611.
86. Advani, S.M., et al., *Clinical, Pathological, and Molecular Characteristics of CpG Island Methylator Phenotype in Colorectal Cancer: A Systematic Review and Meta-analysis.* Transl Oncol, 2018. **11**(5): p. 1188-1201.
87. Markowitz, S.D. and M.M. Bertagnolli, *Molecular origins of cancer: Molecular basis of colorectal cancer.* N Engl J Med, 2009. **361**(25): p. 2449-60.
88. Guinney, J., et al., *The consensus molecular subtypes of colorectal cancer.* Nat Med, 2015. **21**(11): p. 1350-6.
89. Watson, A.J. and P.D. Collins, *Colon cancer: a civilization disorder.* Dig Dis, 2011. **29**(2): p. 222-8.
90. Leggett, B. and V. Whitehall, *Role of the serrated pathway in colorectal cancer pathogenesis.* Gastroenterology, 2010. **138**(6): p. 2088-100.
91. De Palma, F.D.E., et al., *The Molecular Hallmarks of the Serrated Pathway in Colorectal Cancer.* Cancers (Basel), 2019. **11**(7).
92. Del Vecchio Blanco, G., et al., *Adenoma, advanced adenoma and colorectal cancer prevalence in asymptomatic 40- to 49-year-old subjects with a first-degree family history of colorectal cancer.* Colorectal Dis, 2013. **15**(9): p. 1093-9.
93. Del Vecchio Blanco, G., et al., *Familial colorectal cancer screening: When and what to do?* World J Gastroenterol, 2015. **21**(26): p. 7944-53.
94. Hassen, S., N. Ali, and P. Chowdhury, *Molecular signaling mechanisms of apoptosis in hereditary non-polyposis colorectal cancer.* World J Gastrointest Pathophysiol, 2012. **3**(3): p. 71-9.
95. Chung, D.C. and A.K. Rustgi, *The hereditary nonpolyposis colorectal cancer syndrome:*

genetics and clinical implications. Ann Intern Med, 2003. **138**(7): p. 560-70.

96. Miller, S. and S. Steele, *Novel molecular screening approaches in colorectal cancer.* J Surg Oncol, 2012. **105**(5): p. 459-67.

97. Levin, B., et al., *Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology.* CA Cancer J Clin, 2008. **58**(3): p. 130-60.

98. Wolf, A.M.D., et al., *Colorectal cancer screening for average-risk adults: 2018 guideline update from the American Cancer Society.* CA Cancer J Clin, 2018. **68**(4): p. 250-281.

99. Ladabaum, U., et al., *Cost-Effectiveness and National Effects of Initiating Colorectal Cancer Screening for Average-Risk Persons at Age 45 Years Instead of 50 Years.* Gastroenterology, 2019. **157**(1): p. 137-148.

100. Hewitson, P., et al., *Screening for colorectal cancer using the faecal occult blood test, Hemoccult.* Cochrane Database Syst Rev, 2007(1): p. CD001216.

101. Zorzi, M., et al., *Divergent Long-Term Detection Rates of Proximal and Distal Advanced Neoplasia in Fecal Immunochemical Test Screening Programs: A Retrospective Cohort Study.* Ann Intern Med, 2018. **169**(9): p. 602-609.

102. Baumgart, D.C. and S.R. Carding, *Inflammatory bowel disease: cause and immunobiology.* Lancet, 2007. **369**(9573): p. 1627-40.

103. de Souza, H.S. and C. Fiocchi, *Immunopathogenesis of IBD: current state of the art.* Nat Rev Gastroenterol Hepatol, 2016. **13**(1): p. 13-27.

104. Mak, W.Y., et al., *The epidemiology of inflammatory bowel disease: East meets west.* J Gastroenterol Hepatol, 2020. **35**(3): p. 380-389.

105. Flynn, S. and S. Eisenstein, *Inflammatory Bowel Disease Presentation and Diagnosis.* Surg Clin North Am, 2019. **99**(6): p. 1051-1062.

106. Annese, V., *Genetics and epigenetics of IBD.* Pharmacol Res, 2020. **159**: p. 104892.

107. Ahluwalia, B., et al., *Immunopathogenesis of inflammatory bowel disease and mechanisms of biological therapies.* Scand J Gastroenterol, 2018. **53**(4): p. 379-389.

