Published OnlineFirst August 31, 2021; DOI: 10.1158/1078-0432.CCR-21-1982

EpiPanGI Dx: A Cell-free DNA Methylation Fingerprint for the Early Detection of Gastrointestinal Cancers

Raju Kandimalla[1], Jianfeng Xu[2,3], Alexander Link[4], Takatoshi Matsuyama[5], Kensuke Yamamura[6], M. Iqbal Parker[7], Hiroyuki Uetake[8], Francesc Balaguer[9], Erkut Borazanci[10], Susan Tsai[11], Douglas Evans[11], Stephen J. Meltzer[12], Hideo Baba[6], Randall Brand[13], Daniel Von Hoff[10,14], Wei Li[2,3], and Ajay Goel[1,15,16]

ABSTRACT

Purpose: DNA methylation alterations have emerged as front-runners in cell-free DNA (cfDNA) biomarker development. However, much effort to date has focused on single cancers. In this context, gastrointestinal (GI) cancers constitute the second leading cause of cancer-related deaths worldwide; yet there is no blood-based assay for the early detection and population screening of GI cancers.

Experimental Design: Herein, we performed a genome-wide DNA methylation analysis of multiple GI cancers to develop a pan-GI diagnostic assay. By analyzing DNA methylation data from 1,781 tumor and adjacent normal tissues, we first identified differentially methylated regions (DMR) between individual GI cancers and adjacent normal, as well as across GI cancers. We next prioritized a list of 67,832 tissue DMRs by incorporating all significant DMRs across various GI cancers to design a custom, targeted bisulfite sequencing platform. We subsequently validated these tissue-specific DMRs in 300 cfDNA specimens and applied machine learning algorithms to develop three distinct categories of DMR panels

Results: We identified three distinct DMR panels: (i) cancer-specific biomarker panels with AUC values of 0.98 (colorectal cancer), 0.98 (hepatocellular carcinoma), 0.94 (esophageal squamous cell carcinoma), 0.90 (gastric cancer), 0.90 (esophageal adenocarcinoma), and 0.85 (pancreatic ductal adenocarcinoma); (ii) a pan-GI panel that detected all GI cancers with an AUC of 0.88; and (iii) a multi-cancer (tissue of origin) prediction panel, EpiPanGI Dx, with a prediction accuracy of 0.85–0.95 for most GI cancers.

Conclusions: Using a novel biomarker discovery approach, we provide the first evidence for a cfDNA methylation assay that offers robust diagnostic accuracy for GI cancers.

Introduction

Despite improved overall survival rates due to recent advancements in cancer therapies, cancer remains the second leading cause of mortality worldwide (1). At present in the United States, average-risk or asymptomatic population screening is recommended for only colorectal, breast, cervical, lung, and prostate cancers (2). Population screening for low prevalence cancers is challenging due to a lack of cost-effective diagnostic tools (3). Thus, to facilitate population screening and thereby eradicate the mortality associated with cancer, a universal cancer screening test that is noninvasive, simple, and robust is urgently needed.

Circulating tumor DNA released into the bloodstream by a tumor cell carries both a genetic and an epigenetic signature of the cell of origin, and is therefore becoming a key tool in developing liquid biopsy–based biomarkers for early detection and treatment monitoring (4). Unfortunately, the diversity of genetic mutations across cancers and the prevalence of these mutations across large genomic regions makes it challenging to develop mutation-based, pan-cancer diagnostic tests (5). In contrast, epigenetic DNA methylation changes occur in specific genomic regions called CpG islands and can be consistently measured using bisulfite sequencing in various biological fluids, including plasma, serum, urine, and saliva. Because of their high cancer specificity, and their appearance during the earliest phases of cancer development, aberrant DNA methylation alterations provide an excellent avenue by which to develop pan-cancer liquid biopsy–based diagnostic markers (6, 7). However, most recent studies investigating plasma cell-free DNA

[1]Center for Gastrointestinal Research, Center for Translational Genomics and Oncology, Baylor Scott & White Research Institute, Charles A Sammons Cancer Center, Baylor University Medical Center, Dallas, Texas. [2]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, Texas. [3]Department of Biological Chemistry, School of Medicine, University of California, Irvine, Irvine, California. [4]Department of Gastroenterology, Hepatology and Infectious Diseases, Otto-von-Guericke University Hospital, Magdeburg, Germany. [5]Department of Gastrointestinal Surgery, Tokyo Medical and Dental University Graduate School of Medicine, Tokyo, Japan. [6]Department of Gastroenterological Surgery, Graduate School of Medical Sciences, Kumamoto University, Kumamoto, Japan. [7]Division of Medical Biochemistry and Structural Biology, Institute for Infectious Disease and Molecular Medicine, University of Cape Town, Cape Town, South Africa. [8]Department of Specialized Surgery, Tokyo Medical and Dental University Graduate School of Medicine, Tokyo, Japan. [9]Gastroenterology Department, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Hospital Clínic, University of Barcelona, Barcelona, Spain. [10]HonorHealth Research Institute, Scottsdale, Arizona. [11]Department of Surgery, Medical College of Wisconsin, Milwaukee, Wisconsin. [12]Department of Medicine, Division of Gastroenterology, The Johns Hopkins University School of Medicine, Baltimore, Maryland. [13]Department of Medicine, Division of Gastroenterology, Hepatology, and Nutrition, University of Pittsburgh, Pittsburgh, Pennsylvania. [14]Translational Genomics Research Institute, an Affiliate of City of Hope, Phoenix, Arizona. [15]Department of Molecular Diagnostics and Experimental Therapeutics, Beckman Research Institute of City of Hope, Monrovia, California. [16]City of Hope Comprehensive Cancer Center, Duarte, California.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

R. Kandimalla and J. Xu contributed equally to this article, and W. Li and A. Goel jointly supervised this article.

Corresponding Authors: Ajay Goel, Translational Genomics, Beckman Research Institute, City of Hope, Monrovia, CA 91016. Phone/Fax: 626-218-4783; E-mail: ajgoel@coh.org; and Wei Li, wei.li@uci.edu

Clin Cancer Res 2021;27:6135–44

doi: 10.1158/1078-0432.CCR-21-1982

Published OnlineFirst August 31, 2021; DOI: 10.1158/1078-0432.CCR-21-1982

Translational Relevance

Gastrointestinal (GI) cancers constitute the second leading cause of cancer-related deaths worldwide, yet there is no blood-based assay for early detection and/or population screening of all GI cancers. Because of their generally low prevalence and lack of cost-effective screening tools, except for colorectal cancer, most GI cancers present at a late stage, leading to a high mortality rate, and underscoring the need for improved screening tools. Owing to their high cancer specificity, DNA methylation alterations have emerged as front-runners in cell-free DNA biomarker development. Herein, we performed a genome-wide DNA methylation analysis of multiple GI cancer tissues, and subsequently validated the tissue-specific DMRs in 300 cell-free DNA specimens by designing a custom, targeted bisulfite sequencing platform. In summary, we developed a robust, sensitive, targeted methylation-based assay for multi-GI cancer detection. Our EpiPanGI Dx assay needs further validation in completely independent retrospective and prospective datasets for clinical translation as early diagnostic markers.

(cfDNA) methylation patterns for biomarker development have focused on only individual cancers (8–10), whereas few investigated multiple cancers (11, 12).

