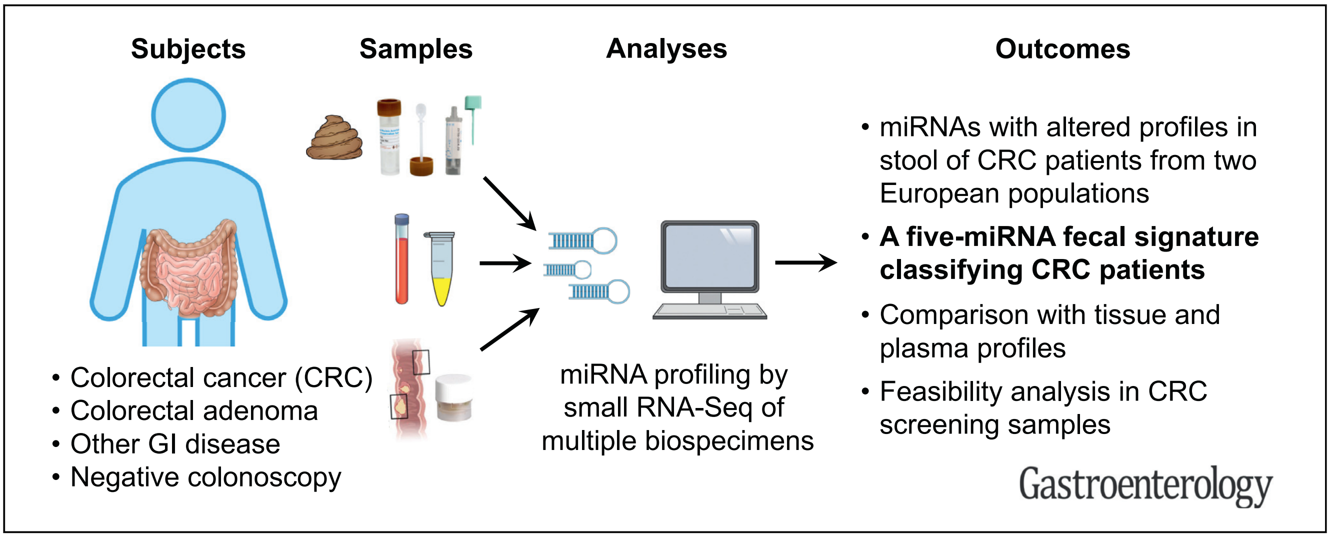
Gastroenterology 2023;165:582–599

A Fecal MicroRNA Signature by Small RNA Sequencing Accurately Distinguishes Colorectal Cancers: Results From a Multicenter Study

Barbara Pardini,[1,2,*] **Giulio Ferrero**,[3,4,*] **Sonia Tarallo**,[1,2,*] Gaetano Gallo,[5,6] Antonio Francavilla,[1] Nicola Licheri,[4] Mario Trompetto,[6] Giuseppe Clerico,[6] Carlo Senore,[7] Sergio Peyre,[8] Veronika Vymetalkova,[9,10,11] Ludmila Vodickova,[9,10,11] Vaclav Liska,[11,12] Ondrej Vycital,[11,12] Miroslav Levy,[13] Peter Macinga,[14] Tomas Hucl,[14] Eva Budinska,[15] Pavel Vodicka,[9,10,11] **Francesca Cordero**,[4,§] and **Alessio Naccarati**[1,2,§]

[1]Italian Institute for Genomic Medicine, Turin, Italy; [2]Candiolo Cancer Institute, FPO-IRCCS, Turin, Italy; [3]Department of Clinical and Biological Sciences, University of Turin, Turin, Italy; [4]Department of Computer Science, University of Turin, Turin, Italy; [5]Department of Surgery, Sapienza University of Rome, Rome, Italy; [6]Department of Colorectal Surgery, Clinica S. Rita, Vercelli, Italy; [7]Epidemiology and Screening Unit-CPO, University Hospital Città della Salute e della Scienza, Turin, Italy; [8]LILT (Lega Italiana Lotta contro i Tumori), associazione provinciale di Biella, Biella, Italy; [9]Department of Molecular Biology of Cancer, Institute of Experimental Medicine of the Czech Academy of Sciences, Prague, Czech Republic; [10]Institute of Biology and Medical Genetics, 1st Medical Faculty, Charles University, Prague, Czech Republic; [11]Biomedical Center, Faculty of Medicine in Pilsen, Charles University, Pilsen, Czech Republic; [12]Department of Surgery, University Hospital and Faculty of Medicine in Pilsen, Charles University, Pilsen, Czech Republic; [13]Department of Surgery, First Faculty of Medicine, Charles University and Thomayer Hospital, Prague, Czech Republic; [14]Department of Gastroenterology and Hepatology, Institute for Clinical and Experimental Medicine, Prague, Czech Republic; and [15]RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic



- Colorectal cancer (CRC)
- Colorectal adenoma
- Other GI disease
- Negative colonoscopy

miRNA profiling by small RNA-Seq of multiple biospecimens

- miRNAs with altered profiles in stool of CRC patients from two European populations
- **A five-miRNA fecal signature classifying CRC patients**
- Comparison with tissue and plasma profiles
- Feasibility analysis in CRC screening samples

BACKGROUND & AIMS: Fecal tests currently used for colorectal cancer (CRC) screening show limited accuracy in detecting early tumors or precancerous lesions. In this respect, we comprehensively evaluated stool microRNA (miRNA) profiles as biomarkers for noninvasive CRC diagnosis. **METHODS:** A total of 1273 small RNA sequencing experiments were performed in multiple biospecimens. In a cross-sectional study, miRNA profiles were investigated in fecal samples from an Italian and a Czech cohort (155 CRCs, 87 adenomas, 96 other intestinal diseases, 141 colonoscopy-negative controls). A predictive miRNA signature for cancer detection was defined by a machine learning strategy and tested in additional fecal samples from 141 CRC patients and 80 healthy volunteers. miRNA profiles were compared with those of 132 tumors/adenomas paired with adjacent mucosa, 210 plasma extracellular vesicle samples, and 185 fecal immunochemical test leftover samples. **RESULTS:** Twenty-five miRNAs showed altered levels in the stool of CRC patients in both cohorts (adjusted $P < .05$). A 5-miRNA signature, including miR-149-3p, miR-607-5p, miR-1246, miR-4488, and miR-6777-5p, distinguished patients from control individuals (area under the curve [AUC], 0.86; 95% confidence interval [CI], 0.79–0.94) and was validated in an independent cohort (AUC, 0.96; 95% CI, 0.92–1.00). The signature classified control individuals from patients with low-/high-stage tumors and advanced adenomas (AUC, 0.82; 95% CI, 0.71–0.97). Tissue miRNA profiles mirrored those of stool samples, and fecal profiles of different gastrointestinal diseases highlighted miRNAs specifically dysregulated in CRC. miRNA profiles in fecal immunochemical test leftover samples showed good correlation with those of stool collected in preservative buffer, and their alterations could be detected in adenoma or CRC patients. **CONCLUSIONS:** Our comprehensive fecal miRNome analysis identified a signature accurately discriminating cancer aimed at improving noninvasive diagnosis and screening strategies.

Keywords: Stool MicroRNAs; Noninvasive Diagnosis; Small RNA Sequencing; Colorectal Cancer; Precancerous Lesions; Machine Learning.

In the last 30 years, we have witnessed a dramatic increase in understanding the epidemiology, etiology, molecular biology, and various clinical aspects of colorectal cancer (CRC).[1] However, approximately 1.8 million new cases are annually diagnosed worldwide, posing CRC as the third most common incident cancer. Moreover, although early-stage tumors can be efficiently treated, CRC is still the second-leading cause of cancer-related death, with 900,000 deaths in 2018.[2,3] Hence, the early detection of preclinical cancers or precursor lesions is a desirable objective, because it may strongly increase the chances for successful treatment and cure.

Most European countries have implemented CRC screening programs based on noninvasive stool tests for detecting fecal occult blood, mainly the fecal immunochemical test (FIT).[4,5] FIT selects individuals showing a higher prevalence of CRC and advanced benign neoplasia but has limited sensitivity to recognize advanced colorectal adenomas (AAs).[6] Colonoscopy is also used in an opportunistic screening setting and detects both cancer and premalignant lesions but is bothersome and invasive, as well as costly for the health system.[7] Despite the fact that FIT-based screening programs are undeniably efficient in detecting premalignant growths and providing an earlier diagnosis, successfully reducing CRC burden, only approximately 5% of individuals who receive a colonoscopy based on FIT results will end up with a significant lesion (CRC or AA). Stool tests show a relatively low specificity, resulting in a high number of false positives and a considerable number of unnecessary colonoscopies.[8] Complementing traditional screening stool tests with other noninvasively detectable fecal molecular biomarkers could improve the triage of individual for colonoscopy, reducing the costs for the health systems in terms of the number of examinations and decreasing the risks and discomfort for patients.[9,10]

Identifying reliable biomarkers is not trivial, given the ensemble of hidden interactions between molecules and patient-specific clinical/anamnestic characteristics. However, machine learning (ML) algorithms have been defined to reveal significant features able to accurately discriminate groups of individuals. In particular, explainable ML approaches allow the identification of novel molecular biomarker signatures to improve early CRC diagnosis, as recently demonstrated for fecal microbial species[11] and urinary proteins.[12]

WHAT YOU NEED TO KNOW

BACKGROUND AND CONTEXT

Current screening programs for the noninvasive detection of colorectal cancer (CRC) are based on fecal tests with limited accuracy for early malignancies or precancerous lesions. Evaluating microRNA (miRNA) profiles in stool could improve the screening strategy.

NEW FINDINGS

Investigating the whole miRNome in stool and with ad hoc explainable machine learning, we identified in 2 independent cohorts 5 miRNAs that could accurately classify CRC patients from control individuals. The signature was validated in a third cohort and assayed in fecal immunochemical test leftover samples from the screening.

LIMITATIONS

Despite the large number of samples overall collected and sequenced, the disease subtypes investigated were still not exhaustive of the heterogeneity in CRC and adenomas. Although we showed the feasibility of the molecular analysis, the investigation on screening samples still represents a pilot approach.

CLINICAL RESEARCH RELEVANCE

The investigation of the whole miRNome in all of the cohorts led to a comprehensive overview of the fecal miRNA profiles, providing the possibility to accurately single out those signals that may enhance the accuracy of the screening. The identified miRNA signature accurately discriminates different stages of CRC development, and it constitutes a coadjuvant to current screening programs for a noninvasive, accurate diagnosis.

BASIC RESEARCH RELEVANCE

New and previously reported miRNAs altered in CRC are detectable in stool and may highlight a novel role of these molecules released in the gut in physiologic and pathologic conditions.

The analysis of small noncoding RNAs in fecal samples has attracted interest with an excellent biological and analytic rationale for its application in large-scale clinical investigations.[13] Tumor-secreted small noncoding RNAs are directly and continuously released into the intestinal lumen,

* Authors share co-first authorship; § Authors share co-senior authorship.

Abbreviations used in this paper: AA, advanced colorectal adenoma; AUC, area under the curve; BMI, body mass index; CRC, colorectal cancer; CZ, Czech; DEmiRNA, differentially expressed microRNA; EV, extracellular vesicle; FIT, fecal immunochemical test; GI, gastrointestinal; IBD, inflammatory bowel disease; IT, Italian; miRNA, microRNA; ML, machine learning; nAA, nonadvanced adenoma; qPCR, quantitative polymerase chain reaction; RNA-seq, RNA sequencing; RT, reverse transcription.

Most current article

0016-5085
https://doi.org/10.1053/j.gastro.2023.05.037

and their profiles may be altered in concomitance with the presence of CRC and precancerous lesions. Moreover, small noncoding RNAs, such as microRNAs (miRNAs), are remarkably stable, enabling their accurate detection in stool without the need for special stabilization or logistic requirements.[14] miRNAs are suitable biomarkers in surrogate tissues and biofluids because their levels are altered in specific pathologic states,[15] in the presence of precursor lesions,[16] and in CRC development.[17–19] In addition, specific fecal miRNA alterations have been associated with the gut microbiome composition[20] and proposed as noninvasive CRC biomarkers.[21]

So far, comprehensive miRNA profiling by small RNA sequencing (small RNA-seq) has been mainly performed on tumor tissues or plasma.[21,22] In contrast, studies on fecal samples investigated few miRNAs in relation to CRC, typically in small cohorts and without taking into account their demographic characteristics.[23] In this respect, studies on the whole fecal miRNome showed that different lifestyles and dietary habits might critically affect specific miRNA levels.[24,25] In addition, limited evidence is available on stool miRNA profiles in relation to patient clinicopathologic characteristics, such as specific CRC stages, precancerous lesions or other gastrointestinal (GI) diseases, except for the reported pleiotropic dysregulation of miR-21-5p in several diseases.[26] Therefore, an miRNA signature for CRC detection derived from a comprehensive fecal miRNome analysis across multiple populations is currently lacking.

This multicenter study aimed to explore, by deep sequencing, the miRNA profiles in stool samples that best characterize CRC patients from control individuals and distinguish colorectal adenomas or other GI diseases. The analyses were performed in different independent cohorts adopting the same protocol for participant recruitment, sample collection, and small RNA-seq experiments/analyses. An integrated explainable ML strategy identified a fecal miRNA signature distinguishing CRC patients from control individuals, and the results were validated in an additional cohort. Finally, altered miRNAs in stool were also investigated in FIT-positive leftover samples collected within a population-based CRC screening program.

Methods

Stool Study Cohorts

Italian cohort. Stool specimens as well as clinical and demographic data were collected from 317 individuals recruited in a hospital-based study in Vercelli, Italy (Table 1). Based on the results of complete colonoscopy examination, participants were classified into (1) 89 sporadic CRC patients, (2) 74 polyp patients (6 hyperplastic polyps, 20 nonadvanced adenomas [nAAs] and 48 AAs; serrated lesions were excluded because there were too few), (3) 49 individuals with a GI disease (6 Crohn's disease, 9 ulcerative colitis, 14 diverticulitis, 7 diverticulosis, 13 hemorrhoidal disease), and (4) 105 colonoscopy-negative control individuals. AAs were defined based on the presence of high-grade dysplasia, villous component, or lesion size of >1 cm as defined by Zarchy and Ershoff.[27] Of this cohort, 93 stool samples (from 29 CRC patients, 27 polyps, 13 patients with a GI disease, and 24 colonoscopy-negative control individuals) have been used and described previously in other studies.[11,28,29]

Czech cohort. Stool specimens as well as clinical and demographic data were collected from 162 Czech individuals recruited in 2 hospitals in Prague and 1 in Plzen, Czech Republic (Table 1). Based on colonoscopy results, participants were divided in (1) 66 CRC patients, (2) 28 polyp patients (9 hyperplastic polyps, 13 nAAs, 6 AAs; no serrated lesions were collected), (3) 32 patients with other GI diseases (3 Crohn's disease, 11 ulcerative colitis, 17 diverticulosis, 1 unclassified inflammatory bowel disease [IBD]); and (4) 36 colonoscopy-negative individuals.

Validation cohort. Stool specimens from 141 CRC patients recruited in a hospital in Brno, Czech Republic, and 80 stool samples of healthy volunteers contributing to science were included. These participants were previously described in other studies: the CRC population is described by Zwinsova et al[30] but here is sequenced for the first time for small RNA-seq; healthy volunteers are a part of the cohorts described and sequenced for small noncoding RNAs by Tarallo et al[24] and Francavilla et al.[31]

Fecal immunochemical test cohort. FIT buffer leftover samples from 185 individuals with a positive test result were collected within the CRC screening for the Piedmont Region (Italy). Based on colonoscopy results, participants were classified as control individuals ($n = 53$), AA ($n = 80$), nAA ($n = 30$), or CRC ($n = 22$). Among them, 57 individuals also provided stool samples before undergoing colonoscopy.

More details on the cohorts included in the study are given in the Supplementary Materials. The local ethics committees of Azienda Ospedaliera SS. Antonio e Biagio e C. Arrigo of Alessandria (Italy, protocol no. Colorectal miRNA CEC2014), AOU Città della salute e della Scienza di Torino (Italy), the Institute of Experimental Medicine of Prague (Czech Republic), Masaryk Memorial Cancer Institute (protocol no. 2018/865/MOU), and Masaryk University of Brno (Czech Republic, protocol no. EKV-2019-044) approved the study. All patients gave written informed consent following the Declaration of Helsinki before participating in the study.

Other Analyzed Biospecimens

For 132 patients having surgery at the Vercelli hospital, primary tissues (102 CRC and 30 adenomas) paired with adjacent colonic mucosa were collected.