108. Miner-Williams, W.M. and P.J. Moughan, *Intestinal barrier dysfunction: implications for chronic inflammatory conditions of the bowel.* Nutr Res Rev, 2016. **29**(1): p. 40-59.

109. Muehler, A., et al., *Clinical relevance of intestinal barrier dysfunction in common gastrointestinal diseases.* World J Gastrointest Pathophysiol, 2020. **11**(6): p. 114-130.

110. Yeshi, K., et al., *Revisiting Inflammatory Bowel Disease: Pathology, Treatments, Challenges and Emerging Therapeutics Including Drug Leads from Natural Products.* J Clin Med, 2020. **9**(5).

111. Stidham, R.W. and P.D.R. Higgins, *Colorectal Cancer in Inflammatory Bowel Disease.* Clin Colon Rectal Surg, 2018. **31**(3): p. 168-178.

112. Vermeire, S., G. Van Assche, and P. Rutgeerts, *C-reactive protein as a marker for inflammatory bowel disease.* Inflamm Bowel Dis, 2004. **10**(5): p. 661-5.

113. Gisbert, J.P. and A.G. McNicholl, *Questions and answers on the role of faecal calprotectin as a biological marker in inflammatory bowel disease.* Dig Liver Dis, 2009. **41**(1): p. 56-66.

114. Chen, P., et al., *Serum Biomarkers for Inflammatory Bowel Disease.* Front Med (Lausanne), 2020. **7**: p. 123.

115. Schaefer, J.S., *MicroRNAs: how many in inflammatory bowel disease?* Curr Opin Gastroenterol, 2016. **32**(4): p. 258-66.

116. Palazzo, A.F. and E.S. Lee, *Non-coding RNA: what is functional and what is junk?* Front Genet, 2015. **6**: p. 2.

117. Clement, T., V. Salone, and M. Rederstorff, *Dual luciferase gene reporter assays to study miRNA function.* Methods Mol Biol, 2015. **1296**: p. 187-98.

118. Romano, G., et al., *Small non-coding RNA and cancer.* Carcinogenesis, 2017. **38**(5): p. 485-491.

119. Watson, C.N., A. Belli, and V. Di Pietro, *Small Non-coding RNAs: New Class of Biomarkers and Potential Therapeutic Targets in Neurodegenerative Disease.* Front Genet, 2019. **10**: p. 364.

120. Bartel, D.P., *MicroRNAs: target recognition and regulatory functions.* Cell, 2009. **136**(2): p. 215-33.

121. Fabian, M.R., N. Sonenberg, and W. Filipowicz, *Regulation of mRNA translation and stability by microRNAs.* Annu Rev Biochem, 2010. **79**: p. 351-79.

122. Dragomir, M.P., E. Knutsen, and G.A. Calin, *SnapShot: Unconventional miRNA Functions.* Cell, 2018. **174**(4): p. 1038-1038 e1.

123. Ha, M. and V.N. Kim, *Regulation of microRNA biogenesis.* Nat Rev Mol Cell Biol, 2014. **15**(8): p. 509-24.

124. Di Leva, G., M. Garofalo, and C.M. Croce, *MicroRNAs in cancer.* Annu Rev Pathol, 2014. **9**: p. 287-314.

125. Romaine, S.P., et al., *MicroRNAs in cardiovascular disease: an introduction for clinicians.* Heart, 2015. **101**(12): p. 921-8.

126. Quinlan, S., et al., *MicroRNAs in Neurodegenerative Diseases.* Int Rev Cell Mol Biol, 2017. **334**: p. 309-343.

127. Okugawa, Y., Y. Toiyama, and A. Goel, *An update on microRNAs as colorectal cancer biomarkers: where are we and what's next?* Expert Rev Mol Diagn, 2014. **14**(8): p. 999-1021.

128. Ren, A., et al., *Detection of miRNA as non-invasive biomarkers of colorectal cancer.* Int J Mol Sci, 2015. **16**(2): p. 2810-23.

129. Vaira, V., et al., *microRNA profiles in coeliac patients distinguish different clinical phenotypes and are modulated by gliadin peptides in primary duodenal fibroblasts.* Clin Sci (Lond), 2014. **126**(6): p. 417-23.

130. Comincini, S., et al., *Identification of Autophagy-Related Genes and Their Regulatory miRNAs Associated with Celiac Disease in Children.* Int J Mol Sci, 2017. **18**(2).

