Supplementary Tables

Supplementary Table 1. Features employed in model development.

Features	
seRNA Biomarkers	**8 seRNA Biomarkers:** *GAPDH, ACY1, AREG, EGLN2, TNFRSF10B, KRAS, SMAD4,* and *CDH1*
Smoking Status	**1 Enumerated Status:** Current smoker, Previous smoker, or Never smoker
FIT Status	**1 Enumerated Status:** Positive, Negative, or Invalid

Supplementary Table 2. Demographics associated with the prospective cohort (n = 1,305).

Demographics	Total (n = 1,305)	Training Set (n = 939)	Testing Set (n = 366)
Age			
45-54	817 (62.6%)	597 (63.6%)	220 (60.1%)
55-64	418 (32.0%)	288 (30.7%)	130 (35.5%)
65-74	65 (5.0%)	52 (5.5%)	13 (3.6%)
75+	5 (0.4%)	2 (0.2%)	3 (0.8%)
Smoking			
No, I have never smoked	756 (57.9%)	551 (58.7%)	205 (56.0%)
No, I have smoked in the past, but quit	411 (31.5%)	282 (30.0%)	129 (35.2%)
Yes, I currently smoke	138 (10.6%)	106 (11.3%)	32 (8.7%)
Sex			
Female	816 (62.5%)	607 (64.6%)	209 (57.1%)
Male	482 (36.9%)	328 (34.9%)	154 (42.1%)
No Answer	7 (0.5%)	4 (0.4%)	3 (0.8%)
Ethnic Background			
African American	170 (13.0%)	120 (12.8%)	50 (13.7%)
Asian	29 (2.2%)	19 (2.0%)	10 (2.7%)
Hispanic / non-white	73 (5.6%)	46 (4.9%)	27 (7.4%)
White	984 (75.4%)	722 (76.9%)	262 (71.6%)
Other	40 (3.1%)	28 (3.0%)	12 (3.3%)
Prefer not to answer	9 (0.7%)	4 (0.4%)	5(1.4%)
Average Income			
$200,000 or More	32 (2.5%)	23 (2.4%)	9 (2.5%)
$150,000-$199,999	52 (4.0%)	37 (3.9%)	15 (4.1%)
$100,000-$149,999	185 (14.2%)	134 (14.3%)	51 (13.9%)
$75,000-$99,999	175 (13.4%)	130 (13.8%)	45 (12.3%)
$50,000-$74,999	241 (18.5%)	168 (17.9%)	73 (19.9%)
$30,000-$49,999	237 (18.2%)	177 (18.9%)	60 (16.4%)
Under $29,999	279 (21.4%)	198 (21.0%)	81 (22.1%)
Prefer not to answer	104 (8.0%)	72 (7.7%)	32 (8.7%)
Insurance			
No Insurance	1 (0.1%)	0 (0.0%)	1 (0.3%)
Private Insurance	974 (74.6%)	693 (73.8%)	281 (76.8%)
Public Insurance (Medicaid)	142 (10.9%)	109 (11.6%)	33 (9.0%)
Public Insurance (Medicare Advantage)	11 (0.8%)	9 (0.9%)	2 (0.5%)
Public Insurance (Medicare)	161 (12.3%)	116 (12.3%)	45 (12.3%)
Self-Insured	16 (1.2%)	12 (1.3%)	4 (1.1%)

Supplementary Methods

Eligibility requirements (Prospective and Retrospective)

All participants recruited for this study underwent extensive evaluation to ensure eligibility. It was required that patients meet the following criteria to be considered eligible:

- Participant is male or female, 45-84 years of age, inclusive.
- Participant is able to understand the study procedures, is able to provide verbal consent to participate in the study, and authorizes release of relevant protected health information through reviewing and verbally consenting to a HIPAA-compliant medical release form.
- Participant is able and willing to provide a stool sample and subsequently undergo a screening colonoscopy.

Exclusionary criteria (Prospective)

For participants recruited to the prospective cohort, participants were excluded from analysis if any of the following applied:

- Participant's stool sample was received >96 hours after stool sample production.
- Participant's FIT was invalid based on lack of signal in the control strip.
- Participant had a colonoscopy in the past 10 years.
- Participant had findings on any previous colonoscopy including hyperplastic polyps of any size. (Note: Tissue biopsies that resulted in no histopathological findings were acceptable).
- Participant has a history or recent diagnosis of CRC or adenoma.
- Participant has a history of aerodigestive tract cancer.
- Participant has had a positive fecal occult blood test or FIT within the previous six (6) months.
- Participant has had a positive FIT-DNA test within the previous 3 years.
- Participant has had a prior colorectal resection for any reason other than sigmoid diverticular disease.
- Participant has had overt rectal bleeding (e.g., hematochezia or melena) within the previous 30 days. (Blood on toilet paper after wiping does not constitute rectal bleeding).
- Participant has a diagnosis or personal history of any of the following high-risk conditions for colorectal cancer:
    - Inflammatory bowel disease (IBD) including chronic ulcerative colitis (CUC) and Crohn's disease.
    - Greater than or equal to (>=) 2 first-degree relatives who have been diagnosed with colon cancer. (Note: first-degree relatives include parents, siblings, and offspring).
    - One first-degree relative with CRC diagnosed before the age of 60.
    - Familial adenomatous polyposis (also referred to as "FAP", including attenuated FAP).
    - Hereditary non-polyposis colorectal cancer syndrome (also referred to as "HNPCC" of "Lynch Syndrome").
    - Other hereditary cancer syndromes including but not limited to Peutz-Jeghers Syndrome, MYH-Associated Polyposis (MAP), Gardner's Syndrome, Turcot's (or Crail's) Syndrome, Cowden's Syndrome, Juvenile Polyposis, Cronkhite-Canada Syndrome, Neurofibromatosis, and Familial Hyperplastic Polyposis.

