Multimodal Epigenetic Sequencing Analysis (MESA) of Cell-free DNA for Non-invasive Cancer Detection

Wei Li (wei.li@uci.edu)
University of California, Irvine https://orcid.org/0000-0001-9931-5990

Yumei Li
University of California, Irvine

Jianfeng Xu
Helio Genomics, Inc.

Chaorong Chen
University of California, Irvine

Yang-kui Gu
Department of Minimally Invasive Interventional Radiology

Zhenhai Lu
Sun Yat-sen University Cancer Center

Diange Li
Laboratory of Advanced Medicine and Health

Jason Li
University of California, Irvine

Allison Sorg
Helio Genomics, Inc.

Curt Roberts
Helio Genomics, Inc.

Shivani Mahajan
Helio Genomics, Inc.

Maxime Gallant
Helio Genomics, Inc.

David Taggart
Helio Genomics, Inc.

Article

Keywords:

Posted Date: August 2nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1889126/v1

Abstract

Multimodal characterization of cell-free DNA (cfDNA) in blood can enable the sensitive and non-invasive detection of human cancers but remains technically challenging and costly. Here, we developed Multimodal Epigenetic Sequencing Analysis (MESA), a flexible and sensitive method of capturing and integrating multimodal epigenetic information of cfDNA using a single experimental assay, i.e., non-disruptive bisulfite-free methylation sequencing, such as Enzymatic Methyl-seq (EM-seq) and TET-assisted pyridine borane sequencing (TAPS). MESA can simultaneously infer four epigenetic modalities, namely cfDNA methylation, nucleosome occupancy, nucleosome fuzziness, and fragmentation profile for regions surrounding gene promoters and polyadenylation sites (PASs). When applied to 462 cfDNA samples from 2 independent clinical cohorts for colon cancer, new modalities (e.g., nucleosome fuzziness) and genomic features (e.g., PASs) introduced in MESA are highly complementary or superior to conventional ones, such as promoter DNA methylation, for cancer detection. Furthermore, MESA's integrated analysis of multimodal epigenetic features significantly improved the detection accuracy for colon, liver, and pancreatic cancers compared to single modality models. Together, MESA captures additional and highly complementary epigenetic information from cfDNA without additional experimental assays, highlighting the importance and clinical potential of using multimodal epigenetic features for non-invasive cancer detection.

Introduction

Cancer has long been a leading cause of death worldwide. While research on cancer treatment continues to make progress in reducing cancer mortality, early detection provides the best opportunity to improve patient survival and lower treatment cost[1]. Recently, the analysis of circulating cfDNA — degraded DNA fragments in blood plasma originating primarily from the apoptosis of normal and diseased cells — has shown great potential for early cancer detection[2-4]. The use of these liquid biopsies (non-invasive blood cfDNA-based detection methods) in routine screening is central to increasing surveillance adherence, identifying cancers in early curable stages, and ultimately reducing worldwide cancer mortality. One such approach is the whole-genome sequencing (WGS) of cfDNA, which provides genetic information, such as somatic mutations and copy number variations[5,6]. However, detecting cancer-specific genetic alterations is challenging due to the limited number of detectable changes and low fraction of circulating tumor DNA in patient blood samples[2,5,7,8].

Aside from genetic alterations, cfDNA methylation has been shown as a promising biomarker for early cancer detection, as aberrant DNA methylation has been frequently reported in cancer cells and may occur early in tumorigenesis[9-12]. Currently, the gold standard for the detection of DNA methylation is bisulfite sequencing. However, this harsh bisulfite treatment degrades a large fraction of the DNA resulting in biased genome coverage and increased sequencing cost[13]. Recently, the development of bisulfite-free DNA methylation sequencing methods, such as Enzymatic Methyl-seq (EM-seq) and TET-assisted pyridine borane sequencing (TAPS), have improved methylation sequencing quality and reduced sequencing cost[14-16].

Circulating cfDNA primarily consists of nucleosome-associated fragments that largely retain the chromatin structure information of the cells from which they originate[17,18]. As cfDNA is degraded by endonucleases before being released into the bloodstream, closed chromatin regions with dense nucleosomes are particularly well-protected against enzymatic degradation, while open chromatin regions are more sensitive to

endonuclease activity[17]. Several studies have developed methods utilizing chromatin-associated features for the non-invasive detection or monitoring of cancers, including nucleosome occupancy[19,20], window protection score[17], and fragmentation profile[18,21]. However, these methods rely on WGS and thus do not provide any further epigenetic information.

Recently, the non-destructive nature of EM-seq and TAPS enabled the combination of two epigenetic modalities based on low-coverage whole-genome methylation sequencing (**Supplementary Table 1**). In particular, TAPS-based cfTAPS[22] provided DNA methylation and fragmentation for 85 cancer/cirrhosis/pancreatitis/control samples. Similarly, EM-seq-based cfNOME[23] measured DNA methylation and nucleosome occupancy for 12 chronic kidney disease/control samples. Despite this progress, these two methods are largely limited by small sample sizes and fail to utilize the full spectrum of epigenetic information from cfDNA. Here, we introduce a four-in-one multimodal epigenetic sequencing analysis (MESA) of cfDNA (**Fig. 1**) for 462 colon cancer/control samples from two independent cohorts with deep targeted methylation sequencing. MESA can simultaneously infer four highly-complementary epigenetic modalities, namely 1) cfDNA methylation, 2) nucleosome occupancy, 3) nucleosome fuzziness, and 4) fragmentation profile across gene promoters and polyadenylation sites (PASs). MESA's integrated analysis of multimodal epigenetic features significantly improved the cancer detection accuracy compared to single modality models.

Results

MESA cohorts

To systematically demonstrate the performance of MESA, we designed two targeted EM-seq panels of different scales for two independent clinical cohorts, Cohort 1 (n = 130) and Cohort 2 (n = 332) (**Fig. 1, Supplementary Tables 2 and 3**). The target regions included a custom-designed methylation panel and a nucleosome organization panel with regions surrounding both transcription start sites (TSSs) and polyadenylation sites (PASs) of cancer-related genes (**Materials and methods**; **Supplementary Tables 4-7)**. Novel to our panel design is the introduction of PASs, whose alternative regulation is frequently reported to be involved in tumorigenesis[24-27]. Since nucleosome occupancy around PASs has also been reported to be associated with alternative polyadenylation regulation[28-30], we predicted that its inclusion would contribute to the improved performance of the cancer detection model. In contrast to low-pass whole-genome methylation sequencing such as cfTAPS[22] (mean coverage of 11.6×), this targeted design allowed us to perform deeper sequencing with a mean coverage of 74.2× (range from 41× to 123×) for Cohort 1 and a mean coverage of 200.3× (range from 76× to 570×) for Cohort 2 at a comparatively low cost. Next, we assessed the quality of our sequencing data based on non-human, internal spike-in controls with known unmethylated CpG sites (CpG-unmethylated lambda DNA). Less than 1% methylation was detected in unmethylated lambda, indicating a conversion efficiency of at least 99%.

cfDNA methylation in MESA enables accurate detection of colon cancer

As a baseline, we first explored the effectiveness of cfDNA methylation features alone in distinguishing between cancer patients and non-cancer controls. We observed that the average methylation level of all target CpG sites was elevated in cancer samples compared to non-cancer controls (**Fig. 2A** and **B**). This observation is

consistent with the fact that the targeted CpG sites are primarily located in promoter regions, which are known to be frequently hypermethylated in cancers[31]. To further investigate whether these methylation signatures can discriminate between cancer patients and non-cancer controls, we performed principal component analysis (PCA) for cfDNA methylation levels in all target CpG sites. Methylation at these sites showed reasonable separation in PC1 and PC2 (**Fig. 2C** and **D**). Next, we investigated the performance of these methylation features for colon cancer prediction using machine learning methods with leave-one-out cross-validation (LOOCV) (**Materials and methods**). Methylation alone achieved an impressive prediction of colon cancer for both cohorts based on an ensemble classifier incorporating random forest, logistic regression, and deep forest methods (**Fig. 2E** and **F**, AUC (area under the curve) = 0.8602 for Cohort 1 and AUC = 0.8422 for Cohort 2). These results suggested that cfDNA methylation in MESA can be used to detect colon cancer with reasonable accuracy.

