Journal Pre-proof

A fecal miRNA signature by small RNA sequencing accurately distinguishes colorectal cancers: results from a multicentric study

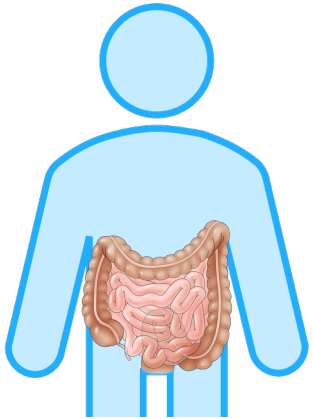
Barbara Pardini, Giulio Ferrero, Sonia Tarallo, Gaetano Gallo, Antonio Francavilla, Nicola Licheri, Mario Trompetto, Giuseppe Clerico, Carlo Senore, Sergio Peyre, Veronika Vymetalkova, Ludmila Vodickova, Vaclav Liska, Ondrej Vycital, Miroslav Levy, Peter Macinga, Tomas Hucl, Eva Budinska, Pavel Vodicka, Francesca Cordero, Alessio Naccarati

PII: S0016-5085(23)00811-9
DOI: https://doi.org/10.1053/j.gastro.2023.05.037
Reference: YGAST 65746

To appear in: *Gastroenterology*
Accepted Date: 24 May 2023

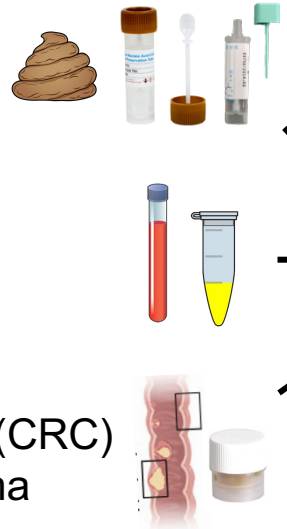
Please cite this article as: Pardini B, Ferrero G, Tarallo S, Gallo G, Francavilla A, Licheri N, Trompetto M, Clerico G, Senore C, Peyre S, Vymetalkova V, Vodickova L, Liska V, Vycital O, Levy M, Macinga P, Hucl T, Budinska E, Vodicka P, Cordero F, Naccarati A, A fecal miRNA signature by small RNA sequencing accurately distinguishes colorectal cancers: results from a multicentric study, *Gastroenterology* (2023), doi: https://doi.org/10.1053/j.gastro.2023.05.037.

Subjects



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

Samples



Analyses

miRNA profiling by small RNA-Seq of multiple biospecimens

Outcomes

- 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

A fecal miRNA signature by small RNA sequencing accurately distinguishes colorectal cancers: results from a multicentric study

Short Title: Stool miRNA signature and colorectal cancer

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#], Alessio Naccarati[1,2#+]

[1] Italian Institute for Genomic Medicine (IIGM), c/o IRCCS Candiolo, Candiolo 10060, Turin, Italy.
[2] Candiolo Cancer Institute, FPO IRCCS, Candiolo 10060, Turin, Italy.
[3] Department of Clinical and Biological Sciences, University of Torino, Turin 10100, Italy.
[4] Department of Computer Science, University of Torino, Turin 10100, Italy.
[5] Department of Surgical Sciences, "La Sapienza" University of Rome, Rome 00185, Italy.
[6] Department of Colorectal Surgery, Clinica S. Rita, Vercelli 13100, Italy.
[7] Epidemiology and Screening Unit-CPO, University Hospital Città della Salute e della Scienza, Turin 10100, Italy.
[8] LILT Lega Italiana Lotta contro i Tumori, associazione provinciale di Biella, Biella 13900, Italy.
[9] Department of Molecular Biology of Cancer, Institute of Experimental Medicine of the Czech Academy of Sciences, Prague 14220, Czech Republic.
[10] Institute of Biology and Medical Genetics, 1st Medical Faculty, Charles University, Prague 12800, Czech Republic.
[11] Biomedical Centre, Faculty of Medicine in Pilsen, Charles University, Pilsen 32300, Czech Republic.
[12] Department of Surgery, University Hospital and Faculty of Medicine in Pilsen, Charles University, Pilsen 32300, Czech Republic.
[13] Department of Surgery, First Faculty of Medicine, Charles University and Thomayer Hospital, Prague 14059, Czech Republic.
[14] Department of Gastroenterology and Hepatology, Institute for Clinical and Experimental Medicine, Prague 14021, Czech Republic.
[15] RECETOX, Faculty of Science, Masaryk University, Brno 62500, Czech Republic

+Corresponding authors:
Alessio Naccarati, PhD
Italian Institute for Genomic Medicine (IIGM)
c/o IRCCS Candiolo,
SP 142, Km 3.95,
10060 Candiolo,
Torino, Italy
alessio.naccarati@iigm.it

Barbara Pardini, PhD
Italian Institute for Genomic Medicine (IIGM)
c/o IRCCS Candiolo,
SP 142, Km 3.95,
10060 Candiolo,
Turin, Italy
barbara.pardini@iigm.it

*, # These authors equally contributed to the manuscript.

Acknowledgements
This work was supported by the Italian Institute for Genomic Medicine (IIGM) and Compagnia di San Paolo Torino, Italy (to AN, BP, and ST), by Lega Italiana per La Lotta contro i Tumori (to FC, BP and AN), by the Grant Agency of the Czech Republic (17-16857S to AN, PV, LV, VV and 22-05942S to VV, LV, TH and VL), by Grant Agency of the Ministry of Health of the Czech Republic (AZV NV18-03-00199 to VV, PV, LV, TH, and PM); by the Fondazione CRT (grant number 2017.2025 to FC) and by the COST action TRANSCOLONCAN (CA17118). PV and LV acknowledge support from the National Operation Programme (Natl. Institute for Cancer Research LX22NPO 05120). This project has received funding from the European Union`s Horizon 2020 research and innovation program under grant agreement No 825410 (ONCOBIOME project to AN, BP, FC and ST). The research leading to these results has received funding from AIRC under IG 2020 - ID. 24882 – P.I. Naccarati Alessio Gordon (to AN).

Authors contribution
Concept and design (BP, ST, and AN); collection and assembly of the data (BP, GF, ST, GG, VV, LV, VL, OV, ML, PM, EB, and AN); data analyses and interpretation (BP, GF, ST, AF, NL, CS, EB, FC, and AN); provisions of study materials or patients (GG, MT, GC, SP, VV, LV, VL, OV, ML, PM, TH, and PV); manuscript writing (BP, GF, FC, and AN); manuscript editing (ST, GG, AF, GG, NL, MT, GC, CS, SP, VV, LV, VL, OV, ML, PM, TH, EB, PV, and EB); all authors gave final approval of the version.

Conflict of interests
No relevant conflicts of interest exist for all the authors. This publication reflects only the author’s view and the European Commission is not responsible for any use that may be made of the information it contains.

Abstract

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 non-invasive CRC diagnosis.

Methods. A total of 1,273 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 CRC, 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 and 80 healthy volunteers. miRNA profiles were compared with those of 132 tumor/adenoma paired with adjacent mucosa, 210 plasma extracellular vesicles samples, and 185 fecal immunochemical tests (FIT) leftover samples.

Results. Twenty-five miRNAs showed altered levels in stool of CRC patients in both cohorts (adj. P<.05). A five-miRNA signature, including miR-149-3p, miR-607-5p, miR-1246, miR-4488, and miR-6777-5p, distinguished patients from controls (AUC=0.86, 95% CI=0.79-0.94) and was validated in an independent cohort (AUC=0.96, 95% CI=0.92-1.00). The signature classified controls from low-/high-stage tumors, and advanced adenomas (AUC=0.82, 95% CI=0.71-0.97). Tissue miRNA profiles mirrored those of stool samples, while fecal profiles of different gastrointestinal diseases highlighted miRNAs specifically dysregulated in CRC. miRNA profiles in FIT leftover samples showed good correlation with those of stool collected in preservative buffer and their alterations can be detected in adenoma or CRC patients.

Conclusions. Our comprehensive fecal miRNome analysis identified a signature accurately discriminating cancer aimed at improving a non-invasive diagnosis and screening strategies.

Keywords: stool microRNAs, non-invasive diagnosis, small RNA sequencing, colorectal cancer, precancerous lesions, machine learning

Introduction

In the last 30 years, we have witnessed a dramatic increase in understanding of the epidemiology, etiology, molecular biology, and various clinical aspects of colorectal cancer (CRC)[1]. However, ~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, as it may strongly increase the chances for successful treatment and cure.