Gastrointestinal (GI) cancers, including colorectal cancer, hepatocellular carcinoma (HCC), esophageal squamous cell carcinoma (ESCC), gastric cancer, esophageal adenocarcinoma (EAC), and pancreatic ductal adenocarcinoma (PDAC) constitute the second leading cause of cancer-related deaths worldwide, yet there is no blood-based assay for early detection and/or population screening of GI cancers. Because of their generally low prevalence and lack of cost-effective screening tools, except for colorectal cancer (13), most GI cancers present at a late stage, leading to a high mortality rate, and underscoring the need for improved screening tools. Most studies to date investigated genome-wide methylation patterns at the tissue level in individual cancers, subsequently selecting the most significant tissue markers for testing in the cfDNA of the corresponding cancer type. In this way, these single-cancer studies failed to analyze DNA methylation patterns in an unbiased and comprehensive manner, and thereby lack the ability to discover pan-cancer-specific markers. To address this challenge and to identify methylation markers across GI cancers, we performed a genome-wide DNA methylation analysis of multiple GI cancers, which we used to develop a novel cfDNA methylation biomarker panel for the early detection of individual GI cancers, a pan-GI diagnostic panel, and a multi-GI cancer prediction panel (EpiPanGI Dx).

Materials and Methods

Patients and clinicopathologic data

Whole-genome 450K tissue DNA methylation data across six GI cancers (colorectal cancer, HCC, ESCC, gastric cancer, EAC, PDAC) and adjacent normal tissues were obtained from The Cancer Genome Atlas (TCGA) and GSE72872 dataset (14). Complete clinical, epidemiologic, molecular, and histopathologic data are available at TCGA website: https://tcga-data.nci.nih.gov/tcga/. Retrospective plasma cfDNA specimens collected from 300 patients with the six GI cancers and healthy age-matched controls were collected from various institutes. Written informed consent was obtained from all patients and the study was approved by the Institutional Review Boards of all participating institutions. The study adhered to Declaration of Helsinki ethical guidelines.

Specimen processing of patient plasma samples

Plasma samples were transferred to 2-mL microcentrifuge tubes and centrifuged at 16,000 × *g* for 10 minutes at 4°C to remove any cellular debris. Circulating cfDNA (10–100 ng) was extracted from 1–2 mL plasma using the QIAamp Circulating Nucleic Acid kit (Qiagen) and quantified using the Quant-iT high-sensitivity Picogreen double-stranded DNA Assay Kit (Invitrogen by Thermo Fisher Scientific). For targeted methylation sequencing, 10 ng plasma cfDNA was first bisulfite treated using the ZYMO Gold Kit. A Swift Bioscience Methyl-Seq library preparation kit was adapted to generate individual libraries incorporating 13 PCR cycles and overnight ligation. Custom-targeted CpG methylation probes were designed using the Roche Nimblegen target capture kit, Custom SeqCap Epi Choice 30 MB. Libraries were quantified using the Quant-iT high-sensitivity Picogreen double-stranded DNA Assay Kit before equimolarly pooling 10 individual libraries per capture consisting of 2 μg total DNA. Hybridization and capture were performed using VK SeqCap Epi Reagent Kit Plus and SeqCap EZ hybridization/wash kit from Roche Nimblegen. For blocking, a universal blocker (IDT technologies) was used. Pooled libraries were sequenced on an Illumina NovaSeq S4 using paired-end, 100-bp reads, incorporating 150 individual libraries per lane. Sequencing matrices including the coverage distribution and methylation ratio distribution of gitBS in all plasma samples are included in Supplementary Figs. S1 and S2.

Plasma-targeted bisulfite data processing, differentially methylated region calling, and visualization

For each plasma sample, after trimming adaptor and low-quality bases, BSMAP (2.90) was used to align bisulfite sequencing reads to the hg19 human genome assembly. The methylation ratio of CpG sites was calculated using the methratio.py script (from BSMAP package). CpG methylation ratios supported by less than 4 reads were discarded before downstream analysis. Metilene (0.2–7) was used for calculating *de novo* differentially methylated regions (DMR) between two conditions, for example, normal versus cancer. For each CpG site, at least three samples of each condition must have a nonmissing value. Missing values were imputed using Metilene during DMR calling. Because the methylation difference between normal and cancer tissues is typically diluted in plasma, we selected DMRs based on a relatively loose cut-off (absolute methylation difference more than 0.1 and *P* value less than 0.05) for downstream analysis. The methylation level of a DMR was represented as the mean methylation ratio of its CpG sites. The z-score of each DMR methylation level was used for heatmap visualization. Ward clustering and Euclidean distance were used for heatmap plotting.

Machine learning methods used for developing various GI cancer detection panels

Feature selection for individual GI cancer detection and pan-GI cancer detection

For individual GI cancer prediction, normal and cancer plasma samples were randomly partitioned into a training set (70%) and a test set (30%). Within the training set only, DMR identification and feature selection (using the "Boruta" R package to select the top 200 informative DMRs) were performed in normal and cancer plasma samples for each GI cancer. For pan-GI cancer detection, samples from the

Published OnlineFirst August 31, 2021; DOI: 10.1158/1078-0432.CCR-21-1982

training sets or testing sets for each GI cancer were pooled into a single pan-GI training set or testing set, respectively. Using the training set, DMRs identified from each GI cancer were also pooled, for a total of approximately 8,000 DMRs for feature selection (using the Boruta R package to select the top 200 informative DMRs).

Feature selection for multi-GI cancer classification

Plasma samples from six GI cancers and healthy people were used for classification analysis. ESCC and EAC were combined as one class, given their high similarity. Plasma samples from each class were randomly partitioned into a training set (70%) and a test set (30%) independently. Using the training set, class-specific DMRs were identified by one-versus-rest comparisons. Approximately 4,000 DMRs identified from all classes were pooled together and the top 200 informative DMRs were selected (using Boruta R package with default parameters) for downstream GI cancer classification.

Feature selection using the Boruta R package

After splitting the data into training and test sets, the Boruta R package was used to select the most informative DMRs from the training set for cancer detection. Given the randomness introduced by missing value imputation and random forest construction, we repeated the feature selection step 50 times and finally choose the top 200 DMRs that were most frequently selected by the Boruta algorithm for subsequent analyses.

Prediction model training and evaluation

Training sets were used to train random forest (R package "ranger") models for individual GI cancer prediction, pan-GI cancer prediction, and multi-GI cancer classification, respectively. The hyperparameters were tuned by 10-fold cross-validation. For model evaluation, the remaining 30% test sets were used to plot the ROC curve and calculate the AUC scores for each random forest model. The training-test set split, DMR calling, and feature selection were repeated 10 times to avoid overestimating model performance.

Independent cohort validation

PDAC patient samples were from two independent cohorts [58 samples from University of Pittsburgh (Pittsburgh, PA) and 16 samples from Medical College of Wisconsin (MCW, Milwaukee, WI)]. The PDAC Pittsburg cohort, which has more patient samples, was used for DMR calling, feature selection (top 200 informative DMRs), and model training. The AUC scores for this model in detecting cancer were calculated using the PDAC MCW cohort.

Early-stage cancer prediction

For colorectal cancer, HCC, gastric cancer, and PDAC, cancer stage information was available and therefore we looked at the early-stage cancer prediction accuracy in these four cancers. For this, we took all late-stage (stage IV) cancer samples, along with 70% of the normal plasma samples for DMR calling, feature selection (top 200 informative DMRs), and model training. The performance of the trained model was then evaluated using the early-stage (stage I–III) cancer samples and the remaining 30% of normal samples.

