scientific reports

Check for updates

OPEN

Stool shed cell transcriptomics mirrors tumor biology and enables colorectal cancer diagnosis

Keren Bahar Halpern[1,5]✉, Ilan Kent[2,3,5], Oran Yakubovsky[1,2,3,5], Shani Ben-Moshe[1], Tal Barkai[1,4], Michal Fine[1], Roy Novoselsky[1], Alon Israeli[2,3], Ido Nachmany[2,3] & Shalev Itzkovitz[1]✉

Screening and molecular characterization of human intestinal pathologies such as colorectal cancer (CRC) currently depends on colonoscopy, an invasive procedure associated with risks and poor adherence. A non-invasive method that captures host molecular changes could improve early detection and monitoring of intestinal diseases. Transcriptomic profiling of shed intestinal cells in stool has shown potential in neonates but is limited in adults by the dominance of bacterial RNA. To address this, we combined microbial ribosomal RNA (rRNA) depletion with unique molecular identifier (UMI)-based RNA sequencing to enrich and quantify human transcripts in stool. Applying this method to samples from 54 CRC patients and 24 healthy controls, we profiled thousands of human genes per sample. Stool-derived gene expression distinguished CRC from control samples with high accuracy (AUC = 0.86) and strongly correlated with matched tumor tissue signatures. Notably, stool transcriptomes reverted to control-like patterns after tumor resection. Our approach offers a powerful, non-invasive alternative to current CRC diagnostics and enables molecular insights into tumor biology. This method could complement or replace existing screening tools and may be applicable to other gastrointestinal diseases.

Keywords Colorectal cancer, Microbial depletion, Non-invasive, Stool

Abbreviations

AUC	Area under the receiver operating curve
CRC	ColoRectal cancer
CT	Control
FIT	Fecal immunochemical test
IBD	Inflammatory bowel diseases
PCA	Principal Component Analysis
scRNAseq	Single-cell RNA sequencing
ST	Spatial transcriptomics

The current gold standard for diagnosing gastrointestinal diseases—including inflammatory bowel disease (IBD), celiac disease, and colorectal cancer (CRC)—relies on endoscopic procedures[1]. While effective, endoscopy is invasive, expensive, and requires burdensome preparation. Colonoscopy is recommended for all adults over age 45 to detect and remove malignant or pre-malignant lesions[2], yet adherence rates remain low (~ 50–60%)[3,4]. Non-invasive alternatives like the fecal immunochemical test (FIT) are widely used, but suffer from limited sensitivity and specificity, particularly for early-stage disease[5,6].

Molecular stool-based tests—including measurements of microRNAs[7], DNA mutations[8], methylation patterns[9], and microbiome composition[10]—have offered new diagnostic avenues. However, these methods do not capture the broader transcriptomic shifts associated with altered host cell states. A non-invasive approach that profiles human gene expression in stool could offer greater diagnostic power and deeper molecular insight.

[1]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [2]Department of General Surgery and Transplantation, Sheba Medical Center, Ramat Gan, Israel. [3]Faculty of Medicine and Health Sciences, Tel Aviv University, Tel Aviv, Israel. [4]Sheba Medical Center, Ramat Gan, Israel. [5]Keren Bahar Halpern, Ilan Kent and Oran Yakubovsky contributed equally to this work. ✉email: kerenb@weizmann.ac.il; shalev.itzkovitz@weizmann.ac.il

The gut continuously sheds large numbers of epithelial, immune, and stromal cells[11–14]. We previously showed that transcriptomic profiling of human cells from fecal washes can reveal molecular changes in IBD[13,14]—but these studies required bowel prep, which reduces bacterial load. In neonates, where stool is less microbially dense[15–17], host transcriptomes have been measurable. In contrast, the high bacterial content of adult stool has limited detection of human transcripts to a few dozen genes[18–21].

Microbial ribosomal RNA (rRNA) are the most abundant RNA species and can therefore compromise readout of mRNA. While microbial rRNA depletion has improved bacterial metatranscriptomic profiling[22–25], its potential for enhancing human transcript detection in stool has not been explored. Here, we combine rRNA depletion with UMI-based RNA sequencing to profile shed human cells from adult stool, enabling accurate, non-invasive CRC diagnosis and molecular characterization.

Results

Microbial ribosomal RNA depletion enables deep human mRNA profiling from stool

We developed a method to perform deep transcriptomic profiling of human mRNAs from stool by depleting abundant bacterial ribosomal RNA (rRNA) (Fig. 1A). Stool samples were self-collected at home, preserved in RNAlater[26], and total RNA was extracted. To enrich for human transcripts, we hybridized the RNA with DNA probes targeting microbial rRNA sequences and treated the samples with RNase H, which selectively degrades RNA strands in RNA–DNA hybrids. This process efficiently removes bacterial rRNA and enriches human mRNA (Methods).

We then applied poly(A)-capture, unique molecular identifier (UMI)-based RNA sequencing[27] to quantify gene expression in the shed human cells contained in the stool samples. rRNA depletion increased the number of detected human genes by an average of 6.1-fold (Fig. 1B, Supplementary Table 1), raising the median gene count from 392 ± 541 without depletion to 1745 ± 1050 with depletion ($p = 8.3e-5$). Gene yield remained stable in samples stored at − 20 °C for up to five days (Fig. 1C), indicating robustness for home collection. Our approach therefore enables non-invasive stool-based deep transcriptomic profiling of the intestinal molecular states in diverse pathophysiological conditions.

Colorectal cancer patients stool transcriptomics mirror tissue states in CRC

To demonstrate the utility of our method, we recruited a cohort of 54 patients with CRC (Fig. 2A,B, Supplementary Table 2, Methods), and 24 healthy controls. Patients provided stool samples before undergoing surgery for the removal of the tumor. For each patient, we acquired tissue samples from the tumor and from an adjacent non-malignant colonic tissue, 3–5 cm from the tumor margin (Fig. 2A,B, Supplementary Table 3). For a subset of 11 patients, we also acquired stool samples at least 1 month after surgery, to assess whether the molecular signals resemble those of healthy controls. Both tissue samples and stool samples strongly clustered according to the patients' clinical states (Fig. 2C,D). Notably, gene expression in the combined dataset revealed modules of genes that were either higher in the tissue samples compared to stool samples (*PIGR*, *CEACAM5* and *EPCAM*, Fig. 2C), or specific to the disease state, regardless of the sample type (tissues or stool). For example, the genes *CCL20*, *LYZ*, *IFITM2* and *CD44* were more highly expressed in both CRC tissue samples and stools compared to adjacent healthy tissue and control stools, whereas the genes *FABP1*, *CA1*, *KRT20*, *GUCA2A*, *AQP8* and *MUC2* were more highly expressed in the healthy tissues and in the control stools (Fig. 2C). We identified large sets of differentially expressed genes between the tumor tissues and the adjacent healthy tissues (Fig. 3A, Supplementary Table 4) and between the CRC patients' and control patients' stool samples (Fig. 3B, Supplementary Table 4). Tissue gene expression changes correlated with those previously measured in a single-cell RNA sequencing (scRNAseq) study[28] (Supplementary Fig. 1).