131. Capuano, M., et al., *MicroRNA-449a overexpression, reduced NOTCH1 signals and scarce goblet cells characterize the small intestine of celiac patients.* PLoS One, 2011. **6**(12): p. e29094.

132. Magni, S., et al., *miRNAs affect the expression of innate and adaptive immunity proteins in celiac disease.* Am J Gastroenterol, 2014. **109**(10): p. 1662-74.

133. Bascunan, K.A., et al., *A miRNA-Based Blood and Mucosal Approach for Detecting and Monitoring Celiac Disease.* Dig Dis Sci, 2020. **65**(7): p. 1982-1991.

134. Amr, K.S., et al., *Circulating microRNAs as potential non-invasive biomarkers in pediatric patients with celiac disease.* Eur Ann Allergy Clin Immunol, 2019. **51**(4): p. 159-164.

135. Buoli Comani, G., et al., *miRNA-regulated gene expression differs in celiac disease patients according to the age of presentation.* Genes Nutr, 2015. **10**(5): p. 482.

136. Feng, Y., et al., *MicroRNAs, intestinal inflammatory and tumor.* Bioorg Med Chem Lett, 2019. **29**(16): p. 2051-2058.

137. Thorlacius-Ussing, G., et al., *Expression and Localization of miR-21 and miR-126 in Mucosal Tissue from Patients with Inflammatory Bowel Disease.* Inflamm Bowel Dis, 2017. **23**(5): p. 739-752.

138. James, J.P., et al., *MicroRNA Biomarkers in IBD-Differential Diagnosis and Prediction of Colitis-Associated Cancer.* Int J Mol Sci, 2020. **21**(21).

139. Rashid, H., et al., *Fecal MicroRNAs as Potential Biomarkers for Screening and Diagnosis of Intestinal Diseases.* Front Mol Biosci, 2020. **7**: p. 181.

140. Schonauen, K., et al., *Circulating and Fecal microRNAs as Biomarkers for Inflammatory Bowel Diseases.* Inflamm Bowel Dis, 2018. **24**(7): p. 1547-1557.

141. Ahadi, A., *The significance of microRNA deregulation in colorectal cancer development and the clinical uses as a diagnostic and prognostic biomarker and therapeutic agent.*

Noncoding RNA Res, 2020. **5**(3): p. 125-134.

142. Slaby, O. and G.A. Calin, *Non-coding RNAs in Colorectal Cancer Preface.* Non-Coding Rnas in Colorectal Cancer, 2016. **937**: p. V-V.

143. Hernandez, R., et al., *Downregulated microRNAs in the colorectal cancer: diagnostic and therapeutic perspectives.* BMB Rep, 2018. **51**(11): p. 563-571.

144. Liu, G. and B. Li, *Role of miRNA in transformation from normal tissue to colorectal adenoma and cancer.* J Cancer Res Ther, 2019. **15**(2): p. 278-285.

145. Rapado-Gonzalez, O., et al., *Circulating microRNAs as Promising Biomarkers in Colorectal Cancer.* Cancers (Basel), 2019. **11**(7).

146. Ragusa, M., et al., *Non-coding landscapes of colorectal cancer.* World J Gastroenterol, 2015. **21**(41): p. 11709-39.

147. Ahmed, F.E., et al., *Diagnostic microRNA markers for screening sporadic human colon cancer and active ulcerative colitis in stool and tissue.* Cancer Genomics Proteomics, 2009. **6**(5): p. 281-95.

148. Link, A., et al., *Fecal MicroRNAs as novel biomarkers for colon cancer screening.* Cancer Epidemiol Biomarkers Prev, 2010. **19**(7): p. 1766-74.

149. Wu, C.W., et al., *Detection of miR-92a and miR-21 in stool samples as potential screening biomarkers for colorectal cancer and polyps.* Gut, 2012. **61**(5): p. 739-45.

150. Yau, T.O., et al., *microRNA-221 and microRNA-18a identification in stool as potential biomarkers for the non-invasive diagnosis of colorectal carcinoma.* Br J Cancer, 2014. **111**(9): p. 1765-71.

151. Rotelli, M.T., et al., *Fecal microRNA profile in patients with colorectal carcinoma before and after curative surgery.* Int J Colorectal Dis, 2015. **30**(7): p. 891-8.

152. Chang, P.Y., et al., *MicroRNA-223 and microRNA-92a in stool and plasma samples act as complementary biomarkers to increase colorectal cancer detection.* Oncotarget, 2016. **7**(9): p. 10663-75.