Blood samples were collected from 210 out of 317 Italian (IT) cohort participants, stratified into patients with CRC ($n = 52$), AAs ($n = 19$), nAAs ($n = 15$), hyperplastic polyps ($n = 6$), and other GI diseases ($n = 39$), and control individuals ($n = 79$).

Sample Collection

Naturally evacuated fecal samples were obtained from participants previously instructed to self-collect the specimen at home. Samples were collected in nucleic acid collection and transport tubes with RNA stabilizing solution (Norgen Biotek Corp). Stool aliquots (200 μL) were stored at –80°C until RNA extraction.[20] For the validation cohort of CRC patients from Brno, stool samples were collected from untreated patients before the scheduled surgery with DNA-free swabs (Deltalab). Patients performed the collection at home and returned the samples to the hospital, where they were immediately frozen at –80°C until further processing.

Table 1. Study Population Characteristics

	IT cohort (n = 317)					CZ cohort (n = 162)				
Covariate	Controls (n = 105)	Other GI disease (n = 49)	Polyps (n = 74)	CRC (n = 89)	*P* value	Controls (n = 36)	Other GI disease (n = 32)	Polyps (n = 28)	CRC (n = 66)	*P* value
Age, *y*										
Average ± SD	59.6 ± 10.7	56.7 ± 13.6	66.2 ± 9.1	70.6 ± 9.7	7.34E–13	57.8 ± 10.5	58.7 ± 9.4	63.1 ± 8.4	68.0 ± 11.2	8.34E–06
Range	39–84	30–82	42–93	50–88		40–76	41–75	48–82	40–88	
Sex, n										
Male	52	23	41	52	4.83E–01	14	16	14	46	1.74E–02
Female	53	26	33	37		22	16	14	20	
BMI, kg/m^2										
Average	25.3 ± 4.5	25.0 ± 3.4	25.0 ± 3.7	25.8 ± 5.1	9.02E–01	28.2 ± 6.1	28.8 ± 7.0	29.0 ± 3.5	27.1 ± 5.4	1.61E–01
Range	15.4–40.0	19.5–33.7	19.5–36.0	16.0–44.1		21.0–43.9	22.0–60.9	22.6–34.7	16.9–47.6	
Smoking status, n										
Nonsmoker	31	17	18	35	2.16E–01	25	24	13	32	2.53E–02
Ex-smoker	16	6	20	15		3	0	8	12	
Smoker	38	12	22	31		8	8	6	18	
NA	20	14	14	7		0	0	1	4	
Localization, n[a]										
Proximal			19	37				16	16	
Distal			11	20				11	15	
Rectum			18	28				6	34	
NA			32	6				0	1	
Polyp type, n										
Tubular adenoma			18					19		
Tubulovillous adenoma			12					0		
Tubular sessile			5					0		
Hyperplastic polyp			6					9		
NA			31					0		
Adenoma type, n										
AA			48					6		
nAA			20					13		
pT (combined), n										
T1–T2				27					20	
T3–T4				54					43	
Tis				0					1	
NA				7					2	

Table 1. Continued

	IT cohort (n = 317)					CZ cohort (n = 162)				
Covariate	Controls (n = 105)	Other GI disease (n = 49)	Polyps (n = 74)	CRC (n = 89)	*P* value	Controls (n = 36)	Other GI disease (n = 32)	Polyps (n = 28)	CRC (n = 66)	*P* value
AJCC staging, n										
I				18					16	
II				24					16	
III				29					15	
IV				5					14	
NA				13					5	
Grade, n										
G1–G2				39					44	
G3				38					18	
NA				12					4	
Metastasis (lymph node or distal), n										
No				49					52	
Yes				31					11	
NA				9					3	
Other GI diseases, n										
Crohn’s disease		6					3			
Ulcerative rectocolitis		9					11			
Diverticulosis		7					17			
Diverticulitis		14					0			
Hemorrhoidal disease		13								
NA		0					1			

AJCC, American Joint Committee on Cancer; NA, not available; pT, post-operatory tumor size; SD, standard deviation.
[a]Totals may be different from the total number of individuals in each category because of the presence of multiple lesions.

For the FIT cohort, leftovers from FIT tubes (~1.2 mL) used for automated tests (OC-sensor, Eiken Chemical Co) for hemoglobin quantification were stored at –80°C until use.

Plasma samples were obtained from 8 mL of blood centrifuged for 10 minutes at 1000 revolutions/minute, and aliquots were stored at –80°C until use. Plasma extracellular vesicles (EVs) were precipitated from 200 μL of plasma using ExoQuick (System Biosciences) according to Sabo et al.[32]

Paired tumor/adenoma tissue and adjacent nonmalignant mucosa (at least 20 cm distant) were obtained from CRC and adenoma patients during surgical resection and immediately immersed in RNAlater solution (Ambion). All samples were stored at –80°C until use.

Total RNA Extraction, Small RNA Sequencing Library Preparation, and Quantitative Real-Time Polymerase Chain Reaction

Total RNA from stool and FIT leftover samples was extracted using the Stool Total RNA Purification Kit (Norgen Biotek Corp) as previously described.[20] Total RNA from plasma EVs was extracted as described in Sabo et al.[32] For tissue samples, total RNA was extracted using QIAzol (Qiagen) according to the manufacturer's instructions.

Small RNAs were converted into barcoded complementary DNA libraries for Illumina single-end sequencing (75 cycles on HiSeq4000 or NextSeq500, Illumina Inc) as previously described.[24]

Candidate miRNA biomarkers were replicated in stool samples using the miRCURY LNA miRNA PCR Assays (Qiagen). Reverse transcription (RT) was performed using the miRCURY LNA RT kit (Qiagen) according to the manufacturer's instructions. All reactions were run on an ABI Prism 7900 Sequence Detection System (Applied Biosystems). Analyses were performed as described by Moisoiu et al.[33] More details are provided in the Supplementary Materials.

Computational and Statistical Analyses

Small RNA-seq analyses were performed as described by Tarallo et al,[20] considering a curated miRNA reference based on miRBase v22 and including a characterization of novel miRNAs (Supplementary Table 1*A*). Differential expression analyses were performed using DESeq2 v1.22.2.[34] Functional enrichment analysis was performed with RBiomirGS v0.2.12,[35] considering the validated miRNA-target interactions. A generalized linear model was defined by considering the miRNA levels as the dependent variable and participant age, sex, body mass index (BMI), smoking habit, and cohort as independent variables.

An ML strategy was implemented to identify the optimal fecal miRNA signature to accurately classify CRC patients from control individuals. The ML approach is composed of 3 phases: data preparation, feature selection, and classification. (More details are provided in the Supplementary Materials.) The signature was determined by considering an increasing number of miRNAs prioritized by filter and classifier-embedded methods applied to the training set (70% of the IT/Czech [CZ] cohorts). The optimal set of miRNAs providing the highest area under the curve (AUC) was selected and further tested by 100 stratified 10-fold cross-validations, first on the remaining 30% of the IT/CZ cohorts excluded from the training set and then on the validation cohort. The training and test sets were defined by a stratified selection to maintain the same proportion of participants characterized by specific covariates (ie, age, sex, cohort, disease status, and tumor staging).

Other statistical tests were performed using the Wilcoxon-Mann-Whitney and Kruskal-Wallis (continuous variables) or chi-square (categorical variables) methods. The Benjamini-Hochberg method was used for multiple-testing correction. Results were considered significant at $P < .05$.

Study Design

This study was designed to define and characterize a fecal miRNA signature that accurately distinguishes CRC patients from control individuals (Figure 1). The applied analysis strategy included the following phases.

Fecal miRNome profiling and biomarker discovery.

- Detection of stool miRNAs with altered levels in CRC: miRNA profiles from small RNA-seq and metadata were used for a differential expression analysis between CRC patients and control individuals of both the IT cohort and CZ cohort, independently. The overlapping differentially expressed miRNAs (DEmiRNAs) from both cohorts were the input of the next step.

- Feature selection and definition of an miRNA predictive signature: An ML strategy identified an miRNA signature composed of the minimal set of DEmiRNAs that better distinguished CRC patients from control individuals by a stratified cross-validation procedure.

- Validation of the miRNA predictive signature. The signature performance was estimated in the validation cohort by a stratified cross-validation procedure.

Fecal differentially expressed microRNA characterization in different sample types and diseases.

- Assessment of DEmiRNA profiles in different biospecimens and clinical situations: DEmiRNA levels were evaluated in (1) tumor/adenoma tissue and adjacent mucosa, (2) plasma EVs of CRC patients and control individual, and (3) fecal samples from patients with a GI disease or precancerous lesions to identify CRC-specific or commonly altered miRNAs. In particular, the miRNA signature from (1) was also tested in the discrimination of patients with precancerous lesions (AA or nAA), alone or in combination with CRC, from control individuals.

- Testing the DEmiRNA levels in samples from a CRC screening program: DEmiRNA profiles were explored in parallel in FIT buffer leftovers and in stool collected in tubes with RNA stabilizing solution. Subsequently, stool DEmiRNA levels were analyzed in the leftover samples of the FIT cohort by stratifying participants based on the colonoscopy results.

A detailed description of the methods is provided in the Supplementary Materials.

Results

Stool MicroRNA Profiles Are Altered in Colorectal Cancer Patients: Evidence From 2 European Populations

In agreement with previous studies,[20,24,31] an average of 479 (range, 86–1516) miRNAs were detected in each stool sample by small RNA-seq (further details in the Supplementary Materials and Supplementary Table 1*B* and *C*). The age- and sex-adjusted differential expression analysis between CRC patients and control individuals was performed independently on both the IT cohort and CZ cohort identifying, respectively, 250 and 29 DEmiRNAs (median expression, >20 reads; adjusted $P < .05$) (Figure 2*A* and Supplementary Table 2*A*).

Twenty-five stool DEmiRNAs were in common between both cohorts (Figure 2*B*, Table 2, and Supplementary Table 2*A*), all with a coherent expression trend (20 up-regulated and 5 down-regulated; rho = 0.75; $P < .001$) (Figure 2*B*). The alteration of these fecal miRNA levels in relation to CRC was further supported by a generalized linear model analysis adjusted for cohort, age, sex, BMI, and smoking habits: 22 out of the 25 DEmiRNAs remained significantly associated ($P < .05$) (Supplementary Table 2*B*). DEmiRNA profiles were further explored in relation to CRC patient clinical data (Figure 2*C* and *D*). The levels of 3 down-regulated miRNAs (miR-607-5p, miR-677-5p, and miR-922-5p) significantly decreased with increasing tumor size (Figure 2*D*). miR-922-5p also significantly decreased in patients with advanced disease stages or lymph node invasion (Figure 2*D* and Supplementary Table 2*C*). Conversely, increasing levels of 19 out of the 20 up-regulated miRNAs in CRC were observed along with tumor size, with miR-1246, miR-1290, miR-148-3p, and miR-194-5p significantly related to this parameter. The levels of 11 CRC–up-regulated miRNAs significantly increased in patients with lymph node invasion. In addition, the levels of 11 miRNAs were significantly higher in samples from patients with rectal compared to colon cancers (Figure 2*D*).

Functional analysis of DEmiRNA target genes showed their involvement in cancer-related processes, including cell cycle regulation and DNA repair, particularly for up-regulated miRNA targets (Supplementary Table 2*D* and *E*).

A Fecal MicroRNA Signature Distinguishes Colorectal Cancer Patients From Control Individuals

An explainable ML strategy was implemented to identify the minimal set of miRNAs as a signature for CRC detection (Supplementary Figure 1 and Supplementary Materials). The pipeline was applied on the 25 DEmiRNA profiles and considering 70% of the IT cohort and CZ cohort as the training set (Supplementary Table 3*A*). The best miRNA signature distinguishing CRC patients from control individuals included miR-607-5p, miR-6777-5p, miR-4488, miR-149-3p, and miR-1246 (AUC, 0.87 ± 0.01) (Figure 2*E*). This set of 5 miRNAs represented the best combination of noncorrelated molecules with the highest discriminative power. Moreover, they showed a good performance in the classification of the 30% of participants excluded from the training set (AUC, 0.81 ± 0.01) (Figure 2*F*). The classification improved after the inclusion of sex and age in the model (AUC, 0.86 ± 0.01) (Table 3 and Supplementary Table 3*B*). The performance of the signature was again tested in the validation cohort, where it remained fairly similar, irrespective (AUC, 0.91 ± 0.01) or not (AUC, 0.96 ± 0.01) of age and sex (Figure 2*F*, Table 3, and Supplementary Table 3*B*).

By stratifying patients for CRC stage, the same 5-miRNA signature accurately distinguished patients with stages III–IV CRC (validation cohort: AUC, 0.96 ± 0.01 and 0.94 ± 0.01, respectively, including or not age and sex), or CRC stages I–II from control individuals (validation cohort: AUC, 0.95 ± 0.01 and 0.87 ± 0.01, respectively, including or not age and sex) (Table 3 and Supplementary Table 3*B*).

The panel of 5 miRNAs of the signature identified by sequencing was tested by RT quantitative polymerase chain reaction (qPCR) in RNA isolated from a subset of 96 stool samples equally distributed among IT and CZ cohort participants, with a balanced number of CRC patients and control individuals (Supplementary Figure 2*A*). The 5 miRNAs were detected in all samples, also using this second method. The normalized levels from RT-qPCR showed patterns comparable to those provided by sequencing, except for miR-4488 (Supplementary Figure 2*A*). In particular, miR-1246 and miR-149-3p levels were significantly increased in patient samples. The same method was used to test the 5 miRNA levels in RNA from 8 FIT leftover samples of participants with a positive FIT result at the CRC screening: all miRNAs were also detected in this biospecimen (data not shown).

For 4 signature miRNAs, a concordant expression pattern was observed between small RNA-seq and RT-qPCR normalized levels, particularly for miR-1246 (rho = 0.63, $P < .001$) and miR-149-3p (rho = 0.26, $P < .05$) (Supplementary Table 3*C* and Supplementary Figure 2*B*). Only the levels of miR-4488 were characterized by a negative correlation (rho = –0.48, $P < .001$) in CRC patients only.

Stool Differentially Expressed MicroRNA Profiles Mirror Those of Primary Colorectal Cancer and Adenoma Tissues

A paired differential expression analysis was performed between tumor tissues and matched adjacent mucosa collected from 102 CRC patients. Among the 25 stool DEmiRNAs, 14 were differentially expressed (adjusted $P < .05$) in this comparison (Figure 3*A* and Supplementary Table 4*A*), with 7 miRNAs (miR-21-5p, miR-1246, miR-1290, miR-148a-3p, miR-4488, miR-149-3p, miR-12114) up-regulated in tumor tissues coherently with their increase in CRC patient stool. Among them, 3 (miR-1246, miR-4488, miR-149-3p) were included in our miRNA signature. The 5 miRNAs significantly down-regulated in CRC patient stool (miR-607-5p, miR-6777-5p, included in the 5-miRNA signature; miR-6076; miR-922-5p; and miR-9899) were poorly expressed (normalized reads, <20) in both tumor and adjacent tissues (Supplementary Table 4*A*).

The differential analysis performed on 30 adenoma tissues matched with adjacent mucosa showed miR-21-5p, miR-1290, miR-148a-3p, and miR-200b-3p as significantly up-regulated in adenoma tissues (adjusted $P < .001$), whereas let-7i-5p and miR-4508 were down-regulated (Figure 3*A* and Supplementary Table 4*A*).