We next examined the gene expression changes between CRC and non-CRC in the tissues and in the stool samples (Fig. 3C). We found that the ratios of expression between CRC and healthy control stools significantly correlated with the ratios of expression between tumor and adjacent healthy tissue ($r = 0.29$, $p < 5e^{-112}$). The ratios between CRC and healthy control stools also strongly correlated with the ratios of expression in stools of CRC patients before and after tumor resection (Fig. 3D, $r = 0.42$, $p < 5.7e^{-75}$). Principal component analysis showed that post surgery stool samples cluster within the control stool samples (Supplementary Fig. 2A). Gene expression signature of post surgery stool resembled the one observed in control stool (Supplementary Fig. 2B,C). Stool shed cell transcriptomics therefore reflect a return to healthy molecular states after the removal of the tumor.

Stool transcriptomes reflect spatial gene expression patterns in tumor tissue

To further evaluate whether gene expression changes observed in stool samples reflect spatially localized tumor biology, we analyzed a high-resolution spatial transcriptomics (ST) dataset of colorectal cancer tissues[29] Genes upregulated in CRC patient stool samples were similarly elevated in tumor regions compared to adjacent non-malignant tissue (Supplementary Fig. 3, Supplementary Table 5) suggesting that stool-derived RNA captures spatially defined expression patterns within the tumor microenvironment.

For example, genes associated with healthy colonocytes—such as CA2 (Fig. 4A), ZG16, and AQP8 (Supplementary Fig. 4A,B)—were enriched in non-tumor tissue regions, elevated in stool samples from healthy controls, and decreased in CRC patient stool samples. These expression patterns reverted toward healthy levels in post-resection stool samples (Fig. 4B,C). Conversely, tumor-enriched genes such as IFITM2 (Fig. 4D), ARPC1B, and S100A11 (Supplementary Fig. 4C,D) were elevated in both tumor tissue and pre-surgery CRC stool samples but decreased after tumor removal (Fig. 4E,F). These findings demonstrate that stool transcriptomics of shed human cells faithfully mirrors spatial gene expression dynamics associated with malignant transformation in colorectal tissue.

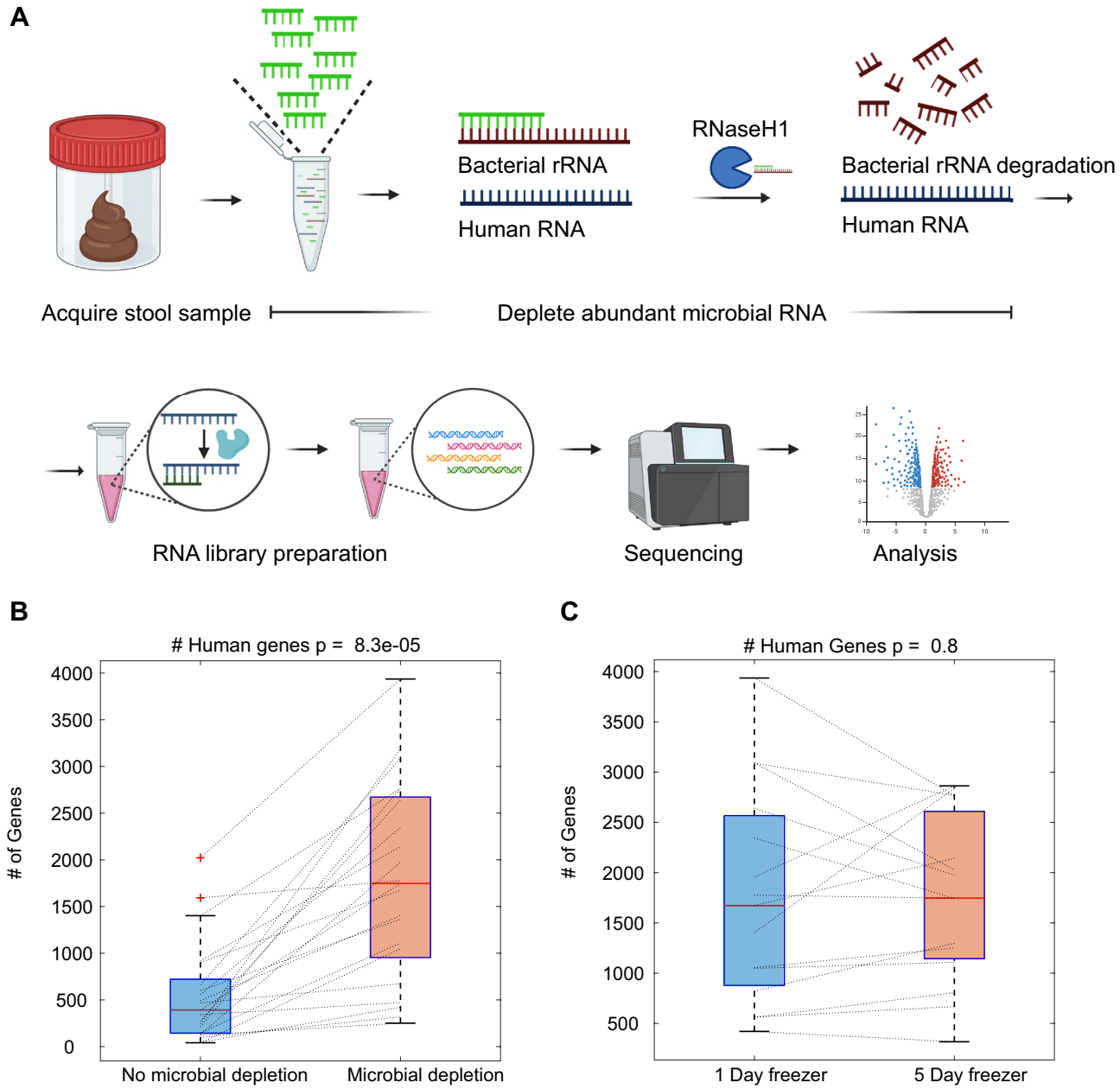
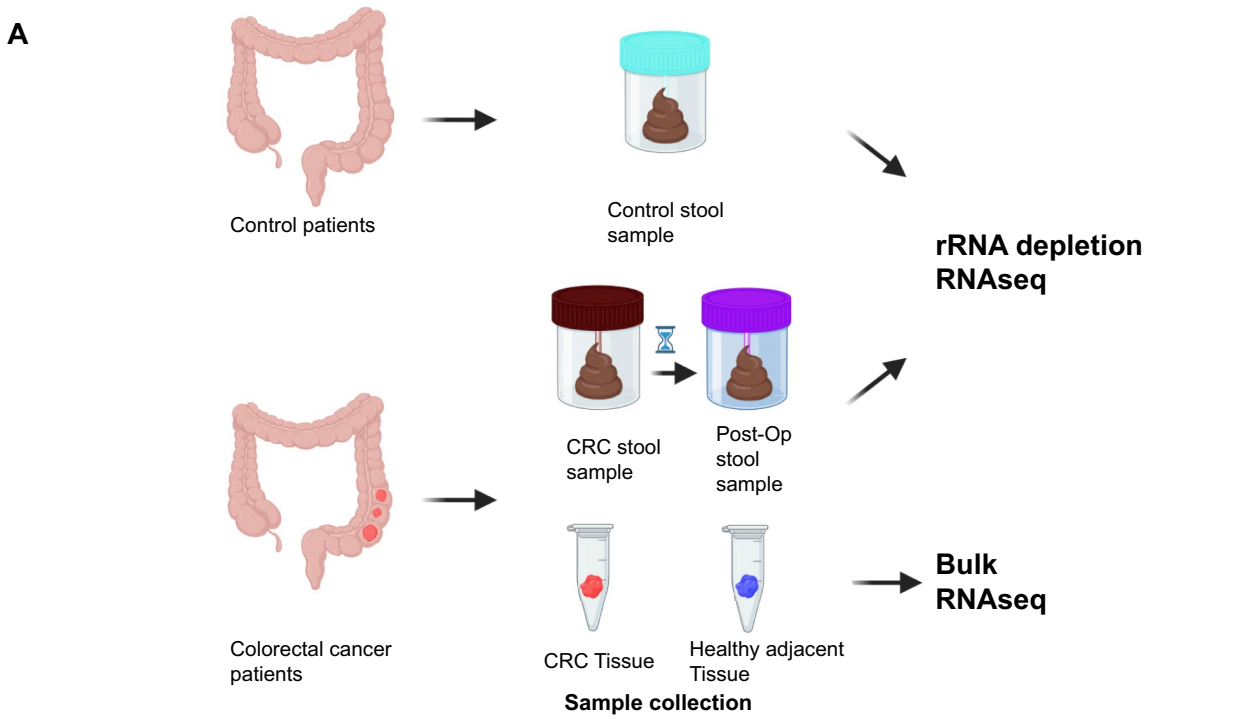
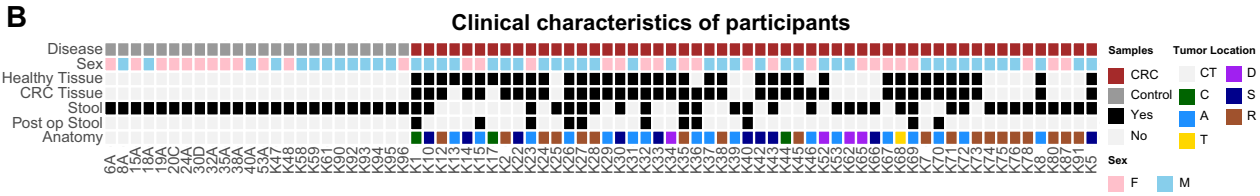