153. Duran-Sanchon, S., et al., *Identification and Validation of MicroRNA Profiles in Fecal Samples for Detection of Colorectal Cancer.* Gastroenterology, 2020. **158**(4): p. 947-957 e4.

154. Siomi, M.C., et al., *PIWI-interacting small RNAs: the vanguard of genome defence.* Nat Rev Mol Cell Biol, 2011. **12**(4): p. 246-58.

155. Weick, E.M. and E.A. Miska, *piRNAs: from biogenesis to function.* Development, 2014. **141**(18): p. 3458-71.

156. Kiss, T., *Small nucleolar RNAs: An abundant group of noncoding RNAs with diverse cellular functions.* Cell, 2002. **109**(2): p. 145-148.

157. Mannoor, K., J. Liao, and F. Jiang, *Small nucleolar RNAs in cancer.* Biochim Biophys Acta, 2012. **1826**(1): p. 121-8.

158. Shen, Y., et al., *Transfer RNA-derived fragments and tRNA halves: biogenesis, biological functions and their roles in diseases.* J Mol Med (Berl), 2018. **96**(11): p. 1167-1176.

159. Zhu, L., et al., *tRNA-derived fragments and tRNA halves: The new players in cancers.* Cancer Lett, 2019. **452**: p. 31-37.

160. Lopez, J.P., et al., *Biomarker discovery: quantification of microRNAs and other small non-coding RNAs using next generation sequencing.* BMC Med Genomics, 2015. **8**: p. 35.

161. Cheng, J., et al., *piRNA, the new non-coding RNA, is aberrantly expressed in human cancer cells.* Clin Chim Acta, 2011. **412**(17-18): p. 1621-5.

162. Herrera, M., et al., *Differential distribution and enrichment of non-coding RNAs in exosomes from normal and Cancer-associated fibroblasts in colorectal cancer.* Mol Cancer, 2018. **17**(1): p. 114.

163. Mei, Y.P., et al., *Small nucleolar RNA 42 acts as an oncogene in lung tumorigenesis.* Oncogene, 2012. **31**(22): p. 2794-804.

164. Xiong, W., et al., *Identification of tRNAderived fragments in colon cancer by comprehensive small RNA sequencing.* Oncol Rep, 2019. **42**(2): p. 735-744.

165. Li, S., et al., *Angiogenin promotes colorectal cancer metastasis via tiRNA production.* Int J Cancer, 2019. **145**(5): p. 1395-1407.

166. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.

167. Qin, J., et al., *A human gut microbial gene catalogue established by metagenomic sequencing.* Nature, 2010. **464**(7285): p. 59-65.

168. Caporaso, J.G., et al., *QIIME allows analysis of high-throughput community sequencing data.* Nat Methods, 2010. **7**(5): p. 335-6.

169. Durazzi, F., et al., *Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota.* Sci Rep, 2021. **11**(1): p. 3030.

170. Bharti, R. and D.G. Grimm, *Current challenges and best-practice protocols for microbiome analysis.* Brief Bioinform, 2021. **22**(1): p. 178-193.

171. Heintz-Buschart, A. and P. Wilmes, *Human Gut Microbiome: Function Matters.* Trends Microbiol, 2018. **26**(7): p. 563-574.

172. Rooks, M.G. and W.S. Garrett, *Gut microbiota, metabolites and host immunity.* Nat Rev Immunol, 2016. **16**(6): p. 341-52.

173. Lakhan, S.E. and A. Kirchgessner, *Gut inflammation in chronic fatigue syndrome.* Nutr Metab (Lond), 2010. **7**: p. 79.

174. Kostic, A.D., et al., *Genomic analysis identifies association of Fusobacterium with colorectal carcinoma.* Genome Res, 2012. **22**(2): p. 292-8.

175. Fang, P., et al., *The Microbiome as a Modifier of Neurodegenerative Disease Risk.* Cell Host Microbe, 2020. **28**(2): p. 201-222.

176. Dabke, K., G. Hendrick, and S. Devkota, *The gut microbiome and metabolic syndrome.* J Clin Invest, 2019. **129**(10): p. 4050-4057.

177. Bascunan, K.A., et al., *Dietary Gluten as a Conditioning Factor of the Gut Microbiota in Celiac Disease.* Adv Nutr, 2020. **11**(1): p. 160-174.