MESA successfully captures nucleosome organization information

EM-seq preserves the integrity of cfDNA as compared to bisulfite conversion, which enabled us to capture additional epigenetic information. From the merged non-cancer controls, we observed a peak around 167 bp (corresponding to the length of DNA associated with a nucleosome and a linker histone) in the cfDNA fragment length distribution **(Fig. 3A** for Cohort 1 and **Supplementary Fig. 1A** for Cohort 2), which is consistent with that from cfDNA WGS data[17,19]. Further supporting the association between cfDNA and nucleosomes, the dinucleotide frequency of these fragments showed a ~10 bp periodicity **(Fig. 3B** for Cohort 1 and **Supplementary Fig. 1B** for Cohort 2), which recapitulates key features of nucleosome-associated fragments digested by micrococcal nuclease (MNase)[32]. Next, to accurately measure nucleosome occupancy profiles from cfDNA, we used the quantification method DANPOS2[33,34], a tool widely used for processing MNase-seq (a technique used for profiling nucleosome landscape) data[35]. The occupancy profiles reported by DANPOS2 were concordant with nucleosome profiles from lymphoblastoid cells **(Fig. 3C**), indicating the targeted EM-seq successfully captured nucleosome information. Moreover, profiles reported by DANPOS2 had lower background noise compared with raw read coverage measurements, as shown by example regions **(Fig. 3C**) and the typical well-positioned nucleosomes around TSSs **(Fig. 3D**). Interestingly, we also observed a nucleosome-depleted region (NDR) around PASs and well-positioned nucleosomes flanking NDR (**Fig. 3E**). These results indicated that MESA was able to capture nucleosome organization information in both TSSs and PASs.

Nucleosome occupancy and fuzziness in MESA enable accurate detection of colon cancer

Based on our findings that DANPOS2 could accurately measure nucleosome organization features from targeted EM-seq, we then investigated whether these features could be used for cancer detection. We derived two types of features from nucleosome organization: 1) nucleosome occupancy, which reflects the frequency with which nucleosomes occupy a given DNA region in a cell population; 2) nucleosome fuzziness, which is defined as the deviation of nucleosome positions within a region in a cell population and could reflect cell heterogeneity at the chromatin level (**Fig. 4A**). Both features were defined for each nucleosome organization target region (TSS and PAS target regions) by DANPOS2 (**Materials and methods**). We hypothesized that nucleosome occupancy and fuzziness might capture non-overlapping changes between cancer and control samples. Genome browser track visualization of four regions showed examples of either occupancy or

fuzziness changes between cancer and control samples in Cohort 1 (**Fig. 4B**). Particularly, these changes were found in both TSS (**Fig. 4B**, top panels) and PAS (**Fig. 4B**, bottom panels) regions, emphasizing the importance of introducing PAS target regions in the MESA panel design.

We then investigated the predictive potential of nucleosome occupancy and fuzziness using the LOOCV method. Consistent with previous work[19], our model based solely on nucleosome occupancy of TSS target regions achieved an AUC of 0.8055 for Cohort 1 and 0.9160 for Cohort 2 (**Fig. 4C** and **D**). Interestingly, adding PAS target regions further improved model performance, as demonstrated by the enhanced AUC after combining nucleosome occupancy features of TSS and PAS target regions (**Fig. 4C** and **D**; AUC = 0.8464 for Cohort 1 and AUC = 0.9497 for Cohort 2). To our knowledge, this is the first time that nucleosome occupancy around PAS regions from cfDNA has been utilized in cancer detection. Another novelty of our design is the introduction of nucleosome fuzziness, which reflects cell heterogeneity at the chromatin level[33,36]. Nucleosome fuzziness based on cfDNA may differentiate cancer from controls, as cancerous tissue is typically more heterogeneous than normal tissue[37,38]. Indeed, nucleosome fuzziness alone achieved exceptional cancer classification (**Fig. 4E** and **F**; AUC = 0.7569 for Cohort 1 and AUC = 0.9238 for Cohort 2). Moreover, the combination of the two modalities (nucleosome occupancy and fuzziness) further improved model performance for at least one of the two cohorts (**Fig. 4E** and **F**; AUC = 0.8457 for Cohort 1 and AUC = 0.9574 for Cohort 2). These results suggested that the new modality (nucleosome fuzziness) and genomic feature (PASs) introduced in MESA are effective for cancer detection.

Integrating multimodal epigenetic features in MESA enhances cancer detection

We next investigated the integration of multimodal features captured by MESA for cancer detection. In addition to DNA methylation, nucleosome occupancy, and nucleosome fuzziness features we previously introduced, we also included fragmentation profile which has been widely used for cancer detection[18]. Using LOOCV, we found that the integrated model has the highest AUC of 0.8962 for Cohort 1 and 0.9562 for Cohort 2 than four single modality models (**Fig. 5A** and **B; Supplementary Table 8**), highlighting the benefits of incorporating multimodal information in cancer prediction. When evaluating models based on cancer stage, the multimodal model still outperformed single modality models (**Fig. 5C** and **D; Supplementary Fig. 2** and **Fig. 3**). By visualizing the predicted probability of classifying each sample to the cancer group, we found a similar pattern for the four single modality models (**Fig. 5E** and **F**), suggesting that each modality concordantly predicted the same classification for most samples. Additionally, when examining the correlations between the probabilities of different single modality models, we found correlations as low as 0.19 (**Fig. 5G** and **H**), indicating that single modality models may capture complementary information for cancer detection. The observed improved performance of the multimodal model is consistent with the fact that the integration of single modalities into multimodal features combines the complementary information. Together, MESA's integrated analysis of multimodal epigenetic features significantly improved the performance of cancer detection models relative to the usage of a single modality alone.

MESA with cfTAPS data

As MESA took advantage of the non-disruptive nature of EM-seq to capture multimodal epigenetic information from a single assay, the multimodal approach was predicted to effectively perform on any cfDNA methylation

sequencing assay of a similar nature. We tested this hypothesis on another bisulfite-free cfDNA sequencing method cfTAPS[22], which was applied to a cohort including 21 hepatocellular carcinoma (HCC) patients, 23 pancreatic ductal adenocarcinoma (PDAC) patients, and 30 non-cancer controls. As shown by a well-studied nucleosome array, the occupancy reported by DANPOS2 for cfTAPS data was consistent with nucleosome profiles from lymphoblastoid cells (**Fig. 6A**), indicating cfTAPS could capture nucleosome information as targeted EM-seq did. Despite the low sequence depth (mean coverage of 11.6×), we still observed occupancy changes between cancer and control samples for regions surrounding either TSSs or PASs (**Fig. 6B**). Then, we extracted three types of features, including DNA methylation, nucleosome occupancy, and fragmentation profile. Next, we applied the same model training method as for Cohort 1 and Cohort 2 to the cohort of cfTAPS data (HCC vs. control; PDAC vs. control). Here, we did not include nucleosome fuzziness because it was inaccurate to calculate the fuzziness score when the sequencing depth was low. In line with results from Cohort 1 and Cohort 2, we found that the multimodal model has the highest AUC compared with three single modality models (**Fig. 6C** and **D**; AUC = 0.9381 for HCC cohort and AUC = 0.8449 for PDAC cohort). Since there were two cancer types in this dataset, we also trained three-class models to distinguish HCC, PDAC, and controls. Using LOOCV, we found the multimodal model achieved an overall accuracy of 0.7297 (**Fig. 6D** and **Supplementary Fig. 4**), which outperformed the three single modality models. Together, MESA's integrated analysis of multimodal epigenetic features is widely applicable in multiple non-disruptive methylation sequencing protocols.

Discussion

In this study, we present a comprehensive epigenetic analysis of cfDNA, aiming at improving the non-invasive early detection of human cancers. Our integrated model shows enhanced detection accuracy for colon, liver and pancreatic cancers compared to single modality models in three independent cohorts with either EM-seq or cfTAPS. A significant advantage of the multimodal assay is its flexibility. Based on each unique dataset and patient cohort input, each of the four modalities of epigenetic information may be either included or excluded in an unbiased manner. For example, cancer types in which nucleosome occupancy is relatively unchanged may benefit only from integrating the remaining modalities. Removal of nucleosome occupancy features, in this case, could prevent confounding and unnecessary complexity. Therefore, this multimodal approach allows for the development of an unbiased combinatorial prediction model. Furthermore, all four modalities are simultaneously captured in a single assay, offering full flexibility without the need to perform multiplex assays while minimizing potential batch effects and other technical biases in multiplex and separated assays.