Informative DMR validation using cancer tissue data

Calculated beta values of 450K methylation array data for TCGA-COAD, TCGA-LIHC, TCGA-ESCA, TCGA-STAD, and TCGA-PAAD were downloaded from the UCSC Xena database. Calculated beta values of 450K methylation array data for EAC was downloaded from Gene Expression Omnibus (GEO; GSE72872). The 450K CpG sites were mapped to the informative DMRs selected for individual GI cancer detection, pan-GI cancer detection, and multi-GI cancer classification. The methylation level of the informative DMRs for each cancer tissue sample was calculated by taking the mean of the mapped CpG site beta values. The normal and cancer tissue samples were randomly partitioned into a training (70%) and test set (30%) manner. We trained a random forest model with the training set and calculated the AUC scores of the model with the remaining test set.

Data availability

All data associated with this study are presented in the article or Supplementary Materials and Methods. The raw plasma cfDNA gitBS sequencing data reported in this article have been deposited into the GEO, under accession number GSE149438.

Results

Development of a GI-targeted bisulfite sequencing panel

The study design describing tissue discovery, followed by plasma cfDNA validation, is illustrated in **Fig. 1**; Supplementary Fig. S1A. We first analyzed 450K methylation array data from 1,781 tumor and adjacent normal tissues from six different GI cancers: colorectal cancer, HCC, ESCC, GC, EAC, and PDAC. By comparing data from tumor versus normal tissues within each GI cancer, as well as across all GI cancers, we identified a total of 67,832 regions of interest (ROI), based on significant differentially methylated probes with a $P < 0.001$ and an absolute delta beta of 0.20 across all the comparisons (Supplementary Table S1 and S2). The covered regions were highly enriched for promoters as well as gene body regions (Supplementary Fig. S3B), which are more susceptible to aberrant methylation during oncogenesis. We merged overlapping tissue-level ROIs from the various GI cancers to design a targeted SeqCap Epi-based bisulfite sequencing platform, which we termed the "GI-targeted bisulfite sequencing (gitBS)" panel (Supplementary Table S3). Compared with a previously reported strategy (10), we used a meticulous analysis of every significant probe identified via a 450K tissue analysis across six GI cancers to build our gitBS panel, which included a much broader genomic region (∼30 MB) covering approximately 1% of the human genome.

Evaluation of gitBS in plasma cfDNA

To evaluate the comprehensive list of tissue-specific markers in plasma cfDNA, we performed gitBS on 300 total plasma samples collected from patients with colorectal cancer, HCC, ESCC, gastric cancer, EAC, or PDAC, and age-matched controls (Supplementary Table S4). In comparing the individual GI cancers with controls, we identified a total of 216,887 differentially methylated CpGs consisting of 10,677 DMRs, in colorectal cancer (5,689), HCC (1,072), ESCC (1,063), gastric cancer (949), EAC (1,177), and PDAC (727; Supplementary Table S5). To confirm the diagnostic power of the identified DMR panels across each GI cancer, we performed hierarchical clustering based on the identified DMRs for each GI cancer type. For most GI cancers, we observed a clear separation of two clusters representing cancer versus normal samples (Supplementary Figs. S4–S8). For PDAC, although the boundary between cancer and normal clusters was less clear, most PDAC samples clustered together (Supplementary Fig. S9). Overall, our results indicate that these DMRs could be used as potential biomarkers for GI cancer detection.

Published OnlineFirst August 31, 2021; DOI: 10.1158/1078-0432.CCR-21-1982

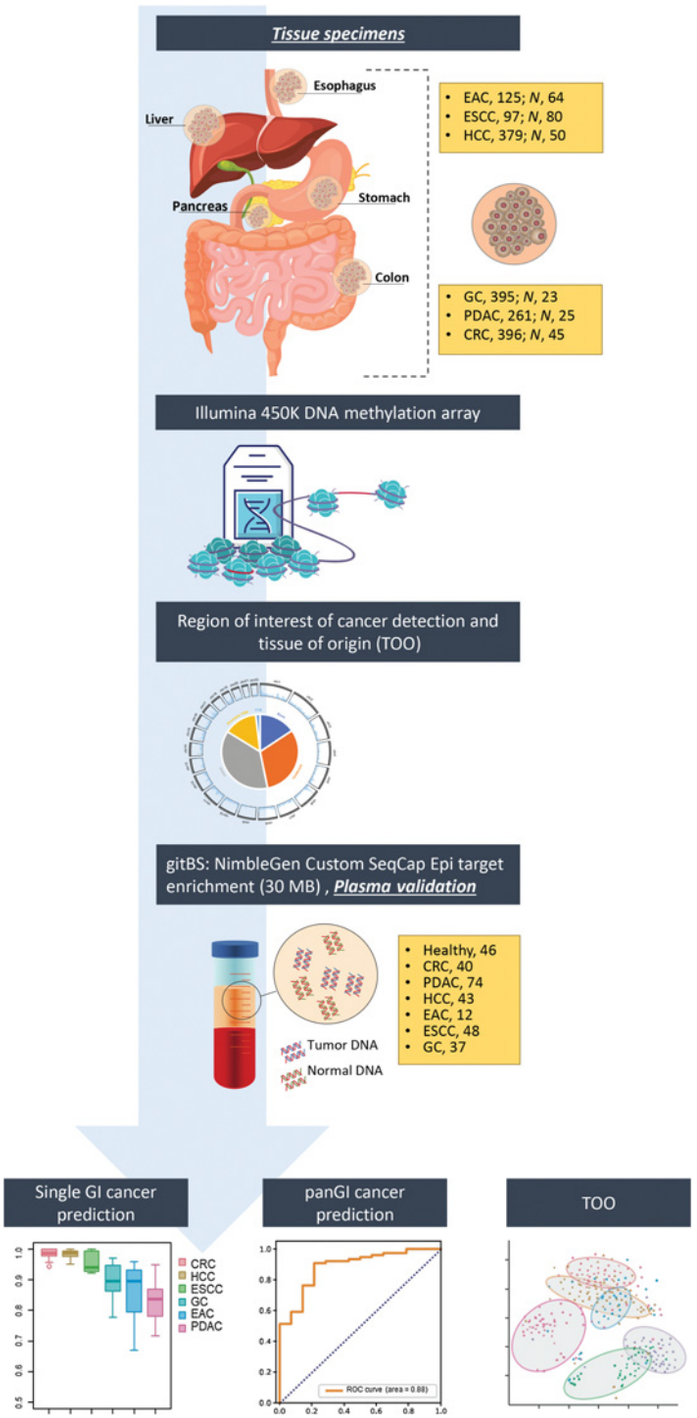

Figure 1.

Study design depicting tissue discovery and plasma validation of EpiPanGI Dx. Genome-wide 450K DNA methylation analysis on individual GI cancers versus adjacent normal tissues and across six GI cancers resulted in the identification of 67,832 DMRs of interest. Subsequently, we developed a custom plasma-specific gitBS target enrichment panel to evaluate in plasma cfDNA (n = 300). This resulted in the identification of plasma DMR panels for the detection of individual GI cancers, pan-GI cancers, and tissue of origin using machine learning algorithms.