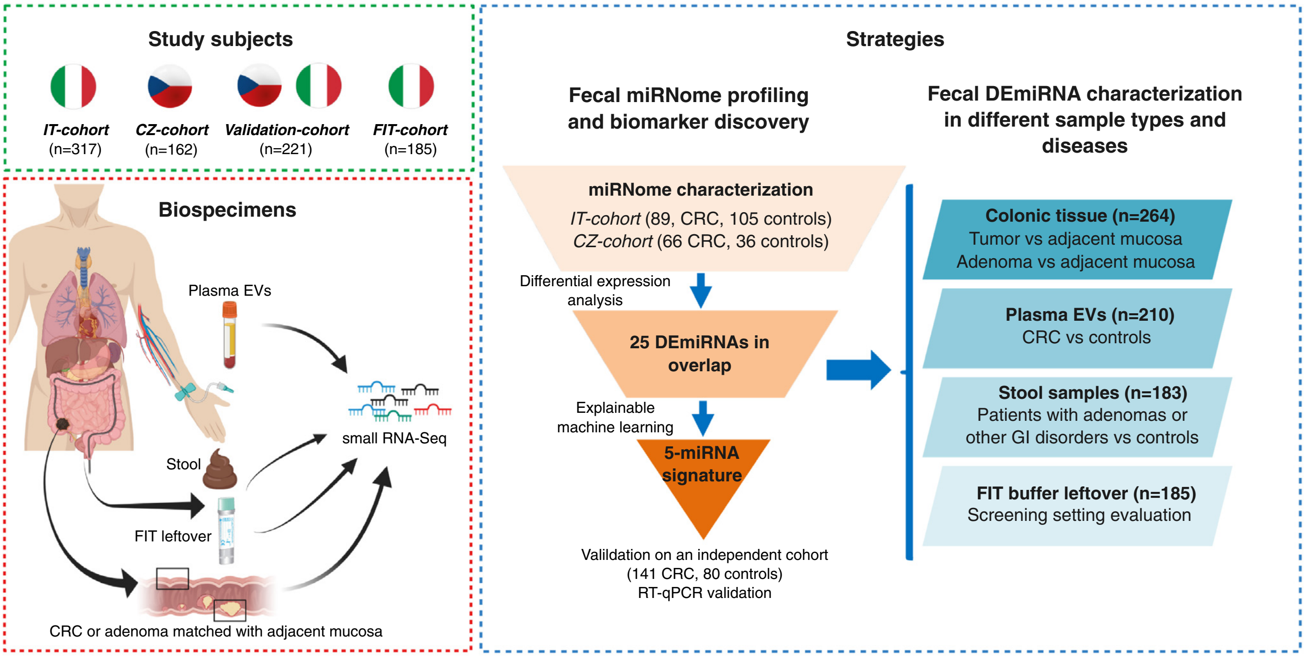

Study subjects
IT-cohort (n=317)
CZ-cohort (n=162)
Validation-cohort (n=221)
FIT-cohort (n=185)
Biospecimens
Plasma EVs
Stool
FIT leftover
small RNA-Seq
CRC or adenoma matched with adjacent mucosa
Strategies
Fecal miRNome profiling and biomarker discovery
miRNome characterization
IT-cohort (89, CRC, 105 controls)
CZ-cohort (66 CRC, 36 controls)
Differential expression analysis
25 DEmiRNAs in overlap
Explainable machine learning
5-miRNA signature
Valildation on an independent cohort (141 CRC, 80 controls) RT-qPCR validation
Fecal DEmiRNA characterization in different sample types and diseases
Colonic tissue (n=264)
Tumor vs adjacent mucosa
Adenoma vs adjacent mucosa
Plasma EVs (n=210)
CRC vs controls
Stool samples (n=183)
Patients with adenomas or other GI disorders vs controls
FIT buffer leftover (n=185)
Screening setting evaluation

Figure 1. Representation of the study design.

Few MicroRNA Levels Are Dysregulated in Circulating Extracellular Vesicles of Colorectal Cancer Patients

Small RNA-seq was performed on RNA isolated from plasma EVs collected from 210 participants in the IT cohort, detecting an average of 309 (range, 252–1213) miRNAs in these samples (Supplementary Table 4*B*). Among the 25 DEmiRNAs identified in stool samples of CRC patients, both miR-1246 and miR-4488 emerged as coherently significantly dysregulated in plasma EVs, although the latter was associated with low levels (normalized reads, <20) (Supplementary Table 4*B*). Another miRNA (miR-150-5p) was differentially expressed between CRC patients and control individuals (Supplementary Table 4*B*).

A Subset of Stool Differentially Expressed MicroRNAs Is Specifically Dysregulated in Colorectal Cancer Patients but Not in Those With Other GI Diseases

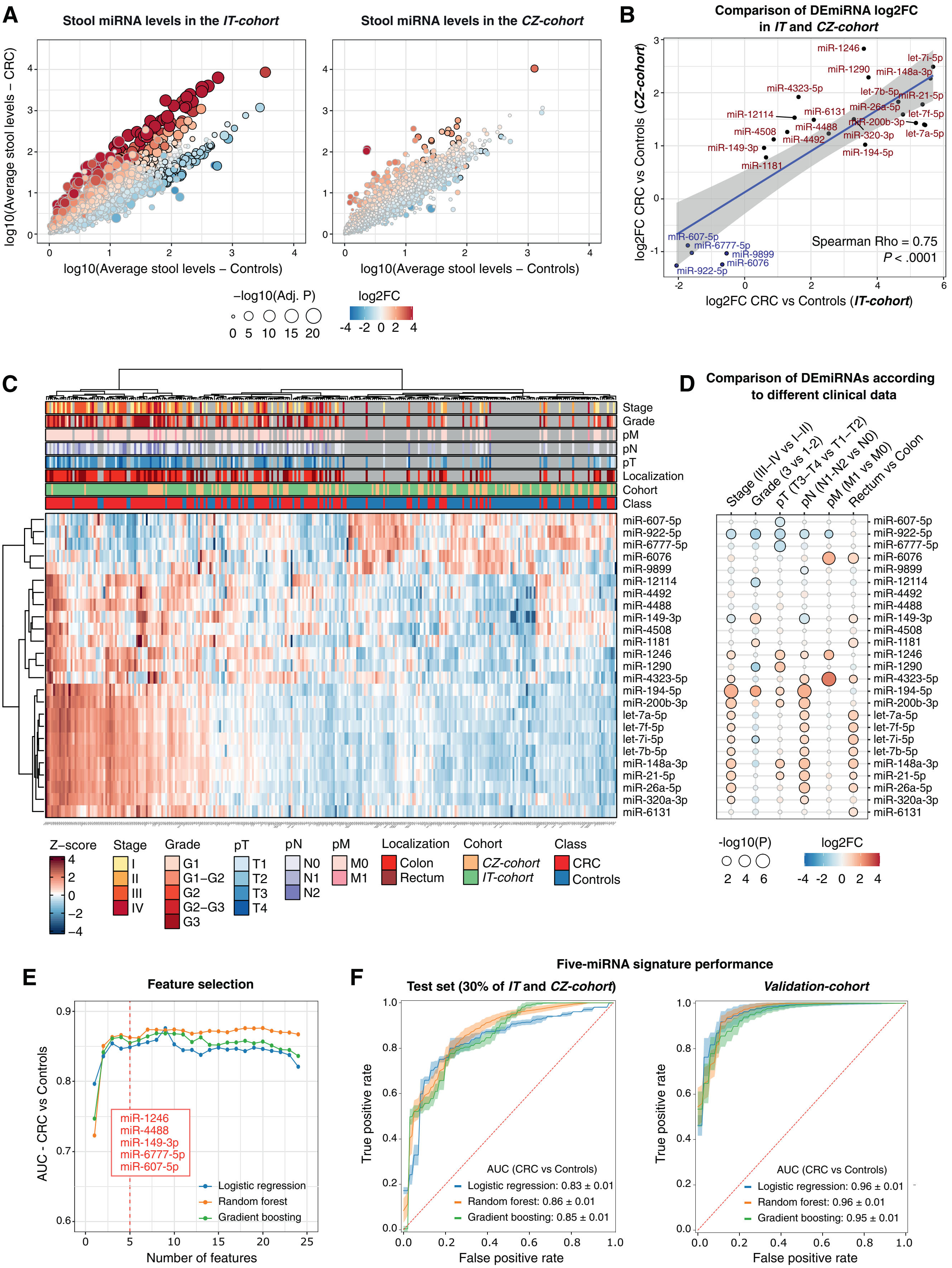
The CRC DEmiRNAs were further compared with those from patients with GI disorders and other precancerous lesions in both the IT and CZ cohorts. The age-, sex-, and cohort-adjusted differential expression analysis between each disease category and control individuals showed that the levels of 21 out of the 25 CRC DEmiRNAs were significantly altered in at least another GI disease (Figure 3*B*). Notably, in patients with ulcerative colitis, diverticulitis, nAA, or AA, 60% of the CRC DEmiRNAs were also dysregulated (Figure 3*B* and Supplementary Table 4*C*). The lowest number of dysregulated miRNAs was observed in patients with Crohn's disease (2 miRNAs) or diverticulosis (5 miRNAs), whereas no DEmiRNAs were found in patients with hyperplastic polyps.

Considering the 5 miRNAs constituting our predictive signature to distinguish CRC patients from control individuals, miR-6777-5p was not differentially expressed (compared to control individuals) in any other GI disease, miR-149-3p was significantly up-regulated only in patients with AA, and miR-607-5p was significantly down-regulated in patients with AA or ulcerative colitis compared to control individuals (Figure 3*B* and Supplementary Table 4*C*). Conversely, miR-4488 and miR-1246 stool levels significantly increased in patients with diverticulosis, ulcerative colitis, diverticulitis, or AA, with the latter miRNA also increased in Crohn's disease patients.

The identified signature was also used to classify AA and nAA patients from control individuals. Specifically, the miRNA signature was able to distinguish AA from control participants, both including (AUC, 0.82 ± 0.01) or not (AUC, 0.77 ± 0.02) age and sex in the analysis, as well as nAA (AUC, 0.80 ± 0.03 and 0.77 ± 0.02, respectively, including or not age and sex). Finally, patients with either CRC or AA were accurately distinguished from control individuals (including or not age and sex: AUC, 0.84 ± 0.01 and 0.81 ± 0.01, respectively) but not between them (CRC vs AA: AUC, 0.68 ± 0.02) (Table 3 and Supplementary Table 3*B*).

MicroRNAs Are Detectable in Fecal Immunochemical Test Leftover Samples by Small RNA Sequencing

The sequencing analysis was extended to 185 available leftover samples of the FIT cohort, still detecting an average of 618 miRNAs in each sample (Supplementary Table 1*B*). All of the 25 stool DEmiRNAs were detected in this type of sample. Considering the threshold adopted by our pipeline (ie, a minimum of 20 reads), 4 (miR-607-5p, miR-1246, let-


A
Stool miRNA levels in the IT-cohort
Stool miRNA levels in the CZ-cohort
log10(Average stool levels – CRC)
log10(Average stool levels – Controls)
0
1
2
3
4
−log10(Adj. P)
0 5 10 15 20
log2FC
-4 -2 0 2 4
B
Comparison of DEmiRNA log2FC in IT and CZ-cohort
log2FC CRC vs Controls (CZ-cohort)
log2FC CRC vs Controls (IT-cohort)
-2
0
2
4
6
-1
3
miR-1246
let-7i-5p
miR-1290
miR-148a-3p
miR-4323-5p
let-7b-5p
miR-21-5p
miR-26a-5p
miR-12114
miR-6131
let-7f-5p
miR-200b-3p
miR-4488
miR-4508
miR-320-3p
let-7a-5p
miR-4492
miR-149-3p
miR-194-5p
miR-1181
miR-607-5p
miR-6777-5p
miR-9899
miR-6076
miR-922-5p
Spearman Rho = 0.75
P < .0001
C
Stage
Grade
pM
pN
pT
Localization
Cohort
Class
miR-607-5p
miR-922-5p
miR-6777-5p
miR-6076
miR-9899
miR-12114
miR-4492
miR-4488
miR-149-3p
miR-4508
miR-1181
miR-1246
miR-1290
miR-4323-5p
miR-194-5p
miR-200b-3p
let-7a-5p
let-7f-5p
let-7i-5p
let-7b-5p
miR-148a-3p
miR-21-5p
miR-26a-5p
miR-320a-3p
miR-6131
Z-score
4
2
0
−2
−4
Stage
I
II
III
IV
Grade
G1
G1−G2
G2
G2−G3
G3
pT
T1
T2
T3
T4
pN
N0
N1
N2
pM
M0
M1
Localization
Colon
Rectum
Cohort
CZ-cohort
IT-cohort
Class
CRC
Controls
D
Comparison of DEmiRNAs according to different clinical data
Stage (III–IV vs I–II)
Grade (3 vs 1-2)
pT (T3–T4 vs T1–T2)
pN (N1-N2 vs N0)
pM (M1 vs M0)
Rectum vs Colon
miR-607-5p
miR-922-5p
miR-6777-5p
miR-6076
miR-9899
miR-12114
miR-4492
miR-4488
miR-149-3p
miR-4508
miR-1181
miR-1246
miR-1290
miR-4323-5p
miR-194-5p
miR-200b-3p
let-7a-5p
let-7f-5p
let-7i-5p
let-7b-5p
miR-148a-3p
miR-21-5p
miR-26a-5p
miR-320a-3p
miR-6131
-log10(P)
2 4 6
log2FC
-4 -2 0 2 4
E
Feature selection
AUC - CRC vs Controls
Number of features
0
5
10
15
20
25
0.6
0.7
0.8
0.9
miR-1246
miR-4488
miR-149-3p
miR-6777-5p
miR-607-5p
Logistic regression
Random forest
Gradient boosting
F
Five-miRNA signature performance
Test set (30% of IT and CZ-cohort)
Validation-cohort
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC (CRC vs Controls)
Logistic regression: 0.83 ± 0.01
Random forest: 0.86 ± 0.01
Gradient boosting: 0.85 ± 0.01
AUC (CRC vs Controls)
Logistic regression: 0.96 ± 0.01
Random forest: 0.96 ± 0.01
Gradient boosting: 0.95 ± 0.01

7a-3p, miR-922) were detected in all samples, and 18 were detected in more than half (Figure 3*C* and Supplementary Table 4*D*). Three miRNAs included in our signature (miR-607-5p, miR-1246, miR-6777-5p) were detected in more than 95% of samples (Figure 3*C*), whereas miR-149-3p and miR-4488 were detected in 112 (57.4%) and 57 (30.8%) samples, respectively.

Then, miRNA levels in FIT cohort samples were explored by stratifying participants according to the colonoscopy results. Comparing the levels of the 25 stool DEmiRNAs between 46 participants with a negative colonoscopy result (excluding 7 participants with high hemoglobin levels) and 22 patients with CRC, 8 (let-7a-5p, let-7i-5p, miR-148a-3p, let-7b-5p, miR-320a-3p, miR-12114, miR-21-5p, miR-607-5p) were significantly different (adjusted $P < .05$) (Supplementary Table 4*E* and Figure 3*C*). Correlating the miRNA levels in FIT leftovers with the hemoglobin levels, only let-7b-5p showed a significant but limited correlation (rho = 0.16, $P < .05$) (Supplementary Table 4*F*).

Interestingly, miR-1246 and miR-607-5p were characterized, respectively, by increasing and decreasing levels, from colonoscopy-negative participants to CRC patients, as observed in the stool of the 3 case-control cohorts initially investigated for the miRNA signature identification (Figure 3*D*).

Comparable miRNA expression levels and variability were observed between paired FIT leftover/stool samples from 57 individuals analyzed by small RNA-seq (rho = 0.70, $P < .001$) (Supplementary Table 1*B* and Supplementary Figure 2*C*). Considering the levels of 468 miRNAs detected in at least half of FIT leftover samples, 99.6% were coherent with those in stool, with 282 miRNAs significantly correlated (average rho = 0.39, $P < .05$) (Figure 3*C*, Supplementary Figure 2*C*, and Supplementary Table 4*D*). In both sample types, miR-3125-3p, miR-6075-5p, and miR-1246 were characterized by the highest levels, and miR-3125-3p was detected in all samples and associated with the lowest expression variability, in agreement with our previous findings in stool samples of 335 control individuals[25] (Supplementary Figure 3*A* and Supplementary Table 4*D*). The levels of all 25 stool DEmiRNAs positively correlated between the 2 specimens, with 13 of them reaching statistical significance (including miR-607-5p, miR-1246, miR-149-3p, and miR-4488 from the 5-miRNA signature; $P < .05$) (Figure 3*C* and Supplementary Table 4*D*).

The 5-miRNA signature analyzed in FIT buffer leftovers was finally tested for the classification of patients with CRC from control individuals considering the signature alone or in combination with patient age, sex, and FIT hemoglobin levels. The 5-miRNA signature alone showed comparable classification performance (AUC, 0.85) as using age, sex, and hemoglobin levels (AUC, 0.87), and the combination of both data provided the best classification results (AUC, 0.93) (Supplementary Table 3*D*).

Discussion

In the present study, to our knowledge, we performed the first large-scale profiling of the stool miRNome by deep sequencing of samples from patients with CRC, colorectal polyps, or other GI diseases and control individuals. Given the pervasive detection across multiple cohorts, we confirmed previous findings about fecal miRNA potential use as noninvasive molecular biomarkers[23] (Supplementary Table 1*C* and Supplementary Figure 3*A*). We also reported novel evidence on specific markers across different disease conditions. Notably, a fecal miRNA signature was able to accurately distinguish CRC patients from control individuals: both its ability to distinguish AA and its detection in FIT leftovers support future investigations for a use in CRC screening implementation.