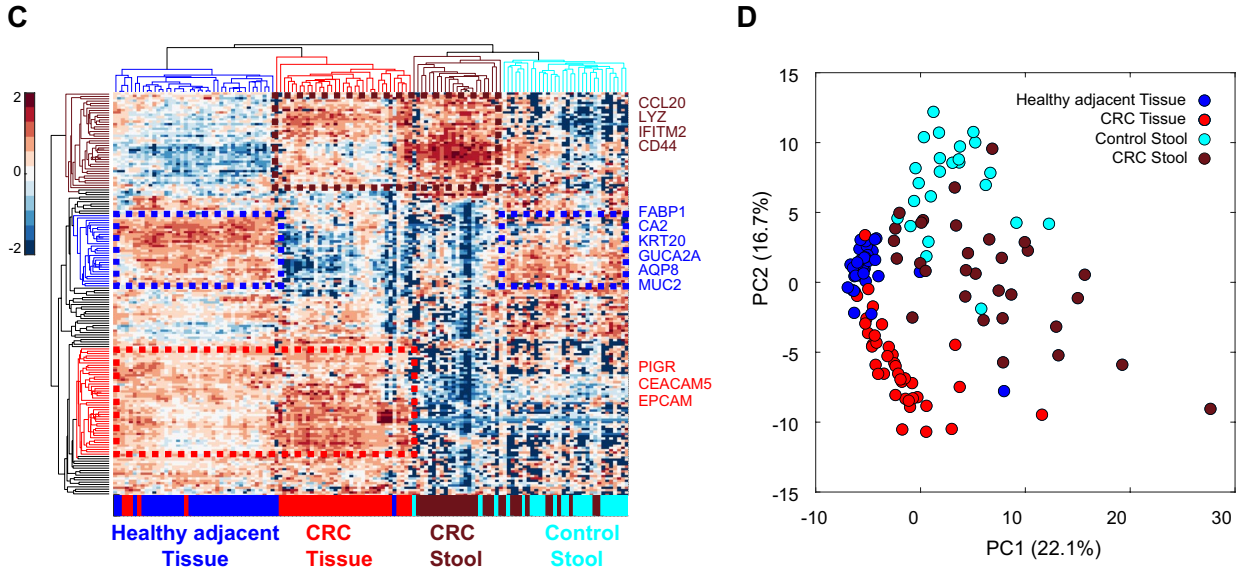
Fig. 1. Microbial ribosomal depletion increases human genes representation in stool samples. (**A**), Illustration of RNA library preparation from stool samples demonstrating the use of ribosomal depletion to eliminate abundant microbial RNAs. Illustrations in the figure were created using BioRender. (**B**) Ribosomal depletion increases the number of human genes by 6.1 fold on average (paired comparison). The median number of detected human genes increases from 392 ± 541 with no ribosomal depletion to 1745 ± 1050 with ribosomal depletion ($p = 8.3e^{-5}$). (**C**) Storage of stool samples at − 20 °C over 5 days does not change the number of detected human genes (median of 1674 ± 1073 genes after 1 day of storage, 1745 + 834 for 5 day storage, p = 0.8). In (**B**, **C**) p-values were computed using Kruskal–Wallis tests. Dotted lines connect samples from same patients, horizontal lines are medians, boxes delineate 25-75 percentiles.

Untargeted stool transcriptomics accurately detects CRC

To evaluate whether stool transcriptomes can accurately diagnose colorectal cancer (CRC), we trained a linear classifier on human gene expression from stool-derived shed cells, using a training and test set approach (Fig. 5A; Methods). The classifier achieved an area under the receiver operating characteristic curve (AUC) of 0.86—approaching the performance of tissue-based classification (AUC = 0.96, Fig. 5B). This untargeted cancer score matches the sensitivity reported for the multi-factor mt-sRNA test that combines 8 stool-derived eukaryotic RNA (seRNA) biomarkers, patient demographic information (smoking status), and a fecal immunochemical







test (FIT/iFOBT)[30] (mt-sRNA; Sensitivity 94%) and the three-marker multitarget FIT assay[31] (mtFIT; Sensitivity 78.7%) and exceeds that of the single-analyte FIT test (≈ 75%)[32].We next identified the most informative stool biomarkers contributing to CRC classification (Fig. 5C; Supplementary Table 6). Using these biomarkers, we calculated a cancer score for each of the stool samples (methods). Patient specific comparisons of cancer score pre and post-surgery showed a constant decrease (Supplementary Fig. 2C, Paired Wilcoxon p = 0.0039). Stool-based cancer score was independent of disease stage, tumor invasion score (T score), tumor histological grade, tumor size and tumor location (Supplementary Fig. 5, Supplementary Table 2). Stool biomarkers included several myeloid-associated genes such as *IL1B, CXCL8*, and *S100A8*, the latter encoding calprotectin—a well-established marker of gut inflammation[26]. Consistent with this, computational deconvolution[33] revealed a significant increase in shed myeloid cells in CRC stool compared to controls (Fig. 5D; Supplementary Table 7).