A potential concern of this multimodal approach is that modalities might be highly correlated with one another, thus not necessarily reflecting complementary information. In this paper, we showed that the predicted probabilities of individual modalities are not highly correlated. For example, although the nucleosome organization is weakly related to the fragmentation profile[18], nucleosomes can provide information other than fragmentation. Nucleosome organization focuses on the position-specific cfDNA fragments, while fragmentation profile concentrates only on the size of the cfDNA fragments globally[18,21]. Even if two samples have the same fragment size distribution, these samples may possess dramatically different nucleosome organization in most regions. Therefore, they can still provide complementary information for the prediction model. We further note that, to our knowledge, this study introduces the measurement of nucleosome fuzziness

and PAS regions for the first time in cfDNA sequencing data analysis. As shown by our results, they both contribute to a better performance of the cancer detection model.

One limitation of our study is its relatively small sample size. Follow-up studies will be needed to strengthen the application of MESA in a wide variety of human cancers. However, despite the limitations, our study demonstrates a salient example of how targeted EM-seq of cfDNA captures multimodal epigenetic information and enables accurate detection of cancer at a low relative cost. Our design provides a clinically practical method for liquid biopsy, especially for cancer types with few or no genetic changes. Moreover, for Cohort 1, we observed better performances of the multimodal model for early stage (I and II) than late stage (III and IV) patients (**Fig. 5C**). Although this observation may be biased by the relatively small sample size of each stage, it shows the possible advantages of MESA on early cancer detection. As cfDNA methylation-based liquid biopsies garner more attention and clinical use, MESA represents a widely applicable platform for improving non-invasive cancer detection.

Materials And Methods

Study cohort

Cohort 1 comprised 70 patients diagnosed with colorectal cancer and 60 control individuals without colorectal cancer. Cohort 1 subjects were recruited at clinical sites within the United States through the ELITE Study (NCT05181826) or were obtained through the following contract research organizations: BioIVT (Westbury, NY), BioOptions (Brea, CA), Discovery Life Sciences (Boston, MA) and DX Biosamples (San Diego, CA). Cohort 2 comprised 129 patients diagnosed with colorectal cancer and 203 control patients without colorectal cancer. Cohort 2 subjects were enrolled at the Sun Yat-sen University Cancer Center (Guangzhou, China). Subjects diagnosed with colorectal cancer were diagnosed according to current clinical practices. Control subjects did not have any clinical history or symptoms of colorectal cancer. All specimen collection protocols were approved by the respective Institutional Review Board (IRB). For both Cohort 1 and Cohort 2, informed consent was obtained from all patients, in accordance with the Declaration of Helsinki Ethical Principles for medical research involving human subjects.

Collection and preparation of samples

Cohort 1 specimens were drawn into PAXgene cfDNA tubes (PreAnalytiX) and shipped to a central Helio Genomics laboratory (USA) using custom specimen collection and shipping kits (Helio Genomics). The whole blood specimens were then processed to cleared plasma by centrifugation and stored at approximately -80°C until analysis. Cohort 2 specimens were drawn into KANGJIAN blood collection tubes at the corresponding hospital. Samples were shipped to a central LAMH laboratory (Laboratory of Advanced Medicine and Health, China) with dry ice and stored at approximately -80°C until analysis.

Targeted sequencing panel design

TCGA-COAD and TCGA-READ 450K methylation array data were downloaded from the UCSC Xena database (https://tcga.xenahubs.net) and processed by a custom script to identify CpG sites with significant methylation differences between cancerous and adjacent normal tissues. A total of 9599 significantly differentially

methylated CpG sites in the colorectal cancer samples were selected, and a list of 150 bp genomic regions centered on each of the selected CpG sites was designed for targeted sequencing. Additionally, 912 promoter regions (TSS ± 1 kb) and 365 polyadenylation regions (PAS ± 1 kb) of the curated cancer-related genes were added to the targeted panel. With the repeat elements and ENCODE blacklist regions removed[39], the size of the version 1 colorectal cancer targeted panel (used on Cohort 1 subjects) was about 4.6 Mb (**Supplementary Table 4**). The shrinking version 2 colorectal cancer targeted panel (used on Cohort 2 subjects) was about 220 kb (**Supplementary Table 5**). Both targeted panels were synthesized by Twist Bioscience (USA).

Targeted EM-seq of cfDNA

The Helio ECLIPSE™ platform was used to analyze cfDNA extracted from patient specimens as previously described[40]. Briefly, total cfDNA was isolated from specimens by using either (Cohort 1) a QIAsymphony DSP Circulating DNA Kit (QIAGEN, USA) or (Cohort 2) the EliteHealth cfDNA Extraction Kit (EliteHealth, China). Spike-in control unmethylated Lambda DNA was sheared down into about 170 bp by sonication. A total of 5 ng cfDNA along with 0.2 pg of unmethylated Lambda DNA per specimen was used to prepare the barcoded NGS libraries using the NEB Next Enzymatic Methyl-seq Kit (New England Biolabs, USA) according to the manufacturer's instructions. The libraries were then hybridized with a custom set of capture probes (Twist Bioscience, USA) to capture the targeted library sequences using the Twist Fast Hybridization and Wash Kit, along with the Twist Universal Blocker. The captured libraries were then supplemented with a 20% PhiX genomic DNA library to increase base calling diversity and submitted for sequencing using Illumina NovaSeq 6000 instruments as 2 × 150 bp reads.

Targeted EM-seq data processing and quality control

Raw sequencing reads were first trimmed by TrimGalore (v0.6.5, https://github.com/FelixKrueger/TrimGalore) to remove low-quality reads and potential adaptor contamination. Then the remaining reads were aligned to the hg19 human genome reference using BSMAP (v2.90)[41]. The aligned reads were further processed by Samtools (v0.1.19)[42] and Bedtools (v2.29.1)[43] to only keep primarily mapped reads with fragment sizes between 80 bp and 200 bp to remove potential genomic DNA contamination from normal blood cells. This final file served as the input file for all the following processes except fragment size distribution analysis, which used reads without a size filter. Spike-in unmethylated lambda DNA was used to control for C to T conversion efficiency. Samples with lambda methylation levels of more than 1% (CT conversion rate less than 99%) were removed from the downstream analysis.

Multimodal feature extraction from targeted EM-seq of cfDNA

We extracted four types of features: cfDNA methylation, nucleosome occupancy, nucleosome fuzziness, and fragmentation profile.

- cfDNA methylation: Conventional methylation ratio was calculated by Methratio.py (BSMAP, v2.90) from aligned bam files for the target CpG sites.
- Nucleosome occupancy: Occupancy values were calculated using DANPOS2 (v2.2.2), a tool widely used for processing MNase-seq data. For Cohort 2, the average value for each nucleosome organization target region was calculated using bigWigAverageOverBed from UCSC tools (v393)[44]. Due to the relatively long

target regions of Cohort 1 (2 kb), we split each target region into 1 kb sliding windows with 10 bp steps. Then, for each sliding window, we calculated the average nucleosome occupancy.

- Nucleosome fuzziness: Fuzziness values were calculated using DANPOS2. For each nucleosome organization target region (1 kb sliding windows for Cohort 1), we calculated the average fuzziness of all the nucleosomes whose center is located within the region.
- Fragmentation profile: Fragmentation profile was defined as the ratio of short (80 - 150 bp) to long cfDNA fragments (151 - 200 bp) in a target region.

Single modality machine learning models for cancer detection

We trained machine learning models for Cohort 1 and Cohort 2 using the same procedure. All the models were trained and evaluated using the leave-one-out cross-validation (LOOCV) method. Briefly, all the N samples were divided into training samples and test samples for N iterations, where the number of test samples = 1 and the number of training samples = N-1. Since missing values could reduce the accuracy of the machine learning model, we imputed the missing values in each iteration. Firstly, we removed features with missing values in more than 10% of the training sample. Then we imputed the missing values with the average valid values of training samples for the corresponding feature. Next, feature selection and model training were performed in the training samples. Then predictive power of the features and performance of the model were evaluated on the test sample. Finally, the results of all the N iterations were aggregated together to calculate performance metrics.