Development of cfDNA methylation panels for individual GI cancer detection

To develop plasma-specific DMR panels for individual GI cancer detection, we used machine learning algorithms. Briefly, we have split plasma samples from patients with GI cancer and healthy controls into training (70%) and test sets (30%). *De novo* DMRs between GI cancer and healthy controls were identified only with samples from training sets. Next, we performed feature selection based on the Boruta algorithm, which is known to be powerful for biological features (15). We then used the chosen DMRs to train a random forest model, which outperformed several other machine learning techniques for GI cancer detection, such as logistic regression model, support vector machine and K-nearest neighbor models (Supplementary Fig. S10). We used PDAC for this comparison because PDAC plasma samples are not well separated from healthy controls based on the clustering result (Supplementary Fig. S9). Finally, we evaluated prediction model

Published OnlineFirst August 31, 2021; DOI: 10.1158/1078-0432.CCR-21-1982

performance by calculating area under the ROC curve (AUC) scores using the test set samples. We repeated the entire process 10 times to prevent biases due to dataset splitting. Our cancer prediction models achieved the best performance for colorectal cancer and HCC, with median AUC scores of 0.98; prediction models for the other GI cancers had median AUC scores of 0.94 (ESCC), 0.90, (gastric cancer), 0.90 (EAC), and 0.85 (PDAC), which is higher or comparable to previous reports (**Fig. 2A**; refs. 11, 16).

We subsequently applied the plasma derived DMR panels established using machine learning to distinguish GI cancer tissues from

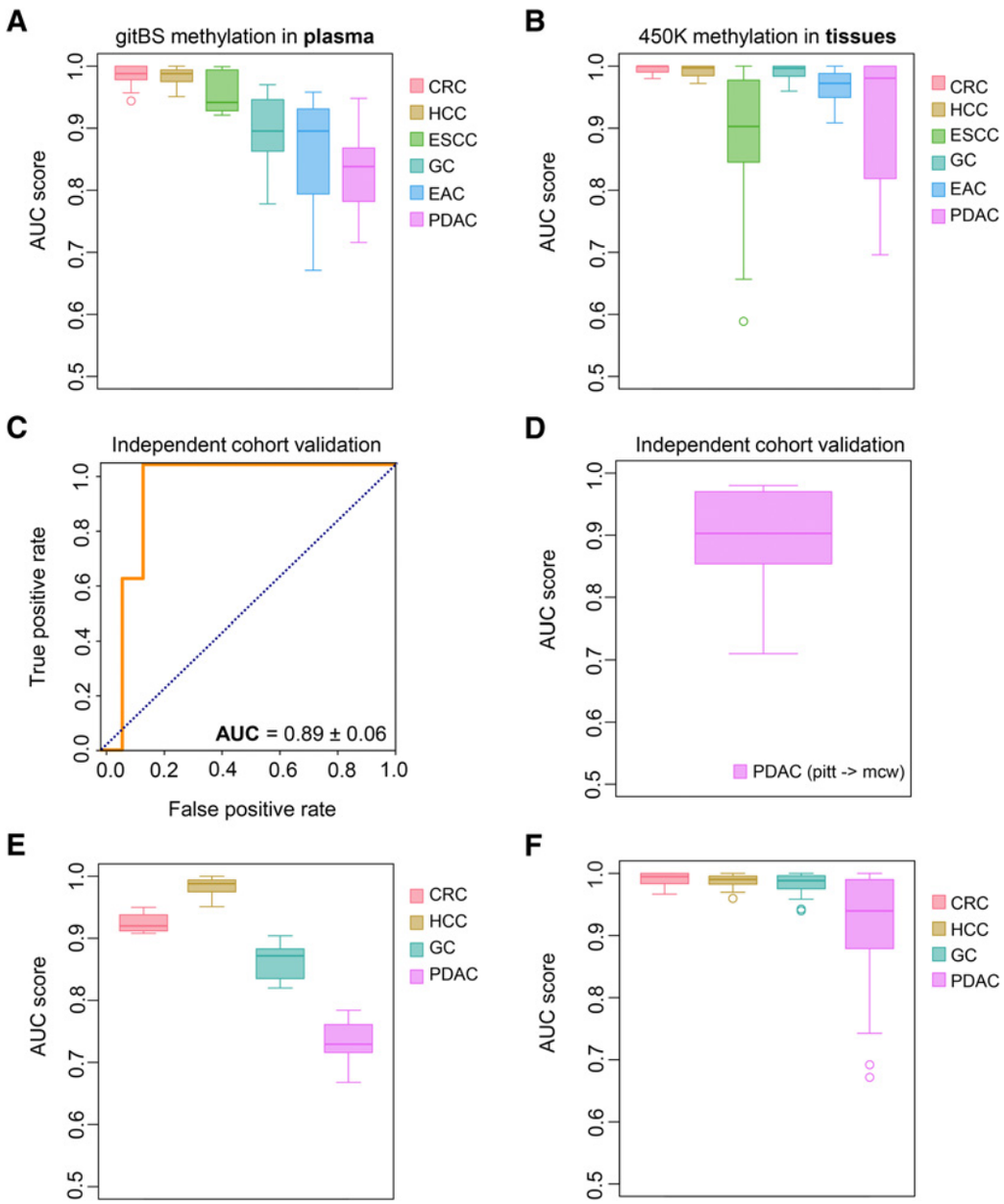

Figure 2.

Individual GI cancer detection accuracy using informative plasma DMRs identified from gitBS. **A,** Prediction accuracy of the machine learning model trained for each GI cancer. Samples ($n = 300$) were randomly partitioned into a training set (70%) and a test set (30%) 10 times. DMR calling, feature selection, and model training were performed on the training sets. Boxplots show prediction model AUC scores calculated in test sets for each GI cancer. Sample size: colorectal cancer (40), PDAC (74), HCC (43), EAC (12), ESCC (48), gastric cancer (37), normal (46). **B,** Use of the informative plasma DMRs from **A** to predict cancer in GI cancer tissues dataset ($n = 1,781$). Boxplots show AUC scores of 10 independent runs. **C** and **D,** Representative ROC curve (left) and AUC score (right) for the PDAC independent validation set (10 runs). **E,** Late-stage (stage IV) plasma samples, along with randomly selected 70% normal plasma sample, were used for DMR calling, feature selection, and model training. This whole process was repeated for 10 times to avoid bias due to sample selection. Boxplots show AUC scores of prediction models on early-stage (stage I–III) plasma samples (colorectal cancer: 29; HCC: 36; GC: 16; PDAC: 35). **F,** Use of informative plasma DMRs from **E** to predict cancer in early-stage GI cancer tissues ($n = 1,257$). GC, gastric cancer.

Published OnlineFirst August 31, 2021; DOI: 10.1158/1078-0432.CCR-21-1982

adjacent normal tissues. As expected, the median AUC scores of models for each of the GI cancers were 0.99 (colorectal cancer), 0.99 (HCC), 0.90 (ESCC), 1.00 (gastric cancer), 0.97 (EAC), and 0.94 (PDAC). Consistent with the performance of the PDAC model in plasma, the model performed relatively poorly at predicting PDAC in tissue (**Fig. 2B**). Therefore, we tested the PDAC DMRs in another independent plasma cohort. Interestingly, the machine learning model, trained and tested with PDAC plasma samples from the first cohort, achieved even higher prediction accuracy in the independent PDAC cohort, with an AUC of 0.89 (**Fig. 2C** and **D**).

Given that the ultimate goal of cancer screening is to identify cancer at an early stage, we evaluated the ability of the plasma DMRs to detect early-stage GI cancers in colorectal cancer (29), HCC (36), gastric cancer (16), and PDAC (35). We did not have access to early-stage EAC and ESCC and hence are not tested. Our models achieved median AUC scores of 0.92 (colorectal cancer), 0.99 (HCC), 0.87 (gastric cancer), and 0.73 (PDAC) for predicting early-stage plasma samples in the test set (**Fig. 2E**). When applied to early-stage tumor tissues in the same four GI cancers, the DMR panels achieved median AUC values of 0.99 (colorectal cancer), 0.99 (HCC), 0.99 (gastric cancer), and 0.94 (PDAC; **Fig. 2F**). Altogether, these results indicate that DNA methylation aberrations we identified have great potential for detecting individual GI cancers along with early-stage cancers.