◀**Fig. 2**. Stool transcriptomics mirror tissue states in CRC. (**A**) Illustration of sample collection for the colorectal cancer and control cohorts. Illustration in the figure were created using BioRender. (**B**) Clinical characteristics of participants. Sample type nomenclature: F—female, M—male. Anatomy—location of the tumor: CT—control stool (no anatomy location), C—Cecum, A—ascending colon, T—transverse colon, D—descending colon, S—sigmoid colon, R—rectum. (**C**) Hierarchical clustering of stool samples and tissues from colorectal cancer and control patients. Representative genes are shown on the right colored by the group color with highest expression. Red outline dashed box highlights genes expressed more highly in tissues compared to stool, brown/blue outline dashed boxes highlight genes expressed more highly/lowly in CRC regardless of the sample type (tissue or stool). (**D**) Principal component analysis (PCA) of transcriptomic profiles. Numbers in parentheses show percent of explained variance by each PC. Analysis in (**C**, **D**) based on highly expressed (normalized expression > $5e^{-5}$), highly variable genes previously shown to be differentially expressed between colorectal cancer and control based on single cell data[28] (Methods).

To assess whether stool classification relies solely on inflammation, we repeated the analysis after excluding myeloid cell-associated genes or IBD related genes (Methods). Remarkably, classification performance remained robust for both myeloid exclusion (stool AUC = 0.84; tissue AUC = 0.96; Supplementary Fig. 6A,B), or IBD exclusion (stool AUC = 0.82; tissue AUC = 0.96; Supplementary Fig. 6C,D) indicating that stool transcriptomes encode CRC-specific signatures that extend beyond inflammatory cell signals.

Discussion

Stool samples contain a mixture of cells shed from both healthy colonic epithelium and tumor tissue. Our findings demonstrate that localized colorectal tumors shed sufficient cells into the lumen to generate a distinct and detectable transcriptomic signal, despite the background of normal colonocyte-derived RNA. This may reflect the heightened proliferative activity of tumor cells, resulting in increased turnover and shedding, or a greater post-shedding RNA stability of malignant cells within the gut environment.

By combining microbial ribosomal RNA depletion with sensitive RNA sequencing, our method enables comprehensive profiling of human gene expression from stool. Stage-stratified performance shows that the cancer score remains discriminative even in early disease (Stage 1; Supplementary Fig. 5C), indicating that detection is not confined to advanced tumor burden. Empirically, a stable classification and biological interpretability were retained with as few as ~ 1500 detected human genes (≈ 6000 UMIs), a read depth readily achievable from a single routine stool aliquot. The high AUC (Fig. 5B) indicates strong diagnostic potential, suggesting that transcriptome profiling offers a competitive alternative to existing modalities and may facilitate more comprehensive, molecularly informed CRC screening in the future. This not only supports accurate CRC detection, but also allows inference of biological pathways, tumor subtypes, and potentially, prediction of individualized treatment responses. Such molecular resolution has relevance for non-invasive monitoring of rectal cancer—especially in patients undergoing non-operative management or surveillance following curative therapy[34].

Looking forward, Stool shed cell transcriptomics offers a scalable and accessible tool for studying human gut physiology and pathophysiology across contexts including development, aging, and other gastrointestinal diseases.

Methods

Patient population

Human samples and clinical data were obtained from patients undergoing colectomy due to the presence of malignancy (average age 64.2 ± 13.9). Control stool samples were obtained from healthy individuals (average age 39.8 ± 13.6). A summary of clinical characteristics of all participants can be found in Fig. 2B and Supplementary Table 2. All surgeries were performed at Sheba Medical Center, Israel.

Ethics statement for human participant

This study was approved by the Sheba Medical Center Institutional Review Board (approval number SMC-8665-21). All participants provided written informed consent prior to inclusion. The study was conducted in accordance with the Declaration of Helsinki and relevant local regulations.

Sample collection

Pre-op stool samples were collected either at home or at the hospital 1 to 7 days prior to tumor resection, post-op samples were collected at home at least 1 month after tumor resection. Stools were collected directly to RNA Later (AM7021, Invitrogen), placed at − 20 °C for up to 5 days and transferred to the lab for long storage at − 80 °C. Tissues were collected at the operation room as soon as the specimen was excised. Tissues were snap frozen directly on dry ice.

RNA extraction

For tissue samples—snap frozen tissues were thawed in 300 µl Tri-reagent and mechanically homogenized with bead beating, followed by a short centrifugation step to pull down beads and any tissue left-overs. Following this, ethanol was added in a ratio of 1:1 to the supernatant from the previous step and continued according to the manufacturer instructions of Direct-zol mini prep kit (ZYMO research, R2052). For stool samples—stools in RNA Later were thawed on ice for 30 min until buffer was completely thawed. Stool sample was extracted from the buffer and wiped on kim-wipe paper. Pea size stool sample was placed in lysis buffer containing 1 ml