178. Forsberg, G., et al., *Presence of bacteria and innate immunity of intestinal epithelium in childhood celiac disease.* Am J Gastroenterol, 2004. **99**(5): p. 894-904.

179. De Palma, G., et al., *Intestinal dysbiosis and reduced immunoglobulin-coated bacteria associated with coeliac disease in children.* BMC Microbiol, 2010. **10**: p. 63.

180. Nadal, I., et al., *Imbalance in the composition of the duodenal microbiota of children with coeliac disease.* J Med Microbiol, 2007. **56**(Pt 12): p. 1669-1674.

181. Caminero, A., et al., *Duodenal bacterial proteolytic activity determines sensitivity to dietary antigen through protease-activated receptor-2.* Nat Commun, 2019. **10**(1): p. 1198.

182. Bonder, M.J., et al., *The influence of a short-term gluten-free diet on the human gut microbiome.* Genome Med, 2016. **8**(1): p. 45.

183. Iaffaldano, L., et al., *Oropharyngeal microbiome evaluation highlights Neisseria abundance in active celiac patients.* Sci Rep, 2018. **8**(1): p. 11047.

184. Tian, N., et al., *Salivary Gluten Degradation and Oral Microbial Profiles in Healthy Individuals and Celiac Disease Patients.* Appl Environ Microbiol, 2017. **83**(6).

185. Sacchetti, L. and C. Nardelli, *Gut microbiome investigation in celiac disease: from methods to its pathogenetic role.* Clin Chem Lab Med, 2020. **58**(3): p. 340-349.

186. Glassner, K.L., B.P. Abraham, and E.M.M. Quigley, *The microbiome and inflammatory bowel disease.* J Allergy Clin Immunol, 2020. **145**(1): p. 16-27.

187. Bernstein, C.N. and J.D. Forbes, *Gut Microbiome in Inflammatory Bowel Disease and Other Chronic Immune-Mediated Inflammatory Diseases.* Inflamm Intest Dis, 2017. **2**(2): p. 116-123.

188. Burns, M.B., et al., *Colorectal cancer mutational profiles correlate with defined microbial communities in the tumor microenvironment.* PLoS Genet, 2018. **14**(6): p. e1007376.

189. Thomas, A.M., et al., *Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation.* Nat Med, 2019.

25(4): p. 667-678.

190. Gupta, A., R. Madani, and H. Mukhtar, *Streptococcus bovis endocarditis, a silent sign for colonic tumour.* Colorectal Dis, 2010. **12**(3): p. 164-71.

191. Boleij, A., et al., *Clinical Importance of Streptococcus gallolyticus infection among colorectal cancer patients: systematic review and meta-analysis.* Clin Infect Dis, 2011. **53**(9): p. 870-8.

192. Wang, T., et al., *Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers.* ISME J, 2012. **6**(2): p. 320-9.

193. Balamurugan, R., et al., *Real-time polymerase chain reaction quantification of specific butyrate-producing bacteria, Desulfovibrio and Enterococcus faecalis in the feces of patients with colorectal cancer.* J Gastroenterol Hepatol, 2008. **23**(8 Pt 1): p. 1298-303.

194. Wang, X. and M.M. Huycke, *Extracellular superoxide production by Enterococcus faecalis promotes chromosomal instability in mammalian cells.* Gastroenterology, 2007. **132**(2): p. 551-61.

195. Wang, X., et al., *Enterococcus faecalis induces aneuploidy and tetraploidy in colonic epithelial cells through a bystander effect.* Cancer Res, 2008. **68**(23): p. 9909-17.

196. Tsoi, H., et al., *Peptostreptococcus anaerobius Induces Intracellular Cholesterol Biosynthesis in Colon Cells to Induce Proliferation and Causes Dysplasia in Mice.* Gastroenterology, 2017. **152**(6): p. 1419-1433 e5.

197. Long, X., et al., *Peptostreptococcus anaerobius promotes colorectal carcinogenesis and modulates tumour immunity.* Nat Microbiol, 2019. **4**(12): p. 2319-2330.

198. Castellarin, M., et al., *Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma.* Genome Res, 2012. **22**(2): p. 299-306.

199. Bashir, A., et al., *Fusobacterium nucleatum: an emerging bug in colorectal tumorigenesis.* Eur J Cancer Prev, 2015. **24**(5): p. 373-85.