For each of the four modalities, sequential backward selection (SBS) was used for feature selection[45], in which features are sequentially removed from the complete feature set, and the feature subset with the highest ROC AUC is finally selected. Specifically, in each iteration of LOOCV, we calculated rankings for all features in the training set based on their contribution to the prediction using the Boruta algorithm in the BorutaPy package (v0.3)[46]. Then SBS was used to refine the top 1000 features from Boruta with ROC AUC based on support vector machine (SVM), which outputs a selected feature subset. Next, we applied a voting classifier for the model training, which ensembled the predictive probabilities from three methods: Logistic Regression, Random Forest[47], and Deep Forest[48]. Briefly, the voting prediction result on the i th testing sample was given by

$$P_{cancer}^{i} = \frac{1}{3}\sum_{m=1}^{3} P_{m,cancer}^{i}$$

Where P_{cancer}^{i} denotes the final predictive probability of classifying the i th testing sample to cancer group and $P_{m,cancer}^{i}$ denotes the predictive probability of classifying the i th testing sample to the cancer group using model m. This procedure was done by ensemble.VotingClassifier from scikit-learn package (v0.24.2)[49]. Predictive results of all the LOOCV iterations were aggregated together to get an overall AUC value.

Multimodal machine learning model for cancer detection

The multimodal machine learning model was built following the model-based multimodal integration strategy[50]. A feature subset was selected for each of the four modalities based on the training procedure. The multimodal prediction model was then trained on the concatenation of z-score standardized selected features of the four modalities and predicted on the corresponding testing sample. This approach could preserve unique information from different modalities and provide complementary information across different types of features.

cfTAPS data processing and machine learning models for cancer detection

The cfTAPS data was processed in the same manner as the original paper[22]. Raw sequenced reads were trimmed using TrimGalore (v0.6.5, https://github.com/FelixKrueger/TrimGalore) to remove adapter and low-quality bases. Trimmed reads were aligned to the hg19 human reference genome using bwa mem (v0.7.17)[51]. The alignment files were filtered to remove low mapping quality (MAPQ < 20) as well as duplicate reads using alignmentSieve from deepTools (v3.5.0)[52]. MethylDackel extract (v0.6.1, https://github.com/dpryan79/MethylDackel) was used for methylation calling. CpG sites that overlapped common single-nucleotide polymorphism (SNP)[53] (https://ftp.ncbi.nih.gov/snp/organisms/human_9606_b151_GRCh37p13), blacklisted regions[39], centromeres, and sex chromosomes were excluded from downstream analysis.

Next, we extracted three types of features: DNA methylation, nucleosome occupancy, and fragmentation profile. (1) DNA methylation: The methylation ratio was calculated using the number of methylated CpGs divided by the total number of sequenced CpGs for each promoter and enhancer region. The promoter and enhancer regions were downloaded from Ensemble[54] (http://ftp.ensembl.org/pub/grch37/release-100/regulation/homo_sapiens/homo_sapiens.GRCh37.Regulatory_Build.regulatory_features.20191101.gff.gz). (2) Nucleosome occupancy: Occupancy values were calculated using DANPOS2. Average values of the 1 kb regions surrounding TSSs and PASs of all RefSeq annotated genes[55] were calculated. The locations of PASs were downloaded from PolyA_DB (version 3)[56]. Due to the relatively low coverage of cfTAPS data, we removed

features that had occupancy values lower than the mean of all values in at least one sample. (3) Fragmentation profile: Fragmentation profiles were calculated as the fraction of cfDNA fragments (300 – 500 bp) at 10-bp length range bins, which were used in the original cfTAPS paper[22].

We then trained both two-class (distinguishing cancer (HCC or PDAC) and control samples) and three-class models (distinguishing HCC, PDAC, and control samples) using the same procedure as for targeted EM-seq data. The only difference was that we applied SBS to the top 300 features from Boruta because of the relatively low sample size. For the three-class models, we used accuracy instead of ROC AUC as the performance metric for SBS.

Declarations

Availability of data

All processed data used to generate the results are available at Zenodo https://doi.org/10.5281/zenodo.6812875. The raw sequencing data will be available at the European Genome-phenome Archive (EGA) (accession number in processing).

Availability of codes

MESA source code: https://github.com/ChaorongC/MESA.

All the codes and data used to reproduce all the major results in this manuscript: https://rpubs.com/LiYumei/926228.

Consent for publication

Not applicable

Competing interests

Wei Li is a consultant for Helio Genomics and ChosenMed. Jianfeng Xu, Allison J. Sorg, Curt C. Roberts, Shivani Mahajan, Maxime A. Gallant, and David J. Taggart are employees of Helio Genomics. Diange Li is an employee of Laboratory of Advanced Medicine and Health.

Funding

This work was supported by the following grants: UC Irvine Setup fund and Grace B. Bell Endowed Chair fund to W.L. Y.L. is supported by the Hewitt Foundation for Medical Research Postdoctoral Fellowship.

Author contributions

W.L., Y.L., and J.X. conceived and supervised this project. Y.L., J.X., and C.C. performed the data analysis. Y. G. and Z. L. collected and prepared the samples. All authors interpreted the data and wrote the manuscript.

Acknowledgements

The authors thank the patients who generously participated in the study and the principal investigators and institutions who oversaw their enrollments. We also thank members of Wei Li lab for helpful discussions.

Supplementary information

Supplementary Figs. 1 to 4 and Tables 1 to 8.

References

1 Hawkes, N. Cancer survival data emphasise importance of early diagnosis. *BMJ* **364**, l408, doi:10.1136/bmj.l408 (2019).

2 Bettegowda, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med* **6**, 224ra224, doi:10.1126/scitranslmed.3007094 (2014).

3 Lui, Y. Y. *et al.* Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation. *Clin Chem* **48**, 421-427 (2002).

4 Phallen, J. *et al.* Direct detection of early-stage cancers using circulating tumor DNA. *Sci Transl Med* **9**, doi:10.1126/scitranslmed.aan2415 (2017).

5 Leary, R. J. *et al.* Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. *Sci Transl Med* **4**, 162ra154, doi:10.1126/scitranslmed.3004742 (2012).

6 Zviran, A. *et al.* Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nat Med* **26**, 1114-1124, doi:10.1038/s41591-020-0915-3 (2020).

7 Liebs, S. *et al.* Detection of mutations in circulating cell-free DNA in relation to disease stage in colorectal cancer. *Cancer Med* **8**, 3761-3769, doi:10.1002/cam4.2219 (2019).

8 Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333-339, doi:10.1038/nature12634 (2013).

9 Chan, K. C. *et al.* Noninvasive detection of cancer-associated genome-wide hypomethylation and copy number aberrations by plasma DNA bisulfite sequencing. *Proc Natl Acad Sci U S A* **110**, 18761-18768, doi:10.1073/pnas.1313995110 (2013).

10 Liu, M. C. *et al.* Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann Oncol* **31**, 745-759, doi:10.1016/j.annonc.2020.02.011 (2020).

11 Shen, S. Y. *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579-583, doi:10.1038/s41586-018-0703-0 (2018).

12 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* **36**, 350-368, doi:10.1016/j.ccell.2019.09.003 (2019).

13 Tanaka, K. & Okamoto, A. Degradation of DNA by bisulfite treatment. *Bioorg Med Chem Lett* **17**, 1912-1915, doi:10.1016/j.bmcl.2007.01.040 (2007).

14 Liu, Y. *et al.* Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat Biotechnol* **37**, 424-429, doi:10.1038/s41587-019-0041-2 (2019).

15 Schutsky, E. K. *et al.* Nondestructive, base-resolution sequencing of 5-hydroxymethylcytosine using a DNA deaminase. *Nat Biotechnol*, doi:10.1038/nbt.4204 (2018).

16 Vaisvila, R. *et al.* Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res* **31**, doi:10.1101/gr.266551.120 (2021).

17 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).

18 Cristiano, S. *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385-389, doi:10.1038/s41586-019-1272-6 (2019).

19 Ulz, P. *et al.* Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nat Genet* **48**, 1273-1278, doi:10.1038/ng.3648 (2016).

20 Ulz, P. *et al.* Inference of transcription factor binding from cell-free DNA enables tumor subtype prediction and early detection. *Nat Commun* **10**, 4666, doi:10.1038/s41467-019-12714-4 (2019).

21 Mathios, D. *et al.* Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat Commun* **12**, 5060, doi:10.1038/s41467-021-24994-w (2021).

22 Siejka-Zielinska, P. *et al.* Cell-free DNA TAPS provides multimodal information for early cancer detection. *Sci Adv* **7**, eabh0534, doi:10.1126/sciadv.abh0534 (2021).

23 Erger, F. *et al.* cfNOMe - A single assay for comprehensive epigenetic analyses of cell-free DNA. *Genome Med* **12**, 54, doi:10.1186/s13073-020-00750-5 (2020).

