Hydroxymethylation profile of cell free DNA is a biomarker for early colorectal cancer

Nicolas Walker
Cambridge Epigenetix
Mamunur Rashid
Cambridge Epigenetix
Shirong Yu
Cambridge Epigenetix
Helen Bignell
Cambridge Epigenetix
Casper Lumby
Cambridge Epigenetix https://orcid.org/0000-0001-8329-9228
Carmen Livi
Cambridge Epigenetix
Kate Howell
Cambridge Epigenetix https://orcid.org/0000-0001-8069-920X
David Morley
Cambridge Epigenetix
Sandro Morganella
European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus,vHinxton, Cambridgeshire, CB10 1SD
Daniel Barrell
Cambridge Epigenetix
Shabhonam Caim
Cambridge Epigenetix
Walraj Gosal
Cambridge Epigenetix
Jens Fullgrabe
Cambridge Epigenetix
Tom Charlesworth
Cambridge Epigenetix
Louella Vasquez
Cambridge Epigenetix
Miika Ahdesmaki
Cambridge Epigenetix

Jordan Eizenga
Amanita Informatics
Parul Prabhat
Cambridge Epigenetix
Vitali Proutski
Cambridge Epigenetix
Marie Murat-Onana
Cambridge Epigenetix
Catherine Greenwood
Cambridge Epigenetix
Lisa Kirkwood
Cambridge Epigenetix
Meeta Maisuria-Armer
Cambridge Epigenetix
Mengjie Li
Cambridge Epigenetix
Emma Coats
Cambridge Epigenetix
Victoria Winfield
Cambridge Epigenetix
Lachlan Macbean
Cambridge Epigenetix
Toby Stock
Cambridge Epigenetix
Alice Tome-Fernandez
Cambridge Epigenetix
Yat Chan
Cambridge Epigenetix
Nasir Sheikh
Cambridge Epigenetix
Paula Golder
Cambridge Epigenetix
Tobias Ost
Cambridge Epigenetix
Michael Steward
Cambridge Epigenetix
Douglas Stewart
Cambridge Epigenetix
Albert Vilella

Cambridge Epigenetix
Mojtaba Noursalehi
Biostatistics and Research, LLC
Benedict Paten
Amanita Informatics LLC
Deborah Lucarelli
Cambridge Epigenetix
Joanne Mason
Cambridge Epigenetix
Gareth Ridge
Cambridge Epigenetix
Jason Mellad
Cambridge Epigenetix
Suman Shirodkar
Cambridge Epigenetix
Shankar Balasubramanian
Cancer Research UK Cambridge Institute https://orcid.org/0000-0002-0281-5815
Joanna Holbrook (Joanna.Holbrook@cegx.co.uk)
Cambridge Epigenetix

Article

Keywords: colorectal cancer (CRC), screening, early detection, early-stage cancer

Posted Date: July 20th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-667874/v1

Version of Record: A version of this preprint was published at Scientific Reports on October 4th, 2022. See the published version at https://doi.org/10.1038/s41598-022-20975-1.

Abstract

Early detection of colorectal cancer (CRC) will improve survival rates. We created a classifier to detect CRC, based on 5-hydroxymethylcytosine levels in cell free DNA isolated from blood samples of 2198 individuals. Our classifier discriminated CRC samples from controls with an area under the receiver operating characteristic curve (AUC) of 90% (sensitivity was 55% at 95% specificity). Performance was similar for early stage 1 (AUC 89%) and late stage 4 CRC (AUC 94%). Performance was independent of the proportion of tumor-DNA in the cell free DNA. We expanded the classifier to include information about cell free DNA fragment size and abundance across the genome. Overall performance was similar (AUC 91%), with gains in sensitivity (63% at 95% specificity). The 5-hydroxymethylcytosine signal allows detection of CRC, even in cell free DNA samples with undetectable tumor DNA. Including 5-hydroxymethylcytosine in multi-analyte screening, will improve sensitivity for early-stage cancer.

1 Introduction

Detection and treatment of cancer when the disease is still at early stage could save many lives, reduce morbidity, and relieve the burden of cancer on healthcare systems[1,2].

Liquid biopsy holds much promise as a minimally invasive method to detect early cancer in body fluids such as blood and urine[3–5]. There are many challenges to developing a liquid biopsy test that is sufficiently powerful to detect cancer in blood samples, especially at early- stage disease when the tumor is still small and releasing only minute amounts of biomarkers into the blood stream. Tumor DNA (ctDNA) represents just 0.1-1% of overall cell free DNA

(cfDNA) in early disease[6–8]. Multi-analyte approaches that measure multiple biomarkers from the same blood sample could be useful for detection of early-stage cancer[9,10].

Changes in the epigenome may precede genetic changes in tumorgenesis[11]. Therefore, there is growing interest in utilizing the epigenome of cfDNA for cancer detection. Several groups have investigated whether cancer can be detected via epigenetic changes in the tumor fraction of cfDNA such as DNA methylation[8,12], DNA hydroxymethylation[13–15] and cfDNA characteristics which may reveal chromatin structure[16].

Methylation of cystosine bases to produce methylcytosine (5mC) is a well-known epigenetic mechanism controlling gene expression. 5mC is oxidized to form hydroxymethylcytosine (5hmC), formylcytosine (5fC) and carboxylcytosine (5caC). 5mC and 5hmC have different functional roles: 5mC is present in heterochromatin and euchromatin and generally represses gene expression[17] whereas 5hmC is mainly present in euchromatin and is associated with the mostly highly transcribed gene bodies and their enhancers[18,19], as well as with poised enhancers about to transition from inactive to active[20]. 5mC and 5hmC have differential affinity to epigenetic readers; for instance, methyl binding proteins (MBD) preferentially bind 5mC. UHRF2 has been reported to have preferential affinity for 5hmC[21] and there are a

limited number of proteins that bind both modifications, for example MeCAP2[22–24]. Both marks are actively replaced after mitosis[25].

The different functional roles and distribution in cancer samples suggest 5mC and 5hmC have independent utility as biomarkers[22,26]. However, to date, the hunt for epigenetic markers of cancer has been constrained by available technologies, with limited options available to distinguish between 5mC and 5hmC. Efforts have traditionally focused on the use of bisulfite to sequence both 5mC and 5hmC, without distinguishing one from another[8,27].

Several techniques have recently emerged to quantify and utilize 5-hydroxymethylome epigenetic signatures for cancer detection via liquid biopsy[13–15, 28,29]. However, these studies have been limited by their small sample size and limited quantitative performance of the methodology used. The information provided about cancer by 5hmC profiles has been shown to be orthogonal and additive to 5mC[14,30].

In this study, we used a new method for precisely measuring 5hmC levels in the cell free genome and applied machine learning to train a classifier to distinguish individuals with colorectal cancer (CRC) versus controls (and other cancers) for 2,483 samples. We assessed the dependence of the classifier on cancer stage and on ctDNA levels. We evaluated how classifier performance evolved when orthogonal information about cfDNA fragment depth size and breakpoints characteristics were added to 5hmC information. The study showed 5hmC profiles of cfDNA are a strong predictor of cancer and particularly sensitive for detection of early-stage cancer. The performance achieved in this well controlled study - the largest of its kind to date - is at least comparable to performance reported for other analytes in smaller studies[8,12,31–33] and for 5hmC in smaller studies[13,14]. Our data suggest that 5hmC signal derived from epigenetic changes in non-tumor cells may be responsible for the sensitivity at early stage. This sensitivity is retained when 5hmC signal is combined with fragmentomics to produce an additive signal.

2 Results

2.1 Study Population

We isolated cfDNA from plasma. Blood samples were donated by 2483 individuals prior to undergoing colonoscopy. These double-spun plasma samples were purchased from multiple vendors and included biobanked and prospectively collected samples. Experimental batches for cfDNA extraction, hydroxymethylome library preparation and sequencing were balanced

for key sample characteristics (vendor, age, sex, ethnicity and diagnosis) (Fig. 1, Table 1, Table 2). The cfDNA from each individual was processed to generate two sequencing libraries: a whole genome library (denoted “input”) and a hydroxymethylome library.

Table 1
Demographics of whole cohort (n = 2483), training set (n = 781 control and CRC samples) and validation set (n = 384 control and CRC samples

Participating Individuals				
Study	**Total**	**Control***	**Other cancers****	**CRC**
N	2483	573	1113	797
Age, years, mean (SD)	64.19(9.11)	62.24(8.99)	63.97(8.79)	65.70(9.37)
Female gender, n (%)	1283 (51.67 %)	318 (55.50 %)	559 (50.22 %)	406 (50.94 %)
Ethnicity				
Asian	270 (10.87 %)	61 (10.65 %)	122 (10.96 %)	87 (10.92 %)
Black/African American	52 (2.09 %)	12 (2.09 %)	39 (3.50 %)	1 (0.13 %)
Pacific Islander	55 (2.22 %)	0 (0.00 %)	1 (0.09 %)	54 (6.78 %)
Other	274 (11.04 %)	104 (18.15 %)	62 (5.57 %)	108 (13.55 %)
Unknown	23 (0.93 %)	0 (0.00 %)	16 (1.44 %)	7 (0.88 %)
White	1809 (72.86 %)	396 (69.11 %)	873 (78.44 %)	540 (67.75 %)

Training Set						
	Control*	**CRC**	**CRC Stage 1**	**CRC Stage 2**	**CRC Stage 3**	**CRC Stage 4**
N	316	465	91	189	135	50
Age, years, mean (SD)	61.74 (8.89)	65.51 (9.13	66.20 (8.20)	65.78 (9.31)	65.19 (9.78)	64.06 (8.22)
Female gender, n (%)	174 (55.06 %)	244 (52.47 %)	50 (54.95 %)	98 (51.85 %)	70 (51.85 %)	26 (52.00 %)
Ethnicity						

* Control samples were from individuals with conditions including rheumatoid arthritis, COPD and peptic ulcer.