In CRC patients, 25 fecal miRNAs emerged coherently altered in 2 independent cohorts. The profile of these miRNAs in stool reflected their altered expression in tumor tissue or adjacent colonic mucosa. More than half of such DEmiRNAs were already reported as altered in CRC, either in tissue or in various biofluids, including the up-regulated miR-21-5p, miR-148a-3p, miR-149-3p, miR-194-5p, miR-200b-3p, and miR-320a-3p (Supplementary Table 5*A*).[23,36] Other miRNAs were associated with a disease for the first time by us; thus, further in vitro studies are needed to characterize the functional activity of these molecules and their involvement in CRC. Moreover, 3 DEmiRNAs identified in our study (miR-4323-5p, miR-607-5p, and miR-922-5p) are not currently annotated in the miRbase but were quantified based on the read mapping position within the miRNA hairpin. This is consistent with the need for

Figure 2. (*A*) Scatterplot reporting the stool miRNA average levels in CRC patients (*y*-axis) or control individuals (*x*-axis) from the IT cohort (*left*) or CZ cohort (*right*). The dot color represents the log2 fold change (log2FC) from the differential expression analyses between CRC and healthy individuals, and the size is proportional to the age, sex, and multiple-testing adjusted *P* values. (*B*) Scatterplot reporting the correlations of log2FC of the 25 DEmiRNAs from the comparison between CRC and control individuals and in common between the IT cohort (*x*-axis) and the CZ cohort (*y*-axis). The up-regulated and down-regulated miRNAs are reported in red and blue, respectively. (*C*) Heatmap of stool DEmiRNA levels in CRC and control individuals of both cohorts. For each participant, the CRC stage and grade based on the American Joint Committee on Cancer system, presence of metastasis, lymph node invasion status (pN), tumor size (pT), tumor localization, cohort of origin, and disease status (CRC or control) are reported. (*D*) DEmiRNA levels comparing CRC patients stratified for clinical data. The dot color represents the log2FC, and the dot size is proportional to the statistical significance. Black borders represent tests with *P* < .05. (*E*) Line plot reporting the ability of different combinations of feature selection methods and classifiers to perform the classification of CRC and control individuals. Each dot represents an AUC obtained using a different number of fecal DEmiRNAs in input. (*F*) Receiver operating characteristic curves obtained for the classification of CRC and control individuals using the identified miRNA signature. Data are reported for the 30% of participants excluded from the training set (*left*) and for the validation cohort (*right*). Adj., adjusted.

Table 2. Expression Levels and Fold Changes of the 25 Stool DEmiRNAs in Common Between the IT and CZ Cohorts

				Median levels, controls		Median levels, CRC		log2FC		Benjamini-Hochberg adjusted *P* value[a]	
ID	miRNA gene ID	Chromosome	Genomic context	IT cohort	CZ cohort	IT cohort	CZ cohort	IT cohort	CZ cohort	IT cohort	CZ cohort
let-7a-5p	MIRLET7A3	chr22	Intergenic	52.18	28.12	717.25	50.53	5.44	1.39	2.51E–24	1.05E–02
let-7b-5p	MIRLET7B	chr22	Intergenic	20.19	12.94	474.50	26.28	4.63	1.83	3.04E–19	6.54E–03
let-7f-5p	MIRLET7F1/MIRLET7F2	chr9/chrX	Intergenic/intron (*HUWE1*)	54.93	33.83	513.72	38.72	5.41	1.40	2.27E–27	1.05E–02
let-7i-5p	MIRLET7I	chr12	Partial overlap (*LINC01465*)	16.75	10.68	577.93	27.38	5.68	2.49	1.25E–23	6.54E–04
miR-1181	MIR1181	chr19	Exon (*CDC37*)	72.46	38.12	83.60	65.61	0.64	0.78	1.12E–02	4.63E–02
miR-12114	MIR12114	chr22	Intron (*PPP6R2*)	126.48	43.52	266.97	67.67	1.50	1.53	1.06E–07	4.71E–03
miR-1246	MIR1246	chr2	Intron (*LINC01117*)	909.33	568.34	2970.91	2364.91	3.59	2.83	9.63E–17	3.98E–06
miR-1290	MIR1290	chr1	Intron (*ALDH4A1*)	46.70	33.77	231.36	82.25	3.73	2.29	1.71E–21	4.13E–04
miR-148a-3p	MIR148A	chr7	Intergenic	19.17	11.56	425.27	25.82	5.60	2.27	4.19E–22	1.92E–03
miR-149-3p	MIR149	chr2	Intron (*GPC1*)	30.82	16.15	34.55	36.97	0.58	0.96	1.89E–02	3.92E–02
miR-194-5p	MIR194-1 / MIR194-2	chr1 / chr11	Intron (*IARS2*)/intergenic	69.85	59.45	206.31	68.59	3.63	1.02	3.44E–20	2.38E–02
miR-200b-3p	MIR200B	chr1	Intergenic	22.03	20.39	204.93	23.29	5.16	1.43	2.85E–23	2.01E–02
miR-21-5p	MIR21	chr17	Exon (*VMP1*)	37.68	42.23	557.19	63.56	5.36	1.78	1.15E–22	1.22E–02
miR-26a-5p	MIR26A1 / MIR26A2	chr3 / chr12	Intron (*CTDSPL*)/intron (*CTDSPL2*)	36.78	33.23	425.88	44.01	4.77	1.59	2.85E–23	1.68E–02
miR-320a-3p	MIR320A	chr8	Intergenic	27.26	16.26	271.19	33.93	3.29	1.50	1.01E–15	5.33E–03
miR-4323-5p	MIR4323	chr19	Intron (POU2F2-AS1)	67.11	29.50	73.39	58.96	1.62	1.92	8.88E–07	5.12E–03
miR-4488	MIR4499	chr11	Intergenic	113.12	50.73	342.90	73.67	2.53	1.23	2.94E–19	2.91E–02
miR-4492	MIR4492	chr11	Exon/intron (*BCL9L*)	25.04	14.50	34.76	22.24	1.28	1.26	1.62E–06	7.47E–03
miR-4508	MIR4508	chr15	Intergenic	94.44	34.09	98.33	86.36	0.87	1.12	3.85E–04	2.56E–02
miR-607-5p	MIR607	chr10	Intergenic	222.53	132.30	51.44	87.13	–1.72	–0.88	2.17E–18	6.54E–03
miR-6076	MIR6076	chr14	Intron (*LINC01588*)	32.14	23.14	15.10	15.54	–0.68	–1.24	1.05E–02	1.83E–02
miR-6131	MIR6131	chr5	Intergenic	31.05	15.50	103.66	22.39	2.08	1.49	2.19E–12	3.31E–03
miR-6777-5p	MIR6777	chr17	Intron (*SREBF1*)	235.14	140.02	42.53	80.22	–1.60	–1.02	4.60E–08	1.29E–02
miR-922-5p	MIR922	chr3	Exon (*RUBCN*)	335.74	206.43	71.51	89.57	–2.06	–1.26	1.99E–11	3.92E–02
miR-9899	MIR9899	chr2	Intron (*LYPD6*)	71.25	50.86	33.99	26.40	–0.55	–1.03	1.09E–02	4.00E–02

chr, chromosome; ID, identifier; log2FC, log2 fold change.

[a] Age- and sex-adjusted analysis.

Table 3. Performance of the 5-miRNA Predictive Signature in the Different Comparisons

Analysis details[a]							Precision		F1 score	
Comparison	Validation set	AUC (Mean ± SD)	95% CI	Accuracy	Sensitivity	Specificity	Disease	Control	Disease	Control
CRC vs control individuals	IT cohort + CZ cohort[b]	0.86 ± 0.01	0.79–0.94	0.78	0.78	0.78	0.82	0.74	0.80	0.76
CRC vs control individuals	Validation cohort	0.96 ± 0.01	0.92–1.00	0.89	0.90	0.88	0.93	0.83	0.91	0.85
Stage I–II CRC vs control individuals	IT cohort + CZ cohort[b]	0.86 ± 0.01	0.76–0.96	0.81	0.65	0.90	0.79	0.82	0.71	0.86
Stage I–II CRC vs control individuals	Validation cohort	0.95 ± 0.01	0.90–1.00	0.86	0.82	0.91	0.90	0.83	0.86	0.87
Stage III–IV CRC vs control individuals	IT cohort + CZ cohort[b]	0.88 ± 0.01	0.78–0.98	0.83	0.66	0.92	0.82	0.83	0.73	0.88
Stage III–IV CRC vs control individuals	Validation cohort	0.96 ± 0.01	0.91–1.00	0.85	0.75	0.94	0.91	0.82	0.82	0.88
CRC + AA vs control individuals	IT cohort + CZ cohort[b]	0.84 ± 0.01	0.77–0.91	0.77	0.83	0.67	0.81	0.70	0.81	0.69
AA vs control individuals	IT cohort + CZ cohort[b]	0.82 ± 0.01	0.71–0.97	0.79	0.61	0.86	0.62	0.85	0.62	0.85
AA + nAA vs control individuals	IT cohort + CZ cohort[b]	0.77 ± 0.02	0.65–0.89	0.73	0.62	0.81	0.67	0.77	0.64	0.79
nAA vs control individuals	IT cohort + CZ cohort[b]	0.80 ± 0.01	0.63–0.97	0.82	0.13	0.99	0.79	0.82	0.22	0.90
CRC vs AA	IT cohort + CZ cohort[b]	0.68 ± 0.02	0.54–0.82	0.76	0.92	0.25	0.80	0.49	0.85	0.33

[a] Analysis includes age and sex covariates.
[b] Thirty percent of samples were excluded from the training and matched by age, sex, cohort, and CRC stage.

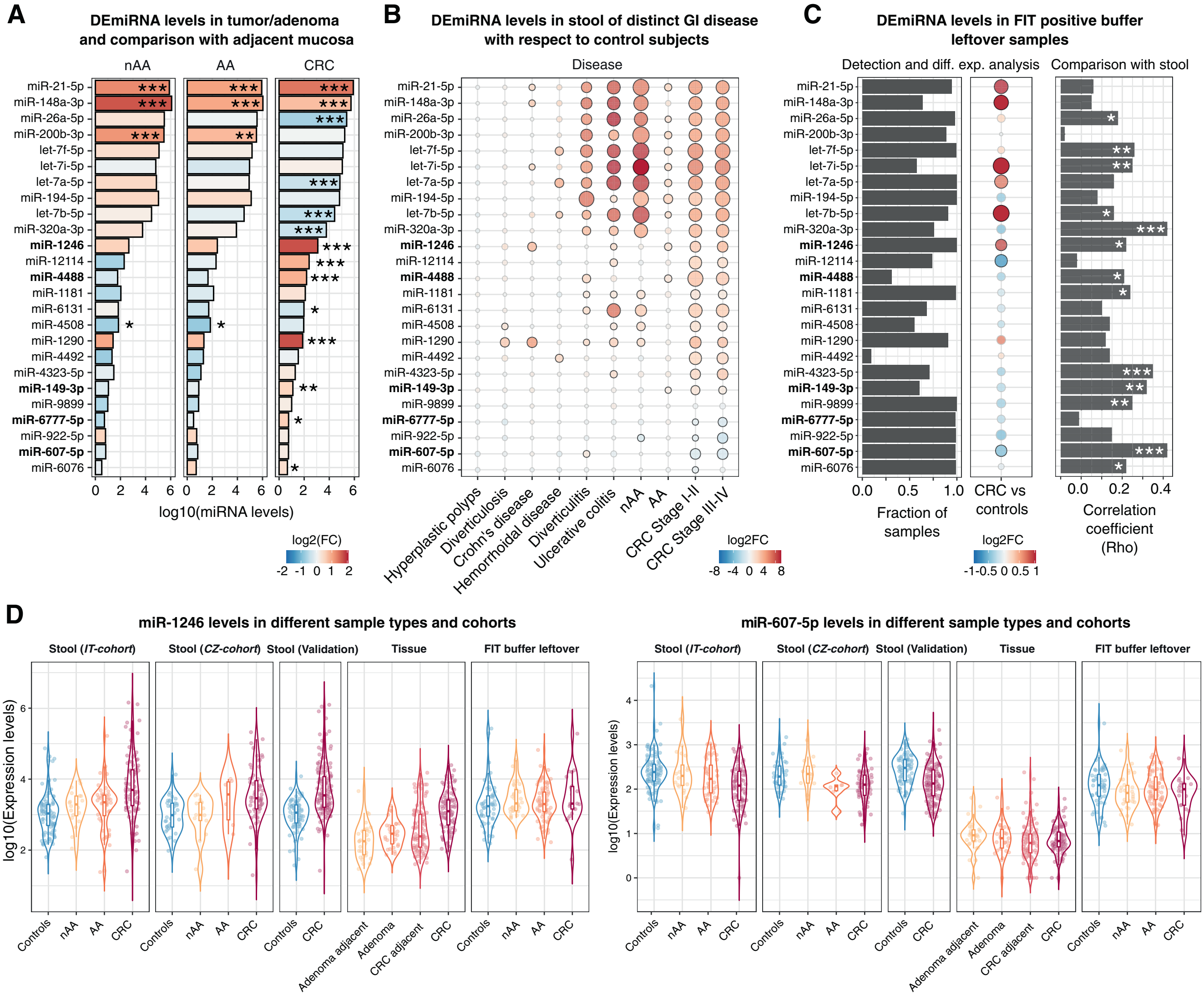


Figure 3. Characterization of the 25 fecal DEmiRNAs in different sample types. (*A*) Bar plot reporting the median levels in tumor, AA, and nAA tissues. The color code represents the log2 fold change (log2FC) from the paired differential expression analysis between CRC/adenoma tissues and matched adjacent mucosa. ***Adjusted $P < .001$, **adjusted $P < .01$, *adjusted $P < .05$. (*B*) Comparison of miRNA levels in the stool of patients with CRC, colorectal adenomas, hyperplastic polyps, or other GI disorders with respect to control individuals. The dot color represents the log2FC, and the dot size is proportional to the analysis significance. Black borders represent results with an adjusted $P < .05$. (*C*) DEmiRNA analysis in FIT leftover samples from CRC screening. (*Left*) The fraction of FIT cohort samples in which each miRNA was detected and (*center*) results of the differential expression analysis between FIT-positive patients with CRC diagnosis based on colonoscopy outcome and those with a negative one. The dot color represents the log2FC, and the dot size is proportional to the analysis significance. Black borders represent a DESeq2 Benjamini-Hochberg adjusted $P < .05$. (*Right*) Correlation coefficients between miRNA levels in stool and FIT buffer leftover samples from the same individuals (***$P < .001$, *$P < .05$). (*D*) Box plots reporting miR-1246 and miR-607-5p levels in all study cohorts and biospecimens.

continuous refinement of miRBase annotations[37] and with evidence of new miRNAs reported by different groups.[38,39]

Consistent with their overall higher/lower levels in the stool of CRC patients with respect to that of control individuals, the 25 DEmiRNA levels also increased/decreased with tumor size and stage. On the other hand, they were characterized by coherent altered levels when patients were stratified by tumor localization (proximal, distal, rectum) (Supplementary Table 4*C*). This further supports the importance of these miRNAs in relationship with the disease, as confirmed by the overrepresentation of cancer-related processes involving their validated target genes (Supplementary Table 2*D* and *E*).

Based on this initial evidence, we implemented an integrated explainable ML strategy to explore, among the 25 DEmiRNAs, the minimal set of stool miRNAs able to accurately discriminate CRC patients from control individuals. Our approach generated a signature composed of 5 miRNAs (namely, miR-1246, miR-607-5p, miR-6777-5p, miR-4488, miR-149-3p) that was clinically validated in an additional independent cohort of cases compared to healthy volunteers and technically validated by another methodology (ie, RT-qPCR). The accurate discrimination of both participants in early and late cancer stages from control individuals confirmed the robustness of these 5 miRNAs for CRC detection. Although based on a small sample set, the signature could also accurately discriminate participants with AA from control individuals (AUC, 0.86), and in all analyses, high performances were obtained, irrespectively, by adjusting or not for sex and age, 2 relevant risk factors for this cancer.[40] To the best of our knowledge, this is the first signature based on fecal miRNAs whose efficiency was proven in populations from 2 countries characterized by different lifestyle and dietary habits[41] and CRC incidence.[42] Notably, such populations also show different trends in early-onset CRC,[43] the incidence of which is linked to unhealthy individual habits, such as a sedentary lifestyle.[44]

Similar to the functional analysis of all 25 DEmiRNAs, focused research on the 5-signature miRNA target genes evidenced a prevalence of genes involved in cancer-related processes, including regulation of the cell cycle, programmed cell death, and DNA damage response. Interestingly, functional analysis of predicted target genes of miR-607-5p highlighted terms/processes related to nuclear cell cycle DNA replication and showed *TRIM66*, *HIPK2*, *GRIN2B*, and *WTIP* as the targets with the highest number of miR-607-5p binding sites (Supplementary Table 5*B* and *C*).