Inclusion / Exclusion criteria (Retrospective)

For participants to be recruited to the retrospective cohort it was required that the patient have been diagnosed with colorectal cancer via colonoscopy and histopathology. It was also required that the patient provide a stool sample prior to being treated via surgery or chemotherapy. For participants recruited to the retrospective cohort, participants were excluded from analysis if any of the following applied:

- Participant's stool sample was received >96 hours after stool sample production.
- Participant's FIT was invalid based on lack of signal in the control strip.

Transcript expression quantification from QXDx data

Once QXDx files were generated using the droplet reader, a threshold was determined using internal controls. The threshold determined for each biomarker was subsequently employed to all experimental wells across the plate. Concentration for each marker, after employing the predefined threshold, was determined by the QXDx software (Quantalife v1.7). Samples were considered to have failed quality metrics if the housekeeping gene (GAPDH) was below the limit of detection based on plate controls recommended by the manufacturer.[38,39]

Performance evaluation metrics for training and testing sets

To assess classification performance, a receiver operator characteristic (ROC) curve was generated. This curve iterates through the model output to assess sensitivity and specificity at varied thresholds. The ROC area under the curve (AUC) represents the total accuracy of the model whereby increased AUC implies increased sensitivity / specificity. Several metrics are used to assess the performance of a classification model:

- *Sensitivity / specificity*: Sensitivity is defined as the number of participants in a given category identified as positive by the FIT-RNA test divided by the total number of participants in a given category. Sensitivity is defined for participants with colorectal cancer (CRC), advanced adenomas (AA), and other non-advanced adenomas (ONA). Specificity is defined as the number of participants in a given category identified as negative by the FIT-RNA test divided by the total number of participants in a given category. Specificity is defined for participants with hyperplastic polyps and no findings on a colonoscopy.
- *Confidence interval*: 95% confidence intervals estimated the proportion with a dichotomous result or finding in a single sample (positive or negative). The confidence interval used was the binomial exact calculation of the proportion.[40]
- *Median accuracy*: For each fold within internal cross validation (ICV), a ROC AUC is generated with point sensitivities and specificities determined by a threshold identified in the sub-training set using the Youden's J statistic.[28] The median accuracy cites the point sensitivities or specificities of the fold (20% of the total training set) that attained the median ROC AUC across all five folds during internal cross validation.
- *Concatenated accuracy*: To summarize the performance of the models from all folds of internal cross validation, the predictions of the evaluation samples in all 5 sub-testing sets, each of which was evaluated only once, were concatenated and used to construct a single ROC AUC (*concatenated AUC)* or a single performance (*concatenated accuracy*). It should be noted that predictions were made using five unique models and thresholds that were intrinsic to the performance of the ICV-training sets (subsets sampled from the entire training cohort).

Model development using the training cohort

The inputs for the RNA-FIT algorithm include: 1) concentrations for the 8 seRNA biomarkers, 2) the patient's smoking status, and 3) the fecal immunochemical test (FIT) result (**Supplementary Table 1**). The ordinal regression classifier was trained to capture the ordinality of the three categories (negative findings, non-advanced adenomas, and advanced neoplasias) which reflect the progression of disease. An implementation of the ordinal logistic model using All-Threshold variant in mord (version 0.3) was employed.[41,42] During training, L-BFGS-B, which is an extension of the limited-memory BFGS algorithm, was used to minimize the logistic loss. The output from the model provides a predicted score for advanced neoplasias (advanced adenomas or colorectal cancer).

The ultimate binarization threshold for predicting advanced neoplasia was calculated by maximizing Youden's J statistic[28] which accounts for model specificity and sensitivity with constraints. The specificity is defined for the training cohort with hyperplastic polyps and no findings on a colonoscopy, and the sensitivity is defined for the training cohort with advanced neoplasms. To retain high specificity, a constraint was applied to the

specificity, which was allowed to vary between 0.85 and 1.0. The optimal threshold corresponding to the maximal Youden’s J statistic was picked from all possible points on the segment of the ROC curve that was bounded by the constrained specificity.