200. Yang, Y., et al., *Fusobacterium nucleatum Increases Proliferation of Colorectal Cancer Cells and Tumor Development in Mice by Activating Toll-Like Receptor 4 Signaling to Nuclear Factor-kappaB, and Up-regulating Expression of MicroRNA-21.* Gastroenterology, 2017. **152**(4): p. 851-866 e24.

201. Yu, T., et al., *Fusobacterium nucleatum Promotes Chemoresistance to Colorectal Cancer by Modulating Autophagy.* Cell, 2017. **170**(3): p. 548-563 e16.

202. Veziant, J., et al., *Association of colorectal cancer with pathogenic Escherichia coli: Focus on mechanisms using optical imaging.* World J Clin Oncol, 2016. **7**(3): p. 293-301.

203. Cougnoux, A., et al., *Bacterial genotoxin colibactin promotes colon tumour growth by inducing a senescence-associated secretory phenotype.* Gut, 2014. **63**(12): p. 1932-42.

204. Dong, J., J.W. Tai, and L.F. Lu, *miRNA-Microbiota Interaction in Gut Homeostasis and Colorectal Cancer.* Trends Cancer, 2019. **5**(11): p. 666-669.

205. Cooks, T., et al., *Mutant p53 cancers reprogram macrophages to tumor supporting macrophages via exosomal miR-1246.* Nat Commun, 2018. **9**(1): p. 771.

206. Ajouz, H., D. Mukherji, and A. Shamseddine, *Secondary bile acids: an underrecognized cause of colon cancer.* World J Surg Oncol, 2014. **12**: p. 164.

207. Zhu, Q.D., et al., *MiR-199a-5p Inhibits the Growth and Metastasis of Colorectal Cancer Cells by Targeting ROCK1.* Technol Cancer Res Treat, 2018. **17**: p. 1533034618775509.

208. Teng, Y., et al., *Plant-Derived Exosomal MicroRNAs Shape the Gut Microbiota.* Cell Host Microbe, 2018. **24**(5): p. 637-652 e8.

209. Mohan, M., et al., *Dietary Gluten-Induced Gut Dysbiosis Is Accompanied by Selective Upregulation of microRNAs with Intestinal Tight Junction and Bacteria-Binding Motifs in Rhesus Macaque Model of Celiac Disease.* Nutrients, 2016. **8**(11).

210. Riboli, E., et al., *European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection.* Public Health Nutr, 2002. **5**(6B): p. 1113-24.

211. Ferrero, G., et al., *Small non-coding RNA profiling in human biofluids and surrogate tissues*

from healthy individuals: description of the diverse and most represented species. Oncotarget, 2018. **9**(3): p. 3097-3111.

212. Tarallo, S., et al., *Altered Fecal Small RNA Profiles in Colorectal Cancer Reflect Gut Microbiome Composition in Stool Samples.* mSystems, 2019. **4**(5).

213. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

214. Zhang, P., et al., *MicroRNA-143a-3p modulates preadipocyte proliferation and differentiation by targeting MAPK7.* Biomed Pharmacother, 2018. **108**: p. 531-539.

215. Franzosa, E.A., et al., *Species-level functional profiling of metagenomes and metatranscriptomes.* Nat Methods, 2018. **15**(11): p. 962-968.

216. Wirbel, J., et al., *Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox.* Genome Biol, 2021. **22**(1): p. 93.

217. Singh, A., et al., *DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays.* Bioinformatics, 2019. **35**(17): p. 3055-3062.

218. Witten, I.H., et al., *Data mining : practical machine learning tools and techniques*. Fourth edition. ed. 2017, Amsterdam: Morgan Kaufmann. xxxii, 621 pages.

219. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.

220. Huang, W., *MicroRNAs: Biomarkers, Diagnostics, and Therapeutics.* Methods Mol Biol, 2017. **1617**: p. 57-67.

221. Morris, N.L. and M.A. Choudhry, *Maintenance of gut barrier integrity after injury: Trust your gut microRNAs.* J Leukoc Biol, 2021.

222. Pardini, B., et al., *MicroRNAs as markers of progression in cervical cancer: a systematic review.* BMC Cancer, 2018. **18**(1): p. 696.

223. Gomez Castro, M.F., et al., *p31-43 Gliadin Peptide Forms Oligomers and Induces NLRP3 Inflammasome/Caspase 1- Dependent Mucosal Damage in Small Intestine.* Front Immunol, 2019. **10**: p. 31.