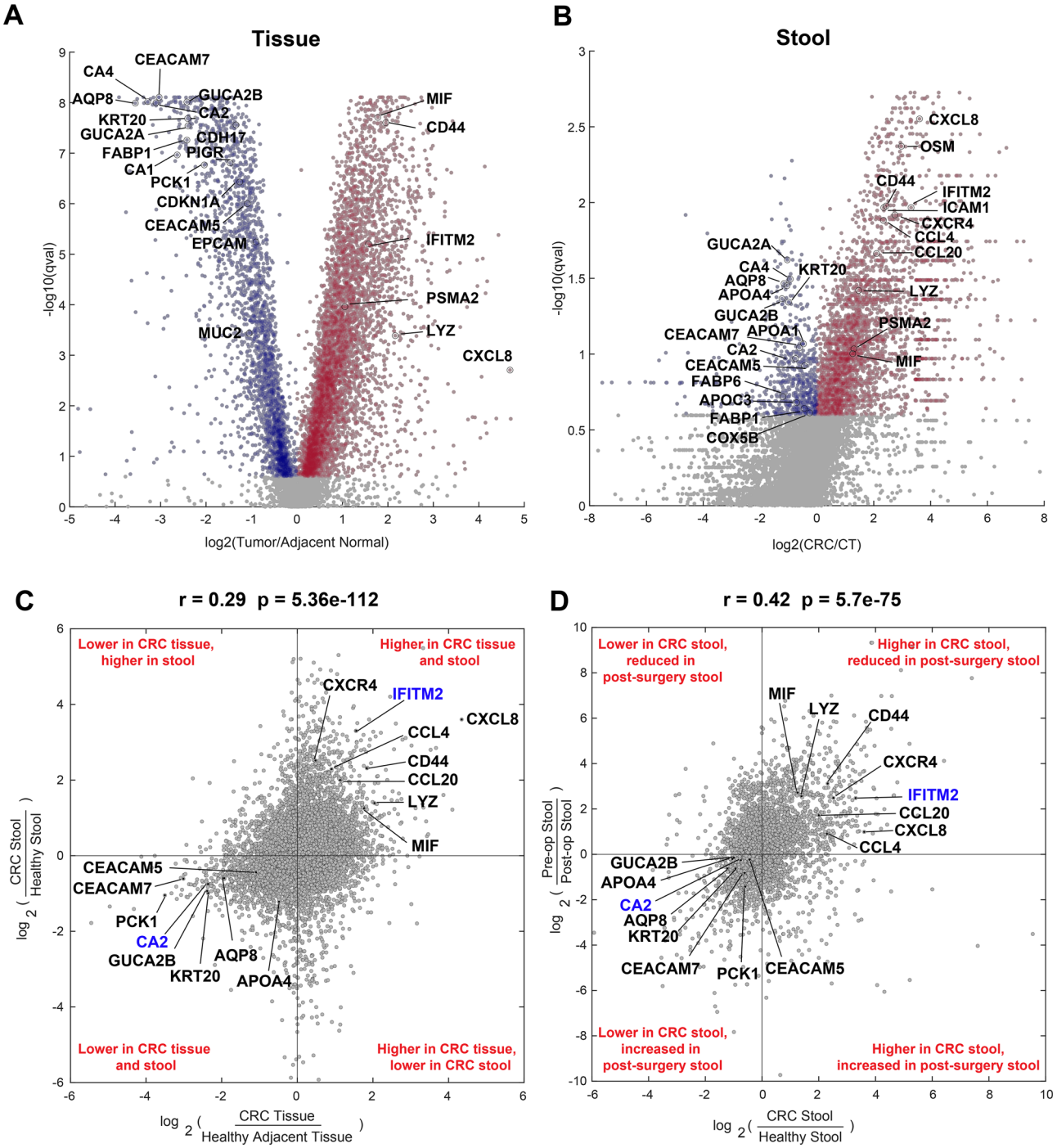


Fig. 3. Colorectal cancer patients stool samples capture tissue molecular states. (**A**) Differentially expressed genes between the CRC tissue and the adjacent healthy tissue. (**B**) Differentially expressed genes between the CRC stool and healthy stool. In (**A**, **B**), Genes with normalized expression above 1e − 5 are shown. Genes with corresponding q-values below 0.25 are marked in red or blue for CRC or control elevated genes respectively. p-values were computed using Kruskal–Wallis tests and Benjamini-Hochberg method was applied to correct for multiple hypotheses. Names of representative genes are shown. (**C**) Scatter plot of $\log_2$ gene expression ratio between CRC stool and control stool (y axis) and $\log_2$ gene expression ratio between CRC and healthy tissues (x axis). Spearman correlation 0.29, $p = 5.36e^{-112}$ (**D**) Scatter plot of $\log_2$ expression ratio between patient matched pre-op and post-op stool (y axis) and gene expression ratio between CRC and healthy stool (x axis). Spearman correlation 0.42 $p = 5.7e^{-75}$. Black font genes in (C,D) are representative genes, blue font genes shown in spatial transcriptomics in Fig. 4.

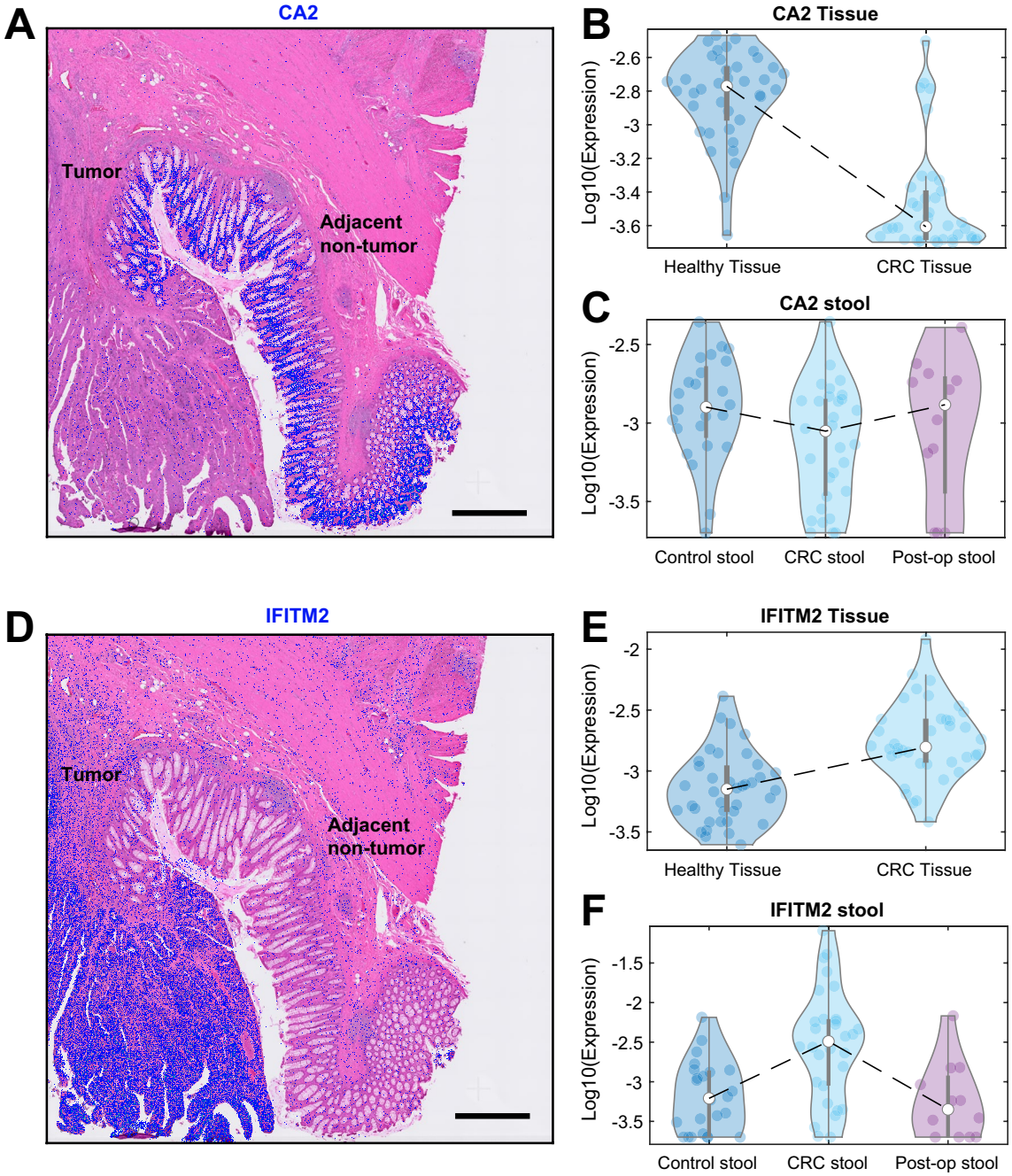
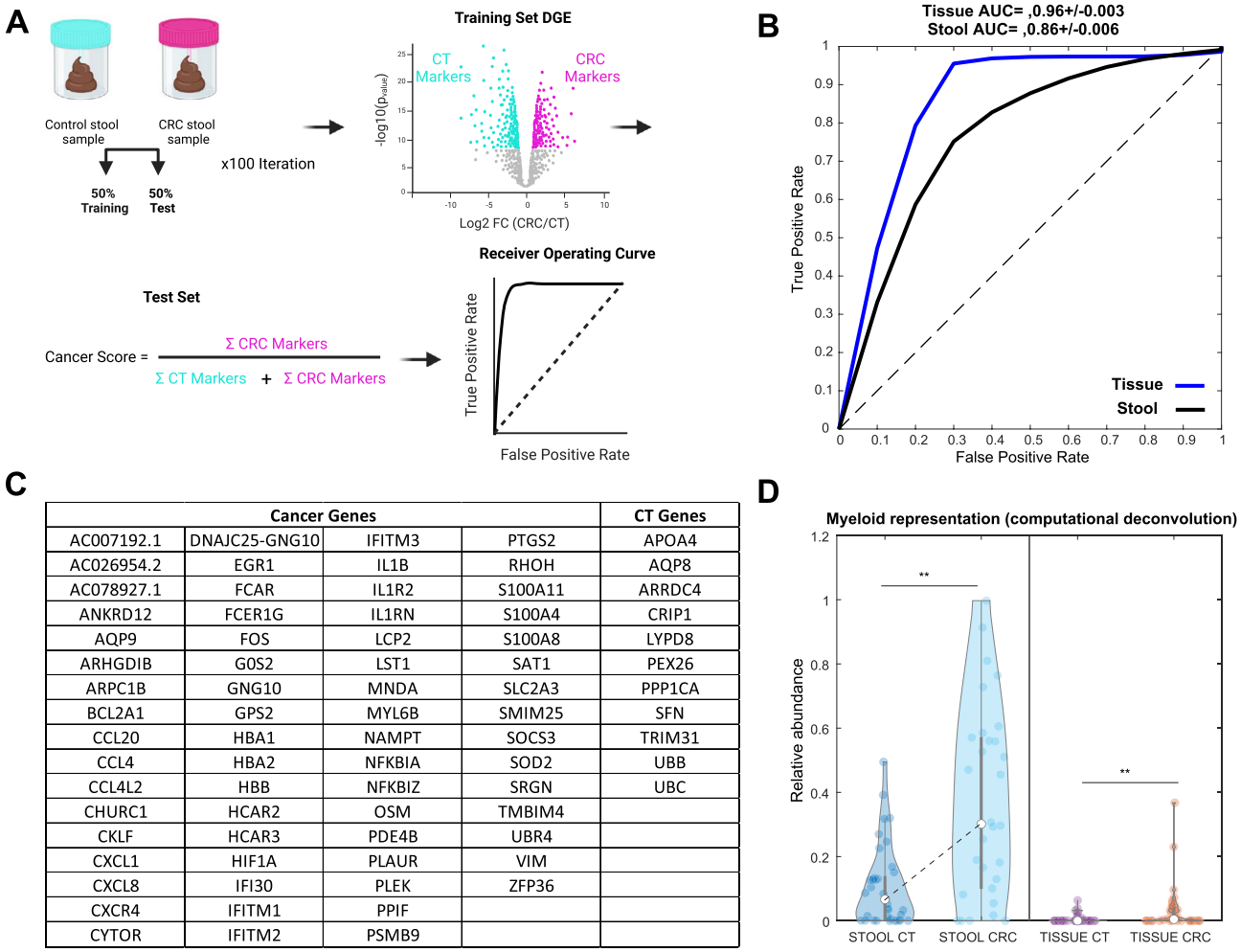