* Other cancer samples were from individuals with the following cancers: advanced adenoma, breast cancer, lung cancer, non-advanced adenoma, ovarian cancer, prostate cancer, stomach cancer, urinary cancer

Participating Individuals						
Asian	15 (4.75 %)	51 (10.97 %)	16 (17.58 %)	13 (6.88 %)	18 (13.33 %)	4 (8.00 %)
Black/African American	10 (3.16 %)	0 (0.00 %)	0 (0.00 %)	0 (0.00 %)	0 (0.00 %)	0 (0.00 %)
Pacific Islander	0 (0.00 %)	26 (5.59 %)	3 (3.30 %)	23 (12.17 %)	0 (0.00 %)	0 (0.00 %)
Other	71 (22.47 %)	73 (15.70 %)	2 (2.20 %)	28 (14.81 %)	30 (22.22 %)	13 (26.00 %)
Unknown	0 (0.00 %)	5 (1.08 %)	0 (0.00 %)	5 (2.65 %)	0 (0.00 %)	0 (0.00 %)
White	**220 (69.62 %)**	**310 (66.67 %)**	**70 (76.92 %)**	**120 (63.49 %)**	**87 (64.44 %)**	**33 (66.00 %)**
Validation Set						
N	164	220	45	87	61	27
Age, years, mean (SD)	65.88 (9.64)	62.46 (9.24)	64.31 (7.54)	66.41 (10.05)	65.49 (10.29)	67.63 (10.00)
Female gender, n (%)	95 (57.93 %)	116 (52.73 %)	21 (46.67 %)	49 (56.32 %)	35 (57.38 %)	11 (40.74 %)
Ethnicity						
Asian	10 (6.10 %)	24 (10.91 %)	5 (11.11 %)	8 (9.20 %)	6 (9.84 %)	5 (18.52 %)
Black/African American	1 (0.61 %)	1 (0.45 %)	0 (0.00 %)	1 (1.15 %)	0 (0.00 %)	0 (0.00 %)
Pacific Islander	0 (0.00 %)	15 (6.82 %)	2 (4.44 %)	13 (14.94 %)	0 (0.00 %)	0 (0.00 %)
Other	30 (18.29 %)	24 (10.91 %)	0 (0.00 %)	5 (5.75 %)	14 (22.95 %)	5 (18.52 %)
Unknown	0 (0.00 %)	1 (0.45 %)	0 (0.00 %)	1 (1.15 %)	0 (0.00 %)	0 (0.00 %)

* Control samples were from individuals with conditions including rheumatoid arthritis, COPD and peptic ulcer.

* Other cancer samples were from individuals with the following cancers: advanced adenoma, breast cancer, lung cancer, non-advanced adenoma, ovarian cancer, prostate cancer, stomach cancer, urinary cancer

Participating Individuals						
White	123 (75.00 %)	155 (70.45 %)	38 (84.44 %)	59 (67.82 %)	41 (67.21 %)	17 (62.96 %)

* Control samples were from individuals with conditions including rheumatoid arthritis, COPD and peptic ulcer.

* Other cancer samples were from individuals with the following cancers: advanced adenoma, breast cancer, lung cancer, non-advanced adenoma, ovarian cancer, prostate cancer, stomach cancer, urinary cancer

Table 2
Disease characteristics of whole cohort

	Total cohort		Validation set		Training set	
N	**2483**	**%**	**691**	**%**	**1415**	**%**
Colorectal cancer	797	32.10	220	31.84	465	32.86
CRC-1	161	6.48	45	6.51	91	6.43
CRC-2	319	12.85	87	12.59	189	13.36
CRC-3	222	8.94	61	8.83	135	9.54
CRC-4	95	3.83	27	3.91	50	3.53
Other cancers						
Advanced adenoma	306	12.32	82	11.87	169	11.94
Breast cancer	114	4.59	34	4.92	68	4.81
Lung cancer	140	5.64	51	7.38	80	5.65
Non-advanced adenoma	373	15.02	88	12.74	218	15.41
Ovarian cancer	32	1.29	10	1.45	15	1.06
Prostate cancer	91	3.66	25	3.62	50	3.53
Stomach cancer	32	1.29	8	1.16	21	1.48
Bladder cancer	25	1.01	9	1.30	13	0.92
Control						
Standard	531	21.39	149	21.56	291	20.57
Rheumatoid arthritis	20	0.81	7	1.01	11	0.78
Peptic ulcer	12	0.48	6	0.87	6	0.42
COPD	10	0.40	2	0.29	8	0.57

Libraries were successfully sequenced for 2106 individuals (Fig. 1) (685 with CRC, 480 controls and 941 with other conditions such as adenoma and other cancers). Of the 2106 individuals, mean age was 64 years and 52% were female. Samples that did not produce successful libraries were excluded; in most cases this was due to poor cfDNA yield (Fig. 1), the individuals who donated these samples had similar characteristics to the 2106 individuals who comprised the study population (Table 1, Table 2).

2.2 5hmC is efficiently captured from cfDNA to produce hydroxymethylome libraries

We developed a technology for hydroxymethylome capture of cfDNA fragments containing 5hmC residues, using just 5 ng of cfDNA. The high-throughput methodology was automated on 96-well plates using liquid handlers. Briefly, cfDNA extracted from 2 ml double spun plasma and quantified. Illumina compatible sequencing libraries were prepared using 5ng input cfDNA, a portion of this 'input' library was reserved for sequencing. The remaining sequencing library was denatured, and the single-stranded library was copied to create a double-stranded library where only one strand retained epigenetic information. 5hmC residues were enzymatically labelled with a modified glucose group, which was then biotinylated. 5hmC-containing double-stranded DNA fragments were captured using streptavidin beads. The copied strand without epigenetic modifications was recovered from the 5hmC-captured libraries and amplified to form the hydroxymethylome-enriched sequencing library (Fig. 2A).
Input and hydroxymethylome libraries were paired-end sequenced with an average of 62M reads per library. The input libraries covered on average 85% of the human genome with at least one read and ~ 17% with a read depth of more than five reads. In contrast the

hydroxymethylome libraries were more localized to distinct genomic regions and tended to form peaks that could be characterized at both broad and narrow resolution and covered only 31% of the genome with at least one read and ~ 5% at more than five reads. This is consistent with previous reports describing the genomic distribution of 5hmC[13,15,34].

5hmC was enriched in genic regions with an average of 1.8 as many reads falling in the genic regions compared to intergenic regions. In contrast, input libraries had an average ratio of 0.8 between genic and intergenic regions. This is consistent with previous reports that cfDNA is preferentially hydroxymethylated in genic regions[14,15]. These metrics varied little between plates and processing batches (median absolute deviation was 0.02 for intra-plate technical replicates and 0.04 for inter-plate technical replicates).

Control DNAs were included in all samples to report the efficacy and quantitative nature of hydroxymethylome capture. The positive controls, containing 1, 3 or 6 5hmC residues, were enriched in the hydroxymethylome versus input libraries with 88, 267 and 658-fold more reads, respectively. The negative controls, containing 6 mC residues or unmodified cytosines, were not enriched (1 and 0.38-fold respectively) (Fig. 2B).

2.3 5hmC quantification and distribution

Machine learning feature sets were generated from the hydroxymethylome capture data by producing normalized ratios of read counts in the hydroxymethylome versus input sequencing libraries within coordinates of genes and enhancers.

An unsupervised t-distributed stochastic neighbour embedding (t-SNE) analysis showed some separation of CRC and control sample across projected dimension 1, but with

substantial overlap between the classes (Fig. 3). No separation was observed for covariates such as gender and age (Fig. 3).

2.4 5hmC classifier detects CRC, even at early stage

5hmC levels in enhancer regions were used to train a supervised classifier algorithm with an ensemble of 50 learners to distinguish CRC from controls (405 CRC samples and 296 controls, Fig. 1). The area under the receiver operator curve (AUC) achieved in cross- validation within the training set was 90%, with 63% sensitivity at 95% specificity.

The 5hmC classifier was then applied to previously unseen samples from the validation set (220 CRC and 164 controls). Overall, AUC in the validation set was 90% for CRC. AUC was highest at stage 4 (94%) and declined only slightly to an AUC of 89% at stage 1 (Fig. 4A).

The 5hmC classifier achieved a specificity of 84%, with sensitivity of 77% for CRC (78%, 83%, 75% and 93% for stages 1, 2, 3 and 4, respectively). The performance at stage 1 is comparable with other recently reported classifiers for colorectal cancer (Fig. 7C). In the validation samples, we detected CRC (all stages combined) with 55% sensitivity when specificity was fixed at 95%.

The 5hmC classifier reported in Fig. 4 was trained on a sample set that balanced age, ethnicity, sex and processing batch samples evenly across case and controls (Table 1). We trained another model that accounted for an extended set of clinical covariates (age, ethnicity, sex, diabetes, and use of statins, alcohol, tobacco and NSAIDs) via propensity score weighting of samples and found there was no substantial difference in classifier performance (Table 3).

Table 3
(A) Validation performance of 5hmC classifiers for other cancers and conditions

Condition	AUC (%)	Sensitivity at 90%	Sensitivity at 95%
CRC vs controls	90 (87–93)	71 (60–81)	55 (36–71)
CRC vs controls (weighted for additional clinical covariates)	90 (87–93)	72 (59–81)	59 (38–72)
CRC vs controls and adenomas	83 (80–86)	61 (46–73)	31 (23–44)
Lung cancer vs controls	80 (73–87)	37 (21–73)	18 (11–38)
Breast cancer vs controls	82 (75–89)	33 (22–56)	16 (11–29)
Prostate cancer vs controls	79 (71–88)	30 (20–54)	15 (10–28)
CRC vs other cancers	78 (74–82)	34 (26–47)	17 (13–24)

Table 3
(B) Comparison of the validation performance of classifiers trained with different feature types

Feature type	AUC (%)	Sensitivity (%)	Sensitivity (%)	Sensitivity at 95%
CRC vs controls	90 (87–93)	81 (71–88)	84 (76–91)	55 (36–71)
CRC vs controls (weighted for additional clinical covariates)	85 (81–89)	75 (61–82)	82 (74–89)	45 (29–59)
CRC vs controls and adenomas	81 (79–87)	80 (73–86)	65 (52–77)	61 (42–71)

When adenoma samples (both individuals with non-advanced and advanced disease) were added to the control samples for 5hmC classifier training, AUC in the validation set was 83%, a poorer performance than the classifier trained on CRC and control samples only

(Table 3). This suggests that adenomas represent a source of false positives to the classifier. We also trained classifiers that successfully distinguished lung (AUC 80%), breast (AUC 82%), and prostate (AUC 79%) cancers from controls, and distinguished CRC from other cancers (AUC 78%) (Table 3).