Most European countries have implemented CRC screening programs based on non-invasive stool tests for detecting fecal occult blood, mainly the fecal immunochemical test (FIT)[4, 5]. FIT selects subjects showing a higher prevalence of CRC and advanced benign neoplasia but has limited sensitivity to recognize advanced colorectal adenomas (AA)[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 ~5% of individuals that receive a colonoscopy based on FIT results will end 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 non-invasively detectable fecal molecular biomarkers could improve the triage of subjects for colonoscopy, reducing the costs for the health systems in terms of the number of examinations and decreasing the risks and discomfort for the 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 subjects. In particular, explainable ML approaches allow identifying novel molecular biomarker signatures to improve early CRC diagnosis, as recently demonstrated for fecal microbial species[11] and urinary proteins[12].

The analysis of small non-coding RNAs in fecal samples has attracted interest with an excellent biological and analytical rationale for its application in large-scale clinical investigations[13]. Tumor-secreted small non-coding RNAs are directly and continuously released into the intestinal lumen, and their profiles may be altered in concomitance with the presence of CRC and precancerous lesions. Moreover, small non-coding 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 since their levels are altered in specific pathological 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 non-invasive 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 on 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 impact specific

miRNA levels[24, 25]. In addition, limited evidence is available on stool miRNA profiles in relation to patient clinicopathological characteristics such as specific CRC stages, in precancerous lesions or other gastrointestinal (GI) diseases, except for the reported pleiotropic dysregulation of miR-21-5p in several diseases[26]. Therefore, a miRNA signature for CRC detection derived from a comprehensive fecal miRNome analysis across multiple populations is currently lacking.

This multicentric study aimed to explore by deep sequencing miRNA profiles in stool samples that best characterize CRC patients from controls and distinguish colorectal adenomas or other GI diseases. The analyses were performed in different independent cohorts adopting the same protocol for subject recruitment, sample collection, and small RNA-Seq experiments/analyses. An integrated explainable ML strategy identified a fecal miRNA signature distinguishing CRC patient from controls 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 (IT) cohort – Stool specimens, and clinical and demographic data were collected from 317 subjects recruited in a hospital-based study at Vercelli, Italy (**Table 1**). Based on results of complete colonoscopy examination, participants were classified into: (i) 89 sporadic CRC patients; (ii) 74 polyp patients (6 hyperplastic polyps, 20 non-advanced adenomas (nAA), and 48 AA; serrated lesions were excluded since too few); (iii) 49 subjects with a GI disease (6 Crohn's disease, 9 ulcerative colitis, 14 diverticulitis, 7 diverticulosis, 13 hemorrhoidal disease); and (iv) 105 colonoscopy-negative control subjects. AA were defined based on the presence of high-grade dysplasia, villous component, or lesion size >1 cm as defined by[27]. Of this cohort, 93 stool samples (from 29 CRC patients, 27 polyps, 13 subjects with a GI disease, and 24 colonoscopy-negative controls) have been employed and described previously in other studies[11, 28, 29].

Czech (CZ) cohort – Stool specimens, and clinical and demographic data were collected from 162 Czech individuals recruited in two hospitals in Prague and one in Plzen, Czech Republic (**Table 1**). Based on colonoscopy results, subjects were divided in: (i) 66 CRC patients, (ii) 28 polyp patients (9 hyperplastic polyps, 13 nAA, 6 AA; no serrated lesions were collected); (iii) 32 patients with other GI diseases (3 Crohn's disease, 11 ulcerative colitis, 17 diverticulosis, 1 unclassified inflammatory bowel disease, IBD); and (iv) 36 colonoscopy-negative subjects.

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 subjects were previously described in other studies: the CRC population is described in[30], but here sequenced for the first time for small RNA-seq; healthy volunteers are a part of the cohorts described and sequenced for small non-coding RNAs in[24, 31].

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

More details on the cohorts included in the study are given in **Supplementary Materials**.

Local ethics committees of Azienda Ospedaliera SS. Antonio e Biagio e C. Arrigo of Alessandria (Italy, protocol N. 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 number 2018/865/MOU) and Masaryk University of Brno (Czech Republic, protocol number EKV-2019-044) approved the study. All patients gave written informed consent following the Declaration of Helsinki prior to participating in the study.

Other analyzed bio-specimens

For 132 subjects surgically operated at 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 *IT-cohort* subjects stratified in patients with CRC (n=52), AAs (n=19), nAAs (n=15), hyperplastic polyps (n=6), other GI diseases (n=39), and control subjects (n=79).

Sample collection

Naturally evacuated fecal samples were obtained from subjects 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 (200µl) 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.

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

Plasma samples were obtained from 8ml of blood centrifuged for 10min at 1000rpm, and aliquots were stored at -80°C until use. Plasma EVs were precipitated from 200µl of plasma using the ExoQuick (System Biosciences) according to[32].

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

Total RNA extraction, small RNA-Seq library preparation and Quantitative Real-Time PCR

Total RNA from stool and FIT leftovers 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[32]. For tissue samples, total RNA was extracted using QIAzol (Qiagen), according to the manufacturer's instructions.

Small RNAs were converted into barcoded cDNA 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 the reactions were run on an ABI Prism 7900 Sequence Detection System (Applied Biosystems). Analyses were performed as described in[33]. More details are provided in **Supplementary Materials**.

Computational and statistical analyses

Small RNA-Seq analyses were performed as described in[20], considering a curated miRNA reference based on miRBase v22 and including a characterization of novel miRNAs (**Supplementary Table 1A**). 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 (GLM) was defined by considering the miRNA levels as dependent variable and subject age, sex, Body-Mass Index (BMI), smoking habit, and cohort as independent variables.

A ML strategy was implemented to identify the optimal fecal miRNA signature to accurately classify CRC patients from control subjects. The ML approach is composed of three-phases: data preparation, feature selection, and classification (more details are provided in **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 *IT/CZ-cohort*). 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 *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 subjects characterized by specific covariates (i.e., age, sex, cohort, disease status, and tumor staging).

Other statistical tests were performed using Wilcoxon-Mann-Whitney and Kruskal-Wallis's (continuous variables) or chi-squared (categorical variables) methods. Benjamini-Hochberg (BH) 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 subjects (**Figure 1**). The applied analysis strategy included the following phases:

1) 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 and control subjects of both *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 a miRNA predictive signature.* A ML strategy identified a miRNA signature composed of the minimal set of DEmiRNAs that better distinguished CRC patients from controls 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.

2) Fecal DEmiRNA characterization in different sample types and diseases:

- *Assessment of DEmiRNA profiles in different biospecimens and clinical situations.* DEmiRNA levels were evaluated in: (i) tumor/adenoma tissue and adjacent mucosa; (ii) plasma EVs of CRC patients and control subjects; and (iii) 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 controls.
- *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 subjects based on the colonoscopy results.

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

Results

Stool miRNA profiles are altered in CRC patients: evidence from two 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 **Supplementary Materials** and **Supplementary Table 1B** and **C**). The age and sex-adjusted differential expression analysis between CRC and control subjects was performed independently on both *IT-cohort* and *CZ-cohort* identifying, respectively, 250 and 29 DEmiRNAs (median expression >20 reads and adj. *P*<.05) (**Figure 2A** and **Supplementary Table 2A**).

Twenty-five stool DEmiRNAs were in common between both cohorts (**Figure 2B**, **Table 2,** and **Supplementary Table 2A)**, all with a coherent expression trend (20 up-regulated and five down-regulated, rho=0.75, *P*<.001, **Figure 2B**). The alteration of these fecal miRNA levels in relation to CRC was further supported by a GLM analysis adjusted for cohort, age, sex, BMI, and smoking habits: 22 out the 25 DEmiRNAs remained significantly associated (*P*<.05) (**Supplementary Table 2B)**. DEmiRNA profiles were further explored in relation to CRC patient clinical data (**Figure 2C** and **D**). The levels of three down-regulated miRNAs (miR-607-5p, miR-677-5p, and miR-922-5p) significantly decreased with increasing tumor size (T) (**Figure 2D**). miR-922-5p also significantly decreased in patients with advanced disease stages or lymph node invasion (**Figure 2D** and **Supplementary Table 2C)**. 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 2D**).

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 2D** and **E**).