Fig. 4. Colorectal cancer patients stool samples capture spatial transcriptomics patterns. (**A**) Visium HD spatial transcriptomics slide[29] demonstrating the expression of *CA2* in the non-tumor domain of the tissue section. Scale bar 1 mm. (**B**) Expression of the *CA2* gene in the tissues analyzed in our study. (**C**) Expression of the *CA2* gene in stool samples. (**D**) Visium HD spatial transcriptomics slide[29] demonstrating the expression of *IFITM2* in the tumor domain of the tissue section. Scale bar 1 mm. (**E**) Expression of the *IFITM2* gene in the tissues analyzed in our study. (**F**) Expression of the *IFITM2* gene in stool samples.

RLT buffer with 40mM DTT and about 50 µl of beads. Stool was mechanically homogenized with bead beating for 3 min followed by 3 min centrifugation at 5000 rpm step to pull down beads and any stool solids left-overs. Following this, ethanol was added in a ratio of 1:1 to the supernatant from the previous step and continued according to the manufacturer instructions of Qiagen RNeasy Micro Kit (74004, Qiagen). Stool was eluted in 30 µl RNAse free water.

Bacterial genomic depletion and RNA sequencing

Stool samples RNA was depleted from bacterial ribosomal RNA using the NEBNext rRNA Depletion Kit (Bacteria, New England Biolabs, E7850X) according to the manufacturer instruction. Briefly: one third of total



Cancer Genes				CT Genes
AC007192.1	DNAJC25-GNG10	IFITM3	PTGS2	APOA4
AC026954.2	EGR1	IL1B	RHOH	AQP8
AC078927.1	FCAR	IL1R2	S100A11	ARRDC4
ANKRD12	FCER1G	IL1RN	S100A4	CRIP1
AQP9	FOS	LCP2	S100A8	LYPD8
ARHGDIB	G0S2	LST1	SAT1	PEX26
ARPC1B	GNG10	MNDA	SLC2A3	PPP1CA
BCL2A1	GPS2	MYL6B	SMIM25	SFN
CCL20	HBA1	NAMPT	SOCS3	TRIM31
CCL4	HBA2	NFKBIA	SOD2	UBB
CCL4L2	HBB	NFKBIZ	SRGN	UBC
CHURC1	HCAR2	OSM	TMBIM4	
CKLF	HCAR3	PDE4B	UBR4	
CXCL1	HIF1A	PLAUR	VIM	
CXCL8	IFI30	PLEK	ZFP36	
CXCR4	IFITM1	PPIF		
CYTOR	IFITM2	PSMB9		

Fig. 5. Linear classifier for CRC diagnosis using stool biomarkers. (**A**) Diagram showing the linear classification process. Illustration in the figure was created using BioRender. (**B**) Classifier trained to differentiate between colorectal cancer patients and healthy patients based on the human shed cell transcriptomics of tissues (blue) and of stool samples (black) including myeloid genes. The classifier was evaluated using a 50% test set in each iteration. Receiver operating characteristic (ROC) curve is an average of 100 iterations. Area under the curve values are the mean values and standard errors of the mean over the 100 iterations. (**C**) Stool biomarkers extracted using the linear classifier based on all genes (normalized expression > $5e^{-5}$). (**D**) Estimated fractions of myeloid cells in stool and tissue samples inferred by computational deconvolution[33] (**p < 1e − 3).