Among all the miRNAs of the signature, miR-1246 has been previously widely studied in CRC. Altered levels of this miRNA have been found in circulating exosomes in relation to cancer metastasis and prognosis.[45,46] Exosomal miR-1246 levels were induced by *Fusobacterium nucleatum* in in vitro and in vivo CRC models with an increase of tumor cell metastatic potential.[47] These results align with more recent observations on the relationship between intratumor levels of *F nucleatum* and the aggressiveness of colon and breast cancers.[48] An intratumor increase in this well-known CRC-related bacteria might induce the release of exosomal miR-1246 in the gut lumen, with the subsequent detection of this miRNA in stool samples. Similar considerations could be drawn from another study investigating a model of enterotoxigenic *Bacteroides fragilis* that induced up-regulation of exosomal miR-1246 in CRC cell lines.[49] Interestingly, in the same study, this microbial species reduced the exosomal levels of another fecal miRNA included in our signature, miR-149-3p, that was demonstrated to regulate tumor-infiltrating CD4$^+$ T-helper type 17 differentiation.[49]

Similar findings were observed when analyzing the fecal miRNome and gut metagenome data from a previous study by our group in which we investigated the miRNA-microbiota relationships in stool samples.[20] Specifically, by reanalyzing the data from that study, miR-1246 levels emerged as significantly related to both *F nucleatum* and *B fragilis* abundances, whereas miR-149-3p was inversely related to *B fragilis* abundances (Supplementary Figure 3*B*). This pervasive relationship between in vitro exosomal miRNA levels and microbial infections suggests that the most informative stool biomarkers for CRC might reflect the dysregulated interactions between colonic tissue and the gut microbiota. Interestingly, in the miRNA-microbiota correlation analysis, 2 down-regulated fecal miRNAs (miR-607-5p and miR-6777-5p), included in the predictive signature and so far scantly investigated in the literature, were inversely related not only to *F nucleatum* and *B fragilis* abundances but also to *Escherichia coli*, another species related to CRC onset[50] (Supplementary Figure 3*B*).

To further explore the stool results, we tested DEmiRNA patterns in tumor and adenoma tissues paired with nonmalignant adjacent mucosa from patients of the IT cohort. Stool generally mirrored the altered miRNA expression levels of these tissues. Only the levels of miR-21-5p and miR-148a-3p increased in both CRC and adenoma compared to matched adjacent mucosa, whereas the other DEmiRNAs (including miR-1246, miR-4488, and miR-149-3p of the signature) showed a CRC-specific dysregulation. miR-607-5p and miR-6777-5p, decreasing in patients' fecal samples, were characterized by low expression levels in both tumor/adenoma and adjacent mucosa, suggesting their deletion or epigenetic silencing. In The Cancer Genome Atlas,[51] both miRNAs are frequently deleted in CRC (Supplementary Table 5*D*), supporting the down-regulation in stool and tumor tissues observed by us. In agreement with our findings, previous studies have demonstrated that the down-regulation of miRNAs seems to be a premature step in the development of several cancers.[52,53] Surprisingly, miR-320a, let-7b-5p, and let-7a-3p, more abundant in stool of CRC patients, were more expressed in adjacent mucosa than in tumor tissue. miR-320a has been widely reported as down-regulated in CRC,[54] whereas its circulating levels increased in relation to gut inflammation in IBD patients,[55] coherent with our data in stool samples. Interestingly, miR-320a has been described as a key regulator of intestinal barrier formation.[56] Similarly, the expression of let-7 family members has been observed in the healthy gut epithelium, whereas their genetic depletion induced tumorigenesis in CRC mouse models.[57] Thus, the analysis of stool miRNAs is relevant to identify not only markers of the tumor small noncoding transcriptome but may also unveil an intestinal response of the stromal component to the presence of a tumor mass.

We also explored the miRNome of plasma EVs from a subset of the study population using the same experimental approach as in stool and tissue samples. However, in this circulating biospecimen, only a few miRNAs showed similar trends as in feces. For instance, among the miRNAs of the signature, miR-1246 and miR-4488 levels significantly increased in plasma EVs of CRC patients compared with control individuals. These results are consistent with previous findings reported by us, supporting stool miRNAs as more sensitive than plasma miRNAs in reflecting intestinal changes driven by a long-term dietary pattern.[24] Although more data are needed to compare the stool and plasma EV miRNome, given the reported relationships between miR-1246 levels in EVs and CRC metastasis,[45] these circulating molecules may be more informative for advanced stages of the disease, which is beyond the scope of our investigation.

In this study, we sought to compare the stool DEmiRNA profiles of CRC patients with those of patients with other bowel inflammatory diseases of different severity confirmed by colonoscopy. Besides different polyp types, we included samples from several GI diseases, like different types of IBDs and diverticulitis. Notably, although the CRC-specific miRNAs were down-regulated, most of the altered miRNAs in common with adenomas and inflammatory diseases were up-regulated: miR-21-5p was the clearest example, confirming the literature.[26] As an exception, miR-607-5p was down-regulated in the stool miRNA profiles of patients with AA and ulcerative colitis. Accordingly, recent studies showed altered miRNA profiles in the fecal samples of patients with inflammation,[58,59] even in relation to microbiota.[60] We can therefore conclude that altered stool miRNA profiles reflect either the intestinal response to an inflammatory process or the transcriptional alterations related specifically to CRC development. Importantly, we clearly demonstrated that well-known CRC-related miRNAs, such as miR-21-5p, show dysregulated fecal levels in several disease contexts, suggesting that other miRNAs, such as miR-6777-5p and miR-149-3p, should be investigated to design CRC-specific molecular signatures. This is the first evidence from a large-scale analysis of individuals with different gastrointestinal diseases of stool miRNAs specifically altered in CRC. It also highlights an extensive reflection of the gut inflammation on the fecal miRNA levels.

The fact that specific dysregulated fecal miRNAs could distinguish individuals with CRC or precursor lesions from control individuals and that, at least for cancer, data were confirmed in different cohorts, encouraging their use to complement the existing noninvasive screening tests. In this respect, we also investigated whether miRNAs can be detected in buffer-diluted stool leftovers from FIT tubes used in a context of a population-based screening program, and we found a remarkable similarity between the profiles detected in the stool samples collected in nucleic acid preservative medium tubes from the same participants. Despite data on a larger cohort being needed, this pilot small RNA-seq–based quantification of miRNAs in FIT buffer leftovers is consistent with previous evidence measuring miRNAs in this sample type by RT-qPCR,[22] as well as by us. By exploring miRNA profiles within FIT-positive patients, we observed a subset of miRNAs differentially expressed between individuals with a positive or a negative colonoscopy outcome. In addition, miR-1246 and miR-607-5p from the 5-miRNA signature deserve further investigation because they were detected in most of the samples, and their levels respectively increased and decreased progressively, going from individuals with negative colonoscopy results, to those with adenomas of different severity, to CRC patients. Although these data confirm that miRNAs can be widely detected in FIT leftovers, the comparative results between individuals must be carefully considered given the small group size analyzed so far; the lack of samples from FIT-negative individuals; and the fact that we cannot rule out the role of confounding factors, including subclinical diseases in the colonoscopy-negative patients.

Most likely, by including hemoglobin levels evaluated by FIT, the discrimination capability of the present stool miRNA predictive signature would be further improved, as already reported in the past (FIT/FOBT + microbiome,[11,61] FIT + miRNAs,[21] and FIT + methylation markers[62]). The sensitivity and specificity of our 5-miRNA signature suggest that it could show a similar diagnostic performance as the multitarget stool DNA test[63] when used as a screening test in average-risk populations. Duran-Sanchon et al[21] proposed a 2-stool miRNA-based classification signature (namely, miR-27a-3p and miR-421) combined with hemoglobin levels, age, and sex of FIT-positive individuals. The signature accurately classified CRC (AUC, 0.93) from control individuals but was less efficient when AA patients were included (AUC, 0.70).[62] Different from us, the researchers initially selected miRNAs based on their differential expression between tumor tissue and adjacent mucosa and included in all models sex and age, 2 important risk factors for CRC. Hereby, we demonstrated the robustness of our signature because its performance remained similar even without the inclusion of age and sex covariates. In addition, despite the study not being designed for identifying stool biomarkers for adenomas, the 5-miRNA signature was able to accurately distinguish AA alone or in combination with CRC (AUC, 0.84), suggesting its use to detect precancer lesions at risk. In our study, miR-27a-3p and miR-421 were detected in tissue samples but not in stool, where only the former miRNA was measurable. In search of reproducible fecal molecular biomarkers for the noninvasive diagnosis of CRC and adenomas,[11] a hypothesis-free miRNome-wide approach, such as the small RNA-seq analysis in stool performed in multiple independent populations, overcomes these issues.

The present study has several strengths: (1) the inclusion of independent cohorts from 2 countries with different diet and lifestyle habits as well as CRC rates; (2) the fact that the cohorts were different for CRC clinical characteristics, allowing the identification of accurate biomarkers independent of the disease stage; (3) the adoption of the same protocol for the collection of stool in both training cohorts; (4) the validation of the signature on a cohort with a different stool collection protocol, showing its robustness; (5) the miRNome-wide approach in different biospecimens and different GI disease contexts, which has allowed us to discriminate miRNAs specifically dysregulated in the stool of CRC patients; 6) the implementation of an explainable ML approach able to provide an unbiased method for identifying the minimal set of predictive biomarkers.

However, we are also aware of several limitations. Although there was a similar study design for recruitment, the 2 cohorts were heterogeneous for individual cancer categories. This heterogeneity could be responsible for the observed differences in the median stool miRNA levels and expression differences between the 2 cohorts. Given the difference in the clinical characteristics of CRC patients, the main driver of such a difference may be the higher proportion of low-grade and low-stage tumors in the CZ cohort. However, the fact that the results are reproducible between cohorts further supports the robustness of the signature identified in this study.

Despite the large number of analyzed samples, the variegated spectrum of CRC, adenomas, and other precancerous lesions needs to be more exhaustively represented and deserves further investigation. For example, we did not investigate serrated lesions or deeply explore the alterations in CRC stratified based on molecular or clinical data. In addition, even though the observed DEmiRNAs were not reported to be modulated by dietary habits,[24] the lack of dietary/lifestyle information of analyzed individuals may represent a limitation of the study. Follow-up studies with additional cohorts representing patients with different ethnicities, dietary patterns, and lifestyle habits are required, but this is beyond the scope of this study, which, to our knowledge, represents the largest sequencing-based analysis of stool miRNAs so far.

In conclusion, this multicenter and international study based on small RNA-seq allowed us to comprehensively detect in stool several miRNAs differentially expressed in CRC. Furthermore, the implemented ML approach identified a minimal number of miRNAs whose combined profiles showed a good discriminating power for the presence of a tumor or AA, independent of age and sex. This may represent a fecal signature for improving the effectiveness of current noninvasive screening programs, potentially increasing sensitivity and maintaining high specificity, and applicable on a large scale, with a reasonable cost/time required.

In this respect, for FIT implementation, in the near future miRNA profiles will be investigated in additional cohorts, possibly from different countries, increasing the number/types of precancer lesions and also including FIT-negative samples, with the chance to explore the role of diet and lifestyle habits on an adequate scale. Furthermore, the inclusion of FIT-negative samples will allow the possibility to prospectively test miRNA profiles in subsequent rounds of CRC screening, collecting multiple samples per individual. In parallel, the analysis of the microbiome composition of stool/leftover FIT samples will help deepen the research on gut-host crosstalk with small noncoding RNAs. Finally, even if small RNA-seq and RT-qPCR currently represent the most commonly used approaches for miRNA analyses, we must consider that more rapid, practical, but reliable approaches, such as biosensors, may provide an alternative for testing the miRNA signature in a large clinical setting.

Supplementary Material

Note: To access the supplementary material accompanying this article, visit the online version of *Gastroenterology* at www.gastrojournal.org, and at https://doi.org/10.1053/j.gastro.2023.05.037.

References

1. Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. Nat Rev Gastroenterol Hepatol 2019;16:713–732.
2. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021;71:209–249.
3. Ferlay J, Colombet M, Soerjomataram I, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer 2019;144:1941–1953.
4. Kral J, Kojecky V, Stepan M, et al. The experience with colorectal cancer screening in the Czech Republic: the detection at earlier stages and improved clinical outcomes. Public Health 2020;185:153–158.
5. Lauby-Secretan B, Vilahur N, Bianchini F, et al. The IARC perspective on colorectal cancer screening. N Engl J Med 2018;378:1734–1740.
6. Senore C, Basu P, Anttila A, et al. Performance of colorectal cancer screening in the European Union member states: data from the second European screening report. Gut 2019;68:1232–1244.
7. Rabeneck L, Chiu HM, Senore C. International perspective on the burden of colorectal cancer and public health effects. Gastroenterology 2020;158:447–452.
8. **Robertson DJ, Lee JK**, Boland CR, et al. Recommendations on fecal immunochemical testing to screen for colorectal neoplasia: a consensus statement by the US Multi-Society Task Force on Colorectal Cancer. Gastroenterology 2017;152:1217–1237.
9. Loktionov A. Biomarkers for detecting colorectal cancer non-invasively: DNA, RNA or proteins? World J Gastrointest Oncol 2020;12:124–148.
10. **Weng M, Wu D, Yang C**, et al. Noncoding RNAs in the development, diagnosis, and prognosis of colorectal cancer. Transl Res 2017;181:108–120.
11. **Thomas AM, Manghi P**, Asnicar F, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med 2019;25:667–678.
12. **Sun Y, Guo Z**, Liu X, et al. Noninvasive urinary protein signatures associated with colorectal cancer diagnosis and metastasis. Nat Commun 2022;13(1):2757.
13. Francavilla A, Turoczi S, Tarallo S, et al. Exosomal microRNAs and other non-coding RNAs as colorectal cancer biomarkers: a review. Mutagenesis 2020;35:243–260.
14. Hombach S, Kretz M. Non-coding RNAs: classification, biology and functioning. Adv Exp Med Biol 2016;937:3–17.
15. Di Leva G, Croce CM. miRNA profiling of cancer. Curr Opin Genet Dev 2013;23:3–11.
16. Moridikia A, Mirzaei H, Sahebkar A, et al. MicroRNAs: potential candidates for diagnosis and treatment of colorectal cancer. J Cell Physiol 2018;233:901–913.
17. Dragomir MP, Kopetz S, Ajani JA, et al. Non-coding RNAs in GI cancers: from cancer hallmarks to clinical utility. Gut 2020;69:748–763.
18. Pardini B, Sabo AA, Birolo G, et al. Noncoding RNAs in extracellular fluids as cancer biomarkers: the new frontier of liquid biopsies. Cancers (Basel) 2019;11(8):1170.
19. Cervena K, Novosadova V, Pardini B, et al. Analysis of MicroRNA expression changes during the course of therapy in rectal cancer patients. Front Oncol 2021;11:702258.
20. **Tarallo S, Ferrero G**, Gallo G, et al. Altered fecal small RNA profiles in colorectal cancer reflect gut microbiome composition in stool samples. mSystems 2019;4(5):e00289-19.
21. Duran-Sanchon S, Moreno L, Auge JM, et al. Identification and validation of microRNA profiles in fecal samples for detection of colorectal cancer. Gastroenterology 2020;158:947–957.
22. Zhao Z, Zhu A, Bhardwaj M, et al. Fecal microRNAs, fecal microRNA panels, or combinations of fecal microRNAs with fecal hemoglobin for early detection of colorectal cancer and its precursors: a systematic review. Cancers (Basel) 2021;14(1):65.