24 Lopez de Silanes, I., Quesada, M. P. & Esteller, M. Aberrant regulation of messenger RNA 3'-untranslated region in human cancer. *Cell Oncol* **29**, 1-17, doi:10.1155/2007/586139 (2007).

25 Mayr, C. & Bartel, D. P. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**, 673-684, doi:10.1016/j.cell.2009.06.016 (2009).

26 Lembo, A., Di Cunto, F. & Provero, P. Shortening of 3'UTRs correlates with poor prognosis in breast and lung cancer. *PLoS One* **7**, e31129, doi:10.1371/journal.pone.0031129 (2012).

27 Xia, Z. *et al.* Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat Commun* **5**, 5274, doi:10.1038/ncomms6274 (2014).

28 Spies, N., Nielsen, C. B., Padgett, R. A. & Burge, C. B. Biased chromatin signatures around polyadenylation sites and exons. *Mol Cell* **36**, 245-254, doi:10.1016/j.molcel.2009.10.008 (2009).

29 Khaladkar, M., Smyda, M. & Hannenhalli, S. Epigenomic and RNA structural correlates of polyadenylation. *RNA Biol* **8**, 529-537, doi:10.4161/rna.8.3.15194 (2011).

30 Huang, H., Chen, J., Liu, H. & Sun, X. The nucleosome regulates the usage of polyadenylation sites in the human genome. *BMC Genomics* **14**, 912, doi:10.1186/1471-2164-14-912 (2013).

31 Herman, J. G. & Baylin, S. B. Gene silencing in cancer in association with promoter hypermethylation. *N Engl J Med* **349**, 2042-2054, doi:10.1056/NEJMra023075 (2003).

32 Gaffney, D. J. *et al.* Controls of nucleosome positioning in the human genome. *PLoS Genet* **8**, e1003036, doi:10.1371/journal.pgen.1003036 (2012).

33 Chen, K. *et al.* DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res* **23**, 341-351, doi:10.1101/gr.142067.112 (2013).

34 Chen, K. *et al.* Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nat Genet* **47**, 1149-1157, doi:10.1038/ng.3385 (2015).

35 Johnson, S. M., Tan, F. J., McCullough, H. L., Riordan, D. P. & Fire, A. Z. Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. *Genome Res* **16**, 1505-1516, doi:10.1101/gr.5560806 (2006).

36 Pugh, B. F. A preoccupied position on nucleosomes. *Nat Struct Mol Biol* **17**, 923, doi:10.1038/nsmb0810-923 (2010).

37 Meacham, C. E. & Morrison, S. J. Tumour heterogeneity and cancer cell plasticity. *Nature* **501**, 328-337, doi:10.1038/nature12624 (2013).

38 Sun, X. X. & Yu, Q. Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. *Acta Pharmacol Sin* **36**, 1219-1227, doi:10.1038/aps.2015.92 (2015).

39 Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354, doi:10.1038/s41598-019-45839-z (2019).

40 Lin, N. *et al.* A multi-analyte cell-free DNA-based blood test for early detection of hepatocellular carcinoma. *Hepatol Commun*, doi:10.1002/hep4.1918 (2022).

41 Xi, Y. & Li, W. BSMAP: whole genome bisulfite sequence MAPping program. *BMC Bioinformatics* **10**, 232, doi:10.1186/1471-2105-10-232 (2009).

42 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

43 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842, doi:10.1093/bioinformatics/btq033 (2010).

44 Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204-2207, doi:10.1093/bioinformatics/btq351 (2010).

45 Cotter, S. F., Kreutz-Delgado, K. & Rao, B. D. Backward sequential elimination for sparse vector subset selection. *Signal Processing* **81**, 1849-1864, doi:10.1016/S0165-1684(01)00064-0 (2001).

46 Kursa, M. B. & Rudnicki, W. R. Feature Selection with the Boruta Package. *Journal of Statistical Software* **36**, 1-13, doi:10.18637/jss.v036.i11 (2010).

47 Breiman, L. Random Forests. *Machine Learning volume* **45**, 5-32, doi:10.1023/A:1010933404324 (2001).

48 Zhou, Z.-H. & Feng, J. Deep Forest: Towards an Alternative to Deep Neural Networks∗. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 3553-3559, doi:10.24963/ijcai.2017/497 (2017).

49 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830, doi:10.5555/1953048.2078195 (2011).

50 Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A. & Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genet* **16**, 85-97, doi:10.1038/nrg3868 (2015).

51 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

52 Ramirez, F., Dundar, F., Diehl, S., Gruning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**, W187-191, doi:10.1093/nar/gku365 (2014).

53 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311, doi:10.1093/nar/29.1.308 (2001).

54 Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The ensembl regulatory build. *Genome Biol* **16**, 56, doi:10.1186/s13059-015-0621-5 (2015).

55 Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-65, doi:10.1093/nar/gkl842 (2007).

56 Wang, R., Nambiar, R., Zheng, D. & Tian, B. PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res* **46**, D315-D319, doi:10.1093/nar/gkx1000 (2018).

Figures

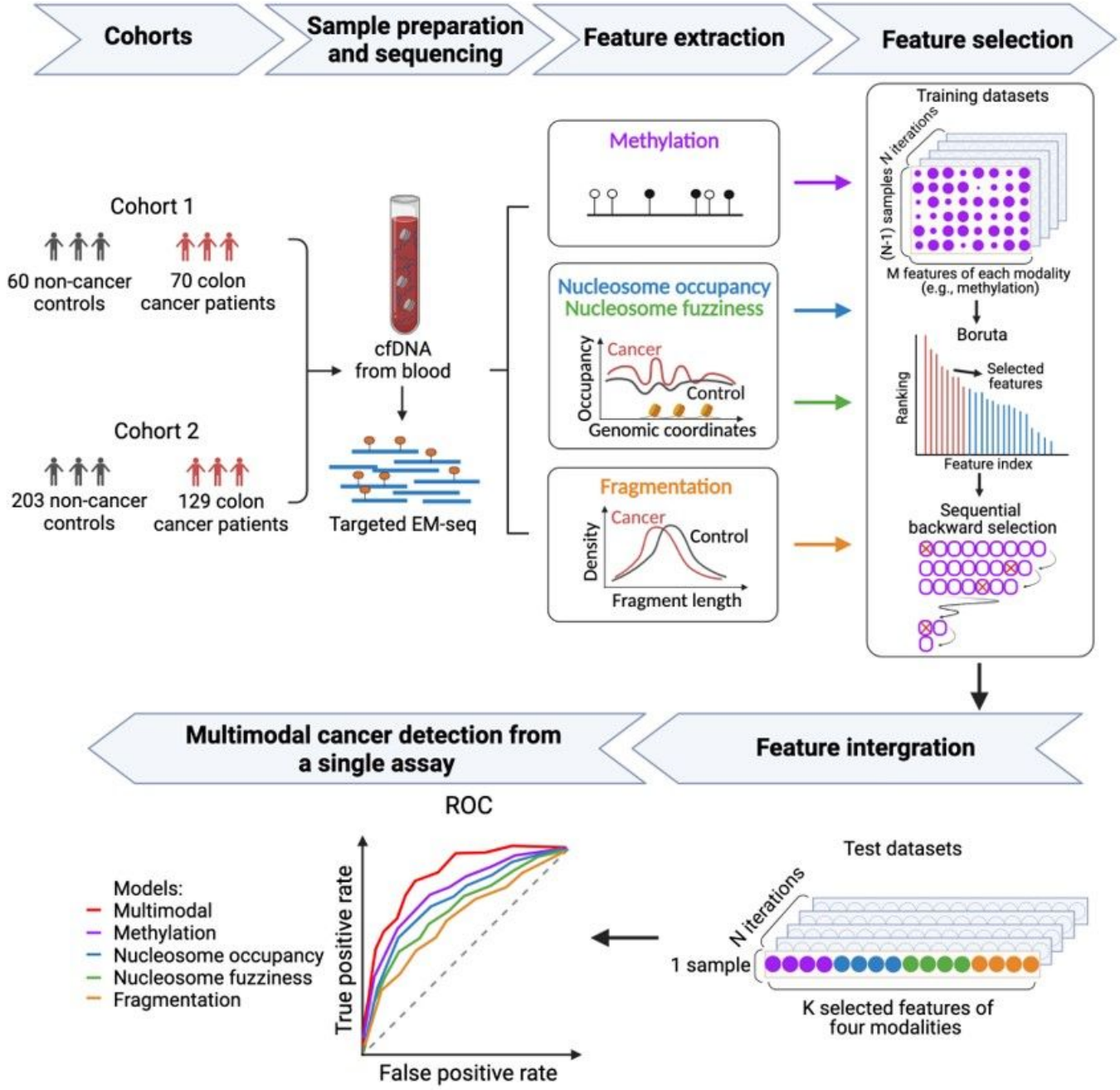

Figure 1

Schematic diagram displaying the design of MESA. cfDNA is isolated from blood samples of two independent cohorts and then processed to generate targeted EM-seq libraries. Analysis of the EM-seq data enables the extraction of four modalities: cfDNA methylation (purple), nucleosome occupancy (blue), nucleosome fuzziness (green), and fragmentation profile (orange). Then, the feature selection is performed for each modality separately. The Boruta algorithm is used for feature ranking, and the top-ranking features (shown in red) are selected for the sequential backward selection procedure. Selected features are standardized and combined for training a multimodal machine learning model, which outperforms the single modality models in the detection of cancer based on leave-one-out cross-validation.