Development of a pan-GI cancer detection model

Having performed this study in individual GI cancers, we next used our DMR data to identify a pan-GI classifier. To do this, we pooled the training sets and test sets used for each individual GI cancer prediction model together as a pan-GI training set and test set, respectively. We also pooled the DMRs identified from each GI cancer for pan-GI cancer feature selection and model training. We achieved a median AUC of 0.88 for the pan-GI cancer prediction model in the test-set plasma cohort (**Fig. 3A**). Similarly, the plasma DMRs achieved an excellent AUC of 0.98 in distinguishing pan-GI cancer tissues from normal tissues (**Fig. 3B**).

Development of multi-GI cancer classification model EpiPanGIDx

Finally, we have developed a plasma multi-GI cancer prediction model EpiPanGIDx using random forest that in addition to identify all GI cancers, also have the ability to reveal the tissue of origin. Given that ESCC and EAC both develop from the esophagus, we treated them as the same class in our model. For each class versus the other GI cancers, we identified class-specific plasma DMRs (Supplementary Table S6), which we then pooled for feature selection and model training. In the test set, our models classified samples into normal plasma, colorectal cancer, PDAC, HCC and ESCC/EAC with higher accuracy than previous studies (**Fig. 4A**; ref. 16). Clustering the data using a t-SNE plot also showed clear separation of most GI cancers from healthy samples and from one another (**Fig. 4C**). The class-specific plasma DMRs also successfully classified GI cancer and normal tissues with high accuracy (**Fig. 4B** and **D**). Collectively, these results prove the feasibility of utilizing cfDNA methylation markers for not only GI cancer detection, but also for identifying the tissue of origin of GI cancers.

Identification of minimum DMRs needed to achieve optimal accuracy across all GI cancers

Finally, to advance the development of powerful and cost-effective cfDNA methylation biomarker panels for GI cancer detection, we also evaluated the performance of our models when varying number of informative DMRs were selected for model training. For individual GI cancer prediction models, the top 50 DMRs were sufficient for achieving optimal accuracy for each GI cancer. Even with as few as 10 DMRs, models for HCC or colorectal cancer prediction still showed excellent performance, with AUC scores >0.95 (**Fig. 5**; Supplementary Figs. S11–S16; Supplementary Table S7). For both the pan-GI and multi-GI classification models, at least the top 150 informative DMRs were required to achieve the optimal performance (**Fig. 5**; Supplementary Figs. S17–S19; Supplementary Table S7).

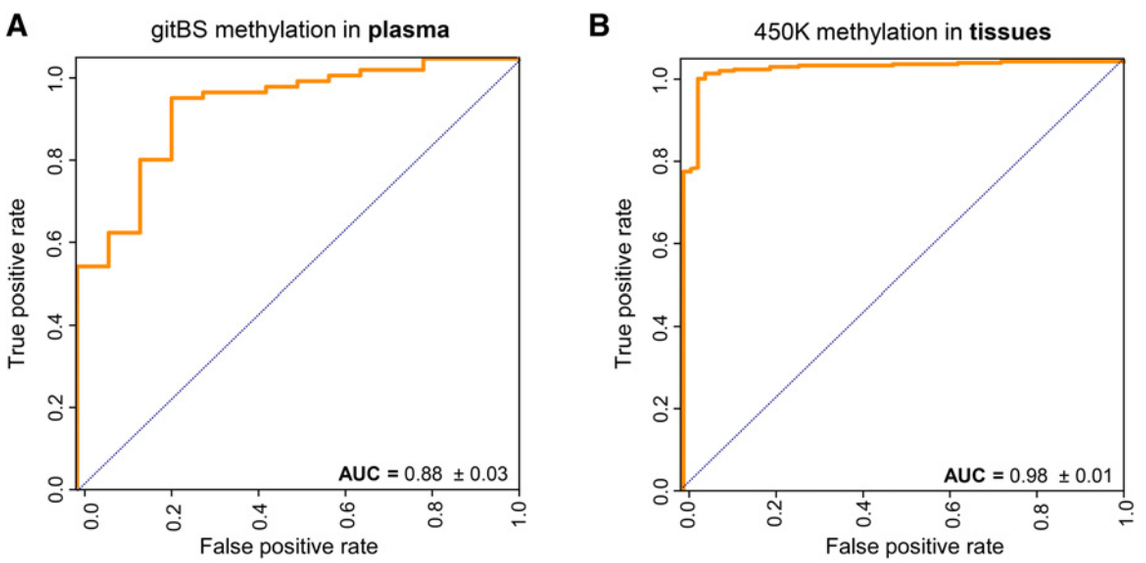

Figure 3.

pan-GI cancer detection accuracy using informative plasma DMRs identified from gitBS. **A,** Plasma samples of each GI cancer were randomly partitioned into a training set (70%) and test set (30%) 10 times. Training sets of all GI cancers were pooled for training a pan-GI cancer prediction model. Representative ROC curve and AUC scores for the combined test sets are shown. **B,** Use of informative plasma DMRs from **A** to predict pan-GI cancer in tissue samples.