23. Francavilla A, Tarallo S, Pardini B, et al. Fecal microRNAs as non-invasive biomarkers for the detection of colorectal cancer: a systematic review. Minerva Biotecnol 2019;31:30–42.
24. **Tarallo S, Ferrero G, De Filippis F**, et al. Stool microRNA profiles reflect different dietary and gut microbiome patterns in healthy individuals. Gut 2021;71:1302–1314.
25. **Francavilla A, Gagliardi A, Piaggeschi G**, et al. Faecal miRNA profiles associated with age, sex, BMI, and lifestyle habits in healthy individuals. Sci Rep 2021;11(1):20645.
26. Jenike AE, Halushka MK. miR-21: a non-specific biomarker of all maladies. Biomark Res 2021;9(1):18.
27. Zarchy TM, Ershoff D. Do characteristics of adenomas on flexible sigmoidoscopy predict advanced lesions on baseline colonoscopy? Gastroenterology 1994;106:1501–1504.
28. **Wirbel J, Pyl PT**, Kartal E, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med 2019;25:679–689.
29. Lin Y, Lau HC, Liu Y, et al. Altered mycobiota signatures and enriched pathogenic *Aspergillus rambellii* are associated with colorectal cancer based on multicohort fecal metagenomic analyses. Gastroenterology 2022;163:908–921.
30. Zwinsová B, Petrov VA, Hrivňáková M, et al. Colorectal tumour mucosa microbiome is enriched in oral pathogens and defines three subtypes that correlate with markers of tumour progression. Cancers (Basel) 2021;13(19):4799.
31. **Francavilla A, Ferrero G, Pardini B**, et al. Gluten-free diet affects fecal small non-coding RNA profiles and microbiome composition in celiac disease supporting a host-gut microbiota crosstalk. Gut Microbes 2023;15(1):2172955.
32. **Sabo AA, Birolo G, Naccarati A**, et al. Small non-coding RNA profiling in plasma extracellular vesicles of bladder cancer patients by next-generation sequencing: expression levels of miR-126-3p and piR-5936 increase with higher histologic grades. Cancers (Basel) 2020;12(6):1507.
33. **Moisoiu T, Dragomir MP, Iancu SD**, et al. Combined miRNA and SERS urine liquid biopsy for the point-of-care diagnosis and molecular stratification of bladder cancer. Mol Med 2022;28(1):39.
34. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550.
35. Zhang J, Storey KB. RBiomirGS: an all-in-one miRNA gene set analysis solution featuring target mRNA mapping and expression profile integration. PeerJ 2018;6:e4262.
36. Slaby O. Non-coding RNAs as biomarkers for colorectal cancer screening and early detection. Adv Exp Med Biol 2016;937:153–170.
37. **Alles J, Fehlmann T**, Fischer U, et al. An estimate of the total number of true human miRNAs. Nucleic Acids Res 2019;47:3353–3364.
38. **Jima DD, Zhang J**, Jacobs C, et al. Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs. Blood 2010;116:e118–e127.
39. **Friedlander MR, Lizano E, Houben AJ**, et al. Evidence for the biogenesis of more than 1,000 novel human microRNAs. Genome Biol 2014;15(4):R57.
40. Wei EK, Giovannucci E, Wu K, et al. Comparison of risk factors for colon and rectal cancer. Int J Cancer 2004;108:433–442.
41. Imamura F, Micha R, Khatibzadeh S, et al. Dietary quality among men and women in 187 countries in 1990 and 2010: a systematic assessment. Lancet Glob Health 2015;3(3):e132–e142.
42. **Wong MCS, Huang J, Lok V**, et al. Differences in incidence and mortality trends of colorectal cancer worldwide based on sex, age, and anatomic location. Clin Gastroenterol Hepatol 2021;19:955–966.
43. Vuik FE, Nieuwenburg SA, Bardou M, et al. Increasing incidence of colorectal cancer in young adults in Europe over the last 25 years. Gut 2019;68:1820–1826.
44. Patel SG, Karlitz JJ, Yen T, et al. The rising tide of early-onset colorectal cancer: a comprehensive review of epidemiology, clinical features, biology, risk factors, prevention, and early detection. Lancet Gastroenterol Hepatol 2022;7:262–274.
45. Desmond BJ, Dennett ER, Danielson KM. Circulating extracellular vesicle microRNA as diagnostic biomarkers in early colorectal cancer—a review. Cancers (Basel) 2019;12(1):52.
46. Cooks T, Pateras IS, Jenkins LM, et al. Mutant p53 cancers reprogram macrophages to tumor supporting macrophages via exosomal miR-1246. Nat Commun 2018;9(1):771.
47. Guo S, Chen J, Chen F, et al. Exosomes derived from *Fusobacterium nucleatum*-infected colorectal cancer cells facilitate tumour metastasis by selectively carrying miR-1246/92b-3p/27a-3p and CXCL16. Gut 2021;70:1507–1519.
48. **Fu A, Yao B, Dong T**, et al. Emerging roles of intratumor microbiota in cancer metastasis. Trends Cell Biol 2023;33:583–593.
49. **Cao Y, Wang Z, Yan Y**, et al. Enterotoxigenic *Bacteroides fragilis* promotes intestinal inflammation and malignancy by inhibiting exosome-packaged miR-149-3p. Gastroenterology 2021;161:1552–1566.
50. Clay SL, Fonseca-Pereira D, Garrett WS. Colorectal cancer: the facts in the case of the microbiota. J Clin Invest 2022;132(4):e155101.
51. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487(7407):330–337.
52. Esquela-Kerscher A, Slack FJ. Oncomirs – microRNAs with a role in cancer. Nat Rev Cancer 2006;6:259–269.
53. **Vila-Navarro E, Vila-Casadesus M**, Moreira L, et al. MicroRNAs for detection of pancreatic neoplasia: biomarker discovery by next-generation sequencing and validation in 2 independent cohorts. Ann Surg 2017;265:1226–1234.
54. Liang Y, Li S, Tang L. MicroRNA 320, an anti-oncogene target miRNA for cancer therapy. Biomedicines 2021;9(6):591.
55. Cordes F, Demmig C, Bokemeyer A, et al. MicroRNA-320a monitors intestinal disease activity in patients with inflammatory bowel disease. Clin Transl Gastroenterol 2020;11(3):e00134.

56. Muenchau S, Deutsch R, de Castro IJ, et al. Hypoxic environment promotes barrier formation in human intestinal epithelial cells through regulation of microRNA 320a expression. Mol Cell Biol 2019;39(14):e00553-18.

57. Madison BB, Jeganathan AN, Mizuno R, et al. Let-7 represses carcinogenesis and a stem cell phenotype in the intestine via regulation of Hmga2. PLoS Genet 2015; 11(7):e1005408.

58. Wohnhaas CT, Schmid R, Rolser M, et al. Fecal microRNAs show promise as noninvasive Crohn's disease biomarkers. Crohns Colitis 360 2020;2(1):otaa003.

59. **Verdier J, Breunig IR**, Ohse MC, et al. Faecal microRNAs in inflammatory bowel diseases. J Crohns Colitis 2020;14:110–117.

60. **Ambrozkiewicz F, Karczmarski J**, Kulecka M, et al. In search for interplay between stool microRNAs, microbiota and short chain fatty acids in Crohn's disease—a preliminary study. BMC Gastroenterol 2020;20(1):307.

61. **Xie YH, Gao QY, Cai GX**, et al. Fecal *Clostridium symbiosum* for noninvasive detection of early and advanced colorectal cancer: test and validation studies. EBioMedicine 2017;25:32–40.

62. Bosch LJ, Oort FA, Neerincx M, et al. DNA methylation of phosphatase and actin regulator 3 detects colorectal cancer in stool and complements FIT. Cancer Prev Res (Phila) 2012;5(3):464–472.

63. Imperiale TF, Ransohoff DF, Itzkowitz SH, et al. Multitarget stool DNA testing for colorectal-cancer screening. N Engl J Med 2014;370:1287–1297.

Author names in bold designate shared co-first authorship.

Received August 8, 2022. Accepted May 24, 2023.

Correspondence
Address correspondence to: Alessio Naccarati, PhD, Italian Institute for Genomic Medicine, c/o IRCCS Candiolo, SP 142, Km 3.95, 10060 Candiolo, Turin, Italy. e-mail: alessio.naccarati@iigm.it; or Barbara Pardini, PhD, Italian Institute for Genomic Medicine, c/o IRCCS Candiolo, SP 142, Km 3.95, 10060 Candiolo, Turin, Italy. e-mail: barbara.pardini@iigm.it.

Acknowledgments
The authors are very grateful to all volunteers who participated with enthusiasm in the study. They are also thankful to Dr Vladimir Benes (Genecore Facility at EMBL, Heidelberg, Germany) for his kind collaboration. The artwork for this study has been created with BioRender.com.

CRediT Authorship Contributions
Barbara Pardini, PhD (Data curation: Equal; Formal analysis: Supporting; Funding acquisition: Supporting; Investigation: Lead; Methodology: Lead; Supervision: Lead; Writing – original draft: Lead; Writing – review & editing: Equal).

Giulio Ferrero, PhD (Conceptualization: Equal; Data curation: Lead; Formal analysis: Lead; Investigation: Equal; Methodology: Lead; Software: Lead; Supervision: Supporting; Validation: Equal; Visualization: Lead; Writing – original draft: Lead; Writing – review & editing: Lead).

Sonia Tarallo, PhD (Conceptualization: Lead; Data curation: Lead; Formal analysis: Supporting; Investigation: Lead; Methodology: Lead; Validation: Lead; Visualization: Equal; Writing – original draft: Lead; Writing – review & editing: Equal).

Gaetano Gallo, MD, PhD (Conceptualization: Supporting; Data curation: Lead; Investigation: Supporting; Resources: Lead; Writing – review & editing: Equal).

Antonio Francavilla, PhD (Conceptualization: Supporting; Data curation: Supporting; Investigation: Supporting; Methodology: Equal; Validation: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Nicola Licheri, MSc (Data curation: Equal; Formal analysis: Equal; Resources: Equal; Software: Equal; Writing – review & editing: Supporting).

Mario Trompetto, MD, PhD (Data curation: Equal; Investigation: Supporting; Methodology: Supporting; Visualization: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Giuseppe Clerico, MD, PhD (Data curation: Supporting; Investigation: Supporting; Project administration: Supporting; Visualization: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Carlo Senore, MD, PhD (Conceptualization: Supporting; Data curation: Supporting; Formal analysis: Supporting; Methodology: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Sergio Peyre, MD (Funding acquisition: Lead; Investigation: Supporting; Methodology: Supporting; Project administration: Supporting; Writing – original draft: Supporting; Writing – review & editing: Supporting).

Veronika Vymetalkova, PhD (Data curation: Equal; Formal analysis: Supporting; Funding acquisition: Supporting; Methodology: Equal; Project administration: Equal; Validation: Equal; Writing – original draft: Supporting; Writing – review & editing: Equal).

Ludmila Vodickova, PhD (Data curation: Supporting; Formal analysis: Supporting; Funding acquisition: Supporting; Investigation: Supporting; Project administration: Supporting; Validation: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Vaclav Liska, MD, PhD (Data curation: Supporting; Investigation: Supporting; Resources: Equal; Writing – original draft: Supporting; Writing – review & editing: Equal).

Ondrej Vycital, MD, PhD (Conceptualization: Supporting; Data curation: Supporting; Investigation: Supporting; Project administration: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Miroslav Levy, MD, PhD (Conceptualization: Supporting; Data curation: Supporting; Investigation: Supporting; Project administration: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Peter Macinga, MD, PhD (Conceptualization: Supporting; Data curation: Supporting; Investigation: Supporting; Project administration: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Tomas Hucl, MD, PhD (Conceptualization: Supporting; Data curation: Supporting; Investigation: Supporting; Project administration: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Eva Budinska, PhD (Data curation: Supporting; Investigation: Supporting; Resources: Equal; Validation: Equal; Writing – original draft: Supporting; Writing – review & editing: Equal).

Pavel Vodicka, MD, PhD (Conceptualization: Supporting; Data curation: Supporting; Formal analysis: Supporting; Funding acquisition: Supporting; Investigation: Supporting; Resources: Supporting; Writing – original draft: Supporting; Writing – review & editing: Equal).

Francesca Cordero, PhD (Conceptualization: Equal; Data curation: Lead; Formal analysis: Lead; Investigation: Equal; Methodology: Equal; Supervision: Supporting; Validation: Supporting; Visualization: Supporting; Writing – original draft: Lead; Writing – review & editing: Lead).

Alessio Naccarati, PhD (Conceptualization: Lead; Data curation: Lead; Formal analysis: Lead; Funding acquisition: Lead; Investigation: Supporting; Methodology: Equal; Project administration: Lead; Resources: Equal; Supervision: Lead; Validation: Equal; Visualization: Lead; Writing – original draft: Lead; Writing – review & editing: Lead).

Conflicts of interest
The authors disclose no conflicts.

Funding
This work was supported by the Italian Institute for Genomic Medicine and Compagnia di San Paolo Torino, Italy (to Alessio Naccarati, Barbara Pardini, and Sonia Tarallo), by Lega Italiana per La Lotta contro i Tumori (to Francesca Cordero, Barbara Pardini, and Alessio Naccarati), by the Grant Agency of the Czech Republic (17-16857S to Alessio Naccarati, Pavel Vodicka, Ludmila Vodickova, and Veronika Vymetalkova and 22-05942S to Veronika Vymetalkova, Ludmila Vodickova, Tomas Hucl, and Vaclav Liska), by the Grant Agency of the Ministry of Health of the Czech Republic (AZV NV18-03-00199 to Veronika Vymetalkova, Pavel Vodicka, Ludmila Vodickova, Tomas Hucl, and Peter Macinga), by the Fondazione CRT (grant no. 2017.2025 to Francesca Cordero), and by the COST action TRANSCOLONCAN (CA17118). Pavel Vodicka and Ludmila Vodickova acknowledge support from the National Operation Program (National Institute for Cancer Research LX22NPO 05120). Eva Budinska thanks the RECETOX Research Infrastructure (No. LM2023069) financed by the Ministry of Education, Youth and Sports, and the OP RDE (the CETOCOEN EXCELLENCE project No. CZ.02.1.01/0.0/0.0/17_043/0009632 and the CETOCOEN PLUS project No. CZ.02.1.01/0.0/0.0/15_003/0000469) for supportive background. This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 825410 (ONCOBIOME project to Alessio Naccarati, Barbara Pardini, Eva Budinska, Francesca Cordero, and Sonia Tarallo). The research leading to these results has received funding from AIRC under IG 2020 - ID 24882 (PI: Alessio Naccarati) to Alessio Naccarati; and IG 2019 - ID 23473 (PI: Carlo Senore) to Carlo Senore. This publication reflects only the authors' views, and the European Commission is not responsible for any use that may be made of the information it contains.

Data Availability
All data relevant to the study are included in the article or in the Supplementary Material. Raw data are available upon request to the corresponding author.

Supplementary Methods

Stool Study Cohorts

Italian cohort. Stool specimens as well as clinical and demographic data were collected from 317 individuals recruited in a hospital-based study at 1 hospital in Vercelli, Italy (Table 1 and Figure 1*A*). Based on the results of a completed colonoscopy examination with adequate bowel preparation, participants were classified into (1) 89 CRC patients (individuals with newly diagnosed sporadic CRC); (2) 74 polyps patients, stratified as hyperplastic polyps (n = 6), nAA (n = 20), or AA (n = 48); (3) 49 patients with GI disease, such as IBD (including Crohn's disease and indeterminate or ulcerative colitis) or diverticular disease; and (4) 105 control individuals.

AAs were defined based on the presence of high-grade dysplasia, villous component, or lesion length of >1 cm as defined by Zarchy and Ershoff.[1] Of this cohort, 93 stool samples (from 29 CRC patients, 27 polyps, 13 individuals with a GI disease, and 24 colonoscopy-negative control individuals) were used and have been described previously.[2–4]

Czech cohort. Stool specimens as well as clinical and demographic data were collected from a cohort of 162 Czech individuals recruited in 2 hospitals in Prague and 1 in Plzen, Czech Republic (Table 1 and Figure 1*A*). Based on colonoscopy results, participants were divided into (1) 66 CRC patients; (2) 28 individuals with colorectal polyps, grouped as hyperplastic polyps (n = 9), nAA (n = 13), and AA (n = 6); (3) 32 patients with other GI disorders; and (4) 36 colonoscopy-negative control individuals.

In both studies, colonoscopy was recommended for 2 main reasons: (1) because of the recommendation of the family doctor for various reasons (age of the patient, complaints in the gut, etc) or (2) because the patient had a positive FIT result (ie, there was blood in the stool at the time of the test, and therefore the individual was invited to have a colonoscopy to further investigate the reason for blood in stool). In any case, individuals with major GI diseases other than cancer were considered apart from those control individuals with a negative colonoscopy finding.

Validation cohort. Stool specimens from 141 CRC patients recruited in the hospital in Brno, Czech Republic,[5] and 80 stool samples of healthy volunteers contributing to science[6] were included as an independent validation cohort. Stool specimens from 141 CRC patients were obtained at a hospital in Brno, Czech Republic: these individuals were previously described by Zwinsova et al[5] and here are sequenced for the first time for small RNA-seq.