2.4.1 5hmC classifier is robust under low ctDNA fraction

To further investigate the encouraging finding that performance of the 5hmC classifier was similar across cancer stages, we estimated the amount of ctDNA in cfDNA samples in the validation set using the ichorCNA tumor fraction statistic. ichorCNA is reported to have a limit of detection of ~ 3% at mean sequencing depth of 0.1x[35].

At this threshold we detected tumor fractions of 3–4% in the cfDNA of ~ 6% of control samples. We interpreted this as a false positive rate of the ichorCNA method, although it is possible that a small number of “control” individuals could have undiagnosed non-CRC tumors, given the age range. Using the 3% threshold as a limit of detection, the 5hmC classifier correctly called 97% of CRC samples with detectable tumor fraction and 75% of CRC samples with tumor fraction below the 3% detection threshold (Fig. 4C). This demonstrates that the 5hmC classifier is robust under low tumor fraction.

Lowering the limit of detection to the median tumor fraction observed in the controls (implying that 50% of controls have detectable tumor fraction) still resulted in 76% of samples being classified correctly in the undetectable class (with 83% being classified correctly in the detectable class). In contrast to the higher limit of detection at this level, CRC classification and tumor fraction class (detectable/undetectable) are statistically independent (Fisher Exact test, p = 0.285). As expected, mean

ctDNA fraction was greater in the CRC samples versus controls (0.046 versus 0.018, Mann-Whitney U test p = 0.0026) and ctDNA fractions were correlated with reported cancer stage (Spearman's rho = 0.25, p = 0.0002, Fig.

4B). Not only did the 5hmC classifier correctly call samples with higher tumor fraction (Supplementary Fig. 1), but it was also highly robust for samples where tumor fraction was below the detection threshold of 3% (Fig. 4C). Therefore, the 5hmC classifier is not solely dependent on the ctDNA fraction.

2.5 Interleukin signaling may drive 5hmC-based classification

To investigate the biological signals driving the 5hmC classifier's performance in CRC samples with low ctDNA fraction, the classifier's feature enhancers were mapped to gene names and queried in the Key Pathway Advisor software (Clarivate). IL11 signaling to the PIK3CA cascade was the top ranked pathway. This indicates that cfDNA fragments from the immune system may be driving detection (Fig. 4D, Supplementary Fig. 2). We further assessed features identified from training a classifier using only early-stage CRC samples (stages 1 and 2) and compared these with the features from a classifier trained only on late- stage CRC samples (stages 3 and 4). We found that the interleukin signal is present in late- stage CRC. Evidence pointed to an association with microRNA in early-stage CRC (Supplementary Fig. 3).

2.6 Performance of a region-based fragmentomics classifier has greater dependence on cancer stage and ctDNA fraction than a 5hmC classifier

To investigate the apparent lack of stage dependence observed for the 5hmC classifier, we interrogated the cfDNA fragment characteristics from the same samples (the validation set). These fragmentomic characteristics were observed from the input libraries generated as a control to the hydroxymethylome capture. Therefore, the data was readily available from exactly the same sample set but without information about 5hmC levels. Fragmentomics has previously been reported as being stage dependent[36].Using a technique similar to the DELFI method[36] (Supplementary Methods), *in silico* analysis of read depth and estimated DNA

fragment size was performed on sequencing reads from the input libraries, comparing the number of long to short fragments in 5Mb windows.

A classifier was produced using the same machine learning methodology to that used for the 5hmC classifiers. This fragmentomics classifier distinguished CRC samples from controls with an AUC of 83% and 62% sensitivity at 95% specificity in the validation set.

Performance of the fragmentomics classifier decreased from 91% for stage 4 CRC samples to 80% for stage 1 samples (Fig. 5A). This represents a higher loss of performance to detect early-stage CRC than the 5hmC classifier, which retained performance at early-stage (Fig. 7A and 7B). Correlation between AUC and CRC stage was higher for the fragmentomics classifier (Spearman's rho = 0.95, $p = 0.05$) than the 5hmC classifier (Spearman's rho = 0.80, p = 0.333). In addition, the fragmentomics classifier score was more highly correlated with ctDNA content in late-stage tumors (stage 4 Spearman's rho = 0.77, p =

$4.7x10^{-6}$) compared to the 5hmC classifier (Spearman's rho = 0.66, p = 0.00028), potentially explaining the performance gain in these late tumors (Fig. 5B).

We also trained a classifier with features based on the positioning of nucleosomes (the Nucleosome Presence (NPS) method; see Supplementary Methods.

We demonstrated that the enhancer-based 5hmC classifier (median AUC 90.3%) outperforms both the fragmentomics (median AUC 83.1%) and NPS classifiers (median AUC 85.2%) across all cancer stages on sensitivity and specificity (Fig. 6A-C, Table 3B).

In summary, we produced a fragmentomics based classifier, used the same machine learning procedure used to train the 5hmC classifier. In contrast to the 5hmC classifier, the fragmentomics classifier was dependent on CRC stage and ctDNA fraction. We hypothesize

that genome-wide 5hmC profiles can capture additional signal from the host that aids early- stage detection.

2.7 5hmC is additive to orthogonal sample characteristics

In an effort to capture all information yielded during sample processing, we trained a classifier including genome-wide 5hmC data, genome-wide region-based fragmentomics data (as above), and further sample characteristics such as library yield, genome-wide fragment size distribution and copy number related quantities. The performance of the resultant classifier to detect CRC in the validation set increased to an AUC of 91%, and 63% sensitivity at 95% specificity (Fig. 6B-D). This gain in performance was evident in stages 2–4 of CRC.

3 Discussion

Patients diagnosed with early-stage CRC have markedly better survival than patients diagnosed with late-stage CRC[37]. Here, we report that epigenetic profiling of cfDNA demonstrated the biomarker 5hmC to be a powerful biomarker for early CRC in liquid biopsy.

We used genome-wide 5hmC profiles to create and train a classifier. When this classifier was applied to unseen 5-hydroxymethylome data, CRC was detected with an AUC of 90%, with 71% sensitivity at 90% specificity, and 55% sensitivity at 95% specificity. Further we show that 5hmC successfully detected early-stage CRC (Fig. 7A). Importantly, this classifier detected stage 1 CRC with an AUC of 89% and a sensitivity of 56% at 95% specificity. Our operational performance (classifying each sample in the validation set without fixing specificity) on the validation set for stage 1 CRC (78% sensitivity at ~ 85% specificity) is statistically equivalent to several other reports using both non-5hmC and 5hmC-based classifiers for mixed stages of CRC including later stage [8, 12–15,29] (Fig. 7C). For example, in a study from Wan et al., the cfDNA whole genome classifier had a mean AUC of 92% (95% CI 0.91–0.93)[12] as cross validation performance in a training set (n = 817 samples). In further work from the same group, Putcha et al. reported validation (not yet peer reviewed) of the classifier in a small validation set (n = 17) with a mix of

stage 1 and 2 samples [33]. We include both the training and validation reports in Fig. 7C as it demonstrates how a classifier that performs very similarly to the one described here (with a similar training regime designed to emphasis bias reduction) may be tuned for higher specificity with a resulting trade off with sensitivity; for which there is substantial uncertainty in reported accounts due to the limited sizes of CRC stage-1 validation sets. Given further work, we believe that 5hmC based classifiers as described in this work could perform at least as equivalently at high specificity for early-stage CRC as those previously reported.

In another example, Liu and colleagues report a classifier based on methylation in cfDNA. In the validation set (n = 610), this classifier achieved specificity of 99.3% (95% CI 98.3–99.8%) but sensitivity was just 54.9% across a range of cancer types and stages[8]. The cfDNA methylation-based classifier reported by Kim et al. correctly classified 94% of CRC samples (stage 1 to 3) with 94% specificity in a validation set (n = 72 CRC and 35 controls)[32]. Guler et al. used a 5hmC-based cfDNA classifier to detect pancreatic cancer, achieving an AUC of 92–94% in two small independent validation sets comprising 228 and 17 subjects, respectively[15]. In further work from the same group, Li et al. used a cfDNA 5hmC classifier to detect CRC with an AUC of 94% (88% sensitivity, 89% specificity) in a small validation population (69 subjects)[14] The AUC of 90% that we achieved with our 5hmC classifier is comparable to these previous studies, and was estimated using a robust large validation set of 220 CRC samples and 164 controls.

Our method uses just 5 ng input cfDNA to generate 5hmC-enriched profiles across the entire genome, in contrast to other studies that are restricted to just small regions of the genome such as the Liu *et al* study[8], potentially missing signal with discriminatory power to detect cancer signatures. This genome-wide approach revealed substantial signal from non-tumor sources of cfDNA.

Performance of the 5hmC classifier did not appear to rely on the estimated ctDNA fraction in blood and likely used signal from cfDNA sourced from non-cancerous cells that undergo state change in response to tumor genesis and progression. This is supported by the finding that the classifier used 5hmC enhancer regions significantly enriched in inflammatory immune responses such as interleukin signaling. Indeed ~ 55% of cfDNA derives from white blood cells, according to whole genome analysis.[38] Despite this inflammatory component, we have separately shown the ability of classifiers to discriminate between different types of cancer, suggesting cancer type specificity in the classifier (Table 1).

Pathway analysis demonstrated the presence of immune system related pathways, particularly in features present in classifiers trained to discriminate between CRC stage 3 and stage 4 cancer and control samples (“late-stage features”). A classifier trained to discriminate between CRC stage 1 and 2, and control samples (“early-stage features”) did not use inflammatory pathways features, but significant gene hubs relating to microRNA expression common to both late stage and early-stage features were identified (Supplementary Fig. 3).

This implies that mechanisms related to microRNA expression may partly power the early- stage performance of the 5hmC CRC classifier.

Limitations and future work

Although we and others have reported results at higher specificities, we interpret performance at these levels cautiously, due to the comparatively small population of cancer-negative individuals that is unlikely to fully account for sample heterogeneity, clinical and demographic biases present in an asymptomatic CRC screening population.

Limits of the study included the possibility of selection bias in the study population since blood samples were collected from individuals presenting for colonoscopy. None of the clinical covariates we tested impacted classifier performance (Table 1), but not all possible risk factors were collected so we can not exclude the effects of some covariates on classifier performance.