A fecal miRNA signature distinguishes CRC 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 3A**). The best miRNA signature distinguishing CRC patients from controls included miR-607-5p, miR-6777-5p, miR-4488, miR-149-3p, and miR-1246 (AUC=0.87±0.01; **Figure 2E**). This set of five miRNAs represented the best combination of not correlated molecules with the highest discriminative power. Moreover, they showed a good performance in the classification of the 30% of subjects excluded from the training set (AUC=0.81±0.01; **Figure 2F**). The classification improved after the inclusion of sex and age in the model (AUC=0.86±0.01; **Table 3** and **Supplementary Table 3B**). The performance of the signature was again tested in the *Validation-cohort*, where it remained fairly similar, irrespectively (AUC=0.91±0.01) or not (AUC=0.96±0.01) of age and sex (**Table 3**, **Figure 2F**, and **Supplementary Table 3B**).

By stratifying patients for CRC stage, the same five-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 subjects (*Validation-cohort*, AUC=0.95±0.01 and 0.87±0.01, respectively including or not age and sex; **Table 3,** and **Supplementary Table 3B**).

The panel of five miRNAs of the signature identified by sequencing was tested by RT-qPCR in RNA isolated from a subset of 96 stool samples equally distributed among *IT* and *CZ-cohort* subjects, with a balanced number of CRC cases and controls (**Supplementary Figure 2A**). The five 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 2A**). In particular, miR-1246 and miR-149-3p levels were significantly increased in patient samples. The same method was used to test the five miRNA levels in RNA from eight FIT leftover samples of subjects with positive FIT at the CRC screening: all miRNAs were detected also in this biospecimen (data not shown).

For four 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 3C** and **Supplementary Figure 2B**). Only the levels of miR-4488 were characterized by a negative correlation (rho=-0.48, $P<.001$) in CRC patients only.

Stool DEmiRNA profiles mirror those of primary CRC 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 resulted differentially expressed (adj. $P<.05$) in this comparison (**Figure 3A** and **Supplementary Table 4A),** with seven 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, three (miR-1246, miR-4488, miR-149-3p) were included in our miRNA signature. The five miRNAs significantly down-regulated in CRC patient stool (miR-607-5p, miR-6777-5p, included in the five-miRNA signature, miR-6076, miR-922-5p, and miR-9899) were poorly expressed (normalized reads <20) in both tumor and adjacent tissues (**Supplementary Table 4A**).

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 (adj. $P<.001$), while let-7i-5p and miR-4508 were down-regulated (**Figure 3A** and **Supplementary Table 4A**).

Few miRNA levels are dysregulated in circulating EVs of CRC patients

Small RNA-Seq was performed on RNA isolated from plasma EVs collected from 210 subjects in the *IT-cohort*, detecting an average of 309 (range=252-1213) miRNAs in these samples (**Supplementary Table 4B)**. 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, albeit the latter was associated with low levels (normalized reads <20) (**Supplementary Table 4B**). Another miRNA (miR-150-5p) was differentially expressed between CRC and control subjects (**Supplementary Table 4B**).

A subset of stool DEmiRNAs is specifically dysregulated in CRC 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 *IT*- and *CZ-cohorts*. The age-, sex-, and cohort-adjusted differential expression analysis between each disease category and control subjects showed that the levels of 21 out of the 25 CRC DEmiRNAs were significantly altered in at least another GI disease (**Figure 3B).** Notably, in patients with ulcerative colitis, diverticulitis, nAA, or AA, 60% of the CRC DEmiRNAs were also dysregulated (**Figure 3B** and **Supplementary Table 4C**). The lowest number of dysregulated miRNAs was observed in patients with Crohn's disease (two miRNAs) or diverticulosis (five miRNAs), while no DEmiRNAs were found in patients with hyperplastic polyps.

Considering the five miRNAs constituting our predictive signature to distinguish CRC from control subjects, miR-6777-5p was not differentially expressed (compared to control) in any other GI disease, miR-149-3p was significantly up-regulated only in patients with AA, while miR-607-5p was significantly down-regulated in patients with AA or ulcerative colitis, compared to control subjects (**Figure 3B** and **Supplementary Table 4C**). 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 subjects. Specifically, the miRNA signature was able to distinguish AA from controls, 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 controls (including or not age and sex, AUC=0.84±0.01 and 0.81±0.01, respectively) but not between them (CRC versus AA, AUC=0.68±0.02; **Table 3** and **Supplementary Table 3B**).

miRNAs are detectable in FIT leftover samples by small RNA-Seq

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 1B**). All the 25 stool DEmiRNAs were detected in this type of samples. Considering the threshold adopted by our pipeline (i.e., a minimum of 20 reads), four (miR-607-5p, miR-1246, let-7a-3p, miR-922) were detected in all samples and 18 in more than half (**Figure 3C** and **Supplementary Table 4D**). Three miRNAs included in our signature (miR-607-5p, miR-1246, miR-6777-5p) were detected in over 95% of samples (**Figure 3C**) while 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 stratifying subjects according to the colonoscopy results. Comparing the levels of the 25 stool DEmiRNAs between 46 subjects with a negative colonoscopy (excluding seven subjects with high hemoglobin levels) and 22 patients with CRC, eight (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 (adj. $P<.05$, **Supplementary Table 4E** and **Figure 3C**). 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 4F**).

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

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 1B** and **Supplementary Figure 2C**). 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 3C, Supplementary Figure 2C,** and **Supplementary Table 4D**). 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 subjects[25] (**Supplementary Figure 3A** and **Supplementary Table 4D**). The levels of all 25 stool DEmiRNAs positively correlated between the two specimens, with 13 of them reaching the statistical significance (including miR-607-5p, miR-1246, miR-149-3p, and miR-4488 from the five-miRNA signature, *P*<.05; **Figure 3C** and **Supplementary Table 4D**).

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

Discussion

In the present study, 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 controls. Given the pervasive detection across multiple cohorts, we confirmed previous findings about fecal miRNA potential use as non-invasive molecular biomarkers[23] (**Supplementary Table 1C** and **Supplementary Figure 3A)**. We also reported novel evidence on specific markers across different disease conditions. Notably, a fecal miRNA signature was able to distinguish CRC patients from controls accurately: 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 two 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 5A**)[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, three DEmiRNAs identified in our study (miR-4323-5p, miR-607-5p, and miR-922-5p) are not currently annotated in miRbase, but were quantified based on the read mapping position within the miRNA hairpin. This is consistent with the need for continuous refinement of miRBase annotations[37] and with evidence of new miRNAs reported by different groups[38, 39].

Consistently with their overall higher/lower levels in stool of CRC patients with respect to control subjects, the 25 DEmiRNA levels also increased/decreased with tumor size (T) 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 4C**). 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 2D** 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 five 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 (i.e., RT-qPCR). The accurate discrimination of both early and late cancer stages from controls confirmed the robustness of these five miRNAs for CRC detection. Although based on a small sample set, the signature could also discriminate AA from controls accurately (AUC=0.86), and in all analyses, high performances were obtained irrespectively by adjusting or not for sex and age, two 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 two 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], whose incidence is linked with unhealthy individual habits, such as a sedentary lifestyle[44].

Similarly to the functional analysis of all the 25 DEmiRNAs, a focused research on the five-signature miRNA target genes evidenced a prevalence of genes involved in cancer-related processes, including regulation of 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 5B** 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 research 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 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, while miR-149-3p was inversely related to *B. fragilis* abundances (**Supplementary Figure 3B**). 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, two 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 3B**).

To further explore the stool results, we tested DEmiRNA patterns in tumor and adenoma tissues paired with non-malignant 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, while 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 5D**), 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], while its circulating levels increased in relation to gut inflammation in IBD patients[55], coherently 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, while 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 non-coding 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 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 subjects. 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 EVs miRNome, given the above-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 stool DEmiRNA profiles of CRC patients with those of subjects 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, while 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 stool miRNA profiles of patients with AA and ulcerative colitis. Accordingly, recent studies showed altered miRNA profiles in fecal samples of patients with inflammations[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. These are the first evidence from a large-scale analysis of subjects 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 subjects with CRC or precursor lesions from controls and that, at least for cancer, data were confirmed in different cohorts, encourage their use to complement the existing non-invasive 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 subjects. 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 subjects with a positive or a negative colonoscopy outcome. In addition, miR-1246 and miR-607-5p from the five-miRNA signature deserve further investigation since they were detected in most of the samples and their levels respectively increased and decreased progressively going from colonoscopy negative subjects, 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 subjects must be carefully considered given the small group size analyzed so far, the lack of samples from FIT-negative subjects, and the fact that we cannot rule out the role of confounding factors, including subclinical diseases in the colonoscopy negative patients.