Stool samples of healthy volunteers contributing to science are a part (about 20%) of the cohort described and sequenced for small RNA-seq by Tarallo et al[6] and Francavilla et al.[7] The healthy volunteers are derived from a subgroup of healthy individuals (no cancer, no precancer lesions) nested from the omnivorous group described by Tarallo et al.[6] and Francavilla et al.[7] Only individuals with age >30 years were considered for the analysis.

Fecal immunochemical test cohort. FIT leftover samples collected from 185 participants with a positive result from FIT analysis in the CRC screening for the general population of Piedmont Region (Italy) were added to the study. Based on the results of a completed colonoscopy examination with adequate bowel preparation, the individuals were classified as control individuals (n = 53) or individuals with AA (n = 80) or nAA (n = 30) and with CRC (n = 22). Among the 185 participants, 57 also provided stool samples before undergoing colonoscopy.

Colonoscopy was recommended because the patients had abnormal or positive FIT results (ie, there was blood in the stool at the time of the test), and therefore they were invited to have a colonoscopy to further investigate the reason for blood in stool.

Other Analyzed Biospecimens

For 132 patients (102 CRC patients and 30 patients with colorectal adenoma) primary CRC/adenoma tissues paired with adjacent colonic mucosa were collected in the same hospital as IT cohort. Among these patients, 69 (51 CRC and 18 colorectal adenoma) donated their stool and plasma samples and were included in the IT cohort.

Blood samples were collected from 210 participants of the IT cohort, stratified as 52 patients with CRC, 19 with AAs, 15 with nAAs, 6 with hyperplastic polyps, 34 with other GI disorders, and 79 control individuals.

Sample Collection

Naturally evacuated fecal samples were obtained from all participants previously instructed to self-collect the specimen at home. For all cohorts, stool samples were collected in nucleic acid collection and transport tubes with RNA stabilizing solution (Norgen Biotek Corp) and returned to the endoscopy unit. Stool aliquots (200 μL) were stored at –80°C until RNA extraction.[6,8] The only exception was represented by the validation cohort of CRC patients from Brno, for which stool samples were collected from untreated patients before the scheduled surgery with DNA-free swabs (Deltalab). Patients performed the collection at home before their hospitalization for the surgery and brought the samples to the hospital, where they were immediately frozen at –80°C until further processing.

For the FIT cohort, leftovers from FIT tubes (~ 1.2 mL) used for automated tests (OC-sensor, Eiken Chemical Co) for hemoglobin quantification were also collected and stored at –80°C until use.

Plasma samples were obtained from 8 mL of blood centrifuged for 10 minutes at 1000 revolutions/minute, and aliquots were stored at –80°C until use. Plasma exosomes/EVs were isolated from 200 μL of plasma using the ExoQuick exosome precipitation solution (System Biosciences, Mountain View), according to the manufacturer's instructions.[9,10] Briefly, plasma was mixed with 50.4 μL of ExoQuick solution and refrigerated at 4°C overnight (at least 12 hours). The mixture was then further centrifuged at 1500*g* for 30 minutes. The EV pellet was dissolved in 200 μL of nuclease-free water, and RNA was extracted immediately from the solution.

Paired primary tumor/adenoma tissue and nonmalignant adjacent mucosa were obtained from CRC and adenoma patients (at least 20 cm distant), collected during surgical resection and immediately immersed in RNAlater solution (Ambion). All tissues samples were stored at –80°C until use.

Extraction of Total RNA

Total RNA was extracted from all stool samples using the Stool Total RNA Purification Kit (Norgen Biotek Corp) as previously described.[8,10] Total RNA from plasma EVs was extracted as described by Sabo et al[9] and Ferrero et al.[10] For tissue samples, total RNA was isolated using QIAzol (Qiagen) after tissue homogenization performed with ULTRA-TURRAX Homogenizer (IKA), followed by phenol/chloroform extraction according to the manufacturer's standard protocol.

Library Preparation for Small RNA Sequencing

Small RNA-seq libraries were prepared from RNA extracted from tissues, stool, and plasma EVs as previously described by Tarallo et al.[6] Briefly, the NEBNext Multiplex Small RNA Library Prep for Illumina (New England Biolabs, Inc) kit was used to convert small RNA transcripts into barcoded complementary DNA (cDNA) libraries. For each library, 6 μL of RNA (35 ng for EV RNA and 250 ng for tissue/stool RNA) was processed as the starting material. Each library was prepared with a unique indexed primer. Multiplex adapter ligations, RT primer hybridization, RT reaction, and PCR amplification were performed according to the manufacturer's protocol. After PCR amplification, the cDNA constructs were purified with the QIAQuick PCR Purification Kit (Qiagen), following the modifications suggested by the NEBNext Multiplex Small RNA Library Prep for Illumina protocol. Final libraries were loaded on the Bioanalyzer 2100 (Agilent Technologies) using the DNA High Sensitivity Kit (Agilent Technologies) according to the manufacturer's protocol. Libraries were pooled together (in 24-plex or 30-plex) and further purified with a gel size selection. A final Bioanalyzer 2100 run with the High Sensitivity DNA Kit (Agilent Technologies) allowed us to assess DNA library quality regarding size, purity, and concentration. The obtained libraries were subjected to the Illumina sequencing pipeline on Illumina HiSeq4000 and NextSeq500 sequencers (Illumina Inc).

Quantitative Real-Time Polymerase Chain Reaction

Five miRNAs of the final signature (miR-607-5p, miR-6777-5p, miR-4488, miR-149-3p, and miR-1246) were validated with a different technique in 2 subsets of stool RNA from the IT cohort (n = 51), the CZ cohort (n = 45), and the FIT cohort (n 8) using the miRCURY LNA SYBR Green PCR kit (Qiagen) according to the manufacturer's instructions for plasma/serum. RT was performed using the miRCURY LNA RT kit (Qiagen) according to the manufacturer's instructions with the addition of 1 spike-in (UniSp6) to the RT reaction.

For qPCR, complement cDNA was diluted 1:30; 3 μL of 1:30 water-diluted cDNA products were mixed at 7 μL of miRCURY SYBR Green Mastermix and 1 μL of specific miRNA probe (Qiagen). All cDNA products were prepared in triplicate PCR reactions following the manufacturer's instructions. For quality control purposes, 1 RNA sample was measured twice, and a sample containing nuclease-free water and carrier RNA was profiled as the negative control. All the reactions were run on the ABI Prism 7900 Sequence Detection System (Applied Biosystems). A melt curve analysis was performed for the amplification specificity of each individual target per sample.

GenEx software (Multi-D) was used for data preprocessing, including interplate calibration, evaluation of isolation and RT efficiency, setting specific cutoffs for negative control miRNA cycle threshold (Ct) values, and triplicate averaging. The analyses were performed by calculating ΔCt values by global mean. The fold change was calculated as log2 – ΔΔCT between CRC and control samples. miRNAs with a Ct value of >38 were deemed to be not detected. To avoid biased inference due to qPCR nondetects (Ct value = 40), a left-censoring approach was used. Ct values of 40 were in fact substituted with the highest observed Ct value for a given miRNA.[11] Ct values were then normalized by subtracting the Ct value of the selected endogenous controls or the global mean Ct from each of the 5 miRNAs of interest. Differential miRNA expression was determined by logistic regression adjusted for age and smoking. The unadjusted *P* values of <.05 were considered as statistically significant because these analyses were hypothesis driven.

Bioinformatics and Statistical Analysis

Small RNA-seq pipeline analyses were performed using a previously published Docker-embedded software to guarantee the computational reproducibility of the analysis.[8] Trimmed reads were mapped against an in-house curated reference of human miRNAs based on miRbase v22 (Supplementary Table 1*A*). The alignment was performed using BWA algorithm v0.7.12.[12] miRNA levels were quantified using 2 methods called the "knowledge-based" and "position-based" methods, as described by Tarallo et al.[8] The sequences of the mature miRNAs were compared and, in the case of mature miRNAs characterized by identical sequences, the associated read counts were summed. An miRNA was considered as detected if supported by at least 20 normalized reads.

The age- and sex-adjusted differential expression analysis was performed using DESeq2 R package v1.22.2[13] using the likelihood ratio test method. For tissue samples, to test the significance of miRNA differential expression levels between CRC/adenoma tissue and matched adjacent nonmalignant colonic mucosa, a paired DESeq2 analysis was applied. An miRNA was considered differentially expressed (DEmiRNA) if associated with an adjusted *P* value of <.05 and a median number of reads of >20 in at least 1 study group. In each analysis in which the IT and CZ cohorts were analyzed together, the cohort variable was added to the DESeq2 model to adjust for the cohort batch effect.

Statistical analysis between continuous variables was performed using the Wilcoxon rank sum test or Kruskal-Wallis test. Statistical analysis between categorical variables was performed using the chi-square test.

Functional enrichment analysis was performed with RBiomirGS v0.2.12[14] in default settings and considering the validated miRNA-target interactions from miRTarBase and miRecord. A term was considered enriched if associated with an adjusted $P < .05$ and at least 2 target genes. The analysis was performed on the Kyoto Encyclopedia of Genes and Genomes (c2.cp.kegg.v7.5.1), Reactome (c2.cp.reactome.v7.5.1), WikiPathways (c2.cp.wikipathways.v7.5.1), Gene Ontology Biological Processes (c5.go.bp.v7.5.1), and Hallmark gene set libraries (h.all.v7.5.1) from MSigDB v7.5.1.[15] The analysis input was the average log2 fold change and combined adjusted P value computed by the differential expression analysis between the CRC and control groups of the IT cohort and CZ cohort.

Analysis of the copy number variation data from the COAD cohort of The Cancer Genome Atlas was performed by retrieving the GISTIC score from CBioPortal v4.1.15 (https://www.cbioportal.org/) considering the dataset named "Colorectal Adenocarcinoma (TCGA, PanCancer Atlas)."

Functional analysis of signature miRNA target genes was performed using Enrichr (version March 29th, 2021)[16] considering the validated targets provided by miRTarBase. A Gene Ontology Biological Process was considered enriched if associated with a $P < .001$. Because miR-607-5p was a novel miRNA identified in this study, its putative targets were predicted using miRanda v3.3a.[17] to scan the human 3′ untranslated region sequences from Ensembl v109. Among the 3807 potential targets identified, the top 100 genes characterized by the highest binding score were used for the analysis.

The correlation analysis between fecal miRNA levels and microbial abundances was performed by reanalyzing the small RNA-seq and shotgun metagenomic data from Thomas et al.[2] Preprocessing of metagenomic data was performed following the procedures described by Thomas et al[2] and Wirbel et al.[3] Specifically, raw reads quality controlled, adapter removal, and removal of human and PhiX reads were performed using the pipeline available at https://github.com/SegataLab/preprocessing. Then, taxonomic profiling was performed with MetaPhlAn3 in default settings with mpa_v30_CHOCOPhlAn_201901 as the markers database. Correlation analysis was performed using the Spearman method and graphically represented using the *corrplot* R package.

Explainable Machine Learning Approach

The 3-phase explainable ML approach to identify the minimal miRNA predictive signature is shown in Supplementary Figure 1. The 3 phases of the workflow were data preparation, feature selection and classification.

The data preparation phase has been designed to make the data usable to the ML approach and consists of (1) dataset loading and encoding, (2) dataset splitting in training and test sets, and finally (3) feature z-score normalization. The input data consist of a list of N individuals associated with the pathologic category, characterized by a set of covariates (eg, age and sex) and by a count matrix of dysregulated miRNAs. Once loaded and encoded, the dataset is represented by a matrix X paired with a vector Y. Matrix X is composed of N $\times$ M real numbers, where N is the number of individuals that are described by M features, which are either miRNAs or covariates. Vector Y is of length N as well and contains the encoded pathologic category of each participant represented in X.

The dataset is divided into training and test sets (with a given proportion of individuals, eg, 70% vs 30%). The former set is used to train ML models, and the latter is used only to evaluate the model performances. During the dataset split, a stratification of the participants according to the pathologic category and specific confounding covariates (eg, sex, age, disease stage) is performed. This guarantees that the proportion of pathologic categories of the whole dataset is maintained in both the training and test sets.

Finally, a z-score normalization is applied. The mean and standard deviation of all the features of the training set are estimated and used to normalize both the training and the test set.

The feature selection phase identifies the most relevant and nonredundant features in the distinction of the participants between groups of interest. To identify the k-best features from a given dataset, multiple selection criteria are available.[18] Specifically, filter methods assess feature relevance by computing a score between each feature and the target variable, whereas embedded strategies are based on learning algorithms that have built-in feature selection mechanisms. Hereby, the analysis of variance F test and mutual information were adopted as filter methods, whereas the embedded methods were based on logistic regression and random forest.

A repeated stratified k-fold cross-validation setting is adopted to apply the selection criteria on different subsamples of the training set to avoid—or at least reduce—data overfitting.

For this study, the whole procedure was repeated 30 times for any k from 1 to 25 to test feature sets composed of an increased number of DEmiRNAs. Each feature set was evaluated by a classification procedure, described later in this section, to identify its average performance.

The final selection was performed by means of a utility function, peak of the AUC(k), that guarantees the best balance between the AUC and the number of features selected—namely, to select the minimal number of miRNAs providing the best performance—that ultimately constitutes the miRNA predictive signature.

The classification phase is used to predict the qualitative response for a given individual to a category, according to the miRNA signature previously identified. Hereby, 3 classifiers were selected and applied independently: random forest,[19] logistic regression,[20] and gradient boosting.[21] The classifiers were applied with default values for the hyperparameters. Specifically, for the random forest classifier, the parameters were num_trees = 100 and criterion = entropy, penalty = l2 was selected for the logistic regression, and num_trees = 100 was set for the gradient boosting classifier. The set of patients to be classified was partitioned using a stratified 10-fold cross validation. For each classifier, 100 independent runs were performed. The performance metrics for each classifier—AUC, accuracy, precision, and recall—were computed as average metrics among all runs performed.

This approach was implemented in Python 3 using the following libraries: scikit-learn,[18] pandas, and matplotlib library[22] for ML algorithms, dataset representation, and data visualization, respectively.

Overview of the MicroRNA Content in the Analyzed Sample Types

Fecal samples from the Italian cohort and Czech cohort. From the analysis of small RNA-seq experiments, an average of 86.50% ± 10.03% of reads passed the preprocessing phase, and an average of 1.32% ± 2.22% of reads were aligned to human miRNAs. The observed percentage of aligned reads is in line with previous small RNA-seq analyses of fecal miRNA content.[6,8] Despite all miRNA annotations that were used for the differential expression analysis, a threshold of 20 normalized reads was used to define an miRNA as detected in a specific sample. Using this threshold, on average, 421.97 ± 222.07 (range, 86–1516) miRNAs were detected in each sample.

Fecal samples from the validation cohort. From the analysis of small RNA-seq experiments on the validation cohort, an average of 95.58% ± 2.88% of reads passed the preprocessing phase, and an average of 1.14% ± 1.34% of reads were aligned to human miRNAs. An average of 440.73 ± 217.94 (range, 75–1713) fecal miRNAs were detected in these samples.

Plasma extracellular vesicle samples. From the small RNA-seq experiments on plasma EV samples, an average of 91.41% ± 9.85% of sequencing reads passed the preprocessing phase and, on average, 20.12% ± 11.56% were assigned to human miRNA annotations. The average number of miRNAs detected in these samples was 309.69 ± 90.40 (range, 252–1213).