isolated stool RNA was used as input. Abundant bacteria ribosomal RNA-specific probes were hybridized to rRNA molecules in the total RNA stool sample by temperature ramp down from 95 to 22 °C over the course of 30 min. The rRNA-probe complexes were removed using RNase H1 enzyme. Next, DNase I was used to clean up the remaining sample by eliminating any residual DNA probes or genomic DNA. RNA was cleaned using magnetic beads and eluted in 9 µl nuclease free water. Eluted RNA was used directly in the RT reaction. RNA was processed by the mcSCRBseq protocol[27] with minor modifications. For tissue RNA, RT reaction was applied on 20 ng of total RNA (2 µl 10 ng/µl RNA, 6.4 nuclease free water). For stool RNA 8.4 µl from bacterial ribosomal RNA depleted RNA were used. Final RT reaction volume was 20 µl (1 × Maxima H Buffer, 1 mM dNTPs, 2 µM TSO* E5V6NEXT, 7.5% PEG8000, 20U Maxima H enzyme, 1 µM barcoded RT primer, 8.4 µl RNA sample). Subsequent steps were applied as mentioned in the protocol. Library preparation was done using Nextera XT kit (Illumina) on 0.6 ng amplified cDNA from tissue or 1 ng from stool. Library final concentration of 2nM was loaded on NovaX (Illumina) sequencing machine aiming at 20 M reads per sample with the following setting: Read1—26 bp, Index1—10 bp, Index2—10 bp, Read2—66 bp.

Bioinformatics and computational analysis

Illumina output files were demultiplexed and aligned to the human HG38 genome with UTAP[35], using CUTADAPT with the default parameters.

Statistical analyses were performed with MATLAB R2023a. Mitochondrial and ribosomal genes as well as non-protein coding genes were removed from the analysis. Protein coding genes were extracted using the

annotation in the Ensembl database (BioMart) for reference genome GRCm38 version 91. Gene expression for each sample was consequently normalized by the sum of the UMIs of the remaining genes that individually take up less than 5% of the sample sum. Samples with less than 1500 genes over the remaining genes were removed from the analyses.

Differential gene expression was performed using Wilcoxon ranksum tests and Benjamini–Hochberg FDR corrections.

Computational deconvolution was performed using cellanneal[33] using signature tables extracted from a single cell atlas[28] (Supplementary Table 7).

Analysis of single cell data

Clustering was performed with the MATLAB function clustergram over the Zscore-transformed expression matrix, using Spearman distances and including highly variable genes (log–log regression residuals > 0 of noise (std/mean) vs. mean) with maximal expression above $5e^{-5}$ of normalized summed UMIs. These genes were crossed with differentially expressed genes of cancer and control single cells from Pelka et al.[28] extracted as follows. From each of the clusters at the ClusterMidway level at the Pelka UMAP[28], 1000 cells were randomly selected and divided into 2 groups based on the specimen type (N for normal, T for Tumor). Differential gene expression was performed using Wilcoxon ranksum tests and Benjamini–Hochberg FDR corrections, and a list of 2185 genes (FDR q-value less than 0.05 and fold change over 1.5) was extracted.

For the comparison of gene expression ratio between CRC single cells and healthy single cells to gene expression ratio between CRC and healthy tissue (Supplementary Fig. 1) we used data from Pelka et al.[28]. From each of the Epi and EpiT clusters, 5000 cells were randomly selected for the analysis. The mean expression for each gene was calculated for each group and used for the ratio of CRC to control single cells. All p-values were computed using Kruskal–Wallis tests.

Analysis of spatial transcriptomics data

Data was taken from Oliveira et al.[29]. in this preprint, 3 samples of colorectal cancer tissues were analyzed on visium HD slide. For each slide, K-means clustering was used with K = 5. Next, the clusters that represent the tumor region or the adjacent non-tumor region were selected (as illustrated in Supplementary Fig. 3C). Differential gene expression between the clusters for each slide was performed using the loupe browser (Supplementary Table 5). Genes were divided into 2 groups for tumor or adjacent non-tumor related genes and plotted against the tissue or stool fold change between the CRC and control samples.

Linear classifier for cancer score prediction

Samples used to build the classifier were stool or tissue samples from the CRC or control cohort (Fig. 2). Classification was based on a parameter that represented the fractional summed expression of CRC-related genes, extracted as follows. For 100 iterations, the data set was split into a training set (50%) and a test set (50%). Differential gene expression between the CRC and control training set samples was performed. Genes with a maximal normalized expression higher than $5^{*}e^{-4}$ over all the included samples that were expressed in more than 5% of the samples were further considered. The Kruskal–Wallis test followed by the Benjamini–Hochberg FDR correction was performed. Genes, the mean expression of which in CRC samples were more than twofold higher than the mean in control samples, with FDR q-value less than 0.25 were selected as CRC markers. Inversely, genes, the mean expression of which in CRC samples was less than two fold lower than the mean in control samples, with FDR q-value less than 0.25 were chosen as control markers. Those genes were identified based on the training set only (Supplementary Table 6).

Next, for the test set samples, the sums of CRC markers and control markers were calculated, after the gene expression levels for each gene were normalized internally by their maximal values across the test set. Cancer score was calculated for each test set sample, as follows: (sum of normalized CRC markers)/(sum of normalized CRC markers + sum of normalized control markers). The receiver operating characteristic curve[36] and subsequent false-positive rate, true-positive rate, and AUC were recorded for each test set to examine the classification of samples of CRC patients and control patients based on the cancer score.

In a classifier excluding myeloid cells genes, we omitted genes that showed expression levels in the myeloid cell type that were at least 1.5 times greater than their maximum expression in any other cell type (Supplementary Table 6). In a classifier excluding IBD related genes, we omitted genes that were differentially expressed in fecal washes between IBD and control samples (Ungar et al.[14], Online Supplemental Table 6 AUC wash > 0.85).

Data availability

The datasets generated and/or analysed during the current study are available in the Zenodo public depository under https://doi.org/10.5281/zenodo.15621495. And in the ArrayExpress, BioStudies repository under https://www.ebi.ac.uk/biostudies/studies/S-BSST2083. Data taken for single cells analysis was downloaded from the Broad Institute single cell portal. Pelka et al. https://singlecell.broadinstitute.org/single_cell/study/SCP1162/human-colon-cancer-atlas-c295#study-download. Data taken for spatial transcriptomics analysis was downloaded from Oliveira et al. https://www.10xgenomics.com/products/visium-hd-spatial-gene-expression/dataset-human-crc.

Code availability

All code used in the paper is available upon request.