Most likely, 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 five-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 and colleagues proposed a two-stool miRNA-based classification signature (namely miR-27a-3p and miR-421) combined with hemoglobin levels, age, and sex of FIT-positive subjects. The signature accurately classified CRC (AUC=0.93) from control subjects, but was less efficient when AA patients were included (AUC=0.70)[62]. Differently from us, the authors initially selected miRNAs based on their differential expression between tumor tissue and adjacent mucosa and included in all models sex and age, two important risk factors for CRC. Hereby, we demonstrated the robustness of our signature since 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 five-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 non-invasive 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, overcome these issues.

The present study has several strengths: 1) the inclusion of independent cohorts from two countries with different diet and lifestyle habits as well as CRC rates; 2) the fact that cohorts were different for CRC clinical characteristics allowing the identification of accurate biomarkers independent on 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 different stool collection protocol showing its robustness; 5) the miRNome-wide approach in different biospecimens and different GI diseases contexts that has allowed us to discriminate miRNAs specifically dysregulated in 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 two cohorts were heterogeneous for individual

cancer categories. This heterogeneity could be responsible of the observed differences in the median stool miRNA levels and expression differences between the two cohorts. Given the difference in the clinical characteristics of CRC patients, the main driver of such 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 need 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 the dietary habits[24], the lack of dietary/lifestyle information of analyzed subjects may represent a limitation of the study. Follow-up studies with additional cohorts representing patients with different ethnicity, dietary patterns, and lifestyle habits are required but this is beyond the scope of this study, which represents the largest sequencing-based analysis of stool miRNAs, so far.

In conclusion, this multicentric 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, independently from age and sex. This may represent a fecal signature for improving the effectiveness of current non-invasive 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 the 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 including also 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 at 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 to deepen the research on gut-host crosstalk with small non-coding 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 to test the miRNA signature in a large clinical setting.

Figure legends

Figure 1. Representation of the study design.

Figure 2. A) Scatter plot reporting the stool miRNA average levels in CRC patients (y-axis) or control subjects (x-axis) from *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, while the size is proportional to the age, sex, and multiple testing adjusted *P*-values. **B)** Scatter plot reporting the correlations of log2FC of the 25 differentially expressed fecal miRNAs (DEmiRNAs) from the comparison between CRC and control subjects and in common between the *IT-cohort* (x-axis) and the *CZ-cohort* (y-axis). In red and blue are reported the up-regulated and down-regulated miRNAs, respectively. **C)** Heat map of stool DEmiRNA levels in CRC and control subjects of both cohorts. For each subject are reported: CRC stage and grade based on the AJCC system, presence of metastasis, lymph-node invasion status (pN), tumor size (pT), tumor localization, the cohort of origin, and disease status (CRC or controls). **D)** DEmiRNA levels comparing CRC patients stratified for clinical data. The dot color represents the log2FC, while the dot size is proportional to the statistical significance. Black borders represent tests with a $P<.05$. **E)** Line plot reporting the ability of different combinations of feature selection methods and classifiers to perform the classification of CRC and control subjects. Each dot represents an Area Under the Curve (AUC) obtained using a different number of fecal DEmiRNAs in input. **F)** Receiver operating characteristic (ROC) curves obtained for the classification of CRC and control subjects using the identified miRNA signature. Data are reported for the 30% of subjects excluded from the training set (**left**) and for the *Validation-cohort* (**right**).

Figure 3. Characterization of the 25 fecal DEmiRNAs in different sample types. **A**) Bar plot reporting the median levels in tumor, advanced adenomas (AA), and non-advanced adenomas (nAA) tissues. The color code represents the log2FC from the paired differential expression analysis between CRC/adenoma tissues and matched adjacent mucosa, ***adj. $P<.001$; **adj. $P<.01$; *adj. $P<.05$. **B)** Comparison of miRNA levels in stool of patients with CRC, colorectal adenomas, hyperplastic polyps, or other GI disorders with respect to control subjects. The dot color represents the log2FC, while dot size is proportional to the analysis significance. Black borders represent results with an adj. $P<.05$. **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 while the dot size is proportional to the analysis significance. Black borders represent a DESeq2 BH adj. $P<.05$; (*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.

References

1. Keum N, Giovannucci E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. Nature reviews. Gastroenterology & hepatology 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. International journal of 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. The New England journal of medicine 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 e3.
9. Loktionov A. Biomarkers for detecting colorectal cancer non-invasively: DNA, RNA or proteins? World journal of gastrointestinal oncology 2020;12:124-148.
10. **Weng M, Wu D, Yang C**, et al. Noncoding RNAs in the development, diagnosis, and prognosis of colorectal cancer. Translational research : the journal of laboratory and clinical medicine 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. Nature medicine 2019;25:667-678.
12. **Sun Y, Guo Z,** Liu X, et al. Noninvasive urinary protein signatures associated with colorectal cancer diagnosis and metastasis. Nature communications 2022;13:2757.
13. Francavilla A, Turoczi S, Tarallo S, et al. Exosomal microRNAs and other non-coding RNAs as colorectal cancer biomarkers: a review. Mutagenesis 2019.
14. Hombach S, Kretz M. Non-coding RNAs: Classification, Biology and Functioning. Advances in experimental medicine and biology 2016;937:3-17.
15. Di Leva G, Croce CM. miRNA profiling of cancer. Current opinion in genetics & development 2013;23:3-11.
16. Moridikia A, Mirzaei H, Sahebkar A, et al. MicroRNAs: Potential candidates for diagnosis and treatment of colorectal cancer. Journal of cellular physiology 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 2019;11.
19. Cervena K, Novosadova V, Pardini B, et al. Analysis of MicroRNA Expression Changes During the Course of Therapy In Rectal Cancer Patients. Frontiers in oncology 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.

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 e4.
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.
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 Biotecnologica 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.
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;11:20645.
26. Jenike AE, Halushka MK. miR-21: a non-specific biomarker of all maladies. Biomark Res 2021;9:18.
27. Zarchy TM, Ershoff D. Do characteristics of adenomas on flexible sigmoidoscopy predict advanced lesions on baseline colonoscopy? Gastroenterology 1994;106:1501-4.
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. 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.
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;15:2172955.
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 2020;12.
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;28:39.
34. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology 2014;15:550.
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. Advances in experimental medicine and biology 2016;937:153-70.
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:e118-27.
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:R57.
40. Wei EK, Giovannucci E, Wu K, et al. Comparison of risk factors for colon and rectal cancer. International journal of cancer 2004;108:433-42.
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:e132-42.

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. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association 2021;19:955-966 e61.
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 2019;12.
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;9:771.
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 2020.
48. **Fu A, Yao B, Dong T,** et al. Emerging roles of intratumor microbiota in cancer metastasis. Trends Cell Biol 2022.
49. **Cao Y, Wang Z, Yan Y,** et al. Enterotoxigenic Bacteroidesfragilis Promotes Intestinal Inflammation and Malignancy by Inhibiting Exosome-Packaged miR-149-3p. Gastroenterology 2021;161:1552-1566 e12.
50. Clay SL, Fonseca-Pereira D, Garrett WS. Colorectal cancer: the facts in the case of the microbiota. J Clin Invest 2022;132.
51. Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. Nature 2012;487:330-7.
52. Esquela-Kerscher A, Slack FJ. Oncomirs - microRNAs with a role in cancer. Nat Rev Cancer 2006;6:259-69.
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.
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: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.
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: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:otaa003.
59. **Verdier J, Breunig IR,** Ohse MC, et al. Faecal Micro-RNAs 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;20:307.
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 prevention research 2012;5:464-72.

63. Imperiale TF, Ransohoff DF, Itzkowitz SH, et al. Multitarget stool DNA testing for colorectal-cancer screening. N Engl J Med 2014;370:1287-97.