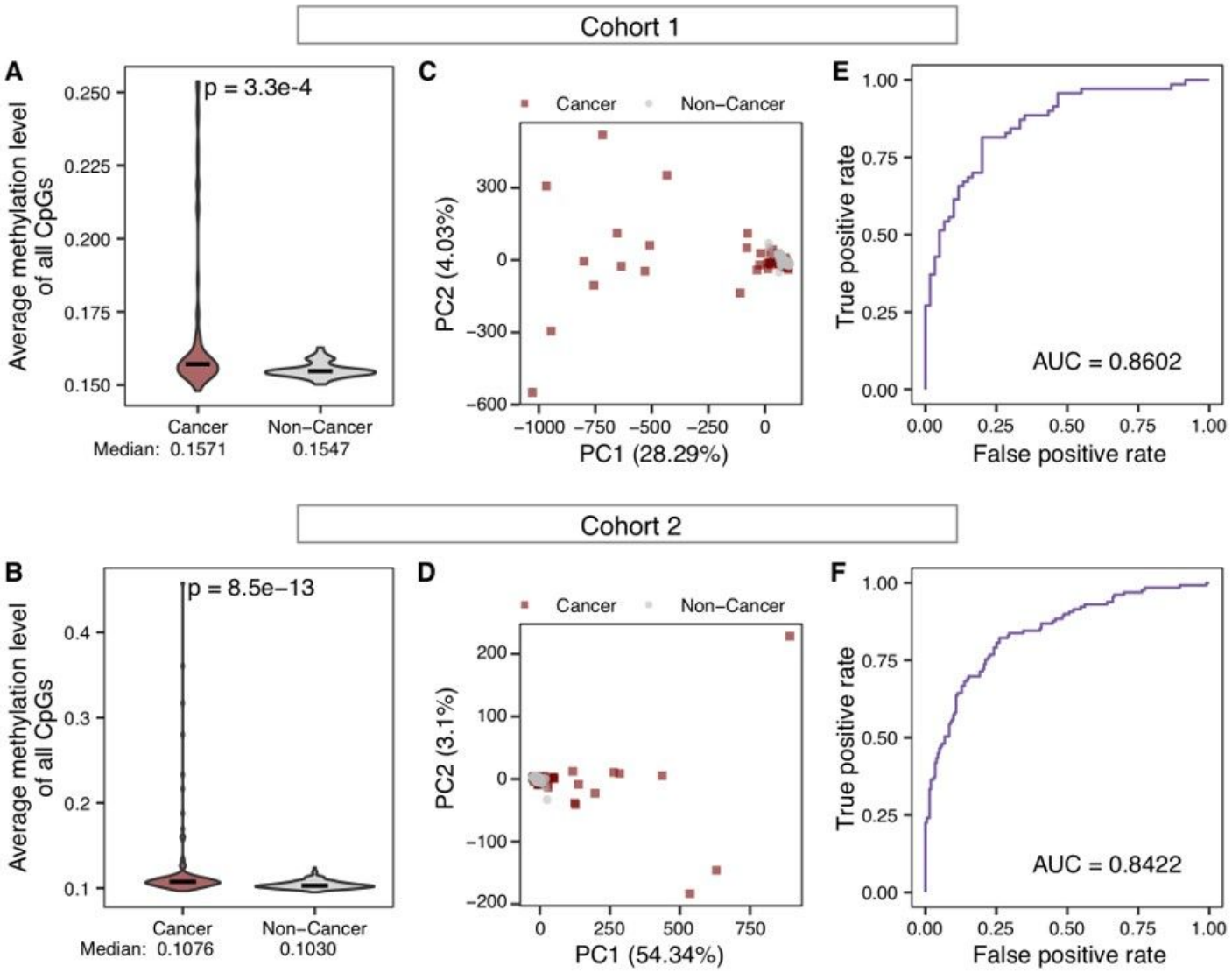

Figure 2

Differential cfDNA methylation between cancer and non-cancer samples enables accurate cancer detection.

The average methylation level of all target CpG sites in cancer patients (Cancer) and controls (Non-Cancer) from Cohort 1 **(A)** and Cohort 2 **(B)**. Scatter plots showing PC1 and PC2 from PCA of methylation level of all target CpG sites in cancer patients (Cancer) and controls (Non-cancer) of Cohort 1 **(C)** and Cohort 2 **(D)**. The percentage of variances explained by each PC is shown in the parentheses. ROC curves of model performance based on the methylation level of CpG sites for Cohort 1 **(E)** and Cohort 2 **(F)**.