Conclusions

This genome-wide approach to cfDNA profiling was enabled by our novel hydroxymethylome capture platform, which profiles the hydroxymethylome across the entire genome. This allowed selection of the most relevant 5hmC-enriched regions on which to base our classifier. The 5hmC classifier successfully detected CRC samples, regardless of cancer stage, with performance comparable to previously described classifiers for cancer detection based on cfDNA patterns of methylation and somatic mutation. We demonstrate that the performance of the 5hmC classifier improved when orthogonal fragmentomics information was added. Since epigenetic profiles are additive to genetic profiles[32], it may be optimal to consider multi-analyte approaches to cancer detection, combining epigenetic information with fragmentomics and genetic profiling of mutations.

In conclusion, this is the first study that has demonstrated the power of 5hmC to detect early- stage colorectal cancer via blood cfDNA in a heterogenous, well-balanced, well powered cohort, employing both internal validation and cross-validation data sets to verify performance. We conclude that 5hmC is a powerful and interpretable biomarker that can be used to power and enhance non-invasive diagnostic tools for the detection of early-stage and treatable cancer.

4 Methods

4.1 Study design

This multicenter, case-control study aimed to create and assess a 5hmC-based classifier for detecting CRC, even at early stage. Blood samples were obtained from 2483 individuals from 12 commercial suppliers covering 56 individual sites, sourced from both biobanks (~ 40%) and prospective collection (~ 60%). Control samples were from people aged 45–85 years who were at average risk for CRC and had been assessed by colonoscopy with results that showed no presence of CRC or adenomatous polyps (adenomas). Cancer samples were from people aged 45–85 years who underwent colonoscopy and were

diagnosed with CRC, lung, breast, bladder, prostate, ovarian, stomach cancer or adenomas. Advanced adenoma was defined as high-grade dysplasia or with ≥ 25% villous histologic features OR measuring ≥ 1 cm in the greatest dimension, in agreement with Imperiale et al.[39]

We recorded the following data for each blood sample donor: age, sex, ethnicity, smoking, diabetes status, previous medical history, and concomitant medications (including daily NSAID use) and type of diagnosis.

Extraction of cfDNA was attempted from double spun plasma donated by 2483 individuals (Fig. 1). cfDNA from 2198 individuals was passed forward to sequencing. 5hmC was quantified across the cfDNA genome for all individuals via comparison of region read depth in sequencing libraries enriched by highly sensitive capture of 5hmC, to a non-enriched library. The hydroxymethylome of 701 participants was interrogated by machine learning algorithms to produce classifiers distinguishing CRC from a range of other conditions. The classifiers were then tested on previously unseen 5-hydroxymethylome data from 691 other individuals (the validation set).

The sample size of the validation set was computed based on demonstrating improvement over the multitarget stool DNA test (sensitivity 92.3% and specificity 86.6% [95% CI: 85.9–87.2%]) and fecal immunochemical test (FIT) (sensitivity 73.8% and specificity 94.9% [95% CI: 94.4–95.3%])[39]. Achieving a desired 1-sided sensitivity for CRC, we initially chose a lower bound of 83.0% for the confidence interval and a point estimate of 92.3% for this study. Without accounting for gender, age and stages of disease differences, 95% power,

0.025 alpha and 5% dropout rate, a sample size of at least 330 CRC confirmed cases was deemed required for sensitivity characterization. To achieve approximately the same power, a similar sample size of at least 340 would be required to demonstrate the desired specificity. However, we chose to trade off model training error with validation error, reducing the validation sample size to 220 CRC and 164 controls. Consequently, the ability of 5hmC to detect CRC may be underestimated by this study.

The study was performed in accordance with the Declaration of Helsinki and was approved by the relevant independent ethics committee or institutional review board for each commercial supplier of blood samples. Written informed consent was obtained for all donors of samples.

4.2 Method details

4.2.1 Blood sample collection

Where possible, sampling was performed before colonoscopy. Blood (10 ml) was drawn into a K2 EDTA blood tube, placed on ice and processed within 4 hours. Samples were centrifuged (2000 g for 10 minutes at room temperature). The plasma layer was transferred to a clean tube and was again centrifuged (2000 g for 10 minutes at room temperature), to remove any remaining cellular material. Double-spun plasma was aliquoted into tubes in at least 1 or 2 ml volumes, then immediately frozen and stored at -80°C before shipment to the central laboratory for investigation.

4.2.2 Sample balancing

To avoid confounding the biological signal, the OSAT algorithm[40] was utilized to achieve an even distribution of disease state and potential confounders across experimental plates for both cfDNA extraction and hydroxymethylome capture. The associations of these factors with batch were tested using a Chi-square test and the design was modified where necessary. The distribution of disease, sex, ethnicity, and age group did not show statistically significant variation ($p < 0.05$) over the 96-well plates that were subject to the automated library processing. This ensured that any plate related processing biases were distributed evenly across sample characteristics.

4.2.3 cfDNA extraction and library creation

cfDNA was extracted using the NextPrep-Mag™ kit on the Chemagic Prime platform (Perkin Elmer chemagen Technologie GmbH, Baesweiler, Germany) using 2 ml of plasma, in 48- well plates. Two plates were extracted simultaneously and combined in a single 96-well plate at the end of the extraction process. cfDNA concentration was pre-quantified by PicoGreen (Life Technologies) assay on a CLARIOstar plate reader. cfDNA that reached a threshold concentration by PicoGreen was further quantified and assessed for cfDNA purity by gel electrophoresis (Fragment Analyser, Agilent, Santa Clara, CA, USA). cfDNA samples with yield ≥5 ng were normalized and 5 ng was plated using the Chemagic Prime instrument into 96-well plates ready for library preparation. Five 166bp synthetic controls were included in every sample of the experiment to control the quality of hydroxymethylome capture. The positive controls contain 1, 3 or 6 of 5hmC residues, and negative controls contain 6 of 5mC and unmodified Cs, respectively.

cfDNA samples were end repaired, adenylated (Kapa Hyper Prep kit, Roche Sequencing and Life Science), ligated to unique dual index (UDI) adaptors (Illumina TruSeq DNA Unique

Dual (UD) Indexes, Illumina, San Diego, CA, USA) and purified using SpeedBeads™ magnetic carboxylate modified particles.

Part of each sample (1 µl) was used to create an “input library” by directly PCR amplifying (9 cycles) ligated cfDNA. The remaining 12 µl of each sample was used for hydroxymethylome capture.

4.2.4 Hydroxymethylome capture

After the adapter ligation, the cfDNA strands were denatured and copied using a primer complementary to the sequence in the Illumina adapter using DNA polymerase I Klenow fragment (3'→5'exo) (Enzymatics, QIAGEN). Consequently, all the DNA fragments in the library comprised duplexes where one strand represented the original native genomic DNA, complete with epigenetic marks, and the other strand was an unmarked complimentary copy. 5hmC residues in the original genomic strand were selectively labelled with an azide- modified UDP-Glucose by incubation with UDP-6-N3-Glu (Jena Bioscience, Jena, Germany) and T4-beta-glucosyltransferase (Thermo Fisher Scientific, MA, USA). In turn, the azide groups were biotinylated with DBCO-PEG4-Biotin (Click Chemistry Tools, AZ, USA).

Samples were then purified using the DNA Clean & Concentrator kit (Zymo Research, Irvine, CA, USA). The 5hmC biotin conjugates were selectively bound to streptavidin beads (Dynabeads M-270, Invitrogen, Carlsbad, CA, USA). Finally, the single strand copies of the hydroxymethylome library were liberated from the beads by 0.1M NaOH and were PCR amplified (16 cycles). (Fig. 2A). The input and hydroxymethylome libraries were purified using SpeedBeads™ magnetic carboxylate modified particles (Sigma-Aldrich).

Concentration was determined using PicoGreen, library size and concentration were also determined using Fragment Analyser data.

4.2.5 Sequencing

We prepared 3 nM of non-hydroxymethylome enriched libraries (“input”) and hydroxymethylome library pools, respectively. Libraries were sequenced on the

NovaSeq platform using 100bp paired-end mode, yielding approximately 60 million reads per sample.

4.2.6 Bioinformatic data processing and quality control

Demultiplexing and trimming was achieved using bcl2fastq Read (Illumina Basespace). Reads were aligned to the human genome (GRCh38) using BWA-MEM[41], those with a BWA mapping quality score (MAPQ score) less than 1 were filtered. Sequence duplicates were removed using Picard MarkDuplicates. Libraries were scored for quality on 30 criteria. A cumulative quality score of 0 indicates perfect library quality and the absence of quality issues. Libraries scoring over 15 were discarded (along with their mate pair). Libraries scored 5 points if they had < 10M reads post deduplication, or a ratio of reads per kilobase of genebody, per million mapped reads (RPKM) across all for gene bodies divided by the RPKM in intergenic regions of < 1 (for the hydroxymethylome libraries only), or a lack of spike-in control amplification (< 1 for ratio of hydroxymethylation over methylation and cytosines), or a mitochondrial RPKM > 1000, or < 10% of reads mapping to peaks, or a median insert size > 200nt, or uniformity < 0.8 (for input libraries only). Libraries scored 3 points if they had > 1.5x the interquartile range for 26 quality metrics. Hydroxymethylome libraries scored 1 point if they deviated by > 2 standard deviations from ranges of gene body versus intergenic enrichment, duplication rate and coverage of the previously observed in other in- house studies, and to input libraries which deviated by > 2 standard deviations in sequence diversity score to the observed ranges from previous studies.

4.2.7 Feature Definition

To calculate 5hmC levels at gene enhancers, we first calculated read counts using Bam readcounts v0.01. RPKM were calculated over candidate gene-enhancers (adapted from[42]) downloaded from GeneCards v4.4. 5hmC enrichment was computed as the log2 ratio between the hydroxymethylome library RPKM and the input library RPKM after the inclusion of pseudocounts.

We produced a set of cfDNA fragment features inspired by the DELFI approach adapting the computational methodology, available via GitHub[36]. Briefly, we divided the genome into 100KB bins and quantified cfDNA fragment sizes per bin. We removed blacklisted regions, genomic gaps (UCSC table)

and non-standard chromosomes *a priori*. We excluded outlier bins in fragment size, only retaining fragments between 100nt to 220nt length. Finally, we split the genome into 100KB bins (in total 26170 non-overlapping genomic regions) and calculated the following characteristics of fragment size distribution per genomic bin: number of short fragments (100-150nt), number of long fragments (151-220nt), ratio between short and long fragments and the total number of fragments. This approach generates 26170 features per metric and per sample. The last step is the averaging of the 100 KB bins into larger non-overlapping genomic regions of 5 MB (in total 512 bins).