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	Average ± St.Dev.	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	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	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	Non-smoker	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*	Proximal			19	37				16	16	
	Distal			11	20				11	15	
	Rectum			18	28				6	34	
	na			32	6				0	1	
Polyp type	Tubular adenoma			18					19		
	Tubulo-villous adenoma			12					0		
	Tubular sessile			5					0		
	Hyperplastic polyp			6					9		
	na			31					0		

Adenoma type (AA/nAA)	AA			48					6		
	nAA			20					13		
pT (Combined)	T1-T2				27					20	
	T3-T4				54					43	
	Tis				0					1	
	na				7					2	
AJCC Staging	I				18					16	
	II				24					16	
	III				29					15	
	IV				5					14	
	na				13					5	
Grade	G1-G2				39					44	
	G3				38					18	
	na				12					4	
Metastasis (lymph node or distal)	No				49					52	
	Yes				31					11	
	na				9					3	
Other GI diseases	Crohn's disease		6					3			
	Ulcerative rectocolitis		9					11			
	Diverticulosis		7					17			
	Diverticulitis		14					0			
	Hemorrhoidal disease		13								
	na		0					1			

BMI=body mass index; CRC= colorectal cancer; GI=gastrointestinal;AA=advanced adenoma; nAA=non-advanced adenoma; AJCC=American Joint Committee on Cancer. *Numbers may be different from those of the of the subjects in each category due to the presence of multiple lesions.

Table 2. Expression levels and fold changes of the 25 stool DEmiRNAs in common between the IT- and CZ-cohorts.

				Median levels Healthy		Median levels CRC		log2FC		BH Adjusted p-value*	
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

*Age and sex adjusted analysis

Table 3. Performance of the five-miRNA predictive signature in the different comparisons.

Analysis details*							Precision		F1-score	
Comparison	**Validation set**	**AUC (Mean ± St.Dev.)**	**95% CI**	**Accuracy**	**Sensitivity**	**Specificity**	**Disease**	**Controls**	**Disease**	**Controls**
CRC vs Controls	IT-cohort + CZ-cohort**	0.86 ± 0.01	0.79-0.94	0.78	0.78	0.78	0.82	0.74	0.80	0.76
CRC vs Controls	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 Controls	IT-cohort + CZ-cohort**	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 Controls	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 Controls	IT-cohort + CZ-cohort**	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 Controls	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 Controls	IT-cohort + CZ-cohort**	0.84 ± 0.01	0.77-0.91	0.77	0.83	0.67	0.81	0.70	0.81	0.69
AA vs Controls	IT-cohort + CZ-cohort**	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 Controls	IT-cohort + CZ-cohort**	0.77 ± 0.02	0.65-0.89	0.73	0.62	0.81	0.67	0.77	0.64	0.79
nAA vs Controls	IT-cohort + CZ-cohort**	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**	0.68 ± 0.02	0.54-0.82	0.76	0.92	0.25	0.80	0.49	0.85	0.33

*Analysis includes age and sex covariates.

**30% of samples excluded from the training and matched by age, sex, cohort, and CRC stage

Study subjects

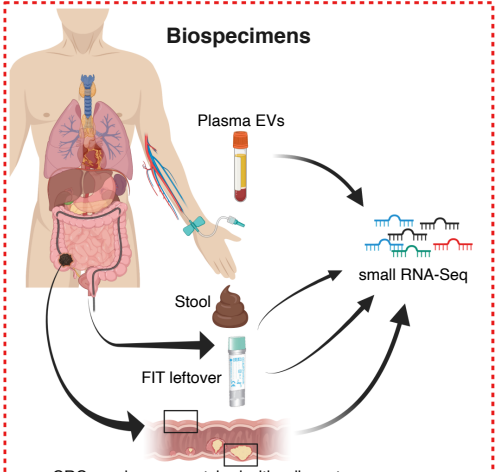
IT-cohort (n=317)

CZ-cohort (n=162)

Validation-cohort (n=221)

FIT-cohort (n=185)

Biospecimens



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

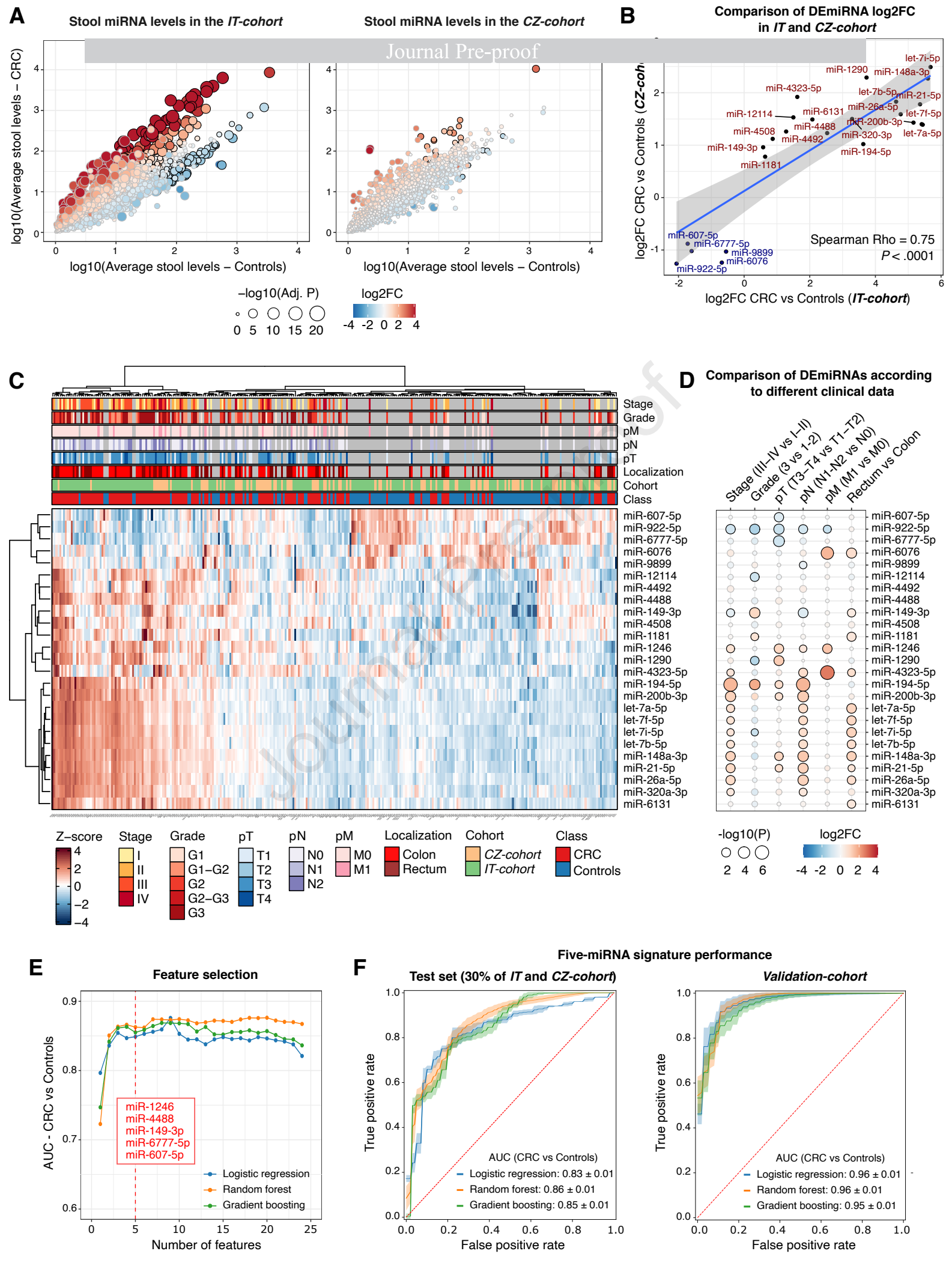
Fecal DEmiRNA characterization in different sample types and diseases