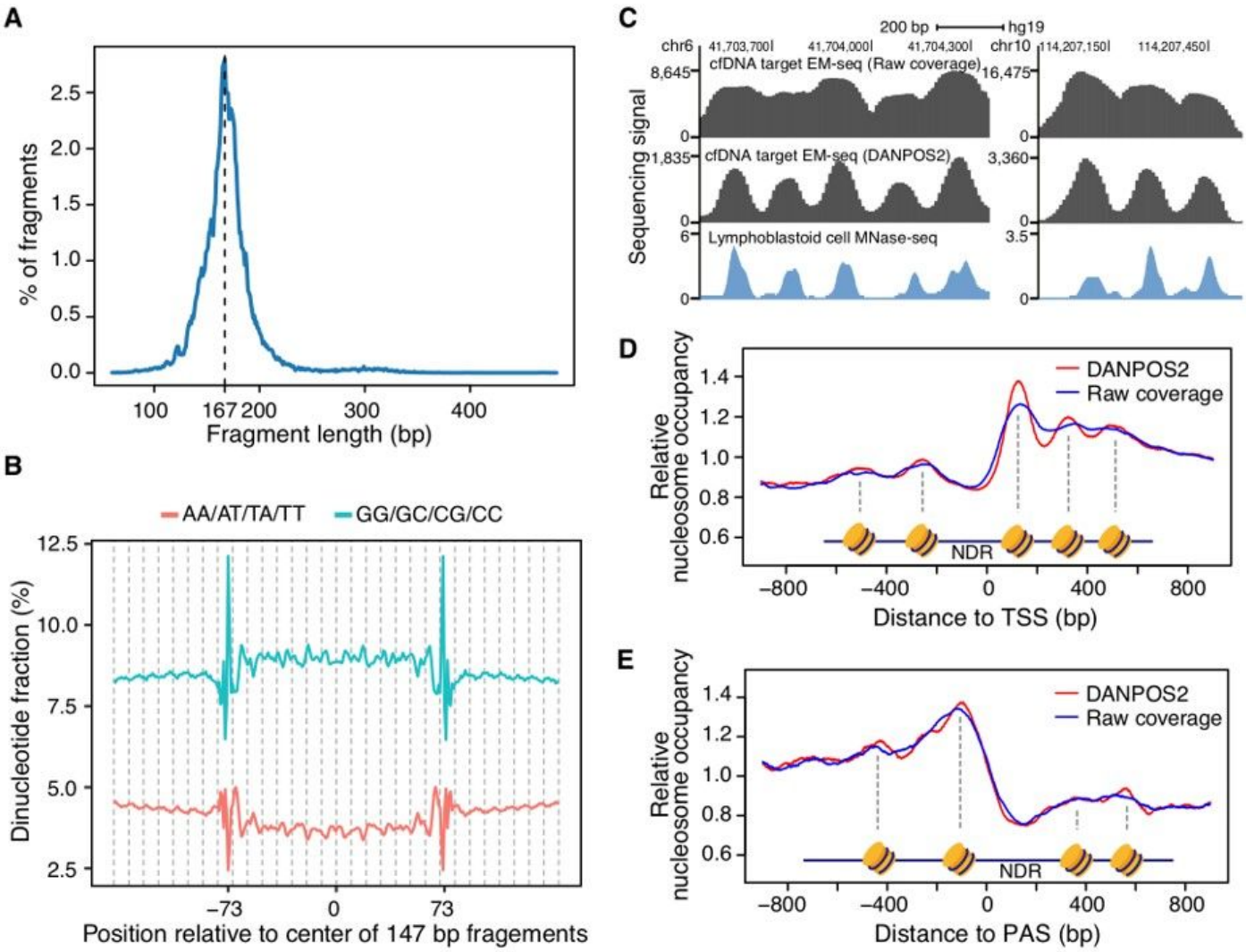

Figure 3

Nucleosome organization information from targeted EM-seq of cfDNA. (A) Fragment length distribution of sequenced cfDNA fragments. A peak value at 167 bp (black dashed line) is consistent with the association with nucleosomes. **(B)** The distribution of dinucleotide fraction across 147 bp fragments and the flanking genomic regions. **(C)** Genome browser tracks showing sequencing signals of targeted EM-seq of healthy cfDNA (cfDNA targeted EM-seq) and nucleosome calls from a published dataset (Lymphoblastoid cell MNase-seq). DANPOS2, occupancy values reported by DANPOS2. Raw coverage, occupancy values estimated by read coverage. Aggregate lines showing nucleosome occupancy profiles across TSSs **(D)** and PASs **(E)** of target genes. Relative nucleosome occupancy represents nucleosome occupancy normalized by the average value of the plotted regions. NDR, nucleosome depleted regions. Results in this figure are based on targeted EM-seq of 60 healthy controls from Cohort 1.

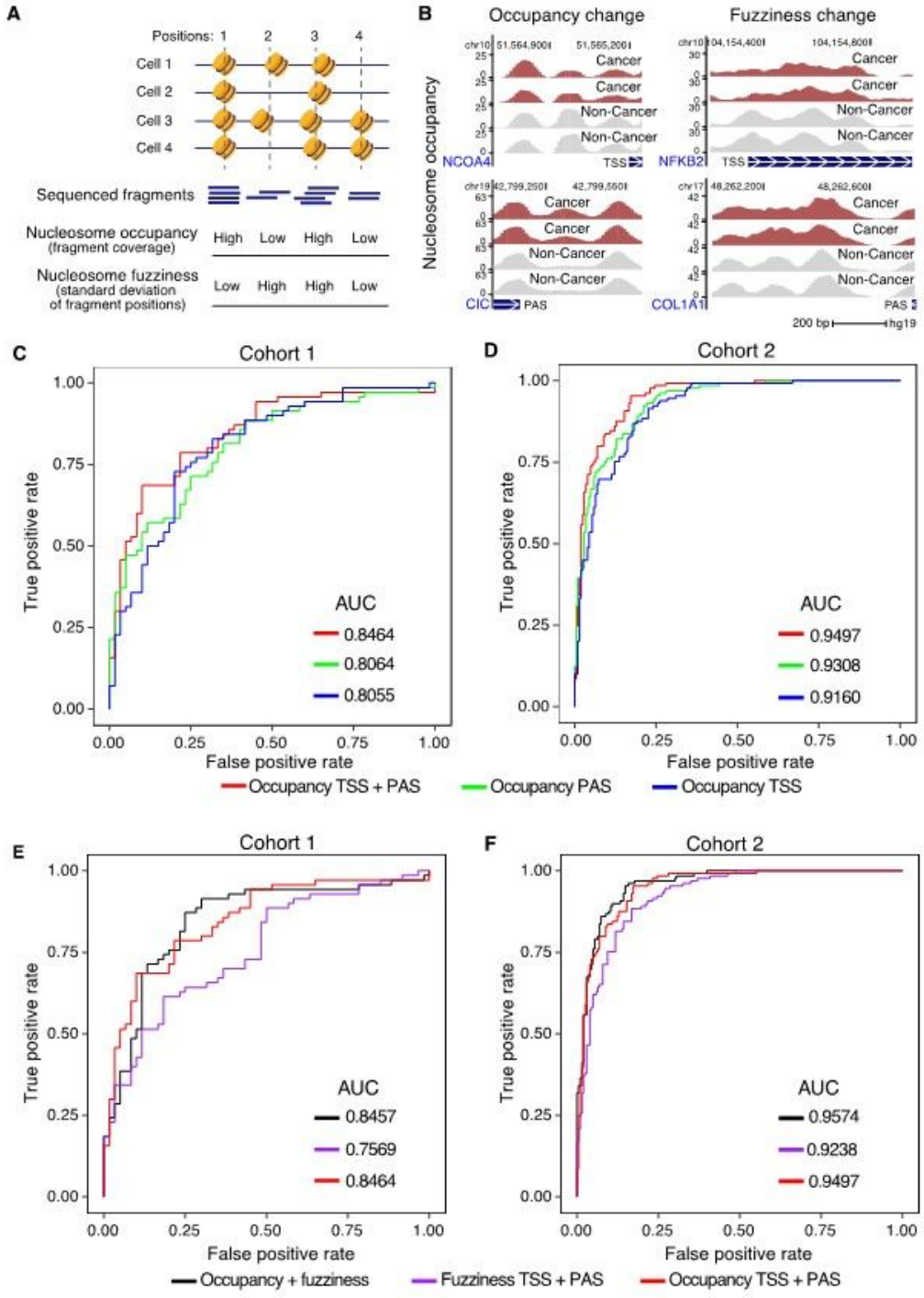

Figure 4

Accurate detection of cancer based on nucleosome occupancy and fuzziness. (A) A schematic diagram showing the differences between nucleosome occupancy and fuzziness for four example positions in four cells. **(B)** Genomic regions showing nucleosome occupancy (left panel) and fuzziness changes (right panel) between cancer and non-cancer samples. The top panel shows genome browser tracks of TSS target regions, and the bottom panel shows PAS target regions. For each panel, two example cancer and non-cancer samples are displayed. The blue boxes show the gene bodies with white arrows indicating the transcription directions. **(C-D)** ROC curves showing the model performances based on the nucleosome occupancy of TSS target regions

(Occupancy TSS), PAS target regions (Occupancy PAS) or combination of the two (Occupancy TSS + PAS). **(E-F)** ROC curves showing the model performances based on nucleosome occupancy (Occupancy TSS + PAS), fuzziness (Fuzziness TSS + PAS) or combination of the two (Occupancy + fuzziness).