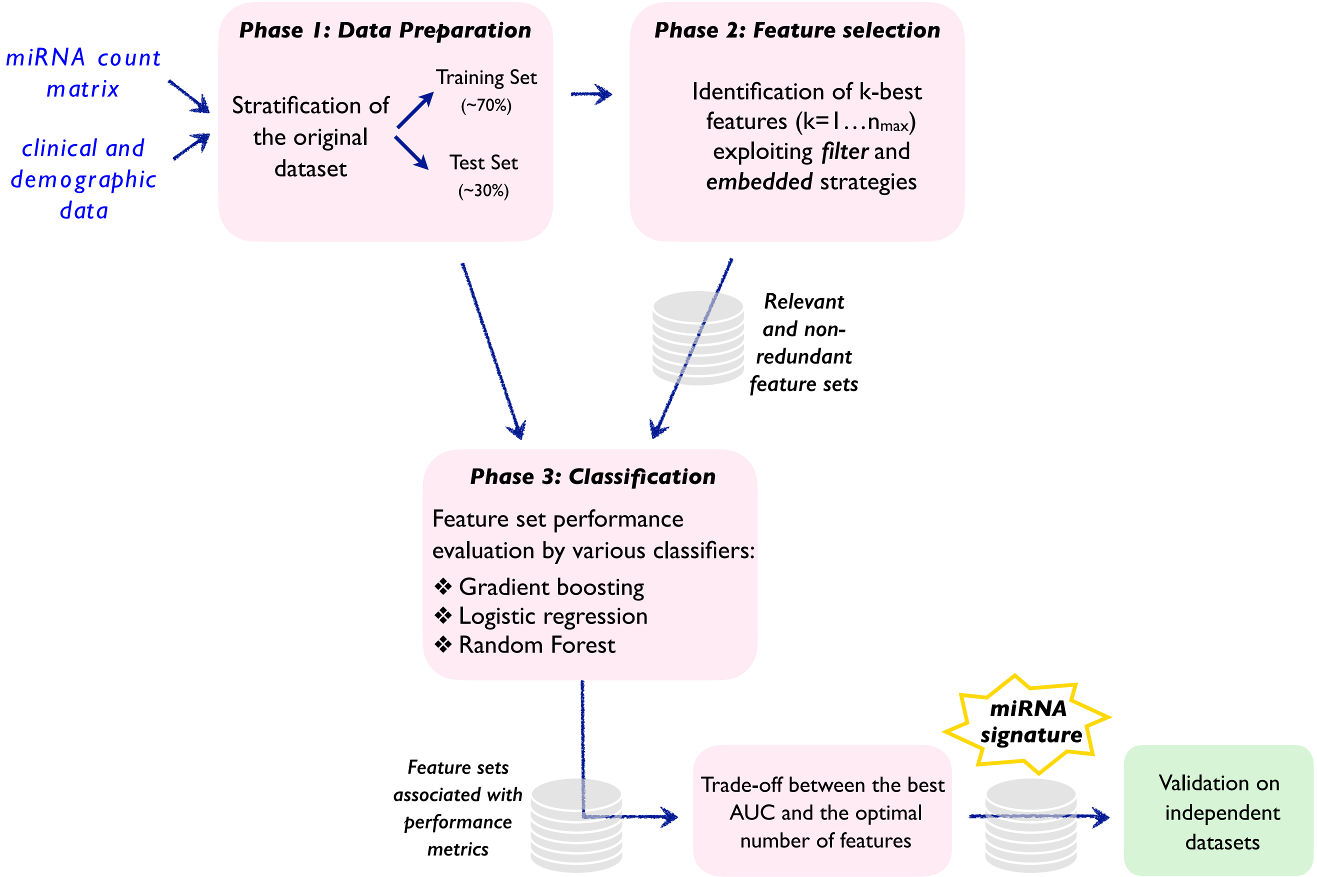
Tissue samples. In tissue samples, an average of 81.75% ± 13.01% sequencing reads were obtained from the preprocessing step, and among them, 68.56% ± 18.01% aligned on human miRNA annotations. On average, 581.84 ± 173.34 (range, 403–1997) miRNAs were detected in each sample.

Fecal immunochemical test leftover samples. From the small RNA-seq experiments on FIT leftover samples, an average of 90.30% ± 6.04% of sequencing reads passed the preprocessing phase, and, on average, 1.18% ± 0.49% were assigned to human miRNA annotations. The average number of miRNAs detected in these samples was 633.81 ± 41.07 (range, 541–744).

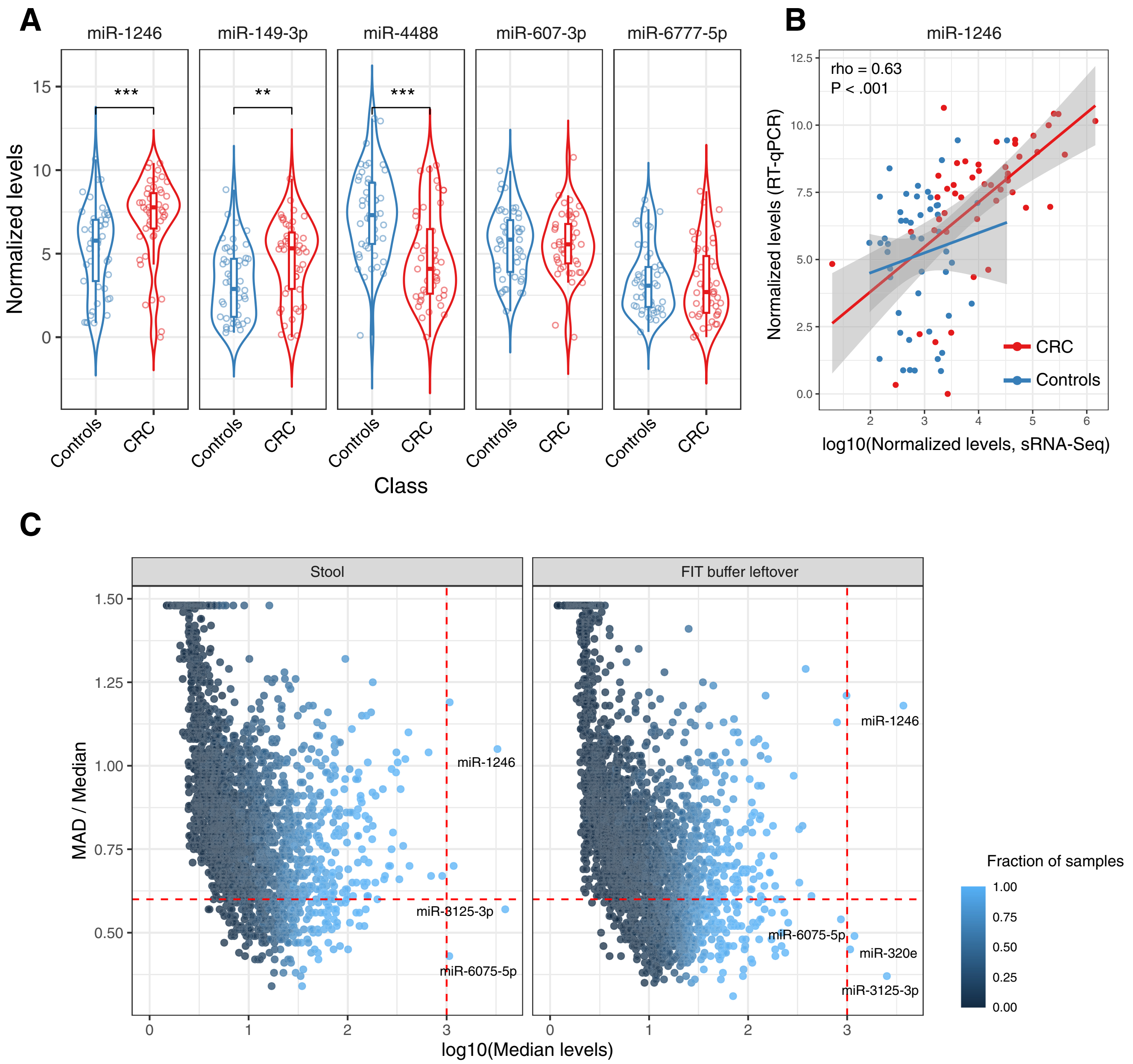
Supplementary References

1. Zarchy TM, Ershoff D. Do characteristics of adenomas on flexible sigmoidoscopy predict advanced lesions on baseline colonoscopy? Gastroenterology 1994;106:1501–1504.
2. Thomas AM, Manghi P, Asnicar F, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med 2019;25:667–678.
3. Wirbel J, Pyl PT, Kartal E, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med 2019;25:679–689.
4. Lin Y, Lau HC, Liu Y, et al. Altered mycobiota signatures and enriched pathogenic *Aspergillus rambellii* are associated with colorectal cancer based on multicohort fecal metagenomic analyses. Gastroenterology 2022;163:908–921.
5. Zwinsova B, Petrov VA, Hrivnakova M, et al. Colorectal tumour mucosa microbiome is enriched in oral pathogens and defines three subtypes that correlate with markers of tumour progression. Cancers (Basel) 2021;13(19):4799.
6. Tarallo S, Ferrero G, De Filippis F, et al. Stool microRNA profiles reflect different dietary and gut microbiome patterns in healthy individuals. Gut 2022;71:1302–1314.
7. Francavilla A, Ferrero G, Pardini B, et al. Gluten-free diet affects fecal small non-coding RNA profiles and microbiome composition in celiac disease supporting a host-gut microbiota crosstalk. Gut Microbes 2023;15(1):2172955.
8. Tarallo S, Ferrero G, Gallo G, et al. Altered fecal small RNA profiles in colorectal cancer reflect gut microbiome composition in stool samples. mSystems 2019;4(5):e00289-19.
9. Sabo AA, Birolo G, Naccarati A, et al. Small non-coding RNA profiling in plasma extracellular vesicles of bladder cancer patients by next-generation sequencing: expression levels of miR-126-3p and piR-5936 increase with higher histologic grades. Cancers (Basel) 2020;12(6):1507.
10. Ferrero G, Cordero F, Tarallo S, 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;9:3097–3111.
11. McCall MN, McMurray HR, Land H, et al. On non-detects in qPCR data. Bioinformatics 2014;30:2310–2316.
12. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754–1760.
13. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550.
14. Zhang J, Storey KB. RBiomirGS: an all-in-one miRNA gene set analysis solution featuring target mRNA mapping and expression profile integration. PeerJ 2018;6:e4262.
15. Liberzon A, Birger C, Thorvaldsdottir H, et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 2015;1:417–425.

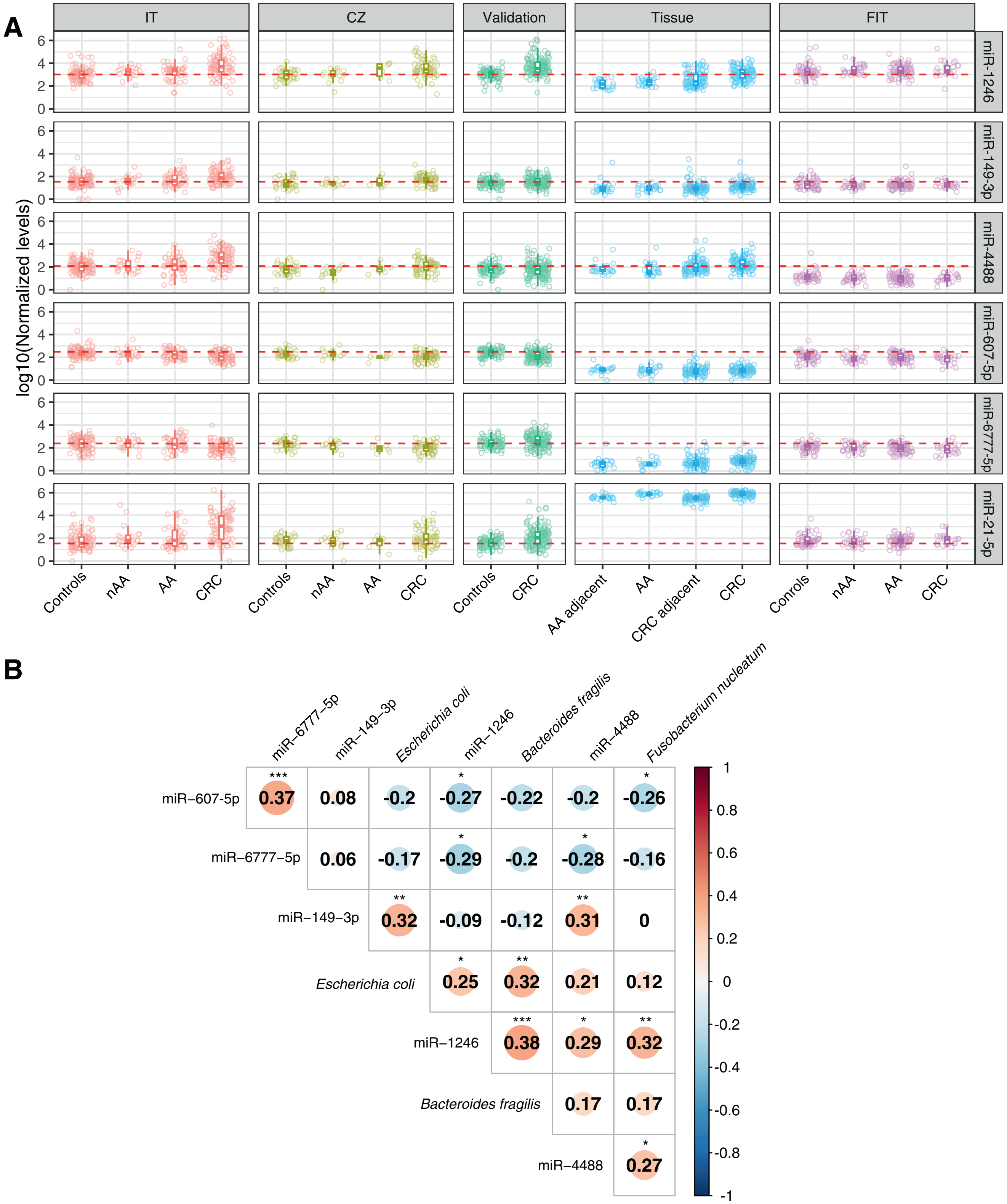
16. Xie Z, Bailey A, Kuleshov MV, et al. Gene set knowledge discovery with Enrichr. Curr Protoc 2021;1(3):e90.
17. Betel D, Koppal A, Agius P, et al. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome Biol 2010;11(8):R90.
18. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Machine Learn Res 2011;12:2825–2830.
19. Breiman L. Random forests. Mach Learn 2001;45:5–32.
20. Fan RE, Chang KW, Hsieh CJ, et al. LIBLINEAR: a library for large linear classification. J Mach Learn Res 2008; 9:1871–1874.
21. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189–1232.
22. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng 2007;9:90–95.


miRNA count matrix
clinical and demographic data
Phase 1: Data Preparation
Stratification of the original dataset
Training Set (~70%)
Test Set (~30%)
Phase 2: Feature selection
Identification of k-best features (k=1…n$_{max}$) exploiting *filter* and *embedded* strategies
Relevant and non-redundant feature sets
Phase 3: Classification
Feature set performance evaluation by various classifiers:
❖ Gradient boosting
❖ Logistic regression
❖ Random Forest
Feature sets associated with performance metrics
Trade-off between the best AUC and the optimal number of features
miRNA signature
Validation on independent datasets

Supplementary Figure 1. Schematic representation of the 3-phase explainable ML approach. An miRNA count matrix and the clinical/demographic data are the input data, and the best performing miRNA signature is the output.



Supplementary Figure 2. (*A*) Box plot showing the RT-qPCR normalized levels of the 5 miRNAs of the stool signature. *P* value by Wilcoxon rank sum test. ****P* < .001, ***P* < .01. (*B*) Scatterplot comparing the stool levels of miR-1246 measured by small RNA-seq (*x*-axis) and RT-qPCR (*y*-axis). The coefficient and significance of the Spearman correlation analysis is also reported. (*C*) Scatterplot reporting the median levels (*x*-axis) and the expression variability (as the ratio between median absolute deviation [MAD] and median, *y*-axis) of miRNAs measured in stool samples (*left plot*) or FIT buffer leftover (*right plot*) from the same subjects.



	miR-6777-5p	miR-149-3p	*Escherichia coli*	miR-1246	*Bacteroides fragilis*	miR-4488	*Fusobacterium nucleatum*
miR-607-5p	*** 0.37	0.08	-0.2	* -0.27	-0.22	-0.2	* -0.26
miR-6777-5p		0.06	-0.17	* -0.29	-0.2	* -0.28	-0.16
miR-149-3p			** 0.32	-0.09	-0.12	** 0.31	0
Escherichia coli				* 0.25	** 0.32	0.21	0.12
miR-1246					*** 0.38	* 0.29	** 0.32
Bacteroides fragilis						0.17	0.17
miR-4488							* 0.27

Supplementary Figure 3. (*A*) Box plots reporting, for each study cohort, the normalized levels of the 5 stool miRNAs belonging to our CRC-predictive signature. At the bottom, the levels of miR-21-5p are also reported. The red dashed lines refer to the median miRNA level measured in control individuals of the IT cohort. (*B*) Correlation plot representing the results of the Spearman correlation analysis between the levels of the 5 fecal miRNAs and *F nucleatum*, *E coli*, and *B fragilis* abundances by the reanalysis of data from Supplementary Reference.[2] The size of the dot is proportional to the absolute correlation coefficient. ****P* < .001; ***P* < .01; **P* < .05.