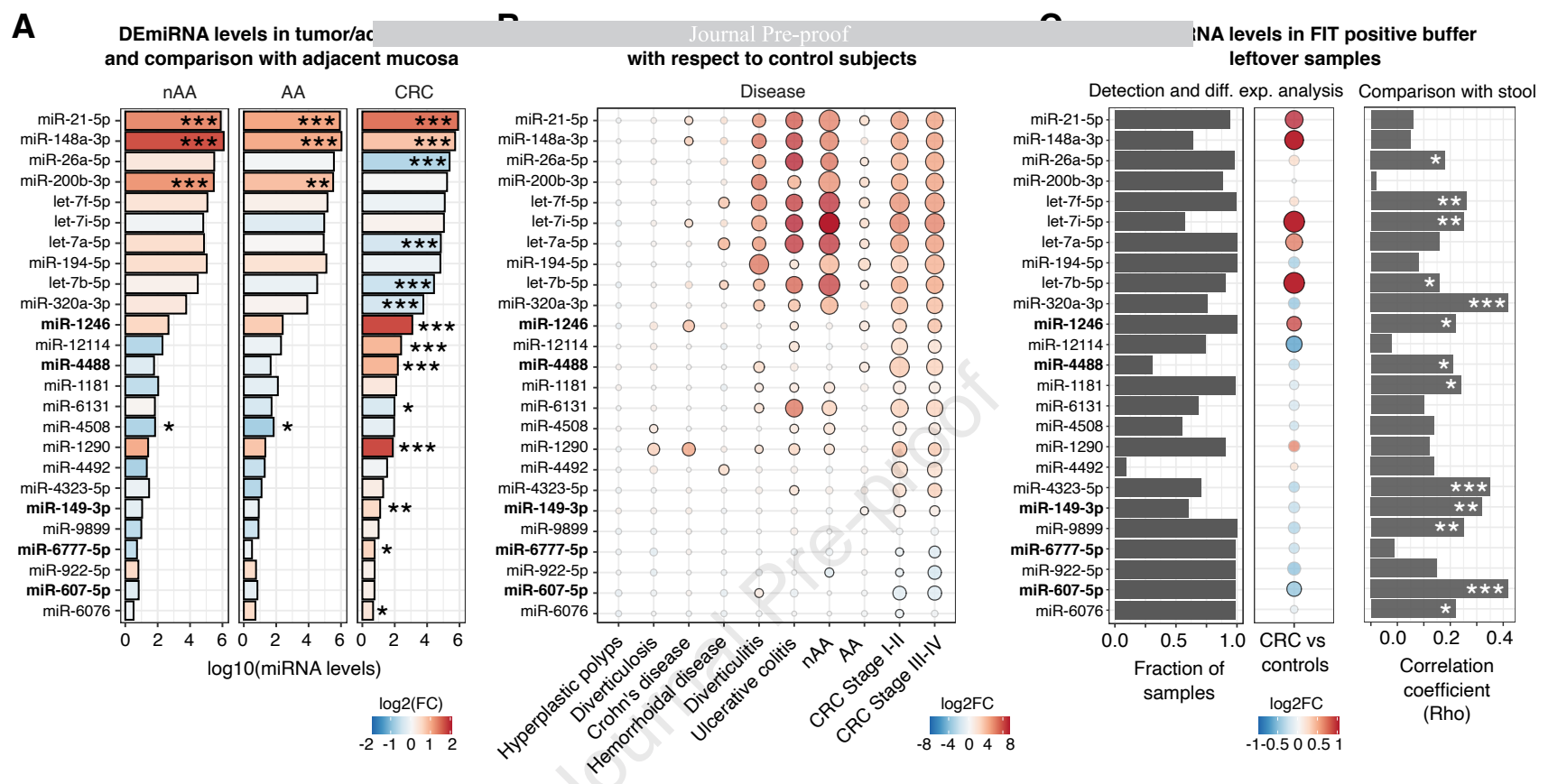
Colonic tissue (n=264)
Tumor vs. adjacent mucosa
Adenoma vs. adjacent mucosa

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





Supplementary methods

Stool study cohorts

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

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

Czech (CZ) cohort – Stool specimens, clinical and demographic data were collected for a cohort of 162 Czech individuals recruited in two hospitals in Prague and one in Plzen, Czech Republic (**Table 1, Figure 1A**). Based on colonoscopy results subjects were divided in: (i) 66 CRC patients, (ii) 28 individuals with colorectal polyps grouped in hyperplastic polyps (n=9), nAA (n=13), and AA (n=6); (iii) 32 patients with other GI disorders; and (iv) 36 colonoscopy-negative control subjects.

In both studies, colonoscopy was recommended for two main reasons: 1) because of the recommendation of the family doctor for various reasons (age of the patient, complaints in the gut, etc.); 2) because the patient had positive fecal immunochemical test (FIT; i.e., there was blood in the stool at the time of the test and therefore it was invited to have a colonoscopy in order to further investigate the reason for blood in stool). In any case, subjects with other major GI diseases than cancer were considered apart from those controls with negative colonoscopy.

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 recruited in a hospital in Brno, Czech Republic: these subjects were previously described in [5] and here 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 in [6, 7]. The healthy volunteers are derived from a subgroup of healthy subjects (no cancer, no precancer lesions) nested from the omnivorous group described in Tarallo et al.[6] and from Francavilla et al.[7]. Only subject with age >30 years were considered for the analysis.

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

Colonoscopy was recommended because the patients resulted with an abnormal or positive FIT levels (i.e., 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 bio-specimens

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 of *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 subjects of *IT-cohort* stratified in 52 patients with CRC, 19 with AAs, 15 with nAAs, 6 with hyperplastic polyps, 34 with other GI disorders, and 79 control subjects.

Sample collection
Naturally evacuated fecal samples were obtained from all subjects previously instructed to self-collect the specimen at home. For all the cohorts, stool samples were collected in nucleic acid collection and transport tubes with RNA stabilizing solution (Norgen Biotek Corp.) and returned to the endoscopy unit. Stool aliquots (200µl) were stored at -80°C until RNA extraction [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, Spain). Patients performed the collection at home, the morning of 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 (~1.2ml) 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 8ml of blood centrifuged for 10min at 1000rpm, and aliquots were stored at -80°C until use. Plasma exosomes/EVs were isolated from 200µl of plasma using the ExoQuick exosome precipitation solution (System Biosciences, Mountain View, CA, USA), according to the manufacturer's instructions [9, 10]. Briefly, plasma was mixed with 50.4µl of ExoQuick solution and refrigerated at 4°C overnight (at least 12h). The mixture was then further centrifuged at 1500g for 30min. The EVs pellet was dissolved in 200µl of nuclease-free water and RNA was extracted immediately from the solution.
Paired primary tumor/adenoma tissue and non-malignant adjacent mucosa were obtained from CRC and adenoma patients (at least 20cm distant) collected during surgical resection and immediately immersed in RNA later 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 in [9, 10]. For tissue samples, total RNA was isolated using QIAzol (Qiagen) after tissue homogenization performed with ULTRA-TURRAX® Homogenizer, followed by phenol/chloroform extraction according to the manufacturer's standard protocol.

Library preparation for small RNA sequencing (small RNA-seq)
Small RNA-seq libraries were prepared from RNA extracted from tissues, stool, and plasma EVs.as previously described in [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 cDNA libraries. For each library, 6 µl of RNA (35 ng for EVs RNA and 250 ng for tissue/stool RNA) were processed as starting material. Each library was prepared with a unique indexed primer. Multiplex adapter ligations, reverse transcription primer hybridization, reverse transcription 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 to assess DNA libraries quality regarding size, purity, and concentration. The obtained libraries were subjected to the Illumina sequencing pipeline on Illumina HiSeq4000 and NextSeq500 sequencers (Illumina Inc., USA).

Quantitative Real-Time PCR

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 two subsets of stool RNA from the *Italian (IT) cohort* (n=51), from the *Czech (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. Reverse transcription (RT) was performed using the miRCURY LNA RT kit (Qiagen) according to the manufacturer's instructions with the addition of one spike-in (UniSp6) to the RT reaction.

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

GenEx software (Multi-D) was used for data pre-processing including inter-plate calibration, evaluation of isolation and reverse transcription efficiency, setting specific cut-offs for negative control miRNA Ct values, and triplicates averaging. The analyses were performed calculating delta Ct (ΔCt) values by global mean. The fold-change was calculated as log 2–ΔΔCT between CRC (or BC subcategories) and control samples. miRNAs with a Ct value > 38 were deemed to be not detected. To avoid biased inference due to qPCR non-detects (Ct value = 40) a left-censoring approach was employed. 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< .05 were considered as statistically significant, since 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 1A**). The alignment was performed using BWA algorithm v0.7.12 [12]. miRNA levels were quantified using two methods called the "knowledge-based" and "position-based" methods as described in [8]. The sequences of the mature miRNAs were compared and in case of mature miRNAs characterized by identical sequences, the associated read counts were summed. A 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 (LRT) method. For tissue samples, to test the significance of miRNA differential expression levels between CRC/adenoma tissue and

matched adjacent non-malignant colonic mucosa, a paired DESeq2 analysis was applied. A miRNA was considered differentially expressed (DEmiRNA) if associated with an adjusted P-value< .05 and a median number of reads >20 in at least one study group. In each analysis in which IT and CZ-cohorts were analyzed together the cohort variable was added to the DESeq2 model to adjusted for cohort batch effect.

Statistical analysis between continuous variables was performed using Wilcoxon Rank Sum test or Kruskal-Wallis's test. Statistical analysis between categorical variables was performed using 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 adj. P<.05 and at least two target genes. The analysis was performed on KEGG (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 log2FC and combined adjusted P-value computed by the differential expression analysis between CRC and control groups of IT-cohort and CZ-cohort.