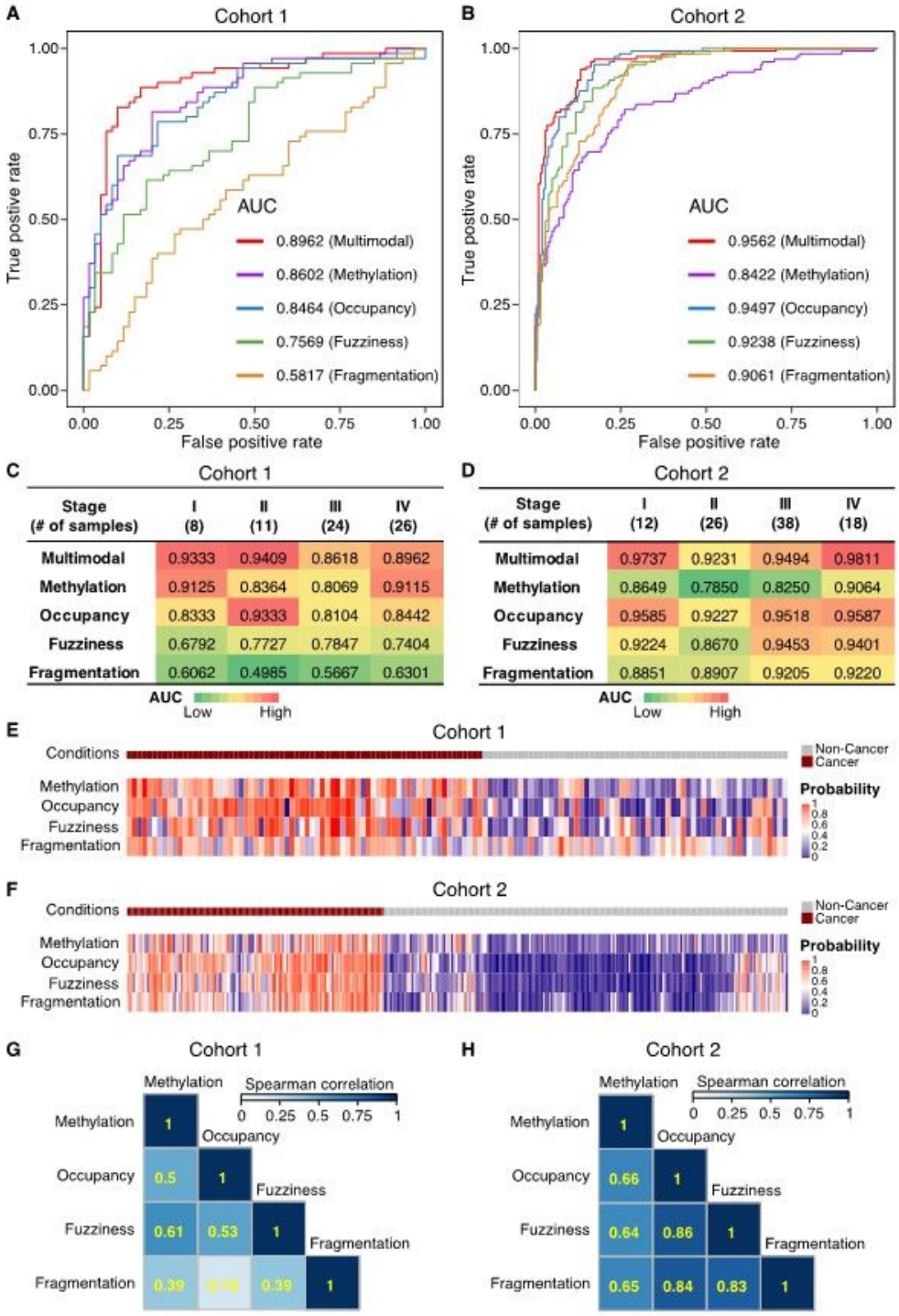

Figure 5

Multimodal epigenetic analysis from MESA improves the performance of cancer detection model. (A-B) ROC curves showing model performances based on different modalities. Methylation, methylation ratio of all target CpG; Occupancy, nucleosome occupancy of all TSS and PAS target regions; Fuzziness, nucleosome fuzziness

of all TSS and PAS target regions; Fragmentation, the ratios of short (80-150 bp) to long fragments (151-200 bp) for all target regions; Multimodal, the combination of all four types of features. **(C-D)** Tables showing the AUC values of ROC curves for different models and cancer patients in different stages. Cancer patients without stage information are removed from this analysis. The values are colored from low to high by a green-yellow-red color scale. The shade of the color represents higher, middle, or lower values. **(E-F)** Heatmaps showing the predicted probabilities of single modality models for each sample. The probability represents the predicted probability of classifying the sample to the cancer group. **(G-H)** Heatmaps showing pairwise Spearman correlations of the predicted probability of all samples between different types of features. The Spearman correlation values are labeled on the heatmaps.

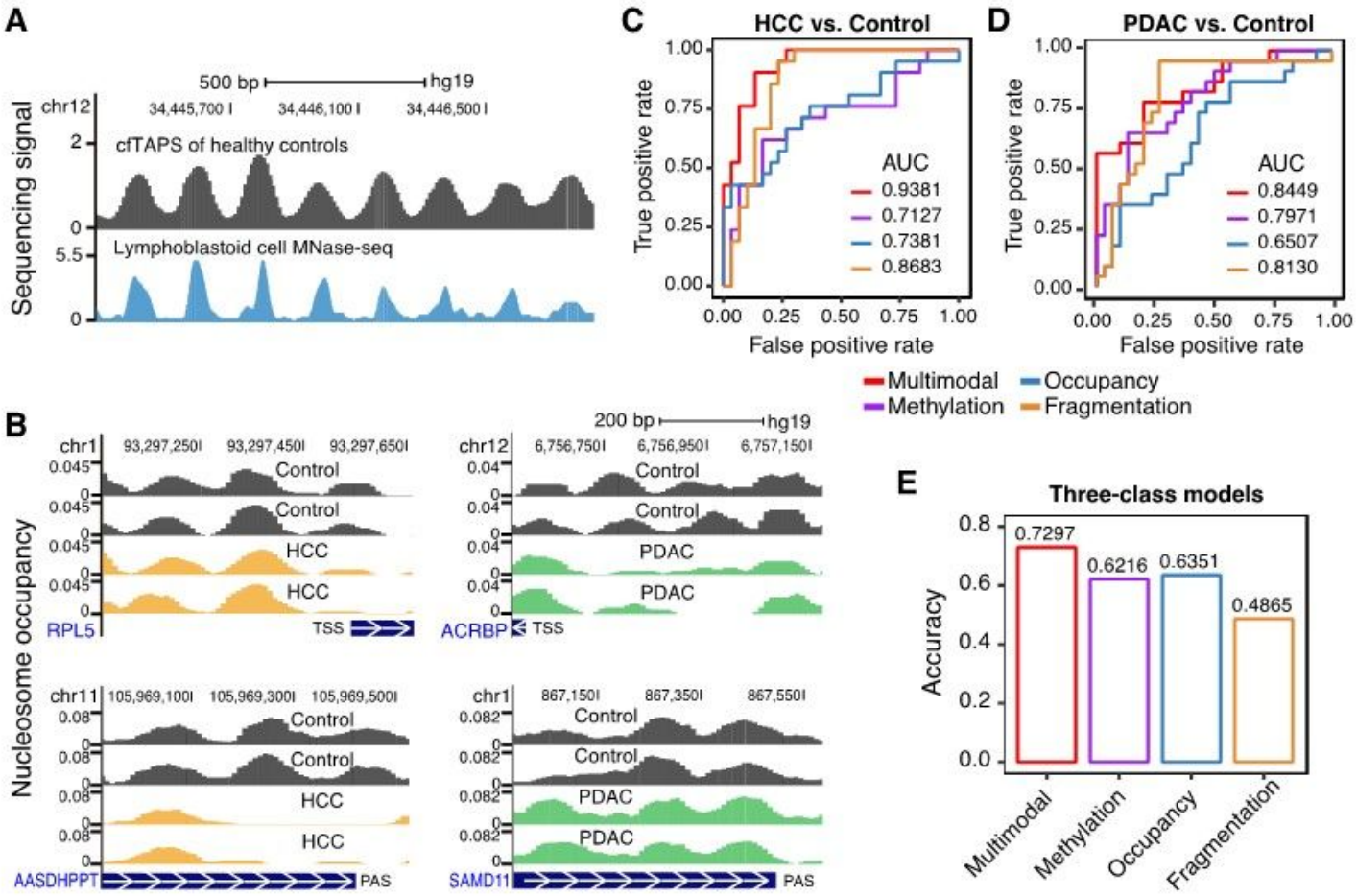

Figure 6

Multimodal epigenetic analysis of cfTAPS improves the performance of cancer detection model. (A) Genome browser tracks showing sequencing signals of cfTAPS of controls (cfTAPS of healthy controls) and nucleosome calls from a published dataset (Lymphoblastoid cell MNase-seq). Sequencing signals from cfTAPS are calculated by DANPOS2. **(B)** Genomic regions showing nucleosome occupancy changes between HCC (left panel) or PDAC (right panel) and control samples. Nucleosome occupancy is calculated by DANPOS2. The top panel shows tracks of regions surrounding TSSs, and bottom panel shows regions surrounding PASs. For each panel, two example cancer and control samples are displayed. The blue boxes

show the gene bodies with white arrows indicating the transcription directions. **(C-D)** ROC curves showing the performances of two-class models which distinguish HCC **(C)** or PDAC **(D)** from control samples. **(E)** Barplot showing the overall accuracy of three-class models which distinguish HCC, PDAC, and control samples. Methylation, methylation ratio of promoter and enhancer regions; Occupancy, nucleosome occupancy of 1 kb regions surrounding TSSs and PASs; Fragmentation, the ratios of fragments (300 to 500 bp) in 10-bp length range bins; Multimodal, the combination of all three types of features.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials.pdf
- SupplementaryTables.xlsx