Published OnlineFirst August 31, 2021; DOI: 10.1158/1078-0432.CCR-21-1982

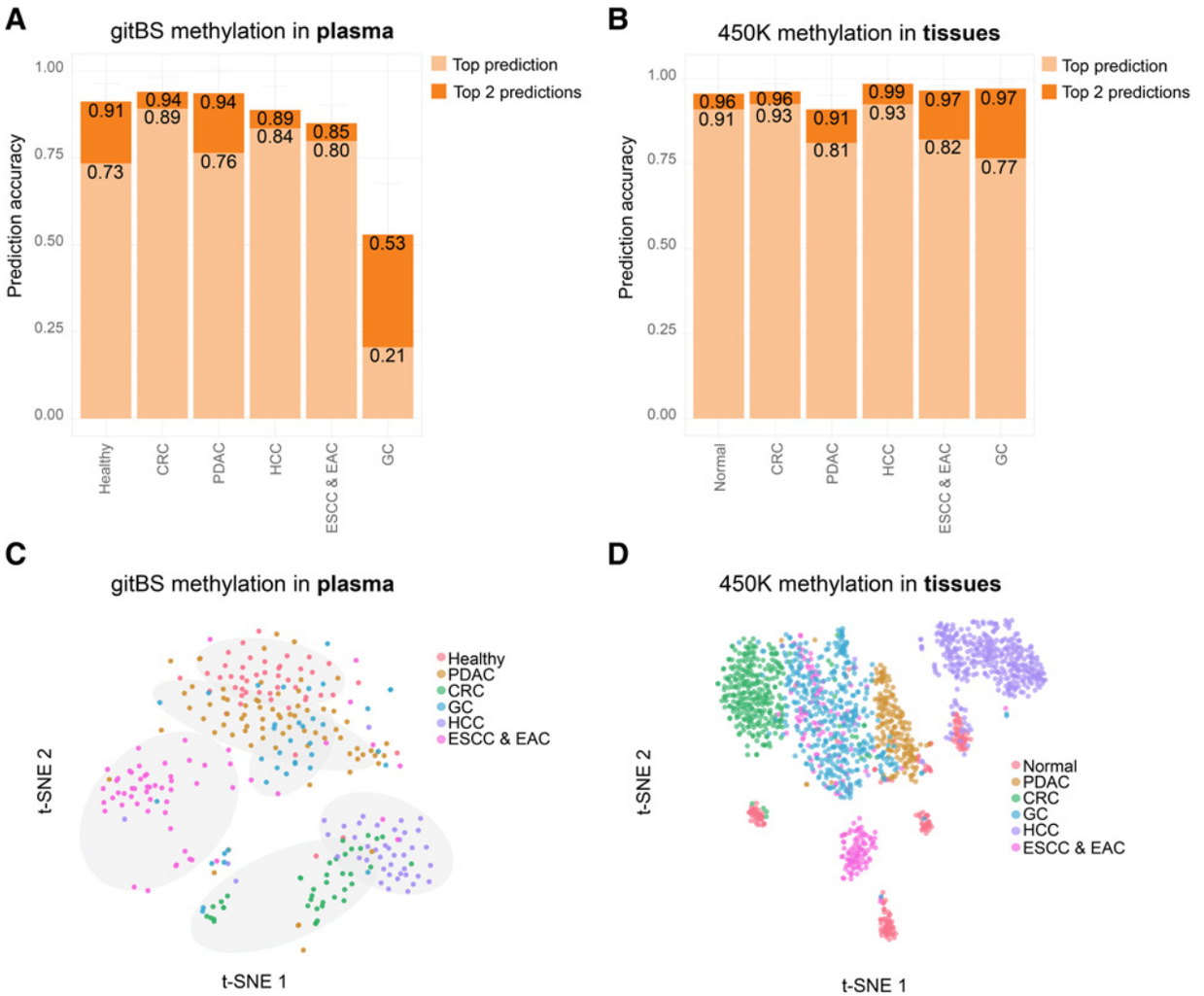

Figure 4.

Multi-GI cancer tissue of origin classification using informative plasma DMRs identified from gitBS. **A,** Classification accuracy of the plasma samples from patients with GI cancer. The number on the *y*-axis indicates the ratio of samples being correctly predicted. Light orange: sample labels were the same as the top prediction. Dark orange: sample labels were among the top 2 predictions. **B,** Use of informative plasma DMRs from **A** to classify GI cancer tissues. **C** and **D,** t-SNE plots for plasma samples ($n = 300$) and GI cancer tissue samples (1,781) generated using informative plasma DMRs.

Discussion

The lack of population-based screening for all cancers is attributed to the low prevalence of many cancers in the general population (3, 17). However, by developing sensitive multi-cancer or multi-organ diagnostic tests, population screening could be implemented, even for low-prevalence cancers. In this regard, GI cancers, which encompass a variety of cancer types, provide a unique opportunity for developing a pan-GI diagnostic assay. Ahlquist and colleagues, showed that using a pan-GI diagnostic assay, only 83 patients need to be screened to diagnose 1 positive patient with GI cancer (3). Herein, we performed a comprehensive genome-wide DNA methylation study across six GI cancers to identify a noninvasive plasma DMR panel "EpiPanGI Dx" that predicted tissue of origin of all GI cancers with high accuracy.

Most previous studies either studied individual GI cancers (9, 10, 18) or selected a panel of significant tissue markers and subsequently validated them in cfDNA using PCR-based methods (19, 20). Thus, cancer specificity was not well studied, and those studies failed to build multi-organ diagnostic assays to implement cost-effective population screening tests. In contrast, we first identified every tissue-significant CpG across six GI cancers, followed by development of plasma-specific diagnostic panels for the accurate detection of GI cancer tissues of origin using a single targeted methylation test, EpiPanGI Dx. Compared with previous studies (11), we selected fewer DMRs for prediction, which makes our model more feasible for large-scale validation studies and clinical practice (**Fig. 5**; Supplementary Figs. S11–S19). In addition, a low cost per sample, as well as a low (10 ng) input required for cfDNA, makes our targeted methylation assay very feasible for clinical use.

Recent plasma cfDNA methylation studies showed that targeted methylation sequencing is quite robust in discovering multi-tissue

Published OnlineFirst August 31, 2021; DOI: 10.1158/1078-0432.CCR-21-1982

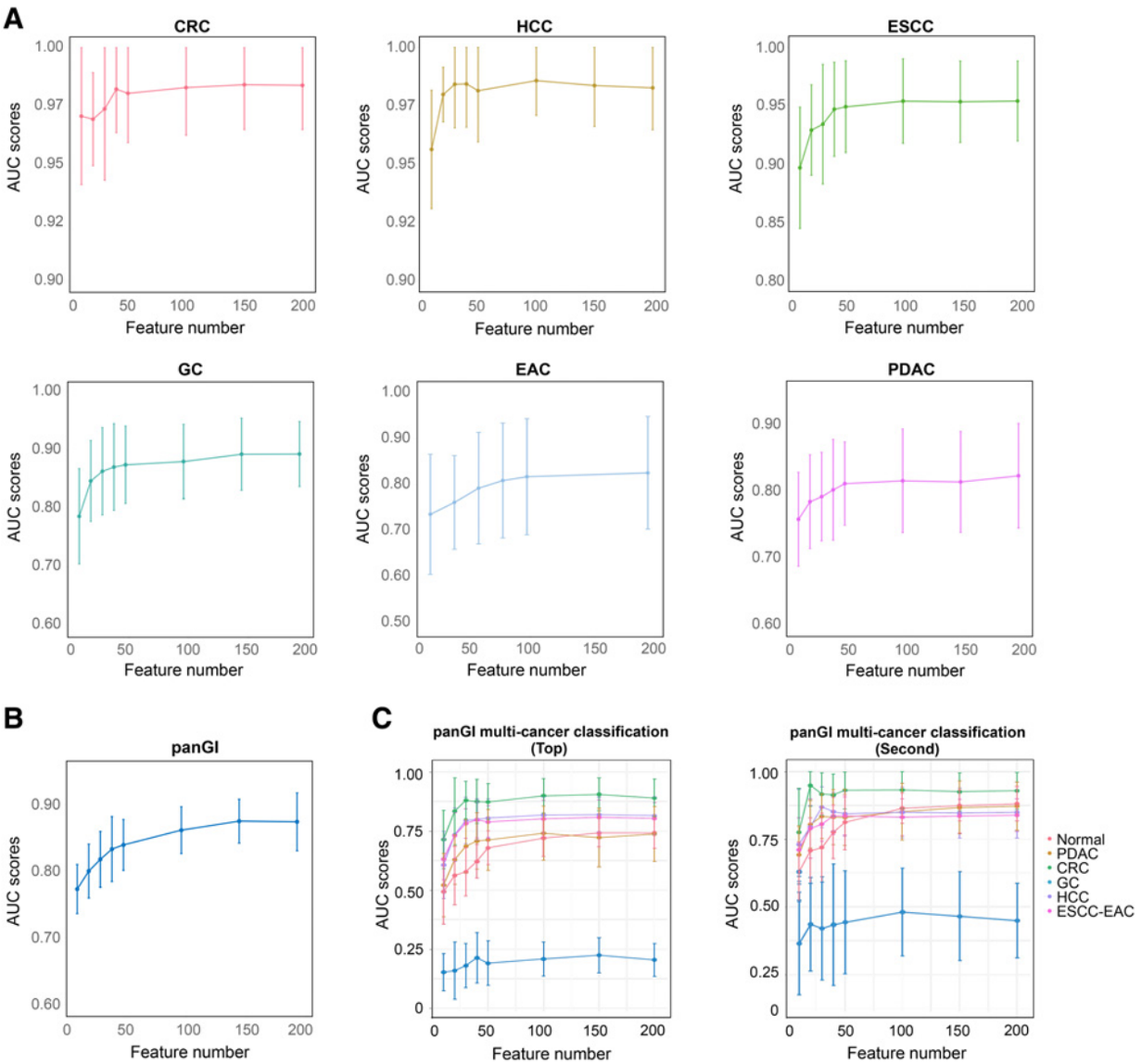

Figure 5.
AUC plots with variable numbers of informative DMRs across GI cancers. **A,** Individual GI cancer prediction models. **B,** pan-GI cancer prediction model. **C,** Multi-GI cancer tissue of origin classification model.