Analysis of the copy number variation data from 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 Processes was considered enriched if associated with a *P*<.001. Since 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'UTR sequences from Ensembl v109. Among the 3,807 potential targets identified, the top 100 gene characterized by the highest binding score were used for the analysis.

The correlation analysis between faecal miRNA levels and microbial abundances was performed by reanalyzing the small RNA-Seq and shotgun metagenomic data from [2]. Preprocessing of metagenomic data was performed following the procedures described in [2, 3]. Specifically, raw reads quality-controlled, adapter removal, and removal of human and PhiX reads was 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 markers database. Correlation analysis was performed using the Spearman method and graphically represented using the *corrplot* R package.

Explainable Machine Learning approach

The three-phases explainable machine learning approach to identify the minimal miRNA predictive signature is shown in the **Supplementary Figure 1**. The three 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 machine learning (ML) approach and consists in: a) dataset loading and encoding, b) dataset splitting in training and test sets, and finally c) feature Z-score normalization. The input data consists of a list of N subjects associated with the pathological category, characterized by a set of covariates (e.g., 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. The matrix X is composed of N x M real numbers, where N is the number of subjects that are described by M features, which are

either miRNAs or covariates. The vector Y is of length N as well and contains the encoded pathological category of each subject represented in X.
The dataset is divided into training and test sets (with a given proportion of individuals, e.g., 70% versus 30%). The former set is used to train ML models, while the latter is used only to evaluate the model performances. During the dataset split a stratification of the subjects according to the pathological category and specific confounding covariates (e.g., sex, age, disease stage) is performed. This guarantees that the proportion of pathological 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 were estimated, and used to normalize both the training and the test set.

The *feature selection* phase identifies the most relevant and non-redundant features in the distinction of the subjects between groups of interest. To identify the k-best features from a given dataset multiple selection criteria are available[18]. Specifically, *filter* methods assess features relevance by computing a score between each feature and the target variable, while *embedded strategies* are based on learning algorithms that have built-in feature selection mechanisms. Hereby, ANOVA 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 (CV) 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 below, 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 Area Under the Curve (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 subject to a category, according to the miRNA signature previously identified. Hereby, three 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 random forest classifier, the parameters were num_trees=100 and criterion=entropy while 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 CV. For each classifier, 100 independent runs were performed. The performance metrics for each classifier, AUC, accuracy, precision, and recall) were computed as an average metrics among all runs performed.

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

Overview of the miRNA content in the analyzed sample types
Faecal samples from IT-cohort and CZ-cohort. From the analysis of small RNA-seq experiments an average of 86.50±10.03% of reads passed the preprocessing phase while 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 faecal miRNA content [6, 8]. Despite all miRNA annotations were used for the differential expression analysis, a threshold of 20 normalized reads was used to define a 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.

Faecal 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 while an average of 1.14±1.34% of reads were aligned to human miRNAs. An average of 440.73±217.94 (range: 75-1,713) faecal miRNAs were detected in these samples.

Plasma EV 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: 252-1,213).

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: 403-1,997) miRNAs were detected in each sample.

FIT 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 Figure legends

Supplementary Figure 1. Schematic representation of the three-phase explainable machine learning approach. A miRNA count matrix and the clinical/demographic data are the input data, while the best performing miRNA signature is the output.

Supplementary Figure 2. A. Box plot showing the RT-qPCR normalized levels of the five miRNAs of the stool signature. P-value by Wilcoxon Rank-Sum test. ***P<.001, **P<.01. **B**. Scatter plot 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**. Scatter plot reporting the median levels (x-axis) and the expression variability (as the ratio between MAD and median, y-axis) of miRNAs measured in stool samples (*left plot*) or FIT buffer leftover (*right plot*).

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

References

1. Zarchy TM, Ershoff D. Do characteristics of adenomas on flexible sigmoidoscopy predict advanced lesions on baseline colonoscopy? Gastroenterology 1994;106:1501-4.
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.
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;15:2172955.
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.
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 2020;12.
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;30:2310-6.
12. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754-60.
13. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology 2014;15:550.
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 systems 2015;1:417-425.
16. Xie Z, Bailey A, Kuleshov MV, et al. Gene Set Knowledge Discovery with Enrichr. Curr Protoc 2021;1: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:R90.
18. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011;12:2825-2830.
19. Breiman L. Random forests. Machine Learning 2001;45:5-32.
20. Fan RE, Chang KW, Hsieh CJ, et al. LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research 2008;9:1871-1874.
21. Friedman JH. Greedy function approximation: A gradient boosting machine. Annals of Statistics 2001;29:1189-1232.
22. Hunter JD. Matplotlib: A 2D graphics environment. Computing in Science & Engineering 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:

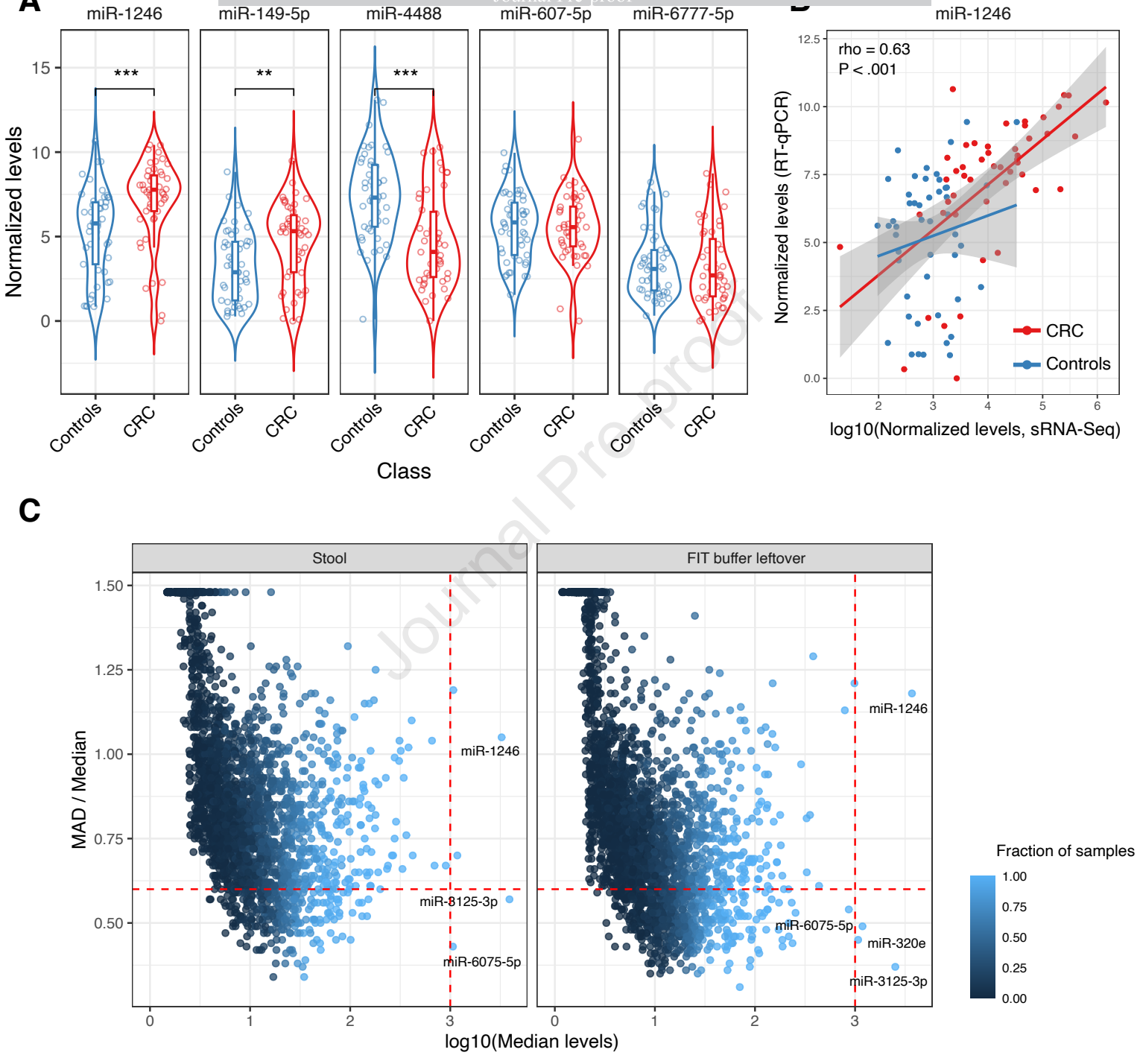
- ❖ Gradient boosting
- ❖ Logistic regression
- ❖ Random Forest