The final set of fragmentomics features, referred to as Nucleosome Presence Score (NPS) features, consists of metrics related to nucleosome presence and is capturing information at a highly localized scale. Our approach is inspired by the windowed protection score method of Snyder *et al*[16] but includes some key differences. Firstly, 40 samples (20 CRC and 20 HV) from the training cohort were reserved for developing the NPS approach. Subsequent models were never trained using these 40 samples. Based on 5hmC pulldown libraries in these 40 samples, a total of about 235,000 regions were identified by merging peaks produced by the MACS2 and EPIC2 peak callers. Using the bedtools coverage tool, average per-base coverages were computed for each region for each of the 40 samples. By sorting regions by the median coverage across all 40 samples, the 200 regions with the highest median coverage were chosen. In total, these regions covered about 4.5Mb of the genome. Next, the reads from the input libraries of the 40 samples were pooled, thus producing a single .bam file of depth 110.65X. This pooled sample was used for identifying nucleosome positions in the 200 regions defined above. Nucleosome calling was done by computing NPS profiles and using a simplistic peak calling approach, this approach assigning nucleosomes to NPS maxima in a 151bp sliding window. If multiple maxima existed within 76bp of each other, these were assumed to represent a single nucleosome position located at the midpoint between the maxima. NPS profiles were computed using fragment size data describing the start and end position of fragments. Fragment data were generated from deduplicated bam files with non- properly paired reads removed using Samtools (-f2 flag). Pairs of reads were collapsed into fragments using the bedtools bamtobed command and fragments of length more than 1000bp were removed, these were assumed to be errors. To compute the NPS profile in a given region, a sliding window approach was employed with a window size of 121bp and with NPS values defined for the midpoint of the window. With a view to capturing single nucleosome configurations, fragments less than 120bp or larger than 250bp were discarded. For each window, the NPS was defined as the ratio of the number of fragments spanning the window (n_span) to the number of fragments with at least one endpoint inside the window (n_within). As a result, the metric takes positive values and is independent of read depth. To limit cases of divergence, a pooling approach was applied wherein the NPS at position i was defined using information from +/- 5 neighboring positions:

$$NPS(i) = \frac{\sum_{i-5}^{i+5} n_{span}}{\sum_{i-5}^{i+5} n_{within}}$$

In events where, $\sum_{i-5}^{i+5} n_{within} = 0$ the NPS was set as NA and subsequently imputed. Imputation was achieved using a simple “fill in the gaps” strategy where missing values were assigned such as to linearly bridge the nearest non-NA values. Samples where more than 90% of the NPS profile of a given region was NA were categorized as “undefined” for this specific region. Such incidences were addressed by the feature-level imputation strategy. Finally, with a view to constructing a clear nucleosome signal, NPS profiles were smoothed using a Savitzky-Golay filter of degree 2 with a window size of 151bp.

Feature matrices were constructed for all samples using the nucleosomes identified from the 40 left out samples. Features were defined as the minimum NPS value in a +/-50bp neighborhood around the midpoint between two nucleosome positions, provided the nucleosomes were no more than 300bp apart. The midpoint between nucleosomes were found to be marginally more informative than the actual nucleosome positions. Samples where more than 90% of the NPS profile of a given region was NA were categorized as “undefined” for this specific region. Such incidences were addressed by the feature-level imputation strategy. No feature had more than 2% missingness.

4.2.8 Classifier development and internal cross-validation

We trained a Support Vector Machine (SVM) models using a linear kernel function on feature scaled (z-score normalization) 5hmC levels of enhancers quantile-normalized over samples (see Supplementary Fig. 4). An ensemble of 50 models were trained. Each model in the ensemble was trained on a randomly selected 80% of the samples in the training set and the trained models were used to predict 20% of the remaining samples in an internal cross- validation procedure within the training set. We identified features significantly correlated with technical covariates such as age, sex and vendor using linear regression and ANOVA F-

test for continuous and categorical variables, respectively. These features were then excluded. Model hyperparameters (the C parameter) were optimized for the highest AUC within a 10- fold cross validation strategy. Performance was averaged across the 50 individual learners, and the unique features selected by all 50 were retained (Supplementary Fig. 4). 95% confidence intervals of various performance metrics (e.g., AUC, sensitivity) for each of the classifier ensembles were computed using 2000 bootstrap replications of predicted samples. Classifier development and cross-validation within the training set was performed by the model development team, who logged trained and timestamped classifiers in a registry, along with auxiliary information on the training procedures for each model. A MD5 checksum was then computed for each classifier, functioning as a unique identifier.

4.2.9 Assessing classifier performance in validation set

The cross-validated classifier was then assessed in the “held-out” validation set of 691 samples, locked prior to model development process.

To facilitate a blinded model validation strategy a separate team performed the model validation. Version history on the classifier registry verified that hash keys had not been modified since initial logging. The validation team generated feature matrices and metadata files for the validation set and subsequently

applied the models on the validation data. The validation team operated on virtual machines and storage belonging to a separate cloud project which was inaccessible to the model development team. During the model validation process, the hash key from each applied model was compared to the logged hash key to ensure model integrity. Here again, prediction probabilities from every learner within the classifier ensemble were averaged to compute the final prediction probability for each sample. Final performance metrics such as the AUC, sensitivity and specificity were computed based on the averaged prediction probabilities. The performance results were automatically uploaded to a cloud database without any intervention from the model development team.

4.2.10 Assessing Tumor Fraction with ichorCNA

We ran the ichorCNA workflow on input libraries sequenced for the internal validation sample set.

This involved first running readCounter from the hmmcopy version 0.1.1 with the following command with the parameters window size set to 1000000 and quality set to 20. We then further ran ichorCNA, using the recommended settings for low tumor fraction samples as per below:

- -centromere GRCh38.GCA_000001405.2_centromere_acen.txt \
- estimateNormal True --estimatePloidy True --estimateScPrevalence False \
- scStates 'c()' --txnE 0.9999 --txnStrength 10000 --normal 'c(0.95, 0.99, 0.995, 0.999)' \
- ploidy 'c(2)' --maxCN 3 --normalPanel HD_ULP_PoN_hg38_1Mb_median_normAutosome_median.rds
- chrs 'c(1:22)' --chrTrain 'c(1:22)'
- gcWig $gc_hg38_1000kb.wig

4.2.11 Functional analysis of 5hmC enhancer regions driving classifier performance

To develop a better mechanistic understanding of the classifier we ran pathway analysis using the Key Pathway Analysis (KPA) tool on top discriminatory enhancers. Enhancers were ordered based on their average contribution (averaged across the models) to the classifier and from the top 500 enhancers that appeared in at least 5 individual models were selected for the pathway analysis. The top scoring gene target for each 500 enhancers was taken from the ‘connected_gene’ field in the GeneHancer database[42]and used as the input for pathway analysis.

Declarations

Competing interests statement

SC, MR, CML, CKL, KH, SM, NJW, LV, MA, MN, TO, HB, SY, JDH, DL, EC and MS are named inventors to patent applications filed by Cambridge Epigenetix Limited pertaining to one or more aspects of the technologies described herein.

Author contributions

Authors who made substantial contribution to the conception and design of the study were: AV, NJW, TC, GR, MS, TWBO, VW, JM, JM, LV, HB, MN, SY, DL, SS, WG, VP

Acknowledgements

Dr Fiona Dunlevy, Axcience, provided editorial assistance in the preparation of this article. Dr Isabel Calvo, Morphology provided assistance in the preparation of figures.

References

1. Siegel, R. L. *et al.* Colorectal cancer statistics, 2020. *CA Cancer J. Clin.* **70**, 145-164, doi:10.3322/caac.21601 (2020).
2. Kakushadze, Z., Raghubanshi, R. & Yu, W. Estimating Cost Savings from Early Cancer *Data* **2**, doi:https://doi.org/10.3390/data2030030 (2017).
3. Nakamura, Y. & Shitara, K. Development of circulating tumour DNA analysis for gastrointestinal *ESMO Open* **5**, doi:10.1136/esmoopen-2019-000600 (2020).
4. Heitzer, E., Haque, I. S., Roberts, C. E. S. & Speicher, M. R. Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nature Reviews Genetics* **20**, 71-88, doi:10.1038/s41576-018-0071-5 (2019).
5. Chen, X. *et al.* Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. *Nature Communications* **11**, 1-10, doi:10.1038/s41467- 020-17316-z (2020).
6. Wan, J. C. M. *et al.* Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Rev. Cancer* **17**, 223-238, doi:10.1038/nrc.2017.7 (2017).
7. Haque, I. & Elemento, O. Targeted ctDNA mutation-detection panels require infeasibly large input volumes for early detection. *bioRxiv*, doi:https://doi.org/10.1101/237578 (2017).
8. Liu, M. C. *et al.* Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Oncol.* **31**, 745-759, doi:10.1016/j.annonc.2020.02.011 (2020).
9. Hofmann, L. *et al.* A Multi-Analyte Approach for Improved Sensitivity of Liquid Biopsies in Prostate Cancer. *Cancers (Basel)* **12**, doi:10.3390/cancers12082247 (2020).
10. Cohen, J. D. *et al.* Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science (New York, N.Y.)* **359**, 926-930, doi:10.1126/science.aar3247 (2018).
11. Feinberg, A. P., Ohlsson, R. & Henikoff, S. The epigenetic progenitor origin of human *Nature Reviews Genetics* **7**, 21-33, doi:10.1038/nrg1748 (2006).
12. Wan, N. *et al.* Machine learning enables detection of early-stage colorectal cancer by whole-genome sequencing of plasma cell-free DNA. *BMC Cancer* **19**, 832, doi:10.1186/s12885-019-6003-8 (2019).
13. Song, C. X. *et al.* 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. *Cell Res.* **27**, 1231-1242, doi:10.1038/cr.2017.106 (2017).