224. Tarallo, S., et al., *MicroRNA expression in relation to different dietary habits: a comparison in stool and plasma samples.* Mutagenesis, 2014. **29**(5): p. 385-91.

225. Lee, M., et al., *An association between crypt apoptotic bodies and mucosal flattening in celiac disease patients exposed to dietary gluten.* Diagn Pathol, 2019. **14**(1): p. 98.

226. Shalimar, D.M., et al., *Mechanism of villous atrophy in celiac disease: role of apoptosis and epithelial regeneration.* Arch Pathol Lab Med, 2013. **137**(9): p. 1262-9.

227. Pardini, B., et al., *Noncoding RNAs in Extracellular Fluids as Cancer Biomarkers: The New Frontier of Liquid Biopsies.* Cancers (Basel), 2019. **11**(8).

228. Caio, G., et al., *Effect of Gluten-Free Diet on Gut Microbiota Composition in Patients with Celiac Disease and Non-Celiac Gluten/Wheat Sensitivity.* Nutrients, 2020. **12**(6).

229. Pecora, F., et al., *Gut Microbiota in Celiac Disease: Is There Any Role for Probiotics?* Front Immunol, 2020. **11**: p. 957.

230. Laparra, J.M., et al., *Bifidobacterium longum CECT 7347 modulates immune responses in a gliadin-induced enteropathy animal model.* PLoS One, 2012. **7**(2): p. e30744.

231. Olivares, M., M. Laparra, and Y. Sanz, *Influence of Bifidobacterium longum CECT 7347 and gliadin peptides on intestinal epithelial cell proteome.* J Agric Food Chem, 2011. **59**(14): p. 7666-71.

232. Francavilla, R., et al., *Salivary microbiota and metabolome associated with celiac disease.* Appl Environ Microbiol, 2014. **80**(11): p. 3416-25.

233. Esquela-Kerscher, A. and F.J. Slack, *Oncomirs - microRNAs with a role in cancer.* Nat Rev Cancer, 2006. **6**(4): p. 259-69.

234. Vila-Navarro, E., et al., *MicroRNAs for Detection of Pancreatic Neoplasia: Biomarker Discovery by Next-generation Sequencing and Validation in 2 Independent Cohorts.* Ann Surg, 2017. **265**(6): p. 1226-1234.

235. Wohnhaas, C.T., et al., *Fecal MicroRNAs Show Promise as Noninvasive Crohn's Disease Biomarkers.* Crohns Colitis 360, 2020. **2**(1): p. otaa003.

236. Verdier, J., et al., *Faecal Micro-RNAs in Inflammatory Bowel Diseases.* J Crohns Colitis, 2020. **14**(1): p. 110-117.

237. Li, X.L., et al., *P53 mutations in colorectal cancer - molecular pathogenesis and pharmacological reactivation.* World J Gastroenterol, 2015. **21**(1): p. 84-93.

238. Wang, S.W. and Y.M. Sun, *The IL-6/JAK/STAT3 pathway: potential therapeutic strategies in treating colorectal cancer (Review).* Int J Oncol, 2014. **44**(4): p. 1032-40.

239. Desmond, B.J., E.R. Dennett, and K.M. Danielson, *Circulating Extracellular Vesicle MicroRNA as Diagnostic Biomarkers in Early Colorectal Cancer-A Review.* Cancers (Basel), 2019. **12**(1).

240. Jin, Y., et al., *MiR-182-5p inhibited proliferation and metastasis of colorectal cancer by targeting MTDH.* Eur Rev Med Pharmacol Sci, 2019. **23**(4): p. 1494-1501.

241. Gaedcke, J., et al., *The rectal cancer microRNAome--microRNA expression in rectal cancer and matched normal mucosa.* Clin Cancer Res, 2012. **18**(18): p. 4919-30.

242. Sayagues, J.M., et al., *Genomic characterization of liver metastases from colorectal cancer patients.* Oncotarget, 2016. **7**(45): p. 72908-72922.

243. Rounge, T.B., et al., *Circulating small non-coding RNAs associated with age, sex, smoking, body mass and physical activity.* Sci Rep, 2018. **8**(1): p. 17650.

244. Santoru, M.L., et al., *Cross sectional evaluation of the gut-microbiome metabolome axis in an Italian cohort of IBD patients.* Sci Rep, 2017. **7**(1): p. 9523.