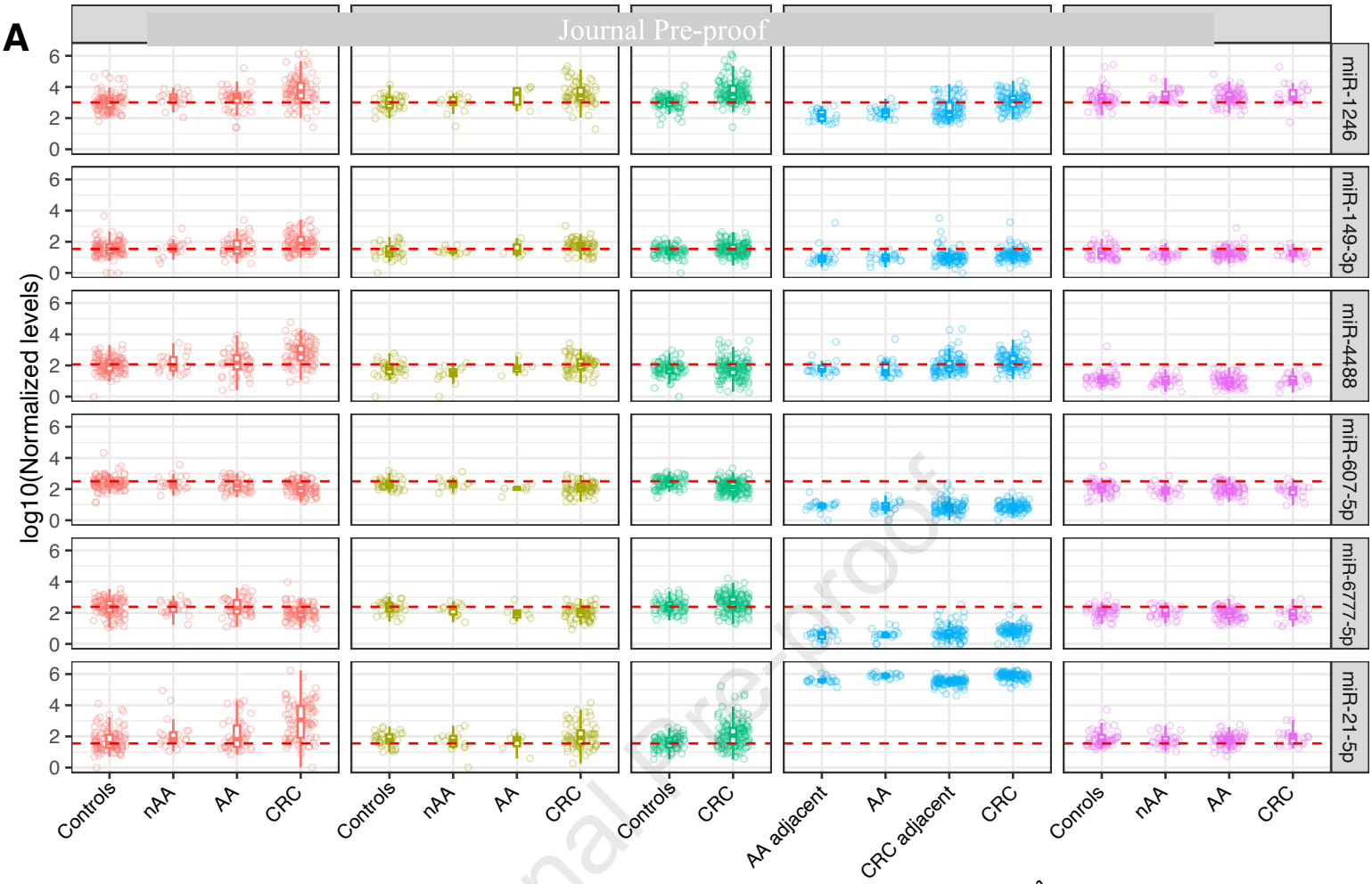
Feature sets associated with performance metrics

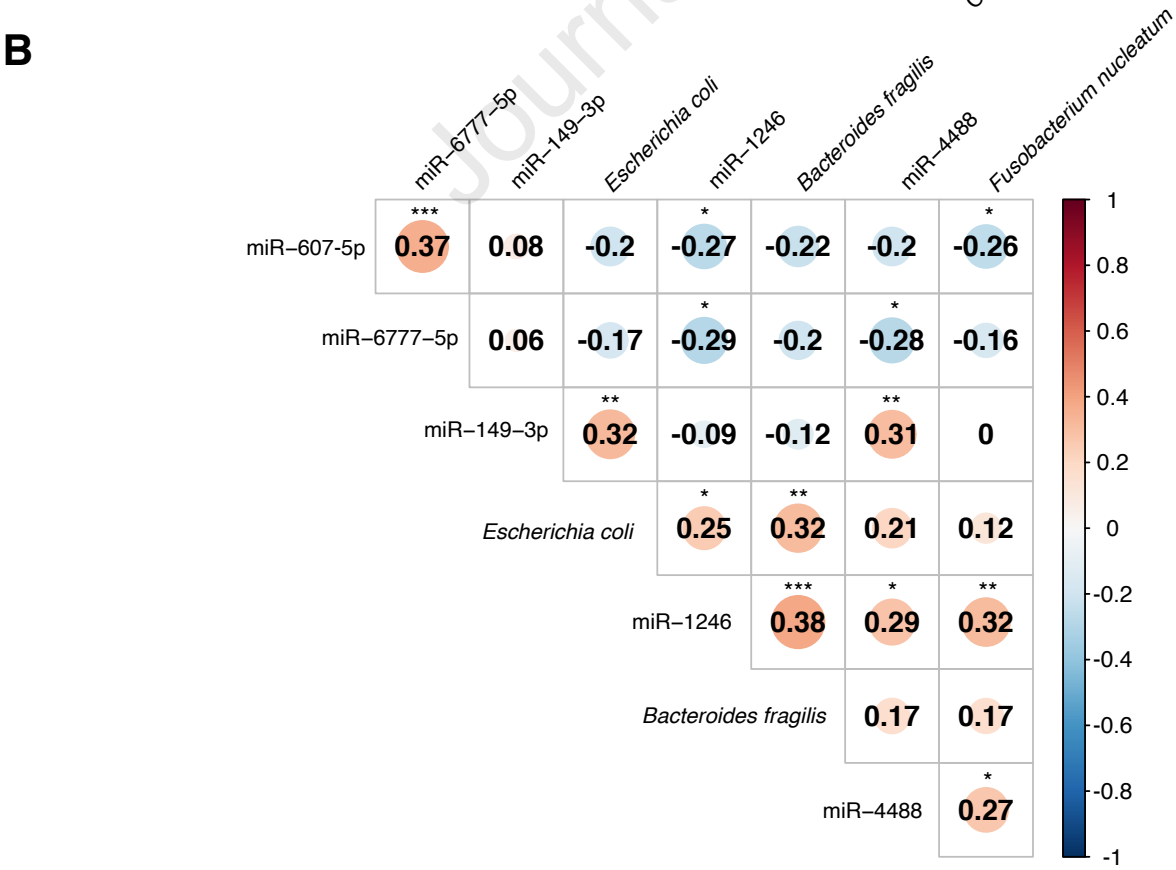
Trade-off between the best AUC and the optimal number of features

miRNA signature

Validation on independent datasets







What You Need to Know.

BACKGROUND AND CONTEXT

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

NEW FINDINGS

Investigating the whole miRNome in stool and with an ad hoc explainable machine learning, we identified in two independent cohorts five miRNAs that could accurately classify CRC from control subjects. 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, still the disease subtypes investigated were not exhaustive of heterogeneity in CRC and adenomas. Although we showed the feasibility of the molecular analysis, the investigation on screening samples represents still a pilot approach.

CLINICAL RESEARCH RELEVANCE

The investigation of the whole miRNome in all the cohorts led to obtain 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 discriminate different stages of CRC development and it constitutes a co-adjuvant to current screening programs for a non-invasive accurate diagnosis.

Lay summary

miRNAs are involved in colorectal carcinogenesis. This study detected their altered profiles in stool samples of cancer patients and identified a signature capable to accurately discriminate cancer patients