14. Li, W. *et al.* 5-Hydroxymethylcytosine signatures in circulating cell-free DNA as diagnostic biomarkers for human cancers. *Cell Res.* **27**, 1243-1257, doi:10.1038/cr.2017.121 (2017).
15. Guler, G. D. *et al.* Detection of early stage pancreatic cancer using 5- hydroxymethylcytosine signatures in circulating cell free DNA. *Nat Commun* **11**, 5270, doi:10.1038/s41467-020-18965-w (2020).
16. Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M. & Shendure, J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell* **164**, 57-68, doi:10.1016/j.cell.2015.11.050 (2016).
17. Bird, A. Perceptions of epigenetics. *Nature* **447**, 396-398, doi:10.1038/nature05913 (2007).
18. Song, C.-X. *et al.* Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Biotechnol.* **29**, 68-72, doi:10.1038/nbt.1732 (2011).
19. Wilkins, O. M. *et al.* Genome-wide characterization of cytosine-specific 5- hydroxymethylation in normal breast tissue. *Epigenetics* **15**, 398-418, doi:10.1080/15592294.2019.1695332 (2020).
20. Yu, M. *et al.* Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian *Cell* **149**, 1368-1380, doi:10.1016/j.cell.2012.04.027 (2012).
21. Spruijt, C. G. *et al.* Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized *Cell* **152**, 1146-1159, doi:10.1016/j.cell.2013.02.004 (2013).
22. Uribe-Lewis, S. *et al.* 5-hydroxymethylcytosine marks promoters in colon that resist DNA hypermethylation in cancer. *Genome Biol.* **16**, 69, doi:10.1186/s13059-015- 0605-5 (2015).
23. Mellén, M., Ayata, P., Dewell, S., Kriaucionis, S. & Heintz, N. MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell* **151**, 1417-1430, doi:10.1016/j.cell.2012.11.022 (2012).
24. Hashimoto, H., Hong, S., Bhagwat, A. S., Zhang, X. & Cheng, X. Excision of 5- hydroxymethyluracil and 5-carboxylcytosine by the thymine DNA glycosylase domain: its structural basis and implications for active DNA demethylation. *Nucleic Acids* **40**, 10203-10214, doi:10.1093/nar/gks845 (2012).
25. Bachman, M. *et al.* 5-Hydroxymethylcytosine is a predominantly stable DNA *Nat. Chem.* **6**, 1049-1055, doi:10.1038/nchem.2064 (2014).
26. Li, X., Liu, Y., Salz, T., Hansen, K. D. & Feinberg, A. Whole-genome analysis of the methylome and hydroxymethylome in normal and malignant lung and liver. *Genome* **26**, 1730-1741, doi:10.1101/gr.211854.116 (2016).
27. Healthcare, I. & Health, O. Grail Multi-Cancer Test Meets Validation Goals; Patients to Receive Results Under New Pilot | 3-5 (2020).
28. Cai, J. *et al.* Genome-wide mapping of 5-hydroxymethylcytosines in circulating cell- free DNA as a non-invasive approach for early detection of hepatocellular carcinoma. *Gut*, gutjnl-2019-318882, doi:10.1136/gutjnl-2019-318882 (2019).
29. Gao, P. *et al.* 5-Hydroxymethylcytosine profiling from genomic and cell-free DNA for colorectal cancers *J. Cell. Mol. Med.* **0**, doi:10.1111/jcmm.14252 (2019).

30. Cao, *et al.* Integrated epigenetic biomarkers in circulating cell-free DNA as a robust classifier for pancreatic cancer. *Clin. Epigenetics* **12**, 112, doi:10.1186/s13148- 020-00898-2 (2020).
31. Dean, J. *et al.* Sa1651 PLASMA BASED CELL-FREE CIRCULATING TUMOR DNA (CTDNA) ASSESSMENT FOR NON-INVASIVE DETECTION OF COLORECTAL CANCER (CRC). *Gastroenterology* **158**, S-369, doi:https://doi.org/10.1016/S0016-5085(20)31616-4 (2020).
32. Kim, S.-T. *et al.* Abstract 916: Combined genomic and epigenomic assessment of cell-free circulating tumor DNA (ctDNA) improves assay sensitivity in early-stage colorectal cancer (CRC). 916-916, doi:10.1158/1538-7445.sabcs18-916 (2019).
33. Putcha, G. *et al.* Blood-based detection of early-stage colorectal cancer using multiomics and machine learning. *Clin. Oncol.* **38**, 66-66, doi:10.1200/JCO.2020.38.4_suppl.66 (2020).
34. Hohos, N. M. *et al.* DNA cytosine hydroxymethylation levels are distinct among non- overlapping classes of peripheral blood leukocytes. *Immunol. Methods* **436**, 1-15, doi:10.1016/j.jim.2016.05.003 (2016).
35. Adalsteinsson, V. A. *et al.* Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nature Communications* **8**, doi:10.1038/s41467-017-00965-y (2017).
36. Cristiano, *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*, doi:10.1038/s41586-019-1272-6 (2019).
37. Cancer Research *Bowel cancer survival statistics*<https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by- cancer-type/bowel-cancer/survival#heading-Three> (2020).
38. Kustanovich, A., Schwartz, R., Peretz, T. & Grinshpun, A. Life and death of circulating cell-free DNA. *Cancer Biol. Ther.* **20**, 1057-1067, doi:10.1080/15384047.2019.1598759 (2019).
39. Imperiale, F. *et al.* Multitarget stool DNA testing for colorectal-cancer screening. *Engl. J. Med.* **370**, 1287-1297, doi:10.1056/NEJMoa1311194 (2014).
40. Yan, *et al.* OSAT: A tool for sample-to-batch allocations in genomics experiments.*BMC Genomics* **13**, doi:10.1186/1471-2164-13-689 (2012).
41. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA- *arXiv e-prints*, arXiv:1303.3997 (2013).
42. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database : the journal of biological databases and curation* **2017**, 1-17, doi:10.1093/database/bax028 (2017).

Figures

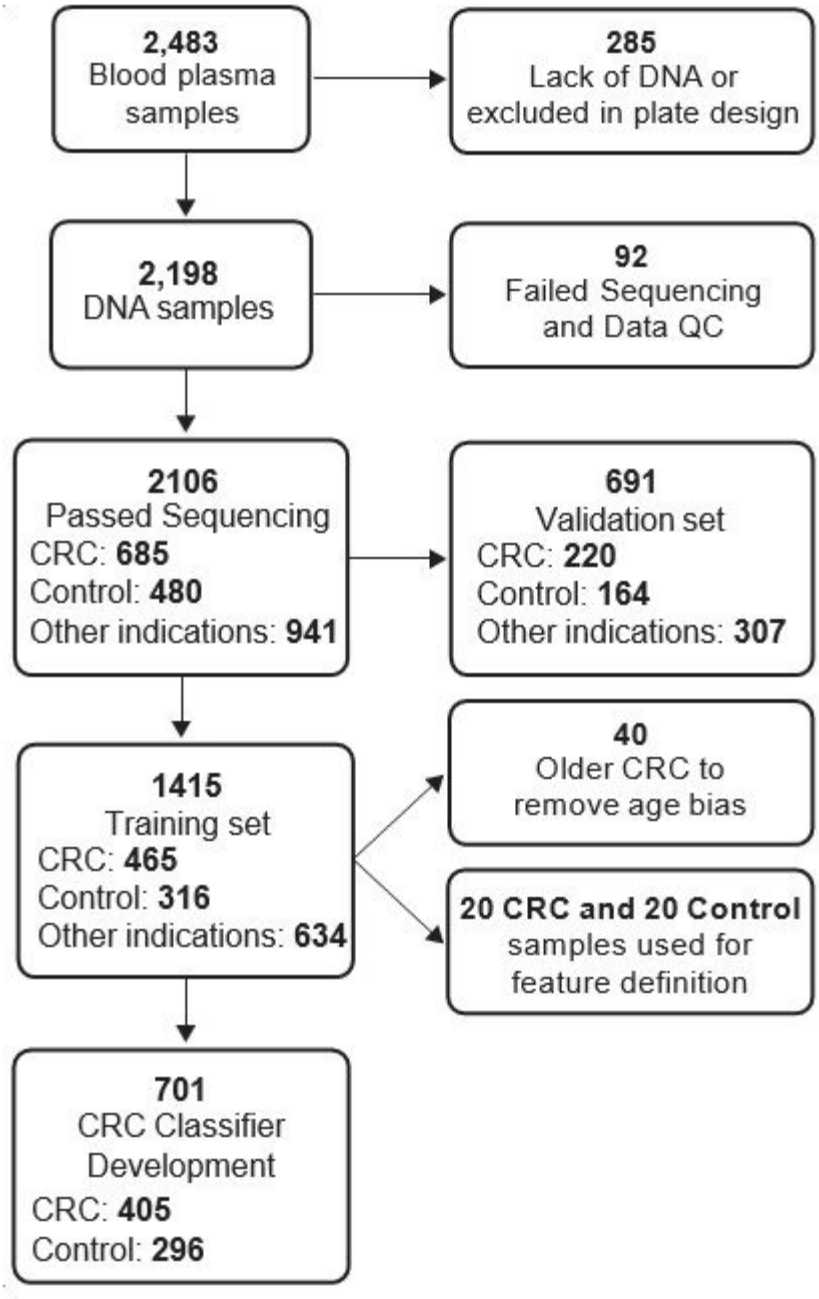


Figure 1

Flow chart of subjects included in the study. Control samples were made up of individuals who were CRC and adenomatous polyp negative (colonoscopy confirmed). 8.3% of the control individuals were diagnosed with peptic ulcers, arthritis or COPD.