cfDNA methylation markers. Most notably, Liu and colleagues (21). identified tissue of origin methylation markers across 50 different cancers. Another study used targeted methylation sequencing to identify plasma cfDNA markers using that differentiate between colorectal cancer, non–small cell lung cancer, breast cancer, and melanoma (22). Shen and colleagues used a cfMeDIP-seq method to discover plasma DMRs that differentiate between multiple solid cancers including pancreatic, colorectal cancer, breast, lung, renal, and bladder cancers (11). However, ours is the first study in which organ-specific methylation markers were used to develop a multi-GI cancer cfDNA assay. Excitingly, the detection accuracy of our EpiPanGI Dx assay, with as few as 50 DMRs, was quite high across all GI cancers, considering it is a multi-cancer diagnostic test. Furthermore, our EpiPanGI Dx assay developed from plasma cfDNA showed excellent diagnostic accuracy (AUC, 0.91–0.99) when applied back to GI cancer tissue cohorts. Thus, the markers we trained and validated in plasma cfDNA are highly cancer specific. PDAC showed somewhat lower accuracy and this could be attributed to the tumor purity and further validation of our markers can help us refining the signatures in PDAC.

The unique strengths of our study are: (i) Comprehensive profiling of all GI cancer tissue methylation markers followed by the development of a targeted plasma cfDNA panel for the development of EpiPanGI Dx. (ii) Use of machine learning algorithms with training and validation sets, as well as using 10× cross-validation, to compute the accuracy of the EpiPanGI Dx assay across GI cancers. (iii) In

addition, the assay is quite cost-effective as our models require fewer biomarkers than previously reported studies (11) and therefore will be more feasible for the development of diagnostic panels for large-scale clinical usage. (iv) Our assay can be performed using as little as 10 ng cfDNA. (v) Although the plasma samples were collected from several different parts of the world, the detection accuracy of the EpiPanGI Dx assay in cfDNA, as well as the performance of the test in tissue data, shows the robustness of our markers.

Our study also has several limitations. First, the study is retrospective; therefore, we could not test the true population screening ability of our models. Second, although we showed our assay to be quite robust in identifying early-stage cancers at both the tissue and plasma level, the number of samples used to represent each stage was limited. Third, although many previous studies showed the superiority of cfDNA methylation markers over genomic mutations for cancer detection, we did not have the mutation profiles of our cfDNA samples to be able to directly compare (or even combine) the diagnostic performance of our methylation assay relative to genomic mutations. However, in future studies, we expect a combination of epigenomic and genomic markers to further improve the accuracy and robustness of cfDNA-based early detection markers.

In summary, we developed a robust, sensitive, targeted methylation-based cfDNA test for multi-GI cancer detection. Our EpiPanGI Dx assay needs further validation in completely independent retrospective and prospective datasets for clinical translation as early diagnostic markers. Further large-scale prospective validation will pave a way to test the performance of our assay in population-level screening for GI cancers. Nevertheless, these findings underscore the potential utility of cfDNA methylation markers for noninvasive, cost-effective, and early detection of GI cancers and serve as a platform for future organ-specific methylation studies for multi-cancer detection.

Authors' Disclosures

J. Xu reports patent number 63/233,957 pending for Compositions and methods for cell-free DNA epigenetic gastrointestinal cancer detection and treatment and is a full-time employee at Helio Health. A. Link reports grants from EFRE (ZS/2018/11/95324) outside the submitted work. M.I. Parker reports grants from Medical Research Council during the conduct of the study. H. Uetake reports grants from Taiho, Takeda, and Chugai outside the submitted work. F. Balaguer reports other support from Elsevier, Sysmex, Norgine, and CPP outside the submitted work. E. Borazanci reports consulting with BioNTech and Imaging Endpoints; received speaker's bureau honoraria from Ipsen; and reports institutional support from Helix, Fujifilm, Nektar, Stand up to Cancer, Seena Magowitz Foundation, Minneamrita Therapeutics, Idera, Eli Lilly, BioNTech, and Bioline. S.J. Meltzer reports other support from Capsulomics, Inc. during the conduct of the study. R.E. Brand reports grants from NCI during the conduct of the study as well as grants from Immunovia and Freenome, Inc. outside the submitted work. D. Von Hoff reports stock and other ownership interests in Medtronic, CerRX, SynDevRx, UnitedHealthcare, Anthem Inc., Stromatis Pharma, Systems Oncology, Stingray Therapeutics, FORMA Therapeutics, Orpheus Bioscience, and AADi; consulting or advisory roles with DNAtrix, Esperance Pharmaceuticals, Imaging Endpoints, Immodulon Therapeutics, Senhwa Biosciences, Tolero Pharmaceuticals, Alpha Cancer Technologies, CanBas, Lixte Biotechnology, Oncolyze, RenovoRx, TD2, Aptose Biosciences, CV6 Therapeutics, EMD Serono, Fujifilm, Kura Oncology, Phosplatic Therapeutics, SOTIO, Strategia Therapeutics, Synergene, 7 Hills Pharma, Cancer Prevention Pharmaceuticals, Geistlich Pharma, HUYA Bioscience International, Immunophotonics, Genzada Pharmaceuticals, L.E.A.F. Pharmaceuticals, Oncology Venture, TP Therapeutics, Verily, Athenex, Novita Pharmaceuticals, Nucana, Vicus Therapeutics, Codiak Biosciences, Agenus, Kelun, RadImmune, Samumed, BioXCel Therapeutics, Bryologyx, Sirnaomics, AiMed, Corcept Therapeutics, Erimos Pharmaceuticals, Gimbal, Pfizer, GiraFpharma, Axis Therapeutics, DrugCendR, ImmuneOncia, Orphagen Pharmaceuticals, Array BioPharma, MaveriX Oncology, Northern Biologics, Viracta Therapeutics, Varian Biopharma, Xerient, AlaMab Therapeutics, Avesta76 Therapeutics, Bessor Pharma, NeoTx, Xerient, Decoy Biosystems, Noxxon Pharma, Reflexion Medical, Reglagene, Lycia Therapeutics, NGM Biopharmaceuticals, Coordination Pharmaceuticals, EXACT Therapeutics, Nirogy Therapeutics, Seagen, Agastiya Biotech, Amunix, Atalion Therapeutics, Cytocom, GlaxoSmithKline, ImaginAb, SignaBlok, SonaCare Medical, Caribou Biosciences, and Xenter; and research funding (to institution) from Lilly, Genentech, Celgene, Incyte, Merrimack, Plexxikon, Minneamrita Therapeutics, Abbvie, Aduro Biotech, Cleave Biosciences, CytRx Corportation, Daiichi Sankyo, Deciphera, Endocyte, Exelixis, Five Prime Therapeutics, Gilead Sciences, Merck, Pfizer, Pharamcyclics, Phoenix Biotech, Samumed, Strategia, and Halozyme outside the submitted work. D. Von Hoff also reports patents, royalties, and other intellectual property on the following topics: intramedullary catheter; methods of human prostate cancer; use of 5,6-dihydro-5-azacytidine in the treatment of prostate cancer; targeting site-2 protease (S2P) for the treatment of pancreatic cancer (pending); targeting ecto-5-nucleotidase (Cd73) for the treatment of pancreatic cancer; targeting a protein tyrosine phosphotase-PRL-1 for the treatment of pancreatic cancer (pending); targeting a protein PRC1 for the treatment of pancreatic cancer (pending); protein kinase inhibitors (pending); methods, compounds, and compositions with genotype selective anticancer activity (pending); methods and kits to predict therapeutic outcome of BTK inhibitors (pending); muscle fatigue substance cytokines and methods of inhibiting tumor growth therewith (pending); and 2-aryl-pyridylazoles for the treatment of solid tumors such as pancreatic cancer (pending). W. Li reports grants from NIH during the conduct of the study; in addition, W. Li has a patent for EpiPanGI Dx pending and is a consultant for Helio Health and ChosenMed. A. Goel reports patent number 63/233,957 pending for Composition and methods for cell-free DNA epigenetic gastrointestinal cancer detection and treatment. No disclosures were reported by the other authors.