245. Dai, Z., et al., *Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers.* Microbiome, 2018. **6**(1): p. 70.

246. Wassenaar, T.M., *E. coli and colorectal cancer: a complex relationship that deserves a critical mindset.* Crit Rev Microbiol, 2018. **44**(5): p. 619-632.

247. Vogtmann, E., et al., *Colorectal Cancer and the Human Gut Microbiome: Reproducibility with Whole-Genome Shotgun Sequencing.* PLoS One, 2016. **11**(5): p. e0155362.

248. Agramonte-Hevia, J., et al., *Gram-negative bacteria and phagocytic cell interaction mediated by complement receptor 3.* FEMS Immunol Med Microbiol, 2002. **34**(4): p. 255-66.

249. Miller, S.I., R.K. Ernst, and M.W. Bader, *LPS, TLR4 and infectious disease diversity.* Nat Rev Microbiol, 2005. **3**(1): p. 36-46.

250. Neal, M.D., et al., *Enterocyte TLR4 mediates phagocytosis and translocation of bacteria across the intestinal barrier.* J Immunol, 2006. **176**(5): p. 3070-9.

251. Giuffrida, P. and A. Di Sabatino, *MicroRNAs in Celiac Disease Diagnosis: A miR Curiosity or Game-Changer?* Dig Dis Sci, 2020. **65**(7): p. 1877-1879.

252. Roszkowska, A., et al., *Non-Celiac Gluten Sensitivity: A Review.* Medicina (Kaunas), 2019. **55**(6).

7 PUBLICATIONS

List of scientific publications from the candidate (in bold those related to the PhD work)

Published:

- **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). pii: e00289-19. doi: 10.1128/mSystems.00289-19.**

- **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. Combined metagenomic analysis of colorectal cancer datasets defines 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.**

- **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 March;31(1):30-42 DOI: 10.23736/S1120-4826.18.02495-3.**

- Pardini B, De Maria D, **Francavilla A**, Di Gaetano C, Ronco G, Naccarati A. MicroRNAs as markers of progression in cervical cancer: a systematic review. BMC Cancer. 2018 Jun 27;18(1):696. doi: 10.1186/s12885-018-4590-4.

- **Francavilla A, Turoczi S, Tarallo S, Vodicka P, Pardini B, Naccarati A. Exosomal microRNAs and other non-coding RNAs as colorectal cancer biomarkers: a review. Mutagenesis. 2019 Nov 30. pii: gez038. doi: 10.1093/mutage/gez038.**

- Wang S, Romanak K, Tarallo S, **Francavilla A,** Viviani M, Vineis P, Rothwell J, Mancini FR, Cordero F, Naccarati A, Severi G, Venier M. The use of silicone wristbands to evaluate personal exposure to semivolatile organic chemicals (SVOCs) in France and Italy. Environmental Pollution. 2020 Aug 21, 0269-7491, doi.org/10.1016/j.envpol.2020.115490.

In preparation:

- Tarallo S, Ferrero G, De Filippis F, **Francavilla A**, Pasolli E, Panero V, Cordero F, Segata N, Grioni S, Pensa R, Pardini B, Ercolini D, Naccarati A, Stool microRNA profiles reflect different dietary and gut microbiome patterns in healthy individuals" under revision.

- **Pardini B, Ferrero G, Tarallo S, Gaetano G, Francavilla A, Licheri N, Trompetto M, Clerico G, Senore C, Peyre S, Vymetalkova V, Vodickova L, Liska V, Vycital O, M Levy**

M, Macinga P, Hucl T, Vodicka P, Cordero F, Naccarati A. Small non-coding-RNA profiling in two independent cohorts and in multiple biospecimens identifies a fecal miRNA predictive model to accurately distinguish colorectal cancer and adenomas.

- Francavilla A, Gagliardi A, Piaggeschi G, Tarallo S, Cordero F, Pensa R, Impeduglia A, Caviglia G, Ribaldone D, Gallo G, Grioni S, Pardini B, Ferrero G, Naccarati A. Specific microRNA profiles in fecal samples are related to age, sex, body mass index and lifestyle habits in healthy individuals.

- **Francavilla A, Ferrero G, Pardini B, Tarallo S, Bruno M, Caviglia G, Cordero F, Vineis P, Ribaldone D, Naccarati A. sncRNAs and gut microbiome profiling in stool of subjects affected by celiac disease.**