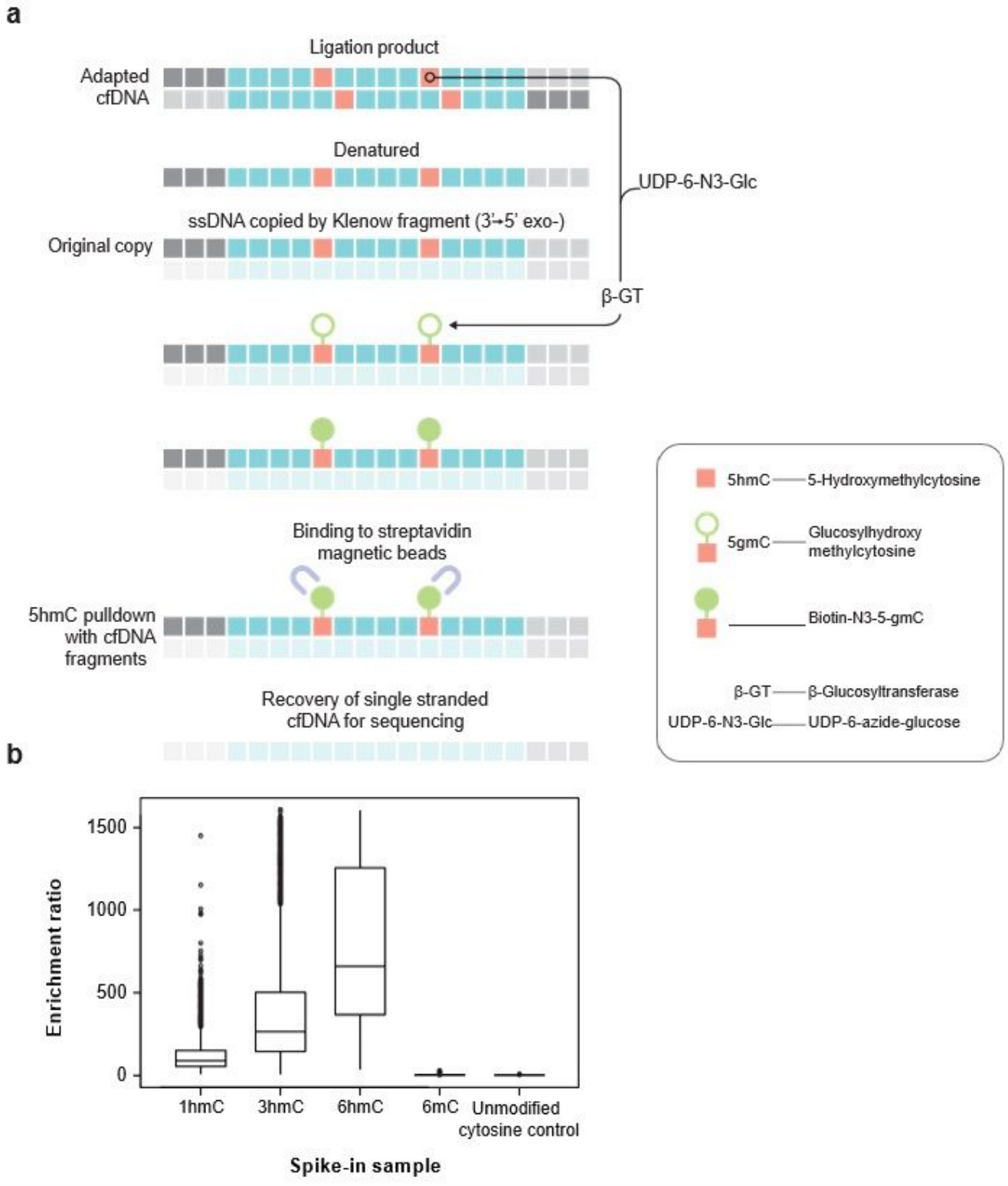


Figure 2

(A) Hydroxymethylome capture procedure (B) 166bp synthetic spike-in controls with 1,3,6 5hmC residues demonstrate that the hydroxymethylome enriches for 5hmC over controls containing 6 5mC residues and unmodified cytosines

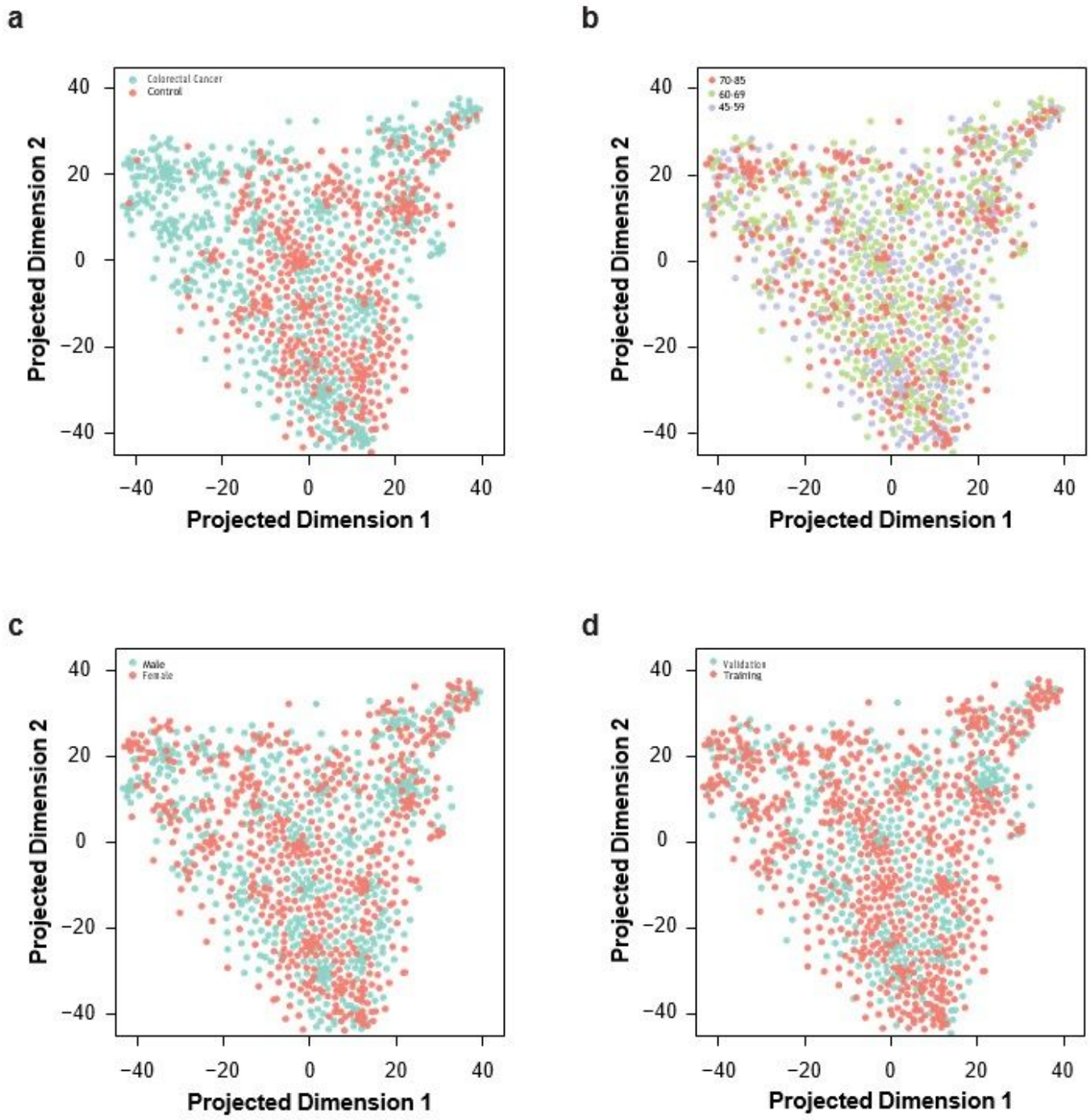
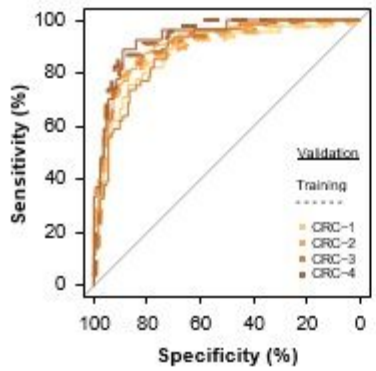


Figure 3

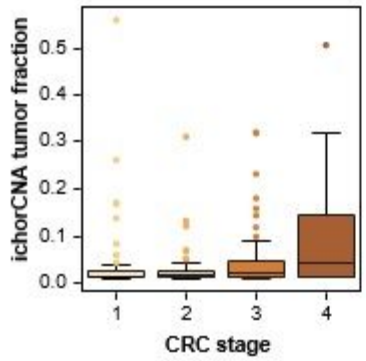
A two-dimensional representation of 5hmC quantified within gene enhancers over the training and validation samples displays evidence of clustering by disease status (CRC=red, control=blue), with little bias for gender (open for male, closed for female) or age (45-59,60-69, 70-85 years) (t-SNE parameters: perplexity = 20, theta = 0.5).

a



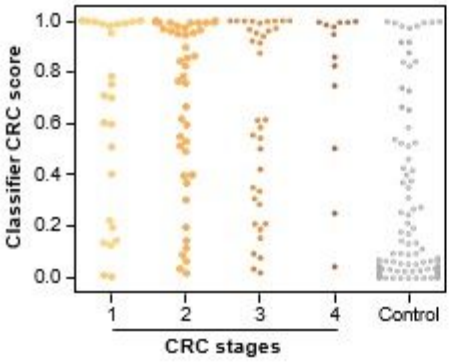
	Validation	
	AUC	Sens. at 95% specificity
CRC-1	89 (83-94)	55 (29-73)
CRC-2	91 (88-95)	60 (38-77)
CRC-3	89 (84-93)	44 (25-69)
CRC-4	94 (89-98)	66 (37-85)
Overall	90 (87-93)	55 (36-71)

b



	Correlation between CRC score and ichorCNA on Validation Samples	
Stage	Spearman's Rho	p-value
CRC-1	0.40	5.136 E-4
CRC-2	0.08	4.570 E-1
CRC-3	0.48	1.117 E-4
CRC-4	0.66	2.800 E-4
Overall	0.34	2.717 E-6

c



	% Samples above or below the classifier CRC score	
	<0.5	≥0.5
CRC-1	27.78	72.22
CRC-2	18.06	81.94
CRC-3	38.46	61.54
CRC-4	16.67	83.33
Control	84.21	15.79

d

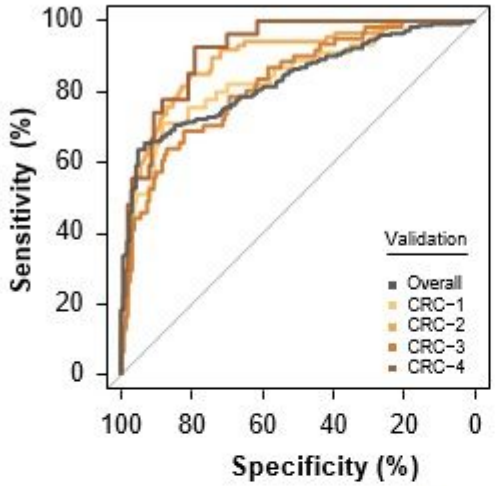
Name	Input Data p-value	Key Hubs p-value
Immune response IL-11 signaling pathway via MEK/ERK and PI3K/AKT cascades	5.490 E-5	5.322 E-18
Chemotaxis Lysophosphatidic acid signaling via GPCRs	6.750 E-5	3.375 E-17
Development PTHR1 in bone and cartilage development	3.250 E-5	2.338 E-16
Signal transduction Angiotensin II/ AGTR1 signaling via p38, ERC and PI3K	3.556 E-4	3.943 E-15
Oxidative stress ROSS-induced cellular signaling	31.113 E-4	1.327 E-15