Authors' Contributions

R. Kandimalla: Formal analysis, investigation, writing–original draft, writing–review and editing. **J. Xu:** Data curation, formal analysis, investigation, writing–original draft, writing–review and editing. **A. Link:** Resources, writing–review and editing. **T. Matsuyama:** Resources. **K. Yamamura:** Resources, writing–original draft. **M.I. Parker:** Resources. **H. Uetake:** Resources, supervision. **F. Balaguer:** Resources, writing–original draft. **E. Borazanci:** Resources. **S. Tsai:** Resources. **D. Evans:** Resources. **S.J. Meltzer:** Resources. **H. Baba:** Resources, supervision. **R. Brand:** Resources. **D. Von Hoff:** Resources, supervision, writing–original draft. **W. Li:** Data curation, formal analysis, supervision, writing–original draft, writing–review and editing. **A. Goel:** Conceptualization, supervision, funding acquisition, writing–review and editing.

Acknowledgments

This work was supported by R01 (CA72851, CA181572, CA184792, CA202797) and U01 (CA187956, CA214254) grants from the NCI, NIH; RP140784 from the Cancer Prevention Research Institute of Texas; grants from the Sammons Cancer Center and Baylor Foundation, as well as funds from the Baylor Scott & White Research Institute, Dallas, TX awarded to A. Goel. This work was also supported by R01 (HG007538, CA193466, CA228140) grants awarded to W. Li. M.I. Parker was jointly supported by the SAMRC with funds received from the National Department of Health and the MRC UK with funds from the UK Government's Newton Fund and GSK. S.J. Meltzer was supported by the ACS Clinical Research Professorship, DK118250 and CA211457, and endowed professorship Myerberg/Hendrix Professor of Gastroenterology.

We thank Feng Gao for participating and sharing his expertise in analyzing the 450K tissue data.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received May 29, 2021; revised July 24, 2021; accepted August 26, 2021; published first August 31, 2021.

Published OnlineFirst August 31, 2021; DOI: 10.1158/1078-0432.CCR-21-1982

References

1. Torre LA, Siegel RL, Ward EM, Jemal A. Global cancer incidence and mortality rates and trends–an update. Cancer Epidemiol Biomarkers Prev 2016;25:16–27.
2. Smith RA, Andrews KS, Brooks D, Fedewa SA, Manassaram-Baptiste D, Saslow D, et al. Cancer screening in the United States, 2019: a review of current American Cancer Society guidelines and current issues in cancer screening. CA Cancer J Clin 2019;69:184–210.
3. Ahlquist DA. Universal cancer screening: revolutionary, rational, and realizable. NPJ Precis Oncol 2018;2:23.
4. van der Pol Y, Mouliere F. Toward the early detection of cancer by decoding the epigenetic and environmental fingerprints of cell-free DNA. Cancer Cell 2019; 36:350–68.
5. Mroz EA, Rocco JW. The challenges of tumor genetic diversity. Cancer 2017;123: 917–27.
6. Lam K, Pan K, Linnekamp JF, Medema JP, Kandimalla R. DNA methylation based biomarkers in colorectal cancer: a systematic review. Biochim Biophys Acta 2016;1866:106–20.
7. Kandimalla R, van Tilborg AA, Zwarthoff EC. DNA methylation-based biomarkers in bladder cancer. Nat Rev Urol 2013;10:327–35.
8. Moss J, Magenheim J, Neiman D, Zemmour H, Loyfer N, Korach A, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat Commun 2018;9:5068.
9. Xu RH, Wei W, Krawczyk M, Wang W, Luo H, Flagg K, et al. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. Nat Mater 2017;16:1155–61.
10. Luo H, Zhao Q, Wei W, Zheng L, Yi S, Li G, et al. Circulating tumor DNA methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer. Sci Transl Med 2020;12:eaax7533.
11. Shen SY, Singhania R, Fehringer G, Chakravarthy A, Roehrl MHA, Chadwick D, et al. Sensitive tumour detection and classification using plasma cell-free DNA methylomes. Nature 2018;563:579–83.
12. Guo S, Diep D, Plongthongkum N, Fung HL, Zhang K, Zhang K. Identification of methylation haplotype blocks aids in deconvolution of heterogeneous tissue samples and tumor tissue-of-origin mapping from plasma DNA. Nat Genet 2017;49:635–42.
13. Provenzale D, Gupta S, Ahnen DJ, Markowitz AJ, Chung DC, Mayer RJ, et al. NCCN guidelines insights: colorectal cancer screening, version 1.2018. J Natl Compr Canc Netw 2018;16:939–49.
14. Krause L, Nones K, Loffler KA, Nancarrow D, Oey H, Tang YH, et al. Identification of the CIMP-like subtype and aberrant methylation of members of the chromosomal segregation and spindle assembly pathways in esophageal adenocarcinoma. Carcinogenesis 2016;37:356–65.
15. Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. Brief Bioinform 2017;20:492–503.
16. Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. Science 2018;359:926–30.
17. Cole P, Morrison AS. Basic issues in population screening for cancer. J Natl Cancer Inst 1980;64:1263–72.
18. Qin Y, Wu CW, Taylor WR, Sawas T, Burger KN, Mahoney DW, et al. Discovery, validation, and application of novel methylated DNA markers for detection of esophageal cancer in plasma. Clin Cancer Res 2019;25: 7396–404.
19. Eissa MAL, Lerner L, Abdelfatah E, Shankar N, Canner JK, Hasan NM, et al. Promoter methylation of ADAMTS1 and BNC1 as potential biomarkers for early detection of pancreatic cancer in blood. Clin Epigenetics 2019;11:59.
20. Freitas M, Ferreira F, Carvalho S, Silva F, Lopes P, Antunes L, et al. A novel DNA methylation panel accurately detects colorectal cancer independently of molecular pathway. J Transl Med 2018;16:45.
21. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden M. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. Ann Oncol 2020;31:745–59.
22. Liu L, Toung JM, Jassowicz AF, Vijayaraghavan R, Kang H, Zhang R, et al. Targeted methylation sequencing of plasma cell-free DNA for cancer detection and classification. Ann Oncol 2018;29:1445–53.

Published OnlineFirst August 31, 2021; DOI: 10.1158/1078-0432.CCR-21-1982

Clinical Cancer Research

EpiPanGI Dx: A Cell-free DNA Methylation Fingerprint for the Early Detection of Gastrointestinal Cancers

Raju Kandimalla, Jianfeng Xu, Alexander Link, et al.

Clin Cancer Res 2021;27:6135-6144. Published OnlineFirst August 31, 2021.