Received: 3 June 2025; Accepted: 25 August 2025

Published online: 02 October 2025

References

1. Li, H. et al. Advanced endoscopic methods in gastrointestinal diseases: A systematic review. *Quant. Imaging Med. Surg.* **9**, 905–920 (2019).
2. Carethers, J. M. Commencing colorectal cancer screening at age 45 years in US racial groups. *Front. Oncol.* **12**, 966998 (2022).
3. Shaukat, A. & Levin, T. R. Current and future colorectal cancer screening strategies. *Nat. Rev. Gastroenterol. Hepatol.* **19**, 521–531 (2022).
4. Kowalkowski, H., Austin, G., Guo, Y., Miller-Wilson, L.-A. & DaCosta Byfield, S. Patterns of colorectal cancer screening and adherence rates among an average-risk population enrolled in a national health insurance provider during 2009–2018 in the United States. *Prev. Med. Rep.* **36**, 102497 (2023).
5. Cao, L.-J. et al. A fecal-based test for the detection of advanced adenoma and colorectal cancer: A case-control and screening cohort study. *BMC Med.* **19**, 250 (2021).
6. Wassie, M. M. et al. Accuracy, acceptability, and application: Fecal immunochemical tests for early detection of advanced neoplasia in colonoscopy-based surveillance. *Dig. Dis. Sci.* **69**, 2621–2630 (2024).
7. Coleman, D. & Kuwada, S. miRNA as a biomarker for the early detection of colorectal cancer. *Genes* **15**, 338 (2024).
8. Imperiale, T. F. et al. Next-generation multitarget stool DNA test for colorectal cancer screening. *N. Engl. J. Med.* **390**, 984–993 (2024).
9. Porcaro, F., Voccola, S., Cardinale, G., Porcaro, P. & Vito, P. DNA methylation biomarkers in stool samples: Enhancing colorectal cancer screening strategies. *Oncol. Rev.* **18**, (2024).
10. Chen, G. *et al.* Exploring the gut microbiome's role in colorectal cancer: Diagnostic and prognostic implications. *Front. Immunol.* **15**, (2024).
11. Bahar Halpern, K. et al. The cellular states and fates of shed intestinal cells. *Nat. Metab.* **5**, 1858–1869 (2023).
12. Sender, R. & Milo, R. The distribution of cellular turnover in the human body. *Nat. Med.* **27**, 45–48 (2021).
13. Dan, S. *et al.* Distal fecal wash host transcriptomics identifies inflammation throughout the colon and terminal ileum. *Cell. Mol. Gastroenterol. Hepatol.* (2023).
14. Ungar, B. et al. Host transcriptome signatures in human faecal-washes predict histological remission in patients with IBD. *Gut* https://doi.org/10.1136/gutjnl-2021-325516 (2022).
15. Senn, V. *et al.* Microbial colonization from the fetus to early childhood—A comprehensive review. *Front. Cell. Infect. Microbiol.* **10**, (2020).
16. Kundu, P., Blacher, E., Elinav, E. & Pettersson, S. Our gut microbiome: The evolving inner self. *Cell* **171**, 1481–1493 (2017).
17. Hul, M. V. et al. What defines a healthy gut microbiome?. *Gut* **73**, 1893–1908 (2024).
18. Schlaberg, R. et al. Fecal host transcriptomics for non-invasive human mucosal immune profiling: Proof of concept in clostridium difficile infection. *Pathog. Immun.* **3**, 164 (2018).
19. Barnell, E. K. et al. Noninvasive detection of high-risk adenomas using stool-derived eukaryotic RNA sequences as biomarkers. *Gastroenterology* **157**, 884-887.e3 (2019).
20. Herring, E., Tremblay, É., McFadden, N., Kanaoka, S. & Beaulieu, J.-F. Multitarget stool mRNA test for detecting colorectal cancer lesions including advanced adenomas. *Cancers* **13**, 1228 (2021).
21. Knight, J. et al. Non-invasive analysis of intestinal development in preterm and term infants using RNA-Sequencing. *Sci. Rep.* **4**, 5453 (2014).https://doi.org/10.1038/srep05453
22. Zhang, Y. et al. Metatranscriptomics for the human microbiome and microbial community functional profiling. *Annu. Rev. Biomed. Data Sci.* **4**, 279–311 (2021).
23. Mouradov, D. et al. Oncomicrobial community profiling identifies clinicomolecular and prognostic subtypes of colorectal cancer. *Gastroenterology* **165**, 104–120 (2023).
24. Gosalbes, M. J. et al. Metatranscriptomic approach to analyze the functional human gut microbiota. *PLoS ONE* **6**, e17447 (2011).
25. Wahl, A., Huptas, C. & Neuhaus, K. Comparison of rRNA depletion methods for efficient bacterial mRNA sequencing. *Sci. Rep.* **12**, 5765 (2022).
26. Khaki-Khatibi, F. et al. Calprotectin in inflammatory bowel disease. *Clin. Chim. Acta.* **510**, 556–565 (2020).
27. Bagnoli, J. W. et al. Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.* **9**, 2937 (2018).
28. Pelka, K. et al. Spatially organized multicellular immune hubs in human colorectal cancer. *Cell* **184**, 4734-4752.e20 (2021).
29. Oliveira, M. F. *et al.* Characterization of immune cell populations in the tumor microenvironment of colorectal cancer using high definition spatial profiling. https://doi.org/10.1101/2024.06.04.597233 (2024).
30. Barnell, E. K. et al. Multitarget stool RNA test for colorectal cancer screening. *JAMA* **330**, 1760–1768 (2023).
31. de Klaver, W. et al. Clinical Validation of a Multitarget Fecal Immunochemical Test for Colorectal Cancer Screening : A Diagnostic Test Accuracy Study. *Ann. Intern. Med.* **174**(9), 1224–1231. https://doi.org/10.7326/M20-8270 PMID: 34280333 (2021).
32. Katsoula, A., Paschos, P., Haidich, A.-B., Tsapas, A. & Giouleme, O. Diagnostic accuracy of fecal immunochemical test in patients at increased risk for colorectal cancer: A meta-analysis. *JAMA Intern. Med.* **177**, 1110–1118 (2017).
33. Buchauer, L. & Itzkovitz, S. Cellanneal: A user-friendly deconvolution software for transcriptomics data. *J. Open Source Softw.* **9**, 5610 (2024).
34. Goffredo, P., Quezada-Diaz, F. F., Garcia-Aguilar, J. & Smith, J. J. Non-operative management of patients with rectal cancer: Lessons learnt from the OPRA trial. *Cancers* **14**, 3204 (2022).
35. Kohen, R. et al. UTAP: User-friendly transcriptome analysis pipeline. *BMC Bioinform.* **20**, 154 (2019).
36. Fan, J., Upadhye, S. & Worster, A. Understanding receiver operating characteristic (ROC) curves. *Can. J. Emerg. Med.* **8**, 19–20 (2006).

Acknowledgements

S.I. is supported by the Helen and Martin Kimmel Award for Innovative Investigation, the Yad Abraham Research Center for Cancer Diagnostics and Therapy, the Moross Integrated Cancer Center, the Minerva Stiftung grant, a Weizmann-Sheba grant, the Israel Science Foundation Grants no. 908/21 and 3663/21, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme Grant no. 768956 and a Grant from the Ministry of Innovation, Science & Technology, Israel.

Author contributions

K.B.H., I.K., O.Y., S.B.M., I.N. and S.I. conceived the study. I.K. preformed all surgeries and provided tumor samples. K.B.H., O.Y., T.B., M.F. and A.I. collected and processed samples. K.B.H. and S.I performed computational analysis and wrote the manuscript. R.N. contributed to computational analysis. All the authors discussed the results and commented on the manuscript.

Declarations

Competing interests

The authors declare the following competing interests: SI is a co-founder of Tracells. KBH, SBM, IK, OY, TB, IN, SI are inventors on provisional patents which cover aspects of the findings reported in this manuscript. MF, RN, AI have no conflicts of interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-17411-5.

Correspondence and requests for materials should be addressed to K.B.H. or S.I.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.