Figure 4

(A) Classifier trained on 5hmC levels in enhancer regions show equivalent performance on the training (dotted line) and validations sets (solid line), and high performance across all stages, with AUC ranging from 88.6% to 93.6%. (B) IchorCNA tumour fraction is positively correlated with tumour stage in validation samples. Correlation with the CRC classifier score is lower in early stage (stage 1 and 2) with higher p-values than later stage samples. (C) CRC classifier score on validation samples with ichorCNA values ≤3% tumour fraction, demonstrating that the 5hmC based classifier maintains robust performance on

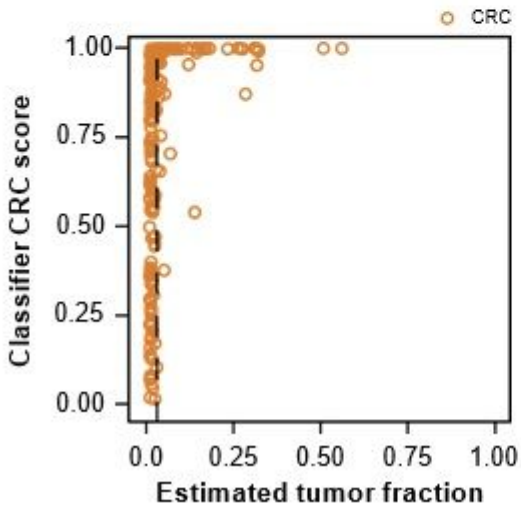
samples with low tumour fraction. The corresponding table presents the percentage of samples either side of the classification threshold (0.5) demonstrating that the classifier performs similarly across CRC stage (D) Top significant biological pathways identified that relate pathway relationship of genes regulated by enhancer features in the 5hmC classifier indicate a global immune response to tumorigenesis

a



	Validation	
	AUC	Sens. at 95% specificity
CRC-1	80 (73-88)	58 (38-73)
CRC-2	80 (74-86)	51 (30-64)
CRC-3	86 (79-93)	74 (49-85)
CRC-4	91 (85-97)	70 (48-89)
Overall	83 (79-87)	62 (44-70)

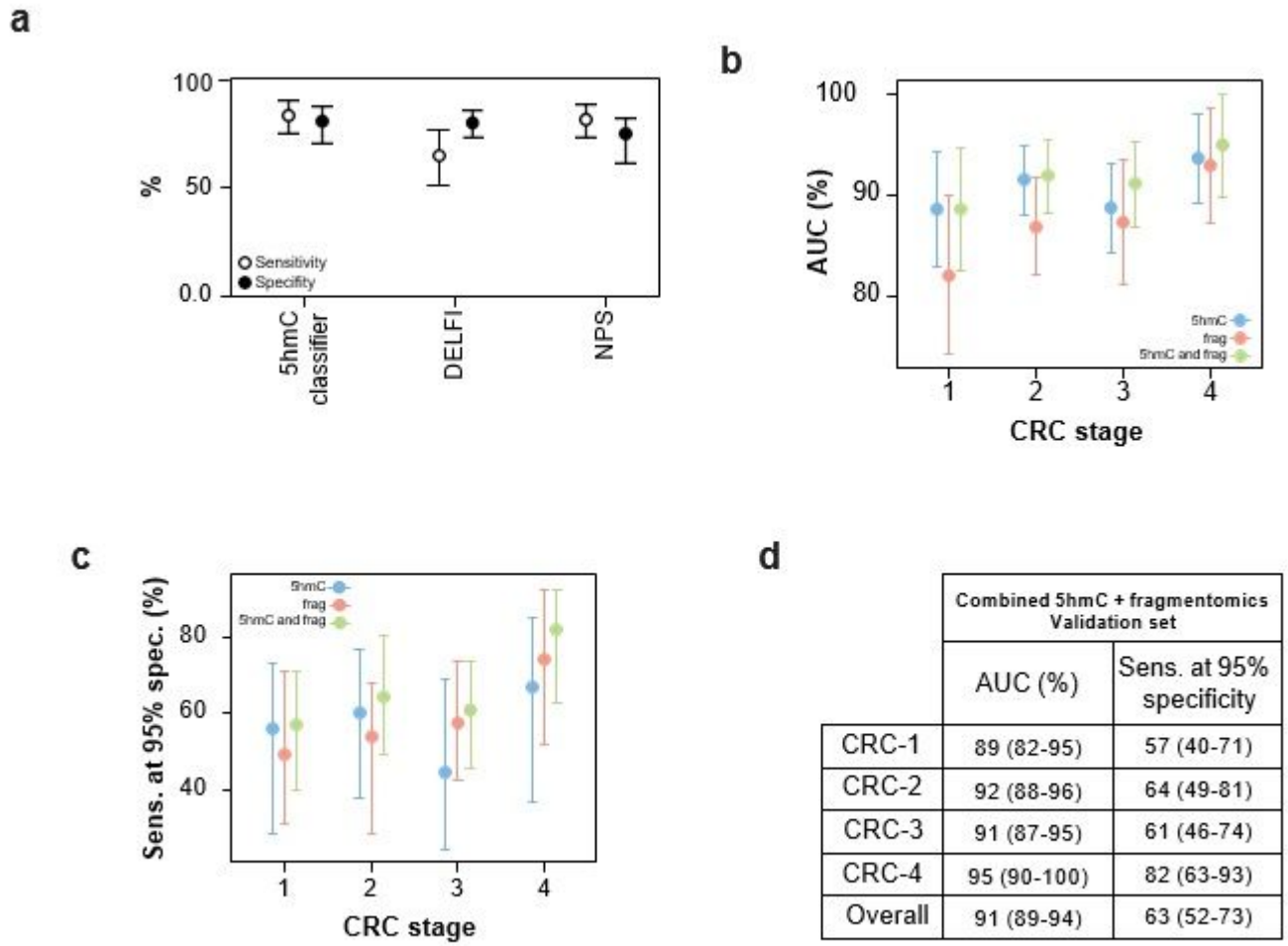
b



	Validation	
Stage	Spearman's Rho	p-value
CRC-1	0.66	1.140 E-1
CRC-2	0.33	1.600 E-3
CRC-3	0.59	1.196 E-6
CRC-4	0.77	4.704 E-6

Figure 5

(A) classifier trained using a DELFI like approach demonstrates CRC stage dependent performance in validation samples (B) The classifier prediction probability shows strong concordance with the estimated tumour fraction (ichorCNA) particularly in late stages in both training and validation samples.



	Combined 5hmC + fragmentomics Validation set	
	AUC (%)	Sens. at 95% specificity
CRC-1	89 (82-95)	57 (40-71)
CRC-2	92 (88-96)	64 (49-81)
CRC-3	91 (87-95)	61 (46-74)
CRC-4	95 (90-100)	82 (63-93)
Overall	91 (89-94)	63 (52-73)

Figure 6

(A) Overall, the median performance estimate was higher for the 5hmC classifier compared to the DELFI and NPS classifiers. (B-D) Median AUC and sensitivity at 95% specificity of 5hmC, DELFI like fragmentomics approach and combined classifier. 5hmC classifier performs better than the DELFI-like fragmentomics classifier in early CRC stages (1 & 2) while in late stages (3 & 4) 5hmC shows significant additivity at higher specificity (95% specificity).

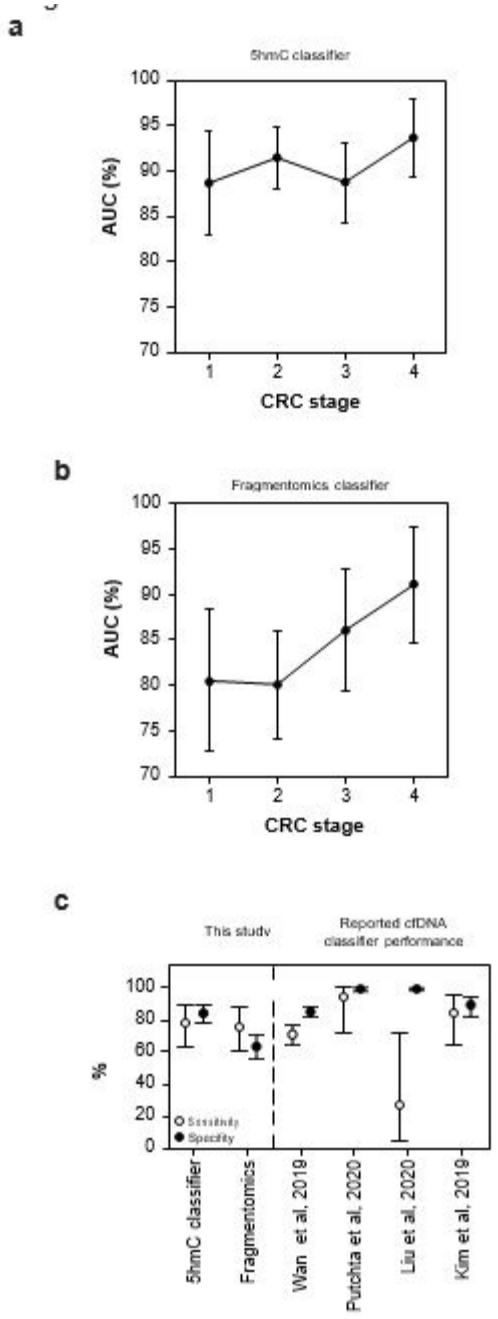


Figure 7

(A-B) A classifier trained on 5hmC levels in enhancer regions maintains performance at early-stage cancer compared to a model trained on cfDNA fragment size and coverage (DELFI-like approach). (C) 5hmC based classifier performs comparably to reported classifiers for Stage I colorectal cancer. To gain approximately comparable confidence intervals, 95% binomial confidence intervals have been computed for all classifiers using publicly available information8,12,32,33. The CRC classifier from Putcha et al, 2020 contains both Stage I and Stage II sample